AD_____

Award Number: W81XWH-04-1-0472

TITLE: Genome-Wide Chromosomal Targets of Oncogenic Transcription Factors

PRINCIPAL INVESTIGATOR: Vishwanath R. Iyer

CONTRACTING ORGANIZATION: The University of Texas at Austin
Austin, TX 78712-0159

REPORT DATE: April 2008

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-04-2008	Final	31 Mar 2004 – 30 Mar 2008

4. TITLE AND SUBTITLE

Genome-Wide Chromosomal Targets of Oncogenic Transcription Factors

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0472

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Vishwanath R. Iyer

Email: vishy@mail.utexas.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Texas at Austin
Austin, TX 78712-0159

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT We proposed to develop a new genomic method named STAGE (Sequence Tag Analysis of Genomic Enrichment) to identify the direct downstream targets of transcription factors important in breast cancer. STAGE was based on high-throughput sequencing of concatamerized tags derived from DNA associated with transcription factors isolated by chromatin immunoprecipitation. We have successfully accomplished the original goals of the project. The advent of next-generation sequencing technologies (Solexa, 454) after the inception of this project provided new opportunities to enhance our original idea of developing sequencing based methods of target identification. We have taken advantage of the power of next generation sequencing, and applied it to several transcription factors important in cancer and cell proliferation. Since the binding and function of transcription factors is strongly governed by chromatin and positions of nucleosomes, we have also adapted the sequencing approach to identify positioned nucleosomes genome-wide at high resolution.

15. SUBJECT TERMS
ONCOGENES, GENOMICS, TRANSCRIPTION FACTORS, CHROMOSOMAL TARGETS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	40	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

Introduction

The objective of this Idea Development project was to develop a novel, unbiased, sequencing-based genomic approach for identifying the direct chromosomal targets of transcription factors that are important in breast cancer. Cancer involves, at least in part, aberrant programs of gene expression mediated by oncogenic transcription factors activating downstream target genes. Distinguishing between direct and indirect targets of transcription factors is important for reconstructing the transcriptional regulatory networks that underlie complex gene expression programs that are activated in cancer. Transcription factors have been proposed as targets of anti-cancer therapy [1]. Identification of the target genes of oncogenic transcription factors is therefore of great interest and an area of intensive investigation. The binding of oncogenic transcription factors to their cognate sites in vivo is strongly influenced by chromatin structure and the positions of nucleosomes along the promoter and relative to potential binding sites.

The direct in vivo binding targets of a transcription factor can be identified using the technique of chromatin immunoprecipitation (ChIP), where DNA bound by a transcription factor in vivo is first isolated after crosslinking and immunoprecipitation. Genomic identification of these binding sites is customarily accomplished by hybridization to a comprehensive whole-genome microarray that includes all potential regulatory elements (ChIP-chip). We proposed to develop STAGE (Sequence Tag Analysis of Genomic Enrichment) as an alternative to whole-genome hybridization for target identification. STAGE was based on high-throughput sequencing of short sequence tags from DNA isolated by ChIP. These tags are mapped back to the reference human genome sequence and computational analysis of the localization and clustering of the tags enables identification of the binding sites of the transcription factor.

We have accomplished the original objectives of the project in terms of developing the technology. We have further extended it to take advantage of next-generation sequencing methodologies that were introduced since the inception of this project. The first part of this Final Report describes the research accomplishments with reference to the approved Statement of Work, previous Annual Reports, and publications arising out of this project. Subsequently, we describe progress over the last year since the last Annual Report, including unpublished data.

Body

Task 1 *Develop STAGE to identify direct chromosomal targets of transcription factors.*

A) Our development of STAGE (Sequence Tag Analysis of Genomic Enrichment) first involved a proof-of-principle experiment in yeast where we piloted the ability of our tag-sequencing based method to identify targets of the general transcription factor TBP (TATA-box binding protein). We then applied this technique for identifying novel targets in human cells of the transcription factor E2F4, a member of the E2F family of transcription factors involved in cell proliferation. Concurrent with the development of the experimental methods for carrying out STAGE, we also developed and implemented computational pipelines for aligning the sequence tags back to the human genome

sequence, as well as for quantitating the statistical enrichment of neighboring tags so as to signify enrichment in the ChIP. These accomplishments were described in the first Annual report (2005) and the report of our development of STAGE as a novel technology was published as an Article in *Nature Methods* [2], which is included in the Appendix.

B) Although we had initially planned to carry out STAGE analysis of targets of ER in breast cancer, competing genome-wide analyses by other labs prompted us to focus our analysis on c-Myc and E2F4, which are both relevant in breast cancer (Annual Report 2006), as well as consider other transcription factors and applications of sequencing technologies. This was detailed in the 2006 Annual report. At this time we were still using standard sequencing technology, which involved making concatamers of STAGE sequence tags, cloning into plasmid vector, and sequencing by standard Sanger chemistry.

The introduction of 454 sequencing technology in 2005 [3] offered us the opportunity to tremendously increase the depth of sequencing in STAGE, and at the same time bypass the laborious concatamerization and cloning steps required for the original protocol. We worked on adapting STAGE to the 454 platform. We succeeded in using 454 technology to identify novel targets of the transcription factor STAT1 as well as Myc. Our success with these efforts was described in the 2006 and 2007 Annual Reports, and was published as a report in *Genome Research* [4] which is included in the Appendix.

Task 2 *Validation, analysis and interpretation of direct targets identified by STAGE*

A) For validation, analysis, and interpretation of transcription factor targets identified by STAGE, we have adopted several parallel approaches. *First*, we carry out computational simulations to estimate the false discovery rate (FDR) of our tag clustering algorithms. *Second* we perform real-time quantitative PCR (qPCR) comparing STAGE-identified targets between ChIP and input samples. These methods of analysis have been described in previous Annual reports as well as in the publications of the initial STAGE results for E2F4, STAT1 and Myc [2,4]. *Third*, we have used tiling microarrays for a more global comparison of targets identified by STAGE with targets identified by ChIP-chip. Our joint data on Myc targets identified by STAGE and by ChIP-chip formed part of the ENCODE Consortium publication last year in *Nature* [5], with the Myc ChIP-chip data serving both as a verification of STAGE results, as well as providing insights into the binding of this transcription factor near promoters. *Fourth*, we have examined the enrichment of DNA sequence motifs in the regions identified as binding sites by STAGE – here we consider both the enrichment over background of motifs expected for the transcription factor that was under study (Myc or STAT1), and also the co-enrichment of motifs for other transcription factors that may bind cooperatively and regulate gene expression together with the chosen factor. This analysis of motif co-enrichment was presented earlier for STAT1 and Myc (with Solexa sequencing) in our 2007 Annual Report, and the analysis for STAT1 is part of the Genome Research publication included in the Appendix. A similar analysis for our recent Solexa sequencing data for E2F4 is described below. We have also developed a computational

method to identify statistically and biologically significant pairs of transcription factor binding sites in the human genome, which could be applied to data from our ChIP-seq results [6]. *Finally*, we consider enrichment of functional categories of genes identified as targets by STAGE and the pathways mediated by them as providing insights into the functions targeted by the transcription factor. Again, this analysis was included in our 2007 Annual Report for STAT1 and in the *Genome Research* paper in the Appendix, and is included below for new data with E2F4.

B) Use of Solexa ChIP-seq. Although whole-genome tiling arrays are becoming increasingly available from Affymetrix and NimbleGen, their use for ChIP-chip of oncogenic transcription factors remains challenging. Two key limitations of genome-wide ChIP-chip are first, that tiling arrays still cover only the non-repetitive portions of the genome (50-70%). Since a sequencing method such as STAGE is an unbiased sampling of the entire genome, it relieves this limitation to a considerable extent. Second, ChIP DNA, which is low in yield, needs to be amplified significantly in order to hybridize it to several tiling arrays which are required to cover the genome. Next-generation sequencing technologies (454, Illumina/Solexa, ABI-SOLiD, and others on the horizon such as Helicos) make it possible to sequence hundreds of thousands to millions of sequence tags from small ChIP sample sizes without the need for cloning in plasmid vectors as we originally conceived in our proposal, and they require only limited amplification. As such, the use of Solexa and 454 type sequencing technologies (generally termed ChIP-seq) represent the natural evolution of the STAGE method for ChIP analysis that was originally conceived in this project.

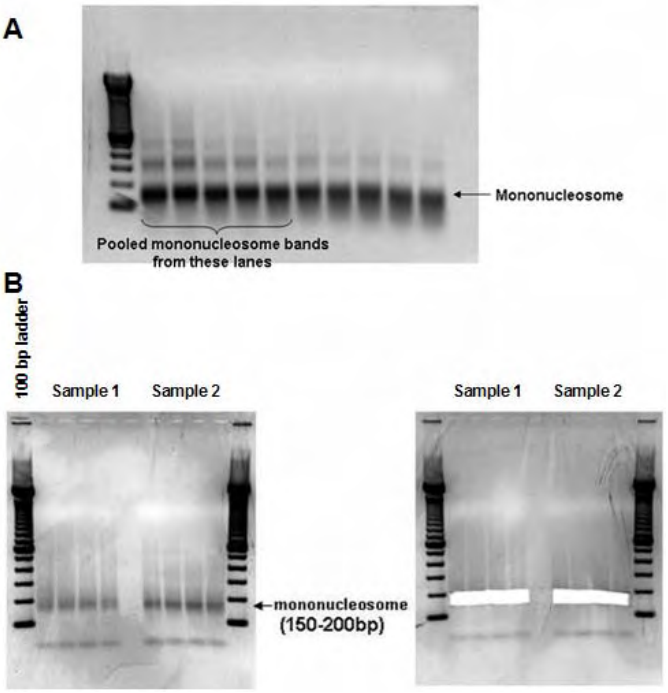


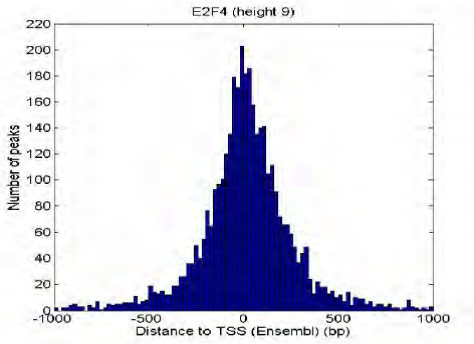
Fig. 1 Nucleosomes isolated by array capture. (A) Representative example of mononucleosomes generated by micrococcal nuclease (MNase) digestion of K562 erythroleukemia cell chromatin. (B) Nucleosomes from two independent samples as in A were captured using ENCODE region tiling arrays (NimbleGen), amplified by PCR and analyzed on a gel. The left panel shows captured mononucleosomes and the right panel is after gel extraction of the mononucleosomal product. This product is being sequenced using Solexa technology.

The sequencing approach has a further advantage over microarray based approaches when single nucleotide resolution is desired for identifying the ends of the DNA fragments. Although this is not useful for ChIP-chip where the ends of the isolated DNA fragments are generated by random ultrasonication, single nucleotide resolution is extremely useful for mapping the position of individual nucleosomes whose positions along

the genome to form chromatin strongly influence where oncogenic transcription factors bind in vivo. Since the last Annual Report, our main focus has been to extend STAGE ChIP-seq (using the Solexa platform) to additional transcription factors that are relevant to cell proliferation and breast cancer, as well as to develop the methodology to map single nucleosomes in chromatin using the sequencing approach.

C) Nucleosome/chromatin analysis For single-nucleosome mapping, we have carried out a successful proof of principle study initially in yeast by using our ultra-high throughput sequencing approach to map all the single nucleosomes and their dynamic remodeling in response to a transcriptional perturbation. This work was recently published as an article in *PLoS Biology* [7] (included in Appendix). Importantly, this work allowed us to develop the computational infrastructure to define the ends of individual nucleosomes from ultra high-throughput sequencing data, which will be applicable to nucleosome mapping in the human genome. In order to accomplish nucleosome mapping in human, the required depth of sequencing to cover all nucleosomes still makes it impractical to carry out on a large enough scale. However, we are now developing a solution by using array capture methodology [8] to target in-depth sequencing to a desired region of the genome. Our initial tests using array capture to recover DNA from selected loci have yielded positive results (Fig. 1), and we have now prepared nucleosomal DNA from the ENCODE regions as a proof of principle and submitted these samples for Solexa sequencing.

The transcription factors that we are currently analyzing by ChIP-seq using Illumina/Solexa include the immediate early oncogenic factor SRF (Serum Response Factor), its co-factors ELK1 and ELK4, the factor E2F4, and its cofactor p130. SRF has been implicated in immediate-early gene regulation in MCF7 breast cancer cells [9] while ELK1 and ELK4 are *ets* family transcription factors implicated in breast and other cancers [10,11]. E2F4 is also known to be involved in breast cancer [12,13]. Although we had initially developed our STAGE method using E2F4 as an example (also described in Annual Report 2005-2006), to date there is no truly unbiased whole-genome ChIP analysis of this important transcription factor. Below is described some of our recent analysis of E2F4 ChIP-seq data. We have just received similar ChIP-seq data for SRF, ElK1 and ELK4, and our p130 ChIP-seq sample is currently being processed.

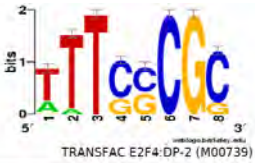


Distance to TSS	-50K<	-50K:-10K	-10K:-1K	-1K:-0.2K	-0.2K:0	TSS	0:0.2K	0.2K:1K	>1K
N peaks	272 5.3%	225 4.4%	223 4.4%	496 9.8%	1233 24.2%		1321 26.0%	628 12.3%	688 15.3%

Fig. 2 Binding distribution of genome-wide E2F4 targets in relation to the transcription start sites (TSS). The top panel is a histogram of peak heights ≥ 9 relative to the TSS. The bottom table shows the fraction of binding peaks in each indicated interval relative to the TSS in Kb.

D) ChIP-seq for additional transcription factors In collaboration with the British Columbia Genome Sequencing Center in Vancouver, we have generated ChIP-seq

datasets for several transcription factors including E2F4. In some cases (SRF, ELK1, ELK4), these ChIP experiments were carried out to examine differential binding in quiescent, serum-starved primary fibroblasts versus proliferating, serum-treated cells. For E2F4 ChIP carried out in GM06990 lymphoblastoid cells, we obtained 11.6 million Solexa reads averaging 32 bp each, of which 6.1 million were uniquely aligning to the reference human genome sequence. Since binding of any transcription factor to its target sites in vivo is expected to vary continuously over a large range of affinities, it is not trivial to segregate genomic loci cleanly into "targets" and "non-targets". However, the strength of binding, and the enrichment of a given locus in the ChIP correlates with the number of overlapping reads, indicated by the "height" of a sequence peak. At a False Discovery Rate (FDR) of 0.1%, corresponding to a peak height of 9, we could identify 5086 binding targets for E2F4 in the genome. This represents the first truly unbiased whole-genome identification of E2F4 binding targets in the human genome. Our preliminary analysis of subsets of the reads indicates that we have identified the vast majority (~90%) of actual E2F4 binding sites in the genome in this experiment.



>Motif no. 1

Motif (IUPAC notation):
5' NNNNGCGNNAAANNNN 3' (sense)
5' NNNNTTTNNCGCNNNN 3' (anti-sense)

p-value < 2.21^{-016}
Bonferroni corrected p-value < 1.81^{-012}
Enrichment: 1.52 fold

Fig. 3 Motif analysis of genome-wide E2F4 targets. The top shows the consensus motif for E2F4 as designated by TRANSFAC. The bottom shows a matching motif that was discovered by the program DRIM in the binding sites for E2F4 identified by ChIP-seq. The binding sites are statistically significant as indicated.

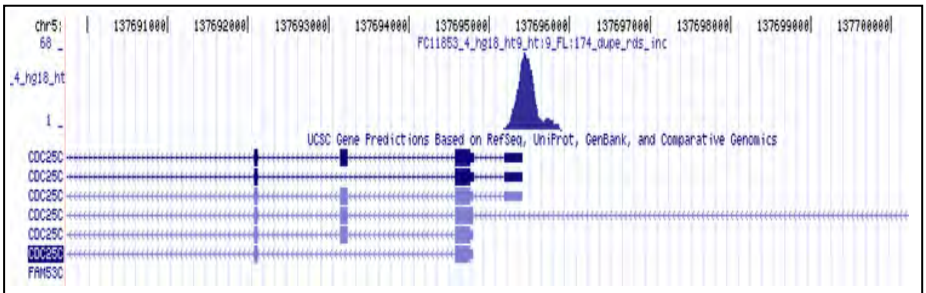


Fig. 4 Sample of Solexa ChIP-seq data showing a strong binding site for E2F4 (peak height 68) upstream of the cell cycle gene CDC25.

As we have observed earlier with Myc, there was a strong tendency for E2F4 binding to be near the transcriptional start sites of genes (Fig. 2). Analysis of the DNA sequences corresponding to the ChIP-seq peaks using the pattern discovery program DRIM, revealed the strong enrichment of the E2F binding motif (Fig. 3). Motif co-occurrence analysis indicated that out of 363 high-quality vertebrate transcription factor binding motifs available in the TRANSFAC database, 86 sequence motifs were found to occur ≤ 6 bases from an E2F4 motif. These include the motifs for transcription factors such as AP-2, ELK1, NF-κB, Myc, Egr-4, C/EBP and others, which could potentially be co-regulators with E2F4 of its target genes. However the statistical and biological significance of this type of co-factor occurrence remains to be determined and will be analyzed in the coming months. Analysis of enrichment of functional gene categories among identified targets revealed functions such as cell-cycle regulation (Fig. 4), DNA

replication and damage, cell-cycle checkpoints, chromosome segregation, etc, which were all statistically significantly enriched, consistent with a role for E2F4 in regulating cell-cycle progression.

We have currently started receiving similar ChIP-seq data for the transcription factors SRF, ELK1 and ELK2, as well as the E2F4 co-factor p130. Fig. 5 shows the verification of ChIP for the first three transcription factors, using qPCR to show binding of these proteins to target promoters in quiescent and serum-stimulated cells. Analysis of the targets of these oncogenic transcription factors will shed considerable light on the regulatory networks mediated by them in proliferating cells and in cancer. We anticipate preparing these new results for publication in the coming months.

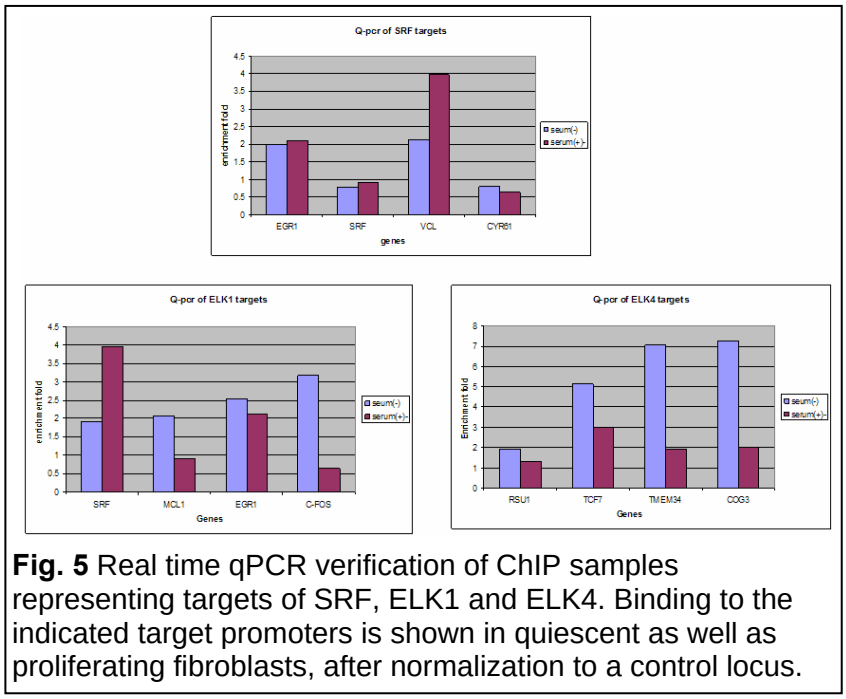


Fig. 5 Real time qPCR verification of ChIP samples representing targets of SRF, ELK1 and ELK4. Binding to the indicated target promoters is shown in quiescent as well as proliferating fibroblasts, after normalization to a control locus.

