Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data

Anton Valouev[1,4], David S Johnson[2,4], Andreas Sundquist[3], Catherine Medina[2], Elizabeth Anton[2], Serafim Batzoglou[3], Richard M Myers[2] & Arend Sidow[1,2]

Molecular interactions between protein complexes and DNA mediate essential gene-regulatory functions. Uncovering such interactions by chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-Seq) has recently become the focus of intense interest. We here introduce quantitative enrichment of sequence tags (QuEST), a powerful statistical framework based on the kernel density estimation approach, which uses ChIP-Seq data to determine positions where protein complexes contact DNA. Using QuEST, we discovered several thousand binding sites for the human transcription factors SRF, GABP and NRSF at an average resolution of about 20 base pairs. MEME motif-discovery tool–based analyses of the QuEST-identified sequences revealed DNA binding by cofactors of SRF, providing evidence that cofactor binding specificity can be obtained from ChIP-Seq data. By combining QuEST analyses with Gene Ontology (GO) annotations and expression data, we illustrate how general functions of transcription factors can be inferred.

Chromatin immunoprecipitation (ChIP) has become an important assay for the genome-wide study of protein-DNA interactions and gene regulation[1–3]. In a typical ChIP experiment, protein complexes that contact DNA are cross-linked to their binding sites, the chromatin is sheared into short fragments, and then the specific DNA fraction that interacts with the protein of interest is isolated by immunoprecipitation. A genome-wide readout of the protein binding sites is produced either by hybridization of the DNA pool to a tiling array (ChIP-chip[4]) or by end-sequencing of millions of different DNA fragments (ChIP-Seq[5–9]). In higher organisms, particularly mammals, ChIP-chip data tend to have low resolution and are often quite noisy[10], two shortcomings that ChIP-Seq promises to surmount. As a consequence, ChIP-chip is being rapidly displaced by ChIP-Seq in genome-wide discovery of mammalian transcription factor binding sites (TFBSs).

The goal of ChIP-Seq data analyses is to find those genomic regions that are enriched in a pool of specifically precipitated DNA fragments. Regions of high sequencing read density are referred to as peaks. The output of software implementing peak-finding methodology is a list of 'peak calls' that comprises the genomic locations of sites inferred to be occupied by the protein. To date, studies that have presented ChIP-Seq data[5,6] used peak-finding methodology that heuristically quantifies read density but does not take full advantage of certain important properties of the data such as the directionality of sequencing reads. The growing importance of ChIP-Seq demands development of rigorous and transparent statistical approaches that fully leverage the inherent advantages of ChIP-Seq.

We here describe QuEST, a ChIP-Seq data analysis method that is based on realistic statistical modeling of the ChIP-Seq experimental approach. QuEST generates peak calls with substantial power and resolution by leveraging key attributes of the sequencing data such as directionality of reads and the size of fragments that were sequenced (which, notably, is estimated from the data themselves rather than provided by the user). QuEST achieves the desired balance between sensitivity and specificity by calculating false-discovery rates from controls that are routinely conducted as part of ChIP experiments. Underlying QuEST's statistical framework is the kernel density estimation approach[11], which facilitates aggregation of signal originating from densely packed sequencing reads at the TFBSs, leading to statistically robust peak calls.

To demonstrate the power and resolution of analyses facilitated by QuEST, we generated ChIP-Seq data for three functionally different human transcriptional regulatory proteins that have well-defined binding specificities and regulatory roles. Growth-associated binding protein (GABP) and serum response factor (SRF) are thought to function primarily as transcriptional activators[12–18], and neuron-restrictive silencer factor (NRSF) is a transcriptional repressor[19,20]. We applied QuEST to these data as part of a larger workflow that also included MEME-based motif discovery and, in the case of SRF, identification of cofactor motifs that were indicative of cofactor interactions. Finally, we analyzed the ChIP-Seq data together with microarray results and GO terms to describe the function of the transcription factors.

[1]Department of Pathology and [2]Department of Genetics, Stanford University Medical Center, 300 Pasteur Drive, Stanford, California 94305, USA. [3]Department of Computer Science, Stanford University, 318 Campus Drive, Stanford, California 94305, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.S. (arend@stanford.edu).

