

Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment

Jonghwan Kim, Akshay A Bhinge, Xochitl C Morgan & Vishwanath R Iyer

Identifying the chromosomal targets of transcription factors is important for reconstructing the transcriptional regulatory networks underlying global gene expression programs. We have developed an unbiased genomic method called sequence tag analysis of genomic enrichment (STAGE) to identify the direct binding targets of transcription factors *in vivo*. STAGE is based on high-throughput sequencing of concatemeric tags derived from target DNA enriched by chromatin immunoprecipitation. We first used STAGE in yeast to confirm that RNA polymerase III genes are the most prominent targets of the TATA-box binding protein. We optimized the STAGE protocol and developed analysis methods to allow the identification of transcription factor targets in human cells. We used STAGE to identify several previously unknown binding targets of human transcription factor E2F4 that we independently validated by promoter-specific PCR and microarray hybridization. STAGE provides a means of identifying the chromosomal targets of DNA-associated proteins in any sequenced genome.

Determining the binding sites of regulatory proteins on the genome is important for reconstructing transcriptional regulatory networks^{1–3}. The binding of a transcription factor to its genomic targets can be assayed by combining chromatin immunoprecipitation (ChIP) and microarray (chip) hybridization. This ChIP-chip method was first developed for yeast⁴, where it has been used to define the targets of more than 100 transcription factors^{2,5,6}.

Although ChIP-chip has also enabled the identification of transcription factor targets in human cells^{7,8}, it is challenging to apply this approach comprehensively to study large and complex genomes. Human promoter microarrays based on core promoters⁷ or CpG islands⁸ cover a subset of all potential regulatory regions and may not adequately represent regions that are distant from genes or within introns. Tiling arrays of polymerase chain reaction (PCR) products⁹ or oligonucleotides¹⁰ have been made for the smallest human chromosomes, but extending such arrays to cover the entire genome is expensive, and the arrays are currently unavailable to most researchers. Although these efforts are underway for the human genome and some model organisms, the development of similar platforms for the mouse, plants, prokaryotes and many other model organisms is lagging.

Here, we address some of these limitations by developing an unbiased genomic method to identify the chromosomal targets of transcription factors. We term this method STAGE, and it is based on high-throughput sequencing of concatemeric tags derived from DNA enriched by ChIP. Cloning and sequencing of ChIP DNA has been carried out previously¹¹, but these efforts did not constitute a high-throughput genomic approach. As a demonstration of its utility, we first used STAGE to map the targets of TATA-box binding protein (TBP) in yeast. We then optimized STAGE and developed analysis algorithms that enabled us to successfully use STAGE to identify several known and new binding targets of transcription factor E2F4 in human cells.

RESULTS

STAGE identifies chromosomal targets in yeast

STAGE is conceptually derived from serial analysis of gene expression (SAGE)^{12,13}, but the template for STAGE consists of genomic loci enriched by ChIP. Briefly, transcription factors are cross-linked to their target sites *in vivo* with formaldehyde. After ChIP with a specific antibody against a given transcription factor, the recovered DNA fragments are amplified by PCR using biotinylated degenerate primers and digested with the four-base cutter (5'-CATG) restriction endonuclease *Nla*III. The biotinylated fragments are isolated using streptavidin beads and ligated to linkers containing a recognition site for *Mme*I, a type IIS restriction enzyme. Digestion with *Mme*I releases 21-base-pair (bp) tags containing *Nla*III sites from DNA fragments enriched after ChIP. Multiple tags are concatemericized, cloned and sequenced. STAGE generates 21-bp tags derived from ChIP DNA (Fig. 1). Mapping these tags to the genome can identify the loci represented in the ChIP sample and thus identify protein-binding locations.

We first used STAGE to identify the targets of yeast TATA-box binding protein (TBP). Out of a total of 1,344 sequenced tags, 294 (22%) did not match any sequence in the yeast genome. The total number of sequenced tags and the number of orphan and ambiguous tags are provided in **Supplementary Table 1** online. Out of 1,050 valid STAGE tags, 433 showed multiple hits on the genome and could not be assigned to a single gene; 77 tags had single hits but had no annotated genes within one kilobase (kb). The remainder comprised 437 distinct tags, each of which had

