Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing

Gordon Robertson[1], Martin Hirst[1], Matthew Bainbridge[1], Misha Bilenky[1], Yongjun Zhao[1], Thomas Zeng[1], Ghia Euskirchen[2], Bridget Bernier[1], Richard Varhol[1], Allen Delaney[1], Nina Thiessen[1], Obi L Griffith[1], Ann He[1], Marco Marra[1], Michael Snyder[2] & Steven Jones[1]

We developed a method, ChIP-sequencing (ChIP-seq), combining chromatin immunoprecipitation (ChIP) and massively parallel sequencing to identify mammalian DNA sequences bound by transcription factors *in vivo*. We used ChIP-seq to map STAT1 targets in interferon-γ (IFN-γ)–stimulated and unstimulated human HeLa S3 cells, and compared the method's performance to ChIP-PCR and to ChIP-chip for four chromosomes. By ChIP-seq, using 15.1 and 12.9 million uniquely mapped sequence reads, and an estimated false discovery rate of less than 0.001, we identified 41,582 and 11,004 putative STAT1-binding regions in stimulated and unstimulated cells, respectively. Of the 34 loci known to contain STAT1 interferon-responsive binding sites, ChIP-seq found 24 (71%). ChIP-seq targets were enriched in sequences similar to known STAT1 binding motifs. Comparisons with two ChIP-PCR data sets suggested that ChIP-seq sensitivity was between 70% and 92% and specificity was at least 95%.

Methods for profiling *in vivo* sites of protein-DNA interactions in mammalian genomes use ChIP or a methylation-based tagging technique (DamID) followed by either microarray hybridization or sequencing[1,2]. For ChIP with microarray hybridization (ChIP-chip), tiled oligonucleotide microarrays offer higher spatial resolution than spotted arrays, although the latter typically have better coverage of repetitive regions.

ChIP-sequencing methods directly offer whole-genome coverage. Additionally, they offer both low analytical complexity and sensitivity that increases with sequencing depth. Methods have been reported that use SAGE-like tags[3–6] and 36-bp paired-end tags (PETs)[7–9]. ChIP-PET maps both ends of an immunoprecipitated DNA fragment to a genome, and has been reported with both Sanger and multiplex 454 (Roche) sequencing[10]. The Illumina 1G system (1G) provides a 1–2 orders of magnitude increase in the amount of sequence that can be cost-effectively generated[11]. We reasoned that given this high throughput, a method based on deep 1G sequencing of short-read single-end tags (SETs), which are simpler to prepare than PETs, may be effective for profiling mammalian protein-DNA interactions. Thus we appraised the 1G system as a platform for ChIP with tag sequencing.

As a test system, we selected the mammalian transcription factor STAT1, whose cellular biology is relatively well characterized, and whose use permits a comparison of unstimulated and stimulated cellular states[12–16]. In both resting and stimulated cells, STAT proteins shuttle continuously between cytoplasm and nucleus[12,13,15]. Signaling by several cytokines, growth factors and hormone receptors leads to activation of receptor-associated JAK family kinases that phosphorylate a substantial fraction of cytoplasmic STAT1 proteins[12,15,17–20]. Phosphorylated STAT1 forms specific homodimers, heterodimers and heterotrimers that bind DNA with high affinity, and thus accumulate in the nucleus. STAT1 complexes activate or repress transcription primarily by the homodimer binding to IFN-γ activation site (GAS) elements, but also to interferon-stimulated response elements (ISREs)[16,17]. The regulatory activity of STAT1 also depends on other bound transcription factors[17,21], and bound STAT1 can recruit histone acetyltransferases and transcriptional coactivators[13]. Activated STATs have a short nuclear half-life, and are rapidly dephosphorylated and returned to the cytoplasm[13,15]. Independent of activation, unphosphorylated STAT1 constitutively regulates the basal expression of certain genes[17,19,22].

Using ChIP-seq, we compared STAT1 DNA binding in IFN-γ–stimulated and unstimulated human HeLa S3 cells, by generating approximately 47 million reads (27 bp), totaling 1.26 Gb of sequence data, from immunoprecipitated DNA fragments. For each sample, we transformed reads that mapped to unique genomic locations into a DNA fragment overlap profile. We identified significant peaks by thresholding profiles at a height equivalent to an estimated false discovery rate (FDR) ≤ 0.001. The global properties of the resulting two profiles were distinct and consistent with high ChIP enrichment. We compared the results of ChIP-seq with data obtained using ChIP-PCR and ChIP-chip.

[1]British Columbia Cancer Agency Genome Sciences Centre, 675 West 10th Avenue, Vancouver, British Columbia V5Z 4S6, Canada. [2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. Correspondence should be addressed to S.J. (sjones@bcgsc.ca).

RECEIVED 11 MAY; ACCEPTED 5 JUNE; PUBLISHED ONLINE 11 JUNE 2007; DOI:10.1038/NMETH1068

Table 1 | ChIP-seq reads (SETs) and peaks thresholded at FDR = 0.001

	Parameter	Stimulated	Unstimulated
Reads	Total sequenced (millions)	24.1	22.7
	Total, uniquely mapped (millions)	15.1	12.9
	In peaks (millions)	2.71 (17.9%)	0.54 (4.2%)
	Peak coverage (Mb)	44.3 (1.44%)	10.3 (0.33%)
	Median width (bp)	916	815
	Enrichment	12.1	12.2
Peaks	Peak height at FDR threshold	11	11
	Number of peaks	41,582	11,004
	Average height	29.2	21.0
	Median height	16	13

Peak coverage percentages assume a 3.1 Gb genome.

