Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson, *et al.*

Science **316**, 1497 (2007);
DOI: 10.1126/science.1141319

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

sponse in cutaneous CHS. The finding that topical application of Δ^9-THC reduced allergic inflammation points to the promising potential of developing pharmacological treatments (*24*) with the use of selective CB receptor agonists or FAAH inhibitors.

References and Notes

1. S. Grabbe, T. Schwarz, *Immunol. Today* **19**, 37 (1998).
2. T. Bisogno, A. Ligresti, V. Di Marzo, *Pharmacol. Biochem. Behav.* **81**, 224 (2005).
3. M. M. Ibrahim *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 3093 (2005).
4. M. Maccarrone *et al.*, *J. Biol. Chem.* **278**, 33896 (2003).
5. Materials and methods are available as supporting material on *Science* Online.
6. A. Zimmer, A. M. Zimmer, A. G. Hohmann, M. Herkenham, T. I. Bonner, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5780 (1999).
7. N. E. Buckley *et al.*, *Eur. J. Pharmacol.* **396**, 141 (2000).
8. S. Werner *et al.*, *Science* **266**, 819 (1994).
9. J. Knop, R. Stremmer, C. Neumann, E. De Maeyer, E. Macher, *Nature* **296**, 757 (1982).
10. R. I. Lehrer, J. Hanifin, M. J. Cline, *Nature* **223**, 78 (1969).
11. C. Nathan, *Nat. Rev. Immunol.* **6**, 173 (2006).
12. S. Oka *et al.*, *J. Immunol.* **177**, 8796 (2006).
13. Y. Ueda, N. Miyagawa, T. Matsui, T. Kaya, H. Iwamura, *Eur. J. Pharmacol.* **520**, 164 (2005).
14. V. Di Marzo *et al.*, *J. Neurochem.* **75**, 2434 (2000).
15. S. Stander, M. Schmelz, D. Metze, T. Luger, R. Rukwied, *J. Dermatol. Sci.* **38**, 177 (2005).
16. C. A. Lunn *et al.*, *J. Pharmacol. Exp. Ther.* **316**, 780 (2006).
17. B. F. Cravatt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9371 (2001).
18. B. F. Cravatt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10821 (2004).
19. J. S. Lee, G. Katari, R. Sachidanandam, *BMC Bioinformat.* **6**, 189 (2005).
20. M. Ashburner *et al.*, *Nat. Genet.* **25**, 25 (2000).
21. M. F. Bachmann, M. Kopf, B. J. Marsland, *Nat. Rev. Immunol.* **6**, 159 (2006).
22. A. de Paulis *et al.*, *Int. Arch. Allergy Immunol.* **124**, 146 (2001).
23. D. D. Taub *et al.*, *J. Clin. Invest.* **95**, 1370 (1995).
24. T. W. Klein, *Nat. Rev. Immunol.* **5**, 400 (2005).

25. This work was supported by grants from the Deutsche *Forschungsgemeinschaft* [SFB645 and GRK804 (to M.K. and A.Z.) and Tu90/5-1 (to T.T.)], by a Bonfor stipend to E.G., and a grant from Epitech S.r.l. to V.D.M. We thank M. Krampert for her help in wound healing experiments, L. Cristino for her help with immunohistochemistry, and J. Essig, A. Zimmer, E. Erxlebe, and I. Heim for technical assistance.

Supporting Online Material

www.sciencemag.org/cgi/content/full/316/5830/1494/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 and S2
References

8 March 2007; accepted 4 May 2007
10.1126/science.1142265

Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,[1]* Ali Mortazavi,[2]* Richard M. Myers,[1]† Barbara Wold[2,3]†

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position [±50 base pairs (bp)], which facilitated our finding motifs and allowed us to identify noncanonical NRSF-binding motifs. These ChIPSeq data also have high sensitivity and specificity [ROC (receiver operator characteristic) area ≥ 0.96] and statistical confidence ($P < 10^{-4}$), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.

Although much is known about transcription factor binding and action at specific genes, far less is known about the composition and function of entire factor-DNA interactomes, especially for organisms with large genomes. Now that human, mouse, and other large genomes have been sequenced, it is possible, in principle, to measure how any transcription factor is deployed across the entire genome for a given cell type and physiological condition. Such measurements are important for systems-level studies because they provide a global map of candidate gene network input connections. These direct physical interactions between transcription factors or cofactors and the chromosome can be detected by chromatin immunoprecipitation (ChIP) (*1*). In ChIP experiments, an immune reagent specific for a DNA binding factor is used to enrich target DNA sites to which the factor was bound in the living cell. The enriched DNA sites are then identified and quantified.

For the gigabase-size genomes of vertebrates, it has been difficult to make ChIP measurements that combine high accuracy, whole-genome completeness, and high binding-site resolution. These data-quality and depth issues dictate whether primary gene network structure can be inferred with reasonable certainty and comprehensiveness, and how effectively the data can be used to discover binding-site motifs by computational methods. For these purposes, statistical robustness, sampling depth across the genome, absolute signal and signal-to-noise ratio must be good enough to detect nearly all in vivo binding locations for a regulator with minimal inclusion of false-positives. A further challenge in genomes large or small is to map factor-binding sites with high positional resolution. In addition to making computational discovery of binding motifs feasible, this dictates the quality of regulatory site annotation relative to other gene anatomy landmarks, such as transcription start sites, enhancers, introns and exons, and conserved noncoding features (*2*). Finally, if high-quality protein-DNA interactome measurements can be performed routinely and at reasonable cost, it will open the way to detailed studies of interactome dynamics in response to specific signaling stimuli or genetic mutations. To address these issues, we turned to ultrahigh-throughput DNA sequencing to gain sampling power and applied size selection on immuno-enriched DNA to enhance positional resolution.