Personnel receiving pay from this project

Simon Ascher	Undergraduate Research Assistant
Akshay Bhinge	Graduate Student
Patrick Killion	Graduate Student
Ryan McDaniell	Graduate Student
Bum-Kyu Lee	Graduate Student
Zheng Liu	Post-doctoral Fellow
Laura Tu	Undergraduate Research Assistant

Key Research Accomplishments

- Developed STAGE (Sequence Tag Analysis of Genomic Enrichment), a novel method to identify the chromosomal targets of transcription factors including oncogenic transcription factors.
- Optimized protocols and developed novel computational analysis methods for successful application of STAGE in mammalian cells.
- Verified target promoters predicted by STAGE through independent means such as promoter specific PCR and microarray hybridization (ChIP-chip).

- Successfully adapted STAGE for sequencing using 454 bead-based pyrosequencing technology and applied it in proof-of-principle experiment to identify targets of Stat1, important in breast cancer.
- Successfully used Solexa next-generation sequencing technology for identifying targets of the oncogenic transcription factor c-Myc, also important in breast cancer.
- Extended STAGE ChIP-seq approach for several additional transcription factors, E2F4, p130, SRF, ELK1, ELK4 that are important in breast and other cancers.
- Applied sequencing approach for identifying positions of nucleosomes which govern transcription factor binding. Completed proof-of-principle experiment in yeast to map nucleosome positions and their remodeling.
- Developed modified approach to apply nucleosome sequencing to human promoters, by combining array capture of promoter nucleosomes with deep Solexa sequencing.

Reportable Outcomes

- Kim, J., Bhinge, A. A., Morgan, X. C., and Iyer, V. R. (2005). Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment, *Nat Methods* **2**, 47-53.
- Kim J. & Iyer V.R. Identifying Chromosomal Targets of DNA-Binding Proteins by Sequence Tag Analysis of Genomic Enrichment (STAGE), in Current Protocols in Molecular Biology Unit 21.10, (Ausubel F.M. *et al*, eds.) John Wiley & Sons.
- Ph.D. awarded to J. Kim for his thesis "Genome-wide mapping of DNA protein interactions in eukaryotes" University of Texas at Austin, December 2005. Dr. Kim was the first author on our published report on STAGE.
- Platform Presentation: "Genome-wide mapping of DNA -protein interactions in large genomes by STAGE — Sequence Tag Analysis of Genomic Enrichment" at the Cold Spring Harbor meeting on "Systems Biology: Genomic Approaches to Transcriptional Regulation", March 2004
- Platform Presentation at the American Society of Microbiology (Texas) meeting, Houston, November 2004
- Grant award R01 HG003532-01 V. Iyer (PI) from NIH/NHGRI "STAGE and FAIRE for Regulatory Element Identification" This is a technology development project funded under the ENCODE Consortium. It is a collaboration between my lab and the lab of Dr. Jason Lieb at University of North Carolina at Chapel Hill. The objective was to develop and combine STAGE and sequencing methods with other methods for isolating open chromatin elements in the human genome.
- ENCODE Teleconference Presentation: "Identifying the Chromosomal Targets of Proteins by STAGE (Sequence Tag Analysis of Genomic Enrichment)" April 21 2006.

- University Continuing Fellowship awarded to Patrick Killion (2005-2006), who is responsible for developing ArrayPlex, used for analysis of transcription factor target data (described in 2006 Annual Report).
- Bhinge, A. A., Kim, J., Euskirchen, G. M., Snyder, M., and Iyer, V. R. (2007). Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE), *Genome Res* **17**, 910-6.
- Grant award U54 HG004563-01 (sub-contract) G. Crawford (PI), V. Iyer co-investigator. "Comprehensive Identification of Active Functional Elements in Human Chromatin". This was an scale-up of the ENCODE consortium project to the whole genome. The objective was to combine methods for open chromatin isolation (DNaseI HS, FAIRE) with ChIP-chip of selected transcription factors in the ENCODE sanctioned cell lines. Next-generation sequencing is used extensively as a readout in this project.
- The ENCODE Project Consortium (2007). (309 co-authors including V. Iyer, A. Bhinge and J. Kim from the Iyer lab) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447, 799-816.
- Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V. R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLoS Biol* **6**, e65.

Conclusions

This Idea Development project was successful in developing STAGE as a high-throughput, sequencing based approach for identifying transcription factor targets. STAGE was applied to the oncogenic transcription factors Myc and E2F4, which are important in breast cancer. The advent of next-generation sequencing technology during the course of this project, and the 1 year no-cost extension allowed us to adapt our procedures and analysis algorithms to these new methods, and take advantage of the increased throughput provided by them. We have also been able to use the same sequencing based approach to examine chromatin structure and nucleosome positioning in unprecedented detail. Recent data from several transcription factors will be analyzed in the coming months and published.

As an outcome of this project, it is possible to comprehensively identify the binding targets of any transcription factor in breast cancer samples. One can now potentially identify targets of transcriptional regulators like Myc, ER, E2F4, etc. in primary breast cancer biopsy samples from cancer patients. It will then become possible to elucidate differences in the molecular pathways and networks mediated by these oncogenic regulators among different cancer patients. One potential utility of this could be that such binding analysis could form the basis for a new type of class discovery approach and sub-typing among cancers, with distinct diagnostic and prognostic value, in a manner similar to, but independent from what has previously been possible with gene expression or proteomic profiling.

References

1) Darnell, J. E., Jr. (2002). Transcription factors as targets for cancer therapy, *Nat Rev Cancer* **2**, 740-9.
2) Kim, J., Bhinge, A. A., Morgan, X. C., and Iyer, V. R. (2005). Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment, *Nat Methods* **2**, 47-53.
3) Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A.*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376-80.
4) Bhinge, A. A., Kim, J., Euskirchen, G. M., Snyder, M., and Iyer, V. R. (2007). Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE), *Genome Res* **17**, 910-6.
5) The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature* **447**, 799-816.
6) Morgan, X. C., Ni, S., Miranker, D. P., and Iyer, V. R. (2007). Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining, *BMC Bioinformatics* **8**, 445.
7) Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V. R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLoS Biol* **6**, e65.
8) Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., and Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing, *Nat Methods* **4**, 907-9.
9) Duan, R., Xie, W., Li, X., McDougal, A., and Safe, S. (2002). Estrogen regulation of c-fos gene expression through phosphatidylinositol-3-kinase-dependent activation of serum response factor in MCF-7 breast cancer cells, *Biochem Biophys Res Commun* **294**, 384-94.
10) Buchwalter, G., Gross, C., and Wasylyk, B. (2004). Ets ternary complex transcription factors, *Gene* **324**, 1-14.
11) Kim, C. G., Choi, B. H., Son, S. W., Yi, S. J., Shin, S. Y., and Lee, Y. H. (2007). Tamoxifen-induced activation of p21Waf1/Cip1 gene transcription is mediated by Early Growth Response-1 protein through the JNK and p38 MAP kinase/Elk-1 cascades in MDA-MB-361 breast carcinoma cells, *Cell Signal* **19**, 1290-300.
12) Ho, G. H., Calvano, J. E., Bisogna, M., and Van Zee, K. J. (2001). Expression of E2F-1 and E2F-4 is reduced in primary and metastatic breast carcinomas, *Breast Cancer Res Treat* **69**, 115-22.
13) Rakha, E. A., Pinder, S. E., Paish, E. C., Robertson, J. F., and Ellis, I. O. (2004). Expression of E2F-4 in invasive breast carcinomas is associated with poor prognosis, *J Pathol* **203**, 754-61.

Appendices

Publications arising out of this project and referenced in this Final Report.

1) Kim, J., Bhinge, A. A., Morgan, X. C., and Iyer, V. R. (2005). Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment, *Nat Methods* **2**, 47-53.

2) Bhinge, A. A., Kim, J., Euskirchen, G. M., Snyder, M., and Iyer, V. R. (2007). Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE), *Genome Res* **17**, 910-6.

3) Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V. R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLoS Biol* **6**, e65.

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 concatemerized 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 yeast[4], 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 promoters[7] or CpG islands[8] 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) products[9] or oligonucleotides[10] 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 concatemerized tags derived from DNA enriched by ChIP. Cloning and sequencing of ChIP DNA has been carried out previously[11], 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 concatemerized, 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

Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology & Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas 78712-0159, USA. Correspondence should be addressed to V.R.I. (vishy@mail.utexas.edu).

PUBLISHED ONLINE 21 DECEMBER 2004; DOI:10.1038/NMETH726

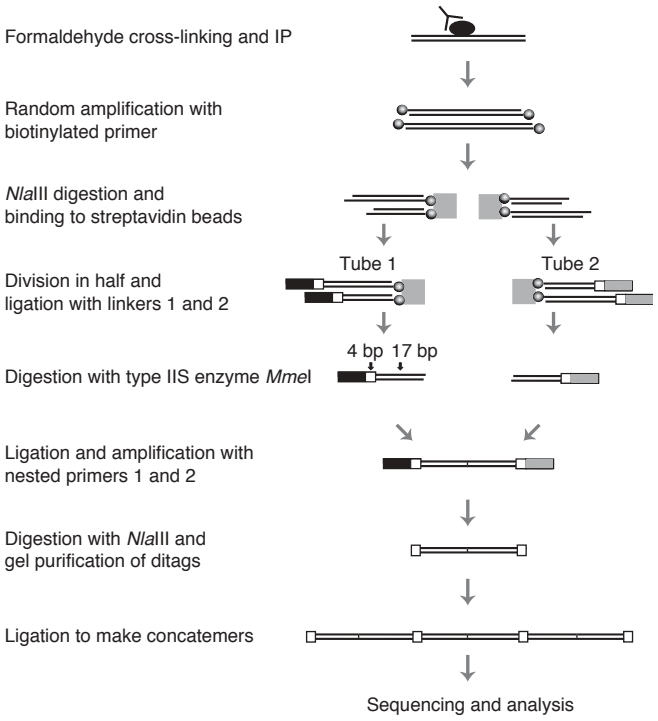


Figure 1 | The STAGE strategy. STAGE is based on high-throughput sequencing of concatemerized tags of defined length that are derived from DNA enriched by ChIP. Proteins were cross-linked to their binding sites *in vivo* with formaldehyde and chromatin was extracted and sheared. The cross-linked protein-DNA 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	*nocc*	Target gene
CATGATGGAAACGAAGACGAC	10	tF(GAA)B
CATGAGAATGTGCTTCAGTAT	8	tF(GAA)B
CATGAAGGTGACAAAATGATT	5	tK(CUU)E1
CATGATCAAATTCTGTGAAGC	5	tL(CAA)A
CATGCAAATCTAAATAAAAAC	5	tH(GUG)H
CATGTATACTTAACAGATATG	5	RDN 5-1
CATGAGATATGCTGTTTCAAG	4	tL(CAA)A
CATGTATATATTGCACTGGCT	4	RDN 5-1
CATGAAACTAGGAAAACGTAC	3	tE(UUC)J
CATGAAGATGATTCGATACCG	3	tV(AAC)M1
CATGATGAAGTTTAGATCTGC	3	tW(CCA)K
CATGATGGCAGACTTCCATCG	3	tV(AAC)G2
CATGATGTCGCTATTTCTAAT	3	tY(GUA)J2
CATGCAAGATGTAGACCCAAC	3	YGRWσ5
CATGCAATCCCAGTAGTAGGT	3	SCR1
CATGCAGCTGTTGTATCAAGA	3	tV(AAC)G1
CATGCATGTTTTACGTTGTGG	3	tP(AGG)N
CATGGAATGTGCAATTAAGAC	3	tT(AGU)N2
CATGTGGTGTAAAAAGATAAC	3	tT(AGU)J
CATGTTATCCTGAGCATCCAC	3	tG(GCC)O2
CATGTTTACCCTCAAACAAAG	3	tV(AAC)K1
CATGTTTCCTCTAAAGATGGT	3	tR(UCU)B
CATGAAAACCTCTCAAACCTT	2	tH(GUG)E1
CATGAAAAGGTTTAATGACTT	2	tT(AGU)O1
CATGAAGACCTATTCGCTTAT	2	tV(AAC)G3
CATGAAGCGCACAAGATTGGA	2	tR(UCU)G3
CATGAATGGCGCAGATTTATT	2	tV(AAC)M1
CATGAGGCGCACTTTTGATTT	2	tY(GUA)F2
CATGAGTTGCCATTAGAAACG	2	tW(CCA)G1
CATGATACTGACTTATTGGGC	2	tD(GUC)L1
CATGCAAGACGTAGACCCAAC	2	tI(AAU)I2
CATGCAAGTGTGGCATAAAAG	2	tK(CUU)E2
CATGCAGAAAAGATAAGATGC	2	YPL029W
CATGCCTGTGCAACGCCGCAG	2	tE(UUC)J
CATGCTCGGCAATAGCTTCAA	2	tG(CCC)D
CATGCTTTGTCTTCCTGTTAG	2	tP(UGG)O2
CATGGAAAAACGAATGGAGAC	2	tA(AGC)K1
CATGGAAATCGAACCTTTCAC	2	tN(GUU)N2
CATGGAGTCTAACTTTGTTGT	2	tN(GUU)O2
CATGGAGTCTTTTATTTCCGA	2	tN(GUU)L
CATGGCAAAAACTGTAAAGTT	2	tR(UCU)G2
CATGGCGAATTTTTCACATAT	2	tV(UAC)D
CATGGCGATTATTTCATTATG	2	tR(UCU)G3
CATGGCTAGTCAAATAAGTGG	2	YGL080W
CATGGGGTAAGTTCCGATGGC	2	tV(AAC)E2
CATGGGTTCAAACACTTCCAA	2	tY(GUA)F1
CATGGTGAAAGTTTAATCTTT	2	tR(ACG)K
CATGTAAACCATCCCTTTTCA	2	YJL005W
CATGTATAAAACCTACCGCTT	2	tS(CGA)C
CATGTATCAAATTGCACGTGA	2	YPRCδ22
CATGTATGAAACTGGGAATTC	2	tS(AGA)B
CATGTCAATGTCCATTTCTTT	2	tT(AGU)I2
CATGTCTTTTGTGGATTATTT	2	tS(CGA)C
CATGTGAGGCTTAGGTGATTT	2	tN(GUU)N2
CATGTGTTTGAATTAGCGATC	2	tL(CAA)A
CATGTTACAATTCCTTTCCAT	2	tG(UCC)G
CATGTTATGTTCAATTGGCAG	2	YELCτ1
CATGTTCAAGGACGGCTTGGT	2	tD(GUC)J1
CATGTTTTCGTTATTTCATAA	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.

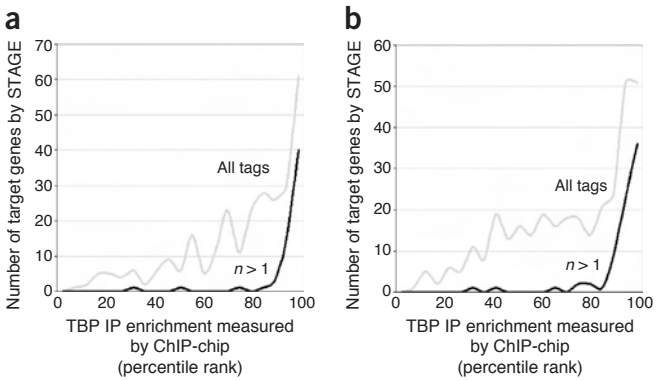


Figure 2 | Correlation between yeast targets predicted by STAGE and ChIP-chip. The enrichment value of yeast TBP targets after ChIP was determined by microarray hybridization. The percentile rank (0–100) of the ratio of ChIP-enriched fragments to genomic DNA was used to determine the ChIP enrichment 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.

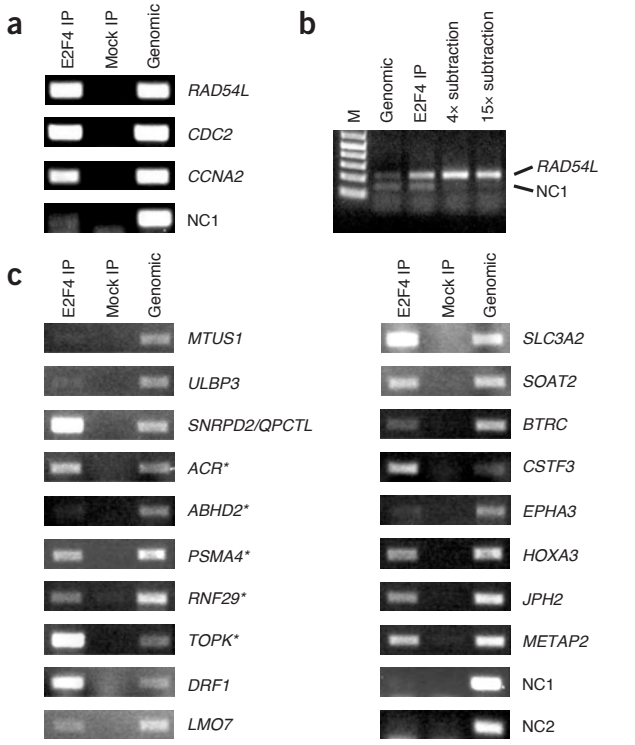


Figure 3 | ChIP of E2F4 targets and validation of STAGE targets by ChIP-PCR. (**a**) Binding of human E2F4 to known target promoters in fibroblasts. PCR was performed using primers corresponding to the promoters of the indicated genes. The ninth exon of *CCNB1* was used as a negative control for ChIP 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.

of each promoter by TBP was indicated by the rank of its enrichment in ChIP relative to the reference[16].

