Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire

Ali Mortazavi, Evonne Chen Leeper Thompson, Sarah T. Garcia, et al.

Genome Res. 2006 16: 1208-1221
Access the most recent version at doi:10.1101/gr.4997306

Letter

Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire

Ali Mortazavi,[1] Evonne Chen Leeper Thompson,[2] Sarah T. Garcia,[2] Richard M. Myers,[2] and Barbara Wold[1,3]

[1]*Division of Biology, California Institute of Technology, Pasadena, California 91125, USA;* [2]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA*

We constructed and applied an open source informatic framework called Cistematic in an effort to predict the target gene repertoire for transcription factors with large binding sites. Cistematic uses two different evolutionary conservation-filtering algorithms in conjunction with several analysis modules. Beginning with a single conserved and biologically tested site for the neuronal repressor NRSF/REST, Cistematic generated a refined PSFM (position specific frequency matrix) based on conserved site occurrences in mouse, human, and dog genomes. Predictions from this model were validated by chromatin immunoprecipitation (ChIP) followed by quantitative PCR. The combination of transfection assays and ChIP enrichment data provided an objective basis for setting a threshold for membership and rank-ordering a final gene cohort model consisting of 842 high-confidence sites in the human genome associated with 733 genes. Statistically significant enrichment of NRSE-associated genes was found for neuron-specific Gene Ontology (GO) terms and neuronal mRNA expression profiles. A more extensive evolutionary survey showed that NRSE sites matching the PSFM model exist in roughly similar numbers in all fully sequenced vertebrate genomes but are notably absent from invertebrate and protochordate genomes, as is NRSF itself. Some NRSF/REST sites reside in repeats, which suggests a mechanism for both ancient and modern dispersal of NRSEs through vertebrate genomes. Multiple predicted sites are located near neuronal microRNA and splicing-factor genes, and these tested positive for NRSF/REST occupancy in vivo. The resulting network model integrates post-transcriptional and translational controllers, including candidate feedback loops on NRSF and its corepressor, CoREST.

[Supplemental material is available online at www.genome.org. The Cistematic source code and associated databases are available at http://cistematic.caltech.edu.]

Specific repressors, such as canonical zinc finger transcription factors, stand out in vertebrate genomes because of their large number, significant expansion in mammals, and diversity of cellular and organismic functions affected (Hamilton et al. 2003). The Krab family of zinc finger sequence-specific DNA-binding repressors, for example, numbers over 400 in rodent and human genomes (Dehal et al. 2001; Shannon et al. 2003). For the vast majority of these, nothing is known about their target-gene repertoire or binding motif. A few, studied in more detail, play important roles in diverse cellular and organismic functions ranging from regulation of rodent male-specific genes by the Rsl (regulator of sex limitation) Krab repressors (Krebs et al. 2005) to lipid metabolism and possible predisposition to hypoalphalipoproteinemia by znf202 (Wagner et al. 2000). Much more is known about NRSF/REST, a zinc finger repressor famous for negative regulation of neuronal genes in non-neuronal cell types and in neuronal stem cells and progenitors prior to differentiation (Chong et al. 1995; Schoenherr and Anderson 1995; Chen et al. 1998). The main isoform of NRSF represses transcription by recruiting cofactors such as CoREST (Andres et al. 1999), CTD phosphatases (Yeo et al. 2005), mSin3A, and histone deacetylases (Huang et al. 1999). Another isoform, REST4, is thought to act in a dominant negative fashion (Hersh and Shimojo 2003). In addition to neuronal development, NRSF/REST may have other roles in cardiac development (Kuwahara et al. 2003), pancreatic islet development (Atouf et al. 1997; Abderrahmani et al. 2001), and perhaps B- or T-cell lineages (Scholl et al. 1996). Little is known about which genes affecting these non-neuronal lineages are direct NRSF/REST targets or how many overlap with the neuronal set.

A first step toward understanding how a regulator fits into the design logic and function of a gene network is to define its genome-wide target gene set. In multicellular animals and plants, this is not easily done by direct experimental measurements, because the matrix of all possible target DNA sites, across many tissues and developmental states, is so vast. An alternate starting point is to use comparative genomics, constrained by some smaller sets of functional data, to generate a computational genome-wide model that can then be tested directly and interrogated to develop new focused hypotheses.

Two considerations make the NRSF/REST repressor a superior candidate for this analysis. First, factors with tandem arrays of zinc fingers can recognize relatively long and specific target motifs, and this makes computational approaches for finding target genes more feasible. Specifically, NRSF has a 21-bp binding site (NRSE or RE-1), and much is known about where and how NRSEs function. They can direct repression from positions within

[3]Corresponding author.
E-mail woldb@caltech.edu; fax: (626) 449-0756.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4997306.

16:1208–1221 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

5′-UTRs, in introns and at intron/exon junctions, as well as upstream of the transcription start and downstream of the coding stop (Schoenherr et al. 1996; Thiel et al. 1998). One study also reported that repression can extend to neighboring genes at one locus, although it is not clear whether this is general or not (Lunyak et al. 2002). NRSF transcriptional repression also appears to be tuned in vivo for strength and timing at different target genes during the progression from pluripotent stem cell to differentiated neuron or glial cell (Kuwabara et al. 2004; Ballas et al. 2005). It is not known whether these distinctions, so far studied for a few genes, reflect differences in the sequence, number, or organization of NRSE sites.

The second virtue of NRSF/REST for genome-wide target prediction is that a collection of NRSF sites has been quantitatively assayed for activity in vivo (Schoenherr et al. 1996; Bruce et al. 2004). These assays, which include sequences that resemble the consensus binding site, but lack function, are invaluable for calibrating and interpreting any model of NRSF binding derived by other criteria, including evolutionary conservation of NRSE occurrences.

In addition to direct transcriptional regulation, post-transcriptional and translational mechanisms mediated by microRNAs are implicated in neurogenesis. Cells undergoing terminal differentiation express tissue-specific microRNAs that are currently thought to modulate translation and/or degradation of large networks of target mRNAs (for review, see Kosik and Krichevsky 2005). miR-124a, for example, is neuron specific and can target hundreds of genes when expressed in HeLa cells (Lim et al. 2005). A broad survey of microRNA expression in brain and neuronal cell culture (Sempere et al. 2004) suggests that there are at least a dozen different microRNAs that are predominantly expressed in the brain. While prediction of likely target sites in 3′-UTRs of known mRNAs has been very active (John et al. 2004; Krek et al. 2005; Lewis et al. 2005), little is known about how microRNAs are themselves transcriptionally regulated, except that microRNAs located within introns of protein-coding genes tend to be expressed along with their "host" gene (for review, see Ying and Lin 2004). This emerging picture raises the question of how transcriptional regulators are connected to and coordinated with the post-transcriptional ones.

In the first part of this work, we use NRSF/REST as an amenable test case to build a comprehensive genome-wide model for the corresponding gene cohort. To do this, we develop a set of generally applicable algorithms and open-source software tools (Cistematic) to make and refine site predictions and enumerate the target gene cohort. We show that it is possible to begin with a single biologically defined, evolutionarily conserved NRSF/REST site, then use conservation among mouse, human, and dog genomes to develop a refined model for NRSF sites. The resulting model is compared and contrasted with prior ones (Schoenherr et al. 1996; Bruce et al. 2004), and we show that the major known functions of NRSF can be deduced computationally by using RNA expression and GO analysis modules in Cistematic. We test our model by experimentally measuring in vivo binding at 113 loci by chromatin immunoprecipitation followed by Q-PCR. In the second part of the study, we use the PSFM model to investigate evolution of the NRSF network over much greater evolutionary distances, and to develop and test specific hypotheses about links between NRSF/REST and post-transcriptional regulatory pathways. High-confidence candidate sites near neuronal microRNAs and splicing factors are identified, and in vivo interaction of NRSF at these loci is experimentally verified.