The ChIPSeq assay shown here differs from other large-scale ChIP methods such as ChIPArray, also called ChIPchip (*1*); ChIPSAGE (SACO) (*3*); or ChIPPet (*4*) in design, data produced, and cost. The design is simple (Fig. 1A) and, unlike SACO or ChIPPet, it involves no plasmid library construction. Unlike microarray assays, the vast majority of single-copy sites in the genome is accessible for ChIPSeq assay (*5*), rather than a subset selected to be array features. For example, to sample with similar completeness by an Affymetrix-style microarray design, a nucleotide-by-nucleotide sliding window design of roughly 1 billion features per array would be needed for the nonrepeat portion of the human genome. In addition, ChIPSeq counts sequences and so avoids constraints imposed by array hybridization chemistry, such as base composition constraints related to T_m, the temperature at which 50% of double-stranded DNA or DNA-RNA hybrids is denatured; cross-hybridization; and secondary structure interference. Finally, ChIPSeq is feasible for any sequenced genome, rather than being restricted to species for which whole-genome tiling arrays have been produced.

ChIPSeq illustrates the power of new sequencing platforms, such as those from Solexa/Illumina and 454, to perform sequence census counting assays. The generic task in these applications is to identify and quantify the molecular

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305–5120, USA. [2]Biology Division, California Institute of Technology, Pasadena, CA 91125, USA. [3]California Institute of Technology Beckman Institute, Pasadena, CA 91125, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: woldb@its.caltech.edu (B.W.); myers@shgc.stanford.edu (R.M.M.)

contents of a nucleic acid sample whose genome of origin has been sequenced. The very large numbers of short individual sequence reads produced by these instruments (currently ~400,000 reads of 200 nucleotides (nt), or ~40 million reads of 25 nt, per instrument run, depending on the platform used) are extremely well suited to making direct digital measurements of the sequence content of a nucleic acid sample. By determining a short sequence read from each of many (10^5 to 10^7) randomly selected molecules from the sample and then informatically mapping each sequence read onto the reference genome, the identity of each starting molecule is learned, and its frequency in the sample is calculated. Desired levels of sensitivity and statistical certainty needed to detect rare molecular species can be achieved, in principle, by adjusting the total number of sequence reads. Sequence census assays do not require knowing in advance that a sequence is of interest as a promoter, enhancer, or RNA-coding domain, as most current microarray designs do. Below, we use the Solexa/Illumina platform, because high-read numbers contribute to high sensitivity and comprehensiveness in large genomes.

We used ChIPSeq to build a high-resolution interactome map for human neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor). This zinc finger repressor negatively regulates many neuronal genes in stem and progenitor cells and in nonneuronal cell types, such as the Jurkat T cell line studied here (*6*). A primary reason for selecting NRSF as a test case is that prior studies provide a large set of known "gold-standard" target genes, including more than 80 in vivo binding sites defined by ChIP-QPCR (quantitative real-time fluorescence polymerase chain reaction) (*7*). A subset of these genes has also been tested for regulatory function by transfection assays (*8*). In addition, the DNA motif bound by NRSF, called NRSE (also called RE1), is long (21 bp) and well-specified (*8*). This has led to a rich group of computational models for the site and for all its occurrences in the human genome (*7*, *9*–*11*). These sites provide a framework of explicit predictions that can now be tested by measuring repressor binding globally. Finally, there is a high-quality monoclonal antibody (*12*) that recognizes NRSF efficiently in ChIP experiments (*7*).

We prepared two DNA samples for each ChIPSeq experiment: an NRSF/REST–enriched ChIP sample and a companion control sample of the same fixed chromatin, but without immunoenrichment. In an effort to increase positional precision and to provide optimal substrate for the Solexa/Illumina sequencing platform, we introduced a size-selection step after cross-link reversal (Fig. 1A and fig. S2) (*5*). DNA sequencing of each sample was performed by the Solexa/Illumina protocol (*13*). Two to 5 million 25-nt sequence reads were produced per sample, of which about half mapped to single sites in the human genome (table S1). Sequence reads that map to multiple sites in the genome were removed from subsequent analysis. This eliminates sequences in simple repeats, some complex repeats, and also 25-nt segments that are not unique by chance. The location of each remaining unique sequence read in the genome was recorded. To accommodate polymorphisms relative to the reference genome, up to two mismatches were allowed. The resulting sequence read distribution was processed with a ChIPSeq peak locator algorithm developed for this purpose (*5*). The algorithm finds a local concentration of sequence hits (a location cluster) and, within that location, calls a peak. We then required of these a minimum fivefold enrichment of sequence reads in the ChIP sample relative to the corresponding location in the control. Fivefold enrichment is a conservative choice among enrichment thresholds commonly used in contemporary large-scale ChIP studies. A location that passed these criteria and also had 13 or more independent sequence reads (a threshold value selected based on the sensitivity and specificity analysis described below) was called an NRSF-positive binding event.