RESULTS

Detection of ChIP-enriched sequences

We refer to peaks, profiles, data and reads resulting from stimulated or unstimulated cells as stimulated or unstimulated peaks, profiles, data and reads, respectively. After ChIP for STAT1 in IFN-γ–stimulated HeLa S3 cells, we used the 1G system to produce a total of 24.1 million 27-bp sequence reads (SETs; Table 1). Of these, 15.1 million (63%) aligned uniquely to the non-repeat-masked human genome sequence. For unstimulated cells, we generated 22.7 million reads and uniquely mapped 12.9 million reads (57%). In total, we generated 46.8 million reads that represented over 1.26 Gb of sequence data. We discarded all reads that could not be uniquely mapped to the genome.

We extended SETs directionally, transforming each into a 174-bp extended SET (XSET), and then calculated XSET overlap profiles. We established a threshold for stimulated and unstimulated overlap profiles at peak heights of 11, which corresponded to estimated FDRs < 0.001 for each profile (**Table 1**, **Fig. 1**, **Supplementary Fig. 1** and **Supplementary Data 1** online). A peak's 'height' was the maximum number of overlapped XSETs for that peak; height and FDR were inversely related (**Supplementary Fig. 2** online). The resulting 41,582 stimulated peaks contained 17.9% of mapped reads, whereas the 11,004 unstimulated peaks contained a 4.3-fold smaller fraction of mapped reads (4.2%). This means that, of uniquely mapped reads, ~82% from stimulated cells and ~96% from unstimulated cells aligned to the genome in a manner that was indistinguishable from random expectation at an FDR threshold of 0.001. Approximately 85% of the unstimulated peaks overlapped stimulated peaks (**Supplementary Fig. 1b** and **Supplementary Fig. 3** online). Over a large range of peak heights, the stimulated profile contained at least ten times more peaks and covered at least 20 times more genomic sequence than the unstimulated profile (**Fig. 2**). We used a computational resampling approach to estimate that the sequencing depth likely had 'saturated' or identified all peaks available to the method for unstimulated cells, but not for the stimulated cells (data not shown). This indicates that the differences in the number of significant stimulated and unstimulated peaks was not simply due to the 15% difference in the number of uniquely mapped reads between the two sets of results. Global fold-enrichment, estimated as the ratio of the fraction of tags in peaks to the fraction of the genome covered by peaks, was approximately 12 for both samples. Stimulated and unstimulated peak width distributions were significantly different (Kolmogorov-Smirnov test, $P < 10^{-20}$; **Supplementary Fig. 4** online). Median peak widths were 916 bp and 815 bp, respectively.

The stimulated and unstimulated profiles contained 0.6% (243/41,582) and 2.5% (278/11,004) of narrow peaks with lengths ≤200 bp, which spanned a wide range of heights and may have derived from the same immunoprecipitated DNA fragment. These may represent artifacts of the experimental approach and could be filtered out using heuristic approaches.

Peak locations relative to genomic annotations

A comparison of stimulated and unstimulated peaks to different types of annotated genomic features is shown in **Table 2**. Because STAT1 is a transcriptional regulator, we compared peak locations to annotated Ensembl genes[23]. Peaks were concentrated between approximately –600 and +500 bp relative to the transcriptional start site, with the highest density at approximately –100 bp for both stimulated and unstimulated peaks (**Supplementary Fig. 5** online). Because evolutionary conservation is a useful, though incomplete, indication that a sequence may be functional[24], we also determined the fraction of peaks that overlapped conserved genomic elements[25]. Finally, we noted that a number of peaks overlapped genomic regions annotated as repetitive[25] and were present in pericentromeric regions (**Fig. 1**).

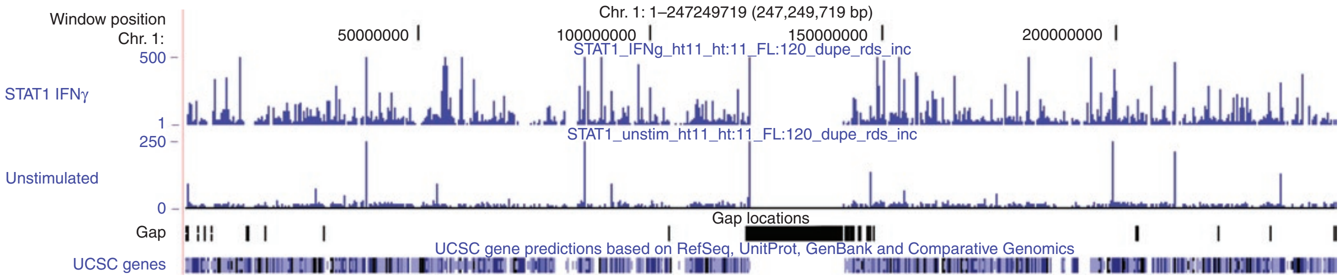

Figure 1 | FDR-thresholded XSET profiles and peaks. Stimulated and unstimulated FDR-thresholded XSET profiles for the 247 Mb chromosome 1 (UCSC hg18 genome browser[25]). Profiles were clipped at peak heights of 500 and 250.