RECEIVED 6 MARCH; ACCEPTED 23 JULY; PUBLISHED ONLINE 17 AUGUST 2008; DOI:10.1038/NMETH.1246

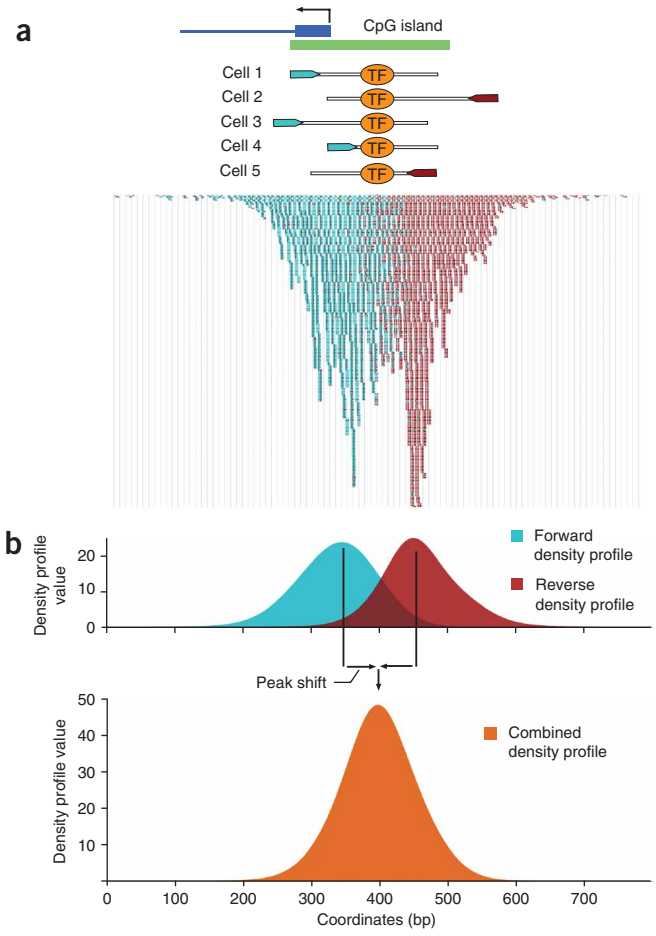

Figure 1 | QuEST's representation of ChIP-Seq data using density profiles. (**a**) GABP ChIP-Seq reads from the promoter and CpG island of the gene encoding nitric oxide synthase–interacting protein. Hypothetical schematic of GABP binding in five cells to the corresponding DNA fragments with sequencing reads marked in blue (forward) or red (reverse). Actual read data are shown below. (**b**) Forward and reverse read density profiles derived from the read data (top) contribute to the CDP (bottom). The zero *x*-coordinate corresponds to position 54775300 of human chromosome 19 (US National Center for Biotechnology Information (NCBI) build 36). One area of sequence read enrichment from the genome-wide profiles is shown for illustration.

RESULTS

Analytical framework

QuEST requires data in the form of genome coordinates ('tags') obtained from mapping several million sequencing reads to a reference genome. Tags from forward and reverse reads cluster on opposite sides of the TFBS (**Fig. 1a**) This is because sequencing proceeds from one end of the fragment toward its middle in a strand-specific manner, which leads to an underrepresentation of tags in the immediate proximity of the TFBS.

QuEST first constructs two separate profiles, one for forward and one for reverse tags. These profiles are characterized by areas of strong enrichment where tags are particularly dense (**Fig. 1**). The distance between forward and reverse profiles is not known a priori, but it is important to account for it and to correctly combine the two separate profiles into one. As this distance may vary considerably from experiment to experiment, QuEST estimates it from a particularly robust subset of the data. We refer to half of this distance as the 'peak shift'.

Once the experiment-specific peak shift has been estimated, the forward and reverse profiles are shifted and summed to produce the combined density profile (CDP) on which all subsequent analyses are carried out (**Fig. 1b** and **Supplementary Fig. 1** online). By combining the profiles in this manner, QuEST accomplishes two key aspects of ChIP-Seq analysis: first, the signals from reverse and forward reads are represented by a single classifier; second, local maxima of this classifier correspond to protein-DNA cross-linking points, providing an estimate for the location of the TFBS.