An example of primary ChIPSeq data from two independent experiments is shown in Fig. 1B for the *NEUROD1* locus. This positive signal, which has intermediate signal intensity and statistical certainty ($P = 8 \times 10^{-6}$), identifies a novel NRSF-binding target. The NRSF/REST sequence-tag distribution centers directly over the only canonical NRSE motif in a 4-kb region, which is located in the open reading frame of the *NEUROD1* gene. This site was called a

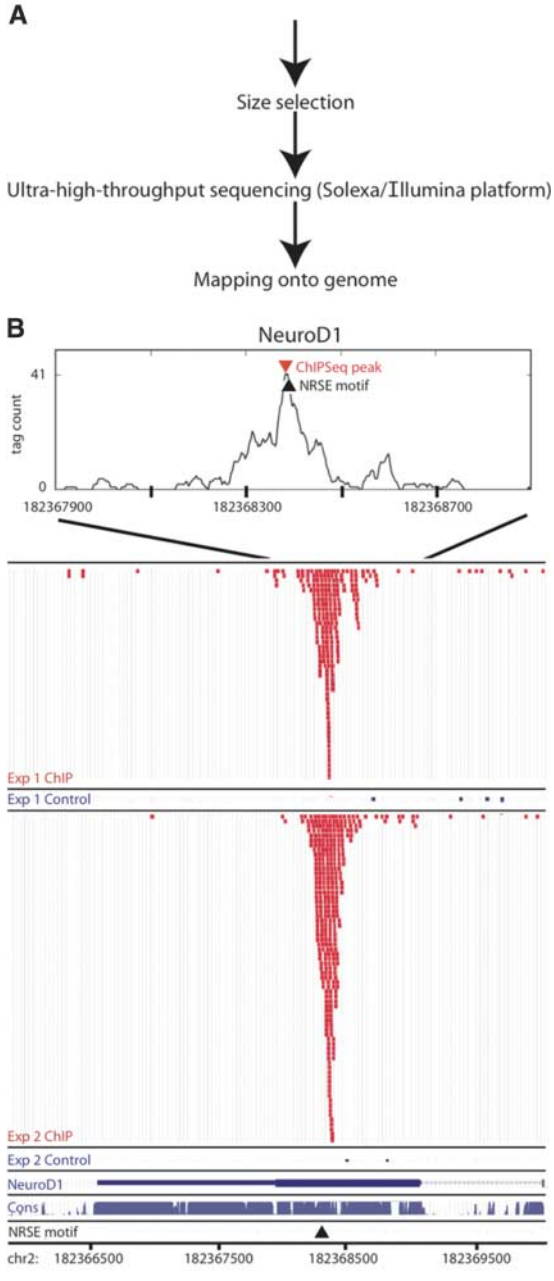

Fig. 1. ChIPSeq discovers NRSF/REST protein-DNA binding events with high resolution on a genome-wide scale. (**A**) Generalized scheme of ChIPSeq begins with ChIP, followed by size selection for recovered material (*5*), followed by standard preparation for Solexa/Illumina sequencing. An optional preamplification step after immuno-enrichment and before size selection can be inserted to work from smaller cell number input (used in experiment 2 below) (*5*). (**B**) Close-up view of ChIPSeq data mapped to a novel NRSF-binding site (NRSE; black arrowhead) located in the *NEUROD1* gene. Experiment 1 (no preamplification) and experiment 2 (with preamplification). Output from the peak-call algorithm is shown for this locus (red arrowhead), and corresponds closely with the sole NRSE in the *NEUROD* locus (black arrowhead).

ChIPSeq peak by the locator algorithm [open source available at (*14*)]. A previous study had implicated NRSF in repression of *NEUROD1*, but had failed to find a local site computationally. The authors theorized long-distance repression to explain the effect (*15*), but our results suggest a simpler explanation of a degenerate site within the *NEUROD1* gene itself. Over the entire primary data set (tables S2 and S3), the distribution of sequence-tag number per location ranged from the threshold value of 13 sequence tags to a maximum of 6718 tags at the highest signal (Fig. 2A and fig. S3). The two ChIPSeq experiments produced similar results (fig. S2), mapping 1946 shared enriched locations, most of which occur in or near 1020 genes.

NRSF-binding sites previously identified by QPCR or transfection assays (*7*, *8*) plus a set of known negatives (*5*) were used to measure sensitivity (successful detection of true positives) and specificity (successful rejection of true negatives) of the ChIPSeq assay. A ROC (receiver operator characteristic) analysis provides a way of measuring and graphically portraying sensitivity (fraction true positive on the *y* axis) versus specificity (1 – the fraction of false-positives, displayed on the *x* axis) (Fig. 2B). Perfect sensitivity and specificity would produce an ROC curve that traces the *y* axis to a value of 1.0, which would extend across all possible *x* values to produce an area under the curve of 1.0. In contrast, entirely random classification by chance would produce an ROC area of ~0.5. The observed ROC areas are high at 0.96 (shown) and 0.97 for the two independent experiments. The selected threshold of 13 sequence reads per region required for inclusion in the ChIPSeq interactome corresponds to a sensitivity of 87% and a specificity of 98%. We conclude that the ChIPSeq NRSF interactome measurements are accurate and, as suggested by *P* values (table S2), statistically robust. For rough comparison, a recent ChIPPet study of the p53 interactome did not measure these parameters, but investigators estimated that less than 35% of the largest group of positive signals (Pet2 sites, which were defined by two paired end tags) are true sites, whereas the much more certain and smaller class (Pet3-and-above sites) likely misses more than half of true positives (*4*). In general, we expect that differences in immune reagent quality, epitope availability, and other aspects of design that affect all ChIP experiments, as well as interactome structure itself, will contribute to differences between studies.

