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 2007

TYPE OF REPORT: Annual

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE 01-04-2007	**2. REPORT TYPE** Annual	**3. DATES COVERED** 31 Mar 2006 – 30 Mar 2007

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 originally 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 was based on high-throughput sequencing of concatamerized tags derived from DNA associated with transcription factors isolated by chromatin immunoprecipitation. Over the last year, we have significantly improved the efficiency of our methodology by modifying the initial method to take advantage of new developments in sequencing technology. We first used a bead-based, pyrosequencing method developed by 454/Roche to identify the targets of STAT1. We developed new methods to score target genes and independently verified targets using quantitative real time PCR. Secondly, we have adapted our approach to use even more high-throughput sequencing technology developed by Solexa to identify the targets of c-Myc, by sequencing millions of tags. We will also use Solexa sequencing to improve the coverage of E2F4 targets by significantly deeper sequencing.

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** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	12	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction

The objective of this Idea Development award was to develop 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 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. This DNA is then identified by hybridization to a comprehensive whole-genome microarray that includes all potential regulatory elements (ChIP-chip). We proposed to develop an alternative to whole-genome hybridization for target identification. Our method, called STAGE (Sequence Tag Analysis of Genomic Enrichment), 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.

Body

At the time this project was conceived (2004), whole-genome tiling microarrays covering the entire human genome were not available for ChIP-chip. This was a large part of the motivation for developing STAGE, which we have successfully accomplished. In the last couple of years, whole-genome tiling arrays are becoming increasingly available, and their prices are coming down. For example, it is possible to perform a single whole-genome scan using oligonucleotide tiling arrays from Affymetrix or NimbleGen for approximately $3000-$8000. At the same time sequencing technology has made rapid strides such that it is possible to sequence hundreds of thousands to millions of sequence tags from small sample sizes without the need for cloning in plasmid vectors as we originally conceived in our proposal. At this point, rather than view STAGE and similar sequence based methods (which have been developed simultaneously in a few labs) as a competing method to ChIP-chip, it makes sense to think of them as complementary methods. First of all, one of the limitations of ChIP-chip remains that the coverage of the genome even on whole-genome tiling arrays is incomplete. Only the non-repetitive regions of the human genome are represented on these microarrays, which can be as high as 50%. ChIP sequencing methods like STAGE do not suffer from this limitation because the entire genome is potentially sampled by sequencing, and only the depth of sequencing limits the extent of coverage. Secondly, whole-genome tiling arrays and sequencing provide the only large-scale quantitative validation of each other; it is very impractical to carry out real-time quantitative PCR on a large enough scale to accomplish this. Third, the sequencing approach has a clear 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 or DNaseI hypersensitive sites, which are correlated with where oncogenic transcription factors bind in vivo.

Over the last year our emphasis has been to exploit the revolutionary developments in sequencing technology enabled by 454 and Solexa. We have completed the analysis and validation of STAT1 targets (begun in last year's annual report) and begun deep sequencing of c-Myc targets using Solexa technology. We are also beginning the application of our sequencing approach to identifying the positions of single nucleosomes in vivo. In order to best take advantage of the novel high-throughput sequencing technologies that were completely unanticipated at the time of our original proposal, as well as recent developments in the availability of high resolution tiling oligonucleotide microarrays, we have obtained a no-cost extension of our project until March 31 2008.

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

A) In last year's annual report we described the rationale for initiating the identification of STAT1 targets using the 454 methodology. Here we describe the completion of that analysis, including validation of identified targets by quantitative real-time PCR. We used 454 technology [2] to sequence approximately 160,000 tags comprising the STAGE library for STAT1. In order to identify STAT1 targets with high confidence, we developed a slightly modified and more stringent algorithm compared to what was described earlier for c-Myc targets in the 2006 report (see Task 2). Since approximately 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. 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 less than 5% for STAT1 (Fig. 1). We used a window of 500 bp for all further analysis and used our

