Mapping of Transcription Factor Binding Regions in Mammalian Cells by ChIP: Comparison of Array and Sequencing Based Technologies

Ghia M. Euskirchen[1]*, Joel Rozowsky[2]*, Chia-Lin Wei[3], Wah Heng Lee[3], Zhengdong D. Zhang[2], Stephen Hartman[1], Olof Emanuelsson[2+], Viktor Stolc, [5] Sherman Weissman[4], Mark Gerstein[2], Yijun Ruan[3], Michael Snyder[1,2]

1
Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven CT 06520-8103, USA
2
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven CT 06520-8114, USA
3
Genome Institute of Singapore, Singapore 138672
4
Department of Genetics, Yale University School of Medicine, New Haven, CT 06520–8005,
USA
5
Center for Nanotechnology, NASA Ames Research Center, Moffett Field, CA 94035, USA
*
To whom correspondence should be addressed: E-mail: michael.snyder@yale.edu (M.S.)
+
Present address: Stockholm Bioinformatics Center, AlbaNova University Center, Stockholm
University, SE-10691 Stockholm, Sweden

*These authors contributed equally

Running title: STAT1 ChIP-chip and ChIP-PET targets in ENCODE

Key words: chromatin immunoprecipitation, ChIP-chip, ChIP-PET, ENCODE, STAT1, transcription factor binding regions

Abstract

Mapping of transcription factor (TF) binding regions has provided tremendous insight into our understanding of gene expression regulatory networks. Recent progress in this field can largely be credited to the application of chromatin immunoprecipitation (ChIP) technologies. We compared strategies for mapping TF binding regions in mammalian cells using two different ChIP schemes: ChIP followed by DNA microarray analysis (ChIP-Chip) and ChIP followed by DNA sequencing (ChIP-PET). In these studies we first investigated parameters central to obtaining robust ChIP-chip datasets through the analysis of STAT1 targets in the ENCODE-designated regions of the human genome, and then compared ChIP-chip to ChIP-PET. We devised methods for scoring and comparing results among various tiling arrays and examined parameters such as DNA microarray format (oligonucleotide or PCR product elements), oligonucleotide length, hybridization conditions, and the use of competitor Cot-1 DNA to determine which among these variables enhances ChIP-chip performance in the detection of TF binding regions. The best performance was achieved with high density oligonucleotide arrays, oligonucleotides ≥50 b, the presence of competitor Cot-1 DNA and hybridizations conducted in microfluidics stations. Furthermore when target identification was evaluated as a function of array number, 80-86% of targets were identified with 3 or more arrays. Comparison of ChIP-chip with ChIP-PET revealed a strong concurrence of results for the highest ranked targets with less overlap for the low ranked targets. With advantages and disadvantages unique to each approach, we found that ChIP-chip and ChIP-PET are frequently complementary in their relative abilities to detect STAT1 targets for the lower ranked targets; each method was able to detect validated targets that were missed by the other method. Thus the most comprehensive list of STAT1 binding regions is obtained by merging the results from ChIP-chip and ChIP-sequencing. Overall, this study provides information for robust identification, scoring and validation of TF targets using ChIP-based technologies.

Introduction

Identification of transcription factor binding sites is essential for understanding the regulatory circuits that control cellular processes such as cell division and differentiation as well as metabolic and physiological balance. Traditionally the pursuit of transcription factor targets has exposed only a few binding regions at a time. However, a population of any given transcription factor molecules is expected to coordinate the activity of many regulatory regions. Hence the identification of additional targets is critical not only to elucidate transcription factor function, but also to fully understand the manner in which specific genes are regulated and ultimately to derive gene regulatory networks.

Recent years have witnessed several new approaches for the global mapping of transcriptional regulatory regions. Such approaches include computational methods that predict target binding sites by way of motif and gene expression analyses (see for example Bailey and Elkan 1995; Liu et al. 2001; Liu et al. 2002; Wasserman and Sandelin 2004) as well as more direct *in vivo* approaches that require isolation of target DNA through chromatin immunoprecipitation (ChIP) of the transcription factor of interest. These ChIP-based strategies identify target binding regions by using the immunoprecipitated DNA to either probe a DNA microarray that tiles significant regions of the human genome (ChIP-chip: Horak et al. 2002; Weinmann et al. 2002; Ren et al. 2002; Martone et al. 2003; Euskirchen et al. 2004; Cawley et al. 2004; Odom et al. 2004) or for direct DNA sequencing (ChIP sequencing: Wei et al. 2006; Impey et al. 2004; Kim et al. 2005a; Chen and Sadowski 2005; Roh et al. 2005). In ChIP-chip experiments the DNA associated with a transcription factor of interest is compared to a reference sample, generally either genomic DNA or any DNA that might be immunoprecipitated with a negative control antibody. ChIP-chip experiments entail the use of DNA tiling microarrays that are prepared either by deposition of PCR products or by oligonucleotide synthesis. These arrays may tile promoter regions, large genomic segments, entire

chromosomes, or in some cases an entire genome (Martone et al. 2003; Cawley et al. 2004; Kim et al. 2005b; Boyer et al. 2005; Lee et al. 2006). ChIP sequencing experiments, on the other hand, do not require the use of a reference sample. Sequencing is performed from individually cloned ChIP fragments (Weinmann et al. 2001; Hug et al. 2004), from concatenations of single 'tags' where each tag is signature derived from a ChIP DNA fragment (STAGE) (Impey et al. 2004; Kim et al. 2005a; Chen and Sadowski 2005; Roh et al. 2005) or from concatenations of *p*aired-*e*nd-di*t*ags cloned from the 5′ and 3′ ends of each ChIP DNA fragment (ChIP-PET) (Wei et al. 2006; Loh et al. 2006).

Although ChIP-based technologies have demonstrated widespread utility, many experimental parameters important for enhancing the performance of ChIP have not been adequately explored for mammalian cells. Moreover, a direct comparison of ChIP-chip and ChIP sequencing has not been performed. Such information is crucial for the large number of experiments that are performed on subsets of mammalian genomes and will become even more crucial as these experiments expand to cover entire genomes.

Our goal here was to first evaluate parameters central to ChIP-chip experiments and then compare targets derived from our most robust ChIP-chip methods to those targets obtained by ChIP-PET. While many microarray parameters for ChIP-chip appear to translate well from previously established microarray protocols (see for example Hegde et al. 2000; Oberley and Farnham 2003; Buck and Lieb 2004; Wu et al. 2006), other variables are more tenuous. In particular, we focused on addressing the following questions: 1) Are ChIP-chip results influenced by factors such as array platform and oligonucleotide length? We note that various research groups have used a range of microarray formats and feature types such as arrays fabricated by contact printing PCR products or by using photolithography to synthesize oligonucleotides. Not only may various lengths of arrayed DNA directly affect hybridization of ChIP DNA, but probe

density also affects data analysis, array cost and quality control in array production, all of which are important considerations for genome-scale experiments. 2) Does the inclusion of competitor Cot-1 DNA improve the accuracy of ChIP-chip results? We note it is difficult to completely exclude all repetitive elements from array tiling paths. 3) Lastly when does one encounter diminishing returns in adding biological replicates to a ChIP-chip dataset?

To explore these questions and to compare ChIP technologies we performed chromatin immunoprecipitations using the sequence specific transcription factor STAT1 (*S*ignal *T*ransducer and *A*ctivator of *T*ranscription). The STAT1 transcription factor is a cytoplasmic protein that translocates to the nucleus and becomes active as a DNA-binding factor when cells encounter interferons or other peptide signals (reviewed in Boehm et al. 1997; Bromberg and Chen 2001; Levy and Darnell 2002). STAT1-dependent transcription is important for determining immune and inflammatory responses, antiviral effects, proliferation, apoptosis and differentiation (Levy and Darnell 2002; Boehm et al. 1997). STAT1 was selected by the ENCODE consortium (ENCODE Project Consortium 2004) as an ideal factor to test the performance of ChIP DNA across platforms and is a model factor for two main reasons: 1) STAT1 ChIP experiments show less enrichment than those with more general DNA-binding proteins such as PolII or chromatin modifications and hence would be expected to more thoroughly test the performance of various platforms, and 2) STAT1 is inducible, and therefore it offers a direct biologic control in the form of STAT1 ChIP samples prepared from control cells not treated with interferon-γ (IFN-γ). STAT1 ChIP-chip studies have been conducted previously on a chromosome 22 PCR product tiling array (Hartman et al. 2005). In the study presented herein, chromatin immunoprecipitations were performed to find many previously unidentified binding regions for STAT1 under IFN-γ stimulation by first comparing various ChIP-chip experimental parameters and then using the most robust of

these tiling array experimental designs to assess the performance of ChIP-chip relative to ChIP-PET.

Our STAT1 mapping studies focus on the ENCODE regions which represent 1% of the human genome in 44 subregions that sum to 30 Mb (ENCODE Project Consortium 2004). The 44 subregions range in length from 500 kb to 1.9 Mb and were selected to include loci of biological interest and regions that stratify both gene density and non-exonic conservation with mouse. The final results from the datasets described here have also been included in the meta-analysis conducted by the ENCODE Transcriptional Regulatory Elements Subgroup (The ENCODE Consortium 2006). Our studies are expected to provide useful information for comparing and integrating data generated from the ENCODE group as well as for future genome-scale studies that map transcription factor binding regions using ChIP-based methods.

Results

Exploring ChIP-chip Performance: Longer Oligonucleotides Yield Better Signals

In the first phase of these studies we investigated ChIP-chip performance on oligonucleotide arrays synthesized by maskless photolithography (Nuwaysir et al. 2002). The results from these maskless arrays could then be compared with datasets generated using other microarray platforms (PCR product arrays) and a ChIP sequencing technology (ChIP-PET). For these studies we used chromatin immunoprecipitated STAT1 DNA, which produces modest signal enrichments relative to ChIP DNA isolated to study other transcription factors and chromatin modifications. HeLaS3 cells were treated with IFN-γ to induce STAT1 binding, and then incubated briefly with 1% formaldehyde to crosslink protein to DNA. The cells were lysed, nuclei were prepared, and chromatin was sheared to approximately 1 kb final DNA size. STAT1 and its associated DNA were immunoprecipitated using an anti-STAT1 antibody. Crosslinks

were reversed and the success of each immunoprecipitation was examined by PCR analysis using primers to a known STAT1 binding region in the promoter of IRF1 (Interferon Regulatory Factor 1; Hartman et al. 2005) whose locus is included in the ENCODE regions.

Using this assay we investigated the effects of varying a number of parameters on the performance of ChIP-chip. These parameters included the type of beads used in the immunoprecipitation step (magnetic or sepharose), various labeling technologies and array hybridization conditions. The final ChIP and microarray conditions selected are reported in the Materials and Methods. No difference in immunoprecipitation efficiency was observed using magnetic as opposed to sepharose beads. However signal enrichment and array uniformity were significantly improved when the hybridization solution was continuously circulated over the array surface using microfluidic chambers; thus all arrays were subjected to this procedure. We also included unlabeled Cot-1 competitor DNA in all hybridizations as noted below, except for the experiment where biological replicates were split and hybridized either in the presence or absence of Cot-1 repetitive DNA (Supplementary Table 1).

Arrays with oligonucleotides of different lengths (25–60 b) are currently used for ChIP-chip experiments (Cawley et al, 2004; Kim et al., 2005b; Boyer et al. 2005). We systematically examined the contribution of array oligonucleotide length to ChIP-chip performance. Custom arrays with 36 b, 50 b, 60 b or 70 b oligonucleotides tiling most or all of the ENCODE regions were designed and synthesized by maskless photolithography (See Materials and Methods). The oligonucleotides were designed to comprehensively cover nonrepetitive regions and are tiled end-to-end such that immediately adjacent genomic DNA segments are represented on the arrays. Thus the short oligonucleotide arrays have more probes per region, but are expected to exhibit lower signals and

increased cross hybridization relative to arrays with longer oligonucleotides, depending upon the exact conditions employed. Lower signals will reduce accuracy when enrichment ratios are determined (see Discussion). STAT1 ChIP DNA prepared from nuclear extracts of IFN-γ treated cells was labeled with Alexa647 and hybridized to the arrays along with Alexa555-labeled STAT1 ChIP DNA isolated in parallel from the nuclear extracts of uninduced (STAT1-nuclear excluded) cells. Each biological replicate was labeled and hybridized independently of all other biological replicates in a given dataset. The 36 b array dataset contained two biological replicates; all other ChIP-chip datasets contained three or more biological replicates (see Supplementary Table 1). Array signals representing enrichments in ChIP DNA samples from IFN-γ-treated cells relative to those ChIP DNA samples prepared from untreated control cells were scored using a sliding window approach (see Materials and Methods). As shown in Figure 1, significantly higher signal enrichments for STAT1 target regions were observed from the arrays containing 50 b oligonucleotides relative to the arrays containing 36 b oligonucleotides. Increases in oligonucleotide lengths to 60 b and 70 b improved array performances only marginally compared to the 50 b arrays (data not shown). The reduced performance of the 36 b arrays was not due to the use of two replicates; a comparison of signal enrichments from datasets comprised of two biological replicates hybridized to each of two 36 b and 50 b arrays yielded a similar outcome (see later section and Figure 6). The lower signal enrichments observed with the 36 b arrays were also not likely due to suboptimal hybridization affinities as several different hybridization conditions were tested for the arrays at each oligonucleotide length and improved signal enrichments were not apparent with any of these alternative conditions (See Materials and Methods). Moreover the expected difference in hybridization temperature for the 36 b array relative to 50 b array is calculated to be 4°C or less (Bertone et al. 2006). Thus, longer oligonucleotides enhance performance and for all remaining experiments 50 b arrays were used.