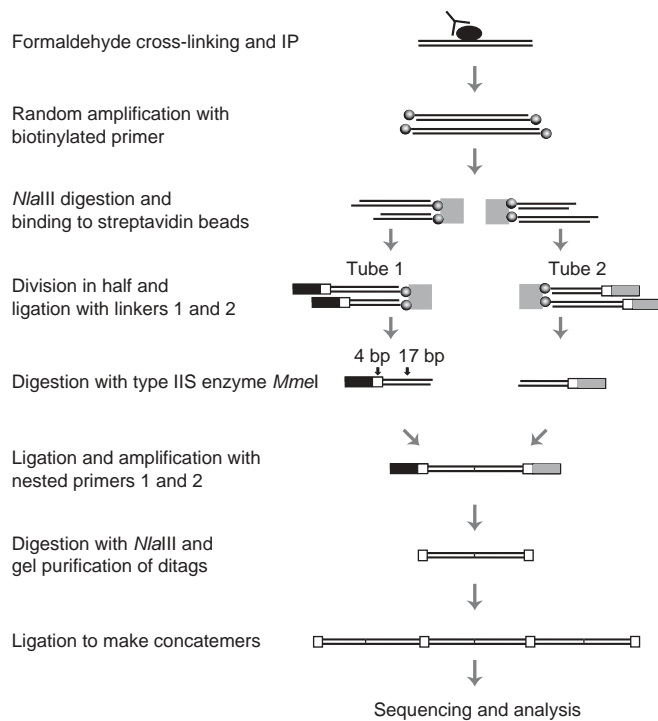


Figure 1 | The STAGE strategy. STAGE is based on high-throughput sequencing of concatemerized tags of defined length that are derived from DNA enriched by ChIP. Proteins were cross-linked to their binding sites *in vivo* with formaldehyde and chromatin was extracted and sheared. The cross-linked protein-DNA complexes were immunoprecipitated, cross-links were reversed and ChIP DNA was amplified by PCR using biotinylated primers. Amplified DNA fragments were digested with *Nla*III, which cuts at 5'-CATG sites. Fragments with ends containing the *Nla*III site were isolated by binding to streptavidin beads. They were separately ligated to one of two linkers containing a *Mme*I site, then incubated with *Mme*I, which cleaves 21 bp away from its recognition site. The 21-bp tags attached to linkers were isolated and ligated to create ditags. Ditags were amplified by PCR using nested primers and trimmed by digesting with *Nla*III. Trimmed ditags were gel purified, concatemerized by ligation, cloned and sequenced.

only one hit on the yeast genome and was located within 1 kb of the start of a gene.

Of the 437 tags, 378 occurred only once in the STAGE pool and 59 occurred multiple times. Seventy-nine putative targets were represented by more than one tag occurrence. The one notable feature of the abundant tags was that a substantial majority mapped within 1 kb of an RNA polymerase III (pol III) promoter. Based on this observation and on the fact that pol III promoters are prominent targets of TBP^{14,15}, we assigned the gene with a pol III promoter as the putative target when a tag mapped near it. In other cases, the nearest gene was assigned as the putative target. Tags that occurred multiple times in the STAGE pool, as well as their putative targets, are listed (Table 1). Sixty-eight of 79 targets represented by multiple tags were genes with an RNA pol III promoter. STAGE thus identified many prominent chromosomal targets of TBP in yeast.

Validation of STAGE targets by microarray hybridization

To compare our STAGE targets to those identified by microarray hybridization, ChIP DNA samples were fluorescently labeled and cohybridized to whole-genome (ORFs + intergenic regions) microarrays with an amplified genomic DNA reference. The occupancy

Table 1 | High-abundance yeast TBP STAGE tags

Tag sequence	<i>nocc</i>	Target gene
CATGATGGAACGAAGACGAC	10	tF(GAA)B
CATGAGAATGTGCTTCAGTAT	8	tF(GAA)B
CATGAAGTGCACAAAATGATT	5	tK(CUU)E1
CATGATCAAATCTGTGAAGC	5	tL(CAA)A
CATGCAAATCTAAAATAAAAC	5	tH(GUG)H
CATGTATACTAACAGATATG	5	RDN 5-1
CATGAGATATGCTGTTTCAAG	4	tL(CAA)A
CATGTATATATTGCACTGGCT	4	RDN 5-1
CATGAAACTAGGAAAACGTAC	3	tE(UUC)J
CATGAAGATGATTCGATACCG	3	tV(AAC)M1
CATGATGAAGTTAGATCTGC	3	tW(CCA)K
CATGATGGCAGACTTCCATCG	3	tV(AAC)G2
CATGATGTCGCTATTCTAAT	3	tY(GUA)J2
CATGCAAGATGTAGACCCAAC	3	YGRW σ 5
CATGCAATCCCAGTAGTAGGT	3	SCR1
CATGCAGCTGTTGTATCAAGA	3	tV(AAC)G1
CATGCATGTTTACGTTGTTGG	3	tP(AGG)N
CATGGAATGTGCAATTAAGAC	3	tT(AGU)N2
CATGTGTTGTA AAAAGATAAC	3	tT(AGU)J
CATGTTATCCTGAGCATCCAC	3	tG(GCC)O2
CATGTTTACCTCAAACAAG	3	tV(AAC)K1
CATGTTTCTCTAAAGATGGT	3	tR(UCU)B
CATGAAAACCTCTCAAACCTT	2	tH(GUG)E1
CATGAAAAGGTTAATGACTT	2	tT(AGU)O1
CATGAAGACCTATTCGCTTAT	2	tV(AAC)G3
CATGAAGCGCACAAGATTGGA	2	tR(UCU)G3
CATGAATGCGCCAGATTATT	2	tV(AAC)M1
CATGAGGCGCACTTTTGATTT	2	tY(GUA)F2
CATGAGTTGCCATTAGAAAACG	2	tW(CCA)G1
CATGATACTGACTTATTGGGC	2	tD(GUC)L1
CATGCAAGACGTAGACCCAAC	2	tI(AAU)I2
CATGCAAGTGTGGCATAAAAAG	2	tK(CUU)E2
CATGCAGAAAAGATAAGATGC	2	YPL029W
CATGCCTGTCAACGCCGACG	2	tE(UUC)J
CATGCTCGGCAATAGCTTCAA	2	tG(CCC)D
CATGCTTTGTCTTCTGTTAG	2	tP(UGG)O2
CATGGAAAACGAATGGAGAC	2	tA(AGC)K1
CATGGAAATCGAACCTTTCAC	2	tN(GUU)N2
CATGGAGTCAACTTTGTTGT	2	tN(GUU)O2
CATGGAGTCTTTTATTTCCGA	2	tN(GUU)L
CATGGCAAAAACGTAAAGTT	2	tR(UCU)G2
CATGGCGAATTTTTCACATAT	2	tV(UAC)D
CATGGCGATATTTCATTATG	2	tR(UCU)G3
CATGGCTAGTCAAATAAGTGG	2	YGL080W
CATGGGGTAAGTCCGATGGC	2	tV(AAC)E2
CATGGGTTCAAACACTTCCAA	2	tY(GUA)F1
CATGGTGAAAGTTAATCTTT	2	tR(ACG)K
CATGTAAACCATCCCTTTTCA	2	YJL005W
CATGTATAAAAACCTACCGCTT	2	tS(CGA)C
CATGTATCAAATGCGACGTGA	2	YPRC δ 22
CATGTATGAAAACGTGGAAATTC	2	tS(AGA)B
CATGCAATGTCCATTTCTTT	2	tT(AGU)I2
CATGCTTTTGTGGATTATTT	2	tS(CGA)C
CATGTGAGGCTTAGGTGATTT	2	tN(GUU)N2
CATGTGTTTGAATTAGCGATC	2	tL(CAA)A
CATGTTACAATTCCTTCCAT	2	tG(UCC)G
CATGTTATGTTCAATGGCAG	2	YELC τ 1
CATGTTCAAGGACGGCTTGGT	2	tD(GUC)J1
CATGTTTTCGTTAATTCATAA	2	tR(UCU)B

