Dissertation Advisor: Professor Kevin Struhl Joseph Daniel Fleming

Genome-wide integrative analysis of transcription factor occupancy and gene regulation in models of human cancer and cellular differentiation

ABSTRACT

Few transcription factors (TFs) have been studied in the context of an integrative analysis incorporating genomic datasets from diverse genome regulatory mechanisms. Such an analysis allows the testing of specific regulatory associations in an unbiased and comprehensive manner. The promoter binding TF complex NF-Y regulates a diverse set of constitutive, inducible, developmental, oncogenic and tissue-specific genes. Using cancer models, ChIP-Seq, shRNA, and genomics, I have undertaken a genome-wide study of NF-Y. NF-Y binds to not only promoters but also extensively to enhancers, select classes of repetitive elements, inactive chromatin domains and insulators. NF-Y is a "pioneer"-like factor able to access its motif within closed, transcriptionally inactive chromatin domains. NF-Y pervasively associates with FOS, usually in the absence of JUN and the AP-1 motif, and with a group of growth controlling oncogenic TFs. I also show that NF-Y asymmetrically binds to its motif and stereo-aligns with specific TFs and their motifs. My results indicate that NF-Y is not merely a commonly-used, proximal promoter TF, but rather functions at a more diverse set of genomic elements.

The dynamics of TF occupancy, *cis*-regulatory element (CRE) usage and their linkage to gene expression during a differentiation process, from a genome-wide perspective, is poorly understood and is critical to the understanding of fundamental aspects of development and disease. I utilize a model of inflammation-mediated oncogenic transformation, siRNA, ChIP-

Seq, FAIRE-Seq, and microarrays to study the genomic aspects of transformation driven by Src-mediated activation of the inflammatory TF STAT3. I show that CRE usage is static, even in the presence of induced STAT3 activity, and large-scale transcriptional and phenotypic changes. STAT3 induced occupancy is tightly associated with FOS, pre-existing CREs, and does not create CREs *de novo*. I also highlight a putative role of TSC22D3 in inhibiting an epigenetic switch and in STAT3 and AP-1 factors driving the embryonic-like and bone-like phenotypes of breast cancer. The research presented here suggests that phenotypic alterations occurring during disease are not driven by large-scale perturbations of CRE usage.

Overall, this dissertation provides an invaluable resource of genome-scale datasets within cancer models that will assist in future endeavors of scientific discovery.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS AND ALTERNATIVE NAMES

ChIP	Chromatin immuno-precipitation
Chip	DNA microarray chip
Cys	Cysteine
DNase	Deoxyribonuclease
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ENCODE	Encyclopedia of DNA elements
EP300	p300, also known as
FACS	Fluorescence-assisted Cell Sorting
FAIRE	Formaldehyde-assisted isolation of regulatory elements
GO	Gene ontology
HAT	Histone acetyltransferase
HERV	Human endogenous retrovirus
HEPES	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid
HFD	Histone fold domain
KAT2A	GCN5, also known as
KAT2B	p300/PCAF, also known as
KAT3A	p300/CBP, also known as
lncRNA	Long non-coding intergenic RNA
LTR	Long terminal repeat
Met	Methionine
MNase	Mononuclease
NFKB1	NFκB p105/p50, also known as
NF-Y	Nuclear factor Y
PcG	Polycomb group complex
PMSF	Phenylmethylsulfonyl fluoride
Pol	Polymerase
POU2F1	OCT1, also known as
POU2F2	OCT2, also known as
Pro	Proline
PTM	Post translational modification
Q	Glutamine
QPCR	Quantitative polymerase chain reaction
SDS	Sodium dodecyl sulfate
SELEX	Systematic evolution of ligands by exponential enrichment
Seq	Sequencing, massively parallel
Ser	Serine
shSCM	Small hairpin scrambled control RNA
shRNA	Small hairpin RNA
TAD	Transcription activation domain
TAF	TBP-associated factor
TF	Transcription factor
TNFRSF11A	RANK, also known as
TP53	p53, also known as

TP63	p63, also known as
TSS	Transcription start site
TTS	Transcription termination site
Tyr	Tyrosine
UCSC	University of California, Santa Cruz
U-STAT3	Unphosphorylated STAT3

ACKNOWLEDGEMENTS

I want to thank: Steve Buratowski, Alan Cantor, and Danesh Moazed for sitting on my DAC committee and for their scientific advice over the years; X. Shirley Liu, Zhiping Weng and Timur Yusufzai for sitting on my exam committee and providing a stimulating, and thankfully short, conversation; and, Fevzi Ozbek and Josh Forest at WQCG for extensive IT support throughout my project.

Special thanks go out to Carol Baisden for providing a comfortable lab experience, and to Kevin Struhl who never thought me how to do an experiment, but did teach me how to be a scientist, which I am greatly appreciative of, and who allowed me to pursue my dissertation in his lab. I also want to thank those teachers who I've crossed paths with over the years who have helped to nurture my scientific curiosity and put up with my incessant questioning: Dr. Gerard Colleran, Dr. Michael Power and Prof. Peter Hauschka.

To all of the members of the Struhl Lab over the years, I want to express my gratitude for the many suggestions, conversations, proof-reading, scientific tools, and experimental help that you have given me over the years and for making the lab a fun, productive, and thought provoking place to study. Especially Marianne Lindahl-Allen, Nathan Lamarre-Vincent, Koon Ho (Chris) Wong, Heather Hirsch, Yi Jin, Zarmik Moqtaderi, Joe Geisberg, and finally, but not least, Elsa Beyer and Rajani Gudipatti who have given me more help than their short time in the Struhl Lab has warranted.

I want to thank Walter, Ron, and David who have been my close "parental" support group during my time in Boston, and have provided many a fun and relaxing break from the lab and a warm delicious meal almost every week.

Lastly, to my partner Brian, for being so patient, loving and supportive throughout my PhD, and to my Mom and Dad who never put a barrier in my way, and supported me as I followed my interest in science.

To all of you, thank you!

Dedicated to the loving memory of my father, Oliver Fleming, who

never got the chance to see this day

1945 – 2007

CHAPTER 1: Introduction

The NF-Y transcription factor complex

General: NF-Y (Nuclear Factor Y, also known as CBF, CP1) is a highly conserved [1] heterotrimeric transcription factor (TF) complex originally identified in *Homo sapiens* [2] and yeast [3]. It is composed of: NFYA, NFYB and NFYC in *H. sapiens*; Hap2 [4], Hap3 [5], and Hap5 [6] in *Saccharomyces cerevisiae*. In *S. cerevisiae* there is an additional 4th subunit, Hap4 [7], that contains transactivation domains that have been incorporated into other NF-Y subunits in mammals. NFYB and NFYC form a stable heterodimer, via their histone fold domains (HFDs) (discussed below), to which NFYA binds with high specificity. The heterotrimeric complex is then fully capable of binding to its DNA sequence motif, the CCAAT box, which is a common eukaryotic promoter element. All three subunits are required for DNA binding *in vitro* [8] and *in vivo*, though there are limited exceptions (see below), and this has not been tested in an unbiased manner by genome-wide studies. All three subunits make contact with DNA [8], and the affinity for the CCAAT box is extremely high (Kd 10^{-10}-10^{-11} [9, 10]). In general, NF-Y is considered a mild transactivator, on a level similar to other glutamine-rich (Q-rich) domain containing TFs (SP1, POU2F1 (OCT1), POU2F2 (OCT2)), but ~500 fold less than VP16, which contains an acidic activation domain [11].

NF-Y is ubiquitously expressed, found in all human and murine tissues, cell lines and tumors assayed. There are a few exceptions, however. NFYA protein is not detectable by Western blot in myocytes [12, 13], heart muscle [12, 13] or circulating monocytes [14]. This regulation is not at the mRNA level, but rather happens post transcriptionally as *NFYA* mRNA is clearly present. Both myocytes and heart muscle are post-mitotic and terminally differentiated,

and, as such, are not actively dividing. NF-Y has a role in regulating many cell-cycle genes, and, on one level, it is not surprising that NFYA is not required in these non-cycling cells. However, NF-Y regulates many other critical cellular processes such as DNA repair, apoptosis and cholesterol metabolism. It is intriguing to postulate that NFYB and NFYC may bind DNA and regulate transcription in the absence of NFYA.

The importance of NF-Y is underscored by the early embryonic lethality of an NFYA knockout mouse model [15] due to defects in cell proliferation and massive apoptosis, and a *D. melanogaster* knockout also shows early embryonic lethality [16]. In addition, functional inactivation of NF-Y subunits or the use of a dominant negative NFYA mutant, indicates that NF-Y-DNA binding is important for transcriptional activation and the pattern of histone modifications at promoters (reviewed by [17]).

Structure of the NF-Y complex

NFYA: The 3D structure of NFYA is unknown as the crystal structure has not yet been published. However structural modeling and biochemical experimentation has shed light on its structure and the function of specific domains. The N-terminal region of NFYA contains a large Q-rich domain, rich in hydrophobic residues, but lacking in charged residues. NFYC has a similar Q-rich domain at its C-terminus. As assayed by LexA and Gal4 fusion proteins, these Q-rich domains serve as the transcriptional activation domains of NF-Y [11, 18-20]. In addition, NFYA has two small juxtaposed domains, which are highly conserved, in its C-terminus that mediate NFYB-NFYC dimer interaction and DNA binding [21]. While NFYB and NFYC are known to contact DNA, the sequence specific CCAAT recognition domain is contained within the C-terminus of NFYA [22-25].

NFYB and NFYC: NFYB and NFYC were found to be related to the core histones, H2B and H2A, respectively, as they contain conserved HFDs [26, 27]. NFYB is 30% identical to H2B and NFYC is 21% identical to H2A, though similarity is a lot higher [26]. The NF-Y subunits are part of a small disparate family of non-histone HFD containing proteins in humans: *e.g.* NC2α and NC2β (for review see [28]); TAF3/-4/-6/-9/-10/-11/-12/-13 of the TFIID and SAGA complexes (for review see [29, 30]); and CHRAC16/-14, YEATS2 and POLE3 of the DNA Polymerase ε, ATAC and CHRAC complexes [31-33]. The sequence identity within the HFDs of histones (14-18%) [34] is comparable to that between the HFDs of NF-Y and histones (~15%) [26, 35] and key residues within it are well conserved. In general, DNA and protein sequence similarity is low, but secondary and tertiary structural similarity is exceptionally well conserved. In histones, the HFDs are responsible for both octameric protein–protein complex formation and non-sequence-specific protein-DNA interactions. They have a similar function within the NF-Y complex, as elucidated by many biochemical and mutagenic experiments over the years. The HFDs mediate both dimerization of NFYB and NFYC, and non-sequence-specific protein-DNA interactions [22-25]. The crystal structure of NFYB-NFYC has been solved, and confirmed the presence and role of HFDs mediating heterodimer formation, and, via modeling, the HFD-DNA interaction [25].

The CCAAT box

NF-Y binds to the core pentanucleotide sequence CCAAT, commonly referred to as the CCAAT box. An estimated 30-60% of human proximal promoters contain CCAAT boxes [36-39], which is similar to the frequency of the TATA element (~35-70%), but less than the GC box (~95%), which is ubiquitous in mammalian promoters [36, 40]. The CCAAT box is found both in TATA-containing and TATA-less promoters. As found by *in silico* studies [36-40], and confirmed by

limited genomic ChIP experiments [41, 42], the CCAAT box is highly positioned approximately 80 bp upstream of the transcriptional start site (TSS), with the motif in either orientation, which is similar to the TATA element which observes a closer distribution (approximately -35 bp) to the TSS [40]. This biased distribution suggests that the location of NF-Y in relation to the TSS is important for the function of NF-Y, though this has not been specifically tested [43]. At promoters, specific distance and orientation requirements [43-45] between cooperating TFs [44-49], adjacent NF-Y binding sites [50], and the TSS is required for optimal transcriptional activation by NF-Y. In essentially all promoters tested, mutation of the CCAAT box reduces or eliminates both constitutive and/or inducible transcriptional activity [51]. In this regard, NF-Y can be thought of as having an "architectural" role in positioning protein factors in the correct location at promoters in respect to the transcriptional machinery, though this has only been observed on single promoter studies with a limited number of motifs and TFs. It is not known if this architectural function of NF-Y translates genome-wide and if it occurs outside of core promoters within other genomic contexts.

Though mutation of any of the core pentanucleotides greatly reduces NF-Y binding to the CCAAT box and associated transcription [52-54], the specific flanking sequences are also important for NF-Y binding. Many *in silico* [36-40], *in vivo* foot-printing [10, 54], SELEX [10], and ChIP-chip [41, 42, 55] studies defining the CCAAT box have repeatedly found specific flanking sequences to be preferred, and mutation of these sequences affect NF-Y occupancy and transcription from the associated promoter [10, 43, 54]. Purines are favored at positions 3 and 4, pyrimidines predominate at 2, C/G at 10 and 12, and purines at 11 (Figure 1). In limited instances, NF-Y can associate with chromatin in the absence of a recognizable CCAAT box: mainly recruited by hormone receptors: estrogen receptor (only NFYA was present) [56]; and

mineralocorticoid receptor (only NFYC was present) [57]. The lack of unbiased genome-wide maps of NF-Y subunit binding sites hinders the study of the biology of NF-Y binding in the absence of the CCAAT box.

How does NF-Y regulate transcription?

Given that CCAAT boxes are found in 30-60% of human proximal promoters, including inducible, constitutive and cell-type specific promoters, it is not surprising that NF-Y has been documented interacting with a plethora of transcriptional regulators. Sequence specific DNA binding TFs, co-activators and co-repressors, and, given its close proximity to the TSS, many general RNA Pol II factors interact with NF-Y. Indeed, a recent review [17] listed 42 transcriptional regulators that interact with at least one NF-Y subunit, and this count doesn't include kinases (CDK2 [58]), splicing factors (*e.g.* SF1 and YBX1 [59]), structural molecules (ACTIN4 [60]) and polymerases (PAPOLG [59]). It is not known how extensive the association of NF-Y is with any one particular TF, let alone 10s of TFs and the complexes they form at NF-Y bound sites. One exception is that of SP1 which has been documented in a promoter ChIP-chip study closely associating with NF-Y [61].

I will use the example of the MHC class II gene promoters, which have been extensively studied, to highlight a mode of transactivation by NF-Y, *i.e.* TF cooperativity mediated by conserved spacing between DNA motifs for the activation of transcription.

Motif and TF cooperativity: NF-Y was originally discovered as a factor binding to the Y box motif (a CCAAT box), one of multiple conserved motifs common to all MHC class II gene promoters [2, 54] and controlling their expression. A second of these motifs, which partners intimately with the CCAAT box, is the X box. NF-Y binding to the CCAAT box cooperates synergistically with the TF RFX binding to the X box [46, 62, 63], to recruit the co-activator

CIITA [46, 64]. Mutation of the CCAAT box inhibits interaction of the X box with RFX, but the opposite is not true, suggesting that NF-Y is a prerequisite for RFX recruitment. There is a well conserved distance preference (19-20 bp) between the X and CCAAT boxes across multiple MHC class II promoters. Altering the distance between the motifs by half helical turns (~5 bp), but not by full helical turns, severely disrupts gene expression [44, 45, 47, 65]. Both NF-Y and RFX are required for the recruitment of CIITA, which in turn is required for the recruitment of the histone acetyltransferases (HATs) KAT3A (p300/CBP) and KAT2B (p300/PCAF) [66, 67]. BRG1, of the SWI/SNF chromatin remodeling complex, is also recruited by CIITA and is required for CIITA mediated expression of MHC class II genes. This was shown by the failure of exogenous CIITA to induce gene expression in cells lacking BRG1 (SMARCA4) [68, 69]. NF-Y/RFX binding also facilitates the direct interaction of CIITA with the basal transcription factors TAF9, TAF6 and TFIIB [70, 71]. These orchestrated events ultimately lead to the induction of MHC class II gene expression in a time and cell-type specific manner that is absolutely dependent on NF-Y and the CCAAT box.

Direct interaction with the basal transcriptional machinery: NF-Y and RNA Pol II basal factor interactions are important for promoter activation. CCAAT boxes and the TBP binding TATA and Initiator (Inr) motifs are common promoter elements, with biased and transcriptionally important spacing requirements, both with respect to the TSS and between each other. Thus, at most CCAAT box containing promoters, NF-Y, TBP and TBP-associating factors (TAFs) are located in close physical proximity. The experimental dissection of the MHC class II *Eα* promoter in mice showed that the CCAAT box was required for the correct use of the +1 TSS [72] and NF-Y binding increases the affinity of holo-TFIID to the promoter [73]. *In vitro*, studies have shown that NFYA is required for pre-initiation complex formation at the *Eα* promoter, but

not once it is formed, nor for the re-initiation of transcription [74]. Additionally, NF-Y binding to the CCAAT box in the *γ-globin* promoter is required for TBP-TFIIB recruitment *in vivo* [75]. The HFDs of NF-Y allow it to interact with the HFD containing subunits of the basal transcriptional machinery. Many of the TAFs contain HFDs [76, 77] which allow them to interact with NF-Y (TAF4/-11/-12/-13) [73] and other HFD containing TFs (*e.g.* NC2). NFYB and NFYC, but not NFYA, co-immuno-precipitate with TBP and TAF5 [35] in solution and the NFYB-NFYC-TBP interaction domains have been identified [35] and are the same as those that interact with NC2. The Q-rich transactivation domain of NFYA has also been shown to interact with TAF5 *in vitro* [78].

Interaction with co-activators: NF-Y can also associate with co-activators (KAT2A (GCN5), KAT2B (P/CAF), EP300 (p300), SUB1 (PC4)) and co-repressors (HDAC1, PcG complex) which are functionally important for transcription from CCAAT box containing promoters. The histone acetylase complex KAT2B physically interacts with NF-Y *in vitro* and CCAAT box mutations and dominant negative NFYA constructs prevent induction of the *MDR1* gene promoter by KAT2B overexpression [79]. NF-Y was also found in a complex composed of SP1/EP300/KAT2B/HDAC1, that induces the transcriptional activity of the *TGFBR2* promoter upon trichostatin A treatment and is modulated by a mechanism where KAT2B (a histone acetylase that activates) or HDAC1 (a histone deacetylase than inhibits) predominate in the complex [80]. The HFDs of NFYB/NFYC dimer can stably associate with KAT2A *in vitro* and *in vivo* and overexpression of KAT2A potentiates NF-Y activation of the *collagen, type I, α2* (*COL1A2*) promoter [81].

Interactions with nucleosomes: NF-Y, either the HFD dimer or the trimer, have been shown to directly interact with chromatin *in vitro* and/or *in vivo* and that this function is important for

transcription. The NF-Y trimer can successfully bind to a CCAAT box containing promoter both during and after reconstitution of nucleosomal DNA using purified histones [82]. NF-Y forms higher-ordered NF-Y-nucleosome-DNA structures *in vitro*, in a CCAAT box dependent manner, importantly, with preformed nucleosomes and even in the presence of free naked non-nucleosomal CCAAT box containing DNA fragments [82]. In a similar study using a more purified system of recombinant histones, the NFYB/NFYC HFD containing dimer, associates with H3-H4 tetramers *in vitro* both in the presence of DNA and in solutions free of DNA [83]. The same study also tracked down the H3-H4 interaction region to the HFD of NFYB. However, nucleosomal, octameric-like structures on DNA were not formed with clear differences in the DNase I, MNase and exonuclease III digestion patterns. The NF-Y-nucleosome interaction affects choice of TSS used in an *in vitro* transcription system [84], and in general, lack of NF-Y at the core promoter of CCAAT box regulated genes is associated with a closed nucleosomal structure *in vivo* [85, 86]. How NF-Y-CCAAT-nucleosome interactions behave outside of core promoters *in vivo* is incompletely understood. The association between CCAAT boxes, their occupancy by NF-Y, the degree of nucleosome depletion and the histone modifications present in the immediate vicinity has never been studied *in vivo* or *in vitro*.

NF-Y and repression: NF-Y can also repress transcription, and the mechanisms are quite varied. A previous partial genomic study utilizing ChIP-chip, found NF-Y bound to the promoters of genes with the repressive H3K27me3 and H4K20me3 histone post translational modifications (PTMs) [42] which confirmed previous studies on single promoters showing a role of NF-Y in repression [87-92]. A report with *Caenorhabditis elegans* showed a role of NF-Y in maintaining the repression of the Hox gene *egl-5* during development [93]. This repression was dependent on the CCAAT box in the *egl-5* promoter and NF-Y directly interacted with the MES-

2/MES-6 PcG repression complex, therefore implicating a direct role for NF-Y in repression. Another mechanism of NF-Y repression involves a multi-protein complex composed of NF-Y/HMGA1/KLF9/SIN3A that forms on, deacetylates, and inhibits the *GHR* promoter [90]. In response to DNA damage, G(2)/M promoters are repressed by direct association of acetylated TP53 (p53) with NF-Y and the CCAAT box, resulting in the recruitment of HDAC1/4/5 and transcriptional repression [94]. A complete map of NF-Y binding sites in the human genome, the associated histone PTMs, and RNA expression level, would greatly increase our understanding of NF-Y's repressive function.

Probably the most interesting aspects of NF-Y repression are those that involve CCAAT-less promoters. As mentioned earlier, there are reports of hormone receptor recruitment of NF-Y to CCAAT-less promoters and the induction of transcriptional repression. It was shown that the mineralocorticoid receptor recruits NFYC, but not NFYA or NFYB, to the *ENaC* promoter, as a cofactor, which prevents the N- and C-termini of mineralocorticoid receptor from interacting upon hormone binding which prevents activation of transcription [57]. ChIP-Seq for NFYC was not undertaken in this dissertation due to the poor performance of the antibody. Another study [56] showed that NFYA mutants, devoid of DNA binding or trimer formation ability, can inhibit the ERα mediated transcriptional induction of the *F12* (*FXII)* and *VIT* promoters, via a mechanism that involves interaction of the NFYA C-terminus directly with ERα and not with the CCAAT box. Neither promoter contains a canonical CCAAT motif and NFYB may be involved. These methods of repression are probably limited to hormone receptors and the genes they target, however they show an interesting, and largely unexplored, aspect of NF-Y biology, that of the individual subunits functioning outside of the trimer complex.

It is clear that NF-Y, sitting at -80 bp upstream of the TSS in between upstream transactivating motifs and the TATA/Inr elements, efficiently penetrates nucleosomal structures, via its HFDs. From this position NF-Y recruits and cooperates with upstream transactivators, to recruit TFIID and the pre-initiation complex to CCAAT containing core promoters to initiate transcription. The associated HATs serve to modulate NF-Y transactivation potential by aiding disruption of local chromatin structure thereby enhancing transcription.

Regulation of NF-Y

Splicing: NFYA and NFYC are known to express multiple isoforms, while NFYB is not known to be alternatively spliced in *H. sapiens*, which has been confirmed by GENCODE (the ENCODE related group annotating human genes) [95]. NFYA has at least two confirmed isoforms (GENCODE, [18, 95, 96]), NFYAs (short) and NFYAl (long), which differ in only 28 amino acids within the Q-rich transcriptional activation domain. There are known instances in the literature where these isoforms switch during differentiation [97] or show cell type specific biases [18, 98]. Indeed, these isoforms have different functions as shown by the ability of a specific isoform to drive a specific phenotype and co-operate in transcription with a specific partnering TF [98, 99].

NFYC is much more complex, with the human genome encoding 13 splice isoforms (GENCODE, [95]). By northern blots, 4 isoforms have been observed in *Rattus norvegicus* tissues [100], and 2 in *H. sapiens* tissues [101]. The functions of some of these isoforms are starting to be described [102]. In keeping with the findings from NFYA, two *H. sapiens* NFYC splice isoforms differ in the Q-rich transactivation domain and have cell type specific biases. There functions are unknown, however their RNA levels do differ in their response to DNA damage [102]. A third splice variant, which lacks the HFD and therefore cannot interact with

NFYB, has a specific function in acting as a negative regulator of TGFβ signaling by interacting with SMAD2/-3 [103].

Expression and post translational modification of NF-Y: NFYA protein levels are known to fluctuate during the cell-cycle [104] and in certain cells during differentiation [12-14], while the protein levels of NFYB and NFYC (and the mRNA of all three subunits) remain constant. This fluctuation in NFYA protein levels modulates the complex's transactivation function, making NFYA the regulatory subunit. A post-transcriptional process involving NFYA acetylation [105] by EP300 (p300) [81] on conserved lysine residues (K283, K289) located in the trimerization and DNA binding domains, has been shown to increase NFYA protein stability by preventing the poly-ubiquitination of overlapping lysine residues (K283, K289, K292, K296) [106]. This acetylation-ubiquitination dynamic regulates proteasomal degradation and accumulation of NFYA protein in the cell. In *Xenopus*, NFYB can also be acetylated, by EP300, though the residues are unknown and the function unclear [107].

A second common modification of TFs is phosphorylation. NFYA contains two CDK2 phosphorylation sites in its C-terminus (S292 and S298) [58, 108, 109], near the trimerization and DNA binding domains, that are phosphorylated in a cell-cycle dependent manner. CDK2 interacts with and phosphorylates NFYA *in vitro* and *in vivo*. Phosphorylation does not impair heterotrimer formation [58] but does prevent NF-Y-DNA interaction [58]. This is functionally important as CDK2-dependent phosphorylation of NFYA is essential for expression of cell-cycle genes (*e.g. CDC2*, *CDK2*) and cell-cycle progression [108]. NFYB and NFYC are not known to be phosphorylated.

Cellular redox: A common process of reversible regulation of proteins is the post-translational reduction of cysteine (Cys) thiol groups (-SH) to moieties such as disulphide (-S–

S-) and S-nitrosothiol (SNO), that can alter protein structure and effect signaling, respectively (for review see [62]). The activity of many TFs are known to be modulated by cellular redox state at Cys residues (*e.g.* FOS [110], JUN [110], NFκB [111], and MYB [112]), one of the best studied being the *S. cerevisiae* Yap1 protein [113, 114]. Pertinently, all known NFYB orthologues have three conserved Cys residues in the HFD that are not present in histone HFDs, and histones are not known to be regulated by cellular redox. Nakshatri *et al.*1996 [115] showed by mutagenesis studies and by the alteration of redox potential, that reduction of two of these Cys residues modulates NFYB covalent-multimerization and NF-Y DNA binding ability. Further work by Thon *et al.* 2010 [116] on the NF-Y orthologue AnCF of *Aspergillus nidulans* confirmed these findings. Given the highly conserved nature of this phenomenon, redox regulation of NF-Y is likely a general mechanism.

Cellular localization: Many TFs are actively transported into the nucleus upon an activating stimulus (usually a kinase cascade), however, NF-Y nuclear import seems to be constitutive and not a major method of regulation. There are reports of TGFβ signaling, through SMAD interactions and TSC2 dependent [117-119] regulation of NFYA nuclear import. This latter mechanism could be important for tumorigenesis as *TSC2* (also known as tuberin) is a tumor suppressor gene. The proteins responsible for nuclear import have been identified: importin β for NFYA; and importin 13 for NFYB and NFYC, which are imported together [120-124].

NF-Y and disease

There are no known mutations within the NF-Y subunits that manifest in disease. This is likely due to the absolute requirement for NF-Y for cellular growth and development, as seen by the early embryonic lethality of NFYA deficient mice. However, indirect perturbations of NF-Y function and *in silico* findings have associated NF-Y with disease.

Impressively, a number of studies have found that the CCAAT box is enriched, in concert with the E2F motif, in cancer signature genes [125-132]. Three methodologically independent studies of gene expression sets from 1,000s of diverse *H. sapiens* tumor samples found the CCAAT box and E2F motifs, either separately or in combination, to be commonly enriched across many tumor types in the promoters of genes misregulated in cancer [133-135]. This is not surprising, as NF-Y is required for cellular proliferation and transcriptionally controls, along with E2Fs, many cell-cycle genes, at some of which NF-Y is required for E2F binding [136]. Tabach *et al.* 2005 [125] found a promoter module of p21/NF-Y/E2F motifs in the transcriptional response of transformation induced genes by the inactivation of TP53 (p53) and p16(INK4A) tumor suppressors. Many of the genes were cell-cycle genes. In a meta-analysis of 8 breast cancer metastasis gene expression datasets, Thomassen *et al.* 2008 [137] identified E2F/NF-Y/YY1 as the TFs involved in metastasis, with cell-cycle and metabolism related genes being significantly enriched in metastasizing tumors. NF-Y is known to be involved in regulating metabolic biosynthetic processes [61].

NF-Y has also been linked to polyglutamine-based neurological disorders, Leigh syndrome, schizophrenia and diabetes; however, these will not be discussed here as they are not relevant to this dissertation.

The biology of STAT3

The signal transducer and activator of transcription 3 (STAT3) was originally identified as a factor that regulated the acute phase response genes, an inflammatory process of transcriptional induction upon treatment with the inflammatory cytokine IL6 [138]. As such, STAT3 has been intimately linked to inflammation from its discovery, yet the direct transcriptional targets and the

genomic sites of occupation of STAT3, especially during inflammation-mediated oncogenic transformation, are poorly characterized.

Structure and function of STAT3

STAT3 functions as a latent monomeric cytoplasmic TF whose activation is tightly regulated by tyrosine^{705} (Tyr) phosphorylation and its subsequent homo- or hetero-dimerization and nuclear localization. STAT3 contains a SRC homology 2 (SH2) domain that recognizes phosphotyrosine residues on other molecules (*e.g.* STAT3, STAT1, EGF [139]) and mediates reciprocal SH2-phosphotyrosine dimerization. A second function of the SH2 domain is to serve as a specificity domain as the peptide sequence adjacent to the phosphotyrosine residue is recognized and affects affinity [140]. STAT3 contains a DNA binding domain based on the immunoglobulin fold and is structurally similar to that of NFκB and TP53 [141]. STAT3 phosphorylation and dimerization is obligatory for DNA binding to the consensus motif TTCCNGGAA. STAT dimers can themselves dimerize, and these dimer-dimer interactions allow STATs to strongly interact with adjacent low affinity motifs that would poorly mediate occupancy of a single STAT dimer [142-145].

STAT3 interacts with TFs and co-factors to mediate transcriptional activation. The C-terminal of STAT3 contains the transactivation domain (TAD) and a small peptide motif within it (Pro-Met-Ser-Pro) is highly conserved and its phosphorylation (serine^{727}) is required for maximal transactivation by STAT3. Almost all STATs interact with and recruit to promoters the histone HAT EP300 via their TAD [146]. HATs acetylate conserved residues in histones to create negatively charged residues that form a repulsive force which opens the chromatin structure to facilitate transcription. STAT3 itself is reversibly acetylated on lysine^{685} by KAT3A [147] which is required for DNA binding and transcriptional activation of STAT3 target genes

[148]. Not all STAT3 interactions are mediated by the C-terminal TAD. The STAT3-JUN interaction have been well characterized, mapped to the N-terminal regions of STAT3, and the complex specifically targets a subset of STAT3 regulated genes [149, 150]. Negative interactions have also been documented. SIN3A [151] and HDAC1 [152] have been shown to interact with and deacetylate STAT3 to promote its nuclear exclusion.

New mechanisms of STAT3 gene regulation are emerging, particularly involving repression and a body of research on unphosphorylated STAT3 (U-STAT3). A recent study has shown that acetylated STAT3 plays a role in methylating and repressing the promoters of tumor suppressors by interaction with DNA methyltransferase-1 [153]. Additionally, *Drosophila melanogaster* STAT has been implicated in HP1 localization and heterochromatic gene silencing [154, 155]. Traditionally, all of the regulatory potential of STAT3 has been thought to be due to phosphorylated STAT3, however, there is now a large body of evidence showing that U-STAT3 can bind DNA [156], regulate genes distinct from phosphorylated STAT3 (44), and interact with NFκB [157, 158] to activate [157] or inhibit [159] transcription [157, 160]. These new aspects of STAT3 biology complicate the interpretation of results involving STAT3, as it can't be thought of as just an inducible TF involved in gene activation. To this regards, the ChIP-Seq derived STAT3 genomic locations produced by dissertation and the integration with ChIP-Seq datasets of other TFs, can be used to explore these aspects of STAT3 biology.

STAT3, inflammation and cancer

There are many cytokines and external stimuli that can signal to and activate STAT3. A recent review [161] has catalogued 19 cytokines (*e.g.* IL6, IFNγ, and TNFα), 5 growth factors (*e.g.* EGF, CSF2 (GMCSF) and PDGF) and 16 miscellaneous external stimuli/chemicals (*e.g.* UVB, tobacco and diesel exhaust particles). STAT3 is phosphorylated by receptor tyrosine kinases (*e.g.*

PDGF, EGF, CSF1R) and by members of the JAK family of kinases (JAK1, JAK2, TYK2) that are resident on cell surface receptors. Upon stimulating cells with a cytokine, STATs are phosphorylated within 10mins. The oncoproteins Src [162] and Ras ([163]) have also been shown to phosphorylate and/or activate STAT3. STAT3 activity is modulated by Ser^{727} phosphorylation which enhances transactivation [164, 165] and by EP300 mediated acetylation which stabilizes dimer formation [148]. Ultimately, STAT3 serves to induce transcription, and the transactivation domain located within the C-terminus is essential for this function, as when it is deleted, STAT3 cannot activate transcription [166] and acts as a dominant negative.

Inflammation and the production of an inflammatory milieu, composed of cytokines, chemokines, reactive oxygen species and growth factors, is commonly associated with many different types of cancer [167-169]. While acute inflammation is important to critical bodily functions such as pathogen defense, wound healing and tissue repair, chronic inflammation, which has no known normal physiological role, has now been linked to various disease states including cancer, rheumatoid arthritis, atherosclerosis, multiple sclerosis, asthma, and Alzheimer's disease [170-172]. Many pro-inflammatory cytokines and chemokines are released by cancer cells (TNFα, IL1β, IL6, IL8, IL17, CSF2 (GMCSF)) and act as potent proliferative and survival factors. Their receptors are prevalent on cancer cells, for example the chemokine receptors CXCR4 and CCR7 are highly expressed in breast cancer cells and mediate the invasion phenotype [173]. The inhibition of CXCL12 and CXCR4 receptor dependent inflammatory signaling reduces pulmonary metastases in a murine model of breast cancer [174]. STAT3 is an important inflammatory TF, and along with NFκB, is a major effector of the inflammatory signaling pathways. STAT3 has been found to be a central mediator of the transcriptional changes in many different types of cancers, including breast cancer [175], pancreatic cancer

[176, 177], prostate cancer [178], liver cancer [179], melanoma [180], among others (for a review see [181, 182]). In this regard STAT3 and NFκB link inflammation to carcinogenesis and their biology within cancer cells is an important avenue of research for potential therapeutic intervention.

Rational for dissertation project

While this thesis does not explicitly study aspects of the biological sciences of dental medicine, it does, however, take a broader view of transcription regulation and how it pertains to cellular regulation. Cell differentiation and organismal development are all mediated by cell-type specific TFs interacting with DNA motifs, the transcriptional machinery and, ultimately, the regulation of gene expression. As such, understanding the biology of TFs is critical to the study of developmental biology and the pathology of disease.

As we pass from the single gene/transcript/promoter/*cis*-regulatory element view of the molecular biology of transcription regulation that was imposed on the scientific community by the available biological techniques, we are now entering the era of transcription regulation from the viewpoint of whole genome analysis. This is largely driven by rapidly advancing technology in the area of massively parallel DNA sequencing and its application to traditional molecular biology techniques. Because of this shift from a narrow and biased view of the molecular biology of cell regulation, we can now ask very broad, genome-scale, highly integrative, "big picture" questions about fundamental aspects of cellular regulation. This dissertation and the many genome scale views of transcription regulation just now being published, especially with the help of the ENCODE Project Consortium [183-185] (See Appendices 1 and 2), are re-asking old questions but with newly developed biological techniques that can interpret data from the entire 3,137,161,264 bp (at last count from the hg19 version) of the genome of *H. sapiens*. In addition,

new questions, that were not approachable even 5 years ago, can now be asked and answered due to the massive increase of genome scale datasets for chromatin bound transcriptional regulators, histone modifications, histone occupancy, *cis*-regulatory elements, gene expression, isoform transcripts, protein-protein interactions, and protein modifications within a single cell-type, but also across cell states and even species. This thesis, as a small part of the ENCODE project, has undertaken a genomic study of the NF-Y transcription factor complex and the transcriptional regulation of a cellular differentiation process, inflammation-mediated oncogenic transformation.

To date there is a lack of data on the genomic profile of NF-Y binding in the *H. sapiens* genome and how this profile changes with cell type. Even more so, there is a fundamental lack of unbiased knowledge regarding the TFs and histone modifications present at NF-Y locations. In addition, NF-Y-CCAAT box-chromatin interactions are largely unexplored *in vivo*. Many studies over the last two decades have provided biased views of individual loci, and a handful of studies have tried to provide a broader perspective. However, all have been limited to promoter-like genomic elements, non-repetitive regions, and/or < 2% of the *H. sapiens* genome with limited integration of other datasets. None have explored TF-NF-Y and/or chromatin state-NF-Y interactions in an unbiased genomic study. This dissertation aims to address these questions and has confirmed many known aspects of NF-Y biology. An unbiased genome-scale study can contribute greatly to the scientific knowledge pertaining to NF-Y and holds the promise of defining new aspects of NF-Y biology and gene regulation in general.

While the bulk of this dissertation centered on NF-Y, a second, related area, explored here is the genome-scale view of transcription regulation dynamics during a cellular differentiation process. Here I use, with the help of many friends both past and present, an inflammation-mediated oncogenic transformation model of an immortalized breast epithelial cell

line to explore aspects of gene regulation during phenotypic changes. This cell line is discussed further in the Introduction to Chapter 3 and I will not repeat that discussion here. STAT3 is absolutely require for transformation in this model system and the dynamics of STAT3 mediated transcriptional regulation have not been explored and would provide invaluable insight into the inflammatory transformation pathways regulated by STAT3. Whether the *cis*-regulatory element usage of a cell undergoing oncogenic transformation is dynamic or stable has also never been explored. It is a fundamental question of cancer biology, and is related to similar events during development and disease progression, and is undertaken by this dissertation.

CHAPTER 2: NF-Y co-associates with FOS at promoters, enhancers, repetitive elements and inactive chromatin regions, and is stereo-aligned with growth-controlling transcription factors

ABSTRACT

NF-Y is a trimeric transcription factor (TF) composed of two histone-like subunits (NFYB and NFYC) and a sequence-specific subunit (NFYA). NF-Y binds to the CCAAT box, a common promoter element. We have identified the location of NFYA and NFYB across the *H. sapiens* genome in three cell types and annotated the sites with respect to chromatin states, 78 chromatin associating factors, *cis*-regulatory elements, DNA sequence motifs, genic features, RNA, and gene ontologies. Approximately 25% of NF-Y sites are in promoters and an equally large proportion are in enhancers, which tend to be tissue specific, and NFYA and NFYB bind asymmetrically with respect to the CCAAT box. Surprisingly, a large portion of NF-Y sites are in select subclasses of HERV LTR repeats that appear to be transcriptionally inactive. Unexpectedly, NF-Y extensively co-localizes with FOS in all genomic contexts, and at promoters and enhancers this often occurs in the absence of JUN and the AP-1 DNA motif. Unlike most TFs, NF-Y can access the CCAAT box within "non-modified" inactive chromatin domains and H3K27me3^{+} repressed domains. NF-Y was associated with a select cluster of growth-controlling, potentially oncogenic TFs, which helps explain the abundance of CCAAT boxes in the promoters of genes overexpressed in cancer. Our results indicate that NF-Y is not merely a commonly-used, proximal promoter TF, but rather performs a more diverse set of biological functions, many of which are likely to involve co-association with FOS.

AUTHOR CONTRIBUTIONS

Joseph D. Fleming, Giulio Pavesi, Paolo Benatti, Carol Imbriano, Roberto Mantovani and Kevin Struhl.

J.F. conceived the project, designed experiments, performed biological experiments and bioinformatically analyzed all data, except for: G.P. analyzed the motif and transcription factor stereo-chemical positioning data; P.B. and C.I. performed shRNA biological experiments and J.F. generated RNA for hybridization. J.F. wrote the chapter with substantial input from R.M. and K.S.

INTRODUCTION

Transcriptional regulatory proteins and the RNA polymerase (Pol) II machinery recruit chromatin-modifying activities to their target loci, thereby determining the genomic pattern of histone modifications and nucleosome occupancy [186]. Activator proteins, functioning combinatorially at distal enhancers and in proximity to core promoters, recruit nucleosome-remodeling and histone acetylase complexes, thereby generating nucleosome-depleted regions that nevertheless have peaks of histone acetylation [187-189]. The RNA Pol II machinery recruits H3-K4 histone methylases near the core promoter and upon transcriptional elongation recruits H3-K36 and H3-K79 histone methylases to active coding regions. Although less well defined, other DNA-binding proteins and nascent RNA can recruit H3-K27 or H3-K9 methylases to other genomic regions, resulting in heterochromatic silencing by polycomb complexes or HP1, respectively [190, 191].

As a consequence of the above and other mechanistic relationships between TFs and chromatin-modifying activities, the genome-wide pattern of histone modifications and nucleosome occupancy can be used to classify promoters, enhancers, insulators, and distinct types of heterochromatic regions in a given cell type under a given physiological condition. Using chromatin immuno-precipitation (ChIP), formaldehyde-assisted isolation of regulatory elements (FAIRE), and DNase I hypersensitivity techniques coupled to massively parallel DNA sequencing, such classification of functional genomic regions has been done in several cell lines in the context of the ENCODE consortium [183-185, 192]. In addition, the ENCODE consortium has performed genome-wide mapping of binding sites for 80 chromatin associating factors (at the time of writing), most notably in the erythroid cancer cell line K562. These genome-wide maps provide an invaluable resource for uncovering new functional aspects of individual TFs.

NF-Y (also known as CBF, CP1) is a heterotrimeric, DNA-binding TF that is conserved in all eukaryotes [25]. NF-Y binds specifically to the CCAAT box [8, 10] that is frequently found in eukaryotic promoters [36, 38]. The NFYB and NFYC subunits contain histone-fold domains (HFDs) structurally related to H2B and H2A, respectively [26], which mediate formation of a stable histone-like heterodimer [25], to which NFYA binds, whereupon the resulting heterotrimeric complex can bind to DNA [8]. NFYA contains the sequence specific CCAAT recognition domain, and NFYB and NFYC also contact DNA through their HFDs [22-24]. All bases of the core pentanucleotide are critical for NF-Y binding, with immediate flanking sequences on both ends also being important for efficient DNA binding *in vitro* [193, 194] and *in vivo* [41, 42, 55].

At many promoters, the CCAAT box is highly positioned ~80 bp upstream of the transcriptional start site (TSS), in either orientation, suggesting that its location is important for gene expression. In essentially all promoters tested, mutation of the CCAAT box reduces or eliminates transcriptional activity [51]. In addition, functional inactivation of NF-Y subunits or the use of a dominant negative NFYA mutant indicates that NF-Y binding is important for the pattern of histone modifications at promoters (reviewed by [17]). Interestingly, bioinformatic studies comparing gene expression patterns in tumors vs normal tissues indicate that NF-Y sites are highly enriched in promoters of genes overexpressed in tumors [133-135], particularly in the most aggressive cohorts. The importance of NF-Y is further underscored by the early embryonic lethality of an NFYA mouse knockout model due to defects in cell proliferation and extensive apoptosis [15].

Here we describe the genome-wide analysis of NF-Y binding in three tumor cell lines. Using data generated by the ENCODE consortium, we analyze the bound loci with respect to

chromatin states and binding by other TFs. Our results uncover many new and unexpected aspects of NF-Y biology.

RESULTS AND DISCUSSION

Unbiased genome-wide identification of NF-Y binding sites

We performed ChIP with anti-NFYA and anti-NFYB antibodies in three cell types (K562, GM12878 and HeLaS3) followed by massively parallel DNA sequencing. Antibodies [51] were validated by Western blot and IP-WB showing that NFYA and NFYB were specifically recognized (Supplemental Figure 1, A, B). Immuno-precipitated DNA was validated using QPCR to known NF-Y targets (Supplemental Figure 1, C, D) and the reproducibility between biological replicates was high (Pearson correlations > 0.8).

Using a stringent cut-off (*P*-value <= 10^{-9}), we identified 12655, 7932 and 5457 NFYB binding sites and 4726, 289 and 3726 NFYA binding sites in K562, GM12878 and HeLaS3 cells, respectively (Figure 1, A). Applying the *de novo* motif discovery tool *MEME* to NFYB peaks in K562 cells, we identified the typical NF-Y binding motif (Figure 1, C) that corresponded well to the motif derived from ChIP-chip experiments [51]. Similar NF-Y binding motifs were found in all datasets (data not shown). These high-confidence binding sites, 83% of which had at least one CCAAT box within each site (with a mean of 1.7 motifs per site), were used for subsequent bioinformatic analyses. At lower stringency, we identified 14772 (*P* <= 10^{-7}) and 18523 (*P* <= 10^{-5}) NFYB sites in K562, 81% and 77% of which, respectively, had CCAAT boxes. The subset of NFYB sites with relatively high *P*-values in the range of 10^{-5} to 10^{-7} contained CCAAT boxes at a rate of ~60%, whereas the genomic background is ~5% for similarly sized regions (Supplemental Figure 2, A). Based on these observations, and a peak saturation analysis (Supplemental Figure 2, B), we estimate that there are an additional ~4000 low affinity NF-Y binding sites in the genome of K562 cells.

Figure 1: ChIP-Seq of two components of the NF-Y complex in three cell types

A. *MACS* peak analysis indicating peak numbers, mean peak lengths and standard deviations, at three different *P*-value thresholds for NFYA and NFYB ChIP-Seq datasets in GM12878, HeLaS3, and K562.

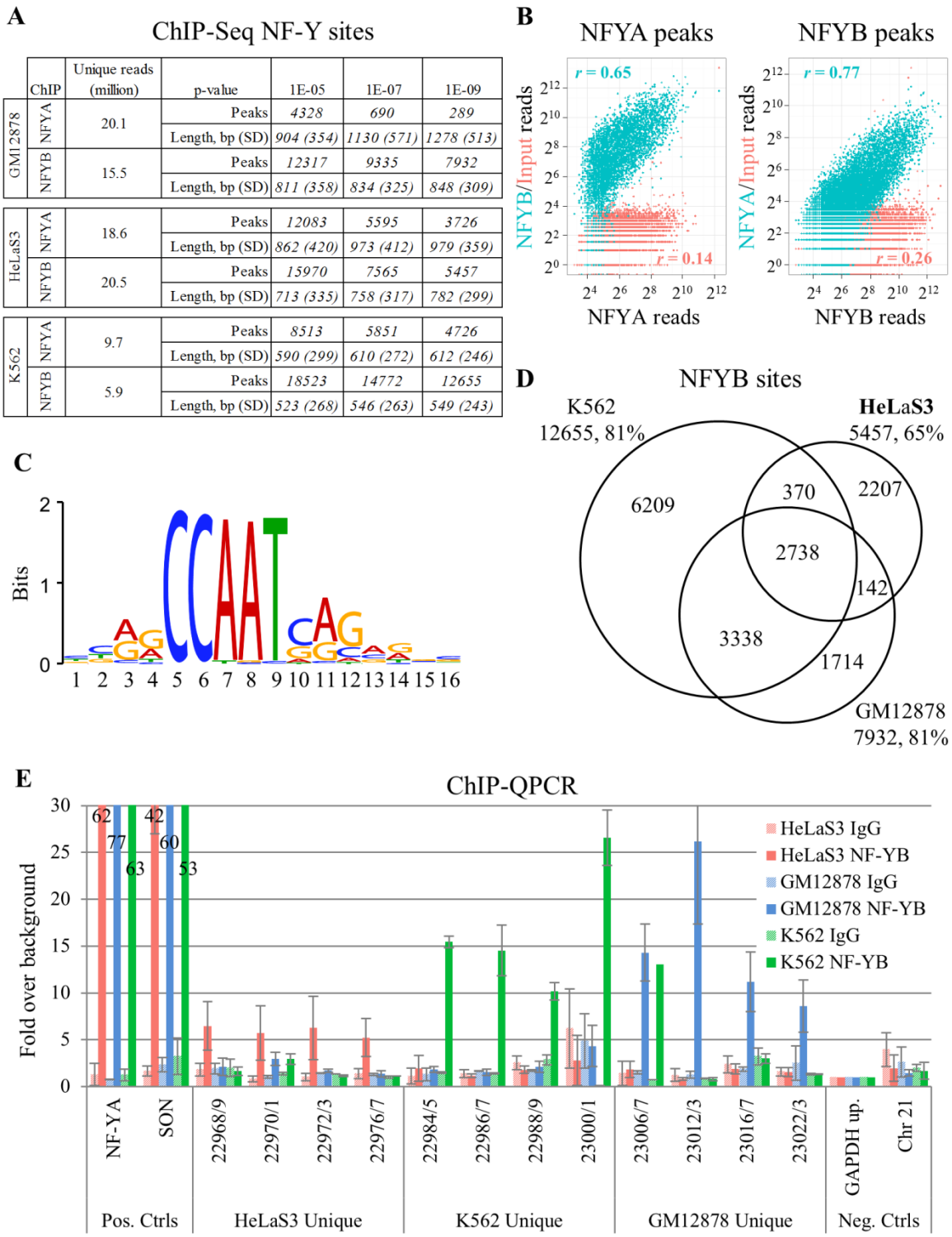
B. Scatter plots of NFYA, NFYB and input read counts at NFYA or NFYB sites in K562 showing correlation between datasets. Blue shading represents correlation amongst NFYA and NFYB. Orange shading represents NFYA or NFYB correlation with input.

C. Identification of the NF-Y DNA binding site motif *de novo* from 12655 K562 NFYB peaks depicted as a sequence logo [246].

D. Venn diagrams depicting the overlap between NFYB peak populations in GM12878, HeLaS3, and K562. Integers represent peak numbers called at the 10^{-9} *P*-value threshold. The percentages of peaks with CCAAT boxes are indicated (%).

E. ChIP-QPCR validation of NFYB peaks unique to each cell type. Error bars represent standard deviation of 3 biological replicates. "Pos. Ctrls." are loci known to be bound by NF-Y. "Neg Ctrls." are loci known to be devoid of NF-Y. Data represents a fold over background measurement compared to a non-NF-Y bound region (GAPDH up.). Solid and striped bars are ChIPs performed with NFYB specific antibody and non-specific rabbit IgG, respectively.

Figure 1 (Continued)



	ChIP	Unique reads (million)	p-value	1E-05	1E-07	1E-09
GM12878	NFYA	20.1	Peaks	*4328*	*690*	*289*
			Length, bp (SD)	*904 (354)*	*1130 (571)*	*1278 (513)*
	NFYB	15.5	Peaks	*12317*	*9335*	*7932*
			Length, bp (SD)	*811 (358)*	*834 (325)*	*848 (309)*
HeLaS3	NFYA	18.6	Peaks	*12083*	*5595*	*3726*
			Length, bp (SD)	*862 (420)*	*973 (412)*	*979 (359)*
	NFYB	20.5	Peaks	*15970*	*7565*	*5457*
			Length, bp (SD)	*713 (335)*	*758 (317)*	*782 (299)*
K562	NFYA	9.7	Peaks	*8513*	*5851*	*4726*
			Length, bp (SD)	*590 (299)*	*610 (272)*	*612 (246)*
	NFYB	5.9	Peaks	*18523*	*14772*	*12655*
			Length, bp (SD)	*523 (268)*	*546 (263)*	*549 (243)*

The apparently higher number of NFYB sites with respect to NFYA sites could be due to target loci bound only by NFYB. In this regard, in nuclei, NFYB is more abundant than NFYA, and NFYB is present in certain post-mitotic cells whereas NFYA is not detectable [12-14] . However, the NFYA and NFYB datasets were highly correlated (Pearson correlation 0.7-0.8; Figure 1, B), and quantitative PCR analysis of individual sites revealed 3-fold higher enrichments for NFYB than for NFYA. Furthermore, analysis of 21 NFYB sites that appeared to lack NFYA showed low occupancy of NFYB such that an NFYA peak was below the detection limit (Supplemental Figure 3, A and B). These results indicate that the NFYB antibody was more "immuno-efficient" than the NFYA antibody and that there were few, if any, genomic sites that were bound by NFYB but not NFYA. For this reason, we used the NFYB dataset to define NF-Y binding sites in subsequent analyses.

Approximately 39% of NF-Y sites were occupied in at least 2 cell types, whereas the remaining 61% of NF-Y-bound sites were cell-type specific (Figure 1, D). In accord with this observation, examination of 14 NF-Y target genes identified previously in different cell lines [42] revealed that 13 were bound in K562 and 8 were bound in HelaS3. We validated the cell type specificity of a small number of these loci by ChIP-QPCR (Figure 1, E). The lower number of NFYB bound loci in GM12878 and HeLaS3 was most likely due to the higher efficiency of the ChIP assay in K562 cells, rather than to biological differences.

Asymmetric binding of NFYA and NFYB to the CCAAT box

Linking the high-resolution positioning data of NF-Y subunits to the CCAAT box location, we confirmed that NFYA binds directly over the CCAAT sequence (Figure 2, A). Interestingly, the NF-Y complex is asymmetric, with NFYB binding ~15 bp downstream from the CCAAT box, as defined by the CCAAT strand (Figure 2, A). This asymmetry fits extremely well with the

Figure 2: Annotation of NF-Y peaks to genomic features

A. The average position of NFYB is upstream of the CCAAT box. Kernel density estimate of the distribution of the 5'-CCAAT-3' and 5'-ATTGG-3' sequences under NFYA and NFYB peaks in relation to the peak summit centered at 0 bp. Only the position of best matching CCAAT box within 100 bp of the peak summit was considered and plotted. Transparent lines indicate raw data; solid lines indicate Gaussian smoothed data.

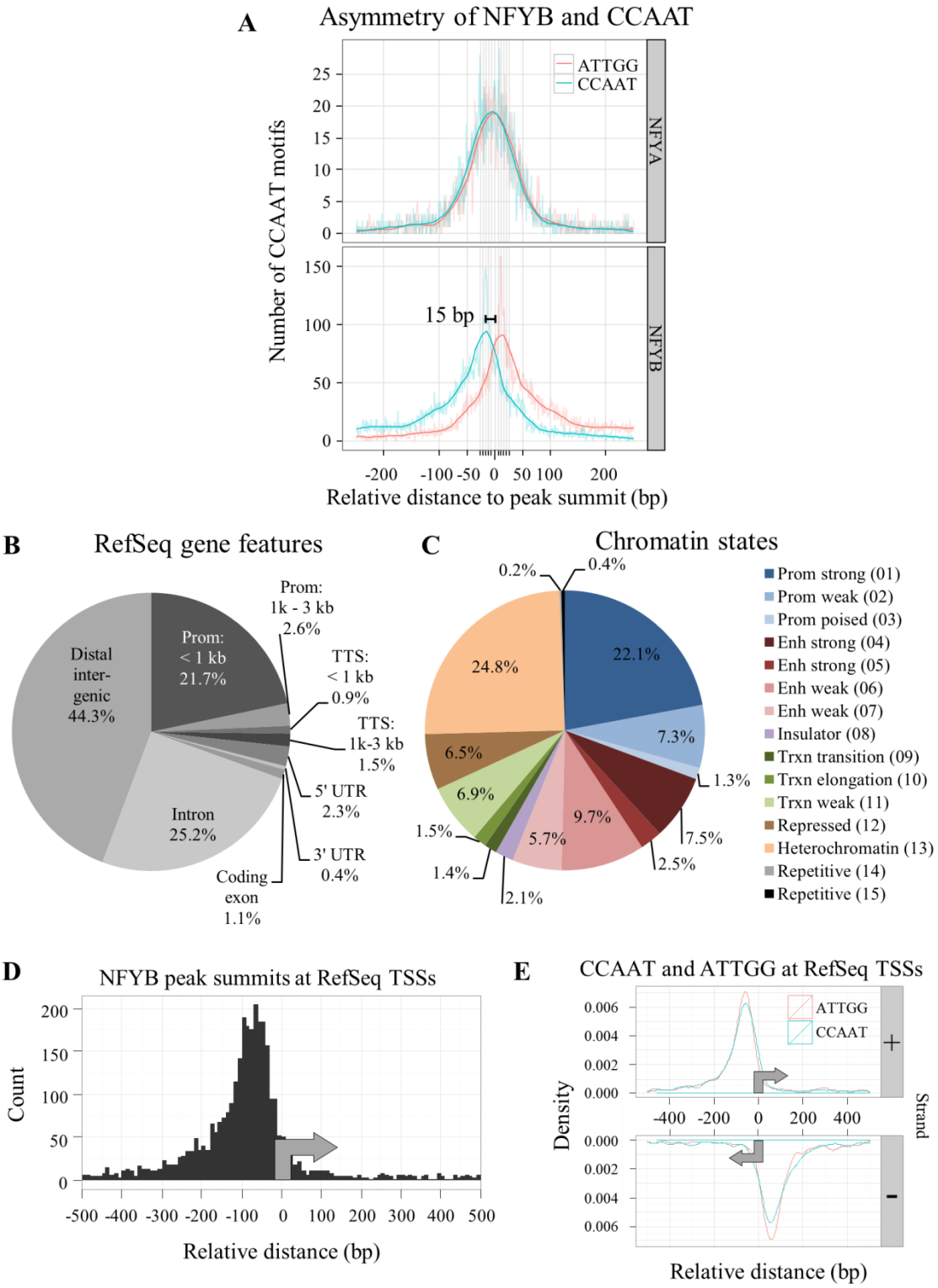
B. Annotation of K562 NFYB peak summits to RefSeq gene features.

C. As in A, except chromatin state maps were used. Abbreviations are: "Prom" = promoter, "enh" = enhancer, "trxn" = transcription. Numbering is from the chromatin state maps of [205].

D. Frequency distribution of K562 NFYB peak summits at RefSeq TSSs showing a preferential location between -50 and -100 bp upstream of the TSS.

E. Gaussian kernel density estimate of the distribution of positive and negative strand 5'-CCAAT-3' and 5'-ATTGG-3' motifs at K562 NFYB bound RefSeq TSSs. Only the best motif per region was considered. Bandwidth was equal to the standard deviation of the smoothing kernel. Gray arrows indicate the direction of transcription.

Figure 2 (Continued)



available biochemical knowledge of NF-Y/DNA contacts (Dolfini et al., 2009) and with the crystal structure of trimer interactions with DNA ([25]; M. Nardini, M. Bolognesi, R. Mantovani, in preparation). The high resolution of protein-DNA positioning achievable through ChIP-Seq, and the large number of datasets available through ENCODE, urges the detailed and expansive analyses of chromatin bound protein complex prediction among transcriptional regulators (see below).

NF-Y targets cell signaling, DNA repair, cell-cycle, metabolic and gene expression genes

GREAT gene ontology analysis of NFYB bound loci from K562, GM12878 and HeLaS3 revealed a strong enrichment of genes involved in cell signaling pathways ("*Integrin alpha2beta3 signaling*", "*Signaling mediated by p38-gamma and p38-delta*"), cell cycle ('*G2/M checkpoints*", "*Regulation of DNA replication*"), DNA repair ("*Homologous recombination repair"* and *"Base excision repair*") and metabolism ("*Superpathway of cholesterol biosynthesis*", "*Metabolism of polyamines*") (Table 1). Cell cycle and metabolism terms are in line with previous findings, and further stress the central role of NF-Y in growth controlling decisions.

In addition, just below our fold enrichment cutoff, we found a preponderance of GO terms associated with gene expression in all three cell lines. Upon further analysis, it was apparent that NF-Y significantly targeted genes involved in "*Transcription*", "*mRNA splicing*", "*mRNA editing*", "*mRNA 3'-end processing*", and "*mRNA transport*". Included were a large and diverse set of TFs, including the NF-Y genes themselves, members of the transcriptional machinery, and co-activators and co-repressors (Supplemental Figure 4, A and B). Thus, the picture emerging is one of NF-Y as a regulator of gene expression regulators.

Table 1: NF-Y binds to genes involved in cell signaling, DNA repair, cell-cycle, and gene expression

Gene ontology analyses of NFYB bound loci in K562, GM12878 and HeLaS3. Only the top 10 terms with a fold enrichment > 2 are shown. Observed region hits correspond to the number of regulatory regions, of genes in that gene ontology term, that had >= 1 NFYB sites. Highly redundant categories are not shown. For a full list of significant GO terms see Supplemental Data.

	GO term	*P*-value	FDR q-value	Fold enrichment	Observed hits
GM12878	G2/M DNA damage checkpoint	7.7E-11	1.1E-08	2.1	91
	M/G1 Transition	1.1E-09	8.5E-08	2.0	91
	Homologous recombination repair	1.8E-09	1.0E-07	2.0	89
	Polo-like kinase mediated events	4.5E-07	9.7E-06	2.0	62
	APC/C:Cdc20 mediated degradation of Securin	8.8E-07	1.7E-05	2.1	54
	Ubiquitin-dependent degradation of Cyclin D	1.7E-06	2.9E-05	2.4	39
	Superpathway of cholesterol biosynthesis	2.3E-06	3.7E-05	2.4	36
	Signaling mediated by p38-gamma and p38-delta	1.4E-05	1.7E-04	2.8	23
	Activation of ATR in response to replication stress	1.6E-05	1.8E-04	2.1	39
	Integrin alphaIIbbeta3 signaling	1.9E-05	2.1E-04	2.0	44
HeLaS3	G2/M checkpoints	2.3E-09	6.6E-07	2.1	79
	Homologous recombination repair	2.9E-07	1.7E-05	2.0	62
	RNA polymerase I chain elongation	4.6E-06	1.2E-04	3.2	21
	Retrograde neurotrophin signalling	5.5E-06	1.3E-04	3.6	18
	Regulation of DNA replication	6.8E-08	7.9E-06	2.6	43
	Integrin alphaIIbbeta3 signaling	1.7E-07	1.2E-05	2.6	39
	Alpha6Beta4Integrin	7.5E-07	2.9E-05	2.3	44
	Synthesis of DNA	2.8E-06	9.0E-05	2.0	52
	Cyclin E associated events during G1/S transition	3.1E-06	9.9E-05	2.0	52
K562	Regulation of DNA replication	2.9E-09	9.5E-08	2.1	81
	Unwinding of DNA	6.0E-09	1.9E-07	4.7	22
	Nucleosome assembly	1.6E-07	3.4E-06	2.4	46
	Ubiquitin-dependent degradation of Cyclin D	7.0E-07	1.2E-05	2.1	55
	Signaling events mediated by PRL	1.3E-06	2.0E-05	2.3	41
	RNA polymerase I transcription	2.0E-06	2.7E-05	2.1	49
	Response to elevated platelet cytosolic Ca2+	2.1E-06	2.7E-05	2.8	28
	Signaling by Rho GTPases	5.7E-06	5.8E-05	4.8	13
	Base excision repair	9.5E-06	9.1E-05	3.1	21
	Metabolism of polyamines	2.1E-05	1.9E-04	3.1	19

In a separate analysis (*IPA*, Ingenuity Systems) of signaling pathways, we found that NF-Y preferentially associates with genes involved in the inter-related TP53 (p53) and TRAIL apoptotic (death receptor) pathways (Supplemental Figure 4, C and D). This observation reinforces the notion of a direct and indirect NF-Y/TP53 interplay, with opposing functional consequences depending on the *TP53* status of the cell, *i.e.* proliferation or apoptosis (reviewed in [195]). In addition, it is consistent with anecdotal evidence about the role of NF-Y in apoptosis [196, 197], which helps explain the phenotypes of NFYA overexpression and inactivation experiments [85], and point to specific molecules as areas of future investigation.

NF-Y binds to a diverse set of genomic features including non-genic regions

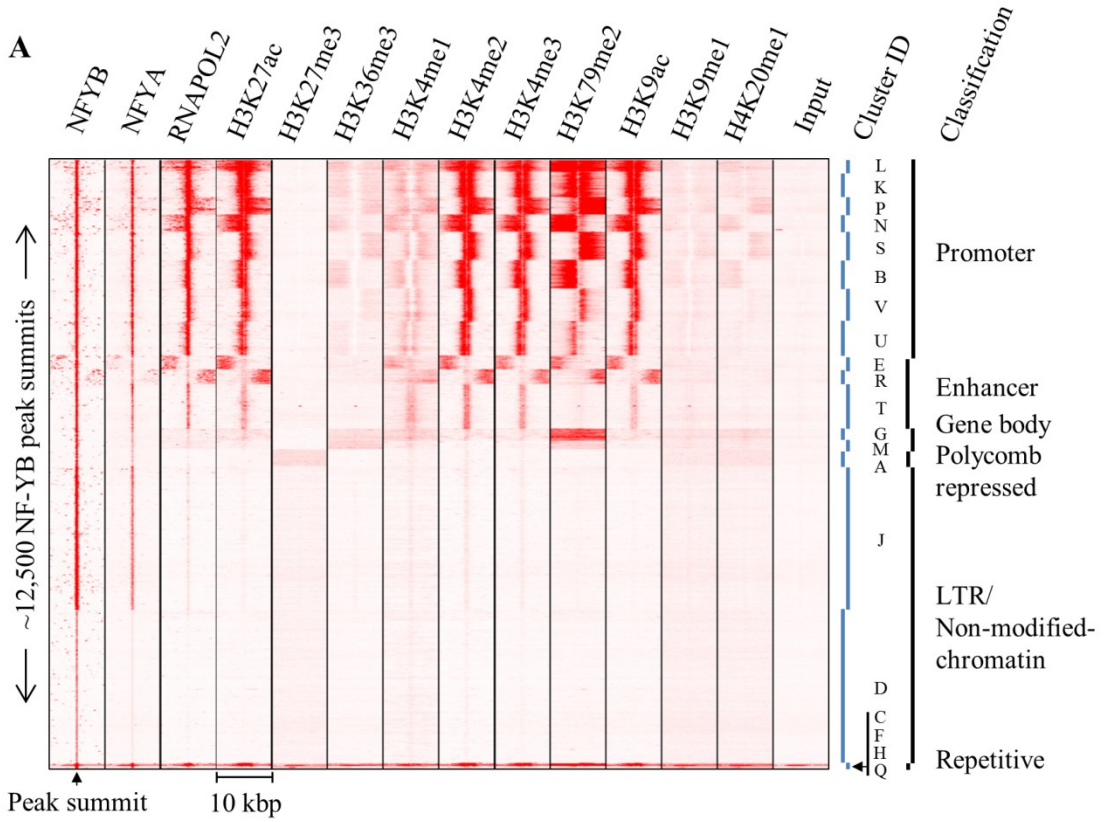
We annotated the NFYB bound regions in K562 to RefSeq genes (Figure 2, B; Supplemental Figure 5), maps of histone modifications (Figure 2, C) and nucleosome-depleted regions, and RNA levels (Figure 3, A and B). Unexpectedly, ~25% of the NF-Y binding sites were not situated near RefSeq promoters, genic regions (lncRNAs [198]; miRBASE [199]; UCSC RNA genes [200]; NONCODEdb [201]) or loci bound by RNA Pol II or Pol III [202] (not shown). These sites were not false positives as the vast majority (88%) contained CCAAT boxes, and 46% of them were present in at least one other cell type. Based on the patterns of co-localized histone modifications, and RNA Pol II, NF-Y bound regions in K562 and HeLaS3 reproducibly partitioned into 20 clusters that could be grouped into five major classes (Figure 3, A and B; Supplemental Figure 6): promoter, enhancer, gene body, PcG repressed, and LTR/non-modified-chromatin. As discussed below, these results indicate that NF-Y binding was prevalent in tissue-specific enhancers and specific types of repetitive sequences, in addition to proximal promoters, where NF-Y has traditionally been observed.

Figure 3: NF-Y bound loci resided within 5 epigenetic domains

A. K-means clustering of K562 NFYB loci based on the distribution of histone PTM, RNA Pol II, NFYB and NFYA ChIP-Seq reads within a region spanning +/-5 kbp from the summit of NFYB peaks (centered at 0 bp). Clustering was carried out on transformed rank normalized read counts. Raw read count intensity is depicted in red. The interpretation and classification of clusters into functional categories are shown to the right.

B. NFYB summits from clusters derived from A were annotated to genomic features: chromatin states, LTRs, dbTSS, RefSeq promoters, and FAIRE-Seq regions. The percentage of peak summits within each cluster overlapping a specific feature is indicated. Overlap with LTRs was assayed within a window of +/-250 bp from the ends of the LTR feature. RefSeq promoters were considered within a window of -2500:+500 bp from the TSS. A direct overlap with FAIRE-Seq regions and chromatin states was used. Long polyA purified RNA reads were counted within a window of +/-500 bp about the NFYB peak summit and the median value of that cluster is shown.

Figure 3 (Continued)



B

Chromatin state\Cluster	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	T	U	V
Promoter strong (01)	0	71	0	0	2	35	1	0	0	76	63	0	63	58	0	3	73	12	55	60
Promoter weak (02)	3	8	0	2	5	2	5	5	4	5	4	3	8	9	19	7	8	29	17	13
Promoter inactive/poised (03)	16	0	0	1	0	0	0	0	2	0	0	0	0	0	14	0	0	1	0	0
Enhancer strong (04)	0	5	0	0	17	6	25	0	0	10	23	4	16	21	2	17	6	29	14	13
Enhancer strong (05)	0	0	0	1	11	0	11	0	1	0	1	8	2	1	5	13	1	11	2	1
Enhancer weak (06)	4	14	0	7	6	0	3	5	12	8	9	1	9	9	14	9	11	15	10	12
Enhancer weak (07)	5	0	0	10	16	0	9	0	7	0	0	6	1	1	2	20	1	2	1	0
Insulator (08)	4	1	0	5	1	0	0	0	3	0	0	0	0	0	0	2	0	0	0	0
Transcriptional transition (09)	1	0	0	1	6	4	24	0	1	0	0	25	0	0	12	3	0	0	0	0
Transcriptional elongation (10)	0	0	0	1	4	0	13	0	1	0	0	39	0	0	0	4	0	0	0	0
Transcribed weak (11)	3	1	0	12	18	0	9	5	12	0	0	10	0	1	2	14	0	0	0	0
Repressed (12)	60	0	100	11	2	0	0	0	8	0	0	0	0	0	12	1	0	0	0	0
Non-modified chromatin (13)	3	1	0	50	11	0	0	0	49	0	0	1	0	0	0	7	0	0	0	0
Repetitive (14)	1	0	0	0	0	0	0	33	0	0	0	0	0	0	14	0	0	0	0	0
Repetitive (15)	0	0	0	0	0	53	0	52	0	0	0	2	0	0	2	0	0	0	0	0
LTR (+-250bp)	54	6	100	58	51	2	47	29	82	4	7	43	12	10	40	52	4	30	7	8
dbTSS (+-250bp)	27	88	0	24	42	71	65	43	21	92	92	61	86	86	48	46	90	55	80	85
RefSeq (-2500, +500bp)	11	88	0	8	17	53	2	19	7	76	67	2	76	75	24	19	90	30	70	73
FAIRE-Seq	25	86	0	11	36	69	44	24	24	87	83	31	77	84	43	37	85	78	83	88
RNA (median, +-500bp)	0	536	0	0	7	1161	9	5	0	1003	885	41	391	727	5	10	769	6	185	346
n =	355	589	1	3,187	285	55	245	21	2,982	536	240	181	365	355	42	309	582	931	708	686

Percent overlap

Reads

Only a minority of NF-Y binding sites are located at proximal promoter regions

Although NF-Y is typically described as a factor that binds to proximal promoter regions, only 22% of NF-Y sites were located within 1 kbp upstream of a RefSeq TSS (Figure 2, B; Supplemental Figure 5), consistent with our previous analysis of 2% of the *H. sapiens* genome [42]. For such proximal promoter binding sites, a frequency distribution plot of NFYB peak summits indicated that NF-Y was highly positioned upstream of the TSS at -40 to -100 bp (Figure 2, D), in line with the position of the CCAAT box at TSSs (Figure 2, E), in agreement with previous observations [51]. Though NFYA and NFYB bound asymmetrically to the CCAAT box, the orientation with respect to the TSS was largely irrelevant for transcription, as only a small difference in the frequency of CCAAT and its complement ATTGG were noticed on the same strand (Figure 2, E). More generally, only a third of NF-Y loci (clusters L, K, P, N, S, B, V, U; n = 4061; Figure 3, A) were associated with active promoters, as defined by high levels of di- and tri-methylated H3-K4, acetylated H3-K27 and H3-K9, RNA Pol II, and nucleosome depletion (defined by a "valley" of low enrichment of mono-methylated H3-K4 at NFYB summits and a positive FAIRE signal; Figure 3, A and B).

A subset of NF-Y sites was located at tissue-specific enhancers

NF-Y binding to enhancers has been rarely described, *e.g.* the 5' upstream regions of the MHC class II genes [203] and the intronic enhancer of the Hoxb4 gene [204]. Of NF-Y peak summits, 25% were located within a region demarcated by Ernst et al. [205] to be an enhancer chromatin state (Figure 2, C). Our analysis, using similar datasets, found a lower percentage of NF-Y sites to be located within regions consistent with known histone modifications of enhancers (12%; clusters E, R and T; n = 1525; Figure 3, A). This discrepancy is likely due to our more conservative definition of enhancer and the wider genomic region used for interpretation. NF-Y

sites adjacent to Ernst et al promoter states, though still within the histone modifications defining that state derived from the nearby active TSS, were designated by us to be "promoter", however, they would be classified as "enhancer" by Ernst et al. Clusters E and R are exceptional, in that they represent NF-Y sites located close to (~2.5 kb), but not within regions of high enrichment for H3-K27ac, H3-K9ac, H3K4me1/-2/-3 (strong actively transcribing promoters), unlike all other clusters from the enhancer and promoter groups where NF-Y is directly within the enriched domains.

Interestingly, cell type specific NF-Y sites were enriched for enhancers and were, on average, located further away from TSSs as compared with NF-Y sites common to all cell types (Supplemental Figure 7, A and B). GO analysis of cell-type specific NFYB loci reveals categories enriched in individual cell types: "*NF-κB cascade and regulation of IL12*" was enriched in GM12878, a cell type where NFκB is constitutively active [206, 207]; HeLaS3 showed enrichment for "*Epidermis morphogenesis*" and "*Establishment of tissue polarity*", commonly associated with cells of epithelial origin (Supplemental Figure 7, C).

Functional inactivation of NF-Y indicates a transcriptional role for NF-Y located distally to TSSs

Given the preponderance of NF-Y locations distal to TSSs, we decided to identify the direct transcriptional targets of NF-Y by performing expression array analysis on HeLaS3 cells depleted for NFYA by lentiviral small hairpin RNA (shRNA) (Supplemental Figure 8, A and B) and correlating these changes to the location of NF-Y. At a *P*-value cutoff of 10^{-4}, 84 genes were down-regulated and 252 genes were up-regulated (Table 2) upon NFYA knockdown. Of these, only 11% (n = 9) and 39% (n = 98) had NF-Y bound to their proximal promoters, respectively. The topmost differentially down- and up-regulated genes both

Table 2: shRNA knockdown of NFYA

Differentially expressed genes upon NFYA knockdown in HeLaS3 and the number that was bound by NF-Y as determined by ChIP-Seq. Windows are in relation to RefSeq TSSs. Adjusted *P*-value is Bonferroni corrected.

Cutoff *P*-value (adjusted)	Genes differentially regulated		NFYA bound				NFYB bound			
			-2.5 kbp, +500 bp		+/-50 kbp		-2.5 kbp, +500 bp		+/-50 kbp	
			#	p-value	#	p-value	#	p-value	#	p-value
1.00E-06 (7.5E-04)	Down	9	2	1.8E-01	2	2.0E-01	2	1.0E+00	2	7.3E-01
	Up	25	3	4.7E-01	3	5.0E-01	10	1.7E-01	11	1.9E-01
1.00E-05 (2.0E-03)	Down	27	2	1.0E+00	3	7.3E-01	3	8.1E-02	3	2.2E-02
	Up	91	15	1.3E-02	15	2.6E-02	45	3.3E-06	50	3.4E-06
1.00E-04 (6.7E-03)	Down	84	3	1.2E-01	4	1.9E-01	9	6.8E-04	9	1.5E-05
	Up	252	34	8.5E-03	37	4.1E-03	98	1.3E-05	110	2.6E-05
1.00E-03 (2.7E-02)	Down	220	5	2.2E-04	6	2.3E-04	19	4.3E-11	28	2.4E-10
	Up	629	101	5.9E-10	108	1.3E-10	233	1.6E-09	264	4.5E-09
1.00E-02 (1.1E-01)	Down	513	12	4.9E-09	18	7.6E-07	54	3.7E-19	79	1.8E-16
	Up	1518	223	9.8E-17	245	9.7E-20	536	4.3E-16	624	4.4E-18

trended towards having a higher percentage of their promoters occupied by NF-Y than non-differentially regulated genes (Supplemental Figure 8, D). Of the 1059 NF-Y peaks in HeLaS3 located within 250 bp of a RefSeq TSS, only 5.2% were differentially regulated at a *P*-value of 10^{-4} (n = 55). The low percentage of differentially regulated genes bound by NFYB (or NFYA) was similar to that found with other TFs [208-210] and could be exacerbated by the incomplete functional inactivation of NF-Y.

The above observations suggest that NF-Y located more distally may be important transcriptionally for the differentially regulated genes. In this regard, we ranked NF-Y sites by the fold change in RNA expression of the nearest associated gene upon NF-Y inactivation. The most strongly down-regulated genes had NF-Y sites that were much more distal to the TSS, with the median distance being >10 kb (Supplemental Figure 8, C). This data suggests that NF-Y located at enhancers, was important for transcription of neighboring genes.

LTRs were the most prevalent class of NF-Y sites in the *H. sapiens* genome

Of all NF-Y binding sites in K562, 40% directly overlapped an LTR, the promoter elements of endogenous retroviruses, making LTRs the most prevalent class of NF-Y loci in the *H. sapiens* genome, even more so than core promoters of endogenous genes (Figure 4, A). NF-Y selectively associated with two families of LTRs - MLT1 and LTR12 (Figure 4, B and C). NF-Y did not bind to all LTR families, irrespective of the presence of a CCAAT box in the consensus sequence. The R66 tandem repeat (which is related to LTR12B [211, 212]), MER51A and MER51E also associated with NF-Y. In general, there was no significant cell type specificity in LTR binding.

Most NF-Y bound sites at LTRs lacked any detectable histone modifications within 5 kbp of the NF-Y peak summit (clusters D and J in Figure 3, A and B; Figure 4, D). These NF-Y loci

appeared to be transcriptionally inactive, yet maintained substantial NFYB and NFYA occupancy. Although most NF-Y bound LTRs appeared to be transcriptionally inactive, promoter and enhancer chromatin states with high levels of H3 acetylation and/or H3-K4 methylation contained a sizeable minority (27% K562; 20% GM12878; Figure 4, D) of LTRs. These appeared to be transcriptionally active, thus most likely representing functional *cis*-regulatory elements derived from transposable repetitive elements and regulating endogenous genes.

LTRs function as promoter elements of endogenous retroviruses and they can act as regulatory elements for certain host genes [213]. NF-Y sites abound in viral LTRs [214-218]. The selectivity for the gamma-retrovirus LTR family and within it for certain members, likely reflects the presence of CCAAT in the original viral LTRs. Thus, our results suggest a strong genetic pressure on their genomic transduced copies to maintain NF-Y binding. This is not unprecedented, as evidenced by the preference of particular TFs for specific repetitive sequences [219, 220]. Genetic analysis of the ERV-9/LTR12 element located 5' of the globin locus-control region indicates a crucial role of the 14 CCAAT and GATA containing E3 repeats for expression of the β-globin locus [221, 222]. Because of this precedent, we expected to observe a genomic theme of most, if not all, LTR repetitive sequences bound by NF-Y to be either at enhancers or promoters in regions of H3 acetylation. Instead, the opposite was true. The majority were associated with heterochromatin-like domains apparently devoid of any transcriptional signal, either positive or negative. Since the vast cohort of endogenous LTR proviral sites were under strong control by the host organism and, in most cases, actively repressed by genetic and epigenetic means [213], we are tempted to speculate that NF-Y plays a role in the epigenetic

Figure 4: NF-Y binds extensively to long terminal repeats

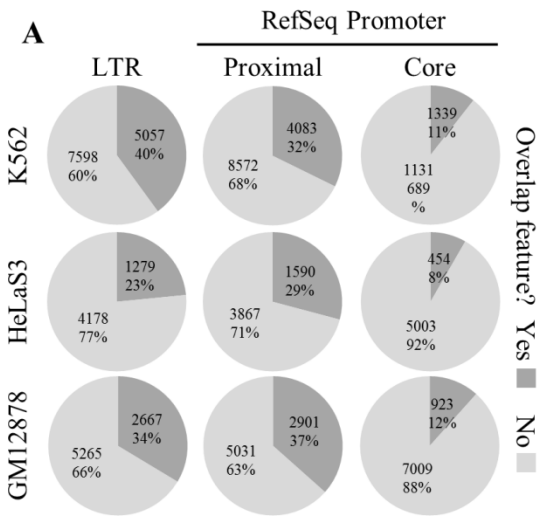
A. NFYB peaks were extensively found at LTRs, more so than core promoters. The percentage of all K562 NFYB peak summits that occupy the indicated feature. Core and proximal promoters are defined as -250:+50 bp and -2500:+500 bp from the TSS of RefSeq promoters, respectively.

B. Mapping of ChIP-Seq reads from K562, GM12878 and HeLaS3 to RepBase consensus sequences showing an abundance of NF-Y specific reads mapping to repetitive elements. Ratios reflect the enrichment of reads in the NFYB ChIP sample as compared to input. Only RepBase entries with a read ratio >= 5 are shown. Orange shading indicates repeat elements present in all cell lines. Green and red shading indicate the presence and absence, respectively, of a CCAAT box match at *P*-value < 10^{-4} in the consensus sequence.

C. Frequency of overlap between NFYB peak summits and the genomic locations of LTR families showing that only a specific subset of LTR families are bound by NF-Y. Only LTR elements that overlapped at least one peak in each cell line are shown. The two most highly overlapping repeat families are indicated, LTR12 and MLTJ1.

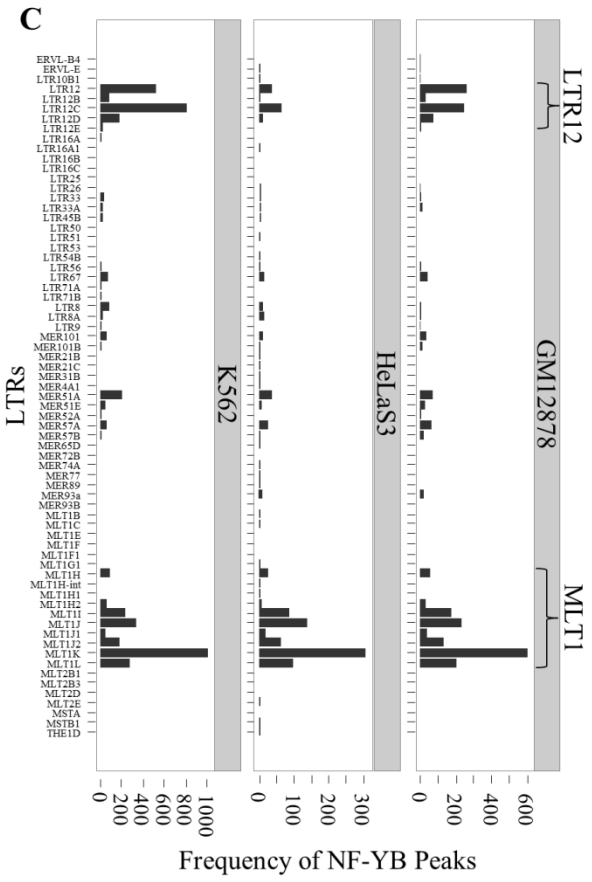
D. NF-Y bound LTRs were mainly situated within heterochromatin-like domains. Distribution of NFYB bound LTRs from K562 and GM12878 at chromatin states. No chromatin state map was available for HeLaS3.

Figure 4 (Continued)



B

Name	Type	NF-YB % Reads	Input % Reads	Ratio IP:Input	
LTR12D	Long terminal repeat	8.10	0.13	62.4	K562
R66	Tandem repeat	0.22	0.00	61.8	
LTR12	Long terminal repeat	4.22	0.08	54.3	
LTR12B	Long terminal repeat	3.33	0.07	47.8	
MER122	Interspersed repetitive element	0.12	0.00	46.8	
LTR12E	Long terminal repeat	6.83	0.16	42.2	
LTR12C	Long terminal repeat	21.67	0.52	41.7	
LTR12F	Long terminal repeat	1.19	0.04	27.4	
MLT1J	Long terminal repeat	0.06	0.00	27.3	
REP522	Repetitive subtelomeric-like	0.07	0.00	26.9	
LTR71B	Long terminal repeat	0.28	0.02	17.0	
MER51E	Long terminal repeat	0.09	0.01	15.7	
A@1	Simple repeat	1.58	0.13	12.1	
U1	snRNA	0.28	0.02	12.1	
MLT1J1	Long terminal repeat	0.01	0.00	8.6	
C@1	Simple repeat	0.06	0.01	8.2	
MER51A	Long terminal repeat	0.24	0.03	8.0	
A@2	Simple repeat	0.58	0.08	7.0	
C@2	Simple repeat	0.02	0.00	6.9	
MER101	Long terminal repeat	0.03	0.00	6.9	
R66	Tandem repeat	0.02	0.00	9.2	HeLaS3
LTR12	Long terminal repeat	0.48	0.06	7.9	
LTR12B	Long terminal repeat	0.40	0.05	7.7	
LTR12D	Long terminal repeat	0.89	0.13	6.9	
LTR12F	Long terminal repeat	0.19	0.03	5.9	
U1	snRNA	0.11	0.02	5.8	
MER51A	Long terminal repeat	0.08	0.01	5.6	
MER51E	Long terminal repeat	0.02	0.00	5.0	
LTR12	Long terminal repeat	1.86	0.07	28.5	GM12878
LTR12D	Long terminal repeat	3.00	0.11	26.9	
LTR12B	Long terminal repeat	1.42	0.06	24.9	
R66	Tandem repeat	0.08	0.00	23.1	
LTR12E	Long terminal repeat	2.90	0.15	19.8	
LTR12C	Long terminal repeat	7.78	0.43	17.9	
LTR12F	Long terminal repeat	0.52	0.04	14.3	
MER51A	Long terminal repeat	0.17	0.03	5.6	
MER51E	Long terminal repeat	0.03	0.01	5.1	



D

	K562	GM12878
Promoter strong (01)	1.4%	1.2%
Promoter weak (02)	2.4%	2.1%
Prom. inactive/poised (03)	1.4%	0.1%
Enhancer strong (04)	5.2%	4.0%
Enhancer strong (05)	3.1%	3.6%
Enhancer weak (06)	6.6%	3.2%
Enhancer weak (07)	7.3%	6.0%
Insulator (08)	1.4%	0.7%
Trxn. transition (09)	1.1%	0.9%
Trxn. elongation (10)	1.9%	2.5%
Trxn. weak (11)	11.5%	9.5%
Repressed (12)	9.0%	2.7%
Non-modified-chr. (13)	47.3%	63.4%
Repetitive (14)	0.2%	0.3%
Repetitive (15)	0.1%	0.0%

repression of these LTRs in somatic tissue, and/or in their activation during embryogenesis, where many repetitive elements are de-methylated and become expressed [223].

NF-Y binds CCAAT boxes in non-modified-chromatin domains *in vivo,* unlike most TFs

The majority of NF-Y sites (n = 6169; 49%) were in 2 similar clusters (D and J, *i.e.* LTR/non-modified-chromatin class; Figure 3, A) that displayed no positive or repressive histone modifications, negligible RNA Pol II and polyA RNA, and overlapped few open regulatory regions (11, 25%) and RefSeq TSSs (7, 11%). Interestingly, most of these loci overlapped LTRs, 58% and 82%, respectively (Figure 3, B). These NF-Y sites are interesting as most TFs are believed to not be able to bind to their DNA motifs within closed, transcriptionally inactive chromatin domains.

To further explore this issue, we calculated the percentage of motifs residing within NFYB peaks within distinct chromatin states, over a range of motif quality scores. Interestingly, and unlike other TFs such as E2Fs and MYC, NFYB was not excluded from any chromatin state assayed (Figure 5, A-C). At strong and weak promoters, > 80% of CCAAT boxes (with scores >= 16) were occupied by NF-Y (Figure 5, A). CCAAT boxes at enhancers and insulators were also well occupied by NF-Y (30-65%, respectively; Figure 5, A) although the percent occupancy was lower than at strong promoters, indicating that binding to these genomic regions was more selective. More generally, CCAAT boxes situated within open chromatin regions, as defined by FAIRE, were exceptionally well occupied to near saturated levels by NF-Y, with 80% occupancy (Figure 5, A). Interestingly, many CCAAT boxes within the non-modified-chromatin (10%), PcG repressed (20%) and transcription elongation states (10-25%) were occupied by NF-Y.

To test whether the substantial occupation of CCAAT boxes within non-modified chromatin and repressed genomic contexts was unique to NF-Y, we performed the same analysis

on 22 TFs, whose binding sites in K562 cells have been determined by the ENCODE consortium [183-185] (Figure 5, B; Supplemental Figure 9). As expected, most of the TFs examined showed motif occupancy at nucleosome-depleted regulatory regions at high levels, comparable to those of NF-Y. In contrast, GATA1 and GATA2, thought to be "pioneer" TFs (for review see [224, 225]), were highly selective and unable to saturate their motifs that resided within these nucleosome-depleted regulatory regions. However, most TFs lacked the ability to occupy even their highest quality motifs within non-modified and repressed chromatin states. For the 23 factors tested, only USF1, MAFK, and NF-Y could bind to motifs in the context of nucleosomes lacking some of the most common "positive" histone modifications or containing the repressive H3-K27me3 mark (Figure 5, B; Supplemental Figure 9).