QuEST then searches the CDP for enriched loci in a process referred to as 'peak calling'. Specifically, QuEST identifies candidates for CDP peaks as positions in the reference genome corresponding to local maxima of the CDP with sufficient enrichment compared to the control data. The strongest of these are likely to be due to real binding events, whereas weaker-scoring peaks may be false positives, requiring the setting of a CDP threshold for peak calling. As this threshold may vary considerably between experiments, the desired balance between sensitivity and specificity can be achieved by a calibration procedure. Briefly, we separated the negative control data into two sets, one of which we used as a pseudo-ChIP sample in which peaks are to be predicted and the other of which we used as a background for this sample. Any peak that is predicted in this comparison is a false positive. Hence, the false-discovery rate estimate is the ratio of the number of peaks predicted in the pseudo-ChIP analysis to the number of peaks identified in the real ChIP experiment. This approach allows the user to set specific thresholds and determine the false-discovery rate, or vary the thresholds until a desired false-discovery rate is achieved (**Supplementary Fig. 2** online).

As a final result, for each peak in the list of high-confidence peak calls, QuEST reports a score quantifying the tag enrichment at the peak and a genome coordinate that corresponds to the position of that peak. Each such coordinate is a predictor of the position of a binding event, likely an endogenous TFBS occupied by the immunoprecipitated transcription factor. The kernel density estimation–derived score QuEST reports for each peak is proportional to the frequency at which the TFBS was present in the sequenced library. Because the score reflects the amount of evidence for the peak, QuEST ranks the final peak calls accordingly.

Performance of QuEST

To evaluate key aspects of the performance of QuEST, we generated five ChIP-Seq libraries from the human Jurkat cell line and sequenced them using the Solexa platform (**Table 1**). One library

Table 1 | ChIP-Seq data and analysis summary

	GABP	SRF	NRSF (polyclonal antibody)	NRSF (monoclonal antibody)
Number of aligned ChIP reads	7,862,231	8,721,730	8,813,398	5,358,147
Number of peaks called by QuEST	6,442	2,429	2,960	2,596
False-discovery rate estimate	1/6,442	1/2,429	<1/2,960	1/2,595
Percent peaks near genes (<2 kb or within genes)	83	72	53	53

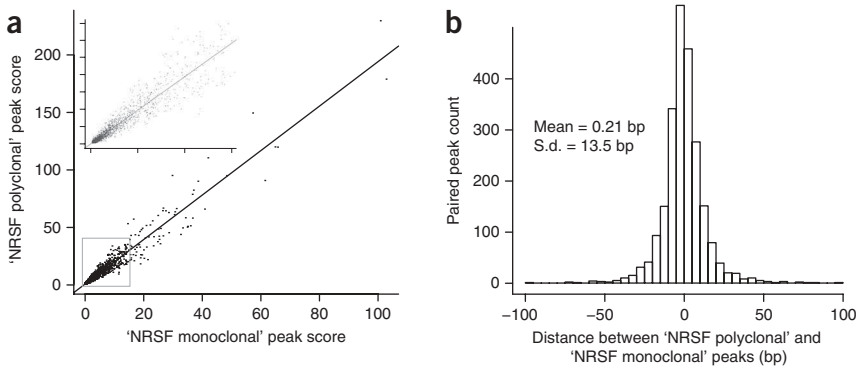

Figure 2 | Reproducibility and robustness of QuEST results assessed by comparison between two independent NRSF datasets. (**a**) Correlation between peak scores for data collected using polyclonal and monoclonal antibodies to NRSF ($r = 0.97$) with the inset expanding the portion near the graph origin. (**b**) Distribution of the distance between 'NRSF polyclonal' and 'NRSF monoclonal' peak call positions.