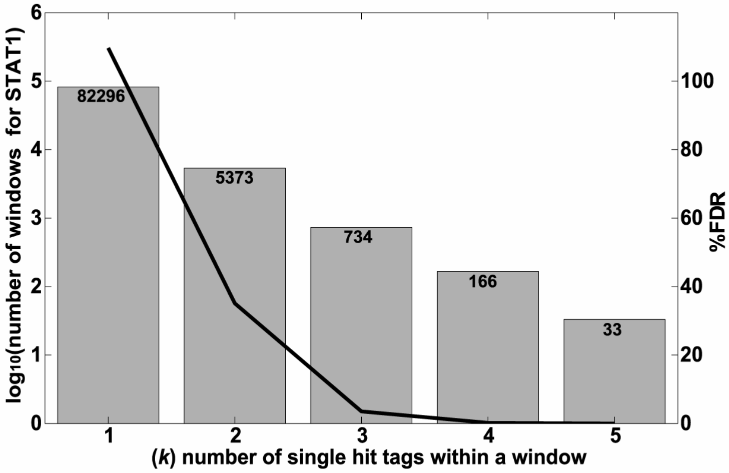


Figure 1. A window size of 500 bp gave an optimal separation between random and real data. The grey bars indicate $\log_{10}$ of the number of windows detected based on STAT1 tags, with actual numbers of windows at each k listed at the top of the column. The black line shows the decline in the false 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.

improved and more stringent analysis algorithm to define STAT1 targets. STAGE detected 381 binding sites for STAT1 in the entire genome at this threshold. Based on annotations in the RefSeq gene database [3], 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).

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

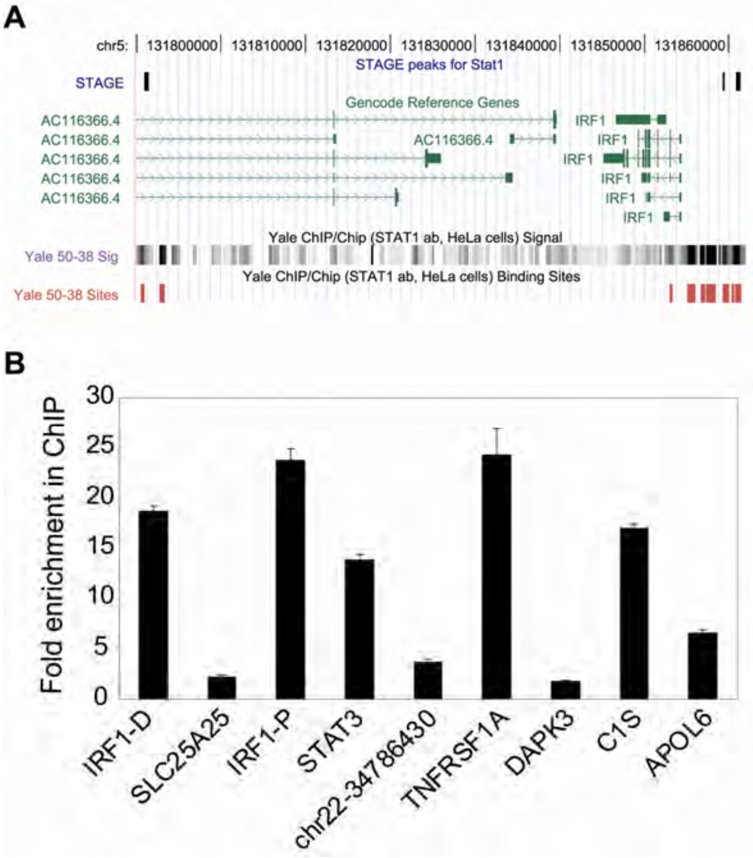
Position of binding sites	**percentage of binding sites**
Relative to transcription start sites of the gene (percentage of total sites)	
Within 50 kb	68%
Within 20 kb	47%
Within 20 kb upstream	24%
Within 20 kb downstream	23%
Sites found internal to genes (percentage of internal sites found within 20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