A smaller fraction of stimulated than unstimulated peaks overlapped or were within 1 kb of an annotated transcription start site (19% versus 29%; **Table 2**). Much smaller fractions of peaks overlapped or were within 1 kb of 3′ ends of annotated genes (~6% and ~7%). Both the stimulated and unstimulated data had approximately the same fraction of intragenic and intergenic peaks (~50% and ~25%, respectively). Similar fractions (~56%) of stimulated and unstimulated peaks overlapped conserved genomic elements. Approximately 16–18% of all peaks overlapped repetitive elements, often those annotated as satellite or tandem repeats[26], and 1–3% of peaks were in pericentromeric regions, within centromeres or in peritelomeric regions.

Peaks overlapped known functional STAT1 binding sites

We compiled 41 (22 GAS plus 19 ISRE) known functional human STAT1 binding sites at 34 genomic loci from the literature, along with the cell types tested and the type of interferon stimulation (α, β or γ; **Supplementary Fig. 6** and **Supplementary Data 2** online).

Stimulated peaks overlapped 24 of the 34 loci in all cell types, suggesting a sensitivity of 71%. For 86% (24/28) of the overlapped sites, the known site was within 70 bp of the location of the peak maximum (**Fig. 3**). An additional four sites, associated with peaks whose heights were less than 20, were within 220 bp of the associated peak maximum. Peaks overlapped 7 of the 34 loci (21%) in unstimulated cells, consistent with unstimulated peaks representing *bona fide* binding sites. For 12 sites at 9 loci that were reported from experiments done with HeLa cells, stimulated peaks overlapped 8 of 9 loci, indicating a sensitivity of 89%. Further, the *CD40* gene, whose known site was not overlapped by a stimulated peak, is reported as not IFN-γ–inducible in HeLa cells (**Supplementary Data 2**). This suggests that the *CD40* site was a true negative rather than a false negative, and that the sensitivity for known STAT1 sites in HeLa cells was 100%. This sensitivity estimate from the previously characterized STAT1 sites in HeLa cells was consistent with our sensitivity estimate from 13 published 'high confidence' HeLa STAT1 sites[27], for which stimulated peaks overlapped 11 sites, and overlapped or were proximal to 12 of the 13 sites. In our results, sites that were not overlapped by a stimulated peak were reported from diverse cell lines other than HeLa.

Peaks were enriched in STAT1 binding motifs

We assessed whether peaks were enriched in sequences similar to those known to be functional STAT1 binding sites, relative to a random expectation. We used a data resource of 41 such 'known' sites in two ways: as GAS and ISRE position weight matrices (PWMs) generated from the binding-site sequences and directly by scanning with site sequences (**Supplementary Data 2** and **Supplementary Figs. 6–8** online). A random expectation was generated from shuffled peak sequences and from 'negative' regions that were tiled in the ChIP-chip data set described below, but had neither ChIP-chip nor ChIP-seq peaks. For PWM scans, we retained only the best score for a peak. Peak sequences were enriched in higher-scoring sequences,

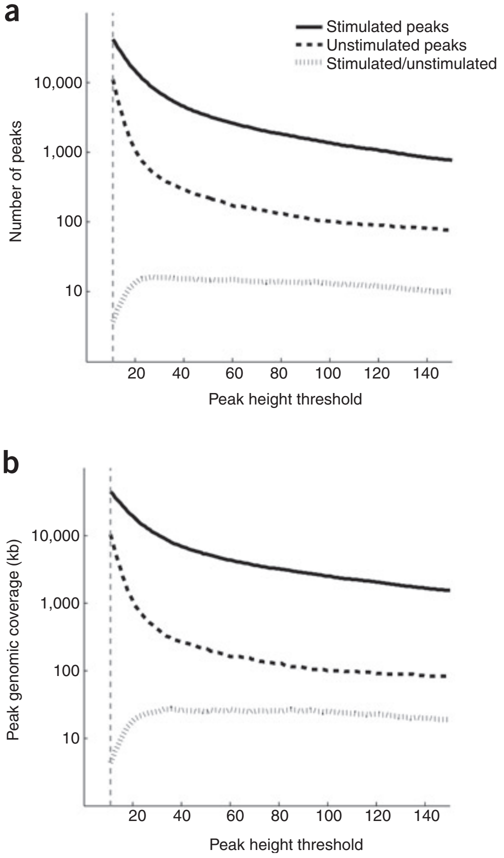

Figure 2 | Stimulated and unstimulated peak sets had distinct global properties. (**a**,**b**) Number of peaks (**a**) and peak genomic coverage (**b**) as a function of peak height. The peak height of 11, which corresponds to an FDR threshold of 0.001 for both stimulated and unstimulated data sets, is shown as dashed vertical gray lines.

Table 2 | Peak locations relative to genomic annotations

Peak location	Stimulated	Unstimulated
Overlaps 5′ end of a gene	5,967 (0.143)	2,647 (0.241)
Within 1 kb of 5′ end of a gene	7,766 (0.187)	3,232 (0.294)
Overlaps 3′ end of a gene	851 (0.020)	289 (0.026)
Within 1 kb of 3′ end of a gene	2,274 (0.055)	729 (0.066)
Intragenic	21,156 (0.509)	5,908 (0.537)
Intergenic	10,714 (0.258)	2,589 (0.235)
Overlaps PhastCons element	23,209 (0.558)	6,171 (0.561)
Overlaps tandem repeat	6,590 (0.158)	1,962 (0.178)
<2 Mb from a centromere	483 (0.012)	363 (0.033)
Total peaks	41,582	11,004

Annotations included Ensembl Build 41 gene models, and UCSC[25] hg18 tandem repeats, conserved PhastCons elements and centromeres. Parentheses indicate fractions of total numbers of stimulated or unstimulated peaks. 'Intergenic' peaks were more than 20 kb from the transcriptional start site of an Ensembl protein-coding gene.