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 G_1 cells[17]. We first used ChIP and promoter-specific PCR to verify the binding of E2F4 to known target promoters[7] (**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 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 microarray[7], 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	*MTUS1*	1971		Mitochondrial tumor suppressor gene 1
2	*ULBP3*	1961		UL16 binding protein 3
3	*SNRPD2**	1933		Small nuclear ribonucleoprotein D2 polypeptide, 16.5 kDa
	*QPCTL**	1923		Hypothetical protein FLJ20084
4	*PXK*	1015		PX domain–containing serine/threonine kinase
5	*FLJ22353*	993		Hypothetical protein FLJ22353
6	*GAJ*	993	Yes	GAJ protein
7	*ACR*	992		Acrosin
8	*RAD54L*	992		RAD54-like (*S. cerevisiae*)
9	*AAMP*	982		Angio-associated migratory cell protein
10	*ABHD2*	982		Abhydrolase domain–containing 2
11	*BLVRB**	982		Biliverdin reductase B (flavin reductase (NADPH))
	*SPTBN4**	982		Spectrin, beta, non-erythrocytic 4
12	*DC2*	982		DC2 protein
13	*FLJ13912*	982	Yes	Hypothetical protein FLJ13912
14	*FLJ25416*	982		Hypothetical protein FLJ25416
15	*FLJ32000*	982	Yes	Hypothetical protein FLJ32000
16	*FLJ90834*	982		Hypothetical protein FLJ90834
17	*MPV17*	982	Yes	MpV17 transgene, murine homolog, glomerulosclerosis
18	*PRCP*	982		Prolylcarboxypeptidase (angiotensinase C)
19	*PSMA4*	982		Proteasome (prosome, macropain) subunit, alpha type, 4
20	*RNF29*	982		Ring finger protein 29
21	*TOPK*	982	Yes	T-LAK cell–originated protein kinase
22	*DRF1*	974		Dbf4-related factor 1
23	*LMO7*	971		LIM domain only 7
24	*SLC3A2*	971	Yes	Solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2
25	*SOAT2*	971		Sterol *O*-acyltransferase 2
26	*ARHGAP11A*	965	Yes	KIAA0013 gene product
27	*ABC1*	961		Amplified in breast cancer 1
28	*BTRC*	961		Beta-transducin repeat containing
29	*GAL3ST1*	961		Cerebroside (3′-phosphoadenylylsulfate:galactosylceramide 3′) sulfotransferase
30	*CSTF3*	961		Cleavage stimulation factor, 3′ pre-RNA, subunit 3, 77 kDa
31	*CTAG3**	961		Cancer/testis antigen 3
	*RIOK1**	961		RIO kinase 1 (yeast)
32	*DNALI1*	961		Dynein, axonemal, light intermediate polypeptide 1
33	*EPHA3*	961		EPH receptor A3
34	*FIBL-6*	961	Yes	Hemicentin
35	*FLJ20712*	961		Hypothetical protein FLJ20712
36	*HIST2H2AC*	961		Histone 2, H2ac
37	*HOXA3*	961		Homeobox A3
38	*JPH2*	961		Junctophilin 2
39	*MAP3K7*	961	Yes	Mitogen-activated protein kinase kinase kinase 7
40	*METAP2*	961	Yes	Methionyl aminopeptidase 2
41	*PDGFA*	961		Platelet-derived growth factor alpha polypeptide
42	*RPL23A*	961	Yes	Ribosomal protein L23a
43	*SNIP1*	961	Yes	Smad nuclear interacting protein
44	*CCRL2*	926		Chemokine (C-C motif) receptor-like 2
45	*C20orf141*	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.

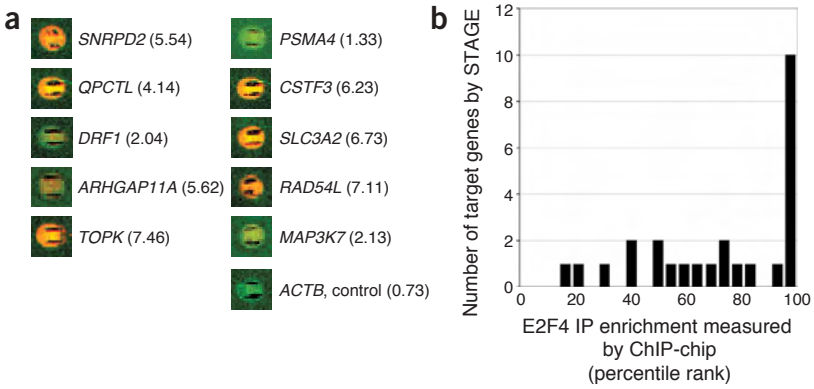


Figure 4 | Validation by ChIP-chip of E2F4 targets predicted by STAGE. DNA from an E2F4 ChIP was amplified and labeled with Cy5, and hybridized to a human core-promoter microarray together with a mock IP sample labeled with Cy3. The ratio of Cy5/Cy3 (red/green) signal is an indicator of the binding of 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 microarray[20] 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-κB[9], 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 STAGE[22]. 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 chromosome[9].

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 TBP[25] 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 previously[26], 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, concatemerized 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 Cruz[19].

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 previously[26]. 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 analysis[27]. 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 promoters[20] 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 described[26]. 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, respectively[6]. 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-μl reaction volume with 1 μl (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/

1. Pollack, J.R. & Iyer, V.R. Characterizing the physical genome. *Nat. Genet.* **32** (Suppl.), 515–521 (2002).
2. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
3. 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).
4. Phimister, B. Getting hip to the chip. *Nat. Genet.* **18**, 195–197 (1998).
5. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
6. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
7. Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
8. 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).
9. Martone, R. *et al.* Distribution of NF-κB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* **100**, 12247–12252 (2003).
10. 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).
11. 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).
12. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
13. Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. Analysing uncharted transcriptomes with SAGE. *Trends Genet.* **16**, 423–425 (2000).
14. 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).
15. 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).
16. 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).
17. Cam, H. & Dynlacht, B.D. Emerging roles for E2F: beyond the G1/S transition and DNA replication. *Cancer Cell* **3**, 311–316 (2003).
18. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* **31**, 34–37 (2003).
19. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
20. Odom, D.T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
21. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512 (2002).
22. Matsumura, H. *et al.* Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. USA* **100**, 15718–15723 (2003).
23. 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).
24. 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).
25. 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).
26. 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).
27. 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).

Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE)

Akshay A. Bhinge,[1,5] Jonghwan Kim,[1,4,5] Ghia M. Euskirchen,[2,3] Michael Snyder,[2,3] and Vishwanath R. Iyer[1,6]

[1]*Institute for Cellular and Molecular Biology, Center for Systems and Synthetic Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas 78712, USA;* [2]*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA;* [3]*Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA*

Identifying the genome-wide binding sites of transcription factors is important in deciphering transcriptional regulatory networks. ChIP-chip (Chromatin immunoprecipitation combined with microarrays) has been widely used to map transcription factor binding sites in the human genome. However, whole genome ChIP-chip analysis is still technically challenging in vertebrates. We recently developed STAGE as an unbiased method for identifying transcription factor binding sites in the genome. STAGE is conceptually based on SAGE, except that the input is ChIP-enriched DNA. In this study, we implemented an improved sequencing strategy and analysis methods and applied STAGE to map the genomic binding profile of the transcription factor STAT1 after interferon treatment. STAT1 is mainly responsible for mediating the cellular responses to interferons, such as cell proliferation, apoptosis, immune surveillance, and immune responses. We present novel algorithms for STAGE tag analysis to identify enriched loci with high specificity, as verified by quantitative ChIP. STAGE identified several previously unknown STAT1 target genes, many of which are involved in mediating the response to interferon-γ signaling. STAGE is thus a viable method for identifying the chromosomal targets of transcription factors and generating meaningful biological hypotheses that further our understanding of transcriptional regulatory networks.

[Supplemental material is available online at www.genome.org.]

The ENCODE project has suggested that a larger fraction of the human genome than previously suspected may be transcriptionally active (The ENCODE Project Consortium 2006). Correspondingly, a significant fraction of the genome is likely to be involved in regulating gene expression and other aspects of human biology. Much of the regulatory potential of *cis*-acting sequences in the genome involves interactions of proteins with DNA. Identifying the genomic binding sites of regulatory proteins such as transcription factors is important for cataloging the regulatory potential encoded in the human genome and reconstructing transcriptional regulatory networks. Chromatin immunoprecipitation (ChIP) combined with microarray hybridization (ChIP-chip) has enabled global mapping of transcription factor binding sites in the human genome (Kim et al. 2005a; Lee et al. 2006). Although whole-genome oligonucleotide tiling arrays are becoming available for ChIP-chip analyses, they remain expensive and entail specialized resources. Another limitation with the use of tiling arrays is that they typically do not cover repetitive sequences, which account for a significant fraction of the genome. For example, recent "whole-genome" tiling arrays included only ~50% of the genome that was nonrepetitive (Kim et al. 2005a; Lee et al. 2006). Binding sites and functional elements that lie near repetitive sequences are therefore likely to be undetected through the use of such arrays. Many tiling array platforms currently need seven to a few dozen arrays to cover the genome, requiring significant scale up of antibody, cell culture material, and effort, especially if replicate experiments are performed.

We have developed an unbiased genomic method to map transcription factor binding sites called STAGE (Sequence Tag Analysis of Genomic Enrichment), based on sequencing "tags" or short oligonucleotide signatures from ChIP-enriched DNA (Kim et al. 2005b). Since it is not constrained by the availability of tiling microarrays for any particular organism, STAGE makes it possible to experimentally determine whether the target genes of a transcription factor in one species are also targets in a related species. Similar sequencing-based approaches for identifying transcription factor targets have recently also been independently developed in other labs (Impey et al. 2004; Roh et al. 2004, 2005; Chen and Sadowski 2005; Loh et al. 2006; Wei et al. 2006).

In order to make STAGE more competitive with genome-wide tiling arrays, we have now developed modifications that exploit new developments in sequencing technology. Here we use STAGE for analysis of the targets of the transcription factor, STAT1. We used bead-based pyrosequencing (454) technology to improve the throughput and cost-effectiveness of sequencing and significantly reduce the time and effort needed to perform STAGE (Margulies et al. 2005). STAT (Signal Transducer and Activator of Transcription) proteins are transcription factors that mediate cytokine and growth factor signaling. Interferons modulate cell proliferation, apoptosis, immune surveillance, and im-

[4]Present address: Children's Hospital Boston, Boston, MA, 02115, USA.
[5]These authors contributed equally to this work.
[6]Corresponding author.
E-mail vishy@mail.utexas.edu; fax (512) 232-3472.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5574907. Freely available online through the *Genome Research* Open Access option.

17:910–916 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

mune responses primarily via the JAK-STAT pathway (Platanias 2005). Interferon (IFNG) specifically activates STAT1 which forms homodimers, translocates to the nucleus, and binds to promoters bearing the gamma-activation sequence (GAS) motif and activates (IFNG) inducible genes (Ramana et al. 2000). ChIP-chip analysis of STAT1 targets on chromosome 22 revealed that STAT1 regulates several genes involved in cell growth, apoptosis, immune responses, and lipid metabolism (Hartman et al. 2005). We used STAGE to identify genome-wide STAT1 binding targets after interferon (IFNG) treatment. We also developed improved analysis algorithms to identify target sites with high specificity. Our results indicate that IFNG-induced STAT1 binds to a large number of sites genome-wide and that many of these sites lie proximal to genes that are involved in biological processes modulated by IFNG.

Results

Identifying STAGE tags for STAT1 by bead-based pyrosequencing

DNA bound by STAT1 in IFNG-treated HeLa cells was isolated by ChIP. We generated ditags as described before (Kim et al. 2005b) and amplified ditags by PCR. Amplified ditags were sequenced by 454 Inc., but without the initial nebulization step normally used in their procedure to shear the DNA. Thus, each read typically contained a complete STAGE ditag, flanked by primer sequences. We sequenced a total of 179,954 reads from the STAT1 STAGE tag library, representing about 17 Mb of sequence from one run. After removing duplicate reads, we were able to extract 162,577 tags; 31,353 tags (19%) could not be matched to any location in the genome and were considered orphans. The remaining 131,224 tags were used for further analysis.

If STAGE tags are derived from ChIP-enriched DNA, then the distribution of tags in the STAGE library should deviate from a randomly selected population of tags. We simulated background tag libraries in silico by randomly selecting the same number of tags (131,224 for STAT1) from the entire genome multiple times. Tags that had more than one hit, i.e., a perfect match, on the genome were ignored. The average frequency distribution of single-hit tags in the random library was compared with the experimental STAT1 STAGE library. For a frequency of occurrence of 1, the numbers of tags in the random and real data were similar. However, for a frequency of occurrence of 2 and more, there was strong enrichment in the STAGE library over background (Fig. 1). Thus, the STAGE tags generated by 454 sequencing represented DNA that was distinct from simulated random genomic DNA.

STAGE targets for STAT1

Since ~50% of the human genome consists of repeat sequences, a given tag in the STAGE library may map to multiple locations in the genome. A tag that is represented in the genome at multiple locations would be more likely to be found in the STAGE library by random chance. Hence, a higher frequency of occurrence of a tag in the STAGE library does not necessarily reflect the enrichment of the tag in the ChIP-enriched DNA. To exclude such ambiguous tags in our analysis, we calculated the probability that a given tag was truly enriched over background by ChIP. Each tag was first assigned a probability of enrichment by assuming that the selection of tags from the genome follows a binomial distribution. Details of the calculations and the algorithm we developed to identify significant targets are included in the Methods. Since STAGE tags are derived from ChIP-enriched DNA, multiple tags can be expected to cluster within short regions in the genome similar in size to the fragments isolated by ChIP, as compared to a random library representing no enrichment, where the tags would be expected to be sampled uniformly across wide regions in the genome. We used this rationale to define binding targets. We performed a simulation where we scanned windows of different sizes across each chromosome and counted the frequencies of windows containing different numbers of single-hit tags. For each window size, we determined whether there were a larger number of windows containing a given number of single-hit tags in the real STAGE library as compared to a simulated random library of STAGE tags. A window of 500 bp gave a false discovery rate (FDR) based on simulations of <5% for STAT1 while the number of targets detected was 734 (Fig. 2). The complete set of data for all window sizes used is given in Supplemental Table 1. We used a window of 500 bp for all further analysis. To improve the specificity of target detection, a window was considered a target only if at least one tag within that window was deemed to be enriched. Thus, for each window we calculated two probabilities, namely, the probability of finding a given number of single-hit tags and the probability that at least one of those tags was statistically likely to be enriched. To avoid assigning high probabilities to windows that contained only a single enriched tag, we gave greater weight to the probability of finding a given number of single-hit tags within a window than to the probability of simply finding any enriched tags in that window. This combined probability calculation gave us a false discovery rate of <1% at a probability threshold of 0.95. It should be noted, however, that this false discovery rate is based on in silico analysis under the assumption that selection of STAGE tags follows a binomial distribution. It is possible that experimental manipulations introduce biases that were not modeled in the simulation. STAGE detected 381 binding sites for STAT1 in the entire genome

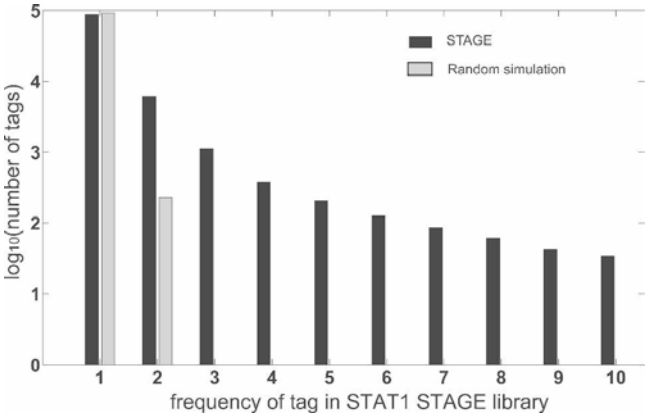


Figure 1. Comparison of the STAT1 STAGE tag library with a simulated randomly generated background library. A background library was generated to simulate STAGE tag libraries by randomly selecting the same number of tags from the genome as the experimental STAGE library. This procedure was repeated 20 times and the values were averaged. Only tags with a single, unique hit on the genome were used in this analysis. The numbers of single-hit tags (Y-axis) were plotted against the frequencies of those tags in the random (gray bars) and experimental (black bars) tag library (X-axis). For frequencies of 2 and above, the STAGE tag library for STAT1 shows a clear enrichment over a randomly generated tag library.

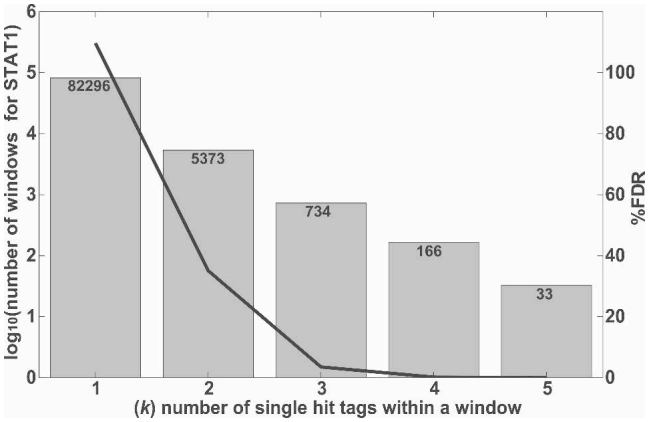


Figure 2. Determination of optimal window size used for target identification. Windows of different sizes (300, 500, 1000, and 2000 bp) were scanned across the entire genome. For each window, we defined *k* as the number of single-hit tags found within the window. The number of windows observed for a given *k* in the STAGE tag data was compared with the number observed in random simulated data. A window size of 500 bp gave an optimal separation between random and real data. Data shown is for a window size of 500 bp. The gray bars indicate $\log_{10}$ of the number of windows detected based on STAT1 tags, with actual numbers of windows at each *k* listed at the *top* of the column. The black line shows the decline in the false discovery rate (FDR) with increasing *k*. The FDR was calculated as the ratio of the number of windows found in the random simulated library to the number of windows detected in the experimental STAT1 library. The raw data for other window sizes is included in Supplemental Table 1.

at this threshold. Based on annotations in the RefSeq gene database (Pruitt et al. 2005), 68% of the STAT1 binding sites found by STAGE were within 50 kb of the transcription start site (TSS) of a gene, 70% of which were found within 20 kb (Table 1).

Verification of STAT1 targets by ChIP-chip and quantitative ChIP

Seven of the 381 STAT1 binding sites identified in the genome by STAGE were within the ENCODE regions. Three of these seven targets overlapped with a ChIP-chip peak where the STAT1 ChIP-chip was performed on ENCODE region tiling oligonucleotide arrays (Fig. 3A; The ENCODE Project Consortium 2007). To obtain a quantitative estimate of the false positive rate of our STAGE analysis, we selected 10 target sites identified by STAGE that had probabilities ranging from 0.95 to 1.0 and assayed their enrichment in a biologically independent STAT1 ChIP sample. Nine out of these 10 sites showed a quantitative enrichment in the ChIP sample relative to the input, with eight of them showing an enrichment of more than twofold (Fig. 3B). Thus, we estimate our true positive rate to be ~90% giving a false positive rate of 0.1. We also compared STAT1 target genes identified by STAGE to STAT1 target promoters that we identified by ChIP-chip using a global core-promoter microarray. The core-promoter microarray included 9764 different promoters where a promoter was defined as 1 kb upstream of and 200 bp downstream from the TSS of a gene. ChIP-chip revealed 157 promoters to be bound by STAT1 at an enrichment ratio greater than threefold. Twenty-nine out of these 9764 promoters had a high-confidence STAT1 binding site, as identified by STAGE, between 1 kb upstream of and 200 bp downstream from the TSS, and 11 out of these 29 were in common with the targets identified by ChIP-chip (Fig. 4A). Under a hypergeometric distribution, this overlap was significant at a *P*-value $<10^{-12}$.

Enrichment of motifs in STAT1 targets

If a STAT1 binding site detected by STAGE occurred within 1 kb upstream of and 200 bp downstream from the TSS of a gene, we considered that gene to be a STAT1 target. STAGE detected 59 genes in RefSeq as STAT1 targets by the above criteria (Supplemental Table 2). Sixty-two percent of these target genes (37/59) had the GAS STAT1 motif TTCNNNGAA within 1 kb upstream of and 200 bp downstream from the TSS of the gene. This represented a motif enrichment among target promoters of more than twofold compared to background. The background in this case was considered as 1 kb upstream of and 200 bp downstream from the TSS of all genes in RefSeq. This enrichment was statistically significant (*P*-value $<10^{-8}$) assuming a hypergeometric distribution.

We applied the same analysis for all STAT1 binding sites in the entire genome. For each window detected as a STAT1 binding site, we searched for the STAT1 GAS motif in that window extending our search to 250 bp on either side of the window. Out of 381 binding sites detected by STAGE, 226 (59.32%) had the GAS consensus sequence. This represents an enrichment of more than twofold over background level of occurrence of the GAS motif in randomly selected windows from the entire genome (*P*-value $<10^{-43}$) (Fig. 4B). Additionally, in accordance with the fact that STAT1 is known to exhibit cooperative binding with other transcription factors like AP1, MYC, and NFKB, we found an enrichment for the STAT1 motif along with motifs for *AP1* (Eferl and Wagner 2003), MYC (Adhikary and Eilers 2005), and NFKB (Martone et al. 2003) (Fig. 4B).