By preventing accessibility to target sites, chromatin is a formidable barrier for binding by most TFs. This creates a dilemma as to how *cis*-regulatory elements and their resident DNA motifs can provide transcriptional competency if they cannot be accessed by trans-acting factors. There are a small number of "pioneer" factors that can efficiently bind to their motif located within non-nucleosome depleted, non-modified chromatin. Once bound, these pioneer TFs can recruit chromatin-modifying activities to generate open chromatin for the subsequent binding of partnering TFs [224, 225]. NF-Y can associate with a CCAAT box after nucleosome assembly *in vitro*, and the NFYB/NFYC HFD dimer can physically interact with H3/H4 in solution and on DNA [83]. Indeed, NF-Y binding is not mutually exclusive with nucleosomes *in vitro*, giving NF-Y the theoretical functional ability to interact efficiently with chromatin bound CCAAT boxes *in vivo*. NF-Y binds to a sizeable number of sites either in functionally "hostile" environments, or sites lacking all the common positive histone modifications. Perhaps, the structural features of the HFD heterodimer are instrumental for this. We propose that NF-Y is a

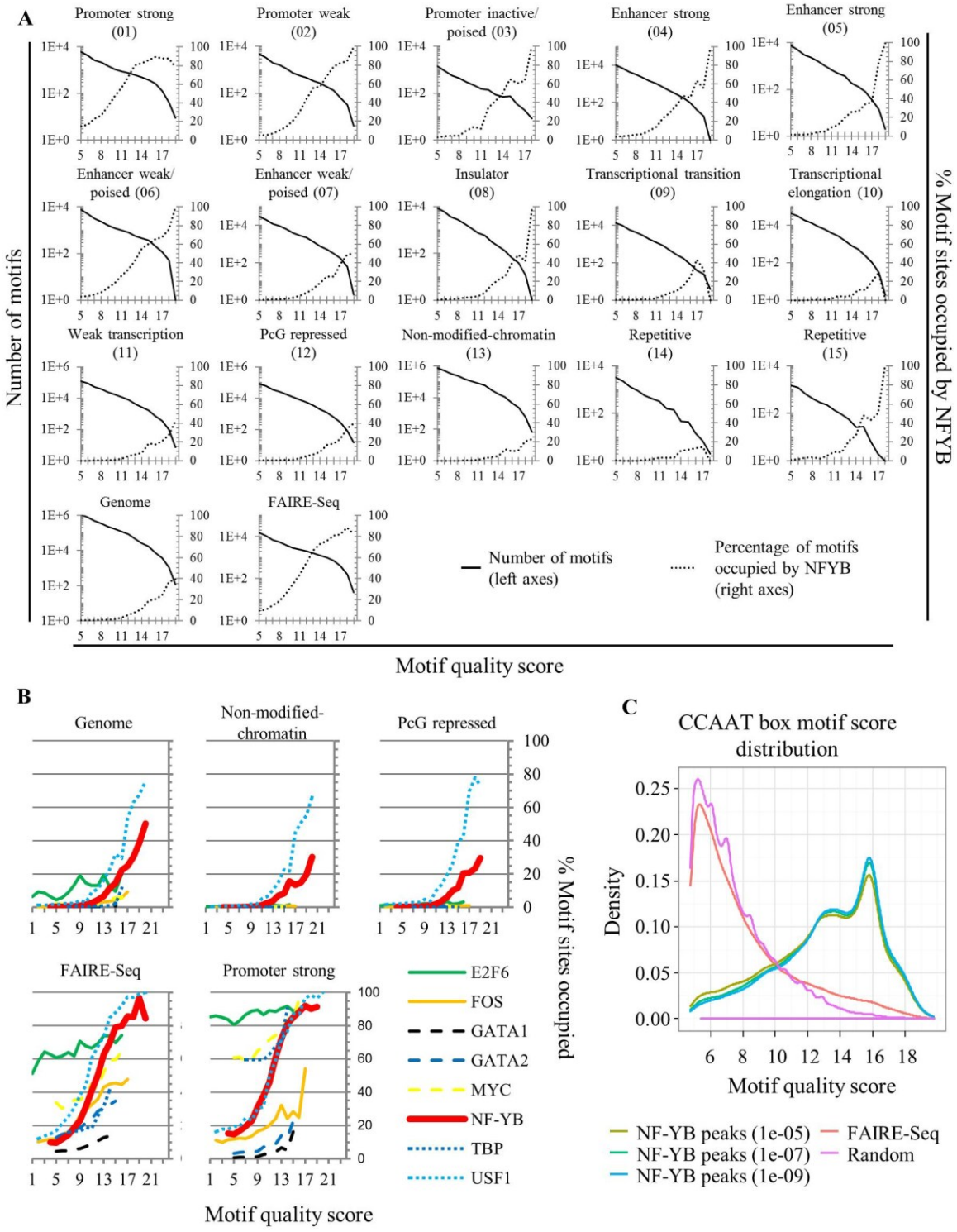
Figure 5: NF-Y can occupy its motif in closed chromatin

A. NF-Y has the ability to bind to its motif in many epigenetic domains, including repressed and non-modified-chromatin regions. The percentage of genome-wide computationally discovered CCAAT boxes within each chromatin state, FAIRE-Seq regions or the entire genome, that directly overlapped NFYB K562 sites plotted as a function of CCAAT box motif quality (right axes). Also shown are the numbers of discovered CCAAT boxes as a function of CCAAT box motif quality (left axes). Numbering was derived from [205].

B. NF-Y was unusual in its ability to bind to closed chromatin CCAAT boxes. Similar to A, except motif sites of different TFs are plotted as a function of motif quality. Only a subset of TFs is shown, see Supplemental Figure 9 for all TFs analyzed.

C. Distribution of CCAAT box quality scores under NFYB K562 peaks, called at 3 different *P*-values, a random genomic background sample set of 400k 500 bp regions and K562 FAIRE-Seq regions.

Figure 5 (Continued)



new type of "pioneer" TF that retains histone-like features, while possessing high sequence-specificity with a remarkable ability to access its motif irrespective of the chromatin state.

NF-Y functions with different TFs based on genomic context, and the prevalence of an association with FOS

Given the availability of 78 ChIP-Seq datasets in K562 for chromatin associated factors involved in diverse functions, we explored their combinatorial genomic interactions with NF-Y and focused on three classes of NF-Y bound sites – promoters, enhancers and LTR/non-modified-chromatin. We statistically tested for co-association between NF-Y and individual factors and found a high number, 44 at promoters and 50 at enhancers (at a *P*-value <= 10^{-10}; Supplemental Figure 10, A). We looked for combinatorial interactions beyond a one-way co-association with NFYB, by performing hierarchical clustering (Supplemental Figure 10, B) and describing the most common sets of 4-, 3-, and 2-way combinations of factors (Supplemental Figure 10, C). 2-way combinations were deemed relevant for enhancers due to the dearth of factors located in those regions. Figure 6 shows a summary of the factors present with NF-Y at promoters and enhancers.

Promoters: Hierarchical clustering revealed a distinct cluster that contained a core group of NF-Y co-associating factors: FOS, CHD2, TBP, RNA Pol II, CCNT2, HMGN3, MYC, and E2F4/6 (Supplemental Figure 10, B). The most common 4-way sets of TF combinations present at NFYB promoters variously included FOS, HMGN3, MYC, E2F4/6, HEY1 and CHD2, verifying this group as highly prevalent and extensively overlapping (Supplemental Figure 10, C). FOS was conspicuous, as it was the factor that most closely clustered with NFYA and extensively associates with NF-Y in multi-way overlaps (Supplemental Figure 10, C; note highlighted FOS entries). When we contrasted NFYB bound to non-bound promoters, FOS and

CHD2 were absent from the latter. FOS was highly specific for promoters, with 59% occupancy, but was largely absent at non-NFYB bound promoters (< 8%). HMGN3, MYC, E2F4/6, and HEY1 were common promoter bound TFs, generally enriched at, but not specific to, NF-Y bound promoters (Supplemental Figure 10, C).

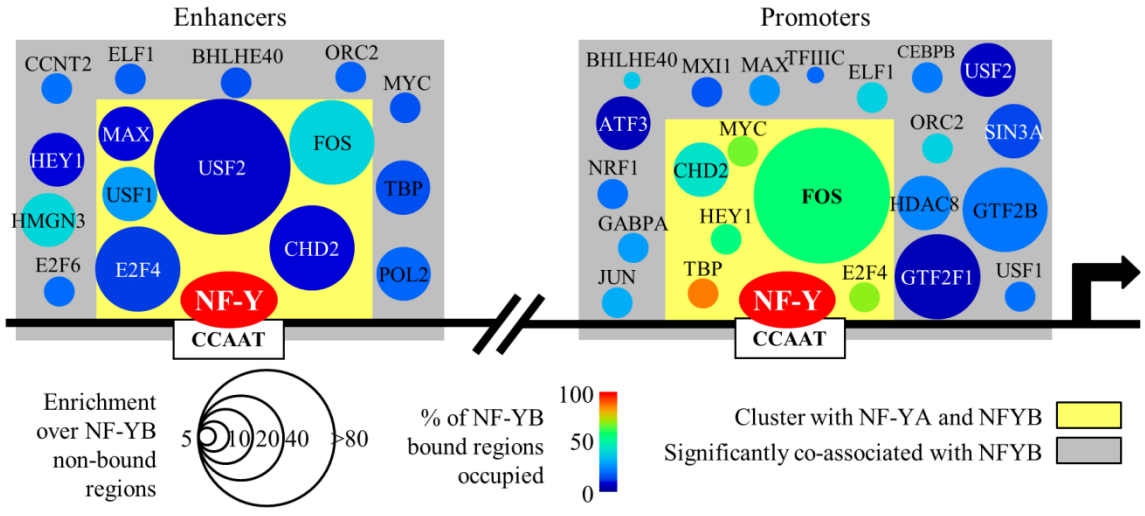
Enhancers: NF-Y formed a well-defined cluster consisting of FOS, USF1/2, MAX, CHD2 and E2F4 (Supplemental Figure 10, B), a slightly different grouping compared to promoters but very similar. When individual and 2-way combinations were assayed, FOS and USF1 were highly prevalent, being present at 39% and 27% of NF-Y enhancers, respectively, and were the most common 2-way overlap at 13% (Supplemental Figure 10, C).

Somewhat expectedly, E2Fs, represented here by E2F4 and E2F6, closely associated with NF-Y. Bioinformatic studies identified CCAAT boxes and E2F motifs as highly enriched in the promoters of genes overexpressed in tumors [133-135], and an enrichment of E2F sites in the proximity of CCAAT boxes in RefSeq promoters has been shown [51]. Importantly, apoptosis mediated by overexpression of NFYA was abolished in $E2F1^{-/-}$ cells [226]. Moreover, E2F4 is part of the DREAM complex [227, 228], which binds to the CDE DNA motif, and co-operates with the CCAAT box to negatively regulate expression of G2/M-specific genes during the cell cycle [229, 230]. CCAAT box and CDE containing G2/M genes were significantly overexpressed in a model of step-wise transformation of primary fibroblasts [125]. These data invite further analysis between the interaction of NF-Y and E2F sites (see below) particularly in cancer signature and cell cycle genes, and the integration with ChIP-Seq data of additional members of the E2F family and the DREAM complex.

Essentially all E box binding TFs present in ENCODE are statistically enriched at NF-Y locations, suggesting a pervasive partnership in *cis* between CCAAT and E boxes. MYC, which

Figure 6: NF-Y co-associates with many factors at promoters and enhancers

Illustration of the factors that significantly associate with NF-Y bound strong promoters and enhancers. Only those factors with greater than the median fold enrichment with respect to NFYB non-bound regions (enrichment indicated by circle size), greater than the median value of percent occupancy of NFYB bound regions (percentage occupied indicated by color), and that significantly co-associate with NF-Y (gray box; see Supplemental Figure 10, A). Factors enclosed within a yellow box are, additionally, the subset of factors that cluster with NFYA and NFYB (see Supplemental Figure 10, B). A black arrow indicates the start of a transcribed region. Two vertical slashes are used to represent being distal to a promoter area.



teams up with MAX to bind to the E box, is a good example. Interestingly, the number of MYC/NF-Y bound promoters exceeds those with MAX/NF-Y, suggesting that either MYC heterodimerizes with another E box binding partner, or that it binds in an E box independent manner, possibly directly to NF-Y [59, 231]. The interaction data detailed above go a long way to explain the importance of NF-Y for growth-regulating genes, and establish that NF-Y makes widespread partnership with a group of TFs - MYC, E2Fs, and FOS - that control cellular proliferation, and, when altered, can lead to cancer.

LTR/non-modified-chromatin: Given NF-Y's ability to bind to closed chromatin we wanted to know what factors could be partnering with NF-Y in these regions. We found extensive co-localization of NF-Y with only four factors, FOS and USF1, and to a lesser degree, USF2 and SP1 (Supplemental Figure 11, A). In addition, specific groupings of these factors occurred when we clustered the regions (clusters HL4, HL5, HL6, HL7, HL8, HL9, HL10; Supplemental Figure 11, A). As most of the non-modified-chromatin NF-Y sites were LTRs, we searched NF-Y-LTR sites located in non-modified-chromatin for known and novel DNA motifs, both in K562 and GM12878. We found that these regions are extensively de-enriched for all known DNA motifs that we assayed for when compared to all non-modified-chromatin residing LTRs (not shown), except for the CCAAT box and, in K562 only, the motif for KLF4 (*P*-value = 1.6×10^{-10}). A complementary *de novo* motif analysis found over-represented motifs that showed little resemblance to known elements, other than the expected CCAAT box and, confirming, a DNA motif similar to that of KLF4 (Supplemental Figure 11, B). ChIP-Seq data for KLF4 is not available in K562 or GM12878, however, its RNA and protein are detectable (not shown, [232]). The TF KLF4 is known to act as a transcriptional activator and repressor [233-238] and may be co-operating with NF-Y to repress LTR elements via a mechanism independent of H3K27me3.

The TFs USF1 and USF2 create barrier elements of acetylated chromatin in intergenic domains, thereby stopping the spread of heterochromatin [239-241]. The NF-Y-USF sites in non-modified-chromatin are unlikely to be canonical USF barrier elements, as these regions are not acetylated (Figure 3, A). However, this does not exclude a unique barrier element function of these NF-Y-USF sites functioning by a different mechanism.

The biological function of the LTR/non-modified-chromatin residing NF-Y sites is truly intriguing. Though we do not know their function, we do know that they are not acting as TSSs (no detectable RNA Pol II, RNA Pol III or polyA RNA, and very few dbTSS entries), DNA replication origins (ORC2 was not present), insulators (CTCF, RAD21 and SMC3 were not detectable), enhancers (no detectable H3K4me1), or canonical USF barrier elements (no detectable H3 acetylation). However, we do know that FOS, USF1, USF2, and SP1 were present, that these loci were depleted for known motifs, and that they were LTRs, which opens up possible avenues of biochemical and genetic experimentation.

It should also be noted that cluster HL2 (Supplemental Figure 11, A), though only representing 147 NF-Y sites, displayed specific enrichment for four members of the CTCF-cohesin insulator complex (CTCF, CTCFL, RAD21, and SMC3), in direct proximity with NF-Y. A similar small cluster was also observed in the PcG repressed class (not shown). There is no known precedent for this chromatin associated interaction in the literature and it raises the question as to what NF-Y-CTCF-cohesion complexes could be doing in the cell.

NF-Y extensively co-associates with FOS at loci lacking an AP-1 motif

The overlap of FOS and NF-Y at all chromatin states, cluster classes and genic features is striking. In fact, genome-wide, 45% of NFYB peaks directly overlapped a FOS peak, and 39% of FOS peaks directly overlapped an NFYB peak (Table 3). The correlation of occupancy between

NFYA and FOS at promoters and enhancers was high (Supplemental Figure 10, B), and even in the LTR/non-modified-chromatin class (clusters D and J) FOS signal was directly coincident with NF-Y at a large subset of NF-Y locations (clusters HL4, HL5, HL6, HL7, HL8, HL9; Supplemental Figure 11, A). The degree of correlation between NFYB and FOS ChIP-Seq reads at NFYB peaks was also exceptionally high (Pearson = 0.74), and only marginally lower than that observed between the NF-Y subunits (Pearson = 0.77) (Figure 7, A). Interestingly, a correlation (Pearson = 0.14) was not observed with JUN (Figure 7, A). These observations raise the question as to whether the NF-Y-FOS co-association involved JUN and the AP-1 motif, or if it could be mediated via NF-Y and the CCAAT box. NF-Y and FOS peaks were located just as close (< 50 bp) as that observed between the NF-Y subunits and between FOS and its dimerization partner JUN (Figure 7, B). Interestingly, most NF-Y-FOS sites lacked detectable AP-1 motifs, either by *de novo* discovery (Figure 7, C; right panels) or by searching for the canonical motif (Figure 7, C; left panel), with one very notable exception being NF-Y-FOS-LTR loci, which will be discussed below. FOS-JUN loci had an AP-1 motif positioned under FOS peak summits, as expected (Figure 7, C; left panel). Only about half of the FOS-NF-Y sites were co-occupied by JUN and another, undetermined, B-Zip partner(s) may mediate FOS binding at NF-Y sites, but this is at odds with the lack of a canonical AP-1 motif. A representative example of the interplay is shown in Figure 7, D.

There are no reports of NF-Y and FOS protein-protein interactions in the literature or in public databases that could explain this novel co-association (BIND, BioGRID, DIP, HPRD, IntAct, and MINT interaction databases via the APID portal [242]). Given that NF-Y and FOS have both been studied for decades and are ubiquitous factors, it is unlikely to have been missed by the scientific community, unless the NF-Y-FOS interaction was highly unique to a specific

Table 3: Overlap between FOS, JUN, MYC and NF-Y genomic binding site populations

Values represent the percentage of the peak population (left row) directly overlapping the peak population of a second factor (top column). All binding sites were called at a *P*-value <= 10^{-9}. All sites were: FOS (n = 14404); JUN (n = 18480); MYC (n = 13693); NFYA (n = 4726); NFYB (n = 12655).

% Overlap

	FOS	JUN	MYC	NFYA	NFYB
FOS		49.5	22.9	25.1	38.5
JUN	38.5		28.1	6.5	9.7
MYC	24.0	37.9		15.2	21.9
NFYA	76.9	25.7	44.6		94.5
NFYB	44.5	14.5	24.3	35.3	

Figure 7: NF-Y and FOS are closely co-associated at loci that lack JUN and the AP-1 motif

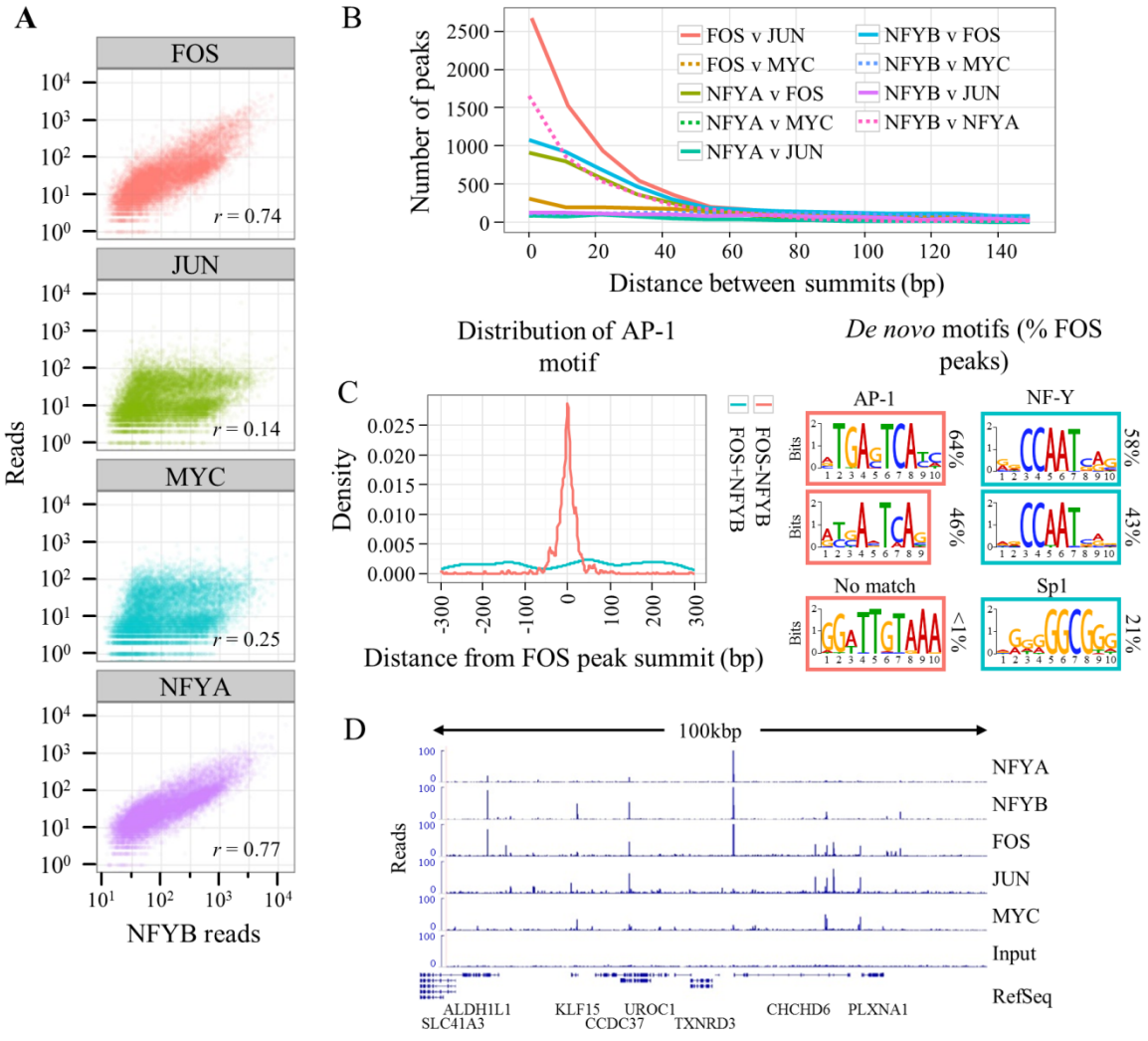
A. Correlation between ChIP-Seq read counts at NFYB peak summits, within a window of +/-500 bp, between NFYB and NFYA, FOS, JUN or MYC in K562 cells. NFYA and FOS were well correlated with NFYB, whereas JUN and MYC were not.

B. FOS-NF-Y peak summits were located just as closely as FOS-JUN and NFYA-NFYB summits. FOS, NFYA and NFYB ChIP-Seq peak summits were mapped to the nearest FOS, JUN, MYC or NFYA peak summit. The number of ChIP-Seq peaks at the indicated distance between adjacent peak summits is plotted. All peaks were called at a 10^{-9} *P*-value threshold in K562, where summit was the local maxima in read counts.

C. The AP-1 motif was not present under FOS sites that overlap NF-Y. The top 1000 K562 FOS ChIP-Seq sites, as ranked by site *P*-value, that directly overlap an NFYB site ("FOS+NFYB") and the top 1000 that do not overlap an NFYB site (10^{-5} *P*-value site list, "FOS-NFYB") were assayed for the distribution of the AP-1 motif in relation to the FOS peak summit centered at 0 bp. Plotted is the Gaussian kernel density estimate of the AP-1 motif using a bandwidth of 0.5 of the standard deviation of the smoothing kernel. The top 3 motifs discovered *de novo* from each FOS peak set, as above, are depicted with the percentage of FOS peaks containing a match to that motif indicated.

D. FOS associated at the same genomic loci as NF-Y, usually in the absence of JUN. Representative view of a locus on chromosome 3 of the K562 ChIP-Seq read counts from NFYA, NFYB, FOS, JUN, and MYC ChIPs, with an input control.

Figure 7 (Continued)



cellular compartment not commonly analyzed for direct protein-protein interactions (*i.e.* chromatin) and/or was cell type specific. Related to our finding, an unexpected result emerged recently from ChIP-Seq analysis of JNKs (Jun N-terminal Kinase). Rather than the predicted AP-1 motif, the only recognizable motif in JNK sites was CCAAT, and indeed NF-Y was shown to be necessary for JNK-DNA association [243]. In light of our data, one possibility is that FOS, directly or indirectly and possibly with JNK, binds to NF-Y, though only on chromatin and/or in specific cell types, forming a novel NF-Y/FOS/JNK complex that does not require the AP-1 motif or JUN, and recruits members of the MAPK family to CCAAT box containing regulatory regions.

NF-Y sites contain positionally biased TFs

To investigate a possible distance bias between NF-Y and TFs on chromatin, we plotted the distribution of the relative position of TATA, E box, E2F and AP-1 motif instances (termed "predicted") at NFYB peaks, in relation to the position of the best scoring CCAAT box (1st C is position 1), while maintaining strandedness. We then plotted the subset of motif instances (termed "verified") that were actually occupied by the TF of interest by ChIP-Seq (Figure 8).

First, we checked the NF-Y-FOS connection (Figure 8, A) and, remarkably, there was a clear AP-1 motif 10-11 bp upstream of 5'-CCAAT-3', which corresponded to FOS ChIP-Seq peaks. However, this positioning was only found in NF-Y-bound LTR sequences, as NF-Y-FOS sites, in general, did not contain an AP-1 motif (Figure 7, C). This finding has no precedent in promoter studies and it is even more surprising as it involves repetitive sequences. The functional nature of this interaction remains to be determined and the precise positioning and distribution of the interplay may be an indication that the two TFs cooperate to keep LTRs repressed, presumably with an unknown B-Zip partner. The TATA, E box and E2F motifs were

also located at remarkably discrete, highly biased positions in a CCAAT orientation specific manner: TATA at +50 bp; the E-box at -10/-11 bp; and the E2F motif at +6/+7, +31, +55 and +72 bp (Figure 8, B). The position of the TATA box was maintained in TBP peak locations at NF-Y loci, albeit with a somewhat reduced frequency. The stereo positioning of the E box location was only maintained when MAX or USF1, but not MYC, loci were considered, suggesting that MYC, when associating with NF-Y, was either not positioned, or did not bind DNA directly. The E2F motif was unusual in that multiple stereo alignments were present and only one, the closest to CCAAT, was maintained at E2F6, but not at E2F4 occupied sites.

The USF1 finding was particularly interesting, as USF1 was one of the few factors that partnered with NF-Y in non-modified-chromatin domains (Supplemental Figure 11, A) and also had the ability to recognize its motif within a repressive nucleosomal structure (Figure 5, B). NF-Y and USF1 may cooperate to penetrate repressive, non-modified chromatin domains containing a CCAAT box through a mechanism that requires precise motif positioning with an E box.

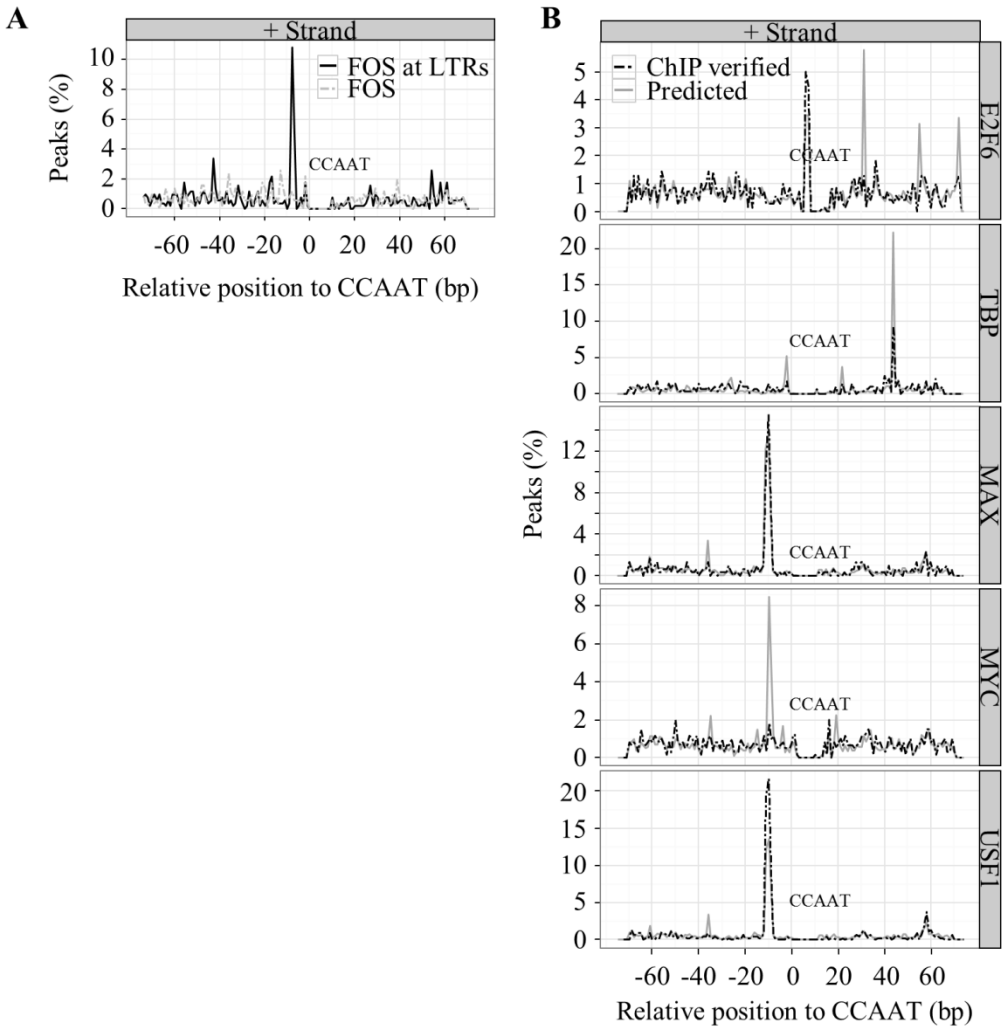
Overall, our data indicate the presence of precise positional bias between NF-Y and some of its most common TF partners, notably, those that play crucial roles in the control of cell proliferation, cell-cycle and metabolism genes. In the vast majority of NF-Y bound promoters, it is known that NF-Y synergizes with neighboring TFs and it appears to be more of a promoter organizer and facilitator of transcription, than a strong activator *per se*. There are three examples in which cooperativity with NF-Y is mediated by precise spacing: the MHC Class II promoters; NF-Y/ATF6 sites in ER stress response promoters [244]; and the multiple CCAAT boxes in G2/M promoters [245]. Several studies reported overlaps between TFs at a genomic level, but, to the best of our knowledge, the mutual TF interplays were never detailed with such a high degree of precision. We establish here that the quality of ChIP-Seq peaks in ENCODE allows one to

Figure 8: Motif pairings with the CCAAT box are stereo-positioned

A. The AP-1 motif was only stereo-positioned, with respect to the CCAAT box, at LTR elements. Similar to B, except FOS peaks directly overlapping LTRs were considered.

B. The percentage of NFYB peaks that have a TATA-box (TBP), E-box (MYC, MAX, USF1), and E2F motif (E2F6) at the specified distance from the best scoring CCAAT box centered at 0 bp of NFYB sites, showing highly precise stereo-positioning of DNA motifs. All NFYB peaks were categorized as "predicted", while those NFYB peaks overlapping the respective ChIP-Seq peaks of the other TF were categorized as "verified". Only the top 500 peaks in each category were plotted. The negative strand plots were near identical mirror images of the positive strand plots and are not shown.

Figure 8 (Continued)



study the precise genomic architectural rules of TF interactions on DNA *in vivo* and within specific genomic contexts.

Conclusions

Our comprehensive analysis of NF-Y confirms many functions including its prevalence at proximal promoters, particularly those of growth controlling genes, at a much higher degree of precision and completion. More interestingly, our analyses uncover several novel and unexpected aspects of NF-Y function. In particular, NF-Y binds asymmetrically at its target sites, plays an important role at many tissue-specific enhancers, is capable of binding "closed" chromatin including at LTRs, co-associates pervasively with FOS, but not other AP-1 factors, and displays precise stereo positioning with a restricted group of TFs involved in cellular proliferation. Lastly, we note that comprehensive bioinformatic analyses of the type performed here have been done on relatively few TFs. Similar analyses on other TFs whose target sites have been or will be defined by ChIP-seq are likely to uncover new functional properties and relationships of biological relevance, in particular to reconstruct regulatory element architectures.

METHODS

Cell culture

K562, GM12878 and HeLaS3 were grown as per standard ENCODE protocols ([183-185]; Appendix C) and a detailed protocol is available at: http://genome.ucsc.edu/ENCODE/.

Chromatin immuno-precipitation

Cells were fixed by the addition of 1% v/v formaldehyde at room temperature for 10 min and quenched with 0.2 M glycine. Cell pellets were washed twice with PBS, lysed in CLB (25 mM HEPES pH7.8, 1.5 mM MgCl2, 10 mM KCl, 0.1% NP-40) with 1 mM DTT added just before use, and nuclei pelleted by centrifugation at 12 kG. Crude nuclei were then lysed in NLB (50 mM HEPES pH7.9, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na deoxycholate) with 1% SDS. Nuclear extract was fragmented using a Branson 450 sonicator and/or Misonix 3000 to yield chromatin of a suitable length for immuno-precipitation. Chromatin was spun at 12 kG for 10 min to remove precipitates and the supernatant was flash frozen and stored at -80^{o}C until use. Chromatin from 2x10^{7}cell equivalents were used per ChIP. Chromatin was diluted 10x in NLB and pre-cleared with Protein A-Sepharose beads for 2 hr at 4^{o}C. The supernatant was incubated with 5-10 µg of the appropriate antibody overnight at 4^{o}C. Protein A-Sepharose beads were added for 2 hr then washed as follows: 2x NLB with 0.1% SDS; 2x NLB with 0.1% SDS and 640 mM NaCl; 2x WB (20 mM Tris-HCL pH8.0, 250 mM LiCl, 1 mM EDTA pH8.0, 0.5% NP-40, 0.5% Na deoxycholate); finally, 2x TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH8.0). Bound protein was eluted in TE with 1% SDS for 15 min at > 60^{o}C. Protease inhibitor cocktail and 1 mM PMSF were added to all buffers just before use. Elutions were digested with 20 µL of 20 mg/mL pronase for 2 hr at 42^{o}C and crosslinks reversed by overnight incubation at 65^{o}C.

DNA was purified using phenol:chloroform:isoamyl alcohol extraction utilizing high density MaXtract tubes (Qiagen, USA) as per manufacturer's protocol. Aqueous phase DNA was precipitated by the addition of 200 mM NaCl, 500 mM NaAc, 80 μg/mL glycogen, and 2 volumes of ethanol, while incubating for > 1 hr at -80^{o}C, followed by centrifugation at 12 kG at 4^{o}C for 20 min. The precipitate was washed in 95% ethanol, resuspended in TE and stored at -20^{o}C until needed.

ChIP-Sequencing

ChIP DNA (2 biological replicates) prepared as above, and immuno-precipitated with anti-NFYB or anti-NFYA antibody (Mantovani, R.), and input DNA (3 biological replicates) were end repaired with calf intestinal alkaline phosphatase (New England Biolabs, USA) and sent for sequencing to the Stanford Center for Genomics and Personalized Medicine or the Department of Molecular, Cellular and Developmental Biology at Yale University. Library preparation and Illumina sequencing were carried out as per manufacturer protocols and ENCODE standards ([183-185], http://genome.ucsc.edu/ENCODE/). Datasets for NF-Y are deposited at UCSC as per ENCODE guidelines. Sequence reads (~28 nucleotides) were mapped to the *H. sapiens* genome (hg18) using *Bowtie* [247], allowing <= 2 mismatches per read and reads with > 10 reportable alignments were discarded. Binding sites were called using *MACS* v1.4 [248] at a *P*-value threshold of 10^{-9} (unless otherwise noted) on non-redundant reads using input to control for local genomic biases. See Supplemental Data for genomic coordinates.

ENCODE Consortium data Sets

ChIP-sequencing datasets for histone PTMs, TFs, and RNA-Seq for K562 and/or HeLaS3 cell lines were provided by the ENCODE Consortium via the UCSC Genome Browser and are

described there and ([183-185]; http://genome.ucsc.edu/ENCODE/; Appendix C). ChIP-Seq datasets were mapped and peaks called as described above. RNA-Seq data was prepared by Helicos (Cambridge, MA) as long (> 200 nt), poly-A enriched, cytosolic RNA and mapped using *rSeq* [249, 250]. Chromatin state maps were also from ENCODE and are described at http://genome.ucsc.edu/ENCODE/ and in [205]. The chromatin state "heterochromatin" was renamed to "non-modified-chromatin".

ChIP-QPCR

Primer pairs were designed to amplify regions within 150 bp of the summit of ChIP-Seq peaks. *Batch primer3* was used for primer design using default parameters [251]. All primers were tested for unique hits to the *H. sapiens* genome using UCSC *In-Silico PCR* (Jim Kent, UCSC) and by dissociation curve analysis. See supplemental data for primer sequences. QPCR was performed on an Applied Biosystems 7900FAST instrument (kindly provided by the HMS ICCB) on ChIP and input DNA (prepared as above, except Qiagen columns were used for purification), using 2x Taq Mix (see [252]), except 250 nM EVA green (Biotium, USA) replaced SYBR green. PCR program was: 95°C 10 min, followed by 40 cycles of 95°C for 5 sec, 60°C for 30 sec. ChIP-QPCR values are represented as fold enrichment over an NF-Y non-bound control region as previously described [253]. Error bars are based on the standard deviation observed in 2-4 biological replicates run in QPCR triplicates.

Lentiviral knockdown and gene expression arrays

Scrambled control (shSCM) and NFYA pLKO.1-shRNAs were designed by Sigma-Aldrich. The puromycin resistance cassette was replaced with an EGFP cassette. Viral production and transduction were carried out as previously described [254]. HeLaS3 cells were transduced with

shSCM or shNFYA viral supernatants, in triplicate, and cells collected after 48 hr of incubation. The distribution of cells within the cell-cycle was checked via FACS as previously described [254]. Knockdown efficiency was assayed by PCR on cDNA to known NFYA target genes and by Western Blot on whole cell protein extracts using anti-NFYA, and anti-Actin antibodies. For arrays, total RNA was prepared by Trizol extraction and Qiagen RNeasy kit purification, converted to biotinylated aRNA and hybridized to U133 Plus 2.0 GeneChip expression arrays using the 3' IVT Express Kit (Affymetrix, USA) following the manufacturer's protocol. Array hybridization was carried out by the Molecular Genetics Core Facility at Children's Hospital Boston supported by NIH-P50-NS40828 and NIH-P30-HD18655. Arrays were RMA normalized [255], gene expression levels calculated, differential expression determined and probes annotated using the following *R* packages from the Bioconductor project: *affy* [256], *limma* [257], and *annaffy* [258].

Annotation of peaks to gene features, GO analysis (*GREAT*/*IPA*)

Genomic locations of peak summits were submitted to the annotation tool *GREAT* [259] using the following parameters: whole genome background set, basal plus extension, proximal upstream = 5 kbp, proximal downstream = 1 kbp, distal = 1 mbp; or whole genome background set, basal, proximal upstream = 5 kbp, proximal downstream = 1 kbp. Molecular signaling pathways were visualized using *IPA* (Ingenuity Systems, USA, http://www.ingenuity.com) were a gray shaded node represents a K562 NFYB binding site located within the putative regulatory region, as defined by *GREAT*, of that molecule. Peak summits were annotated to genomic features using in-house scripts.

De novo motif discovery

DNA sequences corresponding to the regions +/-50 bp of ChIP-Seq peak summits for NFYB in K562 were gathered using *BEDTools* [260] and repeat masked using *RepeatMasker* with –q option [261]. Sequences were searched for *de novo* motifs using parallel *MEME* [262] using the following parameters: zoops, revcomp, minw [range 5-40], maxw [range 10-60]. Background letter frequencies were based on a 5-order Markov model derived from hg18 repeat masked sequences -350:+100 bp about RefSeq TSSs, the non-modified-chromatin or PcG repressed chromatin state maps. For NFYB, a second background model using FAIRE-Seq regions for K562 was also carried out and produced a similar motif (not shown). *Tomtom* [263] was used to compare *de novo* motifs to known motifs in the JASPAR_CORE_2009 database [264]. For TFs other than NF-Y, motifs were discovered as above except the top 1000 ChIP-Seq peaks of each factor were used and the top motif was selected, except for FOS, which produced the CCAAT box and was substituted for the motif derived from JUN ChIP-Seq. Similarly, for motifs in the non-modified-chromatin state *MEME* was run using a motif width range of 10-15, on all K562 NFYB peaks residing within the non-modified-chromatin state, on non-masked sequences, with a background set derived from the entire non-modified-chromatin state of K562 [205].

Motif stereo positioning

NFYB summit locations from K562 were scanned using *Pscan* [265], for matches to the NF-Y matrix in the JASPAR_CORE_2009 database (MA0060.1) [264]. For NF-Y loci with the best matrix match on the positive strand, the first C (of CCAAT) of the best match was set to 0 bp. Genomic sequences +/-75 bp from the motifs were retrieved and scanned with *Pscan* using the collection of matrices in the JASPAR_CORE_2009 database [264]. For each JASPAR matrix, only regions containing a best matrix match > 0.8, computed as described in [265], were

considered for further analyses. This population was deemed “predicted”. For each "predicted” population, the subpopulation of regions that overlapped the relevant TF ChIP-Seq peak dataset were deemed “ChIP verified”. The frequency of the best motif occurrences for each motif matrix at each bp from the CCAAT box was determined for each population and plotted as the percentage of motifs.

Histone PTMs and chromatin associated factor clustering

Density arrays at NFYB peak summits spanning either +/-5 kbp or +/-500 bp representing ChIP-Seq read counts of histone PTMs (H3K79me2, H3K4me3, H3K27me3, H3K4me1, H4K20me1, H3K36me3, H3K4me2, H3K9ac, H3K9me1, H3K27ac), NFYA , NFYB, and RNA Pol II or NFYA, NFYB and 78 chromatin associated factors (see Supplemental Figure 11, A for the full list) with appropriate input samples, were computed using the ranked based correlation method of *seqMINER* v1.2 [266]. Clustering was carried out using the following parameters: T = 10, K-means. Clusters from 3-50 were considered. Non-normalized raw read counts are depicted in Figure 3, A and Supplemental Figure 11, A.

Mapping to repeats

Bowtie [247] was used to map the NFYB and input ChIP-Seq datasets to a reference “genome” composed of Repbase v15.08 [267] entries - simple.ref, humrep.ref, humsub.ref and pseudo.ref – allowing <= 2 mismatches per read and reads with > 1 alignment had one alignment selected at random. Read counts for each Repbase entry were tallied and the ChIP:input ratio calculated. Individual consensus sequences of repeat elements were scored for the presence or absence of the CCAAT box using the matrix derived from this paper and *FIMO* [268] with matches called at a significance *P*-value threshold of 10^{-4}.

Hierarchical clustering of binding events to promoters and enhancers

Regions considered promoters and enhancers were taken from the K562 chromatin state maps of [205]. Regions were considered "bound" if an NFYB peak summit directly overlapped the region. Regions were considered "non-bound" if no NFYB peak overlapped the region of interest and the region had < 1.5x the normalized fold-over-input ChIP-Seq enrichment for NFYB. At all NFYB-bound or NFYB non-bound regions, chromatin associated factors were scored as present (1) or absent (0) based on directly overlapping peak summits. The *R* packages *pvclust* [269] and *snow* (http://cran.r-project.org/web/packages/snow/) were used to cluster the matrices and to calculate *P*-values using multiscale bootstrap resampling. Parameters were: method.dist="binary", method.hclust="ward", nboot=10000. Red and blue numbers in plots indicate the approximately unbiased (AU) *P*-values and the bootstrap probability (BP), respectively, as detailed in [269].

Statistical test of TF co-association with NF-Y

NFYB bound regions were as above. Promoters or enhancers occupied by NFYB we assessed for individual co-occupancy of 78 transcriptional regulators. The significance of the overlap was tested by a 2x2 contingency table using Fisher's exact test and calculated using [270].

Western blot and RT-PCR

As described in [254]. Briefly, total cell protein extracts were prepared by resuspending the cell pellets from shSCM or shNFYA infected cells in lysis buffer (50 mM Tris–HCl pH 8.0, 120 mM NaCl, 0.5% NP-40, 1 mM EDTA, protease and phosphatase inhibitors). An equivalent amount of cellular extracts were resolved by SDS–PAGE, electro transferred to PVDF membrane, and

immuno-blotted with the following antibodies at 1:1000 in TBS containing 1 mg/ml BSA: anti-NFYA (sc-10779), and anti-Actin (sc-1616) from Santa Cruz Biotechnology, USA.

ACKNOWLEDGEMENTS

We thank the members of the Snyder and Gernstein labs; the ENCODE Consortium for support; Hannah Monahan for preparing sequencing libraries; WQCG and RITG for technical support and access to computing facilities; Koon Ho Wong, Nathan Lamarre-Vincent, and Rajani Gudipatti for helping with figures and reading the manuscript; Joseph Geisberg for helpful advice; Benoit Miotto for performing ORC2 ChIP-Seq. This work was supported by grants to K.S. from the National Institutes of Health (GM30186), to R.M. from Lombardy Region (NEPENTE) and AIRC, and to C.I. from AIRC (MFAG 6192).

CHAPTER 3: Genome-wide dynamics of STAT3, FOS and *cis*-regulatory element usage during inflammatory-mediated oncogenic transformation

AUTHOR CONTRIBUTIONS

Joseph D. Fleming, Marianne Lindahl-Allen, Paul Giresi, Elsa Beyer, Asaf Rotem, Jason Lieb and Kevin Struhl.

J.F. wrote the chapter. K.S., J.F., and M.L.A. conceived the project. M.L.A., J.F., P.G. and E.B. designed experiments. M.L.A. and J.F. contributed equally, with assistance from E.B., to performing biological experiments involving: tissue culture, chromatin preparation, and ChIP. P.G. and M.L.A performed FAIRE-Seq biological experiments. E.B. performed siRNA biological experiments. J.F. performed bioinformatics analysis of ChIP-Seq, RNA-Seq and gene expression microarray datasets. P.G. performed bioinformatics analysis of FAIRE-Seq data. J.F. and P.G. performed motif analyses. J.F. and A.R. performed DIC imagery.

J.F. and M.L.A. contributed equally to this work.

ABSTRACT

Oncogenic transformation can be triggered by inflammatory signaling pathways and inflammation has been linked to diverse types of cancer. Here we use an inducible isogenic model of inflammatory oncogenic transformation to track the genomic changes in the inflammatory transcription factor (TF) STAT3, a partnering TF, FOS, and genome-wide *cis*-regulatory element (CRE) usage during a time course of transformation. STAT3 genomic binding is highly induced during transformation, linked to preexisting FOS bound sites, but does not create new CREs, likely due to STAT3s inability to bind to DNA motifs outside of open chromatin. Surprisingly, CRE usage is highly stable during transformation, a process with large scale phenotypic and gene expression changes. STAT3 regulated AP-1 factors are deregulated during transformation and may regulate the embryonic-like and bone-metastasis phenotypes commonly observed in cancer and breast cancer, respectively. Using siRNA we found that direct or indirect regulation by STAT3 accounts for 1/3rd of the gene expression program during transformation and that, a second inflammatory TF, NFκB likely controls the rest. We also highlight putative roles for circadian rhythm related TFs in transformation and the likely inhibitory role of TSC22D3 acting on an epigenetic switch that initiates and maintains the transformed phenotype. This study is one of the first to track a critical determinant TF and CRE usage during a cellular phenotypic change, and the first for transformation. The genome maps of STAT3, FOS and CREs catalogued here will be a valuable asset to the community for future studies of inflammation-mediated oncogenic transformation.

INTRODUCTION

Oncogenic transformation is the phenotypic process a normal cell undergoes to become cancerous. It has long been known to be driven mostly by the perturbation of kinases (*e.g.* SRC, RAS, BCR-ABL, ERBB2), which drive the inappropriate activity of downstream TFs. These factors mediate transcriptional changes within the cell which ultimately mediate the phenotypic qualities observed in the transformed cancerous cell type such as invasion, metastasis, loss of contact mediated growth inhibition, uncontrolled proliferation and formation of tumors in nude mice. This process has been thoroughly studied over the decades and the signal transducer and activator of transcription 3 (STAT3) has been found to be a central mediator of the transcriptional changes in many different types of cancers: breast cancer [175], pancreatic cancer [176, 177], prostate cancer [178], liver cancer [179], melanoma [180], among others (for a review see [181, 182]). STAT3 directly regulates the genes involved in cell proliferation, cell cycle control, metastasis, apoptosis, angiogenesis, and embryogenesis, and as such, is a key factor in the process of transformation.

STAT3 is a DNA binding TF [138, 139], that is part of a larger family consisting of 7 members. STAT3 contains an SH2 domain and is phosphorylated at tyrosine 705 (Tyr^{705}) and serine 727 (Ser^{727}) in response to many cytokines and growth factors [161]. Tyr^{705} phosphorylation is critical for the dimerization, nuclear localization, and gene activation by STAT3, while Ser^{727} phosphorylation plays a more minor role in modulating STAT3 activity [164, 271, 272]. SRC directly phosphorylates STAT3 *in vitro* [273] and co-immuno-precipitates with STAT3 from cellular extracts [274-276], and is itself an oncogenic kinase. Many primary tumors [277-283], tumor derived cell lines, and v-Src or ABL kinase transformed cell types contain constitutively activated STAT3 [162, 273, 281, 284, 285]. Of primary breast cancer

specimens, 30-60% contain Tyr^{705} phosphorylated STAT3 [286-289]. Breast cancer cell lines also have elevated levels of Tyr^{705} STAT3 and inhibition of STAT3 activity impairs proliferation and induces apoptosis [175, 290, 291]. Pertinently, overexpression of a highly active form of STAT3 (STAT3-C) in *H. sapiens* mammary epithelial cell lines [292] or fibroblasts [293] has been shown to be sufficient to induce transformation. This was shown by anchorage independent growth in soft agar and tumor development in nude mice, indicating that the transcriptional output of activated STAT3 is all that is required, at least in breast cancer epithelial cells and fibroblasts, for a transformed phenotype. In addition, inactivation of STAT3, by dominant negative constructs [284], has been shown to inhibit Src induced transformation and reduce tumor size/burden in murine models. However, STAT3 does not act alone, as we have previously shown that inflammation-mediated oncogenic transformation of MCF10A-ER-Src cells is dependent upon a second inflammatory transcription factor, NFκB, as well as STAT3 [289].

In previous work in the Struhl Lab, an inflammatory cancer gene signature was found on the basis of the identification of genes that were differentially expressed in 2 isogenic models of oncogenic transformation [294]. One model involved non-transformed mammary epithelial cells (MCF-10A; [295]) containing ER-Src, a derivative of the Src kinase oncoprotein that is fused to the ligand-binding domain of the estrogen receptor (ER) [296]. Treatment of such cells with tamoxifen (TAM) rapidly induces Src kinase activity which activates an epigenetic switch involving STAT3, NFκB and downstream effectors [289, 297], that initiates and maintains transformation. Inactivation of STAT3 or NFκB prevents ER-Src induced transformation. Upon Src activation, phenotypic transformation is observed within 24 to 36 hours ([289, 298]; Figure 15), thereby making it possible to kinetically follow the transition between non-transformed and transformed isogenic cells.

Here we use this model to further explore the transcriptional regulatory network involved in inflammation-mediated transformation, specifically focusing on STAT3. To gain a comprehensive understanding of the molecular events that occur upon Src activation in MCF10A-ER-Src cells, we performed ChIP-Seq, FAIRE-Seq, and gene expression studies at several time points post Src activation, followed by genomic analyses.

RESULTS

Study design

The cell line MCF10A-ER-Src provides an isogenic model for the study of the time-dependent molecular dynamics that occur during oncogenic transformation. Upon treatment with tamoxifen, cells undergo a rapid, and epigenetically maintainable, phenotypic and morphological differentiation from an immortal non-transformed cellular state to a transformed cancerous cellular state within 24-36 hr [289, 294]. To explore the relationship between a critical determinant TF (STAT3), a partnering TF (FOS), *cis*-regulatory-element usage and changes in gene expression during this pathologic cellular differentiation process, we sampled the differentiation pathway by ChIP-Seq, gene expression microarrays and FAIRE-Seq (Figure 9). Focusing on one of the main TFs mediating this process, we also performed siRNA knockdown of STAT3, prior to transformation, to determine its dependent and independent contribution to the transcriptional signal.

Biological functions of chromatin bound STAT3

The activation of STAT3 is a hallmark of oncogenic transformation in MCF10A-ER-Src cells (Figure 10, A). It has been known since the late 90s that overexpression of constitutively active STAT3 can lead to transformation of a non-transformed, though immortal, cell line and that many different types of cancers have constitutively active STAT3. To identify where STAT3 was binding in the genome, STAT3 ChIP-Seq was performed at 0 hr and 36 hr post EtOH treatment, and 4 hr, 12 hr and 36 hr post ER-Src activation (TAM treatment). We identified 78,293 non-redundant sites of STAT3 occupancy within MCF10A-ER-Src cells and 15,098 genes which contained at least one STAT3 binding site within their putative regulatory domain

Figure 9: Experimental study design

Illustration depicting the basic outline of the experimental design showing tamoxifen or ethanol treatment of MCF10A-ER-Src cells and the harvesting of chromatin or RNA for ChIPs, FAIRE and gene expression analyses.

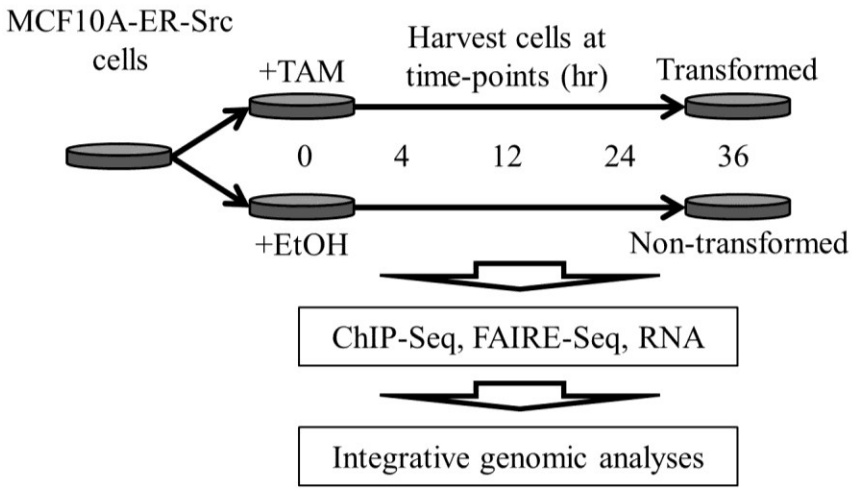
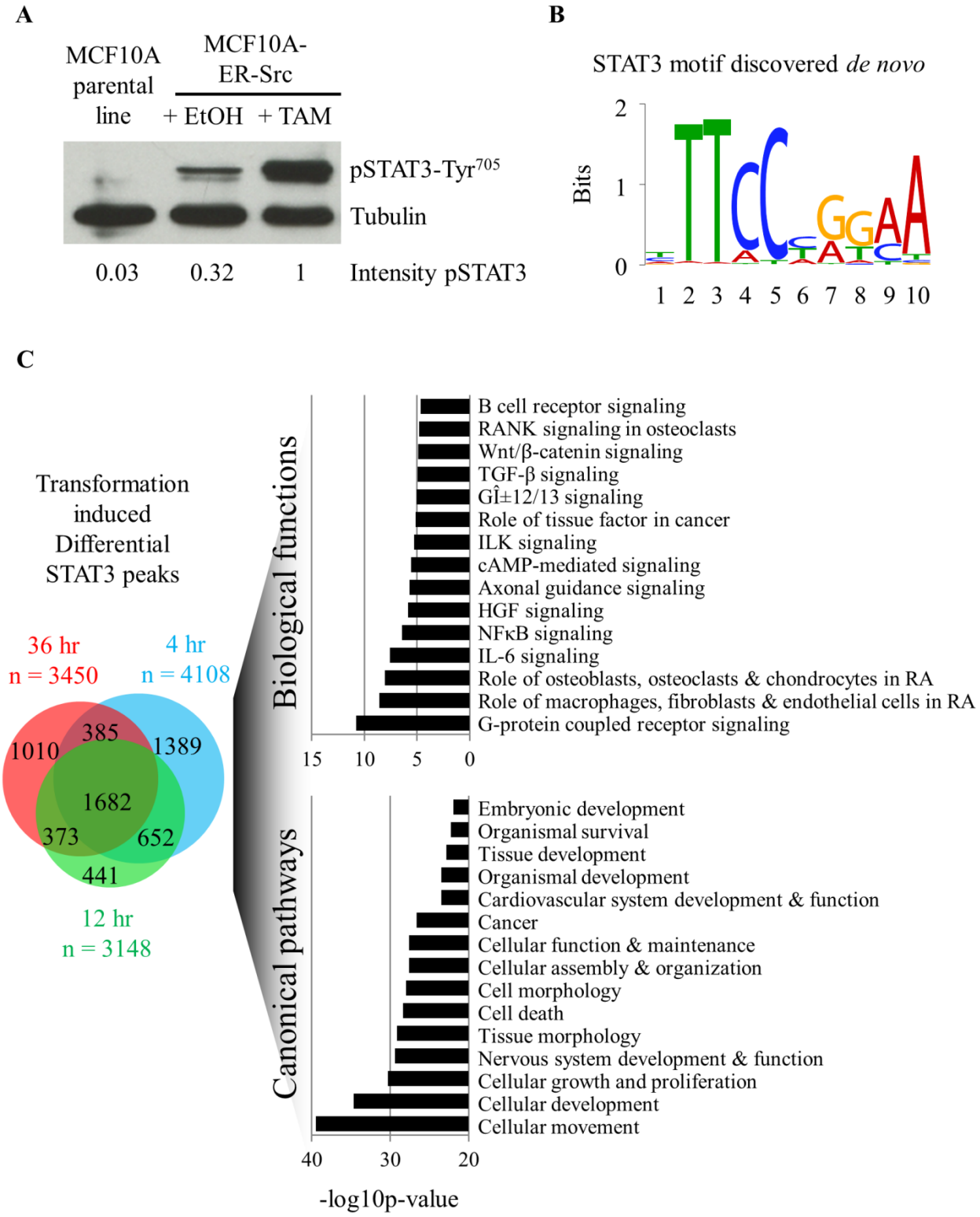


Figure 10: STAT3 during transformation and the GO terms associated with differential binding

A. Western blot of Tyr^{705} phosphorylated STAT3 in MCF10A parental cells and MCF10A-ER-Src cells treated with EtOH or TAM for 24 hr. Tubulin was used as loading control. Intensity values were normalized to tubulin and expressed as relative to the parental cell line.
B. The STAT3 DNA binding motif derived from ChIP-Seq identified STAT3 binding sites.
C. Gene ontology terms significantly bound by transformation dependent differential STAT3 sites.

Figure 10 (Continued)



(as defined by *GREAT*, see Methods). Only 3.3% of STAT3 sites (n = 2,629) were not associated with a RefSeq gene. *De novo* motif analysis of the top 10,000 of these sites revealed the canonical STAT3 motif (Figure 10 , B). The location of differential STAT3 binding sites (see below) in the genome revealed that many genes linked to inflammation were significantly over-represented ("*IL-6 signaling*", "*NFκB signaling*", and "*TGF-beta signaling*") as was expected (Figure 10, C). STAT3 was also located within the regulatory regions of genes from ontologies such as "*cellular movement*", "*growth and proliferation*", "*cell death*" and "*embryonic development*" (Figure 10, C), key processes all linked to cancer, and, as such, confirming the central role of STAT3 in transformation.

Transformation increased STAT3 DNA binding activity

STAT3 RNA levels were increased by ~50% during transformation, and STAT3 activity, as measured by Tyr^{705} phosphorylation, was increased ~3 fold. The increase in STAT3 activity was reflected in an increase in average ChIP signal at STAT3 bound loci (Figure 11, A-E) and the induction of new transformation dependent STAT3 sites. 26,783 STAT3 binding sites (cut off *P*-value <= 10^{-9}) were discovered in non-transformed MCF10A-ER-Src cells, probably representing a basal level of ER-Src signaling (Figure 10, A). In MCF10A-ER-Src cells undergoing Src induced transformation, STAT3 bound sites at 4 hr, 12 hr and 36 hr post induction increased to 77,262, 67,015, and 74,584 sites, respectively. While the increase in transformation induced STAT3 sites is impressive, only 5.4% (n = 4,157), 4.8% (n = 3,200) and 4.7% (n = 3510) of STAT3 sites at these time points, respectively, were not detected in control cells and induced greater than ~5 fold (mean + 1 x standard deviation) in ChIP signal intensity. These STAT3 sites will hitherto be referred to as differential (similarly for FOS bound sites, see below).

Figure 11: Genome view of STAT3 binding during transformation

ChIP-Seq read counts at the *ARNTL2* (A), *SOCS3* (B), *TSC22D3* (C), *NFKB1* (D) and *IL6* (E) loci during transformation of MCF10A-ER-Src cells. "4 hr", "12 hr" and "36 hr" indicate time post ER-Src induction by TAM treatment. EtOH and TAM input samples are single replicates, all others are of 2 biological replicates combined. ChIP-Seq and FAIRE-Seq elements deemed to be transformation dependent differential ("Diff.") sites and all sites derived from TAM and EtOH treated samples are shown. Red arrows highlight differential ("Diff.") ChIP-Seq sites.

Figure 11 (Continued)

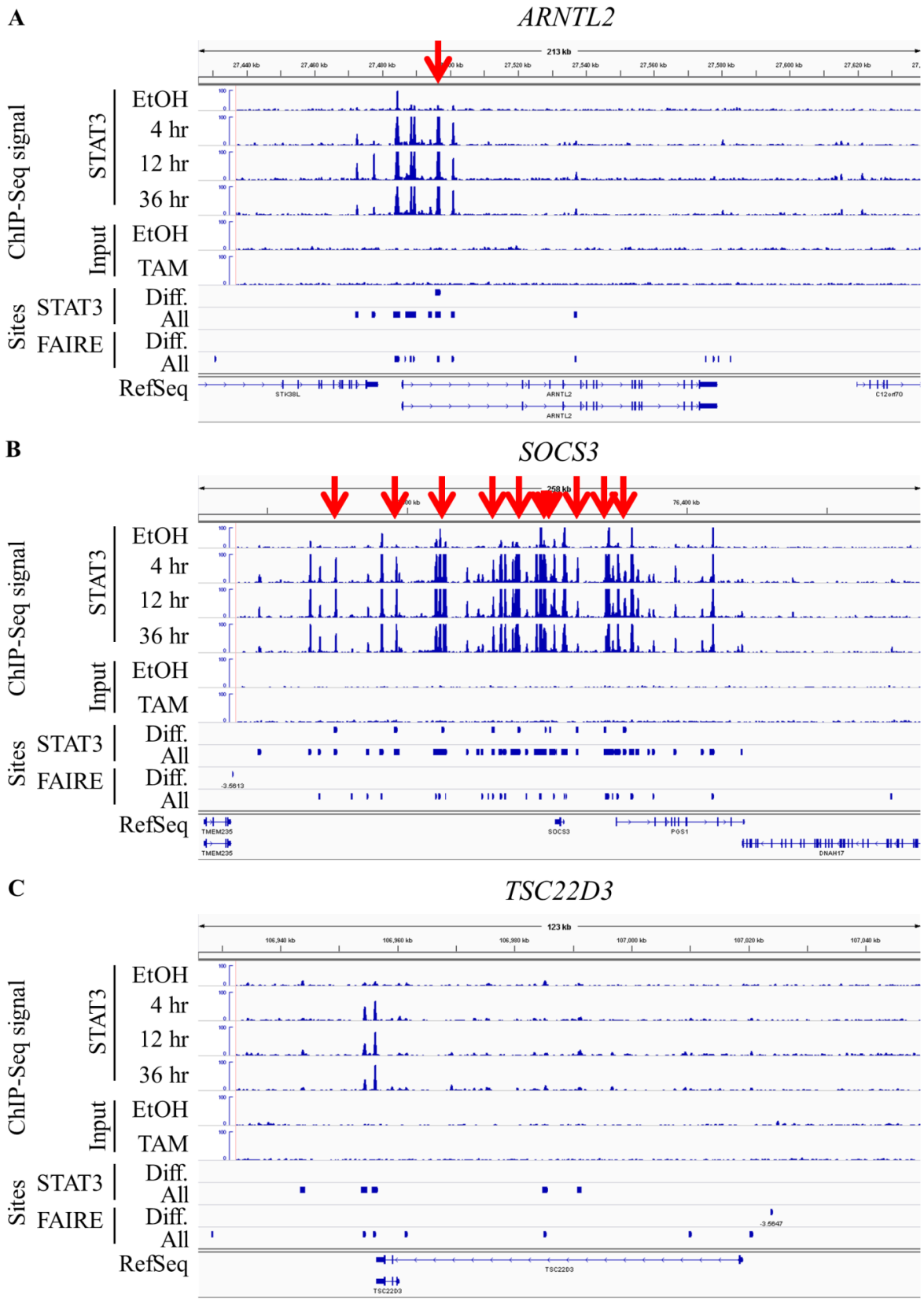
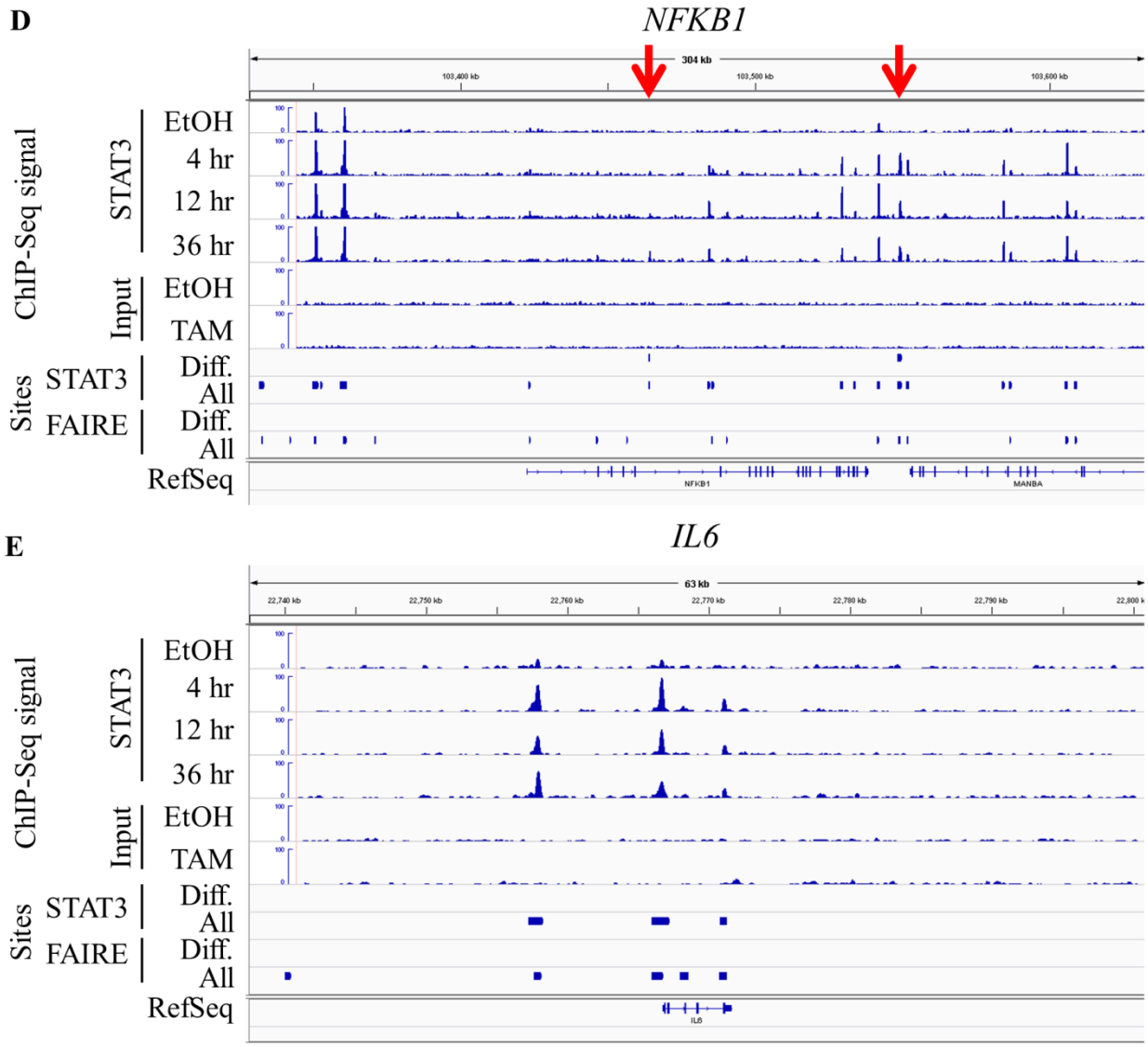


Figure 11 (Continued)



The location of STAT3 during transformation

STAT3 is mainly located at regions outside of proximal promoters, which has been observed previously [299, 300]. In MCF10A-ER-Src cells, only 7% of STAT3 binding sites were located within 2500 bp upstream of a RefSeq transcriptional start site (TSS) (Figure 12, A). Most STAT3 sites were located within introns (40%) and regions distal to RefSeq gene features (42%) (Figure 12, A), most of which were located within CREs and were therefore most likely enhancers. Of all STAT3 sites, 57% directly overlapped a CRE. The locations of differential STAT3 sites were primarily formed at locations distal to RefSeq TSSs. Differential STAT3 sites were found 250 bp upstream of RefSeq TSSs at a rate similar, and not reaching significance, to that of the genomic background (0.4% vs. 0.3%), however, all STAT3 sites were found at a rate of 2.6% (P-value < 10^{-15}; Figure 12, A). This preference is also seen at regions 250 – 2500 bp upstream of RefSeq TSSs (2.6% vs. 4.1%, P-value < 10^{-9}; Figure 12, A). Moreover, there was a statistically significant increase in distal intergenic STAT3 sites in the differential population compared to all STAT3 sites (49% vs. 42%, P-value < 10^{-15}; Figure 12, A). Plots of the density of STAT3 sites in relation to RefSeq TSSs also showed that STAT3 sites that were differential during transformation were preferentially located in regions distal of TSSs (Figure 12, B).

The bias of differential STAT3 sites towards distal intergenic regions was unlikely to be due to STAT3 alone, and most likely reflected the biased location of a co-operating factor(s) and their DNA motif(s) and/or saturation of STAT3 binding to proximal promoter locations. A comparison of known motifs between differential STAT3 sites and all STAT3 sites implicated the AP-1 (occurring as "*AP-1*" and "*NFE2L1::MafG*"), NOBOX and PRXX2 motifs in cooperating with STAT3 at these loci (not shown). Neither NOBOX, an oocyte specific transcriptional activator, nor PRXX2, interestingly involved in mesenchymal cell proliferation

Figure 12: Transformation induced differential STAT3 sites are preferentially located outside of proximal promoters

A. Distribution of STAT3 sites at RefSeq gene features.

B. ChIP-Seq peak density of STAT3 and NF-Y about RefSeq TSSs. Differential STAT3 sites were located more distally than all STAT3 peaks from 4 hr, 12 hr and 36 hr post ER-Src induction. 0 bp represents the TSS.

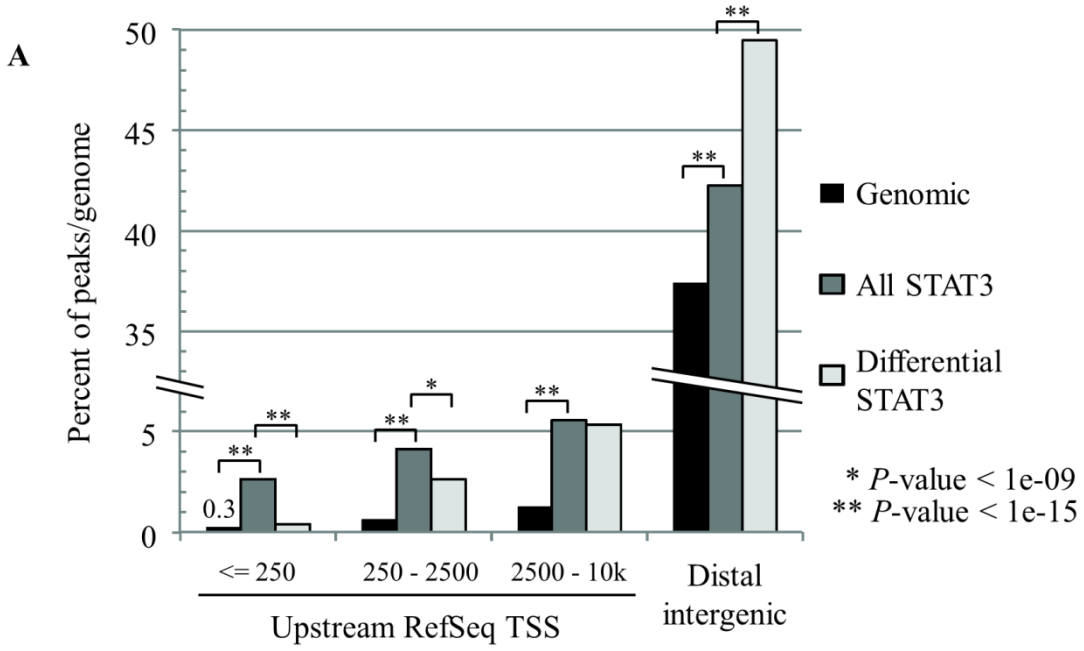
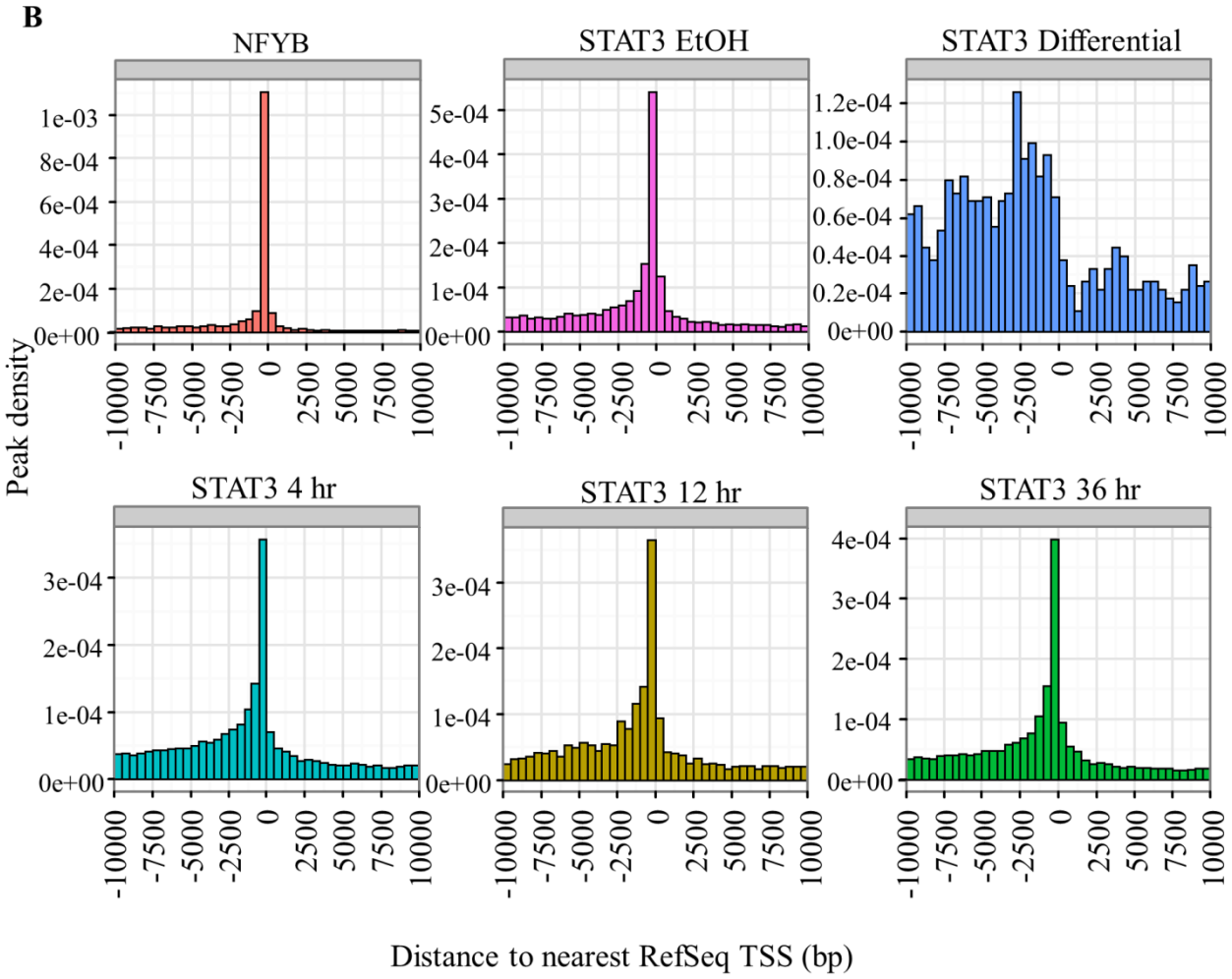


Figure 12 (Continued)



and wound healing, had detectable RNA in MCF10A-ER-Src cells (not shown). All members of the FOS and JUN family (AP-1 factors) were expressed (Figure 17, A).

STAT3 was located at FOS bound sites

The increase in STAT3 occupancy across the genome observed during transformation was closely associated with FOS binding. ChIP-Seq of the AP-1 factor FOS revealed that 82% of all STAT3 sites directly overlapped FOS bound sites. As can be seen in Figure 13, A, there was a large overlap between STAT3 and FOS sites throughout transformation. Specifically observing STAT3 differential sites indicated that nearly all (88%) were associated with a pre-existing FOS bound site, and not with differential FOS sites (25%; Figure 13, B).

***Cis*-regulatory elements were static during cancer transformation**

To identify the CREs utilized during transformation we performed FAIRE (formaldehyde assisted isolation of regulatory elements) followed by massively parallel DNA sequencing [301, 302]. FAIRE allows the identification of nucleosome free genomic regions by utilizing the preferential solubility of fragmented, non-protein bound DNA in the aqueous phase of phenol-chloroform purification. This enriches for CREs which are depleted or free of nucleosomes [303]. Nucleosomes are, by far, the major protein component bound to the genome, and are efficiently cross-linked to DNA by formaldehyde whereas TFs, in general, are not [304, 305]. Thus, nucleosome depleted FAIRE regions are enriched and can be detected by DNA sequencing. Cross-linked chromatin from 0 hr untreated (control), and 4 hr, 12 hr and 36 hr TAM (tamoxifen) treated cells were analyzed to identify CREs used during transformation.

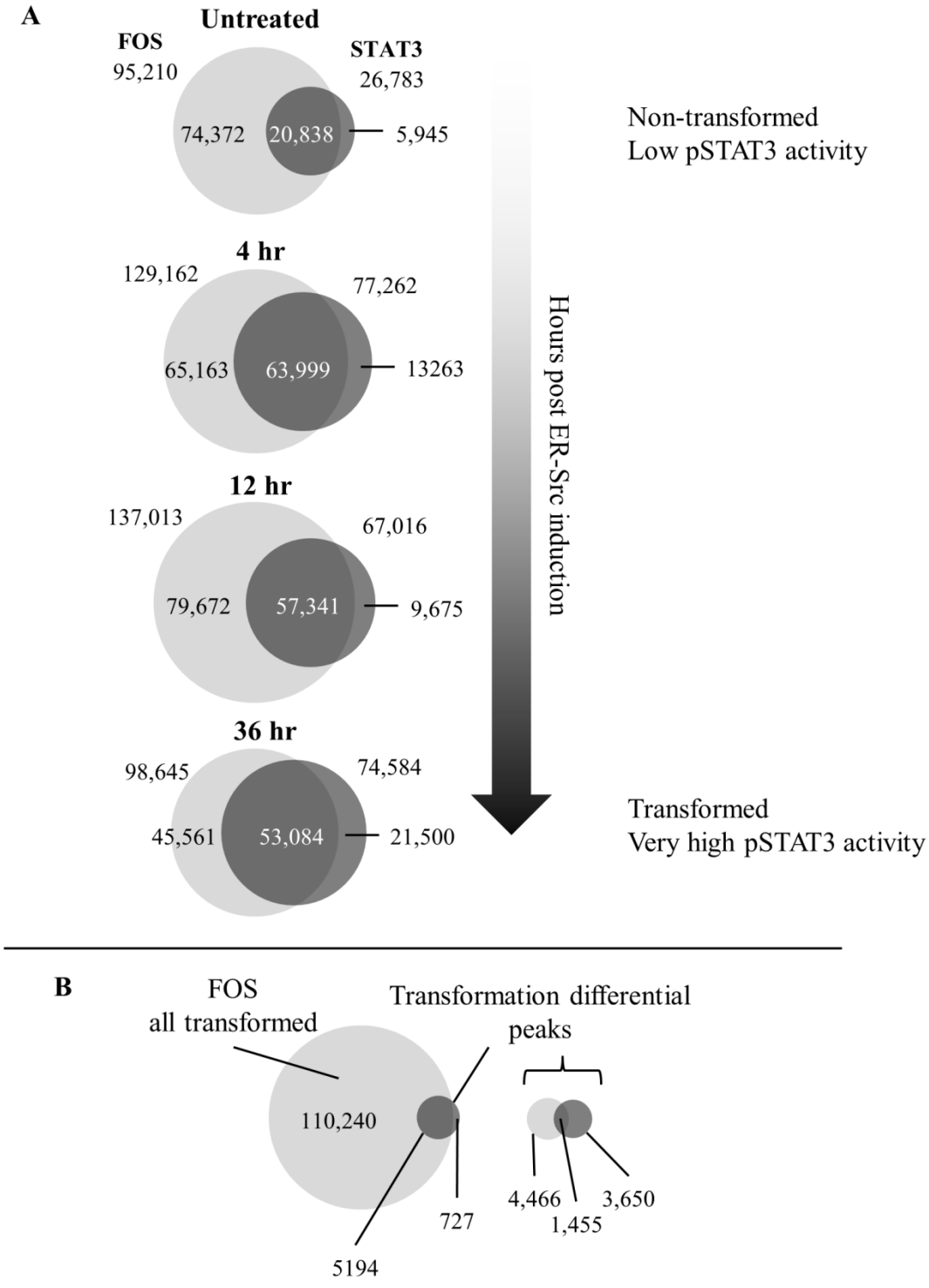
Across all our samples, 100,597 non-redundant CREs were identified in MCF10A-ER-Src cells which was in line with the number of FAIRE-Seq defined CREs found in other cell types

Figure 13: Overlap of STAT3 and FOS sites during transformation

A. STAT3 and FOS sites from each time point that directly overlapped.

B. Transformation dependent differential STAT3 sites directly overlapping all FOS sites from 4 hr, 12 hr and 24 hr post ER-Src induction or transformation-dependent differential FOS sites.

Figure 13 (Continued)



[184, 306]. The locations corresponded well to the known locations of TFs derived from ChIP-Seq experiments in MCF10A-ER-Src cells (Figure 14, A). Surprisingly, given the major phenotypic and transcriptional changes observed during transformation (Figure 15; Figure 22, E), only 6.6% (n = 6617) of CREs were differentially present in at least one time point during transformation (Figure 14, B). An analysis of differential CREs indicated that they were indistinguishable from the genomic background in terms of gene ontologies (not shown), with no significantly enriched terms. The differential CREs were most likely false positives and were not considered further.

Differential STAT3 sites did not create new CREs

In addition to the above findings, the differential CREs were not the preferential location of differential STAT3 or FOS bound loci (not shown; note "Diff." FAIRE track in Figure 11 and Figure 20). STAT3 activity was increased during transformation and would be expected to be preferentially located at newly formed transformation-dependent CREs. The vast majority of STAT3 sites represented a modest, but *en masse* genome-wide accumulation of STAT3 on chromatin which was not reflected in the generation of new CREs. Of differential STAT3 sites, 38% (n = 2266) directly overlapped a stable CRE, however, only 0.24% (n = 14) of differential STAT3 sites occurred at differential CREs and only 0.27% (n = 18) of differential CREs contained differential STAT3 binding sites. Therefore, differential STAT3 sites did not elicit the mass formation of new transformation-dependent CREs, but rather largely utilized the pre-existing population.

Figure 14: Co-localization of FAIRE-Seq regions and TF binding sites

A. FAIRE-Seq regions co-localized with TF binding sites. ChIP-Seq datasets of TFs, RNA Pol II and RPC155 (Pol III subunit) in MCF10A-ER-Src cells at the *EPAS1* locus with the location of FAIRE-Seq sites indicated.

B. The active CREs identified by FAIRE-Seq during transformation and clustering of the differential CREs (highlighted in red) based on their dynamics over the time-course assayed.