B) Although 454 sequencing technology used above for STAT1 offers great advantages over traditional clone-based sequencing, it still does not achieve saturation coverage of all targets contained in a ChIP sample. An alternative which provides even deeper sequencing is the single-molecule sequencing technology developed by Solexa (now Illumina) which uses a "sequencing by synthesis on arrays" approach [4]. A single flow cell (out of 8) on a Solexa instrument can generate up to 5 million sequence reads from the ends of DNA fragments. Only on the order of 100 nanograms of DNA are needed, without any need for cloning and sequence reads from 25-27 bases are generated. We have now used Solexa technology for STAGE, and applied it initially to identify the targets of c-Myc. We carried out ChIP for c-Myc in human HeLa S3 cells and sent the ChIP sample to the Michael Smith Genome Sciences Center in Vancouver, BC. From two flow cells we obtained approximately 10.3 million sequence reads of 27 bp each. Of these about 5.4 million reads could be aligned perfectly to the Human Genome sequence, whereas about 6.5 million reads could be aligned allowing for mismatches. Thus although there is some error rate with the sequencing, the total number of reads from a single run still exceeded what we were able to achieve with 454 sequencing. Analysis of this large sequencing data is ongoing and some preliminary results are presented below.

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

A) To improve the specificity of target detection for STAT1 targets, 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 less than 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

Figure 2. (A) Overlap between STAGE and ChIP-chip. 3 out of the 7 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 verification of binding sites identified by STAGE. 9 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.



modeled in the simulation.

We have used a combination of tiling arrays (ChIP-chip) and quantitative real-time PCR to validate STAGE targets. Seven of the 381 STAT1 binding sites that we identified in the genome using STAGE were within the ENCODE regions. 3 of these 7 targets overlapped with a ChIP-chip peak where the STAT1 ChIP-chip was performed on ENCODE region tiling oligonucleotide arrays (Fig. 2A). In order to obtain a quantitative estimate of the false positive rate of our STAGE analysis, we selected ten 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 these ten sites showed a quantitative enrichment in the ChIP sample relative to the input, with eight of them showing an enrichment of more than 2 fold (Fig. 2B). Thus, we estimate our true positive rate to be approximately 90% giving a false positive rate of 0.1.

If a STAT1 binding site detected by STAGE occurred within 1 kb upstream and 200 bp downstream of 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. 62% of these target 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 a motif enrichment among target promoters of more than 2-fold compared to background. This enrichment was statistically significant (P-value < 10^{-8}). 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%) had the GAS consensus sequence. This represents an enrichment of more than 2-fold over background (P-value < 10^{-43}) (Fig. 3). Additionally, in accordance with the fact that STAT1 is known to exhibit co-operative binding with other transcription factors like AP1, c-Myc and NF-kappaB, we found an enrichment for the STAT1 motif along with motifs for AP1, c-Myc and NF-kappaB (Fig. 3).

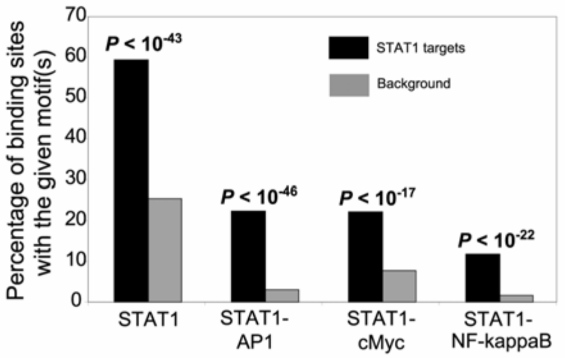


Figure 3. Enrichment of STAT1 motifs. The Y-axis shows the percentage counts of the number of sites bearing the given motif(s) out of the 381 STAT1 binding sites detected by STAGE. Almost 60% of the 381 binding sites had the STAT1 motif TTCNNNGAA as compared to 27% in the background. We also detected an enrichment for the co-occurrence of the binding motifs for STAT1 and AP1 (TGAG/CTCA), STAT1 and c-Myc (CACA/GTG) and STAT1 and NF-kappaB (GGGA/GNNC/TC/TCC).

