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 2006

TYPE OF REPORT: Annual

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
01-04-2006	Annual	31 Mar 2005 – 30 Mar 2006

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

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 that are important in breast cancer. STAGE is based on high-throughput sequencing of concatamerized tags derived from DNA associated with transcription factors that is isolated by chromatin immunoprecipitation. Identification of the direct targets of oncogenic transcription factors important in breast cancer can help elucidate their role in mediating gene regulatory networks underlying the disease. Our work over the past year has successfully identified targets of c-Myc, E2F4 and Stat1, all factors important in breast cancer, and our data supports the idea that it is possible to comprehensively identify the targets of transcription factors important in breast cancer using STAGE, as proposed in our Statement of Work.

15. SUBJECT TERMS
Oncogenes, Genomics, Transcription Factors, Chromosomal Targets

16. SECURITY CLASSIFICATION OF:				18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	11	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

(Slightly modified from previous Progress Report)
This project is aimed at developing a novel, unbiased genome-wide method 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 often 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 chromosomal targets of transcription factors can be identified by the technique of ChIP-chip, where DNA bound by a transcription factor in vivo is first isolated after crosslinking and immunoprecipitation, then identified by hybridization to a comprehensive whole-genome microarray that includes all potential regulatory elements. Although whole-genome tiling arrays suitable for ChIP-chip are just becoming available for the human genome, they remain expensive and challenging to use. Moreover, tiling microarrays are not yet available for many model organisms. In our original Statement of Work, we proposed to develop a new genomic method named STAGE (Sequence Tag Analysis of Genomic Enrichment) that can potentially overcome some of the limitations of ChIP-chip analysis and can be applied to transcription factors important in breast cancer. STAGE is based on high-throughput sequencing of concatamerized tags derived from DNA associated with transcription factors that is isolated by chromatin immunoprecipitation. Identification of the direct targets of oncogenic transcription factors important in breast cancer can help elucidate their role in mediating gene regulatory networks underlying the disease.

Body

Our proof of principle paper demonstrating the successful development and application of STAGE for E2F4, a human transcription factor, was published as an Article in *Nature Methods* [2]. E2F4 is a member of the E2F family of transcription factors that are associated with cell proliferation and cancer, and E2F4 has complex roles in breast cancer. It has been reported to be a tumor suppressor gene [3] as well as an oncogene [4]. It is also reported to mediate the transcription of ER-alpha, which is a key transcriptional regulator in breast cancer [5]. Our work using STAGE to identify E2F4 targets was described in the previous Progress Report (April 2005) and those details are not duplicated here. Below we describe our progress during the current reporting period (31 March 2005 - 30 March 2006).

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

We have now also successfully applied STAGE for identifying c-myc target genes. Using a validated c-myc ChIP-chip sample, we carried out the STAGE strategy to isolate tags representing c-myc targets. We isolated approximately 4500 recombinant clones in *E. coli* for generating the c-Myc tag library. Purified clones were sent to Agencourt (http://www.agencourt.com/) for sequencing. Each clone had on average about 15 to 20 tags. 21 bp tags were extracted from each clone using custom Perl scripts. Out of a total of 127,351 tags extracted for c-Myc, 15% of the sequenced tags did not map to the human genome. The proportion of these tags, termed orphan tags, was consistent with previous estimates for SAGE. Thus we obtained 107,484 tags for c-Myc that were used for all further analysis. The number of sequenced tags and the number of distinct (non-redundant or unique) tags is given in Table 1.

Although we had carried out ChIP for estrogen receptor (ER-α) and begun to analyze targets by ChIP-chip using our core promoter array as described in the last progress report, a recent paper was published that described a similar comprehensive analysis of ER-α targets by ChIP-chip [6]. Because the generation and sequencing of STAGE tags is labor-intensive and expensive, we decided to focus our attention on c-myc and an alternative transcription factor that was more likely to yield novel biological information, in addition to helping the development of STAGE analysis methods. We therefore applied STAGE for the analysis of targets of the Stat1 transcription factor. Stat1 represents a family of transcription factors that have been implicated as being important in breast cancer [7], and have recently been shown to be activated by hypoxia in breast cancer cells [8]. To improve the efficiency of high throughput sequencing required for STAGE, we developed a novel adaptation of our STAGE method in order to exploit a recent development in sequencing technology, namely the bead and emulsion based pyrosequencing technology pioneered by 454, Inc., which obviates the need for a cloning step [9]. We collaborated with the lab of Mike Snyder (Yale University) to obtain Stat1 ChIP samples and prepared STAGE ditags. Ditags in this library were amplified using a 454 specific PCR primer and this library was sent to 454 for sequencing.