Results

The availability of multiple whole-genome DNA sequences raises the possibility of building strong predictive models for the entire binding-site repertoire of a sequence-specific DNA-binding factor by leveraging preferential conservation of functionally important sites. We used a two-part strategy that begins by deriving and refining a PSFM model for the binding site. The starting point is one or more functionally tested and conserved instances to seed a multigenome search for additional conserved instances. The cisMatcher algorithm used to do this is designed to focus on site instances that are embedded in somewhat larger conserved domains. The reasoning is that functional sites are often located within larger conserved *cis*-regulatory modules. At later times in the process, one can exercise an option to recover other instances of the site that do not require conservation beyond the boundaries of the site model. The second process develops a model for the genome-wide cohort of genes associated with sites defined by the fact that they match the PSFM at or above a specified score. At this stage, various conservation and gene geography criteria are selected and applied. They can require, for example, that PSFM match sites occur near orthologs in multiple genomes and that candidate cohort genes be located within a specified distance of a PSFM match site. Thus, the refined PSFM from the first part of the process is used to interrogate the genome(s) to find which genes are located near site instances. The PSFM match score, coupled with archival and new experimental data, is then used to help establish an appropriate threshold for inclusion in target gene cohort.

Deriving and refining a conservation-based PSFM site model

The Cistematic pipeline is outlined in Figure 1 and summarized here, with details in the Methods section. In one experiment, the derivation pipeline was initiated with orthologs from a single gene, *SCG10* (*STMN2*) in human, mouse, and dog genomes (Mori et al. 1992; Schoenherr and Anderson 1995). This seed PSFM was used to run a genome-wide search that used the cisMatcher algorithm. It collected additional similar instances that occur in domains of conservation (here set for 87.5% PSFM match and 85% similarity in a 25–65-bp window) shared by at least two of the three participating genomes (Supplemental Fig. S2 and below). These conserved occurrences (81) of the motif were then used to derive a refined SCG10 PSFM, which we call NRSE2. In a second experiment, in contrast, we began with a collection of 33 different known NRSEs and used them to develop the seed PSFM (nrsePWM33). In a third experiment, we ran the PSFM pipeline on several other individual NRSE instances. The resulting site models were remarkably similar to each other (Fig. 2; Supplemental Table S1). We conclude that cisMatcher, operating over this set of genomes, derives a set of convergent PSFM models for NRSE sites. This means that our refinement process, which draws into the model many additional conserved instances, is robust to the identity of the specific initiating NRSE.

Estimating a membership threshold

How similar to the site model does a sequence need to be to function in vivo? We used multiple kinds of experimental data to iterate toward an informed and increasingly objective membership threshold. Setting a threshold is, at this stage, a useful and necessary simplification, but there is no biochemical or biological reason to expect a crisp boundary between sites that do and do not bind the factor. Figure 3A displays archival data for known

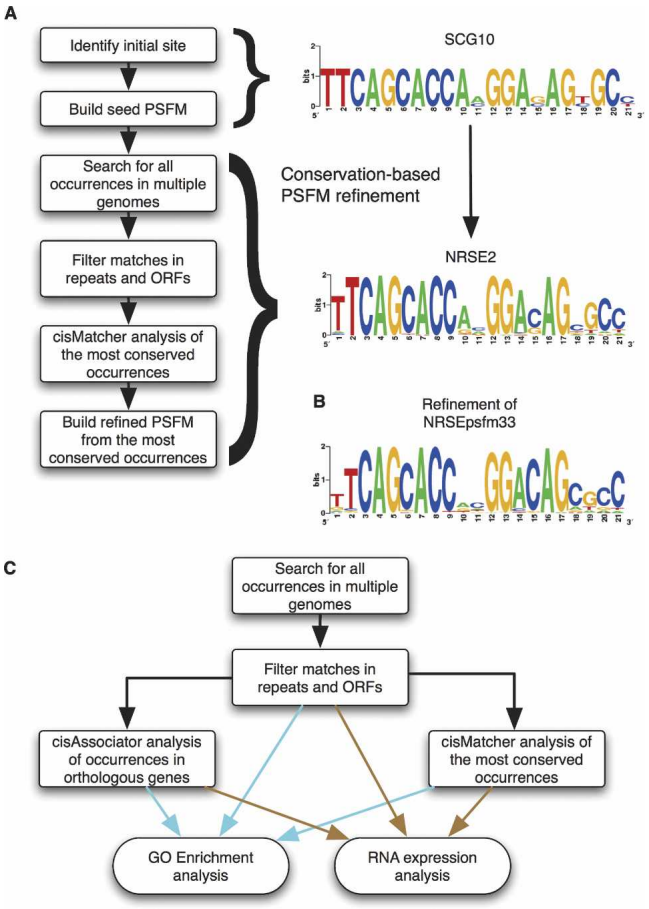

Figure 1. Experimental approach. (*A*) Results from genome-wide matches to the initial NRSE PSFM (SCG10) were analyzed with cisMatcher and used to create a refined NRSE PSFM (NRSE2). (*B*) A refinement starting with a PSFM of 33 known sites (Supplemental Table S2) produces a result very similar to NRSE2. (*C*) The genome was searched for occurrences of NRSE1, using its consensus (TYAGMRCCNNRGMSAG) (Bruce et al. 2004); or NRSE2, using its position-specific frequency matrix (PSFM). Resulting NRSE1 and NRSE2 gene cohorts were then analyzed for Gene Ontology (GO) enrichment and expression analysis as follows: (1) the NRSE2 PSFM was further processed and analyzed for GO enrichment and expression analysis of two subsets; (2) human genes with matches that co-occur in mouse and/or dog, and (3) human genes that are nearest to the "most conserved" matches, as identified by cisMatcher.

NRSE sites, plus a few previously tested negative sites that resemble the NRSE, plotted as a function of PSFM match score. These data suggested starting with an estimated 84% match score threshold. We also asked whether the PSFM match score correlated with the bioactivity of individual instances in a reporter transfection assay, drawing on data from Schoenherr et al. (1996). Remarkably, there was a significant correlation of PSFM match score with repression strength ($R^2 = 0.82$, Fig. 3B). The repression activity data are in general agreement with Figure 3A, and support a threshold value in the low 80's. The relationship of PSFM match score with repression efficiency in the transfection assay may also indicate that both reflect binding affinity.

Assembling and testing target-gene cohort models

The three mammalian genomes were then searched for every match to NRSE2 above a predetermined threshold and genes within a 10-kb radius were grouped into *cis*-regulatory cohorts of genes. This cohort of human NRSE2-associated genes was filtered for evolutionary conservation by requiring that matches also exist within 10 kb of an ortholog in mouse and/or dog genomes using Cistematic's cisAssociator algorithm. Because some known NRSF sites can apparently act in isolation without surrounding conserved elements, cisAssociator deliberately does not require alignment or additional conservation outside of the site. Note that when a match is within 10 kb of more than one gene, cisAssociator includes all genes into the cohort. This choice is based on the report that single NRSE instances can apparently silence multiple nearby genes (Lunyak et al. 2002). However, it also means that, even if the definition of the NRSE2 PSFM is optimal, some genes included in the cohort model will be false positives. We also wanted to collect additional sites that might function from distances greater than 10 kb, but without greatly increasing false positives. The cohort was therefore expanded by using the Cistematic cisMatcher algorithm to identify genes with conserved NRSE2 matches that are distal (Supplemental Table S4).

We then used the cohort model to revisit the threshold issue, evaluating it experimentally by sampling 113 candidate NRSE sites that spanned a range of high-scoring and low-scoring PSFM scores. Chromatin immunoprecipitation (ChIP) was performed and assayed by quantitative PCR (QPCR) (Fig. 4). These in vivo protein:DNA interaction data generally validate the PSFM model (see below) and the Probit model in Figure 4B suggests that a threshold around 84 is reasonable, but also indicates that there is no sharp PSFM boundary. This means that users of this and related models will select membership thresholds, or ranges for thresholds, to best serve different specific uses of the model for which pressure on sensitivity versus selectivity are different.

How do previously identified NRSE cohorts, based on conventional consensus sites, compare with the new PSFM? We compared the NRSE2 PSFM matches with instances found using the original (NRSE0) consensus of Schoenherr et al. (1996) and the recent (NRSE1) consensus used in the genome survey of Bruce et al. (2004). Cistematic recovered the respective gene cohorts corresponding to NRSE0 and NRSE1 instances. Supplemental Figure S3 shows that NRSE1 contains a significant fraction of matches that score poorly with the PSFM model (<80%), with many low-scoring NRSE1 matches occurring in complex repeats. Matches within repeats were excluded from subsequent analyses for both NRSE1 and NRSE2, although we note that individual instances embedded within repeats might be functional.