each was from ChIPs against the transcriptional activators GABP and SRF, two were from ChIPs against the transcriptional repressor NRSF (one using a polyclonal and the other a monoclonal antibody), and the last was a negative control library (reversed cross-links, no immunoprecipitation or RX-noIP). We generated 7.9, 8.7, 8.8 and 5.4 million mapped sequence tags for GABP, SRF, NRSF polyclonal and NRSF monoclonal datasets, respectively (**Table 1**), as well as 17.4 million mapped tags for the RX-noIP library. QuEST identified 6,442 (GABP), 2,429 (SRF), 2,960 (NRSF polyclonal) and 2,596 (NRSF monoclonal) CDP peak positions with significant enrichment of ChIP sequencing reads (we defined significance by the experiment-specific false discovery rates; **Table 1**). Saturation analysis indicated that these libraries were sequenced to sufficient depth to identify the majority of peaks (**Supplementary Fig. 3** online).

To test robustness and reproducibility of QuEST's ability to accurately quantify tag enrichment, we compared QuEST scores of the 2,320 peaks that were in common for the two NRSF datasets. These scores were strongly correlated ($r = 0.97$; **Fig. 2a**). The mean distance between corresponding peaks from the two datasets was 0.2 bp, with a s.d. of 13.5 bp (**Fig. 2b**), demonstrating highly reproducible peak call positions.

We identified previously described transcriptional targets of GABP, SRF and NRSF to provide some validation for the peaks identified by QuEST in this study. These include GABP-regulated interleukin-16 (IL16)[12], cytochrome *c* oxidase subunits IV and Vb[12], and SRF-regulated FHL2 (ref. 21). QuEST also identified three peaks in the autoregulated SRF gene[16], one in the promoter and two in one of the introns. Finally, the genes *Calb1*, *Bdnf*, *Syt4* and *Nav1* are NRSF targets in mouse embryonic stem cells[20], and their orthologs were also 'marked' by peaks in our data.

Theoretically, the genomic coordinate QuEST reports for each of its peaks should be 'marked' by the canonical TFBS motif. We first determined canonical motifs and their corresponding position specific scoring matrices (PSSMs) using the *de novo* motif finder MEME[22]. For each ChIP-Seq experiment, the input data into MEME was the set of 200 bp sequences from around each peak call. The resulting motifs closely corresponded to the previously established canonical recognition sites for each of the three factors[12,15,23]. To then determine the specific positions of motifs within the peak call regions, we searched for matches of the PSSMs in the 200 bp around each peak, using a log-odds-ratio approach and a stringent threshold. The majority of peaks contained one or more significant PSSM matches, which we used to evaluate the resolution of QuEST peak calls. Remarkably, the mean distance between peak call and motif ranged from 0.1 bp in the NRSF monoclonal set to 2.55 bp for GABP, with the s.d. ranging from 13.4 bp to 21.8 bp (**Fig. 3**).

Leveraging QuEST peak calls for biological insight

Our MEME analysis found that the canonical motifs of each transcription factor were most notably enriched in their respective CDP peaks (**Fig. 4**). Canonical motifs explain 71% (GABP), 33% (SRF) and 69% (NRSF) of the peaks after accounting for motifs that are expected to occur by chance (**Supplementary Fig. 4** online), illustrating QuEST's high specificity in TFBS discovery. The comparatively low fraction of CDP peaks in the SRF dataset explained by the presence of a canonical motif is likely due to interactions with cofactors. GABP and SRF, both of which assemble into a complex with a pair of DNA binding subunits[12,16], most frequently contain two motifs (**Fig. 4**), in contrast to NRSF peaks, which typically contain one motif.

For SRF, the initial MEME analysis also revealed the presence of the SP1 motif. Presence of this motif explains a substantial fraction of peaks (48%), providing evidence that the previously suggested interaction[24] between SP1 and SRF is common.

We also conducted a second round of MEME analyses focusing only on those peak-associated sequences that did not contain the canonical SRF motif. Such peaks may be due to indirect DNA binding of the targeted factor via a different, interacting, DNA-binding protein. This analysis yielded an additional significant motif that resembled the recognition site of the Ets family of factors. This motif explained the presence of an additional 17% of the SRF peaks.