The standard sequencing technology we had been using initially costs $3 per read at Agencourt. At this rate, the cost of sequencing is **14c/tag**. With 454 technology, the cost was $ 9000 per run, with each run generating 150,000—200,000 reads (2 tags per read). Thus, our cost for sequencing was approximately **2.5c/tag**. Given that the 454 modification of STAGE does not require cloning and plasmid purification, this is a huge improvement in cost and labor efficiency. The number of tags we sequenced using the modified 454 procedure for STAGE is given in Table 1.

Table 1. Summary of tags analyzed for c-myc and Stat1 target identification. Orphans are tags that don't map to the genome sequence. The remainder are viable tags used for analysis. Distinct refers to the number of unique tags in each category. For both c-myc and Stat1, we identified and excluded duplicate reads that represented the same original ditag in the library that got sequenced more than once.

	c-myc		Stat1	
	Total	Distinct	Total	Distinct
All	127,351	101,114	162,577	133,116
Orphans	19,867	17,925	30,781	27,097
Viable	107,484	83,189	131,224	106,019

The frequency of occurrence of the given tag in the genome was termed *nhit* while the frequency of occurrence of the tag the STAGE sequencing pool was termed *nocc.* If STAGE tags were truly derived from ChIP-enriched DNA then the distribution of tags in the STAGE sequencing pool should deviate from a randomly population. We randomly selected similar numbers of tags from the genome as we obtained after sequencing, and analyzed the distribution of tags that occurred multiple times in our library. As shown in Figure 1, for a frequency (nocc) of 1, the numbers of tags in the random and real data are almost the same. However, for a frequency of 2 and above, the

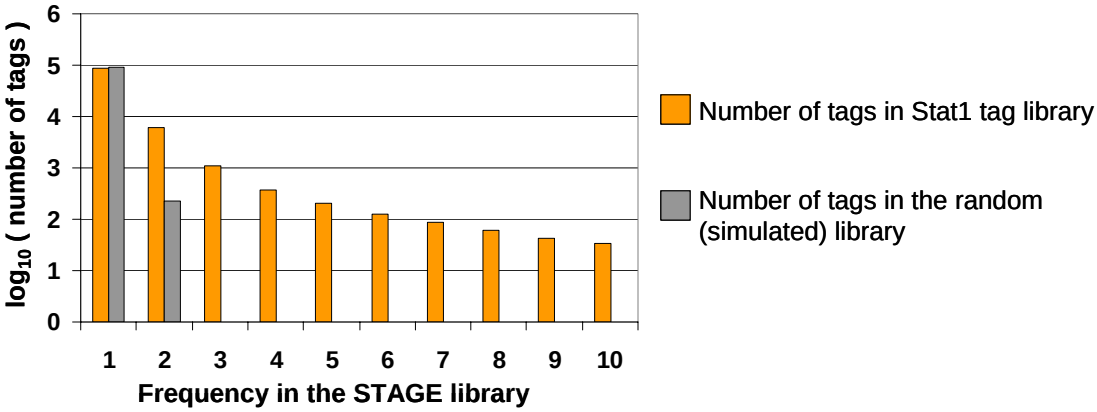

Figure 1. Distribution of tags with different nocc values in the experimental STAGE library compared to a random simulation.

enrichment in the experimental STAGE library is evident.

We developed a statistical method for defining transcription factor targets based on STAGE tags. This method relies on scanning a window of defined size across each chromosome, and scoring the window in terms of its probability of being significantly enriched in the ChIP procedure, rather than being background. Through optimizations involving random simulation and calculation of the false discovery rate (FDR), we found that a window size of 500 bp gave satisfactory separation between our experimental data and randomly simulated data.