Genes proximal to STAT1 binding sites

STAGE identified several previously unknown STAT1 target genes (Supplemental Table 2), many of which are involved in IFNG signaling. One of these was DAPK3 (death-associated protein kinase 3), a positive regulator of programmed cell death. DAPK3 induces apoptosis by associating with the pro-apoptotic protein DAXX. IFNG is known to increase DAPK3–DAXX complex formation and this complex is necessary for induction of caspases and IFNG-mediated apoptosis (Kawai et al. 2003). STAT1 modulation of DAPK3 could thus represent one mechanism by which IFNG can induce apoptosis. DAPK3 phosphorylates MDM2 and (CDKN1A), components of the TP53 pathway (Burch et al. 2004), and its identification as a STAT1 target suggests a novel collaboration between the IFNG/STAT1 apoptotic pathway and the TP53 tumor suppressor pathway. Another possible mechanism for

Table 1. Percentage distribution of STAT1 binding sites in the entire genome that were proximal to RefSeq annotated genes

Position of binding sites	Percentage of binding sites
Relative to transcription start sites of the gene (percentage of total sites)	
Within 50 kb	68%
Within 20 kb	47%
Within 20 kb upstream	24%
Within 20 kb downstream	23%
Sites found internal to genes (percentage of internal sites found within 20 kb)	
First exon	18%
First intron	42%

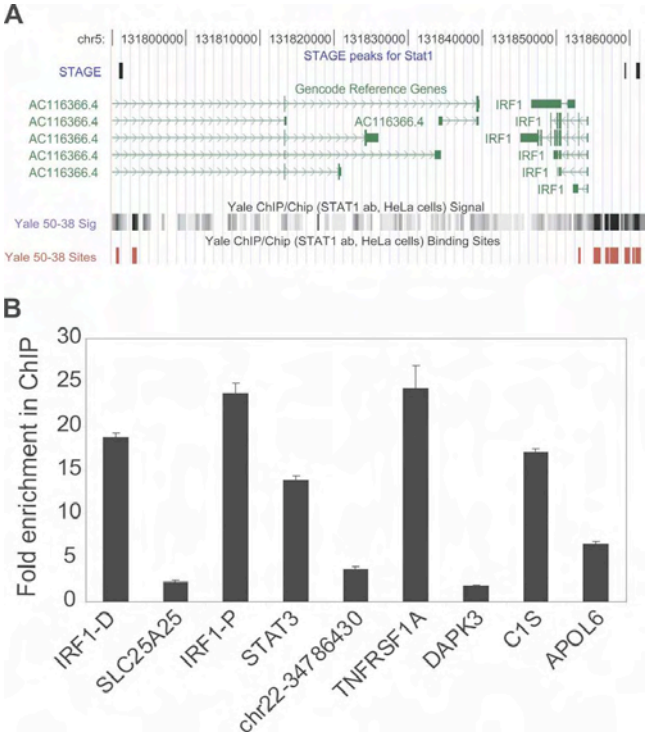


Figure 3. (*A*) STAT1 binding sites in the ENCODE regions. A portion of the ENCODE region ENm002 is shown as displayed in the UCSC Human Genome Browser. Three out of the seven STAT1 binding sites identified by STAGE matched STAT1 binding sites identified by ChIP-chip analysis performed on NimbleGen ENCODE region tiling arrays. Transcripts identified in this region by the GENCODE project are shown in green. The *bottom* shows raw ratio data as well as peak calls for STAT1 binding sites from NimbleGen ChIP-chip data. (*B*) Quantitative ChIP analysis of binding sites identified by STAGE. Nine out of 10 binding sites detected by STAGE were validated as true binding loci by quantitative PCR. Columns show fold enrichment of each locus in the ChIP sample relative to input DNA, normalized to an unrelated control locus. STAGE detected two binding sites separated by >1500 bp in the *IRF1* promoter which are indicated in the figure. *IRF-D* indicates the distal (*IRF1*-distal) and *IRF1-P* indicates the proximal site (*IRF1*-proximal). No genes were found in the proximity of the site indicated as chr22-34786430.

IFNG-mediated apoptosis was suggested by the observation that APOL6, which induces mitochondria-mediated apoptosis characterized by the release of cytochrome-c and activation of caspase-9 (Liu et al. 2005), was also identified as a STAT1 target by STAGE.

STAT3 is anti-apoptotic and induces cell proliferation while STAT1 promotes growth arrest and apoptosis (Stephanou et al. 2000; Stephanou and Latchman 2005). In mouse embryonic fibroblasts, it was shown that IFNG induces high levels of expression of STAT1 while STAT3 levels remain low. However, in the absence of STAT1, i.e., in STAT1$^{-/-}$ cells, IFNG stimulation induces high levels of *STAT3* gene expression (Ramana et al. 2005). Our data implicating *STAT3* as a direct transcriptional target of STAT1 suggest that STAT1 represses *STAT3* during IFNG signaling, further promoting its own apoptotic function.

Tumor necrosis factor (TNF) is cytokine that is involved in a plethora of cellular responses including cell differentiation, survival, and apoptosis. TNF binds to its receptor TNFRSF1A (Tumor Necrosis Factor Receptor Super Family 1A) and causes NFKB activation, which is crucial for the expression of many proinflammatory cytokines, chemokines, and multiple regulators of apoptosis and cell differentiation. In the absence of IFNG stimulation, cytoplasmic STAT1 binds to *TNFRSF1A* and maintains a tight control over TNF-mediated NFKB activation. However, IFNG stimulation was shown to increase sensitivity of cells to further TNF stimulation (Wesemann and Benveniste 2003). STAGE identified a STAT1 binding site in the first intron of *TNFRSF1A*, suggesting the possibility that IFNG dependent increased sensitivity to TNF could be a direct result of activation of *TNFRSF1A* by IFNG-stimulated STAT1. All the target sites and genes described above were verified by quantitative ChIP from an independent ChIP sample (Fig. 3B). We also identified other previously known STAT1 targets such as *IRF1*, *HLA-E*, *ICAM1*, as well as *STAT1* itself, whose expression is known to be induced by IFNG. The complete list of STAT1 targets identified by STAGE is provided in Supplemental Table 2.

Identification of MYC targets within the ENCODE regions by STAGE

We also used STAGE to identify the targets of MYC, an important oncogenic transcription factor. We carried out ChIP using an antibody against MYC in HeLa cells followed by the STAGE procedure. We sequenced ~4500 clones using standard sequencing methodology for generating the MYC STAGE library. Each clone contained on average ~20–30 STAGE tags. Out of a total of 127,351 tags extracted for MYC, 19,867 (15%) were orphans that could not be mapped to the human genome. We used the re-

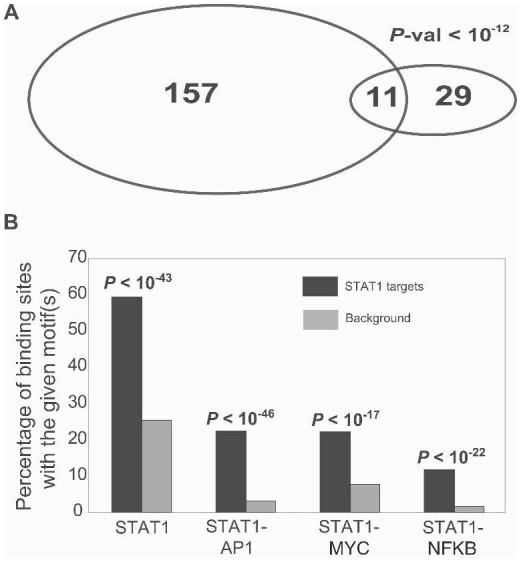


Figure 4. (*A*) Overlap of STAT1 target genes identified by STAGE with STAT1 target genes identified by ChIP-chip using a core promoter array. STAGE identified 29 promoters out of the ~9000 promoters present on the core promoter array as STAT1 target promoters. Eleven out of these 29 overlapped with the 157 promoters identified as STAT1 targets by ChIP-chip analysis at an enrichment ratio greater than threefold. The enrichment ratio refers to the ratio of the fluorescence intensity of ChIP DNA to that of reference DNA at each spot on the core promoter microarray. (*B*) Motif analysis. The Y-axis shows the percentage counts of the number of sites bearing the given motif(s) out of the 381 STAT1 binding sites detected by STAGE. Almost 60% of the 381 binding sites had the STAT1 motif TTCNNNGAA as compared to 27% in the background. We also detected an enrichment for the co-occurrence of the binding motifs for STAT1 and AP1 (TGAG/CTCA), STAT1 and MYC (CACA/GTG), and STAT1 and NFKB (GGGA/GNNC/TC/TCC) in accordance with the fact that STAT1 exhibits cooperative binding with these factors to regulate downstream promoters.

maining 107,484 tags for further analysis. Based on extrapolations from our ChIP-chip data (below) and previous observations (Cawley et al. 2004), MYC is likely to have between 17,000 and 25,000 binding sites on the genome. Because our depth of sequencing of STAGE tags for MYC was slightly lower than for STAT1, and the possibility that MYC may have a larger number of binding targets on the genome, the high specificity algorithm we developed for identifying STAT1 targets did not yield a significant number of binding targets for MYC. We therefore used a more relaxed algorithm as described in Methods to identify 2218 binding sites for MYC in the entire genome at a probability threshold of 0.8. We calculated the false discovery rate based on simulations at this threshold to be 5%. Twenty-six of the MYC binding sites identified by STAGE occurred within the ENCODE region. We also identified MYC binding sites within the ENCODE regions by ChIP-chip using NimbleGen oligonucleotide tiling arrays (The ENCODE Project Consortium 2007). The ChIP-chip analysis included three biological replicates, and we defined MYC binding peaks in the ENCODE regions using NimbleGen SignalMap software. Fourteen out of the 26 MYC binding sites identified by STAGE within the ENCODE regions were within 500 bp of a ChIP-chip peak in at least one of the three biological replicate experiments.

Discussion

Bead-based pyrosequencing technology has several advantages for STAGE over standard sequencing approaches (Margulies et al. 2005). First, there is no requirement for cloning and isolation of independent recombinant clones. Rather, tags generated by the STAGE procedure can be directly sequenced. Potential biases introduced by cloning in bacteria can thus be avoided. Second, the water-in-oil emulsion that is generated in making the library can be stored, and only a portion of this sample is used to generate on the order of 200,000 sequence reads in a single run of the instrument. Thus from a single chromatin immunoprecipitation reaction performed from a normally grown culture of mammalian cells, it is possible to sequence many samples and together generate more than one million sequence reads amounting to >100 Mb of sequence using STAGE, greatly improving the depth of sequencing and coverage of targets enriched in the ChIP sample. Third, bead-based pyrosequencing is more cost-effective. In our experience, the price of sequencing a STAGE tag using 454 Inc.'s service was about one-fifth that of standard clone-based sequencing (2.5 cents per tag for 454 vs. 14 cents per tag for clone-based sequencing). It is possible to modify the STAGE procedure such that each pyrosequencing read covers four tags, improving the cost-effectiveness and coverage by twofold.

We have developed analysis algorithms to detect genomic binding loci with high specificity. A recently developed algorithm, START, is also aimed at detecting transcription factor targets using ChIP-derived tag libraries (Marinescu et al. 2006). START uses a gene-centric approach where a user-defined window upstream of or downstream from a gene is searched to map tags and genes are denoted as targets using a z-score. START is thus limited to detecting binding sites near the 5′ end of a gene. Our approach defines enriched loci in the whole genome and then identifies genes that lie proximal to these binding sites, enabling identification of binding sites that may have long-range effects on the regulated gene. START does not assign statistical significance to clusters of tags that are not centered on a gene. Finally, START does not make any attempt to distinguish tags that are enriched from tags that might simply be noise, while we assign each tag a probability of enrichment to better distinguish noise from signal.

The currently implemented algorithm is an improvement over the previously employed algorithm (Kim et al. 2005b) to assign probabilities to STAGE-detected binding sites. Though the older algorithm assigned probabilities based on tag enrichment and rewarded clustering of tags, it did not make any attempt to differentiate if a given cluster is significant or not. The current algorithm assigns each cluster a probability of significance and employs individual tag enrichment as an additional criterion to compute a combined probability. This enables a more stringent assessment of whether a given window is a binding site or not. Overall, our results indicate that in depth sequencing using STAGE can identify biologically relevant direct binding targets of transcription factors throughout the genome.

Methods

ChIP for STAT1 and MYC

STAT1 ChIP was performed in HeLa S3 cells that were induced with 5 ng/mL human recombinant IFNG (R&D Systems) for 30 min and then fixed with 1% formaldehyde at room temperature for 10 min. Fixation was quenched with 125 mM glycine and cells were lysed in hypotonic lysis buffer (20 mM HEPES, pH 7.9, 10 mM KCl, 1 mM EDTA, pH 8, 10% glycerol, 1 mM DTT, 0.5 mM PMSF, 0.1 mM sodium orthovanadate, and protease inhibitors). Cell lysates were homogenized and nuclear pellets were collected and lysed in RIPA buffer (10 mM Tris-Cl, pH 8.0, 140 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% deoxycholic acid, 0.5 mM PMSF, 1 mM DTT, 0.1 mM sodium orthovanadate, and protease inhibitors). Nuclear lysates were sonicated with a Branson 250 Sonifier (output 20%, 100% duty cycle) to shear the chromatin to ~1 kb in size. Clarified lysates were incubated overnight at 4°C with anti-STAT1 alpha p91 (C-24) rabbit polyclonal antibody (sc-345 from Santa Cruz Biotechnology). Protein–DNA complexes were precipitated by protein A agarose and immunoprecipitates were washed three times in 1× RIPA, once in PBS, and then eluted. Crosslinks were reversed overnight at 65°C, and ChIP DNA was purified by Proteinase K treatment followed by extraction with phenol:chloroform:isoamyl alcohol extraction and precipitation with ethanol. Chromatin immunoprecipitation was performed for MYC in HeLa cells using anti-myc antibody (SC-764x from Santa Cruz Biotechnology) using the same protocol as described previously for E2F4 (Kim et al. 2005b).

STAT1 and MYC tag libraries

The STAGE procedure was modified for generating the STAT1 tag library. All steps leading to the generation of ditags from ChIP-enriched DNA were performed exactly as for MYC below. Gel-purified ditags were amplified by PCR using linker specific primers and sequenced by 454 Inc. Duplicate reads were removed by a Perl script. For MYC, the STAGE procedure was carried out as described previously (Kim et al. 2005b). Purified clones were sequenced by Agencourt Inc. Twenty-one-base-pair tags were extracted from each read using Perl scripts.

Generating hits for STAGE tags on the genome

We used the May 2004 Build 35 Human Genome assembly available at http://genome.ucsc.edu for all analyses. Twenty-one-base-pair tags were matched to the genome as described previously (Kim et al. 2005b). Briefly, an indexed, custom database of all

$CATG(N)_{17}$ sequences in the genome was first created. This represents a database of all possible STAGE tags using NlaIII, where each tag sequence was keyed to its chromosome and nucleotide coordinates. Each STAGE library tag was now mapped to the indexed genome-wide tag database by a simple binary search algorithm (Cormen et al. 2001).

Assigning probabilities for tag enrichment

We defined the number of distinct positions in the genome containing a perfect match to a given tag in the STAGE library as *nhit*. Thus, a tag with a *nhit* of 1 meant that this tag mapped to a single unique location in the human genome. We defined the number of occurrences of the tag in the sequenced STAGE library, that is, the number of times a given tag was observed in the STAGE library, as *nocc*. The selection of *N* tags at random from the entire genome could be modeled as a binomial distribution where the success of an event is defined as selecting a tag with a given *nhit*. The background probability of selection of a tag with a given *nhit* was calculated as p = *nhit*/total number of tags in the genome. If an observed tag with a given *nhit* has a *nocc* = f, we calculated the probability of selecting a tag with the observed *nhit* and *nocc* $\geq f$ under a random model. This probability was calculated as 1 minus the cumulative binomial probability of selecting that particular tag with a frequency $\leq f - 1$, which was calculated as

$$(1 - \sum_{0}^{f-1} \binom{N}{x} p^{x}(1-p)^{N-x})$$

where p is the background probability of selection of the tag and x iterates from 0 to $f - 1$.

Multiplying this probability by the total number of tags found in the genome having the given *nhit* yields the expected frequency of selecting tags with the given *nhit* and *nocc* $\geq f$

Thus, the expected frequency of a tag with a given *nhit* and *nocc* = f when N tags are selected at random was calculated as

$$\text{Expected frequency} = (1 - \sum_{0}^{f-1} \binom{N}{x} p^{x}(1-p)^{N-x})\, M$$

where p = *nhit*/T, and T is the total number of 21 bp $CATG(N_{17})$ tags found in the entire genome (27,429,149). M = number of tags with a given *nhit*.

$$\text{Probability that a given tag is enriched} = \left(1 - \frac{\text{expected frequency}}{\text{observed frequency}}\right).$$

If the expected frequency was greater than the observed frequency, the tag was assigned a low enrichment probability of 0.001.

STAGE target calls for STAT1

A window size of 500 bp was used as described above. For each window, we defined k = number of tags assigned to the window with a single hit on the genome.

Probability that the window is a target = *wt_nhit** P_{hit}+ *wt_nocc* *P_{nocc}, where

$$P_{hit} = 1 - \frac{\text{expected frequency of windows with given k}}{\text{observed frequency of windows with given k}}.$$

The expected frequency of a window with a given k was obtained from random simulations. It is also possible to calculate this expected frequency and avoid time-consuming random simulations.

P_{nocc} was calculated as the probability that at least one tag assigned to the window was not random:

$$P_{nocc} = 1 - \prod_{i}\left(1 - \frac{p(\text{tag}_i)}{nhit_i}\right)$$

where $p(\text{tag}_i)$ is the probability that tag_i was enriched. *wt_nhit* and *wt_nocc* were empirically derived weights and were set to 0.9 and 0.1, respectively.

STAGE target calls for MYC

A window of size 500 bp was scanned across each chromosome, and tags mapping within the window were assigned to the window. For the MYC analysis, we discarded tags that had more than 10 hits on the genome.

$$\text{Probability that the window is a target} = 1 - \prod_{i}(1 - p(\text{tag}_i)).$$

Quantitative ChIP PCR for binding sites identified by STAGE

We performed quantitative PCR on an independent IFN-γ-stimulated STAT1 ChIP DNA sample. We selected 10 sites to test, spanning a range of final STAGE probability scores. For each of the 10 selected binding sites, we extended the site by 100 bp on either side. Primers were designed to amplify 60–100 bp fragments within the extended window. Quantitative PCR reactions were performed in triplicate in a 96-well optical reaction plate (ABI PRISM) using SYBR Green PCR Master Mix (Applied Biosystems) on an ABI 7900 instrument. The –ΔΔCt values for each locus were calculated with respect to the ChIP input DNA, normalized to a reference locus (*GAPDH* gene promoter) as described (Livak and Schmittgen 2001). Data for the nine sites that could be confirmed are shown in Figure 3B. Primer sequences are provided in Supplemental Table 3.

Acknowledgments

We thank R. Green, M. Singer, and N. Jiang at NimbleGen Inc. for facilitating the MYC ChIP-chip analysis; K. Rosenbloom and UCSC browser staff for data upload; and 454 Inc. and Agencourt Inc. for sequencing. This work was supported by an NIH/NHGRI ENCODE Technology Development grant (HG003532) and a grant from the US Army Breast Cancer Research Program to V.R.I.

References

Adhikary, S. and Eilers, M. 2005. Transcriptional regulation and transformation by Myc proteins. *Nat. Rev. Mol. Cell Biol.* **6:** 635–645.