Tags that occurred more than once are listed, including the 4-bp *Nla*III site (5'-CATG). The number of times the tag occurred in the STAGE pool is indicated by *nocc*. Target genes were designated as described in the text.

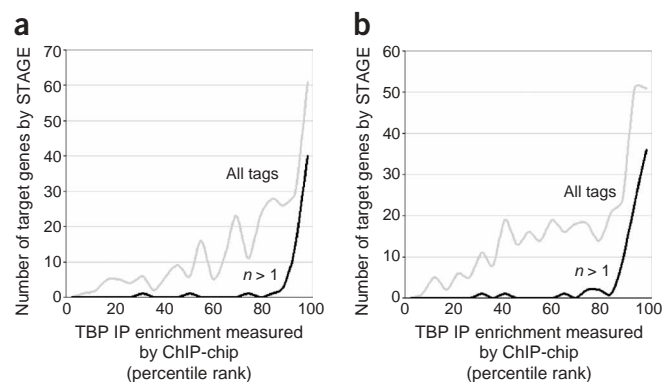


Figure 2 | Correlation between yeast targets predicted by STAGE and ChIP-chip. The enrichment value of yeast TBP targets after ChIP was determined by microarray hybridization. The percentile rank (0–100) of the ratio of ChIP-enriched fragments to genomic DNA was used to determine the ChIP enrichment value for each locus. For each interval of TBP ChIP enrichment values plotted on the x-axis, the number of targets predicted by STAGE is plotted on the y-axis. **(a)** Comparison between STAGE and ChIP-chip when the same sample was analyzed by both methods. The gray line indicates all predicted STAGE targets, whereas the black line indicates only the subset of 79 target genes predicted by multiple tag occurrences. **(b)** Comparison between STAGE and ChIP-chip when different ChIP samples were analyzed.

of each promoter by TBP was indicated by the rank of its enrichment in ChIP relative to the reference¹⁶.

STAGE identified increasing numbers of genes as TBP targets with increasing enrichment in ChIP as measured by microarrays (Fig. 2a). This relationship was more pronounced when we considered only genes that were identified as targets by more than one tag occurrence (Fig. 2a). Among the putative TBP targets represented by at least two tag occurrences, 92% had high enrichment values (>90) in ChIP-chip. When the two ChIP samples were independently generated, 91% of the targets predicted by at least two STAGE tag occurrences showed high ChIP-chip enrichment values (Fig. 2b). Thus, identification of chromosomal targets by STAGE correlates well with that by ChIP-chip, especially when the target genes were designated by multiple occurrences of STAGE tags.

STAGE in human cells

We chose transcription factor E2F4 to test STAGE in human cells. E2F4 is a member of the E2F family of transcriptional regulators that functions as a repressor in quiescent and early G₁ cells¹⁷. We first used ChIP and promoter-specific PCR to verify the binding of E2F4 to known target promoters⁷ (Fig. 3a). We then constructed a human E2F4 STAGE pool from these validated ChIP samples.

To reduce and account for background genomic DNA in ChIP, we introduced two enhancements. First, we tested a subtraction step as a potential means of reducing background from nonspecific genomic loci. Briefly, DNA fragments enriched by ChIP were randomly amplified by PCR with degenerate primers, and, in parallel, sheared genomic DNA fragments were amplified using biotinylated degenerate primers. ChIP DNA was hybridized to an excess of biotinylated genomic DNA and biotin-containing heteroduplexes were removed by binding to streptavidin beads. The remaining DNA was used as the input for STAGE. Details of the subtraction procedure are given in **Supplementary Methods** online. In a ChIP sample where the enrichment of an E2F4 target

