Cell, Vol. 116, 499–509, February 20, 2004, Copyright ©2004 by Cell Press

Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs

Simon Cawley,[1,5] Stefan Bekiranov,[1,5] Huck H. Ng,[2,3,4] Philipp Kapranov,[1] Edward A. Sekinger,[2] Dione Kampa,[1] Antonio Piccolboni,[1] Victor Sementchenko,[1] Jill Cheng,[1] Alan J. Williams,[1] Raymond Wheeler,[1] Brant Wong,[1] Jorg Drenkow,[1] Mark Yamanaka,[1] Sandeep Patel,[1] Shane Brubaker,[1] Hari Tammana,[1] Gregg Helt,[1] Kevin Struhl,[2,*] and Thomas R. Gingeras[1,*]

[1]Affymetrix
3380 Central Expressway
Santa Clara, California 95051
[2]Deptartment of Biological Chemistry
and Molecular Pharmacology
Harvard Medical School
Boston, Massachusetts 02115
[3]Department of Biological Sciences
National University of Singapore
Singapore 117543
[4]Genome Institute of Singapore
Singapore 138672

Summary

Using high-density oligonucleotide arrays representing essentially all nonrepetitive sequences on human chromosomes 21 and 22, we map the binding sites in vivo for three DNA binding transcription factors, Sp1, cMyc, and p53, in an unbiased manner. This mapping reveals an unexpectedly large number of transcription factor binding site (TFBS) regions, with a minimal estimate of 12,000 for Sp1, 25,000 for cMyc, and 1600 for p53 when extrapolated to the full genome. Only 22% of these TFBS regions are located at the 5′ termini of protein-coding genes while 36% lie within or immediately 3′ to well-characterized genes and are significantly correlated with noncoding RNAs. A significant number of these noncoding RNAs are regulated in response to retinoic acid, and overlapping pairs of protein-coding and noncoding RNAs are often coregulated. Thus, the human genome contains roughly comparable numbers of protein-coding and noncoding genes that are bound by common transcription factors and regulated by common environmental signals.

Introduction

The transcriptome of the human genome is composed of discrete RNA molecules serving multiple functions including encoding protein information, signaling, structural support of subcellular elements, and transcriptional and posttranscriptional regulatory agents. A considerable amount of information concerning the protein coding transcripts of the genome has been collected and assembled with the near completed draft of the human genome (Lander et al., 2001; Venter et al., 2001). Recently, there have been several reports based on empirical and computational evidence indicating that the transcriptome is larger and more complex than first considered.

A detailed map has been published providing the locations of RNA transcription along human chromosomes 21 and 22 using cytoplasmic, poly(A)$^+$ RNA from a collection of 11 developmentally diverse cell lines (Kapranov et al., 2002). These maps indicated that there is as much as an order of magnitude more transcription along these two chromosomes than could be accounted for by the then current genomic annotations of protein coding RNAs, suggesting a hitherto hidden population of RNA transcripts. Analysis of a representative collection of these novel transcripts revealed that they possess little protein coding potential and some occupy an antisense orientation relative to well-characterized coding transcripts. Similar results were presented in two reports describing the results of in depth serial analysis of gene expression (SAGE) of a collection of cDNA libraries (Chen et al., 2002; Saha et al., 2002), and most recently using arrays of serial PCR fragments covering the nonrepetitive human chromosome 22 genomic sequences (Rinn et al., 2003). We will refer to this general group of unannotated RNAs as noncoding, although some of these RNAs are likely to encode short proteins.

Interestingly, the comparative sequence analysis of the recently completed mouse genome with the human genome resulted in an unexpected observation that is consistent with these previous empirical observations (Waterston et al., 2002). Comparative analyses of the mouse genome revealed that there is almost two times more evolutionary sequence conservation observed than expected and that these conserved sequence regions are located distal from the well-annotated exons. Dermitzakis et al. have also analyzed a subsection of the conserved murine and human sequences found on human chromosome 21 (mouse chromosomes 10,16,17) for evidence of transcription emanating from these regions (Dermitzakis et al., 2002). Approximately 37% (837) of the conserved sequence regions (>100 bp and >70% identity) on chromosomes 10, 16, and 17 were determined to be transcribed. Most recently, analysis by the FANTOM consortium and the RIKEN genome research group of the mouse transcriptome revealed that 15,923 of the 60,770 full-length clones that they isolated are novel, with 11,665 (73%) of these being noncoding transcripts (Okazaki et al., 2002). A total of 2,431 of these reported transcripts were noted as antisense transcripts. These empirical and computational findings all point to a large and as yet poorly understood population of RNAs in the transcriptome. At present, however, functional attributes of these RNAs have yet to be demonstrated.

The high-density, tiled arrays used to map the RNA transcripts along human chromosomes 21 and 22 contain on average one oligonucleotide pair every 35 bp and represent essentially all nonrepetitive sequences

*Correspondence: tom_gingeras@affymetrix.com (T.R.G.), Kevin@hms.harvard.edu (K.S.)
[5]These authors contributed equally to this work

of these chromosomes. In combination with chromatin immunoprecipitation (ChIP), these arrays should permit the identification of physiological target sites of transcriptional regulatory proteins in an unbiased and comprehensive manner. ChIP has been used to analyze the location of transcriptional regulatory proteins in yeast cells using arrays with PCR products containing all intergenic regions (Horak et al., 2002a; Iyer et al., 2001; Lee et al., 2002; Ng et al., 2002; Ren et al., 2000; Simon et al., 2001; Zeitlinger et al., 2003). However, the few papers combining ChIP and microarrays in mammalian cells have involved selected upstream regions (Horak et al., 2002b; Mao et al., 2003; Ren et al., 2002) and hence do not address the possibility of transcription factor binding sites in other locations. This issue is particularly relevant in mammalian cells because the mRNA and protein coding sequences represent a small percentage of the total genome and because transcriptional regulatory proteins can function at long and variable distances from transcriptional initiation sites.

To further explore properties of the transcriptome and to identify functional attributes of the noncoding transcripts, binding sites for a collection of transcription factors have been mapped along chromosomes 21 and 22 in an unbiased approach, as a means of identifying possible regulatory regions for a wide variety of cellular RNAs. Interestingly, only 22% of the transcription factor binding sites (TFBS) are located at the canonical 5′ termini of well-characterized protein-coding genes, while 36% lie within of immediately 3′ to well-characterized genes and are significantly correlated with noncoding RNAs. A number of these noncoding RNAs are regulated in response to retinoic acid stimulation, and coregulation of overlapping pairs of protein-coding and noncoding RNAs occurs at a frequency significantly greater than chance. These data point to evidence that protein coding and noncoding genes have similar functional attributes regarding (1) the existence of common transcription factors in their promoter regions and (2) their ability to respond to environmental and developmental conditions, which together suggest that that they may be controlled by the same transcriptional regulatory machinery. These functional attributes argue against the idea that these noncoding RNAs merely represent transcriptional noise, but instead suggest that they may have biological functions.

Results

Identification of Transcription Factor Binding Sites along Chromosomes 21 and 22

By combining chromatin immunoprecipitation and high-density oligonucleotide arrays interrogating the nonrepeat genomic sequences of chromosomes 21 and 22 at 35 base pair (bp) resolution (Kapranov et al., 2002), the positions of binding for three human transcription factors (TFs), cMyc, Sp1, and p53, were determined within two cell lines (cMyc and Sp1 in Jurkat, p53 in HCT1116). A total of 353, 756, and 48 high confidence (p value $< 10^{-5}$) binding sites were observed for Sp1, cMyc, and p53, respectively (all data is accessible at http://transcriptome.affymetrix.com/publication/tfbs). At this stringent threshold, it is estimated that on the order of one or less of the TFBS regions is falsely detecting differential hybridization (Supplemental Figure S1 at http://www.cell.com/cgi/content/full/116/4/499/DC1). Considering the stringency of the threshold used, the true number of sites is expected to be even larger. These 1157 TFBS were identified by comparing the results of the binding experiments for each TF to two control experiments. In one set of control experiments, the immunoprecipitation step was omitted, in another set, an antibody to bacterial GST was employed (instead of an antibody against a transcription factor). A total of 64% of the TFBSs were commonly identified using both controls. Extrapolating these findings to the whole genome predicts a total of about 12,000 Sp1 sites, 25,000 cMyc sites, and 1600 p53 sites.