Burch, L.R., Scott, M., Pohler, E., Meek, D., and Hupp, T. 2004. Phage-peptide display identifies the interferon-responsive, death-activated protein kinase family as a novel modifier of MDM2 and p21WAF1. *J. Mol. Biol.* **337:** 115–128.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499–509.

Chen, J. and Sadowski, I. 2005. Identification of the mismatch repair genes *PMS2* and *MLH1* as p53 target genes by using serial analysis of binding elements. *Proc. Natl. Acad. Sci.* **102:** 4813–4818.

Cormen, T., Leiserson, C., and Rivest, R. 2001. *Introduction to algorithms*. MIT Press, Cambridge, MA.

Eferl, R. and Wagner, E.F. 2003. AP-1: A double-edged sword in tumorigenesis. *Nat. Rev. Cancer* **3:** 859–868.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* (in press).

Hartman, S.E., Bertone, P., Nath, A.K., Royce, T.E., Gerstein, M.,

Weissman, S., and Snyder, M. 2005. Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes & Dev.* **19:** 2953–2968.
Impey, S., McCorkle, S.R., Cha-Molstad, H., Dwyer, J.M., Yochum, G.S., Boss, J.M., McWeeney, S., Dunn, J.J., Mandel, G., and Goodman, R.H. 2004. Defining the CREB regulon: A genome-wide analysis of transcription factor regulatory regions. *Cell* **119:** 1041–1054.
Kawai, T., Akira, S., and Reed, J.C. 2003. ZIP kinase triggers apoptosis from nuclear PML oncogenic domains. *Mol. Cell. Biol.* **23:** 6174–6186.
Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005a. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876–880.
Kim, J., Bhinge, A.A., Morgan, X.C., and Iyer, V.R. 2005b. Mapping DNA–protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* **2:** 47–53.
Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. 2006. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* **125:** 301–313.
Liu, Z., Lu, H., Jiang, Z., Pastuszyn, A., and Hu, C.A. 2005. Apolipoprotein l6, a novel proapoptotic Bcl-2 homology 3-only protein, induces mitochondria-mediated apoptosis in cancer cells. *Mol. Cancer Res.* **3:** 21–31.
Livak, K.J. and Schmittgen, T.D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method. *Methods* **25:** 402–408.
Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38:** 431–440.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.
Marinescu, V.D., Kohane, I.S., Kim, T.K., Harmin, D.A., Greenberg, M.E., and Riva, A. 2006. START: An automated tool for serial analysis of chromatin occupancy data. *Bioinformatics* **22:** 999–1001.
Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T.E., Luscombe, N.M., Rinn, J.L., Nelson, F.K., Miller, P., Gerstein, M., et al. 2003. Distribution of NFKB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci.* **100:** 12247–12252.
Platanias, L.C. 2005. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* **5:** 375–386.
Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33:** D501–D504.
Ramana, C.V., Chatterjee-Kishore, M., Nguyen, H., and Stark, G.R. 2000. Complex roles of STAT1 in regulating gene expression. *Oncogene* **19:** 2619–2627.
Ramana, C.V., Kumar, A., and Enelow, R. 2005. STAT1-independent induction of SOCS-3 by interferon-gamma is mediated by sustained activation of Stat3 in mouse embryonic fibroblasts. *Biochem. Biophys. Res. Commun.* **327:** 727–733.
Roh, T.Y., Ngau, W.C., Cui, K., Landsman, D., and Zhao, K. 2004. High-resolution genome-wide mapping of histone modifications. *Nat. Biotechnol.* **22:** 1013–1016.
Roh, T.Y., Cuddapah, S., and Zhao, K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes & Dev.* **19:** 542–552.
Stephanou, A. and Latchman, D.S. 2005. Opposing actions of STAT-1 and STAT-3. *Growth Factors* **23:** 177–182.
Stephanou, A., Brar, B.K., Knight, R.A., and Latchman, D.S. 2000. Opposing actions of STAT-1 and STAT-3 on the Bcl-2 and Bcl-x promoters. *Cell Death Differ.* **7:** 329–330.
Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207–219.
Wesemann, D.R. and Benveniste, E.N. 2003. STAT-1 alpha and IFN-gamma as modulators of TNF-alpha signaling in macrophages: Regulation and functional implications of the TNF receptor 1:STAT-1 alpha complex. *J. Immunol.* **171:** 5313–5319.

Received May 31, 2006; accepted in revised form September 18, 2006.

Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation

Sushma Shivaswamy[1☯], Akshay Bhinge[1☯], Yongjun Zhao[2], Steven Jones[2], Martin Hirst[2], Vishwanath R. Iyer[1*]

1 Institute for Cellular and Molecular Biology, Center for Systems and Synthetic Biology, and Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas, United States of America, **2** Michael Smith Genome Sciences Center, British Columbia Cancer Agency, Vancouver, British Columbia, Canada

The eukaryotic genome is packaged as chromatin with nucleosomes comprising its basic structural unit, but the detailed structure of chromatin and its dynamic remodeling in terms of individual nucleosome positions has not been completely defined experimentally for any genome. We used ultra-high–throughput sequencing to map the remodeling of individual nucleosomes throughout the yeast genome before and after a physiological perturbation that causes genome-wide transcriptional changes. Nearly 80% of the genome is covered by positioned nucleosomes occurring in a limited number of stereotypical patterns in relation to transcribed regions and transcription factor binding sites. Chromatin remodeling in response to physiological perturbation was typically associated with the eviction, appearance, or repositioning of one or two nucleosomes in the promoter, rather than broader region-wide changes. Dynamic nucleosome remodeling tends to increase the accessibility of binding sites for transcription factors that mediate transcriptional changes. However, specific nucleosomal rearrangements were also evident at promoters even when there was no apparent transcriptional change, indicating that there is no simple, globally applicable relationship between chromatin remodeling and transcriptional activity. Our study provides a detailed, high-resolution, dynamic map of single-nucleosome remodeling across the yeast genome and its relation to global transcriptional changes.

Citation: Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, et al. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol 6(3): e65. doi:10.1371/journal.pbio.0060065

Introduction

The eukaryotic genome is compacted into nucleosomal arrays composed of 146-bp DNA wrapped around a core histone octamer complex [1]. The location of nucleosomes affects nearly every cellular process requiring access to genomic DNA, but it is not well understood how nucleosomes are positioned and remodeled throughout any genome. Mapping nucleosome positions using DNA microarrays covering 4% of the yeast genome has shown that a majority of assayable nucleosomes were well positioned [2]. Computational analyses incorporating structural mechanics of nucleosome associated DNA [3–5] and comparative genetics [6] have predicted nucleosome positions in the yeast genome. However, experimental validation and comparison with available in vivo data show that intrinsic signals in genomic DNA determine only 15%–17% of nucleosome positioning above what is expected by chance [3,4]. In vivo nucleosome positions are influenced by the presence of numerous ATP-dependent remodelers, and the transcriptional machinery [7,8].

Recently, chromatin immunoprecipitation (ChIP)-sequencing technology was used to map the positions of nucleosomes containing the variant H2A.Z histone across the yeast genome [9]. H2A.Z nucleosomes are enriched at promoters; therefore, this study mapped about 10,000 nucleosomes. Tiling arrays have been recently used to catalog the positions of nucleosomes at 4–5-bp resolution across the yeast genome and their repositioning by chromatin remodelers [10,11]. However, dynamic changes in individual nucleosome positions in response to physiological perturbations that cause global transcriptional reprogramming have not yet been examined on a genomic scale in any organism.

To map the location of individual nucleosomes on a genomic scale and at high resolution, we used ultra-high–throughput sequencing methodology (Solexa/Illumina) to sequence the ends of nucleosome-associated DNA. Our approach enabled us to map individual nucleosomes nominally at single-nucleotide resolution. Nucleosome density and stability at promoters and over coding regions were correlated specifically with transcription rate rather than absolute transcript levels. Two different modes of chromatin remodeling were associated with transcriptional regulation. Gene activation was mainly accompanied by the eviction of one to two nucleosomes from the promoter, and gene repression was mainly accompanied by the appearance of nucleosomes with varying stability over the promoter. Our work con-

Academic Editor: Oliver J. Rando, University of Massachusetts Medical School, United States of America

Received September 17, 2007; **Accepted** January 30, 2008; **Published** March 18, 2008

Abbreviations: ChIP, chromatin immunoprecipitation; ChIP-chip, chromatin immunoprecipitation–microarray; NPP, nucleosome positioning periodicity; qPCR, quantitative real-time PCR; TSS, transcription start site

* To whom correspondence should be addressed. E-mail: vishy@mail.utexas.edu

☯ These authors contributed equally to this work.

Author Summary

The eukaryotic genome is packed in a systematic hierarchy to accommodate it within the confines of the cell's nucleus. This packing, however, presents an impediment to the transcription machinery when it must access genomic DNA to regulate gene expression. A fundamental aspect of genome packing is the spooling of DNA around nucleosomes—structures formed from histone proteins—which must be dislodged during transcription. In this study, we identified all the nucleosome displacements associated with a physiological perturbation causing genome-wide transcriptional changes in the eukaryote *Saccharomyces cerevisiae*. We isolated nucleosomal DNA before and after subjecting cells to heat shock, then identified the ends of these DNA fragments and, thereby, the location of nucleosomes along the genome, using ultra-high–throughput sequencing. We identified localized patterns of nucleosome displacement at gene promoters in response to heat shock, and found that nucleosome eviction was generally associated with activation and their appearance with gene repression. Nucleosome remodeling generally improved the accessibility of DNA to transcriptional regulators mediating the response to stresses like heat shock. However, not all nucleosomal remodeling was associated with transcriptional changes, indicating that the relationship between nucleosome repositioning and transcriptional activity is not merely a reflection of competing access to DNA.

stitutes the first study of dynamic single-nucleosome remodeling in response to transcriptional perturbation across an entire eukaryotic genome.

Results

Strategy for Identifying Nucleosome Positions Using Ultra-High–Throughput Sequencing

We used micrococcal nuclease to isolate mononucleosome-associated DNA from yeast cells before and after a physiological perturbation (heat shock for 15 min) that causes genome-wide transcriptional changes, and sequenced the ends of the fragments. Only uniquely aligning reads were used to define the ends of nucleosomal DNA. After aligning sequence reads to the genome, we defined nucleosome peaks by first using a Parzen window probability estimation of read densities, then defining a peak of width 146 bp around the centers of appropriately spaced maxima in the density function (Materials and Methods). Our approach yielded nucleosome positions at single-nucleotide resolution. We calculated a score for the position and stability of each nucleosome, which were normalized to account for differences in sequencing depth. Scores in the range of 0.2 to 0.25 and higher indicated nucleosomes whose positions often matched in the two independent biological samples and, hence, indicated bona fide nucleosomes; nucleosomes below this threshold were defined by too few reads to be discernable above background. At a score cutoff of 0.25, we defined the locations of 49,043 nucleosomes in normally growing cells and 52,817 nucleosomes in heat-shocked cells. Assuming that two adjacent nucleosomes cannot be closer than 200 bp, altogether about 73% of the yeast genome is covered by a positioned nucleosome. Since only uniquely aligning reads were used in our analysis, and the yeast genome contains an appreciable fraction of repeated sequence elements, we estimate that about 78% of the genome is covered by positioned nucleosomes.

Recapitulation of Known Nucleosome Positions and Expected Remodeling Events

We assessed the quality and accuracy of our nucleosome sequencing data by examining the nucleosomes known to be positioned at the *PHO5* promoter. The yeast *PHO5* promoter is repressed during growth in rich media by specifically positioned nucleosomes flanking a short, hypersensitive region containing a binding site for the transcription factor Pho4 [12]. These nucleosomes were evident in the alignment of our raw sequence reads, and their precise positions calculated by our analysis algorithm corresponded to the known positions of these nucleosomes. The positions of these three nucleosomes did not vary in the two independent biological samples before and after heat shock, as this perturbation does not affect the *PHO5* promoter (Figure 1A). Quantitative real-time PCR (qPCR) for the three nucleosome peaks and three troughs (linker regions) identified by sequencing provided independent experimental verification of these nucleosome positions and the fact that their positions did not change in the two samples (Figure 1B). At individual promoters where transcription is activated by heat shock, the raw data traces and our inferred nucleosome peaks showed that nucleosomes were displaced at the promoter after the perturbation (Figure 1C). Conversely, at promoters that are repressed, positioned nucleosomes appeared after the perturbation (Figure 1D). The genome-wide nucleosome positions we identified experimentally correspond well with individual nucleosomes mapped on chromosome III as well as nucleosome-bound sequences isolated in previous studies [2,4] (see Figure S1 and Table S1). While this manuscript was in preparation, a catalog of nucleosome positions in yeast was published [13]. Our mapped nucleosome positions also agree well with this recent study (Figure S1 and Table S1). Thus, our mononucleosome preparations and the high-throughput sequencing assay recapitulated bona fide in vivo nucleosome positions and rearrangements.

Lower Nucleosome Occupancy at Promoters Compared to Coding Regions

Low-resolution analysis using PCR microarrays has shown that promoters are nucleosome-poor relative to coding regions [14,15]. In accord with these findings, we found that both the number and the stability of nucleosomes were significantly lower at promoters than over coding regions ($p < 2.2 \times 10^{-16}$). We plotted the average nucleosome profile over all yeast genes to get an idea of how individual nucleosomes were distributed in relation to promoters and coding regions. Several features of chromatin organization were evident from this plot (Figure 2A). First, as noted before, promoters showed a lower probability of nucleosomes as compared to coding sequences. Second, the apparent nucleosome-free region immediately upstream of the transcription start site (TSS) is only approximately the width of a single nucleosome. Third, there is a strongly positioned nucleosome, likely an H2A.Z-containing nucleosome, that marks the start of the transcribed region immediately downstream of the TSS [9]. Fourth, positioned nucleosomes continue at periodic intervals downstream of the TSS, with decreasing probabilities. These characteristics of nucleosome positioning corroborate results based on mapping nucleosomes across a single yeast chromosome [2].

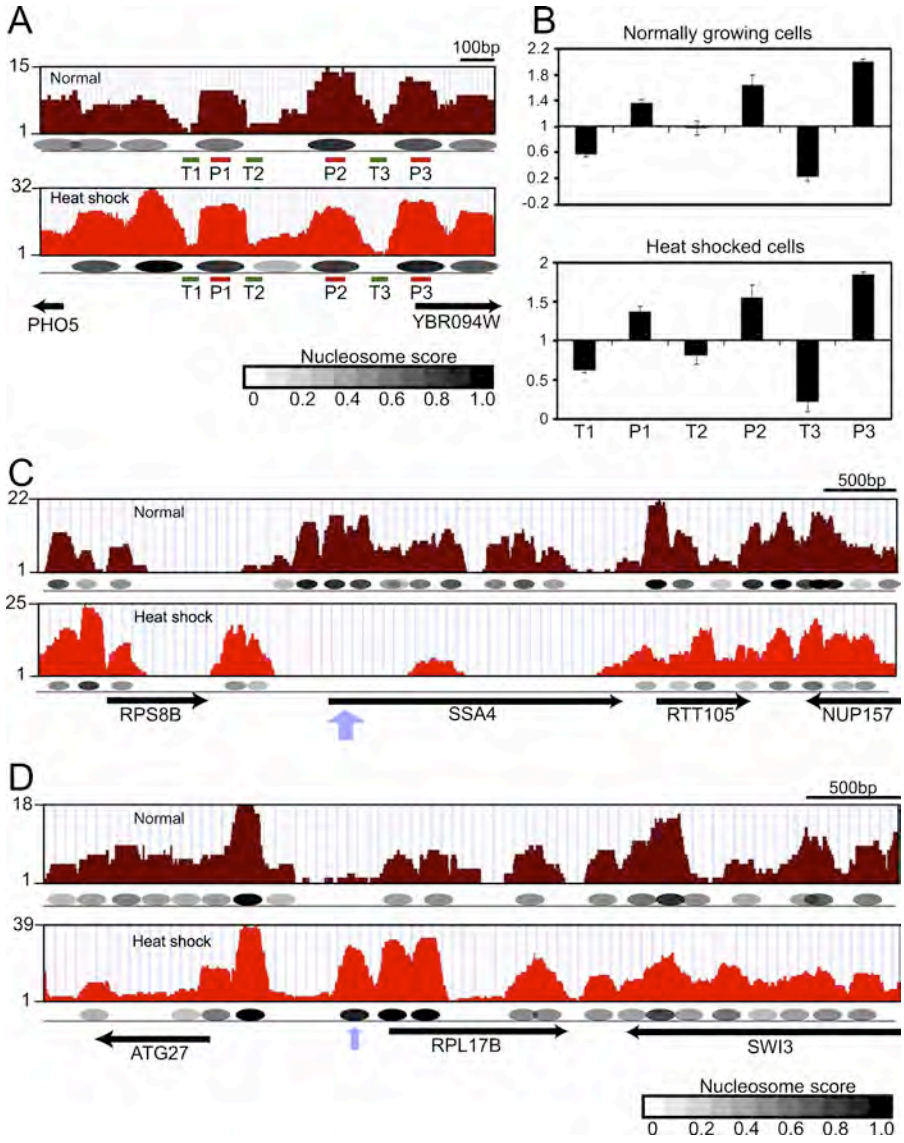


Figure 1. Ultra-High–Throughput Sequencing Recapitulates In Vivo Nucleosome Positions

(A) Detailed view of the *PHO5* locus showing the raw sequence reads (brown and red profiles). The nucleosome positions calculated using our analysis algorithm are shown as ovals, shaded according to their nucleosome score as indicated. The positions of the amplicons used for qPCR analysis are marked as red (peaks) and green (troughs) lines below. The black arrows indicate the positions of genes in that region.

(B) qPCR verification of the three nucleosome peaks and three troughs identified by sequencing confirm that their positions remain the same before and after heat shock.

(C) The heat-shock–induced SSA4 gene and flanking regions, showing that nucleosomes are displaced specifically at the SSA4 promoter and coding region after heat shock (thick purple arrow).

(D) The heat-shock–repressed ribosomal protein gene *RPL17B* and flanking regions, showing that a single positioned nucleosome appears after heat shock specifically at the RPL17B promoter (thin purple arrow). The nucleosome positions calculated using our analysis algorithm are indicated as in (A).

doi:10.1371/journal.pbio.0060065.g001

We obtained nearly identical results in the independent heat-shocked cells (Figure S2). Interestingly, we also observed a strongly positioned nucleosome at the 3′ end of the coding region followed by a relatively nucleosome-free region, which has not been noted before. This 3′ nucleosomal mark does not reflect the boundary of a downstream promoter, because it was evident even at the 3′ end of convergently transcribed genes lacking another promoter immediately downstream of their 3′ end (Figures 2B and S2). This 3′ end chromatin feature was not biased towards convergently transcribed genes, but we noted a modest association with genes that were expressed at low levels and with long genes (unpublished data). Our data also established that although the internucleosomal linker length could vary widely, the linker length is commonly about 30 bp in the yeast genome (Figure S3).