We next assessed the precision of ChIPSeq site location relative to 771 computationally high-scoring NRSE motifs in the genome that also have positive ChIPSeq signals, by measuring the distance from the experimental ChIPSeq peak to the center of the computational NRSE sequence motif. In this group, 754 sites were ChIPSeq-positive in both experiments, and the center of a 21-bp NRSE motif was within ± 50 bp of the called ChIPSeq peak (*5*) for 94% of these (Fig. 2C). The resolution, which depends in part on size selection of sheared chromatin after immuno-enrichment (*5*), is much higher than is typical for ChIPchip or ChIPSAGE (± 500 to 1000 bp) (*3*, *16*).

How comprehensive are the NRSF ChIPSeq measurements? Several lines of evidence address this question. First, as shown in Fig. 2D, virtually all strong canonical NRSF motif instances across the human genome were detectably occupied. We defined strong sites as those having ≥90% match to a previously developed motif model (a position-specific frequency matrix), which is based on evolutionarily conserved site instances across multiple mammalian genomes (*5*, *7*). This high representation of detectable binding suggests that no strong sites were missed by undersampling. It also implies that all sites are accessible for NRSF/REST-binding in Jurkat cells, at least part-time in some individual cells, although the degree of accessibility might vary and may account for wide differences in the number of tags per site (Fig. 2A). Second, we observed ChIPSeq-positive signals for sites previously studied in detail by transfection analysis (*8*), and they correspond to a wide range of ChIPSeq signals, with all but one scoring positive in both ChIPSeq experiments. Taken together with the sensitivity results (Fig. 2B), these observations suggest that the NRSF/REST interactome measurements are genome-comprehensive and have been sampled deeply enough to include most sites known by any other criteria to be biologically active, even if relatively weakly. This level of genome completeness is attributable to the depth of Solexa/Illumina sequence sampling and is substantially greater than in prior studies of the adenosine 3′,5′-monophosphate (cAMP) response element–binding protein (CREB) interactome measured by SACO (*3*) and the p53 interactome measured by ChIPPet (*4*).

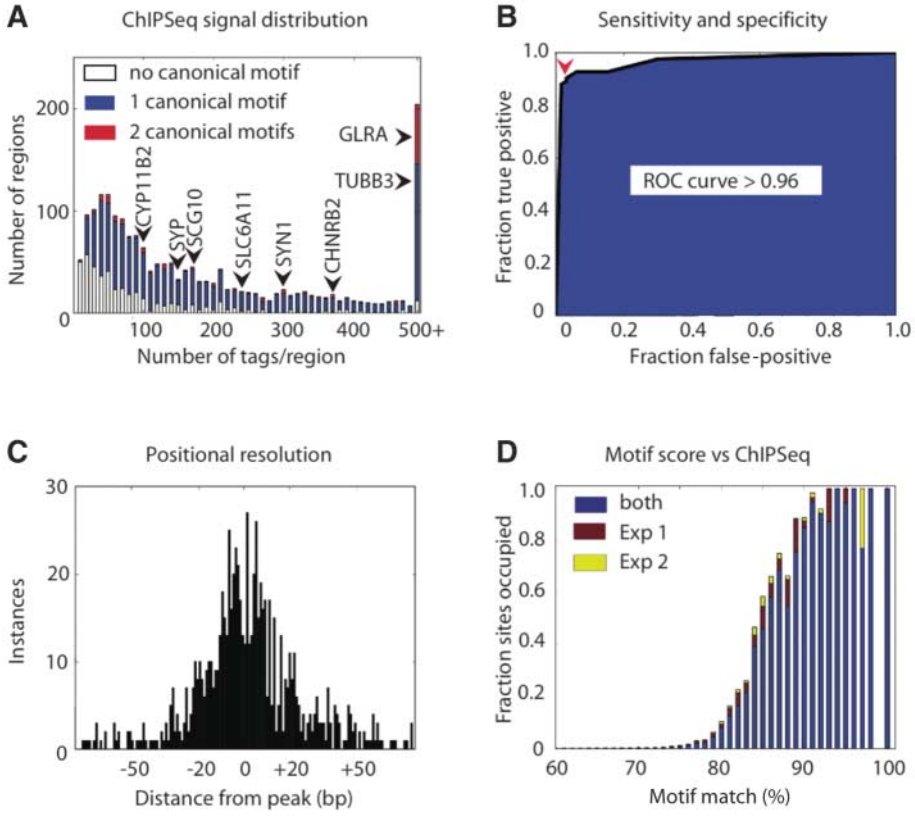

Fig. 2. (**A**) Histogram of all ChIPSeq-positive regions, as a function of sequence read number, that map within that region; zero (white), one (blue), or two or more (red) canonical motif instances defined as those scoring ≥70% match to the position-specific frequency matrix (PSFM) model of the NRSF-binding site (*7*). (**B**) ROC analysis with area under the curve >0.96 for experiment 2 (shown) and >0.97 for experiment 1. This measures the performance of ChIPSeq in detecting previously validated true positives (83) and true negatives (130), as described in the text and (*5*). The threshold used for subsequent analysis corresponds to 87% and specificity 98%, as indicated by the arrowhead. ChIPSeq false-negatives corresponded to the lowest QPCR-positive validation values. (**C**) Distribution of the distance from the center of 771 high-scoring canonical NRSE motifs [84% or higher match to the PSFM motif model (*7*)] in ChIPSeq-enriched regions (experiment 1). In this example, 46% of peaks fall within the boundaries of NRSE (here, +10 to −10 bp), and 94% of the canonical NRSEs fall within 50 bp of the called experimental peak. (**D**) Site occupancy detected by ChIPSeq for NRSE motifs in Jurkat cells as a function of PSFM score (*7*).