Validation of Targets from the 50 b Oligonucleotide Array

Signal enrichment maps are suggestive of binding regions, but in order for array performance to be properly assessed it is essential to validate the targets identified from the ChIP-chip experiments. Therefore, we devised a scheme to measure sensitivity and specificity of the experiments. STAT1 targets were ranked according to their signal enrichments and fractions of these targets were sampled across the rankings and tested for enrichment by ChIP-PCR analysis. Primer pairs were designed to different loci and STAT1 ChIP DNA was used as template in ChIP-PCR validation studies; a two-fold or greater enrichment in each of at least two STAT1 biological replicate ChIP-PCR experiments was chosen as a threshold for enrichment. Target validation was plotted as a function of rank order for each ChIP-chip dataset. As shown in Figure 2A targets at the top of the rank list validated as true positives whereas the frequency of target validation diminishes further down the list of rankings. Thus, most of the first 75 targets are expected to be bona fide targets whereas most of the regions below 100 on the rank list are negative. Extrapolation of the confirmed positives as a function of rank order for the entire list suggests that there are approximately 124 positives in the top 200 targets listed (Table 1). This figure is expected to be an overestimation because many targets lie immediately adjacent to one another and likely represent enrichments from a single common target region in many, if not most, cases. If targets are combined into 10 kb regions then the total number of STAT1 targets is approximately 67 for the ChIP-chip dataset using the arrays with 50 b end-to-end tiling.

We also compared the accuracy of target detection for the 50 b ChIP-chip dataset as a function of signal enrichment. As shown in Figure 3 the fraction of validated positives decreases and the fraction of false positives increases at a very sharp signal enrichment threshold. Thus there is a very sharp transition at a particular signal enrichment

(approximately 0.25 on a log2 scale) above which most targets validate as positives.

We next compared the accuracy of STAT1 targets identified from the 50 b array ChIP-chip dataset to those identified from the 36 b array ChIP-chip dataset. We selected the highest 75 ranked targets from the 50 b array data, corresponding to a false positive rate of 0.26, and cross referenced these with the entire list of 39 targets identified from the 36 b array data. For the 36 b arrays only the top ranked 39 regions had positive signals at a statistically significant cutoff. We suspect this low number is due to diminished signal through nonspecific or cross hybridization with shorter oligonucleotides on the 36 b arrays. The targets of the 50 b and 36 b arrays combined into 84 distinct target regions (see 'Comparison of Target Lists' in Materials and Methods); 18 were common to both lists and most of these (8 of the 11 tested by ChIP-PCR analysis) validated as bona fide targets. The 36 b oligonucleotide array failed to identify 68% (51/75) of the targets detected with the 50 b oligonucleotide array. Of these 51 targets, 27 were tested by ChIP-PCR analysis and the majority of these (20/27) could be validated. In contrast 15 targets were unique to the 36 b array. 7 of these 15 were tested by ChIP-PCR analysis and none showed enrichment. If we restrict analysis of the 36 b array to the top 25 targets, thereby reducing its false positive rate from 0.52 to 0.38, a similar trend is observed (Supplementary Table 2) and fewer targets specific to the 36 b array are identified indicating a greater overlap of the top ranked targets between the 50 b and 36 b lists. In conclusion, based on chromosomal maps of signal enrichments (Figure 1) and target validations, the 50 b arrays outperformed the shorter 36 b arrays under the conditions we employed.

Comparison of Oligonucleotide and PCR Product Arrays

Both PCR product arrays and oligonucleotide arrays are used extensively for ChIP-chip experiments. A recent study analyzing ChIP-chip datasets from yeast revealed that an

oligonucleotide array format provided better sensitivity and resolution than arrays containing PCR products (Borneman et al. unpublished). However, the yeast genome is of much lower complexity and is expected to exhibit much higher signal-to-noise relative to mammalian DNA. Thus it is conceivable, if not likely, that PCR product arrays could perform better than oligonucleotide arrays for experiments using mammalian cells. Six independent biological replica STAT1 ChIP samples were isolated and hybridized to six PCR product arrays (Supplementary Table 1) in the presence of Cot-1 DNA and compared to those targets obtained from a dataset using three biological replicates hybridized with Cot-1 DNA to 50 b oligonucleotide arrays; in many cases the same ChIP samples were used. As shown in Figure 1, the signal enrichments appeared better for the oligonucleotide array dataset relative to the PCR product array dataset.

To determine whether targets identified by oligonucleotide and PCR product arrays overlap, the results from the oligonucleotide arrays and the PCR product arrays were scored using a sliding window scheme and the on-line microarray processing tool ExpressYourself, respectively (see Materials and Methods). The top 75 ranked targets taken from the 50 b oligonucleotide array data and the top 75 ranked targets taken from the PCR product array data were merged to form a union of regions that could be used as the basis for comparing the two ChIP-chip datasets (Table 2 and see 'Comparison of Target Lists' in Materials and Methods). 6 targets overlapped between the 50 b oligonucleotide array and the PCR product array target lists (Table 2, lower panel). The different platforms were compared as a function of their rank order on the target lists. As shown in Figure 4, the positives at the very top of the rank order lists largely agree and less concurrence is observed for targets with lower rankings. If we restrict analysis of the PCR product array dataset to the top 33 targets, thereby reducing its false positive rate from 0.64 to 0.40, a similar trend is observed (Supplementary Table 3).

To ascertain if targets from the PCR product array and the 50 b oligonucleotide array validated at a similar rates, and to determine if the two platforms exhibited similar sensitivities and specificities, targets were selected and tested for validation across a wide range of rank orders using ChIP-PCR analysis. Two or more independent biological replicates were required to show enrichment with each biological replicate assayed separately (See Materials and Methods). As shown in Figure 2B, the frequency of validated targets (i.e. the positive predictive value) from the PCR product array data was diminished relative to the 50 b oligonucleotide array data (Figure 2A) indicating that the PCR product array dataset contains more false positives. In addition the sensitivity of the PCR product array format was lower.

To investigate these differences in array performance, we examined regions that were specific to one of the target lists and that were tested for enrichment by ChIP-PCR (Table 2). 7 targets that were identified by the PCR product array dataset and validated by ChIP-PCR analysis were not present on the target list from the 50 b oligonucleotide array dataset. Inspection of these regions revealed 6 of the 7 targets contained a combination of repetitive elements and AT-rich sequences that likely resulted in low signal enrichments on the oligonucleotide arrays. In contrast, 21 targets identified from the oligonucleotide array dataset and validated by ChIP-PCR analysis were not found using the PCR product arrays. 2 of the 21 were adjacent to positive regions detected by the PCR product arrays, but we could not identify aspects of sequence composition that might cause the other 19 targets to escape detection in the ChIP-chip experiments performed with the PCR product arrays.

The Presence of Competitor Cot-1 DNA in the Hybridization Improves Signal-to-Noise

Highly repetitive sequences comprise 50% of mammalian genomes and can be potential

targets as well as a source of noise. We therefore investigated the value of including unlabeled Cot-1 repetitive DNA in the hybridizations because the addition of Cot-1 DNA might be expected to decrease nonspecific hybridization (DeRisi et al. 1996) and improve the accurate detection of transcription factor targets. To make this comparison, 6 biological replicates were divided after labeling and hybridized on 12 arrays in plus and minus Cot sets, employing 50 b arrays with 38 b spacing (See Materials and Methods). The addition of an excess of Cot-1 DNA produced a modest improvement in signal-to-noise. Figure 5 illustrates this point for a region on chromosome 15 where several peaks and often the overall background were noticeably reduced. The 3 false positives in this region (pink arrows) had high signal enrichments in the experiment lacking Cot-1 DNA, but had low signal enrichments in the experiment containing Cot-1 DNA. Target b in Figure 5 (which lies in a region containing a gene duplication, orange bars) was confirmed by ChIP-PCR analysis. It was ranked 22nd on the target list for STAT1 ChIP DNA hybridized in the absence of Cot-1 DNA and slipped just below the threshold on the ranked target list (to 95th) when the matching sample pairs were hybridized in the presence of Cot-1 DNA.

An inspection of 22 targets specific to the Cot-absent dataset revealed that 13 targets had highly repetitive elements in their regions and 8 targets had segmental duplications. When the same sliding window scoring method was applied to the Cot-absent and Cot-present datasets, a significant number of additional targets was found in the Cot-absent ranked target list (181 targets) relative to the Cot-present ranked target list (3 targets) at the equivalent threshold of 3.5 fold enrichment. Importantly validation of targets revealed a much higher accuracy for the STAT1-associated regions identified in the presence of Cot-1 DNA than in the absence of Cot-1 DNA. Targets specific to either the Cot-present or Cot-absent datasets were sampled from among the top 75 ranked targets identified (Table 3) and tested for enrichment by ChIP-PCR analysis. The experiment containing

Cot-1 DNA detected 15 validated positive regions specific to that dataset at a false positive rate of 0.25, whereas the experiment lacking Cot-1 DNA detected only 2 validated positive regions specific to that dataset at a false positive rate of 0.83 (Table 3). Thus, more accurate results can be obtained through inclusion of Cot-1 DNA in ChIP-chip hybridizations.

The Value of Adding More Biological Replicate Experiments

Researchers typically perform multiple biological replicate experiments for microarray datasets, although a systematic analysis of how replicas improve accuracy and reproducibility of targets has not been previously investigated. We therefore examined the value of performing multiple replicate experiments. The top 50, 100 and 200 targets were taken from the 50 b every 38 b experiment performed in the presence of Cot-1 DNA, using the targets determined in the previous section for 6 biological replicates corresponding to 6 arrays (Supplementary Table 1). As noted above, the top 50 targets have the highest frequency of enrichment in ChIP-PCR validations and those near the bottom of the list (e.g. ranked 150-200) have the lowest frequency of positive validation. The efficiency of target detection from among all targets identified in this Cot-present dataset was determined using a single biological replicate on one array, and then progressively increasing the number of biological replicates, with each replicate hybridized to a separate array. As shown in Figure 6, 50-70% of all targets from the 6-array Cot-present dataset can be identified even with a single array. As expected a higher fraction of the targets are identified using the top 50 target list relative to the top 200 target list since the largest fraction of positive regions resides at the highest rankings as shown in the ChIP-PCR validation studies. The scheme of 3 arrays corresponding to 3 independent biological replicates, which is typical for most published ChIP-chip experiments, identified most (80-86%) of the final targets included in the 6-array dataset.

Comparison of ChIP-Chip to ChIP-PET

DNA sequencing is an alternative readout for detecting immunoprecipitated DNA (Wei et al. 2006; Impey et al. 2004; Kim et al. 2005a; Chen and Sadowski 2005; Roh et al. 2005). In ChIP sequencing, a ChIP-enriched fragment is represented by either a single internal 20 bp tag sequence (ChIP-STAGE) or a 36 bp paired-end-ditag (ChIP-PET in which the ditag is constructed from 18 bp 5′ and 3′ signature sequences extracted from each end of the ChIP DNA fragment, thus demarcating the full length of the sonicated ChIP fragment). The binding sites are then deduced by the frequency with which tags are extracted from ChIP DNA fragments relative to the background expectation. The advantage of using paired-end-ditags over single tags is that the PETs mark the start and end of each ChIP fragment. When PET fragments are mapped to the reference genome (e.g. the human NCBIv35 [hg17] build of the genome sequence), the identity of each individual ChIP fragment can be inferred by the PET mapping location and binding sites can be accurately defined by the common regions within clusters of overlapping PETs. Furthermore, duplicate PET fragments arising from random fragment amplification events during cloning can be easily distinguished and removed by treating these multiple PETs that map to an identical location as a single fragment.

725,877 PETs were sequenced from STAT1 ChIP DNA isolated from IFN-γ induced cells. 66% of the PETs map to unique locations in the genome and represent 327,838 distinct ChIP DNA fragments ranging from 0.1 to 6 kb. Of these unique paired-end-ditags only those PET fragments with 5′ and 3′ ends <6 kb apart were considered. The PET-defined ChIP fragments that overlapped with each other were grouped into clusters: clusters of two overlapping fragments are termed as PET-2, clusters of 3 overlapping fragments as PET-3, and clusters of 3 or more overlapping fragments as PET3+ etc. The frequency of each cluster throughout the ENCODE regions is shown in Table 4. The ENCODE region with the most overlapping fragments lies upstream of IRF1 and is a

PET-33 cluster (Figure 7A). To assess the reliability of each class of PET-clusters as a readout for ChIP enrichment, Monte Carlo simulation was performed to determine the frequency of clusters expected by random chance (Table 4; see Materials and Methods). Based on the frequency of PET clusters generated at random, greater than 46% of PET-3 clusters and greater than 88% of PET4+ clusters are likely to represent bona fide binding targets.

Comparison of signal maps derived from ChIP-chip and Chip-PET data reveals appreciable agreement between the two approaches (Figure 7) and the concurrence is highest for those targets with the highest signal (Table 5). Since the ChIP-PET sequencing experiment inherently covers all of the ENCODE regions, we only considered those 75 PET3+ clusters whose sequence was represented on the 50 b every 50 b array tilepath (Supplementary Table 1 and Materials and Methods) for a true comparison between the two platforms. Of these 75 PET3+ clusters, there were 11 PET5+ clusters (those with the highest enrichment), 9 of which were also identified in the 50 b every 50 b array ChIP-chip dataset (Table 5). For the remaining 64 PET-3 and PET-4 clusters, only 5 overlap the targets lists for the 50 b every 50 b array dataset, giving an overall concurrence of 14 targets (Table 6).