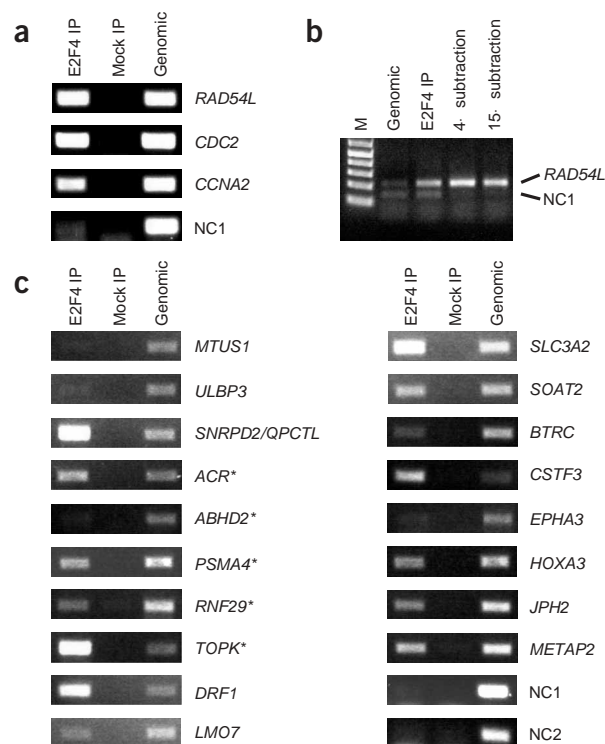


Figure 3 | ChIP of E2F4 targets and validation of STAGE targets by ChIP-PCR. **(a)** Binding of human E2F4 to known target promoters in fibroblasts. PCR was performed using primers corresponding to the promoters of the indicated genes. The ninth exon of *CCNB1* was used as a negative control for ChIP enrichment (NC1). **(b)** The subtraction procedure leads to improved enrichment of the *RAD54L* promoter in ChIP. 'M' is a size ladder. **(c)** Validation of STAGE targets by ChIP-PCR. A subset of 18 promoters out of the 45 predicted by STAGE were randomly chosen. E2F4 binding to the promoters of the indicated genes was assayed by promoter-specific PCR. NC1 is the ninth exon of *CCNB1* and NC2 is the promoter of *ACTB*; both are negative controls. *SNRPD2* and *QPCTL* are divergently transcribed. The putative targets of E2F4 predicted by SubSTAGE are marked by an asterisk.

over background was originally suboptimal, we observed improved enrichment after subtraction (Fig. 3b). Tags from this E2F4 subtraction STAGE (SubSTAGE) pool were combined for analysis with STAGE tags obtained without the subtraction step.

Additionally, we performed STAGE on normal, unselected human genomic DNA to profile tags arising from background genomic DNA that was not enriched by ChIP. This background STAGE pool would thus serve as an analysis control to account for sampling of STAGE tags from highly repetitive regions of the genome. We analyzed approximately 3,500 valid tags to identify targets of E2F4 in human cells.

Targets of human transcription factor E2F4

To overcome the ambiguity inherent in mapping many 21-bp tags to specific locations on the human genome, we developed an algorithm to score tags and genes as putative targets. Each distinct tag was assigned a tag score based on the number of its hits on the genome and the number of its occurrences in the STAGE pool. Details of the scoring method are described (see **Methods** and **Supplementary Methods** online). A higher number of hits on the genome lowered the tag score, and a higher occurrence number in the STAGE pool raised the tag score. For each human gene in

RefSeq^{18,19}, a final STAGE enrichment score was generated that was indicative of the enrichment of its promoter in ChIP. The final STAGE enrichment score for each gene was calculated by dividing its raw score from the ChIP STAGE library by its raw score from the appropriate background genomic STAGE library.

There were 48 putative targets of E2F4 with STAGE enrichment scores greater than a threshold of 900 in either of the two STAGE

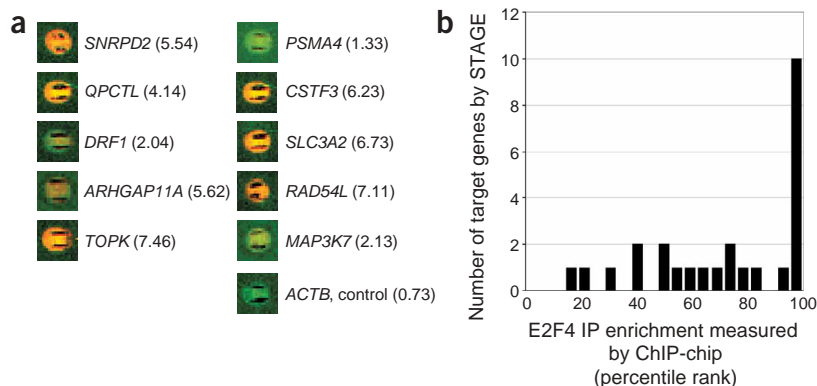
pools (Table 2). Raw scores and final STAGE enrichment scores are available (Supplementary Table 2 online). Most targets were designated by at least one tag with a single hit on the human genome. In addition to previously known targets of E2F4 such as *RAD54L*, *SLC3A2* and *MAP3K7*, which had been identified using a human core promoter microarray⁷, our analysis identified several new targets that had not been identified in previous studies. We also