We next asked whether NRSF binding in or near promoters is correlated with low levels of transcription, as expected for a transcriptional repressor. To answer this question, we looked for high-confidence promoter predictions (*5*) that occur within 1 kb of a ChIPSeq peak. We then assessed genome-wide transcript levels in Jurkat cells by hybridizing labeled RNA to Illumina RefSeq8 Sentrix arrays. The subset of 230 transcripts corresponding to promoters near NRSF-binding sites had significantly lower ($P = 1 \times 10^{-11}$) transcript signals than the full set of 20,589 transcripts (fig. S5). This argues that NRSF binding near promoters is significantly associated with transcriptional repression in these cells.

The positional resolution and low number of false-positives in these experiments can greatly facilitate motif-finding algorithms. The effect is to elevate the frequency of occurrence of true motifs within the input DNA relative to background sequences. This can improve signal-to-noise and also greatly reduce the run times for many algorithms. Much is known about the canonical NRSF-binding site (NRSE motif) (*7*, *10*, *17*), and this allowed us to ask if that site emerges when a sample of the experimental interactome peak domains are used. We first applied the motif-finding algorithm MEME (*18*) to all sites in the top 10% of signal intensity (100-bp segments from 198 regions having 500 reads or more). MEME returned the full previously known motif (table S4). Single or multiple matches to this canonical motif, using a 70% match threshold, account for 75% of all ChIPSeq regions mapped in this study.

We next focused attention on those remaining ChIPSeq-positive regions that have 300 or more ChIPSeq reads, yet have no canonical motif match. There are 22 such locations, and when they were run in MEME, only two candidate motifs stood out (table S5). By inspection, the large canonical NRSF-binding motif of 21 bp is naturally subdivided into two prominent, nonidentical, nonpalindromic half sites (Fig. 3A). The two motifs from the MEME search correspond directly to the separate left and right sides of the canonical motif. We next asked if these motifs occur at other ChIPSeq-binding locations and if they are organized in any discernible pattern. A distinctive pattern was discovered within 50 bp of many ChIPSeq peaks, in which left and right half-site motifs are separated by additional "spacer" sequence that increases the center-to-center distance from the canonical 11 bp to 16 to 19 bp, or decreases it by 1 bp to 10 bp (Fig. 3B and fig. S6). Thus, the canonical site has two central positions that have no sequence specificity, and the noncanonical group is similarly oriented but has increased the separation distance by an additional 5 to 9 bp (Fig. 3C). These linked half sites, oriented with respect to each other in the same way as in the canonical site, occur in NRSF ChIPSeq binding domains in a statistically significant manner relative to random sequence windows in the genome ($\chi^2 = 1309$ for the half-site distance of 17, P value of 0) and account for 197 regions lacking a canonical motif (Fig. 3B and fig. S4). We also found that some binding locations have multiple clustered occurrences of noncanonical motif(s) along with a canonical one.

There are no structural data available for NRSF, so we cannot relate this new family of binding-site motifs to a known DNA binding structure. However, the protein has eight zinc fingers in its DNA binding domain, and other C2H2 zinc finger proteins such as Zif268 bind DNA with three fingers per 10-bp turn, but they show considerable strain when binding with six fingers (*19*). This makes simultaneous binding of one molecule of NRSF to these noncanonical half-site configurations plausible, but it is also possible that the protein is bound to only one half site at a time by using a subset of its fingers in these cases. It will be interesting to learn if there are other functional and molecular characteristics that set these sites apart. For example, do the different NRSF co-repressors differ in their interactions at noncanonical sites compared with canonical ones (*20, 21*)?

We also asked whether half sites are significantly enriched in our ChIPSeq neighborhoods, without regard to orientation or spacing, relative to expectations based on their occurrence in the genomes, and found that these regions are greatly enriched for left-side half sites ($\chi^2 = 3070$) and right-side half sites ($\chi^2 = 11{,}674$). This range of configurations, from concentrated half sites to the noncanonical 16- to 19-bp–spaced left and right sites, to the canonical 11-bp–spaced full site, is quite striking. Significant NRSF binding occurs in vivo, according to our data, at all three kinds of loci. Because the half sites are much shorter than the full 21-bp NRSE motif, they also occur widely over the genome, presumably mainly by chance. This would mean that there is a rich pool of