To further investigate the targets that were unique to either the ChIP-ChIP or ChIP-PET target lists, validation experiments were performed. 10 of the targets identified by ChIP-PET3+ cluster regions and missed in the 50 b every 50 b array ChIP-chip dataset were selected for ChIP-PCR validation and shown to be bona fide targets (Table 6). Repetitive DNA elements appeared to obstruct the identification of 6 of these 10 targets in the 50 b every 50 b ChIP-chip data experiments. These repetitive regions had the following characteristics: 1) 4 regions did not have the area of highest PET signal measured on the tiling arrays because highly repetitive elements were centered on the PET overlap spans

and hence these nucleotides were removed from the array tile path. An example of this is shown in Supplementary Figure 1A for a PET-5 target on Chr21. 2) Repetitive regions were similarly noted for 2 target regions where a combination of AT-rich and RepeatMasked sequences were congruent with the PET overlap spans. This case includes the other validated PET-5 cluster (Supplementary Figure 1B). The remaining 4 PET3+ targets not detected by the 50 b array were missed for no apparent reasons.

Investigation of the 15 confirmed targets that were detected in the 50 b every 50 b array ChIP-chip dataset but that were not on the PET3+ list (Table 6) revealed that 7 resided nearby a ChIP-PET target but were on the shoulder relative to the site of maximal signal. 5 of the 7 targets corresponded to the IRF1 locus, which has one of the strongest signals in the genome (Figure 7A). Thus these array targets correspond to a single common target region. 4 of the remaining 8 ChIP-chip targets from the 50 b every 50 b array dataset intersected PET-2 clusters; we presume increased sequencing depth would have detected these STAT1 binding regions.

We also inspected those regions that did not show enrichment by ChIP-PCR analysis (11 negatives specific to the 50 b every 50 b array dataset and 5 negatives specific to the ChIP-PET experiment, Table 6) to ascertain what sequence features might contribute to the identification of these targets as false positives. Of the 11 false positives from the 50 b every 50 b array ChIP-chip dataset, 6 are either largely or entirely comprised of simple repeats, one additional target region occurs as a segmental duplication, another lies near a strong target in the IRF1 5′ noncoding region and no unusual features that may be uniquely attributable to ChIP-chip performance could be established for the other 3. All 5 ChIP-PETs that were not enriched in ChIP-PCR validation experiments (Table 6) were PET-3 clusters. As indicated by the Monte Carlo simulation (Table 4), approximately 50% of PET-3 clusters are expected to be false positives arising from random

background. Another possible explanation for the ChIP-PET false positives could be nearby repetitive genomic regions that lead to mapping artifacts. One of the 5 ChIP-PET false positives does reside in a repetitive region and may have been misassigned during mapping to the hg17 reference sequence. In another example (shown in Supplementary Figure 2), the false positive in the region chr5:131963298-131964597 [hg17] was initially called a PET-3, although of the 3 PETs that overlapped, 2 of the DNA fragments in this cluster have an almost identical mapping (within 2 bp from each end) and were likely derived from the same ChIP fragment. In summary, these results indicate that ChIP-chip and ChIP-PET exhibit considerable agreement, particularly on the strongest targets. Each approach is capable of identifying validated targets not found by the other technique.

Discussion

The combination of sequenced genomes and ChIP-based technologies has inspired progress for the comprehensive detection of mapping transcription factor binding regions *in vivo*. While most efforts have focused on ChIP-chip strategies, ChIP-sequencing is gaining popularity as a parallel method. In this manuscript we performed STAT1 chromatin immunoprecipitations from IFN-γ stimulated cells and used the resulting ChIP DNA to map STAT1 binding regions by both microarray hybridizations and DNA-sequencing. Furthermore we investigated parameters for executing robust ChIP-chip experiments and devised methods for comparing results across array and sequencing platforms. Based on the outcome of these studies, we determined that reliable ChIP-chip results can be obtained using maskless high density arrays containing longer rather than shorter oligonucleotides and also by including Cot-1 DNA as a competitor to improve hybridization accuracy. In cross referencing STAT1 targets obtained by ChIP-chip with those detected by ChIP-PET we found regions that overlapped between ChIP-chip and ChIP-PET, as well as enriched regions specific to only one of these methods. Thus the sequencing of ChIP DNA fragments is shown to be a valuable and alternative strategy for

target identification.

The ChIP-chip conditions applied here for STAT1 can be extended to other transcription factors or DNA-interacting proteins that are constitutively present in the nucleus. In these experiments the hybridization reference samples are either total genomic DNA or ChIP DNA prepared using normal serum. Examples of other factors we have analyzed by ChIP-chip on 50 b maskless ENCODE tiling arrays include the chromatin remodeling proteins BAF155 and BAF170, as well as the transcription factor c-Jun; the binding profiles all three of these proteins are part of the ENCODE meta-analyses and their tracks are available in the UCSC browser (The ENCODE Consortium 2006). As with STAT1 we labeled unamplified ChIP samples of BAF155, BAF170 and c-Jun in order to avoid possible biases that may arise during PCR or other amplification methods, and these unamplified ChIP samples exhibited good signal enrichments in our hybridizations. ChIP-chip experiments using unamplified samples have been performed by others (Weinmann and Farnham 2002; Kondo et al. 2004; Testa et al. 2005; Rada-Iglesias et al. 2005).

For the maskless array platforms, longer oligonucleotides most likely improve performance because more specific hybridization signals can be obtained relative to array designs with shorter oligonucleotides. Increased specificity through reduced cross hybridization and potentially stronger signals can be achieved with longer oligonucleotides. This in turn should lead to more accurate measurements and thus more accurate ratios of immunoprecipitated DNA relative to control DNA. Extending this logic, PCR product arrays have even longer DNA fragments as array elements and in theory should provide superior results to oligonucleotide arrays. This is not the case, probably for several reasons. First, multiple probes on high density oligonucleotide arrays allow for a number of independent measurements across a region of interest. If

any individual probe performs poorly (for example due to secondary structure, cross hybridization or AT-rich regions) then sampling over multiple probes using a sliding widow approach (see Materials and Methods) can still provide useful signals. Indeed, we have found that signals generated by one or a few oligonucleotides are not usually trustworthy. Second, PCR product arrays probably have many fewer copies of a region of interest attached to the arrays which may reduce signal and in turn should affect the accuracy of the measurements. Third, repetitive sequences on PCR product arrays may reduce signal-to-noise ratios. Finally, a small fraction of PCR products (5-10%) amplify from regions other than those intended (Rinn et al. 2003). This will lead to misassignment and the targets will not be validated.

A recent comparison of tiling oligonucleotide array platforms for mapping transcribed regions revealed that high density 25 b arrays outperformed longer oligonucleotide arrays that were tiled at a lower density (Emanuelsson et al. 2006). The ChIP-chip study presented here reveals better performance for the longer oligonucleotide arrays when the arrays have similar densities. The better performance of the 25 b array format for mapping transcribed regions relative to the longer oligonucleotide arrays is more directly attributable to the higher density rather than oligonucleotide length.

Our validation strategy involved analyzing regions sampled across a range of targets ranked by signal enrichments, as well as analyzing sequences that were ranked below the applied target threshold. This approach allows for a reasonable estimate of the accuracy of the results. By extending the validation frequency as a function of rank, we can extrapolate and determine the sensitivity of the experiment at a particular threshold. It should be noted however, that positives which are unable to be detected by a specific protocol cannot be assessed for sensitivity using this validation method. Nonetheless, this strategy is expected to provide the best approach available for determining these

measurements.

Our study reveals that ChIP-chip and ChIP-PET generally yield similar results, particularly for the strongest signals. However, targets that are uniquely identified by one of these technologies are also captured and many of these targets could be validated as positives by ChIP-PCR analysis. Targets exclusive to either ChIP-chip or ChIP-PET fall into several classes: 1) Many unique targets arise from the manner in which positives are scored. Current ChIP-chip scoring methods merge stretches of probes showing signal enrichments into short windows (we used 1.3 kb) and thus adjacent segments are often part of a single larger target region (> 1.3 kb), whereas ChIP-PET clusters were connected if the PETs share 1 bp overlap with no restriction on the length of each cluster region. Grouping adjacent ChIP-chip targets will alleviate this problem, particularly for highly enriched segments where these incidents occur most frequently. 2) Other targets solely identified by one platform can often be attributed to neighboring repetitive sequences. RepeatMasked sequences are eliminated during the array design process in the ChIP-chip experiments. Consequently targets that lie within or immediately adjacent to genomic repeats are more likely to be missed; real binding sites across short repeat regions may be missed due to the array design but detected by PET sequencing. Conversely, repetitive regions may also lead to false positives through cross hybridization to real targets (in ChIP-chip) or by the occasional misassignment of a tag containing repetitive DNA elements (in ChIP-PET). 3) Simple repeats can also lead to problems for reasons that are not understood. Since both ChIP-chip and ChIP-PET identify unique, validated targets, the use of these technologies in an integrated fashion is anticipated to produce optimal sensitivity and specificity for detecting binding targets.

In principle, ChIP-chip should be more sensitive than ChIP-PET. At a positive predictive value >95%, a computer simulation of targets reveals that 10 million tag reads are

necessary to identify 55% of four-fold enriched targets (unpublished, a more thorough analysis will appear in a subsequent paper). ChIP-chip is capable of detecting two-fold enrichments. One main advantage of ChIP-PET is that it can potentially identify targets that lie in repetitive, but not identical regions; these may be able to be distinguished by DNA sequencing but not by array methods due to cross hybridization effects. However, if a very small unique target region is flanked by large adjacent repeats, these binding regions may be difficult to detect using ChIP-PET since the ends of immunoprecipitated fragment may reside in repetitive DNA. ChIP-chip should give direct and accurate readout of immunoprecipitation efficiencies. Interestingly, enrichment signals derived from microarray experiments tend to be significantly depressed relative to values confirmed by quantitative ChIP-PCR analysis, a phenomenon observed previously (Bernstein et al. 2005).

Our studies suggest that a number of design parameters can be modified to enhance the performances of ChIP-chip and ChIP-PET. For ChIP-chip future generations of array design may incorporate the following improvements: 1) It should be possible to more accurately retrieve targets that lie next to repetitive sequences by increasing the number of oligonucleotide tiles adjacent to repeats. This increased density should facilitate their detection. 2) The judicious choice of nonidentical oligonucleotides should improve array performance. 3) Finally the use of isothermal arrays, where the oligonucleotides on the array vary in length to give a more uniform annealing temperature, should improve performance (Urban et al., 2006). For ChIP-PET, slight modifications to the mapping algorithm should eliminate those few instances where nearly identical ChIP fragments were double counted in determining the ChIP-PET cluster number (see example in Supplementary Figure 2).

Another desirable feature of ChIP-PET is that it is inherently whole genome and can

theoretically find all targets present in genomic sequences. Currently both ChIP-PET and whole genome ChIP-chip are expensive due to the considerable cost of high throughput sequencing and whole genome oligonucleotide arrays. However, both of these technologies are expected to exhibit dramatic decreases in cost in the near future as new sequencing technologies become available (Margulies et al 2005; Shendure et al 2005; Service 2006) and as array densities continue to increase. Thus, both ChIP-chip and ChIP-sequencing technologies will become substantially more cost-effective and their mutual combination would maximize accuracy.

In summary, insights garnered from our experiments should be useful for robust mapping of transcription factor binding sites, both by implementing methods and reagents that currently are available as well as applying the findings of our results to improve future generations of ChIP-based technologies. Given the ever increasing number of transcription factor mapping experiments that are performed (Bieda et al. 2006; Scacheri et al. 2006; Beima et al. 2006; Loh et al. 2006; Lee et al. 2006) the proceeds reported in this work are expected to be invaluable for the large number of studies that are anticipated.

Materials and Methods

STAT1 chromatin immunoprecipitations

HeLaS3 cells (ATCC) were cultured in suspension in spinner flasks in S-MEM (GIBCO-Invitrogen #11380-037) supplemented with 2 mM L-glutamine (GIBCO-Invitrogen #25030-081), 10% fetal bovine serum (GIBCO-Invitrogen #16140-071), and antibiotics (Antibiotic-Antimycotic, GIBCO-Invitrogen #15240-062) at 37°C and 5% CO_2. ChIP samples for a given biological replicate were prepared from distinct cell cultures grown, harvested and processed on separate days from all other biological replicates. For each biological replicate we grew on the order of 12 x10^8 cells which were split into IFN-γ treated and untreated halves for STAT1 ChIPs. These sample preparations yielded enough DNA to be distributed across many of the platforms and ChIP-PCR validations. STAT1 ChIP samples were prepared from IFN-γ stimulated HeLaS3 cells and ChIP DNA quality was verified as previously described (Hartman et al. 2005). Briefly HeLaS3 cultures were divided in half and were either induced with 5 ng/ml human recombinant IFN-γ (R&D Systems #285-IF), or left untreated, for 30 min at 37°C, 5% CO_2 and then fixed with 1% formaldehyde final concentration at room temperature for 10 min. Fixations were quenched by addition of glycine to 125 mM final concentration (from 2 M glycine stock in 1 x PBS) and cells were washed twice in cold 1 x Dulbecco's PBS (GIBCO-Invitrogen #14190-144). Cells were swelled for 10 min 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 Roche protease inhibitors #11-697-498-001) and lysed by dounce homogenization (using pestle B). Nuclear pellets were collected and lysed in 1 x 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 Roche protease inhibitors). Nuclear lysates were sonicated with a Branson 250 Sonifier (Output 20%, 100% duty cycle) to shear the chromatin to approximately 1 kb in size. Clarified lysates were incubated overnight at 4°C with anti-

STAT1 alpha p91 (C-24) rabbit polyclonal antibody (Santa Cruz Biotechnology #sc-345). Protein-DNA complexes were precipitated with RIPA-equilibrated protein A agarose beads (Upstate #16-156) and immunoprecipitates were washed three times in 1 x RIPA, once in 1 x PBS, and then eluted from the beads by addition of 1% SDS, 1 x TE (10 mM Tris-Cl at pH 7.6, 1 mM EDTA at pH 8), and incubation for 10 min at 65°C. Crosslinks were reversed overnight at 65°C. All samples were purified by treatment first with 200 μg/ml RNase A (Qiagen #19101) for 1 h at 37°C, then with 200 μg/ml Proteinase K (Ambion #2548) for 2 h at 45°C, followed by extraction with phenol:chloroform:isoamyl alcohol and precipitation at -70°C with 0.1 volume of 3 M sodium acetate, 2 volumes of 100% ethanol and 1.5 μl of pellet paint co-precipitant (Novagen #69049-3). ChIP DNA prepared from 1 x 10^8 cells was resuspended in 50 μl of ultrapure water (GIBCO-Invitrogen #10977-015).