We identified several previously unknown STAT1 target genes, many of which are involved in interferon γ (IFN-γ) signaling. IFN-γ increases DAPK3-Daxx complex formation which is necessary for induction of caspases and IFN-γ mediated apoptosis [5]. STAT1 targets DAPK3 and this could represent one mechanism by which IFN-γ can induce apoptosis. Another possible mechanism for IFN-γ mediated apoptosis was suggested by the fact that APOL6, which induces mitochondria-mediated apoptosis characterized by the release of cytochrome-c and activation of caspase-9 [6], was also identified as a STAT1 target by STAGE.

STAT3 is anti-apoptotic and induces cell proliferation while STAT1 promotes growth arrest and apoptosis. IFN-γ induces high levels of expression of STAT1 while STAT3 levels remain low. However, in the absence of STAT1 i.e., in STAT1-/- cells, IFN-γ stimulation induces high levels of STAT3 gene expression [7]. We found that STAT3 is a direct transcriptional target of STAT1, suggesting that STAT1 represses STAT3 during IFN-γ signaling, further promoting its own apoptotic function. STAGE identified a STAT1 binding site in the first intron of TNFR1 (tumor necrosis factor-α receptor) suggesting the possibility that IFN-γ dependent increased sensitivity to TNF- α [8] could be a direct result of activation of TNFR1 by IFN-γ stimulated STAT1. All the target sites and genes described above were verified by quantitative ChIP from an independent ChIP sample (Fig. 2B). 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 IFN-γ.

B) For the analysis of our c-Myc sequencing results from Solexa, hits on the genome are called by mapping each read back to the genome (considering only uniquely mapping reads). Each read is then extended by the length of the gel-purified ChIP DNA that was subjected to sequencing, which is 150 bp. The number of overlapping fragments at each region is essentially the height of the peak. This part of the analysis was included in the data we obtained from the Vancouver sequencing center. At a peak height of 5, we observed approximately 4600 peaks, whereas at a height of 3, there were more than 200,000 peaks. We are currently examining the overlap between targets identified by sequencing at different thresholds and targets identified by whole-genome tiling microarrays (ChIP-chip) using NimbleGen microarrays. These newly developed tiling arrays cover the non-repetitive portion of the