For the c-myc STAGE data the probability that a given window is a target is given by

$$1-\prod_i\left(1-p(\mathrm{tag}_i)\right)$$

Where $p(\mathrm{tag}_i)$ = Probability that tag *i* was enriched, and is given by

$$\left(1-\frac{\text{expected frequency}}{\text{observed frequency}}\right)$$

Modeling the selection of tags from the genome at random as a binomial distribution,

Expected frequency = $\left(1-\sum_{0}^{nocc-1}\binom{N}{x}p^{x}(1-p)^{N-x}\right)\mathrm{M}$

where $p = nhit/\mathrm{T}$, T is the total number of 21 bp CATG(N_{17}) tags found in the entire genome (2.742×10^7), and M = number of tags with a given $nhit$.

Based on our validation results comparing STAGE data to tiling microarray ChIP-chip data (see below), we estimate significantly higher numbers of targets for c-myc than Stat1, indicating that our sampling of true c-myc targets is more sparse than for Stat1. To define Stat1 targets with high confidence, we used a slightly modified algorithm from the one described above.

STAGE detected 2242 targets for c-Myc in the entire genome at a probability threshold of 0.8. The expected false discovery rate at this threshold was 0.05. For Stat1, STAGE detected 381 binding sites in the entire genome with a probability greater than 0.9. The false discovery rate at this probability threshold of 0.9 was calculated to be < 0.01.

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

There are three primary ways in which we have attempted to validate our targets determined by STAGE with independent measurements. First we compared the STAGE targets to ChIP-chip tiling array data from within the ENCODE regions which cover about 1% of the human genome. Second, we have compared STAGE data to ChIP-chip data from core promoter microarrays. Third, we have examined the enrichment of consensus motifs in the sites determined to be targets by STAGE. A fourth method of validation, based on qPCR measurements of occupancy is ongoing.

The ENCODE project is a consortium to examine functional elements in the human genome and is currently analyzing in detail, 1% of the human genome [10]. High density tiling oligonucleotide microarrays are available for this 1% of the genome through NimbleGen, and we have been able to use some of this arrays because of our participation in the ENCODE consortium. We compared our STAGE targets to ChIP-chip targets within the ENCODE regions. 26 of the 2242 c-Myc targets that we identified by our STAGE analysis lay within the ENCODE regions. We performed three independent c-Myc ChIP-chip replicates in HeLa cells and identified binding peaks that lay within the ENCODE regions using NimbleGen tiling arrays. 14 of the 26 STAGE detected c-Myc binding sites were within 500 bp of a ChIP-chip peak for at least 1 ChIP-chip replicate (Figure 2). Interestingly, our ChIP-chip analysis revealed that c-Myc is likely to have a very large number of binding sites throughout the genome. Although we detected only 26 binding sites for c-Myc in the ENCODE region, the ChIP-chip analysis indicated that there were 232 binding sites in this 1% of the human genome.