Table 2 | Human E2F4 targets predicted by STAGE

No.	Gene	Gene Score	E2F4 site	Description
1	<i>MTUS1</i>	1971		Mitochondrial tumor suppressor gene 1
2	<i>ULBP3</i>	1961		UL16 binding protein 3
3	<i>SNRPD2</i> *	1933		Small nuclear ribonucleoprotein D2 polypeptide, 16.5 kDa
	<i>QPCTL</i> *	1923		Hypothetical protein FLJ20084
4	<i>PXK</i>	1015		PX domain-containing serine/threonine kinase
5	<i>FLJ22353</i>	993		Hypothetical protein FLJ22353
6	<i>GAJ</i>	993	Yes	GAJ protein
7	<i>ACR</i>	992		Acrosin
8	<i>RAD54L</i>	992		RAD54-like (<i>S. cerevisiae</i>)
9	<i>AAMP</i>	982		Angio-associated migratory cell protein
10	<i>ABHD2</i>	982		Abhydrolase domain-containing 2
11	<i>BLVRB</i> *	982		Biliverdin reductase B (flavin reductase (NADPH))
	<i>SPTBN4</i> *	982		Spectrin, beta, non-erythrocytic 4
12	<i>DC2</i>	982		DC2 protein
13	<i>FLJ13912</i>	982	Yes	Hypothetical protein FLJ13912
14	<i>FLJ25416</i>	982		Hypothetical protein FLJ25416
15	<i>FLJ32000</i>	982	Yes	Hypothetical protein FLJ32000
16	<i>FLJ90834</i>	982		Hypothetical protein FLJ90834
17	<i>MPV17</i>	982	Yes	MpV17 transgene, murine homolog, glomerulosclerosis
18	<i>PRCP</i>	982		Prolylcarboxypeptidase (angiotensinase C)
19	<i>PSMA4</i>	982		Proteasome (prosome, macropain) subunit, alpha type, 4
20	<i>RNF29</i>	982		Ring finger protein 29
21	<i>TOPK</i>	982	Yes	T-LAK cell-originated protein kinase
22	<i>DRF1</i>	974		Dbf4-related factor 1
23	<i>LM07</i>	971		LIM domain only 7
24	<i>SLC3A2</i>	971	Yes	Solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2
25	<i>SOAT2</i>	971		Sterol O-acyltransferase 2
26	<i>ARHGAP11A</i>	965	Yes	KIAA0013 gene product
27	<i>ABC1</i>	961		Amplified in breast cancer 1
28	<i>BTRC</i>	961		Beta-transducin repeat containing
29	<i>GAL3ST1</i>	961		Cerebroside (3'-phosphoadenylylsulfate:galactosylceramide 3') sulfotransferase
30	<i>CSTF3</i>	961		Cleavage stimulation factor, 3' pre-RNA, subunit 3, 77 kDa
31	<i>CTAG3</i> *	961		Cancer/testis antigen 3
	<i>RIOK1</i> *	961		RIO kinase 1 (yeast)
32	<i>DNALI1</i>	961		Dynein, axonemal, light intermediate polypeptide 1
33	<i>EPHA3</i>	961		EPH receptor A3
34	<i>FIBL-6</i>	961	Yes	Hemicentin
35	<i>FLJ20712</i>	961		Hypothetical protein FLJ20712
36	<i>HIST2H2AC</i>	961		Histone 2, H2ac
37	<i>HOXA3</i>	961		Homeobox A3
38	<i>JPH2</i>	961		Junctophilin 2
39	<i>MAP3K7</i>	961	Yes	Mitogen-activated protein kinase kinase kinase 7
40	<i>METAP2</i>	961	Yes	Methionyl aminopeptidase 2
41	<i>PDGFA</i>	961		Platelet-derived growth factor alpha polypeptide
42	<i>RPL23A</i>	961	Yes	Ribosomal protein L23a
43	<i>SNIP1</i>	961	Yes	Smad nuclear interacting protein
44	<i>CCRL2</i>	926		Chemokine (C-C motif) receptor-like 2
45	<i>C20orf141</i>	913		Chromosome 20 open reading frame 141

An asterisk indicates a bidirectional promoter (a promoter driving the expression of two mRNAs in opposite directions). The presence of consensus E2F4 binding sites in a 3 kb window spanning the start of transcription is also indicated.

Figure 4 | Validation by ChIP-chip of E2F4 targets predicted by STAGE. DNA from an E2F4 ChIP was amplified and labeled with Cy5, and hybridized to a human core-promoter microarray together with a mock IP sample labeled with Cy3. The ratio of Cy5/Cy3 (red/green) signal is an indicator of the binding of E2F4 to the locus at a given spot. **(a)** New targets identified by STAGE (see **Table 2** and **Fig. 3c**) include *SNRPD2*, *QPCTL*, *DRF1*, *ARHGAP11A*, *TOPK*, *CSTF3* and *PSMA4*. Previously known E2F4 targets that were also identified by STAGE are *SLC3A2*, *RAD54L* and *MAP3K7*. The *ACTB* promoter is a negative control. **(b)** Correlation between targets predicted by STAGE and ChIP-chip. The average percentile rank (0–100) from two microarray hybridizations, of the ratio of ChIP-enriched fragments to mock IP control DNA was determined for each spot on the microarray. For each interval of E2F4 ChIP enrichment values plotted on the x-axis, the number of targets predicted by STAGE (total 26) is plotted on the y-axis. Ten STAGE predicted targets rank in the top 5% of all spots on the microarray, corresponding to a red/green ratio >2.0.



calculated a significance value for each STAGE enrichment score. All our putative targets (**Table 2**) had scores with *P* values much lower than 0.01. The score for the *ACTB* (β -actin) gene used as a negative control had a much higher *P* value ($P > 0.5$).