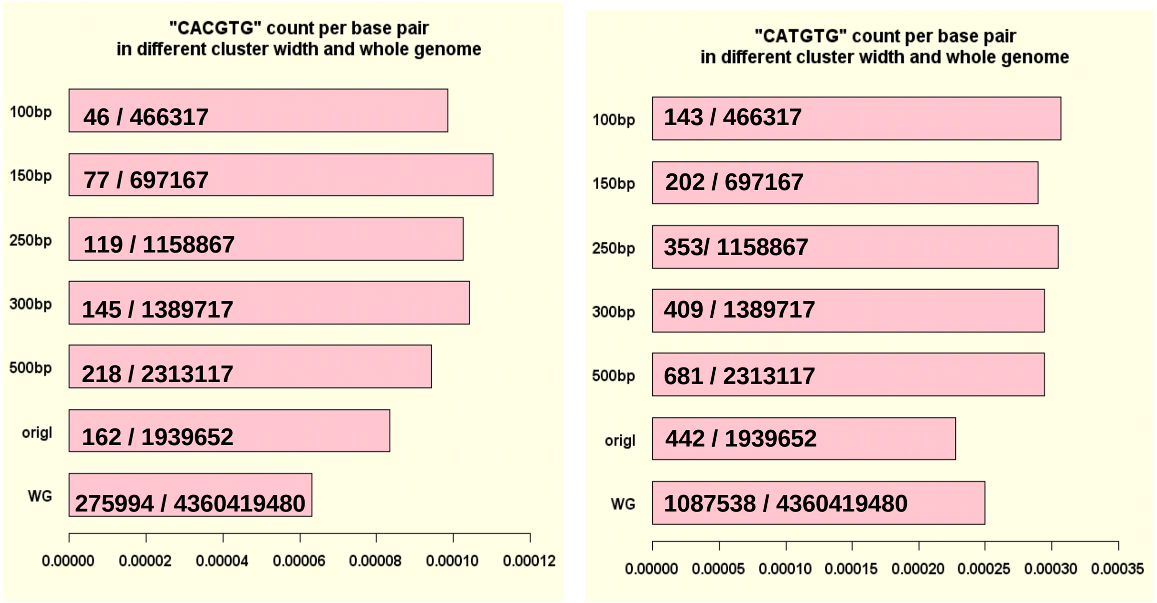


Figure 4. The E-box motif for c-Myc binding is significantly enriched in Solexa sequencing peak clusters compared to the background. The Y axis show peak regions of different widths. The bottom row (WG) shows the representation of c-Myc sites in the background.

human genome in 10 slides at 100 bp resolution, with each slide containing approximately 2.1 million oligos. Comparison of STAGE/Solexa sequencing with tiling arrays offers potentially the best overall global data validation. This global comparison is still in progress as we ascertain the biases and data quality issues of each of these two novel platforms.

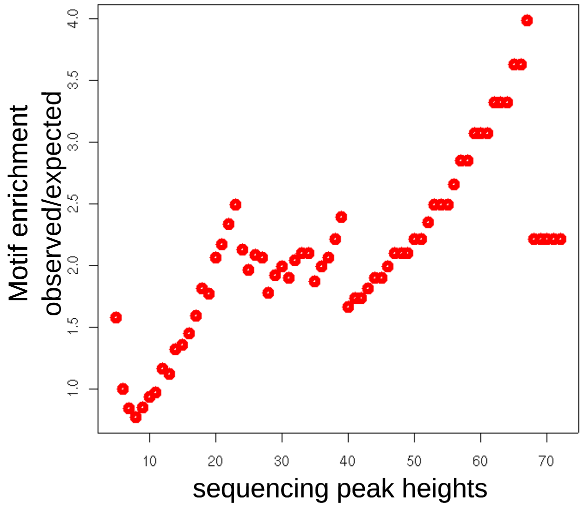


Figure 5. Enrichment of the c-Myc E-box motif within Solexa sequencing peak clusters correlates with the height of the sequencing peaks. The height of the sequencing peaks are generated by overlapping sequence fragments whose ends were sequenced and hence represent enrichment in the ChIP sample.