We then asked how cisMatcher positive sites are distributed relative to gene anatomy (Supplemental Figure S4). Many known instances of the NRSE that have been studied in detail are either intragenic or are located near the promoter, but it is not known how great a role ascertainment bias based on proximity has played in selecting them for study. We mapped the genome-wide cisMatcher set, which is not biased by the method of selection for its position relative to adjacent genes. There is an obvious enrichment of NRSF motifs within 5 kb of gene model start sites (40%), although a full quarter of the conserved matches are more than 10 kb from either the 5′ or 3′ boundary of the nearest gene model and 3′ UTRs have substantial numbers.

Chromatin immunoprecipitation analysis of predicted NRSEs

We tested the NRSE2 cohort experimentally at 113 sites (Fig. 4; Supplemental Table S6), 42 of which fell below our 84% thresh-

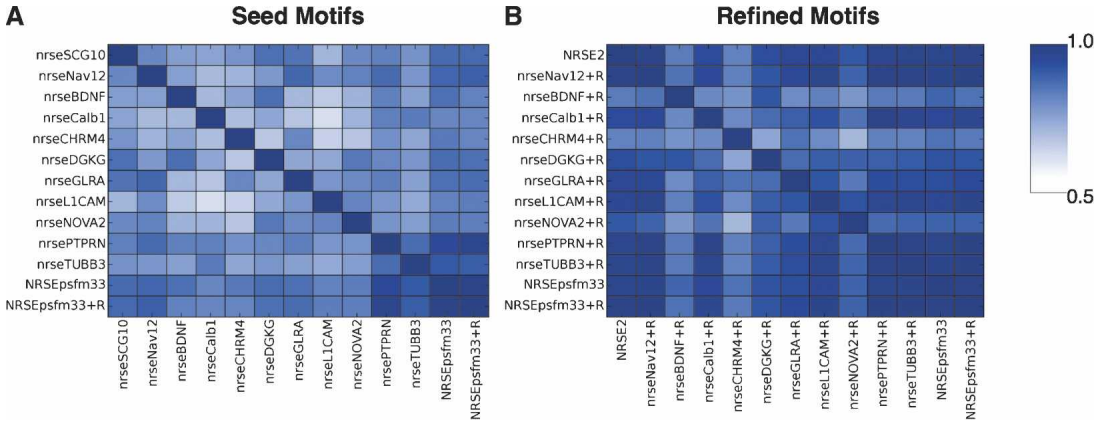

Figure 2. Different seed motifs converge following motif refinement. (*A*) A total of 10 initial seed motifs from known or predicted sites are compared using the motif similarity score (see Methods) to our starting motif (SCG10) as well as a PSFM of 33 known instances (NRSEpsfm33) and its refined version (NRSEpsfm33+R). The correlations median is 0.80. (*B*) Motif refinement of SCG10 (called NRSE2) and of the 10 initial motifs (denoted with a +R) are markedly more similar, with a motif correlations median of 0.91 with several intermotif correlations rising above 0.95.

old, by using chromatin IP coupled with Q-PCR in Jurkat cells (see Methods). Of 71 candidate sites ranking above the 84% threshold, 70 were ChIP positive. In contrast, at slightly lower PSFM match scores, 29 of 42 sites were negative for NRSF ChIP. Thus, predicted sites could be quite effectively partitioned by PSFM score into those that will certainly be ChIP positive and those that are likely to be negative (*P*-value = $8.6*10^{-16}$, Fisher's exact test). The associated Probit analysis allows one to select other thresholds and to consider the confidence limits at any selected threshold. The 84% value is a conservative membership threshold designed to minimize false positives at the cost of accepting some false-negative predictions (13/83, or 16%). Cistematic provides the option of sliding the threshold to provide cohort models that correspond to differing stringencies for false-positive or false-negative members.

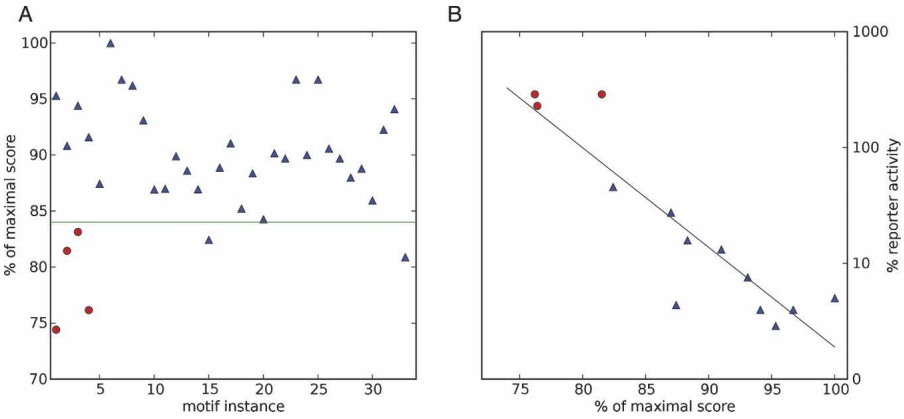

Figure 3. Selection of a threshold for NRSE2 and correlation of score with repression activity. (*A*) The 33 known instances (▲) and four false positives (filled ovals) listed in Table S1 were scored with the NRSE2 PSFM using a consensus score, as described in the text and methods. A threshold of 84% of the best possible score (match #5) was selected conservatively to exclude the known false positives. The PSFMs exclude about 6% of known instances at this relatively high threshold. (*B*) The NRSE2 PSFM score of 10 known instances and three false positives were plotted against their relative repression in a transient transfection of a reporter from Schoenherr et al. (1996), where 100% and above reporter activity represents no repression. The regression shows a marked correlation between PSFM match score and repression (R^2 = 0.82).

GO analysis of the NRSE2 *cis*-regulatory cohort

We next asked whether functions of the NRSF-regulated cohort could be inferred based on enrichment of Gene Ontology (GO) terms. Statistically significant enrichment, subject to Bonferroni correction for multiple hypothesis testing, was observed for each cohort model but not for a large set of randomly scrambled versions of the PSFM (Methods). The NRSE2 PSFM identified a larger cohort (660 human genes within 10 kb of an NRSE2) than the original NRSE0 consensus (362 human genes) or the seed SCG10-based PSFM (192 human genes) with significant enrichments in functional GO categories such as "synaptic transmission," "neurogenesis," and "transporter activity." These functions nicely recapitulate much of the NRSF literature. In contrast, several GO categories significantly enriched in the larger NRSE1 cohort (1270 genes), such as "synaptogenesis" or "calcium-dependent cell-cell adhesion," are conspicuously absent from the other NRSE cohorts. On detailed inspection, the latter results are mainly due to NRSE1 matches within the paralogous protocadherin β cluster. This calls attention to a specific interpretation issue in GO enrichment analysis, which is the power of very similar paralogs in gene families to drive an entire term to significance. Similar paralogy issues do not appear to dominate most significant other terms for any of the NRSE models.

Cistematic's orthology matching function was next used to develop a conserved cohort. NRSE2 instances in human, mouse, and dog were collected and subjected to both cisMatcher and cisAssociator conservation criteria. A total of 505 human genes met at least one of these criteria. GO analysis of the resulting conserved cohort (Fig. 5) shows further enrichment of several GO terms such as "transporter activity," "synapse," and "synaptic vesicle" when com-