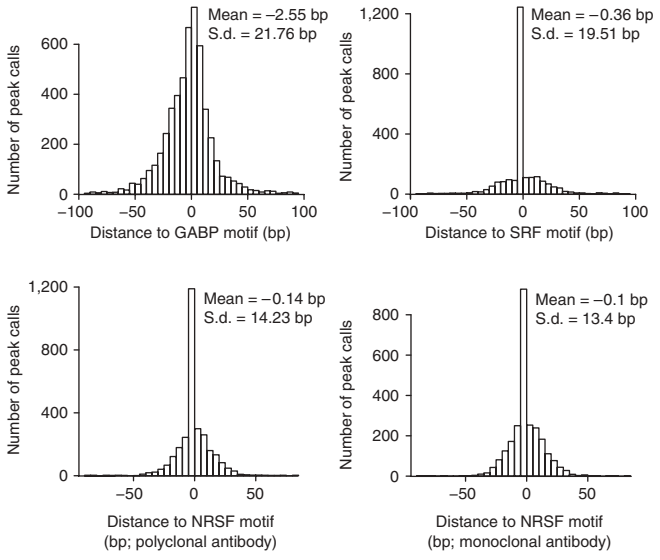

Figure 3 | Resolution of QuEST as quantified by the distance between CDP peak calls and TFBS motif centers. Histograms in each panel represent the distribution of peak distances to the nearest high-scoring motif.

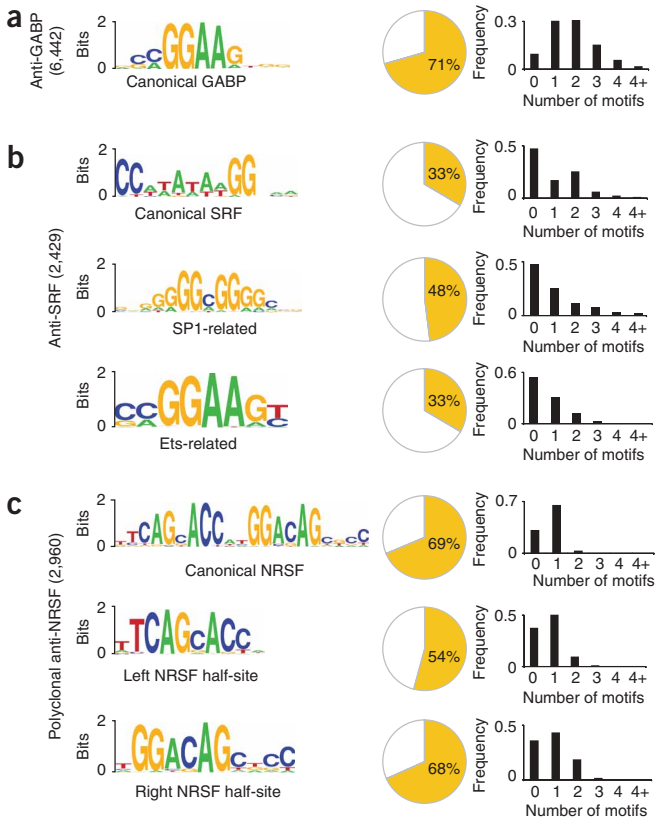

Figure 4 | Motif analysis results. (**a**) GABP. (**b**) SRF. (**c**) NSRF (ChIP with polyclonal antibody). For each of the three transcription factors, significantly overrepresented motifs are graphically depicted (Weblogos[27]). Number of peak-associated sequences on which the analysis was conducted is given in parentheses. Pie charts show the fraction of CDP peaks with motifs in close proximity to the peak (<100 bp). Histograms show the distribution of the motif number within 100 bp of the peak.

The prevalence of an Ets-like motif may be due to the previously described interaction between SRF and the Ets factor ELK4 (refs. 17,25). Notably, the antibody to SRF has no detectable cross-reactivity with other proteins as determined by western blot analysis (data not shown). We applied the same strategy to the NRSF dataset, which reproduced the discovery of NRSF half-sites previously reported (**Fig. 4**), and resulted in an additional 16% of peaks explained. We found no other significant motifs for GABP.

We observed that a large fraction of SRF peaks (29%) occurred within 100 bp of GABP peaks, and NRSF peaks almost never coincided with either SRF or GABP peaks. The close proximity of SRF and GABP peaks might suggest that SRF not only physically interacts with the Ets factor ELK4 (ref. 17) but, in some promoters, with GABP as well.