Validation of STAGE in human cells

From the 45 putative target promoters (**Table 2**), we selected 18 for validation by promoter-specific PCR. Primers were designed to assay a region spanning the ~400 bp upstream of the transcription start site of each gene. We detected E2F4 binding to 15 promoters (**Fig. 3c**). Including *RAD54L* (**Fig. 3a**), we could thus independently verify 16 of 19 (84%) binding targets predicted by STAGE.

We used ChIP-chip to further verify the binding of E2F4 to promoters identified by STAGE. DNA from an independent E2F4 ChIP was amplified and labeled with Cy5 and hybridized to a 9,500-element human core promoter microarray²⁰ together with a mock-immunoprecipitated sample labeled with Cy3. Many previously unknown E2F4 targets that we identified by STAGE were indeed enriched in the independent ChIP-chip as indicated by high red/green (Cy5/Cy3) ratios (**Fig. 4a**). STAGE identified increasing numbers of genes as E2F4 targets with increasing enrichment in ChIP-chip (**Fig. 4b**). Of the 48 E2F4 target genes identified by STAGE, 26 were represented on the microarray. Ten of these (38%) had ChIP-chip enrichment values in the top 5%, indicating they were bona fide targets. The overlap between the targets identified by STAGE and by ChIP-chip, although modest, was highly significant ($P < 10^{-7}$ based on sampling permutation), showing that STAGE enables the identification of target loci in human cells. This overlap between the targets identified by the two different technologies is comparable to the 43% overlap we observed between our ChIP-chip targets and the set of E2F4 targets previously reported in the literature also using ChIP-chip^{7,8}.

In addition to the identification of E2F4 targets based on the occurrence of tags within a 3-kb window proximal to annotated genes, we separately scored genes as putative targets based on the presence of tags within a region from -10 kb to -6 kb or from -6 kb to -2 kb relative to the start of transcription or within the first intron. These analyses identified 48, 43 and 17 additional putative targets, respectively (**Supplementary Tables 3,4** and **5**). Some of these additional putative targets, such as *ACR*, *FLJ22353* and *ULBP3*, had also been identified in our analysis based on the

3-kb proximal region (**Table 2**). It is possible that E2F4 binds to multiple sites at varying distances upstream of some of its target genes. Approximately 1,400 unique STAGE tags were derived from regions of the genome that were not within 10 kb upstream of, or in, the first intron of any gene. Although we have not validated these as true E2F4 binding sites, binding to sites outside promoters would be consistent with recent reports describing such binding by NF- κ B⁹, c-myc and Sp1 (ref. 10).

DISCUSSION

Our results demonstrate the utility of STAGE as an unbiased genomic method for identifying the chromosomal binding targets of proteins. STAGE identified many new target genes of E2F4 in human fibroblasts that had not been identified in previous studies using targeted core promoter microarrays or CpG island microarrays^{7,8}.

The fraction of orphan STAGE tags that did not match any genomic sequence was generally 15–19%, similar to what has been observed for SAGE^{13,21}. Orphan tags likely arise from a combination of PCR and sequencing errors and cross-contamination from unrelated DNA samples. Half of the 22% orphan tags we observed in one instance in yeast consisted of repeated occurrences of just two distinct tags. We did not observe these two tags in any other STAGE pools. Although it is desirable to minimize the occurrence of such orphan tags, they do not present a problem for STAGE, as they are excluded from analysis.

Although there was significant overlap ($P < 10^{-7}$) between the E2F4 targets that we identified by ChIP-chip and by STAGE, the agreement between the two technologies was not perfect. ChIP-chip involves a complex hybridization step and can be affected by the presence of repetitive DNA, poor PCR product in the microarray spot, differential amplification of ChIP DNA during fluorescent labeling and hence low sensitivity or specificity at certain loci. For example, we identified *PSMA4* as an E2F4 target by STAGE and validated it by ChIP-PCR, but it showed only marginal enrichment in ChIP-chip (**Fig. 4a**). However, *MAP3K7*, a previously known target of E2F4 that we also identified by STAGE, likewise did not show enrichment in our ChIP-chip, indicating that ChIP-chip is not infallible. For this reason, we believe that the standard low-throughput ChIP-PCR assay is a more reliable measure of whether a locus is a true binding target.

Based on our ChIP-PCR analysis (Fig. 3a,c), we estimate the true positive rate of STAGE in human cells is ~84% in our experiments. This success rate can potentially be improved by enhancements to the analysis algorithms as well as improvements to the ChIP procedure to reduce nonspecific DNA background. Subtraction is one potential means of reducing background. However, it is possible that the subtraction step may be effective only when the initial ChIP enrichment is poor (Fig. 3b). The use of new type III restriction enzymes generating 26-bp tags rather than 21-bp tags may also improve the specificity of STAGE²². However, 70% of all *Nla*III-anchored 21-bp tags in the human genome were unique, whereas 76% of all such 26-bp tags were unique. The improvement in the ability to uniquely localize tags by increasing their lengths from 21 to 26 bp is therefore not likely to be dramatic.