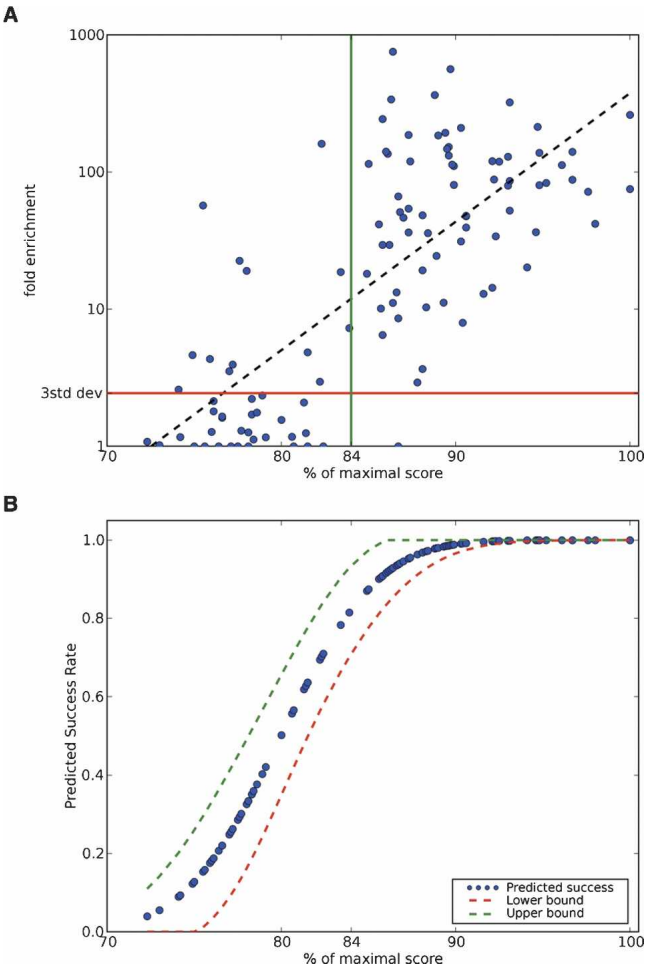

Figure 4. Quantitative analysis of chromatin immunoprecipitation of NRSF. (*A*) A total of 113 potential NRSE2 matches, 42 of which fell below our threshold of 84% (green vertical line), were assayed using ChIP followed by quantitative PCR. Fold enrichments were calculated by dividing the absolute number of genomic equivalents of each NRSE by the mean of the recovered amounts of five random nongenic, nonconserved regions. Fold enrichments that were above three standard deviations from the mean of the five random nongenic amounts (red line, 2.44× enrichment), were considered to be occupied sites. An exponential regression (black line in this semilog plot), which would correspond to the regression in Figure 2B, accounts for about half of the data's variation ($R^2 = 0.56$). A total of 13 of the 83 occupied sites (16%) fell below our 84% threshold. (*B*) Cumulative normal distribution function of probit coefficient vs. score with 95% confidence levels shown by dashes. The estimated chance of a success match goes up by nearly half between 80% and 84%.

pared with the larger NRSE2 cohort, but these effects were not substantial.

To test the robustness of the GO analysis, the columns of the NRSE2 PSFM were scrambled repeatedly and the entire analysis pipeline was repeated 100 times (data not shown). Only two scrambled motifs recovered any significantly enriched GO term, and they found just one each. No scrambled motif recovered significant GO terms when either of our conservation criteria was applied. These results argue that enrichment of specific GO terms for NRSE2 is statistically sound.

Comparative expression analysis of NRSE2

We asked whether the NRSE2 cohort is enriched in genes with a specific RNA expression pattern. One prediction from prior studies of NRSF is that genes expressed predominantly in neurons will be enriched among true biological targets of NRSF (Chen et al. 1998; Ballas et al. 2005). The GNF gene atlas (http://symatlas.gnf.org) (Su et al. 2004) of mRNA expression across 79 human tissues was used to investigate the expression profile of our gene cohorts. CompClust (Hart et al. 2005) was used to cluster the NRSE2 cohort with k-means, k-medians, and DiagEM for k = 5, 10, and 15. While all three algorithms returned similar pan-neuronal clusters, k-medians with a Pearson correlation metric and k = 5 performed best qualitatively and was used for all subsequent analyses. The NRSE1 cohort was also clustered for comparison and produced similar clusters (Supplemental Fig. S5). The NRSE2 clustering is shown in Figure 6A (NRSE2 cluster members are listed in Supplemental Table S3). In every clustering, one or more clusters had a distinctly brain-specific expression pattern whose medoid weights are shown in Figure 6B. The percentage of each cohort falling within these brain-specific clusters ranged from 21% for NRSE1 to 40% for NRSE2. These reactions are significantly higher than the percentage of genes in GNF that have a Pearson correlation coefficient >0.4 with our pan-neuronal medoid vector (1482 of 16,054 genes with current NCBI Gene ID's, or about 9%), which gives a *P*-value of $8.0 * 10^{-71}$ ($\chi^2 = 316.58$, χ^2 test for equality of distributions) for the neuronal enrichment of the NRSE2 cohort. Nevertheless, the majority of neuronal genes are not associated with a recognizable NRSE. As would be predicted if many NRSE2 genes are regulated by NRSF in a neuronal context, there is a large (greater than fourfold) enrichment for brain expression to 40% of all NRSE2-associated genes (Fig. 6C; Supplemental Fig. S5).

Figure 6D gives match-score distributions for the subset of genes that display a predominantly brain-specific RNA expression pattern. Genes within the brain-specific expression clusters share a similar scoring distribution pattern with the entire population of matches for both NRSE0 and NRSE2, whereas NRSE1 pan-neuronal matches show a bimodal distribution with a local minimum at 77%, which is below our predicted cut-off for repression activity (Fig. 3B). Based on the PSFM score and its relation to functional assays, these NRSE1 instances are unlikely to be biologically active on their own.

Confusion matrices (Hart et al. 2005) were used as a generalized Venn diagram to compare the overlap of the genes and expression pattern of the different cohorts. Supplemental Figure S5 shows the confusion matrix for NRSE1 versus NRSE2; while both motifs agree on about 323 genes, both cohorts have large sets of nonoverlapping genes (also known as outersects or relative complements; see Methods). The outersect of NRSE1 is comprised of 615 additional genes not present in the NRSE2 cohort, whereas the corresponding NRSE2 outersect includes 172 genes. Neuronal genes comprise 34% of the NRSE2 outersect, but only 14% of the NRSE1 outersect (*P*-value = $5*10^{-9}$, $\chi^2 = 34.07$, χ^2 test for equality of distributions), suggesting that consensus-based approaches like NRSE1 likely miss neuronal, NRSE-associated genes (Supplemental Fig. S6) as also suggested by Zhang et al. (2006).

NRSE2 PSFM matches in multiple vertebrate and invertebrate genomes

NRSE2 matches were sought in genomes representing four invertebrate phyla (arthropod, nematode, echinoderm, and urochor-