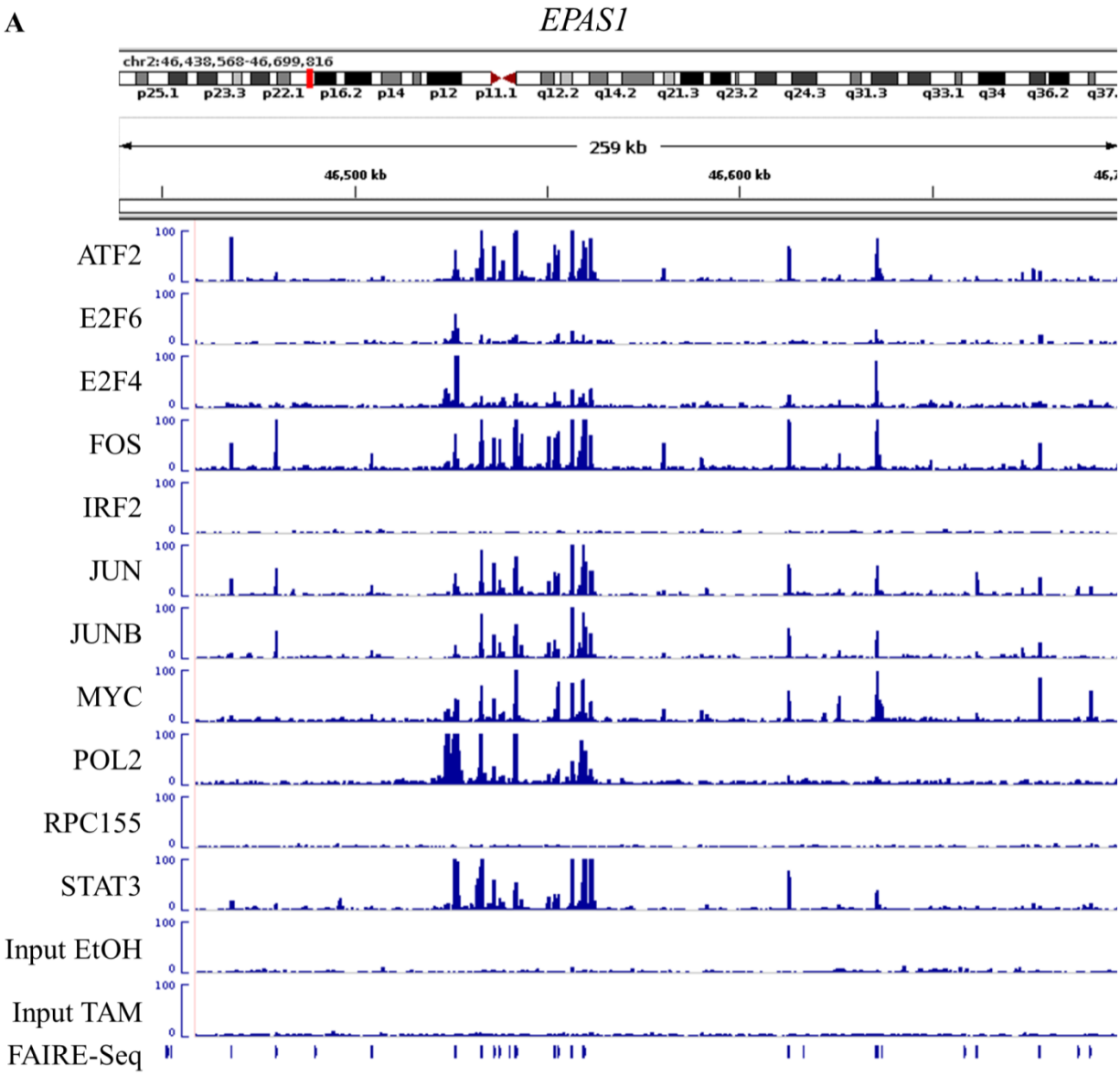


Figure 14 (Continued)

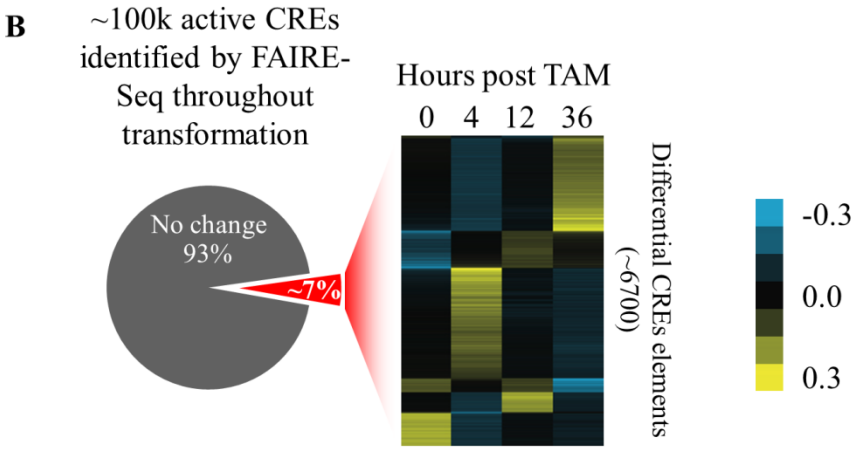
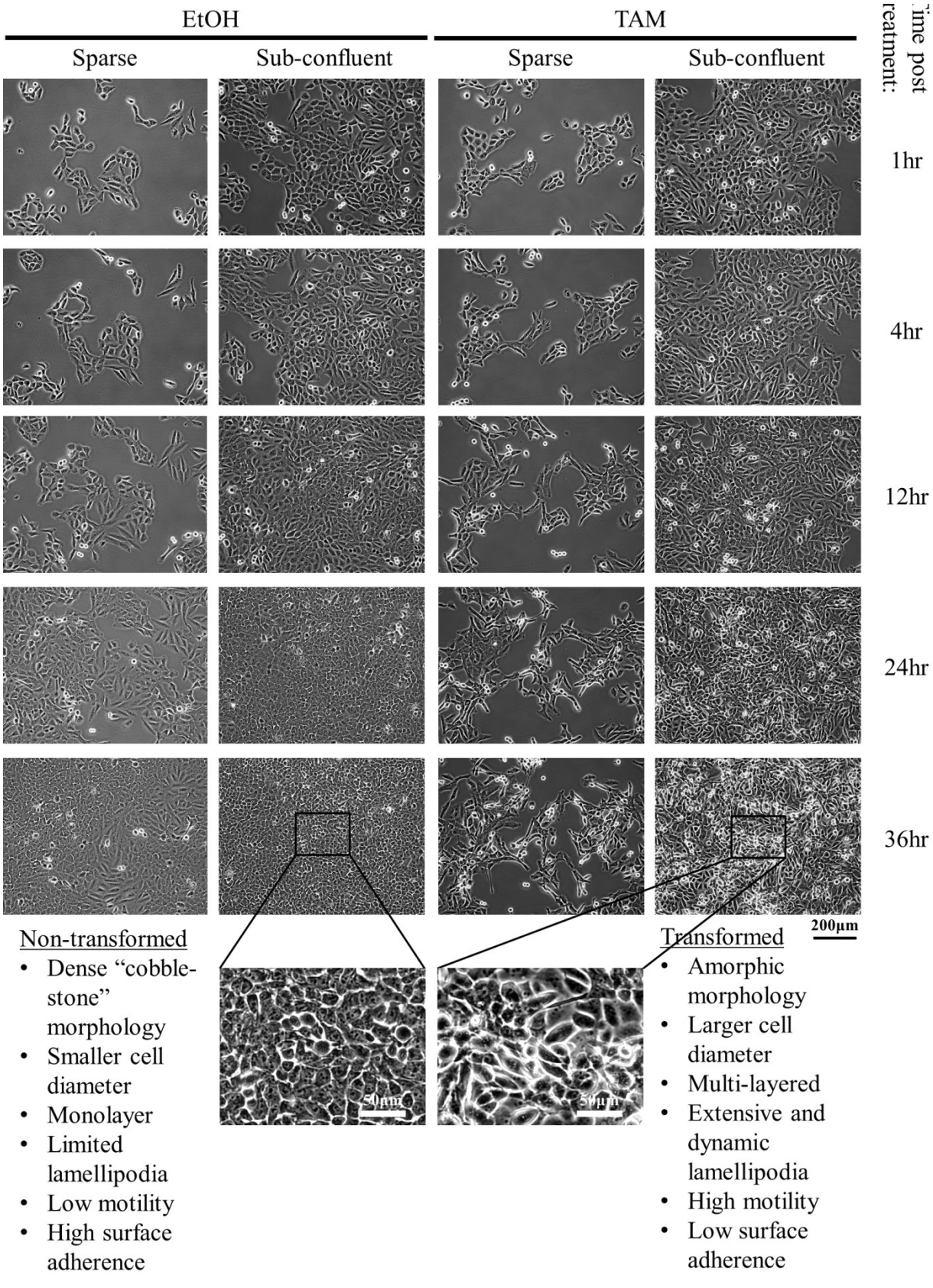


Figure 15: Morphological changes of MCF10A-ER-Src cells undergoing transformation

MCF10A-ER-Src cells, at two cell densities, were treated with EtOH or TAM (induces ER-Src) at time 0 hr and tracked by DIC time-lapse microscopy until 36 hr post treatment. The key phenotypic differences between transformed and non-transformed cells are highlighted. All images were taken at the same magnification. A video of the time-lapse is available in Supplementary Videos.

Figure 15 (Continued)



STAT3 had a limited ability to bind to its motif outside of nucleosome depleted CREs

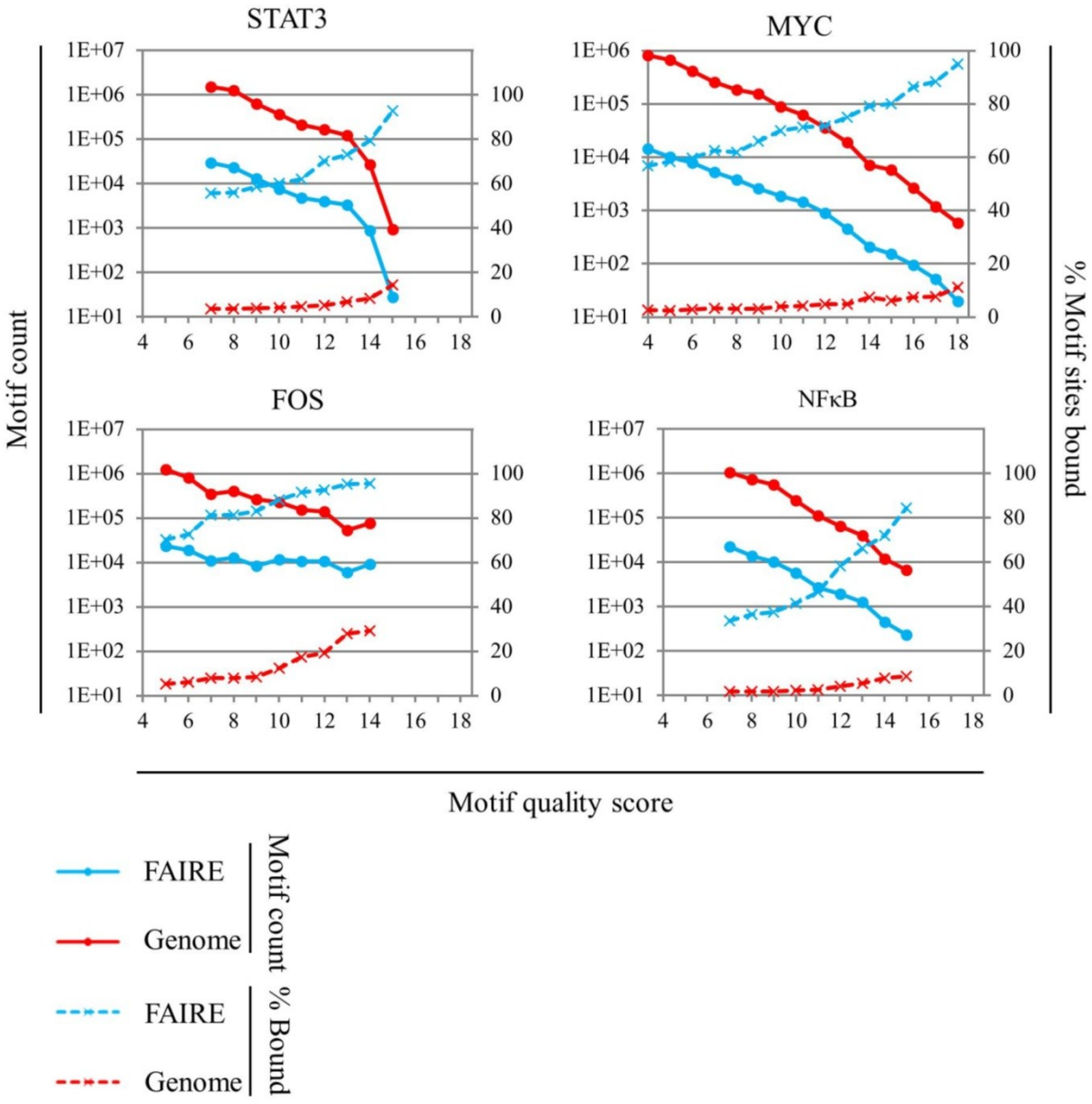
To determine if STAT3 was able to bind to its motif outside of nucleosome-depleted CREs, we compared STAT3 bound sites, the genomic locations of STAT3 DNA motifs and FAIRE-Seq regions. CREs containing a STAT3 motif where exceptionally well occupied by STAT3. At a motif quality score of 14, there were ~26,000 potential STAT3 binding sites within the *H. sapiens* genome of which only 8% were occupied by STAT3, whereas 80% of those falling within FAIRE-Seq regions were bound by STAT3 (Figure 16). Access of STAT3 to its motif was largely limited to nucleosome-depleted open genomic loci and similar results were found for NFκB and MYC. MYC is known to bind only to H3 acetylated loci [307] which are nucleosome depleted open regions. In comparison, genome-wide FOS binding sites were well occupied (30%), which may be due to cooperation with NF-Y (this dissertation, Chapter 2) or the biased prevalence of AP-1 motifs at constitutively open regions. In this regard, STAT3 bound opportunistically, present at most if not all CREs that contained a suitable DNA motif for binding, with the limiting step being the post-translational activation of STAT3 (*i.e.* phosphorylation).

STAT3 regulated AP-1 factors were likely the predominant transcriptional regulators during the later stages of transformation

Temporally, FOS RNA levels peaked at 24 hr post induction of ER-Src and this response was STAT3-dependent (Figure 17, A). FOS is part of the AP-1 TF complex, a heterodimeric regulatory complex composed of FOS (FOS, FOSL1, FOSL2, FOSB) and JUN family (JUN, JUNB, JUND) members, many of which were significantly differentially expressed during transformation (*FOS, FOSL1, FOSL2, JUNB, JUND*; Figure 17, A). Importantly, in a STAT3-dependent manner, *FOSL2* and *JUNB* were activated during transformation, whereas *JUND* was

Figure 16: Occupancy of TF DNA binding site motifs in CREs

The DNA binding site motifs of STAT3, MYC, FOS and NFκB, at varying motif quality scores, were computationally discovered genome-wide and the percentage that resided within a ChIP-Seq peak from the respective TF were calculated and plotted. For FAIRE-Seq regions, only those motifs that directly overlapped a FAIRE element were considered. The number of motifs at each quality score within the genome or FAIRE elements is also shown.



repressed (Figure 17, A). An analysis of the regulatory regions bound by FOS sites that were differential during transformation highlighted its importance to many STAT3 regulated processes such as "*G-protein coupled receptor signaling*", "*NFκB signaling*", "*Cellular movement*", and "*Cell death*" (Figure 17, B). In this regard, FOS cooperates with STAT3 in many key processes of transformation. *De novo* motif analysis of the top 10,000 FOS bound sites revealed the canonical AP-1 motif (Figure 17, C).

FOS bound to embryonic stem cell and bone metastasis related genes and pathways

Differential FOS sites were enriched for embryonic stem cell and development associated gene ontology terms: "*Role of NANOG in ESC pluripotency*", "*Human ESC pluripotency*", "*Organismal development*", "*Embryonic development*", and "*Organ development*" (Figure 17, B). Moreover, *IPA* analysis (see Methods) of the STAT3-dependent transcriptional program at 24 hr post ER-Src induction found FOS, FOSL2 and JUND, as key downstream effectors in a pathway which was regulated during transformation that linked cellular assembly and organization, embryonic development and organ development genes, via TGFβ3 signaling through the guanine nucleotide exchange factor SOS, and the extra-cellular matrix protein, tenascin C (TNC) (Figure 18, B). High TNC expression is a biomarker for poor prognosis in breast cancer [308, 309] and was found to be essential for breast cancer metastasis [310-312]. The genes within the ontology "*Human embryonic stem cell pluripotency*" containing a differential FOS binding site are illustrated and listed in Figure 19, A and B. In addition, other TFs linked to stem cell function were significantly differentially regulated at the RNA level during transformation: *PAX8*, *SOX4/9*, *STAT3*, *ZBTB16*, and *LEF1* (Figure 28).

A second gene ontology category whose genes were significantly enriched for FOS differential sites, and which is pertinent to the clinical manifestation of breast cancer, were

Figure 17: Deregulation of AP-1 factors and the GO terms associated with differential FOS binding during transformation

A. Expression array analysis of AP-1 factors during transformation and their transcriptional dependence on STAT3. Shown are the normalized RNA expression levels at 4 hr and 24 hr post EtOH or TAM treatment in samples transfected with siSCM (scrambled control) or siSTAT3. RNA levels are expressed as fold change over the 4 hr EtOH and siSCM treated sample.

B. Significant gene ontology terms bound by transformation-dependent differential FOS sites.

C. The FOS DNA binding site motif derived *de novo* from FOS ChIP-Seq.

Figure 17 (Continued)

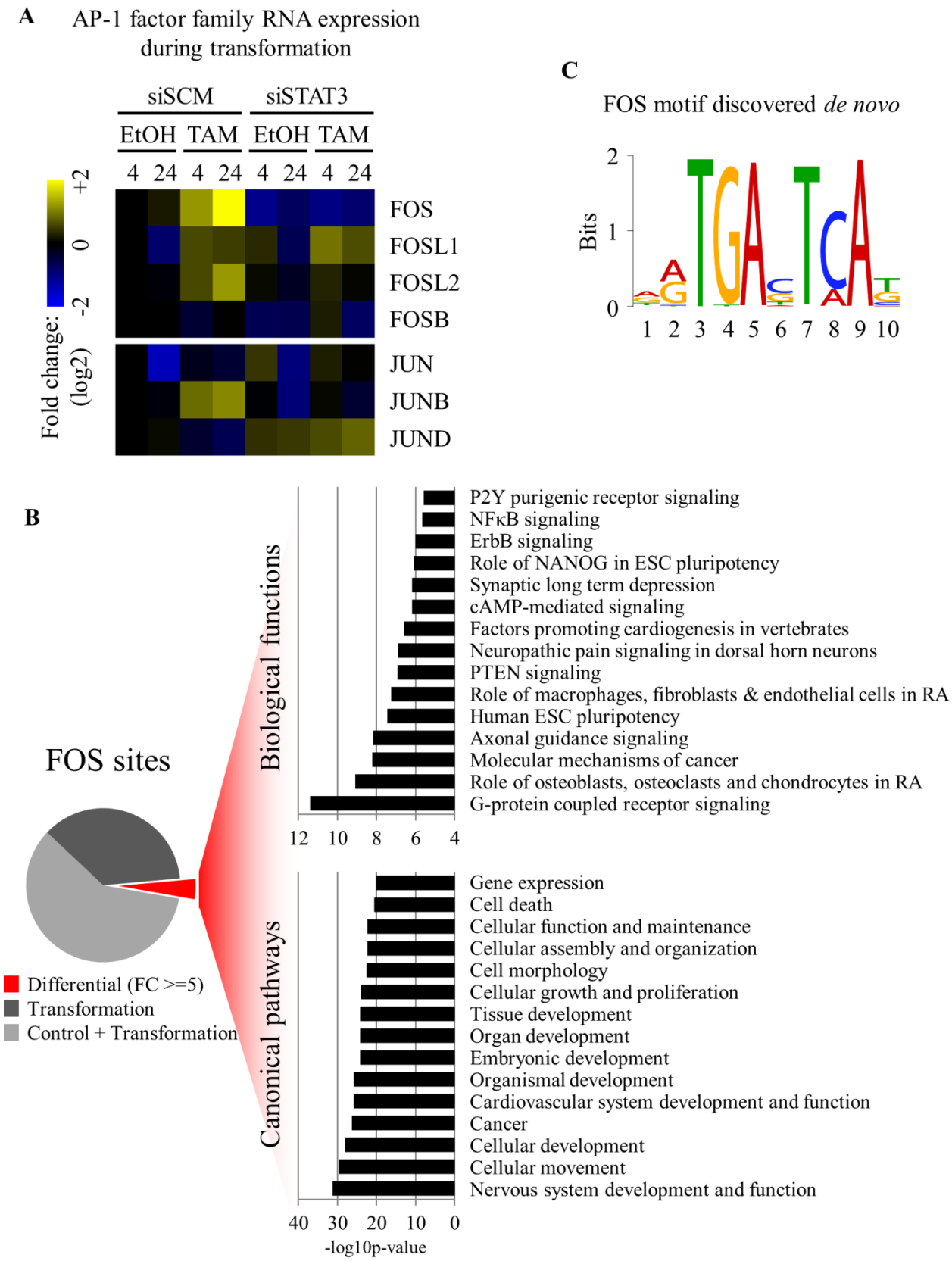


Figure 18: Ingenuity Pathway Analysis of genes differentially regulated during transformation

A. The top interaction network derived from STAT3-dependent and transformation-dependent differentially regulated genes at 4 hr post induction of ER-Src. Green and red shading indicates down- and up-regulated by siSTAT3 treatment, respectively. Only genes that were differentially regulated by transformation and by siSTAT3 were considered.

B. Similar to A except at 24 hr post ER-Src induction.

C. The top interaction network derived from STAT3-independent and transformation-dependent differentially regulated genes at 4 hr post induction of ER-Src. Green and red shading indicates down- and up-regulated during transformation, respectively. Only genes that are differentially regulated by transformation and not by siSTAT3 were considered.

D. Similar to C except at 24 hr post ER-Src induction.

Lines between two genes indicate a known or predicted, direct or indirect interaction.

Figure 18 (Continued)

A

Transformation- and STAT3-dependent differentially expressed genes – 4hr

Pathway: Organismal Injury and Abnormalities, Cellular Movement, Nervous System Development and Function

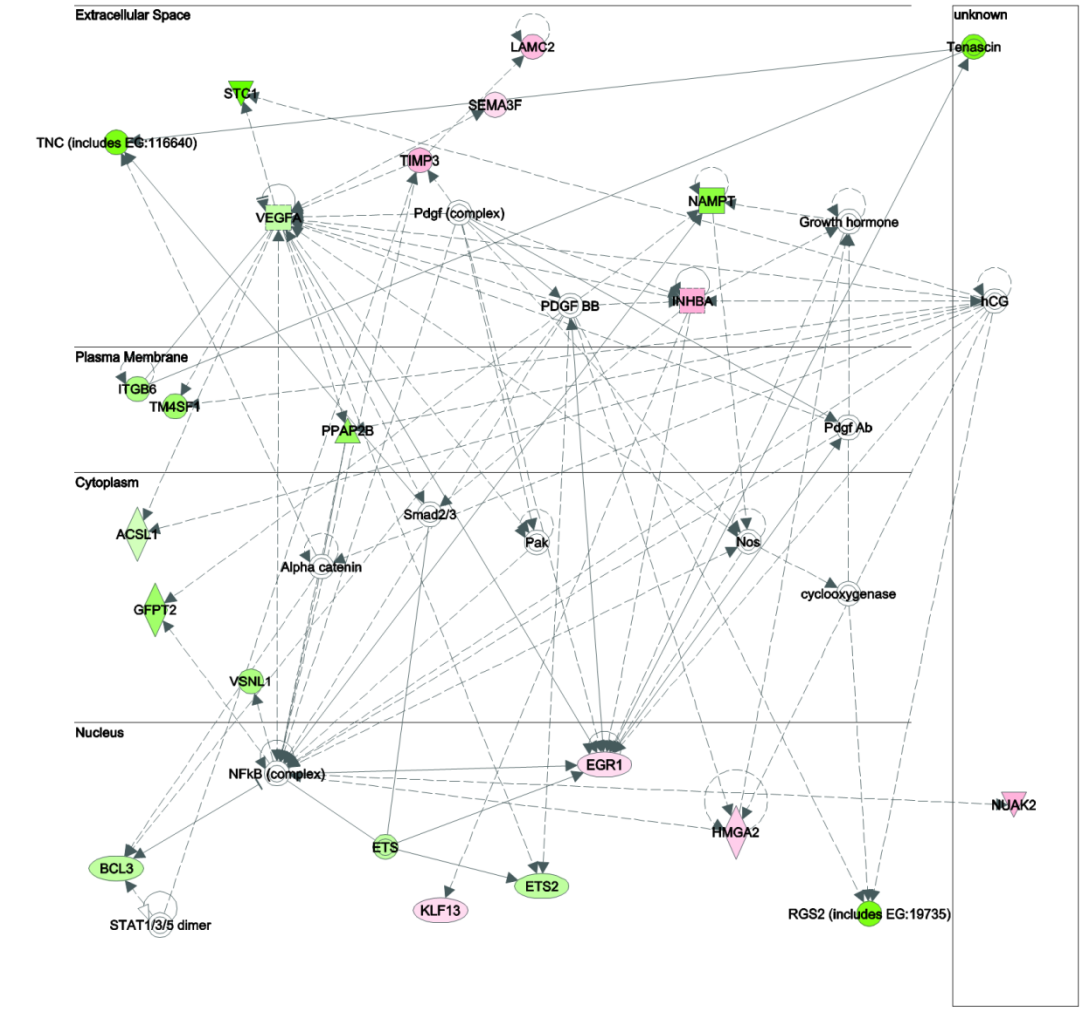


Figure 18 (Continued)

B

Transformation- and STAT3-dependent differentially expressed genes – 24 hr

Pathway: Cellular Assembly and Organization, Embryonic Development, Organ Development

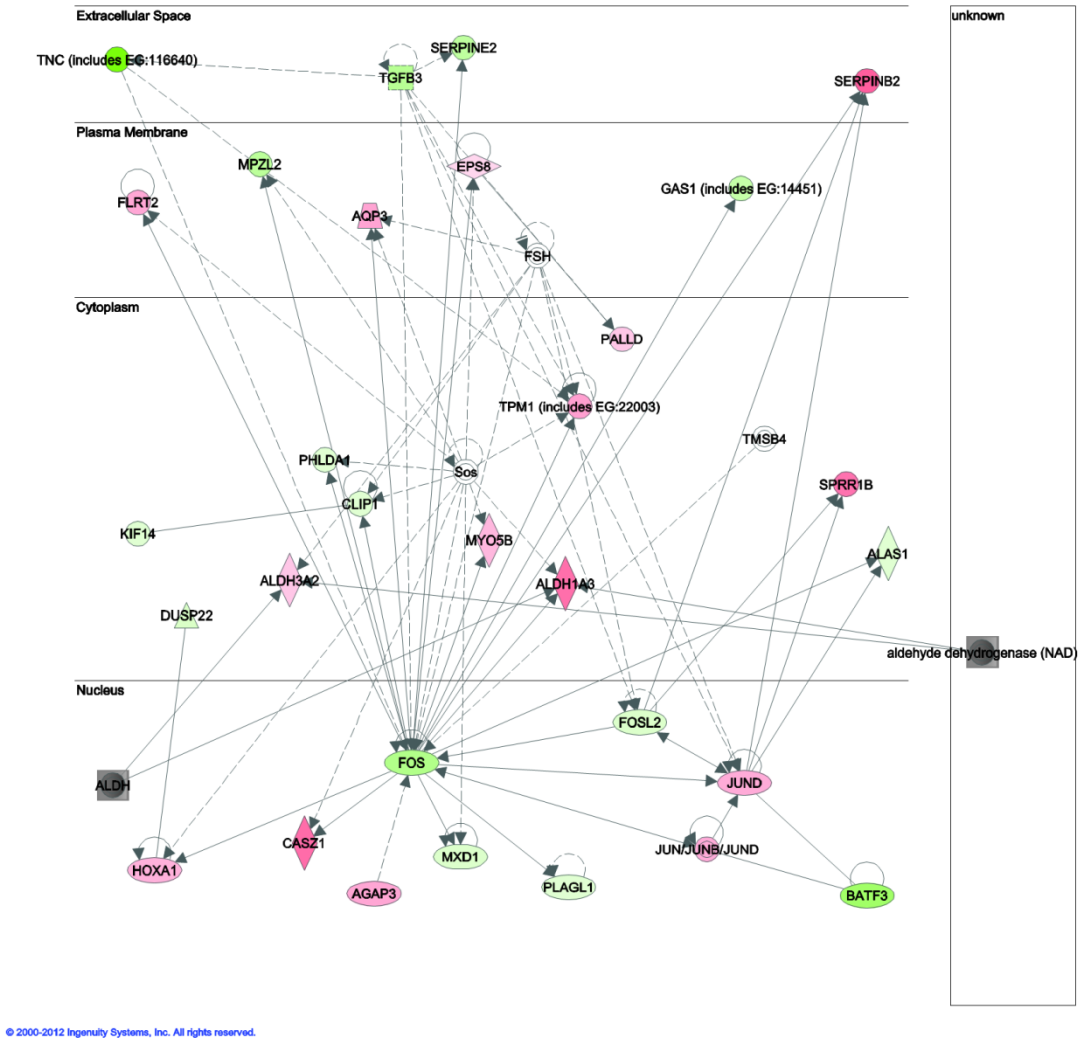


Figure 18 (Continued)

C

Transformation-dependent and STAT3-independent differentially expressed genes – 4 hr

Pathway: Cancer, Hereditary Disorder, Reproductive System Disease

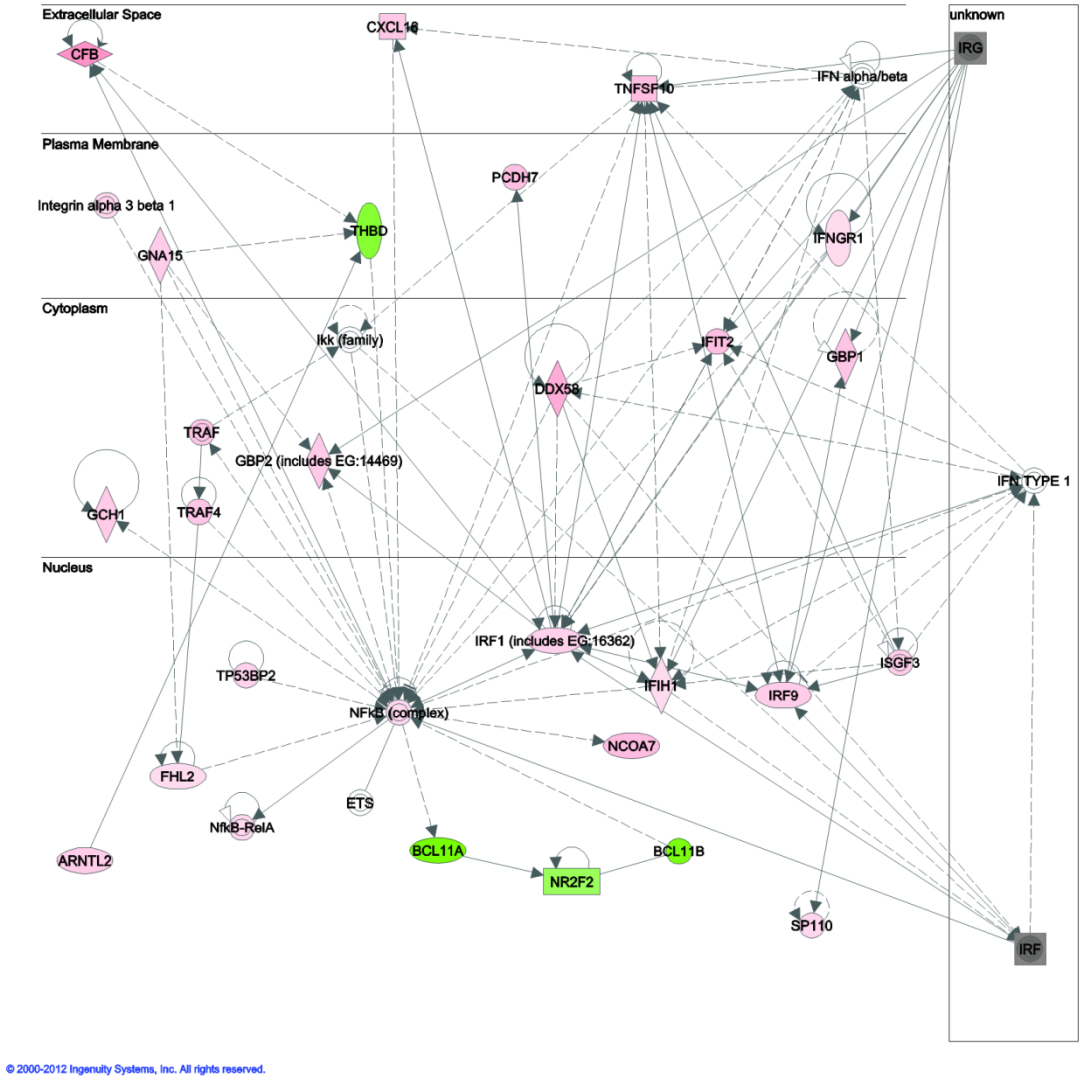


Figure 18 (Continued)

D Transformation-dependent and STAT3-independent differentially expressed genes – 24 hr

Pathway: Carbohydrate Metabolism, Drug Metabolism, Small Molecule Biochemistry

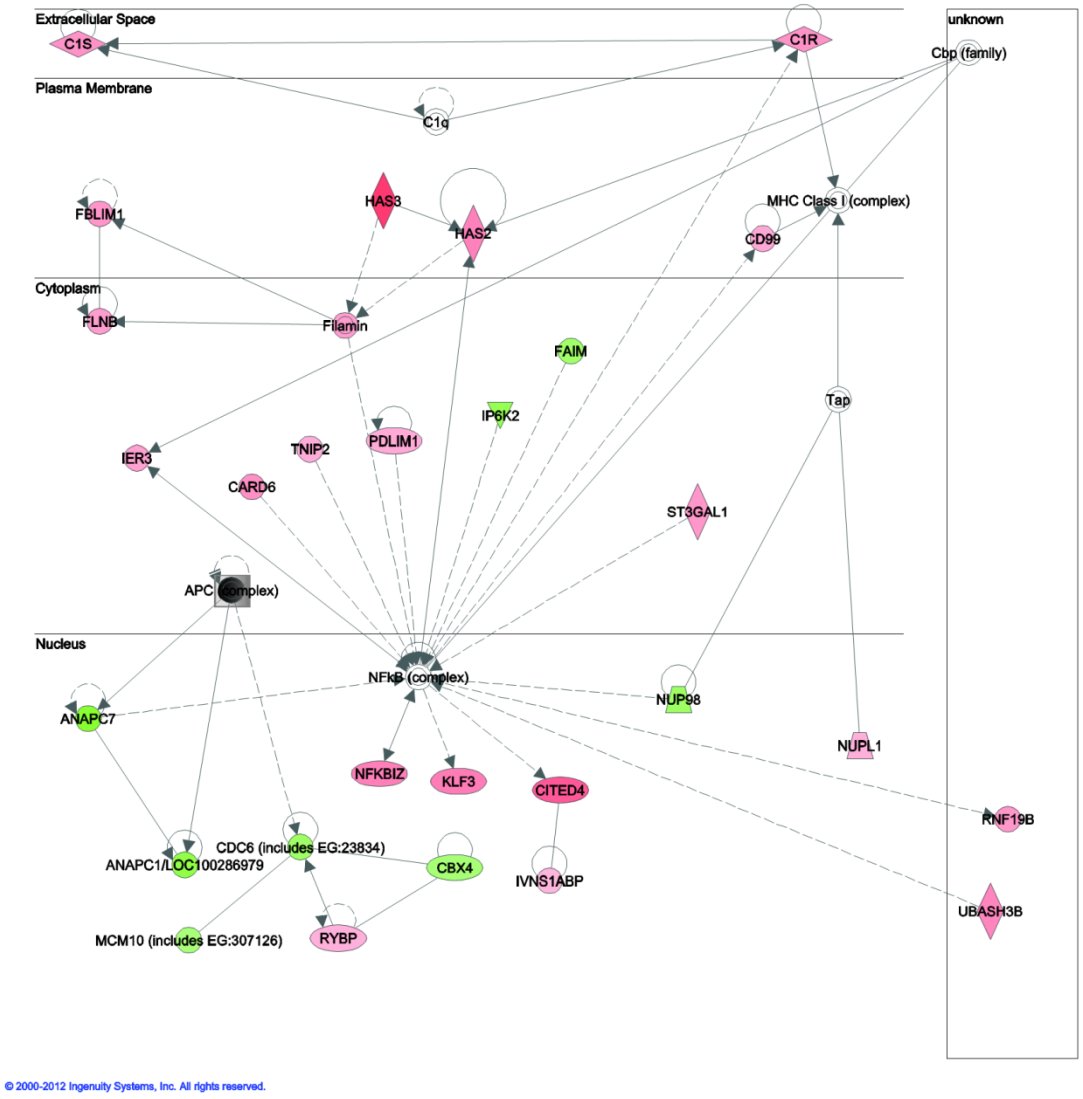


Figure 19: Embryonic-related genes bound by FOS in MCF10A-ER-Src cells

A. The Ingenuity Systems canonical pathway "*Human embryonic pluripotency*" showing genes whose regulatory domain (see *GREAT*, Methods) was bound by transformation-dependent differential FOS sites (highlighted in gray shading).

B. Similar to A. Shows the differential gene expression changes upon ER-Src activation at 4 hr and 24 hr post induction. Genes not showing expression changes were not significantly altered during transformation.

Figure 19 (Continued)

A

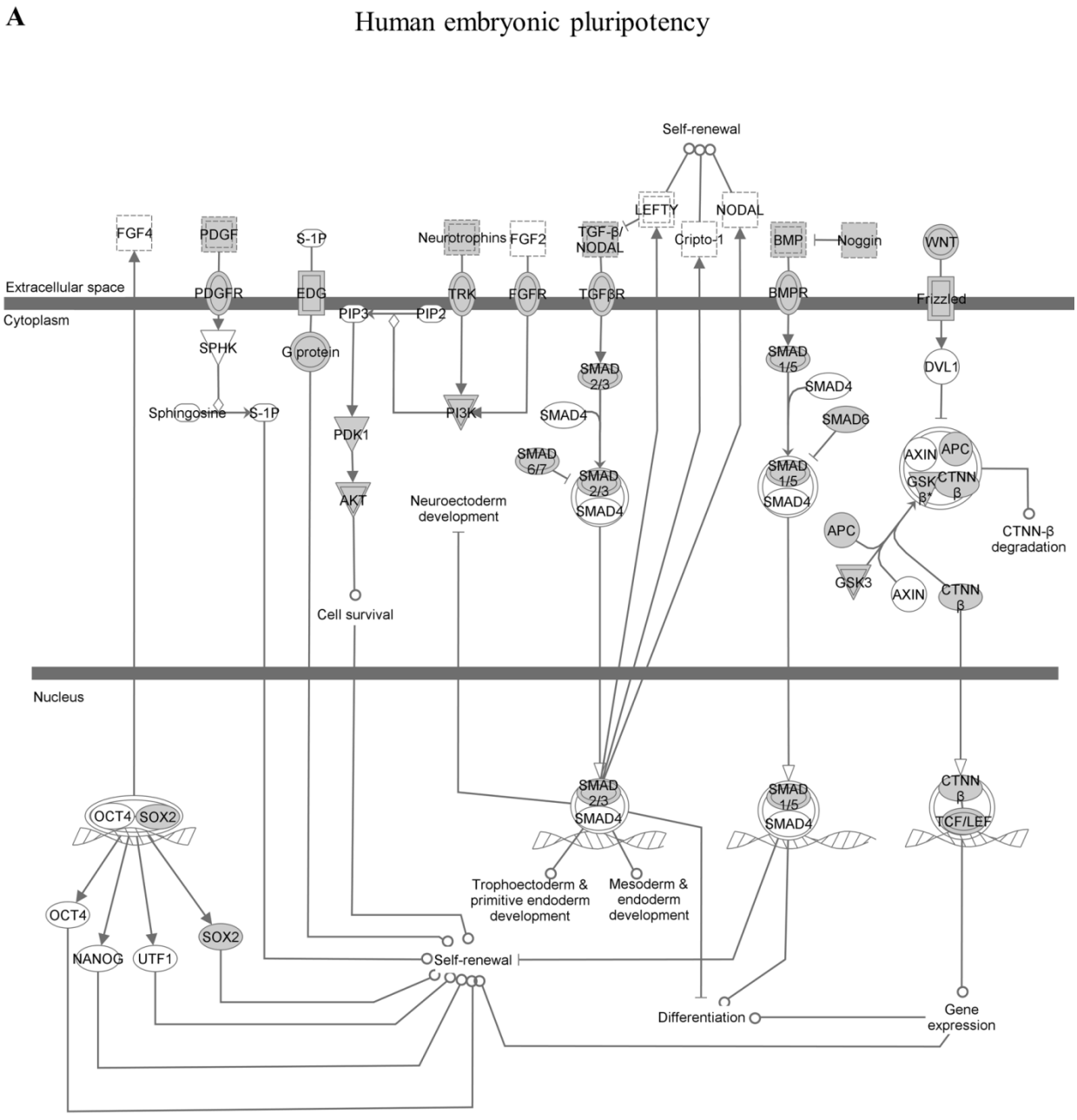


Figure 19 (Continued)

B

	Differentially expressed during transformation			
	4 hr		24 hr	
Gene	log2 FC	*P*-value	log2 FC	*P*-value
AKT3	-	-	-	-
APC	-	-	-	-
BDNF	-	-	-0.62	8.3E-05
BMP15	-	-	-	-
BMP2	-	-	1.56	5.7E-09
BMP3	-	-	-	-
BMP4	-	-	-	-
BMP5	-	-	-	-
BMP7	-	-	-	-
BMP8A	-	-	-	-
BMPR1B	-	-	-	-
CTNNB1	-	-	-	-
FGFR1	-	-	-	-
FGFR2	-	-	-3.21	6.9E-05
FZD1	-	-	-	-
FZD10	-	-	-	-
FZD4	-	-	-	-
FZD7	-	-	-	-
FZD8	-	-	-	-
GNAS	-	-	-	-
GSK3B	-	-	-	-
LEF1	-	-	-0.65	9.8E-05
NGF	-	-	-	-
NOG	-	-	-	-
NTRK2	-	-	-	-
NTRK3	-	-	-	-
PDGFB	-	-	-	-
PDGFC	-	-	-	-
PDGFRA	-	-	-	-
PDGFRB	-	-	-	-
PDPK1	-	-	-	-

	Differentially expressed during transformation			
	4 hr		24 hr	
Gene	log2 FC	*P*-value	log2 FC	*P*-value
PIK3C2A	-	-	-	-
PIK3C2G	-	-	-	-
PIK3C3	-	-	-	-
PIK3CB	-	-	-	-
PIK3CG	-	-	-	-
PIK3R1	-0.61	3.9E-05	-	-
PIK3R3	-	-	-1.05	9.0E-07
PIK3R4	-	-	-	-
PIK3R5	-	-	-	-
S1PR1	-	-	-	-
SMAD1	-	-	-	-
SMAD2	-	-	-	-
SMAD3	0.74	2.2E-06	0.74	2.0E-06
SMAD6	-	-	-	-
SMAD7	-	-	-	-
SOX2	-	-	-	-
TCF4	-	-	-	-
TCF7L2	-	-	-	-
TGFB2	0.74	8.4E-05	-	-
TGFB3	-	-	1.49	1.4E-05
TGFBR1	-	-	-	-
TGFBR2	-0.96	7.5E-05	-	-
WNT11	-	-	-	-
WNT16	-	-	-	-
WNT2	-	-	-	-
WNT2B	-	-	-	-
WNT3	-	-	-	-
WNT7A	-	-	-	-
WNT7B	-	-	-	-
WNT8B	-	-	-	-
WNT9B	-	-	-	-

mechanisms of osteo-pathologies and associated inflammation: "*Role of osteoblasts, osteoclasts and chondrocytes in RA*", and "*Role of macrophages, fibroblasts and endothelial cells in RA*" (Figure 17, B). Breast cancers frequently metastasize to and reoccur within bone, where it induces pathologic osteoclast-mediated bone resorption leading to osteolytic lesions, which are the main causes of pain and disability in breast cancer patients. Here we find that key genes known to be involved in this process (BMP2, MMPs and TNFRSF11A (RANK)) contained differential FOS binding sites (Figure 20, A-C) and were differentially regulated during transformation. The genes within the ontology "*Role of osteoblasts, osteoclasts and chondrocytes in RA*" containing a transformation dependent differential FOS binding site are illustrated and listed in Figure 21, A-D.

Functional inactivation of STAT3 during transformation

To understand if STAT3 alone can explain the gene transcription program observed during transformation and to further define its direct targets, we knocked down STAT3 by siRNA and tracked perturbations in gene expression by microarray expression analysis. siRNAs specific to *STAT3* or a control scrambled siRNA (siSCM) were transiently transfected into non-confluent MCF10A-ER-Src cells. After two days in culture, cells were treated with either tamoxifen (TAM, to induce ER-Src and transformation) or EtOH (control for cell growth and crowding). After an additional 4 or 24 hr in culture, RNA was extracted and assayed by 3'-biased Affymetrix *H. sapiens* whole genome gene expression arrays.

To assess the quality of the STAT3 knockdown, parallel protein samples were harvested 72 hr post siRNA treatment (*i.e.* 24 hr post TAM treatment) and assayed by Western blot (Figure 22, A). A robust knockdown of STAT3 protein levels by > 95% was observed in two replicates. In addition, microarray expression values showed a > 10 fold decrease in *STAT3* RNA level only

Figure 20: Genome view of FOS binding during transformation

ChIP-Seq read counts at the *TNFRSF11A*, *BMP2* and *MMP* loci during transformation of MCF10A-ER-Src cells. 4 hr, 12 hr and 36 hr indicate time post ER-Src induction by TAM treatment. EtOH and TAM input samples are a single replicate, all others are of 2 biological replicates combined. ChIP-Seq and FAIRE-Seq elements deemed to be transformation dependent differential ("Diff.") sites and all sites derived from TAM and EtOH treated samples are shown. Red arrows highlight differential ("Diff.") ChIP-Seq sites.

Figure 20 (Continued)

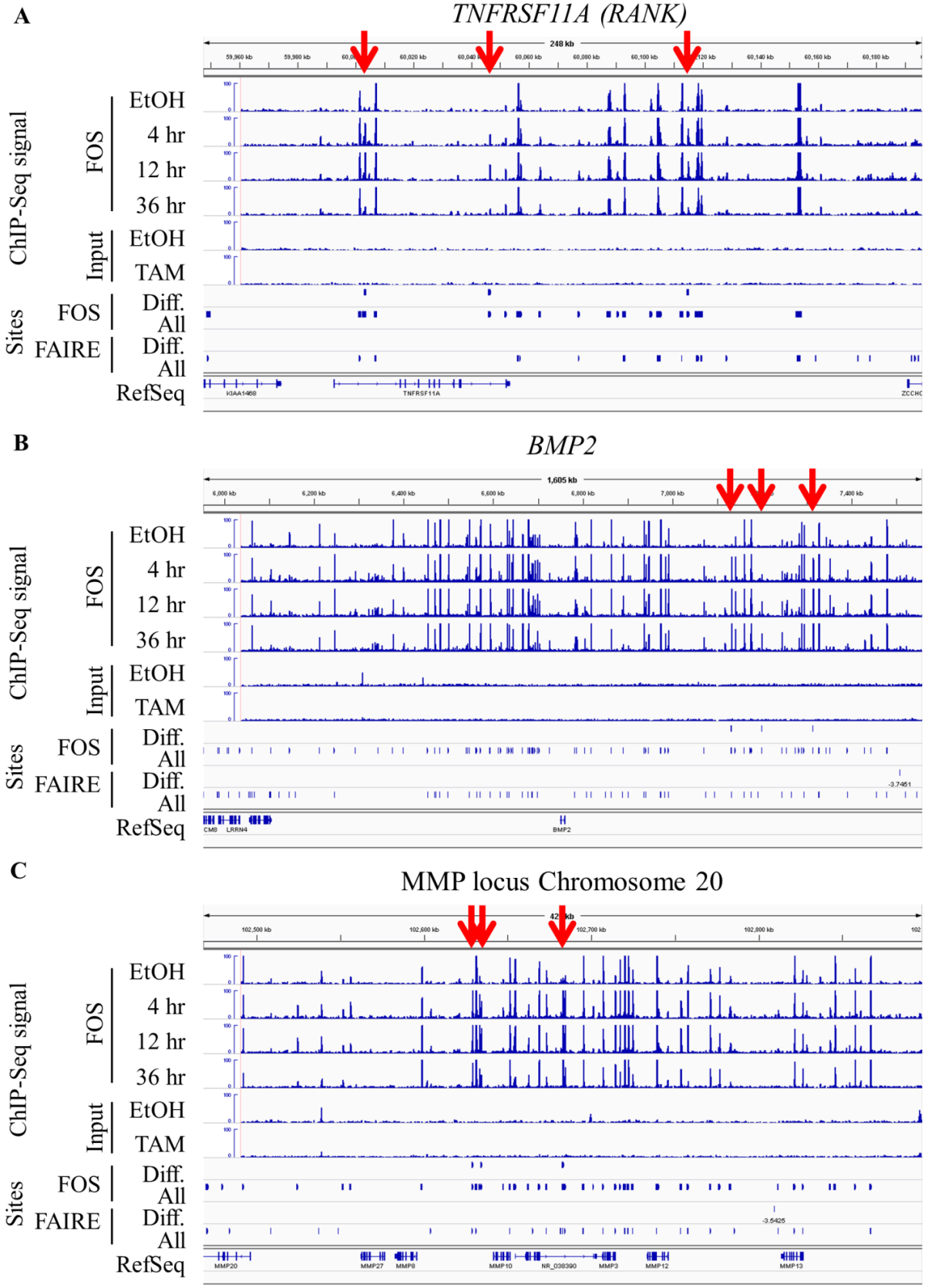


Figure 21: Bone metastasis related genes bound by differential FOS sites

A. B. C. The Ingenuity Systems canonical pathway "*Role of osteoblasts, osteoclasts and chondrocytes in rheumatoid arthritis*" showing genes whose regulatory domain (see *GREAT*, Methods) was bound by transformation-dependent differential FOS sites (highlighted in gray shading).

D. Similar to A. Shows the differential gene expression changes upon ER-Src activation at 4 hr and 24 hr post induction, with *P*-values. Genes not showing expression changes were not significantly altered during transformation.

Figure 21 (Continued)

A

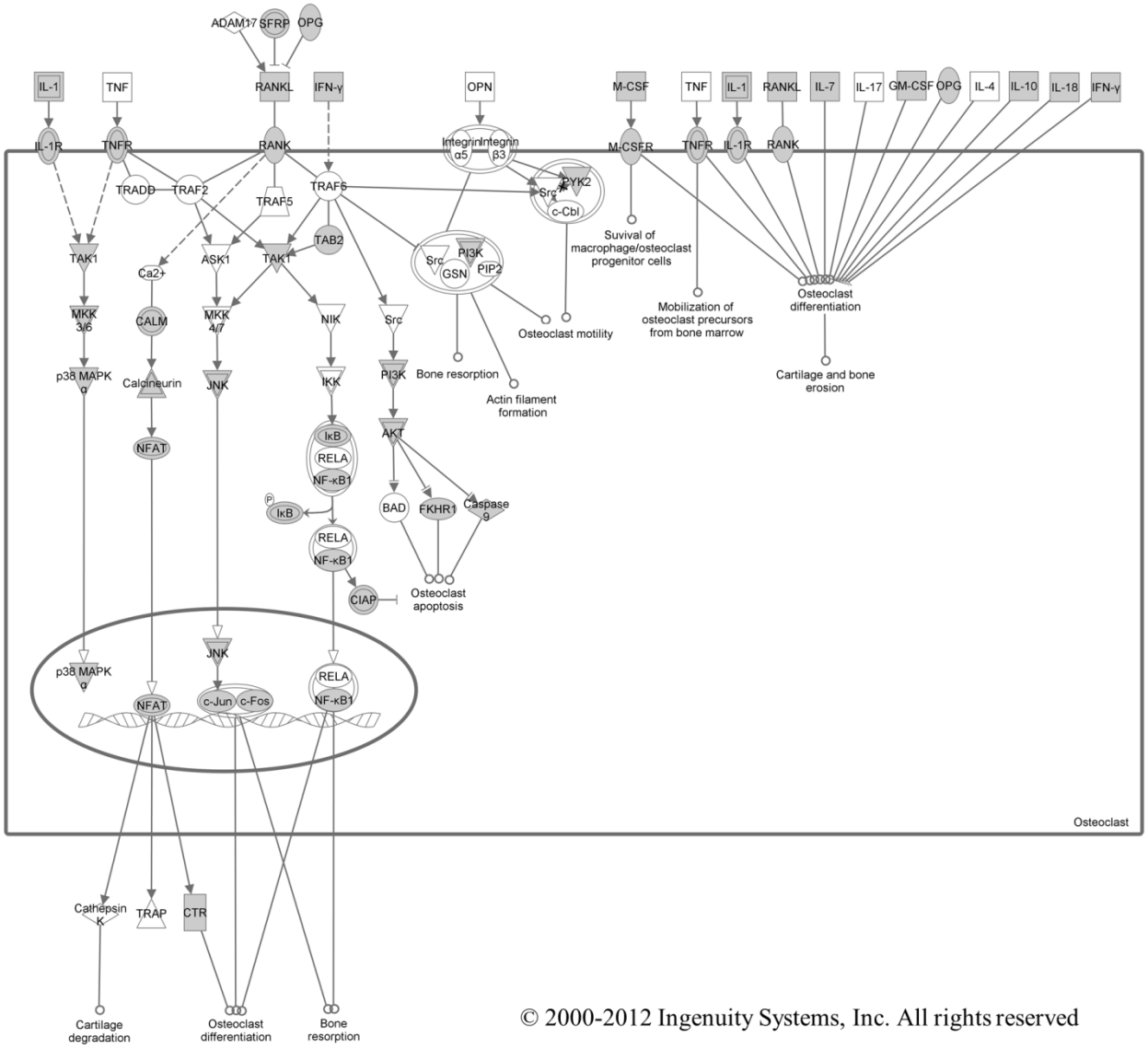


Figure 21 (Continued)

B Role of osteoblasts, osteoclasts and chondrocytes in rheumatoid arthritis

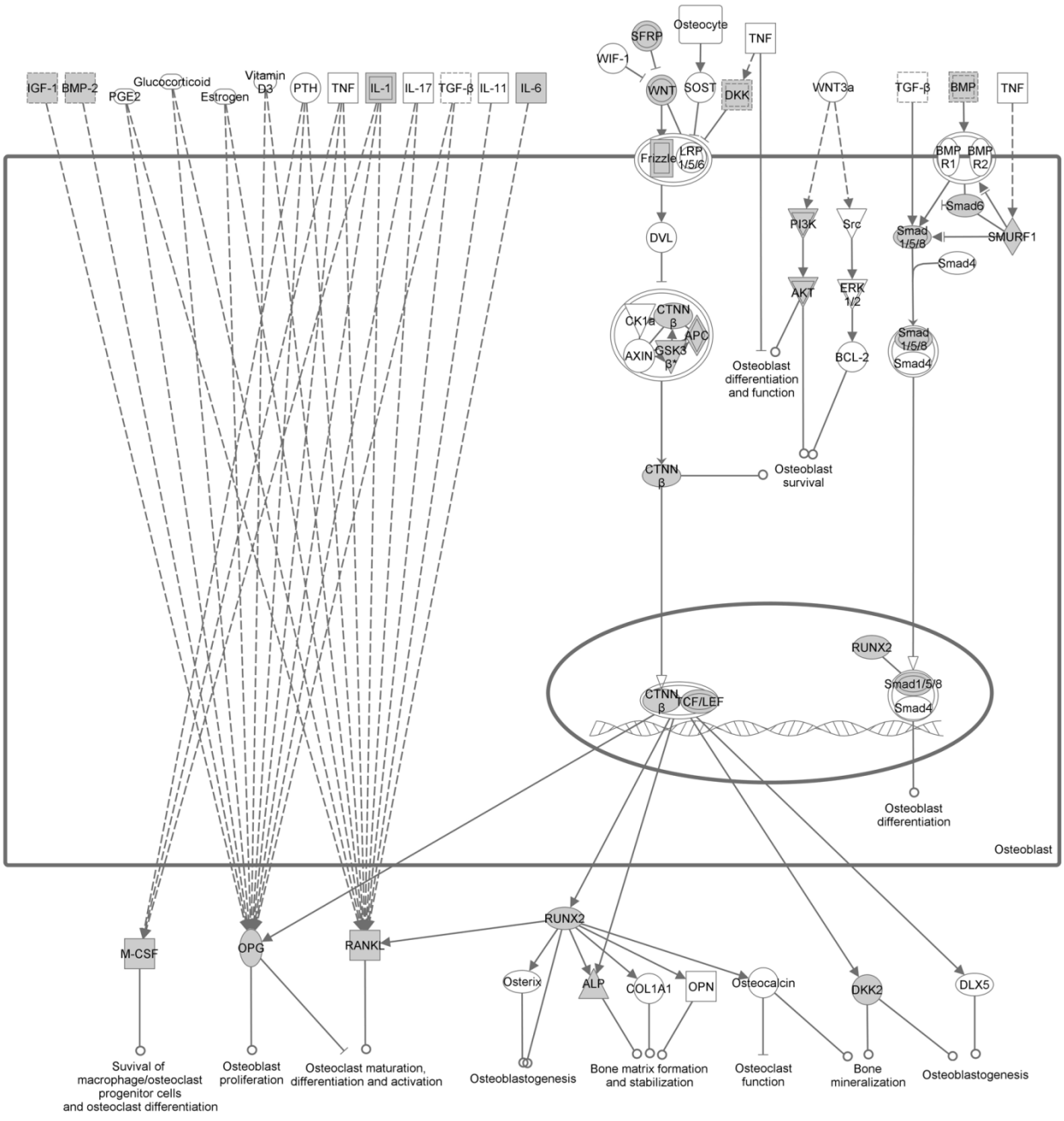


Figure 21 (Continued)

C Role of osteoblasts, osteoclasts and chondrocytes in rheumatoid arthritis

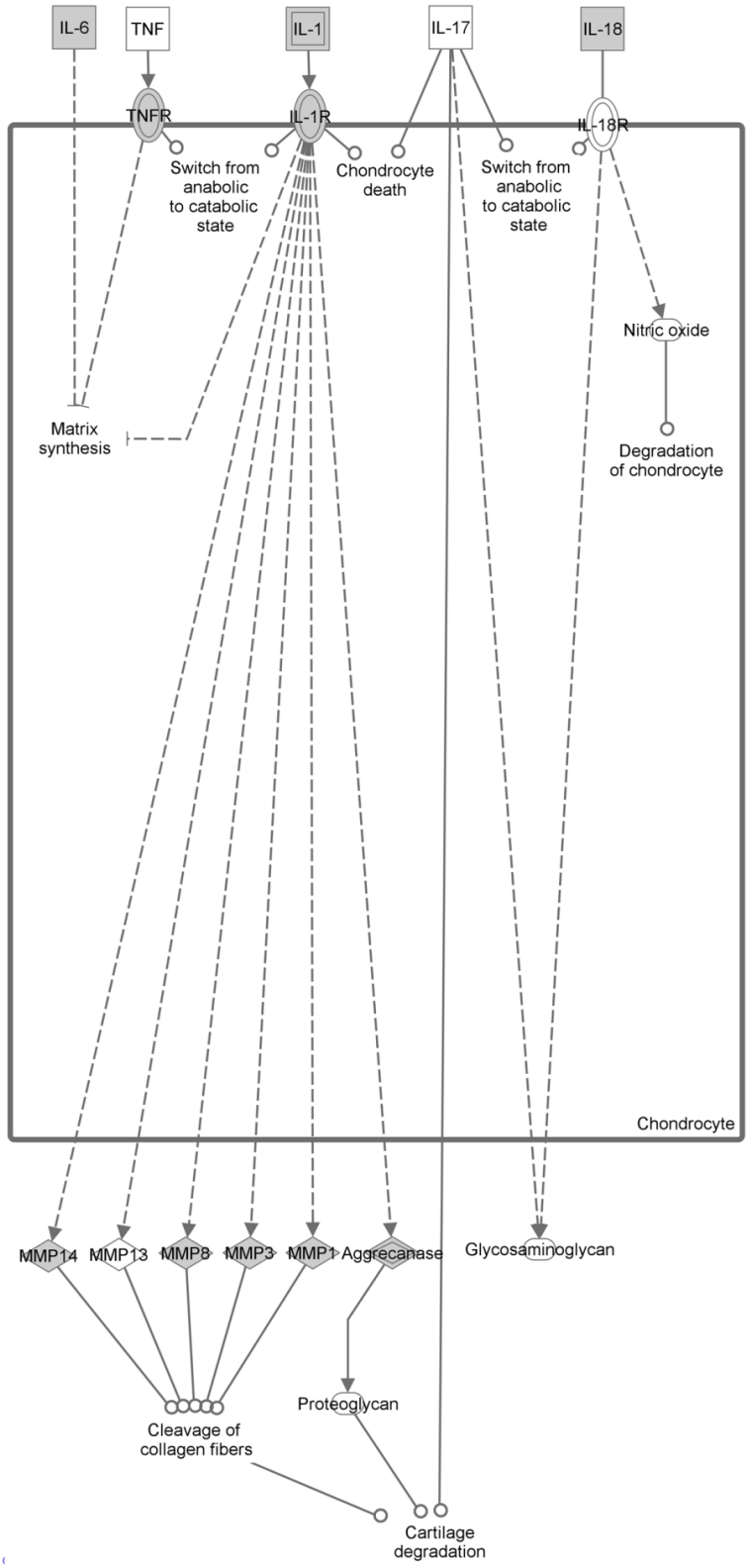


Figure 21 (Continued)

D

Gene	Differentially expressed during transformation			
	4 hr		24 hr	
	log2 FC	*P*-value	log2 FC	*P*-value
ADAMTS4	-	-	-	-
ADAMTS5	-	-	-	-
AKT3	-	-	-	-
ALPL	-	-	-	-
APC	-	-	-	-
BIRC2	-	-	-	-
BIRC3	1.16	1.2E-08	-	-
BMP15	-	-	-	-
BMP2	-	-	1.56	5.7E-09
BMP3	-	-	-	-
BMP4	-	-	-	-
BMP5	-	-	-	-
BMP7	-	-	-	-
BMP8A	-	-	-	-
BMPR1B	-	-	-	-
CALCR	-	-	-	-
CALM1	-	-	-	-
CASP9	-	-	-	-
CSF1	-	-	-	-
CSF1R	-	-	-	-
CSF2	-	-	-	-
CTNNB1	-	-	-	-
DKK1	-0.83	5.5E-08	-1.82	4.4E-13
DKK2	-	-	-	-
FOS	1.20	6.6E-09	2.04	2.3E-12
FOXO1	-0.50	2.4E-05	-	-
FZD1	-	-	-	-
FZD10	-	-	-	-
FZD4	-	-	-	-
FZD7	-	-	-	-
FZD8	-	-	-	-
GSK3B	-	-	-	-
IFNG	-	-	-	-
IGF1	-	-	-	-
IL10	-	-	-	-
IL18	-	-	-1.30	8.3E-10
IL1A	-	-	-	-
IL1B	-	-	-	-
IL1R1	0.56	1.1E-05	1.37	4.3E-11
IL1R2	-	-	2.34	3.3E-10
IL1RAP	0.72	9.9E-06	1.23	4.5E-05
IL1RAPL2	-	-	-	-
IL1RL1	-	-	2.68	3.1E-11
IL1RL2	-	-	-	-
IL33	-	-	6.82	3.3E-09
IL37	-	-	-	-
IL6	2.70	2.5E-08	-	-
IL7	-	-	0.98	3.9E-05
ITGA2	0.65	2.2E-05	1.32	2.0E-09
ITGB1	-	-	-	-

Gene	Differentially expressed during transformation			
	4 hr		24 hr	
	log2 FC	*P*-value	log2 FC	*P*-value
JUN	-	-	1.03	1.2E-08
LEF1	-	-	-0.65	9.8E-05
MAP2K6	-	-	-	-
MAP3K7	-	-	-	-
MAPK14	-	-	0.55	2.1E-05
MAPK8	-	-	-	-
MMP1	-	-	-	-
MMP14	-	-	1.51	6.7E-08
MMP3	-	-	2.24	6.3E-07
MMP8	-	-	-	-
NFAT5	-	-	-	-
NFATC1	-	-	-	-
NFATC2	-	-	-	-
NFKB1	0.60	6.7E-05	-	-
NFKBIA	-	-	-	-
PIK3C2A	-	-	-	-
PIK3C2G	-	-	-	-
PIK3C3	-	-	-	-
PIK3CB	-	-	-	-
PIK3CG	-	-	-	-
PIK3R1	-0.61	3.9E-05	-	-
PIK3R3	-	-	-1.05	9.0E-07
PIK3R4	-	-	-	-
PIK3R5	-	-	-	-
PPP3CA	-	-	-	-
PPP3CC	-	-	-	-
PTK2B	-	-	-	-
RUNX2	-	-	0.69	4.6E-05
SFRP1	-0.55	3.5E-05	-1.59	1.7E-11
SMAD1	-	-	-	-
SMAD6	-	-	-	-
SMAD9	-	-	-	-
SMURF1	-	-	-	-
TAB2	-	-	-	-
TCF4	-	-	-	-
TCF7L2	-	-	-	-
TNFRSF11A	-	-	-	-
TNFRSF11B	-	-	2.10	3.9E-09
TNFSF11	-	-	-	-
WNT11	-	-	-	-
WNT16	-	-	-	-
WNT2	-	-	-	-
WNT2B	-	-	-	-
WNT3	-	-	-	-
WNT7A	-	-	-	-
WNT7B	-	-	-	-
WNT8B	-	-	-	-
WNT9B	-	-	-	-
ADAMTS9	1.46	7.0E-05	2.24	5.2E-11

upon siSTAT3 treatment (Figure 22, B), and the expression of known STAT3 activated genes (*CD46*, *FOS*, *SERPIN3A*, *SOCS3* and *VEGFA*) were effectively perturbed (*CD46* less so) (Figure 22, C). To assess the transformation process, genes known to be induced during transformation in MCF10A-ER-Src cells were checked in the microarray data (Figure 22, B). In addition, microscopic examination of cell morphology confirmed that transformation had occurred in the siSCM TAM treated cells, but not in the siSTAT3 TAM or any EtOH treated cells (not shown). A list of the top 20 genes at 4 hr or 24 hr that were STAT3-dependent or STAT3-independent and transformation-dependent are listed in Figure 23, A and B.

A comparison of the gene ontology terms that were significantly enriched in those genes that were differentially regulated during transformation in a STAT3-dependent or independent manner revealed that STAT3 was more important for regulating genes involved in the inflammatory response and less important in regulating genes that were involved in cellular metabolism, especially at the 24 hr time point. Note the presence of "*Butanoate metabolism*", "*Galactose metabolism*", "*Starch and sucrose metabolism*" and "*Aminosugars metabolism*" in the STAT3-independent 24 hr time point (Figure 24, B) and the absence of metabolism related terms in the STAT3-dependent 24 hr time point (Figure 24, A). *IPA* pathway analysis (see Methods) revealed that NFκB may be controlling expression of the metabolism genes as a network linking carbohydrate metabolism, drug metabolism, and small molecule biochemistry was the top network (Figure 18, D) linking the differentially expressed genes at 24 hr post ER-Src induction that were STAT3-independent, with NFκB as a central effector.

STAT3 activity accounts for a large proportion of differential gene regulation

During transformation, 384 and 1472 genes were differentially regulated (*P*-value <= 10^{-4}, log2 fold change of >= 0.5) at 4 hr and 24 hr post ER-Src activation. Specific functional inactivation

Figure 22: Knockdown of STAT3 during transformation

A. Western blots of protein extracts from TAM treated MCF10A-ER-Src cells done in parallel to RNA samples used for expression microarray analysis. STAT3 protein levels were reduced > 25 fold upon siSTAT3 knockdown by 24 hr.

B. Normalized RNA microarray expression values of STAT3 and four genes known to be differentially regulated during transformation.

C. Similar to A, except showing RNA levels of genes known to be regulated by STAT3 indicating functional inactivation of STAT3 was achieved.

D. Similar to A, except showing expression levels of 4 "housekeeping" genes, expressed at different levels, indicating that arrays were normalized.

E. The numbers of genes differentially up or down regulated during transformation at 4 hr and 24 hr post ER-Src activation and their dependence on STAT3 (see Methods).

Figure 22 (Continued)

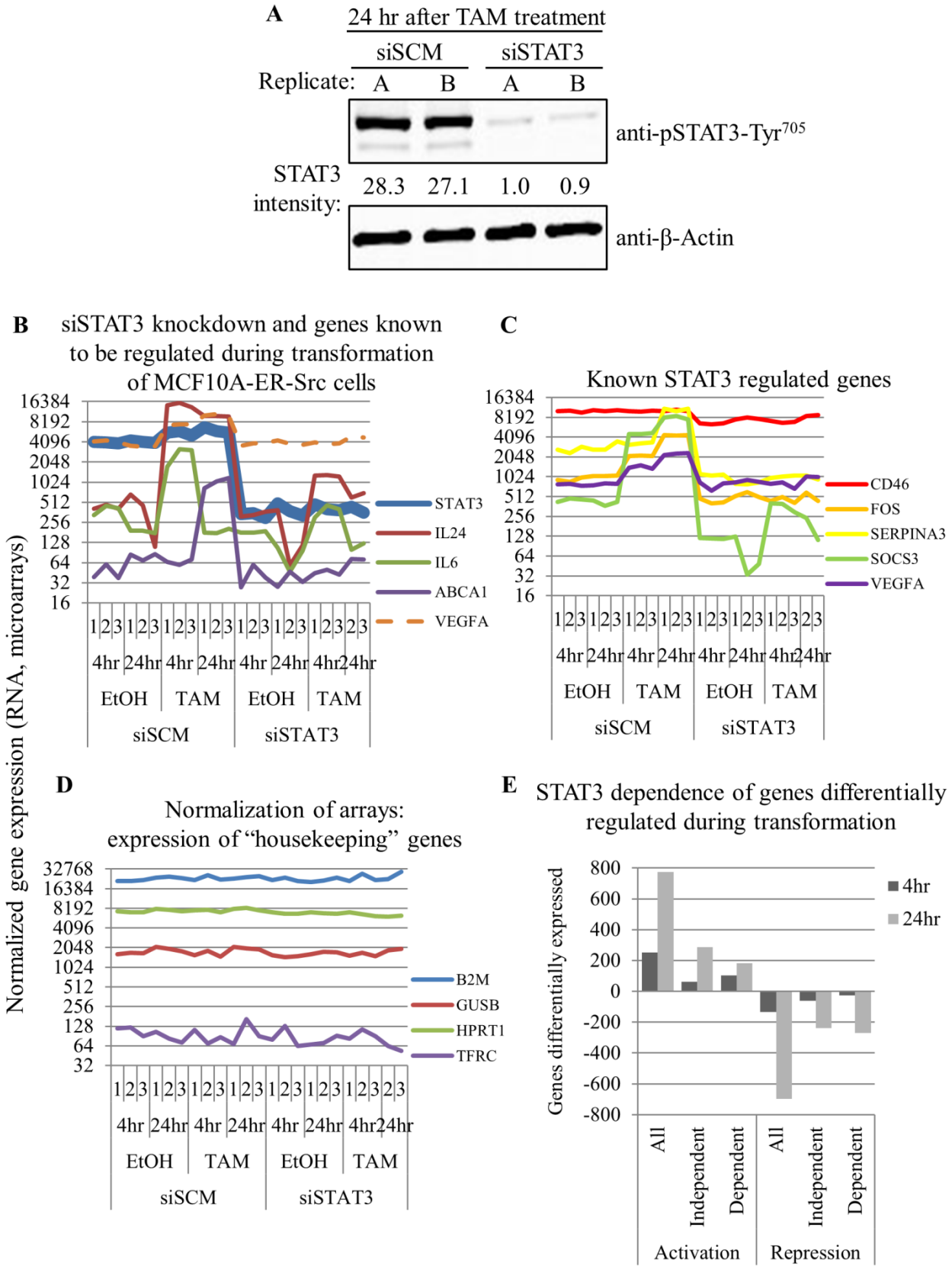


Figure 23: Top 20 differentially regulated genes during transformation and siSTAT3 treatment

A. At 4 hr post ER-Src activation (+TAM).

B. At 24 hr post ER-Src activation (+TAM).

Figure 23 (Continued)

A

Top 20 genes activated by STAT3 at 4 hr

Symbol	Name	Molecular function	TAM + siSCM log2 FC	TAM + siSTAT3 log2 FC
SOCS3	Suppressor of cytokine signaling 3	Protein kinase inhibitor activity	3.2	-2.8
SERPINA1	Serpin peptidase inhibitor, clade A, member 1	Endopeptidase inhibitor activity	2.8	-2.8
STC1	Stanniocalcin 1	Hormone activity	3.6	-2.7
SERPINB13	Serpin peptidase inhibitor, clade B, member 13	Endopeptidase inhibitor activity	3.0	-2.5
TNC	Tenascin C	Unknown	2.5	-2.0
RGS2	Regulator of G-protein signaling 2, 24kDa	GTPase activator activity	2.0	-2.0
SLC43A2	Solute carrier family 43, member 2	Amino acid transmembrane transporter activity	0.8	-1.8
TRIM15	Tripartite motif-containing 15	Zinc ion binding	2.1	-1.7
SERPINB1	Serpin peptidase inhibitor, clade B, member 1	Endopeptidase inhibitor activity	2.0	-1.7
LOC346887	(similar to solute carrier family 16, member 14)	Unknown	1.5	-1.7
NAMPT	Nicotinamide phosphoribosyltransferase	Cytokine activity, nicotinamide phosphoribosyltransferase activity,	1.9	-1.6
FAM46C	Family with sequence similarity 46, member C	Unknown	1.7	-1.6
CLEC2B	C-type lectin domain family 2, member B	Carbohydrate binding	1.7	-1.6
SAA4	Serum amyloid A4, constitutive	Unknown	1.9	-1.5
FLJ36031	Hypothetical protein FLJ36031	Unknown	1.7	-1.4
MCC	Mutated in colorectal cancers	Unknown	1.1	-1.4
NEDD4L	Neural precursor cell expressed, developmentally down-regulated 4-like	Ion channel inhibitor activity	1.7	-1.4
PPAP2B	Phosphatidic acid phosphatase type 2B	Lipid phosphatase activity	1.1	-1.4
GPC6	Glypican 6	Proteoglycan binding	1.1	-1.3
SH3TC1	SH3 domain and tetratricopeptide repeats 1	Unknown	1.2	-1.3

Top 20 genes repressed by STAT3 at 4 hr

Symbol	Name	Molecular function	TAM + siSCM log2 FC	TAM + siSTAT3 log2 FC
ABCB9	ATP-binding cassette, sub-family B (MDR/TAP), member 9	ATPase activity, coupled to transmembrane movement of substances	-2.1	4.1
ARPP21	Cyclic AMP-regulated phosphoprotein, 21 kD	Unknown	-1.2	2.7
LOC100129166	N/A	Unknown	-0.8	2.6
AJAP1	Adherens junctions associated protein 1	Unknown	-1.2	2.3
SIRPD	Signal-regulatory protein delta	Unknown	-1.4	2.0
SEMA3C	Sema domain, Ig domain, short basic domain, secreted, 3C	Semaphorin receptor binding,	-1.2	1.7
ZNF238	Zinc finger protein 238	Transcription factor activity	-1.1	1.6
EXOC6B	Exocyst complex component 6B	Unknown	-1.1	1.6
SOX9	SRY-box 9	Transcription factor activity	-1.7	1.4
INHBA	Inhibin, beta A	Cytokine activity	-0.6	1.3
C11ORF17	Chromosome 11 open reading frame 17	Unknown	-0.1	1.2
PAX8	Paired box 8	Transcription factor activity	-0.2	1.2
TIMP3	TIMP metallopeptidase inhibitor 3	Metalloendopeptidase inhibitor activity	-1.5	1.2
TXNIP	Thioredoxin interacting protein	Enzyme inhibitor activity	-0.7	1.1
LAMC2 5	Laminin, gamma 2	Glycosaminoglycan binding	-0.7	1.1
(237435_at)	N/A	Unknown	-1.3	1.0
CPEB2	Cytoplasmic polyadenylation element binding protein 2	ssRNA binding	-0.9	1.0
ZBTB16	Zinc finger and BTB domain containing 16	Transcription factor activity	-1.1	1.0
SLC46A1	Solute carrier family 46 (folate transporter), member 1	Vitamin transporter activity	-0.2	1.0
ERRFI1	ERBB receptor feedback inhibitor 1	GTPase activator activity	-0.2	0.9

Figure 23 (Continued)

B

Top 20 genes activated by STAT3 at 24 hr

Symbol	Name	Molecular function	siSCM + TAM log2 FC	siSTAT3 + TAM log2 FC
RASD1	RAS, dexamethasone-induced 1	GTPase activity	6.9	-6.4
GZMB	Granzyme B	Peptidase activity	7.6	-6.3
PAEP	Progestagen-associated endometrial protein	Unknown	5.6	-6.2
IL33	Interleukin 33	Cytokine activity	6.8	-5.9
CRP	C-reactive protein, pentraxin-related	Calcium ion binding, lipoprotein binding, choline binding	5.1	-5.4
MYLK3	Myosin light chain kinase 3	Protein kinase activity	3.1	-5.3
OLFM4	Olfactomedin 4	Cell adhesion	5.6	-5.1
STC1	Stanniocalcin 1	Hormone activity	6.8	-5.0
ECSCR	Endothelial cell-specific chemotaxis regulator	Unknown	4.9	-4.7
RHOU	Ras homolog gene family, member U	GTPase activity	4.7	-4.5
CUEDC1	CUE domain containing 1	Ubiquitin system component	5.0	-4.3
TNC	Tenascin C	Cell adhesion	5.2	-4.2
NPAS1	Neuronal PAS domain protein 1	Transcription factor activity	0.9	-4.1
ADAM19	ADAM metallopeptidase domain 19	Metalloendopeptidase activity	5.4	-4.0
HBA2	Hemoglobin, alpha 2; hemoglobin, alpha 1	Heme binding	3.5	-3.8
FGL1	Fibrinogen-like 1	Unknown	4.5	-3.8
CLEC2B	C-type lectin domain family 2, member B	Carbohydrate binding,	5.1	-3.7
CTSL1P8	Cathepsin L1 pseudogene 8	Unknown	3.9	-3.6
CHRNA9	Cholinergic receptor, nicotinic, alpha 9	Ligand-gated ion channel activity	3.6	-3.5
ENTPD3	Ectonucleoside triphosphate diphosphohydrolase 3	Nucleoside-diphosphatase activity	5.1	-3.5

Top 20 genes repressed by STAT3 at 24 hr

Symbol	Name	Molecular function	TAM + siSCM log2 FC	TAM + siSTAT3 log2 FC
DBNDD2	Dysbindin domain containing 2	Unknown	-2.0	4.9
MUC16	Mucin 16, cell surface associated	Cell adhesion	-4.1	4.7
ATP6V1C2	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C2	Hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	-2.7	4.2
SPRR3	Small proline-rich protein 3	Structural molecule activity	-1.8	3.6
HTRA4	HtrA serine peptidase 4	Endopeptidase activity	-3.7	3.3
CLDN1	Claudin 1	Structural molecule activity	-2.4	3.2
IGFBP3	Insulin-like growth factor binding protein 3	Protein tyrosine phosphatase activator activity	-1.0	3.2
TNF	Tumor necrosis factor (TNF superfamily, member 2)	Cytokine activity	-0.8	3.2
CTXN1	Cortexin 1	Transmembrane protein	-3.0	3.2
KRT15	Keratin 15	Structural molecule activity	-3.0	3.1
SPRR2B	Small proline-rich protein 2B	Unknown	-1.0	3.1
EPHB3	EPH receptor B3	Protein kinase activity	-2.8	3.0
BDKRB1	Bradykinin receptor B1	Bradykinin receptor activity	-2.7	2.9
CD24L4	CD24 molecule; CD24 molecule-like 4	Protein kinase activator activity	-2.6	2.9
MACC1	Metastasis associated in colon cancer 1	Growth factor activity	-2.5	2.8
IFIT1	Interferon-induced protein with tetratricopeptide repeats 1	Unknown	-1.4	2.8
GRHL3	Grainyhead-like 3	Transcription regulation	-2.8	2.8
IFI27	Interferon, alpha-inducible protein 27	Unknown	0.4	2.8
MUM1	Melanoma associated antigen (mutated) 1	Unknown	-2.5	2.8
CXCL11	Chemokine (C-X-C motif) ligand 11	Cytokine activity	-1.9	2.7

Figure 24: Gene ontology terms associated with STAT3-dependent genes during transformation

A. Genes significantly differentially regulated during transformation that are significantly affected by STAT3 knockdown.

B. Genes significantly differentially regulated during transformation that are not significantly affected by STAT3 knockdown.

A

STAT3-dependent

4 hr		24 hr	
Pathway	***P*-value**	**Pathway**	***P*-value**
Acute Phase Response Signaling	4.0E-03	IL-6 Signaling	7.9E-08
Clathrin-mediated Endocytosis	6.2E-03	IL-10 Signaling	2.2E-07
IL-6 Signaling	6.3E-03	Biosynthesis of Steroids	2.6E-06
ILK Signaling	6.3E-03	PPAR Signaling	3.0E-06
IL-10 Signaling	9.1E-03	LXR/RXR Activation	3.5E-06
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid	1.2E-02	Acute Phase Response Signaling	5.5E-06
IL-9 Signaling	1.9E-02	Hepatic Fibrosis / Hepatic Stellate Cell Activation	6.5E-06
Oncostatin M Signaling	1.9E-02	LPS/IL-1 Mediated Inhibition of RXR Function	9.1E-06
IL-17A Signaling in Fibroblasts	2.0E-02	Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	4.1E-05
p53 Signaling	2.1E-02	Oncostatin M Signaling	2.5E-04
Atherosclerosis Signaling	3.9E-02	Hepatic Cholestasis	2.8E-04
LXR/RXR Activation	4.2E-02	PPAR/RXR Activation	2.8E-04
LPS/IL-1 Mediated Inhibition of RXR Function	4.8E-02	Cholecystokinin/Gastrin-mediated Signaling	4.7E-04
IL-12 Signaling and Production in Macrophages	5.2E-02	Sertoli Cell-Sertoli Cell Junction Signaling	4.8E-04
Erythropoietin Signaling	6.8E-02	VDR/RXR Activation	7.4E-04
JAK/Stat Signaling	6.8E-02	Type I Diabetes Mellitus Signaling	9.3E-04
Growth Hormone Signaling	6.9E-02	IGF-1 Signaling	1.3E-03
Renal Cell Carcinoma Signaling	7.2E-02	p38 MAPK Signaling	1.5E-03
Prolactin Signaling	7.8E-02	Graft-versus-Host Disease Signaling	1.7E-03
VEGF Family Ligand-Receptor Interactions	8.3E-02	Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	1.8E-03

Figure 24 (Continued)

B

STAT3-independent

4 hr		24 hr	
Pathway	***P*-value**	**Pathway**	***P*-value**
IL-12 Signaling and Production in Macrophages	1.9E-04	Butanoate Metabolism	1.6E-03
Activation of IRF by Cytosolic Pattern Recognition Receptors	2.5E-04	Complement System	5.2E-03
Pancreatic Adenocarcinoma Signaling	3.8E-04	Bile Acid Biosynthesis	5.9E-03
iNOS Signaling	6.3E-04	IL-8 Signaling	6.2E-03
PI3K/AKT Signaling	6.6E-04	Androgen and Estrogen Metabolism	6.6E-03
Colorectal Cancer Metastasis Signaling	1.3E-03	Actin Nucleation by ARP-WASP Complex	6.9E-03
Role of IL-17A in Arthritis	1.4E-03	Tumoricidal Function of Hepatic Natural Killer Cells	1.4E-02
p53 Signaling	1.5E-03	Pentose and Glucuronate Interconversions	1.4E-02
Chronic Myeloid Leukemia Signaling	1.5E-03	Starch and Sucrose Metabolism	1.4E-02
Hepatic Fibrosis / Hepatic Stellate Cell Activation	1.6E-03	p53 Signaling	1.5E-02
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	1.7E-03	NRF2-mediated Oxidative Stress Response	1.5E-02
TNFR2 Signaling	2.1E-03	Galactose Metabolism	1.7E-02
CD40 Signaling	2.5E-03	PDGF Signaling	1.8E-02
Retinoic acid Mediated Apoptosis Signaling	2.5E-03	HGF Signaling	2.0E-02
Inositol Phosphate Metabolism	2.7E-03	Aminosugars Metabolism	2.0E-02
Type I Diabetes Mellitus Signaling	3.0E-03	Arginine and Proline Metabolism	2.0E-02
PTEN Signaling	3.3E-03	Cholecystokinin/Gastrin-mediated Signaling	2.1E-02
Interferon Signaling	3.4E-03	Role of Tissue Factor in Cancer	3.0E-02
Molecular Mechanisms of Cancer	3.5E-03	Estrogen-Dependent Breast Cancer Signaling	4.0E-02
Small Cell Lung Cancer Signaling	3.7E-03	Extrinsic Prothrombin Activation Pathway	4.5E-02

of STAT3 revealed a large proportion, at least 34% (n = 129) and 31% (n = 451) respectively, to be directly or indirectly transcriptionally dependent on STAT3 (Figure 22, E). Genes activated during transformation, but not those that were repressed, correlated with both the number of differential STAT3 bound loci and the fold change in STAT3 ChIP signal, indicating that STAT3 directly regulated only genes that were activated during transformation and not those that were repressed (Figure 25, A-H).

Motif analysis of the promoters of STAT3 regulated and/or transformation specific differentially regulated genes did not find the STAT3 motif significantly enriched in the promoters of early (4 hr) or late (24 hr) responsive genes (Figure 26). This may reflect the general promoter distal positioning of STAT3 in the genome. STAT3 may be cooperating with factors that bind the RXRα DNA motif to activate genes early during transformation. The RXRα motif was found significantly enriched in transformation activated promoters, and in STAT3-dependent promoters, and was the top motif in the promoters of genes differentially regulated at 4 hr, though it did not reach statistical significance in this latter category (Figure 26).

Candidate TFs regulating transformation repressed genes

Based on the above observations from STAT3 knockdown experiments, and the fact that the STAT3 motif was conspicuously absent from the promoters of repressed genes, STAT3 indirectly regulated at least 35% (n = 295) of those genes that were repressed during transformation. The only motif significantly enriched within STAT3-dependent transformation repressed genes was that for INSM1 (Figure 26). Neither *INSM1* nor *INSM2* RNA was expressed in MCF10A-ER-Src cells (not shown) and were therefore unlikely candidates. A novel approach to discern the transcriptional regulators of a gene set, introduced by Ingenuity Systems, looks for a statistically significant overlap with the genes that are experimentally validated functional

Figure 25: Association of transformation induced chromatin bound STAT3 with transformation-dependent differential gene expression

A. All genes differentially expressed during transformation were considered, separated into up- and down-regulated genes by siSTAT3 treatment, and sorted by the probability of differential gene regulation by siSTAT3. Plotted are: the number of transformation differential STAT3 loci (per kb, per region) that occurred at 4 hr post ER-Src induction within proximal promoter regions (+/- 2.5 kbp about TSS) or distal regions (+/- 50 kbp from TSS, excluding the proximal promoter region); and, the associated fold change in gene expression upon siSTAT3 treatment. Pie charts indicate the percentage of the top 500 regions that contained a differential STAT3 site at 4 hr post ER-Src induction.
B. Similar to A, except fold change in STAT3 ChIP-Seq signal at each region over EtOH treated samples is plotted.
C. Similar to A, except 36 hr STAT3 ChIP samples were used and gene expression data was for the 24 hr time-point.
D. Similar to C, except fold change in 36 hr STAT3 ChIP-Seq signal at each region over EtOH treated samples is plotted and gene expression data was for the 24 hr time-point.
E. Similar to A, except genes are sorted by probability of transformation-dependent differential gene expression.
F. Similar to B, except genes are sorted by probability of transformation-dependent differential gene expression.
G. Similar to A, except 36 hr STAT3 ChIP samples were used, gene expression data was for the 24 hr time-point, and genes are sorted by probability of transformation-dependent differential gene expression.

H. Similar to B, except 36 hr STAT3 ChIP samples were used, gene expression data was for the 24 hr time-point, and genes are sorted by probability of transformation-dependent differential gene expression.