Verification by quantitative PCR was performed on a sample of the TFBS, confirming 11 of 11 p53, 6 of 6 Sp1, and 4 of 4 cMyc binding sites (Supplemental Table S1 on *Cell* website). An additional 3 of 3 p53 sites were confirmed using a different antibody, p53_DO1, which reacts with an N-terminal epitope (Supplemental Table S1 online). The TFBS detected with p53_DO1 strongly overlap those detected with the p53 full-length antibody (20 of the 48 p53 TFBS were also detected with p53_DO1). These results provide direct experimental evidence that very few of the identified TFBS are false positives.

Properties of Transcription Factor Binding Sites

For each TFBS, we examined the immunoprecipitated DNA fragments for characteristics associated with transcriptional regulatory regions (proximity to CpG islands and annotated 5′ exons and the presence of known binding motifs). A total of 43%, 24%, and 17% of the Sp1, cMyc, and p53 TFBS, respectively, were within 1 Kb of an annotated CpG island (Table 1A), constituting approximately 5.5-, 3.1-, and 2.1-fold enrichment, respectively, over what would be expected at random. Enrichment for location proximal to 5′ exons was found for Sp1 and cMyc, but not for p53 (Table 1B). Analyses of the TFBS for the presence of known TF binding motifs demonstrated substantial enrichment for all three TFs. In the case of p53, because of the almost complete lack of matches to the exact consensus and with the availability of a weight matrix from an independent study (Hoh et al., 2002), a search for inexact matches was implemented. As much as 62% of the p53 sites contain evidence for a p53 consensus sequence (greater than 3-fold enrichment over what would be expected at random). Using the motif finder, MDscan (Liu et al., 2002), the exact Sp1 consensus is recovered along with many slightly weaker variants that may well bind Sp1 (data not shown). The lack of identifiable binding motifs at some of the sites may be explained by the existence of as yet unidentified binding motifs, indirect interactions with DNA mediated by other proteins, or, potentially, crossreactivity of antibodies with other DNA binding proteins.

Other observations that follow from the mapping of TFBS for Sp1, cMyc, and p53 along chromosomes 21 and 22 are: (1) 61% of the 353 Sp1 sites overlapped predicted cMyc sites and 29% of the 756 cMyc sites overlapped an Sp1 site, suggesting coincident binding and possible coregulation at these sites. (2) The large majority of the 22% TFBS regions located at the 5′ end

Table 1. Enrichment for Promoter Characteristics Found at Predicted TFBS

(A) CpG Islands

TFBS	# Sites	# Sites within 1 Kb of CpG	% Sites within 1 Kb of CpG	Fold-Enrichment
Sp1	353	152	43	5.5
cMyc	756	182	24	3.1
P53	48	8	17	2.1

(B) 5′ Exons

TFBS	# Sites	# Sites within 1 Kb of 5′ Exon	% Sites within 1 Kb of 5′ Exon	Fold-Enrichment
Sp1	353	95	27	5.5
cMyc	756	137	18	3.7
P53	48	0	0	0

(C) Consensus Binding Motifs

TFBS	# Sites	# Motifs	# Sites with ≥ 1 Motif	% Sites with ≥ 1 Motif	# Motifs Expected at Random	Fold-Enrichment
Sp1	353	124	76	22	22	5.6
cMyc	756	611	259	33	269	2.3
P53	48	1	1	2	0.06	16.6

The proximity of predicted TFBS to CpG islands and well-characterized 5′ exons is summarized in Tables 1A and 1B. CpG island annotations were taken from the June 2002 assembly of the human genome at UCSC and 5′ exon annotations were taken from RefSeq and GenBank mRNA records annotated as having "complete CDS." Distance between a predicted TFBS and a CpG island or a 5′ exon is defined as the separation between their nearest ends. The fold enrichment is calculated by comparison with sites generated at random uniformly over the nonrepetitive regions of chromosomes 21 and 22. 7.9% of the randomly generated sites were located within 1 Kb of CpG islands, 4.9% were located within 1 Kb of 5′ exons. Table 1C presents the enrichment for the overall number of exact matches to consensus binding motifs found in the TFBS. Note that some TFBS contained more than 1 binding motif. An additional study considered inexact matches to the p53 sites.

of well-characterized genes represent novel roles for the monitored TFs. The previously described Sp1 regulation of the superoxide dismutase gene (SOD1) (Minc et al., 1999), the cMyc-regulated macrophage migration inhibitory factor (MIF) (Watson et al., 2002), a collection of 17 cMyc-responsive genes (Menssen and Hermeking, 2002), and a set of putative p53 targets (Kannan et al., 2001) represent the sum of the previously characterized genes associated with these TFs along chromosomes 21 and 22. (3) The observation of 69 Genscan-predicted genes with cMyc and/or Sp1 binding sites at their 5′ ends, along with RNA transcription data, supports the existence of these ab initio predictions. (4) Sites located outside of known annotations are likely to represent regulatory regions for novel transcripts based on the RNA transcription and EST data. Out of 21 (10%) of such regions tested, 19 are associated with transcribed regions by RT-PCR (data not shown). (5) Despite the enrichment in the predicted TFBS for characteristics typically associated with promoter regions (proximity to CpG islands and 5′ exons, presence of known binding motifs), many of the remaining TFBS identified in this study lack such properties, suggesting that TFs may be tethered to these genomic locations independently of their sequence-specific DNA binding properties.

Relationship of TFBS to RNA Transcripts

Given that many of the TFBS sites were clustered together, they were grouped into TFBS regions, defined as maximal sets of TFBS such that neighboring sites are separated by less than 1 Kb. The 1157 TFBS sites yielded 866 regions, 305 and 561 on chromosomes 21 and 22, respectively. Interestingly, 36% of these regions are situated within known genes or proximal to the 3′ most exon of a gene (Figure 1). Only 22% of the TFBS regions are located in a canonical position at the 5′ end of well-characterized genes. The location of TFBS regions within or 3′ to well-characterized genes suggests that they may represent distal regulatory elements (e.g., enhancers or silencers) or promoters for noncoding transcripts.

The 33 TFBS regions positioned either within or downstream of the last exon of annotated genes on chromosomes 21 and 22 are interesting due to the potential for antisense transcripts to overlap the 3′ UTRs of these genes (Table 2). This observation is related to the presence of short (~100 bp) conserved stretches within the 5′ and 3′ UTRs of a large number of mammalian genes and the resulting suggestion that these conserved sequences are involved in mRNA stability and possibly form duplexes with antisense transcripts (Lipman, 1997). Similarly, 1600 sense/antisense gene pairs in the genome, most of which primarily overlap 5′ and 3′ noncoding regions, have recently been identified using a computational approach (Yelin et al., 2003). Most importantly, all but two of the 33 TFBS located within or at the 3′ ends of well-characterized genes (Table 2) are positioned just upstream of possible novel transcripts based either on our earlier reported RNA transcription mapping data (Kapranov et al., 2002) or mapped ESTs. Of the 3′ TFBS regions listed in Table 2, 27% (9/33) are situated in genes having one or more of the same transcription factors at the 5′ end.