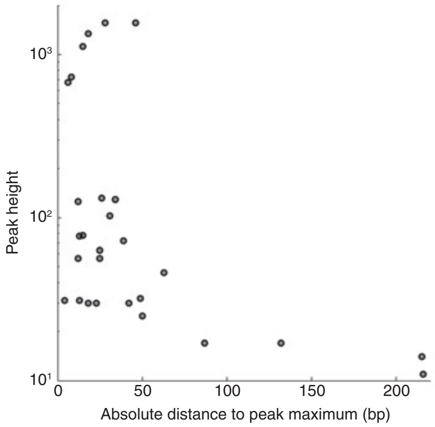

Figure 3 | Relationship of peak height to absolute distance from the peak maximum to an overlapped known STAT1 binding site. Results shown are for the 24 stimulated peaks that overlapped at least one of the 41 human GAS or ISRE sites (see **Supplementary Data 2**).

with stronger enrichment for stimulated than unstimulated peaks (Kolmogorov-Smirnov $P = 0.0$; **Fig. 4a**,**b**). Taller peaks were significantly enriched in higher-scoring sequences in stimulated but not in unstimulated peaks (Kolmogorov-Smirnov test, stimulated $P = 0.0$, unstimulated $P < 10^{-3}$; **Fig. 4a**,**b**, **Supplementary Figs. 9**, **10** and **Supplementary Table 1** online). For scans with known binding-site sequences, ChIP-negative regions indicated that the least stringent condition (2 mismatches for GAS and 3 for ISRE) was uninformative (data not shown). Genomic sequences for both stimulated and unstimulated peaks were enriched in sequences similar to known functional sequences, with higher enrichment for more stringent scans and closer to peak maxima (**Fig. 4c**,**d**, **Supplementary Fig. 11** and **Supplementary Table 2** online). Enrichment was 2.0–5.3 for peak sequences and 4.4–19.1 for sequences within 100 bp of a peak maximum. A two- to threefold enrichment has been previously reported for low-stringency scans with a 9-bp consensus sequence for a PCR product microarray[28]. In both PWM and site-sequence scans, GAS-like sites were more frequent than ISRE-like sites. This is consistent with expectations for IFN-γ–stimulated cells, in which STAT1 complexes activate or repress transcription primarily by the homodimer binding to GAS elements, but also, to a smaller extent, to ISREs[16,17]. Consistent with distances of known STAT1 binding sites to peak maxima (**Fig. 3**), GAS- and ISRE-like sequences were also enriched within approximately 100 bp of stimulated peak maxima (**Fig. 4e**,**f**).

Overlap between ChIP-seq and ChIP-chip peaks

We compared the ChIP-seq peaks with ChIP-chip targets detected on chromosomes 20, 21, 22 and X by a 50-mer high-density tiled oligonucleotide microarray. The regions compared represented approximately 10% of the total human genome, and thus were approximately six times larger than the region assessed with a spotted PCR product array[28] and ten times larger than the ENCODE regions[29]. Site-sequence scanning indicated that ChIP-chip peaks were enriched in sequences similar to known functional STAT1 binding sites; for the 1% FDR peaks, enrichment ratios were 7.75 and 1.9 for the two more stringent scan conditions (**Supplementary Table 3** online). There was a striking correspondence between peak sets generated by the two platforms, particularly for ChIP-seq peaks with lower FDRs and ChIP-chip peaks with higher scores (**Supplementary Fig. 12** online). To quantify peak overlap, adjacent 1% FDR ChIP-chip peaks that were separated by gaps shorter than 125 bp were merged, resulting in 803 peaks (**Supplementary Fig. 13** and **Supplementary Table 4** online). In contrast, these four chromosomes contained 3,090 ChIP-seq peaks. Between 64% and 71% of the merged ChIP-chip peaks overlapped at least one ChIP-seq peak, depending on the overlap required (**Supplementary Table 5** online). Because ChIP-seq reported more peaks than ChIP-chip, a smaller fraction of ChIP-seq peaks overlapped merged ChIP-chip peaks (17–19%). For 5% FDR ChIP-chip peaks, the corresponding overlap was 34–42% and 30–36%. For both ChIP-chip and ChIP-seq peaks, peaks that overlapped between platforms had significantly higher array signal enrichment scores or heights than peaks that were unique to either platform (Kolmogorov-Smirnov $P < 10^{-10}$; **Supplementary Fig. 14** online).