The comprehensiveness of STAGE, by analogy to SAGE, is limited in principle only by the extent of sequencing. We identified dozens of new E2F4 targets after sequencing a few thousand STAGE tags, but we believe our coverage is not saturating for two reasons. First, we observed minimal overlap between the tags generated from the two independent STAGE pools and saw no significant overlap between their predicted targets, even though we verified targets from each pool. Thus, our sampling of tag space, although valid, is relatively sparse. Second, a substantial fraction of the tags in all the combined human STAGE pools was observed only once. These observations suggest that E2F4 STAGE tags generated by further sequencing are likely to be unique and will help predict additional target genes. One way to estimate the false negative rate in future studies would be to compare the predictions from STAGE after saturation sequencing, with predictions made by analyzing ChIP on complete tiling microarrays for a given chromosome⁹.

STAGE has many advantages for the analysis of genome-wide DNA protein interactions, especially in large genomes. First, it does not make assumptions about the location of protein binding sites on the genome. 98% of the human genome is within 1 kb of an *Nla*III site, so binding sites anywhere can potentially be sampled by STAGE. Second, it does not require expensive infrastructure. We estimate that sequencing 30,000 tags, which should allow for extensive coverage of the targets of a single protein, will entail sequencing about 1,200 clones, a cost-efficient option. Third, STAGE is readily applicable to any sequenced organism. Finally, STAGE is not restricted to a specific annotation of a genome; as new transcriptional units are discovered and existing ones become defunct^{23,24}, the same STAGE tag data can be reanalyzed to identify targets based on revised genome annotations.

We envision STAGE as a useful complement to ChIP-chip for analyzing the binding distribution of proteins on the genome. Although STAGE is a high-throughput genomic method, it is less suited than ChIP-chip for repeated quantitative measurements of the binding of a protein under a range of physiological conditions. However, the binding loci predicted by STAGE can be represented on focused microarrays for ChIP-chip. Thus, an initial comprehensive survey of direct binding targets by STAGE, followed by extensive ChIP-chip analysis, can accelerate the discovery of protein-binding regulatory elements in genomes.

METHODS

Cells and antibodies. Yeast cells with a 3×hemagglutinin (HA)-tagged TBP²⁵ were grown at 25 °C in synthetic complete medium minus uracil, collected by centrifugation, resuspended in an

equal volume of prewarmed 39 °C medium and returned to 39 °C. After 10 min, cells were cross-linked by adding formaldehyde (final 1%). Anti-HA antibody (Santa Cruz) at a 1:100 dilution was used for ChIP.

Human foreskin fibroblasts (ATCC CRL 2091) were grown to 60% confluence in 15 cm plates in DMEM containing glucose (1 g/l), antibiotics, and 10% FBS (Hyclone). Cells were washed twice with the same medium lacking FBS and low-serum medium (0.1% FBS) was added. After 72 h, cells were cross-linked with formaldehyde (final 1%). Anti-E2F4 antibody (sc-1082x, Santa Cruz) at a 1:100 dilution was used for ChIP.

STAGE and SubSTAGE. Cross-linking, ChIP, and amplification of ChIP DNA was performed as described previously²⁶, except using a 5'-biotinylated primer during amplification. Further details of ChIP protocols are described in **Supplementary Methods** online. We then followed the LongSAGE protocol (<http://www.sagenet.org/>), except used amplified, biotinylated ChIP DNA as the starting material. Briefly, amplified DNA (1–2 µg) was digested with *Nla*III. The terminal DNA fragments were bound to streptavidin-coated magnetic beads (Dynal) and separated into two tubes. After ligation with linker 1 or 2, which contain recognition sites for *Mme*I, the DNA fragments were released by *Mme*I digestion. The released tags were ligated to generate ditags. Ditags were amplified with nested primers, gel purified and trimmed by *Nla*III digestion. Trimmed ditags were gel purified, concatenated by ligation and cloned into the pZero 1.0 vector (Invitrogen). Insert sizes were assayed in recombinant clones and clones containing at least ten ditags were sequenced. Details of the subtraction step are provided in **Supplementary Methods** online. For the mock immunoprecipitation control and reference samples (Figs. 3 and 4, respectively), the antibody was omitted. For the genomic control STAGE pool, sheared normal human genomic DNA was used as input into STAGE.

Data analysis and scoring. STAGE yields a list of tags with their number of occurrences in the pool. This number is termed *nocc*. Each valid STAGE tag has anywhere between one and several thousand matches on the human genome. This number is termed *nhit*. Our algorithm for defining target genes was as follows. (1) Map the tags to the human genome. (2) Assign a score to each tag based on *nocc* and *nhit*. (3) For each human gene, identify tags within a user-defined window. (4) Calculate a cumulative score for the gene based on the scores of all tags in the given window. (5) Compare these scores to the experimental and computational control. (6) Genes that show a substantially higher score than the control are putative targets. Further details of the scoring algorithm are provided in **Supplementary Methods** online. For all analyses, we used the July 2003 build of the Human Genome sequence assembly available at <http://genome.ucsc.edu>. Genes used in our analysis were based on the RefSeq Genes annotation at the University of California, Santa Cruz¹⁹.