Figure 25 (Continued)

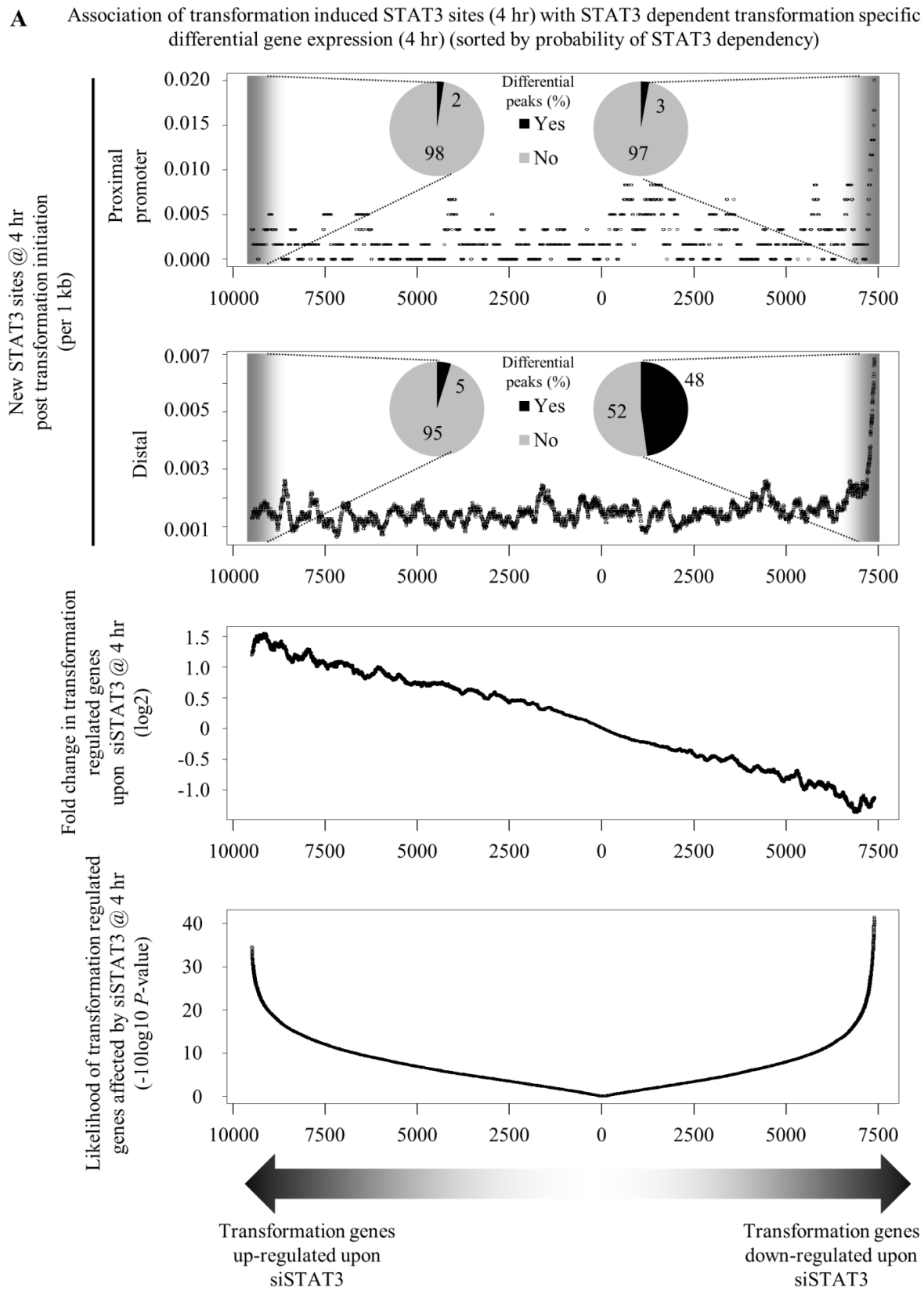


Figure 25 (Continued)

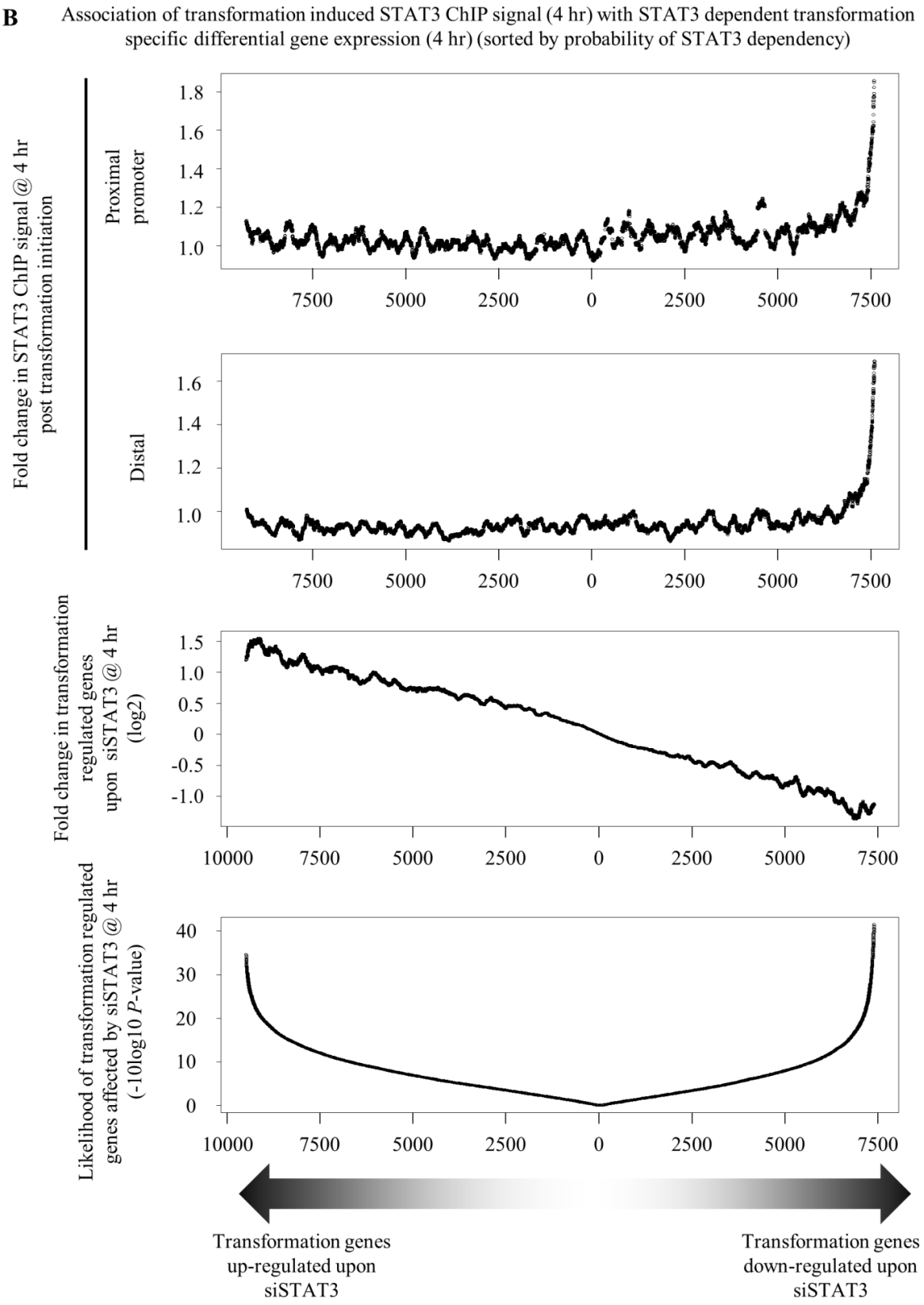


Figure 25 (Continued)

C Association of transformation induced STAT3 sites (36 hr) with STAT3 dependent transformation specific differential gene expression (24 hr) (sorted by probability of STAT3 dependency)

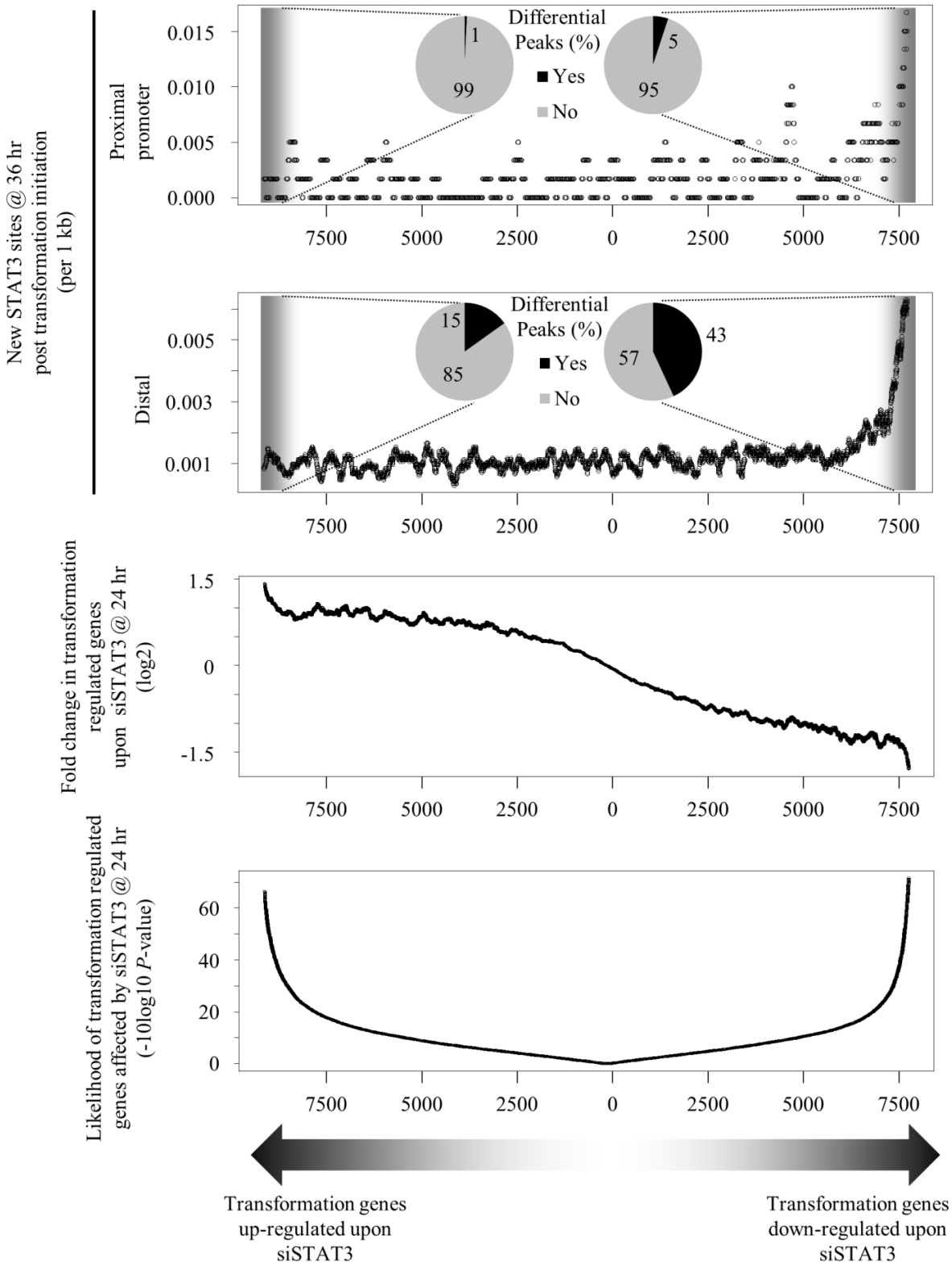


Figure 25 (Continued)

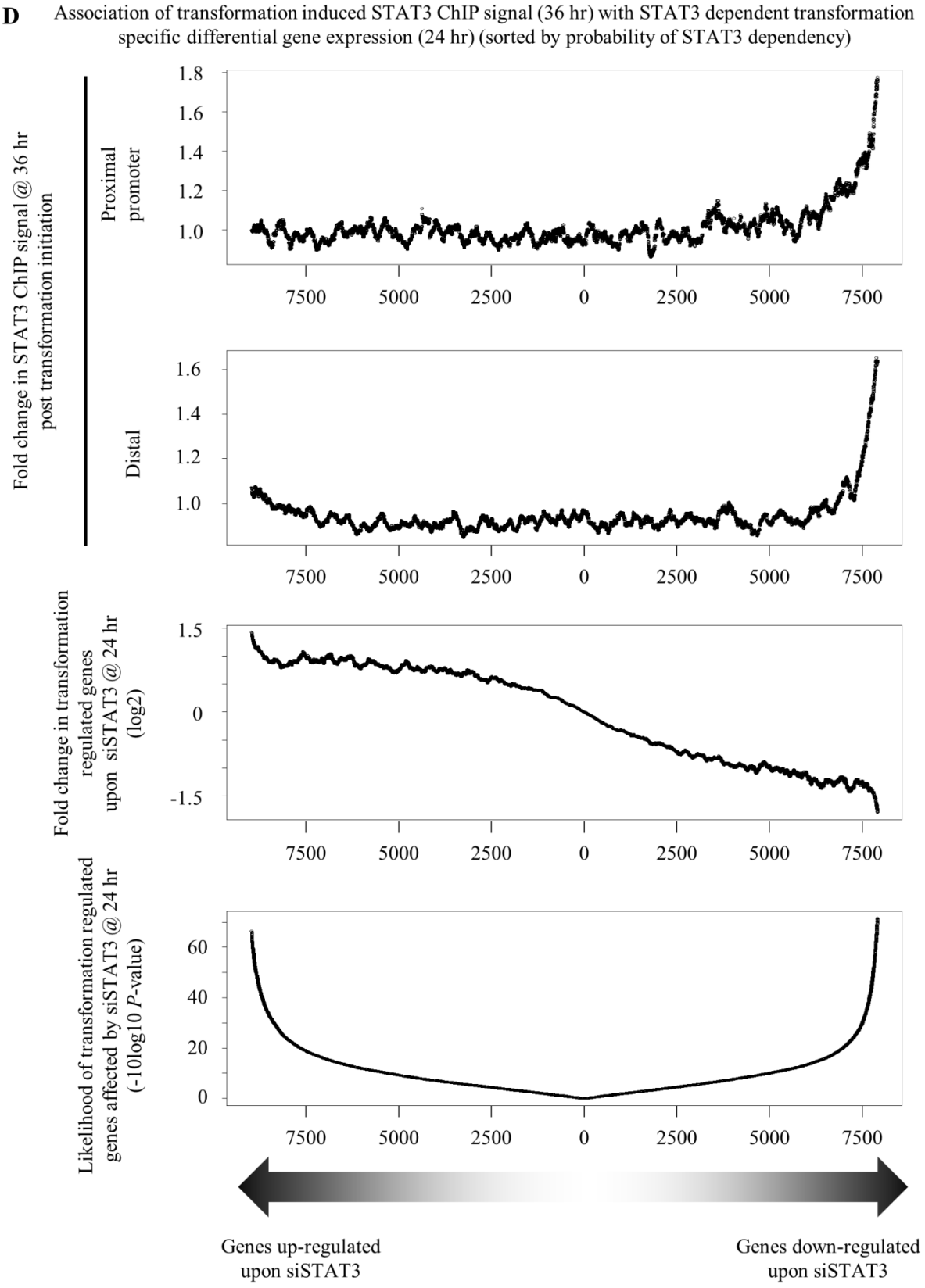


Figure 25 (Continued)

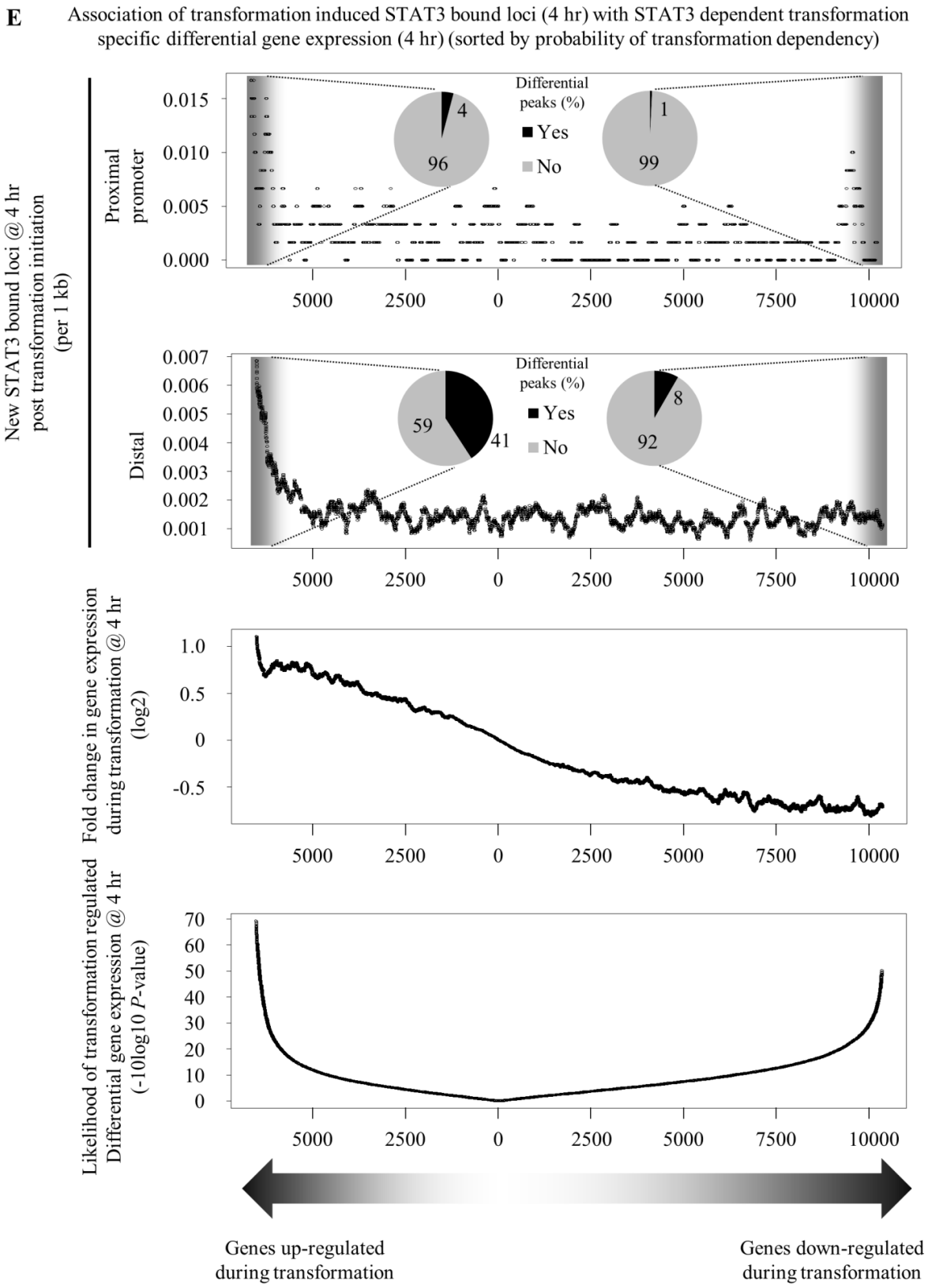


Figure 25 (Continued)

F Association of transformation induced STAT3 ChIP signal (4 hr) with STAT3 dependent transformation specific differential gene expression (4 hr) (sorted by probability of transformation dependency)

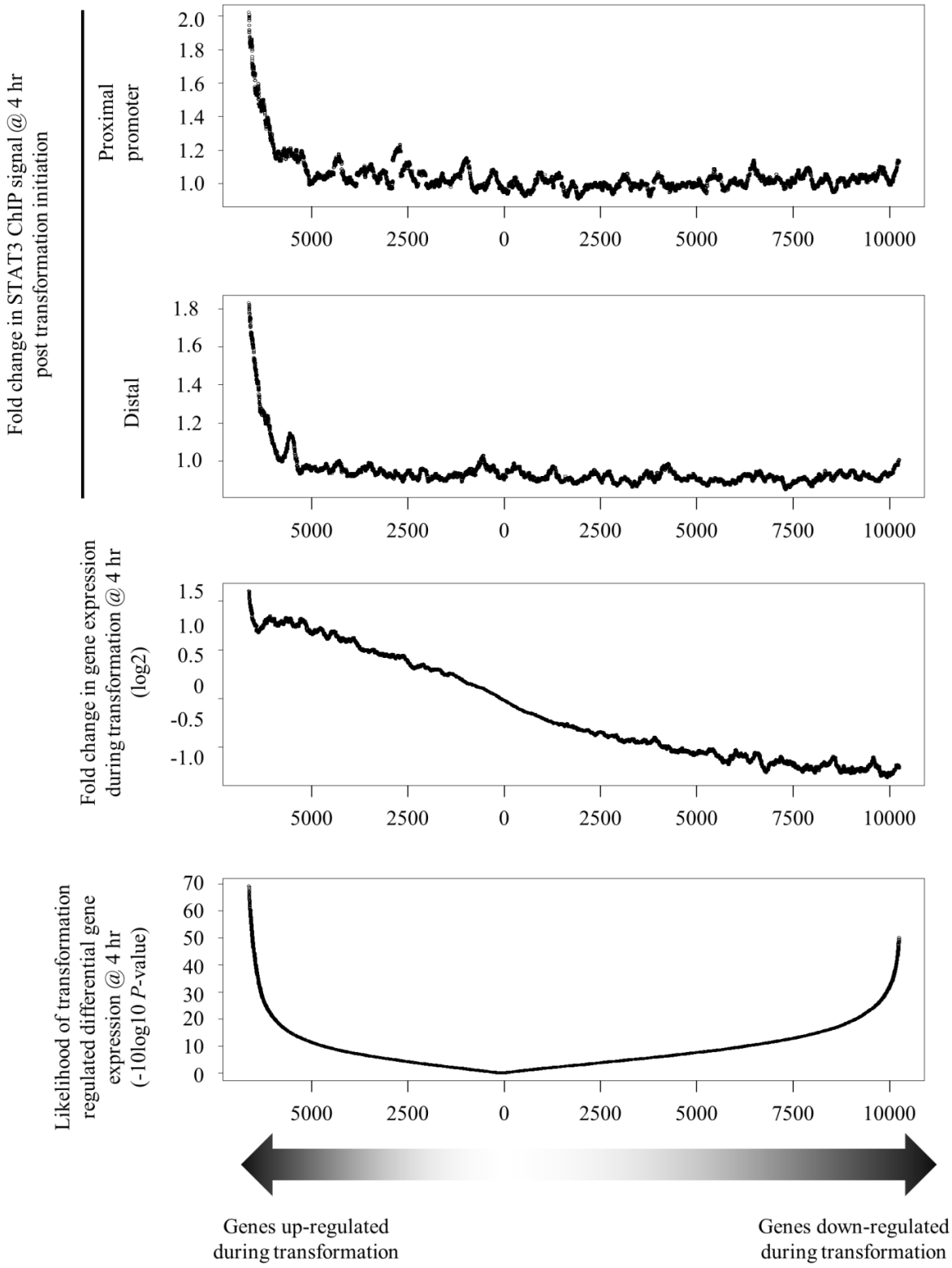


Figure 25 (Continued)

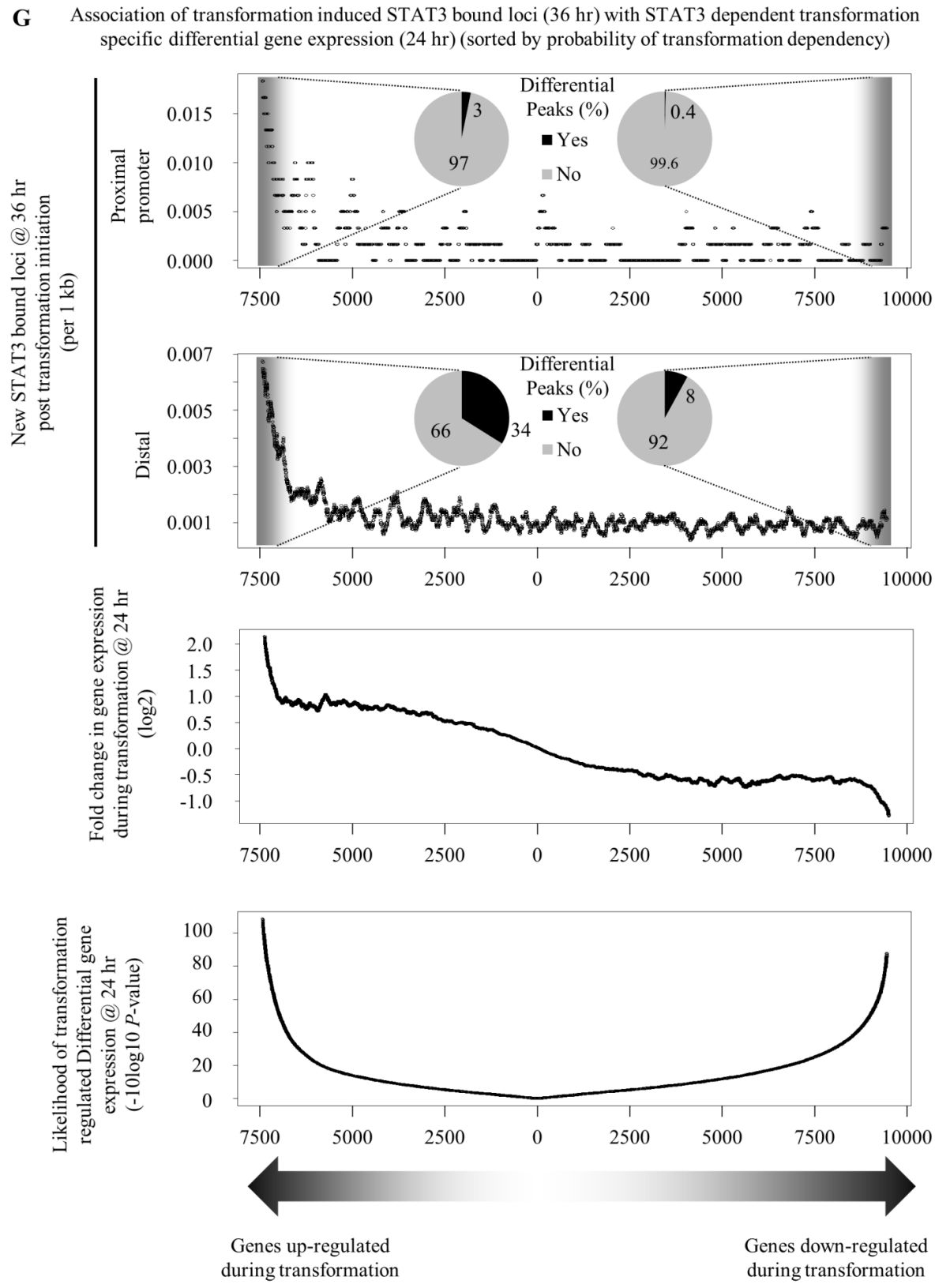


Figure 25 (Continued)

H Association of transformation induced STAT3 ChIP signal (36 hr) with STAT3 dependent transformation specific differential gene expression (24 hr) (sorted by probability of transformation dependency)

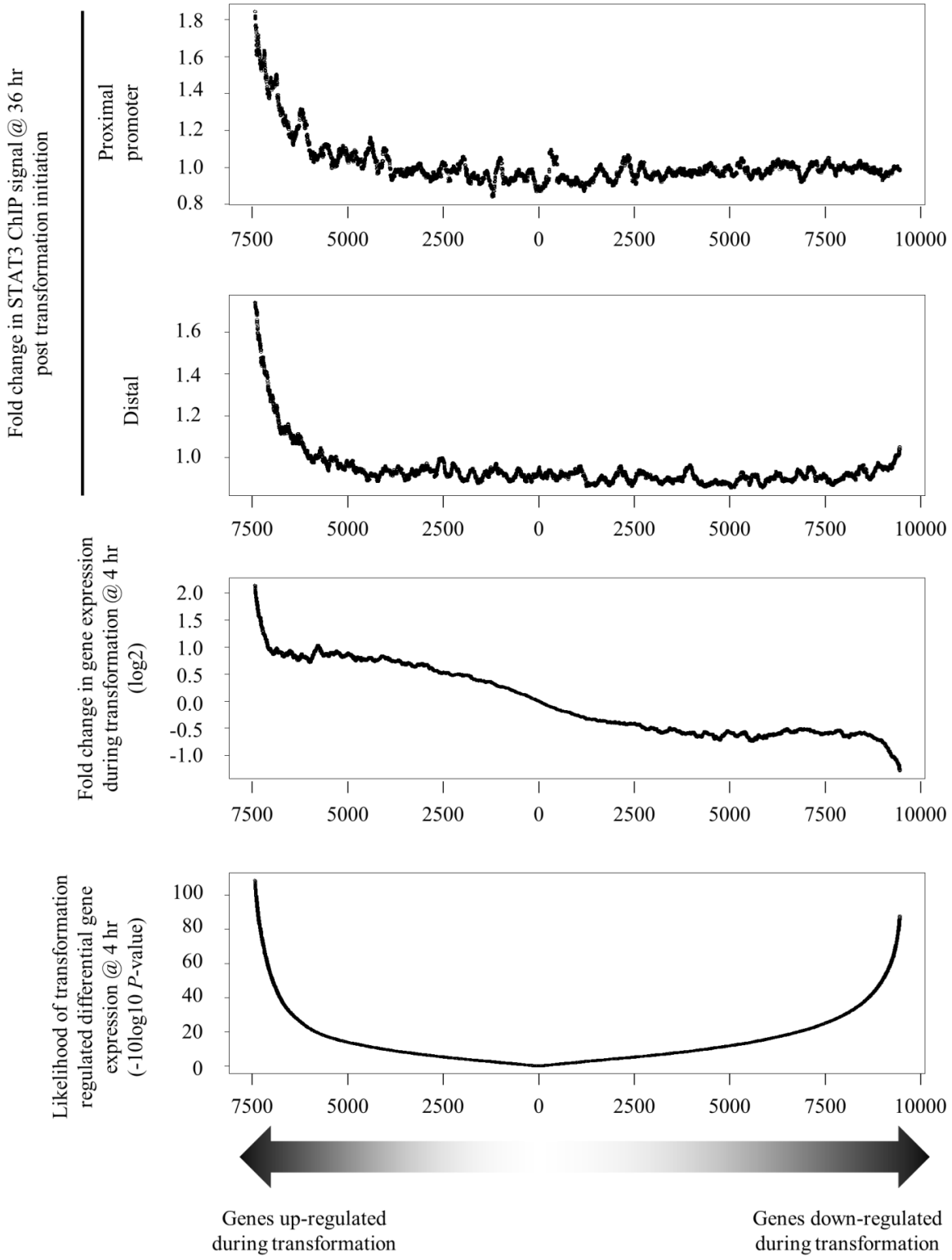


Figure 26: DNA motifs enriched in promoters of differentially expressed genes during transformation

The promoter regions (-1000 bp:+0 bp from TSS) of transformation-dependent differentially expressed genes were searched for known DNA motifs contained in the JASPAR_CORE_2009 database. Different background sets were used depending on the specific question: "vs. RefSeq promoters" highlights motifs enriched compared to all promoters; "vs. 24 hr" highlights motifs that are enriched early in transformation; "vs. 4 hr" highlights motifs that are enriched late in the transformation; "vs. repressed" highlights motifs that are enriched in activated promoters; "vs. activated" highlights motifs that are enriched in repressed promoters; "vs. independent" highlights motifs that are enriched in STAT3- and transformation-dependent promoters; "vs. dependent" highlights motifs that are enriched in STAT3-independent and transformation-dependent promoters. *P*-values considered significant are highlighted in yellow. STAT family motifs are highlighted in green.

Figure 26 (Continued)

All 4 hr vs. RefSeq Promoters

Motif	Matrix ID	*P*-value
NFIC	MA0161.1	1.68E-03
INSM1	MA0155.1	3.45E-03
Stat3	MA0144.1	6.69E-03
Myc	MA0147.1	1.19E-02
Arnt::Ahr	MA0006.1	1.59E-02
SP1	MA0079.2	2.56E-02
Mycn	MA0104.2	2.95E-02
MZF1_1-4	MA0056.1	5.48E-02
RXRA::VDR	MA0074.1	5.49E-02
NFYA	MA0060.1	5.59E-02

All 24 hr vs. RefSeq Promoters

Motif	Matrix ID	*P*-value
SP1	MA0079.2	1.91E-08
INSM1	MA0155.1	9.84E-08
NFKB1	MA0105.1	1.37E-07
Klf4	MA0039.2	1.39E-07
TFAP2A	MA0003.1	4.26E-07
Egr1	MA0162.1	5.78E-06
NF-kappaB	MA0061.1	7.47E-06
Pax5	MA0014.1	1.60E-05
CTCF	MA0139.1	6.83E-05
Tcfcp2l1	MA0145.1	9.10E-05

All 4 hr vs. 24 hr

Motif	Matrix ID	*P*-value
RXRA::VDR	MA0074	2.00E-02
NFIC	MA0161	3.09E-02
Ar	MA0007	6.19E-02
Arnt::Ahr	MA0006	6.33E-02
Stat3	MA0144	7.32E-02
Myc	MA0147	1.05E-01
MIZF	MA0131	1.12E-01
MZF1_1-4	MA0056	1.33E-01
Hltf	MA0109	1.37E-01
SRF	MA0083	1.39E-01

All 24 hr vs. 4 hr

Motif	Matrix ID	*P*-value
Hand1::Tcfe2a	MA0092	6.25E-12
CTCF	MA0139	2.72E-07
RUNX1	MA0002	5.87E-06
ZEB1	MA0103	2.43E-05
RORA_1	MA0071	3.21E-05
Esrrb	MA0141	9.58E-05
SPI1	MA0080	1.01E-04
NFKB1	MA0105	1.35E-04
GATA2	MA0036	1.60E-04
Zfp423	MA0116	3.24E-04

All activated vs repressed

Motif	Matrix ID	*P*-value
NR1H2::RXRA	MA0115	3.15E-06
Foxd3	MA0041	2.31E-04
Stat3	MA0144	6.74E-04
SOX10	MA0442	2.56E-03
STAT1	MA0137	4.05E-03
SPIB	MA0081	4.79E-03
Foxq1	MA0040	1.06E-02
ARID3A	MA0151	1.42E-02
HNF1B	MA0153	1.46E-02
GABPA	MA0062	1.64E-02

All repressed vs activated

Motif	Matrix ID	*P*-value
INSM1	MA0155	8.58E-05
ESR2	MA0258	9.24E-04
CTCF	MA0139	2.27E-03
GATA3	MA0037	7.02E-03
NFYA	MA0060	7.55E-03
NHLH1	MA0048	1.25E-02
Myf	MA0055	2.19E-02
RREB1	MA0073	2.34E-02
Zfp423	MA0116	2.74E-02
Pax5	MA0014	3.54E-02

All STAT3 dependent vs independent

Motif	Matrix ID	*P*-value
Tcfcp2l1	MA0145	3.58E-06
PPARG::RXRA	MA0065	1.03E-05
TLX1::NFIC	MA0119	1.01E-04
NFIC	MA0161	9.55E-04
Stat3	MA0144	5.73E-03
HNF4A	MA0114	9.14E-03
AP1	MA0099	1.67E-02
EWSR1-FLI1	MA0149	3.47E-02
TAL1::TCF3	MA0091	3.64E-02
Ar	MA0007	3.91E-02

All STAT3 independent vs dependent

Motif	Matrix ID	*P*-value
E2F1	MA0024	4.90E-06
HIF1A::ARNT	MA0259	5.03E-04
Sox2	MA0143	5.56E-04
REST	MA0138	1.36E-03
Zfx	MA0146	1.43E-03
Arnt::Ahr	MA0006	1.97E-03
Pdx1	MA0132	2.21E-03
Foxd3	MA0041	2.85E-03
FOXC1	MA0032	2.99E-03
MIZF	MA0131	3.06E-03

targets of a TF based on literature mining. Using this approach, and using all transformation repressed genes, AR, PPARG, NR3C1,TP53, KDM5B, SNAI1, PDX1 and TP63 were found to be significantly associated with transformation repressed genes (Figure 27), all of which are known transcriptional repressors. However, *AR* and *PDX1* do not have detectable RNA in MCF10A-ER-Src cells (not shown). TP53 and NR3C1 are interesting, as they were previously identified as common nodes linking inflammatory signals and cancer transformation to metabolic syndrome [294] within the MCF10A-ER-Src model.

STAT3 cooperates with NFκB in an epigenetic switch that links inflammation to transformation

NFκB is activated rapidly upon ER-Src induction (as measured by RELA/p65 nuclear localization) and mediates an epigenetic switch through indirect induction of IL6 that is necessary for transformation ([289]; Figure 30). STAT3 is also known to be part of this switch, via its direct transcriptional targets *MIR21* and *MIR181b* which cooperate to activate NFκB via posttranslational mechanisms [297]. However, *NFKB1* (p105/p50) was also a direct transcriptional target of STAT3. *NFKB1* RNA levels were increased early during transformation and this response was STAT3-dependent (Figure 28) and, likely, direct. Transformation induced STAT3 ChIP-Seq sites were found within an intron of *NFKB1* and just downstream of the gene, with additional non-differential sites located upstream (Figure 11, D). In addition, IL6, which was not known to be downstream of STAT3 in this switch, was a direct STAT3 target gene as STAT3 was present at its promoter (Figure 11, E) and its transcriptional induction during transformation was STAT3-dependent (Figure 22, B). Hence, an additional positive feedback loop exists in which STAT3 transcriptionally upregulates NFκB and IL6, which can both activate STAT3, thus maintaining rather than initiating the epigenetic switch.

Figure 27: Ingenuity Pathway Analysis prediction of TFs involved in transformation

Transformation-dependent differentially regulated genes, at the indicated times and treatments, and associated RNA expression fold changes, were submitted to Ingenuity Systems' Transcription Factor Analysis tool. This tool matches gene expression changes with the known effects mediated by upstream TFs based on experimental findings from the literature.

Figure 27 (Continued)

STAT3-dependent

Differentially expressed @ 4 hrs		Differentially expressed @ 24 hrs	
Transcription Regulator	*P*-value	Transcription Regulator	*P*-value
STAT3	1.8E-08	TP53	1.2E-12
STAT5A	7.7E-05	NFkB	1.4E-11
SMAD4	8.4E-05	NR3C1	3.8E-11
ER	8.6E-05	ER	1.9E-10
NR3C1	9.4E-05	STAT3	2.0E-10
BRCA1	9.4E-05	FOXO4	6.7E-10
NFkB	1.0E-04	NR3C2	2.1E-08
CREB1	1.4E-04	TP63	6.8E-08
SP3	1.6E-04	SMAD3	7.9E-08
STAT6	1.9E-04	FOS	2.6E-07

STAT3-independent

Differentially expressed @ 4 hrs		Differentially expressed @ 24 hrs	
Transcription Regulator	*P*-value	Transcription Regulator	*P*-value
TP53	7.9E-06	TP53	1.3E-10
ER	9.0E-06	STAT3	1.0E-08
TP63	1.3E-05	NR3C1	1.2E-06
PGR	1.8E-04	Ikb	4.7E-06
RB1	7.7E-04	NFKB1	7.7E-06
TLX1	2.1E-03	IRF1	8.7E-06
ELK1	2.1E-03	NFkB	9.2E-06
Atf	2.3E-03	EGR1	1.0E-05
Betacatenin/TCF	2.3E-03	HDAC3	1.2E-05
URI1	2.3E-03	SP1	1.4E-05

Transformation-dependent

Differentially activated @ 4 and/or 24 hrs		Differentially repressed @ 4 and/or 24 hrs	
Transcription Regulator	*P*-value	Transcription Regulator	*P*-value
STAT3	1.8E-21	ER	7.8E-08
NFkB	1.7E-17	AR	7.3E-07
TP53	4.4E-13	PPARG	1.8E-06
ER	2.6E-12	NR3C1	6.0E-06
NR3C1	3.7E-12	NR3C	9.2E-06
CEBPA	6.9E-12	TP53	1.3E-05
HIF1A	1.3E-11	KDM5B	3.7E-05
SMAD3	2.2E-11	SNAI1	6.1E-05
EPAS1	6.8E-11	PDX1	8.0E-05
JUN	1.1E-10	TP63	8.0E-05

In addition, *IPA* network analysis of the 4 hr and 24 hr time points of genes that were STAT3-independent and transformation regulated both contained NFκB as the central effector of the perturbed signaling pathways, with IRF1 and IRF9 as likely co-operating partners (Figure 18, C and D). This suggests that STAT3 and NFκB cooperate transcriptionally early during transformation with unique and, likely, overlapping transcriptional targets.

STAT3 transcriptionally induced SOCS3 is an inhibitor of inflammatory transformation

SOCS3 was identified as a highly up-regulated gene during transformation of MCF10A-ER-Src cells (and confirmed here) and shown to be an inhibitor of transformation as siSOCS3 led to a modest increase in colony formation in soft agar upon ER-Src induction [289]. The *SOCS3* locus contains numerous differential STAT3 sites (Figure 11, B) and its transcriptional induction during transformation is STAT3-dependent (Figure 23, A). In this regard, STAT3 activation of *SOCS3* represents an auto-inhibitory signal acting against the persistent inflammation observed during transformation of MCF10A-ER-Src cells (Figure 30).

Identification of TFs linked to transformation and their dependency on STAT3 – TSC22D3 and ARNTL2

To cast a broad net, Figure 28 details the RNA expression changes of all significantly differentially regulated TFs during transformation and the effect of siSTAT3 on their expression. Many of these TFs have been previously linked to tumorigenesis and/or inflammation, though many have not and are novel in this regard. *ETS2*, *BCL3*, *FOS*, *ATF3*, *ARNTL2*, and *TSC22D3* were the topmost differentially regulated TFs and all were STAT3-dependent and deserve follow up experimentation. ARNTL2 and TSC22D3 are of particular interest, for disparate reasons, and will be discussed in the following sections.

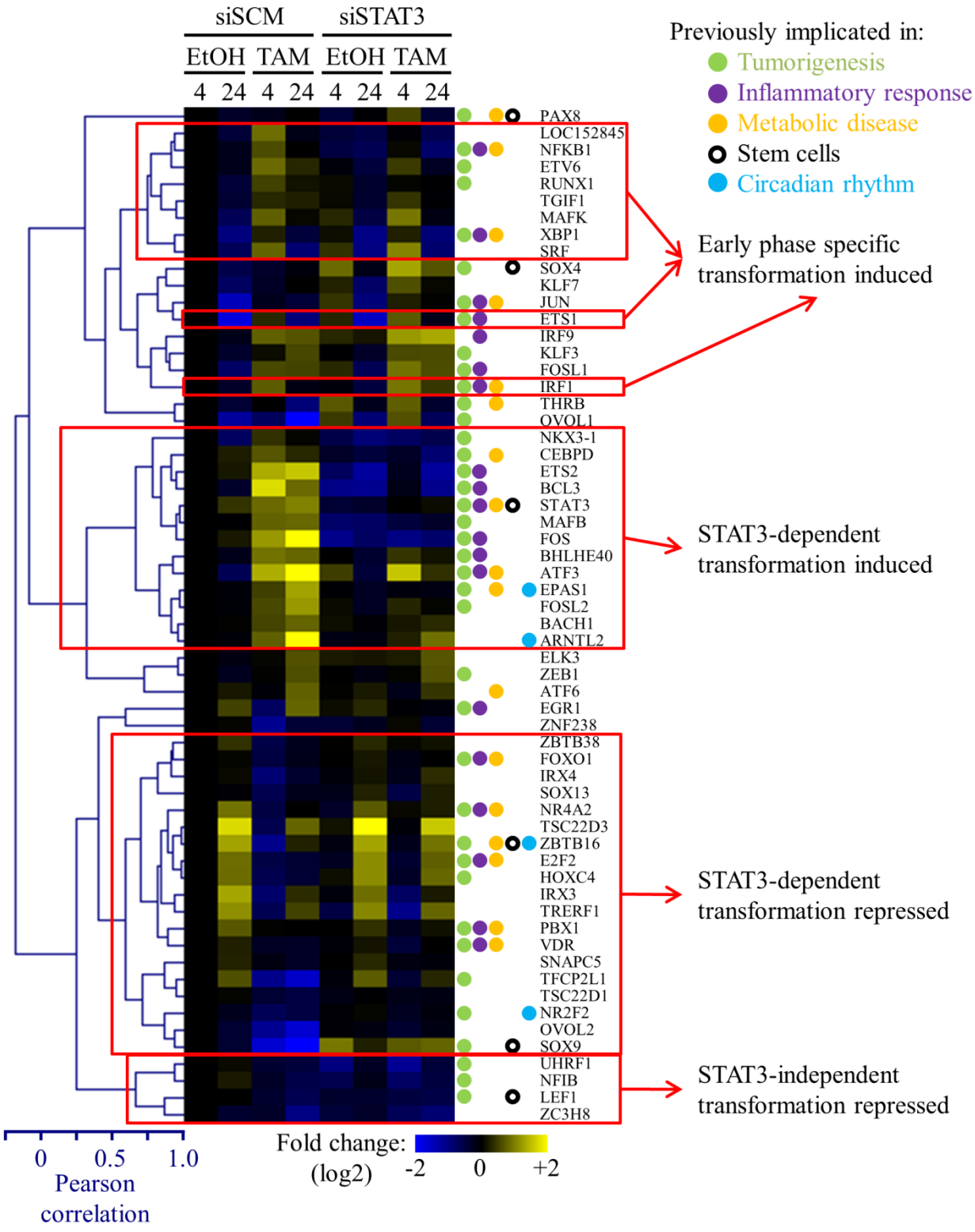
Surprisingly, a large group of these TFs increased in expression at 24 hr post EtOH treatment, and were repressed both early (4 hr) and/or late (24 hr) during transformation (Figure 28). These TFs are very interesting, of which TSC22D3 was the most dynamic, as they were highly expressed as cells became more confluent, which is a stressful growth situation for cells, yet were rapidly repressed during transformation. These TFs could be acting to inhibit cellular transformation during normal growth of somatic cells, or they could be associated with the transcriptional program of contact inhibition. The increased expression observed at 24 hr in EtOH treated cells still occurred after knockdown of STAT3 activity, indicating that during somatic cell growth, their transcription is STAT3-independent. However, for most, during transformation in the absence of STAT3 activity, the repression observed during transformation was reduced, indicating STAT3-dependent transcriptional repression during transformation.

While circadian rhythm related GO terms were not significantly enriched in genes differentially regulated during transformation in MCF10A-ER-Src cells using the gene ontology tool *DAVID* [313, 314], RNA expression of a core member of the circadian clock, *ARNTL2*, was differential. *ARNTL2* RNA was induced during transformation in a STAT3-dependent manner (Figure 28) and the *ARNTL2* locus contained a large increase in STAT3 ChIP signal and a STAT3 differential binding site (Figure 11, A). The expression level of a second core clock gene, *ARNTL*, was reduced upon siSTAT3 treatment in MCF10A-ER-Src cells, though its RNA level was not altered during transformation (not shown). ARNTL2 shares structural and functional homology with CYCLE and can heterodimerize with CLOCK, both core circadian clock members, to regulate gene expression during the circadian cycle [315, 316]. The RNA levels of the TFs *EPAS1* and *ZBTB16*, which are known to be involved in circadian rhythms, were also found to be deregulated in a STAT3-dependent manner upon transformation (Figure 28). EPAS1

Figure 28: Normalized relative RNA expression levels of all TFs differentially expressed during transformation

Shown are all TFs differentially regulated (*P*-value < 10^{-4}, > log2 0.5 fold change) during transformation at 4 hr or 24 hr post TAM treatment in siSCM transfected cells. Shown are the RNA expression levels at 4 hr and 24 hr post EtOH or TAM treatment in samples transfected with siSCM (scrambled control) or siSTAT3. RNA levels are expressed as fold change over 4 hr EtOH and siSCM treated sample. TFs were clustered and red boxes indicate groups of TFs whose transcriptional response to treatment was correlated. Those TFs known to be involved in tumorigenesis, inflammatory response, metabolic disease, "stemness" or the circadian rhythm are indicated by colored circles.

Figure 28 (Continued)



(also known as HIF2α) is a hypoxia inducible TF which can heterodimerize with ARNTL2 to regulate gene [59, 315, 317]. *ZBTB16* is repressed during transformation, and while little is known about the functions of this TF, its expression was shown to be circadian [318, 319] and it has been genetically linked to metabolic syndrome, an inflammatory disease, in rats [320]. A list of all circadian regulated genes that were differentially regulated during transformation can be found in Figure 29.

Figure 29: Transformation and circadian rhythm associated genes

Genes differentially regulated upon ER-Src induced transformation of MCF10A-ER-Src cells that are known to be expressed in a circadian oscillation in mouse peripheral tissues as assayed by gene expression microarrays. Only genes with a Bonferroni corrected *P*-value of $< 10^{-4}$, based on the JTK_CYCLE algorithm [319], were considered as circadian rhythm genes.

Figure 29 (Continued)

Genes differentially regulated during transformation that occur in the circadian rhythm dataset of Hughes ME, *et al* :

Transformation			Transformation and siSTAT3	
ABHD6	GRB14	RBMS1	ABHD6	PNP
ACSL4	HSD17B2	RCL1	ACAT2	PNRC1
ARNTL	HSPA4	RGS16	ACOX2	PPP1R3B
ARRDC3	HSPB1	RNF125	ARNTL	PSMB10
BBOX1	IFITM1	RORA	ARRDC3	PTP4A1
BLCAP	ING2	S100A10	BBOX1	PTPRK
BNIP3	INHBA	SAA4	BNIP3	RCL1
BNIP3L	IVNS1ABP	SAP30L	CCNG2	RGS16
BTG1	KHK	SCARB1	CDH1	SAA4
CABC1	KLF13	SERPINE2	CEBPD	SERPINE2
CAMK1D	KYNU	SLC39A8	CGN	SLC39A8
CASP6	LIPG	SNX10	CHN2	SNX10
CCNG2	LMO7	SORBS1	CLDN1	SORT1
CDH1	LONRF1	SORBS2	CRIM1	ST6GAL1
CEBPD	LONRF3	SORT1	CTGF	SVIL
CGN	LPIN1	SPSB1	CXADR	TGM2
CHN2	LSS	ST3GAL1	DHRS3	TIMP3
CLDN1	MCM10	ST3GAL5	FLNA	TSC22D3
COL27A1	METTL7A	ST6GAL1	GCLC	UGCG
COL4A1	MOCOS	SVIL	GLDC	ZBTB16
CPT1A	MTHFD1L	TBC1D15	GLRX	
CREM	MTHFR	TGM2	GPD1L	
CRIM1	MYO1B	TIMP3	GRAMD3	
CRIP2	NDRG1	TJP2	HMGCS1	
CTGF	NEDD4L	TJP3	HSD17B2	
CXADR	NET1	TMEM97	HSPB1	
CXXC5	NFIL3	TNFAIP2	INHBA	
CYP39A1	NR1D2	TSC22D3	IRF6	
DHRS3	NR2F2	UGCG	IRS1	
DTX4	ODC1	UGP2	KLF13	
EFHD2	PDLIM1	WDR45	LMO7	
EFNA1	PGK1	ZBTB16	LONRF3	
FBXO9	PLAT	ZC3H12A	LPIN1	
FGFR2	PLXNA2		LSS	
FKBP5	PNKD		MMD	
GCH1	PNP		MOCOS	
GCLC	PNRC1		NDRG1	
GLDC	PPP1R3B		NEDD4L	
GLRX	PSEN2		PGK1	
GPD1L	PTPRK		PLXNA2	

DISCUSSION

STAT3 during transformation

STAT3 is a well-known central mediator of inflammation-mediated oncogenic transformation, however, only a small number of its direct transcriptional targets have been identified in such a model. Dechow *et al.* [292] reported 199 genes whose expression was affected by overexpression of a constitutively active STAT3 construct in MCF10A cells and Hutchins *et al.* [299] found < 2500 genes bound by STAT3 within 20 kbp of their TSS in macrophages. We have shown here that at least 1/3rd of the transcriptional program of transformation is mediated by STAT3 activity, either directly or indirectly, and that NFκB likely mediates the rest. Previously, we identified NFκB as a second central mediator of transformation that cooperates with STAT3 [289, 297]. However, the transcriptional program during transformation is more complex with downstream effector TFs likely taking part at later stages of transformation (*e.g.* FOS).

Here we report the genomic locations of 78,293 and 129,192 non-redundant STAT3 and FOS binding sites, respectively, during a time course of inflammation-mediated oncogenic transformation and relate these findings to the STAT3- and/or transformation-dependent transcriptional program of transformation. Curiously, Hutchins *et al.* [299] only found 1,352 STAT3 sites during IL10 stimulation of macrophages, and why the great disparity in the number of sites found compared to ours is unknown. We identified 5921 non-redundant STAT3 sites as differential during transformation and that these sites are found near key genes (*e.g.* MMP locus on Chr11 q22.1-q22.3) and processes (*e.g.* "*Cellular movement*") that mediate transformation. Differential STAT3 sites tended to be situated outside of proximal promoters in distal CREs that were occupied by FOS.

Deregulation of STAT3-dependent circadian clock related genes during transformation

Nearly all mammalian cells have a biological clock that is linked to the circadian day-night cycle. In mammals, the circadian clock is controlled centrally by the hypothalamic suprachiasmatic nucleus of the brain, and through neuroendocrine (*e.g.* pertinently anti-inflammatory glucocorticoids [321]) and external stimuli (*e.g.* light, feeding), is synced with peripheral tissue circadian clocks (for review see [322, 323]). The molecular mechanism of the circadian clock is based on transcriptional-translational feed-back loops, of which, many of the core clock genes are TFs (*CYCLE, CLOCK, NPAS2*) whose activity is negatively regulated by CYR and PER proteins. The circadian clock is linked to many critical cellular processes such as proliferation, apoptosis, DNA damage response, and metabolism with a growing body of evidence elucidating a role for the circadian clock in cancer, both through epidemiological and molecular studies [324]. Mutation of the mouse *Per2* gene leads to cancer and mutation of *CRY* in *p53*-null mice delays the onset of cancer [325] due to NFκB mediated apoptosis [326]. Pertinently, to transformation in the MCF10A-ER-Src breast cancer cell line: women who work night shifts are modestly more prone to developing breast cancer [327-330]; the *PER* genes are deregulated in breast cancers [331]; and, *NPAS2* mutations are associated with an increased risk of breast cancer [332]. In general, patients with a perturbed circadian rhythm are known to be more prone to cancer and have a poorer prognosis [333-335].

Interestingly, *PER2* overexpression can inhibit the transcription of ERα regulated genes [336], a key mammary epithelial cell TF (for a review see [337]). Mutations in *ERα* have long been associated with an increased risk of breast cancer [338] and a key determinant of treatment options [339]. A link between inflammation and the circadian rhythm is starting to be elucidated. The DNA binding activity NFκB/REL complexes [340, 341], which are critical transcriptional

mediators of inflammation, and an NFκB inhibitor protein (NFKBIA) [319] have been found to be circadian regulated. However, *NFKB1* null mice do not manifest defects in daily locomotor activity during the circadian cycle [342], though the main focus of that study was not the circadian clock. Using chemical inhibitors of NFκB activity and the *IL6* knockout mouse, Monje *et al.* [343], showed a molecular link between inflammatory signaling through NFκB and IL6 on the core clock genes *Per2* and *Npas2* in a mouse model of circadian disruption through light deprivation. These literature findings and the STAT3-dependent deregulation of *ARNTL2* during transformation as discovered here, raise the interesting questions as to if and how the circadian clock influences inflammatory transformation in our model, and ultimately the clinical manifestation of breast cancer, and deserve further biological experimentation.

AP-1 factors in inflammation-mediated oncogenic transformation

FOS is one of the most differentially expressed TFs during transformation of MCF-10A-ER-Sr cells, being highly expressed late at 24 hr, its expression is STAT3-dependent, and transformation induced STAT3 sites preferentially form at FOS bound sites. Also, we have characterized the CREs and STAT3 binding sites present at the *FOS* proximal promoter, which is a known target of STAT3, and have found other members of the FOS and JUN family to be deregulated (*FOS, FOSL1, FOSL2, JUNB, JUND*) during transformation. Previously, we identified FOS as a node within the transformation-dependent transcriptional program that is common between two different models of transformation and was also linked to metabolic syndrome, an inflammation-based family of diseases [294]. However, its importance and role within transformation was not expanded upon, and here we show a putative role for FOS in regulating embryonic, stem cell and bone related genes. During transformation, cancerous cells lose their differentiated cellular state and revert to a more embryonic-like state. There are also

known stem cell-like cell populations within transformed cell lines, including MCF10A-ER-Src [344], and primary tumors including AML [345], breast [346], brain [347], multiple myeloma [348], pancreatic [349], and colon [350-352] cancers, among others. The deregulation of FOS and JUN family members is common in cancer (reviewed in [353]). Indeed, overexpression of *FOS* can transform fibroblasts [354], and overexpression of *FOS* [355, 356], and *FOSL1* [356] in the breast cancer cell line MCF7 increases cell motility, invasion and proliferation. Here we present evidence of FOS/JUN family members as downstream effectors of STAT3 with a putative role in regulating aspects of "stemness" and bone metastasis during oncogenic transformation.

The lack of CRE dynamics during transformation

We have also discovered the genomic repertoire of CREs used during transformation and found that this set of CREs does not change and are static, even though large scale phenotypic and gene expression alterations are taking place. It is known from genomic studies comparing different cell types that large scale differences are seen in CREs, including those that are defined by FAIRE-Seq [306]. These cell-type specific CREs tend to be enhancers and are significantly enriched for the DNA motifs of TFs that are pertinent to the establishment/maintenance of that specific cell type. In MCF10A-ER-Src cells, the activation of STAT3 and its important role in establishing the transformed cell state is akin to these cell type specific TFs. However, STAT3 does not elicit the formation of new CREs during transformation. This can be explained by the fact that most cell-type specific determinant TFs are "pioneer" factors with the ability to access their DNA motif, usually in the context of a nucleosome, before cooperating TFs and H3 acetylation occurs (*e.g.* FOX and GATA family members). Though ~35% of STAT3 binding events do not occur within CREs, this is most likely due to a false negative result on the part of

CRE detection. It can be reasoned that the over-arching cell-type, of non-transformed and transformed MCF10A-ER-Src cells, is still that of mammary breast epithelial, and, therefore, the changes that ER-Src induced transformation create, though extensive, does not fundamentally alter the cell-type. All transcriptional changes are mediated through pre-existing CREs and the CRE population is recycled to accommodate new phenotypes.

Does TSC22D3 inhibit the epigenetic switch during somatic cell growth?

TSC22D3 was the most dynamic of the STAT3-dependent transformation repressed TFs (Figure 28). TSC22D3 is particularly interesting as it is known to be a negative regulator of Ras/Raf signaling pathways by directly interacting [357, 358] and inhibiting NFκB [359]. Both Ras and NFκB are mediators of the epigenetic switch in MCF10A-ER-Src cells [289], and thus, could be inhibited by TSC22D3 in somatic cells, before transformation. TSC22D3 was originally identified as a gene up-regulated by the glucocorticoid dexamethasone [360], and has anti-inflammatory and immuno-modulatory properties (for reviews see [361-363]; original articles [364-366]). In inflammation-mediated oncogenic transformation of MCF10A-ER-Src cells, Ras activation of NFκB helps mediate the induction of IL6, which in turn drives transformation [289]. In this regard, it is tempting to speculate that the high level of anti-inflammatory TSC22D3 down-regulates Ras signaling and NFκB activity during somatic cell growth, ultimately preventing run-away pathologic inflammation, in our case, oncogenic transformation. This repressive anti-inflammatory signal is relieved during transformation by STAT3, whose activation represses TSC22D3 transcription.

Updating the epigenetic switch

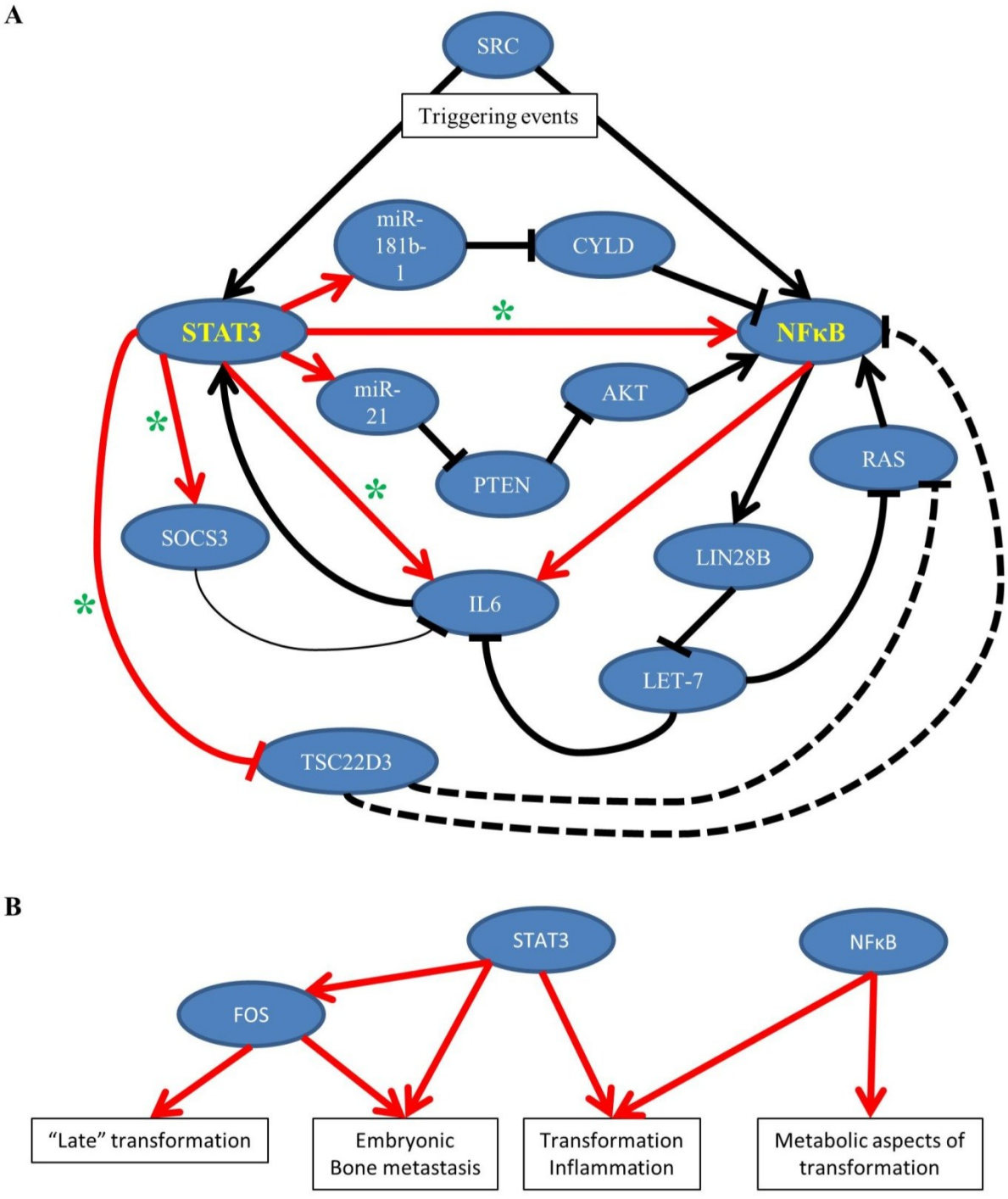
Previously the Struhl Lab discovered an epigenetic switch that is initiated by Src activation of STAT3 and NFκB in MCF10A-ER-Src cells [289, 297]. Here we provide evidence of a transcriptional feedback loop linking STAT3 to the direct transcriptional activation of *IL6* and *NFKB1*, who in turn can activate *STAT3* transcription. This aspect of the epigenetic switch is probably active during the later stages of transformation as it is dependent on new protein production, and while an increase in *STAT3* mRNA is not strictly required for transformation, the maintenance of its expression is important. We have also provided evidence of two inhibitory feedback loops, one acting late through SOCS3 inhibition of IL6/STAT3, the other acting before transformation through TSC22D3 and Ras/NFκB. This later feedback loop is speculative, but very well supported by the literature, and may act to inhibit the epigenetic switch in normal, dividing, proliferating and/or stressed cells. A modified version of the epigenetic switch incorporating these findings is detailed in Figure 30.

Figure 30: The epigenetic switch that initiates and maintains transformation of MCF10A-ER-Src cells

A. Model of the epigenetic switch that mediates transformation in MCF10A-ER-Src cells. Dashed lines indicate predicted interactions base on literature findings. Red lines indicate direct transcriptional regulation. Black lines ending in an arrow or perpendicular slash indicate known positive and inhibitory interactions, respectively, within MCF10A-ER-Src based on ([289, 294, 297]; this dissertation). A thin line indicates weak activity. A green asterisk represents interactions based on data presented in this dissertation.

B. A summary of the major phenotypic effects and processes mediated by STAT3, NFκB, and FOS.

Figure 30 (Continued)



METHODS

Tissue culture and chromatin immuno-precipitation (ChIP)

MCF10A-ER-Src cells were grown and ChIP DNA was isolated as per standard ENCODE protocols ([183-185], Appendix C) and a detailed protocol is available at: http://genome.ucsc.edu/ENCODE/. Cultures were grown to 70% confluency, then treated with either EtOH or tamoxifen (TAM) for 4, 12, 24 or 36 hr, as detailed in [294], and harvested for DNA, protein or RNA as detailed below.

FAIRE-Seq

Cells were grown as above and a full detailed FAIRE-Seq protocol is available at: http://genome.ucsc.edu/ENCODE/ and [301]. See Supplemental Data for genomic coordinates.

ChIP-Seq and peak calling

ChIP DNA (2 biological replicates) prepared as above, and immuno-precipitated with anti-phospho-STAT3 (Cell Signaling, 9131), anti-FOS antibody (Santa Cruz, SC-7202x), or anti-POL2 (Covance, 8WG16) and input DNA (3 biological replicates) were end repaired with calf intestinal alkaline phosphatase (New England Biolabs, USA) and sent for sequencing to the Stanford Center for Genomics and Personalized Medicine. ChIPs for anti-ATF2 (SC-6233x), anti-E2F6 (SC-22823x), anti-E2F4 (SC-866x), anti-IRF2 (SC-13042), anti-JUN (SC-1694), anti-MYC (SC-764), RPC155 (R. White) were prepared as above but using one biological replicate. Library preparation and Illumina (USA) sequencing were carried out as per Illumina protocols and a detailed protocol is available at http://genome.ucsc.edu/ENCODE/. Sequence reads (32 nucleotides) were mapped to the *H. sapiens* genome (hg19) using *Bowtie* [247], allowing <= 2

mismatches per read and reads with > 10 reportable alignments were discarded. Binding sites were called using *MACS* v1.4 [248] at a *P*-value threshold of 10^{-09}, "auto" redundant read setting, using input to control for local genomic biases. *PeakSplitter* [367] was used to split *MACS* called peaks into subpeaks of local maxima using default settings. STAT3 subpeaks (referred to as "sites" in Results and Discussion) were called as differential if they did not overlap a STAT3 bound subpeak in the non-transformed control population (peaks called at 10^{-9} *P*-value) and had a fold change greater than the mean + 1x standard deviation of all peaks within the population. Fold change in STAT3 ChIP-Seq signal was calculated as read counts per region per million mapped reads divided by the corresponding control ChIP signal of that region. A smoothing value of 10 was added to the read count of each region in the transformed and control samples. See Supplemental Data for genomic coordinates.

Annotation of peaks to gene features, GO analysis (*GREAT*/*IPA*)

Genomic locations of subpeak summits were submitted to the annotation tool *GREAT* [259] using the following parameters: whole genome background set, basal plus extension, proximal upstream = 5 kbp, proximal downstream = 1 kbp, distal = 1 mbp; or, whole genome background set, basal, proximal upstream = 5 kbp, proximal downstream = 1 kbp. For *IPA* (Ingenuity Systems, USA; http://www.ingenuity.com) gene probe IDs, with the corresponding log2 fold change, were uploaded into and analyzed by *IPA* (build: 140500, content version: 12710793) using default settings. Molecular signaling pathways were visualized using *IPA* where a gray shaded node represented a subpeak binding site located within the putative regulatory region, as defined by *GREAT*, of that gene/molecule. The biological relationship between two molecules is represented as a line and is based on professionally curated literature findings. The relationships can be direct or indirect.

siRNA transfections

MCF10A-ER-Src cells were reverse transfected with Dharmacon siRNAs (Thermo Scientific, USA) according to the manufacturer's protocol. Briefly, cells were seeded in 6-well plates containing 25 nM ON-TARGETplus SMARTpool STAT3 or ON-TARGETplus Non-targeting Pool and after 48 hr of growth were treated with EtOH or tamoxifen (1 μM) for 4 hr or 24 hr as per [294]. For Western blots, cells were lysed after 72 hr (24 hr post EtOH/TAM treatment). For expression profiling RNA was harvested after 4 or 24 hr post treatment.

Gene expression microarrays

RNA (from 3 biological replicates) was prepared for arrays using 3' IVT Express kit (Affymetrix, USA) as per manufacturer protocol – 100 ng RNA, 15 amplification cycles. Amplified RNA was given to the Children's Hospital Boston microarray core facility for hybridization to GeneChip Human Genome U133 plus 2.0 gene expression arrays (Affymetrix, USA) for hybridization and imaging as per manufacturer protocols.

Western blots

Cells were lysed in 0.5 mL lysis buffer [50 mM Tris-HCl, 1% NP-40 (v/v), 5 mM EDTA, 1 mM NaF, pH 8.0 supplemented with 10 mM β-glycerol phosphate, 1 mM phenylmethanesulfonyl fluoride, 1 mM sodium orthovanadate, 1% Phosphatase Inhibitor Cocktail II (Sigma, USA), and 1 Complete-Mini Protease Inhibitor Cocktail tablet (Roche Applied Science, USA) per 10 mL]. Lysates were cleared by centrifugation at 20 kg, 15 min, 4 °C. Prior to immuno-blotting, lysates were boiled in standard SDS gel-loading buffer and loaded into a 10% polyacrylamide gel. After separation by electrophoresis, the proteins were transferred to nitrocellulose and the membranes were blocked with 5% nonfat dry milk (w/v) in Tris-buffered saline (20 mM Tris, 150 mM NaCl,

pH 7.6) containing 0.1% Tween-20 (v/v). Membranes were probed using mouse-derived anti-STAT3 antibody from Cell Signaling Technologies, USA and anti-β-actin from Sigma, USA. Bands were detected with IRDye 800-labelled goat-anti-mouse IgG (LI-COR Biosciences; USA) and imaged using an Odyssey Infrared Imaging System (LI-COR Biosciences, USA).

Determination of differential gene expression

Gene expression arrays were analyzed using the R packages: *limma* [257], *affy* [256]. Arrays were background corrected, normalized and probe set expression values determined by the mas5 algorithm. Probe sets were annotated to RefSeq gene IDs using *GREAT* [259] or *DAVID* [313, 314]. Genes determined to be transformation regulated/differential were derived from siSCM treated samples comparing EtOH to TAM treatments with a *P*-value < 10^{-4} and with an absolute log2 fold change > 0.5. Those genes determined to be STAT3 and transformation regulated were determined by comparing EtOH to TAM samples under both siSCM and siSTAT3 conditions. Genes were selected as STAT3-independent if their differential expression was statistically insignificant upon siSTAT3 and had an absolute log2 fold change of < 0.5 upon siSTAT3. The number of STAT3-dependent and STAT3-independent genes does not equal the total number of genes considered differential by transformation as many genes could not be unambiguously defined as “dependent” or “independent”.

Motif analysis of differentially regulated genes

Differentially regulated genes were as above. The non-redundant gene set was used, retaining the probe set with the lowest *P*-value, and probe sets unable to be annotated to RefSeq IDs were not considered. The web based *Pscan* [265] was used to establish significantly enriched motifs, the settings were: *H. sapiens*, JASPAR, region about TSS -1000/+0 bp.

Differentially regulated TFs

All probe set IDs that were differentially regulated during transformation (see above) were submitted to *DAVID* [313, 314] and probe sets annotated to the term "transcription factor activity" (GO:0003700) were selected. Normalized expression values for each gene are expressed as log2 fold change over siSCM 4 hr EtOH treated samples. Hierarchical clustering of the resultant expression matrix was carried out using the Pearson correlation and average linkage using the software package *TMEV* [368, 369]. Genes previously implicated in: "*Tumorigenesis*", "*Inflammatory response*" and "*Metabolic disease*" were determined by *IPA* (Ingenuity Systems, 2010; http://www.ingenuity.com); stem cells ("*stem cell division*" (GO:0017145), "*stem cell development*" (GO:0048864)) from the Gene Ontology Consortium (July 2012, [370]); and, circadian rhythm from [319].

Annotation of STAT3 sites to differentially expressed genes

Differentially expressed genes were as above. Non-redundant probe sets were used, discarding the probe set with the highest *P*-value. "Promoter" regions are defined as -2500 bp to +500 bp from RefSeq TSS. "Upstream/downstream" regions are defined as +/- 50 kbp from the RefSeq TSS, excluding the promoter region. The number of transformation specific differential STAT3 ChIP-Seq peaks were counted within these regions, normalized to peaks per 1 kbp, and plotted using a 1000 gene rolling mean performed using the *zoo* [371] package of *R*.

Annotation of STAT3 sites to RefSeq TSSs

STAT3 and NFYB (K562) peak summits, defined as the local maxima of read counts within a peak, were mapped to the nearest RefSeq TSS, incorporating strandedness, using an in-house script. Histograms were plotted using the *R* package *ggplot2* [372] using 500 bp bins within a

region of +/-10 kbp about the TSS centered at 0 bp. The frequency of sites is represented as the Gaussian smoothed kernel density estimate with a bandwidth of the standard deviation of the smoothing kernel, calculated using the density function in *R*. The percentage of STAT3 peaks in the proximal upstream region of RefSeq TSSs and located in distal intergenic regions (defined as not within the following RefSeq genic features: -10 kbp upstream of a TSS, +10 kbp downstream of a TTS, intronic, exonic, 5' UTR or 3' UTR), was calculated and compared to the percentage of the genome within each category. Significance was calculated using the single sided binomial test as implemented in the *binom* function in *R*.

De novo **motif discovery**

The top 10,000 (as ranked by *P*-value) STAT3 or FOS ChIP-Seq subpeak summit locations were determined and the sequence +/- 50 bp was extracted and repeat masked. A 5 order Markov model was used as the background set and was extracted from the repeat masked, non-redundant set of FAIRE-Seq *cis*-regulatory elements. Parallel *MEME* was run with the following settings: zoops, revcomp, minw = [4-26], and maxw = [6-30]. For STAT3 and FOS, the top motif corresponded to the respective known canonical motif.

CHAPTER 4: Discussion and future directions

I will only discuss and comment on the broader findings and questions raised by the research within this dissertation in this section. Specific aspects of NF-Y and STAT3 biology have already been discussed within those respective chapters and here I want to highlight topics that are of a more general interest and to talk about future avenues of research.

Transcription factors occupying closed chromatin residing DNA motifs

The ability of NF-Y to access the CCAAT box within closed and transcriptionally repressive chromatin domains is truly intriguing. This finding is atypical when compared to many other TFs, both from research presented in this dissertation and from other studies in eukaryotes [373, 374]. In eukaryotes, nucleosome occupancy is a major barrier to motif occupancy by their cognate TFs *in vivo*. Nucleosomes physically prevent the interaction between a TF and its DNA motif and in so doing restrict TF occupancy to regions of depleted nucleosome occupancy which are commonly found at promoters and enhancers in eukaryotes [375]. Cawley *et al.* [253] found that only ~1% of consensus DNA motifs for TP53 (p53), SP1 and MYC were bound *in vivo* in *H. sapiens*, and in yeast, Rap1 preferentially associates with promoters and not to non-coding regions which are nucleosome occupied [376]. In sharp contrast, prokaryotes do not contain histones and do not contain closed chromatin domains. Prokaryotes do possess histone-like DNA binding factors (Fis, H-NS, HU, IHF) that associate with DNA to form higher-order structures called nucleoids [377]. However, the *E. coli* genome is fully accessible to the DNA binding TF LexA [378], indicating that nucleoids do not prevent TF occupancy of DNA motifs in prokaryotes.

It has been known for close to 15 years that NF-Y can form hybrid NF-Y-nucleosomal-DNA complexes and that NF-Y-H3-H4 complexes exist *in vitro*, however, *in vivo* studies are lacking. From the data presented in this dissertation, it is clear that NF-Y can saturate its motifs within open chromatin, but also displays a remarkable ability to bind to CCAAT boxes in closed and inactive chromatin domains that lack detectable amounts of common transcription-dependent histone PTMs (*e.g.* H3 acetylation). This ability is probably due to the HFD subunits of NF-Y that allow either an interaction with H3-H4 in the context of nucleosomes, or the displacement of the local CCAAT box occupying nucleosome. The lack of an open regulatory element by FAIRE and the aforementioned *in vitro* findings support the former model, though this requires testing.

The technique of DNase I hypersensitivity followed by sequencing (DNase-Seq) is more sensitive than FAIRE-Seq in detecting open and accessible non-chromatin bound DNA. Moreover, when the sequencing depth reaches +500 million reads, the individually protected DNA bases within all open region across the entire genome become detectable, just as in traditional DNase I experiments analyzing a single locus. It would be interesting to compare the DNase I hypersensitivity pattern at NF-Y bound CCAAT boxes in both open and closed chromatin contexts to see if the NF-Y-DNA contacts are different. This could also be done using traditional DNase I assays at select loci, and has been done with reconstituted nucleosomes *in vitro* [83], however the ENCODE project is currently generating suitable DNase-Seq datasets and all that is required is the computational skill to analyze the data that is freely available. This could also be expanded to cover all TFs in which there are suitable ChIP-Seq datasets, a corresponding DNA motif of sufficient information quality.

There is another related question that needs to be addressed: are H2A and H2B present at the NF-Y occupied CCAAT boxes within closed chromatin domains *in vivo*? To date only

nucleosome occupancy maps are available. To the best of my knowledge, no maps of individual histones, except for H2A.Z, exist for *H. sapiens*. It would be interesting to use ChIP-Seq to determine the occupancy of all four histones genome-wide in K562 cells. These datasets could then be used to ascertain if H3-H4 and/or H2A-H2B are depleted at NF-Y bound CCAAT boxes in closed chromatin *in vivo*. Obviously, this could be done in a more generic fashion to isolate regions of the genome in which the ratio of H3/H4 to H2A/H2B is disparate. Any DNA motifs enriched within these regions could be computationally determined and the associated TF(s) identified and confirmed by ChIP, potentially identifying hybrid TF-histone complexes, especially if the TF(s) have HFDs (such as NF-Y).

The ability of specific TFs to access their motifs within closed chromatin domains is a distinguishing feature of "pioneer" factors. Pioneer factors are TFs that have the ability to access their motifs in closed *cis*-regulatory elements before the arrival of cooperating TFs, chromatin remodelers and H3 acetylation and/or H3K4 methylation. During this dissertation, I identified NF-Y, USF1 and MAFK as factors that can access their motifs within closed, transcriptionally inactive chromatin domains. Surprisingly, the known pioneer factors, GATA1 and GATA2, had a limited ability to do this which requires explanation by the GATA scientific community. A second, critical, aspect of pioneer biology is their ability to open chromatin. The lack of GATA binding in non-modified-chromatin domains could be explained by GATA constitutively opening loci upon binding. Hence, pioneer factors may never be found present in closed transcriptionally inactive epigenetic domains. It is obvious that NF-Y, USF1 and MAFK do not always elicit open chromatin upon binding. This could be due to the lack of an initiating event(s). A genome-wide view of more pioneer factors would shed light on this issue. It is now possible to expand upon my method implemented here to use the continually growing TF, histone PTM

FAIRE and DNase I datasets available from ENCODE, and other sources, to computationally screen for "pioneer"-like factors. In fact, I have already computed the location of all known (> 500) DNA binding site motifs genome-wide and screened them for occupancy to all TF ChIP-Seq datasets (~150) available from ENCODE cell types in which chromatin states are, or can be, derived. Analysis of these datasets will provide invaluable insight into the chromatin context that determines TF access to DNA motifs in eukaryotes, how this varies by TF family type (HFD, bZip, Zn-finger, etc.) and, hopefully, reveal more "pioneer"-like TFs. As many pioneer factors are important for determining cell-type identity during development, this avenue of research could lead to defining new roles for well-known TFs.

Why does the functional inactivation of a TF elicit a limited transcriptional response?

In this dissertation, and in the dissertation of Annie Yang (Struhl Lab graduate student) [208], we mapped the location of NF-Y, STAT3 and TP63 (p63) genome-wide and also performed siRNA knockdown followed by genome-wide gene expression analysis. In all three cases, many more genes are bound by a TF than seem to be regulated by that TF. In the cases of STAT3 and TP63, the knockdown efficiency was > 90%. The most extreme example is that of STAT3, in which we also induced its expression. STAT3 is present in the vicinity of ~15,000 *H. sapiens* genes, however, STAT3 knockdown only affects the expression of 451 genes as far as we can detect. This raises the question: how does TF binding relate to gene expression? In the classic view of gene regulation, DNA motifs dictate TF occupancy and the presence of a bound TF (and its activation if required) dictates gene expression. This is the case in prokaryotes (see above).

It seems obvious to assume that most of the TF binding events observed for NF-Y, STAT3 and TP63 are non-functional, at least with respect to gene expression, but these sites are usually highly conserved which implies maintenance of a biological function. A simple explanation is

that gene regulation is dictated by cooperativity between multiple TFs (which we know is the case in eukaryotic systems) and therefore highly redundant, so loss of a single TF (*e.g.* by siRNA) does not affect transcriptional output in most cases, as the remaining TFs are still bound, functional and sufficient for transcription. This implies that the transcriptional outputs that are altered are unusually dependent on the specific TF activity that was perturbed. Another reason could be due to the insensitivity of detecting minor alterations in transcriptional output. A 10% change in RNA expression level could be biologically meaningful, but undetectable by gene expression microarrays and therefore those genes would be listed as "not affected" when they are actually false negatives. Another reason could be due to the requirement of post translation activating modifications required by a TF for it to act as a transactivator, but not for DNA binding. The TF CREB is a good example: CREB phosphorylation is not a prerequisite for DNA binding, but it is for transactivation [379] and the recruitment of coactivators. In this situation, large numbers of CREB binding sites could be functionally irrelevant for transcription as CREB is not activated.

The lack of differentially active *cis*-regulatory elements during a phenotypic change

When we tracked the location of active CREs throughout oncogenic transformation of mammary epithelial cells in tissue culture, we noticed a lack of the differential formation or loss of active CREs during the time course of differentiation. This has been noticed by other groups recently [380, 381]. This was surprising as previous studies incorporating FAIRE-Seq analysis of multiple cell lines showed that a great diversity in *cis*-element usage is common across cell types, with most of the differential cell type specific FAIRE sites being enhancers and not promoters [306]. It is well-known that embryonic stem cells contain globally open chromatin that becomes more restricted and transcriptionally repressive domains expand as cells differentiate

[382]. This process is driven by chromatin remodelers (*e.g.* CHD1, BAF, PRC1) and pioneer factors are also intimately involved. Many tissue specific TFs are pioneer factors, *e.g.* GATA4, FoxA1 [225], and can therefore access their motifs within closed chromatin to explicitly drive open chromatin formation, new CRE usage and gene expression patterns, thereby altering cellular phenotypes.

The lack of differential CREs during massive phenotypic and transcriptional alterations raises an interesting observation: the cell's library of functionally active CREs are static during transformation and are, therefore: 1) "recycled" to accommodate new phenotypes (*e.g.* transformed); and, 2) TFs, presumably ones that are not "pioneer"-like factors, function only within the pre-defined existing population of CREs to affect transcription. This observation can be extended to all TFs involved in SRC induced transformation of MCF10A-ER-Src cells: the CREs do not change during transformation, therefore, any regulatory protein that does change in activity can only work within the confines of the pre-existing population of CREs. There are multiple reasons for this phenomenon. STAT3 and NFκB, the two TFs that drive oncogenic transformation of MCF10A cells, are not known to have "pioneer" abilities, and therefore cannot create new open nucleosome free regions, even when their activities are functionally overexpressed (*e.g.* STAT3 phosphorylation). In this regard, new active CREs are not formed, even though large scale gene expression changes are happening. The cells, even though they have changed phenotype, are still mammary epithelial cells both before and after transformation (a comparison to gene expression datasets from other cell lines would be useful to show this). As such, they may not have undergone a differentiation process of the same magnitude as, let us say, bone marrow resident preosteoclast cells which undergo differentiation into giant, multi-

nucleated, fused polykaryons, with astonishing apical/baso-lateral cell polarity, with a highly specialized function devoted absolutely to resorbing bone.

The findings discussed here have important implications for the study of developmental biology, disease, and, in particular, cancer. Scientists trying to push pluripotent stem cells down specific cellular differentiation pathways by overexpressing specific combinations of TFs, or developmental biologists trying to decipher the combinatorial transcriptional network of cell differentiation pathways, have to consider that pioneer factors are required to act early during differentiation and that "regular" TFs are more important during the later stages of differentiation. Only pioneer factors can drive the differentiation of cell types and non-pioneer TFs can drive all other aspects of differentiation. Specifically derived from my findings with MCF10A cells, pioneer factors are unlikely candidates for critical oncogenic phenotypes such as metastasis, homing, proliferation, immune evasion, tumor growth, and invasion. None of these phenotypes require alterations in cell-type identity, which can only be driven by TFs with "pioneer" abilities. However, cancer stem cell populations may require pioneer factor activity to drive differentiation into non-stem cell cancer cells.

The questions answered and raised by the work in this dissertation will push forward the field of NF-Y biology into new areas of experimentation. Likewise, the provocative finding that massive phenotypic and gene expression alterations can occur without changes in CRE usage will undoubtedly frame new questions for future research.

APPENDIX A: Supplemental Figures

Supplemental Figure 1: NF-Y ChIP-Seq

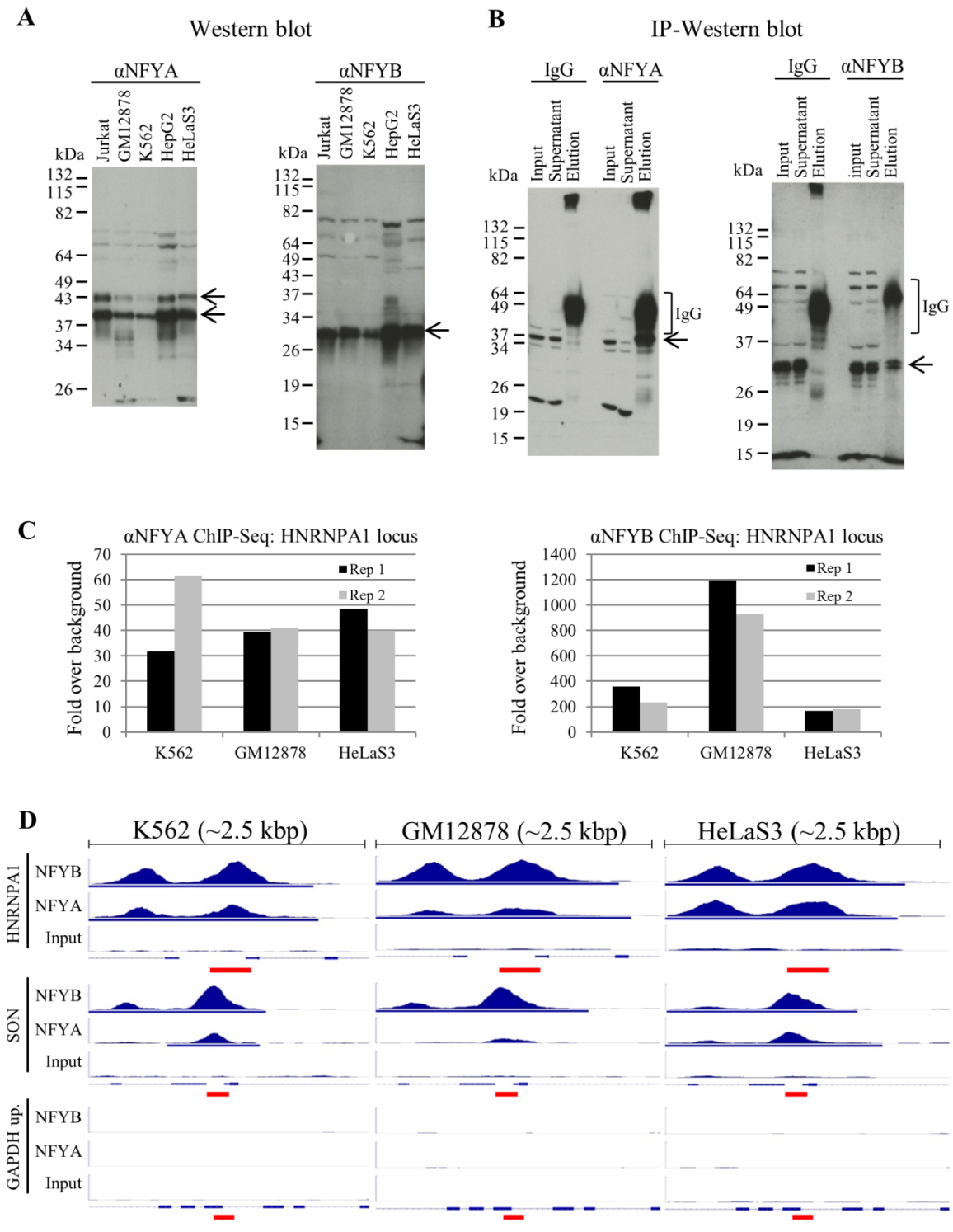
A. Western blots of nuclear extracts from five cell lines probed with anti-NFYA and anti-NFYB antibodies. Arrows highlight doublet bands showing that both isoforms of NFYA and NFYB were detected.

B. Immuno-precipitation Western blot (IP-WB) of nuclear extracts showing enrichment of NFYA and NFYB specific bands in the elution and depletion in the supernatant. An IgG antibody was used as control for non-specific binding.

C. ChIP-QPCR results from anti-NFYA and anti-NFYB IPs performed in K562, GM12878 and HeLaS3, before sequencing, showing enrichment over an NF-Y non-bound control region.

D. Representative loci showing NFYA, NFYB and input control ChIP-Seq data from K562, GM12878 and HeLaS3. Enrichment of reads at the HNRNPA1 and SON promoters were specific to NF-Y and not present in the input dataset. "*GAPDH up.*" and "*TLE6 up.*" were control regions not bound by NF-Y. Red bars indicate ChIP-QPCR primer locations. Blue bars under peaks show *MACS* called peak regions at the 10^{-9} *P*-value. RefSeq genes are illustrated.

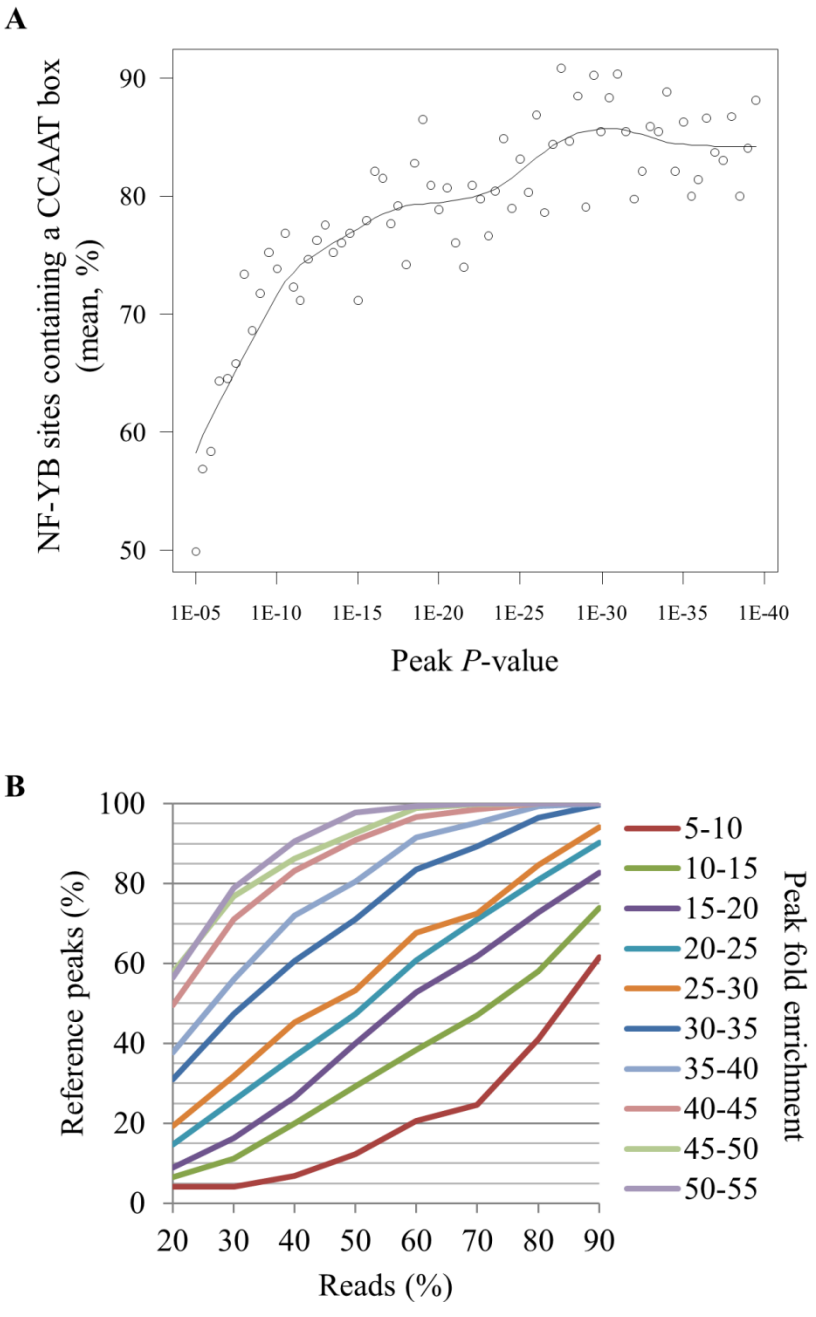
Supplemental Figure 1 (Continued)



Supplemental Figure 2: CCAAT box frequency and saturation analysis

A. Mean percentage of peaks with CCAAT boxes in K562 NFYB peaks called at specific *P*-values. CCAAT boxes were called using *FIMO* at a *P*-value threshold of 10^{-4}. Similarly sized random genomic regions have a CCAAT box rate of 8%.