Peaks overlapped ChIP-PCR–validated STAT1 binding sites

We estimated the method's specificity and sensitivity by comparing locations of stimulated peaks with ChIP-PCR data for STAT1 binding in HeLa S3 cells from two sources[27,30]. The first ChIP-PCR data set[30] consisted of 33 positive and 53 negative experimental regions. We grouped the positive regions into 20 loci by region overlap and proximity, and by overlap and proximity with DNase I hypersensitive sites (DHSs) identified in HeLa S3 cells[27] (**Supplementary Results** online). Seventy percent (14/20) of the loci were overlapped by peaks, and 85% (17/20) of the loci were either overlapped by peaks or were within 250 bp of the location of a peak maximum. Similarly, we grouped the 53 negative regions into 42 loci by region overlap or proximity. Six negative PCR regions overlapped DHSs, and another two negative loci contained PCR regions that were within approximately 150 bp of a DHS. A stimulated peak overlapped a PCR region directly in only one of the 42 negative loci; this peak's height was the minimum permitted by the FDR threshold (11 XSETs), and both the peak and the PCR region overlapped a DHS. For the stimulated sample, these results suggested a sensitivity of 0.70–0.85 (14/20 to 17/20), and a specificity of at least 0.97 (41/42). No negative PCR regions overlapped unstimulated peaks.

The second ChIP-PCR data set consisted of 13 'high-confidence' STAT1 binding sites that were identified by ChIP-chip and validated by quantitative PCR[27]. Three sites were distal and four were proximal to a transcription start site (that is, beyond or within 2.5 kb), one overlapped a coding exon, five were intronic and one, depending on the transcript chosen, was either intronic or proximal. Of these sites, 77% (10/13) overlapped or were within 500 bp of a DHS reported in that work[27]. Stimulated ChIP-seq peaks overlapped 92% (12/13) of the sites, and, consistent with results from known sites, 58% (7/12) of the overlapped sites were within ~150 bp of a ChIP-seq peak maximum.

Overall, these results suggested that the sensitivity of the method in this experiment was at least 70%, but was likely closer to 92%, whereas the specificity was at least 95%.

DISCUSSION

Our results suggest that massively parallel 1G sequencing can generate asymptotically complete genomic DNA binding profiles of mammalian proteins for which ChIP is effective. ChIP-seq appears to offers several advantages over ChIP-chip, particularly for large