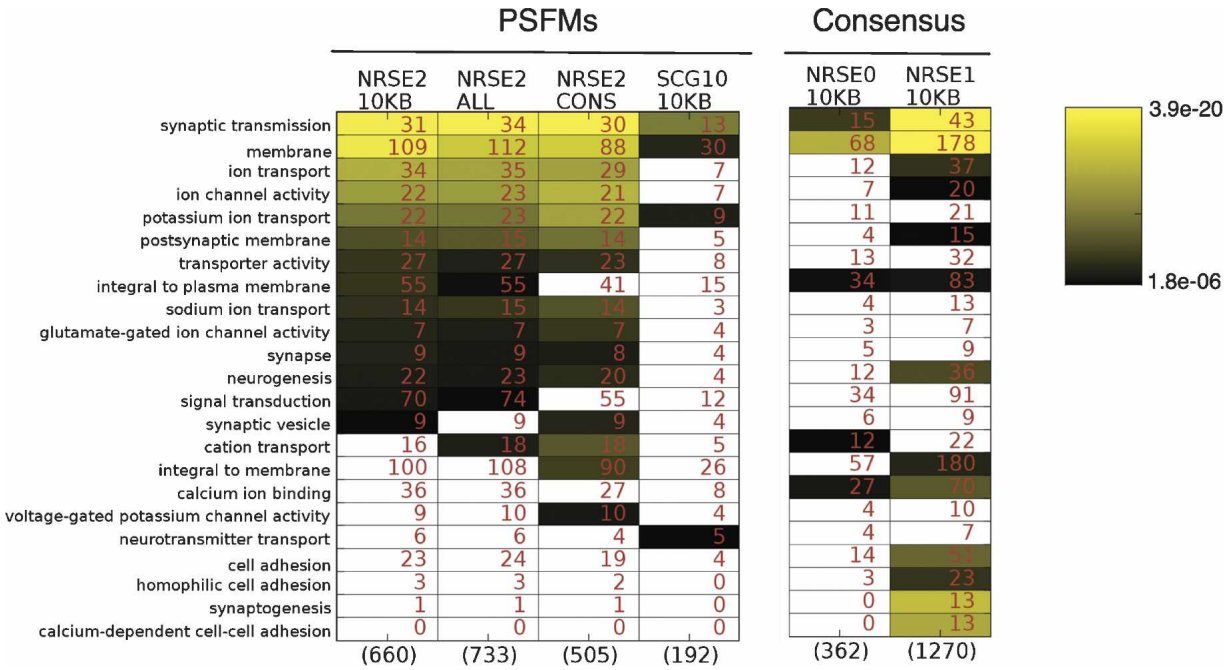

Figure 5. Gene Ontology enrichment comparison of different NRSE *cis*-regulatory cohorts. Cohorts of human genes within 10 kb of a candidate NRSE0 (Schoenherr et al. 1996), NRSE1 (Bruce et al. 2004), SCG10 (the original seed motif), NRSE2, All NRSE2 matches, and conserved NRSE2 matches were filtered of repeat matches and were analyzed for GO term overrepresentation. Significantly enriched GO terms in at least one of the cohorts (of 4576 possible GO terms) are shown. Numbers in cells represent the genes with the term in the cohort, while numbers in parentheses represent the cohort size. Cells shown in color pass the threshold of significance, as determined by a Bonferroni correction. GO terms are sorted in decreasing order by *P*-values of the *leftmost* column. Note that GO enrichments are in terms of decrease in *P*-values, which are directly correlated to the size of the cohorts; the number of genes in the shared association cohort with a particular GO term may go down or stay the same, while its significance increases. The NRSE1 motif behaves differently from the other definitions, as seen in the enrichment of synaptogenesis, which is the result of weak matches within the paralogous protocadherin β family.

date), together with seven additional vertebrate species. The remarkable result is that there are essentially no matches in invertebrate genomes, while all vertebrate genomes have the same order of magnitude of matches, regardless of genome size, with the pufferfish *Tetradon* genome being especially informative (Fig. 7). *Tetraodon* has a highly compressed genome that retains functional sequences such as ORFs at three- to fivefold elevated density. A similar enrichment is seen for NRSE2 occurrences, which suggests that many of them are functional. The notable paucity of NRSE2 sites from the sea urchin, *Drosophila*, and *Ciona* (a urochordate) genomes argues that this repression network is absent up into protochordata, and it calls into question a previous tentative assignment of NRSF orthology to CoREST-interacting zinc fingers in *Caenorhabditis elegans* (Lakowski et al. 2003). We also found that there is only one NRSE2 instance in the entire *C. elegans* genome, and it is not conserved in related worm genomes (*C. briggsae* and *C. remanei*).

NRSE2 PSFM matches in the *Tetraodon* genome were related to matches in the human genome using cisAssociator to identify genes that remain associated with a high-scoring NRSE in both fish and mammals. Table 1 shows the 33 matches that pass our criteria for best reciprocal match of the corresponding gene models. Occurrences of NRSE1 and NRSE2 in human repeats were analyzed using the UCSC RepeatMasker annotations (http://genome.ucsc.edu; Kent et al. 2002; Karolchik et al. 2003) to address whether NRSE instances were found preferentially within the same repeat families. While most NRSE2 matches (285 instances that meet or exceed the 84% match-score threshold of Figure 3) reside mainly in the old vertebrate LINE2 family (226 matches, 79%), the overwhelming majority of NRSE1 consensus matches are in the ERV1 SINE family (1858 of 2339 matches, 79%), which score between 70% and 74%. This dichotomy is particularly striking because there are no NRSE2 matches in the ERV1 family. With two or three strategic chance mutations, many of these repeats could achieve a low-functional match score, upon which selection could operate to favor further optimization.

NRSE2 PSFM matches associated with microRNAs

We proceeded to identify microRNAs in the human genome located within a 25-kb radius of a nonrepeatmasked NRSEs. The search radius was increased from the cisAssociator 10 kb used for the NRSE2 cohort to respond to the observation that some microRNAs are embedded in and expressed as part of primary transcripts from protein-coding genes (Ying and Lin 2004). The sites were mapped against the UCSC entries of the microRNA registry (Griffiths-Jones 2004; Weber 2005). Twenty-one microRNAs were identified (of 326 in the annotations) that represent 16 distinct families. All but one of these microRNAs had been previously characterized in the context of mammalian neuronal differentiation (Sempere et al. 2004; Table 2). MiR-375 was shown separately to be pancreatic β-cell line specific (Poy et al. 2004). It has been shown to target at least one gene (myotrophin) in the murine pancreatic cell line MIN6 in coordination with miR-124a (Krek et al. 2005). Six NRSE-associated miR families also assayed in Sempere belong to 14 families (of 100 surveyed) categorized in Sempere et al. (2004) as "brain specific" or "brain enriched." This

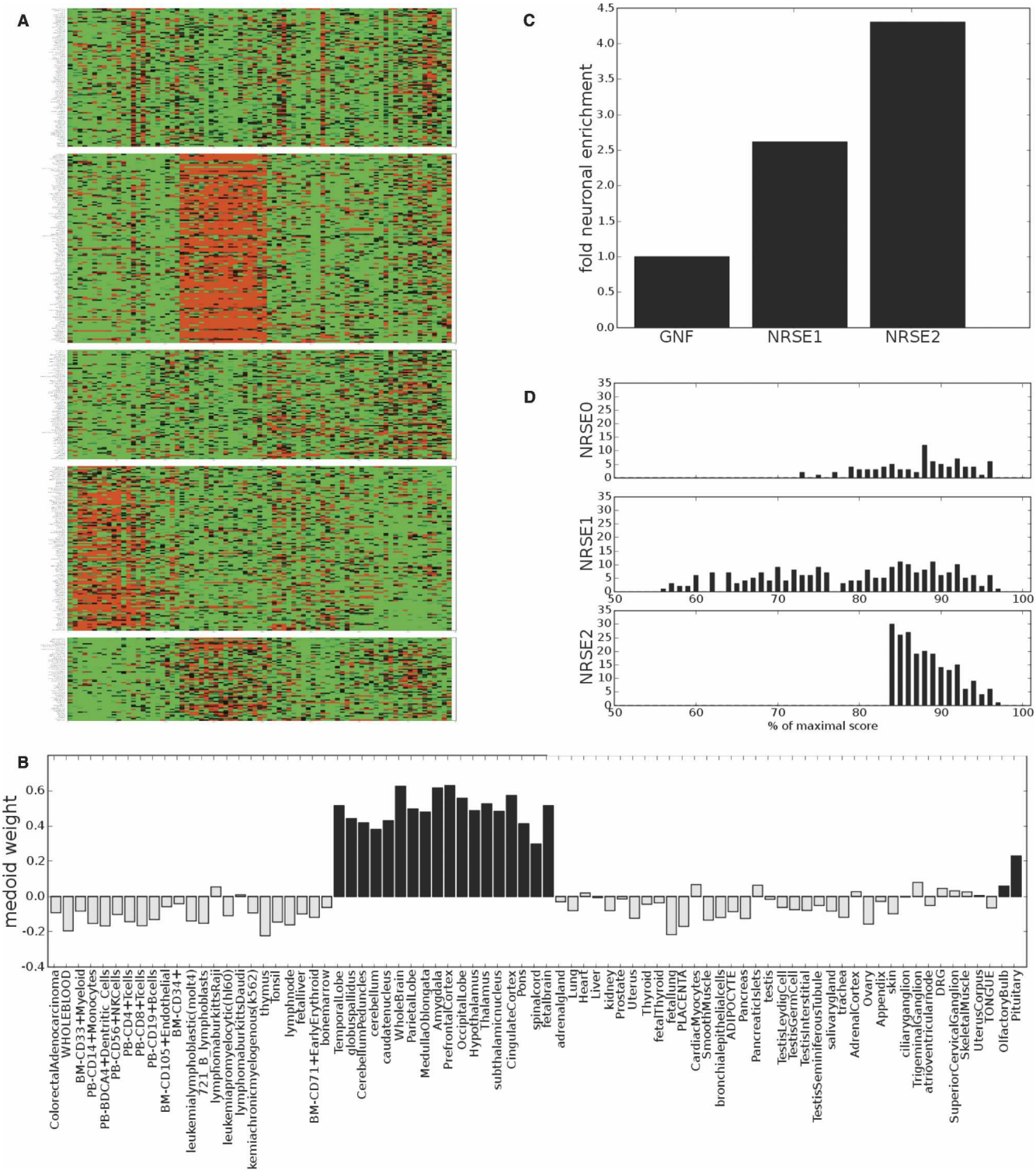

Figure 6. Tissue expression pattern of NRSE associated-genes shows brain-specific expression enrichment. (*A*) Human genes with an NRSE2 (listed in Supplemental Table S2) with an expression pattern in the GNF survey of 79 human tissues, were clustered using the k-medians algorithm as described in the Methods. The second and fifth clusters, which encompass 40% of the NRSE2-associated genes show a clear, brain-specific expression pattern. (*B*) Weights of the k-medoid for cluster 2, with brain tissues highlighted in black. Note that cardiac myocytes and pancreatic islet cells also have positive weights. (*C*) NRSE2 shows a 3.5-fold enrichment of "brain specific" genes (as defined by the medoid in *B*) compared with the GNF data sets, and shows greater enrichment than NRSE1. (*D*) NRSE0 (*top*), NRSE1, and NRSE2 matches associated with genes than have a greater than 0.4 correlation with the medoid vector in *B*. NRSE1 shows a double-humped distribution of matches, with matches weaker than 77% accounting for half of its matches; these low scoring matches are likely false-positives.