Nucleosome Positioning Is Influenced by the Presence of a TATA Box and Is Correlated with Transcription Rate

Although our whole-genome data revealed stereotypical distribution patterns of nucleosomes around promoters, we reasoned that the average profile might conceal several distinct nucleosome occupancy profiles with distinct relation-

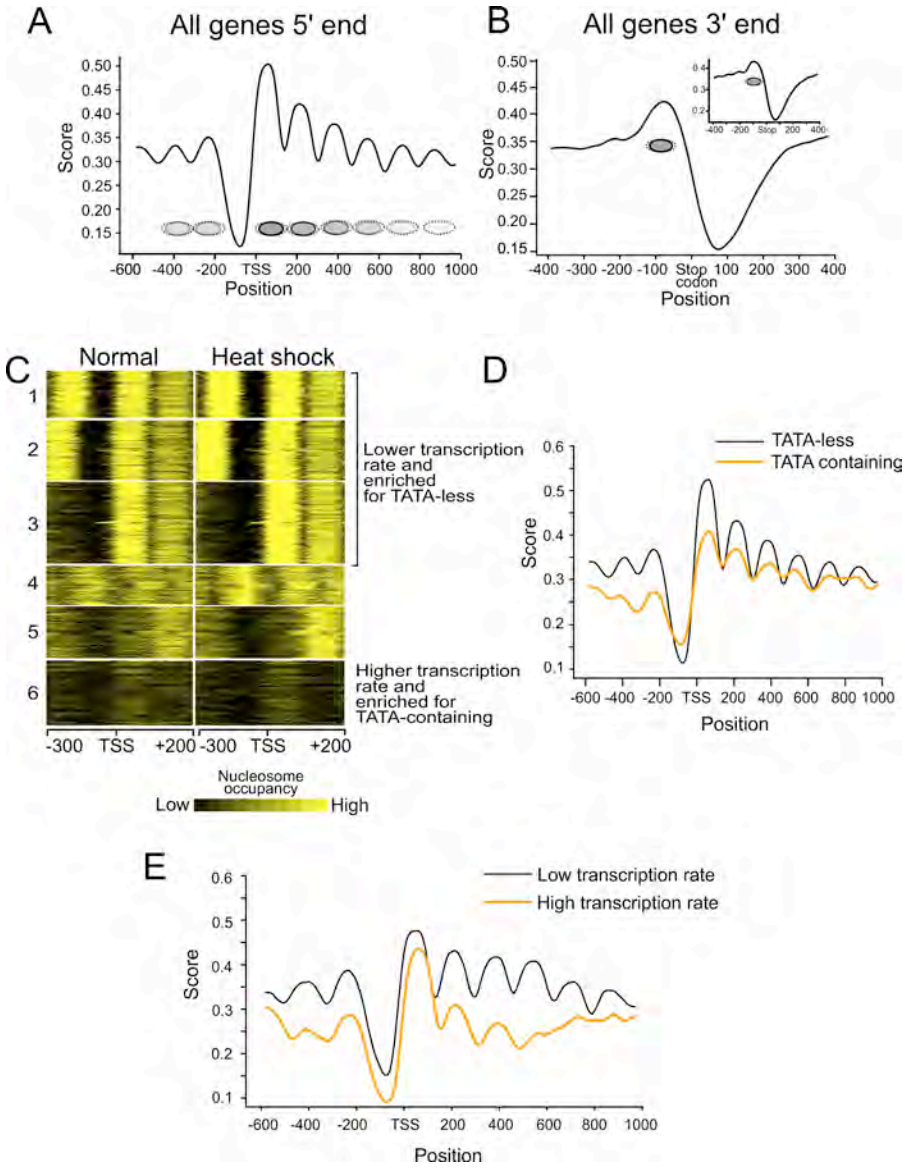


Figure 2. Patterns of Chromatin Organization in the Yeast Genome

(A) Average nucleosome profiles of all genes in the yeast genome from −600 bp to +1,000 bp with respect to the transcription start site (TSS). Nucleosome positions are shown as gray ovals below the profile. The intensity of the filled oval reflects the average probability score of the nucleosomes (see Figure 1 for the color scale), and the dotted oval marks the spread of that nucleosome across all genes.

(B) The 3′ end of genes is marked by a strongly positioned nucleosome, followed by a relatively nucleosome-free region. The inset shows the 3′ end of convergently transcribed genes in which the 3′ end is not followed by another promoter.

(C) Distinct classes of nucleosome profiles revealed by *k*-means clustering of all promoters in the yeast genome. Each row in the clusters shows the position of a nucleosome at an individual promoter. Nucleosomes are colored according to their probability using the shown color scale. Clusters 1–3 showed a significant enrichment for genes with lower transcription rates and for TATA-less genes ($p \leq 10^{-10}$). Cluster 6 showed a significant enrichment for genes with high transcription rates and for TATA-containing genes ($p < 10^{-10}$).

(D) Average nucleosome profiles for TATA-containing (973) and TATA-less (4,382) promoters, aligned with respect to the TSS.

(E) The genes in the yeast genome were sorted in descending order according to their transcription rates [16], and the average promoter nucleosome profiles for the top 500 genes (orange) and the bottom 500 genes (black) are plotted.

doi:10.1371/journal.pbio.0060065.g002

ships to transcriptional activity or promoter sequence characteristics. To reveal such distinctions, we performed *k*-means clustering of the nucleosome peak profiles around all yeast promoters. Indeed, several classes of nucleosome profiles were now evident (Figure 2C). There was no significant distinction between these different promoter classes with respect to either their occupancy by the general transcription factor TBP or their absolute transcript levels (Figure S4). However, there were biases among the clusters with respect to their representation of TATA box–containing and TATA-less promoters, as well as their transcription rates. In general, promoter classes containing a strongly positioned nucleosome were enriched for TATA-less promoters and had lower transcription rates, and conversely, the cluster containing poorly positioned nucleosomes was enriched for TATA-containing promoters and had higher transcription rates [16] (Figure 2C). We ascertained that promoters that appeared to be largely devoid of positioned nucleosomes

were not artificially caused by our exclusion of ambiguous sequence reads. The average nucleosome occupancy profiles for TATA-less and TATA-containing promoters, considered separately, showed that the absence of a consensus TATA element in the promoter was indeed correlated with the stereotypical genome-wide nucleosome profile (Figures 2D and S2). This distinction was not due to the lower number of TATA-containing promoters (unpublished data). Correspondingly, genes with low transcription rates showed stronger nucleosome positioning as compared to genes with higher transcription rates (Figure 2E).

Visual inspection of nucleosome profiles before and after heat shock indicated that the positions of the majority of nucleosomes were closely maintained despite the genome-wide transcriptional perturbation (Figure S1). In general, individual nucleosome positions in each of the promoter classes were largely unchanged in cells after heat shock (Figure 2C). Approximately 65% of all positioned nucleosomes throughout the genome in normally growing cells were within 30 bp of their positions in heat-shocked cells. At a score cutoff of 0.25, less than 10% of the nucleosomes were displaced more than 100 bp after heat shock (Table S1). In addition to the promoter nucleosome classes, we also observed strong, periodically positioned nucleosomes located over the transcribed regions of most genes in the genome. This periodicity was evident when we aligned all coding regions to the first nucleosome downstream of the TSS and ranked all these genes by a nucleosome positioning periodicity (NPP) score applied to the coding region (Figure 3A; Materials and Methods). There was no correlation between NPP and steady-state transcript levels (unpublished data). However, genes with a high NPP score, which had strongly positioned nucleosomes over the coding region, were transcribed at significantly lower rates than genes with a low NPP score (Figure 3B). Correspondingly, genes that were transcribed at low rates showed well-positioned periodic nucleosomes over the coding region relative to genes transcribed at higher rates, which showed weaker nucleosome positioning over the coding region (Figure 3C). Overall, the stereotypical positioning of nucleosomes over coding regions and promoters is consistent with the notion that nucleosome positions in the yeast genome are not random, but rather, are strongly encoded intrinsically through a combination of DNA sequence composition and binding of other proteins.

Sequence-Dependent Positioning of Nucleosomes

Analysis of DNA sequences associated with nucleosomes has indicated that nucleosome positions are intrinsically encoded in DNA [4,6,13]. However, it is not clear to what extent DNA sequence governs nucleosome positions compared to other factors that might also contribute to nucleosome positioning across the genome. One possibility is that when a nucleosome is strongly positioned at one site by virtue of DNA sequence, immediately adjacent nucleosomes are "stacked" against it and therefore show little sequence dependence. In particular, the regular array of nucleosomes we observed over coding regions could reflect sequence-dependent positioning of an H2A.Z nucleosome at the 5′ end of the array corresponding to the TSS, but with the remainder being positioned relative to the first one in a sequence-independent manner. To test this idea, we examined the sequence dependence of successive nucleosome positions in the strongly positioned nucleosomal arrays over the coding region. We first generated a profile of the AA/TT dinucleotide frequency for the sequences associated with the strongest positioned nucleosomes at the first position shown in Figure 3A. Like the profile generated from computational predictions of nucleosome positions [4,6], our profile shows a repeating pattern with an approximate periodicity of ten nucleotides, indicative of the rotational positioning of the nucleosome over a preferred sequence (Figure 3D). Although the information content of our measured dinucleotide profile is modest, it is significantly different from the same dinucleotide profile measured over randomly selected DNA sequences from the genome (Figure 3D). We then measured the average correlation between our nucleosome sequence profile and the same dinucleotide profile for the set of sequences associated with all nucleosomes in each of the positions in the regular array of coding region nucleosomes. As expected, the first position showed the strongest correlation to the positioning sequence, but in general, successive nucleosome positions in the arrays showed lower, but significant, correlations to the positioning DNA profile (Figure 3E). Thus, although the underlying DNA sequence as measured by the dinucleotide profile makes only a modest contribution to the positioning of nucleosomes, in general this contribution is maintained to a large extent even when nucleosomes are adjacent to another well-positioned nucleosome in the coding region.

Nucleosome Remodeling Is Mechanistically Linked to Dynamic Changes in Transcription

In order to examine how dynamic remodeling of individual nucleosomes was globally related to dynamic changes in transcription after the physiological perturbation, we generated nucleosome remodeling profiles for all promoters (Materials and Methods). A positive value in the remodeling profile at a given promoter position indicated that there was a nucleosome covering the position during normal growth, but was depleted or evicted upon heat shock. A negative value indicated the opposite, namely, the appearance of a more strongly positioned nucleosome following heat shock. We grouped the remodeling profiles by k-means clustering and visualized specific patterns of nucleosomal changes at the promoter.

We first analyzed remodeling profiles for promoters that were activated at least 2-fold and promoters that were repressed at least 2-fold by heat shock (Figure 4A and 4B). Two well-defined groups (Group 2 and 4) of activated genes contained promoters in which a single nucleosome that covered the promoter during normal growth was evicted upon heat shock, making the promoter more accessible for binding by transcription factors or the general transcription machinery (Figure 4A). Of these, Group 2 showed a significant enrichment for targets of the activator Msn4 ($p = 0.02$) [17] Promoters in Group 1 had a nucleosome-free region between the TSS and −200 bp both before and after heat shock. This group showed a significant enrichment for targets of the transcriptional activator Hsf1 ($p < 0.02$). Group 3 showed enrichment for the remodeler Swi5 ($p = 0.002$).

The difference in nucleosome profiles between Group 1 and Group 2 genes and the differential enrichment of the two major stress transcription factor targets points to two distinct modes of action by these activators. Hsf1 is constitutively

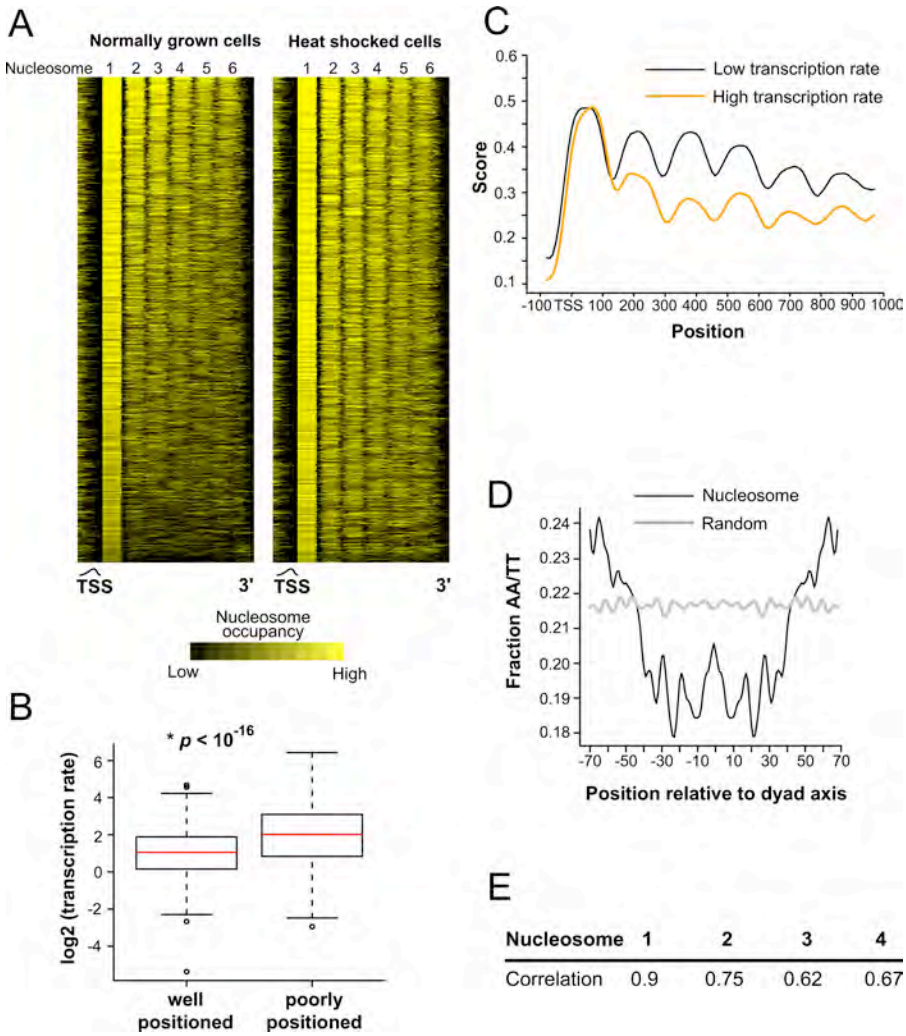


Figure 3. Nucleosome Positioning over Coding Regions Depends on Transcription Rate and Sequence Characteristics

(A) Genes were aligned to the first nucleosome downstream of the TSS and sorted by their nucleosome positioning periodicity (NPP) score (see Materials and Methods). Genes were sorted by their NPP scores in normally growing cells, and the data from heat-shocked cells are shown in the same order. The unaligned TSS is indicated by the approximate curve.

(B) The transcription rate of genes with high NPP scores (well-positioned nucleosomes) is significantly lower than that of genes with low NPP scores (poorly positioned). In these box plots, the red line indicates the median, the upper and lower bounds of the box indicate the interquartile range, the horizontal lines that are connected to the box by a dashed line indicate the upper and lower bounds of nonoutlier values, and the open circles indicate outliers.

(C) Genes were sorted in descending order according to their transcription rates, and the average nucleosome profiles over the coding regions for top 500 genes (orange) and the bottom 500 genes (black) are plotted.

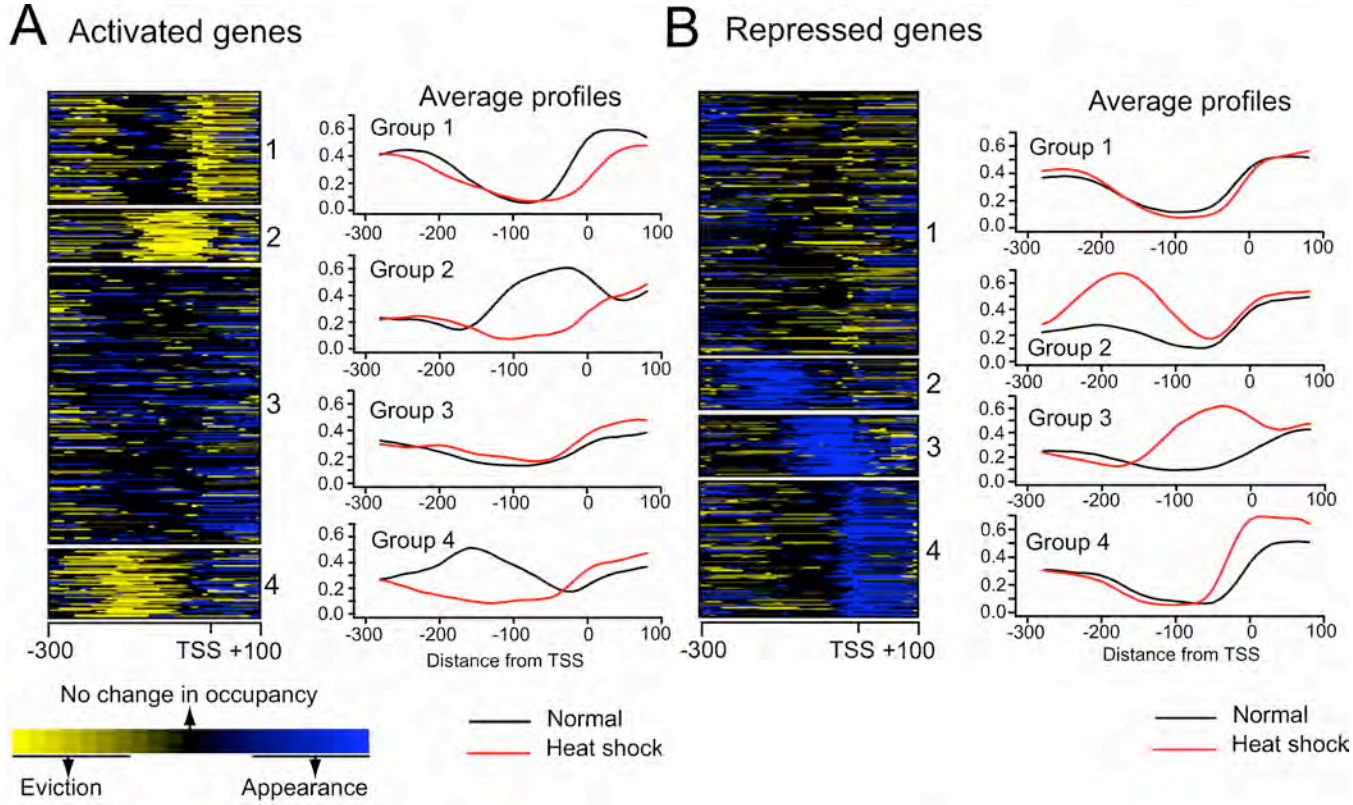
(D) Frequency of AA/TT dinucleotide at each position in the DNA sequence associated with the most strongly positioned first nucleosomes. The frequency profiles for the dinucleotides AA and TT for the first nucleosome shown in (A) were summed and smoothed using a 3-bp moving average. The same analysis was also performed for a comparable set of randomly chosen DNA sequences from the yeast genome.

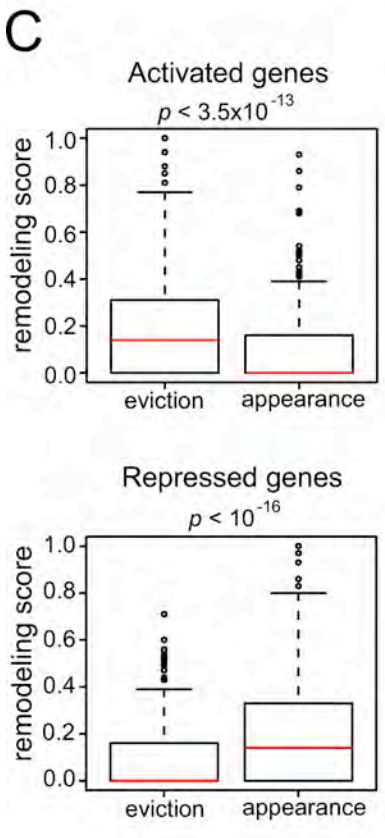
(E) Correlation coefficients of the AA/TT profiles for the DNA sequences underlying each of the indicated coding nucleosome positions from (A), with the positioning profile derived earlier. Each of the correlation values was significantly higher than background.