BAF155, BAF170 and c-Jun chromatin immunoprecipitations

ChIP DNA samples for BAF155, BAF170 and c-Jun were prepared and hybridized according to protocols nearly identical to those used for the STAT1 ChIP samples. Antibodies used were anti-BAF155 (H-76) rabbit polyclonal antibody (Santa Cruz Biotechnology #sc-10756), anti-BAF170 (H-116) rabbit polyclonal antibody (Santa Cruz Biotechnology #sc-10757) and anti-c-Jun rabbbit polyclonal antiserum (Upstate #06-225). Reference DNA samples were ChIP DNA prepared using normal rabbit IgG (Santa Cruz Biotechnology #sc-2027) for the BAF155 and BAF170 experiments and total genomic DNA for the c-Jun experiment. Biological replicates were prepared from distinct cell cultures, were labeled separately and were hybridized without pooling as one biological replicate per array. Datasets were created using maskless arrays (Nuwaysir et al. 2002) with 50 b every 38 b spacing (NimbleGen Systems of Iceland, LLC). ChIP DNA was directly labeled (per manufacturer's protocol) by Klenow random priming with Cy5 nonamers (ChIP DNA prepared using protein specific antibody) or Cy3 nonamers

(total genomic DNA or ChIP DNA prepared using normal rabbit IgG). Full experimental details and microarray data series are available at the Gene Expression Omnibus <http://www.ncbi.nlm.nih.gov/geo/> with the following accession numbers: BAF155 {GSM1496, GSM1497, GSM1498, GSM1499}; BAF170 {GSM1500, GSM1501, GSM1502, GSM1503}; and c-Jun {GSM78080, GSM78081, GSM78082}. Ranked target lists are available at <http://encode.gersteinlab.org/data/Euskirchen_etal/>.

ChIP sample preparation and labeling

Biological replicates are defined as STAT1 ChIP DNA prepared from distinct cell cultures grown, harvested and processed on separate days. ChIP DNA samples from individual biological replicates were labeled separately and hybridized separately (without pooling) as one biological replicate per array (Supplementary Table 1). In many cases the same biological replicates were hybridized to each of the array platforms. For the experiment comparing hybridizations in the presence and absence of Cot-1 DNA, 6 biological replicates were divided after labeling and hybridized over 12 arrays in plus and minus Cot sets.

For PCR product arrays (gift of Bing Ren, UCSD) and maskless arrays with 50 b every 50 b and 36 b every 36 b spacings (both oligo length arrays manufactured by NASA Ames Research Center), ChIP DNA from 1 x 10^8 cells was random primed with Klenow (enzyme and primers from BioPrime DNA Labeling System, Invitrogen #18094-011) and Aminoallyl-dUTP (Sigma #A0410) was incorporated. Next Alexa Fluor dyes (Invitrogen #A32755; Alexa647 for ChIP DNA isolated from IFN γ-stimulated cells and Alexa555 for ChIP DNA isolated from unstimulated cells) were coupled to the Aminoallyl-dUTP. Coupling reactions were terminated with hydroxylamine. Alexa555- and Alexa647-coupled ChIP DNA samples were combined and recovered using a CyScribe GFX Purification Kit (Amersham #27-9606-02) according to the manufacturer's protocol. The

recovered probe was further purified by ethanol precipitation with 0.1 volume of 3M sodium acetate (pH 5.2).

During the course of our studies we tested a number of different labeling technologies including the MICROMAX tyramide signal amplification method (NEN Life Science Products), 3DNA dendrimer technology (Genisphere) and anti-biotin and anti-fluorescein coated Resonance Light Scattering (RLS) particles (Genicon Sciences). These were tested primarily using PCR product arrays. For detection of STAT1 targets, the labeling methods reported here were the most consistently positive in terms of signal, array uniformity, reproducibility, time efficiency and cost effectiveness.

For maskless arrays (Nuwaysir et al. 2002) with 50 b every 38 b spacing (NimbleGen Systems of Iceland, LLC) ChIP DNA from 1 x 10^8 cells was directly labeled (per manufacturer's protocol) by Klenow random priming with Cy5 nonamers (ChIP DNA isolated from IFN γ-stimulated cells) or Cy3 nonamers (ChIP DNA isolated from unstimulated cells).

Microarray hybridizations

All arrays were hybridized with mixing in MAUI hybridization stations from BioMicro Systems (Salt Lake City, UT) for 16–18 h at 42°C. Before deciding on the hybridization protocols described below we tested a number of experimental parameters. The oligonucleotide arrays were hybridized at temperatures ranging from 14 to 33°C below their estimated melting temperatures (T_m), using the formula described in (Sambrook et al. 1989) and assuming 44% GC content for the ENCODE tiling arrays. Hybridization buffers varied from 0.825 to 1.0 M [Na^+] and from 0–40% formamide, final concentrations. Note that optimal hybridizations are performed at ~25°C below the

estimated melting temperatures although hybridization rates are only modestly affected by conditions 15–30°C below the T_m (Wetmur and Davidson 1968).

PCR product arrays were prehybridized in 5x SSC/ 25% formamide/ 0.05% SDS/1% BSA for 1 h at 42°C. Labeled ChIP DNA was precipitated and resuspended in 60 μl of 5x SSC/ 25% formamide/ 0.05% SDS with 5 μg of human Cot-1 DNA (Invitrogen #15279-011) per array. The PCR product arrays were washed in 42°C 2x SSC/0.1% SDS, room temperature 0.1x SSC/0.1% SDS, and 0.1x SSC.

Labeled ChIP DNA for maskless arrays (Nuwaysir et al. 2002) with 50 b every 50 b and 36 b every 36 b spacings (both oligo length arrays manufactured by NASA Ames Research Center) was precipitated with 30 μg of human Cot-1 DNA (Invitrogen #15279-011) per array and pellets were resuspended in 45 μl of hybridization buffer (final concentrations: 40% formamide, 5x SSC, 0.1% SDS, and 0.2x TE). Arrays were washed once with 42°C 0.2% SDS/0.2x SSC, once with room temperature NSWB (6x SSPE, 0.01% Tween-20, 1 mM DTT), twice with 0.2x SSC, and twice with 0.05x SSC.

For maskless arrays (Nuwaysir et al. 2002) with 50 b every 38 b spacing (NimbleGen Systems of Iceland, LLC) labeled ChIP DNA was hybridized in buffer containing 20% formamide, 1.2 M Betaine, and 0.1 μg/μl herring sperm DNA per manufacturer's protocol. The plus Cot-1 experiments included 10 μg of human Cot-1 DNA (Invitrogen #15279-011) per array. Arrays were washed in 42°C 0.2% SDS/0.2x SSC, room temperature 0.2x SSC, and 0.05x SSC.

ChIP-PET experiment

The STAT1 ChIP-PET library was constructed as previously described (Wei et al. 2006). Briefly, the ChIP enriched DNA fragments were cloned into the cloning vector pGIS3 to

generate the ChIP DNA library. Purified plasmid from the ChIP DNA library was digested with MmeI to release the internal fragments and a signature tag from each terminal of the original ChIP DNA insert were self-ligated to form a 'single-ditag library'. 50 bp paired-end-ditags (PETs) were released by BamHI, PAGE-purified and concatenated to clone into pZErO-1 to form the final ChIP-PET library for sequencing.

PET sequences were extracted from the raw reads and mapped to human genome sequence assembly [hg17]. The process of PET extraction and mapping is essentially the same as previously described for cDNA analysis (Ng et al. 2005). The mapping criteria are that both the 5′ and 3′ signatures must have a minimal 17 bp match, be present on the same chromosome and same strand, in the correct orientation (5′→3′), and within 6 kb of genomic distance.

Mapping simulation of overlapping PET clusters

A Monte Carlo simulation was performed to assess the background level of overlapping PET sequences when mapped to the genome. In the simulation, we first randomly selected 4007 unique genomic DNA segments from 44 ENCODE regions (similar to the average fragment size from STAT1 ChIP DNA) and then determined how many fragments overlapped with others. This process was repeated 10,000 times to compute the percentage of randomly selected DNA fragments that overlapped. The results are summarized in Table 4. Based on this simulation, we estimated that 463 PETs (97% of total) would result in two overlapping PETs (PET-2), 47 in PET-3, 3 in PET-4, and so forth due to random chance. In contrast, the numbers of experimentally generated overlapping PETs are significantly higher than the estimated background. Therefore, it is highly likely that overlapping PETs resulted from the immunoprecipitation of STAT1-associated DNA rather than from random events.

STAT1 Target Validations

Primers were designed to amplify 200-350 bp fragments from regions throughout the rank ordered target lists as well as regions where array signals were below cut-off values. ChIP DNA from either 4 x10^6 IFN γ-stimulated or unstimulated cells was amplified in 40 μl reactions with 1 μM of target specific primer pairs and 1 x Qiagen Master Mix (Qiagen # 201203). For each primer pair parallel reactions were run with 0.2 μg HeLaS3 genomic DNA to ensure that a sample set would yield a single band of the expected size. Some primer pairs required addition of PCR additives, either Betaine or Qiagen Q solution at varying concentrations. Cycling conditions were as follows: 5 min at 94°C, 29 cycles of 30 sec at 94°C, 30 sec at 52°C, 30 sec at 72°C, and a final extension period of 10 min at 72°C. The entire completed PCR reactions were loaded on 1.5% agarose gels and only those primer sets in which entire sample volumes were loaded were analyzed further. Each plate of PCR reactions included positive and negative controls, and all reactions from a plate were loaded on the same gel. Densitometric analyses were made using ImageJ software <http://rsb.info.nih.gov/ij/>. For each primer pair, enrichments were calculated for yield from IFN γ-stimulated cells relative to yield from unstimulated cells. To qualify as a validated region, enrichments had to be consistently greater than 2-fold from each of two or more biological replicates. In many cases more than two biological replicates were tested and for some regions validation results were quantified from multiple primer pairs (in separate reactions) to eliminate any primer artifacts. In total 280 regions were tested for validation. Primer sequences used in the ChIP-PCR assays are available at <http://encode.gersteinlab.org/data/Euskirchen_etal/>.

Design of Genomic Tiling Microarrays

All tiling arrays were designed using the sequence from the ENCODE regions based on human genome build NCBIv34 [hg16]. For analysis coordinates from all array designs were remapped on to human genome build NCBIv35 [hg17] using liftOver from the

UCSC Genome Browser (Hinrichs et al. 2006). The 50 b every 50 b tiling array was custom designed with 192,040 50 b oligoneucleotides tiling the forward strand approximately every 50 b (end-to-end) across the following ENCODE regions; ENm001-014, ENr114, ENr132, ENr233, ENr321, ENr331 and ENr333. The 36 b every 36 b tiling array was also custom designed using 382,454 36 b oligonucleotides tiling one strand approximately every 36 b (end-to-end) across all the ENCODE regions excluding ENr112, ENr121, ENr131, ENr211, ENr222, ENr313, ENr324, ENr334. Both of these arrays were designed using the tiling array design tool http://tiling.gersteinlab.org (Bertone et al. 2006). The 50 b every 38 b array uses 382,885 50 b oligonucleotides to tile the forward strand of all ENCODE regions with average overlap between oligonucleotides of 38 b. The PCR product array (supplied by Bing Ren, UCSD) uses 24,341 PCR amplicons of average size 620 bp to tile all of the nonrepetitive ENCODE regions. Relative coverage of the ENCODE regions for each of the array formats is shown in Supplementary Table 1.

Analysis of Microarray Data

For each hybridization the files (in .pair file format) to the two channels corresponding to the ChIP DNA and reference DNA, were uploaded to the TileScope pipeline for high-density tiling array data analysis <http://tilescope.gersteinlab.org> for normalization and scoring (Zhang et al. submitted 2006). The pipeline first performs intra- and inter-slide scaling (between biological replicates) using quantile normalization, the results of which are then integrated using a sliding window approach (a window of size 1000 bp in genomic space was used to integrate neighboring probes from replicate arrays). For each window centered at the genomic coordinate of each oligonucleotide probe, the pseudomedian signal (median of pairwise averages of the $\log_2$ ratio of test to reference signals for all oligonucleotide probes within the window), as well as a p-value measuring the likelihood that the region is bound by the transcription factor (using a Wilcoxon

paired signed rank test comparing test signal against reference signal for all oligonucleotide probes in the window) are computed. In an iterative fashion regions with the highest signal (and p-values less than 10^{-4}) were selected. In order to ensure that the shoulders of other target regions are not identified as distinct binding regions we required that the centers of target regions be spaced at least 1300 bp apart. This procedure generated a ranked list of non-overlapping target regions of size 1300bp. For each set of arrays the top 200 regions were scored in this fashion where possible. (For the 36 b array dataset we were only able to extend the rank list to the top 39 targets, beyond which the regions did not show enriched signal at the statistically significant cutoff used). The data from the PCR product arrays were analyzed with the on-line microarray processing tool ExpressYourself (Luscombe et al. 2003). This microarray data series is available at the Gene Expression Omnibus <http://www.ncbi.nlm.nih.gov/geo/> with accession number GSE2714. The ranked target lists are available at both <http://dart.gersteinlab.org/> as well as at <http://encode.gersteinlab.org/data/Euskirchen_etal/>.