Table 1. NRSE2 matches in human genes that are associated with NRSE2 matches in orthologous genes in the pufferfish *Tetraodon nigroviridis*

GeneID	Symbol	Description
1620	*DBC1*	deleted in bladder cancer 1
1756	*DMD*	dystrophin
2259	*FGF14*	Fibroblast growth factor 14
2566	*GABRG2*	γ-aminobutyric acid (GABA) A receptor, γ 2
2903	*GRIN2A*	glutamate receptor, ionotropic, N-methyl D-aspartate 2A
5579	*PRKCB1*	protein kinase C, β 1
6860	*SYT4*	Synaptogamin IV
7432	*VIP*	vasoactive intestinal peptide
8022	*LHX3*	LIM homeobox 3
8514	*KCNAB2*	potassium voltage-gated channel, shaker-related subfamily, β member 2
8693	*GALNT4*	UDP-N-acetyl-α-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 4
9152	*SLC6A5*	solute carrier family 6 (neurotransmitter transporter, glycine), member 5
25983	*NGDN*	Neuroguidin, EIF4E binding protein
51046	*ST8SIA3*	ST8 α-N-acetyl-neuraminide α-2,8-sialyltransferase 3
51151	*SLC45A2*	membrane associated transporter
51289	*RXFP3*	relaxin/insulin-like family peptide receptor 3
55800	*SCN3B*	sodium channel, voltage-gated, type III, β
57468	*SLC12A5*	solute carrier family 12, (potassium-chloride transporter) member 5
57578	*KIAA1409*	
57583	*GPR178*	G protein-coupled receptor 178
64211	*LHX5*	LIM Homeobox 5
79446	*WDR25*	pre-mRNA splicing factor-like
84335′	*GPR123*	G protein-coupled receptor 123
84623	*KIRREL3*	kin of IRRE like 3
91608	*RASL10B*	RAS-like, family 10, member B
118427	*OLFM3*	olfactomedin 3
128434	*C20orf102*	
146664	*MGAT5B*	mannosyl (α-1,6-)-glycoprotein β-1,6-N-acetyl-glucosaminyltransferase, isozyme B
164633	*CABP7*	Calcium-binding protein 7
222662	*LHFPL5*	Lipoma HMGIC fusion partner-like 5
266743	*NPAS4*	neuronal PAS domain protein 4
286046	*XKR6*	XK, Kell blood group complex subunit-related family, member 6
401647	*C10orf132*	

Best reciprocal blasts between the human and *Tetraodon* gene models were used to relate the NRSE2 matches found in human and *Tetraodon*.

pattern of coherent tissue specificity in expression is significant by the criterion of *P*-value of 0.02 (Fisher's exact test). Seven of these microRNAs are located in introns of genes in the NRSE2 cohort, i.e., miR-153 in *PTPRN*, miR-139 in *PDE2A*; miR-9-1 in *CROC4*; miR-7-3 in *C19orf30*; and miR-24-1, miR-27b, as well as miR-23b in *C9orf3*. In the case of miR-153, miR-139, and miR-9-1, the RNA expression pattern of the "host" gene falls in the brain-specific cluster (Fig. 6A). We assayed NRSEs from 11 of these by ChIP, and 10 scored positive for NRSF/REST occupancy (Table 2). Our results for miR-124a and miR-9 agree with those reported by Conaco et al. (2006).

By inspecting lists of predicted target RNAs for NRSE-associated MicroRNAs (Lewis et al. 2005) we found that CoREST (GenBank D31888) is a candidate target for three of our 16 microRNA families (miR-29b, miR-124a, miR-153), and that NRSF itself (GenBank U22680) is a prospective target of miR-153, which has recently been shown to be brain specific in the zebrafish embryo (Kloosterman et al. 2006). These postulated interactions create a potential feedforward loop that might have the effect of more quickly or definitively down-regulating NRSF mRNA, as NRSF activity begins to fall (Fig. 8). This is given additional impetus by the observations that miR-153 is the microRNA with the best-scoring NRSE site (Table 2), and that its NRSE is embedded in *PTPRN*, a gene expressed strongly and widely in the nervous system.

Discussion

Our effort to model the conserved NRSF-binding site and its target gene cohort differs substantially in design, tools, and outcome from prior attempts (Lunyak et al. 2002; Bruce et al. 2004). We show that a successful PSFM site model can be derived from a single-starting conserved NRSE by using iterations of motif refinement that incorporate additional site instances based on their conservation in multiple mammalian genomes. Prior designs started from collections of multiple genes and produced conventional consensus sites. The NRSF PSFM model, unlike standard consensus motif, captures more information about site structure and affords a way to rank score matches according to how well they match the model site. We then tested the model experimentally across a range of PSFM match scores, including below-threshold borderline values, by ChIP/QPCR experiments. This allowed us to assess the predictive qualities of the model relative to PSFM score. These results encourage us to think that other relatively large and well-specified motifs could be usefully modeled in the same manner. However, it is important to recognize that shorter or less well-specified motifs—those with lower information content—will be difficult or even impossible to treat in this manner without additional algorithms to help discriminate functional occurrences from chance occurrences.

The PSFM site model captures more information about site preferences at each position than does a basic consensus. We showed that the PSFM score correlated well with repression activity in transient transfection assays, arguing that it is a good first-order predictor of function. Our ChIP independently showed that a high PSFM match score is predictive of in vivo NRSF occupancy at a given locus. In most prior attempts to develop genome-wide target site models, including NRSF/REST studies, thresholds for membership were set arbitrarily. Based on NRSF/REST results, we think that integration of functional data in this manner is a natural way to bound computational models, establish confidence limits, and then further refine them. However, the apparent intensity of the ChIP interaction differed greatly from one positive locus to another, and we do not yet know what modulates levels of ChIP signal. Obvious biological possibilities include chromatin structure, the presence or absence of various collaborating factors, and contributions from weaker NRSE sites near strong ones.

Cistematic permitted us to efficiently generate and compare families of related models by varying parameters for conservation, position of sites relative to gene anatomy, PSFM match stringency, and initiating seed sites. The ability to do this in an automated manner is useful for finding out whether a model is vulnerable to changes in input parameters. In one pertinent example, we ran the pipeline beginning with different individual starting-site instances as well as a starting-site pool, and found the results are robust to these variations in the initial seed site.