QuEST analyses can be combined with orthogonal genome-wide data or resources such as GO to provide general insights into the functions of proteins targeted by ChIP-Seq experiments. For both GABP and SRF, a majority of peak calls (83% and 72%, respectively) were within 2 kb of a gene. By contrast, only 53% of NRSF peak calls were within 2 kb of a gene, suggesting that NRSF's effects on gene regulation are, on average, exerted over longer distances than those of GABP and SRF. Having obtained a set of peak-associated genes, we then conducted gene expression profiling and GO analyses.

Gene expression profiling revealed that NRSF-associated genes were expressed at significantly lower relative levels than the average of all genes (Wilcoxon test, $P < 2.2 \times 10^{-16}$; $n_{NRSF} = 1{,}274$, $n_{all} = 20{,}588$). This result is consistent with NRSF's known general function as a transcriptional repressor and with previous results[5]. By contrast, both SRF-associated genes and GABP-associated genes were expressed significantly higher than the average gene (Wilcoxon test, $P < 2.2 \times 10^{-16}$; $n_{SRF} = 1{,}936$, $n_{GABP} = 5{,}394$, $n_{all} = 20{,}588$; **Supplementary Fig. 5** online), which is consistent with their activator functions[12,15].

GO analysis[26] (**Supplementary Tables 1–3** online) revealed that NRSF-associated genes are mostly involved in neuronal function, which is consistent with previous results[5]. Both SRF and GABP had significant enrichment of genes that are involved in basic cellular processes, particularly those related to gene expression. These results are consistent with both GABP and SRF being fundamental regulators of basic cell biology, rather than specialized factors with specific physiological roles. GABP is the more broadly acting of the two factors, as reflected by its almost threefold larger number of peaks and associated genes.

DISCUSSION

The high resolution of QuEST peak calls is noteworthy. For example, 89% of CDP peaks that contained a matching canonical TFBS motif in the NRSF polyclonal data were within 25 bp of the motif center and 56% were within 10 bp (**Fig. 3**). QuEST thereby brings within reach the ability to identify at high confidence the precise locations at which DNA binding proteins interact with the genome.

The score QuEST generates for each peak, according to which the peaks are ranked, is directly proportional to the amount of tag enrichment in the set of DNA fragments that yielded the sequences. Thus, a peak with a score of 50 is due to a TFBS that was twice as abundant in the DNA sample as a TFBS with a peak score of 25. Although both scores may be above the reporting cutoff chosen (by the desired false-discovery rate), and are therefore considered real, there is twice the support for (and hence the confidence in) the stronger peak.

One potential drawback of QuEST is that it does not convert peak scores into definitive *P* values. Instead, the stringency of peak calls is determined by the score threshold at which the peaks are reported, and the false-discovery rate is calculated for this threshold. Users can either use the default threshold or specify their own and assess the stringency through the false-discovery rate.

Model-free analysis as implemented in QuEST may be considered less powerful than approaches that leverage the additional power of an explicit model for the ChIP-Seq data. However, such explicit modeling will likely be elusive in the near future because of the many experimental and biological factors that influence the eventual enrichment signal that is detected by ChIP-Seq. Some part of the enrichment signal should reflect occupancy by the transcription factor, but confounding factors such as antibody specificity, epitope accessibility and susceptibility of TFBS-adjacent DNA to shearing will be difficult to model explicitly. Furthermore, downstream manipulation necessary for library building, especially library amplification and sequencing, introduce additional biases into the enrichment signal. Together, these factors contribute to increased variance of signal strength across the binding sites and complicate detection of weak binding signals. Application of

QuEST or similar approaches will enhance our empirical understanding of ChIP-Seq data during the next few years.