Comparison of Target Lists

As described above target lists are rank lists of non-overlapping target regions of uniform size 1300 bp. In order to fairly compare the ChIP-chip data against the ChIP-PET experiment, the ChIP-PET targets were likewise converted into 1300 bp regions centered on the ChIP-PET cluster. Also, comparisons were done for targets identified in regions common to both platforms because the 50 b every 50 b array does not tile all of the ENCODE regions (Supplementary Table 1). When computing the overlap between any two lists of regions (whether the data are from ChIP-chip or ChIP-PET), the number of entries in the first list intersecting the second is not necessarily the same as the number of the second list intersecting the first (this discrepancy typically happens in loci where multiple target sites are located in a short genomic span). In order to avoid this ambiguity we chose to first merge the two lists under comparison to form a list of union regions

comprising the union of targets from both lists. Then using the union set of regions as a basis, one can compute the number of regions belonging to only one of the two original lists, or union regions that came from both lists. One important note is that some union target regions occurring in more complicated loci tend to be longer and might only contribute one joint region to the counts of number of union regions shared by both lists, even though the region might correspond to multiple entries on each of the original two lists. Regions that have been tested for validation can also be compared against these union target regions to assess validation rates for union regions that were detected on only one of the two lists or by both datasets. This is how the data in Tables 1, 2, 3, 6, S2 and S3, were generated.

Acknowledgements

The authors are very grateful to Bing Ren (UCSD) for sharing the PCR product arrays. We thank Janine Mok and Mike Hudson for critical reading of the manuscript. This work was funded by NIH ENCODE grants.

Figure Captions

Figure 1

Comparison of signal tracks. Signal enrichment tracks are plotted for the 50 b, 36 b and PCR product array platforms for two different loci. Signals of STAT1 bound regions in IFN-γ-stimulated cells relative to untreated cells were derived from multiple biological replicates with one replicate hybridized per array (Materials and Methods and Supplementary Table 1). Annotations above the coordinate axis are for genes on the forward strand and those below are for reverse strand genes Signal enrichment tracks are plotted to the same scale for the platforms displayed, from 0 to 2.5 in panel A and from 0 to 1.3 in panel B. **A**. The IRF1 locus and flanking regions on chromosome 5 (coordinates 131,770,000 to 131,870,000 from build NCBIv35 [hg17]). **B**. The loci on chromosome 21, which contain the cytokine receptors, IFNAR2, IL10RB and IFNAR1 (coordinates 33,500,000 to 33,680,000). **C**. Targets which have been validated by ChIP-PCR (shown) are indicated by symbols a through f. The lanes are labeled for ChIP DNA from IFN-γ-stimulated cells, ChIP DNA from unstimulated cells and for HeLaS3 genomic DNA. Fold enrichments, as calculated for several biological replicates (see Materials and Methods) are indicated for each target (a through f) and for a negative control region (–).

Figure 2

Validation frequency as a function of rank order for the 50 b every 50 b array and PCR product array platforms. For each ChIP-chip dataset (derived from multiple biological replicates with one replicate hybridized per array, Supplementary Table 1) we identified the target regions above a threshold. The targets were tested and divided into true positives (TP) and false positives (FP) based on a ChIP-PCR validation assay (as

described in Materials and Methods). Sensitivity (defined as TP/(TP+FN) where FN are the number of false negatives) and specificity (defined as TN/(TN+FP) where TN are the number of true negatives) of the target list at this threshold are difficult to accurately estimate since the total number of actual binding sites (TP+FN) as well as the number of true negatives (TN) are not known, and other methods for direct, *in vivo* identification binding independent of ChIP methods do not exist. Nonetheless, the positive predictive value (defined as TP/(TP+FP)) for ChIP-chip experiments can be estimated using data from ChIP-PCR validations as described in the Materials and Methods. **A.** The number of targets confirmed by validation, true positives (green line) as well as the number targets that did not validate, false positives (red line) is plotted as a function of target rank (ordered by signal enrichment) for the 50 b every 50 b array platform (from 3 biological replicate arrays). The positive predictive value (blue line) is also shown as a function of rank. **B.** The number of true positives (green line), the number of false positives (red line) and the positive predictive value (blue line) are shown as a function of rank for the PCR product array platform (data from 6 biological replicate arrays).

Figure 3

Validation frequency as a function of signal for the 50 b every 50 b array dataset. The data from Figure 2A were analyzed as a function of array signal enrichment for the 50 b every 50 b array platform. Signal enrichment is defined as the $\log_2$ ratio of signal intensity of the ChIP DNA over the signal intensity of the reference DNA sample (for STAT1 this is the $\log_2$ ratio of intensities for IFN-γ stimulated against unstimulated ChIP DNA samples). A target region is identified as a 'peak' in a signal enrichment track (see Figure 1; for details see Materials and Methods) and is assigned its maximal signal enrichment, the height of the peak. The number of true positives is the green line, the number of false positives is the red line and positive predictive value is the blue line as in Figure 2A. The horizontal scale in this figure is in the opposite orientation to the

horizontal scale displayed in Figure 2; high signal enrichment which appears to the right-hand side in Figure 3, corresponds to higher rank which is to the left in Figure 2. At a $\log_2$ signal of approximately 0.25 the number of false positives increases sharply to the left.

Figure 4

Agreement between the ranked target lists for the 50 b every 50 b array and the PCR product array platforms. Each dataset is comprised of multiple biological replicates with one replicate hybridized per array (Supplementary Table 1). The vertical axis is the number of targets common between the two rank lists up to a certain rank (the horizontal axis). The agreement increases steeply for the highest ranked targets and then starts to plateau.

Figure 5

The effect of Cot-1 DNA in determining STAT1 targets. Signal enrichment tracks are shown for datasets of paired samples (see Materials and Methods and Supplementary Table 1) that were hybridized either in the presence of Cot-1 DNA (top track) and or in the absence of Cot-1 DNA (lower track), both on the 50 b every 38 b array platform. Annotations above the coordinate axis are for genes on the forward strand, and annotations below the coordinate axis are for genes on the reverse strand. Signal enrichment tracks are plotted to the same scale from 0 to 3.3 for the Cot-present and Cot-absent datasets. Targets with labels 'a' and 'b' are identified by both experiments (ChIP-PCR gel images are shown). Three targets (pink arrows) appeared only in the Cot-absent experiment and were identified as false positives by ChIP-PCR validation (gel images not shown). The orange bars indicate a region of segmental duplication, which is a potential cause of the false positives (due to cross hybridization with confirmed target 'b'). For the ChIP-PCR validations displayed, the lanes are labeled for ChIP DNA from IFN-γ

stimulated cells, ChIP DNA from unstimulated cells and for HeLaS3 genomic DNA. The fold enrichments are indicated and were calculated for several biological replicates (see Materials and Methods).

Figure 6

The value of adding biological replicates to a ChIP-chip dataset. For the six 50 b every 38 b arrays that were hybridized in the presence of Cot DNA the reproducibility of target lists for the top 50 (green), 100 (red) and 200 (blue) binding regions was examined as a function of the number of biological replicates analyzed. Each biological replicate is hybridized to a separate array. The agreement is compared against the target list identified by using all six arrays. We see that greater than 80% agreement is obtained when three or more biological replicates are used.

Figure 7

Comparison of ChIP-chip and ChIP-PET. Signal tracks for the ChIP-chip dataset from the 50 b every 50 b platform are shown and compared to ChIP-PET signals (the vertical axis corresponds to the number of overlapping ditags at a given genomic coordinate). **A.** The IRF1 locus on chromosome 5 (coordinates 131,842,000 to 131,865,000 from build NCBIv35 [hg17]). The orange arrows indicate the validated ChIP-chip targets from the 50 b array experiment that were on the shoulders of ChIP-PET clusters in the IRF1 region. **B.** The region on chromosome 21 (coordinates 33,500,000 to 33,700,000) containing the cytokine receptors IFNAR2, IL10RB and IFNAR1. Significant concurrence is observed between the signal readouts from each methods.

Table Legends

Table 1

Comparison of ranked target lists for the top 75 targets from the 50 b every 50 b array dataset with the top 39 targets from the 36 b every 36 b array dataset. The upper panel displays the false positive rates (FPR) calculated for each list considered separately. The lower panel displays the results after merging the list of the top 75 targets from the 50 b arrays with the list of the top 39 targets from the 36 b arrays. The comparison is performed (see Materials and Methods for full details) by first creating the union of the two separate lists and then counting the number of union target regions specific to either the 50 b array dataset or the 36 b array dataset, or those targets identified by both platforms. In each of these three categories the union regions that were tested for validation are displayed as well as the associated false positive rate (FPR). Supplementary Table 2 is a similar comparison between the 50 b and 36 b array data with a more restrictive list of targets (top 25 targets) from the 36 b array (with a lower false positive rate). The FPR is defined as the TN/(TN+TP) where TN are the true negatives and TP are the true positives from the STAT1 ChIP-PCR analysis for target validations. Note that the FPR plus the PPV (positive predictive value as discussed in Figure 2) sum to 1.

Table 2

Comparison of ranked target lists for the top 75 targets from the 50 b every 50 b array dataset with the top 75 targets from the PCR product array dataset. The upper panel displays the false positive rates (FPR) calculated for each list considered separately. The lower panel displays the results after merging the list of the top 75 targets from the 50 b array dataset with the list of the top 75 targets from the PCR product array dataset. As for Table 1, this comparison is performed, by first creating the union of the two separate lists and then counting the number of union target regions specific to either the 50 b arrays or the PCR product arrays or those targets identified by both platforms. In each of these three categories the union regions that were tested for validation are displayed as well as

the associated false positive rate (FPR). Supplementary Table 3 is a similar comparison between the 50 b and PCR arrays with a more restrictive list of targets (top 33) from the PCR product array (with a lower false positive rate).

Table 3

Comparison of ranked target lists for paired samples hybridized either in the presence or absence of Cot-1 DNA on the 50 b every 38 b array platform. Datasets were generated from 6 biological replicates that were split post labeling and hybridized in parallel in plus and minus Cot-1 DNA sets on 12 arrays using the 50 b every 38 b array platform. The top 75 target regions were then identified for both datasets and compared. The upper panel displays the false positive rates (FPR) calculated for each list considered separately. The lower panel displays the results after merging the target lists of the top 75 ranked regions taken from each dataset. As for Table 1 this comparison is performed, by creating the union of the two separate lists and counting the number of union target regions specific to the Cot-present list, specific to the Cot-absent list, or those targets identified by both. In each of these three categories the union regions that were tested for validation are displayed as well as the associated false positive rate.

Table 4

Monte Carlo simulation of the expected number of PET clusters from the ENCODE regions as a function of the PET cluster size. For overlapping PETs, clusters greater than 5 are expected to have very low false positive rates. PET-3 and PET-4 clusters are simulated to have higher false positive rates.

Table 5

ChIP-PET5+ targets compared to the rank list from the 50 b every 50 b array dataset (considering only those ChIP-PET5+ targets with coverage common to the 50 b array

tilepath, Supplementary Table 1). For each PET cluster, its location as well as the cluster overlap region (coordinates are build NCBIv35 [hg17]) are displayed as well as the ranks of targets from the 50 b array dataset that overlap the PET cluster. Since the PET clusters range in size from 1,376 bp to 13,368 bp they can overlap multiple ChIP-chip targets, all of which are 1,300 bp in size. Only 2 of the 11 PET5+ clusters are not detected by ChIP-chip on this 50 b array platform.

Table 6

Comparison of the ranked target list for the top 75 targets from the 50 b every 50 b array dataset and 75 PET3+ targets from the ChIP-PET experiment. A fair evaluation could only be made for the 75 PET3+ clusters that were covered by the 50 b every 50 b array tilepath (see Materials and Methods and Supplementary Table 1). The upper panel displays the false positive rates (FPR) calculated for each dataset considered separately. The lower panel displays the results after merging the list of the top 75 targets identified by the 50 b arrays with the 75 PET3+ targets identified by ChIP-PET. Additionally the ChIP-PET targets are all sized to be 1,300 bp (centered on the overlap region) in order to perform a fair comparison. As for Table 1 this comparison is performed, by first creating the union of the two separate lists and then counting the number of union target regions specific to either the 50 b array targets or the PET3+ clusters, or those targets identified by both platforms. In each of these three categories the union regions that were tested for validation are displayed as well as the associated false positive rate.

Supplementary Table 1

Summary of datasets used for the analyses presented in platform comparisons. For each array platform the number of features, oligonucleotide length, genomic coverage and number of biological replicates performed is listed. Biological replicates are defined as STAT1 ChIP DNA prepared from distinct cell cultures grown, harvested and processed

on separate days. ChIP DNA samples from individual biological replicates were labeled separately and hybridized separately (without pooling) as one biological replicate per array. In many cases the same biological replicates were hybridized to each of the array platforms.

Supplementary Table 2

Similarly to Table 1 the target lists from the 50 b every 50 b and the 36 b every 36 b array datasets are compared, but with the target list restricted to only the top 25 targets for the 36 b arrays in order to compare lists of higher accuracy. Again the upper panel displays the false positive rates (FPR) calculated for each list considered separately. The lower panel displays the results after merging the list of the top 75 targets from the 50 b arrays with the list of the top 25 targets from the 36 b arrays.

Supplementary Table 3

Similarly to Table 1 the target lists from the 50 b every 50 b and the PCR product array datasets are compared, but with the target list restricted to only the top 33 targets for the PCR product arrays in order to compare lists of higher accuracy. Again the upper panel displays the false positive rates (FPR) calculated for each list considered separately. The lower panel displays the results after merging the list of the top 75 targets from the 50 b arrays with the list of top 33 targets from the PCR product arrays.

Supplementary Table 4

Primer pairs used to validate regions sampled across the various ranked targets lists are available at <http://encode.gersteinlab.org/data/Euskirchen_etal/>. The coordinates listed are based on human genome build NCBIv35 [hg17]. The PCR product sizes shown are the results of In Silico PCR run at <http://genome.ucsc.edu/cgi-bin/hgPcr>.