Biochemical and Computational Verification of Novel Transcripts

The structure and sequence of a subset of these novel transcripts were characterized. Prior to collecting exper-

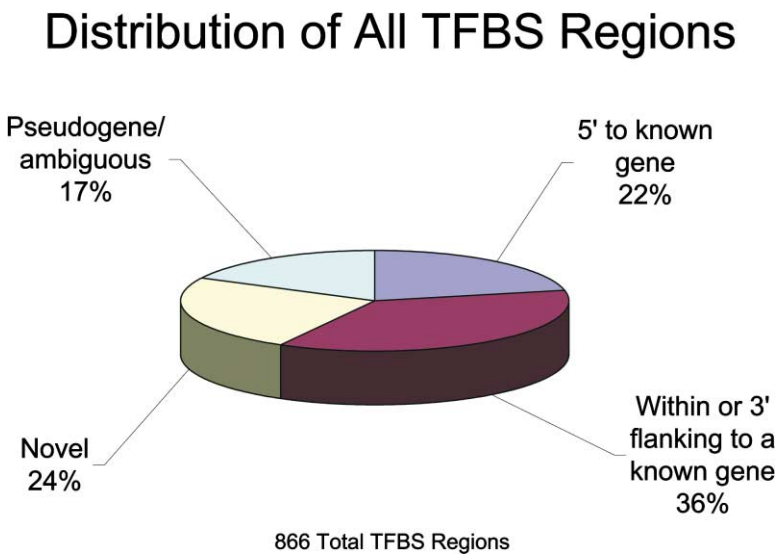

Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

imental data, preliminary evidence for the presence of novel transcripts was derived from chromosome 21 and 22 RNA maps (Kapranov et al., 2002) and from the publicly available EST data. Novel transcripts were verified using RT-PCR analyses in 9/11 regions and were found to have little coding capacity (less then 50 amino acids). Northern hybridization analysis of these isolated transcripts with strand-specific oligonucleotides or riboprobes indicate that they are polyadenylated, in some cases spliced, and are present as single and multi-exon isoforms ranging in size from 800 bp to 9 Kb (Supplemental Figure S3 on *Cell* website). Together with the strand-specific RT-PCR data, this suggests that several of them might also be antisense to known genes, such as, for example, EP300 (Figures 2C and 2D), UBASH3A (Supplemental Figures S2A and S2B online), SEC14L2 (Supplemental Figures S2C and S2D), and others.

The Ewing sarcoma gene (EWSR1) (Plougastel et al., 1993), the tumor suppressor gene, EP300 (Gayther et al., 2000), and mitogen-activated protein kinase MAPK1 (Gonzalez et al., 1992) on chromosome 22 illustrate potential utilization of common TFs to regulate both well-characterized and novel transcripts (Figure 2). Sequence analysis of the novel transcripts that overlap EWSR1 and EP300 indicate that they are spliced RNAs. Interestingly, a conserved region in the 3′ UTR of the EWSR1 gene is consistent with the evidence of antisense regulation of this gene (Lipman, 1997). The EP300 gene is a striking example (Figures 2C and 2D), having a TFBS region 17 kb away from the 3′ end and a novel transcript that splices from this site into the 3′ end of the gene. Additionally, overlapping novel transcripts from the genes encoding nuclear protein UBASH3A (Supplemental Figures S2A and S2B), phosphatidylinositol transfer-like protein SEC14L2 (Supplemental Figures S2C and S2D), TBC/rabGAP domain protein EPI64 (Supplemental Figures S2E and S2F), guanine-nucleotide exchange factor TIAM1 (Supplemental Figures S2G and S2H), KIAA0376 protein (Supplemental Figures S2I and S2J), and GTSE1 (Supplemental Figures S2K and S2L) were verified by RT-PCR and/or Northern blot analyses (Supplemental Figure S3). In many of these cases, the TFBS that are located on the 3′ end of the well-characterized gene appear to be located 5′ of the overlapping novel transcript, which suggests that these transcripts may be regulated by these factors and in precisely the same way as protein coding genes.

Additional supporting evidence that these TFs may be regulating antisense transcripts was found by relating them to full-length mRNAs and ESTs with confidently assignable strandedness (determined from splicing and polyadenylation sites and signals). 1782 clusters of transcripts were formed of well-oriented sequences from public databases aligning to chromosomes 21 or 22. Among these clusters, there was a significant association (chi-square p value $< 10^{-15}$) between the property of proximity to a noncanonical TF and the property of having evidence for transcription on the opposite strand. In this context, a noncanonical TF is one not located at the 5′ end of a known gene and evidence for transcription on the opposite strand is based on public sequence data. Twenty-one percent (363) of these transcript clusters are made up of sense antisense pairs, 44% (161) have an associated noncanonical TF. Of the 161 sense antisense pairs that have a noncanonical TF, 52% contain at least one site conserved between the human and mouse genomes based on BlastZ human-mouse alignments (Schwartz et al., 2003).

Differential Expression Patterns of Novel Transcripts

To address the issue of whether the observed overlapping noncoding transcripts are biologically important, we examined whether some of them exhibited a reproducible and coordinated program of differential expression correlated with the companion coding transcripts. The expression profiles of the poly(A)$^+$ cytosolic RNA fraction were monitored during the response of a pluripotent human germ cell tumor-derived cell line, NCCIT, which undergoes retinoic acid (RA)-induced differentiation into keratin- and neurofilament-positive somatic cells (Damjanov et al., 1993). Empirically derived transcriptional maps of NCCIT using the chromosome 21 and 22 genome tiling arrays during various stages of

Table 2. Location of 3′ TFBS Regions and Proximal Annotated Gene

Start	Stop	3′ TFBS	3′ Flanking Gene	Novel Transcriptome Evidence[a]	Novel EST Evidence[b]	5′ TFBS[c]
Chr. 21						
26953700	26954500	cMyc	C21orf6 gene (NM_016940)	Y	—	cMyc
30550300	30551000	cMyc, Sp1	C21orf59 mRNA (NM_017835)	Y	AI872268	—
31250700	31250800	cMyc	IL10RB gene (NM_000628)	Y	D20421	Sp1, cMyc
32324600	32326500	cMyc	ENST00000290310	Y	—	—
33978900	33980000	cMyc, Sp1	C21orf18 mRNA (NM_017438)	Y	—	—
34099000	34099500	cMyc	CBR3 (NM_001236)	Y	AI183799	Sp1, cMyc
35012900	35013300	Sp1, p53	DSCR5 (NM_016430)	Y	—	cMyc
35170600	35171200	cMyc, Sp1	DSCR3 (NM_006052)	Y	BI769073	Sp1, cMyc
37611800	37612200	cMyc	B3GALT5 (NM_006057)	Y	—	—
40264500	40265500	cMyc	ABCG1 (NM_004915)	Y	—	cMyc
40328700	40329500	cMyc	TFF1 (NM_003225)	Y	AA632099	—
40415000	40418100	cMyc, Sp1	UBASH3A (NM_018961)	Y	—	cMyc
41144500	41145300	cMyc, Sp1	CRYAA (NM_000394)	Y	—	—
42093800	42094500	Sp1	ENST00000291578	Y	BG679497	—
43871400	43871600	cMyc	PCBP3 (NM_020528)	Y	—	—
Chr. 22						
15149900	15150300	cMyc	MIL1 (NM_015367)	Y	—	Sp1[d]
16100000	16100900	cMyc	SLC25A1 (NM_005984)	Y	—	Sp1
18699300	18699600	cMyc	SDF2L1(NM_022044)	N	—	Sp1, cMyc
20933500	20934100	cMyc	MIF (NM_002415)	N	—	—
21036700	21037300	cMyc	HS322B1A (NM_015371)	Y	—	—
22328100	22328500	Sp1	CRYBB2 (NM_000496)	Y	—	—
26259200	26260100	cMyc	KREMEN (NM_032045)	Y	—	—
27354700	27355700	cMyc	OSM (NM_020530)	Y	—	cMyc
28445500	28446000	cMyc	Sanger gene (AC005003.C22.4)	Y	—	—
28775000	28775800	p53	Sanger gene dJ694E4.C22.2	Y	—	Sp1, cMyc
32590700	32592300	Sp1	RASD2 (NM_014310)	Y	—	—
32773900	32774700	Sp1, p53	Sanger gene dJ41P2.C22.1	Y	—	—
36399500	36399700	cMyc	SYNGR1 (NM_004711)	Y	BE348322, H67984	—
38412300	38412400	cMyc	Sanger gene dJ979N1.C22.1	Y	—	cMyc
47554600	47555500	Sp1	MAPK8IP2 (NM_012324)	Y	—	—
47563700	47564300	cMyc	ARSA (NM_000487)	Y	—	Sp1
47689000	47689300	Sp1	ACR (NM_001097)	Y	—	—
47690600	47691600	Sp1, p53	ACR (NM_001097)	Y	—	—