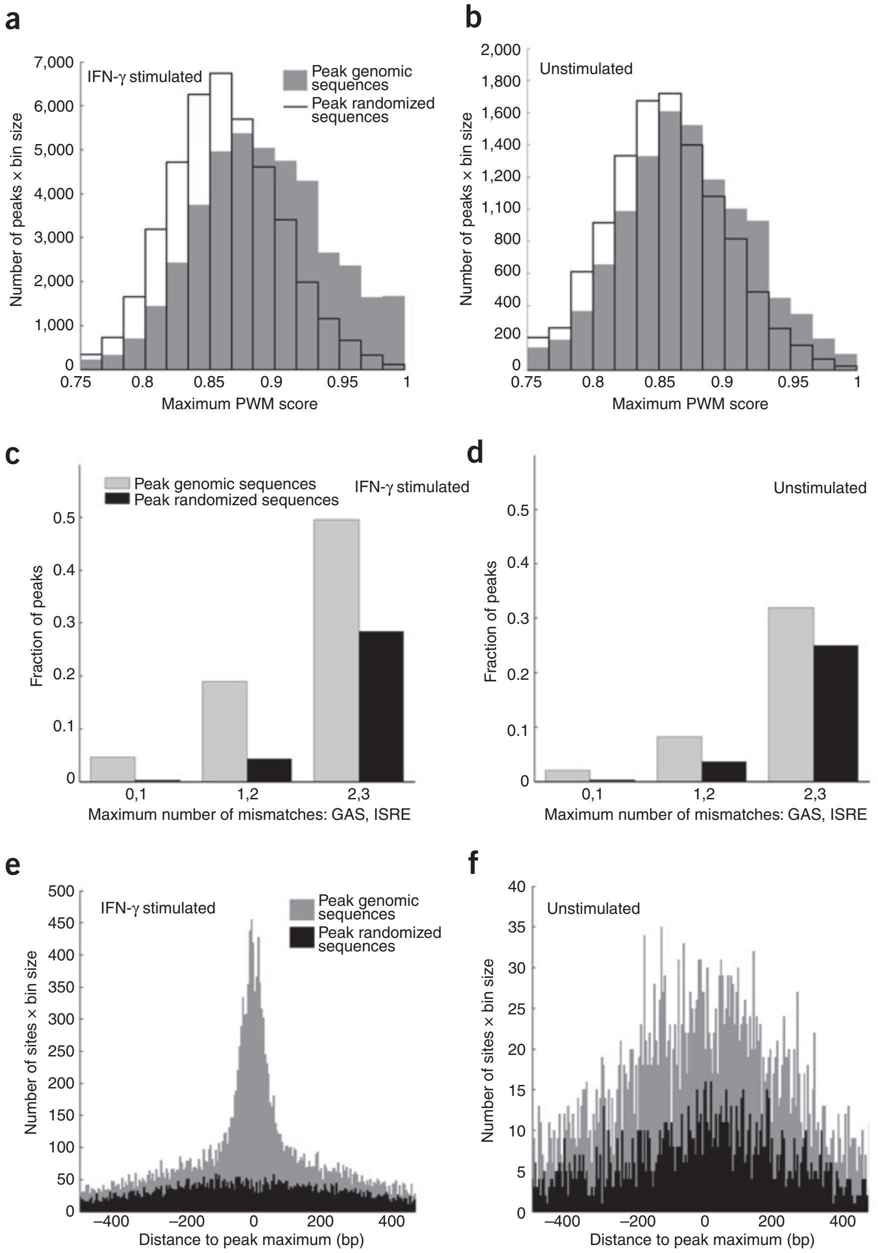

Figure 4 | Peak enrichment in sequences similar to STAT1 binding sites. (**a**,**b**) Scans with GAS- and ISRE-specific PWMs indicated that ChIP-seq peaks were enriched in sequences that were similar to known functional STAT1 binding sites. For 41,582 stimulated (**a**) and 11,004 unstimulated (**b**) peaks, the histograms show the distributions of the best PWM score for genomic and randomized peak sequences. (**c**,**d**) Scans with sequences known to be functional STAT1 binding sites indicated that peaks were enriched in such sequences, particularly close to peak maxima. Fraction of stimulated (**c**) and unstimulated (**d**) peaks whose genomic sequence contained at least one sequence that was similar to a known GAS or ISRE sequence (**Supplementary Data 2** and **Supplementary Figs. 6**, **7**). Results for three scanning mismatch stringencies are shown (left to right): '0, 1', 0 mismatches and up to 1 mismatch for a GAS and an ISRE site, respectively; '1, 2', up to 1 and 2 mismatches, and '2, 3', up to 2 and 3 mismatches. Only peaks with STAT1-like sequences within 100 bp of the peak maximum were considered (see also **Supplementary Fig. 11**). (**e**,**f**) For stimulated (**e**) and unstimulated (**f**) peaks and the medium stringency '1, 2' mismatch case, histograms of the distance from a GAS-like or ISRE-like sequence to the location of a peak's maximum height. Gray and black bars show results for genomic and randomized peak sequences.

genomes. Although it is difficult to compare relative costs of technology platforms, for large genomes we estimate that ChIP-seq may presently be at least an order of magnitude less costly than ChIP-chip. The eight independently loadable lanes in the flow cell of the 1G system make the approach flexible. Handling of physical devices is minimized, as a single flow cell the size of a microscope slide can deliver deeply sequenced whole-genome coverage. Less input material is required for ChIP-seq than for ChIP-chip (nanogram scale for ChIP-seq versus microgram scale for ChIP-chip, with the exact amount varying depending on the number of arrays in the set and on whether the ChIP DNA is amplified). Moreover, because ChIP-seq experiments do not depend on design and manufacture of tiled microarrays, these experiments can be done quickly on any species for which a high-quality genome sequence is available. Whereas reads can be mapped into regions that are annotated as repetitive, it appears that more work will be required to permit estimation of differences relative to ChIP-chip in this area. Deeper sequencing may permit identification of sites with lower binding affinity than is possible with microarrays. ChIP-seq may also permit detection of mutations within binding sites, which may be helpful in understanding changes in transcription factor binding and gene regulation in cancer cells.

A typical ChIP-chip half-height peak width is 500 bp (G.E. and M.S.; unpublished data). Although median widths for ChIP-seq peaks were ~800–900 bp, peak shapes generally resembled a Gaussian distribution, and the spatial resolution suggested by the proximity of sites and motifs to peak maxima is promising for computationally assessing conserved DNA sequence motifs associated with peaks.

We can compare our estimates of the sensitivity and specificity of ChIP-seq with those from ChIP-chip studies that used tiled oligonucleotide microarrays. A platform comparison of STAT1 targets in the ENCODE regions representing 1% of the human genome found a sensitivity of ~72% at a FDR of 0.26 (ref. 30). For RNA polymerase II ChIP in the ENCODE regions, sensitivity has been estimated as 83% and specificity as close to 100% (ref. 26). For CTCF using a whole-genome set of 38 microarrays, sensitivity has been estimated as ~88% at a FDR of 0.05 (ref. 31). We note that accurate estimates of sensitivity and specificity are difficult to obtain, and may vary depending on the biology of a given transcription factor as well as the quality and availability of reference sets of true positives and true negatives that are requisite as a basis for assessing targets. Concurrently with our work, others have also shown high performance of ChIP-seq by analyzing the binding sites of the neuron-restrictive silencer factor[32].

In the present work, we chose a 27-bp read length to balance sequencing throughput, cost, sequencing accuracy and the fraction of reads that can be mapped to a mammalian genome sequence. The data processing that transformed digital sequence reads into profiles was relatively simple. Numerical simulations and analytic calculations predicted that generating overlap profiles from XSETs rather than SETs should improve peak discrimination for low enrichment sites (data not shown). When an experiment offers an appropriate combination of high enrichment and deep sequencing, however, we anticipate that information from SET locations and orientations as well as SET overlap profiles may offer high spatial resolution information that can complement that available from XSET profiles.

To assess the number of false positive peaks that might arise from the immunoprecipitation, we carried out a ChIP-seq experiment in which we uniquely mapped 2 million reads from an immunoprecipitation using immunoglobulin gamma. Thresholding the XSET profile at an FDR of 0.001 identified only 355 peaks (data not shown).

In our hands, 1G experiments for various ChIP targets can now yield 5–9 million sequence reads for each of the eight lanes available in a flow cell, and approximately 60% of these reads will map to unique locations in the reference human genome. Using this methodology, adding more sequence data may identify a progressively larger number of significantly enriched sequence regions, which we anticipate may correspond to some combination of progressively weaker binding sites and sites that are active in a progressively smaller fraction of the input cell population. In support of this, we demonstrated that peaks representing a larger number of overlapped immunoprecipitated fragments were enriched in sequences similar to those known to be functional STAT1 sites, relative to peaks with fewer overlapped fragments. The biological significance of peaks in such deeply sequenced result sets has yet to be fully appreciated or determined.