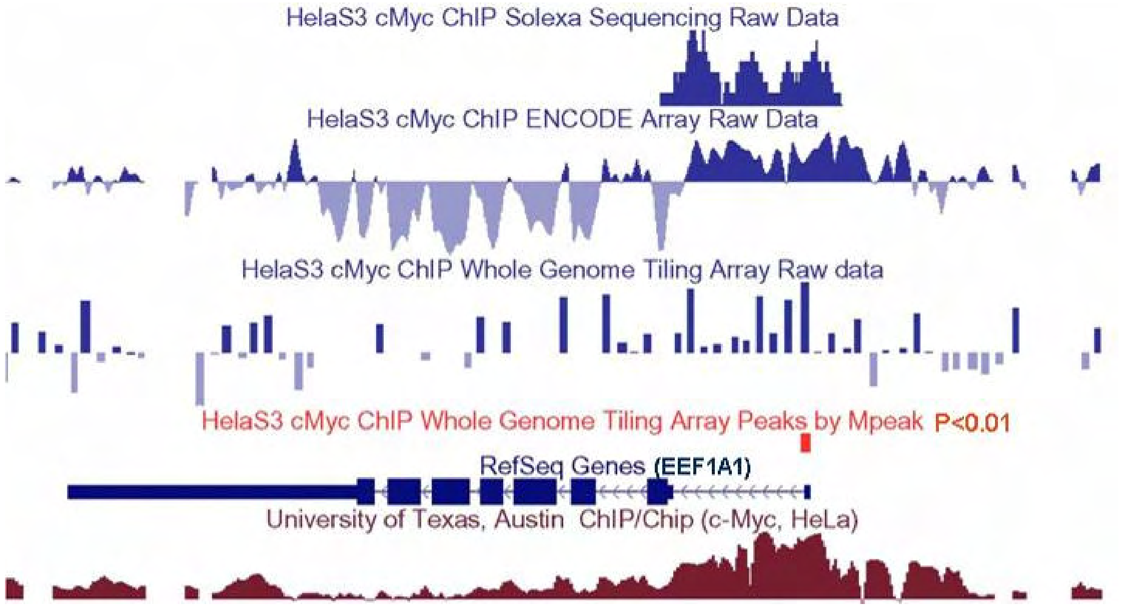


Figure 6. Overlap between Solexa sequencing data (top track) and ChIP-chip data for c-Myc at the promoter of EEF1A1, a known c-Myc target gene. The second track from the top and the bottom track show c-Myc binding as assayed by ENCODE tiling microarrays which cover 1% of the human genome at 38 bp resolution. The middle track shows whole genome tiling array data at 100 bp resolution. The red peak shows a peak at the transcription start site identified using a peak calling algorithm called MPeak.

Nonetheless, we could see a clear enrichment of the binding motif for c-Myc (the E-box: CACGTG or CATGTG) within the peaks called from the whole-genome Solexa sequencing data (Fig. 4). Moreover, there was a correlation between the extent of enrichment for the c-Myc motif and the height of the sequencing peak (Fig. 5). An example of the correspondence of sequencing peaks with ChIP-chip peaks is shown in Figure 6.

Key Research Accomplishments

- Successfully adapted STAGE for sequencing using 454 bead-based pyrosequencing technology

- Completed analysis of STAT1 targets identified by STAGE

- Successfully used Solexa sequencing by synthesis on array approach to sequencing targets of c-Myc.

Reportable Outcomes

- 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.

- Bhinge, A.A., Kim, J., Euskirchen, G., Snyder, M. and Iyer, V.R. Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Research* In press

- The ENCODE Project Consortium (308 authors including Iyer V.R.) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature (2007) In Press

Conclusions

We have successfully used two revolutionary sequencing technologies to identify targets of transcription factors important in oncogenesis, applying 454 to identify targets of STAT1 and Solexa sequencing to identify targets of c-Myc. Analysis of c-Myc targets is underway and it is not clear yet how comprehensively we have been able to identify all the targets of c-Myc. Over the next and final year of the project, we plan to extend our identification of c-Myc and E2F4 targets in additional cell types, achieve deeper sequencing to approach comprehensiveness. Given that the binding of transcription factors in vivo (chromatin) is strongly influenced by the accessibility of chromatin and the positions of nucleosomes, and given that the high throughput sequencing method that was the subject of this project offers unparalleled ability to map the positions of DNA ends at single base resolution, we will also use STAGE for single nucleosome mapping.

References

1) Darnell, J. E., Jr. (2002). Transcription factors as targets for cancer therapy, *Nat. Rev. Cancer* **2**, 740-9.

2) Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376-80.

3) 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**, D501-4.
4) Bentley, D. R. (2006). Whole-genome re-sequencing, *Curr. Opin. Genet. Dev.* **16**, 545-52.
5) Kawai, T., Akira, S., and Reed, J. C. (2003). ZIP kinase triggers apoptosis from nuclear PML oncogenic domains, *Mol. Cell. Biol.* **23**, 6174-86.
6) 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.
7) 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**, 727-33.
8) 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**, 5313-9.

Appendices
None