Supplementary Figure 1

Genomic features of the two PET-5 regions that were not detected in the 50 b every 50 b ChIP-chip dataset (Table 5). These regions validated by ChIP-PCR (gel images shown on the right). **A.** Chromosome 21 between coordinates 32,815,161 and 32,816,460 [hg17]. **B**. Chromosome 20 between coordinates 33,366,674 and 33,367,973 [hg17]. For both chromosomal regions, sizable lengths of repetitive sequence (as identified by RepeatMasker) coincided with the PET overlap spans, thus likely impairing the ability of the arrays to detect these targets due to decreased probe density. The ChIP-PCR lanes are labeled for ChIP DNA from IFN-γ stimulated cells, ChIP DNA from unstimulated cells and for HeLaS3 genomic DNA. Fold enrichments as calculated for several biological replicates (see Materials and Methods) are indicated.

Supplementary Figure 2

Example of a ChIP-PET-3 cluster that was reassigned to a ChIP-PET-2 cluster. This target in the region chr5:131963298-131964597 [hg17] could not be confirmed by ChIP-PCR analysis. Closer inspection revealed an unusual instance of 2 overlapping PETs that have an almost identical mapping (with 2 bp difference) and were likely derived from the same ChIP fragment.

References

Bailey, T.L. and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol Biol. 3: 21-9.

Beima, K.M., Miazgowicz, M.M., Lewis, M.D., Yan, P.S., Huang, T.H-M., and Weinmann, A.S. 2006. T-bet binding to newly identified target gene promoters is cell type-independent but results in variable context-dependent functional effects. J. Biol. Chem. 281: 11992-12000.

Bertone, P., Trifonov, V., Rozowsky, J.S., Schubert, F., Emanuelsson, O., Karro, J., Kao, M.Y., Snyder, M., Gerstein, M. 2006. Design optimization methods for genomic DNA tiling arrays. Genes & Dev. 16: 271-281.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas III, E.J., Gingeras, T.R., et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 120: 169-181.

Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res. 16: 595-605.

Boehm, U., Klamp, T., Groot, M., and Howard, J.C. 1997. Cellular responses to interferon-γ. Annu. Rev. Immunol. 15: 749–795.

Bornemann, A.R., Zhang, Z., Rozowsky, J.S., Seringhaus, M., Gerstein, M., Snyder, M.

2006. A comparison of high density oligonucleotide and PCR-based array platforms for the identification of transcription factor binding sites in yeast. Submitted.

Boyer, L.A. Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947–956.

Bromberg, J. and Chen, X. 2001. STAT proteins: Signal transducers and activators of transcription. Methods Enzymol. 333: 138–151.

Buck, M.J. and Lieb, J.D. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83: 349-360.

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: 4813-8.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. 1996. Use of a cDNA microarray to analyse expression patterns in cancer. Nature Genet. 14: 457-460.

Emanuelsson, O., Nagalakshmi, U., Zheng, D., Rozowsky, J.S., Urban, A.E., Du., J., Lian., Z., Stolc, V., Weissman, S., Snyder, M., Gerstein, M. 2006. Assessing the Performance of Different High-density Tiling Microarray Strategies for Mapping Transcribed Regions of the Human Genome. Genome Research. In press.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636–640.

ENCODE Project Consortium. 2006. The ENCODE pilot project: functional annotation of 1% of the human genome. Nature. Under revison.

Euskirchen, G., Royce, T.E., Bertone, P., Martone, R., Rinn, J.L., Nelson, F.K., Sayward, F., Luscombe, N.M., Miller, P., Gerstein, M., et al. 2004. CREB binds to multiple loci on human chromosome 22. Mol. Cell. Biol. 24: 3804–3814.

Hartman, S. E., Bertone, P., Nath, A.K., Royce, T.E., Gerstein, M., Weissman, S., Snyder, M. 2005. Global changes in STAT target selection and transcription regulation upon interferon treatments. Genes & Dev. 19: 2953-68.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Earle-Hughes, J., Snesrud, E., Lee, N., and Quackenbush, J. 2000. A concise guide to cDNA microarray analysis. Biotechniques, 29: 548-562.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans,M., Furey, T.S., Harte, R.A., Hsu, F., et al. 2006. The UCSC Genome Browser database: update 2006. Nucleic Acids Res. 34 (Database issue): D590-D598.

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M., and Snyder, M. 2002. GATA-1 binding sites mapped in the β-globin locus by using mammalian ChIP-chip analysis. Proc. Natl. Acad. Sci. 99: 2924–2929.

Hug, B.A., Ahmed, N., Robbins, J.A., Lazar, M.A. 2004. A chromatin immunoprecipitation screen reveals protein kinase cβ as a direct RUNX1 target gene. J. Biol. Chem. 279: 825–830.

Impey, S., McCorkle, S.R., Cha-Molstad, H., Dwyer, J.M., Yochum, G.S., Boss, J.M., McWeeney, S., Dunn, J.J., Mandel, G., Goodman, R.H. 2004. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. Cell 119: 1041-54.

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

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005b. A high-resolution map of active promoters in the human genome. Nature 436: 876-80.

Kondo, Y., Shen, L., Yan, P.S., Huang, T. H., and Iss, J.P.J. 2004. Chromatin immunoprecipitation microarrays for identification of genes silenced by histone H3 lysine 9 methylation. Proc. Natl. Acad. Sci. 101: 7398-7403.

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.I., et al. 2006. Control of

developmental regulators by Polycomb in human embryonic stem cells. Cell. 125: 301-313.

Levy, D.E. and Darnell Jr., J.E. 2002. Stats: Transcriptional control and biological impact. Nat. Rev. Mol. Cell Biol. 3: 651–662.

Liu, X,. Brutlag, D.L., Liu, J.S. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp. Biocomput. 127-38.

Liu, X.S., Brutlag, D.L., Liu, J.S. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. 20: 835-9.

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. Genetics 38: 431-40.

Luscombe, N.M., Royce, T.E., Bertone, P., Echols, N., Horak, C.E., Chang, J.T., Snyder, M., and Gerstein, M. 2003. ExpressYourself: A modular platform for processing and visualizing microarray data. Nucleic Acids Res. 31: 3477–3482.

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: 376-80.

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 NF-κB

binding sites across human chromosome 22. Proc. Natl. Acad. Sci. 100: 12247–12252.

Ng, P., Wei, C.L., Sung, W.K., Chiu, K.P., Lipovich, L., Ang, C.C., Gupta, S., Shahab, A., Ridwan, A., Wong, C.H., et al. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat. Methods 2: 105–111.

Nuwaysir, E.F., Huang, W., Albert, T.J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J.P., Ballin, J., et al. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res. 12: 1749–1755.

Oberley, M.J. and Farnham, P.J. 2003. Probing chromatin immunoprecipitates with CpG-island microarrays to identify genomic sites occupied by DNA-biding proteins. Methods Enzym. 371: 577-596.

Odom, D.T., Zizlsperger, N., Gordon, D.B., Bell, G.W., Rinaldi, N.J., Murray, H.L., Volkert, T.L., Schreiber, J., Rolfe, P.A., Gifford, D.K., et al. 2004. Control of pancreas and liver gene expression by HNF transcription factors. Science 303: 1378–1381.

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

Rada-Iglesias, A., Wallerman, O., Koch, C., Ameur, A., Enroth, S., Clelland, G., Wester, K., Wilcox, S., Dovey, O.M., Ellis, P.D. et al. 2005. Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. Human Molecular Genetics. 14: 3435-

3447.

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

Roh, T.Y., Cuddapah, S., Zhao, K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes & Dev. 19: 542–552.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual,* 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Scacheri, P.C., Davis, S., Odom, D.T., Crawford, G.E., Perkins, S., Halawi, M.J., Agarwal, S.K., Marx, S.J., Spiegel, A.M., Meltzer, P.S., et al. 2006. Genome-wide analysis of Menin binding provides insights into MEN1 tumorigenesis. PLoS Genetics 2: 406-419.

Service, R.F. 2006. The race for the $1000 genome. Science 311: 1544-1546.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728–1732.

Testa, A., Donati, G., Yan, P., Romani, F., Huang, T.H.-M., Viganò, M.A., and Mantovani, R. 2005. Chromatin Immunoprecipitation (ChIP) on Chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core

promoters. J. Biol. Chem. 280: 13606-13615.

Urban, A.E., Korbel', J.O., Selzer, R., Richmond, T., Hacker, A., Popescu, G.V., Cubells, J.F., Green, R., Emanuel, B.S., Gerstein, M.B., et al. 2006. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc. Natl. Acad. Sci. 103: 4534-4539.

Wasserman, W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5: 267–287.

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: 207-19.

Weinmann, A.S., Bartley, S.M., Zhang, T., Zhang, M.Q., Farnham, P.J. 2001. Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol. Cell. Biol. 21: 6820–6832.

Weinmann, A.S., Pearlly, S.Y., Oberley, M.J., Huang, T.H.-M., Farnham, P.J. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes & Dev. 16: 235–244.

Weinmann, A.S., and Farnham, P.J. 2002. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. Methods 26: 37-47.

Wetmur, J.G., and Davidson, N. 1968. Kinetics of renaturation of DNA. J. Mol. Biol. 31: 349-370.

Wu, J., Smith, L.T., Plass, C., and Huang, T.H-M. 2006. ChIP-chip comes of age for genome-wide functional analysis. Cancer Res. 66: 6899-6902.

Zhang, Z., Rozowsky, J., Lam, H., Snyder, M., and Gerstein, M. 2006. Tilescope: online analysis pipeline for high-density tiling microarray data. Submitted.

Table 1 Comparison of ranked target lists between the 50 b every 50 b and the 36 b every 36 b array platforms

	50 every 50 dataset	36 every 36 dataset	Union
Count	Top 75	Top 39	84
FPR	0.26	0.52	

	Specific to 50 every 50 set	Specific to 36 every 36 set	Common to both datasets
Count	51	15	18
Positives (ChIP-PCR validation)	20	0	8
Negatives (ChIP-PCR validation)	7	7	3
FPR	0.26 (= 7/27)	1.00 (= 7/7)	0.27 (= 3/11)

Table 2 Comparison of ranked target lists between the 50 b every 50 b and the PCR product array platforms

	50 every 50 dataset	PCR product dataset	Union
Count	Top 75	Top 75	133
FPR	0.26	0.64	

	Specific to 50 every 50 set	Specific to PCR product set	Common to both datasets
Count	65	62	6
Positives (ChIP-PCR validation)	21	7	6
Negatives (ChIP-PCR validation)	11	34	0
FPR	0.34 (= 11/32)	0.83 (= 34/41)	0.00 (= 0/6)

Table 3 Comparison of ranked target lists for paired samples hybridized either in the presence or absence of Cot-1 DNA on the 50 b every 38 b array platform

	Plus Cot dataset	Minus Cot dataset	Union
Count	Top 75	Top 75	104
FPR	0.31	0.57	

	Specific to Plus Cot set	Specific to Minus Cot set	Common to both datasets
Count	34	36	34
Positives (ChIP-PCR validation)	15	2	11
Negatives (ChIP-PCR validation)	5	10	5
FPR	0.25 (= 5/20)	0.83 (= 10/12)	0.31 (= 5/16)

Table 4 Monte Carlo simulation of the expected number of PET clusters from the ENCODE regions as a function of the PET cluster size

	Total PETs	PET-1	PET-2	PET-3	PET-4	PET-5	PET-6	PET-7	PET-8+
ENCODE region	4007	2320	477	88	14	6	4	1	3
Expected at random		2794	463	47	3	0.14	0.0051	0.0002	< 0.01
Estimate % of error		100	97.065	53.409	21.4286	2.3330	0.1285	0.0170	< 1e-5

Table 5

Comparison of ChIP-PET5+ targets and ranked targets from the 50 b every 50 b array ChIP-chip dataset

Cluster Overlap Count	Cluster Location	Cluster Span	Overlap Location	Overlap Span	Rank of Regions on 50 b every 50 b Array
33	chr5:131852871-131866238	13368	chr5:131860666-131860682	17	1, 2, 4, 5, 9, 16, 21, 78, 79
17	chr5:131786288-131794715	8428	chr5:131791111-131791136	26	10, 37, 58, 64, 74
8	chr21:33523431-33526064	2634	chr21:33524385-33524495	111	6
7	chr21:33619380-33622627	3248	chr21:33619552-33619612	61	31
6	chr21:34088872-34092435	3564	chr21:34091417-34091453	37	8
6	chr15:41572478-41573853	1376	chr15:41573016-41573093	78	7
5	chr21:33554904-33562043	7140	chr21:33560526-33560713	188	11
5	chr20:33363723-33369198	5476	chr20:33367242-33367404	163	Not detected
5	chr7:115935369-115942046	6678	chr7:115940692-115940803	112	42
5	chr21:32815193-32821120	5928	chr21:32815780-32815840	61	Not detected
5	chr8:119135229-119140085	4857	chr8:119137560-119137726	167	41

Table 6 Comparison of ranked target lists between the 50 b every 50 b array and ChIP-PET platforms

	50 every 50 dataset	ChIP-PET dataset	Union
Count	Top 75	Top 75	134
FPR	0.26	0.17	

	Specific to 50 every 50 Set	Specific to ChIP-PET	Common to both datasets
Count	59	61	14
Positives (ChIP-PCR validation)	15	10	14
Negatives (ChIP-PCR validation)	11	5	0
FPR	0.42 (= 11/26)	0.33 (= 5/15)	0.00 (= 0/14)