METHODS

Illumina library construction and sequencing. To prepare immunoprecipated DNA for 1G sequencing, we size-fractionated 5–50 ng of immunoprecipitate by 12% PAGE and excised a gel slice containing the 100–300 bp fragments. We eluted DNA from the gel slice overnight at 4 °C in 300 μl of elution buffer (5:1, LoTE buffer (3 mM Tris-HCl (pH 7.5), 0.2 mM EDTA)-7.5 M ammonium acetate) and recovered the DNA using a QIAquick PCR purification kit (Qiagen). Then we repaired the DNA ends using a 1:5 mixture of T4 and Klenow DNA polymerases (Illumina) following the manufacturer's instructions. After a 30-min incubation at 20 °C, we subjected the reaction to phenol–chloroform–isoamyl alcohol (pH 8.0; 100 μl; Fisher) extraction in 0.5-ml phase-lock gel tubes (heavy; Eppendorf) and precipitated the reactions by adding 250 ml of 100% EtOH, 3 ml of mussel glycogen (Invitrogen) and 10 ml of 7.5 M NH_4OAc, and incubating at –20 °C for 20 min. We recovered the precipitate by centrifugation at 20,200*g* for 15 min at 4 °C in an benchtop refrigerated centrifuge (Eppendorf model 5417R). We added a single adenine base to the DNA using Klenow exo– (3′ and 5′ exo minus; Illumina) following the manufacturer's instructions. After a 30-min incubation at 37 °C, we subjected the reaction to phenol–chloroform–isoamyl alcohol extraction and precipitation, as described above. We washed the DNA pellet twice with 1 ml of 70% EtOH and resuspended in 10 ml of LoTE buffer. We ligated adapters (Illumina) to ends of the single adenine–tailed DNA following the manufacturer's instructions with the exception that we diluted the adaptor oligonucleotide mix (Illumina) 1/10 before use. We recovered the DNA using a QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions and using 30 μl of elution buffer heated to 50 °C. We enriched adaptor-modified DNA fragments by PCR using Phusion polymerase (Finnzymes) and PCR primer 1.1 and 2.1 (Illumina) following the manufacturer's instructions (separate 10- and 15-cycle reactions were run). After cycling, we purified the reactions using a QIAquick MiniElute kit (Qiagen) according to the manufacturer's instructions, with 15 μl of elution buffer heated to 50 °C. We quantified the DNA quality with an Agilent DNA 1000 series II assay and a Nanodrop 7500 spectrophotometer using a 1.5-μl aliquot that was diluted to 10 nM. We performed cluster generation and 27 cycles of sequencing on the Illumina cluster station and 1G analyzer following the manufacturer's instructions. We generated libraries for three biological replicates.

Processing 1G data. We extracted sequences from the resulting image files using the open source Firecrest and Bustard applications on a 32-CPU cluster running Red Hat Enterprise Linux 4 and Sun Microsystems Grid Engine 6. Reads were aligned (mapped) to the unmasked human reference genome (NCBI v36, hg18) using the Eland application (Illumina). Eland achieves high throughput by permitting no more than two mismatches per read sequence. We maximized the number of mapped reads by iteratively discarding two bases from the end of a rejected read and resubmitting the truncated read to Eland, until either all reads had been aligned to the genome or the lengths of all unmapped reads were less than 20 bp. Only uniquely mapped reads were retained. DNA fragments were represented by a mapped sequence read of length 27 bp. Because a relatively small number of PCR cycles was used in library preparation and DNA-fragment end locations were weakly clustered rather than random (data not shown), we allowed reads with identical start coordinates to be present in a profile. We combined uniquely aligned reads from the three biological replicates into a single set of reads. We transformed mapped sequence reads into profiles of the number of overlapped DNA fragments at each nucleotide in the reference genome. Because a 27-bp read directionally represents one end of a DNA fragment (SET), we approximated the fragment that produced a read sequence by extending the read to generate an XSET. We chose the XSET length to be the mean fragment length of the size selected DNA. From distances between mapped reads, we estimated this length to be 174 bp. A peak's height was the maximum number of overlapped XSETs for that peak. A random expectation for the probability of observing peaks with a particular height was generated from a numerical background model that generated peaks by randomly placing onto a hypothetical genome several fragments equivalent to the actual uniquely mapped number. Each fragment's length was the estimated mean fragment length, that is, the XSET length. Because 27-bp reads can be mapped uniquely to ~90% of the human genome (data not shown), the background simulations used a mappable genome length that was 90% of 3.08 Gb. For a peak height, we estimated the FDR as the ratio of the number of peaks that the background model indicated should occur by chance, to the number observed (**Supplementary Fig. 2**). For each profile, we chose a threshold peak height as the smallest height that was equivalent to FDR < 0.001 for peaks of that height. All peaks of at least this height were retained in the profile. A set of randomly located mapped DNA fragments is equivalent to a global coverage level λ, whose value is the product of the number and mean length of mapped fragments divided by the mappable genome length. Given a λvalue, the probability of observing a peak with a height of at least H is given by a sum of Poisson probabilities as:

$$1 - \sum_{k=0}^{H-1} \frac{e^{-\lambda}\lambda^k}{k!}$$

For an FDR-segmented set of peaks, we estimated the global fold enrichment as the fraction of tags in peaks divided by the fraction of the genome covered by peaks. We estimated whether the sequencing depth was sufficient to identify all peaks in a sample by computationally resampling random subsets of uniquely mapped reads to generate hypothetical result sets that approximated results from experiments done over a range of sequencing depths. For the peak set generated by each sampled set of reads, we carried out a standard FDR estimate and thresholded the peaks at FDR = 0.001. We then assessed whether any read subset had identified the same number of significant peaks as were identified by the actual total read set. When this occurred, we designated the minimum number of reads required to identify this maximum available number of peaks as the number required to saturate the experiment.