The NRSE2 matches were analyzed for statistically significant functional covariates, from GO and from RNA expression data, using Cistematic modules designed for these purposes. The

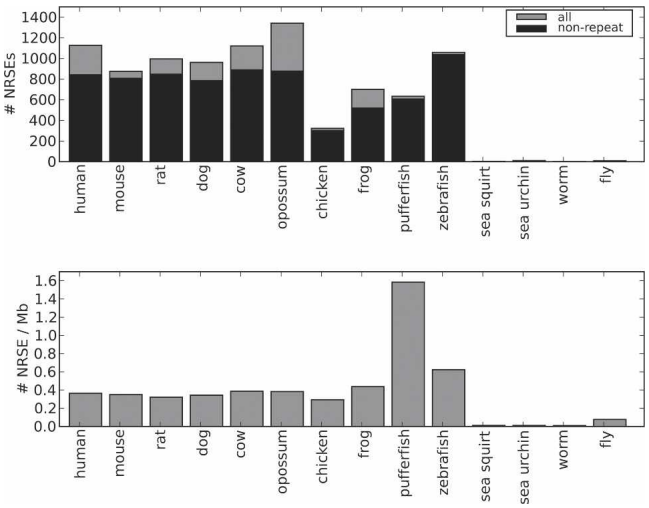

Figure 7. NRSE distribution in vertebrate and invertebrate genomes. (*A*) The number of NRSE2 matches in mammalian genomes is relatively constant and includes a significant number of matches within repeats when compared with other vertebrates and compared with the virtual absence of NRSE2 matches in invertebrates. (*B*) The higher density of all NRSE matches/Mb of genomic sequences in pufferfish and zebrafish when compared with chicken suggest that fish and mammalian NRSE matches may have been expanding independently. Refer to Supplemental Table 5 for the number of matches and the genome size for each genome.

software architecture (Fig. 1; Supplemental Fig. S1) and Open Source license are meant to encourage users to add other analytical modules at will. A key conclusion from these experiments is that the principle function of NRSF could have been inferred solely from analysis of the final NRSE2 cohort model. The enrichment relationships for neuronally expressed RNAs and neuronal GO functions within the NRSE2 cohort model were statistically far above background, despite incompleteness of GO annotations and imperfections in large-scale expression databases. RNA analysis of the NRSE cohort model benefited from strong sampling of brain tissues in the GNF data, and application of this approach to other motifs will be effective as global RNA data sets and GO annotations become more extensive. Had we not already known that NRSF acts as a repressor, this also could have been inferred de novo from the NRSE2 cohort together with expression data for NRSF/REST itself. In mouse and human, the RNA profile for NRSF/REST is in frank opposition to the expression of its direct target repertoire. These inferences show that PSFMs based on evolutionary conservation, and the target gene cohort models derived from them, can successfully predict organismic and molecular functions. The model generates hypotheses at the level of the entire network and also at the level of individual genes (Fig. 8 and below).

We think the approach taken here will be applicable to many transcriptional regulators in vertebrates that meet several criteria. In practical terms, the cardinal requirement is a long and specific binding motif. The length of the NRSE2 PSFM was critical for evading the most dire consequences of Wasserman and Sandelin's "futility theorem," namely, that the vast majority of binding-site instances predicted based on motif knowledge will have no functional significance (Wasserman and Sandelin 2004). Large families of factors whose members are likely to be eligible for Cistematic PSFM models include multifinger zinc finger class regulators that have been expanding rapidly in mammals (Shannon et al. 2003). The second criterion is evolutionary conservation. If a site/factor pair is very new, it will not be possible to leverage conservation, although the addition of increasing num-

Table 2. microRNAs with associated NRSE2 matches in the human genome have a neuronal expression pattern

Name	NRSE2 PSFM (%)	Distance (bp)	Human brain	Mouse brain	P19 + RA	NT2 + RA	ChIP fold enrichment
miR-153–1	97	14,208	Low	Low	Low		87.9
miR-135b	93	10,826	Low	Low	Medium	Low	79.6
miR-124a-2 (*)	92	934	Medium	Medium	Low	Low	
miR-9–1 (*)	91	5,681	High	Medium	High	Low	7.97
miR-29a (clust 1)	91	11,106	Medium	Medium		Low	48.03
miR-29b-1 (clust 1)	91	11,818	Medium	Medium		Low	48.03
miR-212 (clust 2)	88	111				Low	
miR-132 (clust 2)	88	252	Medium	High			
miR-133a-2	88	23,034	Low	Low		Low	10.32
miR-124a-3 (*)	87	487	Medium	Medium	Low	Low	1.00
miR-375	87	9,768	—	—	—	—	8.56
miR-7–3	86	1,097	Medium	Medium	Low	Low	
miR-139	86	2,255	Medium	Medium			29.37
miR-9–3 (*)	86	3,050	High	Medium	High	Low	11.12
miR-124a-1 (*)	86	21,763	Medium	Medium		Low	10.09
miR-124a-3 (*)	86	2,394	Medium	Medium	Low	Low	
mirR-24 (clust 3)	85	1,743					
miR-27b (clust 3)	85	2,319	Medium	Low	Low	Low	
miR-23b (clust 3)	85	2,556	High	Low	Medium	Medium	
miR-203	85	15,684	Low	Low			

MicroRNAs with an NRSE2 match with PSFM score <84% within 25 kb are shown along with their expression pattern from Sempere et al. (2004) in human and mouse brain as well as in mouse P19 and human NT2 cell lines undergoing retinoic acid-induced neuronal differentiation and where several miRs (bold) were categorized as "brain specific" or "brain enriched." Multiple microRNAs that are near the same NRSE are labeled with the same "clust" ID. Entries with asterisks mark members of the same microRNA family that have only one entry in Sempere et al. (2004), and are hence shown with the same expression pattern. miR-375 was found separately to be expressed specifically in pancreatic β cells by Poy et al. (2004). ChIP fold enrichments for those microRNA-associated NRSE2 matches that were part of our 113 sites tested (Supplemental Table S6) that are higher than 2.44 are considered positives.

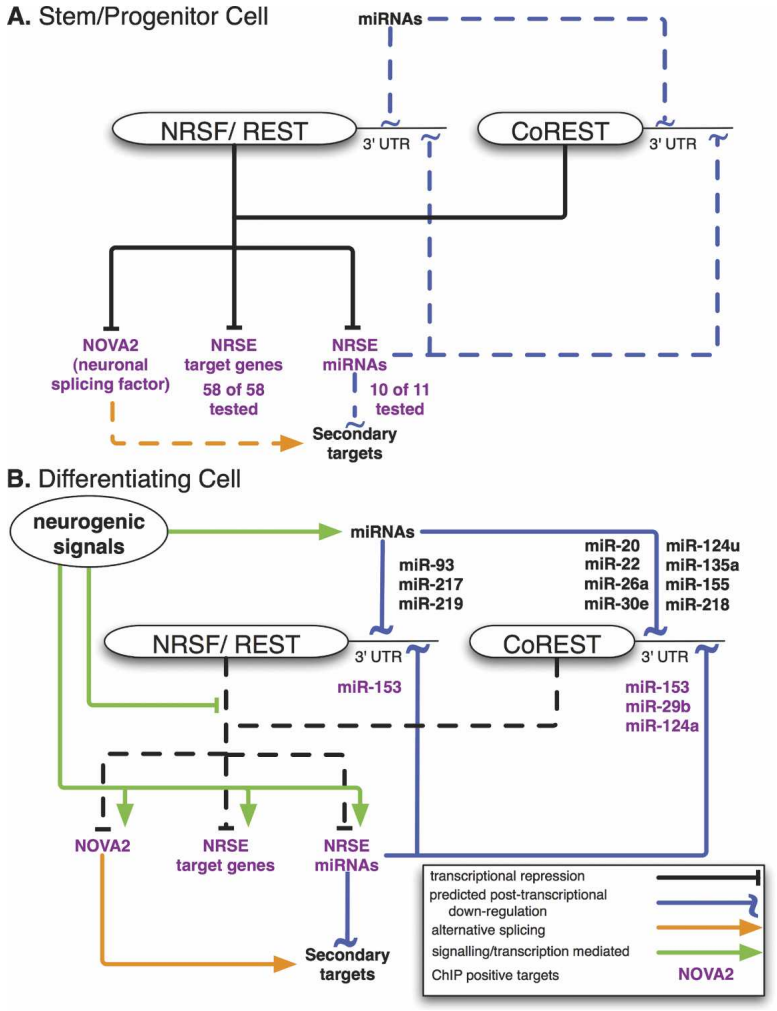

Figure 8. NRSF gene regulatory network model. (*A*) NRSF in conjuction with CoREST and other corepressors prevents the transcription of several hundred targets, including neuronal splicing factors, transcription factors, and microRNAs, as well as many terminal differentiation genes in a stem cell. (*B*) Upon receiving neurogenic signals to terminally differentiate, the NRSF protein is degraded, which leads to derepression of its targets, which are now available to activators. In particular, the NRSE-associated miR-153, which is embedded in the pan-neuronal gene *PTPRN* that has a NRSE in one of its introns, is predicted to down-regulate both NRSF and CoREST mRNAs (which is also the predicted target of the NRSE-associated miR-29b and miR-124a), thus maintaining the derepression.

bers of genomes will provide more branch length and resolution within clades such as the mammals (Boffelli et al. 2004). Finally, whether the data are obtained before the initial PSFM model building or after, quantitative functional analysis of a sample of true positive and true negative sites makes a powerful contribution that can be used to bound model membership and, in the best cases, to predict which instances are likely to be most active in vivo.