Figure 1 Comparison of signal enrichment tracks

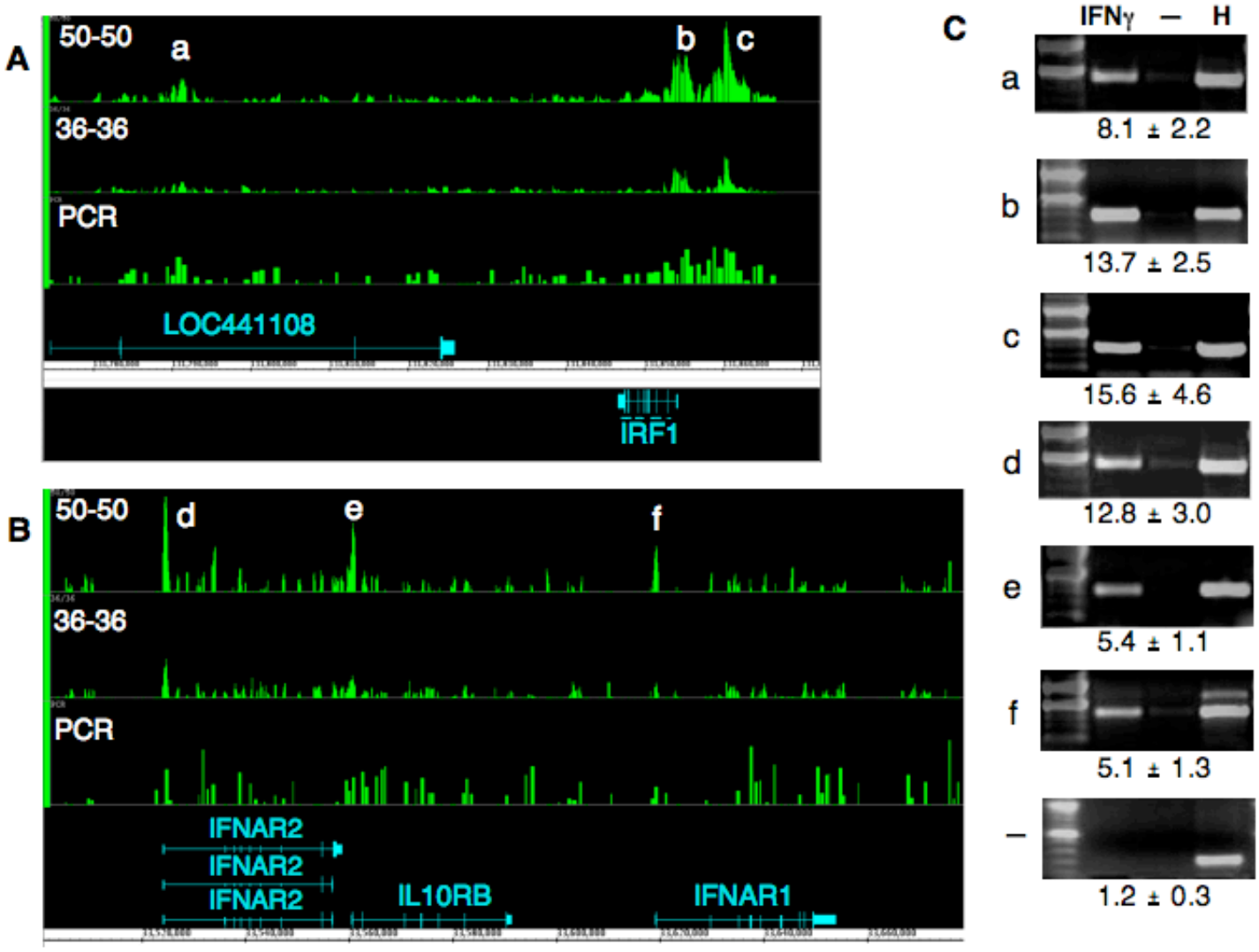


Figure 2 Validation frequency as a function of rank order

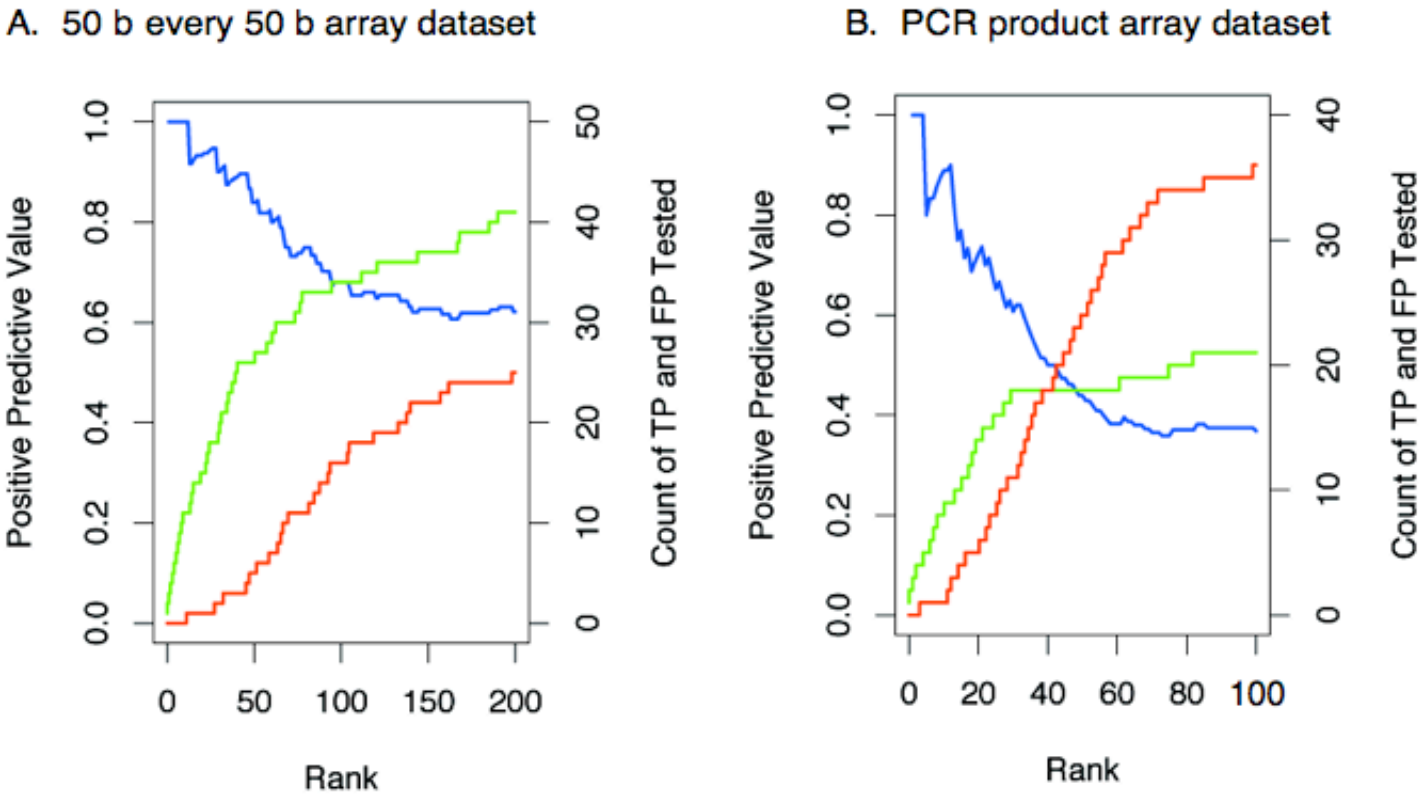


Figure 3 Validation frequency as a function of signal for the 50 b every 50 b array dataset

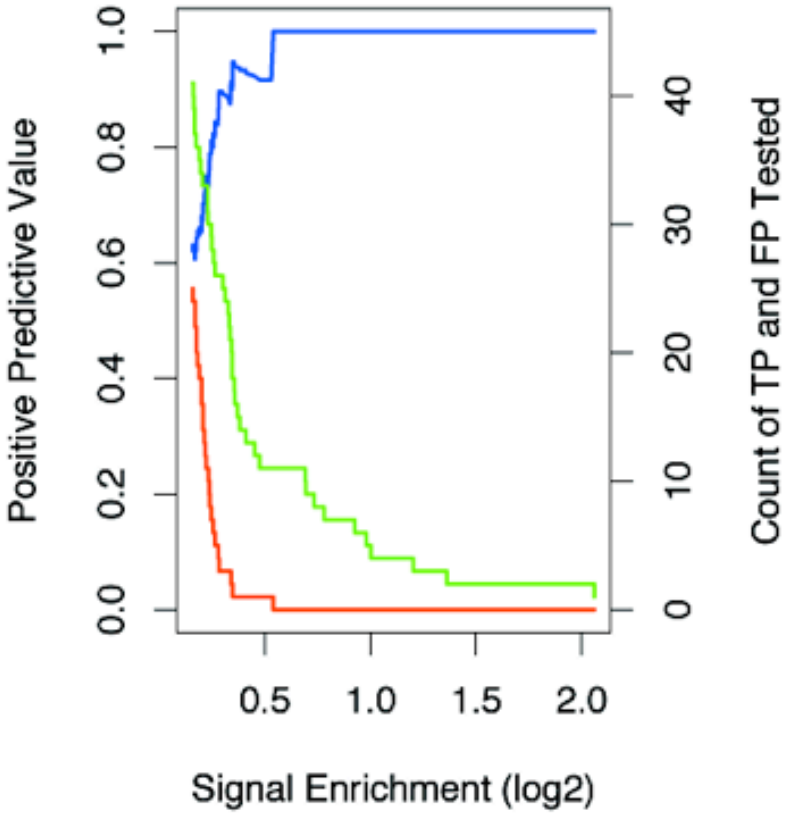


Figure 4 Agreement between the ranked target lists for the 50 b every 50 b array and the PCR product array platforms

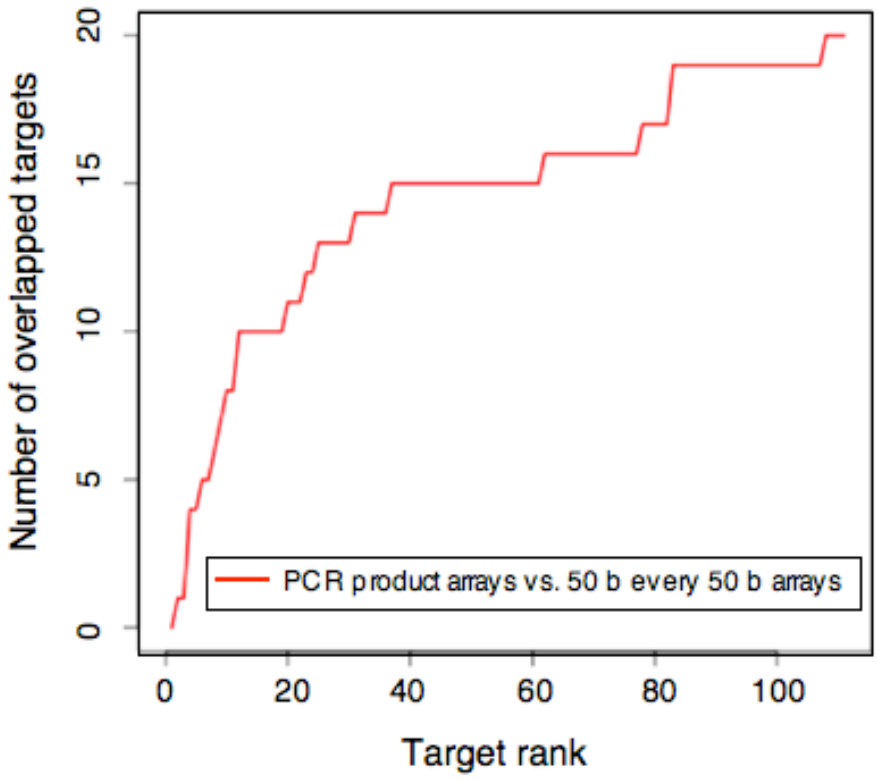
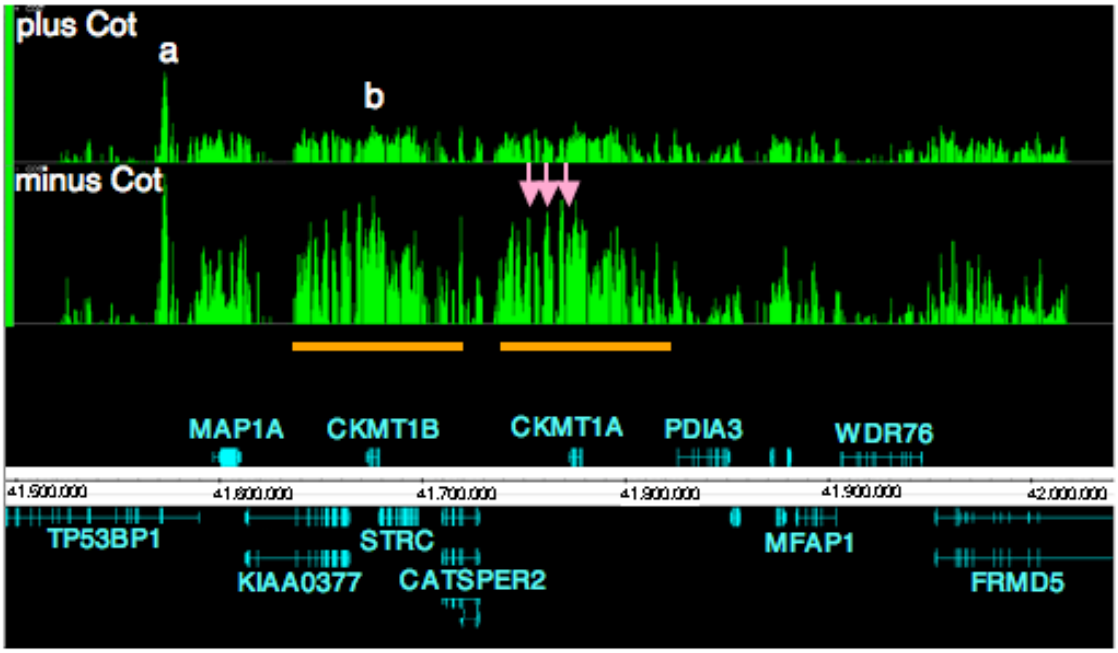