[a] Evidence of novel transcription based upon RNA transcription mapping data (Kapranov et al., 2002).
[b] Evidence of transcription based upon EST data.
[c] Identity of 5′ TFBS of the indicated gene, if present.
[d] Relative to largest annotated transcript.
Coordinates of TFBS regions located 3′ to annotated genes (RefSeq, Sanger, and Ensembl). The TFs detected in these regions and the gene to which the region is proximal are indicated in the third and fourth columns. Evidence of novel transcription based on transcriptome evidence (Kapranov et al., 2002) and on EST data is determined by inspecting for possible transcriptional activity and by the alignment of antisense ESTs outside of annotated exons. The last column indicates if a TFBS was also detected at the 5′ end of the proximal gene.

differentiation were generated with 4, 24, 96, and 366 hr of stimulation with 10 μM of retinoic acid. When all time points are considered, approximately 6% of the protein-coding and 6% of the noncoding RNAs were induced greater than 2-fold in response to retinoic acid. In addition, approximately 9% of protein-coding and 17% of noncoding RNAs are downregulated more than 2-fold in response to retinoic acid. Thus, the noncoding RNA population responds to RA-induced differentiation in a manner very similar to the response of the coding genes.

Probes overlapping coding/noncoding transcript pairs identified from the public databases were used to monitor potential coordinate regulation. For each transcript partner of a coding/noncoding pair, as well as for each time point, the median fold-change with respect to a control for all probes interrogating the transcript was evaluated. From the observed fold-change, a correlation between coding and noncoding paired transcripts was computed.

The average correlation was significantly larger than can be explained by chance alone (Figure 3), suggesting the existence of a subpopulation of transcripts where the coding and noncoding transcripts are positively correlated. The positive correlation is consistent with a coordinated regulation of coding and noncoding transcription. We note that of approximately 10% (21/214) of the overlapping coding/noncoding gene pairs that respond during the timecourse of retinoic acid stimulation, only a few show anti-correlated regulation, which is expected if, for example, the noncoding transcripts are silencing the coding transcripts. The positive coordinated expression of coding and noncoding transcripts points to a supportive function for the expression of the responding

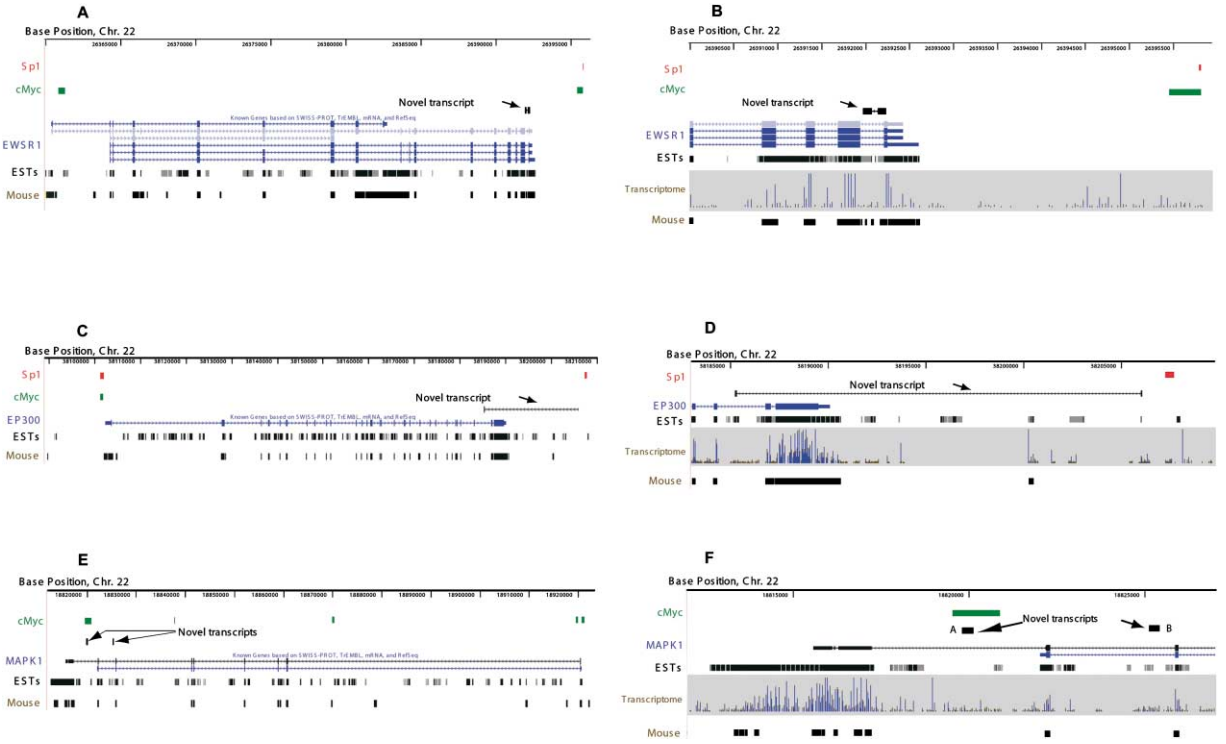

Figure 2. Overlapping Coding and Noncoding Transcripts Potentially Regulated by the Same Transcription Factors

Positions of 5′, 3′, and internal binding sites for Sp1 (red), c-Myc (green), or p53 (blue) are shown for the EWSR1 (A and B), EP300 (C and D), and MAPK1 (E and F) genes. The exon-intron structure of each gene, known isoforms, all currently available human EST annotations, and the tight subset of human-mouse BLASTZ alignments are shown on the June 2002 version of the genome. The locations of experimentally determined novel, noncoding transcripts are indicated by the arrows. Oligonucleotide or RNA probes used in the Northern experiments in Supplemental Figure S3 were derived from these regions. In the cases of EWSR1 and EP300, the presence of novel antisense transcripts is also supported by spliced ESTs AI687358 (EWSR1) and AW511192 and AA889875 (EP300). The enlarged views of the regions containing noncoding transcripts are shown on the right along with earlier reported RNA transcription mapping data (Kapranov et al., 2002).

genes and also shows that such transcription of noncoding genes is not a random and unguided response. A second interesting observation stemming from this result is that the locations of these polyadenylated coding and noncoding transcripts are destined for the cytosol, from which they were isolated. The models recently proposed by Carmichael (Carmichael, 2003) and others point to a nuclear location for the possible negative regulatory functions of long antisense RNAs (i.e., non-siRNAs and non-miRNAs).