METHODS

Density profiles. Individual density profiles for forward and reverse reads at any position i of the genome are given by

$$H_{+;-}(i) = \frac{1}{h} \sum_{j=i-3h}^{i+3h} K((j-i)/h) \times C_{+;-}(j),$$

where h is the kernel density bandwidth (we used $h = 30$ bases), $K(x) = \exp(-x^2 / 2) / (2\pi)^{0.5}$ is the Gaussian kernel density function, and $C_{+;-}(j)$ gives the number of 5′ read ends at position j for forward and reverse reads, respectively. In contrast to the original kernel density estimator[11], our density profiles represent unnormalized density estimates in which the sum is limited to sample points proximal to any given position (within 3 kernel density estimation bandwidths). These modifications were done for computational convenience (**Supplementary Methods** online).

The CDP used in actual peak calling is calculated according to the formula $H(i) = H_+(i - \lambda) + H_-(i + \lambda)$, where λ is a peak shift parameter estimate, and H_+ and H_- are the positive-strand and negative-strand strand density profiles as defined above.

Peak shift estimation. For regions in which the number of tags exceeded 600 in a window of 300 bp, we calculated forward and reverse profiles and recorded local maxima. Regions for which the highest scoring local maximum was 20-fold or greater than the next scoring maximum, for both negative and positive strands, and for which the enrichment in the ChIP sample was at least 20-fold over that for the RX-noIP sample, were selected. The peak shift parameter value was calculated as half of the average distance between peaks on the negative and positive strand. This estimate was robust across all 4 ChIP datasets (**Supplementary Fig. 6** online) and was highly concordant for the two NRSF datasets.

Peak calling. Candidate peaks were identified where the QuEST score profile achieved a local maximum within a 21 bp window, provided their QuEST score was above the ChIP threshold, which we determined in conjunction with the false-discovery rate procedure described below. Within each region, local peaks were identified. A peak was eliminated when the lowest point between it and the adjacent higher peak was greater than 0.9 times the CDP value of the higher peak. The remaining peaks were reported as 'calls' if (i) the value of background CDP was lower than the background CDP threshold or (ii) the ratio of the of ChIP CDP to the background CDP exceeded a specified threshold (referred to as the 'rescue ratio').

False discovery rate estimate for the number of peaks. For each experiment, the RX-noIP data were split into two datasets, one of which served as a pseudo-ChIP dataset (and matched the ChIP data in the number of reads) and the other served as the background set. Then CDPs were calculated for ChIP, pseudo ChIP and background datasets. Using the same score thresholds and rescue ratios, peaks were called in the ChIP and pseudo-ChIP datasets independently by comparison to the background data. The number of called peaks in the pseudo-ChIP data were the false-discovery number, and the false-discovery rate was the false-discovery number divided by the number of peaks called in the ChIP-Seq experiment. For identification of peaks that we used in subsequent MEME analyses, a rescue ratio of 10 was used for all datasets, and for each dataset the score threshold was set such that the false-discovery number was 1.

MEME analyses. For motif identification we extracted, for each dataset separately, 'peak-associated sequences' that comprised the set of 200 bp sequences surrounding each peak call. MEME was then applied with all default parameters to yield overrepresented motifs in each dataset. To identify alternative motifs in the SRF and NRSF data, a log-of-odds threshold of 3.0 was used to remove the peaks containing canonical motifs in the 200 bp window around the peak, after which MEME was applied again (**Supplementary Methods**).

MAST analyses. The number of peaks explained by a particular motif was generated by taking the maximum of the difference between the total number of peaks containing a motif and the number that could be explained by chance at a range of stringencies (*E* values; **Supplementary Methods**) using the MEME tool MAST. For MAST curves, see **Supplementary Figure 4**.

Additional methods. Descriptions of ChIP-Seq library construction and sequencing, gene expression analysis, density profile generation, peak calling and MEME-based motif discovery are available in **Supplementary Methods**.

Software availability. QuEST software is freely available for nonprofit use at http://mendel.stanford.edu/sidowlab/downloads/quest/. All data presented in this study (Rx-noIP and ChIP-Seq data, and peak call coordinates) are available at the same website.

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

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health grants 5 U01 HG003162 and 1 U54-HG004576 to R.M.M., and by funds from the Stanford Pathology and Genetics Ultra-High Throughput Sequencing Initiative. We thank L. Tsavaler for performing the Illumina expression analysis, W.H. Wong, K. McCue and members of Sidow lab for valuable discussions and suggestions.