B. The percentage of peaks, from the K562 NFYB 10^{-5} peak list, that were successfully identified based on a random subsample of ChIP-Seq reads.

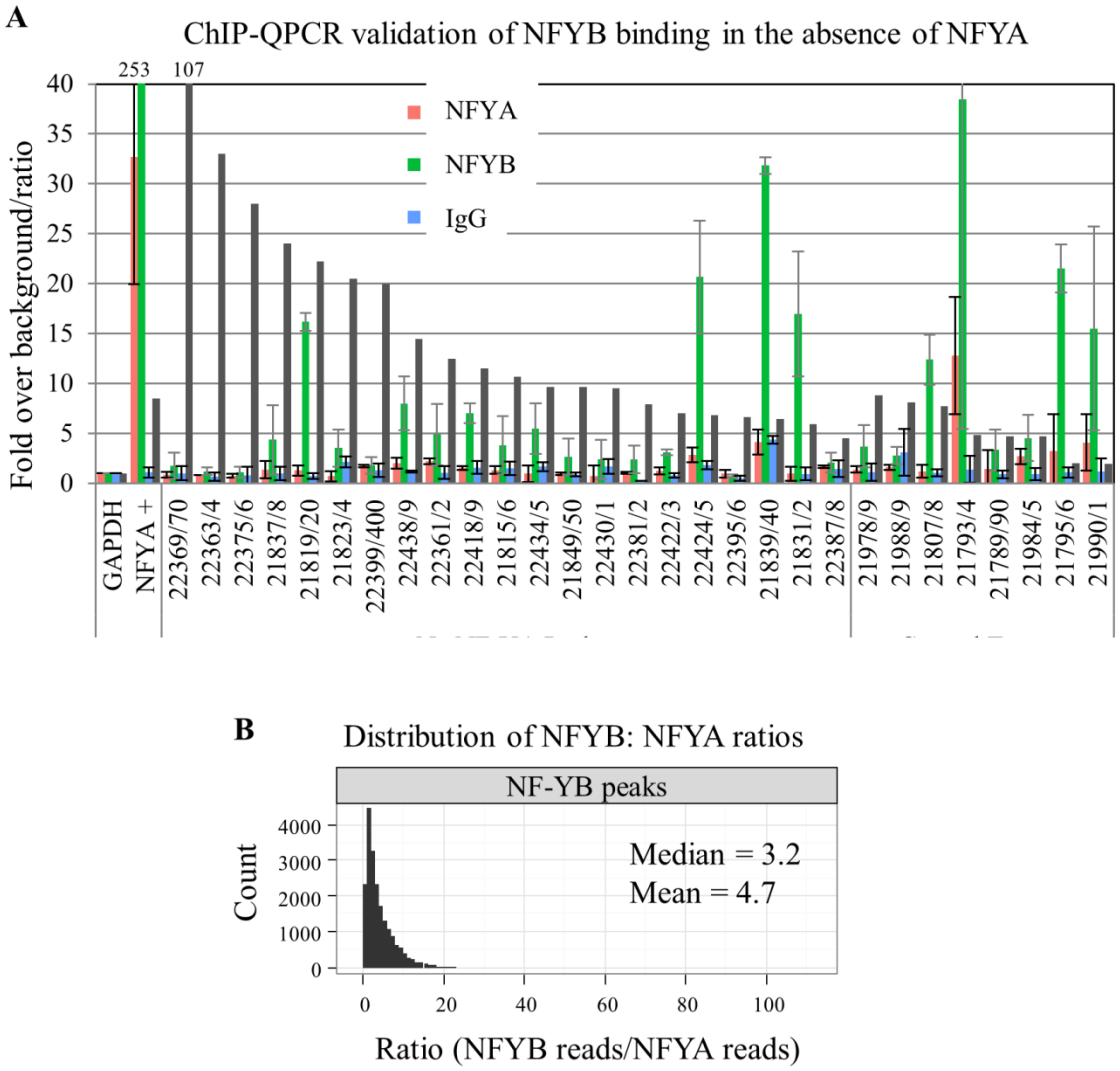


Supplemental Figure 3: ChIP-QPCR validation of NFYB binding in the absence of NFYA

A. NFYB peaks with a high NFYB:NFYA read ratio were assayed by ChIP-QPCR. Selection criteria were: did not overlap an NFYA peak called at a lenient *P*-value threshold of 10^{-5}; and hand-checked by observation of raw ChIP-Seq data and discarded if appreciable NFYA signal was present. A group of control targets that showed similar fold enrichments for NFYB as the test group were selected for comparison. The ratio of NFYB:NFYA reads is shown, and targets are sorted by ratio. The average of 2-4 biological replicates and their associated standard deviations are depicted.

B. Distribution of normalized ratios of NFYB and NFYA ChIP-Seq read counts at NFYB peak regions. Reads were counted within a region spanning +/-100 bp from the summit of NFYB peaks and normalized to the total number of mapped reads.

Supplemental Figure 3 (Continued)



Supplemental Figure 4: NF-Y binds to many genes involved in transcription regulation

A. and B. Transcription regulatory complexes, TFs, RNA Pol II general factors and chromatin associated factors and complexes whose members' putative *cis*-regulatory domains were bound by NFYB. Dark and light green shading indicate NFYB binding within -5 kbp:+1 kbp TSS and -5 kbp:+1 kbp TSS plus up to +/-1 mbp extension, respectively.

C. and D. Ingenuity Pathway Analysis, showing the TP53 (C) and TRAIL (D) signaling pathways. Gray shaded gene terms indicate that that gene's putative *cis*-regulatory domain (-5 kbp:+1 kbp TSS plus up to +/-1 mbp extension) was bound by NFYB in K562.

A

Description	Transcription factor complex			Chromatin remodeling complex		Histone methyltransferase complex	Hetero-chromatin	Histone deacetylase complex	Histone acetyltransferase complex	Transcriptional repressor complex
Term coverage	69%			75%		71%	64%	78%	68%	70%
Genes bound	145			64		42	36	35	34	30
	ABT1	GTF3C6	PROP1	ACTL6A	MORF4L1	ACTB	A1CF	APPL1	ACTB	APPL1
	ALX4	HAND2	PTF1A	ACTL6B	MTA2	AEBP2	BAZ1B	APPL2	ACTL6A	APPL2
	ARNT	HDAC9	PUS1	APPL1	MYSM1	C17orf49	CBX1	CBX5	ATXN7L3	ARID4A
	ARNT2	HELT	RBL1	APPL2	NCOR2	CBX5	CBX3	CHD3	BRD1	C1D
	ARNTL	HES6	RBL2	ARID1A	PHF21A	CHD8	CBX5	CHD4	BRD8	CBX5
	ASCC1	HIF1A	RBM14	ARID1B	RBBP4	E2F6	DNMT1	CIR	BRPF3	CHD3
	ATF1	HMGA1	RELA	ASF1A	RBBP7	EED	DNMT3A	CSNK2A1	C1orf149	CHD4
	ATF4	HNF1A	SATB2	BAHD1	RBM10	EZH1	DNMT3B	EVI1	C20orf20	CORO2A
	ATF7IP	HNF1B	SIX1	BAZ1B	RERE	EZH2	EED	GATAD2A	CREBBP	CSNK2A1
	BSX	HNF4A	SKI	CBX5	RSF1	HDAC9	FOXC1	HDAC1	DMAP1	CTBP2
	CCNH	HOXB13	SMAD2	CHAF1A	SALL1	ING80C	H2AFX	HDAC11	ENY2	DDX20
	CDK2	HOXD12	SMAD3	CHD3	SAP18	JARID2	H2AFY	HDAC4	EP300	ETV3
	CDK4	ING2	SMAD5	CHD4	SAP30	KIAA1076	H2AFY2	HDAC6	EP400	GATAD2A
	CEBPA	IVNS1ABP	SMAD6	CHRAC1	SATB2	KIAA1076	H3F3B	HDAC7	EPC1	HDAC1
	CLOCK	JUN	SMAD7	CIR	SIN3A	KIAA1267	H3F3B	HDAC9	ING3	HDAC4
	CREB1	KAT5	SNAI3	CSNK2A1	SIRT1	LASIL	INCENP	ING2	KAT5	HMGB1
	CREBBP	KLF4	SOX2	DPF1	SIRT2	MAX	MAEL	MBD2	MORF4L1	JAZF1
	CREG1	LBXCOR1	STON1-GTF2A1L	DPF3	SMARCA1	MEN1	MBD2	MBD3	MYST2	MBD3
	CREM	LMO4	SUB1	ESR1	SMARCA2	MLL	MBD3	MORF4L1	MYST4	MTA2
	CRX	MAFB	SUPT3H	EVI1	SMARCA5	MLL3	MECP2	MTA2	PHF15	NCOR2
	CTNNB1	MED17	TAF1	GATAD2A	SMARCC1	MLL5	ORC2L	NCOR2	PHF17	PHF12
	DMBX1	MED27	TAF10	HDAC1	SMARCD1	MYST1	PBX4	PHF21A	RUVBL1	RBBP4
	E2F1	MED7	TAF1A	HDAC11	SMARCD2	OGT	PCGF2	RBBP4	RUVBL2	RBBP7
	E2F2	MEF2B	TAF3	HDAC4	SMARCE1	PAXIP1	RAD18	RBBP7	SAP130	RLIM
	E2F3	MEIS1	TAF4	HDAC6	SUDS3	PELP1	RNF2	RERE	SUPT3H	SALL1
	E2F5	MLXIPL	TAF4B	HDAC7	SUV39H1	PHF20	SALL1	SALL1	TADA1L	SKI
	E2F6	MMS19	TAF5L	HDAC9	TAF6L	PPP1CC	SATB1	SAP18	TAF10	SMARCE1
	E2F7	MYOD1	TAF7	ING2	TAL1	PRPF31	SIRT1	SAP30	TAF4	TBL1XR1
	E2F8	NARG1	TAF8	KIF11	TBL1XR1	RBBP4	SIRT6	SATB2	TAF5L	YWHAB
	ECSIT	NFYA	TBP	MAEL	TOP2B	RBBP7	SMARCC1	SIN3A	TAF6L	ZBTB16
	EDF1	NFYB	TBX2	MBD2	UCHL5	RNF2	SUV39H1	SUDS3	TAF7	
	EP300	NFYC	TCF12	MBD3	ZNF217	RUVBL1	SUZ12	TAF6L	TRRAP	
	EPAS1	NKX2-1	TCF3			RUVBL2	TCP1	TAL1	USP22	
	ERCC2	NKX2-5	TCF4			SETD1A	TNKS1BP1	TBL1XR1	YEATS4	
	ERCC3	NPAS2	TCF7L2			STK38	TOP2B	ZNF217		
	ETS1	NR2E3	TEAD2			SUZ12	UBE2B			
	EYA3	NR6A1	TEAD4			TAF1				
	FOS	ONECUT3	TFAP2D			TAF4				
	FOXE3	PARP1	TFDP2			TAF7				
	FOXF1	PBX1	TFDP3			TEX10				
	GATA6	PBX3	TFEB			UTX				
	GSC	PDLIM1	TP53			WDR5				
	GTF2A1	PITX2	TRRAP							
	GTF2E2	PKNOX1	USF1							
	GTF2F2	PMF1	XRCC6							
	GTF2H3	POU3F1	YY1							
	GTF3C1	POU3F2	ZEB1							
	GTF3C3	PRKDC	ZFHX3							
	GTF3C5									

Supplemental Figure 4 (Continued)

B

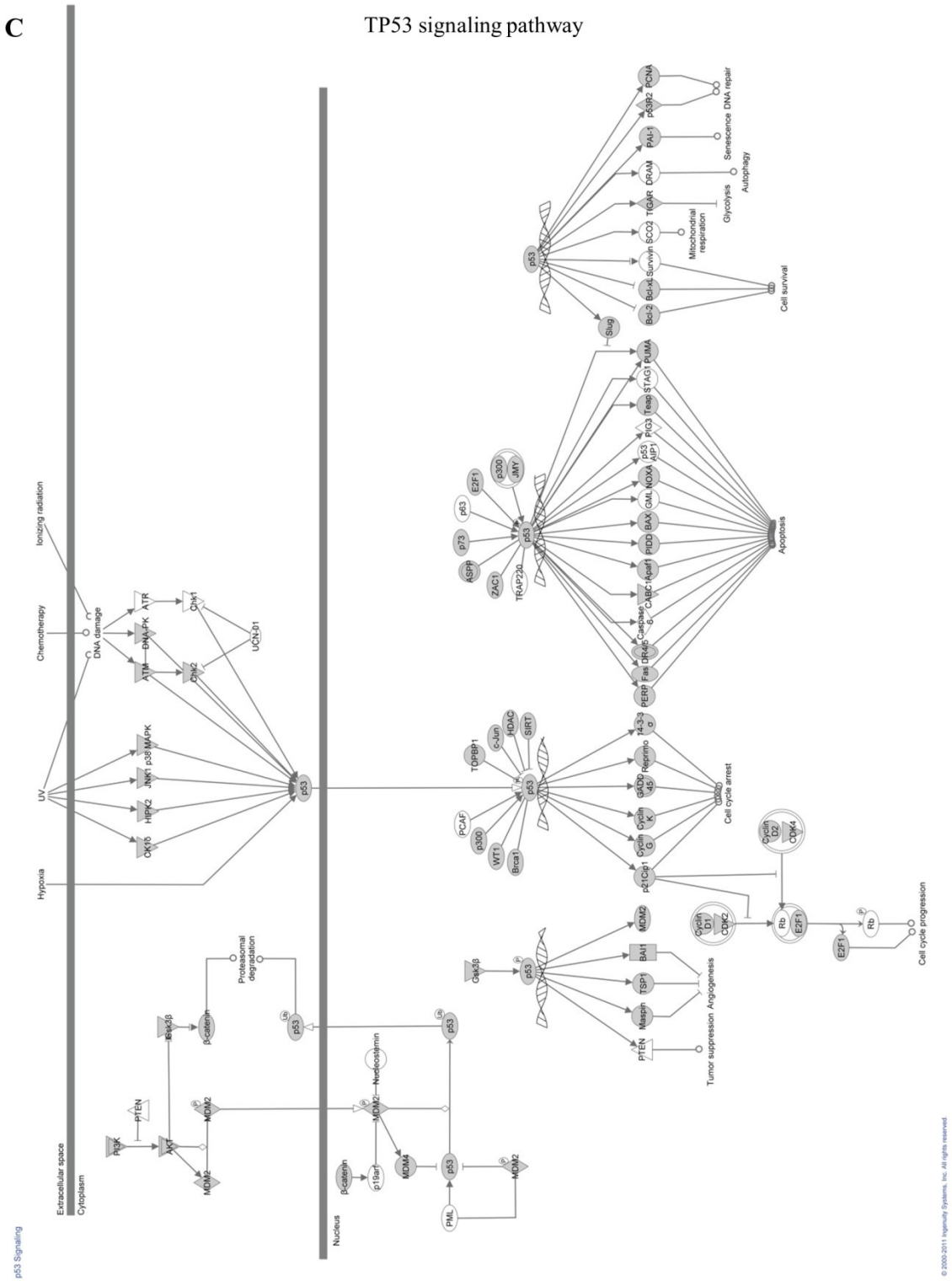
Description	DNA-directed RNA polymerase II, holoenzyme		Mediator complex	MLL1 complex	PcG protein complex	NuA4 histone acetyl-transferase complex	SWI/SNF-type complex	RNA polymerase complex
GO ID	GO:0016591		GO:0016592	GO:0071339	GO:0031519	GO:0035267	GO:0070603	GO:0030880
Term coverage	51%		69%	74%	80%	100%	67%	46%
Genes bound	40		22	20	16	14	14	13
	C19orf2	PAF1	MED9	C17orf49	AEBP2	ACTB	ACTL6A	C19orf2
	CCNH	POLR2A	CDK8	CHD8	ASXL1	ACTL6A	ACTL6B	POLR1A
	CPSF3L	POLR2G	DKFZp434H247	E2F6	BCOR	BRD8	ARID1A	POLR1B
	CTR9	POLR2I	MED10	INO80C	CBX2	C1orf149	ARID1B	POLR1D
	EDF1	PPARGC1A	MED13	KIAA1267	CBX8	DMAP1	BAZ1B	POLR2A
	ELP2	SHFM1	MED13L	LASIL	EED	EP400	CHAF1A	POLR2G
	ELP3	STON1-GTF2A1L	MED14	MAX	EZH1	EPC1	DPF1	POLR2I
	ELP4	SUPT3H	MED15	MLL	EZH2	ING3	DPF3	POLR3A
	ERCC2	TAF1	MED16	MYST1	JARID2	KAT5	SMARCA2	POLR3F
	ERCC3	TAF10	MED17	PELP1	PCGF2	MORF4L1	SMARCC1	POLR3H
	GTF2A1	TAF3	MED18	PHF20	PHC2	RUVBL1	SMARCD1	POLR3K
	GTF2E2	TAF4	MED19	PRPF31	RBBP4	RUVBL2	SMARCD2	PPARGC1A
	GTF2F2	TAF4B	MED21	RNF2	RBBP7	TRRAP	SMARCE1	ZNF768
	GTF2H3	TAF5L	MED24	RUVBL1	RNF2	YEATS4	TOP2B	
	INTS3	TAF7	MED26	RUVBL2	SUZ12			
	INTS4	TAF8	MED29	TAF1	YY1			
	INTS6	TBP	MED30	TAF4				
	INTS7	TP53	MED7	TAF7				
	INTS9	TRRAP	MED8	TEX10				
	MMS19	ZNF768	PPARGC1B	WDR5				
			RBM14					
			THRAP3					

Description	NuRD complex	SAGA-type complex	Transcription elongation factor complex	Transcription factor TFIID complex	SWI/SNF complex	ESC/E(Z) complex	Nup107-160 complex	nBAF complex
GO ID	GO:0016581	GO:0070461	GO:0008023	GO:0005669	GO:0016514	GO:0035098	GO:0031080	GO:0071565
Term coverage	86%	52%	67%	50%	60%	89%	80%	67%
Genes bound	12	12	12	10	9	8	8	8
	APPL1	ATXN7L3	CTR9	EDF1	ACTL6A	AEBP2	AHCTF1	ACTL6B
	APPL2	ENY2	ELL	TAF1	ACTL6B	EED	NUP160	ARID1A
	CHD3	SAP130	ELL2	TAF10	ARID1A	EZH1	NUP37	DPF1
	CHD4	SUPT3H	ELL3	TAF3	ARID1B	EZH2	NUP43	DPF3
	CSNK2A1	TADA1L	ELP2	TAF4	SMARCA2	JARID2	NUP85	SMARCA2
	GATAD2A	TAF10	ELP3	TAF4B	SMARCC1	RBBP4	NUP98	SMARCC1
	HDAC1	TAF4	ELP4	TAF7	SMARCD1	RBBP7	SEC13	SMARCD1
	MBD3	TAF5L	NUFIP1	TAF8	SMARCD2	SUZ12	SEH1L	SMARCE1
	MTA2	TAF6L	PAF1	TBP	SMARCE1			
	RBBP4	TAF7	TAF7	TP53				
	RBBP7	TRRAP	TCEB3					
	SALL1	USP22	TTF2					

Description	Sin3 complex	Catenin complex	STAGA complex	Eukaryotic translation initiation factor 3 complex	npBAF complex	Transcription factor TFTC complex	Chromatin silencing complex	MLL5-L complex
GO ID	GO:0016580	GO:0016342	GO:0030914	GO:0005852	GO:0071564	GO:0033276	GO:0005677	GO:0070688
Term coverage	88%	88%	58%	47%	55%	43%	71%	63%
Genes bound	7	7	7	7	6	6	5	5
	CSNK2A1	APC2	SAP130	EIF3A	ACTL6A	SUPT3H	BAHD1	ACTB
	HDAC1	CDH1	SUPT3H	EIF3D	ARID1A	TAF10	SIRT1	MLL5
	ING2	CTNNA1	TADA1L	EIF3F	SMARCA2	TAF4	SIRT2	OGT
	MORF4L1	CTNNB1	TAF10	EIF3H	SMARCC1	TAF5L	SMARCA5	PPP1CC
	RBBP4	JUP	TAF5L	EIF3J	SMARCD1	TAF7	SUV39H1	STK38
	SIN3A	PVRL1	TAF6L	EIF3L	SMARCE1	TRRAP		
	SUDS3	SMAD7	TRRAP	EIF3M				

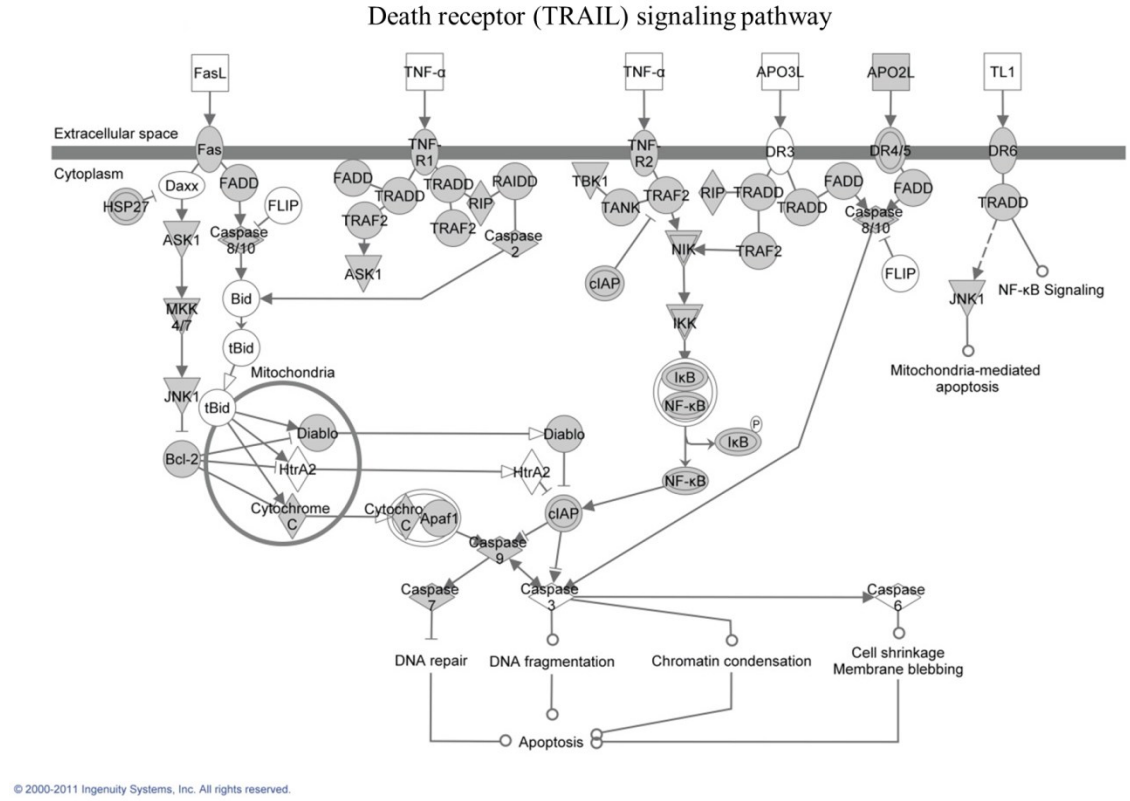
Description	Holo TFIIH complex	CCAAT-binding factor complex	RNA-induced silencing complex	MOZ/MORF histone acetyltransferase complex	Eukaryotic translation initiation factor 4F complex	Eukaryotic translation initiation factor 2B complex	Transcription factor TFIIIC complex	Catenin-TCF7L2 complex
GO ID	GO:0005675	GO:0016602	GO:0016442	GO:0070776	GO:0016281	GO:0005851	GO:0000127	GO:0071664
Term coverage	45%	100%	100%	57%	44%	67%	67%	100%
Genes bound	5	4	4	4	4	4	4	3
	CCNH	ING2	DICER1	BRD1	EIF4A1	EIF2B1	GTF3C1	CTNNB1
	ERCC2	NFYA	EIF2C2	BRPF3	EIF4E	EIF2B2	GTF3C3	JUP
	ERCC3	NFYB	SND1	C1orf149	EIF4G1	EIF2B5	GTF3C5	TCF7L2
	GTF2H3	NFYC	TARBP2	MYST4	EIF4G2	EIF2S1	GTF3C6	
	MMS19							

Supplemental Figure 4 (Continued)



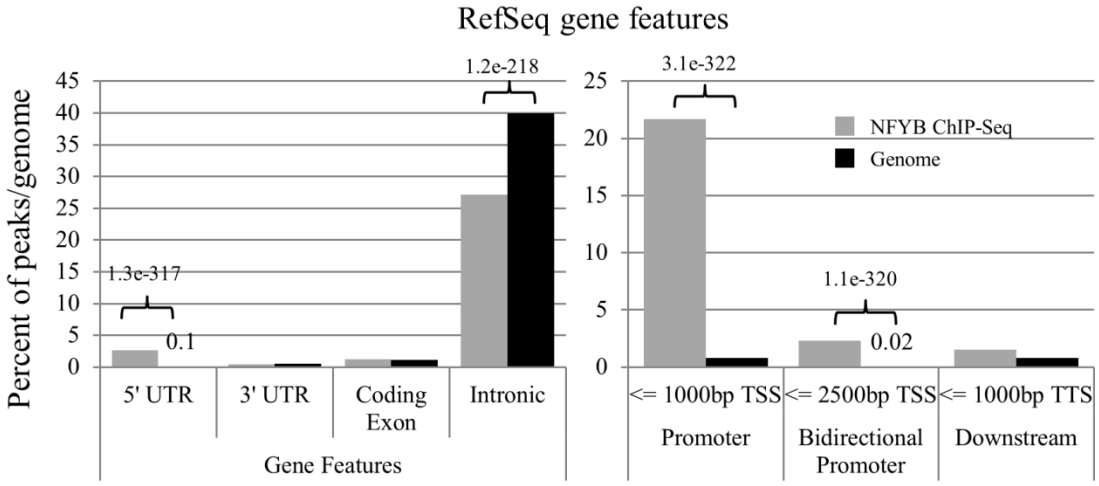
Supplemental Figure 4 (Continued)

D



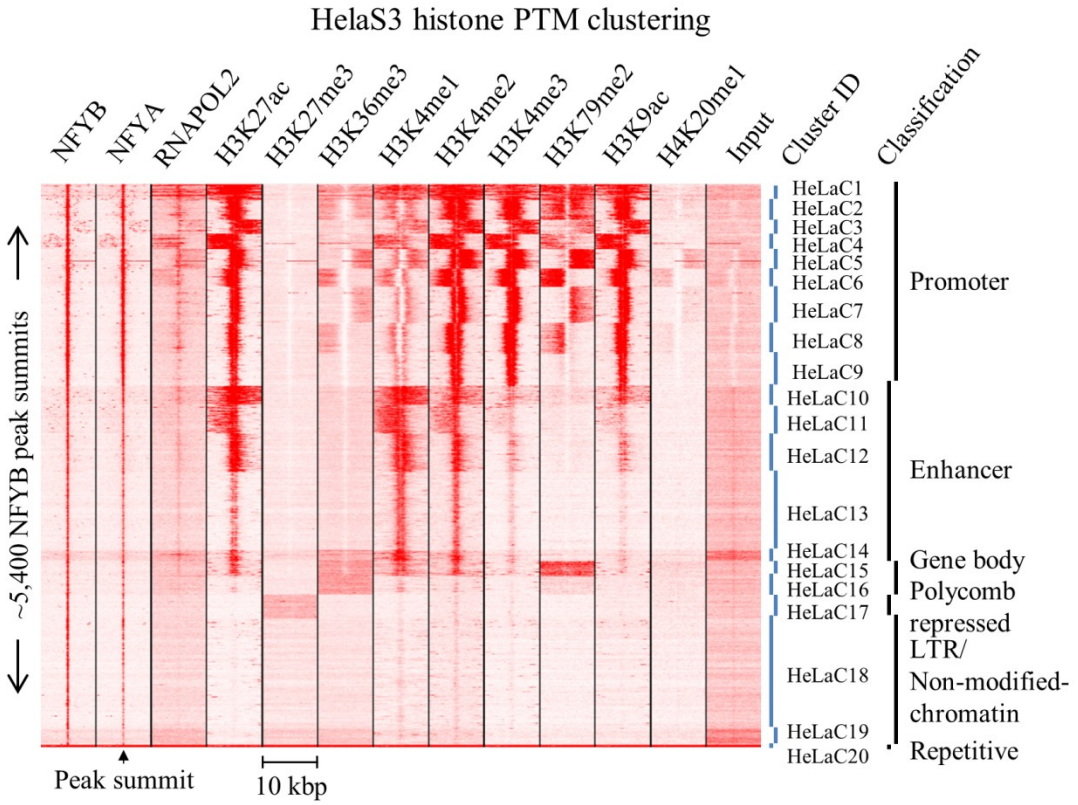
Supplemental Figure 5: Annotation of NF-Y ChIP-Seq peaks to RefSeq gene features

Percent occurrence of K562 NFYB peaks at RefSeq gene features compared to features in the entire genome. *P*-values are indicated.



Supplemental Figure 6: HeLaS3 NF-Y bound loci reside within 5 disparate epigenetic domains

K-means clustering of HeLaS3 NFYB loci based on the distribution of histone PTMs, RNA Pol II, NFYB and NFYA ChIP-Seq reads within a region spanning +/-5 kbp from the summit of NFYB peaks (centered at 0 bp). Clustering was carried out on transformed rank normalized read counts. Raw read count intensity is depicted in red. Similar to Figure 3.



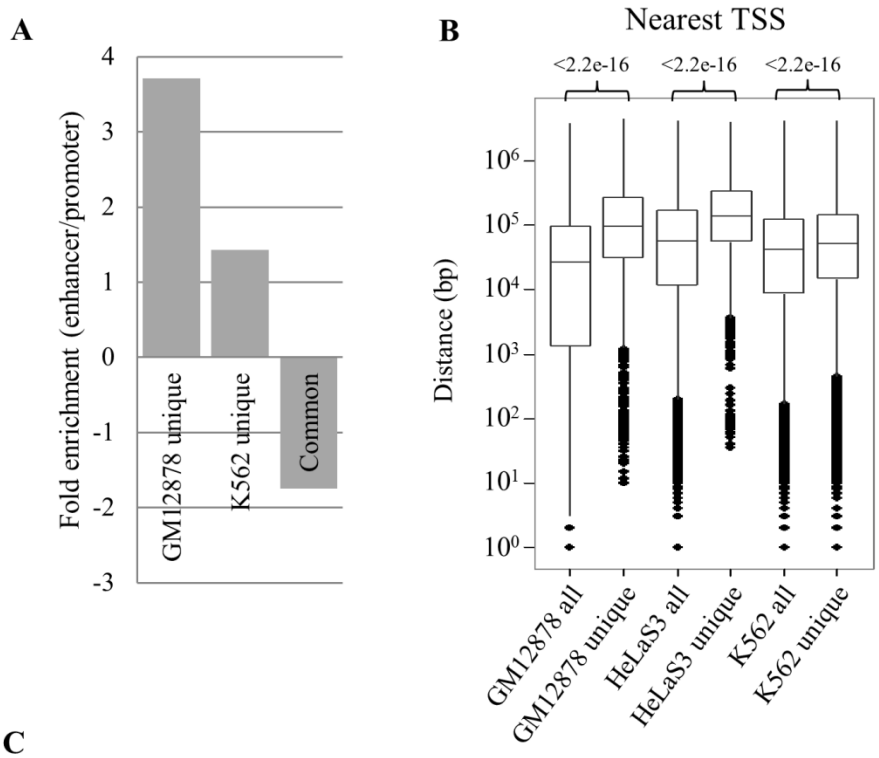
Supplemental Figure 7: NF-Y cell line specific sites are enriched for enhancers and function in cell-type specific biological processes

A. Ratio of enhancer:promoter chromatin states in the GM12878 and K562 cell type specific NFYB binding sites and sites common to all three cell types (K562, GM12878 and HeLaS3). Peaks are considered unique to a cell line if they do not overlap a peak called at the lenient 10^{-5} *P*-value threshold in the other two cell lines.

B. Box plot showing the distance to the nearest RefSeq TSS of NFYB sites. Horizontal edges of the box represent the inter-quartile range. The middle bar represents the median value. Ends of the extensions represent the minimum and maximum datum within 1.5 x inter-quartile range. Outliers are represented as dots. *P*-values represent the significance of the difference in the median value calculated by the Wilcoxon rank sum test.

C. Gene ontology analysis of cell type specific NFYB bound sites unique to K562, GM12878 and HeLaS3. Only the top 10 terms with a fold enrichment > 2 are shown. Observed region hits correspond to the number of NFYB peaks within the regulatory regions of genes in that gene ontology term.

Supplemental Figure 7 (Continued)



C

	GO term	*P*-value	FDR q-value	Fold enrichment	Observed region hits
GM12878	I-kappaB kinase/NF-kappaB cascade	2.1E-06	2.5E-03	3.2	23
	response to metal ion	2.2E-06	2.2E-03	2.0	53
	regulation of interleukin-12 production	4.3E-06	2.6E-03	4.3	15
	toll-like receptor signaling pathway	1.7E-05	6.0E-03	4.6	12
	positive regulation of cytokine production	2.8E-05	8.4E-03	2.3	32
	positive regulation of leukocyte migration	3.3E-05	9.0E-03	3.4	16
	vasodilation	6.4E-05	1.4E-02	3.8	13
	positive regulation of macrophage chemotaxis	8.5E-05	1.5E-02	8.7	6
	regulation of tumor necrosis factor biosynthetic process	8.6E-05	1.5E-02	5.2	9
	immune response-regulating signaling pathway	1.2E-04	1.8E-02	2.4	23
HeLaS3	epidermis morphogenesis	1.2E-10	2.1E-07	4.2	30
	ectoderm and mesoderm interaction	7.7E-10	1.1E-06	9.4	14
	urinary bladder development	1.3E-09	1.2E-06	8.1	15
	establishment of tissue polarity	5.4E-09	3.5E-06	7.3	15
	replicative cell aging	1.3E-08	7.2E-06	7.5	14
	tail morphogenesis	2.1E-08	1.0E-05	7.2	14
	negative regulation of growth	3.2E-08	1.3E-05	2.2	60
	positive regulation of fatty acid metabolic process	3.4E-08	1.4E-05	3.9	24
	hair follicle development	3.6E-08	1.4E-05	2.9	36
	female genitalia development	6.4E-08	1.8E-05	6.5	14
K562	response to cold	1.8E-06	3.8E-04	2.6	32
	negative regulation of Schwann cell proliferation	3.2E-06	6.0E-04	9.4	8
	tRNA processing	1.4E-05	2.1E-03	2.0	44
	negative regulation of activin receptor signaling pathway	1.9E-05	2.5E-03	5.0	11
	ganglioside metabolic process	2.1E-05	2.8E-03	5.0	11
	regulation of histone methylation	3.1E-05	3.8E-03	2.9	20
	rRNA transcription	3.5E-05	4.2E-03	3.8	14
	tRNA splicing, via endonucleolytic cleavage and ligation	8.0E-05	8.3E-03	11.8	5
	cellular response to hypoxia	8.9E-05	8.7E-03	5.1	9
	anterior/posterior axis specification	9.7E-05	9.0E-03	2.2	31

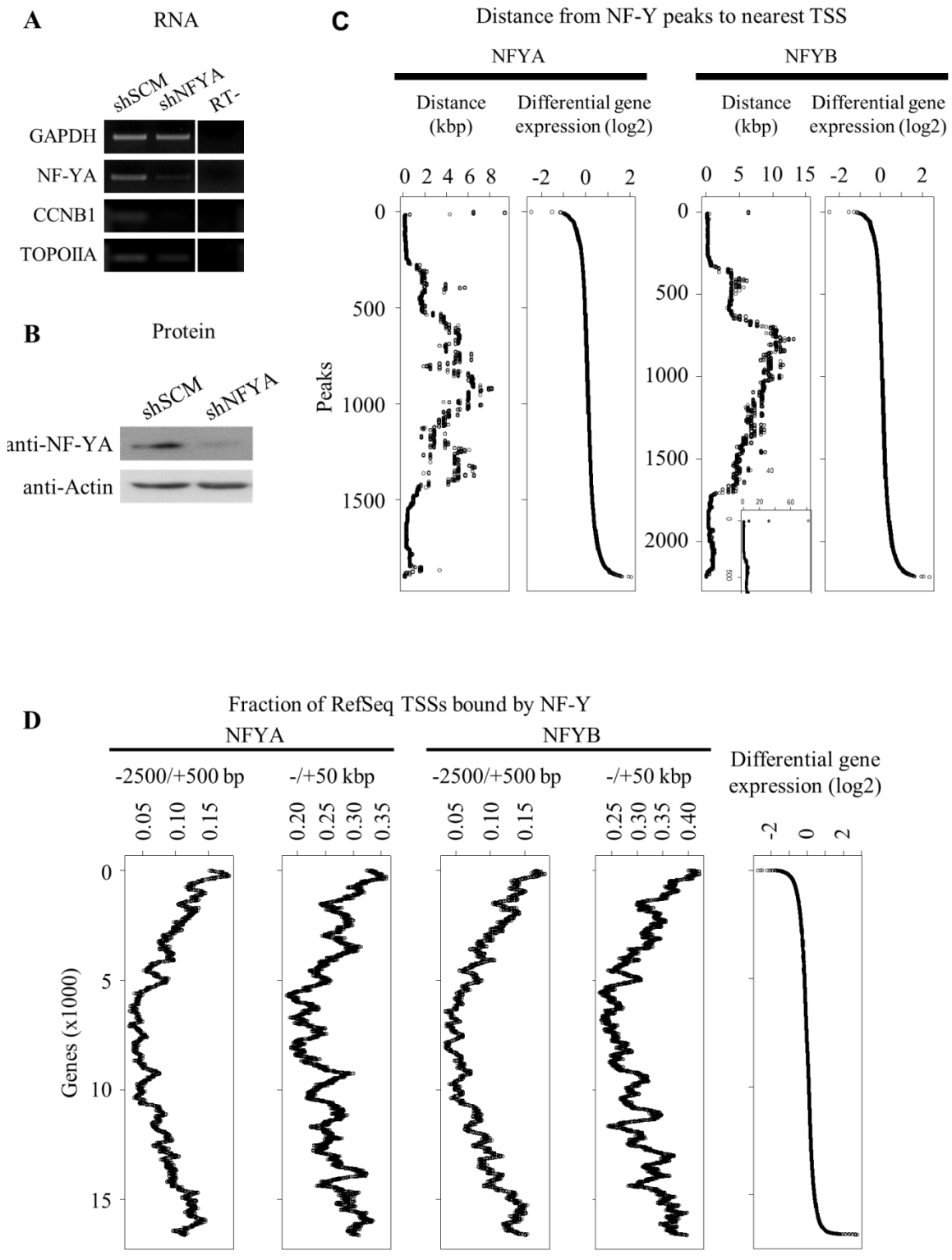
Supplemental Figure 8: Functional inactivation of NFYA and correlation with ChIP-Seq NF-Y sites

A. and B. Representative (A) semi-quantitative PCR and (B) Western blot analysis of an NFYA and scrambled control lentiviral shRNA knockdown in HeLaS3. A. Clear reduction in the mRNA for NFYA is apparent, whereas a control gene, *GAPDH*, was unaffected. *CCNB1* and *TOPOIIA* are known NF-Y regulated genes and are included as positive controls. A reverse transcriptase negative control (RT-) is also shown. B. Membranes were blotted with anti-NFYA or anti-Actin antibodies (control) and show a specific reduction in NFYA protein levels.

C. NF-Y ChIP-Seq peaks were near differentially regulated genes. NFYA or NFYB peaks, excluding peaks that overlapped LTRs, were mapped to the nearest RefSeq TSS and the distance and associated differential gene expression upon shNFYA of that gene determined. Peaks are sorted based on differential gene expression and the median distance of a sliding 200 peak window is shown. Inset, NFYB distance plot rescaled to show data points > 15 kbp.

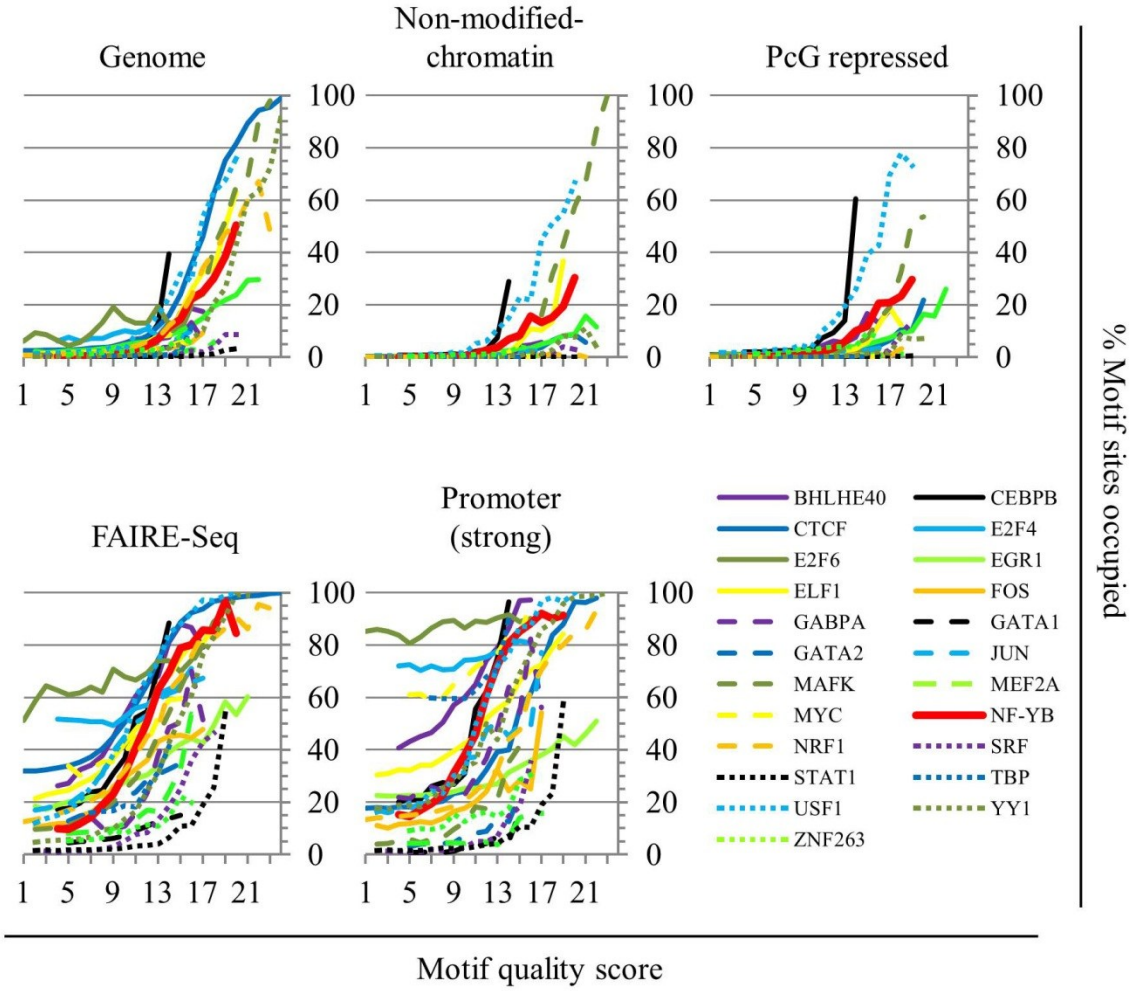
D. The most differentially regulated genes were associated with the highest fraction of TSSs bound by NFYA. The fraction (mean of 500 bp sliding window) of RefSeq TSSs with an NFYA or NFYB ChIP-Seq peak within the indicated range of the TSS, ranked according to differential gene expression upon shNFYA of the associated genes.

Supplemental Figure 8 (Continued)



Supplemental Figure 9: TFs have marked differences in their ability to bind their motif in closed chromatin

The percentage of genome-wide computationally discovered TF binding site motif locations within non-modified-chromatin, PcG repressed and strong promoter chromatin states, FAIRE-Seq regions or the entire genome, that directly overlapped their respective TF sites plotted as a function of motif quality (right axes). Similar to Figure 5, B.



Supplemental Figure 10: NFYB significantly co-associates with many factors at promoters and enhancers

A. The significance of co-association with NFYB at K562 strong promoter and enhancer chromatin states. The number of promoters and enhancers bound by NFYB and one of each individual chromatin associated proteins was assayed by a 2x2 contingency table approach. The significance of the observed overlap was determined by the Fisher exact test. Peak summits from the 10^{-9} peak lists were used to determine occupancy within a given region.

B. Dendrograms depicting the correlation between chromatin associated factors at NFYB bound or NFYB non-bound promoters or enhancers in K562. All promoters and enhancers from the chromatin state maps where scored for the presence/absence of all chromatin associated factors and clustered (see Methods). NFYA and NFYB are indicated by arrows and the cluster they associate with is shaded in yellow.

C. Multi-way overlaps between chromatin associated factors (RNA Pol II/III and associated general factors were not considered) at NFYB bound and NFYB non-bound strong promoters and all enhancers. Only the top 10 combinations are shown. The number and percentage of promoters or enhancers that were simultaneously bound by the indicated factor(s) are shown. Yellow shading represents FOS, which is highly prevalent at NFYB bound promoters.

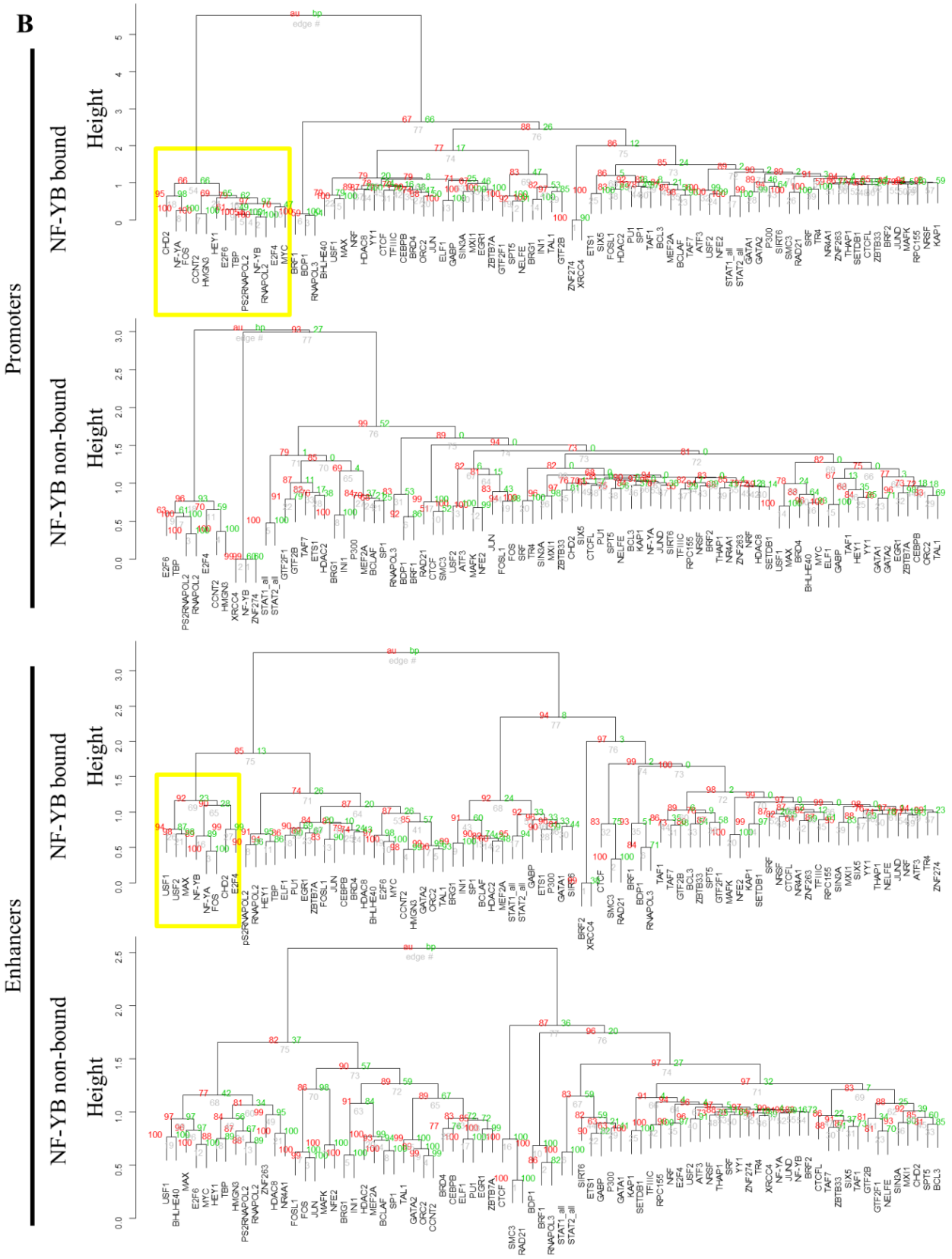
Supplemental Figure 10 (Continued)

A

Promoters (strong)					
Overlapping factor	Genomic peaks	Regions with factor	Expected overlap	Actual overlap	*P*-value
RNAPOL2	23,586	9,809	1,697	2,485	<1.0E-300
FOS	14,404	1,995	345	1,552	<1.0E-300
TBP	14,496	8,225	1,423	2,324	<1.0E-300
CHD2	6,932	2,752	476	1,181	8.7E-280
E2F4	9,862	6,058	1,048	1,875	4.0E-276
pS2RNAPOL2	29,410	8,782	1,520	2,304	1.1E-271
MYC	13,693	6,228	1,078	1,809	3.3E-215
E2F6	20,609	8,036	1,391	2,107	1.6E-212
HEY1	9,229	5,051	874	1,507	8.6E-170
HMGN3	18,815	7,180	1,242	1,810	6.2E-129
ORC2	15,401	3,231	559	1,037	2.0E-122
CCNT2	20,895	6,742	1,167	1,708	2.3E-117
ELF1	17,951	3,519	609	1,038	3.6E-95
GTF2B	2,475	1,501	260	569	1.8E-90
JUN	18,480	2,664	461	831	3.2E-85
MAX	6,402	2,160	374	709	2.2E-81
GABPA	5,025	2,450	424	750	4.3E-71
BHLHE40	16,358	3,568	617	986	6.6E-71
HDAC8	9,860	1,986	344	632	3.4E-65
USF1	21,313	1,631	282	546	1.1E-63
CEBPB	44,168	1,881	325	589	1.9E-57
YY1	5,250	2,658	460	722	6.6E-45
SIN3A	2,701	1,041	180	361	8.2E-45
NRF1	3,328	1,823	315	540	7.5E-44
MXI1	3,020	1,358	235	417	4.0E-37
USF2	1,623	455	79	184	2.7E-32
TFIIIC	10,004	1,964	340	504	1.3E-23
GTF2F1	885	385	67	147	8.5E-23
SP1	5,576	302	52	124	9.5E-23
TFIIF	10,662	4,857	840	1,056	3.4E-22
EGR1	19,094	1,914	331	478	1.1E-19
TAL1	24,841	1,194	207	323	2.2E-18
BRD4	10,746	3,318	574	747	3.0E-18
NELFE	1,136	322	56	120	7.0E-18
JUND	945	262	45	101	2.2E-16
CTCF	46,476	2,860	495	640	1.1E-14
ATF3	939	436	75	141	1.2E-14
SPT5	1,839	248	43	93	2.7E-14
THAP1	1,606	419	73	135	5.1E-14
SIX5	3,397	309	53	99	2.3E-10

Enhancers (all)					
Overlapping factor	Genomic peaks	Regions with factor	Expected overlap	Actual overlap	*P*-value
FOS	14,404	5,969	76	1,246	<1.0E-300
E2F4	9,862	1,587	20	384	<1.0E-300
E2F6	20,609	6,617	85	619	<1.0E-300
RNAPOL2	23,586	5,706	73	587	<1.0E-300
TBP	14,496	3,494	45	471	<1.0E-300
USF1	21,313	6,384	82	869	<1.0E-300
USF2	1,623	517	7	242	<1.0E-300
HMGN3	18,815	4,655	59	503	2.5E-300
CCNT2	20,895	8,639	110	649	1.9E-297
CHD2	6,932	1,205	15	302	1.9E-290
MYC	13,693	4,985	64	478	2.5E-261
MAX	6,402	2,002	26	321	4.7E-243
ELF1	17,951	7,866	100	539	5.4E-225
pS2RNAPOL2	29,410	4,819	62	431	1.4E-222
ORC2	15,401	8,890	114	562	4.0E-218
HEY1	9,229	2,340	30	296	1.9E-193
BHLHE40	16,358	7,291	93	468	2.2E-182
JUN	18,480	10,399	133	526	6.6E-161
HDAC8	9,860	5,473	70	335	4.4E-123
SP1	5,576	2,228	28	204	9.0E-106
EGR1	19,094	8,427	108	386	3.4E-103
GATA2	9,025	5,876	75	314	6.2E-100
TAL1	24,841	15,864	203	545	2.2E-98
ZBTB7A	8,031	3,773	48	235	2.5E-87
SIX5	3,397	330	4	81	2.5E-77
GTF2B	2,475	452	6	88	2.1E-74
GABPA	5,025	1,212	15	128	6.3E-74
TFIIF	10,662	3,308	42	197	5.6E-70
BRD4	10,746	5,039	64	240	1.2E-66
CEBPB	44,168	11,470	146	387	2.0E-66
STAT2_30m	2,514	1,580	20	128	2.0E-60
HDAC2	8,831	4,700	60	219	3.4E-59
PU1	25,479	8,424	108	302	7.6E-57
NRF1	3,328	478	6	74	3.1E-55
GATA1	3,182	1,665	21	120	2.3E-51
FOSL1	11,393	6,742	86	246	1.7E-47
ATF3	939	78	1	35	4.6E-45
CTCF	46,476	6,031	77	224	1.8E-44
BRG1	11,209	5,363	68	196	5.0E-38
NFE2	3,477	2,213	28	118	1.1E-37
MEF2A	10,209	4,337	55	170	1.1E-36
ETS1	2,607	1,153	15	82	4.1E-35
BCLAF1	6,616	2,712	35	127	7.1E-35
MAFK	17,914	6,532	83	212	1.2E-33
MXI1	3,020	517	7	55	1.0E-32
YY1	5,250	791	10	66	1.4E-32
STAT2_6h	2,174	1,149	15	78	4.1E-32
TFIIIC	10,004	3,164	40	127	1.3E-28
NR4A1	5,514	1,685	22	85	6.6E-26
P300	2,969	1,038	13	63	1.2E-23
TAF7	4,536	1,237	16	65	4.5E-21
SIN3A	2,701	318	4	34	6.8E-21
JUND	945	256	3	31	8.8E-21
SIRT6	1,794	1,227	16	64	1.3E-20
GTF2F1	885	90	1	20	2.7E-19
SRF	2,005	666	9	39	9.9E-15
TAF1	4,126	573	7	36	1.3E-14
BCL3	3,924	492	6	32	1.4E-13

Supplemental Figure 10 (Continued)



Supplemental Figure 10 (Continued)

C

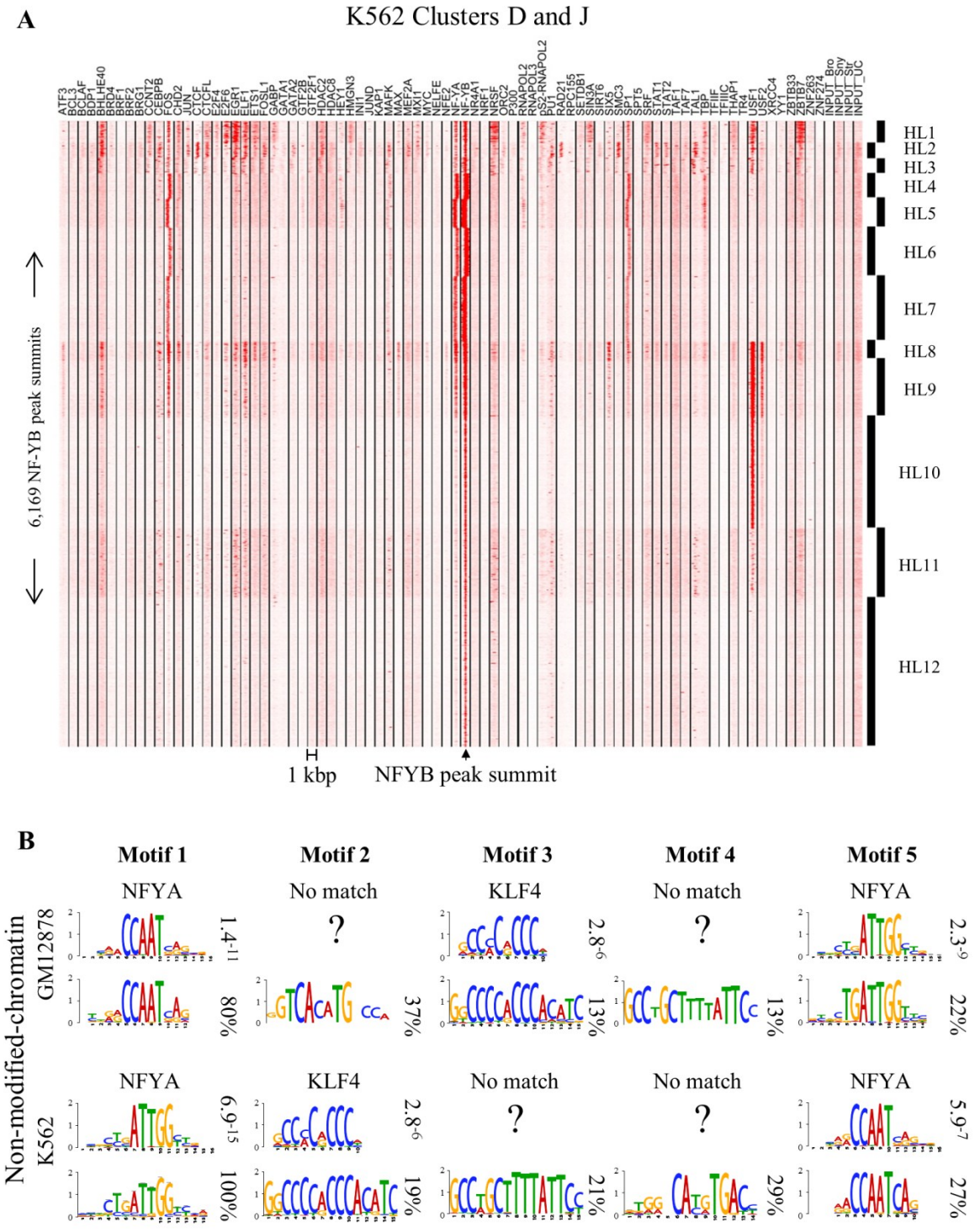
	NF-YB Bound Strong Promoters			NF-YB Non-Bound Strong Promoters			NF-YB Bound Enhancers			NF-YB Non-Bound Enhancers		
		#	%		#	%		#	%		#	%
1-way overlap	E2F6	2107	77.9	E2F6	1749	28.3	FOS	1246	39.0	TAL1	8379	11.7
	E2F4	1875	69.3	HMGN3	1638	26.5	USF1	869	27.2	CEBPB	6198	8.7
	HMGN3	1810	66.9	E2F4	944	15.3	E2F6	619	19.4	JUN	5639	7.9
	MYC	1809	66.9	MYC	866	14.0	ORC2	562	17.6	PU1	4654	6.5
	FOS	1552	57.4	HEY1	762	12.3	TAL1	545	17.1	EGR1	4633	6.5
	HEY1	1507	55.7	BRD4	717	11.6	ELF1	539	16.9	ORC2	4572	6.4
	CHD2	1181	43.7	BHLHE40	656	10.6	JUN	526	16.5	ELF1	4227	5.9
	ELF1	1038	38.4	CTCF	610	9.9	HMGN3	503	15.8	BHLHE40	3906	5.5
	ORC2	1037	38.4	YY1	540	8.7	MYC	478	15.0	CTCF	3844	5.4
	BHLHE40	986	36.5	ELF1	516	8.3	BHLHE40	468	14.7	FOSL1	3757	5.3
2-way overlap	E2F6-E2F4	1654	61.2	HMGN3-E2F6	1074	17.4	USF1-FOS	412	12.9	TAL1-ORC2	3166	4.4
	E2F6-MYC	1616	59.8	HMGN3-E2F4	746	12.1	TAL1-ORC2	346	10.8	FOSL1-JUN	2555	3.6
	HMGN3-E2F6	1578	58.4	E2F6-E2F4	715	11.6	JUN-FOS	339	10.6	TAL1-GATA2	2536	3.6
	HMGN3-E2F4	1489	55.1	E2F6-MYC	679	11.0	E2F6-FOS	307	9.6	JUN-FOS	2476	3.5
	E2F4-MYC	1443	53.4	HMGN3-MYC	596	9.6	ORC2-JUN	296	9.3	TAL1-CEBPB	2371	3.3
	HMGN3-MYC	1351	50.0	E2F6-BHLHE40	450	7.3	E2F6-MYC	291	9.1	ORC2-GATA2	2123	3.0
	HEY1-E2F6	1235	45.7	HMGN3-BRD4	442	7.1	HMGN3-FOS	280	8.8	ORC2-JUN	2078	2.9
	E2F6-FOS	1218	45.0	E2F4-MYC	431	7.0	TAL1-GATA2	275	8.6	TAL1-EGR1	1976	2.8
	HEY1-MYC	1144	42.3	HMGN3-BHLHE40	408	6.6	ORC2-FOS	269	8.4	TAL1-JUN	1945	2.7
	E2F4-FOS	1105	40.9	E2F6-BRD4	401	6.5	ORC2-GATA2	261	8.2	FOSL1-FOS	1824	2.6
3-way overlap	HMGN3-E2F6-E2F4	1353	50.0	HMGN3-E2F6-E2F4	590	9.5	TAL1-ORC2-GATA2	246	7.7	TAL1-ORC2-GATA2	1988	2.8
	E2F6-E2F4-MYC	1343	49.7	HMGN3-E2F6-MYC	512	8.3	ORC2-JUN-FOS	200	6.3	FOSL1-JUN-FOS	1790	2.5
	HMGN3-E2F6-MYC	1253	46.3	E2F6-E2F4-MYC	382	6.2	Usf2-USF1-FOS	189	5.9	TAL1-ORC2-JUN	1322	1.9
	HMGN3-E2F4-MYC	1193	44.1	HMGN3-E2F4-MYC	376	6.1	TAL1-ORC2-JUN	181	5.7	TAL1-ORC2-CEBPB	1255	1.8
	HEY1-E2F6-MYC	1020	37.7	HMGN3-E2F6-BHLHE40	339	5.5	TAL1-ORC2-MYC	173	5.4	TAL1-ORC2-HDAC2	1202	1.7
	HEY1-E2F6-E2F4	982	36.3	HMGN3-E2F6-BRD4	338	5.5	ORC2-MYC-JUN	169	5.3	TAL1-ORC2-BHLHE40	1173	1.6
	E2F6-E2F4-FOS	981	36.3	HMGN3-E2F4-BRD4	296	4.8	TAL1-ORC2-BHLHE40	164	5.1	TAL1-ORC2-EGR1	1168	1.6
	E2F6-MYC-FOS	941	34.8	E2F6-MYC-BHLHE40	272	4.4	ORC2-MYC-BHLHE40	159	5.0	TAL1-FOSL1-JUN	1144	1.6
	HEY1-E2F4-MYC	907	33.5	E2F6-E2F4-BRD4	261	4.2	ORC2-E2F6-MYC	159	5.0	TAL1-HDAC2-GATA2	1122	1.6
	HMGN3-E2F6-FOS	903	33.4	HMGN3-E2F4-BHLHE40	253	4.1	FOSL1-JUN-FOS	158	4.9	ORC2-JUN-FOS	1110	1.6
4-way overlap	HMGN3-E2F4-E2F6-MYC	1124	41.6	HMGN3-E2F6-E2F4-MYC	337	5.4	TAL1-ORC2-GATA2-MYC	138	4.3	TAL1-ORC2-HDAC2-GATA2	953	1.3
	HEY1-E2F4-E2F6-MYC	842	31.1	HMGN3-E2F6-E2F4-BRD4	246	4.0	TAL1-ORC2-GATA2-JUN	128	4.0	TAL1-ORC2-GATA2-CEBPB	898	1.3
	E2F6-MYC-E2F4-FOS	792	29.3	HMGN3-E2F6-E2F4-BHLHE40	224	3.6	TAL1-ORC2-GATA2-BHLHE40	124	3.9	TAL1-ORC2-GATA2-EGR1	863	1.2
	HMGN3-E2F4-E2F6-FOS	787	29.1	HMGN3-E2F6-MYC-BHLHE40	222	3.6	TAL1-ORC2-HMGN3-GATA2	122	3.8	TAL1-ORC2-GATA2-BHLHE40	840	1.2
	HMGN3-E2F6-HEY1-E2F4	769	28.4	HMGN3-E2F6-MYC-BRD4	208	3.4	TAL1-ORC2-HDAC2-GATA2	121	3.8	TAL1-ORC2-GATA2-JUN	836	1.2
	HMGN3-E2F6-HEY1-MYC	753	27.8	HMGN3-E2F4-MYC-BRD4	186	3.0	TAL1-ORC2-JUN-FOS	118	3.7	ORC2-FOSL1-JUN-FOS	826	1.2
	HMGN3-MYC-E2F6-FOS	722	26.7	E2F6-E2F4-MYC-BHLHE40	180	2.9	ORC2-MYC-JUN-FOS	117	3.7	TAL1-ORC2-GATA2-MYC	803	1.1
	HMGN3-E2F4-HEY1-MYC	721	26.7	HMGN3-E2F4-MYC-BHLHE40	176	2.8	TAL1-ORC2-GATA2-ELF1	117	3.7	TAL1-FOSL1-JUN-FOS	779	1.1
	E2F6-MYC-E2F4-CHD2	710	26.3	E2F6-MYC-E2F4-BRD4	172	2.8	Usf2-USF1-Max-FOS	116	3.6	TAL1-ORC2-FOSL1-JUN	763	1.1
	HMGN3-E2F4-E2F6-CHD2	699	25.9	HMGN3-HEY1-E2F6-MYC	138	2.2	ORC2-FOSL1-JUN-FOS	111	3.5	TAL1-ORC2-HMGN3-GATA2	761	1.1

Supplemental Figure 11: NF-Y partners with FOS, USF1, USF2 and SP1 in non-modified-chromatin domains

A. K-means clustering of K562 NFYB loci from the non-modified-chromatin class (clusters D and J; Figure 3, A) based on the distribution of ChIP-Seq reads from chromatin associated factors within a region spanning +/-500 bp from the summit of NFYB peaks (centered at 0 bp). Clustering was carried out on transformed rank normalized read counts. Raw read count intensity is depicted in red.

B. *De novo* motif search of NFYB peaks in the non-modified-chromatin state. Only the top 5 motifs are shown. The respective best match (*P*-value shown to right) to known motifs are shown on top of the discovered motifs. The percentage of NFYB sites containing the discovered motif is indicated to the right. In some instances the very similar Hap3 (yeast NF-Y orthologue) motif was replaced by the NFYA motif which was second to Hap3 in all cases.

Supplemental Figure 11 (Continued)



APPENDIX B: A User's Guide to the Encyclopedia of DNA Elements (ENCODE)

AUTHOR CONTRIBUTIONS

J.F. provided ChIP-Seq datasets as part of the Consortium's effort.

OPEN ACCESS Freely available online PLoS BIOLOGY

A User's Guide to the Encyclopedia of DNA Elements (ENCODE)

The ENCODE Project Consortium¶*

Abstract

The mission of the Encyclopedia of DNA Elements (ENCODE) Project is to enable the scientific and medical communities to interpret the human genome sequence and apply it to understand human biology and improve health. The ENCODE Consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns. In the process, standards to ensure high-quality data have been implemented, and novel algorithms have been developed to facilitate analysis. Data and derived results are made available through a freely accessible database. Here we provide an overview of the project and the resources it is generating and illustrate the application of ENCODE data to interpret the human genome.

Citation: The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046

Academic Editor: Peter B. Becker, Adolf Butenandt Institute, Germany

Received September 23, 2010; **Accepted** March 10, 2011; **Published** April 19, 2011

Funding: Funded by the National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. The role of the NIH Project Management Group in the preparation of this paper was limited to coordination and scientific management of the ENCODE Consortium.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: 3C, Chromosome Conformation Capture; API, application programming interface; CAGE, Cap-Analysis of Gene Expression; ChIP, chromatin immunoprecipitation; DCC, Data Coordination Center; DHS, DNaseI hypersensitive site; ENCODE, Encyclopedia of DNA Elements; EPO, Enredo, Pecan, Ortheus approach; FDR, false discovery rate; GEO, Gene Expression Omnibus; GWAS, genome-wide association studies; IDR, Irreproducible Discovery Rate; Methyl-seq, sequencing-based methylation determination assay; NHGRI, National Human Genome Research Institute; PASRs, promoter-associated short RNAs; PET, Paired-End diTag; RACE, Rapid Amplification of cDNA Ends; RNA Pol2, RNA polymerase 2; RBP, RNA-binding protein; RRBS, Reduced Representation Bisulfite Sequencing; SRA, Sequence Read Archive; TAS, trait/disease-associated SNP; TF, transcription factor; TSS, transcription start site

* E-mail: rmyers@hudsonalpha.org (RMM); jstam@u.washington.edu (JS); mpsnyder@stanford.edu (MS); dunham@ebi.ac.uk (ID); rch8@psu.edu (RCH); bernstein.bradley@mgh.harvard.edu (BEB); gingeras@cshl.edu (TRG); kent@soe.ucsc.edu (WJK); birney@ebi.ac.uk (EB); woldb@caltech.edu (BW); greg.crawford@duke.edu (GEC)

¶ Membership of the ENCODE Project Consortium is provided in the Acknowledgments.

I. Introduction and Project Overview

Interpreting the human genome sequence is one of the leading challenges of 21st century biology [1]. In 2003, the National Human Genome Research Institute (NHGRI) embarked on an ambitious project the Encyclopedia of DNA Elements (ENCODE) aiming to delineate all of the functional elements encoded in the human genome sequence [2]. To further this goal, NHGRI organized the ENCODE Consortium, an international group of investigators with diverse backgrounds and expertise in production and analysis of high-throughput functional genomic data. In a pilot project phase spanning 2003 2007, the Consortium applied and compared a variety of experimental and computational methods to annotate functional elements in a defined 1% of the human genome [3]. Two additional goals of the pilot ENCODE Project were to develop and advance technologies for annotating the human genome, with the combined aims of achieving higher accuracy, completeness, and cost-effective throughput and establishing a paradigm for sharing functional genomics data. In 2007, the ENCODE Project was expanded to study the entire human genome, capitalizing on experimental and computational technology developments during the pilot project period. Here we describe this expanded project, which we refer to throughout as the ENCODE Project, or ENCODE.

The major goal of ENCODE is to provide the scientific community with high-quality, comprehensive annotations of candidate functional elements in the human genome. For the purposes of this article, the term "functional element" is used to denote a discrete region of the genome that encodes a defined product (e.g., protein) or a reproducible biochemical signature, such as transcription or a specific chromatin structure. It is now widely appreciated that such signatures, either alone or in combinations, mark genomic sequences with important functions, including exons, sites of RNA processing, and transcriptional regulatory elements such as promoters, enhancers, silencers, and insulators. However, it is also important to recognize that while certain biochemical signatures may be associated with specific functions, our present state of knowledge may not yet permit definitive declaration of the ultimate biological role(s), function(s), or mechanism(s) of action of any given genomic element.

At present, the proportion of the human genome that encodes functional elements is unknown. Estimates based on comparative genomic analyses suggest that 3% 8% of the base pairs in the human genome are under purifying (or negative) selection [4 7]. However, this likely underestimates the prevalence of functional features, as current comparative methods may not account for lineage-specific evolutionary innovations, functional elements that are very small or fragmented [8], elements that are rapidly evolving or subject to nearly neutral evolutionary processes, or elements that lie in repetitive regions of the genome.

The current phase of the ENCODE Project has focused on completing two major classes of annotations: genes (both protein-coding and non-coding) and their RNA transcripts, and transcriptional regulatory regions. To accomplish these

Author Summary

The Encyclopedia of DNA Elements (ENCODE) Project was created to enable the scientific and medical communities to interpret the human genome sequence and to use it to understand human biology and improve health. The ENCODE Consortium, a large group of scientists from around the world, uses a variety of experimental methods to identify and describe the regions of the 3 billion base-pair human genome that are important for function. Using experimental, computational, and statistical analyses, we aimed to discover and describe genes, transcripts, and transcriptional regulatory regions, as well as DNA binding proteins that interact with regulatory regions in the genome, including transcription factors, different versions of histones and other markers, and DNA methylation patterns that define states of the genome in various cell types. The ENCODE Project has developed standards for each experiment type to ensure high-quality, reproducible data and novel algorithms to facilitate analysis. All data and derived results are made available through a freely accessible database. This article provides an overview of the complete project and the resources it is generating, as well as examples to illustrate the application of ENCODE data as a user's guide to facilitate the interpretation of the human genome.

goals, seven ENCODE Data Production Centers encompassing 27 institutions have been organized to focus on generating multiple complementary types of genome-wide data (Figure 1 and Figure S1). These data include identification and quantification of RNA species in whole cells and in sub-cellular compartments, mapping of protein-coding regions, delineation of chromatin and DNA accessibility and structure with nucleases and chemical probes, mapping of histone modifications and transcription factor (TF) binding sites by chromatin immunoprecipitation (ChIP), and measurement of DNA methylation (Figure 2 and Table 1). In parallel with the major production efforts, several smaller-scale efforts are examining long-range chromatin interactions, localizing binding proteins on RNA, identifying transcriptional silencer elements, and understanding detailed promoter sequence architecture in a subset of the genome (Figure 1 and Table 1).

ENCODE has placed emphasis on data quality, including ongoing development and application of standards for data reproducibility and the collection of associated experimental information (i.e., metadata). Adoption of state-of-the-art, massively parallel DNA sequence analysis technologies has greatly facilitated standardized data processing, comparison, and integration [9,10]. Primary and processed data, as well as relevant experimental methods and parameters, are collected by a central Data Coordination Center (DCC) for curation, quality review, visualization, and dissemination (Figure 1). The Consortium releases data rapidly to the public through a web-accessible database (http://genome.ucsc.edu/ENCODE/) [11] and provides a visualization framework and analytical tools to facilitate use of the data [12], which are organized into a web portal (http://encodeproject.org).

To facilitate comparison and integration of data, ENCODE data production efforts have prioritized selected sets of cell types (Table 2). The highest priority set (designated "Tier 1") includes two widely studied immortalized cell lines K562 erythroleukemia cells [13]; an EBV-immortalized B-lymphoblastoid line (GM12878, also being studied by the 1,000 Genomes Project; http://1000genomes.org) and the H1 human embryonic stem cell line [14]. A secondary priority set (Tier 2) includes HeLa-S3 cervical carcinoma cells [15], HepG2 hepatoblastoma cells [16], and primary (non-transformed) human umbilical vein endothelial cells (HUVEC; [17]), which have limited proliferation potential in culture. To capture a broader spectrum of human biological diversity, a third set (Tier 3) currently comprises more than 100 cell types that are being analyzed in selected assays (Table 2). Standardized growth conditions for all ENCODE cell types have been established and are available through the ENCODE web portal (http://encodeproject.org, "cell types" link).

This report is intended to provide a guide to the data and resources generated by the ENCODE Project to date on Tier 1 3 cell types. We summarize the current state of ENCODE by describing the experimental and computational approaches used to generate and analyze data. In addition, we outline how to access datasets and provide examples of their use.

II. ENCODE Project Data

The following sections describe the different types of data being produced by the ENCODE Project (Table 1).

Genes and Transcripts

Gene annotation. A major goal of ENCODE is to annotate all protein-coding genes, pseudogenes, and non-coding transcribed loci in the human genome and to catalog the products of transcription including splice isoforms. Although the human genome contains ~20,000 protein-coding genes [18], accurate identification of all protein-coding transcripts has not been straightforward. Annotation of pseudogenes and noncoding transcripts also remains a considerable challenge. While automatic gene annotation algorithms have been developed, manual curation remains the approach that delivers the highest level of accuracy, completeness, and stability [19]. The ENCODE Consortium has therefore primarily relied on manual curation with moderate implementation of automated algorithms to produce gene and transcript models that can be verified by traditional experimental and analytical methods. This annotation process involves consolidation of all evidence of transcripts (cDNA, EST sequences) and proteins from public databases, followed by building gene structures based on supporting experimental data [20]. More than 50% of annotated transcripts have no predicted coding potential and are classified by ENCODE into different transcript categories. A classification that summarizes the certainty and types of the annotated structures is provided for each transcript (see http://www.gencodegenes.org/biotypes.html for details). The annotation also includes extensive experimental validation by RT-PCR for novel transcribed loci (i.e., those not previously observed and deposited into public curated databases such as RefSeq). Pseudogenes are identified primarily by a combination of similarity to other protein-coding genes and an obvious functional disablement such as an in-frame stop codon. Because it is difficult to validate pseudogenes experimentally, three independent annotation methods from Yale ("pseudopipe") [21], UCSC ("retrofinder"; http://users.soe.ucsc.edu/~markd/gene-sets-new/pseudoGenes/RetroFinder.html, and references therein), and the Sanger Center [20] are combined to produce a consensus pseudogene set. Ultimately, each gene or transcript model is assigned one of three confidence levels. Level 1 includes genes validated by RT-PCR and sequencing, plus consensus pseudogenes. Level 2 includes manually annotated coding and long non-coding loci that have transcriptional evidence in EMBL/GenBank. Level 3 includes Ensembl gene predictions in regions not yet manually annotated or for which there is new transcriptional evidence.

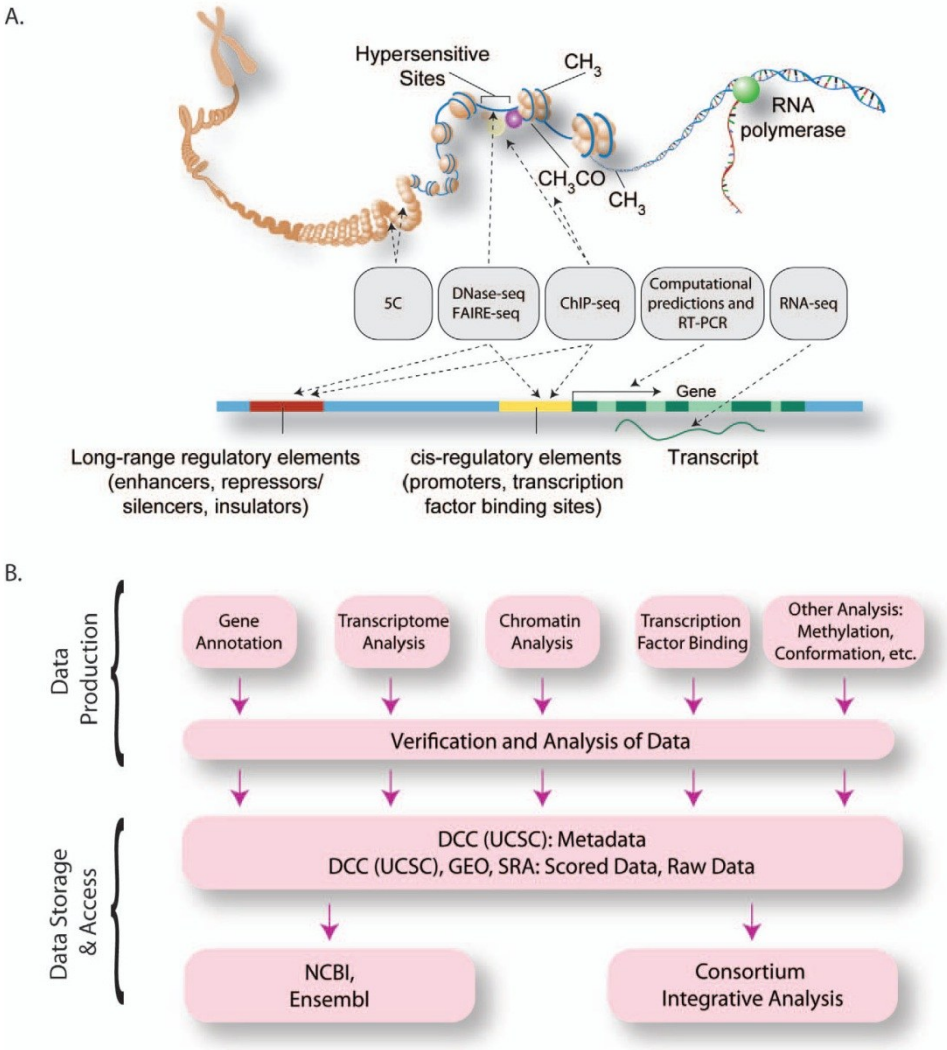


Figure 1. The Organization of the ENCODE Consortium. (A) Schematic representation of the major methods that are being used to detect functional elements (gray boxes), represented on an idealized model of mammalian chromatin and a mammalian gene. (B) The overall data flow from the production groups after reproducibility assessment to the Data Coordinating Center (UCSC) for public access and to other public databases. Data analysis is performed by production groups for quality control and research, as well as at a cross-Consortium level for data integration.
doi:10.1371/journal.pbio.1001046.g001

The result of ENCODE gene annotation (termed "GENCODE") is a comprehensive catalog of transcripts and gene models. ENCODE gene and transcript annotations are updated bimonthly and are available through the UCSC ENCODE browser, distributed annotation servers (DAS; see http://genome.ucsc.edu/cgi-bin/das/hg18/features?segment=21:33031597,33041570;type=wgEncodeGencodeManualV3), and the Ensembl Browser [22].

RNA transcripts. ENCODE aims to produce a comprehensive genome-wide catalog of transcribed loci that characterizes the size, polyadenylation status, and subcellular compartmentalization of all transcripts (Table 1).

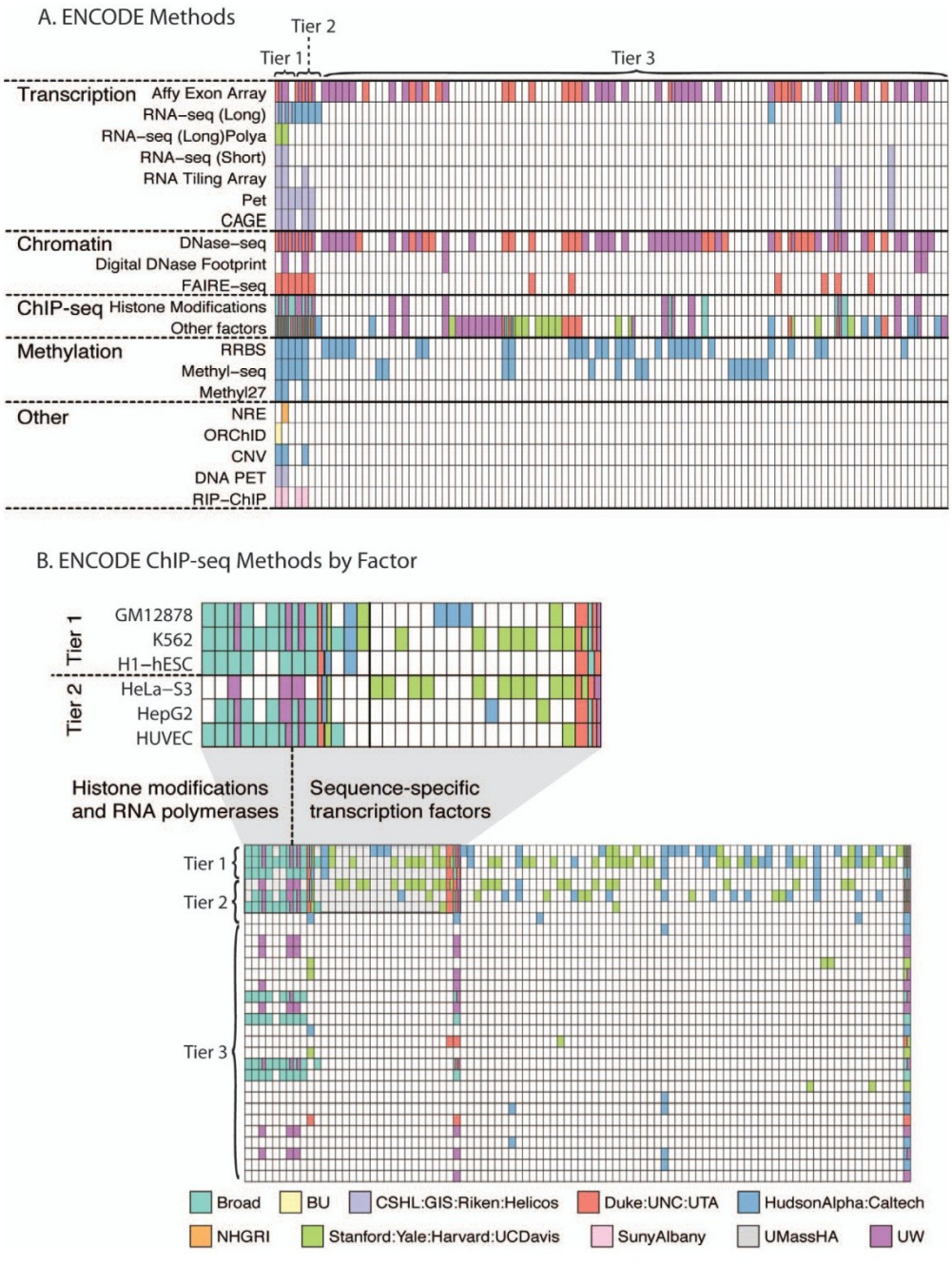

A. ENCODE Methods
Tier 1
Tier 2
Tier 3
Transcription
Affy Exon Array
RNA–seq (Long)
RNA–seq (Long)Polya
RNA–seq (Short)
RNA Tiling Array
Pet
CAGE
Chromatin
DNase–seq
Digital DNase Footprint
FAIRE–seq
ChIP-seq
Histone Modifications
Other factors
Methylation
RRBS
Methyl–seq
Methyl27
Other
NRE
ORChID
CNV
DNA PET
RIP–ChIP
B. ENCODE ChIP-seq Methods by Factor
Tier 1
GM12878
K562
H1–hESC
Tier 2
HeLa–S3
HepG2
HUVEC
Histone modifications and RNA polymerases
Sequence-specific transcription factors
Tier 1
Tier 2
Tier 3
Broad
BU
CSHL:GIS:Riken:Helicos
Duke:UNC:UTA
HudsonAlpha:Caltech
NHGRI
Stanford:Yale:Harvard:UCDavis
SunyAlbany
UMassHA
UW

Figure 2. Data available from the ENCODE Consortium. (A) A data matrix representing all ENCODE data types. Each row is a method and each column is a cell line on which the method could be applied to generate data. Colored cells indicate that data have been generated for that method on that cell line. The different colors represent data generated from different groups in the Consortium as indicated by the key at the bottom of the figure. In some cases, more than one group has generated equivalent data; these cases are indicated by subdivision of the cell to accommodate multiple colors. (B) Data generated by ChIP-seq are split into a second matrix where the cells now represent cell types (rows) split by the factor or histone modification to which the antibody is raised (columns). The colors again represent the groups as indicated by the key. The upper left corner of this matrix has been expanded immediately above the panel to better illustrate the data. All data were collected from the ENCODE public download repository at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC on September 1, 2010.
doi:10.1371/journal.pbio.1001046.g002

ENCODE has generated transcript data with high-density (5 bp) tiling DNA microarrays [23] and massively parallel DNA sequencing methods [9,10,24], with the latter predominating in ongoing efforts. Both polyA+ and polyA− RNAs are being analyzed. Because subcellular compartmentalization of RNAs is important in RNA processing and function, such as nuclear retention of unspliced coding transcripts [25] or snoRNA activity in the nucleolus [26], ENCODE is analyzing not only total whole cell RNAs but also those concentrated in the nucleus and cytosol. Long (>200 nt) and short RNAs (<200 nt) are being sequenced from each subcellular compartment, providing catalogs of potential miRNAs, snoRNA, promoter-associated short RNAs (PASRs) [27], and other short cellular RNAs. Total RNA from K562 and GM12878 cells has been mapped by hybridization to high-density tiling arrays and sequenced to a depth of >500 million paired-end 76 bp reads under conditions where the strand of the RNA transcript is determined, providing considerable depth of transcript coverage (see below).

These analyses reveal that the human genome encodes a diverse array of transcripts. For example, in the proto-oncogene *TP53* locus, RNA-seq data indicate that, while *TP53* transcripts are accurately assigned to the minus strand, those for the oppositely transcribed, adjacent gene *WRAP53* emanate from the plus strand (Figure 3). An independent transcript within the first intron of *TP53* is also observed in both GM12878 and K562 cells (Figure 3).

Additional transcript annotations include exonic regions and splice junctions, transcription start sites (TSSs), transcript 3′ ends, spliced RNA length, locations of polyadenylation sites, and locations with direct evidence of protein expression. TSSs and 3′ ends of transcripts are being determined with two approaches, Paired-End diTag (PET) [28] and Cap-Analysis of Gene Expression (CAGE) [29 31] sequencing.

Table 1. Experimental assays used by the ENCODE Consortium.