Discussion

Expected and Unexpected Locations of Transcription Factor Binding Sites in the Human Genome

An unbiased mapping of Sp1, cMyc, and p53 along human chromosomes 21 and 22 has identified a surprisingly large number of transcription factor binding sites (TFBS) that are occupied in living cells (Figure 4). These TFBS were identified using a stringent p value threshold, and direct verification experiments indicate that few of the identified sites are false positives and that the number of true sites is likely to be higher. Strikingly, extrapolation of these results for only three transcription factors to the entire human genome provides a possible number of regulatory elements (tens of thousands), which approaches the estimated number of annotated coding genes. This unexpected number of TFBS further highlights the presence of the many as yet unannotated genes. However, many of the TFBS are in accord with expectations in that we observed statistically significant enrichment of these TFBS near 5′ promoters of coding genes and CpG islands as well as enrichment of their expected consensus sequence compared to random chance. TFBS tend to cluster along the genome as might be expected of modular genetic control regions. For example, we observe a very strong overlap of cMyc and Sp1 binding regions, in accord with several characterized promoters with both cMyc and Sp1 sites.

A second striking observation is that TFBS bound to classically defined promoter regions represent a clear minority of genomic binding sites in living cells. Although it is highly likely that the specific genomic profile of transcription factor binding will vary considerably among different cell types, we believe that the transcription factors and cells used in the experiments here are representative. Thus, we strongly suspect that the results we have obtained for p53, cMyc, and Sp1 will generally apply to other transcription factors and other cell types. These results emphasize the value of using tiled microarrays representing complete genomic re-

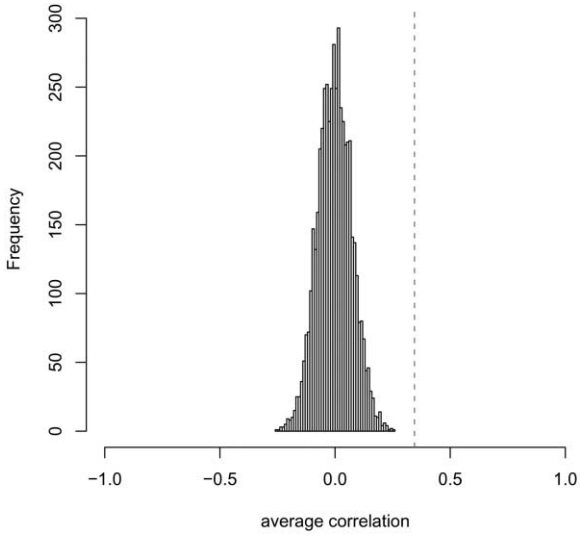

Figure 3. Correlation of Coding/Noncoding Transcript Pairs

Coding/noncoding transcript pairs were formed from public EST and mRNA databases, and correlation in fold-change between the coding and noncoding transcripts was monitored. Among the 29% of the pairs that exhibited differential behavior in the experiment, the average Pearson correlation was 0.34. To assess significance, a bootstrap approach was used in which the time points were randomly permuted and the correlation was recomputed, yielding a two-tailed p value of less than 0.0002. The figure presents a histogram of the 5000 bootstrapped average correlations, and the red line represents the actual observed average correlation.

gions as opposed to arrays restricted to promoters or other selected genomic regions.

Functional Attributes of Noncoding RNAs Suggest that They Are Expressed and Regulated Similarly to Protein-Coding Genes

Although a large number of novel, noncoding RNAs have been detected in the human genome (Chen et al., 2002; Kapranov et al., 2002; Saha et al., 2002), it has been unclear whether these RNAs represent transcriptional noise or functional entities. Here, we demonstrate three functional attributes of these noncoding transcripts. First, many of the noncoding RNAs are associated with nearby TFBS, and we note that Sp1 typically binds and functions at proximal 5′ promoter regions. Second, many of the TFBS regions associated with human noncoding RNAs are conserved in the mouse genome. Third, a significant number of the identified noncoding transcripts are regulated by retinoic acid-induced differentiation. Presumably, much of this transcriptional regulation is mediated by binding of retinoic acid receptors and/or transcription factors induced by these receptors in the promoter regions of the novel transcripts. Importantly, if one compares the sets of non-coding and well-characterized protein-coding RNAs on chromosomes 21 and 22, the number of genes bound by an individual transcription factor and the number of genes regulated by retinoic acid are roughly comparable. Thus, the noncoding transcripts share important functional characteristics of protein-coding genes, and indeed the two sets of transcripts cannot be distinguished by the functional criteria used here.

Although the presence of a TFBS does not necessarily imply a direct effect of the transcription factor on the expression of the gene, it is striking that so many noncoding RNAs are associated with TFBS regions and that roughly comparable numbers of binding sites for a given factor are found at noncoding and protein-coding genes. While the contribution of individual binding sites to the transcription of the nearby gene will certainly vary from gene to gene, this issue is similar for both protein-coding and noncoding genes. Thus, it is difficult to imagine that hundreds of binding sites in the vicinity of the noncoding RNAs represent random and functionally meaningless occurrences. Furthermore, a significant number of noncoding transcripts are regulated by retinoic acid, and coregulation of overlapping noncoding and protein-coding genes occurs at a significantly higher frequency than expected by chance. Such coregulation of overlapping genes was unexpected, and it seems very unlikely that it represents transcriptional noise.

Our analysis was limited to a set of three transcription factors and one environmental induction condition. Thus, it is completely expected that not all noncoding RNAs are associated with the transcription factors tested and that only a minority (20%) of the noncoding RNAs are regulated by retinoic acid. These considerations apply equally to protein-coding and noncoding genes, and these two sets of genes are similar with respect to the functional criteria tested. As the limited set of transcription factors and environmental conditions tested were chosen without preconceived notions of the results, it is highly likely that the general nature of the observations extend to other transcription factors and other conditions of environmental change. Taken together, our results strongly suggest that the large population of noncoding RNAs are expressed and regulated by similar molecular mechanisms that are involved in the control of protein-coding RNAs. Furthermore, the specific functional attributes associated with such a large number of noncoding RNAs on chromosomes 21 and 22 (and by extension the entire human genome) strongly argue that many of these RNAs have biological functions and are not physiological artifacts.

Potential Biological Functions of Noncoding RNAs

There are many possible biological functions of the noncoding RNAs, and elucidating these functions will require detailed molecular and genetic analysis of specific transcripts. One potential function is suggested by the numerous noncoding, antisense transcripts that overlap RNAs corresponding to protein-coding genes. Many reports demonstrate that naturally occurring antisense transcription regulates prokaryotic gene expression (Wagner and Simons, 1994), and there has been an increasing appreciation that antisense transcription plays important functions in eukaryotic cells (Yelin et al., 2003). Examples of this include X chromosome inactivation (Brockdorff et al., 1992; Brown et al., 1992), small noncoding RNA control of gene silencing in *C. elegans* (Ashrafi et al., 2003; Kamath et al., 2003; Lee and Ambros, 2001) and plants (Hamilton and Baulcombe, 1999), gene imprinting (Reik and Walter, 2001; Sleutels and Barlow, 2002), and more recently, evidence of individual gene regulation (Kramer et al., 2003; Solymar et al., 2002). In