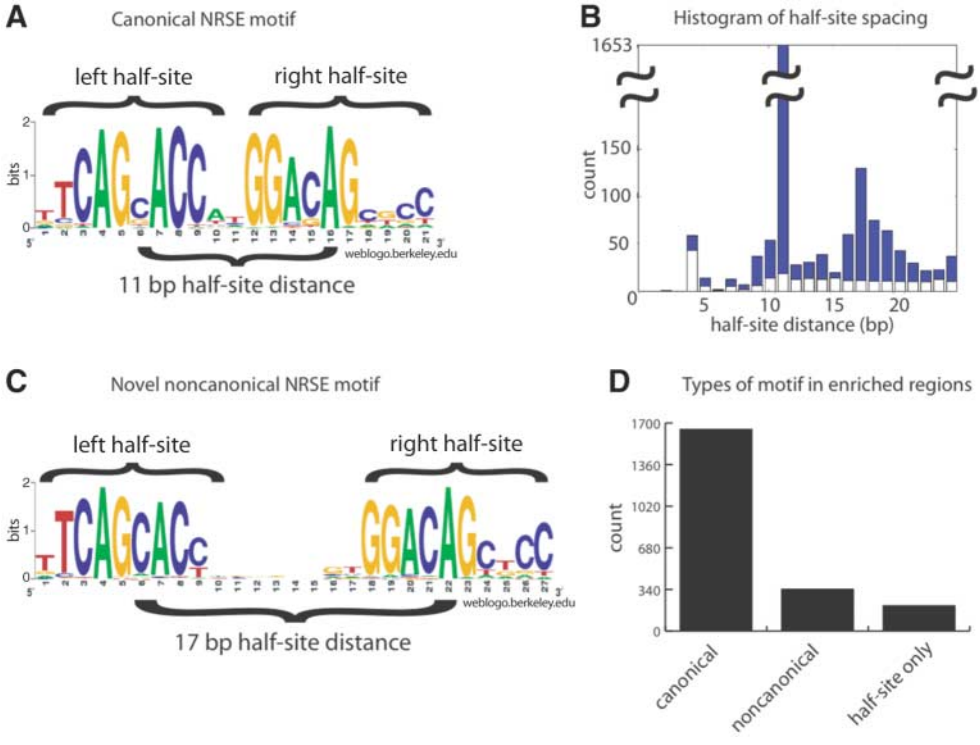

Fig. 3. (**A**) Canonical NRSF-binding motif WebLogo (*26*). Its left and right half sites have a center-to-center distance of 11 bp. (**B**) Histogram of half-site distances in ChIPSeq-enriched regions, showing the observed (blue) and expected (white) counts (based on frequency in the genome). In addition to the expected canonical peak at distance 11 bp, there is also significant enrichment of half sites with noncanonical distances of 16 to 20 bp. (**C**) WebLogo of noncanonical NRSE with half-site distance of 17, showing the lack of conservation in the spacer nucleotides. (**D**) The 2214 NRSF-binding motifs predicted in the 1946 ChIPSeq-positive regions that contain canonical, noncanonical, or only half-site motifs.

possible binding sites from which higher affinity canonical sites could be gradually made and tested in evolution, as suggested previously (*10*). However, these sites were considered unlikely to interact with NRSF specifically (*10*), whereas, within the noncanonical motif family we define here, sites bind on their own, especially when clustered (fig. S6). This suggests a plausible multistep path by which the target-site repertoire could evolve, beginning with clustered partial sites, passing through an intermediate and more specific orientation and spacing (the 10- or 16- to 19-bp–spaced family here), and eventually becoming refined into the canonical site (the 11-bp–spaced classic binding motif). Because there are more than 500 multi–zinc finger transcription factors encoded in the genome (*22*), many of which are evolving rapidly in humans and mice, this strategy might be used by other members of the zinc finger family.