Gene/Transcript Analysis		
Region/Feature	**Method**	**Group**
Gene annotation	GENCODE	Wellcome Trust
PolyA+ coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL; Stanford/Yale//Harvard; Caltech
Total RNA coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL
Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)	PET	CSHL
Small RNAs	short RNA-seq	CSHL
Transcription initiation (5′-end) and termination (3-end′) sites	CAGE; diTAGs	RIKEN, GIS
Full-length RNAs	RACE	University of Geneva; University of Lausanne
Protein-bound RNA coding regions	RIP; CLIP	SUNY-Albany; CSHL
Transcription Factors/Chromatin		
Elements/Regions	**Method(s)**	**Group(s)**
Transcription Factor Binding Sites (TFBS)	ChIP-seq	Stanford/Yale/UC-Davis/Harvard; HudsonAlpha/Caltech; Duke/UT-Austin; UW; U. Chicago/Stanford
Chromatin structure (accessibility, etc.)	DNaseI hypersensitivity; FAIRE	UW; Duke; UNC
Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)	ChIP-seq	Broad; UW
DNaseI footprints	Digital genomic footprinting	UW
Other Elements/Features		
Feature	**Method(s)**	**Group(s)**
DNA methylation	RRBS; Illumina Methyl27; Methyl-seq	HudsonAlpha
Chromatin interactions	5C; CHIA-PET	UMass; UW; GIS
Genotyping	Illumina 1M Duo	HudsonAlpha

doi:10.1371/journal.pbio.1001046.t001

Table 2. ENCODE cell types.

Cell Type	Tier	Description	Source
GM12878	1	B-Lymphoblastoid cell line	Coriell GM12878
K562	1	Chronic Myelogenous/Erythroleukemia cell line	ATCC CCL-243
H1-hESC	1	Human Embryonic Stem Cells, line H1	Cellular Dynamics International
HepG2	2	Hepatoblastoma cell line	ATCC HB-8065
HeLa-S3	2	Cervical carcinoma cell line	ATCC CCL-2.2
HUVEC	2	Human Umbilical Vein Endothelial Cells	Lonza CC-2517
Various (Tier 3)	3	Various cell lines, cultured primary cells, and primary tissues	Various

doi:10.1371/journal.pbio.1001046.t002

Transcript annotations throughout the genome are further corroborated by comparing tiling array data with deep sequencing data and by the manual curation described above. Additionally, selected compartment-specific RNA transcripts that cannot be mapped to the current build of the human genome sequence have been evaluated by 5′/3′ Rapid Amplification of cDNA Ends (RACE) [32], followed by RT-PCR cloning and sequencing. To assess putative protein products generated from novel RNA transcripts and isoforms, proteins may be sequenced and quantified by mass spectrometry and mapped back to their encoding transcripts [33,34]. ENCODE has recently begun to study proteins from distinct subcellular compartments of K562 and GM12878 cells by using this complementary approach.

Cis-Regulatory Regions

Cis-regulatory regions include diverse functional elements (e.g., promoters, enhancers, silencers, and insulators) that collectively modulate the magnitude, timing, and cell-specificity of gene expression [35]. The ENCODE Project is using multiple approaches to identify *cis*-regulatory regions, including localizing their characteristic chromatin signatures and identifying sites of occupancy of sequence-specific transcription factors. These approaches are being combined to create a comprehensive map of human *cis*-regulatory regions.

Chromatin structure and modification. Human *cis*-regulatory regions characteristically exhibit nuclease hypersensitivity [36 39] and may show increased solubility after chromatin fixation and fragmentation [40,41]. Additionally, specific patterns of post-translational histone modifications [42,43] have been connected with distinct classes of regions such as promoters and enhancers [3,44 47] as well as regions subject to programmed repression by Polycomb complexes [48,49] or other mechanisms [46,50,51]. Chromatin accessibility and histone modifications thus provide independent and complementary annotations of human regulatory DNA, and massively parallel, high-throughput DNA sequencing methods are being used by ENCODE to map these features on a genome-wide scale (Figure 2 and Table 1).

DNaseI hypersensitive sites (DHSs) are being mapped by two techniques: (i) capture of free DNA ends at in vivo DNaseI cleavage sites with biotinylated adapters, followed by digestion with a TypeIIS restriction enzyme to generate ~20 bp DNaseI

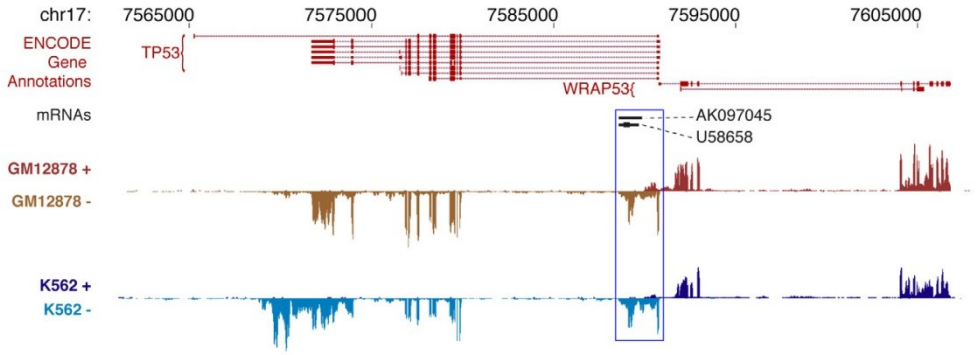


Figure 3. ENCODE gene and transcript annotations. The image shows selected ENCODE and other gene and transcript annotations in the region of the human *TP53* gene (region chr17:7,560,001–7,610,000 from the Human February 2009 (GRCh37/hg19) genome assembly). The annotated isoforms of *TP53* RNAs listed from the ENCODE Gene Annotations (GENCODE) are shown in the top tracks of the figure, along with annotation of the neighboring *WRAP53* gene. In black are two mRNA transcripts (U58658/AK097045) from GenBank. The bottom two tracks show the structure of the *TP53* region transcripts detected in nuclear polyadenylated poly A+ RNAs isolated from GM12878 and K562 cells. The RNA is characterized by RNA-seq and the RNAs detected are displayed according to the strand of origin (i.e. + and −). Signals are scaled and are present at each of the detected p53 exons. Signals are also evident at the U58658 [120] and AK097045 [121] regions located in the first 10 kb intron of the p53 gene (D17S2179E). The U58658/AK097045 transcripts are reported to be induced during differentiation of myeloid leukemia cells but are seen in both GM12878 and K562 cell lines. Finally the p53 isoform observed in K562 cells has a longer 3′UTR region than the isoform seen in the GM12878 cell line.
doi:10.1371/journal.pbio.1001046.g003

cleavage site tags [52,53] and (ii) direct sequencing of DNaseI cleavage sites at the ends of small (<300 bp) DNA fragments released by limiting treatment with DNaseI [54 56]. Chromatin structure is also being profiled with the FAIRE technique [40,57,58], in which chromatin from formaldehyde-crosslinked cells is sonicated in a fashion similar to ChIP and then extracted with phenol, followed by sequencing of soluble DNA fragments. An expanding panel of histone modifications (Figure 2) is being profiled by ChIP-seq [59 62]. In this method, chromatin from crosslinked cells is immunoprecipitated with antibodies to chromatin modifications (or other proteins of interest), the associated DNA is recovered, and the ends are subjected to massively parallel DNA sequencing. Control immunoprecipitations with a control IgG antibody or "input" chromatin sonicated crosslinked chromatin that is not subjected to immune enrichment are also sequenced for each cell type. These provide critical controls, as shearing of crosslinked chromatin may occur preferentially within certain regulatory DNA regions, typically promoters [41]. ENCODE chromatin data types are illustrated for a typical locus in Figure 4, which depicts the patterns of chromatin accessibility, DNaseI hypersensitive sites, and selected histone modifications in GM12878 cells.

For each chromatin data type, the "raw signal" is presented as the density of uniquely aligning sequence reads within 150 bp sliding windows in the human genome. In addition, some data are available as processed signal tracks in which filtering algorithms have been applied to reduce experimental noise. A variety of specialized statistical algorithms are applied to generate discrete high-confidence genomic annotations, including DHSs, broader regions of increased sensitivity to DNaseI, regions of enrichment by FAIRE, and regions with significant levels of specific histone modifications (see Tables 3 and S1). Notably, different histone modifications exhibit characteristic genomic distributions that may be either discrete (e.g., H3K4me3 over a promoter) or broad (e.g., H3K36me3 over an entire transcribed gene body). Because statistical false discovery rate (FDR) thresholds are applied to discrete annotations, the number of regions or elements identified under each assay type depends upon the threshold chosen. Optimal thresholds for an assay are typically determined by comparison to an independent and standard assay method or through reproducibility measurements (see below). Extensive validation of the detection of DNaseI hypersensitive sites is being performed independently with traditional Southern blotting, and more than 6,000 Southern images covering 224 regions in >12 cell types are available through the UCSC browser.

Transcription factor and RNA polymerase occupancy. Much of human gene regulation is determined by the binding of transcriptional regulatory proteins to their cognate sequence elements in *cis*-regulatory regions. ChIP-seq enables genome-scale mapping of transcription factor (TF) occupancy patterns in vivo [59,60,62] and is being extensively applied by ENCODE to create an atlas of regulatory factor binding in diverse cell types. ChIP-seq experiments rely on highly specific antibodies that are extensively characterized by immunoblot analysis and other criteria according

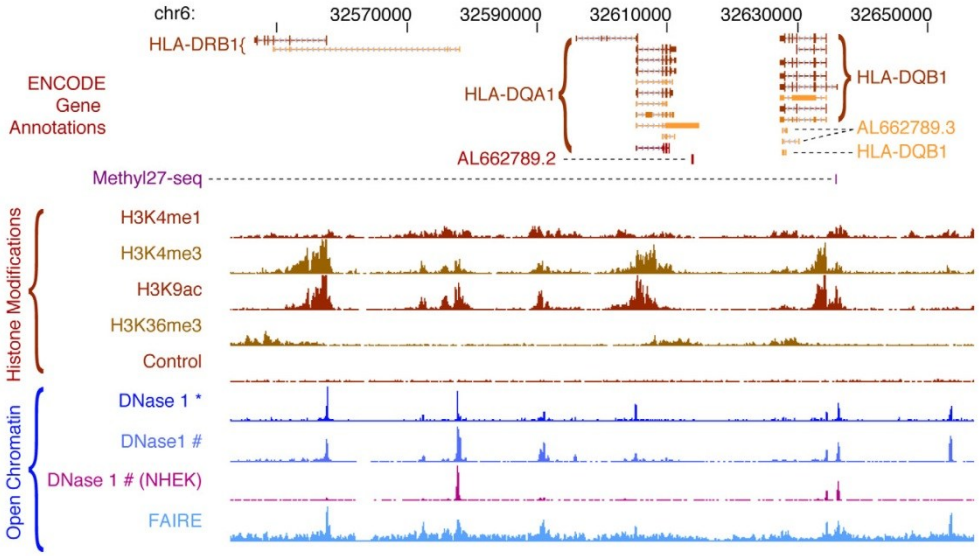


Figure 4. ENCODE chromatin annotations in the *HLA* locus. Chromatin features in a human lymphoblastoid cell line, GM12878, are displayed for a 114 kb region in the *HLA* locus. The top track shows the structures of the annotated isoforms of the *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* genes from the ENCODE Gene Annotations (GENCODE), revealing complex patterns of alternative splicing and several non-protein-coding transcripts overlapping the protein-coding transcripts. The purple mark on the next line shows that a CpG in the promoter of the *HLA-DQB1* gene is partially methylated (assayed on the Illumina Methylation27 BeadArray platform). The densities of four histone modifications associated with transcriptionally active loci are plotted next, along with the input control signal (generated by sequencing an aliquot of the sheared chromatin for which no immunoprecipitation was performed). The last lines plot the accessibility of DNA in chromatin to nucleases (DNaseI) and reduced coverage by nucleosomes (FAIRE); peaks on these lines are DNaseI hypersensitive sites. Note that the ENCODE Consortium generates DNaseI accessibility data by two alternative protocols marked by * and #. The magenta track shows DNaseI sensitivity in a different cell line, NHEK, for comparison.
doi:10.1371/journal.pbio.1001046.g004

Table 3. Analysis tools applied by the ENCODE Consortium.

Class of Software	Description of Task	Examples[a]
Short read alignment	Computationally efficient alignment of short reads to the genome sequence	Bowtie, BWA, Maq, TopHat, GEM, STAR
Peak calling	Converting tag density to defined regions that show statistical properties consistent with binding activity	SPP, PeakSeq, Fseq, MACS, HotSpot
RNA processing	Processing RNA reads into exons and transcripts, with consideration of alternative splicing	Cufflinks, ERANGE, Fluxcapacitor
Integrative peak calling and classification	Jointly considering multiple assay signals to both define the location and character of different genomic regions	ChromHMM, Segway
Statistical tools for specific genomic tasks	Statistical methods developed for replicate-based thresholding, genome-wide-based overlap, and genome-based aggregation	IDR, GSC, ACT
Motif finding tools	Discovering the presence of sequence motifs in enriched peaks	MEME, Weeder
Data analysis frameworks	General frameworks to allow manipulation, comparison, and statistical analysis	R, Bioconductor, MatLab, Galaxy, DART, Genometools
Assign TFBS peaks to genes	Match TFBS to genes they are likely to regulate	GREAT
Compare TF binding and gene expression	Compare binding and expression; compare expressed versus nonexpressed genes	GenPattern, GSEA, Dchip
Conservation	Evaluates conservation of sequences across a range of species	phastCons, GERP, SCONE
Gene Ontology Analysis	Determine types of genes enriched for a given dataset	GO miner, BINGO, AmiGO
Network analysis	Examine relationships between genes	Cytoscape

[a]For full listings and references, see Table S1.
doi:10.1371/journal.pbio.1001046.t003

to ENCODE experimental standards. High-quality antibodies are currently available for only a fraction of human TFs, and identifying suitable immunoreagents has been a major activity of ENCODE TF mapping groups. Alternative technologies, such as epitope tagging of TFs in their native genomic context using recombineering [63,64], are also being explored.

ENCODE has applied ChIP-seq to create occupancy maps for a variety of TFs, RNA polymerase 2 (RNA Pol2) including both unphosphorylated (initiating) and phosphorylated (elongating) forms, and RNA polymerase 3 (RNA Pol3). The localization patterns of five transcription factors and RNA Pol2 in GM12878 lymphoblastoid cells are shown for a typical locus in Figure 5. Sequence reads are processed as described above for DNaseI, FAIRE, and histone modification experiments, including the application of specialized peak-calling algorithms that use input chromatin or control immunoprecipitation data to identify potential false-positives introduced by sonication or sequencing biases (Table 3). Although different peak-callers vary in performance, the strongest peaks are generally identified by multiple algorithms. Most of the sites identified by ChIP-seq are also detected by traditional ChIP-qPCR [65] or are consistent with sites reported in the literature. For example, 98% of 112 sites of CTCF occupancy previously identified by using both ChIP-chip and ChIP-qPCR [66] are also identified in ENCODE CTCF data. Whereas the binding of sequence-specific TFs is typically highly localized resulting in tight sequence tag peaks, signal from antibodies that recognize the phosphorylated (elongating) form of RNA Pol2 may detect occupancy over a wide region encompassing both the site of transcription initiation as well as the domain of elongation. Comparisons among ENCODE groups have revealed that TF and RNA Pol2 occupancy maps generated independently by different groups are highly consistent.

Additional Data Types

ENCODE is also generating additional data types to complement production projects and benchmark novel technologies. An overview of these datasets is provided in Table 1.

DNA methylation. In vertebrate genomes, methylation at position 5 of the cytosine in CpG dinucleotides is a heritable "epigenetic" mark that has been connected with both transcriptional silencing and imprinting [67,68]. ENCODE is applying several complementary approaches to measure DNA methylation. All ENCODE cell types are being assayed using two direct methods for measuring DNA methylation following sodium bisulfite conversion, which enables quantitative analysis of methylcytosines: interrogation of the methylation status of 27,000 CpGs with the Illumina Methyl27 assay [69 72] and Reduced Representation Bisulfite Sequencing (RRBS) [73], which couples *Msp*I restriction enzyme digestion, size selection, bisulfite treatment, and sequencing to interrogate the methylation status of >1,000,000 CpGs largely concentrated within promoter regions and CpG islands. Data from an indirect approach using a methylation-sensitive restriction enzyme (Methyl-seq) [74] are also available for a subset of cell types. These three approaches measure DNA methylation in defined (though overlapping) subsets of the human genome and provide quantitative determinations of the fraction of CpG methylation at each site.

DNaseI footprints. DNaseI footprinting [75] enables visualization of regulatory factor occupancy on DNA in vivo at nucleotide resolution and has been widely applied to delineate the fine structure of *cis*-regulatory regions [76]. Deep sampling of highly enriched libraries of DNaseI-released fragments (see above) enables digital quantification of per nucleotide DNaseI cleavage, which in turn enables resolution of DNaseI footprints on a large scale [55,77,78]. Digital genomic footprinting is being applied on a large scale within ENCODE to identify millions of DNaseI footprints across >12 cell types, many of which localize the specific cognate regulatory motifs for factors profiled by ChIP-seq.

Sequence and structural variation. Genotypic and structural variations within all ENCODE cell types are being interrogated at ~1 million positions distributed approximately every 1.5 kb along the human genome, providing a finely grained map of allelic variation and sequence copy number gains and losses. Genotyping data are generated with the Illumina Infinium

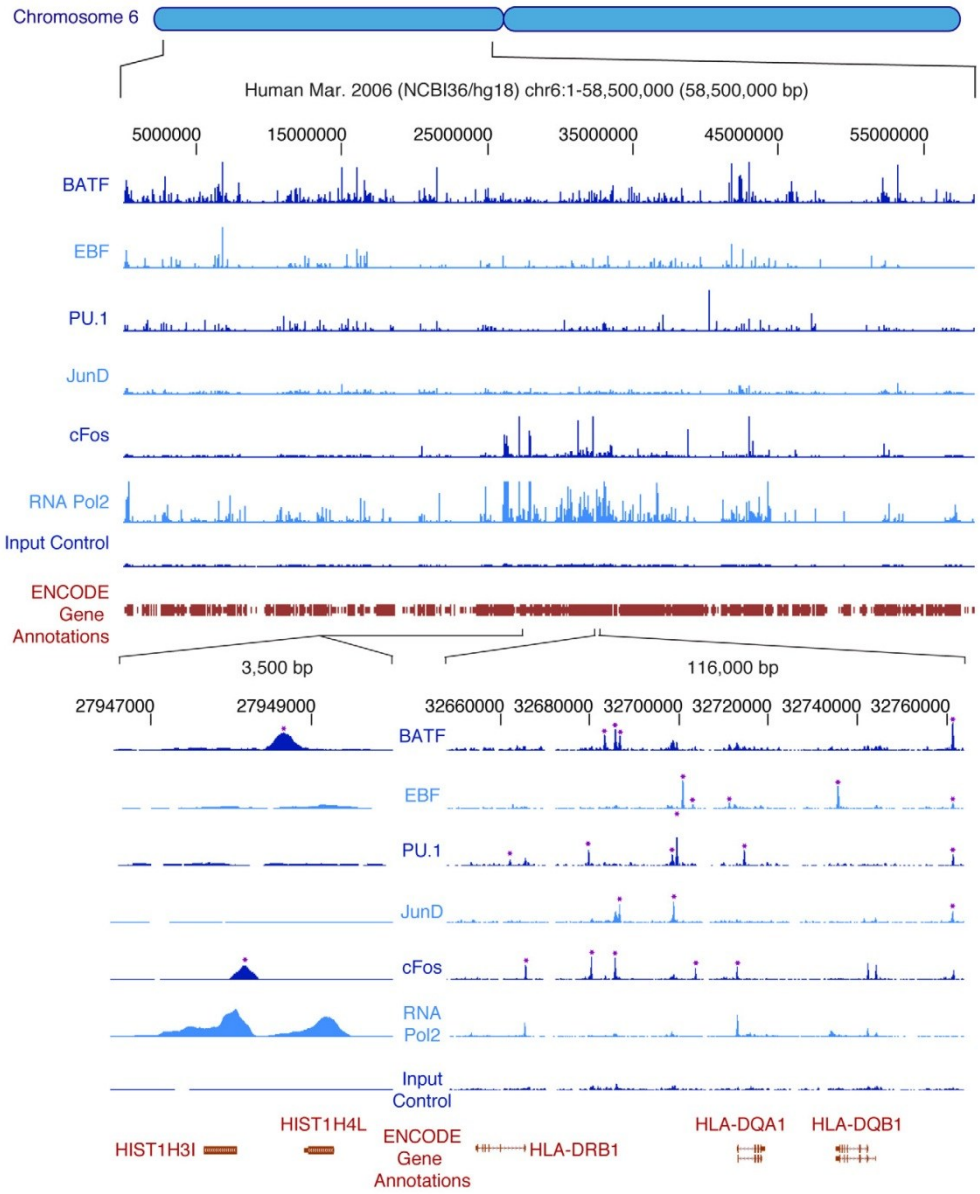


Figure 5. Occupancy of transcription factors and RNA polymerase 2 on human chromosome 6p as determined by ChIP-seq. The upper portion shows the ChIP-seq signal of five sequence-specific transcription factors and RNA Pol2 throughout the 58.5 Mb of the short arm of human chromosome 6 of the human lymphoblastoid cell line GM12878. Input control signal is shown below the RNA Pol2 data. At this level of resolution, the sites of strongest signal appear as vertical spikes in blue next to the name of each experiment ("BATF," "EBF," etc.). More detail can be seen in the bottom right portion, where a 116 kb segment of the HLA region is expanded; here, individual sites of occupancy can be seen mapping to specific regions of the three HLA genes shown at the bottom, with asterisks indicating binding sites called by peak calling software. Finally, the lower left region shows a 3,500 bp region around two tandem histone genes, with RNA Pol2 occupancy at both promoters and two of the five transcription factors, BATF and cFos, occupying sites nearby. Selected annotations from the ENCODE Gene Annotations are shown in each case.
doi:10.1371/journal.pbio.1001046.g005

platform [79], and the results are reported as genotypes and as intensity value ratios for each allele. The genotype and sequence data from GM12878 generated by the 1,000 Genomes Project are being integrated with sequence data from ENCODE chromatin, transcription, TF occupancy, DNA methylation, and other assays to facilitate recognition of functional allelic variation, a significant contributor to phenotypic variability in gene expression [80,81]. The data also permit determination of the sequence copy number gains and losses found in every human genome [82 84], which are particularly prevalent in cell lines of malignant origin.

Long-range Chromatin interactions. Because *cis*-regulatory elements such as enhancers can control genes from distances of tens to hundreds of kb through looping interactions [85], a major challenge presented by ENCODE data is to connect distal regulatory elements with their cognate promoter(s). To map this connectivity, the Consortium is applying the 5C method [86], an enhanced version of Chromosome Conformation Capture (3C) [87], to selected cell lines. 5C has been applied comprehensively to the ENCODE pilot regions as well as to map the interactions between distal DNaseI hypersensitive sites and transcriptional start sites across chromosome 21 and selected domains throughout the genome. Special interfaces have been developed to visualize these 3-dimensional genomic data and are publicly available at http://my5C.umassmed.edu [88].

Protein:RNA interactions. RNA-binding proteins play a major role in regulating gene expression through control of mRNA translation, stability, and/or localization. Occupancy of RNA-binding proteins (RBPs) on RNA can be determined by using immunoprecipitation-based approaches (RIP-chip and RIP-seq) [89 92] analogous to those used for measuring TF occupancy. To generate maps of RBP:RNA associations and binding sites, a combination of RIP-chip and RIP-seq are being used. These approaches are currently targeting 4 6 RBPs in five human cell types (K562, GM12878, H1 ES, HeLa, and HepG2). RBP associations with non-coding RNA and with mRNA are also being explored.

Identification of functional elements with integrative analysis and fine-scale assays of biochemical elements. ChIP-seq of TFs and chromatin modifications may identify genomic regions bound by transcription factors in living cells but do not reveal which segments bound by a given TF are functionally important for transcription. By applying integrative approaches that incorporate histone modifications typical of enhancers (e.g., histone H3, Lysine 4 monomethylation), promoters (e.g., histone H3, Lysine 4 trimethylation), and silencers (e.g., Histone H3, Lysine 27, and Lysine 9 trimethylation), ENCODE is categorizing putative functional elements and testing a subset for activities in the context of transient transfection/reporter gene assays [93 97]. To further pinpoint the biological activities associated with specific regions of TF binding and chromatin modification within promoters, hundreds of TF binding sites have been mutagenized, and the mutant promoters are being assayed for effects on reporter gene transcription by transient transfection assays. This approach is enabling identification of specific TF binding sites that lead to activation and others associated with transcriptional repression.

Proteomics. To assess putative protein products generated from novel RNA transcripts and isoforms, proteins are sequenced and quantified by mass spectrometry and mapped back to their encoding transcripts [33,34,98]. ENCODE has recently begun to study proteins from distinct subcellular compartments of K562 and GM12878 with this complementary approach.

Evolutionary conservation. Evolutionary conservation is an important indicator of biological function. ENCODE is approaching evolutionary analysis from two directions. Functional properties are being assigned to conserved sequence elements identified through multi-species alignments, and conversely, the evolutionary histories of biochemically defined elements are being deduced. Multiple alignments of the genomes of 33 mammalian species have been constructed by using the Enredo, Pecan, Ortheus approach (EPO) [99,100], and complementary multiple alignments are available through the UCSC browser (UCSC Lastz/ChainNet/Multiz). These alignments enable measurement of evolutionary constraint at single-nucleotide resolution using GERP [101], SCONE [102], PhyloP [103], and other algorithms. In addition, conservation of DNA secondary structure based on hydroxyl radical cleavage patterns is being analyzed with the Chai algorithm [7].

Data Production Standards and Assessment of Data Quality

With the aim of ensuring quality and consistency, ENCODE has defined standards for collecting and processing each data type. These standards encompass all major experimental components, including cell growth conditions, antibody characterization, requirements for controls and biological replicates, and assessment of reproducibility. Standard formats for data submission are used that capture all relevant data parameters and experimental conditions, and these are available at the public ENCODE portal (http://genome.ucsc.edu/ENCODE/dataStandards.html). All ENCODE data are reviewed by a dedicated quality assurance team at the Data Coordination Center before release to the public. Experiments are considered to be *verified* when two highly concordant biological replicates have been obtained with the same experimental technique. In addition, a key quality goal of ENCODE is to provide *validation* at multiple levels, which can be further buttressed by cross-correlation between disparate data types. For example, we routinely perform parallel analysis of the same biological samples with alternate detection technologies (for example, ChIP-seq versus ChIP-chip or ChIP-qPCR). We have also compared our genome-wide results to "gold-standard" data from individual locus studies, such as DNase-seq versus independently performed conventional (Southern-based) DNaseI hypersensitivity studies. Cross-correlation of independent but related ENCODE data types with one another, such as DNaseI hypersensitivity, FAIRE, transcription factor occupancy, and histone modification patterns, can provide added confidence in the identification of specific DNA elements. Similarly, cross-correlation between long RNA-seq, CAGE, and TAF1 ChIP-seq data can strengthen confidence in a candidate location for transcription initiation. Finally, ENCODE is performing pilot tests for the biological activity of DNA elements to the predictive potential of various ENCODE biochemical signatures for certain biological functions. Examples include transfection assays in cultured human cells and injection assays in fish embryos to test for enhancer, silencer, or insulator activities in DNA elements identified by binding of specific groups of TFs or the presence of DNaseI hypersensitive sites or certain chromatin marks. Ultimately, defining the full biological role of a DNA element in its native chromosomal location and organismic context is the greatest challenge. ENCODE is beginning to approach this by integrating its data with results from other studies of in situ knockouts and/or knockdowns, or the identification of specific naturally occurring single base mutations and small deletions associated with changes in gene expression. However, we expect that deep insights into the function of most elements will ultimately come from the community of biologists who will build on ENCODE data or use them to complement their own experiments.

Current Scope and Completeness of ENCODE Data

A catalog of ENCODE datasets is available at http://encodeproject.org. These data provide evidence that ~1 Gigabase (Gb; 32%) of the human genome sequence is represented in steady-state, predominantly processed RNA populations. We have also delineated more than 2 million potential regulatory DNA regions through chromatin and TF mapping studies.

The assessment of the completeness of detection of any given element is challenging. To analyze the detection of transcripts in a single experiment, we have sequenced to substantial depth and used a sampling approach to estimate the number of reads needed to approach complete sampling of the RNA population (Figure 6A) [104]. For example, analyzing RNA transcripts with about 80 million mapped reads yields robust quantification of more than 80% of the lowest abundance class of genes (2 19 reads per kilobase per million mapped tags, RPKM) [24]. Measuring RNAs across multiple cell types, we find that, after the analysis of seven cell lines, 68% of the GENCODE transcripts can be detected with RPKM >1.

In the case of regulatory DNA, we have analyzed the detection of regulatory DNA by using three approaches: 1) the saturation of occupancy site discovery for a single transcription factor within a single cell type as a function of sequencing read depth, 2) the incremental discovery of DNaseI hypersensitive sites or the occupancy sites for a single TF across multiple cell types, and 3) the incremental rate of collective TF occupancy site discovery for all TFs across multiple cell types.

For detecting TF binding sites by ChIP-seq, we have found that the number of significant binding sites increases as a function of sequencing depth and that this number varies widely by transcription factor. For example, as shown in Figure 6B, 90% of detectable sites for the transcription factor GABP can be identified by using the MACS peak calling program at a depth of 24 million reads, whereas only 55% of detectable RNA Pol2 sites are identified at this depth when an antibody that recognizes both initiating and elongating forms of the enzyme is used. Even at 50 million reads, the number of sites is not saturated for RNA Pol2 with this antibody. It is important to note that determinations of saturation may vary with the use of different antibodies and laboratory protocols. For instance, a different RNA Pol2 antibody that recognizes unphosphorylated, non-elongating RNA Pol2 bound only at promoters requires fewer reads to reach saturation [105]. For practical purposes, ENCODE currently uses a minimum sequencing depth of 20 M uniquely mapped reads for sequence-specific transcription factors. For data generated prior to June 1, 2010, this figure was 12 M.

To assess the incremental discovery of regulatory DNA across different cell types, it was necessary to account for the non-uniform correlation between cell lines and assays (see Figure 6C legend for details). We therefore examined all possible orderings of either cell types or assays and calculated the distribution of elements discovered as the number of cell types or assays increases, presented as saturation distribution plots (Figure 6C and 6D, respectively). For DNase hypersensitive sites, we observe a steady increase in the mean number of sites discovered as additional cell types are tested up to and including the 62 different cell types examined to date, indicating that new elements continue to be identified at a relatively high rate as additional cell types are sampled (Figure 6C). Analysis of CTCF sites across 28 cell types using this approach shows similar behavior. Analysis of binding sites for 42 TFs in the cell line with most data (K562) also shows that saturation of the binding sites for these factors has not yet been achieved. These results indicate that additional cell lines need to be analyzed for DNaseI and many transcription factors, and that many more transcription factors need to be analyzed within single cell types to capture all the regulatory information for a given factor across the genome. The implications of these trends for defining the extent of regulatory DNA within the human genome sequence is as yet unclear.

III. Accessing ENCODE Data

ENCODE Data Release and Use Policy

The ENCODE Data Release and Use Policy is described at http://www.encodeproject.org/ENCODE/terms.html. Briefly, ENCODE data are released for viewing in a publicly accessible browser (initially at http://genome-preview.ucsc.edu/ENCODE and, after additional quality checks, at http://encodeproject.org). The data are available for download and pre-publication analysis of any kind, as soon as they are verified (i.e., shown to be reproducible). However, consistent with the principles stated in the Toronto Genomic Data Use Agreement [106], the ENCODE Consortium data producers request that they have the first publication on genome-wide analyses of ENCODE data, within a 9-month timeline from its submission. The timeline for each dataset is clearly displayed in the information section for each dataset. This parallels policies of other large consortia, such as the HapMap Project (http://www.hapmap.org), that attempt to balance the goal of rapid data release with the ability of data producers to publish initial analyses of their work. Once a producer has published a dataset during this 9-month period, anyone may publish freely on the data. The embargo applies only to global analysis, and the ENCODE Consortium expects and encourages immediate use and publication of information at one or a few loci, without any consultation or permission. For such uses, identifying ENCODE as the source of the data by citing this article is requested.

Public Repositories of ENCODE Data

After curation and review at the Data Coordination Center, all processed ENCODE data are publicly released to the UCSC Genome Browser database (http://genome.ucsc.edu). Accessioning of ENCODE data at the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html) is underway. Primary DNA sequence reads are stored at UCSC and the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?) and will also be retrievable via GEO. Primary data derived from DNA microarrays (for example, for gene expression) are deposited directly to GEO. The processed data are also formatted for viewing in the UCSC browser. Metadata, including information on antibodies, cell culture conditions, and other experimental parameters, are deposited into the UCSC database, as are results of validation experiments. Easy retrieval of ENCODE data to a user's desktop is facilitated by the UCSC Table Browser tool (http://genome.ucsc.edu/cgi-bin/hgTables?org=human), which does not require programming skills. Computationally sophisticated users may gain direct access to data through application programming interfaces (APIs) at both the UCSC browser and NCBI and by downloading files from http://genome.ucsc.edu/ENCODE/downloads.html.

An overview of ENCODE data types and the location of the data repository for each type is presented in Table 4.

IV. Working with ENCODE Data

Using ENCODE Data in the UCSC Browser

Many users will want to view and interpret the ENCODE data for particular genes of interest. At the online ENCODE portal

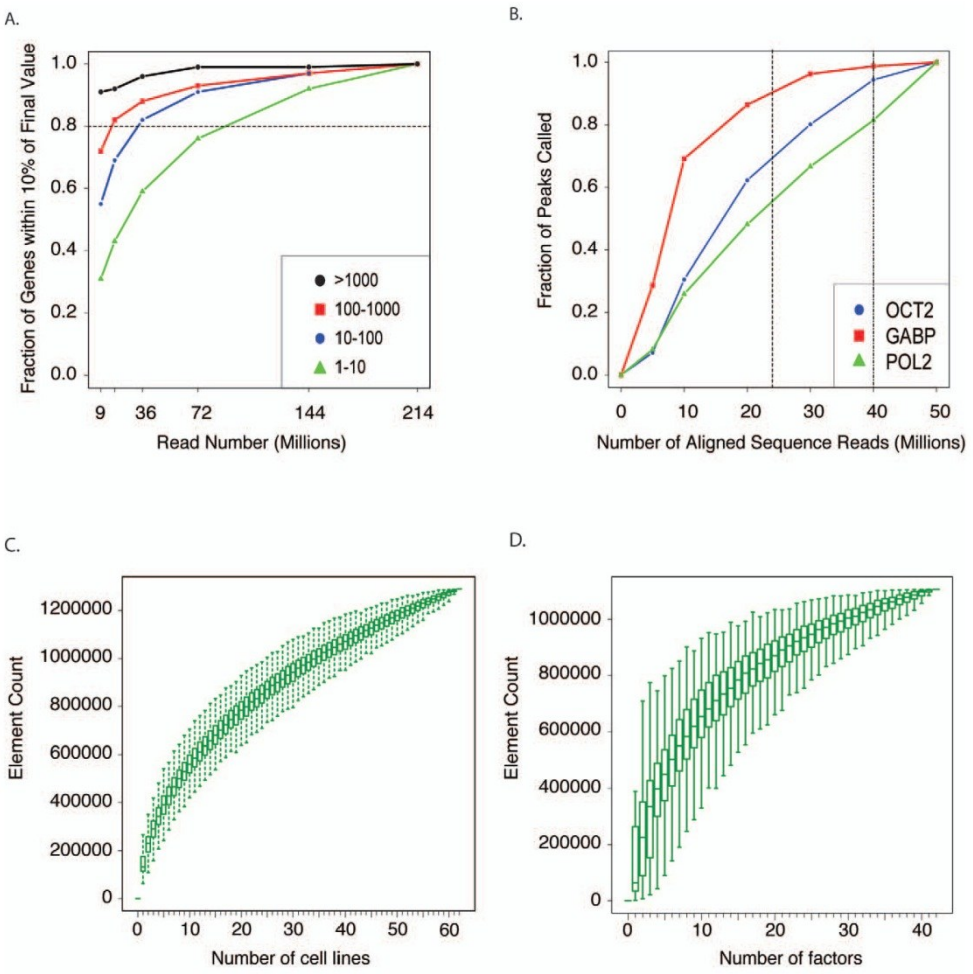


Figure 6. Incremental discovery of transcribed elements and regulatory DNA. (A) Robustness of gene expression quantification relative to sequencing depth. PolyA-selected RNA from H1 human embryonic stem cells was sequenced to 214 million mapped reads. The number of reads (indicated on the *x*-axis) was sampled from the total, and gene expression (in FPKM) was calculated and compared to the gene expression values resulting from all the reads (final values). Gene expression levels were split into four abundance classes and the fraction of genes in each class with RPKM values within 10% of the final values was calculated. At ~80 million mapped reads, more than 80% of the low abundance class of genes is robustly quantified according to this measure (horizontal dotted line). Abundances for the classes in RPKM are given in the inset box. (B) Effect of number of reads on fractions of peaks called in ChIP-seq. ChIP-seq experiments for three sequence-specific transcription factors were sequenced to a depth of 50 million aligned reads. To evaluate the effect of read depth on the number of binding sites identified, peaks were called with the MACS algorithm at various read depths, and the fraction of the total number of peaks that were identified at each read depth are shown. For sequence-specific transcription factors that have strong signal with ChIP-seq, such as GABP, approximately 24 million reads (dashed vertical line) are sufficient to capture 90% of the binding sites. However, for more general sequence-specific factors (e.g., OCT2), additional sequencing continues to yield additional binding site information. RNA Pol2, which interacts with DNA broadly across genes, maintains a nearly linear gain in binding information through 50 million aligned reads. (C) Saturation analysis of ENCODE DNaseI hypersensitivity data with increasing numbers of cell lines. The plot shows the extent of saturation of DNaseI hypersensitivity sites (DHSs) discovered as increasing numbers of cell lines are studied. The plot is generated from the ENCODE DNaseI elements defined at the end of January 2010 (from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC) as follows. We first define a set of DHSs from the overlap of all DHS data across all cell lines. Where overlapping elements are identified in two or more cell lines, these are determined to represent the same element and fused up to a maximum size of 5 kb. Elements above this limit are split and counted as distinct. We then calculate the subset of these elements represented by each single cell line experiment. The distribution of element counts for each single cell line is plotted as a box plot with the median at position 1 on the *x*-axis. We next calculate the element contributions of all possible pairs of cell line experiments and plot this distribution at position 2. We continue to do this for all incremental steps up to and including all cell lines (which is

by definition only a single data point). (D) Saturation of TF ChIP-seq elements in K562 cells. This plot illustrates the saturation of elements identified by TF ChIP-seq as additional factors are analyzed within the same cell line. The plot is generated by the equivalent approach as described in (C), except the data are now the set of all elements defined by ChIP-seq analysis of K562 cells with 42 different transcription factors. The data were from the January 2010 data freeze from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC. For consistency, the peak calls from all ChIP-seq data were generated by a uniform processing pipeline with the Peakseq peak caller and IDR replicate reconciliation.
doi:10.1371/journal.pbio.1001046.g006

(http://encodeproject.org), users should follow a "Genome Browser" link to visualize the data in the context of other genome annotations. Currently, it is useful for users to examine both the hg18 and the hg19 genome browsers. The hg18 has the ENCODE Integrated Regulation Track on by default, which shows a huge amount of data in a small amount of space. The hg19 browser has newer datasets, and more ENCODE data than are available on hg18. Work is in progress to remap the older hg18 datasets to hg19 and generate integrated ENCODE tracks. On either browser, additional ENCODE tracks are marked by a double helix logo in the browser track groups for genes, transcripts, and regulatory features. Users can turn tracks on or off to develop the views most useful to them (Figure 7). To aid users in navigating the rich variety of data tracks, the ENCODE portal also provides a detailed online tutorial that covers data display, data download, and analysis functions available through the browser. Examples applying ENCODE data at individual loci to specific biological or medical issues are a good starting point for exploration and use of the data. Thus, we also provide a collection of examples at the "session gallery" at the ENCODE portal. Users are encouraged to submit additional examples; we anticipate that this community-based sharing of insights will accelerate the use and impact of the ENCODE data.

An Illustrative Example

Numerous genome-wide association studies (GWAS) that link human genome sequence variants with the risk of disease or with common quantitative phenotypes have now become available. However, in most cases, the molecular consequences of disease- or trait-associated variants for human physiology are not understood [107]. In more than 400 studies compiled in the GWAS catalog [108], only a small minority of the trait/disease-associated SNPs (TASs) occur in protein-coding regions; the large majority (89%) are in noncoding regions. We therefore expect that the accumulating functional annotation of the genome by ENCODE will contribute substantially to functional interpretation of these TASs.

For example, common variants within a ~1 Mb region upstream of the c-Myc proto-oncogene at 8q24 have been associated with cancers of the colon, prostate, and breast (Figure 8A) [109 111]. ENCODE data on transcripts, histone modifications, DNase hypersensitive sites, and TF occupancy show strong, localized signals in the vicinity of major cancer-associated SNPs. One variant (*rs698327*) lies within a DNase hypersensitive site that is bound by several TFs and the enhancer-associated protein p300 and contains histone modification patterns typical of enhancers (high H3K4me1, low H3K4me3; Figure 8B). Recent studies have shown enhancer activity and allele-specific binding of TCF7L2 at this site [112], with the risk allele showing greater binding and activity [113,114]. Moreover, this element appears to contact the downstream c-Myc gene in vivo, compatible with enhancer function [114,115]. Similarly, several regions predicted via ENCODE data to be involved in gene regulation are close to SNPs in the *BCL11A* gene associated with persistent expression of fetal hemoglobin (Figure S2). These examples show that the simple overlay of ENCODE data with candidate non-coding risk-associated variants may readily identify specific genomic elements as leading candidates for investigation as probable effectors of phenotypic effects via alterations in gene expression or other genomic regulatory processes. Importantly, even data from cell types not directly associated with the phenotype of interest may be of considerable value for hypothesis generation. It is reasonable to expect that application of current and future ENCODE data will provide useful information concerning the mechanism(s) whereby genomic variation influences susceptibility to disease, which then can then be tested experimentally.

Limitations of ENCODE Annotations

All ENCODE datasets to date are from populations of cells. Therefore, the resulting data integrate over the entire cell population, which may be physiologically and genetically inhomogeneous. Thus, the source cell cultures in the ENCODE experiments are not typically synchronized with respect to the cell cycle and, as with all such samples, local micro-environments in culture may also vary, leading to physiological differences in cell state within each culture. In addition, one Tier 1 cell line (K562) and two Tier 2 cell lines (HepG2 and HeLa) are known to have abnormal genomes and karyotypes, with genome instability. Finally, some future Tier 3 tissue samples or primary cultures may be inherently heterogeneous in cell type composition. Averaging over heterogeneity in physiology and/or genotype produces an amalgamation of the contributing patterns of gene

Table 4. Overview of ENCODE data types.

Data	Description	Location
Metadata	Experimental parameters (e.g., growth conditions, antibody characterization)	UCSC, GEO
Primary data images	CCD camera images from sequencers or microarrays	Not archived
Sequence reads/microarray signal	Minimally processed experimental data; reads and quality information; probe locations and intensities	UCSC, GEO, SRA
Aligned sequence reads	Sequence reads and genomic positions	UCSC, GEO
Genomic signal	Sequence tag density (sliding window); cumulative base coverage or density by sequencing or read pseudo-extension; microarray probe intensity	UCSC, GEO
Enriched region calls/scores/*p* or *q* values	Putative binding or transcribed regions	UCSC, GEO

doi:10.1371/journal.pbio.1001046.t004

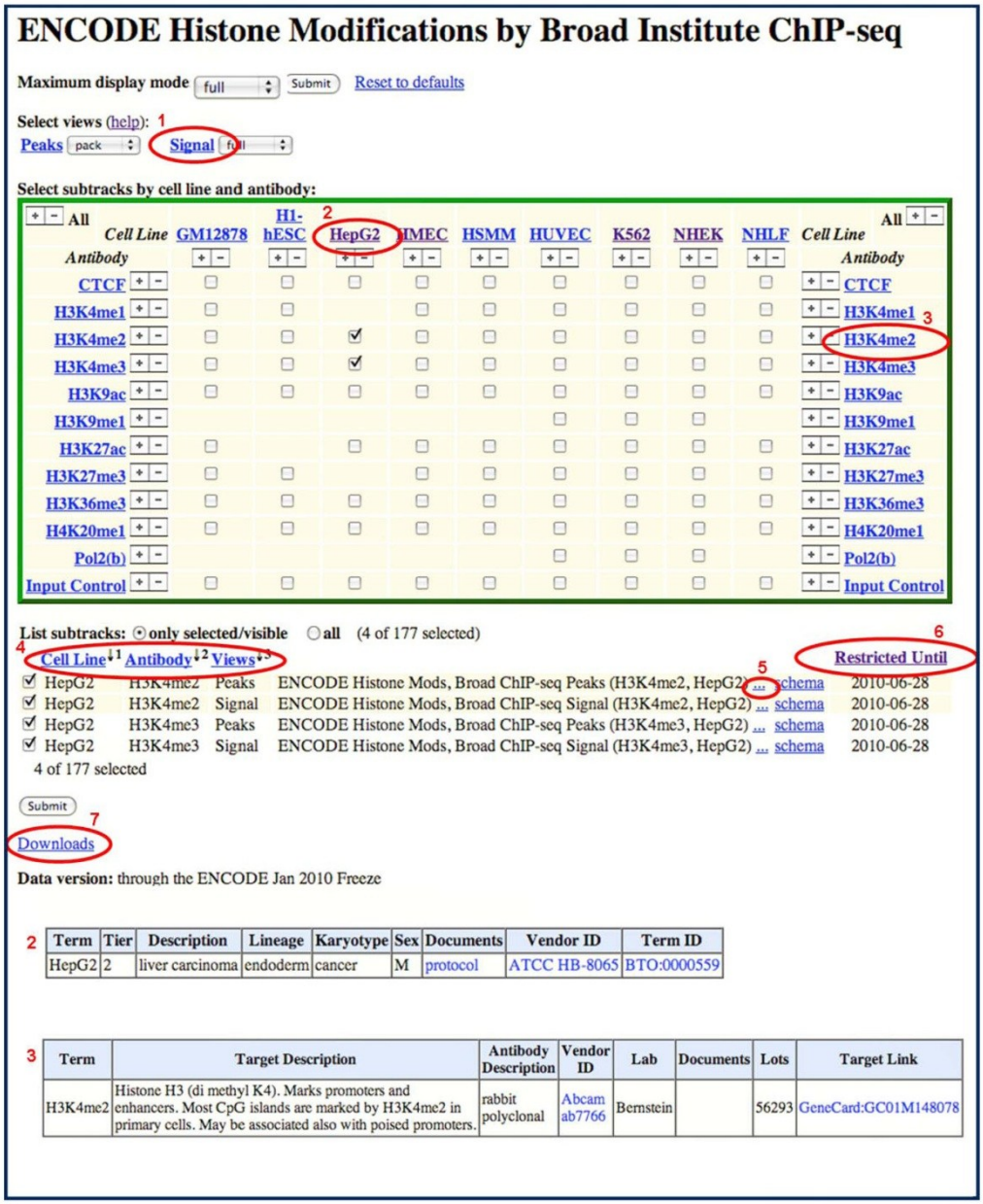


Figure 7. Accessing ENCODE data at the UCSC Portal. Data and results for the ENCODE Project are accessible at the UCSC portal (http://genome.ucsc.edu/ENCODE). "Signal tracks" for the different datasets are selected and displayed in the genome browser to generate images such as those shown in Figures 3–4. The datasets are available from the Track Settings page; an example is shown that illustrates some of the key controls. A dataset is selected and the Signal display plots the values of an assay for a given feature more or less continuously along a chromosome. The height, range for the *y*-axis, windowing function, and many other aspects of the graph are controlled in the Signal Configuration window, accessed by clicking on "Signal" (red oval #1). ENCODE data are commonly generated on multiple cell lines; information about each can be accessed by clicking

on the name of the cell line or antibody (e.g., HepG2, red oval #2). Many ENCODE tracks are actually composites of multiple subtracks; these can be turned on and off by using the boxes in the central matrix or in the subtrack list below. Subtracks can be reordered individually by using drag and drop in the browser image or the Track Settings page, or in logical groups by using the "Cell/Antibody/Views" (red oval #4) ordering controls. Additional information about the feature and the assay, such as the antibody used, can be obtained by clicking on the name of the feature. Some restrictions to the use of ENCODE data apply for a 9-month period after deposit of the data; the end of that 9-month period is given by the "Restricted Until" date. Full data can be downloaded by clicking on the "Downloads" link (red oval #7).
doi:10.1371/journal.pbio.1001046.g007

expression, factor occupancy, and chromatin status that must be considered when using the data. Future improvements in genome-wide methodology that allow the use of much smaller amounts of primary samples, or follow-up experiments in single cells when possible, may allow us to overcome many of these caveats.

The use of DNA sequencing to annotate functional genomic features is constrained by the ability to place short sequence reads accurately within the human genome sequence. Most ENCODE data types currently represented in the UCSC browser use only those sequence reads that map uniquely to the genome. Thus, centromeric and telomeric segments (collectively ~15% of the genome and enriched in recent transposon insertions and segmental duplications) as well as sequences not present in the current genome sequence build [116] are not subject to reliable annotation by our current techniques. However, such information can be gleaned through mining of the publicly available raw sequence read datasets generated by ENCODE.

It is useful to recognize that the confidence with which different classes of ENCODE elements can be related to a candidate function varies. For example, ENCODE can identify with high confidence new internal exons of protein-coding genes, based on RNA-seq data for long polyA+ RNA. Other features, such as candidate promoters, can be identified with less, yet still good, confidence by combining data from RNA-seq, CAGE-tags, and RNA polymerase 2 (RNA Pol2) and TAF1 occupancy. Still other ENCODE biochemical signatures come with much lower confidence about function, such as a candidate transcriptional enhancer supported by ChIP-seq evidence for binding of a single transcription factor.

Identification of genomic regions enriched by ENCODE biochemical assays relies on the application of statistical analyses and the selection of threshold significance levels, which may vary between the algorithms used for particular data types. Accordingly, discrete annotations, such as TF occupancy or DNaseI hypersensitive sites, should be considered in the context of reported p values, q values, or false discovery rates, which are conservative in many cases. For data types that lack focal enrichment, such as certain histone modifications and many RNA Pol2-bound regions, broad segments of significant enrichment have been delineated that encompass considerable quantitative variation in the signal strength along the genome.

V. ENCODE Data Analysis

Development and implementation of algorithms and pipelines for processing and analyzing data has been a major activity of the ENCODE Project. Because massively parallel DNA sequencing has been the main type of data generated by the Consortium, much of the algorithmic development and data analysis to date has been concerned with issues related to producing and interpreting such data. Software packages and algorithms commonly used in the ENCODE Consortium are summarized in Tables 3 and S1.

In general, the analysis of sequencing-based measurements of functional or biochemical genomic parameters proceeds through three major phases. In the first phase, the short sequences that are the output of the experimental method are aligned to the reference genome. Algorithm development for efficient and accurate alignment of short read sequences to the human genome is a rapidly developing field, and ENCODE groups employ a variety of the state-of-the-art software (see Tables 3 and S1). In the second phase, the initial sequence mapping is processed to identify significantly enriched regions from the read density. For ChIP-seq (TFs and histone modification), DNase-seq or FAIRE-seq, both highly localized peaks or broader enriched regions may be identified. Within the ENCODE Consortium, each data production group provides lists of enriched regions or elements within their own data, which are available through the ENCODE portal. It should be noted that, for most data types, the majority of enriched regions show relatively weak absolute signal, necessitating the application of conservative statistical thresholds. For some data, such as those derived from sampling RNA species (e.g., RNA-seq), additional algorithms and processing are used to handle transcript structures and the recognition of splicing events.

The final stage of analysis involves integrating the identified regions of enriched signal with each other and with other data types. An important prerequisite to data integration is the availability of uniformly processed datasets. Therefore, in addition to the processing pipelines developed by individual production groups, ENCODE has devoted considerable effort toward establishing robust uniform processing for phases 1 and 2 to enable integration. For signal comparison, specific consideration has been given to deriving a normalized view of the sequence read density of each experiment. In the case of ChIP-seq for TFs, this process includes in silico extension of the sequence alignment to reflect the experimentally determined average lengths of the input DNA molecules that are sampled by the short sequence tag, compensation for repetitive sequences that may lead to alignment with multiple genomic locations, and consideration of the read density of the relevant control or input chromatin experiment. ENCODE has adopted a uniform standardized peak-calling approach for transcription factor ChIP-seq, including a robust and conservative replicate reconciliation statistic (Irreproducible Discovery Rate, IDR [117], to yield comparable consensus peak calls. As the project continues, we expect further standardizations to be developed.

There are many different ways to analyze and integrate large, diverse datasets. Some of the basic approaches include assigning features to existing annotations (e.g., assigning transcribed regions to annotated genes or Pol2-binding peaks to likely genes), discovery of correlations among features, and identification of particular gene classes (e.g., Gene Ontology categories) preferentially highlighted by a given annotation. Many software tools exist in the community for these purposes, including some developed within the ENCODE Project, such as the Genome Structure Correction statistic for assessing overlap significance [3]. Software tools used for integration by ENCODE are summarized in Tables 3 and S1.

VI. Future Plans and Challenges

Data Production Plans

The challenge of achieving complete coverage of all functional elements in the human genome is substantial. The adult human body contains several hundred distinct cell types, each of which

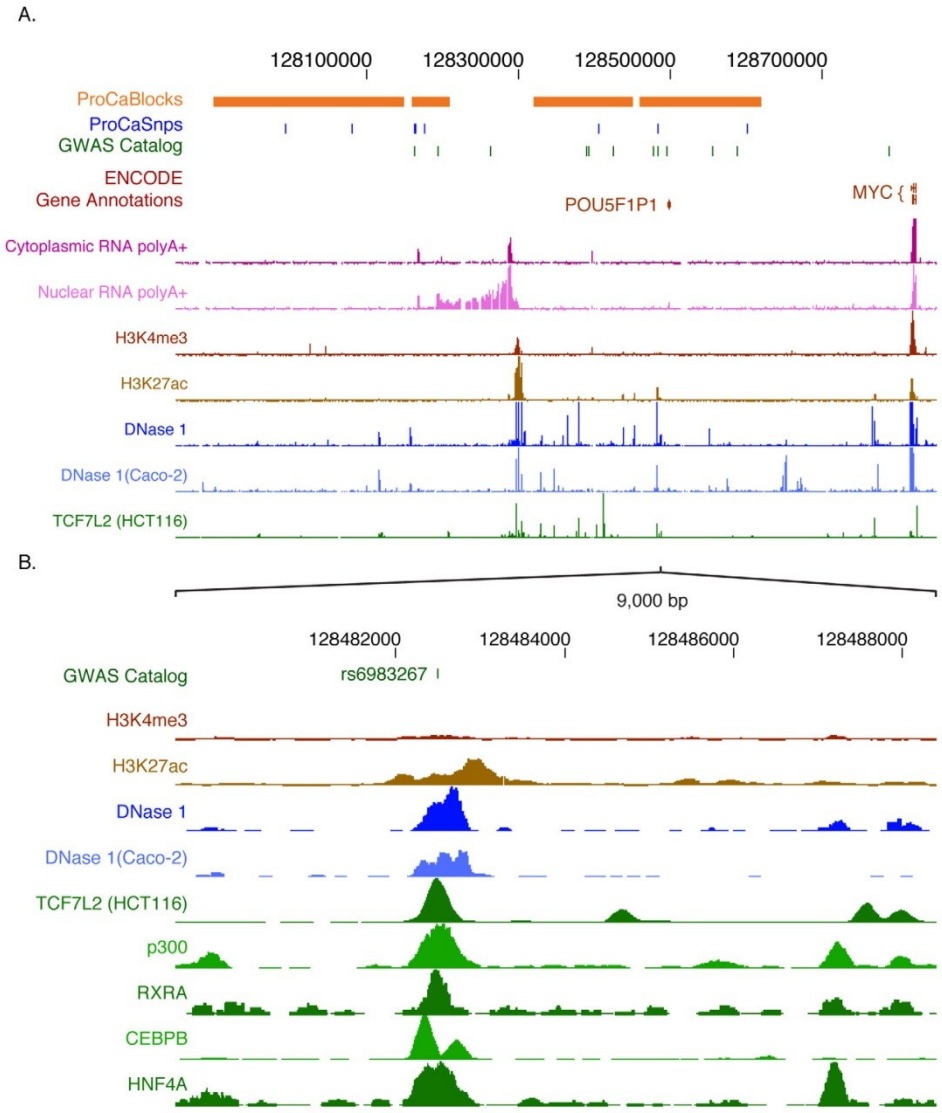


Figure 8. ENCODE data indicate non-coding regions in the human chromosome 8q24 loci associated with cancer. (A) A 1 Mb region including *MYC* and a gene desert upstream shows the linkage disequilibrium blocks and positions of SNPs associated with breast and prostate cancer, with both a custom track based on [121] and the resident track from the GWAS catalog. ENCODE tracks include GENCODE gene annotations, results of mapping RNAs to high-density Affymetrix tiling arrays (cytoplasmic and nuclear polyA+ RNA), mapping of histone modifications (H3K4me3 and H3K27Ac), DNaseI hypersensitive sites in liver and colon carcinoma cell lines (HepG2 and Caco-2), and occupancy by the transcription factor TCF7L2 in HCT116 cells. (B) Expanded view of a 9 kb region containing the cancer-associated SNP *rs6983267* (shown on the top line). In addition to the histone modifications, DNaseI hypersensitive sites and factor occupancy described in (A), the ENCODE tracks also show occupancy by the coactivator p300 and the transcription factors RXRA, CEBPB, and HNF4A. Except as otherwise noted in brackets, the ENCODE data shown here are from the liver carcinoma cell line HepG2.
doi:10.1371/journal.pbio.1001046.g008

expresses a unique subset of the ~1,500 TFs encoded in the human genome [118]. Furthermore, the brain alone contains thousands of types of neurons that are likely to express not only different sets of TFs but also a larger variety of non-coding RNAs [119]. In addition, each cell type may exhibit a diverse array of responses to exogenous stimuli such as environmental conditions or chemical agents. Broad areas of fundamental chromosome function, such as meiosis and recombination, remain unexplored. Furthermore, ENCODE has focused chiefly on definitive cells and cell lines, bypassing the substantial complexity of development and differentiation. A truly comprehensive atlas of human functional elements is not practical with current technologies, motivating our focus on performing the available assays in a range of cell types that will provide substantial near-term utility. ENCODE is currently developing a strategy for addressing this cellular space in a timely manner that maximizes the value to the scientific community. Feedback from the user community will be a critical component of this process.

Integrating ENCODE with Other Projects and the Scientific Community

To understand better and functionally annotate the human genome, ENCODE is making efforts to analyze and integrate data within the project and with other large-scale projects. These efforts include 1) defining promoter and enhancer regions by combining transcript mapping and biochemical marks, 2) delineating distinct classes of regions within the genomic landscape by their specific combinations of biochemical and functional characteristics, and 3) defining transcription factor co-associations and regulatory networks. These efforts aim to extend our understanding of the functions of the different biochemical elements in gene regulation and gene expression.

One of the major motivations for the ENCODE Project has been to aid in the interpretation of human genome variation that is associated with disease or quantitative phenotypes. The Consortium is therefore working to combine ENCODE data with those from other large-scale studies, including the 1,000 Genomes Project, to study, for example, how SNPs and structural variation may affect transcript, regulatory, and DNA methylation data. We foresee a time in the near future when the biochemical features defined by ENCODE are routinely combined with GWAS and other sequence variation driven studies of human phenotypes. Analogously, the systematic profiling of epigenomic features across ex vivo tissues and stem cells currently being undertaken by the NIH Roadmap Epigenomics program will provide synergistic data and the opportunity to observe the state and behavior of ENCODE-identified elements in human tissues representing healthy and disease states.

These are but a few of many applications of the ENCODE data. Investigators focused on one or a few genes should find many new insights within the ENCODE data. Indeed, these investigators are in the best position to infer potential functions and mechanisms from the ENCODE data ones that will also lead to testable hypotheses. Thus, we expect that the work of many investigators will be enhanced by these data and that their results will in turn inform the development of the project going forward.

Finally, we also expect that comprehensive paradigms for gene regulation will begin to emerge from our work and similar work from many laboratories. Deciphering the "regulatory code" within the genome and its associated epigenetic signals is a grand and complex challenge. The data contributed by ENCODE in conjunction with complementary efforts will be foundational to this effort, but equally important will be novel methods for genome-wide analysis, model building, and hypothesis testing. We therefore expect the ENCODE Project to be a major contributor not only of data but also novel technologies for deciphering the human genome and those of other organisms.

Supporting Information

Figure S1 The Organization of the ENCODE Consortium. The geographical distribution of the members of the ENCODE Consortium, with pin colors indicating the group roles as detailed in the text below.
(TIF)

Figure S2 Quantitative trait example (BCL11A). Candidates for gene regulatory features in the vicinity of SNPs at the *BCL11A* locus associated with fetal hemoglobin levels. SNPs associated with fetal hemoglobin levels are marked in red on the top line; those not associated are marked in blue. The phenotype-associated SNPs are close to an antisense transcript (AC009970.1, light orange), shown in the ENCODE gene annotations. This antisense transcript is within a region (boxed in red) with elevated levels of H3K 4me1 and DNase hypersensitive sites. The phenotype-associated region is flanked by two regions (boxed in blue) with multiple strong biochemical signals associated with transcriptional regulation, including transcription factor occupancy. The data are from the lymphoblastoid cell line GM12878, as *BCL11A* is expressed in this cell line (RNA-seq track) but not in K562 (unpublished data).
(TIF)

Table S1 This supplemental table contains additional details of the computational analysis tools used by the ENCODE Consortium that are listed in Table 3. The name of each software tool appears in the first column, and subsequent columns contain the tasks for which the tool is used, the PMID reference number when available, and a web address where the tool can be accessed.
(DOC)

Acknowledgments

We thank Judy R. Wexler and Julia Zhang at the National Human Genome Research Institute for their support in administering the ENCODE Consortium, additional members of our laboratories and institutions who have contributed to the experimental and analytical components of this project, and J. D. Frey for assistance in preparing the figures.

The ENCODE Consortium Authors

Writing Group. Richard M. Myers[1], John Stamatoyannopoulos[2], Michael Snyder[3], Ian Dunham[4], Ross C. Hardison[5], Bradley E. Bernstein[6,7], Thomas R. Gingeras[8], W. James Kent[9], Ewan Birney[4], Barbara Wold[10,11], Gregory E. Crawford[12,13].

Broad Institute Group. Bradley E. Bernstein[6,7], Charles B. Epstein[6], Noam Shoresh[6], Jason Ernst[6,14], Tarjei S. Mikkelsen[6], Pouya Kheradpour[6,14], Xiaolan Zhang[6], Li Wang[6], Robbyn Issner[6], Michael J. Coyne[6], Timothy Durham[6], Manching Ku[6], Thanh Truong[6], Lucas D. Ward[6,14], Robert C. Altshuler[14], Michael F. Lin[6,14], Manolis Kellis[6,14].

Cold Spring Harbor; University of Geneva; Center for Genomic Regulation, Barcelona; RIKEN; University of Lausanne; Genome Institute of Singapore Group. *Cold Spring Harbor I*: Thomas R. Gingeras[8], Carrie A. Davis[8], Philipp Kapranov[15], Alexander Dobin[8], Christopher Zaleski[8], Felix Schlesinger[8], Philippe Batut[8], Sudipto Chakraborty[8], Sonali Jha[8], Wei Lin[8], Jorg Drenkow[8], Huaien Wang[8], Kim Bell[8], Hui Gao[16], Ian Bell[15], Erica Dumais[15], Jacqueline Dumais[15]. *University of Geneva*: Stylianos E. Antonarakis[17], Catherine Ucla[17], Christelle Borel[17]. *Center for Genomic Regulation, Barcelona*: Roderic Guigo[18], Sarah Djebali[18], Julien Lagarde[18], Colin Kingswood[18], Paolo Ribeca[18], Micha Sammeth[18], Tyler Alioto[18], Angelika Merkel[18], Hagen Tilgner[18]. *RIKEN*: Piero Carninci[19], Yoshihide Hayashizaki[19], Timo Lassmann[19], Hazuki Takahashi[19], Rehab F. Abdelhamid[19]. *Cold Spring Harbor II*: Gregory Hannon[20], Katalin Fejes-Toth[8], Jonathan Preall[8], Assaf Gordon[8], Vihra Sotirova[8]. *University of Lausanne*: Alexandre Reymond[21], Cedric Howald[21],

Emilie Aït Yahya Graison[21], Jacqueline Chrast[21]. *Genome Institute of Singapore*: Yijun Ruan[22], Xiaoan Ruan[22], Atif Shahab[22], Wan Ting Poh[22], Chia-Lin Wei[22].

Duke University, EBI, University of Texas, Austin, University of North Carolina–Chapel Hill Group. *Duke University*: Gregory E. Crawford[12,13], Terrence S. Furey[12], Alan P. Boyle[12], Nathan C. Sheffield[12], Lingyun Song[12], Yoichiro Shibata[12], Teresa Vales[12], Deborah Winter[12], Zhancheng Zhang[12], Darin London[12], Tianyuan Wang[12]. *EBI*: Ewan Birney[4], Damian Keefe[4]. *University of Texas, Austin*: Vishwanath R. Iyer[23], Bum-Kyu Lee[23], Ryan M. McDaniell[23], Zheng Liu[23], Anna Battenhouse[23], Akshay A. Bhinge[23]. *University of North Carolina–Chapel Hill*: Jason D. Lieb[24], Linda L. Grasfeder[24], Kimberly A. Showers[24], Paul G. Giresi[24], Seul K. C. Kim[24], Christopher Shestak[24].

HudsonAlpha Institute, Caltech, Stanford Group. *HudsonAlpha Institute*: Richard M. Myers[1], Florencia Pauli[1], Timothy E. Reddy[1], Jason Gertz[1], E. Christopher Partridge[1], Preti Jain[1], Rebekka O. Sprouse[1], Anita Bansal[1], Barbara Pusey[1], Michael A. Muratet[1], Katherine E. Varley[1], Kevin M. Bowling[1], Kimberly M. Newberry[1], Amy S. Nesmith[1], Jason A. Dilocker[1], Stephanie L. Parker[1], Lindsay L. Waite[1], Krista Thibeault[1], Kevin Roberts[1], Devin M. Absher[1]. *Caltech*: Barbara Wold[10,11], Ali Mortazavi[10,11], Brian Williams[10], Georgi Marinov[10], Diane Trout[10], Shirley Pepke[25], Brandon King[10], Kenneth McCue[10], Anthony Kirilusha[10], Gilberto DeSalvo[10], Katherine Fisher-Aylor[10], Henry Amrhein[10], Jost Vielmetter[11]. *Stanford*: Gavin Sherlock[3], Arend Sidow[3,26], Serafim Batzoglou[27], Rami Rauch[3], Anshul Kundaje[26,27], Max Libbrecht[27].

NHGRI Groups. *NHGRI, Genome Informatics Section*: Elliott H. Margulies[28], Stephen C. J. Parker[28]. *NHGRI, Genomic Functional Analysis Section*: Laura Elnitski[29]. *NHGRI, NIH Intramural Sequencing Center*: Eric D. Green[30].

Sanger Institute; Washington University; Yale University; Center for Genomic Regulation, Barcelona; UCSC; MIT; University of Lausanne; CNIO Group. *Sanger Institute*: Tim Hubbard[31], Jennifer Harrow[31], Stephen Searle[31], Felix Kokocinski[31], Browen Aken[31], Adam Frankish[31], Toby Hunt[31], Gloria Despacio-Reyes[31], Mike Kay[31], Gaurab Mukherjee[31], Alexandra Bignell[31], Gary Saunders[31], Veronika Boychenko[31]. *Washington University*: Michael Brent[32], M. J. Van Baren[32], Randall H. Brown[32]. *Yale University*: Mark Gerstein[33,34,35], Ekta Khurana[33,34], Suganthi Balasubramanian[33,34], Zhengdong Zhang[33,34], Hugo Lam[33,34], Philip Cayting[3,33,34], Rebecca Robilotto[33,34], Zhi Lu[33,34]. *Center for Genomic Regulation, Barcelona*: Roderic Guigo[18], Thomas Derrien[18], Andrea Tanzer[18], David G. Knowles[18], Marco Mariotti[18]. *UCSC*: W. James Kent[9], David Haussler[9,36], Rachel Harte[9], Mark Diekhans[9]. *MIT*: Manolis Kellis[6,14], Mike Lin[6,14], Pouya Kheradpour[6,14], Jason Ernst[6,14]. *University of Lausanne*: Alexandre Reymond[21], Cedric Howald[21], Emilie Aït Yahya Graison[21], Jacqueline Chrast[21]. *CNIO*: Alfonso Valencia[37], Michael Tress[37], Jose Manuel Rodriguez[37].

Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UCDavis Group. *Stanford-Yale*: Michael Snyder[3], Stephen G. Landt[3], Debasish Raha[38], Minyi Shi[3], Ghia Euskirchen[3], Fabian Grubert[3], Maya Kasowski[38], Jin Lian[39], Philip Cayting[3,33,34], Phil Lacroute[3], Youhan Xu[38], Hannah Monahan[38], Dorrelyn Patacsil[3], Teri Slifer[3], Xinqiong Yang[3], Alexandra Charos[38], Brian Reed[38], Linfeng Wu[3], Raymond K. Auerbach[33], Lukas Habegger[33], Manoj Hariharan[3], Joel Rozowsky[33,34], Alexej Abyzov[33,34], Sherman M. Weissman[39], Mark Gerstein[33,34,35]. *Harvard*: Kevin Struhl[40], Nathan Lamarre-Vincent[40], Marianne Lindahl-Allen[40], Benoit Miotto[40], Zarmik Moqtaderi[40], Joseph D. Fleming[40]. *University of Massachusetts Medical School*: Peter Newburger[41]. *University of Southern California/UCDavis*: Peggy J. Farnham[42,43], Seth Frietze[42,43], Henriette O'Geen[43], Xiaoqin Xu[43], Kim R. Blahnik[43], Alina R. Cao[43], Sushma Iyengar[43].

University of Washington, University of Massachusetts Medical School Group. *University of Washington*: John A. Stamatoyannopoulos[2], Rajinder Kaul[2], Robert E. Thurman[2], Hao Wang[2], Patrick A. Navas[2], Richard Sandstrom[2], Peter J. Sabo[2], Molly Weaver[2], Theresa Canfield[2], Kristen Lee[2], Shane Neph[2], Vaughan Roach[2], Alex Reynolds[2], Audra Johnson[2], Eric Rynes[2], Erika Giste[2], Shinny Vong[2], Jun Neri[2], Tristan Frum[2], Ericka M. Johnson[2], Eric D. Nguyen[2], Abigail K. Ebersol[2], Minerva E. Sanchez[2], Hadar H. Sheffer[2], Dimitra Lotakis[2], Eric Haugen[2], Richard Humbert[2], Tanya Kutyavin[2], Tony Shafer[2]. *University of Massachusetts Medical School*: Job Dekker[44], Bryan R. Lajoie[44], Amartya Sanyal[44].

Data Coordination Center. W. James Kent[9], Kate R. Rosenbloom[9], Timothy R. Dreszer[9], Brian J. Raney[9], Galt P. Barber[9], Laurence R. Meyer[9], Cricket A. Sloan[9], Venkat S. Malladi[9], Melissa S. Cline[9], Katrina Learned[9], Vanessa K. Swing[9], Ann S. Zweig[9], Brooke Rhead[9], Pauline A. Fujita[9], Krishna Roskin[9], Donna Karolchik[9], Robert M. Kuhn[9], David Haussler[9,36].

Data Analysis Center. Ewan Birney[4], Ian Dunham[4], Steven P. Wilder[4], Damian Keefe[4], Daniel Sobral[4], Javier Herrero[4], Kathryn Beal[4], Margus Lukk[4], Alvis Brazma[4], Juan M. Vaquerizas[4], Nicholas M. Luscombe[4], Peter J. Bickel[45], Nathan Boley[45], James B. Brown[45], Qunhua Li[45], Haiyan Huang[45], Mark Gerstein[32,33,34], Lukas Habegger[33], Andrea Sboner[33,34], Joel Rozowsky[33,34], Raymond K. Auerbach[33], Kevin Y. Yip[33,34], Chao Cheng[33,34], Koon-Kiu Yan[33,34], Nitin Bhardwaj[33,34], Jing Wang[33,34], Lucas Lochovsky[33,34], Justin Jee[33,34], Theodore Gibson[33,34], Jing Leng[33,34], Jiang Du[35], Ross C. Hardison[5], Robert S. Harris[5], Giltae Song[5], Webb Miller[5], David Haussler[9,36], Krishna Roskin[9], Bernard Suh[9], Ting Wang[46], Benedict Paten[9], William S. Noble[2,47], Michael M. Hoffman[2], Orion J. Buske[2], Zhiping Weng[48], Xianjun Dong[48], Jie Wang[48], Hualin Xi[49].

University of Albany SUNY Group. Scott A. Tenenbaum[50], Frank Doyle[50], Luiz O. Penalva[51], Sridar Chittur[50].

Boston University Group. Thomas D. Tullius[52], Stephen C. J. Parker[28,52].

University of Chicago, Stanford Group. *University of Chicago*: Kevin P. White[53], Subhradip Karmakar[53], Alec Victorsen[53], Nader Jameel[53], Nick Bild[53], Robert L. Grossman[53]. *Stanford*: Michael Snyder[3], Stephen G. Landt[3], Xinqiong Yang[3], Dorrelyn Patacsil[3], Teri Slifer[3].

University of Massachusetts Medical School Groups. *University of Massachusetts Medical School I*: Job Dekker[44], Bryan R. Lajoie[44], Amartya Sanyal[44]. *University of Massachusetts Medical School II*: Zhiping Weng[48], Troy W. Whitfield[48], Jie Wang[48], Patrick J. Collins[54], Nathan D. Trinklein[54], E. Christopher Partridge[1], Richard M. Myers[1].

Boise State University/University of North Carolina–Chapel Hill Proteomics Group. Morgan C. Giddings[55,56,57], Xian Chen[58], Jainab Khatun[55], Chris Maier[55], Yanbao Yu[57], Harsha Gunawardena[57], Brian Risk[56].

NIH Project Management Group. Elise A. Feingold[58], Rebecca F. Lowdon[58], Laura A. L. Dillon[58], Peter J. Good[58].

Affiliations

1 HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America,

2 Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America,

3 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America,

4 European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, United Kingdom,

5 Center for Comparative Genomics and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America,

6 Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America,

7 Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America,

8 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America,

9 Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California, United States of America,

10 Biology Division, California Institute of Technology, Pasadena, California, United States of America,

11 Beckman Institute, California Institute of Technology, Pasadena, California, United States of America,

12 Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, United States of America,

13 Department of Pediatrics, Duke University, Durham, North Carolina, United States of America,

14 Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America,

15 Affymetrix, Santa Clara, California, United States of America,

16 Karolinksa Institutet, Huddinge, Sweden,

17 University of Geneva, Geneva, Switzerland,

18 Bioinformatics and Genomics, Centre de Regulacio Genomica, Barcelona, Spain,
19 Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa, Japan,
20 Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America,
21 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland,
22 Genome Institute of Singapore, Singapore,
23 Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas, United States of America,
24 Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,
25 Center for Advanced Computing Research, California Institute of Technology, Pasadena, California, United States of America,
26 Department of Pathology, Stanford University School of Medicine, Stanford, California, United States of America,
27 Department of Computer Science, Stanford University, Stanford, California, United States of America,
28 Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America,
29 National Human Genome Research Institute, Genome Technology Branch, National Institutes of Health, Rockville, Maryland, United States of America,
30 National Human Genome Research Institute, NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland, United States of America,
31 Vertebrate Genome Analysis, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom,
32 Center for Genome Sciences and Department of Computer Science, Washington University in St. Louis, St. Louis, Missouri, United States of America,
33 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America,
34 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America,
35 Department of Computer Science, Yale University, New Haven, Connecticut, United States of America,
36 Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California, United States of America,
37 Structural Computational Biology, Centro Nacional de Investigaciones Oncolùgicas, Madrid, Spain,
38 Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut, United States of America,
39 Department of Genetics, Yale University, New Haven, Connecticut, United States of America,
40 Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, United States of America,
41 Department of Pediatrics, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America,
42 Department of Biochemistry and Molecular Biology, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America,
43 Genome Center, University of California–Davis, Davis, California, United States of America,
44 Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America,
45 Department of Statistics, University of California at Berkeley, Berkeley, California, United States of America,
46 Department of Genetics, Washington University in St. Louis, St. Louis, Missouri, United States of America,
47 Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America,
48 Program in Bioinformatics and Integrative Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America,
49 Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America,
50 College of Nanoscale Sciences and Engineering, University at Albany–SUNY, Albany, New York, United States of America,
51 Children's Cancer Research Institute, Department of Cellular and Structural Biology, San Antonio, Texas, United States of America,
52 Department of Chemistry and Program in Bioinformatics, Boston University, Boston, Massachusetts, United States of America,
53 Institute for Genomics and Systems Biology, The University of Chicago, Chicago, Illinois, United States of America,
54 SwitchGear Genomics, Menlo Park, California, United States of America,
55 Biomolecular Research Center, Boise State University, Boise, Idaho, United States of America,
56 Department of Microbiology and Immunology, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,
57 Biochemistry Department, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,
58 National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422: 835 847.
2. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636 640.
3. ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799 816.
4. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520 562.
5. Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D (2003) The share of human genomic DNA under selection estimated from human-mouse genomic alignments. Cold Spring Harb Symp Quant Biol 68: 245 254.
6. Stone EA, Cooper GM, Sidow A (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. Annu Rev Genomics Hum Genet 6: 143 164.
7. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. Science 324: 389 392.
8. Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA (2007) Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A 104: 12410 12415.
9. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nature Meth 5: 19 21.
10. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57 63.
11. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, et al. (2010) ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res 38 (Database issue): D620 D625.
12. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, et al. (2010) The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38 (Database issue): D613 D619.
13. Lozzio CB, Lozzio BB (1975) Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood 45: 321 334.
14. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, et al. (1998) Embryonic stem cell lines derived from human blastocysts. Science 282: 1145 1147.
15. Gey GO, Coffman WD, Kubicek MT (1952) Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res 12: 264 265.
16. Knowles BB, Howe CC, Aden DP (1980) Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis B surface antigen. Science 209: 497 499.
17. Jaffe EA, Nachman RL, Becker CG, Minick CR (1973) Culture of human endothelial cells derived from umbilical veins: Identification by morphologic and immunologic criteria. J Clin Invest 52: 2745 2756.
18. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931 945.
19. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, et al. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7 Suppl 1: S2 1 31.
20. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7: S4 1 9.

21. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437 1439.
22. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, et al. (2010) Ensembl's 10th year. Nucleic Acids Res 38: D557 D562.
23. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, et al. (2004) Novel RNAs identified from a comprehensive analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14: 331 342.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5: 621 628.
25. Schmid M, Jensen TH (2010) Nuclear quality control of RNA polymerase II transcripts. J Wiley Interdisciplinary Review.
26. Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. Biochimie 84: 775 790.
27. Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8: 413 423.
28. Fullwood MJ, Wei CL, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 19: 521 532.
29. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100: 15776 15781.
30. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626 635.
31. Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, et al. (2009) Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res 19: 255 265.
32. Frohman MA, Dush MK, Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A 85: 8998 9002.
33. Giddings MC, Shah AA, Gesteland R, Moore B (2003) Genome-based peptide fingerprint scanning. Proc Natl Acad Sci U S A 100: 20 25.
34. Merrihew GE, Davis C, Ewing B, Williams G, Käll L, et al. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C elegans gene annotations. Genome Res 18: 1660 1669.
35. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7: 29 59.
36. Wu C (1980) The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286: 854 860.
37. Keene MA, Corces V, Lowenhaupt K, Elgin SC (1981) DNase I hypersensitive sites in Drosophila chromatin occur at the 5′ ends of regions of transcription. Proc Natl Acad Sci U S A 78: 143 146.
38. McGhee JD, Wood WI, Dolan M, Engel JD, Felsenfeld G (1981) A 200 base pair region at the 5′ end of the chicken adult beta-globin gene is accessible to nuclease digestion. Cell 27: 45 55.
39. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57: 159 197.
40. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877 885.
41. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, et al. (2009) Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci U S A 106: 14926 14931.
42. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693 705.
43. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669 681.
44. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching C, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311 318.
45. Liang G, Lin JC, Wei V, Yoo C, Cheng J, et al. (2004) Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A 11: 7357 7362.
46. Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat Cell Biol 6: 73 77.
47. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 28: 169 181.
48. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 18: 349 353.
49. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 21: 301 313.
50. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315 326.
51. Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 6: 41 45.
52. Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, et al. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc Natl Acad Sci U S A 101: 16837 16842.
53. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 25: 311 322.
54. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3: 511 518.
55. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, et al. (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6: 283 289.
56. Sekimata M, Pérez-Melgosa M, Miller SA, Weinmann AS, Sabo PJ, et al. (2009) CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. Immunity 31: 551 564.
57. Giresi PG, Lieb JD (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods 48: 233 239.
58. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, et al. (2010) A map of open chromatin in human pancreatic islets. Nat Genet 42: 255 259.
59. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823 837.
60. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497 1502.
61. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553 560.
62. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651 657.
63. Poser I, Sarov M, Hutchins JR, Hériché JK, Toyoda Y, et al. (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods 5: 409 415.
64. Hua S, Kittler R, White KP (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. Cell 137: 1259 1271.
65. Raha D, Hong M, Snyder M (2010) ChIP-seq: a method for global identification of regulatory elements in the genome. Curr Protoc Mol Biol Chapter 21: Unit 2119 1 14.
66. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231 1245.
67. Jaenisch R (1997) DNA methylation and imprinting: Why bother? Trends Genet 13: 323 329.
68. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes & Dev 16: 6 21.
69. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, et al. (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 17: 510 522.
70. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger SJ, et al. (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 20: 440 446.
71. Laurent L, Wong E, Li G, Tsirigos A, Ong CT, et al. (2010) Dynamic changes in the human methylome during differentiation. Genome Res 20: 320 331.
72. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, et al. (2010) Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res 20: 434 439.
73. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766 770.
74. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, et al. (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res 19: 1044 1056.
75. Galas DJ, Schmitz A (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5: 3157 3170.
76. Strauss EC, Orkin SH (1992) In vivo protein-DNA interactions at hypersensitive site 3 of the human beta-globin locus control region. Proc Natl Acad Sci U S A 89: 5809 5813.
77. Boyle AP, Song L, Lee BK, London D, Keefe D, et al. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res, in press.
78. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res, in press.
79. Lu Y, Dimasi DP, Hysi PG, Hewitt AW, Burdon KP, et al. (2010) Common genetic variants near the Brittle Cornea Syndrome locus ZNF469 influence the blinding disease risk factor central corneal thickness. PLoS Genet 6: e1000947. doi:10.1371/journal.pgen.1000947.
80. McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, et al. (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328: 235 239.

81. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, et al. (2010) Variation in transcription factor binding among humans. Science 328: 232 235.
82. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444 454.
83. Korbel JO, Urban AE, Grubert F, Du J, Royce TE, et al. (2007) Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proc Natl Acad Sci U S A 104: 10110 10115.
84. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, et al. (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet 42: 385 391.
85. Miele A, Dekker J (2008) Long-range chromosomal interactions and gene regulation. Mol Biosyst 4: 1046 1057.
86. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16: 1299 1309.
87. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science 15: 1306 1311.
88. Lajoie BR, van Berkum NL, Sanyal A, Dekker J (2009) My5C: web tools for chromosome conformation capture studies. Nat Methods 6: 690 691.
89. Baroni TE, Chittur SV, George AD, Tennenbaum SA (2008) Advances in RIP-chip analysis: RNA-binding protein immunoprecipitation-microarray profiling. Methods Mol Biol 419: 93 108.
90. Keene JD, Komisarow JM, Friedersdorf MB (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc 1: 302 307.
91. Tenenbaum SA, et al. (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc Natl Acad Sci U S A 97: 14085 14090.
92. Tenenbaum SA, Lager PJ, Carson CC, Keene JD (2002) Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. Methods 26: 191 198.
93. Trinklein ND, Karaöz U, Wu J, Halees A, Force Aldred S, et al. (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. Genome Res 17: 720 731.
94. Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, et al. (2007) Transcription factor binding and histone modifications in human bidirectional promoters. Genome Res 17: 818 827.
95. Collins PJ, Kobayashi Y, Nguyen L, Trinklein ND, Myers RM (2007) The ets-related transcription factor GABP directs bidirectional transcription. PLoS Genet 3(11): e208. doi:10.1371/journal.pgen.0030208.
96. Petrykowska HM, Vockley CM, Elnitski L (2008) Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. Genome Research 18: 1238 1246.
97. Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, et al. (2010) Sequence features that drive human promoter function and tissue specificity. Genome Res 20: 890 898.
98. Khatun J, Hamlett E, Giddings MC (2008) Incorporating sequence information into the scoring function: a hidden Markov model for improved peptide identification. Bioinformatics 24: 674 681.
99. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E (2008a) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res 8: 1814 1828.
100. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P (2008b) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 18: 1829 1843.
101. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901 913.
102. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S (2007) Analysis of sequence conservation at nucleotide resolution. PLoS Comput Biol 3: e254. doi:10.1371/journal.pcbi.0030254.
103. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20: 110 121.
104. Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, et al. (2010) Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. Proc Natl Acad Sci U S A 107: 5254 5259.
105. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66 75.
106. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, et al. (2009) Prepublication data sharing. Nature 461: 168 170.
107. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747 753.
108. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362 9367.
109. Wokolorczyk D, Gliniewicz B, Sikorski A, Zlowocka E, Masojc B, et al. (2008) A range of cancers is associated with the rs6983267 marker on chromosome 8. Cancer Res 68: 9982 9986.
110. Curtin K, Lin WY, George R, Katory M, Shorto J, et al. (2009) Meta association of colorectal cancer confirms risk alleles at 8q24 and 18q21. Cancer Epidemiol Biomarkers Prev 18: 616 621.
111. Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, et al. (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet 41: 1058 1060.
112. Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, et al. (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Gene 5: e1000597. doi:10.1371/journal.pgen.1000597.
113. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41: 882 884.
114. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41: 885 890.
115. Wright JB, Brown SJ, Cole MD (2010) Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol 30: 1411 1420.
116. Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, et al. (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods 7: 365 371.
117. Li Q, Brown JB, Huang H, Bickel P (2011) Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics, in press.
118. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252 263.
119. Nelson SB, Sugino K, Hempel CM (2006) The problem of neuronal cell types: a physiological genomics approach. Trends Neurosci 29: 339 345.
120. Reisman D, Bálint é, Loging WT, Rotter V, Almon E (1996) A novel transcript encoded within the 10-kb first intron of the human p53 tumor suppressor gene (D17S2179E) is induced during differentiation of myeloid leukemia cells. Genomics 38: 364 370.
121. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40 45.

APPENDIX C: An Integrated Encyclopedia of DNA Elements in the Human Genome

AUTHOR CONTRIBUTIONS

J.F. provided ChIP-Seq datasets as part of the Consortium's effort.

ARTICLE

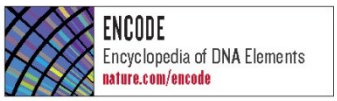
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with regard to non-coding RNAs, alternatively spliced transcripts and regulatory sequences. Systematic analyses of transcripts and regulatory information are essential for the identification of genes and regulatory regions, and are an important resource for the study of human biology and disease. Such analyses can also provide comprehensive views of the organization and variability of genes and regulatory information across cellular contexts, species and individuals.



The Encyclopedia of DNA Elements (ENCODE) project aims to delineate all functional elements encoded in the human genome[1–3]. Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure). Comparative genomic studies suggest that 3–8% of bases are under purifying (negative) selection[4–8] and therefore may be functional, although other analyses have suggested much higher estimates[9–11]. In a pilot phase covering 1% of the genome, the ENCODE project annotated 60% of mammalian evolutionarily constrained bases, but also identified many additional putative functional elements without evidence of constraint[2]. The advent of more powerful DNA sequencing technologies now enables whole-genome and more precise analyses with a broad repertoire of functional assays.

Here we describe the production and initial analysis of 1,640 data sets designed to annotate functional elements in the entire human genome. We integrate results from diverse experiments within cell types, related experiments involving 147 different cell types, and all ENCODE data with other resources, such as candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained regions. Together, these efforts reveal important features about the organization and function of the human genome, summarized below.

- The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.
- Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.
- Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.
- It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.
- Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.
- Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.

ENCODE data production and initial analyses

Since 2007, ENCODE has developed methods and performed a large number of sequence-based studies to map functional elements across the human genome[3]. The elements mapped (and approaches used) include RNA transcribed regions (RNA-seq, CAGE, RNA-PET and manual annotation), protein-coding regions (mass spectrometry), transcription-factor-binding sites (ChIP-seq and DNase-seq), chromatin structure (DNase-seq, FAIRE-seq, histone ChIP-seq and MNase-seq), and DNA methylation sites (RRBS assay) (Box 1 lists methods and abbreviations; Supplementary Table 1, section P, details production statistics)[3]. To compare and integrate results across the different laboratories, data production efforts focused on two selected

*Lists of participants and their affiliations appear at the end of the paper.

BOX 1

ENCODE abbreviations

RNA-seq. Isolation of RNA sequences, often with different purification techniques to isolate different fractions of RNA followed by high-throughput sequencing.