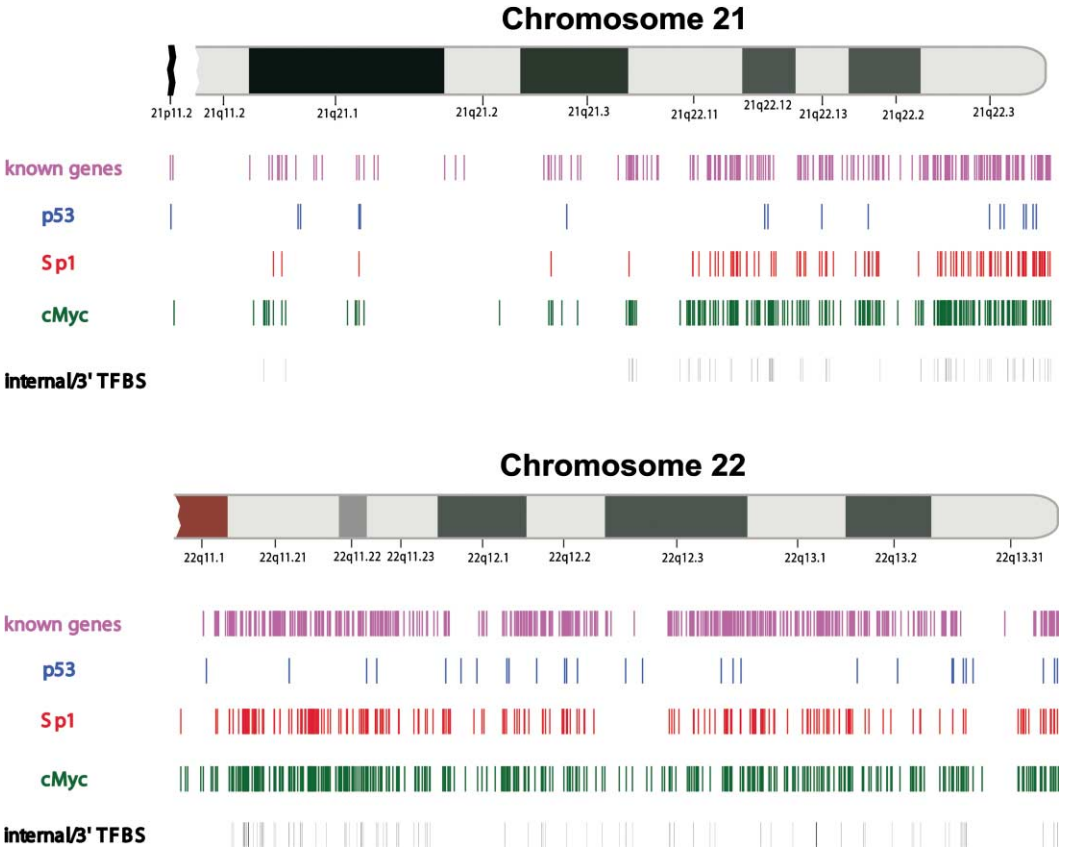

Figure 4. Chromosomal Maps of TFBS

Positions of the TFBS detected for p53, Sp1, and cMyc along chromosomes 21 and 22. The track labeled "known genes" depicts 5′ ends of genes from the "known genes" track on the UCSC genome browser and is based on SWISS-PROT, TrEMBL, GenBank mRNA, and RefSeq annotations. The "internal/3′ TFBS" track presents the locations of the TFBS located within or 3′ to annotated genes, as reported in Table 2 and Supplemental Table S2 online.

addition to these examples, studies undertaken to survey the prevalence of noncoding transcripts on a genome-wide basis have revealed the widespread occurrence of antisense transcripts (Lehner et al., 2002; Patankar et al., 2001; Wagner and Flardh, 2002; Yelin et al., 2003). The empirically derived results of this study indicate a higher proportion of overlapping gene pairs than the 8% based on computational analysis (Yelin et al., 2003). The primary reason for this discrepancy may be that intron regions were not considered in this earlier study and that many regions of the genome are transcribed but are not represented in the annotations or EST databases. As our results indicate that many noncoding, antisense RNA are associated with transcription factors and are transcriptionally regulated, some of these antisense RNAs may have biological functions relating to a classical antisense mechanism. However, the apparent absence of negative correlation in the expression levels of coding-noncoding pairs suggests that the functional relationship of antisense transcripts is more complicated than a relatively simple model of reciprocal inhibition.

Our unexpected observation that coregulation of overlapping pairs of coding and noncoding transcripts occurs more frequently than expected by chance is suggestive of a coordinated expression strategy across the entire genome. Mechanistically, such coregulation may be related to our observation that a common transcription factor is often associated with promoters of both the protein-coding and noncoding transcript of an overlapping pair. Alternatively, coregulation may be achieved via a large-scale chromatin domain that includes both of the overlapping transcripts and that is either accessible or inaccessible in a regulated fashion. Why are these noncoding transcripts found in the cytoplasm if they are synthesized coordinately with the coding transcripts in the nucleus and have the potential to form double-stranded structures? It may be that the noncoding transcripts do not directly regulate the coding gene, but are rather involved in the same pathway as the coding gene itself. Overlapping coding/noncoding gene architecture may serve to facilitate the concordant regulation of both groups of transcripts similar to an RNA-based signaling network model (Mattick, 2001).

The human genome has tens of thousands of noncoding transcripts, and as a general class, they behave similarly to protein-coding genes with respect to the presence of TFBS and the ability to be regulated by environmental signals. By analogy with protein-coding genes, it seems likely that some, and perhaps many, of the noncoding RNAs may have biological functions that are unrelated to those of overlapping or neighboring

protein-coding genes. In this regard, may of the noncoding RNAs do not map near protein-coding genes, yet they have similar properties to noncoding RNAs that are part of overlapping pairs. Lastly, it is possible that biological functions associated with some of these noncoding RNAs might not be due to the RNA products per se, but rather to the transcriptional process itself, which is likely to alter chromatin structure within and adjacent to the transcribed regions. These data encourage future mechanistic investigations and discussions concerning the definition of a gene, the precise role of transcription factor binding proteins, and the possible reconsideration of current concepts of the structure and functioning of a eukaryotic genome.

Experimental Procedures

Chromatin Immunoprecipitation Protocol

Cell lines HCT116 (ATCC CCL-247) and Jurkat (ATTC TIB-152) were grown to a density of 0.4×10^6 cells/ml. HCT116 cells were treated with 5 mU/ml of bleomycin for 24 hr to induce expression of wild-type p53. Cells were fixed with 1% formaldehyde for 10 min at room temperature, with occasional swirling. Glycine was added to a final concentration of 0.2 M and the incubation was continued for an additional 5 min. Cells were collected and washed with ice-cold PBS three times, cell lysis buffer (10 mM Tris-Cl [pH 7.5], 10 mM NaCl, 0.5% NP-40, 1 mM PMSF) three times, and resuspended in SDS lysis buffer (10 mM Tris-HCl [pH 7.5], 150 mM NaCl, 1% SDS, 1 mM EDTA). the cells were disrupted by sonication on ice. The chromatin solution was clarified by centrifugation at 15,000 g at 4°C for 10 min. The average DNA fragment size was 2.5 kb. The chromatin solution was diluted with IP dilution buffer (20 mM Tris-Cl [pH 8], 1 mM EDTA, 1% Triton X-100, and 150 mM NaCl, protease inhibitors) and pre-cleared with protein A Sepharose beads for 15 min. The pre-cleared diluted chromatin sample was incubated with 10 μl of anti-GST, anti-Sp1, anti-c-Myc, anti-p53 (DO1), or anti-p53(FL) for 3 hr followed by the addition of protein A Sepharose beads for an additional 3 hr. The beads were washed once with the IP dilution buffer, twice with 20 mM Tris-Cl (pH 8), 2 mM EDTA, 1% Triton X-100, 150 mM NaCl, 1 mM PMSF, once with 20 mM Tris-Cl (pH 8), 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 500 mM NaCl, 1 mM PMSF, and once with 10 mM Tris-Cl (pH 8), 1 mM EDTA, 0.25M LiCl, 1% NP-40, 1% deoxycholate. The immunoprecipitated material was eluted from the beads by heating for 15 min at 65°C in 25 mM Tris-Cl (pH 7.5), 10 mM EDTA, 0.5% SDS. To reverse the crosslinks, samples were incubated with 1.5 μg/ml Pronase at 42°C for 2 hr followed by 65°C for 5 hr. The samples were then extracted with phenol chloroform isoamyl alcohol followed by chloroform, ethanol precipitated in the presence of glycogen, and resuspended in TE buffer. The resulting precipitated DNA was amplified and hybridized to the chromosome 21 and 22 arrays.