Controls and P values for STAGE enrichment scores. For an experimental control, we performed STAGE on input genomic DNA without ChIP and calculated background gene scores for all genes in the same manner as described above for STAGE from an actual ChIP. Raw gene scores derived from ChIP STAGE were divided by control scores to obtain the final STAGE enrichment

score. To calculate a *P* value for the STAGE enrichment score, 2,000 tags were computationally selected at random from the redundant pool of all CATG 21-mers in the genome and used to generate scores for each gene as described above. This process was iterated 500 times to obtain a distribution of 500 scores for each gene. For each gene, these scores were fitted to a normal distribution. The experimentally determined STAGE enrichment score for a particular gene was compared to this distribution and a *P* value for the score was obtained. Experimental scores with *P* values less than 0.01 were taken to be significant.

Microarrays. Yeast microarrays including all ORFs and intergenic elements were manufactured as described previously^{5,26}. PCR amplification, fluorescent labeling of ChIP DNA fragments and hybridization were performed as described previously²⁶. The reference hybridization probe was generated from sonicated normal yeast genomic DNA processed identically to the probe for ChIP DNA samples. A GenePix 4000B scanner and GenePix Pro 4.0 software (Axon Instruments) were used for scanning and quantitation. Data were uploaded to a local database for analysis²⁷. The enrichment value of TBP ChIP was calculated by ranking genomic loci according to their red/green fluorescence ratios. We determined the percentile rank (0–100) for each array element and either used it directly as a measure of binding (Fig. 2b) or used the average percentile rank for each element from two replicate hybridizations (Fig. 2a). When multiple microarray elements could potentially represent the promoter of a gene, we averaged their percentile ranks.

PCR primer pairs for human core promoters²⁰ were purchased from the Whitehead Institute (Cambridge, Massachusetts, USA). Promoters were amplified by PCR as recommended by the manufacturer, and microarrays were manufactured as previously described²⁶. PCR products corresponding to 33 additional promoter and control loci, including the genes listed in **Supplementary Table 6** online, were included on the array. E2F4 ChIP DNA fragments and the mock IP reference samples were amplified and labeled by ligation-mediated PCR, using Cy5 and Cy3, respectively⁶. The two fluorescently labeled samples were simultaneously hybridized to the promoter microarray and ChIP enrichment of target loci was calculated by ranking the Cy5/Cy3 (red/green) fluorescence ratios.

PCR and primers. Thirty cycles of PCR were performed for the samples in **Figure 4** in a 25- μ l reaction volume with 1 μ l (4%) of immunoprecipitated material. Primers were designed to assay approximately between –400 bp and +1 of the transcription start site. The ninth exon of *CCNB1* and the core promoter of *ACTB* were negative controls NC1 and NC2, respectively (Figs. 3a–c and 4a). Primer sequences are provided in **Supplementary Table 6** online.

Accession numbers. Microarray data have been deposited in NCBI's Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) and are accessible through GEO Series accession number GSE1861.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank K. Struhl for the HA-tagged TBP strain, P. Killion for assistance with the microarray database and T. Hart and members of the Iyer lab for assistance with

microarray production. This work was supported in part by a grant from the Texas State Higher Education Coordinating Board, a US Department of Defense Breast Cancer Idea Award and a National Science Foundation Information Technology Research (ITR) grant.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 14 June; accepted 5 November 2004

Published online at <http://www.nature.com/naturemethods/>

- Pollack, J.R. & Iyer, V.R. Characterizing the physical genome. *Nat. Genet.* **32** (Suppl.), 515–521 (2002).
- Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
- Yu, H., Luscombe, N.M., Qian, J. & Gerstein, M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* **19**, 422–427 (2003).
- Phimister, B. Getting hip to the chip. *Nat. Genet.* **18**, 195–197 (1998).
- Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
- Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
- Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
- Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H. & Farnham, P.J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* **16**, 235–244 (2002).
- Martone, R. *et al.* Distribution of NF- κ B-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* **100**, 12247–12252 (2003).
- 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* **116**, 499–509 (2004).
- Weinmann, A.S., Bartley, S.M., Zhang, T., Zhang, M.Q. & Farnham, P.J. Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol. Cell. Biol.* **21**, 6820–6832 (2001).
- Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
- Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. Analysing uncharted transcriptomes with SAGE. *Trends Genet.* **16**, 423–425 (2000).
- Roberts, D.N., Stewart, A.J., Huff, J.T. & Cairns, B.R. The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. *Proc. Natl. Acad. Sci. USA* **100**, 14695–14700 (2003).
- Kim, J. & Iyer, V.R. Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol. Cell. Biol.* **24**, 8104–8112 (2004).
- Hahn, J.S., Hu, Z., Thiele, D.J. & Iyer, V.R. Genome-wide analysis of the biology of stress responses through heat shock transcription factor. *Mol. Cell. Biol.* **24**, 5249–5256 (2004).
- Cam, H. & Dynlacht, B.D. Emerging roles for E2F: beyond the G1/S transition and DNA replication. *Cancer Cell* **3**, 311–316 (2003).
- Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* **31**, 34–37 (2003).
- Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
- Odom, D.T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
- Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512 (2002).
- Matsumura, H. *et al.* Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. USA* **100**, 15718–15723 (2003).
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
- Hild, M. *et al.* An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* **5**, R3 (2003).
- Kuras, L. & Struhl, K. Binding of TBP to promoters *in vivo* is stimulated by activators and requires Pol II holoenzyme. *Nature* **399**, 609–613 (1999).
- Iyer, V.R. in *DNA Microarrays: A Molecular Cloning Manual* (eds. D. Bowtell & J. Sambrook) 453–463 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2003).
- Killion, P.J., Sherlock, G. & Iyer, V.R. The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics* **4**, 32 (2003).