CAGE. Capture of the methylated cap at the 5′ end of RNA, followed by high-throughput sequencing of a small tag adjacent to the 5′ methylated caps. 5′ methylated caps are formed at the initiation of transcription, although other mechanisms also methylate 5′ ends of RNA.

RNA-PET. Simultaneous capture of RNAs with both a 5′ methyl cap and a poly(A) tail, which is indicative of a full-length RNA. This is then followed by sequencing a short tag from each end by high-throughput sequencing.

ChIP-seq. Chromatin immunoprecipitation followed by sequencing. Specific regions of crosslinked chromatin, which is genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then subjected to high-throughput sequencing to determine the regions in the genome most often bound by the protein to which the antibody was directed. Most often used are antibodies to any chromatin-associated epitope, including transcription factors, chromatin binding proteins and specific chemical modifications on histone proteins.

DNase-seq. Adaption of established regulatory sequence assay to modern techniques. The DNase I enzyme will preferentially cut live chromatin preparations at sites where nearby there are specific (non-histone) proteins. The resulting cut points are then sequenced using high-throughput sequencing to determine those sites 'hypersensitive' to DNase I, corresponding to open chromatin.

FAIRE-seq. Formaldehyde assisted isolation of regulatory elements. FAIRE isolates nucleosome-depleted genomic regions by exploiting the difference in crosslinking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low). FAIRE consists of crosslinking, phenol extraction, and sequencing the DNA fragments in the aqueous phase.

RRBS. Reduced representation bisulphite sequencing. Bisulphite treatment of DNA sequence converts unmethylated cytosines to uracil. To focus the assay and save costs, specific restriction enzymes that cut around CpG dinucleotides can reduce the genome to a portion specifically enriched in CpGs. This enriched sample is then sequenced to determine the methylation status of individual cytosines quantitatively.

Tier 1. Tier 1 cell types were the highest-priority set and comprised three widely studied cell lines: K562 erythroleukaemia cells; GM12878, a B-lymphoblastoid cell line that is also part of the 1000 Genomes project (http://1000genomes.org)[55]; and the H1 embryonic stem cell (H1 hESC) line.

Tier 2. The second-priority set of cell types in the ENCODE project which included HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells and primary (non-transformed) human umbilical vein endothelial cells (HUVECs).

Tier 3. Any other ENCODE cell types not in tier 1 or tier 2.

sets of cell lines, designated 'tier 1' and 'tier 2' (Box 1). To capture a broader spectrum of biological diversity, selected assays were also executed on a third tier comprising more than 100 cell types including primary cells. All data and protocol descriptions are available at http://www.encodeproject.org/, and a User's Guide including details of cell-type choice and limitations was published recently[3].

Integration methodology

For consistency, data were generated and processed using standardized guidelines, and for some assays, new quality-control measures were designed (see refs 3, 12 and http://encodeproject.org/ENCODE/dataStandards.html; A. Kundaje, personal communication). Uniform data-processing methods were developed for each assay (see Supplementary Information; A. Kundaje, personal communication), and most assay results can be represented both as signal information (a per-base estimate across the genome) and as discrete elements (regions computationally identified as enriched for signal). Extensive processing pipelines were developed to generate each representation (M. M. Hoffman *et al.*, manuscript in preparation and A. Kundaje, personal communication). In addition, we developed the irreproducible discovery rate (IDR)[13] measure to provide a robust and conservative estimate of the threshold where two ranked lists of results from biological replicates no longer agree (that is, are irreproducible), and we applied this to defining sets of discrete elements. We identified, and excluded from most analyses, regions yielding untrustworthy signals likely to be artefactual (for example, multicopy regions). Together, these regions comprise 0.39% of the genome (see Supplementary Information). The poster accompanying this issue represents different ENCODE-identified elements and their genome coverage.

Transcribed and protein-coding regions

We used manual and automated annotation to produce a comprehensive catalogue of human protein-coding and non-coding RNAs as well as pseudogenes, referred to as the GENCODE reference gene set[14,15] (Supplementary Table 1, section U). This includes 20,687 protein-coding genes (GENCODE annotation, v7) with, on average, 6.3 alternatively spliced transcripts (3.9 different protein-coding transcripts) per locus. In total, GENCODE-annotated exons of protein-coding genes cover 2.94% of the genome or 1.22% for protein-coding exons. Protein-coding genes span 33.45% from the outermost start to stop codons, or 39.54% from promoter to poly(A) site. Analysis of mass spectrometry data from K562 and GM12878 cell lines yielded 57 confidently identified unique peptide sequences in intergenic regions relative to GENCODE annotation. Taken together with evidence of pervasive genome transcription[16], these data indicate that additional protein-coding genes remain to be found.

In addition, we annotated 8,801 automatically derived small RNAs and 9,640 manually curated long non-coding RNA (lncRNA) loci[17]. Comparing lncRNAs to other ENCODE data indicates that lncRNAs are generated through a pathway similar to that for protein-coding genes[17]. The GENCODE project also annotated 11,224 pseudogenes, of which 863 were transcribed and associated with active chromatin[18].

RNA

We sequenced RNA[16] from different cell lines and multiple subcellular fractions to develop an extensive RNA expression catalogue. Using a conservative threshold to identify regions of RNA activity, 62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons. Of these bases, only 5.5% are explained by GENCODE exons. Most transcribed bases are within or overlapping annotated gene boundaries (that is, intronic), and only 31% of bases in sequenced transcripts were intergenic[16].

We used CAGE-seq (5′ cap-targeted RNA isolation and sequencing) to identify 62,403 transcription start sites (TSSs) at high confidence (IDR of 0.01) in tier 1 and 2 cell types. Of these, 27,362 (44%) are within 100 base pairs (bp) of the 5′ end of a GENCODE-annotated transcript or previously reported full-length messenger RNA. The remaining regions predominantly lie across exons and 3′ untranslated regions (UTRs), and some exhibit cell-type-restricted expression; these may represent the start sites of novel, cell-type-specific transcripts.

Finally, we saw a significant proportion of coding and non-coding transcripts processed into steady-state stable RNAs shorter than 200 nucleotides. These precursors include transfer RNA, microRNA, small nuclear RNA and small nucleolar RNA (tRNA, miRNA, snRNA and snoRNA, respectively) and the 5′ termini of these processed products align with the capped 5′ end tags[16].

Table 1 | Summary of transcription factor classes analysed in ENCODE

Acronym	Description	Factors analysed
ChromRem	ATP-dependent chromatin complexes	5
DNARep	DNA repair	3
HISase	Histone acetylation, deacetylation or methylation complexes	8
Other	Cyclin kinase associated with transcription	1
Pol2	Pol II subunit	1 (2 forms)
Pol3	Pol III-associated	6
TFNS	General Pol II-associated factor, not site-specific	8
TFSS	Pol II transcription factor with sequence-specific DNA binding	87

Protein bound regions

To identify regulatory regions directly, we mapped the binding locations of 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types using ChIP-seq (Table 1, Supplementary Table 1, section N, and ref. 19); 87 (73%) were sequence-specific transcription factors. Overall, 636,336 binding regions covering 231 megabases (Mb; 8.1%) of the genome are enriched for regions bound by DNA-binding proteins across all cell types. We assessed each protein-binding site for enrichment of known DNA-binding motifs and the presence of novel motifs. Overall, 86% of the DNA segments occupied by sequence-specific transcription factors contained a strong DNA-binding motif, and in most (55%) cases the known motif was most enriched (P. Kheradpour and M. Kellis, manuscript in preparation).

Protein-binding regions lacking high or moderate affinity cognate recognition sites have 21% lower median scores by rank than regions with recognition sequences (Wilcoxon rank sum P value $<10^{-16}$). Eighty-two per cent of the low-signal regions have high-affinity recognition sequences for other factors. In addition, when ChIP-seq peaks are ranked by their concordance with their known recognition sequence, the median DNase I accessibility is twofold higher in the bottom 20% of peaks than in the upper 80% (genome structure correction (GSC)[20] P value $<10^{-16}$), consistent with previous observations[21–24]. We speculate that low signal regions are either lower-affinity sites[21] or indirect transcription-factor target regions associated through interactions with other factors (see also refs 25, 26).

We organized all the information associated with each transcription factor—including the ChIP-seq peaks, discovered motifs and associated histone modification patterns—in FactorBook (http://www.factorbook.org; ref. 26), a public resource that will be updated as the project proceeds.

DNase I hypersensitive sites and footprints

Chromatin accessibility characterized by DNase I hypersensitivity is the hallmark of regulatory DNA regions[27,28]. We mapped 2.89 million unique, non-overlapping DNase I hypersensitive sites (DHSs) by DNase-seq in 125 cell types, the overwhelming majority of which lie distal to TSSs[29]. We also mapped 4.8 million sites across 25 cell types that displayed reduced nucleosomal crosslinking by FAIRE, many of which coincide with DHSs. In addition, we used micrococcal nuclease to map nucleosome occupancy in GM12878 and K562 cells[30].

In tier 1 and tier 2 cell types, we identified a mean of 205,109 DHSs per cell type (at false discovery rate (FDR) 1%), encompassing an average of 1.0% of the genomic sequence in each cell type, and 3.9% in aggregate. On average, 98.5% of the occupancy sites of transcription factors mapped by ENCODE ChIP-seq (and, collectively, 94.4% of all 1.1 million transcription factor ChIP-seq peaks in K562 cells) lie within accessible chromatin defined by DNase I hotspots[29]. However, a small number of factors, most prominently heterochromatin-bound repressive complexes (for example, the TRIM28–SETDB1–ZNF274 complex[31,32] encoded by the *TRIM28*, *SETDB1* and *ZNF274* genes), seem to occupy a significant fraction of nucleosomal sites.

Using genomic DNase I footprinting[33,34] on 41 cell types we identified 8.4 million distinct DNase I footprints (FDR 1%)[25]. Our *de novo* motif discovery on DNase I footprints recovered ~90% of known transcription factor motifs, together with hundreds of novel evolutionarily conserved motifs, many displaying highly cell-selective occupancy patterns similar to major developmental and tissue-specific regulators.

Regions of histone modification

We assayed chromosomal locations for up to 12 histone modifications and variants in 46 cell types, including a complete matrix of eight modifications across tier 1 and tier 2. Because modification states may span multiple nucleosomes, which themselves can vary in position across cell populations, we used a continuous signal measure of histone modifications in downstream analysis, rather than calling regions (M. M. Hoffman *et al.*, manuscript in preparation; see http://code.google.com/p/align2rawsignal/). For the strongest, 'peak-like' histone modifications, we used MACS[35] to characterize enriched sites. Table 2 describes the different histone modifications, their peak characteristics, and a summary of their known roles (reviewed in refs 36–39).

Our data show that global patterns of modification are highly variable across cell types, in accordance with changes in transcriptional activity. Consistent with previous studies[40,41], we find that integration of the different histone modification information can be used systematically to assign functional attributes to genomic regions (see below).

DNA methylation

Methylation of cytosine, usually at CpG dinucleotides, is involved in epigenetic regulation of gene expression. Promoter methylation is typically associated with repression, whereas genic methylation correlates with transcriptional activity[42]. We used reduced representation bisulphite sequencing (RRBS) to profile DNA methylation quantitatively for an average of 1.2 million CpGs in each of 82 cell lines and tissues (8.6% of non-repetitive genomic CpGs), including CpGs in intergenic regions, proximal promoters and intragenic regions (gene bodies)[43], although it should be noted that the RRBS method preferentially targets CpG-rich islands. We found that 96% of CpGs exhibited differential methylation in at least one cell type or tissue

Table 2 | Summary of ENCODE histone modifications and variants

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

assayed (K. Varley *et al.*, personal communication), and levels of DNA methylation correlated with chromatin accessibility. The most variably methylated CpGs are found more often in gene bodies and intergenic regions, rather than in promoters and upstream regulatory regions. In addition, we identified an unexpected correspondence between unmethylated genic CpG islands and binding by P300, a histone acetyltransferase linked to enhancer activity[44].

Because RRBS is a sequence-based assay with single-base resolution, we were able to identify CpGs with allele-specific methylation consistent with genomic imprinting, and determined that these loci exhibit aberrant methylation in cancer cell lines (K. Varley *et al.*, personal communication). Furthermore, we detected reproducible cytosine methylation outside CpG dinucleotides in adult tissues[45], providing further support that this non-canonical methylation event may have important roles in human biology (K. Varley *et al.*, personal communication).

Chromosome–interacting regions

Physical interaction between distinct chromosome regions that can be separated by hundreds of kilobases is thought to be important in the regulation of gene expression[46]. We used two complementary chromosome conformation capture (3C)-based technologies to probe these long-range physical interactions.

A 3C-carbon copy (5C) approach[47,48] provided unbiased detection of long-range interactions with TSSs in a targeted 1% of the genome (the 44 ENCODE pilot regions) in four cell types (GM12878, K562, HeLa-S3 and H1 hESC)[49]. We discovered hundreds of statistically significant long-range interactions in each cell type after accounting for chromatin polymer behaviour and experimental variation. Pairs of interacting loci showed strong correlation between the gene expression level of the TSS and the presence of specific functional element classes such as enhancers. The average number of distal elements interacting with a TSS was 3.9, and the average number of TSSs interacting with a distal element was 2.5, indicating a complex network of interconnected chromatin. Such interwoven long-range architecture was also uncovered genome-wide using chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)[50] applied to identify interactions in chromatin enriched by RNA polymerase II (Pol II) ChIP from five cell types[51]. In K562 cells, we identified 127,417 promoter-centred chromatin interactions using ChIA-PET, 98% of which were intra-chromosomal. Whereas promoter regions of 2,324 genes were involved in 'single-gene' enhancer–promoter interactions, those of 19,813 genes were involved in 'multi-gene' interaction complexes spanning up to several megabases, including promoter–promoter and enhancer–promoter interactions[51].

These analyses portray a complex landscape of long-range gene–element connectivity across ranges of hundreds of kilobases to several megabases, including interactions among unrelated genes (Supplementary Fig. 1, section Y). Furthermore, in the 5C results, 50–60% of long-range interactions occurred in only one of the four cell lines, indicative of a high degree of tissue specificity for gene–element connectivity[49].

Summary of ENCODE-identified elements

Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element (detailed in Supplementary Table 1, section Q). The broadest element class represents the different RNA types, covering 62% of the genome (although the majority is inside of introns or near genes). Regions highly enriched for histone modifications form the next largest class (56.1%). Excluding RNA elements and broad histone elements, 44.2% of the genome is covered. Smaller proportions of the genome are occupied by regions of open chromatin (15.2%) or sites of transcription factor binding (8.1%), with 19.4% covered by at least one DHS or transcription factor ChIP-seq peak across all cell lines. Using our most conservative assessment, 8.5% of bases are covered by either a transcription-factor-binding-site motif (4.6%) or a DHS footprint (5.7%). This, however, is still about 4.5-fold higher than the amount of protein-coding exons, and about twofold higher than the estimated amount of pan-mammalian constraint.

Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages, these proportions must be underestimates of the total amount of functional bases. However, many assays were performed on more than one cell type, allowing assessment of the rate of discovery of new elements. For both DHSs and CTCF-bound sites, the number of new elements initially increases rapidly with a steep gradient for the saturation curve and then slows with increasing number of cell types (Supplementary Figs 1 and 2, section R). With the current data, at the flattest part of the saturation curve each new cell type adds, on average, 9,500 DHS elements (across 106 cell types) and 500 CTCF-binding elements (across 49 cell types), representing 0.45% of the total element number. We modelled saturation for the DHSs and CTCF-binding sites using a Weibull distribution ($r^2 > 0.999$) and predict saturation at approximately 4.1 million (standard error (s.e.) = 108,000) and 185,100 (s.e. = 18,020) sites, respectively, indicating that we have discovered around half of the estimated total DHSs. These estimates represent a lower bound, but reinforce the observation that there is more non-coding functional DNA than either coding sequence or mammalian evolutionarily constrained bases.

The impact of selection on functional elements

From comparative genomic studies, at least 3–8% of bases are under purifying (negative) selection[4–11], indicating that these bases may potentially be functional. We previously found that 60% of mammalian evolutionarily constrained bases were annotated in the ENCODE pilot project, but also observed that many functional elements lacked evidence of constraint[2], a conclusion substantiated by others[52–54]. The diversity and genome-wide occurrence of functional elements now identified provides an unprecedented opportunity to examine further the forces of negative selection on human functional sequences.

We examined negative selection using two measures that highlight different periods of selection in the human genome. The first measure, inter-species, pan-mammalian constraint (GERP-based scores; 24 mammals[8]), addresses selection during mammalian evolution. The second measure is intra-species constraint estimated from the numbers of variants discovered in human populations using data from the 1000 Genomes project[55], and covers selection over human evolution. In Fig. 1, we plot both these measures of constraint for different classes of identified functional elements, excluding features overlapping exons and promoters that are known to be constrained. Each graph also shows genomic background levels and measures of coding-gene constraint for comparison. Because we plot human population diversity on an inverted scale, elements that are more constrained by negative selection will tend to lie in the upper and right-hand regions of the plot.

For DNase I elements (Fig. 1b) and bound motifs (Fig. 1c), most sets of elements show enrichment in pan-mammalian constraint and decreased human population diversity, although for some cell types the DNase I sites do not seem overall to be subject to pan-mammalian constraint. Bound transcription factor motifs have a natural control from the set of transcription factor motifs with equal sequence potential for binding but without binding evidence from ChIP-seq experiments—in all cases, the bound motifs show both more mammalian constraint and higher suppression of human diversity.

Consistent with previous findings, we do not observe genome-wide evidence for pan-mammalian selection of novel RNA sequences (Fig. 1d). There are also a large number of elements without mammalian constraint, between 17% and 90% for transcription-factor-binding regions as well as DHSs and FAIRE regions. Previous studies could not determine whether these sequences are either biochemically active, but with little overall impact on the organism, or under lineage-specific selection. By isolating sequences preferentially inserted into

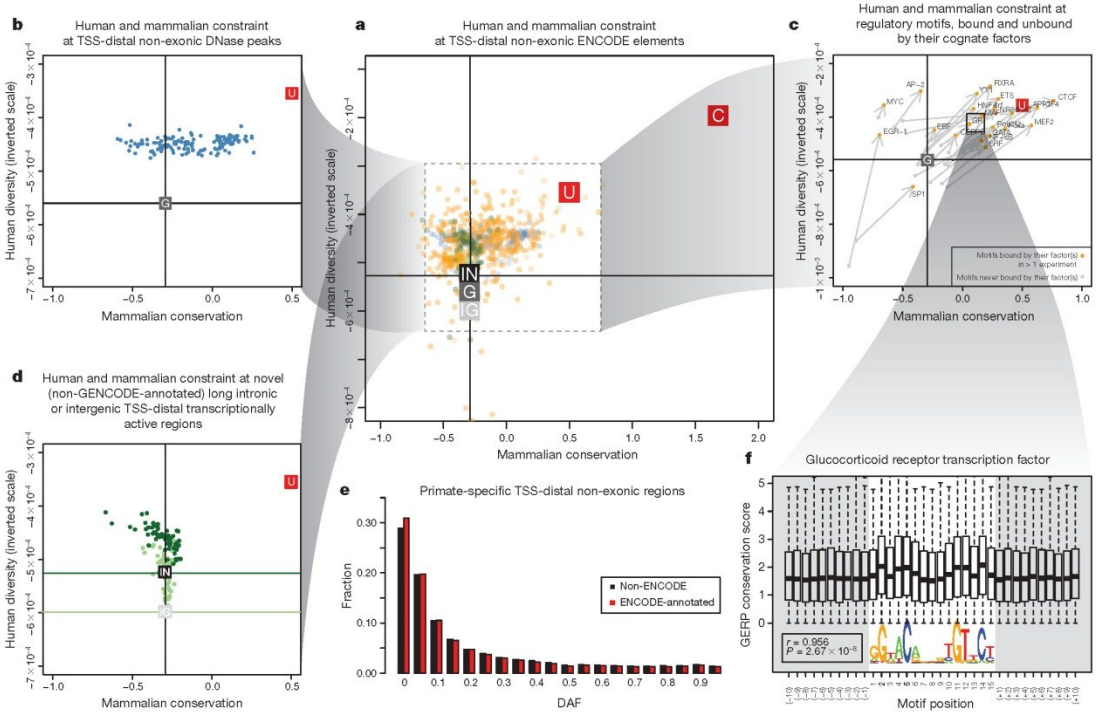


Figure 1 | Impact of selection on ENCODE functional elements in mammals and human populations. **a**, Levels of pan-mammalian constraint (mean GERP score; 24 mammals[8], *x* axis) compared to diversity, a measure of negative selection in the human population (mean expected heterozygosity, inverted scale, *y* axis) for ENCODE data sets. Each point is an average for a single data set. The top-right corners have the strongest evolutionary constraint and lowest diversity. Coding (C), UTR (U), genomic (G), intergenic (IG) and intronic (IN) averages are shown as filled squares. In each case the vertical and horizontal cross hairs show representative levels for the neutral expectation for mammalian conservation and human population diversity, respectively. The spread over all non-exonic ENCODE elements greater than 2.5 kb from TSSs is shown. The inner dashed box indicates that parts of the plot have been magnified for the surrounding outer panels, although the scales in the outer plots provide the exact regions and dimensions magnified. The spread for DHS sites (**b**) and RNA elements (**d**) is shown in the plots on the left. RNA elements are either long novel intronic (dark green) or long intergenic (light green) RNAs. The horizontal cross hairs are colour-coded to the relevant data set in **d**. **c**, Spread of transcription factor motif instances either in regions bound by the transcription factor (orange points) or in the corresponding unbound motif matches in grey, with bound and unbound points connected with an arrow in each case showing that bound sites are generally more constrained and less diverse. **e**, Derived allele frequency spectrum for primate-specific elements, with variations outside ENCODE elements in black and variations covered by ENCODE elements in red. The increase in low-frequency alleles compared to background is indicative of negative selection occurring in the set of variants annotated by the ENCODE data. **f**, Aggregation of mammalian constraint scores over the glucocorticoid receptor (GR) transcription factor motif in bound sites, showing the expected correlation with the information content of bases in the motif. An interactive version of this figure is available in the online version of the paper.

the primate lineage, which is only feasible given the genome-wide scale of this data, we are able to examine this issue specifically. Most primate-specific sequence is due to retrotransposon activity, but an appreciable proportion is non-repetitive primate-specific sequence. Of 104,343,413 primate-specific bases (excluding repetitive elements), 67,769,372 (65%) are found within ENCODE-identified elements. Examination of 227,688 variants segregating in these primate-specific regions revealed that all classes of elements (RNA and regulatory) show depressed derived allele frequencies, consistent with recent negative selection occurring in at least some of these regions (Fig. 1e). An alternative approach examining sequences that are not clearly under pan-mammalian constraint showed a similar result (L. Ward and M. Kellis, manuscript submitted). This indicates that an appreciable proportion of the unconstrained elements are lineage-specific elements required for organismal function, consistent with long-standing views of recent evolution[56], and the remainder are probably 'neutral' elements[2] that are not currently under selection but may still affect cellular or larger scale phenotypes without an effect on fitness.

The binding patterns of transcription factors are not uniform, and we can correlate both inter- and intra-species measures of negative selection with the overall information content of motif positions. The selection on some motif positions is as high as protein-coding exons (Fig. 1f; L. Ward and M. Kellis, manuscript submitted). These aggregate measures across motifs show that the binding preferences found in the population of sites are also relevant to the per-site behaviour. By developing a per-site metric of population effect on bound motifs, we found that highly constrained bound instances across mammals are able to buffer the impact of individual variation[57].

ENCODE data integration with known genomic features

Promoter-anchored integration

Many of the ENCODE assays directly or indirectly provide information about the action of promoters. Focusing on the TSSs of protein-coding transcripts, we investigated the relationships between different ENCODE assays, in particular testing the hypothesis that RNA expression (output) can be effectively predicted from patterns of

chromatin modification or transcription factor binding (input). Consistent with previous reports[58], we observe two relatively distinct types of promoter: (1) broad, mainly (C+G)-rich, TATA-less promoters; and (2) narrow, TATA-box-containing promoters. These promoters have distinct patterns of histone modifications, and transcription-factor-binding sites are selectively enriched in each class (Supplementary Fig. 1, section Z).

We developed predictive models to explore the interaction between histone modifications and measures of transcription at promoters, distinguishing between modifications known to be added as a consequence of transcription (such as H3K36me3 and H3K79me2) and other categories of histone marks[59]. In our analyses, the best models had two components: an initial classification component (on/off) and a second quantitative model component. Our models showed that activating acetylation marks (H3K27ac and H3K9ac) are roughly as informative as activating methylation marks (H3K4me3 and H3K4me2) (Fig. 2a). Although repressive marks, such as H3K27me3 or H3K9me3, show negative correlation both individually and in the model, removing these marks produces only a small reduction in model performance. However, for a subset of promoters in each cell line, repressive histone marks (H3K27me3 or H3K9me3) must be used to predict their expression accurately. We also examined the interplay between the H3K79me2 and H3K36me3 marks, both of which mark gene bodies, probably reflecting recruitment of modification enzymes by polymerase isoforms. As described previously, H3K79me2 occurs preferentially at the 5′ ends of gene bodies and H3K36me3 occurs more 3′, and our analyses support the previous model in which the H3K79me2 to H3K36me3 transition occurs at the first 3′ splice site[60].

Few previous studies have attempted to build qualitative or quantitative models of transcription genome-wide from transcription factor levels because of the paucity of documented transcription-factor-binding regions and the lack of coordination around a single cell line. We thus examined the predictive capacity of transcription-factor-binding signals for the expression levels of promoters (Fig. 2b).

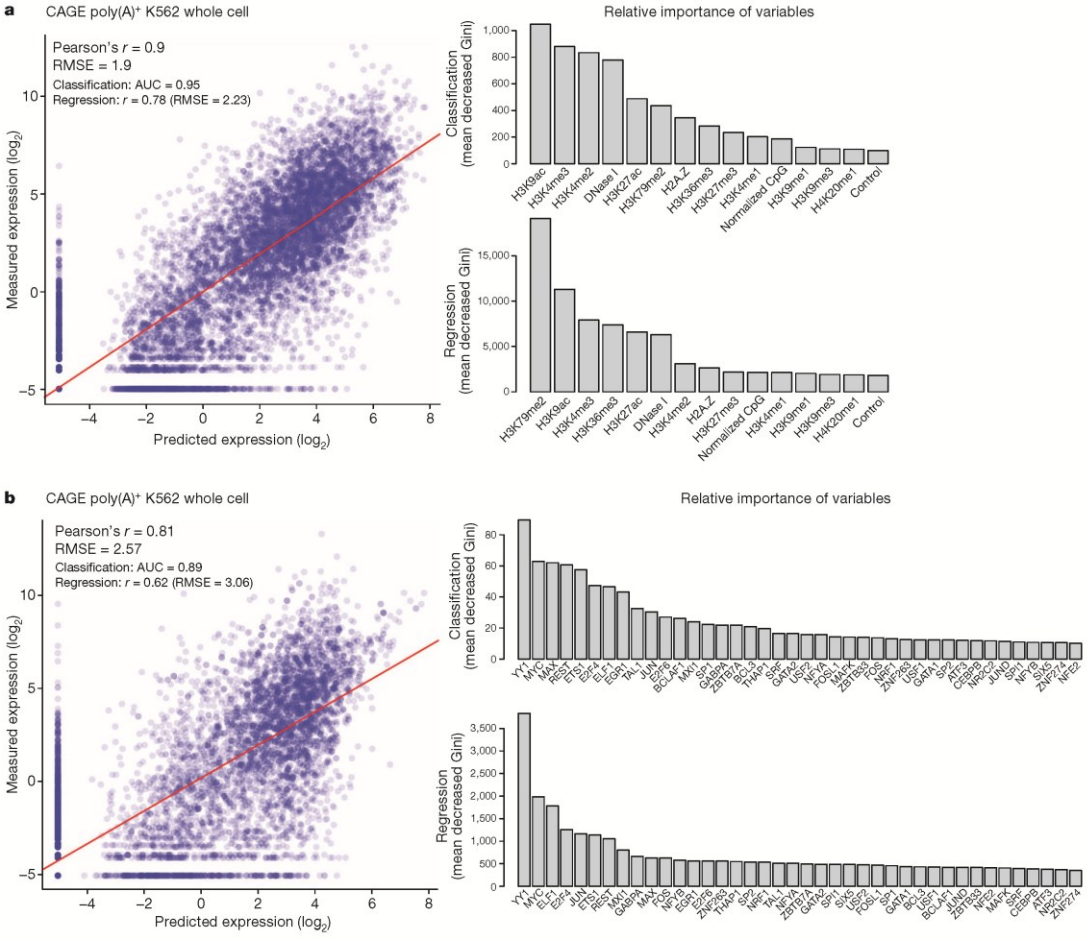


Figure 2 | Modelling transcription levels from histone modification and transcription-factor-binding patterns. **a**, **b**, Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each case the scatter plot shows the output of the correlation models (*x* axis) compared to observed values (*y* axis). The bar graphs show the most important histone modifications (**a**) or transcription factors (**b**) in both the initial classification phase (top bar graph) or the quantitative regression phase (bottom bar graph), with larger values indicating increasing importance of the variable in the model. Further analysis of other cell lines and RNA measurement types is reported elsewhere[59,79]. AUC, area under curve; Gini, Gini coefficient; RMSE, root mean square error.

In contrast to the profiles of histone modifications, most transcription factors show enriched binding signals in a narrow DNA region near the TSS, with relatively higher binding signals in promoters with higher CpG content. Most of this correlation could be recapitulated by looking at the aggregate binding of transcription factors without specific transcription factor terms. Together, these correlation models indicate both that a limited set of chromatin marks are sufficient to 'explain' transcription and that a variety of transcription factors might have broad roles in general transcription levels across many genes. It is important to note that this is an inherently observational study of correlation patterns, and is consistent with a variety of mechanistic models with different causal links between the chromatin, transcription factor and RNA assays. However, it does indicate that there is enough information present at the promoter regions of genes to explain most of the variation in RNA expression.

We developed predictive models similar to those used to model transcriptional activity to explore the relationship between levels of histone modification and inclusion of exons in alternately spliced transcripts. Even accounting for expression level, H3K36me3 has a positive contribution to exon inclusion, whereas H3K79me2 has a negative contribution (H. Tilgner *et al.*, manuscript in preparation). By monitoring the RNA populations in the subcellular fractions of K562 cells, we found that essentially all splicing is co-transcriptional[61], further supporting a link between chromatin structure and splicing.

Transcription-factor-binding site-anchored integration

Transcription-factor-binding sites provide a natural focus around which to explore chromatin properties. Transcription factors are often multifunctional and can bind a variety of genomic loci with different combinations and patterns of chromatin marks and nucleosome organization. Hence, rather than averaging chromatin mark profiles across all binding sites of a transcription factor, we developed a clustering procedure, termed the Clustered Aggregation Tool (CAGT), to identify subsets of binding sites sharing similar but distinct patterns of chromatin mark signal magnitude, shape and hidden directionality[30]. For example, the average profile of the repressive histone mark H3K27me3 over all 55,782 CTCF-binding sites in H1 hESCs shows poor signal enrichment (Fig. 3a). However, after grouping profiles by signal magnitude we found a subset of 9,840 (17.6%) CTCF-binding sites that exhibit significant flanking H3K27me3 signal. Shape and orientation analysis further revealed that the predominant signal profile for H3K27me3 around CTCF peak summits is asymmetric, consistent with a boundary role for some CTCF sites between active and polycomb-silenced domains. Further examples are provided in Supplementary Figs 5 and 6 of section E. For TAF1, predominantly found near TSSs, the asymmetric sites are orientated with the direction of transcription. However, for distal sites, such as those bound by GATA1 and CTCF, we also observed a high proportion of asymmetric histone patterns, although independent of motif directionality. In fact, all transcription-factor-binding data sets in all cell lines show predominantly asymmetric patterns (asymmetry ratio >0.6) for all chromatin marks but not for DNase I signal (Fig. 3b). This indicates that most transcription-factor-bound chromatin events correlate with structured, directional patterns of histone modifications, and that promoter directionality is not the only source of orientation at these sites.

We also examined nucleosome occupancy relative to the symmetry properties of chromatin marks around transcription-factor-binding sites. Around TSSs, there is usually strong asymmetric nucleosome occupancy, often accounting for most of the histone modification signal (for instance, see Supplementary Fig. 4, section E). However, away from TSSs, there is far less concordance. For example, CTCF-binding sites typically show arrays of well-positioned nucleosomes on either side of the peak summit (Supplementary Fig. 1, section E)[62]. Where the flanking chromatin mark signal is high, the signals are often asymmetric, indicating differential marking with histone modifications (Supplementary Figs 2 and 3, section E). Thus, we

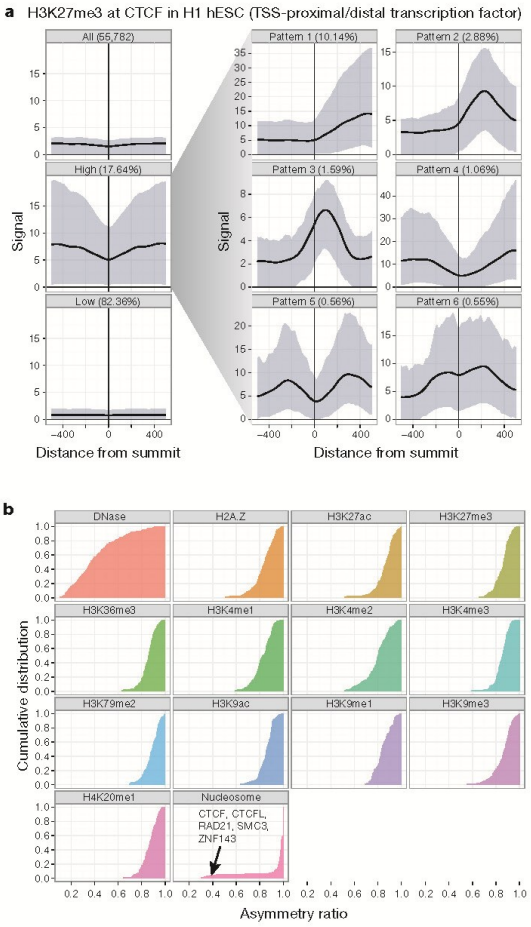


Figure 3 | Patterns and asymmetry of chromatin modification at transcription-factor-binding sites. a, Results of clustered aggregation of H3K27me3 modification signal around CTCF-binding sites (a multifunctional protein involved with chromatin structure). The first three plots (left column) show the signal behaviour of the histone modification over all sites (top) and then split into the high and low signal components. The solid lines show the mean signal distribution by relative position with the blue shaded area delimiting the tenth and ninetieth percentile range. The high signal component is then decomposed further into six different shape classes on the right (see ref. 30 for details). The shape decomposition process is strand aware. **b**, Summary of shape asymmetry for DNase I, nucleosome and histone modification signals by plotting an asymmetry ratio for each signal over all transcription-factor-binding sites. All histone modifications measured in this study show predominantly asymmetric patterns at transcription-factor-binding sites. An interactive version of this figure is available in the online version of the paper.

confirm on a genome-wide scale that transcription factors can form barriers around which nucleosomes and histone modifications are arranged in a variety of configurations[62–65]. This is explored in further detail in refs 25, 26 and 30.

Transcription factor co-associations

Transcription-factor-binding regions are nonrandomly distributed across the genome, with respect to both other features (for example, promoters) and other transcription-factor-binding regions. Within the

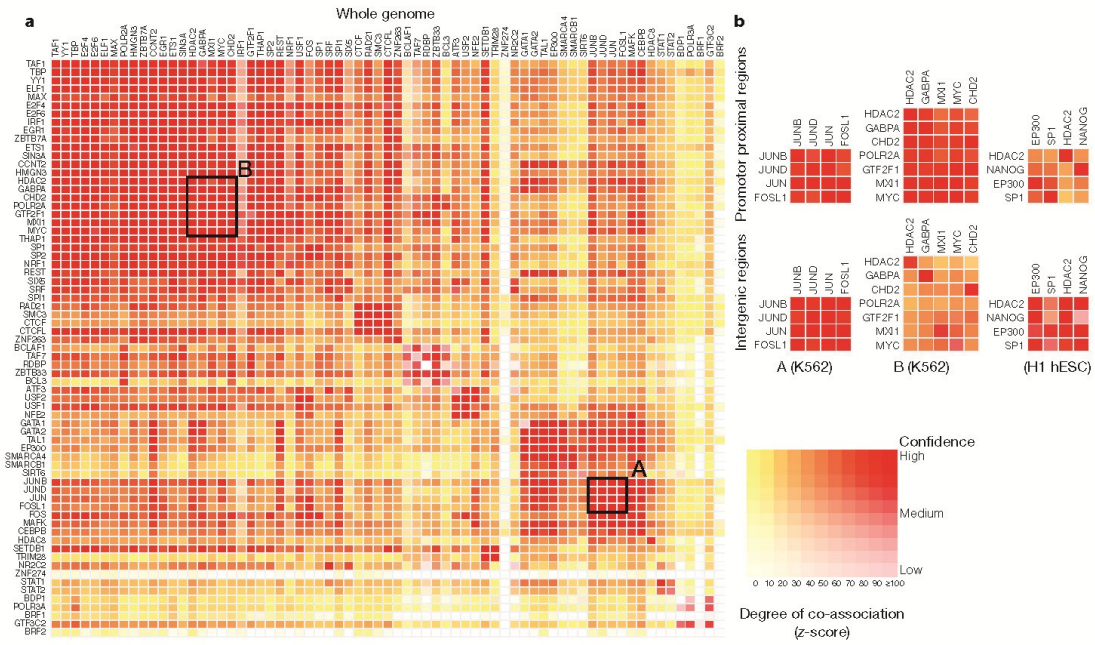


Figure 4 | Co-association between transcription factors. **a**, Significant co-associations of transcription factor pairs using the GSC statistic across the entire genome in K562 cells. The colour strength represents the extent of association (from red (strongest), orange, to yellow (weakest)), whereas the depth of colour represents the fit to the GSC[20] model (where white indicates that the statistical model is not appropriate) as indicated by the key. Most transcription factors have a nonrandom association to other transcription factors, and these associations are dependent on the genomic context, meaning that once the genome is separated into promoter proximal and distal regions, the overall levels of co-association decrease, but more specific relationships are uncovered. **b**, Three classes of behaviour are shown. The first column shows a set of associations for which strength is independent of location in promoter and distal regions, whereas the second column shows a set of transcription factors that have stronger associations in promoter-proximal regions. Both of these examples are from data in K562 cells and are highlighted on the genome-wide co-association matrix (**a**) by the labelled boxes A and B, respectively. The third column shows a set of transcription factors that show stronger association in distal regions (in the H1 hESC line). An interactive version of this figure is available in the online version of the paper.

tier 1 and 2 cell lines, we found 3,307 pairs of statistically co-associated factors ($P < 1 \times 10^{-16}$, GSC) involving 114 out of a possible 117 factors (97%) (Fig. 4a). These include expected associations, such as Jun and Fos, and some less expected novel associations, such as TCF7L2 with HNF4-α and FOXA2 (ref. 66; a full listing is given in Supplementary Table 1, section F). When one considers promoter and intergenic

Table 3 | Summary of the combined state types

Label	Description	Details*	Colour
CTCF	CTCF-enriched element	Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3; CTCF is known to recruit the cohesin complex.	Turquoise
E	Predicted enhancer	Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act at enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be *cis*-regulatory regions. Enriched for sites for the proteins encoded by *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A)− fraction.	Orange
PF	Predicted promoter flanking region	Regions that generally surround TSS segments (see below).	Light red
R	Predicted repressed or low-activity region	This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by *REST* and some other factors (for example, proteins encoded by *BRF2*, *CEBPB*, *MAFK*, *TRIM28*, *ZNF274* and *SETDB1* genes in K562 cells).	Grey
TSS	Predicted promoter region including TSS	Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments.	Bright red
T	Predicted transcribed region	Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A)$^+$ RNA, especially cytoplasmic.	Dark green
WE	Predicted weak enhancer or open chromatin *cis*-regulatory element	Similar to the E state, but weaker signals and weaker enrichments.	Yellow

*Where specific enrichments or overlaps are identified, these are derived from analysis in GM12878 and/or K562 cells where the data for comparison is richest. The colours indicated are used in Figs 5 and 7 and in display of these tracks from the ENCODE data hub.

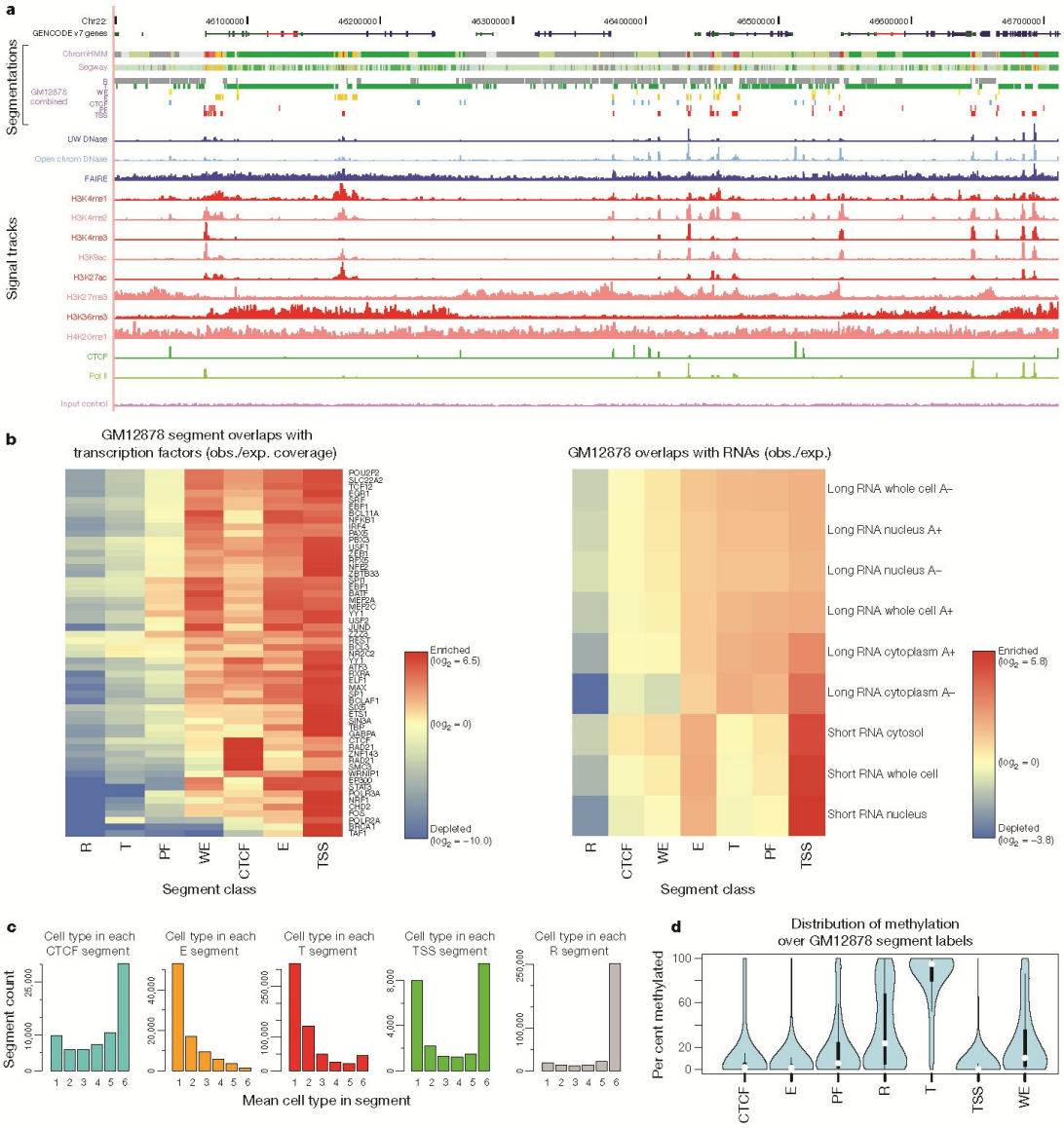


Figure 5 | Integration of ENCODE data by genome-wide segmentation. **a**, Illustrative region with the two segmentation methods (ChromHMM and Segway) in a dense view and the combined segmentation expanded to show each state in GM12878 cells, beneath a compressed view of the GENCODE gene annotations. Note that at this level of zoom and genome browser resolution, some segments appear to overlap although they do not. Segmentation classes are named and coloured according to the scheme in Table 3. Beneath the segmentations are shown each of the normalized signals that were used as the input data for the segmentations. Open chromatin signals from DNase-seq from the University of Washington group (UW DNase) or the ENCODE open chromatin group (Openchrom DNase) and FAIRE assays are shown in blue; signal from histone modification ChIP-seq in red; and transcription factor ChIP-seq signal for Pol II and CTCF in green. The mauve ChIP-seq control signal (input control) at the bottom was also included as an input to the segmentation. **b**, Association of selected transcription factor (left) and RNA (right) elements in the combined segmentation states (*x* axis) expressed as an observed/expected ratio (obs./exp.) for each combination of transcription factor or RNA element and segmentation class using the heat-map scale shown in the key besides each heat map. **c**, Variability of states between cell lines, showing the distribution of occurrences of the state in the six cell lines at specific genome locations: from unique to one cell line to ubiquitous in all six cell lines for five states (CTCF, E, T, TSS and R). **d**, Distribution of methylation level at individual sites from RRBS analysis in GM12878 cells across the different states, showing the expected hypomethylation at TSSs and hypermethylation of genes bodies (T state) and repressed (R) regions.

regions separately, this changes to 3,201 pairs (116 factors, 99%) for promoters and 1,564 pairs (108 factors, 92%) for intergenic regions, with some associations more specific to these genomic contexts (for example, the cluster of HDAC2, GABPA, CHD2, GTF2F1, MXI1 and MYC in promoter regions and SP1, EP300, HDAC2 and NANOG in intergenic regions (Fig. 4b)). These general and context-dependent associations lead to a network representation of the co-binding with many interesting properties, explored in refs 19, 25 and 26. In addition, we also identified a set of regions bound by multiple factors representing high occupancy of transcription factor (HOT) regions[67].

Genome-wide integration

To identify functional regions genome-wide, we next integrated elements independent of genomic landmarks using either discriminative training methods, where a subset of known elements of a particular class were used to train a model that was then used to discover more instances of this class, or using methods in which only data from ENCODE assays were used without explicit knowledge of any annotation.

For discriminative training, we used a three-step process to predict potential enhancers, described in Supplementary Information and ref. 67. Two alternative discriminative models converged on a set of ~13,000 putative enhancers in K562 cells[67]. In the second approach, two methodologically distinct unbiased approaches (see refs 40, 68 and M. M. Hoffman *et al.*, manuscript in preparation) converged on a concordant set of histone modification and chromatin-accessibility patterns that can be used to segment the genome in each of the tier 1 and tier 2 cell lines, although the individual loci in each state in each cell line are different. With the exception of RNA polymerase II and CTCF, the addition of transcription factor data did not substantially alter these patterns. At this stage, we deliberately excluded RNA and methylation assays, reserving these data as a means to validate the segmentations.

Our integration of the two segmentation methods (M. M. Hoffman *et al.*, manuscript in preparation) established a consensus set of seven major classes of genome states, described in Table 3. The standard view of active promoters, with a distinct core promoter region (TSS and PF states), leading to active gene bodies (T, transcribed state), is rediscovered in this model (Fig. 5a, b). There are three 'active' distal states. We tentatively labelled two as enhancers (predicted enhancers, E, and predicted weak enhancers, WE) due to their occurrence in regions of open chromatin with high H3K4me1, although they differ in the levels of marks such as H3K27ac, currently thought to distinguish active from inactive enhancers. The other active state (CTCF) has high CTCF binding and includes sequences that function as insulators in a transfection assay. The remaining repressed state (R) summarizes sequences split between different classes of actively repressed or inactive, quiescent chromatin. We found that the CTCF-binding-associated state is relatively invariant across cell types, with individual regions frequently occupying the CTCF state across all six cell types (Fig. 5c). Conversely, the E and T states have substantial cell-specific behaviour, whereas the TSS state has a bimodal behaviour with similar numbers of cell-invariant and cell-specific occurrences. It is important to note that the consensus summary classes do not capture all the detail discovered in the individual segmentations containing more states.

The distribution of RNA species across segments is quite distinct, indicating that underlying biological activities are captured in the segmentation. Polyadenylated RNA is heavily enriched in gene bodies. Around promoters, there are short RNA species previously identified as promoter-associated short RNAs (Fig. 5b)[16,69]. Similarly, DNA methylation shows marked distinctions between segments, recapitulating the known biology of predominantly unmethylated active promoters (TSS states) followed by methylated gene bodies[42] (T state, Fig. 5d). The two enhancer-enriched states show distinct patterns of DNA methylation, with the less active enhancer state (by H3K27ac/H3K4me1 levels) showing higher methylation. These states also have an excess of RNA elements without poly(A) tails and methyl-cap RNA, as assayed by CAGE sequences, compared to matched intergenic controls, indicating a specific transcriptional mode associated with active enhancers[70]. Transcription factors also showed distinct distributions across the segments (Fig. 5b). A striking pattern is the concentration of transcription factors in the TSS-associated state. The enhancers contain a different set of transcription factors. For example, in K562 cells, the E state is enriched for binding by the proteins encoded by the *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes. We tested a subset of these predicted enhancers in both mouse and fish transgenic models (examples in Fig. 6), with over half of the elements showing activity, often in the corresponding tissue type.

The segmentation provides a linear determination of functional state across the genome, but not an association of particular distal regions with genes. By using the variation of DNase I signal across cell lines, 39% of E (enhancer associated) states could be linked to a proposed regulated gene[29] concordant with physical proximity patterns determined by 5C[49] or ChIA-PET.

To provide a fine-grained regional classification, we turned to a self organizing map (SOM) to cluster genome segmentation regions based on their assay signal characteristics (Fig. 7). The segmentation regions were initially randomly assigned to a 1,350-state map in a two-dimensional toroidal space (Fig. 7a). This map can be visualized as a two-dimensional rectangular plane onto which the various signal distributions can be plotted. For instance, the rectangle at the bottom left of Fig. 7a shows the distribution of the genome in the initial randomized map. The SOM was then trained using the twelve different ChIP-seq and DNase-seq assays in the six cell types previously analysed in the large-scale segmentations (that is, over 72-dimensional space). After training, the SOM clustering was again visualized in two dimensions, now showing the organized distribution of genome segments (lower right of panel, Fig. 7a). Individual data sets associated with the genome segments in each SOM map unit (hexagonal cells) can then be visualized in the same framework to learn how each additional kind of data is distributed on the chromatin state map. Figure 7b shows CAGE/TSS expression data overlaid on the randomly initialized (left) and trained map (right) panels. In this way the trained SOM highlighted cell-type-specific TSS clusters (bottom panels of Fig. 7b), indicating that there are sets of tissue-specific TSSs that are distinguished from each other by subtle combinations of ENCODE

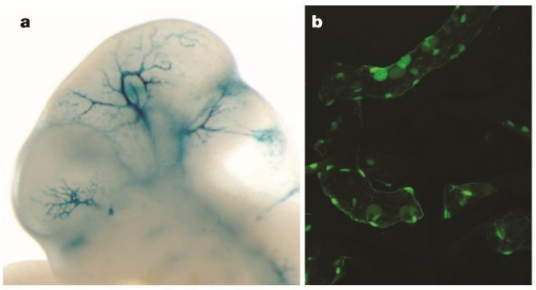


Figure 6 | Experimental characterization of segmentations. Randomly sampled E state segments (see Table 3) from the K562 segmentation were cloned for mouse- and fish-based transgenic enhancer assays. **a**, Representative LacZ-stained transgenic embryonic day (E)11.5 mouse embryo obtained with construct hs2065 (EN167, chr10: 46052882–46055670, GRCh37). Highly reproducible staining in the blood vessels was observed in 9 out of 9 embryos resulting from independent transgenic integration events. **b**, Representative green fluorescent protein reporter transgenic medaka fish obtained from a construct with a basal *hsp70* promoter on meganuclease-based transfection. Reproducible transgenic expression in the circulating nucleated blood cells and the endothelial cell walls was seen in 81 out of 100 transgenic tests of this construct.

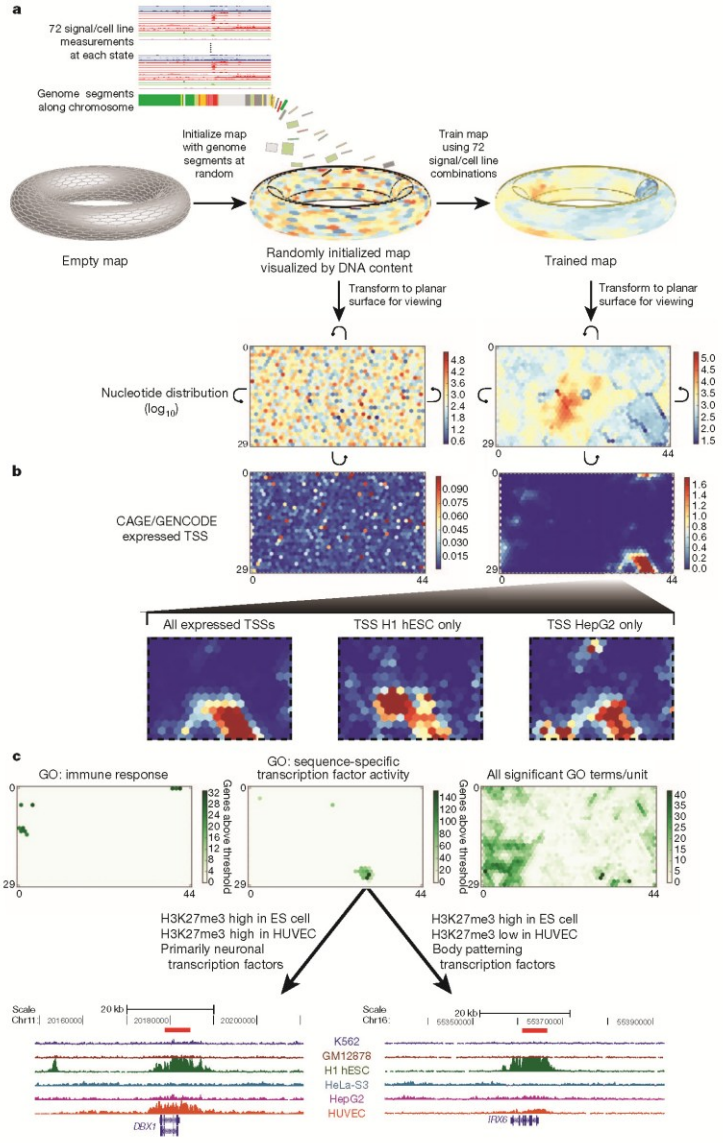


Figure 7 | High-resolution segmentation of ENCODE data by self-organizing maps (SOM). **a–c**, The training of the SOM (**a**) and analysis of the results (**b**, **c**) are shown. Initially we arbitrarily placed genomic segments from the ChromHMM segmentation on to the toroidal map surface, although the SOM does not use the ChromHMM state assignments (**a**). We then trained the map using the signal of the 12 different ChIP-seq and DNase-seq assays in the six cell types analysed. Each unit of the SOM is represented here by a hexagonal cell in a planar two-dimensional view of the toroidal map. Curved arrows indicate that traversing the edges of two dimensional view leads back to the opposite edge. The resulting map can be overlaid with any class of ENCODE or other data to view the distribution of that data within this high-resolution segmentation. In panel **a** the distributions of genome bases across the untrained and trained map (left and right, respectively) are shown using heat-map colours for log_{10} values. **b**, The distribution of TSSs from CAGE experiments of GENCODE annotation on the planar representations of either the initial random organization (left) or the final trained SOM (right) using heat maps coloured according to the accompanying scales. The bottom half of **b** expands the different distributions in the SOM for all expressed TSSs (left) or TSSs specifically expressed in two example cell lines, H1 hESC (centre) and HepG2 (right). **c**, The association of Gene Ontology (GO) terms on the same representation of the same trained SOM. We assigned genes that are within 20 kb of a genomic segment in a SOM unit to that unit, and then associated this set of genes with GO terms using a hypergeometric distribution after correcting for multiple testing. Map units that are significantly associated to GO terms are coloured green, with increasing strength of colour reflecting increasing numbers of genes significantly associated with the GO terms for either immune response (left) or sequence-specific transcription factor activity (centre). In each case, specific SOM units show association with these terms. The right-hand panel shows the distribution on the same SOM of all significantly associated GO terms, now colouring by GO term count per SOM unit. For sequence-specific transcription factor activity, two example genomic regions are extracted at the bottom of panel **c** from neighbouring SOM units. These are regions around the *DBX1* (from SOM unit 26,31, left panel) and *IRX6* (SOM unit 27,30, right panel) genes, respectively, along with their H3K27me3 ChIP-seq signal for each of the tier 1 and 2 cell types. For *DBX1*, representative of a set of primarily neuronal transcription factors associated with unit 26,31, there is a repressive H3K27me3 signal in both H1 hESCs and HUVECs; for *IRX6*, representative of a set of body patterning transcription factors associated with SOM unit 27,30, the repressive mark is restricted largely to the embryonic stem (ES) cell. An interactive version of this figure is available in the online version of the paper.

chromatin data. Many of the ultra-fine-grained state classifications revealed in the SOM are associated with specific gene ontology (GO) terms (right panel of Fig. 7c). For instance, the left panel of Fig. 7c identifies ten SOM map units enriched with genomic regions associated with genes associated with the GO term 'immune response'. The central panel identifies a different set of map units enriched for the GO term 'sequence-specific transcription factor activity'. The two map units most enriched for this GO term, indicated by the darkest green colouring, contain genes with segments that are high in H3K27me3 in H1 hESCs, but that differ in H3K27me3 levels in HUVECs. Gene function analysis with the GO ontology tool (GREAT[71]) reveals that the map unit with high H3K27me3 levels in both cell types is enriched in transcription factor genes with known neuronal functions, whereas the neighbouring map unit is enriched in genes involved in body patterning. The genome browser shots at the bottom of Fig. 7c pick out an example region for each of the two SOM map units illustrating the difference in H3K27me3 signal. Overall, we have 228 distinct GO terms associated with specific segments across

one or more states (A. Mortazavi, personal communication), and can assign over one-third of genes to a GO annotation solely on the basis of its multicellular histone patterns. Thus, the SOM analysis provides a fine-grained map of chromatin data across multiple cell types, which can then be used to relate chromatin structure to other data types at differing levels of resolution (for instance, the large cluster of units containing any active TSS, its subclusters composed of units enriched in TSSs active in only one cell type, or individual map units significantly enriched for specific GO terms).

The classifications presented here are necessarily limited by the assays and cell lines studied, and probably contain a number of heterogeneous classes of elements. Nonetheless, robust classifications can be made, allowing a systematic view of the human genome.

Insights into human genomic variation

We next explored the potential impact of sequence variation on ENCODE functional elements. We examined allele-specific variation using results from the GM12878 cells that are derived from an individual (NA12878) sequenced in the 1000 Genomes project, along with her parents. Because ENCODE assays are predominantly sequence-based, the trio design allows each GM12878 data set to be divided by the specific parental contributions at heterozygous sites, producing aggregate haplotypic signals from multiple genomic sites. We examined 193 ENCODE assays for allele-specific biases using 1,409,992 phased, heterozygous SNPs and 167,096 insertions/deletions (indels) (Fig. 8). Alignment biases towards alleles present in the reference genome sequence were avoided using a sequence specifically tailored to the variants and haplotypes present in NA12878 (a 'personalized genome')[72]. We found instances of preferential binding towards each parental allele. For example, comparison of the results from the POLR2A, H3K79me2 and H3K27me3 assays in the region of *NACC2* (Fig. 8a) shows a strong paternal bias for H3K79me2 and POL2RA and a strong maternal bias for H3K27me3, indicating differential activity for the maternal and paternal alleles.

Figure 8b shows the correlation of selected allele-specific signals across the whole genome. For instance, we found a strong allelic correlation between POL2RA and BCLAF1 binding, as well as negative correlation between H3K79me2 and H3K27me3, both at genes (Fig. 8b, below the diagonal, bottom left) and chromosomal segments (top right). Overall, we found that positive allelic correlations among the 193 ENCODE assays are stronger and more frequent than negative correlations. This may be due to preferential capture of accessible alleles and/or the specific histone modification and transcription factor, assays used in the project.

Rare variants, individual genomes and somatic variants

We further investigated the potential functional effects of individual variation in the context of ENCODE annotations. We divided NA12878 variants into common and rare classes, and partitioned these into those overlapping ENCODE annotation (Fig. 9a and Supplementary Tables 1 and 2, section K). We also predicted potential functional effects: for protein-coding genes, these are either non-synonymous SNPs or variants likely to induce loss of function by frame-shift, premature stop, or splice-site disruption; for other regions, these are variants that overlap a transcription-factor-binding site. We found similar numbers of potentially functional variants affecting protein-coding genes or affecting other ENCODE annotations, indicating that many functional variants within individual genomes lie outside exons of protein-coding genes. A more detailed analysis of regulatory variant annotation is described in ref. 73.

To study further the potential effects of NA12878 genome variants on transcription-factor-binding regions, we performed peak calling using a constructed personal diploid genome sequence for NA12878 (ref. 72). We aligned ChIP-seq sequences from GM12878 separately against the maternal and paternal haplotypes. As expected, a greater

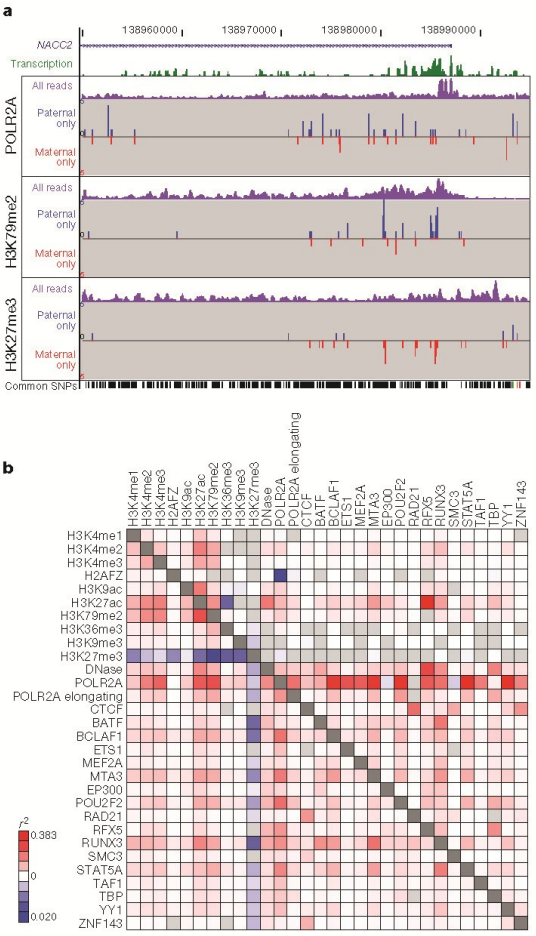


Figure 8 | Allele-specific ENCODE elements. a, Representative allele-specific information from GM12878 cells for selected assays around the first exon of the *NACC2* gene (genomic region Chr9: 138950000–138995000, GRCh37). Transcription signal is shown in green, and the three sections show allele-specific data for three data sets (POLR2A, H3K79me2 and H3K27me3 ChIP-seq). In each case the purple signal is the processed signal for all sequence reads for the assay, whereas the blue and red signals show sequence reads specifically assigned to either the paternal or maternal copies of the genome, respectively. The set of common SNPs from dbSNP, including the phased, heterozygous SNPs used to provide the assignment, are shown at the bottom of the panel. *NACC2* has a statistically significant paternal bias for POLR2A and the transcription-associated mark H3K79me2, and has a significant maternal bias for the repressive mark H3K27me3. **b**, Pair-wise correlations of allele-specific signal within single genes (below the diagonal) or within individual ChromHMM segments across the whole genome for selected DNase-seq and histone modification and transcription factor ChIP-seq assays. The extent of correlation is coloured according to the heat-map scale indicated from positive correlation (red) through to anti-correlation (blue). An interactive version of this figure is available in the online version of the paper.

fraction of reads were aligned than to the reference genome (see Supplementary Information, Supplementary Fig. 1, section K). On average, approximately 1% of transcription-factor-binding sites in GM12878 cells are detected in a haplotype-specific fashion. For instance, Fig. 9b shows a CTCF-binding site not detected using the

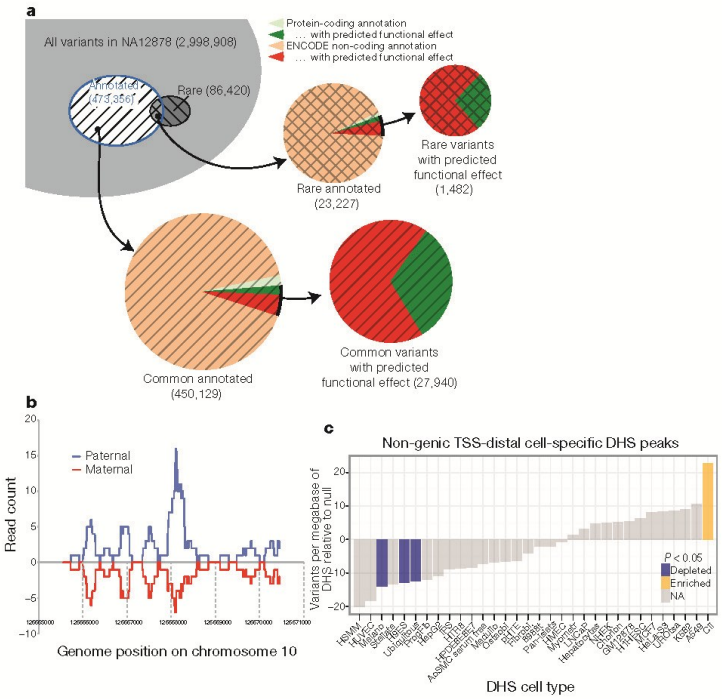


Figure 9 | Examining ENCODE elements on a per individual basis in the normal and cancer genome. **a**, Breakdown of variants in a single genome (NA12878) by both frequency (common or rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project[55])) and by ENCODE annotation, including protein-coding gene and non-coding elements (GENCODE annotations for protein-coding genes, pseudogenes and other ncRNAs, as well as transcription-factor-binding sites from ChIP-seq data sets, excluding broad annotations such as histone modifications, segmentations and RNA-seq). Annotation status is further subdivided by predicted functional effect, being non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations. A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category. **b**, One of several relatively rare occurrences, where alignment to an individual genome sequence (paternal and maternal panels) shows a different readout from the reference genome. In this case, a paternal-haplotype-specific CTCF peak is identified. **c**, Relative level of somatic variants from a whole-genome melanoma sample that occur in DHSs unique to different cell lines. The coloured bars show cases that are significantly enriched or suppressed in somatic mutations. Details of ENCODE cell types can be found at http://encodeproject.org/ENCODE/cellTypes.html. An interactive version of this figure is available in the online version of the paper.

reference sequence that is only present on the paternal haplotype due to a 1-bp deletion (see also Supplementary Fig. 2, section K). As costs of DNA sequencing decrease further, optimized analysis of ENCODE-type data should use the genome sequence of the individual or cell being analysed when possible.

Most analyses of cancer genomes so far have focused on characterizing somatic variants in protein-coding regions. We intersected four available whole-genome cancer data sets with ENCODE annotations (Fig. 9c and Supplementary Fig. 2, section L). Overall, somatic variation is relatively depleted from ENCODE annotated regions, particularly for elements specific to a cell type matching the putative tumour source (for example, skin melanocytes for melanoma). Examining the mutational spectrum of elements in introns for cases where a strand-specific mutation assignment could be made reveals that there are mutational spectrum differences between DHSs and unannotated regions (0.06 Fisher's exact test, Supplementary Fig. 3, section L). The suppression of somatic mutation is consistent with important functional roles of these elements within tumour cells, highlighting a potential alternative set of targets for examination in cancer.

Common variants associated with disease

In recent years, GWAS have greatly extended our knowledge of genetic loci associated with human disease risk and other phenotypes. The output of these studies is a series of SNPs (GWAS SNPs) correlated with a phenotype, although not necessarily the functional variants. Notably, 88% of associated SNPs are either intronic or intergenic[74]. We examined 4,860 SNP–phenotype associations for 4,492 SNPs curated in the National Human Genome Research Institute (NHGRI) GWAS catalogue[74]. We found that 12% of these SNPs overlap transcription-factor-occupied regions whereas 34% overlap DHSs (Fig. 10a). Both figures reflect significant enrichments relative to the overall proportions of 1000 Genomes project SNPs (about 6% and 23%, respectively). Even after accounting for biases introduced by selection of SNPs for the standard genotyping arrays, GWAS SNPs show consistently higher overlap with ENCODE annotations (Fig. 10a, see Supplementary Information). Furthermore, after partitioning the genome by density of different classes of functional elements, GWAS SNPs were consistently enriched beyond all the genotyping SNPs in function-rich partitions, and depleted in function-poor partitions (see Supplementary Fig. 1, section M). GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types (see Supplementary Fig. 2, section M).

Examining the SOM of integrated ENCODE annotations (see above), we found 19 SOM map units showing significant enrichment for GWAS SNPs, including many SOM units previously associated with specific gene functions, such as the immune response regions.

Thus, an appreciable proportion of SNPs identified in initial GWAS scans are either functional or lie within the length of an ENCODE annotation (~500 bp on average) and represent plausible candidates for the functional variant. Expanding the set of feasible functional SNPs to those in reasonable linkage disequilibrium, up to 71% of GWAS SNPs have a potential causative SNP overlapping a DNase I site, and 31% of loci have a candidate SNP that overlaps a binding site occupied by a transcription factor (see also refs 73, 75).

The GWAS catalogue provides a rich functional categorization from the precise phenotypes being studied. These phenotypic categorizations are nonrandomly associated with ENCODE annotations and there is marked correspondence between the phenotype and the identity of the cell type or transcription factor used in the ENCODE assay (Fig. 10b). For example, five SNPs associated with Crohn's disease overlap GATA2-binding sites (P value 0.003 by random permutation or 0.001 by an empirical approach comparing to the GWAS-matched SNPs; see Supplementary Information), and fourteen are located in DHSs found in immunologically relevant cell types. A notable example is a gene desert on chromosome 5p13.1 containing eight SNPs associated with inflammatory diseases. Several are close to or within DHSs in T-helper type 1 (T_H1) and T_H2 cells as well as peaks of binding by transcription factors in HUVECs (Fig. 10c). The latter cell line is not immunological, but factor occupancy detected there could be a proxy for binding of a more relevant factor, such as GATA3, in T cells. Genetic variants in this region also affect expression levels of *PTGER4* (ref. 76), encoding the prostaglandin receptor EP4. Thus, the ENCODE data reinforce the hypothesis that genetic variants in 5p13.1 modulate the expression of flanking genes, and furthermore provide the specific hypothesis that the variants affect occupancy of a GATA factor in an allele-specific manner, thereby influencing susceptibility to Crohn's disease.

Nonrandom association of phenotypes with ENCODE cell types strengthens the argument that at least some of the GWAS lead SNPs are functional or extremely close to functional variants. Each of the associations between a lead SNP and an ENCODE annotation remains a credible hypothesis of a particular functional element

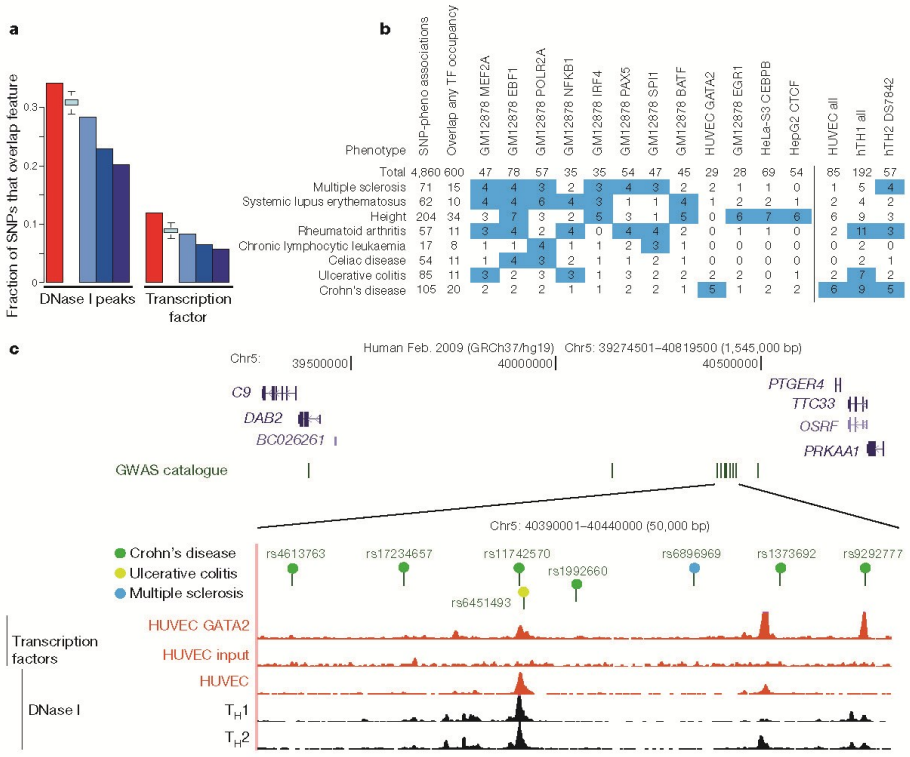


Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data. a, Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at http://main.genome-browser.bx.psu.edu (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. **b**, Aggregate overlap of phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical P-value threshold ≤0.01 (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The P value for the total number of phenotype–transcription factor associations is <0.001. **c**, Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper T_H1 and T_H2 cells. An interactive version of this figure is available in the online version of the paper.

class or cell type to explore with future experiments. Supplementary Tables 1–3, section M, list all 14,885 pairwise associations across the ENCODE annotations. The accompanying papers have a more detailed examination of common variants with other regulatory information[19,25,29,73,75,77].

Concluding remarks

The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome. Our analyses have revealed many novel aspects of gene expression and regulation as well as the organization of such information, as illustrated by the accompanying papers (see http://www.encodeproject.org/ENCODE/pubs.html for collected ENCODE publications). However, there are still many specific details, particularly about the mechanistic processes that generate these elements and how and where they function, that require additional experiments to elucidate.

The large spread of coverage—from our highest resolution, most conservative set of bases implicated in GENCODE protein-coding gene exons (2.9%) or specific protein DNA binding (8.5%) to the broadest, most general set of marks covering the genome (approximately 80%), with many gradations in between—presents a spectrum of elements with different functional properties discovered by ENCODE. A total of 99% of the known bases in the genome are within 1.7 kb of any ENCODE element, whereas 95% of bases are within 8 kb of a bound transcription factor motif or DNase I footprint. Interestingly, even using the most conservative estimates, the fraction of bases likely to be involved in direct gene regulation, even though incomplete, is significantly higher than that ascribed to protein-coding exons (1.2%), raising the possibility that more information in the human genome may be important for gene regulation than for biochemical function. Many of the regulatory elements are not constrained across mammalian evolution, which so far has been one of the most reliable indications of an important biochemical event for the organism. Thus, our data provide orthologous indicators for suggesting possible functional elements.

Importantly, for the first time we have sufficient statistical power to assess the impact of negative selection on primate-specific elements, and all ENCODE classes display evidence of negative selection in these unique-to-primate elements. Furthermore, even with our most conservative estimate of functional elements (8.5% of putative DNA/protein binding regions) and assuming that we have already sampled half of the elements from our transcription factor and cell-type diversity, one would estimate that at a minimum 20% (17% from protein binding and 2.9% protein coding gene exons) of the genome participates in these specific functions, with the likely figure significantly higher.

The broad coverage of ENCODE annotations enhances our understanding of common diseases with a genetic component, rare genetic diseases, and cancer, as shown by our ability to link otherwise anonymous associations to a functional element. ENCODE and similar studies provide a first step towards interpreting the rest of the genome—beyond protein-coding genes—thereby augmenting common disease genetic studies with testable hypotheses. Such information justifies performing whole-genome sequencing (rather than exome only, 1.2% of the genome) on rare diseases and investigating somatic variants in non-coding functional elements, for instance, in cancer. Furthermore, as GWAS analyses typically associate disease to SNPs in large regions, comparison to ENCODE non-coding functional elements can help pinpoint putative causal variants in addition to refinement of location by fine-mapping techniques[78]. Combining ENCODE data with allele-specific information derived from individual genome sequences provides specific insight on the impact of a genetic variant. Indeed, we believe that a significant goal would be to use functional data such as that derived from this project to assign every genomic variant to its possible impact on human phenotypes.

So far, ENCODE has sampled 119 of 1,800 known transcription factors and general components of the transcriptional machinery on a limited number of cell types, and 13 of more than 60 currently known histone or DNA modifications across 147 cell types. DNase I, FAIRE and extensive RNA assays across subcellular fractionations have been undertaken on many cell types, but overall these data reflect a minor fraction of the potential functional information encoded in the human genome. An important future goal will be to enlarge this data set to additional factors, modifications and cell types, complementing the other related projects in this area (for example, Roadmap Epigenomics Project, http://www.roadmapepigenomics.org/, and International Human Epigenome Consortium, http://www.ihec-epigenomes.org/). These projects will constitute foundational resources for human genomics, allowing a deeper interpretation of the organization of gene and regulatory information and the mechanisms of regulation, and thereby provide important insights into human health and disease. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

METHODS SUMMARY

For full details of Methods, see Supplementary Information.

Received 24 November 2011; accepted 29 May 2012.

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
2. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
3. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
5. Chiaromonte, F. *et al.* The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 245–254 (2003).
6. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
7. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324,** 389–392 (2009).
8. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
9. Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. *Genome Res.* **17,** 1245–1253 (2007).
10. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21,** 1769–1776 (2011).
11. Asthana, S. *et al.* Widely distributed noncoding purifying selection in the human genome. *Proc. Natl Acad. Sci. USA* **104,** 12410–12415 (2007).
12. Landt, S. G. *et al.* ChIP-seq guidelines and practices used by the ENCODE and modENCODE consortia. *Genome Res.* http://dx.doi.org/10.1101/gr.136184.111 (2012).
13. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).
14. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* http://dx.doi.org/10.1101/gr.135350.111 (2012).
15. Howald, C. *et al.* Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.* http://dx.doi.org/10.1101/gr.134478.111 (2012).
16. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* http://dx.doi.org/10.1038/nature11233 (this issue).
17. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* http://dx.doi.org/10.1101/gr.132159.111 (2012).
18. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).
19. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* http://dx.doi.org/10.1038/nature11245 (this issue).
20. Bickel, P. J., Boley, N., Brown, J. B., Huang, H. Y. & Zhang, N. R. Subsampling methods for genomic inference. *Ann. Appl. Stat.* **4,** 1660–1697 (2010).
21. Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genet.* **7,** e1001290 (2011).
22. Li, X. Y. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol.* **12,** R34 (2011).

23. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–455 (2011).
24. Zhang, Y. *et al.* Primary sequence and epigenetic determinants of *in vivo* occupancy of genomic DNA by GATA1. *Nucleic Acids Res.* **37,** 7024–7038 (2009).
25. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* http://dx.doi.org/10.1038/nature11212 (this issue).
26. Whitfield, T. W. *et al.* Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13,** R50 (2012).
27. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57,** 159–197 (1988).
28. Urnov, F. D. Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. *J. Cell. Biochem.* **88,** 684–694 (2003).
29. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* http://dx.doi.org/10.1038/nature11232 (this issue).
30. Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.* http://dx.doi.org/10.1101/gr.136366.111 (2012).
31. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. III. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16,** 919–932 (2002).
32. Frietze, S., O'Geen, H., Blahnik, K. R., Jin, V. X. & Farnham, P. J. ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS ONE* **5,** e15082 (2010).
33. Boyle, A. P. *et al.* High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).
34. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6,** 283–289 (2009).
35. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
36. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
37. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
38. Hon, G. C., Hawkins, R. D. & Ren, B. Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.* **18,** R195–R201 (2009).
39. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Rev. Genet.* **12,** 7–18 (2011).
40. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
41. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5,** e1000566 (2009).
42. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnol.* **27,** 361–368 (2009).
43. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
44. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87,** 953–959 (1996).
45. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).
46. Dekker, J. Gene regulation in the third dimension. *Science* **319,** 1793–1794 (2008).
47. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
48. Lajoie, B. R., van Berkum, N. L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nature Methods* **6,** 690–691 (2009).
49. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* http://dx.doi.org/10.1038/nature11279 (this issue).
50. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).
51. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
52. Borneman, A. R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317,** 815–819 (2007).
53. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.* **39,** 730–732 (2007).
54. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
55. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
56. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
57. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biol.* **13,** R49 (2012).
58. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev. Genet.* **8,** 424–436 (2007).
59. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13,** R53 (2012).
60. Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nature Struct. Mol. Biol.* **17,** 1495–1499 (2010).
61. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* http://dx.doi.org/10.1101/gr.134445.111 (2012).
62. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
63. Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16,** 6677–6690 (1988).
64. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
65. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).
66. Frietze, S. *et al.* Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* **13,** R52 (2012).
67. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
68. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9,** 473–476 (2012).
69. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
70. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Struct. Mol. Biol.* **18,** 956–963 (2011).
71. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* **28,** 495–501 (2010).
72. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
73. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* http://dx.doi.org/10.1101/gr.137323.112 (2012).
74. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
75. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* http://dx.doi.org/10.1101/gr.136127.111 (2012).
76. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3,** e58 (2007).
77. Vernot, B. *et al.* Personal and population genomics of human regulatory variation. *Genome Res.* http://dx.doi.org/10.1101/gr.134890.111 (2012).
78. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470,** 264–268 (2011).
79. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* http://dx.doi.org/10.1101/gr.136838.111 (2012).
80. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463,** 943–947 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank additional members of our laboratories and institutions who have contributed to the experimental and analytical components of this project. We thank D. Leja for assistance with production of the figures. The Consortium is funded by grants from the NHGRI as follows: production grants: U54HG004570 (B. E. Bernstein); U01HG004695 (E. Birney); U54HG004563 (G. E. Crawford); U54HG004557 (T. R. Gingeras); U54HG004555 (T. J. Hubbard); U41HG004568 (W. J. Kent); U54HG004576 (R. M. Myers); U54HG004558 (M. Snyder); U54HG004592 (J. A. Stamatoyannopoulos). Pilot grants: R01HG003143 (J. Dekker); RC2HG005591 and R01HG003700 (M. C. Giddings); R01HG004456-03 (Y. Ruan); U01HG004571 (S. A. Tenenbaum); U01HG004561 (Z. Weng); RC2HG005679 (K. P. White). This project was supported in part by American Recovery and Reinvestment Act (ARRA) funds from the NHGRI through grants U54HG004570, U54HG004563, U41HG004568, U54HG004592, R01HG003143, RC2HG005591, R01HG003541, U01HG004561, RC2HG005679 and R01HG003988 (L. Pennacchio). In addition, work from NHGRI Groups was supported by the Intramural Research Program of the NHGRI (L. Elnitski, ZIAHG200323; E. H. Margulies, ZIAHG200341). Research in the Pennachio laboratory was performed at Lawrence Berkeley National Laboratory and at the United States Department of Energy Joint Genome Institute, Department of Energy Contract DE-AC02-05CH11231, University of California.

Author Contributions See the consortium author list for details of author contributions.

Author Information The Supplementary Information is accompanied by a Virtual Machine (VM) containing the functioning analysis data and code. Further details of the VM are available from http://encodeproject.org/ENCODE/integrativeAnalysis/VM. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.B. (birney@ebi.ac.uk).