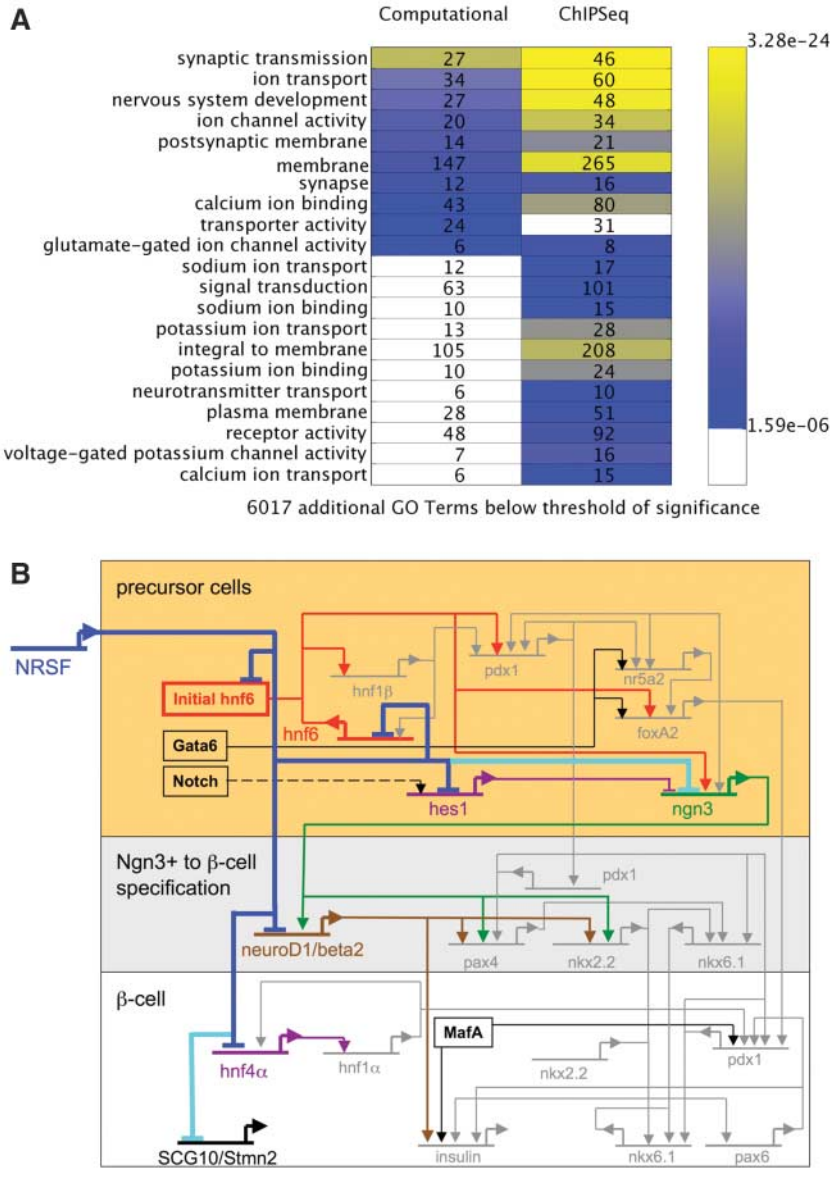

Fig. 4. (**A**) Gene ontology (GO) analysis of the computationally predicted cohort of NRSF target genes [genes scoring as ≥84% match to the previously developed computational model for the NRSE motif (*7*)] compared with ChIPSeq-positive genes (right). *P* values for enrichment of GO terms above the significance threshold of 1.59×10^{-6}, which accounts for multiple hypothesis testing (*7*), are indicated by the color scale; GO terms below the significance threshold are in white boxes. ChIPSeq NRSE target genes are most enriched in synaptic transmission, nervous system development, and ion channel–activity functions (tables S1 and S2). (**B**) ChIPSeq data identified new candidate connections (blue) between NRSF and members of the pancreatic islet β cell–specification gene regulatory network [adapted from (*23*)]. Key transcription factors bound by NRSF, including ngn3 and neuroD1, occupy positions high in the network that govern network activation and progression. ChIPSeq data also confirmed previously known NRSF targets (cyan) that include terminal differentiation genes such as *SCG10/stmn2* (*25*).

We found that genes encoding 110 transcription factors, 22 microRNAs, and five splicing regulators were occupied by NRSF. NRSEs occur prominently in introns (table S6), including a noncanonical site ($P = 4 \times 10^{-5}$) located about 500 bp downstream of the transcription start site of the *NRSF* gene itself, which suggests the possibility of negative autoregulatory feedback. We also found, as expected, that NRSF-bound loci are highly enriched in gene ontology (GO) terms related to neurons and their development (Fig. 4A). The enrichment for the experimentally determined sites exceeded that achieved for any computationally predicted target gene cohort. Synaptic transmission and nervous system development rank in the top three GO terms among 6000, with *P* values for overrepresentation of the NRSF target genes of 10^{-24} and 10^{-17} (*5*) (Fig. 4A). This group includes a set of transcription factors that have not previously been suggested as NRSF targets, but are known to be critical in the gene network that drives islet cell development in the pancreas. The transcription factors NEUROD1/BETA2, hepatocyte nuclear factors HNF4a and HNF6/Onecut1, and Hes1 were all detected here for the first time as in vivo binding targets of NRSF, and together with Neurogenin3, which is a previously identified target (*7*), they are positioned critically in the regulatory network that controls pancreatic β cell development (Fig. 4B) (*23*). Although in vivo binding does not ensure NRSF repression activity, these regulators are known to function as positive drivers of pancreatic neuroendocrine development. If NRSF repression is active at all these sites, as might be the case in progenitor cells, the circuit would be effectively blocked. In this hypothesis, NRSF acts as a permissivity factor gating entry into and progress through the developmental pathway.

These pancreatic network sites are among the more modest ChIPSeq signals, ranging from 55 sequence reads for HNF6 to 202 sequence reads for NeuroD1, values that are comfortably above the significance threshold of 13 (set on the basis of sensitivity/specificity considerations and known regulatory targets of Fig. 2, A and B), yet they fall in the bottom quartile. Thus, these ChIPSeq data were statistically robust enough to map parts of this gene network that might otherwise have gone undetected or been highly uncertain (Fig. 2A and fig. S3). There are precedents in other systems that show that relatively weak sites are biologically important, specifically because they are, in the biochemical binding sense, suboptimal. For example, in *Caenorhabditis elegans*, the Pha4/FoxA factor is the key activator of a large interactome, and a subset of target genes has suboptimal sequences and numbers of sites (*24*). In that system, when binding is suboptimal, it is believed

to help program the temporal order of action during development, with poor binders turning on at later times in the developmental progression, when Pha4 levels are highest. By analogy, the regulators that govern the pancreatic network may be released from NRSF repression relatively early in down-regulation of the repressor to create a permissive state that must be established before the neuroendocrine development program is launched. Also following this logic, *SCG10/Stmn2* is a classic NRSF target gene that is expressed later in development in differentiated islet cells, and it displayed relatively higher ChIPSeq tag scores than most of the transcription factors that are positioned higher and earlier in the network. Independent evidence suggests that *SCG10/Stmn2* expression depends on relief from NRSF-mediated repression in islet cells (*25*). Targets of the regulatory class highlighted here (Fig. 4B) can also participate in positive autoregulatory and cross-regulatory interactions that we expect would stabilize and push forward the circuit once it begins (*23*). This makes a "protective" repressor, active in nonpancreatic cell types or progenitor cells, an attractive piece of regulatory logic.