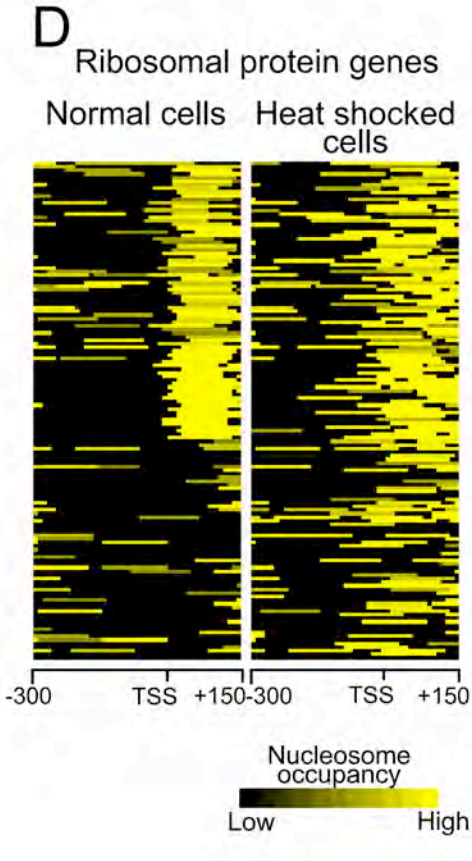
doi:10.1371/journal.pbio.0060065.g003

bound to many heat shock gene promoters [18]. The nucleosome profiles of Group 1 promoters, which showed enrichment for Hsf1 targets, suggest that Hsf1 binding induces eviction of the nucleosome covering the promoter or precludes its occupancy over this region. On the other hand, Msn4 target promoters (enriched in Group 2) had a nucleosome covering the promoter during normal growth. Our data suggest that translocation of Msn4 into the nucleus upon heat shock [19] and its occupancy of the promoter results in eviction of the nucleosome, and thus facilitates activated transcription.

Genes repressed more than 2-fold after heat shock could also be clustered into four major groups based on their nucleosome remodeling profiles (Figure 4B). Group 2 repressed genes had a nucleosome-free region between −200 and −100 bp upstream of the TSS during normal growth, which was covered by the appearance of a single nucleosome after heat shock. Group 3 repressed genes were characterized by the appearance of a single nucleosome between −125 and +50 bp relative to the TSS after heat shock. Group 1 and Group 4 repressed genes had subtle differences between themselves and between normally growing and heat-shocked cells. They both had a nucleosome-free region between −200


A Activated genes
Average profiles
Group 1
Group 2
Group 3
Group 4
1
2
3
4
-300
TSS +100
Distance from TSS
No change in occupancy
Eviction
Appearance
Normal
Heat shock
B Repressed genes
Average profiles
Group 1
Group 2
Group 3
Group 4
Distance from TSS
Normal
Heat shock


C
Activated genes
p < 3.5x10^-13
remodeling score
eviction
appearance
Repressed genes
p < 10^-16
remodeling score
eviction
appearance


D
Ribosomal protein genes
Normal cells
Heat shocked cells
-300
TSS
+150
Nucleosome occupancy
Low
High

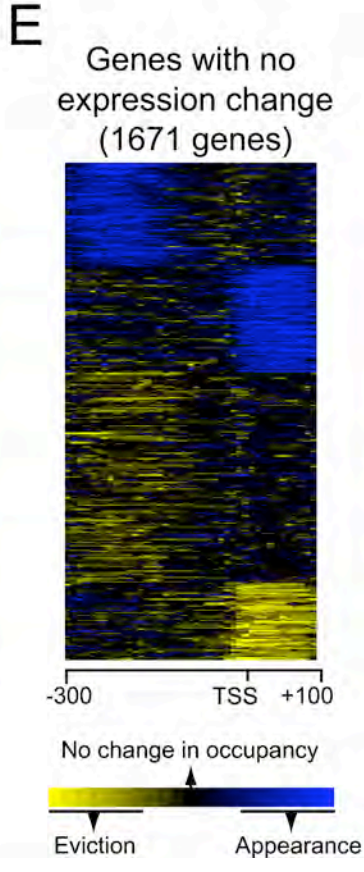

E
Genes with no expression change (1671 genes)
-300
TSS
+100
No change in occupancy
Eviction
Appearance

Figure 4. Classification of Promoter Nucleosome Remodeling Profiles

All profiles are aligned with respect to the TSS.

(A) Remodeling profiles of genes activated greater than 2-fold after heat shock and (B), genes repressed greater than 2-fold by heat shock. Nucleosomes present during normal growth but evicted by heat shock are indicated in yellow, and nucleosomes that appeared after heat shock are shown in blue. The average profiles of nucleosomes in each group before and after heat shock are shown on the right. The *k*-means clustering for (A) and (B) was done based on data from −200 to TSS, but data are shown for −300 to +100.

(C) A remodeling score for eviction and for appearance was separately calculated for activated genes and repressed genes (Materials and Methods), and the data were plotted using box plots similar to Figure 3B. Activated genes showed significantly higher eviction scores than appearance scores, whereas repressed genes showed significantly higher appearance scores than eviction scores.

(D) Nucleosome positions at the promoters of ribosomal protein genes during normal growth and after heat shock, clustered on data from −200 to +100 bp.

(E) Remodeling profiles of genes whose expression changed by less than 1.2-fold after heat shock, clustered based on data from −300 to +100 bp.

doi:10.1371/journal.pbio.0060065.g004

and −100 bp regardless of the transcriptional status of the genes.

The enrichment of transcription factor targets in these four groups based on data from the yeast functional regulatory network [20] and transcription factor ChIP-microarray (ChIP-chip) [17,21] is tabulated in Table S2. Group 3 was significantly enriched for the targets of Rap1, Sfp1, Fhl1, Gcn5, and Esa1, all of which are factors mediating the transcription of ribosomal protein genes during normal growth [22–25]. Consistent with this, ribosomal protein genes were significantly enriched in Group 3 ($p = 2.6 \times 10^{-5}$). In addition, Group 1 was significantly depleted for targets of all the above-mentioned transcription factors, and was also significantly depleted for ribosomal protein genes ($p = 6.5 \times 10^{-4}$).

In order to quantitate whether distinct modes of nucleosome remodeling were generally used for gene activation and repression, we calculated a nucleosome remodeling score for both nucleosome eviction and nucleosome appearance (Materials and Methods). Activated genes showed significantly higher nucleosome eviction than nucleosome appearance, whereas repressed genes showed significantly higher nucleosome appearance than eviction (Figure 4C). Although these general trends are expected, we noted that if we clustered remodeling profiles based on more distal promoter regions (−400 to −200 bp upstream of the TSS), we did observe several apparent nucleosome appearance events at activated promoters (Figure S5). At some promoters, nucleosome eviction proximal to the promoter could occur in conjunction with nucleosome appearance more distally, as would be expected for translational repositioning of nucleosomes.

Since ribosomal protein genes form one of the most prominent classes of genes that are transcriptionally repressed by heat shock, we analyzed nucleosome changes at their promoters separately. Ribosomal protein genes were clustered into three classes based on the presence or absence of a well-positioned nucleosome between −50 and +100 bp in normally grown cells, and the nucleosome score. Upon heat shock, we observed the appearance of medium- to high-scoring nucleosomes between −200 and +100 bp of almost all of these ribosomal protein genes in the three groups (Figure 4D).

This analysis of nucleosomal changes at the promoters of the most strongly regulated genes indicates that chromatin remodeling events accompanying transcriptional regulation are restricted to a small number of discrete patterns involving one or two nucleosomes, rather than encompassing a larger domain around the promoter. We also clustered the nucleosome remodeling profiles for genes whose expression did not change appreciably by the physiological perturbation (less than 1.2-fold change). Surprisingly, we still observed similar specific patterns of single-nucleosome remodeling events at many of these promoters, indicating that specific nucleosome events are not universally associated with transcriptional changes (Figure 4E).

Dynamic Nucleosome Remodeling Causes Changes in the Accessibility of Transcription Factor Binding Sites

Nucleosome positioning can influence the accessibility of the core promoter as well as binding sites for sequence-specific transcriptional regulators [10,26]. About 90% of the sites occupied by transcription factors on chromosome III under normal growth conditions were depleted of nucleosomes [2]. Examination of single-nucleosome remodeling at promoters that were activated or repressed by heat shock in our data revealed instances where the accessibility of the TSS and of experimentally defined transcription factor binding sites was indeed affected by remodeling. For example, at the *UBC4* promoter, which is activated by heat shock, three moderately positioned nucleosomes covering two distinct Hsf1 binding sites as well as the TSS were evicted, whereas a single, well-positioned nucleosome appeared between the two Hsf1 binding sites (Figure 5A). Conversely, at the *RPL17B* promoter, which is repressed by heat shock, one well-positioned nucleosome appeared after heat shock to cover the TSS and a low-confidence proximal Rap1 binding site. Interestingly, another moderate nucleosome upstream was evicted, exposing a higher confidence distal Rap1 binding site as well as an Fhl1 site (Figure 5B). Such eviction and appearance of nucleosomes at adjacent sites could either reflect translational repositioning or independent events; our experiments cannot distinguish between these two possibilities.

Based on these observations and other computational predictions of whole-genome nucleosome positions [4], we hypothesized that chromatin remodeling upon transcriptional perturbation could result in changes in the accessibility of the functional binding sites of stress-related transcription factors. To test this hypothesis, we measured the change in accessibility of transcription factor binding sites upon heat shock, by comparing the overlap between functional binding sites for transcription factors measured by ChIP-chip [17] and nucleosome positions before and after heat shock (Figure 6). Of the 101 factors tested, 46 had fewer than 20 functional binding sites each in the genome, and we therefore excluded them from this analysis. The remaining 55 transcription factors could be stratified into three classes based on the change in accessibility of the functional binding sites after heat shock: factors whose binding sites showed an increase in accessibility after heat shock (Figure 6A), factors whose

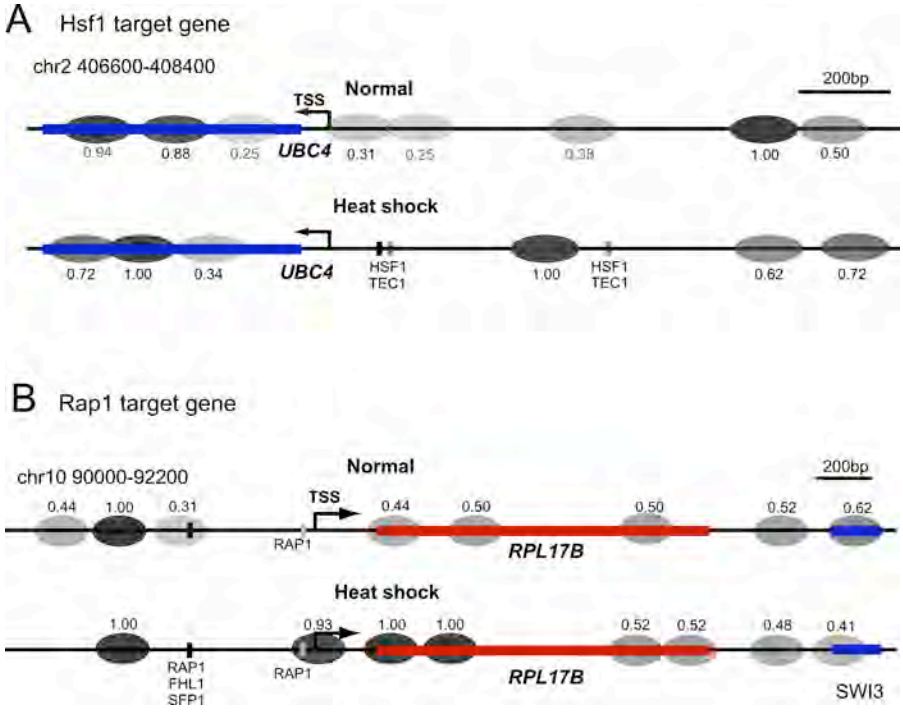


Figure 5. Dynamic Nucleosome Remodeling Affects the Accessibility of Transcription Factor Binding Sites and the TSS

(A) Example of nucleosome eviction at the heat-shock–activated *UBC4* promoter (blue line). Nucleosomes defined by our sequencing data are indicated by ovals, colored according to their stability score. The positions of transcription factor binding sites are from [17] and are shaded according to their confidence. Binding sites for other transcription factors are also affected by remodeling (unpublished data), but these are not known to be related to heat shock.

(B) Example of nucleosome appearance at the heat-shock–repressed *RPL17B* promoter (red line).

doi:10.1371/journal.pbio.0060065.g005

binding sites showed no significant change in accessibility (Figure 6B), and those that showed decreased accessibility after heat shock (Figure 6C). As hypothesized, most of the transcription factors involved in mediating the stress response belonged to the first group. The functional binding sites for several key stress-related transcription factors such as Hsf1, Msn2, Msn4, and Aft2 showed some of the strongest increases in accessibility because of nucleosome repositioning upon heat shock. In addition, binding sites for transcription factors Abf2 and Cbf1, which are involved directly or indirectly in chromatin remodeling [27,28], showed increased accessibility. Surprisingly, we also observed increased accessibility for transcription factors involved in ribosomal protein gene transcription such as Rap1 and Fhl1 (see Figure 5B for an example). These two transcription factors continue to occupy ribosomal gene promoters even during transcriptional repression [29,30], raising the possibility that their occupancy of the promoter under such conditions, facilitated by the increased chromatin accessibility that we observed, could be related to a repressive function. Transcription factors whose binding sites did not show a significant change in accessibility were mainly those involved in the regulation of genes in metabolic pathways.

Discussion

We have mapped the dynamic remodeling of most nucleosomes in the yeast genome during a transcriptional perturbation using a combination of micrococcal nuclease digestion, isolation of mononucleosome associated DNA and Solexa sequencing. Using a Parzen window–based approach, which is a generally applicable method to analyze all similar datasets derived from ultra-high–throughput sequencing, we defined the dynamic remodeling of approximately 50,000 nucleosomes at single-nucleotide resolution in normally growing cells and in cells that were transcriptionally perturbed by heat shock for 15 min. Our study independently confirms expectations about nucleosomal positioning based on previous smaller scale and lower resolution studies, but also reveals novel features about chromatin structure and transcriptional activity, especially given that previous studies have not examined the dynamic repositioning of nucleosomes in response to genome-wide transcriptional reprogramming.

Our results showed that in addition to a positioned nucleosome at the TSS, genes in general tend to also contain a well-positioned nucleosome at the 3′ end of the coding region. Yeast genes are thus demarcated by a well-positioned nucleosome at each end of their transcribed regions, with a nucleosome-free gap just beyond. This could potentially reflect chromatin organization that facilitates RNA polymerase initiation as well as termination. Most coding regions also showed strongly and regularly positioned nucleosomes, although the strength of the nucleosome positioning was weaker in genes transcribed at high rates. Interestingly, the first well-positioned boundary nucleosome downstream of the TSS, which is likely to be an H2A.Z variant–containing nucleosome based on previous studies [9], showed similar stability in genes transcribed at high and low rates (Figure

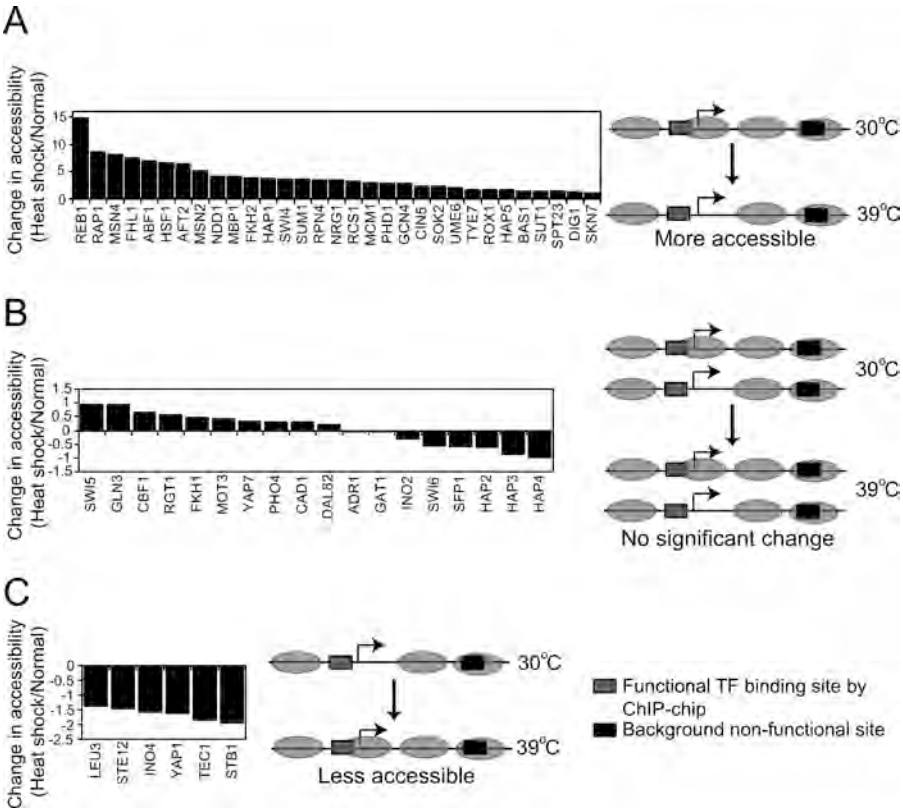


Figure 6. Change in Accessibility of Functional Transcription Factor Binding Sites

Transcription factors were classified into three groups based on the change in accessibility of their functional binding sites because of nucleosome repositioning after heat shock. Graphs of accessibility changes in arbitrary units (see Materials and Methods) are plotted for transcription factor binding sites that (A) showed an increase in accessibility, (B) showed no significant change in accessibility, and (C) showed a decrease in accessibility upon heat shock. The right of each graph shows a schematic of the relationship between nucleosomes and transcription factor binding sites.

doi:10.1371/journal.pbio.0060065.g006

3B), suggesting that this chromatin landmark is important for demarcating promoters.

Upon transcriptional perturbation, the majority of nucleosomes did not change positions, either at promoters or within coding sequences (Figures 2 and 3). Gene-specific remodeling was restricted to the discrete eviction, appearance, or repositioning of one or two nucleosomes localized to promoters. Remodeling events at genes that were activated or repressed upon heat shock could be classified into distinct patterns, indicating that there is no simple rule for nucleosome remodeling at promoters to activate and repress genes. Thus, although activation was generally and quantitatively associated with nucleosome eviction and transcriptional repression with nucleosome appearance (Figure 4C), there were cases in which strongly positioned nucleosomes appeared at activated promoters (Figures 5 and S5). Translational repositioning of nucleosomes would seem like eviction and appearance at different spots in the same promoter. These observations suggest that nucleosome remodeling at promoters is not a trivial consequence of transcriptional activity appearing as overall openness of chromatin at activated promoters and obstruction at repressed promoters, but rather, that the precise placement of individual nucleosomes at promoters mechanistically regulates transcription by modulating access of *trans*-acting factors to specific sites.

In addition to chromatin remodeling specifically at regulated promoters, many promoters however showed dynamic single-nucleosome remodeling during the physiological perturbation even in the absence of any resulting transcriptional change (Figure 4E), indicating that selective, activity-specific remodeling was accompanied by a certain number of background, nonspecific remodeling events. We speculate that these background single-nucleosome remodeling events poise promoters for rapid future transcriptional activity, by either assembling partial preinitiation complexes [31], or by exchanging core histones with one or more histone variants [32]. A recent study showed that nucleosomes are globally positioned by Isw2 acting at the boundary between genes and intergenic regions, and that some of the Isw2-dependent remodeling occurs independent of transcription [11]. Therefore, the background remodeling seen in the absence of transcriptional changes in our study could potentially reflect nonspecific remodeling by ISW-like complexes.