Chromatin immunoprecipitation. We prepared and verified STAT1 ChIP DNA as previously described[30]. The ChIP-chip and ChIP-seq data sets were each derived from three biological replicates with one biological sample used in common between the two methods.

Additional methods. Information about chromatin immunoprecipitation, ChIP-chip assays and comparisons with ChIP-seq, known functional STAT1 binding sites, peak enrichment for sequences similar to known STAT1 binding, peak overlap with conserved regions and tandem repeats, and comparing ChIP-seq peaks to ChIP-PCR results are available in **Supplementary Methods** online.

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

ACKNOWLEDGMENTS
We gratefully acknowledge the expert technical assistance of the Genome Sciences Centre Functional Genomics team. This study received funding support from Genome British Columbia (Platform and MORGEN project), Genome Canada and the BC Cancer Foundation. M.M. and S.J. are Scholars of the Michael Smith Foundation for Health Research. M.M. is a Terry Fox Young Investigator.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at
http://npg.nature.com/reprintsandpermissions

1. Bulyk, M.L. DNA microarray technologies for measuring protein-DNA interactions. *Curr. Opin. Biotechnol.* **17**, 422–430 (2006).
2. Orian, A. Chromatin profiling, DamID and the emerging landscape of gene expression. *Curr. Opin. Genet. Dev.* **16**, 157–164 (2006).
3. Yochum, G.S. *et al.* Serial analysis of chromatin occupancy identifies β-catenin target genes in colorectal carcinoma cells. *Proc. Natl. Acad. Sci. USA* **104**, 3324–3329 (2007).
4. Roh, T.Y., Cuddapah, S., Cui, K. & Zhao, K. The genomic landscape of histone modifications in human T cells. *Proc. Natl. Acad. Sci. USA* **103**, 15782–15787 (2006).
5. Roh, T.Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).
6. Kim, J., Bhinge, A.A., Morgan, X.C. & Iyer, V.R. Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* **2**, 47–53 (2005).
7. Wei, C.L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
8. Loh, Y.H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
9. Zeller, K.I. *et al.* Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc. Natl. Acad. Sci. USA* **103**, 17834–17839 (2006).
10. Ng, P. *et al.* Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.* **34**, e84 (2006).
11. Bentley, D.R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
12. Reich, N.C. & Liu, L. Tracking STAT nuclear traffic. *Nat. Rev. Immunol.* **6**, 602–612 (2006).
13. Lodige, I. *et al.* Nuclear export determines the cytokine sensitivity of STAT transcription factors. *J. Biol. Chem.* **280**, 43087–43099 (2005).
14. Schroder, K., Sweet, M.J. & Hume, D.A. Signal integration between IFNγ and TLR signalling pathways in macrophages. *Immunobiology* **211**, 511–524 (2006).
15. Vinkemeier, U. Getting the message across, STAT! Design principles of a molecular signaling circuit. *J. Cell Biol.* **167**, 197–201 (2004).
16. Brierley, M.M. & Fish, E.N. Stats: multifaceted regulators of transcription. *J. Interferon Cytokine Res.* **25**, 733–744 (2005).
17. Schroder, K., Hertzog, P.J., Ravasi, T. & Hume, D.A. Interferon-gamma: an overview of signals, mechanisms and functions. *J. Leukoc. Biol.* **75**, 163–189 (2004).
18. Pestka, S., Krause, C.D. & Walter, M.R. Interferons, interferon-like cytokines, and their receptors. *Immunol. Rev.* **202**, 8–32 (2004).
19. Ramana, C.V., Chatterjee-Kishore, M., Nguyen, H. & Stark, G.R. Complex roles of Stat1 in regulating gene expression. *Oncogene* **19**, 2619–2627 (2000).
20. Ehret, G.B. *et al.* DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites. *J. Biol. Chem.* **276**, 6675–6688 (2001).
21. Chatterjee-Kishore, M., van den Akker, F. & Stark, G.R. Association of STATs with relatives and friends. *Trends Cell Biol.* **10**, 106–111 (2000).
22. Chatterjee-Kishore, M., Wright, K.L., Ting, J.P. & Stark, G.R. How Stat1 mediates constitutive gene expression: a complex of unphosphorylated Stat1 and IRF1 supports transcription of the LMP2 gene. *EMBO J.* **19**, 4111–4122 (2000).
23. Hubbard, T.J. *et al.* Ensembl 2007. *Nucleic Acids Res.* **35**, D610–D617 (2007).
24. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
25. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
26. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
27. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
28. Hartman, S.E. *et al.* Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes Dev.* **19**, 2953–2968 (2005).
29. ENCODE Project Consortium. The ENCODE (encyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).
30. Euskirchen, G.M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array and sequencing based technologies. *Genome Res.* (in the press).
31. Kim, T.H. *et al.* Analysis of the vertebrate insulator protein CTCF binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
32. Johnson, D.S. *et al.* Genome-wide mapping of the *in vivo* protein-DNA interactions. *Science*; published online 31 May, 2007.