Quantitative PCR Verification of Array-Detected Binding Sites

Quantitative real-time PCR experiments were performed using the Applied Biosystems 7700 sequence detector based on SYBR Green I fluorescence. Reactions were carried out in 10 μl using SYBR green PCR master mix according to the manufacturer's protocol. Cycling was for 10 min at 95°C, followed by 40 cycles of 95°C, 30 s, 60°C, 45 s, 72°C, 45 s. The fold-enrichment value for each transcription factor bound to a particular region of DNA was estimated as V+/V−. V+ was calculated by subtracting the cycle threshold (Ct; defined as the cycle at which the fluorescence signal is statistically significant over background) average of input DNA from the Ct average for the immunoprecipitated DNA; this net Ct value was then used as an exponent for the base 1.9 (1.9 being the mean primer slope). The same procedure was repeated to obtain the negative control region value (V−).

RT-PCR Verification of Noncoding Transcripts

Genomic regions proximal to internal or 3′ TFBS and exibiting evidence of transcription outside of annotated exons based on Affymetrix Transcriptome data (Kapranov et al., 2002) and/or EST evidence were selected for strand-specific RT-PCR. For each such region, 2 RT primers separated by 16–254 bp and 2 pairs of nested PCR primers lying upstream from the RT primers were selected using Oligo 6 (Molecular Biology Insights, Inc.). Jurkat cytosolic poly(A)$^+$ RNA was treated with 0.8 Units of RNase-free DNaseI (Roche) per 1 μg of RNA in 1× 1-PHOR-ALL buffer (Amersham-Pharmacia) for 20 min at 37°C, purified using RNeasy kit (Qiagen), and ethanol-precipitated. For each region, an RT reaction was performed on 100 ng of DNase I-treated poly(A)$^+$ RNA with a pool of 2 RT primers each at 0.9 μM, under the following conditions in the GeneAmp9600 cycler (Perking Elmer): RNA and primers were heated to 70°C for 10 min, followed by a ramp to 42°C or 52°C for 20 min, at which point 5× Superscript II First Strand buffer (Invitrogen), DTT, and four dNTPs were added to the following final concentrations of 1×, 10 mM and 0.5 mM, respectively. After 2 min at this temperature, 200 Units of Superscript II (Invitrogen) were added, followed by a 60 min incubation at 42°C or 52°C. RT was inactivated by a 15 min incubation at 70°C. The mRNA template was degraded using a combination of RNase, DNase-free (Roche), and RNase H (Invitrogen). One-third of each reaction was used for two rounds of PCR with nested primers. Each reaction contained 0.36 μM of each PCR primer, 1× Taq Gold buffer, 0.2 mM dNTPs, 1.5 mM MgCl2, 1.0 U Taq Gold Polymerase (Perkin Elmer) in a final volume of 50 μl. PCR amplification was performed after an initial step of 9 min at 94°C for 40 cycles of 30 s at 94°C, 30 s at 55°C, and 2 min at 60°C, with a final extension of 10 min at 60°C. A second round of PCR amplification was performed with nested primers on 2 μl of the first round reaction under the same conditions. The following controls were performed: no reverse transcriptase to control for DNA contamination, no RT primers to control for self-priming of RNA.

Differentiation of NCCIT Cell Line

NCCIT cell line was obtained from ATCC (CRL-2073), grown in RPMI 1640 supplemented with 10% FBS at 37°C, humidified atmosphere with 5% CO2. Differentiation was induced by 10 μM all-trans-Retinoic acid (Calbiochem, California). In 96 hr and 336 hr experiments, medium with RA was replaced every 72 hr. RNA isolation, labeling, and hybridization to Chrom21_22 arrays were performed as described previously (Kapranov et al., 2002).

Analysis of Tiling Array Data

Arrays were quantile-normalized within treatment/control replicate groups (Bolstad et al., 2003) and then all were scaled to have a median feature intensity of 1000. (PM, MM) intensity pairs were mapped to the genome using exact 25-mer matching. For each genomic position to which a probe pair mapped, a data set was generated consisting of all (PM, MM) pairs mapping within a window of ±500 bp. A Wilcoxon Rank Sum test was applied to the transformation log2(max(PM-MM,1)) for data from the six treatment and six control arrays within the local data set, testing the null hypothesis of equality of the two population distribution functions against the alternative of a positive difference in location between the probability distribution of the treatment and that of the control. The Wilcoxon test was applied in a sliding window across the genome. Genomic positions belonging to TFBS were defined by applying a p value cutoff of 10^{-5}, resultant positions separated by <500 bp were merged to form a predicted TFBS. Predicted TFBS were produced by this method for each TF against each of the controls (skipping the immunoprecipitation step and using an antibody to GST), the resulting two sets of predicted TFBS were merged together to form a nonredundant set of TFBS. All data are accessible at http://transcriptome.affymetrix.com/publication/tfbs.

Calculation of TF Motif Enrichment in Array-Detected Sites

The degree of enrichment for binding motifs in the predicted TFBS was estimated using the patterns GG[G/T]G[C/T]GGG and CA[C/T]G[T/C]G for Sp1 and cMyc, respectively. For p53, which binds as two dimers, the search was for the pattern **X**[0-14N]**X** where **X** = PuPuPuC(A/T)(T/A)GPyPyPy. TFBS less than 1 Kb in length were expanded equally in each direction to have length 1 Kb, and the resulting regions were repeat masked and scanned for the known binding motifs. This observed count of motifs was compared with

the number expected by chance, which was computed by determining the rate of occurrence of binding motifs in the nonrepetitive sequence of chromosomes 21 and 22 and multiplying by the amount of nonrepetitive sequence in the expanded TFBS. 90, 383, and 1 binding sites are found for Sp1, cMyc, and p53, respectively; the expected counts at random are 12, 145, and 0.04, representing an enrichment of 7.5-, 2.6-, and 25-fold for Sp1, cMyc, and p53 (we note that given there was only one detected p53 binding motif the last fold-enrichment will have a higher standard error).

Enrichment of Significant p53 Weight Matrix Scores in Array-Detected Sites

Given the complexity of the p53 motif and the fact that there was only one exact match, a positional weight matrix was used to predict inexact matches to the consensus. A weight matrix, w_{ij}, was derived from a table of n_{ij}, the count at which nucleotide i is observed at position j in $N = 37$ aligned clones of known p53 binding sites (taken from Table 1 of Hoh et al., 2002) using the following formula: $w_{ij} = \ln[(n_{ij} + c)/(p_i N)]$ where $p_i = 0.25$ is used as the background probability of nucleotide i and $c = 0.25$ is used as a pseudocount to compensate for the fact that the weight matrix from Hoh et al. (2002) is based on only 37 motifs. The score S for all possible p53 binding sites along chromosomes 21 and 22 was computed by summing the 20 positional weights corresponding to the p53 double dimer (allowing up to 20 bp spacing between motif pairs). Sequences not containing the fourth C and seventh G in both dimers were filtered out. Using a cutoff of $S > 8$ (corresponding to a frequency of three motifs per 10 kb on the repeat masked sequence of chromosomes 21 and 22), 41 putative consensus sequences were detected in the TFBS, whereas only 12 would be expected at random, yielding >3-fold enrichment. (The sole exact match to the consensus had a score of 8.86.) Applying a more conservative cutoff of 11.6, only 1.4 motifs would be expected at random but 13 were found, giving >9-fold enrichment.

Association of Overlapping Coding and Novel Transcripts to TFBS

dbEST ESTs and GenBank mRNAs annotated as having "complete CDS" were aligned to chromosomes 21 and 22 under stringent alignment criteria (>95% identity over full-length of query with no query gaps tolerated). Only ESTs being of confident strandedness were used, where strandedness evidence comes from splicing and/or having a polyadenylation site, possibly with associated signal. Sequences were projected into genomic space to deal with the large redundancy in the set, and the full-length sequences were used to classify the set into 1782 "transcript clusters," of which 363 show potential for transcription on both strands. Relating this to the detected TFs, the property of a transcript cluster having a noncanonical TF (i.e., one not located at the 5′ end of a known gene) was tested for association with the property of a transcript cluster having evidence for transcription on both strands, yielding a highly significant association (chi-square p value $<10^{-15}$).