We classified transcription factors into three classes based on change in accessibility of their binding sites upon transcriptional perturbation. All the prominent stress-related transcription factors belonged to the category showing a strong increase in accessibility upon transcriptional perturbation. In addition, we found that Rap1 and Fhl1 binding sites showed an increase in accessibility even though the majority of their target genes, namely the ribosomal protein genes, showed a decrease in transcription upon heat-shock

stress. When transcription of the ribosomal protein genes is repressed by heat shock, osmotic shock, or inhibition of the TOR pathway by rapamycin, it is known that Ifh1 leaves the promoter, but Rap1 and Fhl1 remain bound [30]. It is possible that Rap1 and Fhl1 play a role in recruiting chromatin remodelers to bring about a repressive chromatin structure at the ribosomal protein genes. Previous studies have indicated that the primary discriminant between a functional and a nonfunctional transcription factor binding site in vivo is the presence of stably positioned nucleosomes covering the latter [4,9]. Our results above indicate that superimposed on this, there is a second mode of regulation at functional binding sites of stress-related transcription factors brought about by a stimulus-dependent remodeling of one or two nucleosomes, making the site more accessible for stable binding of transcription factors. Alternatively, binding of the transcription factor(s) could result in the remodeling of nucleosomes via the help of chromatin remodelers.

The work described here is the first study of genome-wide dynamic nucleosome remodeling events at single-base resolution. More such studies in yeast and higher eukaryotes will shed light on the relationship between epigenetic changes at high resolution and the global regulation of gene expression.

Materials and Methods

Preparation of mononucleosomes. Yeast S288C cultures were grown in rich medium and subjected to 15-min heat shock as described previously [18,33]. At the end of 15 min, control and heat-shocked cells (200 ml each) were treated with formaldehyde to a final concentration of 1% for 30 min. The reaction was stopped by adding glycine to a final concentration of 125 mM, and cells were harvested by centrifugation. Cells were washed 2× in PBS and resuspended in 20 ml of Zymolyase buffer (1 M sorbitol, 50 mM Tris [pH 7.4], and 10 mM β-mercaptoethanol). Cells were spheroplasted by treating with 25 mg of 20T Zymolyase, and incubated for 40 min at 30 °C with shaking at 200 rpm. The remainder of the steps were carried out using a modified protocol described in [2]. Briefly, cells were spun down, washed 1× with 5 ml of Zymolyase buffer, and resuspended in 2 ml of NP buffer (1 M sorbitol, 50 mM NaCl, 10 mM Tris [pH 7.4], 5 mM $MgCl_2$, 0.075% NP 40, 1 mM β-mercaptoethanol, and 500 μM spermidine). $CaCl_2$ was added to a final concentration of 3 mM, and micrococcal nuclease digestions were carried out at concentrations ranging from 100 U/ml to 600 U/ml for 10 min at 37 °C. The reactions were stopped by adding 100 μl of 5% SDS and 50 mM EDTA. A total of 3 μl of 20 mg/ml proteinase K was added to each tube, and incubated at 65 °C overnight. The DNA was purified by phenol-chloroform-isoamyl alcohol (25:24:1) extraction, and precipitated using ethanol. The DNA was treated with DNase-free RNase, re-extracted with phenol-chloroform-isoamyl alcohol, precipitated with ethanol, and resolved on a 1.25% agarose gel alongside a 100-bp ladder. The mononucleosome size band (approximately 150–200 bp) was excised and purified using the Invitrogen Pure-Link quick gel extraction kit. The purified DNA was sequenced using Solexa sequencing technology.

RNA isolation and expression profiling. S288C cells from 50-ml cultures before and after heat shock at 39 °C for 15 min were resuspended in 8 ml of AE buffer (50 mM sodium acetate [pH 5.2], 10 mM EDTA, 1.7% SDS). RNA extraction, cDNA labeling, and microarray manufacture and hybridizations were done as described previously [18,20,33]. For absolute expression analysis, sheared genomic DNA was labeled with Cy3, and cDNA was labeled with Cy5. For relative expression-change analysis, cDNA from heat-shocked cells was labeled with Cy5, and cDNA from normally grown cells was labeled with Cy3. The labeled cDNAs were mixed and hybridized onto DNA microarrays for 12–16 h. The arrays were washed, dried, and scanned with a Axon 4000B scanner (Molecular Devices). Cy5/Cy3 ratios were quantitated using GenePix Pro software and analyzed using Acuity microarray informatics software after filtering to exclude bad spots.

qPCR validation. Primer pairs used in Figure 1 were designed to cover three peaks and three troughs in the promoter of *PHO5* just upstream of the known Pho4 binding and DNaseI hypersensitive site [12]. Control primers used for normalization were designed in the region between *YCR023C* and *YCR024C*. qPCR was performed using SYBR green chemistry on an ABI 7900 instrument. Enrichment of target loci in the ChIP sample relative to sonicated genomic DNA was calculated for both unstressed cells and cells subjected to heat shock.

Nucleosome position detection. Solexa sequencing reads were mapped back to the Oct 2003 yeast genome assembly obtained from the *Saccharomyces* Genome Database (SGD) (http://www.yeastgenome.org/) and only reads that mapped uniquely to the genome were considered in the majority of our analysis. We generated 514,803 and 1,036,704 uniquely aligning reads for the normal and heat-shock growth conditions, respectively. Reads mapping to the plus and minus strands were processed separately. Reads were clustered using a Parzen window–based approach. Essentially, a Gaussian kernel was centered on each base pair in the genome, and a weighted score was calculated at that position. The mean of the Gaussian was taken as the position under consideration, with the standard deviation (smoothing bandwidth) set at 20 bp. Each read contributed to the mean position based on its kernelized distance from the mean. The weighted score indicated the likelihood of finding an edge of the nucleosome at the position. Thus, the entire genome was converted into a likelihood landscape that was further processed to find local maxima (Figure S6). These maxima were then treated as centers of a cluster. Membership of a read in a cluster was based on its relative contribution to the weighted score of the center. The number of reads assigned to a cluster was defined as the unweighted score of that cluster. We reasoned that a stable nucleosome would be expected to result in a denser clustering of the reads than an unstable one. The denser clustering of the reads results in better concordance of the unweighted score to the weighted score. Hence, each cluster was assigned a stability score that was calculated as the ratio of the unweighted score to the weighted score. Nucleosomes were identified as a plus cluster followed by a minus cluster within 100–200 bp. The nucleosome score was calculated as a sum of the plus and minus cluster unweighted scores. The nucleosome stability score was calculated as a weighted average of the individual stability scores of the participating clusters.

Overlap between unstressed and heat-shock–stressed cells. Whole-genome maps for unstressed and stressed cells were filtered to exclude nucleosomes that had a normalized score less than 0.2 (see normalization procedure below). For each nucleosome in unstressed cells, the distance to the nearest nucleosome after heat shock was calculated. These data are reported in Table S1. Similar analysis was used to determine the overlap between nucleosome positions determined in this study and those from previous studies [2,4].

Random simulations to generate a normalization factor. Reads equal in number to those we obtained from normal and heat-shocked cells were selected at random from the yeast genome assembly Oct 2003, and peak finding was done as described. This process was iterated 20 times. The average maximum score obtained in the simulations was used as a scaling factor to normalize nucleosome peak scores for cells grown at 30 °C. Normalization was done by dividing nucleosome peak scores by the scaling factor. We then calculated a scaling factor for the heat-shock data by multiplying the scaling factor for the 30 °C data by the ratio of the median peak scores for 39 °C to the peak scores for 30 °C. This was done to correct for differences in sequencing depth for the two samples, thus enabling quantitative comparison of nucleosome profiles across the two conditions.

Average nucleosome profiles for TATA-containing and TATA-less genes and separation by transcription rates. The upstream −600 bp to downstream +1,000 bp of each uncharacterized and verified ORF in SGD was binned at 10 bp, and nucleosomes were mapped to each bin. The zero point was the TSS. A nucleosome was said to map to a given bin if it completely overlapped with the 10-bp bin. Each bin was assigned the score of the overlapping nucleosome. In the cases where our algorithm detected overlapping positions for a nucleosome, and more than one nucleosome mapped to a single bin, the bin was assigned the highest score. Genes were separated into TATA-containing or TATA-less [34], and the average nucleosome profiles were generated for each group by averaging the scores for the bin across all the genes (973 and 4,382 promoters, respectively). Genes were similarly separated into the top 500 or bottom 500 with respect to transcription rates [16], and average profiles were plotted for these classes.

Nucleosome positioning periodicity score and dinucleotide positioning profile. The NPP score was generated by calculating the similarity of the experimentally derived nucleosome profile over the coding region of every gene to an artificially generated profile where

six nucleosomes of score 1.0 were regularly placed with 30-bp linker lengths. In general, genes with well-positioned nucleosome had profiles that were most similar to the synthetic profile and hence, had a high NPP score. The first (+1) nucleosome downstream of the TSS is adjacent to a gap and is likely to be more strongly sequence dependent for positioning than a nucleosome that is flanked by other nucleosomes. We therefore derived AA/TT profiles from the sequence underlying the first nucleosome. To derive high-confidence sequence profiles, we aligned all genes to the first nucleosome as shown in Figure 3A. We selected all +1 nucleosomes with a score $\geq$0.9 for the input set. Since nucleosomes show a dyad symmetry in terms of positioning over DNA, the reverse complement of each sequence in the input set was also included before calculating the profile. We calculated frequency profiles for the dinucleotides AA and TT, and summed and smoothed them using a 3-bp moving average. This high-confidence AA/TT profile was then correlated with the AA/TT profiles derived from all nucleosomes at the +1, +2, +3, and +4 positions.

Generation of nucleosome remodeling profiles and remodeling score. Genes that did not a have 200-bp–long promoter region were excluded for this analysis. For all of the genes that passed this filter, the difference between the nucleosome scores in normally grown cells and cells after heat shock was calculated bin-wise from −400 bp upstream to +200 bp downstream of the start codon. For the plots and clusters shown in Figure 4A and 4B, we then created subsets of these data that included either genes that were activated by at least 2-fold, or genes that were repressed at least 2-fold by heat shock. For the cluster in Figure 4E, we selected remodeling profiles that showed a difference in nucleosome score of at least 0.5 between the two growth conditions at three or more positions in the promoter, and also selected genes whose expression did not change by more than 1.2-fold. To calculate the remodeling score, a seven-bin window, corresponding to a distance of 70 bp (approximately half of a nucleosome), was scanned along each profile, and the individual bin scores were averaged for each window. The maximum window score in the positive direction across the entire profile was assigned as the remodeling score for nucleosome eviction while a similar maximum in the negative direction was assigned as the remodeling score for nucleosome appearance.

Increase in accessibility of transcription factor binding sites after stress. Transcription factor motifs were mapped across the entire genome using position-weight matrices derived from [17] using Patser [35] at a *p*-value cutoff of 0.01. These were considered the putative binding sites while the functional ("true") binding sites were derived from published ChIP-chip data [17,18,36]. A functional motif was considered to be occupied, and therefore not accessible, if it overlapped with a nucleosome that had a score $\geq$ 0.5. The occupancy of the ChIP-chip binding sites was compared to that of the putative motif binding sites, and a hypergeometric distribution was used to calculate *p*-values. This analysis was done with data from both normal and heat-shock conditions. To calculate the significance of the change in binding site occupancy upon heat shock, the *p*-values for the heat-shock nucleosome data were divided by the *p*-values derived from the normal condition data.

Supporting Information

Figure S1. Comparison of Nucleosome Positions Before and After Heat Shock, As Well As with Previously Reported Nucleosome Positions

(A) and (B) show different regions of the genome. In each track, the raw sequencing data is on top and consists of uniquely aligning reads extended by the average fragment length selected for sequencing. Below this are the nucleosome positions calculated by our analysis algorithm, with their scores shown next to their positions (see Figure S6). Previously reported nucleosome positions as reported by Yuan et al [2] and Lee et al [13] are indicated.

Found at doi:10.1371/journal.pbio.0060065.sg001 (1.7 MB EPS).

Figure S2. Average Nucleosome Profiles after Heat Shock

(A) Nucleosome profile of all genes in the yeast genome from −600 bp to +1,000 bp with respect to the TSS. Nucleosome positions are shown as gray ovals below the profile. The intensity of the filled oval reflects the average probability score of the nucleosome, and the dotted oval around the filled oval marks the spread of that nucleosome across all genes.
(B) The 3′ end of genes is marked by a strongly positioned nucleosome, followed by a relatively nucleosome-free region. The inset shows the 3′ end of convergently transcribed genes in which the 3′ end is not followed by another promoter.
(C) Average nucleosome profiles for TATA-containing (973) and TATA-less (4,382) promoters, aligned with respect to the TSS.

Found at doi:10.1371/journal.pbio.0060065.sg002 (950 KB EPS).

Figure S3. Internucleosomal Linker Length Distribution in the Yeast Genome

Linker lengths were binned into 10-bp (top) or 5-bp bins (bottom), and their frequency distribution was plotted. The most frequent inter-nucleosomal distance, or linker length, was 25–30 bp. The small peak of linker length at 180 bp in the top graph likely reflects the nucleosome-free region at promoters.

Found at doi:10.1371/journal.pbio.0060065.sg003 (1.5 MB EPS).

Figure S4. TBP Occupancy of Promoters and Absolute Expression Levels of the Different Classes of Genes with Distinct Promoter Nucleosome Profiles Shown in Figure 2C

(A) Box plots showing TBP occupancy using data derived from [33], and (B) box plots showing absolute expression levels before and after heat-shock stress. Absolute expression levels were measured as the $\log_2$ ratio in a DNA+RNA hybridization on genomic microarrays.

Found at doi:10.1371/journal.pbio.0060065.sg004 (1.8 MB EPS).

Figure S5. *k*-Means Clustering of Nucleosome Remodeling Profiles over Heat-Shock–Activated Promoters from −400 to −200 bp Upstream of the TSS

Nucleosome eviction upon heat shock is indicated by yellow, and nucleosome appearance after heat shock is indicated by blue. Clusters 1 and 2 together were significantly enriched for targets of Hsf1 ($p < 0.04$). Although cluster 3 shows nucleosome appearance, this set could include promoters where a nucleosome was evicted from a downstream region and repositioned upstream (e.g., *UBC4* as shown in Figure 5). It could also include promoters where a nucleosome is appearing to cover a repressor site in the heat-shock–activated promoter, or actually appearing at the promoter of another divergently transcribed gene that is repressed by heat shock.

Found at doi:10.1371/journal.pbio.0060065.sg005 (744 KB EPS).

Figure S6. The 0.5-kb Window Showing Parzen Window–Based Peak Detection

(A) Reads mapping to the plus strand (red) and minus strand (blue) were processed separately.
(B) Each base position was assigned a score that was derived from the sum of the relative contributions of all reads in its neighborhood as defined by a Gaussian kernel positioned at that coordinate. A local maximum on the plus strand (red) followed by a corresponding maximum on the minus strand (blue) within a distance of 100 to 200 bp defines a nucleosome. Peaks that were assigned higher Parzen scores defined higher confidence nucleosomes as shown by the grey shading.

Found at doi:10.1371/journal.pbio.0060065.sg006 (1.4 MB EPS).

Table S1. Nucleosome Overlaps

(A) Overlap between nucleosomes mapped in this study with previous studies. Percentages were calculated with reference to the lower of the two numbers considered in the overlap. The threshold for displacement was $\leq$50 bp.
(B) Overlap between nucleosome positions before and after transcriptional perturbation in this study.

Found at doi:10.1371/journal.pbio.0060065.st001 (62 KB DOC).

Table S2. Enrichment and Depletion of Transcription Factor Targets in Nucleosome Profile Clusters from Figure 4B

Found at doi:10.1371/journal.pbio.0060065.st002 (66 KB DOC).

Acknowledgments

Author contributions. SS performed the experiments and wrote the paper. AB analyzed the data and contributed reagents/materials/analysis tools. YZ, SJ, and MH contributed reagents/materials/analysis tools. VRI conceived and designed the experiments and wrote the paper.

Funding. This work was supported in part by funds from the National Institutes of Health and the US Army to VRI.

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

References

1. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251–260.
2. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309: 626–630.
3. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, et al. (2007) Nucleosome positioning signals in genomic DNA. Genome Res 17: 1170–1177.
4. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772–778.
5. Wang JP, Widom J (2005) Improved alignment of nucleosome DNA sequences using a mixture model. Nucleic Acids Res 33: 6743–6755.
6. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006) Nucleosome positions predicted through comparative genomics. Nat Genet 38: 1210–1215.
7. Rando OJ, Ahmad K (2007) Rules and regulation in the primary structure of chromatin. Curr Opin Cell Biol 19: 250–256.
8. Whitehouse I, Tsukiyama T (2006) Antagonistic forces that position nucleosomes in vivo. Nat Struct Mol Biol 13: 633–640.
9. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446: 572–576.
10. Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. Cell 128: 707–719.
11. Whitehouse I, Rando OJ, Delrow J, Tsukiyama T (2007) Chromatin remodelling at promoters suppresses antisense transcription. Nature 450: 1031–1035.
12. Fascher KD, Schmitz J, Horz W (1993) Structural and functional requirements for the chromatin transition at the PHO5 promoter in Saccharomyces cerevisiae upon PHO5 activation. J Mol Biol 231: 658–667.
13. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39: 1235–1244.
14. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL (2004) Global nucleosome occupancy in yeast. Genome Biol 5: R62.
15. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 36: 900–905.
16. Garcia-Martinez J, Aranda A, Perez-Ortin JE (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell 15: 303–313.
17. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.
18. Hahn JS, Hu Z, Thiele DJ, Iyer VR (2004) Genome-wide analysis of the biology of stress responses through heat shock transcription factor. Mol Cell Biol 24: 5249–5256.
19. Gorner W, Durchschlag E, Martinez-Pastor MT, Estruch F, Ammerer G, et al. (1998) Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. Genes Dev 12: 586–597.
20. Hu Z, Killion PJ, Iyer VR (2007) Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet 39: 683–687.
21. Robert F, Pokholok DK, Hannett NM, Rinaldi NJ, Chandy M, et al. (2004) Global position and recruitment of HATs and HDACs in the yeast genome. Mol Cell 16: 199–209.
22. Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28: 327–334.
23. Marion RM, Regev A, Segal E, Barash Y, Koller D, et al. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. Proc Natl Acad Sci U S A 101: 14315–14322.
24. Reid JL, Iyer VR, Brown PO, Struhl K (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. Mol Cell 6: 1297–1307.
25. Rudra D, Zhao Y, Warner JR (2005) Central role of Ifh1p-Fhl1p interaction in the synthesis of yeast ribosomal proteins. EMBO J 24: 533–542.
26. Anderson JD, Widom J (2000) Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. J Mol Biol 296: 979–987.
27. Kent NA, Eibert SM, Mellor J (2004) Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. J Biol Chem 279: 27116–27123.
28. Miyake T, Loch CM, Li R (2002) Identification of a multifunctional domain in autonomously replicating sequence-binding factor 1 required for transcriptional activation, DNA replication, and gene silencing. Mol Cell Biol 22: 505–516.
29. Schawalder SB, Kabani M, Howald I, Choudhury U, Werner M, et al. (2004) Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1. Nature 432: 1058–1061.
30. Wade JT, Hall DB, Struhl K (2004) The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. Nature 432: 1054–1058.
31. Zanton SJ, Pugh BF (2006) Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. Genes Dev 20: 2250–2265.
32. Guillemette B, Bataille AR, Gevry N, Adam M, Blanchette M, et al. (2005) Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol 3: e384. doi:10.1371/journal.pbio.0030384
33. Kim J, Iyer VR (2004) Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. Mol Cell Biol 24: 8104–8112.
34. Basehoar AD, Zanton SJ, Pugh BF (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell 116: 699–709.
35. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563–577.
36. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298: 799–804.