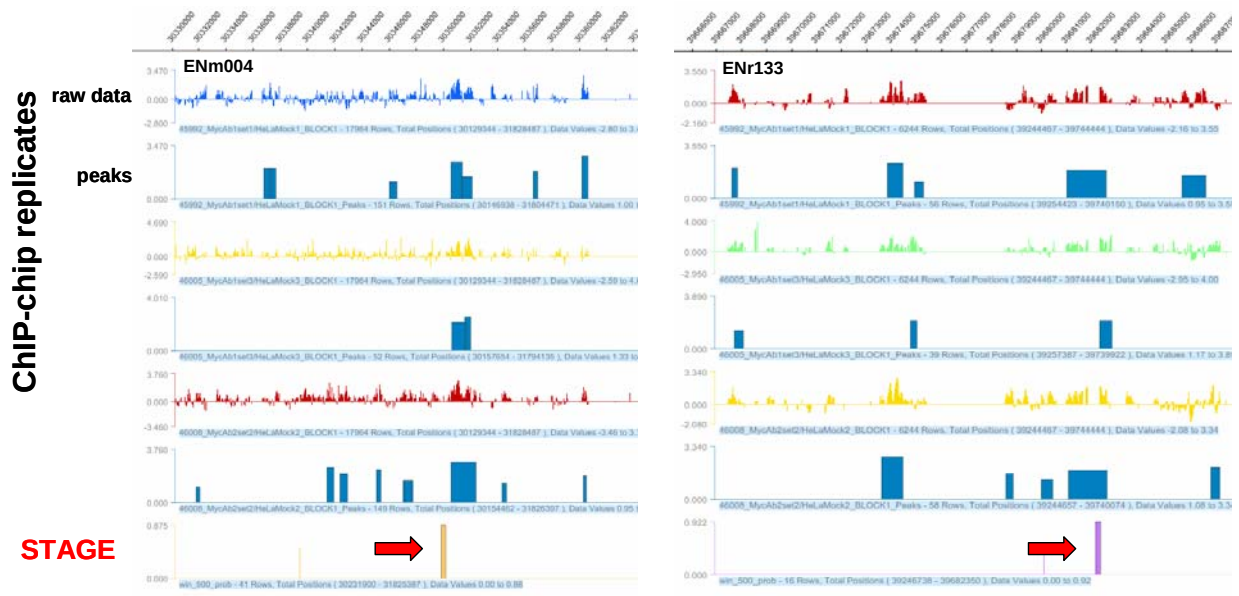

Figure 2. Example of overlap between c-Myc STAGE and ChIP-chip targets within the ENCODE regions. Two ENCODE regions, ENm004 and ENr133 are shown. The top part of each panel shows the raw tiling array data, where the ratio of ChIP DNA to control is shown by a dense bars. Below each track is the peak of binding. The bottom row shows STAGE predicted target sites.

Extrapolating to the entire genome therefore suggests that there could be approximately 25,000 myc binding sites, which is consistent with previous estimates also based on partial tiling arrays .[11]. Thus although STAGE is unbiased and is a genome-wide method in principle, for a factor such as c-Myc, which has a large number of targets, our STAGE analysis has recovered only a subset. Our coverage of c-Myc targets can potentially be improved by additional in-depth sequencing. For the subsequent analysis, in this report, we focus on Stat1, which has fewer targets and therefore better coverage by STAGE.

STAGE detected 7 Stat1 binding sites at a threshold of 0.9 in the ENCODE regions. 3 out of these 7 sites overlapped with a ChIP-chip peak where the Stat1 ChIP-chip was also performed on ENCODE arrays (Figure 3). We next compared results obtained by STAGE with Stat1 ChIP-chip performed on a core-promoter array. The core-promoter array had 9764 different promoters spotted and ChIP-chip revealed 195 promoters to be bound by Stat1 at a red/green log ratio > 2.0. If a Stat1 binding site detected by STAGE was found within 1 kb upstream and 200 bp downstream of the transcription start site (TSS) of a gene from the RefSeq database, that gene was considered a Stat1 target. 20 out of these 9764 promoters were detected as Stat1 targets by STAGE and 12 out of these 20 targets overlapped with the ChIP-chip results. Based on a hypergeometric probability distribution, this overlap was significant at a *P*-value of $< 10^{-12}$. Overall, STAGE detected 59 genes in RefSeq as Stat1 targets according to the above criteria. 62% of these 59 genes (37/59) had the GAS Stat1 promoter motif TTCNNNGAA within 1 kb upstream and 200 bp downstream of the TSS of the gene. This represented an enrichment of more than 2-fold as compared to background. The background was considered as 1 kb upstream and 200 bp downstream of the TSS of all genes in RefSeq. This enrichment was significant at a *P*-val $< 10^{-8}$ assuming a hypergeometric distribution.