The ENCODE Project Consortium

Overall coordination (data analysis coordination) Ian Dunham[1], Anshul Kundaje[2]†; **Data production leads (data production)** Shelley F. Aldred[3], Patrick J. Collins[3], Carrie A. Davis[4], Francis Doyle[5], Charles B. Epstein[6], Seth Frietze[7], Jennifer Harrow[8], Rajinder Kaul[9], Jainab Khatun[10], Bryan R. Lajoie[11], Stephen G. Landt[12], Bum-Kyu Lee[13],

Florencia Pauli[14], Kate R. Rosenbloom[15], Peter Sabo[16], Alexias Safi[17], Amartya Sanyal[11], Noam Shoresh[6], Jeremy M. Simon[18], Lingyun Song[17], Nathan D. Trinklein[3]; **Lead analysts (data analysis)** Robert C. Altshuler[19], Ewan Birney[1], James B. Brown[20], Chao Cheng[21], Sarah Djebali[22], Xianjun Dong[23], Ian Dunham[1], Jason Ernst[19]†, Terrence S. Furey[24], Mark Gerstein[21], Belinda Giardine[25], Melissa Greven[23], Ross C. Hardison[25,26], Robert S. Harris[25], Javier Herrero[1], Michael M. Hoffman[16], Sowmya Iyer[27], Manolis Kellis[19], Jainab Khatun[10], Pouya Kheradpour[19], Anshul Kundaje[2]†, Timo Lassmann[28], Qunhua Li[20]†, Xinying Lin[23], Georgi K. Marinov[29], Angelika Merkel[22], Ali Mortazavi[30], Stephen C. J. Parker[31], Timothy E. Reddy[14]†, Joel Rozowsky[21], Felix Schlesinger[4], Robert E. Thurman[16], Jie Wang[23], Lucas D. Ward[19], Troy W. Whitfield[23], Steven P. Wilder[1], Weisheng Wu[25], Hualin S. Xi[32], Kevin Y. Yip[21]†, Jiali Zhuang[23]; **Writing group** Bradley E. Bernstein[6,33], Ewan Birney[1], Ian Dunham[1], Eric D. Green[34], Chris Gunter[14], Michael Snyder[12]; **NHGRI project management (scientific management)** Michael J. Pazin[35], Rebecca F. Lowdon[35]†, Laura A. L. Dillon[35]†, Leslie B. Adams[35], Caroline J. Kelly[35], Julia Zhang[35]†, Judith R. Wexler[35]†, Eric D. Green[34], Peter J. Good[35], Elise A. Feingold[35]; **Principal investigators (steering committee)** Bradley E. Bernstein[6,33], Ewan Birney[1], Gregory E. Crawford[17,36], Job Dekker[11], Laura Elnitski[37], Peggy J. Farnham[7], Mark Gerstein[21], Morgan C. Giddings[10], Thomas R. Gingeras[4,38], Eric D. Green[34], Roderic Guigó[22,39], Ross C. Hardison[25,26], Timothy J. Hubbard[8], Manolis Kellis[19], W. James Kent[15], Jason D. Lieb[18], Elliott H. Margulies[31]†, Richard M. Myers[14], Michael Snyder[12], John A. Stamatoyannopoulos[40], Scott A. Tenenbaum[5], Zhiping Weng[23], Kevin P. White[41], Barbara Wold[29,42]; **Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis)** Jainab Khatun[10], Yanbao Yu[43], John Wrobel[10], Brian A. Risk[10], Harsha P. Gunawardena[43], Heather C. Kuiper[43], Christopher W. Maier[43], Ling Xie[43], Xian Chen[43], Morgan C. Giddings[10]; **Broad Institute Group (data production and analysis)** Bradley E. Bernstein[6,33], Charles B. Epstein[6], Noam Shoresh[6], Jason Ernst[19]†, Pouya Kheradpour[19], Tarjei S. Mikkelsen[6], Shawn Gillespie[33], Alon Goren[6,33], Oren Ram[6,33], Xiaolan Zhang[6], Li Wang[6], Robbyn Issner[6], Michael J. Coyne[6], Timothy Durham[6], Manching Ku[6,33], Thanh Truong[6], Lucas D. Ward[19], Robert C. Altshuler[19], Matthew L Eaton[19], Manolis Kellis[19]; **Cold Spring Harbor, University of Geneva, Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis)** Sarah Djebali[22], Carrie A. Davis[4], Angelika Merkel[22], Alex Dobin[4], Timo Lassmann[28], Ali Mortazavi[30], Andrea Tanzer[22], Julien Lagarde[22], Wei Lin[4], Felix Schlesinger[4], Chenghai Xue[4], Georgi K. Marinov[29], Jainab Khatun[10], Brian A. Williams[29], Chris Zaleski[4], Joel Rozowsky[21], Maik Röder[22], Felix Kokocinski[8]†, Rehab F. Abdelhamid[28], Tyler Alioto[22,44], Igor Antoshechkin[29], Michael T. Baer[4], Philippe Batut[4], Ian Bell[45], Kimberly Bell[4], Sudipto Chakraborty[4], Xian Chen[43], Jacqueline Chrast[46], Joao Curado[22], Thomas Derrien[22]†, Jorg Drenkow[4], Erica Dumais[45], Jackie Dumais[45], Radha Duttagupta[45], Megan Fastuca[4], Kata Fejes-Toth[4]†, Pedro Ferreira[22], Sylvain Foissac[45], Melissa J. Fullwood[47]†, Hui Gao[45], David Gonzalez[22], Assaf Gordon[4], Harsha P. Gunawardena[43], Cédric Howald[46], Sonali Jha[4], Rory Johnson[22], Philipp Kapranov[45], Brandon King[29], Colin Kingswood[22,44], Guoliang Li[48], Oscar J. Luo[47], Eddie Park[30], Jonathan B. Preall[4], Kimberly Presaud[4], Paolo Ribeca[22,44], Brian A. Risk[10], Daniel Robyr[49], Xiaoan Ruan[47], Michael Sammeth[22,44], Kuljeet Singh Sandhu[47], Lorain Schaeffer[29], Lei-Hoon See[4], Atif Shahab[47], Jorgen Skancke[22], Ana Maria Suzuki[28], Hazuki Takahashi[28], Hagen Tilgner[22]†, Diane Trout[29], Nathalie Walters[46], Huaien Wang[4], John Wrobel[10], Yanbao Yu[43], Yoshihide Hayashizaki[28], Jennifer Harrow[8], Mark Gerstein[21], Timothy J. Hubbard[8], Alexandre Reymond[46], Stylianos E. Antonarakis[49], Gregory J. Hannon[4], Morgan C. Giddings[10], Yijun Ruan[47], Barbara Wold[29,42], Piero Carninci[28], Roderic Guigó[22,39], Thomas R. Gingeras[4,38]; **Data coordination center at UC Santa Cruz (production data coordination)** Kate R. Rosenbloom[15], Cricket A. Sloan[15], Katrina Learned[15], Venkat S. Malladi[15], Matthew C. Wong[15], Galt P. Barber[15], Melissa S. Cline[15], Timothy R. Dreszer[15], Steven G. Heitner[15], Donna Karolchik[15], W. James Kent[15], Vanessa M. Kirkup[15], Laurence R. Meyer[15], Jeffrey C. Long[15], Morgan Maddren[15], Brian J. Raney[15]; **Duke University, EBI, University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis)** Terrence S. Furey[24], Lingyun Song[17], Linda L. Grasfeder[18], Paul G. Giresi[18], Bum-Kyu Lee[13], Anna Battenhouse[13], Nathan C. Sheffield[17], Jeremy M. Simon[18], Kimberly A. Showers[18], Alexias Safi[17], Darin London[17], Akshay A. Bhinge[13], Christopher Shestak[18], Matthew R. Schaner[18], Seul Ki Kim[18], Zhuzhu Z. Zhang[18], Piotr A. Mieczkowski[50], Joanna O. Mieczkowska[18], Zheng Liu[13], Ryan M. McDaniell[13], Yunyun Ni[13], Naim U. Rashid[51], Min Jae Kim[18], Sheera Adar[18], Zhancheng Zhang[24], Tianyuan Wang[17], Deborah Winter[17], Damian Keefe[1], Ewan Birney[1], Vishwanath R. Iyer[13], Jason D. Lieb[18], Gregory E. Crawford[17,36]; **Genome Institute of Singapore group (data production and analysis)** Guoliang Li[48], Kuljeet Singh Sandhu[47], Meizhen Zheng[47], Ping Wang[47], Oscar J. Luo[47], Atif Shahab[47], Melissa J. Fullwood[47]†, Xiaoan Ruan[47], Yijun Ruan[47]; **HudsonAlpha Institute, Caltech, UC Irvine, Stanford group (data production and analysis)** Richard M. Myers[14], Florencia Pauli[14], Brian A. Williams[29], Jason Gertz[14], Georgi K. Marinov[29], Timothy E. Reddy[14]†, Jost Vielmetter[29,42], E. Christopher Partridge[14], Diane Trout[29], Katherine E. Varley[14], Clarke Gasper[29,42], Anita Bansal[14], Shirley Pepke[29,52], Preti Jain[14], Henry Amrhein[29], Kevin M. Bowling[14], Michael Anaya[29,42], Marie K. Cross[14], Brandon King[29], Michael A. Muratet[14], Igor Antoshechkin[29], Kimberly M. Newberry[14], Kenneth McCue[29], Amy S. Nesmith[14], Katherine I. Fisher-Aylor[29,42], Barbara Pusey[14], Gilberto DeSalvo[29,42], Stephanie L. Parker[14]†, Sreeram Balasubramanian[29,42], Nicholas S. Davis[14], Sarah K. Meadows[14], Tracy Eggleston[14], Chris Gunter[14], J. Scott Newberry[14], Shawn E. Levy[14], Devin M. Absher[14], Ali Mortazavi[30], Wing H. Wong[53], Barbara Wold[29,42]; **Lawrence Berkeley National Laboratory group (targeted experimental validation)** Matthew J. Blow[54], Axel Visel[54,55], Len A. Pennacchio[54,55]; **NHGRI groups (data production and analysis)** Laura Elnitski[37], Elliott H. Margulies[31]†, Stephen C. J. Parker[31], Hanna M. Petrykowska[37]; **Sanger Institute, Washington University, Yale University, Center for Genomic Regulation, Barcelona, UCSC, MIT, University of Lausanne, CNIO group (data production and analysis)** Alexej Abyzov[21], Bronwen Aken[8], Daniel Barrell[8], Gemma Barson[8], Andrew Berry[8], Alexandra Bignell[8], Veronika Boychenko[8], Giovanni Bussotti[22], Jacqueline Chrast[46], Claire Davidson[8], Thomas Derrien[22]†, Gloria Despacio-Reyes[8], Mark Diekhans[15], Iakes Ezkurdia[56], Adam Frankish[8], James Gilbert[8], Jose Manuel Gonzalez[8], Ed Griffiths[8], Rachel Harte[15], David A. Hendrix[19], Cédric Howald[46], Toby Hunt[8], Irwin Jungreis[19], Mike Kay[8], Ekta Khurana[21], Felix Kokocinski[8]†, Jing Leng[21], Michael F. Lin[19], Jane Loveland[8], Zhi Lu[57], Deepa Manthravadi[8], Marco Mariotti[22], Jonathan Mudge[8], Gaurab Mukherjee[8], Cedric Notredame[22], Baikang Pei[21], Jose Manuel Rodriguez[56], Gary Saunders[8], Andrea Sboner[58], Stephen Searle[8], Cristina Sisu[21], Catherine Snow[8], Charlie Steward[8], Andrea Tanzer[22], Electra Tapanari[8], Michael L. Tress[56], Marijke J. van Baren[59]†, Nathalie Walters[46], Stefan Washietl[19], Laurens Wilming[8], Amonida Zadissa[8], Zhengdong Zhang[60], Michael Brent[59], David Haussler[61], Manolis Kellis[19], Alfonso Valencia[56], Mark Gerstein[21], Alexandre Reymond[46], Roderic Guigó[22,39], Jennifer Harrow[8], Timothy J. Hubbard[8]; **Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UC Davis group (data production and analysis)** Stephen G. Landt[12], Seth Frietze[7], Alexej Abyzov[21], Nick Addleman[12], Roger P. Alexander[21], Raymond K. Auerbach[21], Suganthi Balasubramanian[21], Keith Bettinger[12], Nitin Bhardwaj[21], Alan P. Boyle[12], Alina R. Cao[62], Philip Cayting[12], Alexandra Charos[63], Yong Cheng[12], Chao Cheng[21], Catharine Eastman[12], Ghia Euskirchen[12], Joseph D. Fleming[64], Fabian Grubert[12], Lukas Habegger[21], Manoj Hariharan[12], Arif Harmanci[21], Sushma Iyengar[65], Victor X. Jin[66], Konrad J. Karczewski[12], Maya Kasowski[12], Phil Lacroute[12], Hugo Lam[12], Nathan Lamarre-Vincent[64], Jing Leng[21], Jin Lian[67], Marianne Lindahl-Allen[64], Renqiang Min[21]†, Benoit Miotto[64], Hannah Monahan[63], Zarmik Moqtaderi[64], Xinmeng J. Mu[21], Henriette O'Geen[62], Zhengqing Ouyang[12], Dorrelyn Patacsil[12], Baikang Pei[21], Debasish Raha[63], Lucia Ramirez[12], Brian Reed[63], Joel Rozowsky[21], Andrea Sboner[58], Minyi Shi[12], Cristina Sisu[21], Teri Slifer[12], Heather Witt[7], Linfeng Wu[12], Xiaoqin Xu[62], Koon-Kiu Yan[21], Xinqiong Yang[12], Kevin Y. Yip[21]†, Zhengdong Zhang[60], Kevin Struhl[64], Sherman M. Weissman[67], Mark Gerstein[21], Peggy J. Farnham[7], Michael Snyder[12]; **University of Albany SUNY group (data production and analysis)** Scott A. Tenenbaum[5], Luiz O. Penalva[68], Francis Doyle[5]; **University of Chicago, Stanford group (data production and analysis)** Subhradip Karmakar[41], Stephen G. Landt[12], Raj R. Bhanvadia[41], Alina Choudhury[41], Marc Domanus[41], Lijia Ma[41], Jennifer Moran[41], Dorrelyn Patacsil[12], Teri Slifer[12], Alec Victorsen[41], Xinqiong Yang[12], Michael Snyder[12], Kevin P. White[41]; **University of Heidelberg group (targeted experimental validation)** Thomas Auer[69]†, Lazaro Centanin[69], Michael Eichenlaub[69], Franziska Gruhl[69], Stephan Heermann[69], Burkhard Hoeckendorf[69], Daigo Inoue[69], Tanja Kellner[69], Stephan Kirchmaier[69], Claudia Mueller[69], Robert Reinhardt[69], Lea Schertel[69], Stephanie Schneider[69], Rebecca Sinn[69], Beate Wittbrodt[69], Jochen Wittbrodt[69]; **University of Massachusetts Medical School Bioinformatics group (data production and analysis)** Zhiping Weng[23], Troy W. Whitfield[23], Jie Wang[23], Patrick J. Collins[3], Shelley F. Aldred[3], Nathan D. Trinklein[3], E. Christopher Partridge[14], Richard M. Myers[14]; **University of Massachusetts Medical School Genome Folding group (data production and analysis)** Job Dekker[11], Gaurav Jain[11], Bryan R. Lajoie[11], Amartya Sanyal[11]; **University of Washington, University of Massachusetts Medical Center group (data production and analysis)** Gayathri Balasundaram[70], Daniel L. Bates[16], Rachel Byron[70], Theresa K. Canfield[16], Morgan J. Diegel[16], Douglas Dunn[16], Abigail K. Ebersol[71], Tristan Frum[71], Kavita Garg[72], Erica Gist[16], R. Scott Hansen[71], Lisa Boatman[71], Eric Haugen[16], Richard Humbert[16], Gaurav Jain[11], Audra K. Johnson[16], Ericka M. Johnson[71], Tattyana V. Kutyavin[16], Bryan R. Lajoie[11], Kristen Lee[16], Dimitra Lotakis[71], Matthew T. Maurano[16], Shane J. Neph[16], Fiedencio V. Neri[16], Eric D. Nguyen[71], Hongzhu Qu[16], Alex P. Reynolds[16], Vaughn Roach[16], Eric Rynes[16], Peter Sabo[16], Minerva E. Sanchez[71], Richard S. Sandstrom[16], Amartya Sanyal[11], Anthony O. Shafer[16], Andrew B. Stergachis[16], Sean Thomas[16], Robert E. Thurman[16], Benjamin Vernot[16], Jeff Vierstra[16], Shinny Vong[16], Hao Wang[16], Molly A. Weaver[16], Yongqi Yan[71], Miaohua Zhang[70], Joshua M. Akey[16], Michael Bender[70], Michael O. Dorschner[73], Mark Groudine[70], Michael J. MacCoss[16], Patrick Navas[71], George Stamatoyannopoulos[71], Rajinder Kaul[9], Job Dekker[11], John A. Stamatoyannopoulos[40]; **Data Analysis Center (data analysis)**; Ian Dunham[1], Kathryn Beal[1], Alvis Brazma[74], Paul Flicek[1], Javier Herrero[1], Nathan Johnson[1], Damian Keefe[1], Margus Lukk[74]†, Nicholas M. Luscombe[75], Daniel Sobral[1]†, Juan M. Vaquerizas[75], Steven P. Wilder[1], Serafim Batzoglou[2], Arend Sidow[76], Nadine Hussami[2], Sofia Kyriazopoulou-Panagiotopoulou[2], Max W. Libbrecht[2]†, Marc A. Schaub[2], Anshul Kundaje[2]†, Ross C. Hardison[25,26], Webb Miller[25], Belinda Giardine[25], Robert S. Harris[25], Weisheng Wu[25], Peter J. Bickel[20], Balazs Banfai[20], Nathan P. Boley[20], James B. Brown[20], Haiyan Huang[20], Qunhua Li[20]†, Jingyi Jessica Li[20], William Stafford Noble[16,77], Jeffrey A. Bilmes[78], Orion J. Buske[16], Michael M. Hoffman[16], Avinash D. Sahu[16]†, Peter V. Kharchenko[79], Peter J. Park[79], Dannon Baker[80], James Taylor[80], Zhiping Weng[23], Sowmya Iyer[27], Xianjun Dong[23], Melissa Greven[23], Xinying Lin[23], Jie Wang[23], Hualin S. Xi[32], Jiali Zhuang[23], Mark Gerstein[21], Roger P. Alexander[21], Suganthi Balasubramanian[21], Chao Cheng[21], Arif Harmanci[21], Lucas Lochovsky[21], Renqiang Min[21]†, Xinmeng J. Mu[21], Joel Rozowsky[21], Koon-Kiu Yan[21], Kevin Y. Yip[21]† & Ewan Birney[1]

[1]Vertebrate Genomics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [2]Department of Computer Science, Stanford University, 318 Campus Drive, Stanford, California 94305-5428, USA. [3]SwitchGear Genomics, 1455 Adams Drive Suite 1317, Menlo Park, California 94025, USA. [4]Functional Genomics, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [5]College of Nanoscale Sciences and Engineering, University ay Albany-SUNY, 257 Fuller Road, NFE 4405, Albany, New York 12203, USA. [6]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [7]Biochemistry and Molecular Biology, USC/Norris Comprehensive Cancer Center, 1450 Biggy Street, NRT 6503, Los Angeles, California 90089, USA. [8]Informatics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [9]Department of Medicine, Division of Medical Genetics, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195, USA. [10]College of Arts and Sciences, Boise State University, 1910 University Drive, Boise, Idaho 83725, USA. [11]Program in Systems Biology, Program in Gene Function and Expression, Department of Biochemistry and Molecular

Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [12]Department of Genetics, Stanford University, 300 Pasteur Drive, M-344, Stanford, California 94305-5120, USA. [13]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, The University of Texas at Austin, 1 University Station A4800, Austin, Texas 78712, USA. [14]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA. [15]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [16]Department of Genome Sciences, University of Washington, 3720 15th Ave NE, Seattle, Washington 98195-5065, USA. [17]Institute for Genome Sciences and Policy, Duke University, 101 Science Drive, Durham, North Carolina 27708, USA. [18]Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, 408 Fordham Hall, Chapel Hill, North Carolina 27599-3280, USA. [19]Computer Science and Artificial Intelligence Laboratory, Broad Institute of MIT and Harvard, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA. [20]Department of Statistics, University of California, Berkeley, 367 Evans Hall, University of California, Berkeley, Berkeley, California 94720, USA. [21]Computational Biology and Bioinformatics Program, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [22]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Doctor Aiguader, 88, Barcelona 08003, Catalonia, Spain. [23]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [24]Department of Genetics, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, CB 7240, Chapel Hill, North Carolina 27599, USA. [25]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, Wartik Laboratory, University Park, Pennsylvania 16802, USA. [26]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 304 Wartik Laboratory, University Park, Pennsylvania 16802, USA. [27]Program in Bioinformatics, Boston University, 24 Cummington Street, Boston, Massachusetts 02215, USA. [28]RIKEN Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [29]Division of Biology, California Institute of Technology, 156-291200 East California Boulevard, Pasadena, California 91125, USA. [30]Developmental and Cell Biology and Center for Complex Biological Systems, University of California Irvine, 2218 Biological Sciences III, Irvine, California 92697-2300, USA. [31]Genome Technology Branch, National Human Genome Research Institute, 5625 Fishers Lane, Bethesda, Maryland 20892, USA. [32]Department of Biochemistry and Molecular Pharmacology, Bioinformatics Core, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [33]Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, 185 Cambridge St CPZN 8400, Boston, Massachusetts 02114, USA. [34]National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, Building 31, Room 4B09, Bethesda, Maryland 20892-2152, USA. [35]National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane, Bethesda, Maryland 20892-9307, USA. [36]Department of Pediatrics, Division of Medical Genetics, Duke University School of Medicine, Durham, North Carolina 27710, USA. [37]National Human Genome Research Institute, National Institutes of Health, 5625 Fishers Lane, Rockville, Maryland 20892, USA. [38]Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [39]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia 08002, Spain. [40]Department of Genome Sciences, Box 355065, and Department of Medicine, Division of Oncology, Box 358081, University of Washington, Seattle, Washington 98195-5065, USA. [41]Institute for Genomics and Systems Biology, The University of Chicago, 900 East 57th Street, 10100 KCBD, Chicago, Illinois 60637, USA. [42]Beckman Institute, California Institute of Technology, 156-29 1200 E. California Boulevard, Pasadena, California 91125, USA. [43]Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Campus Box 7260, 120 Mason Farm Road, 3010 Genetic Medicine Building, Chapel Hill, North Carolina 27599, USA. [44]Centro Nacional de Análisis Genómico (CNAG), C/Baldiri Reixac 4, Torre I, Barcelona, Catalonia 08028, Spain. [45]Genomics, Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [46]Center for Integrative Genomics, University of Lausanne, Genopode Building, 1015 Lausanne, Switzerland. [47]Genome Technology and Biology, Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [48]Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [49]Department of Genetic Medicine and Development, University of Geneva Medical School, and University Hospitals of Geneva, 1 rue Michel-Servet, 1211 Geneva 4, Switzerland. [50]Department of Genetics, The University of North Carolina at Chapel Hill, 5078 GMB, Chapel Hill, North Carolina 27599-7264, USA. [51]Department of Biostatistics, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, 408 Fordham Hall, Chapel Hill, North Carolina 27599-7445, USA. [52]Center for Advanced Computing Research, California Institute of Technology, MC 158-79, 1200 East California Boulavard, Pasadena, California 91125, USA. [53]Department of Statistics, Stanford University, Sequoia Hall. 390 Serra Mall, Stanford, California 94305-4065, USA. [54]DOE Joint Genome Institute, Walnut Creek, California, USA. [55]Genomics Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, MS 84-171, Berkeley, California 94720, USA. [56]Structural Computational Biology, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, 3, 28029 Madrid, Spain. [57]School of Life Sciences, Tsinghua University, School of Life Sciences, Tsinghua University, 100084 Beijing, China. [58]Department of Pathology and Laboratory Medicine, Institute for Computational Biomedicine, Weill Cornell Medical College, 1305 York Avenue, Box 140, New York, New York 10065, USA. [59]Computer Science and Engineering, Washington University in St Louis, St Louis, Missouri 63130, USA. [60]Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Room 353A, Bronx, New York 10461, USA. [61]Center for Biomolecular Science and Engineering, Howard Hughes Medical Institute, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [62]Genome Center, University of California-Davis, 451 Health Sciences Drive, Davis, California 95616, USA. [63]Department of Molecular, Cellular, and Developmental Biology, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06511, USA. [64]Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [65]Biochemistry and Molecular Biology, University of Southern California, 1501 San Pablo Street, Los Angeles, California 90089, USA. [66]Department of Biomedical Informatics, Ohio State University, 3172C Graves Hall, 333 W Tenth Avenue, Columbus, Ohio 43210, USA. [67]Department of Genetics, Yale University, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [68]Department of Cellular and Structural Biology, Children's Cancer Research Institute–UTHSCSA, Mail code 7784-7703 Floyd Curl Dr, San Antonio, Texas 78229, USA. [69]Centre for Organismal Studies (COS) Heidelberg, University of Heidelberg, Im Neuenheimer Feld 230, 69120 Heidelberg, Germany. [70]Basic Sciences Division, Fred Hutchinson Cancer Research Center, 825 Eastlake Avenue East, Seattle, Washington 98109, USA. [71]Department of Medicine, Division of Medical Genetics, Box 357720, University of Washington, Seattle, Washington 98195-7720, USA. [72]Division of Human Biology, Fred Hutchinson Cancer Research Center, 825 Eastlake Avenue East, Seattle, Washington 98109, USA. [73]Department of Psychiatry and Behavioral Sciences, Box 356560, University of Washington, Seattle, Washington 98195-6560, USA. [74]Microarray Informatics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [75]Genomics and Regulatory Systems Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [76]Department of Pathology, Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [77]Department of Computer Science and Engineering, 185 Stevens Way, Seattle, Washington 98195, USA. [78]Department of Electrical Engineering, University of Washington, 185 Stevens Way, Seattle, Washington 98195, USA. [79]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, Massachusetts 02115, USA. [80]Departments of Biology and Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322, USA. †Present addresses: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA (A.K.); UCLA Biological Chemistry Department, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, Jonsson Comprehensive Cancer Center, 615 Charles E Young Dr South, Los Angeles, California 90095, USA (J.E.); Department of Statistics, 514D Wartik Lab, Penn State University, State College, Pennsylvania 16802, USA (Q.L.); Department of Biostatistics and Bioinformatics and the Institute for Genome Sciences and Policy, Duke University School of Medicine, 101 Science Drive, Durham, North Carolina 27708, USA (T.E.R.); Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (K.Y.Y.); Department of Genetics, Washington University in St Louis, St Louis, Missouri 63110, USA (R.F.L.); Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA (L.A.L.D.); National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA (J.Z.); University of California, Davis Population Biology Graduate Group, Davis, California 95616, USA (J.R.W.); Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL, UK (E.H.M.); BlueGnome Ltd., CPC4, Capital Park, Fulbourn, Cambridge CB21 5XE, UK (F.K.); Institut de Génétique et Développement de Rennes, CNRS-UMR6061, Université de Rennes 1, F-35000 Rennes, Brittany, France (T.D.); Caltech, 1200 East California Boulevard, Pasadena, California 91125, USA (K.F.-T.); A*STAR-Duke-NUS Neuroscience Research Partnership, 8 College Road, Singapore 169857, Singapore (M.J.F.); St Laurent Institute, One Kendall Square, Cambridge, Massachusetts 02139, USA (P.K.); Department of Genetics, Stanford University, Stanford, California 94305, USA (H.T.); Biomedical Sciences (BMS) Graduate Program, University of California, San Francisco, 513 Parnassus Avenue, HSE-1285, San Francisco, California 94143-0505, USA (S.L.P.); Monterey Bay Aquarium Research Institute, Moss Landing, California 95039, USA (M.J.v.B.); Department of Machine Learning, NEC Laboratories America, 4 Independence Way, Princeton, New Jersey 08540, USA (R.M.); Neuronal Circuit Development Group, Unité de Génétique et Biologie du Développement, U934/UMR3215, Institut Curie–Centre de Recherche, Pole de Biologie du Développement et Cancer, 26, rue d'Ulm, 75248 Paris Cedex 05, France (T.A.); Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK (M.L.); Unidade de Bioinformatica, Rua da Quinta Grande, 6, P-2780-156 Oeiras, Portugal (D.S.); Department of Genome Sciences, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195-5065, USA (M.W.L.); Center for Bioinformatics and Computational Biology, 3115 Ag/Life Surge Building 296, University of Maryland, College Park, Maryland 20742, USA (A.D.S.).

REFERENCES

1. Li, X.Y., et al., *Evolutionary variation of the CCAAT-binding transcription factor NF-Y.* Nucleic Acids Res, 1992. **20**(5): p. 1087-91.
2. Dorn, A., et al., *Conserved major histocompatibility complex class II boxes--X and Y--are transcriptional control elements and specifically bind nuclear proteins.* Proc Natl Acad Sci U S A, 1987. **84**(17): p. 6249-53.
3. Guarente, L., et al., *Distinctly regulated tandem upstream activation sites mediate catabolite repression of the CYC1 gene of S. cerevisiae.* Cell, 1984. **36**(2): p. 503-11.
4. Pinkham, J.L. and L. Guarente, *Cloning and molecular analysis of the HAP2 locus: a global regulator of respiratory genes in Saccharomyces cerevisiae.* Mol Cell Biol, 1985. **5**(12): p. 3410-6.
5. Hahn, S., et al., *The HAP3 regulatory locus of Saccharomyces cerevisiae encodes divergent overlapping transcripts.* Mol Cell Biol, 1988. **8**(2): p. 655-63.
6. McNabb, D.S., Y. Xing, and L. Guarente, *Cloning of yeast HAP5: a novel subunit of a heterotrimeric complex required for CCAAT binding.* Genes Dev, 1995. **9**(1): p. 47-58.
7. Forsburg, S.L. and L. Guarente, *Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer.* Genes Dev, 1989. **3**(8): p. 1166-78.
8. Sinha, S., et al., *Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A and CBF-B and with yeast HAP2 and HAP3.* Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1624-8.
9. Kim, C.G. and M. Sheffery, *Physical characterization of the purified CCAAT transcription factor, alpha-CP1.* J Biol Chem, 1990. **265**(22): p. 13362-9.
10. Bi, W., et al., *DNA binding specificity of the CCAAT-binding factor CBF/NF-Y.* J Biol Chem, 1997. **272**(42): p. 26562-72.
11. de Silvio, A., C. Imbriano, and R. Mantovani, *Dissection of the NF-Y transcriptional activation potential.* Nucleic Acids Res, 1999. **27**(13): p. 2578-84.
12. Farina, A., et al., *Down-regulation of cyclin B1 gene transcription in terminally differentiated skeletal muscle cells is associated with loss of functional CCAAT-binding NF-Y complex.* Oncogene, 1999. **18**(18): p. 2818-27.
13. Gurtner, A., et al., *Requirement for down-regulation of the CCAAT-binding activity of the NF-Y transcription factor during skeletal muscle differentiation.* Mol Biol Cell, 2003. **14**(7): p. 2706-15.
14. Marziali, G., et al., *The activity of the CCAAT-box binding factor NF-Y is modulated through the regulated expression of its A subunit during monocyte to macrophage differentiation: regulation of tissue-specific genes through a ubiquitous transcription factor.* Blood, 1999. **93**(2): p. 519-26.
15. Bhattacharya, A., et al., *The B subunit of the CCAAT box binding transcription factor complex (CBF/NF-Y) is essential for early mouse development and cell proliferation.* Cancer Res, 2003. **63**(23): p. 8167-72.
16. Yoshioka, Y., et al., *Complex interference in the eye developmental pathway by Drosophila NF-YA.* Genesis, 2007. **45**(1): p. 21-31.
17. Dolfini, D., R. Gatta, and R. Mantovani, *NF-Y and the transcriptional activation of CCAAT promoters.* Crit Rev Biochem Mol Biol, 2012. **47**(1): p. 29-49.
18. Li, X.Y., et al., *Intron-exon organization of the NF-Y genes. Tissue-specific splicing modifies an activation domain.* J Biol Chem, 1992. **267**(13): p. 8984-90.
19. Coustry, F., S.N. Maity, and B. de Crombrugghe, *Studies on transcription activation by the multimeric CCAAT-binding factor CBF.* J Biol Chem, 1995. **270**(1): p. 468-75.

20. Coustry, F., et al., *The transcriptional activity of the CCAAT-binding factor CBF is mediated by two distinct activation domains, one in the CBF-B subunit and the other in the CBF-C subunit.* J Biol Chem, 1996. **271**(24): p. 14485-91.
21. Mantovani, R., et al., *Dominant negative analogs of NF-YA.* J Biol Chem, 1994. **269**(32): p. 20340-6.
22. Kim, I.S., et al., *Determination of functional domains in the C subunit of the CCAAT-binding factor (CBF) necessary for formation of a CBF-DNA complex: CBF-B interacts simultaneously with both the CBF-A and CBF-C subunits to form a heterotrimeric CBF molecule.* Mol Cell Biol, 1996. **16**(8): p. 4003-13.
23. Sinha, S., et al., *Three classes of mutations in the A subunit of the CCAAT-binding factor CBF delineate functional domains involved in the three-step assembly of the CBF-DNA complex.* Mol Cell Biol, 1996. **16**(1): p. 328-37.
24. Zemzoumi, K., et al., *NF-Y histone fold alpha1 helices help impart CCAAT specificity.* J Mol Biol, 1999. **286**(2): p. 327-37.
25. Romier, C., et al., *The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y.* J Biol Chem, 2003. **278**(2): p. 1336-45.
26. Baxevanis, A.D., et al., *A variety of DNA-binding and multimeric proteins contain the histone fold motif.* Nucleic Acids Res, 1995. **23**(14): p. 2685-91.
27. Gusmaroli, G., C. Tonelli, and R. Mantovani, *Regulation of the CCAAT-Binding NF-Y subunits in Arabidopsis thaliana.* Gene, 2001. **264**(2): p. 173-85.
28. Zhou, Z., et al., *Maneuver at the transcription start site: Mot1p and NC2 navigate TFIID/TBP to specific core promoter elements.* Epigenetics, 2009. **4**(1): p. 1-4.
29. Gangloff, Y.G., et al., *The histone fold is a key structural motif of transcription factor TFIID.* Trends Biochem Sci, 2001. **26**(4): p. 250-7.
30. Nagy, Z. and L. Tora, *Distinct GCN5/PCAF-containing complexes function as co-activators and are involved in transcription factor and global histone acetylation.* Oncogene, 2007. **26**(37): p. 5341-57.
31. Bolognese, F., et al., *Cloning and characterization of the histone-fold proteins YBL1 and YCL1.* Nucleic Acids Res, 2000. **28**(19): p. 3830-8.
32. Suganuma, T., et al., *ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding.* Nat Struct Mol Biol, 2008. **15**(4): p. 364-72.
33. Wang, Y.L., et al., *Human ATAC Is a GCN5/PCAF-containing acetylase complex with a novel NC2-like histone fold module that interacts with the TATA-binding protein.* J Biol Chem, 2008. **283**(49): p. 33808-15.
34. Arents, G. and E.N. Moudrianakis, *The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization.* Proc Natl Acad Sci U S A, 1995. **92**(24): p. 11170-4.
35. Bellorini, M., et al., *CCAAT binding NF-Y-TBP interactions: NF-YB and NF-YC require short domains adjacent to their histone fold motifs for association with TBP basic residues.* Nucleic Acids Res, 1997. **25**(11): p. 2174-81.
36. Suzuki, Y., et al., *Identification and characterization of the potential promoter regions of 1031 kinds of human genes.* Genome Res, 2001. **11**(5): p. 677-84.
37. FitzGerald, P.C., et al., *Clustering of DNA sequences in human promoters.* Genome Res, 2004. **14**(8): p. 1562-74.
38. Marino-Ramirez, L., et al., *Statistical analysis of over-represented words in human promoter sequences.* Nucleic Acids Res, 2004. **32**(3): p. 949-58.
39. Mantovani, R., *A survey of 178 NF-Y binding CCAAT boxes.* Nucleic Acids Res, 1998. **26**(5): p. 1135-43.

40. Bucher, P., *Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.* J Mol Biol, 1990. **212**(4): p. 563-78.
41. Testa, A., et al., *Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters.* J Biol Chem, 2005. **280**(14): p. 13606-15.
42. Ceribelli, M., et al., *The histone-like NF-Y is a bifunctional transcription factor.* Mol Cell Biol, 2008. **28**(6): p. 2047-58.
43. Su, M., et al., *Stereochemical analysis of the functional significance of the conserved inverted CCAAT and TATA elements in the rat bone sialoprotein gene promoter.* J Biol Chem, 2006. **281**(15): p. 9882-90.
44. Vilen, B.J., J.F. Penta, and J.P. Ting, *Structural constraints within a trimeric transcriptional regulatory region. Constitutive and interferon-gamma-inducible expression of the HLA-DRA gene.* J Biol Chem, 1992. **267**(33): p. 23728-34.
45. Vilen, B.J., J.P. Cogswell, and J.P. Ting, *Stereospecific alignment of the X and Y elements is required for major histocompatibility complex class II DRA promoter function.* Mol Cell Biol, 1991. **11**(5): p. 2406-15.
46. Zhu, X.S., et al., *Transcriptional scaffold: CIITA interacts with NF-Y, RFX, and CREB to cause stereospecific regulation of the class II major histocompatibility complex promoter.* Mol Cell Biol, 2000. **20**(16): p. 6051-61.
47. Linhoff, M.W., K.L. Wright, and J.P. Ting, *CCAAT-binding factor NF-Y and RFX are required for in vivo assembly of a nucleoprotein complex that spans 250 base pairs: the invariant chain promoter as a model.* Mol Cell Biol, 1997. **17**(8): p. 4589-96.
48. Linhart, C., et al., *Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis.* Cell Cycle, 2005. **4**(12): p. 1788-97.
49. Lucibello, F.C., et al., *Periodic cdc25C transcription is mediated by a novel cell cycle-regulated repressor element (CDE).* EMBO J, 1995. **14**(1): p. 132-42.
50. Liberati, C., et al., *NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices.* J Mol Biol, 1999. **285**(4): p. 1441-55.
51. Dolfini, D., et al., *A perspective of promoter architecture from the CCAAT box.* Cell Cycle, 2009. **8**(24): p. 4127-37.
52. Karsenty, G., P. Golumbek, and B. de Crombrugghe, *Point mutations and small substitution mutations in three different upstream elements inhibit the activity of the mouse alpha 2(I) collagen promoter.* J Biol Chem, 1988. **263**(27): p. 13909-15.
53. Maity, S.N., et al., *Selective activation of transcription by a novel CCAAT binding factor.* Science, 1988. **241**(4865): p. 582-5.
54. Dorn, A., et al., *A multiplicity of CCAAT box-binding proteins.* Cell, 1987. **50**(6): p. 863-72.
55. Ceribelli, M., et al., *Repression of new p53 targets revealed by ChIP on chip experiments.* Cell Cycle, 2006. **5**(10): p. 1102-10.
56. Farsetti, A., et al., *Inhibition of ERalpha-mediated trans-activation of human coagulation factor XII gene by heteromeric transcription factor NF-Y.* Endocrinology, 2001. **142**(8): p. 3380-8.
57. Murai-Takeda, A., et al., *NF-YC functions as a corepressor of agonist-bound mineralocorticoid receptor.* J Biol Chem, 2010. **285**(11): p. 8084-93.
58. Yun, J., et al., *Cdk2-dependent phosphorylation of the NF-Y transcription factor and its involvement in the p53-p21 signaling pathway.* J Biol Chem, 2003. **278**(38): p. 36966-72.
59. Ravasi, T., et al., *An atlas of combinatorial transcriptional regulation in mouse and man.* Cell, 2010. **140**(5): p. 744-52.

60. Poch, M.T., et al., *Two distinct classes of CCAAT box elements that bind nuclear factor-Y/alpha-actinin-4: potential role in human CYP1A1 regulation.* Toxicol Appl Pharmacol, 2004. **199**(3): p. 239-50.
61. Reed, B.D., et al., *Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes.* PLoS Genet, 2008. **4**(7): p. e1000133.
62. Villard, J., et al., *A functionally essential domain of RFX5 mediates activation of major histocompatibility complex class II promoters by promoting cooperative binding between RFX and NF-Y.* Mol Cell Biol, 2000. **20**(10): p. 3364-76.
63. Caretti, G., et al., *Dissection of functional NF-Y-RFX cooperative interactions on the MHC class II Ea promoter.* J Mol Biol, 2000. **302**(3): p. 539-52.
64. Leimgruber, E., et al., *Nucleosome eviction from MHC class II promoters controls positioning of the transcription start site.* Nucleic Acids Res, 2009. **37**(8): p. 2514-28.
65. Wright, K.L., et al., *CCAAT box binding protein NF-Y facilitates in vivo recruitment of upstream DNA binding transcription factors.* EMBO J, 1994. **13**(17): p. 4042-53.
66. Kretsovali, A., et al., *Involvement of CREB binding protein in expression of major histocompatibility complex class II genes via interaction with the class II transactivator.* Mol Cell Biol, 1998. **18**(11): p. 6777-83.
67. Fontes, J.D., et al., *Interactions between the class II transactivator and CREB binding protein increase transcription of major histocompatibility complex class II genes.* Mol Cell Biol, 1999. **19**(1): p. 941-7.
68. Mudhasani, R. and J.D. Fontes, *The class II transactivator requires brahma-related gene 1 to activate transcription of major histocompatibility complex class II genes.* Mol Cell Biol, 2002. **22**(14): p. 5019-26.
69. Pattenden, S.G., et al., *Interferon-gamma-induced chromatin remodeling at the CIITA locus is BRG1 dependent.* EMBO J, 2002. **21**(8): p. 1978-86.
70. Jabrane-Ferrat, N., et al., *MHC class II enhanceosome: how is the class II transactivator recruited to DNA-bound activators?* Int Immunol, 2003. **15**(4): p. 467-75.
71. Wong, A.W., et al., *CIITA-regulated plexin-A1 affects T-cell-dendritic cell interactions.* Nat Immunol, 2003. **4**(9): p. 891-8.
72. Benoist, C. and D. Mathis, *Regulation of major histocompatibility complex class-II genes: X, Y and other letters of the alphabet.* Annu Rev Immunol, 1990. **8**: p. 681-715.
73. Frontini, M., et al., *NF-Y recruitment of TFIID, multiple interactions with histone fold TAF(II)s.* J Biol Chem, 2002. **277**(8): p. 5841-8.
74. Mantovani, R., et al., *Monoclonal antibodies to NF-Y define its function in MHC class II and albumin gene transcription.* EMBO J, 1992. **11**(9): p. 3315-22.
75. Duan, Z., G. Stamatoyannopoulos, and Q. Li, *Role of NF-Y in in vivo regulation of the gamma-globin gene.* Mol Cell Biol, 2001. **21**(9): p. 3083-95.
76. Xie, X., et al., *Structural similarity between TAFs and the heterotetrameric core of the histone octamer.* Nature, 1996. **380**(6572): p. 316-22.
77. Birck, C., et al., *Human TAF(II)28 and TAF(II)18 interact through a histone fold encoded by atypical evolutionary conserved motifs also found in the SPT3 family.* Cell, 1998. **94**(2): p. 239-49.
78. Coustry, F., et al., *The two activation domains of the CCAAT-binding factor CBF interact with the dTAFII110 component of the Drosophila TFIID complex.* Biochem J, 1998. **331 (Pt 1)**: p. 291-7.
79. Jin, S. and K.W. Scotto, *Transcriptional regulation of the MDR1 gene by histone acetyltransferase and deacetylase is mediated by NF-Y.* Mol Cell Biol, 1998. **18**(7): p. 4377-84.

80. Huang, W., et al., *Trichostatin A induces transforming growth factor beta type II receptor promoter activity and acetylation of Sp1 by recruitment of PCAF/p300 to a Sp1.NF-Y complex.* J Biol Chem, 2005. **280**(11): p. 10047-54.
81. Currie, R.A., *NF-Y is associated with the histone acetyltransferases GCN5 and P/CAF.* J Biol Chem, 1998. **273**(3): p. 1430-4.
82. Motta, M.C., et al., *Interactions of the CCAAT-binding trimer NF-Y with nucleosomes.* J Biol Chem, 1999. **274**(3): p. 1326-33.
83. Caretti, G., M.C. Motta, and R. Mantovani, *NF-Y associates with H3-H4 tetramers and octamers by multiple mechanisms.* Mol Cell Biol, 1999. **19**(12): p. 8591-603.
84. Coustry, F., et al., *CBF/NF-Y functions both in nucleosomal disruption and transcription activation of the chromatin-assembled topoisomerase IIalpha promoter. Transcription activation by CBF/NF-Y in chromatin is dependent on the promoter structure.* J Biol Chem, 2001. **276**(44): p. 40621-30.
85. Gatta, R. and R. Mantovani, *NF-Y affects histone acetylation and H2A.Z deposition in cell cycle promoters.* Epigenetics, 2011. **6**(4): p. 526-34.
86. Sun, F., et al., *Nuclear factor Y is required for basal activation and chromatin accessibility of fibroblast growth factor receptor 2 promoter in osteoblast-like cells.* J Biol Chem, 2009. **284**(5): p. 3136-47.
87. Boucher, P.D., M.P. Piechocki, and R.N. Hines, *Partial characterization of the human CYP1A1 negatively acting transcription factor and mutational analysis of its cognate DNA recognition sequence.* Mol Cell Biol, 1995. **15**(9): p. 5144-51.
88. Peng, Y. and N. Jahroudi, *The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases.* J Biol Chem, 2003. **278**(10): p. 8385-94.
89. Peng, Y. and N. Jahroudi, *The NFY transcription factor functions as a repressor and activator of the von Willebrand factor promoter.* Blood, 2002. **99**(7): p. 2408-17.
90. Gowri, P.M., et al., *Recruitment of a repressosome complex at the growth hormone receptor promoter and its potential role in diabetic nephropathy.* Mol Cell Biol, 2003. **23**(3): p. 815-25.
91. Hewetson, A. and B.S. Chilton, *An Sp1-NF-Y/progesterone receptor DNA binding-dependent mechanism regulates progesterone-induced transcriptional activation of the rabbit RUSH/SMARCA3 gene.* J Biol Chem, 2003. **278**(41): p. 40177-85.
92. Bernadt, C.T., et al., *NF-Y behaves as a bifunctional transcription factor that can stimulate or repress the FGF-4 promoter in an enhancer-dependent manner.* Gene Expr, 2005. **12**(3): p. 193-212.
93. Deng, H., et al., *Transcription factor NFY globally represses the expression of the C. elegans Hox gene Abdominal-B homolog egl-5.* Dev Biol, 2007. **308**(2): p. 583-92.
94. Imbriano, C., et al., *Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters.* Mol Cell Biol, 2005. **25**(9): p. 3737-51.
95. Harrow, J., et al., *GENCODE: The reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
96. Vuorio, T., S.N. Maity, and B. de Crombrugghe, *Purification and molecular cloning of the "A" chain of a rat heteromeric CCAAT-binding protein. Sequence identity with the yeast HAP3 transcription factor.* J Biol Chem, 1990. **265**(36): p. 22480-6.
97. Grskovic, M., et al., *Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells.* PLoS Genet, 2007. **3**(8): p. e145.
98. Ge, Y., et al., *Synergistic regulation of human cystathionine-beta-synthase-1b promoter by transcription factors NF-YA isoforms and Sp1.* Biochim Biophys Acta, 2002. **1579**(2-3): p. 73-80.

99. Zhu, J., et al., *NF-Ya activates multiple hematopoietic stem cell (HSC) regulatory genes and promotes HSC self-renewal.* Proc Natl Acad Sci U S A, 2005. **102**(33): p. 11728-33.
100. Sinha, S., et al., *Chromosomal assignment and tissue expression of CBF-C/NFY-C, the third subunit of the mammalian CCAAT-binding factor.* Genomics, 1996. **37**(2): p. 260-3.
101. Bellorini, M., et al., *Cloning and expression of human NF-YC.* Gene, 1997. **193**(1): p. 119-25.
102. Ceribelli, M., et al., *NF-YC complexity is generated by dual promoters and alternative splicing.* J Biol Chem, 2009. **284**(49): p. 34189-200.
103. Chen, F., et al., *Repression of Smad2 and Smad3 transactivating activity by association with a novel splice variant of CCAAT-binding factor C subunit.* Biochem J, 2002. **364**(Pt 2): p. 571-7.
104. Bolognese, F., et al., *The cyclin B2 promoter depends on NF-Y, a trimer whose CCAAT-binding activity is cell-cycle regulated.* Oncogene, 1999. **18**(10): p. 1845-53.
105. Yang, J., et al., *A novel mechanism involving coordinated regulation of nuclear levels and acetylation of NF-YA and Bcl6 activates RGS4 transcription.* J Biol Chem, 2010. **285**(39): p. 29760-9.
106. Manni, I., et al., *Posttranslational regulation of NF-YA modulates NF-Y transcriptional activity.* Mol Biol Cell, 2008. **19**(12): p. 5203-13.
107. Li, Q., et al., *Xenopus NF-Y pre-sets chromatin to potentiate p300 and acetylation-responsive transcription from the Xenopus hsp70 promoter in vivo.* EMBO J, 1998. **17**(21): p. 6300-15.
108. Chae, H.D., et al., *Cdk2-dependent phosphorylation of the NF-Y transcription factor is essential for the expression of the cell cycle-regulatory genes and cell cycle G1/S and G2/M transitions.* Oncogene, 2004. **23**(23): p. 4084-8.
109. Chan, Q.K., et al., *Activation of GPR30 inhibits the growth of prostate cancer cells through sustained activation of Erk1/2, c-jun/c-fos-dependent upregulation of p21, and induction of G(2) cell-cycle arrest.* Cell Death Differ, 2010. **17**(9): p. 1511-23.
110. Abate, C., et al., *Redox regulation of fos and jun DNA-binding activity in vitro.* Science, 1990. **249**(4973): p. 1157-61.
111. Matthews, J.R., et al., *Thioredoxin regulates the DNA binding activity of NF-kappa B by reduction of a disulphide bond involving cysteine 62.* Nucleic Acids Res, 1992. **20**(15): p. 3821-30.
112. Guehmann, S., et al., *Reduction of a conserved Cys is essential for Myb DNA-binding.* Nucleic Acids Res, 1992. **20**(9): p. 2279-86.
113. Wemmie, J.A., S.M. Steggerda, and W.S. Moye-Rowley, *The Saccharomyces cerevisiae AP-1 protein discriminates between oxidative stress elicited by the oxidants H2O2 and diamide.* J Biol Chem, 1997. **272**(12): p. 7908-14.
114. Kuge, S., N. Jones, and A. Nomoto, *Regulation of yAP-1 nuclear localization in response to oxidative stress.* EMBO J, 1997. **16**(7): p. 1710-20.
115. Nakshatri, H., P. Bhat-Nakshatri, and R.A. Currie, *Subunit association and DNA binding activity of the heterotrimeric transcription factor NF-Y is regulated by cellular redox.* J Biol Chem, 1996. **271**(46): p. 28784-91.
116. Thon, M., et al., *The CCAAT-binding complex coordinates the oxidative stress response in eukaryotes.* Nucleic Acids Res, 2010. **38**(4): p. 1098-113.
117. Habib, S.L., *Molecular mechanism of regulation of OGG1: tuberin deficiency results in cytoplasmic redistribution of transcriptional factor NF-YA.* J Mol Signal, 2009. **4**: p. 8.
118. Habib, S.L., et al., *Tuberin regulates the DNA repair enzyme OGG1.* Am J Physiol Renal Physiol, 2008. **294**(1): p. F281-90.
119. Lee, M.R., et al., *Transcription factors NF-YA regulate the induction of human OGG1 following DNA-alkylating agent methylmethane sulfonate (MMS) treatment.* J Biol Chem, 2004. **279**(11): p. 9857-66.

120. Frontini, M., et al., *Cell cycle regulation of NF-YC nuclear localization.* Cell Cycle, 2004. **3**(2): p. 217-22.
121. Kahle, J., et al., *Subunits of the heterotrimeric transcription factor NF-Y are imported into the nucleus by distinct pathways involving importin beta and importin 13.* Mol Cell Biol, 2005. **25**(13): p. 5339-54.
122. Steidl, S., et al., *A single subunit of a heterotrimeric CCAAT-binding complex carries a nuclear localization signal: piggy back transport of the pre-assembled complex to the nucleus.* J Mol Biol, 2004. **342**(2): p. 515-24.
123. Goda, H., et al., *Nuclear translocation of the heterotrimeric CCAAT binding factor of Aspergillus oryzae is dependent on two redundant localising signals in a single subunit.* Arch Microbiol, 2005. **184**(2): p. 93-100.
124. Tuncher, A., et al., *The CCAAT-binding complex of eukaryotes: evolution of a second NLS in the HapB subunit of the filamentous fungus Aspergillus nidulans despite functional conservation at the molecular level between yeast, A.nidulans and human.* J Mol Biol, 2005. **352**(3): p. 517-33.
125. Tabach, Y., et al., *The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation.* Mol Syst Biol, 2005. **1**: p. 2005 0022.
126. Scafoglio, C., et al., *Comparative gene expression profiling reveals partially overlapping but distinct genomic actions of different antiestrogens in human breast cancer cells.* J Cell Biochem, 2006. **98**(5): p. 1163-84.
127. Niida, A., et al., *Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells.* BMC Bioinformatics, 2008. **9**: p. 404.
128. Jurchott, K., et al., *Identification of Y-box binding protein 1 as a core regulator of MEK/ERK pathway-dependent gene signatures in colorectal cancer cells.* PLoS Genet, 2010. **6**(12): p. e1001231.
129. Salvatore, G., et al., *A cell proliferation and chromosomal instability signature in anaplastic thyroid carcinoma.* Cancer Res, 2007. **67**(21): p. 10148-58.
130. Blum, R., et al., *Molecular signatures of prostate stem cells reveal novel signaling pathways and provide insights into prostate cancer.* PLoS One, 2009. **4**(5): p. e5722.
131. Calvo, A., et al., *Molecular characterization of the Ggamma-globin-Tag transgenic mouse model of hormone refractory prostate cancer: comparison to human prostate cancer.* Prostate, 2010. **70**(6): p. 630-45.
132. Forsberg, E.C., et al., *Molecular signatures of quiescent, mobilized and leukemia-initiating hematopoietic stem cells.* PLoS One, 2010. **5**(1): p. e8785.
133. Goodarzi, H., O. Elemento, and S. Tavazoie, *Revealing global regulatory perturbations across human cancers.* Mol Cell, 2009. **36**(5): p. 900-11.
134. Sinha, S., et al., *Systematic functional characterization of cis-regulatory motifs in human core promoters.* Genome Res, 2008. **18**(3): p. 477-88.
135. Rhodes, D.R., et al., *Mining for regulatory programs in the cancer transcriptome.* Nat Genet, 2005. **37**(6): p. 579-83.
136. Zhu, W., P.H. Giangrande, and J.R. Nevins, *E2Fs link the control of G1/S and G2/M transcription.* EMBO J, 2004. **23**(23): p. 4615-26.
137. Thomassen, M., Q. Tan, and T.A. Kruse, *Gene expression meta-analysis identifies metastatic pathways and transcription factors in breast cancer.* BMC Cancer, 2008. **8**: p. 394.
138. Akira, S., et al., *Molecular cloning of APRF, a novel IFN-stimulated gene factor 3 p91-related transcription factor involved in the gp130-mediated signaling pathway.* Cell, 1994. **77**(1): p. 63-71.

139. Zhong, Z., Z. Wen, and J.E. Darnell, Jr., *Stat3: a STAT family member activated by tyrosine phosphorylation in response to epidermal growth factor and interleukin-6.* Science, 1994. **264**(5155): p. 95-8.
140. Ivashkiv, L.B., *Cytokines and STATs: how can signals achieve specificity?* Immunity, 1995. **3**(1): p. 1-4.
141. Becker, S., B. Groner, and C.W. Muller, *Three-dimensional structure of the Stat3beta homodimer bound to DNA.* Nature, 1998. **394**(6689): p. 145-51.
142. Guyer, N.B., et al., *IFN-gamma induces a p91/Stat1 alpha-related transcription factor with distinct activation and binding properties.* J Immunol, 1995. **155**(7): p. 3472-80.
143. Bergad, P.L., et al., *Growth hormone induction of hepatic serine protease inhibitor 2.1 transcription is mediated by a Stat5-related factor binding synergistically to two gamma-activated sites.* J Biol Chem, 1995. **270**(42): p. 24903-10.
144. Xu, X., Y.L. Sun, and T. Hoey, *Cooperative DNA binding and sequence-selective recognition conferred by the STAT amino-terminal domain.* Science, 1996. **273**(5276): p. 794-7.
145. Vinkemeier, U., et al., *DNA binding of in vitro activated Stat1 alpha, Stat1 beta and truncated Stat1: interaction between NH2-terminal domains stabilizes binding of two dimers to tandem DNA sites.* EMBO J, 1996. **15**(20): p. 5616-26.
146. Paulson, M., et al., *Stat protein transactivation domains recruit p300/CBP through widely divergent sequences.* J Biol Chem, 1999. **274**(36): p. 25343-9.
147. Wang, R., P. Cherukuri, and J. Luo, *Activation of Stat3 sequence-specific DNA binding and transcription by p300/CREB-binding protein-mediated acetylation.* J Biol Chem, 2005. **280**(12): p. 11528-34.
148. Yuan, Z.L., et al., *Stat3 dimerization regulated by reversible acetylation of a single lysine residue.* Science, 2005. **307**(5707): p. 269-73.
149. Zhang, X., et al., *Interacting regions in Stat3 and c-Jun that participate in cooperative transcriptional activation.* Mol Cell Biol, 1999. **19**(10): p. 7138-46.
150. Schaefer, T.S., L.K. Sanders, and D. Nathans, *Cooperative transcriptional activity of Jun and Stat3 beta, a short form of Stat3.* Proc Natl Acad Sci U S A, 1995. **92**(20): p. 9097-101.
151. Icardi, L., et al., *The Sin3a repressor complex is a master regulator of STAT transcriptional activity.* Proc Natl Acad Sci U S A, 2012. **109**(30): p. 12058-63.
152. Ray, S., et al., *Requirement of histone deacetylase1 (HDAC1) in signal transducer and activator of transcription 3 (STAT3) nucleocytoplasmic distribution.* Nucleic Acids Res, 2008. **36**(13): p. 4510-20.
153. Lee, H., et al., *Acetylated STAT3 is crucial for methylation of tumor-suppressor gene promoters and inhibition by resveratrol results in demethylation.* Proc Natl Acad Sci U S A, 2012. **109**(20): p. 7765-9.
154. Shi, S., et al., *Drosophila STAT is required for directly maintaining HP1 localization and heterochromatin stability.* Nat Cell Biol, 2008. **10**(4): p. 489-96.
155. Shi, S., et al., *JAK signaling globally counteracts heterochromatic gene silencing.* Nat Genet, 2006. **38**(9): p. 1071-6.
156. Timofeeva, O.A., et al., *Mechanisms of unphosphorylated STAT3 transcription factor binding to DNA.* J Biol Chem, 2012. **287**(17): p. 14192-200.
157. Yang, J., et al., *Unphosphorylated STAT3 accumulates in response to IL-6 and activates transcription by binding to NFkappaB.* Genes Dev, 2007. **21**(11): p. 1396-408.
158. Yoshida, Y., et al., *Interleukin 1 activates STAT3/nuclear factor-kappaB cross-talk via a unique TRAF6- and p65-dependent mechanism.* J Biol Chem, 2004. **279**(3): p. 1768-76.

159. Yu, Z., W. Zhang, and B.C. Kone, *Signal transducers and activators of transcription 3 (STAT3) inhibits transcription of the inducible nitric oxide synthase gene by interacting with nuclear factor kappaB.* Biochem J, 2002. **367**(Pt 1): p. 97-105.
160. Yang, J., et al., *Novel roles of unphosphorylated STAT3 in oncogenesis and transcriptional regulation.* Cancer Res, 2005. **65**(3): p. 939-47.
161. Aggarwal, B.B., et al., *Signal transducer and activator of transcription-3, inflammation, and cancer: how intimate is the relationship?* Ann N Y Acad Sci, 2009. **1171**: p. 59-76.
162. Yu, C.L., et al., *Enhanced DNA-binding activity of a Stat3-related protein in cells transformed by the Src oncoprotein.* Science, 1995. **269**(5220): p. 81-3.
163. Turkson, J., et al., *Requirement for Ras/Rac1-mediated p38 and c-Jun N-terminal kinase signaling in Stat3 transcriptional activity induced by the Src oncoprotein.* Mol Cell Biol, 1999. **19**(11): p. 7519-28.
164. Wen, Z., Z. Zhong, and J.E. Darnell, Jr., *Maximal activation of transcription by Stat1 and Stat3 requires both tyrosine and serine phosphorylation.* Cell, 1995. **82**(2): p. 241-50.
165. Wen, Z. and J.E. Darnell, Jr., *Mapping of Stat3 serine phosphorylation to a single residue (727) and evidence that serine phosphorylation has no influence on DNA binding of Stat1 and Stat3.* Nucleic Acids Res, 1997. **25**(11): p. 2062-7.
166. Caldenhoven, E., et al., *STAT3beta, a splice variant of transcription factor STAT3, is a dominant negative regulator of transcription.* J Biol Chem, 1996. **271**(22): p. 13221-7.
167. Grivennikov, S.I. and M. Karin, *Inflammation and oncogenesis: a vicious connection.* Curr Opin Genet Dev, 2010. **20**(1): p. 65-71.
168. Borrello, M.G., D. Degl'Innocenti, and M.A. Pierotti, *Inflammation and cancer: the oncogene-driven connection.* Cancer Lett, 2008. **267**(2): p. 262-70.
169. Croce, C.M., *Oncogenes and cancer.* N Engl J Med, 2008. **358**(5): p. 502-11.
170. Mantovani, A., et al., *Cancer-related inflammation.* Nature, 2008. **454**(7203): p. 436-44.
171. Mantovani, A., *Cancer: Inflaming metastasis.* Nature, 2009. **457**(7225): p. 36-7.
172. Grivennikov, S.I., F.R. Greten, and M. Karin, *Immunity, inflammation, and cancer.* Cell, 2010. **140**(6): p. 883-99.
173. Muller, A., et al., *Involvement of chemokine receptors in breast cancer metastasis.* Nature, 2001. **410**(6824): p. 50-6.
174. Kim, S.Y., et al., *Inhibition of the CXCR4/CXCL12 chemokine pathway reduces the development of murine pulmonary metastases.* Clin Exp Metastasis, 2008. **25**(3): p. 201-11.
175. Li, L. and P.E. Shaw, *Autocrine-mediated activation of STAT3 correlates with cell proliferation in breast carcinoma lines.* J Biol Chem, 2002. **277**(20): p. 17397-405.
176. DeArmond, D., et al., *Autocrine-mediated ErbB-2 kinase activation of STAT3 is required for growth factor independence of pancreatic cancer cell lines.* Oncogene, 2003. **22**(49): p. 7781-95.
177. Wei, D., et al., *Stat3 activation regulates the expression of vascular endothelial growth factor and human pancreatic cancer angiogenesis and metastasis.* Oncogene, 2003. **22**(3): p. 319-29.
178. Lee, S.O., et al., *Interleukin-6 promotes androgen-independent growth in LNCaP human prostate cancer cells.* Clin Cancer Res, 2003. **9**(1): p. 370-6.
179. Rebouissou, S., et al., *Frequent in-frame somatic deletions activate gp130 in inflammatory hepatocellular tumours.* Nature, 2009. **457**(7226): p. 200-4.
180. Rawat, R., et al., *Constitutive activation of STAT3 is associated with the acquisition of an interleukin 6-independent phenotype by murine plasmacytomas and hybridomas.* Blood, 2000. **96**(10): p. 3514-21.
181. Jove, R., *Preface: STAT signaling.* Oncogene, 2000. **19**(21): p. 2466-7.
182. Bowman, T., et al., *STATs in oncogenesis.* Oncogene, 2000. **19**(21): p. 2474-88.

183. The ENCODE Project Consortium, *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
184. The ENCODE Project Consortium, *A user's guide to the encyclopedia of DNA elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.
185. The ENCODE Project Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
186. Rando, O.J., *Global patterns of histone modifications.* Curr Opin Genet Dev, 2007. **17**(2): p. 94-9.
187. Felsenfeld, G., *Chromatin as an essential part of the transcriptional mechanism.* Nature, 1992. **355**(6357): p. 219-24.
188. Hirschhorn, J.N., et al., *Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure.* Genes Dev, 1992. **6**(12A): p. 2288-98.
189. Kwon, H., et al., *Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex.* Nature, 1994. **370**(6489): p. 477-81.
190. Moazed, D., *Small RNAs in transcriptional gene silencing and genome defence.* Nature, 2009. **457**(7228): p. 413-20.
191. Simon, J.A. and R.E. Kingston, *Mechanisms of polycomb gene silencing: knowns and unknowns.* Nat Rev Mol Cell Biol, 2009. **10**(10): p. 697-708.
192. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
193. Hooft van Huijsduijnen, R.A., et al., *Properties of a CCAAT box-binding protein.* Nucleic Acids Res, 1987. **15**(18): p. 7265-82.
194. Kim, C.G., et al., *Promoter elements and erythroid cell nuclear factors that regulate alpha-globin gene transcription in vitro.* Mol Cell Biol, 1990. **10**(11): p. 5958-66.
195. Imbriano, C., N. Gnesutta, and R. Mantovani, *The NF-Y/p53 liaison: Well beyond repression.* Biochim Biophys Acta, 2012. **1825**(2): p. 131-9.
196. Morachis, J.M., C.M. Murawsky, and B.M. Emerson, *Regulation of the p53 transcriptional response by structurally diverse core promoters.* Genes Dev, 2010. **24**(2): p. 135-47.
197. Hughes, R., et al., *NF-Y is essential for expression of the proapoptotic bim gene in sympathetic neurons.* Cell Death Differ, 2011. **18**(6): p. 937-47.
198. Khalil, A.M., et al., *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(28): p. 11667-72.
199. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic Acids Res, 2011. **39**(Database issue): p. D152-7.
200. Fujita, P.A., et al., *The UCSC Genome Browser database: update 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D876-82.
201. He, S., et al., *NONCODE v2.0: decoding the non-coding.* Nucleic Acids Res, 2008. **36**(Database issue): p. D170-2.
202. Moqtaderi, Z., et al., *Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells.* Nat Struct Mol Biol, 2010. **17**(5): p. 635-40.
203. Dorn, A., et al., *B-cell control region at the 5' end of a major histocompatibility complex class II gene: sequences and factors.* Mol Cell Biol, 1988. **8**(10): p. 3975-87.
204. Gilthorpe, J., et al., *Spatially specific expression of Hoxb4 is dependent on the ubiquitous transcription factor NFY.* Development, 2002. **129**(16): p. 3887-99.
205. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
206. Gubler, U., et al., *Coexpression of two distinct genes is required to generate secreted bioactive cytotoxic lymphocyte maturation factor.* Proc Natl Acad Sci U S A, 1991. **88**(10): p. 4143-7.

207. Wolf, S.F., et al., *Cloning of cDNA for natural killer cell stimulatory factor, a heterodimeric cytokine with multiple biologic effects on T and natural killer cells.* J Immunol, 1991. **146**(9): p. 3074-81.
208. Yang, A., et al., *Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells.* Mol Cell, 2006. **24**(4): p. 593-602.
209. Strub, T., et al., *Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma.* Oncogene, 2011. **30**(20): p. 2319-32.
210. Martynova, E., et al., *Gain-of-function p53 mutants have widespread genomic locations partially overlapping with p63.* Oncotarget, 2012. **3**(2): p. 132-43.
211. Benachenhou, F., V. Blikstad, and J. Blomberg, *The phylogeny of orthoretroviral long terminal repeats (LTRs).* Gene, 2009. **448**(2): p. 134-8.
212. Benachenhou, F., et al., *Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data.* PLoS One, 2009. **4**(4): p. e5179.
213. Bourque, G., *Transposable elements in gene regulation and in the evolution of vertebrate genomes.* Curr Opin Genet Dev, 2009. **19**(6): p. 607-12.
214. Graves, B.J., P.F. Johnson, and S.L. McKnight, *Homologous recognition of a promoter domain common to the MSV LTR and the HSV tk gene.* Cell, 1986. **44**(4): p. 565-76.
215. Dutta, A., M.Y. Stoeckle, and H. Hanafusa, *Serum and v-src increase the level of a CCAAT-binding factor required for transcription from a retroviral long terminal repeat.* Genes Dev, 1990. **4**(2): p. 243-54.
216. Greuel, B.T., L. Sealy, and J.E. Majors, *Transcriptional activity of the Rous sarcoma virus long terminal repeat correlates with binding of a factor to an upstream CCAAT box in vitro.* Virology, 1990. **177**(1): p. 33-43.
217. Faber, M. and L. Sealy, *Rous sarcoma virus enhancer factor I is a ubiquitous CCAAT transcription factor highly related to CBF and NF-Y.* J Biol Chem, 1990. **265**(36): p. 22243-54.
218. Scheef, G., et al., *Transcriptional regulation of porcine endogenous retroviruses released from porcine and infected human cells by heterotrimeric protein complex NF-Y and impact of immunosuppressive drugs.* J Virol, 2002. **76**(24): p. 12553-63.
219. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome Res, 2008. **18**(11): p. 1752-62.
220. Kunarso, G., et al., *Transposable elements have rewired the core regulatory network of human embryonic stem cells.* Nat Genet, 2010. **42**(7): p. 631-4.
221. Yu, X., et al., *The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2.* J Biol Chem, 2005. **280**(42): p. 35184-94.
222. Pi, W., et al., *Long-range function of an intergenic retrotransposon.* Proc Natl Acad Sci U S A, 2010. **107**(29): p. 12992-7.
223. Maksakova, I.A., D.L. Mager, and D. Reiss, *Keeping active endogenous retroviral-like elements in check: the epigenetic perspective.* Cell Mol Life Sci, 2008. **65**(21): p. 3329-47.
224. Magnani, L., J. Eeckhoute, and M. Lupien, *Pioneer factors: directing transcriptional regulators within the chromatin environment.* Trends Genet, 2011. **27**(11): p. 465-74.
225. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene expression.* Genes Dev, 2011. **25**(21): p. 2227-41.
226. Gurtner, A., et al., *Transcription factor NF-Y induces apoptosis in cells expressing wild-type p53 through E2F1 upregulation and p53 activation.* Cancer Res, 2010. **70**(23): p. 9711-20.
227. Litovchick, L., et al., *Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence.* Mol Cell, 2007. **26**(4): p. 539-51.

228. Schmit, F., et al., *LINC, a human complex that is related to pRB-containing complexes in invertebrates regulates the expression of G2/M genes.* Cell Cycle, 2007. **6**(15): p. 1903-13.
229. Muller, G.A., et al., *The CHR promoter element controls cell cycle-dependent gene transcription and binds the DREAM and MMB complexes.* Nucleic Acids Res, 2012. **40**(4): p. 1561-78.
230. Muller, G.A. and K. Engeland, *The central role of CDE/CHR promoter elements in the regulation of cell cycle-dependent gene transcription.* FEBS J, 2010. **277**(4): p. 877-93.
231. Izumi, H., et al., *Mechanism for the transcriptional repression by c-Myc on PDGF beta-receptor.* J Cell Sci, 2001. **114**(Pt 8): p. 1533-44.
232. Kalra, I.S., et al., *Kruppel-like Factor 4 activates HBG gene expression in primary erythroid cells.* Br J Haematol, 2011. **154**(2): p. 248-59.
233. Evans, P.M., et al., *Kruppel-like factor 4 is acetylated by p300 and regulates gene transcription via modulation of histone acetylation.* J Biol Chem, 2007. **282**(47): p. 33994-4002.
234. Yoon, H.S. and V.W. Yang, *Requirement of Kruppel-like factor 4 in preventing entry into mitosis following DNA damage.* J Biol Chem, 2004. **279**(6): p. 5035-41.
235. Oishi, Y., et al., *SUMOylation of Kruppel-like transcription factor 5 acts as a molecular switch in transcriptional programs of lipid metabolism involving PPAR-delta.* Nat Med, 2008. **14**(6): p. 656-66.
236. Turner, J. and M. Crossley, *Cloning and characterization of mCtBP2, a co-repressor that associates with basic Kruppel-like factor and other mammalian transcriptional regulators.* EMBO J, 1998. **17**(17): p. 5129-40.
237. van Vliet, J., J. Turner, and M. Crossley, *Human Kruppel-like factor 8: a CACCC-box binding protein that associates with CtBP and represses transcription.* Nucleic Acids Res, 2000. **28**(9): p. 1955-62.
238. Schuierer, M., et al., *Induction of AP-2alpha expression by adenoviral infection involves inactivation of the AP-2rep transcriptional corepressor CtBP1.* J Biol Chem, 2001. **276**(30): p. 27944-9.
239. West, A.G., et al., *Recruitment of histone modifications by USF proteins at a vertebrate barrier element.* Mol Cell, 2004. **16**(3): p. 453-63.
240. Huang, S., et al., *USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier.* Mol Cell Biol, 2007. **27**(22): p. 7991-8002.
241. Li, X., et al., *Chromatin boundaries require functional collaboration between the hSET1 and NURF complexes.* Blood, 2011. **118**(5): p. 1386-94.
242. Prieto, C. and J. De Las Rivas, *APID: Agile Protein Interaction DataAnalyzer.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W298-302.
243. Tiwari, V.K., et al., *A chromatin-modifying function of JNK during stem cell differentiation.* Nat Genet, 2012. **44**(1): p. 94-100.
244. Yoshida, H., et al., *ATF6 activated by proteolysis binds in the presence of NF-Y (CBF) directly to the cis-acting element responsible for the mammalian unfolded protein response.* Mol Cell Biol, 2000. **20**(18): p. 6755-67.
245. Salsi, V., et al., *Interactions between p300 and multiple NF-Y trimers govern cyclin B2 promoter function.* J Biol Chem, 2003. **278**(9): p. 6642-50.
246. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences.* Nucleic Acids Res, 1990. **18**(20): p. 6097-100.
247. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
248. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
249. Jiang, H. and W.H. Wong, *Statistical inferences for isoform expression in RNA-Seq.* Bioinformatics, 2009. **25**(8): p. 1026-32.

250. Jiang, H. and W.H. Wong, *SeqMap: mapping massive amount of oligonucleotides to the genome.* Bioinformatics, 2008. **24**(20): p. 2395-6.
251. You, F.M., et al., *BatchPrimer3: a high throughput web application for PCR and sequencing primer design.* BMC Bioinformatics, 2008. **9**: p. 253.
252. Aparicio, O., et al., *Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo.* Curr Protoc Mol Biol, 2005. **Chapter 21**: p. Unit 21 3.
253. Cawley, S., et al., *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.* Cell, 2004. **116**(4): p. 499-509.
254. Benatti, P., et al., *Specific inhibition of NF-Y subunits triggers different cell proliferation defects.* Nucleic Acids Res, 2011. **39**(13): p. 5356-68.
255. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
256. Gautier, L., et al., *affy--analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-15.
257. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
258. Smith, C.A., *annaffy: Annotation tools for Affymetrix biological metadata. R package version 1.24.0.* 2010.
259. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nat Biotechnol, 2010. **28**(5): p. 495-501.
260. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
261. Smit, A.F.A., R. Hubley, and P. Green, *RepeatMasker Open-3.0.* 1996-2010.
262. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
263. Gupta, S., et al., *Quantifying similarity between motifs.* Genome Biol, 2007. **8**(2): p. R24.
264. Portales-Casamar, E., et al., *JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2010. **38**(Database issue): p. D105-10.
265. Zambelli, F., G. Pesole, and G. Pavesi, *Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W247-52.
266. Ye, T., et al., *seqMINER: an integrated ChIP-seq data interpretation platform.* Nucleic Acids Res, 2011. **39**(6): p. e35.
267. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.
268. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics, 2011. **27**(7): p. 1017-8.
269. Suzuki, R. and H. Shimodaira, *Pvclust: an R package for assessing the uncertainty in hierarchical clustering.* Bioinformatics, 2006. **22**(12): p. 1540-2.
270. Carlson, J.M., D. Heckerman, and G. Shani, *Estimating false discovery rates for contingency tables.* Microsoft Research Technical Reports, 2009. **MSR-TR-2009-53**.
271. Zhang, X., et al., *Requirement of serine phosphorylation for formation of STAT-promoter complexes.* Science, 1995. **267**(5206): p. 1990-4.
272. Chung, J., et al., *STAT3 serine phosphorylation by ERK-dependent and -independent pathways negatively modulates its tyrosine phosphorylation.* Mol Cell Biol, 1997. **17**(11): p. 6508-16.
273. Cao, X., et al., *Activation and association of Stat3 with Src in v-Src-transformed cell lines.* Mol Cell Biol, 1996. **16**(4): p. 1595-603.

274. Chaturvedi, P., S. Sharma, and E.P. Reddy, *Abrogation of interleukin-3 dependence of myeloid cells by the v-src oncogene requires SH2 and SH3 domains which specify activation of STATs.* Mol Cell Biol, 1997. **17**(6): p. 3295-304.
275. Lu, Y., et al., *Piwil2 suppresses p53 by inducing phosphorylation of signal transducer and activator of transcription 3 in tumor cells.* PLoS One, 2012. **7**(1): p. e30999.
276. Chaturvedi, P., M.V. Reddy, and E.P. Reddy, *Src kinases and not JAKs activate STATs during IL-3 induced myeloid cell proliferation.* Oncogene, 1998. **16**(13): p. 1749-58.
277. Garcia, R. and R. Jove, *Activation of STAT transcription factors in oncogenic tyrosine kinase signaling.* J Biomed Sci, 1998. **5**(2): p. 79-85.
278. Gouilleux-Gruart, V., et al., *STAT-related transcription factors are constitutively activated in peripheral blood cells from acute leukemia patients.* Blood, 1996. **87**(5): p. 1692-7.
279. Karras, J.G., et al., *Signal transducer and activator of transcription-3 (STAT3) is constitutively activated in normal, self-renewing B-1 cells but only inducibly expressed in conventional B lymphocytes.* J Exp Med, 1997. **185**(6): p. 1035-42.
280. Sartor, C.I., et al., *Role of epidermal growth factor receptor and STAT-3 activation in autonomous proliferation of SUM-102PT human breast cancer cells.* Cancer Res, 1997. **57**(5): p. 978-87.
281. Weber-Nordt, R.M., et al., *Constitutive activation of STAT proteins in primary lymphoid and myeloid leukemia cells and in Epstein-Barr virus (EBV)-related lymphoma cell lines.* Blood, 1996. **88**(3): p. 809-16.
282. Zhang, Q., et al., *Activation of Jak/STAT proteins involved in signal transduction pathway mediated by receptor for interleukin 2 in malignant T lymphocytes derived from cutaneous anaplastic large T-cell lymphoma and Sezary syndrome.* Proc Natl Acad Sci U S A, 1996. **93**(17): p. 9148-53.
283. Morikawa, T., et al., *STAT3 expression, molecular features, inflammation patterns, and prognosis in a database of 724 colorectal cancers.* Clin Cancer Res, 2011. **17**(6): p. 1452-62.
284. Bromberg, J.F., et al., *Stat3 activation is required for cellular transformation by v-src.* Mol Cell Biol, 1998. **18**(5): p. 2553-8.
285. Turkson, J., et al., *Stat3 activation by Src induces specific gene regulation and is required for cell transformation.* Mol Cell Biol, 1998. **18**(5): p. 2545-52.
286. Watson, C.J. and W.R. Miller, *Elevated levels of members of the STAT family of transcription factors in breast carcinoma nuclear extracts.* Br J Cancer, 1995. **71**(4): p. 840-4.
287. Berclaz, G., et al., *EGFR dependent expression of STAT3 (but not STAT1) in breast cancer.* Int J Oncol, 2001. **19**(6): p. 1155-60.
288. Dolled-Filhart, M., et al., *Tissue microarray analysis of signal transducers and activators of transcription 3 (Stat3) and phospho-Stat3 (Tyr705) in node-negative breast cancer shows nuclear localization is associated with a better prognosis.* Clin Cancer Res, 2003. **9**(2): p. 594-600.
289. Iliopoulos, D., H.A. Hirsch, and K. Struhl, *An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation.* Cell, 2009. **139**(4): p. 693-706.
290. Burke, W.M., et al., *Inhibition of constitutively active Stat3 suppresses growth of human ovarian and breast cancer cells.* Oncogene, 2001. **20**(55): p. 7925-34.
291. Garcia, R., et al., *Constitutive activation of Stat3 by the Src and JAK tyrosine kinases participates in growth regulation of human breast carcinoma cells.* Oncogene, 2001. **20**(20): p. 2499-513.
292. Dechow, T.N., et al., *Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C.* Proc Natl Acad Sci U S A, 2004. **101**(29): p. 10602-7.
293. Bromberg, J.F., et al., *Stat3 as an oncogene.* Cell, 1999. **98**(3): p. 295-303.
294. Hirsch, H.A., et al., *A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases.* Cancer Cell, 2010. **17**(4): p. 348-61.

295. Soule, H.D., et al., *Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10.* Cancer Res, 1990. **50**(18): p. 6075-86.
296. Aziz, N., H. Cherwinski, and M. McMahon, *Complementation of defective colony-stimulating factor 1 receptor signaling and mitogenesis by Raf and v-Src.* Mol Cell Biol, 1999. **19**(2): p. 1101-15.
297. Iliopoulos, D., et al., *STAT3 activation of miR-21 and miR-181b-1 via PTEN and CYLD are part of the epigenetic switch linking inflammation to cancer.* Mol Cell, 2010. **39**(4): p. 493-506.
298. Iliopoulos, D., H.A. Hirsch, and K. Struhl, *Metformin decreases the dose of chemotherapy for prolonging tumor remission in mouse xenografts involving multiple cancer cell types.* Cancer Res, 2011. **71**(9): p. 3196-201.
299. Hutchins, A.P., S. Poulain, and D. Miranda-Saavedra, *Genome-wide analysis of STAT3 binding in vivo predicts effectors of the anti-inflammatory response in macrophages.* Blood, 2012. **119**(13): p. e110-9.
300. Kwon, H., et al., *Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors.* Immunity, 2009. **31**(6): p. 941-52.
301. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.* Genome Res, 2007. **17**(6): p. 877-85.
302. Nagy, P.L., et al., *Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin.* Proc Natl Acad Sci U S A, 2003. **100**(11): p. 6364-9.
303. Polach, K.J. and J. Widom, *Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation.* J Mol Biol, 1995. **254**(2): p. 130-49.
304. Solomon, M.J. and A. Varshavsky, *Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures.* Proc Natl Acad Sci U S A, 1985. **82**(19): p. 6470-4.
305. Brutlag, D., C. Schlehuber, and J. Bonner, *Properties of formaldehyde-treated nucleohistone.* Biochemistry, 1969. **8**(8): p. 3214-8.
306. Song, L., et al., *Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.* Genome Res, 2011. **21**(10): p. 1757-67.
307. Guccione, E., et al., *Myc-binding-site recognition in the human genome is determined by chromatin context.* Nat Cell Biol, 2006. **8**(7): p. 764-70.
308. Ishihara, A., et al., *Tenascin expression in cancer cells and stroma of human breast cancer and its prognostic significance.* Clin Cancer Res, 1995. **1**(9): p. 1035-41.
309. Jahkola, T., et al., *Tenascin-C expression in invasion border of early breast cancer: a predictor of local and distant recurrence.* Br J Cancer, 1998. **78**(11): p. 1507-13.
310. Oskarsson, T., et al., *Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs.* Nat Med, 2011. **17**(7): p. 867-74.
311. Nagaharu, K., et al., *Tenascin C induces epithelial-mesenchymal transition-like change accompanied by SRC activation and focal adhesion kinase phosphorylation in human breast cancer cells.* Am J Pathol, 2011. **178**(2): p. 754-63.
312. Ilunga, K., et al., *Co-stimulation of human breast cancer cells with transforming growth factor-beta and tenascin-C enhances matrix metalloproteinase-9 expression and cancer cell invasion.* Int J Exp Pathol, 2004. **85**(6): p. 373-9.
313. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
314. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
315. Maemura, K., et al., *CLIF, a novel cycle-like factor, regulates the circadian oscillation of plasminogen activator inhibitor-1 gene expression.* J Biol Chem, 2000. **275**(47): p. 36847-51.

316. Shi, S., et al., *Circadian clock gene Bmal1 is not essential; functional replacement with its paralog, Bmal2.* Curr Biol, 2010. **20**(4): p. 316-21.
317. Takahata, S., et al., *Transcriptionally active heterodimer formation of an Arnt-like PAS protein, Arnt3, with HIF-1a, HLF, and clock.* Biochem Biophys Res Commun, 1998. **248**(3): p. 789-94.
318. Loboda, A., et al., *Diurnal variation of the human adipose transcriptome and the link to metabolic disease.* BMC Med Genomics, 2009. **2**: p. 7.
319. Hughes, M.E., J.B. Hogenesch, and K. Kornacker, *JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets.* J Biol Rhythms, 2010. **25**(5): p. 372-80.
320. Seda, O., et al., *A 14-gene region of rat chromosome 8 in SHR-derived polydactylous congenic substrain affects muscle-specific insulin resistance, dyslipidaemia and visceral adiposity.* Folia Biol (Praha), 2005. **51**(3): p. 53-61.
321. Damiola, F., et al., *Restricted feeding uncouples circadian oscillators in peripheral tissues from the central pacemaker in the suprachiasmatic nucleus.* Genes Dev, 2000. **14**(23): p. 2950-61.
322. Glossop, N.R. and P.E. Hardin, *Central and peripheral circadian oscillator mechanisms in flies and mammals.* J Cell Sci, 2002. **115**(Pt 17): p. 3369-77.
323. Hastings, M.H., A.B. Reddy, and E.S. Maywood, *A clockwork web: circadian timing in brain and periphery, in health and disease.* Nat Rev Neurosci, 2003. **4**(8): p. 649-61.
324. Sahar, S. and P. Sassone-Corsi, *Metabolism and cancer: the circadian clock connection.* Nat Rev Cancer, 2009. **9**(12): p. 886-96.
325. Ozturk, N., et al., *Loss of cryptochrome reduces cancer risk in p53 mutant mice.* Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2841-6.
326. Lee, J.H. and A. Sancar, *Regulation of apoptosis by the circadian clock through NF-kappaB signaling.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 12036-41.
327. Stevens, R.G., *Light-at-night, circadian disruption and breast cancer: assessment of existing evidence.* Int J Epidemiol, 2009. **38**(4): p. 963-70.
328. Stevens, R.G., *Circadian disruption and breast cancer: from melatonin to clock genes.* Epidemiology, 2005. **16**(2): p. 254-8.
329. Schernhammer, E.S., et al., *Rotating night shifts and risk of breast cancer in women participating in the nurses' health study.* J Natl Cancer Inst, 2001. **93**(20): p. 1563-8.
330. Hansen, J., *Increased breast cancer risk among women who work predominantly at night.* Epidemiology, 2001. **12**(1): p. 74-7.
331. Chen, S.T., et al., *Deregulated expression of the PER1, PER2 and PER3 genes in breast cancers.* Carcinogenesis, 2005. **26**(7): p. 1241-6.
332. Hoffman, A.E., et al., *The circadian gene NPAS2, a putative tumor suppressor, is involved in DNA damage response.* Mol Cancer Res, 2008. **6**(9): p. 1461-8.
333. Sephton, S.E., et al., *Diurnal cortisol rhythm as a predictor of breast cancer survival.* J Natl Cancer Inst, 2000. **92**(12): p. 994-1000.
334. Mormont, M.C., et al., *Marked 24-h rest/activity rhythms are associated with better quality of life, better response, and longer survival in patients with metastatic colorectal cancer and good performance status.* Clin Cancer Res, 2000. **6**(8): p. 3038-45.
335. Ticher, A., et al., *The pattern of hormonal circadian time structure (acrophase) as an assessor of breast-cancer risk.* Int J Cancer, 1996. **65**(5): p. 591-3.
336. Gery, S., et al., *The clock gene Per2 links the circadian system to the estrogen receptor.* Oncogene, 2007. **26**(57): p. 7916-20.
337. Couse, J.F. and K.S. Korach, *Estrogen receptor null mice: what have we learned and where will they lead us?* Endocr Rev, 1999. **20**(3): p. 358-417.

338. Shin, A., et al., *Estrogen receptor alpha gene polymorphisms and breast cancer risk.* Breast Cancer Res Treat, 2003. **80**(1): p. 127-31.
339. McGuire, W.L., *Current status of estrogen receptors in human breast cancer.* Cancer, 1975. **36**(2): p. 638-44.
340. Cecon, E., et al., *Daily variation of constitutively activated nuclear factor kappa B (NFKB) in rat pineal gland.* Chronobiol Int, 2010. **27**(1): p. 52-67.
341. Chuang, J.I., et al., *Effect of melatonin on NF-kappa-B DNA-binding activity in the rat spleen.* Cell Biol Int, 1996. **20**(10): p. 687-92.
342. Kassed, C.A. and M. Herkenham, *NF-kappaB p50-deficient mice show reduced anxiety-like behaviors in tests of exploratory drive and anxiety.* Behav Brain Res, 2004. **154**(2): p. 577-84.
343. Monje, F.J., et al., *Constant darkness induces IL-6-dependent depression-like behavior through the NF-kappaB signaling pathway.* J Neurosci, 2011. **31**(25): p. 9075-83.
344. Iliopoulos, D., et al., *Loss of miR-200 inhibition of Suz12 leads to polycomb-mediated repression required for the formation and maintenance of cancer stem cells.* Mol Cell, 2010. **39**(5): p. 761-72.
345. Bonnet, D. and J.E. Dick, *Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell.* Nat Med, 1997. **3**(7): p. 730-7.
346. Al-Hajj, M., et al., *Prospective identification of tumorigenic breast cancer cells.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3983-8.
347. Singh, S.K., et al., *Identification of human brain tumour initiating cells.* Nature, 2004. **432**(7015): p. 396-401.
348. Matsui, W., et al., *Characterization of clonogenic multiple myeloma cells.* Blood, 2004. **103**(6): p. 2332-6.
349. Li, C., et al., *Identification of pancreatic cancer stem cells.* Cancer Res, 2007. **67**(3): p. 1030-7.
350. Dalerba, P., et al., *Phenotypic characterization of human colorectal cancer stem cells.* Proc Natl Acad Sci U S A, 2007. **104**(24): p. 10158-63.
351. O'Brien, C.A., et al., *A human colon cancer cell capable of initiating tumour growth in immunodeficient mice.* Nature, 2007. **445**(7123): p. 106-10.
352. Ricci-Vitiani, L., et al., *Identification and expansion of human colon-cancer-initiating cells.* Nature, 2007. **445**(7123): p. 111-5.
353. Milde-Langosch, K., *The Fos family of transcription factors and their role in tumourigenesis.* Eur J Cancer, 2005. **41**(16): p. 2449-61.
354. Miller, A.D., T. Curran, and I.M. Verma, *c-fos protein can induce cellular transformation: a novel mechanism of activation of a cellular oncogene.* Cell, 1984. **36**(1): p. 51-60.
355. Milde-Langosch, K., et al., *The role of the AP-1 transcription factors c-Fos, FosB, Fra-1 and Fra-2 in the invasion process of mammary carcinomas.* Breast Cancer Res Treat, 2004. **86**(2): p. 139-52.
356. Belguise, K., et al., *FRA-1 expression level regulates proliferation and invasiveness of breast cancer cells.* Oncogene, 2005. **24**(8): p. 1434-44.
357. Ayroldi, E., et al., *Glucocorticoid-induced leucine zipper inhibits the Raf-extracellular signal-regulated kinase pathway by binding to Raf-1.* Mol Cell Biol, 2002. **22**(22): p. 7929-41.
358. Ayroldi, E., et al., *GILZ mediates the antiproliferative activity of glucocorticoids by negative regulation of Ras signaling.* J Clin Invest, 2007. **117**(6): p. 1605-15.
359. Ayroldi, E., et al., *Modulation of T-cell activation by the glucocorticoid-induced leucine zipper factor via inhibition of nuclear factor kappaB.* Blood, 2001. **98**(3): p. 743-53.
360. D'Adamio, F., et al., *A new dexamethasone-induced gene of the leucine zipper family protects T lymphocytes from TCR/CD3-activated cell death.* Immunity, 1997. **7**(6): p. 803-12.

361. Ayroldi, E. and C. Riccardi, *Glucocorticoid-induced leucine zipper (GILZ): a new important mediator of glucocorticoid action.* FASEB J, 2009. **23**(11): p. 3649-58.
362. Beaulieu, E. and E.F. Morand, *Role of GILZ in immune regulation, glucocorticoid actions and rheumatoid arthritis.* Nat Rev Rheumatol, 2011. **7**(6): p. 340-8.
363. Barnes, P.J., *Anti-inflammatory actions of glucocorticoids: molecular mechanisms.* Clin Sci (Lond), 1998. **94**(6): p. 557-72.
364. Beaulieu, E., et al., *Glucocorticoid-induced leucine zipper is an endogenous antiinflammatory mediator in arthritis.* Arthritis Rheum, 2010. **62**(9): p. 2651-61.
365. Yang, Y.H., et al., *Annexin-1 regulates macrophage IL-6 and TNF via glucocorticoid-induced leucine zipper.* J Immunol, 2009. **183**(2): p. 1435-45.
366. Cannarile, L., et al., *Glucocorticoid-induced leucine zipper is protective in Th1-mediated models of colitis.* Gastroenterology, 2009. **136**(2): p. 530-41.
367. Salmon-Divon, M., et al., *PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci.* BMC Bioinformatics, 2010. **11**: p. 415.
368. Saeed, A.I., et al., *TM4 microarray software suite.* Methods Enzymol, 2006. **411**: p. 134-93.
369. Saeed, A.I., et al., *TM4: a free, open-source system for microarray data management and analysis.* Biotechniques, 2003. **34**(2): p. 374-8.
370. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
371. Zeileis, A. and G. Grothendieck, *zoo: S3 Infrastructure for Regular and Irregular Time Series.* Journal of Statistical Software, 2005. **14**(6): p. 1-27.
372. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*, ed. R. Gentleman, K. Hornik, and G. Parmigiani2009: Springer. 212.
373. Li, X.Y., et al., *The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding.* Genome Biol, 2011. **12**(4): p. R34.
374. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
375. Sekinger, E.A., Z. Moqtaderi, and K. Struhl, *Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast.* Mol Cell, 2005. **18**(6): p. 735-48.
376. Lieb, J.D., et al., *Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association.* Nat Genet, 2001. **28**(4): p. 327-34.
377. Dame, R.T., *The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin.* Mol Microbiol, 2005. **56**(4): p. 858-70.
378. Wade, J.T., et al., *Genomic analysis of LexA binding reveals the permissive nature of the Escherichia coli genome and identifies unconventional target sites.* Genes Dev, 2005. **19**(21): p. 2619-30.
379. Hagiwara, M., et al., *Coupling of hormonal stimulation and transcription via the cyclic AMP-responsive factor CREB is rate limited by nuclear entry of protein kinase A.* Mol Cell Biol, 1993. **13**(8): p. 4852-9.
380. Wu, W., et al., *Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration.* Genome Res, 2011. **21**(10): p. 1659-71.
381. John, S., et al., *Chromatin accessibility pre-determines glucocorticoid receptor binding patterns.* Nat Genet, 2011. **43**(3): p. 264-8.
382. Koh, F.M., et al., *Parallel gateways to pluripotency: open chromatin in stem cells and development.* Curr Opin Genet Dev, 2010. **20**(5): p. 492-9.