The initial picture we have of the experimentally determined NRSF/REST interactome has been drawn for one cell type (Jurkat T cell line), and T cells express this factor at relatively high levels. In this cell environment, the interactome is composed of three broad classes of target loci with respect to binding motifs. First are loci defined by near-optimal canonical motifs, and virtually all of these bound the factor detectably in vivo. This suggests that for biochemically optimal sites, there is sufficient chromatin access and high enough DNA affinity to establish measurable occupancy, although the strength of the ChIPSeq signal among these sites varied over wide range. Second are loci containing instances of the noncanonical motif family (Fig. 3) or sites that are weaker matches to the canonical 21-bp site. Binding of sites in this class could not be predicted solely on the basis of their motif sequence. Several hundred instances showed significant binding, whereas thousands of others in this group had no detectable ChIPSeq signal. Binding-motif properties we do not yet appreciate, differential chromatin access, or epitope exposure for subset of NRSF-containing complexes might discriminate the minority ChIP-positives in this group from the majority that are nonbinding. These observations argue that global experimental data are needed to discriminate motif instances occupied in vivo from many others that appear similar in motif quality, but are not similarly occupied, even for a factor like NRSF/REST, which has a well-specified binding-motif family. Finally, there is a third small, but interesting, class that binds NRSF/REST reproducibly, and in some cases quite robustly, but lacks any identifiable NRSE motif except for the presence of half sites (tables S2 and S7). It is uncertain if these are explained by their half-site content, because they make up a tiny minority of loci with superficially similar half-site content. This raises the possibility that NRSF/REST might associate indirectly, rather than directly, with a limited number of specific chromosomal locations.

ChIPSeq, as performed here, is relatively cost-effective; Solexa/Illumina platform sequencing costs per experiment are currently about half that of the most comprehensive human whole-genome tiling arrays. ChIPSeq sampling that is 10 to 20 times deeper than was used here is plausible now, within the general range of microarray costs, and this capacity may be needed for interactomes having many more sites per genome than NRSF. For example, it is possible that some widely used transcription cofactors or chromatin remodeling complexes might have on the order of 10^4 to 10^5 true-positive sites distributed over a wide dynamic range of occupancy levels, and such interactome structures will require correspondingly deeper sequence sampling. Other ultrahigh-throughput sequencing platforms, such as the one from 454 Life Sciences, could also be used to assay ChIP products, but whatever sequencing platform is used, our results indicate that read number capacity and input ChIP DNA size are key parameters.

References and Notes

1. T. H. Kim, B. Ren, *Annu. Rev. Genom. Hum. Genet.* **7**, 81 (2006).
2. ENCODE Project Consortium, *Science* **306**, 636 (2004).
3. S. Impey *et al.*, *Cell* **119**, 1041 (2004).
4. C. L. Wei *et al.*, *Cell* **124**, 207 (2006).
5. Materials and methods are available as supporting material on *Science* Online.
6. N. Ballas, G. Mandel, *Curr. Opin. Neurobiol.* **15**, 500 (2005).
7. A. Mortazavi, E. C. Thompson, S. T. Garcia, R. M. Myers, B. Wold, *Genome Res.* **16**, 1208 (2006).
8. C. J. Schoenherr, A. J. Paquette, D. J. Anderson, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 9881 (1996).
9. A. W. Bruce *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10458 (2004).
10. C. Zhang *et al.*, *Nucleic Acids Res.* **34**, 2238 (2006).
11. J. Wu, X. Xie, *Genome Biol.* **7**, R85 (2006).
12. Z. F. Chen, A. J. Paquette, D. J. Anderson, *Nat. Genet.* **20**, 136 (1998).
13. Solexa sequencing technology, www.illumina.com/pages.ilmn?ID=203.
14. ChIPSeq peak finder, available at http://woldlab.caltech.edu.
15. V. V. Lunyak *et al.*, *Science* **298**, 1747 (2002).
16. S. Cawley *et al.*, *Cell* **116**, 499 (2004).
17. N. C. Jones, P. A. Pevzner, *Bioinformatics* **22**, e236 (2006).
18. T. L. Bailey, C. Elkan, *Mach. Learn.* **21**, 51 (1995).
19. E. Peisach, C. O. Pabo, *J. Mol. Biol.* **330**, 1 (2003).
20. N. Ballas, C. Grunseich, D. D. Lu, J. C. Speh, G. Mandel, *Cell* **121**, 645 (2005).
21. M. Yeo *et al.*, *Science* **307**, 596 (2005).
22. S. Huntley *et al.*, *Genome Res.* **16**, 669 (2006).
23. E. H. Davidson, *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic Press/Elsevier, San Diego, CA, 2006).
24. J. Gaudet, S. E. Mango, *Science* **295**, 821 (2002).
25. F. Atouf, P. Czernichow, R. Scharfmann, *J. Biol. Chem.* **272**, 1929 (1997).
26. Web-based sequence logo generating application; Weblogo.berkeley.edu.

27. We thank G. Schroth and his group at Solexa/Illumina, Inc., for access to the sequencing platform, without which this work would not have been possible. We thank B. Anton, L. Nguyen, C. Medina, and L. Tsavaler for outstanding experimental work and K. F. McCue for invaluable counsel on statistical analyses. We are grateful to D. Anderson of Caltech for the gift of monoclonal antibody against NRSF. This work was supported by NIH grant U01 HG003162 to R.M.M. with a supplement to B.W. and by a grant from the Caltech Beckman Institute. A.M. was supported by NIH/National Research Service Award 5T32GM07616; B.W. by a Bren Chair endowment at Caltech, and R.M.M. by the Stanford W. Ascherman Chair endowment at Stanford University.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1141319/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 to S7
References

14 February 2007; accepted 26 April 2007
Published online 31 May 2007;
10.1126/science.1141319
Include this information when citing this paper.