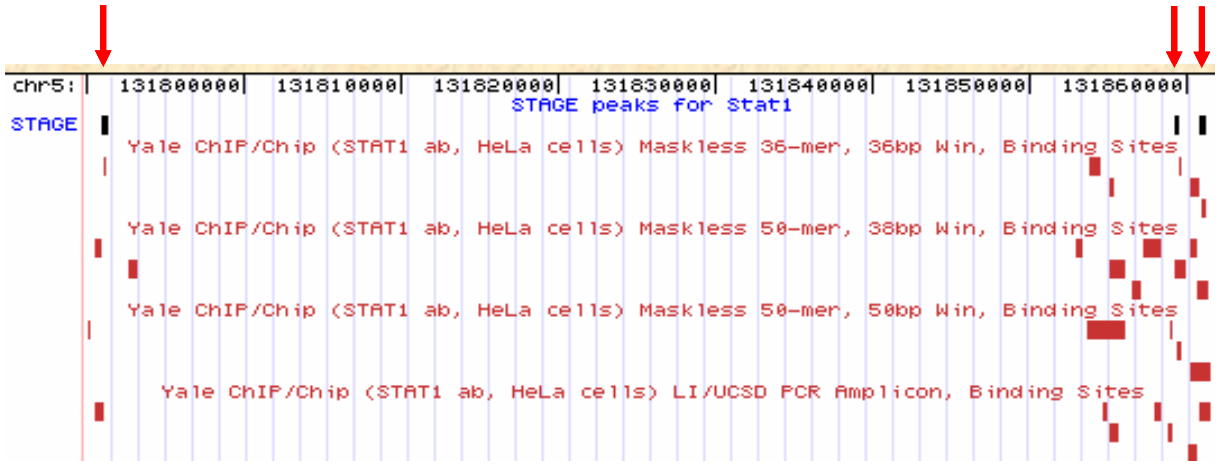

Figure 3. Example of overlap between Stat1 targets identified by STAGE with ChIP-chip peaks within the same ENCODE region (ENm002) using data from our collaborator M. Snyder from Yale University. STAGE peaks are shown in the top row and ChIP-chip peaks from different experiments are shown below.

Key Research Accomplishments

- Used STAGE to identify targets of c-Myc and Stat1, transcription factors that are known to be important in breast cancer.
- Modified the STAGE procedure to take advantage of new high-throughput sequencing technology (454) that is more cost-effective and efficient.
- Developed analysis algorithms for determining STAGE targets in human cells based on tag data.
- Verified STAGE targets by comparison to ChIP-chip data from ENCODE tiling arrays and core promoter microarrays, as well as motif enrichment in STAGE targets for Stat1.

Reportable Outcomes

- 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 {Kim, 2005, 15782160}, described in the previous progress report.
- 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 previous progress report).

Conclusions

Our work over the past year supports the idea that it is possible to comprehensively identify the targets of transcription factors important in breast cancer using STAGE, as proposed in our Statement of Work. However, even with high throughput sequencing technologies, the coverage or comprehensiveness of STAGE remains limited by the extent of sequencing. We had proposed to focus on c-Myc and ER as proof-of-principle examples of transcription factors important in breast cancer, and have now demonstrated positive results using STAGE for c-Myc, E2F4 and Stat1, but not ER. However, data for ER is now available through a core promoter ChIP-chip publication from a different lab. For the next year, we propose to further validate STAGE targets by quantitative PCR (qPCR) and use this validation data to improve out target calling algorithms.

References

1. Darnell, J. E., Jr.: Transcription factors as targets for cancer therapy. *Nat. Rev. Cancer* (2002), **2:** 740-749
2. Kim, J., Bhinge, A. A., Morgan, X. C. & Iyer, V. R.: Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* (2005), **2:** 47-53
3. Ho, G. H., Calvano, J. E., Bisogna, M. & Van Zee, K. J.: Expression of E2F-1 and E2F-4 is reduced in primary and metastatic breast carcinomas. *Breast Cancer Res Treat* (2001), **69:** 115-122
4. Rakha, E. A., Pinder, S. E., Paish, E. C., Robertson, J. F. & Ellis, I. O.: Expression of E2F-4 in invasive breast carcinomas is associated with poor prognosis. *J Pathol* (2004), **203:** 754-761
5. Macaluso, M., Cinti, C., Russo, G., Russo, A. & Giordano, A.: pRb2/p130-E2F4/5-HDAC1-SUV39H1-p300 and pRb2/p130-E2F4/5-HDAC1-SUV39H1-DNMT1 multimolecular complexes mediate the transcription of estrogen receptor-alpha in breast cancer. *Oncogene* (2003), **22:** 3511-3517
6. Laganiere, J. *et al.*: From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proc. Natl. Acad. Sci. USA* (2005), **102:** 11651-11656
7. Clevenger, C. V.: Roles and regulation of stat family transcription factors in human breast cancer. *Am. J. Pathol.* (2004), **165:** 1449-1460
8. Lee, M. Y. *et al.*: Phosphorylation and activation of STAT proteins by hypoxia in breast cancer cells. *Breast* (2005),
9. Margulies, M. *et al.*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005), **437:** 376-380
10. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* (2004), **306:** 636-640
11. Cawley, S. *et al.*: Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs. *Cell* (2004), **116:** 499-509

Appendices
None