Correlation of Overlapping Coding and Novel RNA Expression Levels

363 coding/novel transcript clusters were derived from public EST/mRNA databases. Only clusters with at least one exon unambiguously assigned to each strand were retained, reducing the set to 214. For each of these 214 clusters, a sense and an antisense probeset were formed, consisting of probes spaced approximately every 35 bp along the exons of the transcript. For every probe and for each time point in the retinoic acid-induced differentiation (treatment), the ratio of max(PM-MM,10) for the treatment over max(PM-MM,10) for a control (an untreated NCCIT sample after 4 hr of growth) was determined. The median ratio was computed over the probes in each probeset. A further filter was applied to the set of 214 transcript clusters, retaining only those in which two or more of the time points for both coding and novel transcripts showed differentiation with respect to the control (had a median ratio different to 1), reducing the set to 61 transcript clusters. The Pearson correlation between coding and novel was computed for each of the 61 transcript clusters, the average was 0.34. To assess significance, the time points for each of the 61 transcript clusters were randomly permuted and the correlation was computed; this was repeated 5000 times and a null distribution was estimated. When assessed in light of the null model, the observed correlation had a two-tailed p value of less than 0.0002. The results were very similar in terms of size and significance of the correlation when Spearman correlation was used in place of Pearson correlation.

Acknowledgments

We thank E. Schell, X.-M. Zhu, and M. Mittmann for design of photolithographic masks and Stan Tabor for amplification of DNA samples. H.H.N. was supported by postdoctoral fellowship from the Cancer Research Fund of the Damon Runyon Walter Winchell Foundation, and E.A.S. was supported by a postdoctoral fellowship from the National Institutes of Health. Support for this work was provided in part by Federal Funds from the National Institutes of Health (GM 30186 and GM 53720 to K.S.), National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400, to T.R.G., and by Affymetrix.com.

Received: September 25, 2003
Revised: December 5, 2003
Accepted: January 19, 2004
Published: February 19, 2004

References

Ashrafi, K., Chang, F.Y., Watts, J.L., Fraser, A.G., Kamath, R.S., Ahringer, J., and Ruvkun, G. (2003). Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes. Nature *421*, 268–272.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Carmichael, G.G. (2003). Antisense starts making more sense. Nat. Biotechnol. *21*, 371–372.

Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J.D., and Wang, S.M. (2002). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc. Natl. Acad. Sci. USA *99*, 12257–12262.

Damjanov, I., Horvat, B., and Gibas, Z. (1993). Retinoic acid-induced differentiation of the developmentally pluripotent human germ cell tumor-derived cell line, NCCIT. Lab. Invest. *68*, 220–232.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., and Antonarakis, S.E. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature *420*, 578–582.

Gayther, S.A., Batley, S.J., Linger, L., Bannister, A., Thorpe, K., Chin, S.F., Daigo, Y., Russell, P., Wilson, A., Sowter, H.M., et al. (2000). Mutations truncating the EP300 acetylase in human cancers. Nat. Genet. *24*, 300–303.

Gonzalez, F.A., Raden, D.L., Rigby, M.R., and Davis, R.J. (1992). Heterogeneous expression of four MAP kinase isoforms in human tissues. FEBS Lett. *304*, 170–178.

Hamilton, A.J., and Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950–952.

Hoh, J., Jin, S., Parrado, T., Edington, J., Levine, A.J., and Ott, J. (2002). The p53MH algorithm and its application in detecting p53-responsive genes. Proc. Natl. Acad. Sci. USA *99*, 8467–8472.

Horak, C.E., Luscombe, N.M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M., and Snyder, M. (2002a). Complex transcriptional cir-

cuitry at the G1/S transition in Saccharomyces cerevisiae. Genes Dev. *16*, 3017–3033.

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M., and Snyder, M. (2002b). GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. Proc. Natl. Acad. Sci. USA *99*, 2924–2929.

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533–538.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kannan, K., Amariglio, N., Rechavi, G., Jakob-Hirsch, J., Kela, I., Kaminski, N., Getz, G., Domany, E., and Givol, D. (2001). DNA microarrays identification of primary and secondary target genes regulated by p53. Oncogene *20*, 2225–2234.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916–919.

Kramer, C., Loros, J.J., Dunlap, J.C., and Crosthwaite, S.K. (2003). Role for antisense RNA in regulating circadian clock function in Neurospora crassa. Nature *421*, 948–952.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862–864.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799–804.

Lehner, B., Williams, G., Campbell, R.D., and Sanderson, C.M. (2002). Antisense transcripts in the human genome. Trends Genet. *18*, 63–65.

Lipman, D.J. (1997). Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res. *25*, 3580–3583.

Liu, X.S., Brutlag, D.L., and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. *20*, 835–839.

Mao, D.Y., Watson, J.D., Yan, P.S., Barsyte-Lovejoy, D., Khosravi, F., Wong, W.W., Farnham, P.J., Huang, T.H., and Penn, L.Z. (2003). Analysis of Myc bound loci identified by CpG island arrays shows that max is essential for Myc-dependent repression. Curr. Biol. *13*, 882–886.

Mattick, J.S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. *2*, 986–991.

Menssen, A., and Hermeking, H. (2002). Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. Proc. Natl. Acad. Sci. USA *99*, 6274–6279.

Minc, E., de Coppet, P., Masson, P., Thiery, L., Dutertre, S., Amor-Gueret, M., and Jaulin, C. (1999). The human copper-zinc superoxide dismutase gene (SOD1) proximal promoter is regulated by Sp1, Egr-1, and WT1 via non-canonical binding sites. J. Biol. Chem. *274*, 503–509.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2002). Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. Genes Dev. *16*, 806–819.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L.M., and Wirth, D.F. (2001). Serial analysis of gene expression in Plasmodium falciparum reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. Mol. Biol. Cell *12*, 3114–3125.

Plougastel, B., Zucman, J., Peter, M., Thomas, G., and Delattre, O. (1993). Genomic structure of the EWS gene and its relationship to EWSR1, a site of tumor-associated chromosome translocation. Genomics *18*, 609–615.

Reik, W., and Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. *2*, 21–32.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306–2309.

Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A., and Dynlacht, B.D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev. *16*, 245–256.

Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M., et al. (2003). The transcriptional activity of human Chromosome 22. Genes Dev. *17*, 529–540.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. (2002). Using the transcriptome to annotate the genome. Nat. Biotechnol. *20*, 508–512.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res. *13*, 103–107.

Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. Cell *106*, 697–708.

Sleutels, F., and Barlow, D.P. (2002). The origins of genomic imprinting in mammals. Adv. Genet. *46*, 119–163.

Solymar, D.C., Agarwal, S., Bassing, C.H., Alt, F.W., and Rao, A. (2002). A 3′ enhancer in the IL-4 gene regulates cytokine production by Th2 cells and mast cells. Immunity *17*, 41–50.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Wagner, E.G., and Simons, R.W. (1994). Antisense RNA control in bacteria, phages, and plasmids. Annu. Rev. Microbiol. *48*, 713–742.

Wagner, E.G., and Flardh, K. (2002). Antisense RNAs everywhere? Trends Genet. *18*, 223–226.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Watson, J.D., Oster, S.K., Shago, M., Khosravi, F., and Penn, L.Z. (2002). Identifying genes regulated in a Myc-dependent manner. J. Biol. Chem. *277*, 36921–36930.

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. (2003). Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. *21*, 379–386.

Zeitlinger, J., Simon, I., Harbison, C.T., Hannett, N.M., Volkert, T.L., Fink, G.R., and Young, R.A. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell *113*, 395–404.