AUTHOR CONTRIBUTIONS

A.V., S.B., An.S. and Ar.S. conceived the QuEST peak calling concept and developed the preliminary statistical framework. A.V. further developed and refined the statistical framework, and implemented QuEST. R.M.M. and D.S.J. devised the ChIP experiments. D.S.J., C.M. and E.A. performed the ChIP experiments. A.V. applied QuEST to the sequence data, and produced all quantitative results. A.V. and Ar.S. wrote the manuscript. A.V., D.S.J., S.B., R.M.M. and Ar.S. edited the manuscript.

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

1. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
2. Pokholok, D.K., Zeitlinger, J., Hannett, N.M., Reynolds, D.B. & Young, R.A. Activated signal transduction kinases frequently occupy target genes. *Science* **313**, 533–536 (2006).
3. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
4. Lieb, J.D. Genome-wide mapping of protein-DNA interactions by chromatin immunoprecipitation and DNA microarray hybridization. *Methods Mol. Biol.* **224**, 99–109 (2003).

5. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
6. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
7. Mardis, E.R. ChIP-seq: welcome to the new frontier. *Nat. Methods* **4**, 613–614 (2007).
8. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
9. Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nat. Methods* **5**, 19–21 (2008).
10. Johnson, D.S. *et al.* Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
11. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**, 1065–1076 (1962).
12. Rosmarin, A.G., Resendes, K.K., Yang, Z., McMillan, J.N. & Fleming, S.L. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. *Blood Cells Mol. Dis.* **32**, 143–154 (2004).
13. Lin, J.M. *et al.* Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res.* **17**, 818–827 (2007).
14. Cen, B., Selvaraj, A. & Prywes, R. Myocardin/MKL family of SRF coactivators: key regulators of immediate early and muscle specific gene expression. *J. Cell. Biochem.* **93**, 74–82 (2004).
15. Posern, G. & Treisman, R. Actin' together: serum response factor, its co-factors and the link to signal transduction. *Trends Cell Biol.* **16**, 588–596 (2006).
16. Pipes, G.C., Creemers, E.E. & Olson, E.N. The myocardin family of transcriptional coactivators: versatile regulators of cell growth, migration, and myogenesis. *Genes Dev.* **20**, 1545–1556 (2006).
17. Cooper, S.J., Trinklein, N.D., Nguyen, L. & Myers, R.M. Serum response factor binding sites differ in three human cell types. *Genome Res.* **17**, 136–144 (2007).
18. Collins, P.J., Kobayashi, Y., Nguyen, L., Trinklein, N.D. & Myers, R.M. The ets-related transcription factor GABP directs bidirectional transcription. *PLoS Genet.* **3**, e208 (2007).
19. Schoenherr, C.J. & Anderson, D.J. Silencing is golden: negative regulation in the control of neuronal gene transcription. *Curr. Opin. Neurobiol.* **5**, 566–571 (1995).
20. Ballas, N., Grunseich, C., Lu, D.D., Speh, J.C. & Mandel, G. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell* **121**, 645–657 (2005).
21. Philippar, U. *et al.* The SRF target gene Fhl2 antagonizes RhoA/MAL-dependent activation of SRF. *Mol. Cell* **16**, 867–880 (2004).
22. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* 28–36 (AAAI Press, Menlo Park, California, 1994).
23. Schoenherr, C.J., Paquette, A.J. & Anderson, D.J. Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl. Acad. Sci. USA* **93**, 9881–9886 (1996).
24. Madsen, C.S., Regan, C.P. & Owens, G.K. Interaction of CArG elements and a GC-rich repressor element in transcriptional regulation of the smooth muscle myosin heavy chain gene in vascular smooth muscle cells. *J. Biol. Chem.* **272**, 29842–29851 (1997).
25. Buchwalter, G., Gross, C. & Wasylyk, B. Ets ternary complex transcription factors. *Gene* **324**, 1–14 (2004).
26. Mortazavi, A., Leeper Thompson, E.C., Garcia, S.T., Myers, R.M. & Wold, B. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Res.* **16**, 1208–1221 (2006).
27. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).