Figure 5 The effect of Cot-1 DNA



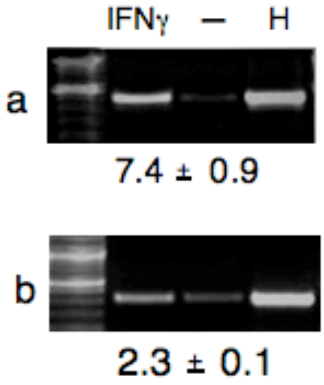


Figure 6 The value of increasing replicate experiments

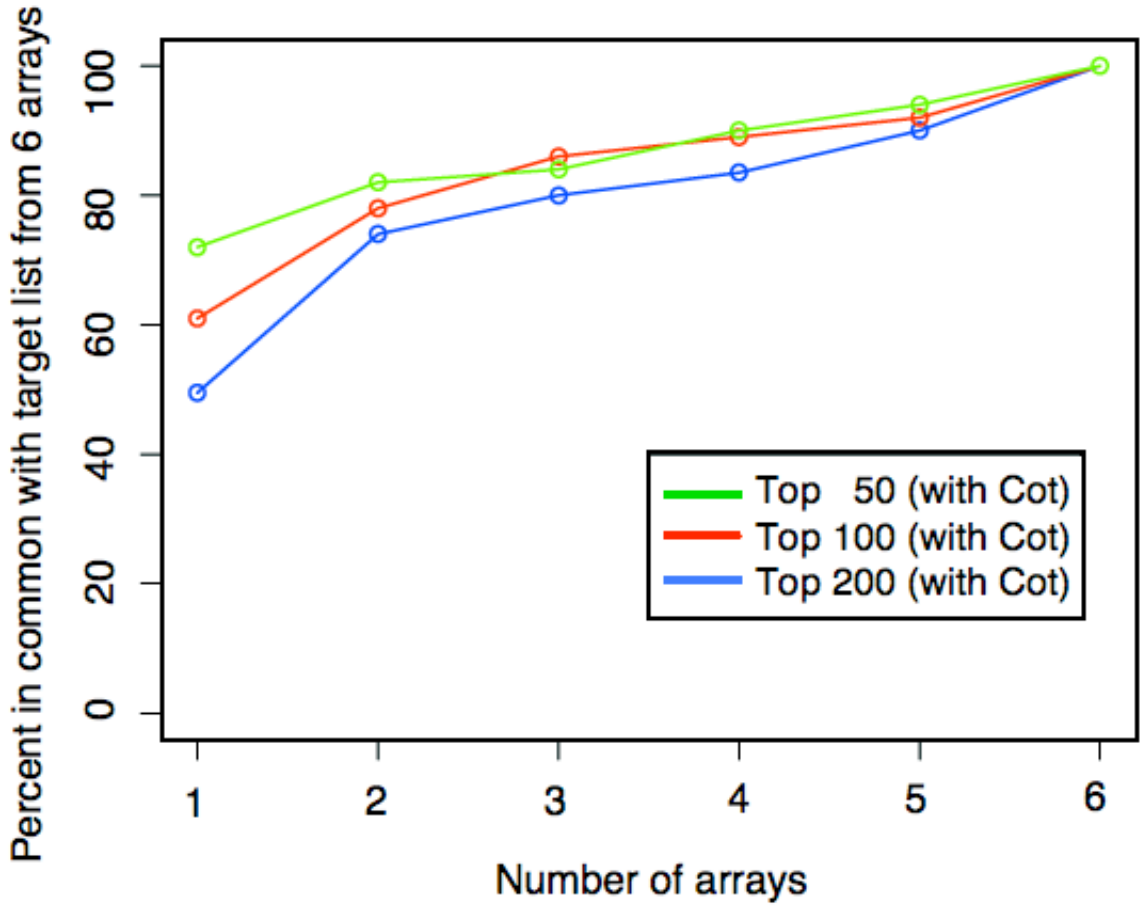
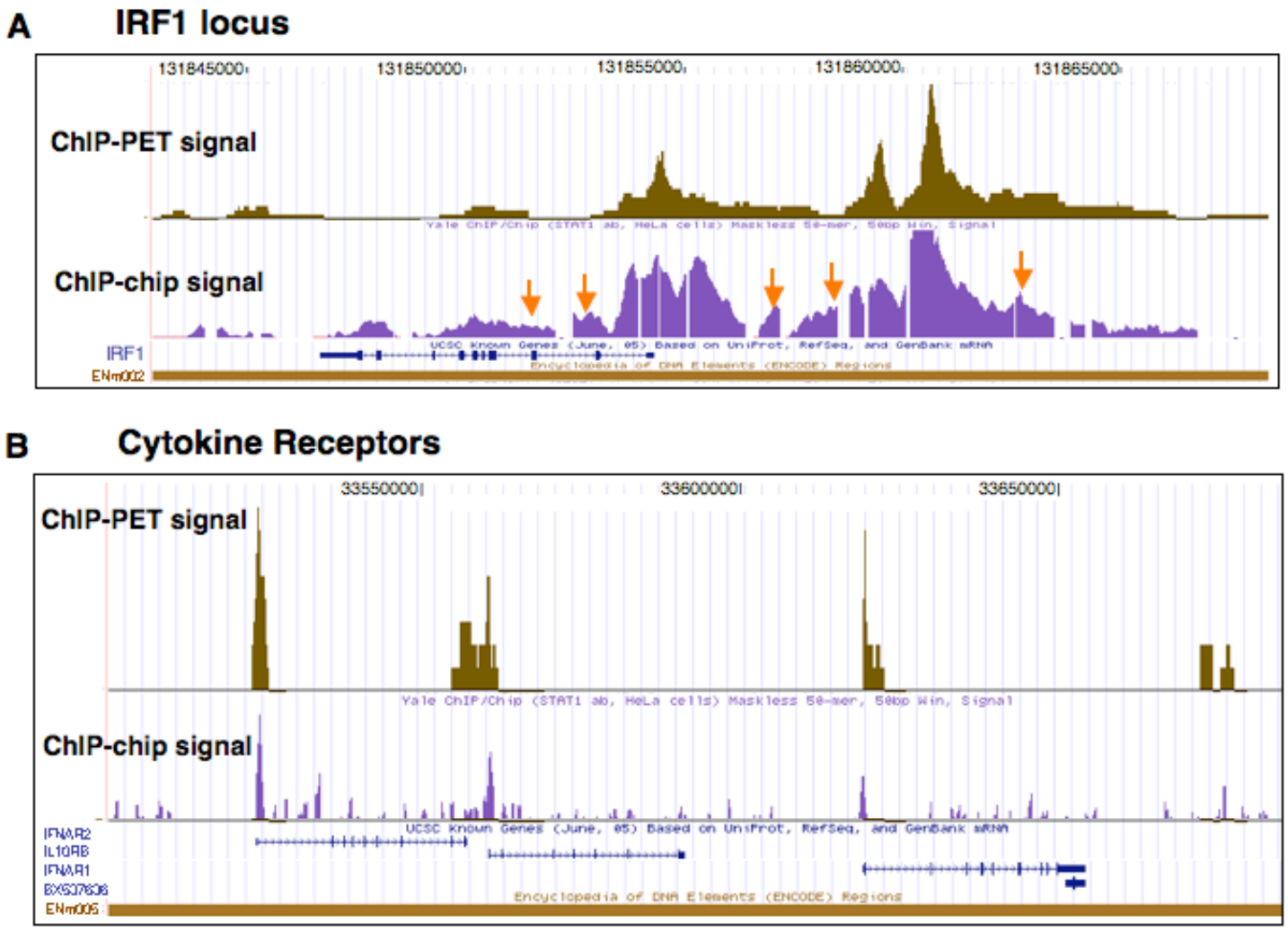


Figure 7 Comparison of ChIP-chip and ChIP-PET



Supplementary Table 1
Summary of datasets used for the analyses presented in platform comparisons

Platform	36every36 array	50every50 array	50every38 array	PCR Product array	ChIP-PET
Number of probe elements	382,454	192,040	382,885	24,341	NA
Resolution	36 bp	50 bp	38 bp	620 bp (average)	< 6 kb
Coverage	36 x 382,454 14.0 Mb	50 x192,040 9.5 Mb	~38 x 382,885 14.6 Mb	620 x 24,341 15.2 Mb	whole genome
Number of arrays in dataset (= number of biological replicates)	2	3	12 arrays = 6 biological replicates split post-labeling and hybridized with and without Cot-1 DNA	6	NA

Supplementary Table 2 Comparison of ranked target lists between the 50 b every 50 b and the 36 b every 36 b array platforms

	50 every 50 dataset	36 every 36 dataset	Union
Count	Top 75	Top 25	74
FPR	0.26	0.38	

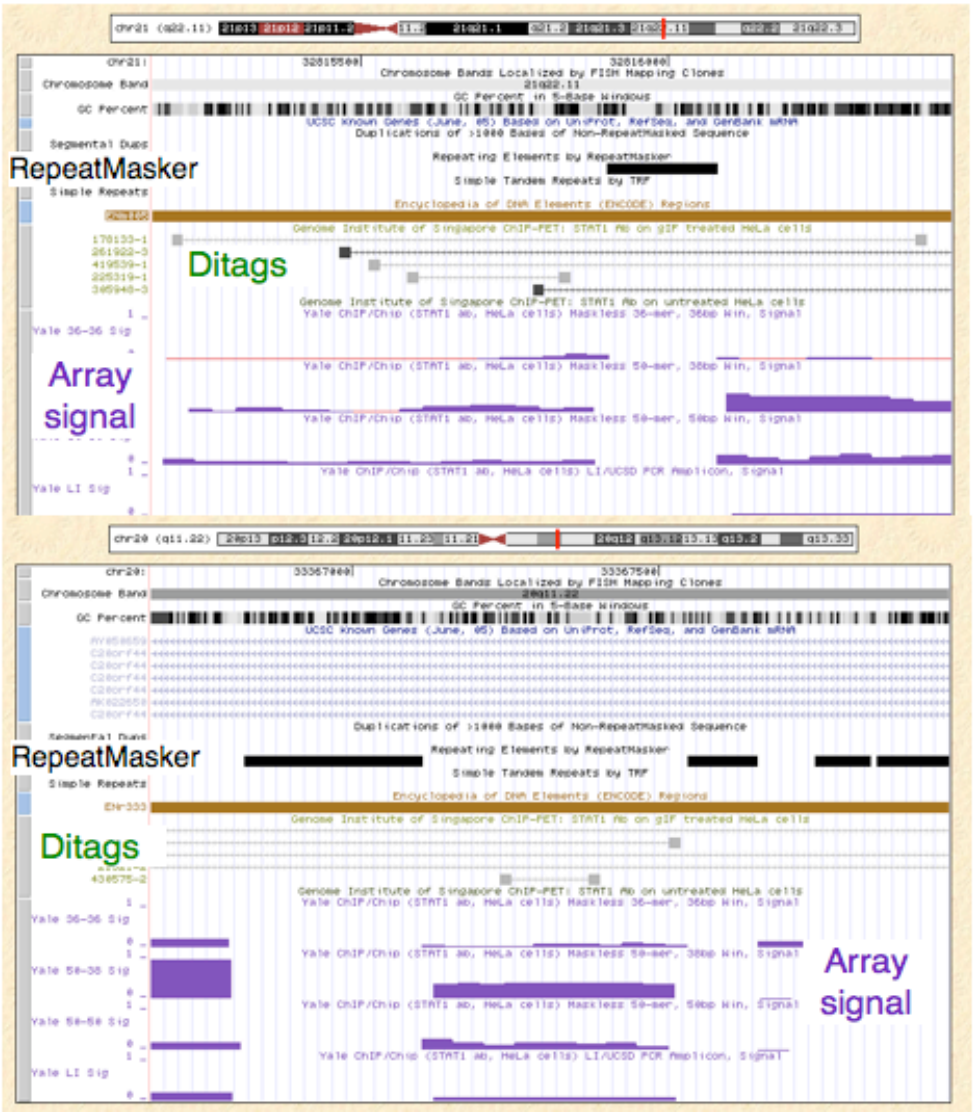
	Specific to 50 every 50 set	Specific to 36 every 36 set	Common to both datasets
Count	55	5	14
Positives (ChIP-PCR validation)	22	0	6
Negatives (ChIP-PCR validation)	8	2	2
FPR	0.27 (= 8/30)	1.00 (= 2/2)	0.25 (= 2/8)

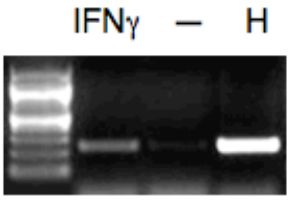
Supplementary Table 3 Comparison of ranked target lists between the 50 b every 50 b and the PCR product array platforms

	50 every 50 dataset	PCR product dataset	Union
Count	75	33	93
FPR	0.26	0.40	

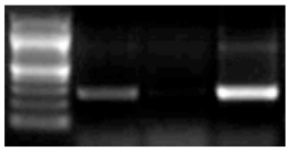
	Specific to 50 every 50 set	Specific to PCR product set	Common to both datasets
Count	65	22	6
Positives (ChIP-PCR validation)	21	7	6
Negatives (ChIP-PCR validation)	11	12	0
FPR	0.34 (= 11/32)	0.63 (= 12/19)	0.00 (= 0/6)

Supplementary Figure 1 The two PET-5 regions detected by ChIP-PET and validated by ChIP-PCR analysis that were undetected by ChIP-chip with the 50 b every 50 b platform





5.2 ± 1.7



7.7 ± 0.5

Supplementary Figure 2
Example of a ChIP-PET-3 cluster that was reassigned to a ChIP-PET-2 cluster