The NRSF/REST network is a chordate invention

All currently available data argue that the neuronal NRSF repression network is a chordate invention. Extending the analysis of NRSE2 matches to an additional 11 available genomes (Fig. 7; Supplemental Table S4) revealed that while NRSE2 is not only absent in *Drosophila* as previously noted (Bruce et al. 2004; Dallman et al. 2004; Yeo et al. 2005), but also is essentially absent from all invertebrate genomes. In sharp contrast, all vertebrate genomes we surveyed have between 302 and 1047 nonrepeat matches, with an average of 750. Within mammals, the average number is modestly higher (842). Furthermore, preliminary surveys of *Amphioxis* (a cephalochordate) and lamprey (a basal vertebrate) whole-genome shotgun traces found that NRSE2 matches are present in both at high densities, while the motif is entirely absent from the urochordate, *Ciona intestinalis*. This, along with the absence of any gene models that are convincingly similar to NRSF in *Ciona* or invertebrate genomes, suggests that NRSF emerged after the time of the last common ancestor shared by vertebrate and urochordates. Paralleling this, NRSF/REST itself is present and highly conserved in all vertebrate genomes but absent from *Ciona* and multiple invertebrate genomes. We did not detect NRSF in searches of sea urchin or *C. elegans*, and others have reported it absent from *Drosophila*, even though its principal corepressors are present there (Dallman et al. 2004; Yeo et al. 2005). We did not detect NRSF in amphioxus trace coverage either, which could be a simple technical issue, but also raises the possibility that the target motif might have emerged ahead of the factor itself.

These data, combined with the existence of high-scoring sites within old LINE2 elements in the human genome, suggest that NRSEs may have first been distributed across vertebrate genomes via repeats at roughly the same time the NRSF DNA-binding factor first appeared. In such a scenario, NRSEs that land near or in genes and also confer some advantage when repressed by NRSF are starting points to expand an NRSF network. The much larger reservoir of weak, probably inert, NRSE1 (Bruce consensus) sites present in other repeat families might provide new NRSF/target gene pairs, given one or two key mutations.

A subset of neuronal genes belong to the NRSE cohort

RNA expression and GO term analyses showed that, under the NRSE2 model, NRSF does not directly act on a majority of genes with broad brain expression or with distinctly neuronal GO classifications. There are roughly 1400 genes preferentially and broadly expressed in adult brain, but only 11% of these have a high-confidence NRSE2 motif. Some of the non-NRSE brain genes are probably glial, while another subset might be explained by weaker NRSEs, functioning individually or multiply. *NeuroD1/BETA2,* for example, is an attractive candidate target based on its expression pattern and function in neurogenesis and pancreatic islet cell genesis (Lee et al. 1995; Huang et al. 2000). It has one NRSE ~4.5 kb upstream that scores above our threshold in

mouse and dog, but slips below threshold in human. However, closer inspection shows that NeuroD1, like the related factors, NeuroG1 and NeuroG2, has additional low-scoring NRSE matches embedded in its ORF. Learning the rules governing use of weaker sites awaits a fully comprehensive experimental mapping of NRSF/REST in vivo interactions, but many neuronal genes probably depend on other factors for their neuronal expression. A corollary is that substantial numbers of additional pan-brain genes present in relaxed-stringency models, including the NRSE1 cohort, are likely neuronal due to other regulatory factors, rather than by the action of a functional NRSE.

The converse is also true. Significant (approximately fourfold) enrichment of the NRSE2 cohort for a brain expression profile leaves 60% unaccounted for. Some reasons for this include incomplete gene annotations, genes restricted to specific kinds of neurons, mRNAs present at levels below microarray threshold, and inclusion of some extra NRSE2 neighborhood genes into the model by the cisAssociator algorithm. For example, several of the NRSE2-associated transcription factors are well known for important functions in specific neuronal populations (Neurogenin-3, POU4F1, POU4F3, LHX3, and LHX5), but none are in the pan-brain cluster, nor is their expression utterly specific to brain. It is also unclear how many genes in this model cohort might be targets of NRSF regulation relevant to its cardiac, pancreatic, or other functions.

NRSF/REST interactions at neuronal transcription factor, microRNA, and RNA-splicing factor loci

The NRSE2 model target gene cohort included other transcription factors, microRNAs, and splicing regulatory factors, all of which could extend the regulatory effects of NRSF/REST. Multiple NRSE instances are associated with transcription factors. Table 1 highlights that, in addition to an expected complement of channels and synaptic proteins, highly conserved NRSE instances shared between human and fish are associated with transcription factors of interest. LHX5 and LHX3 are LIM homeobox factors important for specification and function of distinct neuronal populations. LHX5 also controls regulation of neuronal precursor exit from the cell cycle in the hippocampus (Zhao et al. 1999). Among NRSE2 instances conserved among mammals, there are at least 25 other transcription factors, including NeuroD2 (McCormick et al. 1996), a known mediator of neuronal differentiation; its conserved NRSE is located ~13 kb downstream in mammalian genomes and was validated by the ChIP experiments. Another proneural transcription factor with an NRSE is Neurogenin-3, which marks both a subset of neuronal precursors and the early precursors of pancreatic islet cells (Sommer et al. 1996; Gradwohl et al. 2000). In addition, several genes encoding RNA-binding proteins involved in RNA splicing and editing have NRSEs. Among these, *NOVA2* is especially interesting because it regulates brain-specific RNA splicing for a substantial group of synaptic proteins (Ule et al. 2005). Both of *NOVA2*'s NRSEs (one in the third intron, the other one downstream in a LINE2 repeat) were occupied by NRSF/REST according to the ChIP data.

NRSE2 matches are also associated with multiple neuronal microRNAs, several of which (miR-9-1, miR-9-3, miR-29a/miR-29b, miR-124a-1, miR-133, miR-135b, miR-139, miR-153, miR-375), were validated by ChIP. This suggests the circuit model in Figure 8. In stem cells and progenitors of Figure 8A, NRSF acts by repressing hundreds of protein-coding genes and a handful of microRNA genes. Upon developmental progression to the differentiated state (Fig. 8B), NRSF is down-regulated, first at the protein level and then transcriptionally (Ballas et al. 2005). Thus, its targets are freed—perhaps sequentially according to NRSE strength and number—for induction by various transcription activators. In this model, feedforward connections of microRNAs onto CoREST and NRSF may modulate or accelerate the change from precursor cell to neuron. MicroRNAs and splicing factors can go on to down-regulate other target genes not wanted in differentiating neurons. This extended reach of NRSF from direct negative regulation to indirect positive regulation may also explain why only a fraction of neuronal genes are direct NRSF targets. Embryonic lethality of NRSF null mice at day E10.5, before the onset of neurogenesis (Chen et al. 1998), might therefore result from misexpression of neuronal microRNAs or splicing factors.

Methods

Cistematic

Cistematic is a Python package for automated motif identification in eukaryotic genomes. Cistematic has a three-tiered architecture of objects written in the Python scripting language, which encapsulate the concepts of motifs, genome sequences, and annotations, as well as motif-finding programs (Supplemental Fig. S1). The sequences and annotations that Cistematic uses for vertebrate genomes are derived from the UCSC Genome Database. Primary objectives of Cistematic are to identify, refine, and/or map candidate motifs by determining their genome-wide distribution, their association with potentially coexpressed or coregulated genes, and their GO enrichment.

A typical Cistematic script consists of Python commands that perform a set of operations on certain Cistematic objects. A set of Experiment objects provides ready-made logic to do much of the work for the user. Most of these Experiment objects are designed to handle various aspects of phylogenetic footprinting across multiple metazoan and fungal genomes. Cistematic stores all of its information and results in SQL-queryable databases, using the Sqlite 3.0 database library and the pysqlite 2.0 Python library. Cistematic can also generate tab-delimited files that can be imported into Excel for browsing. Cistematic currently runs on Mac OS X, Linux, and Solaris with Python 2.4 and sqlite installed and is available at http://cistematic.caltech.edu, along with the scripts used to generate the data in this study, which are available at http://cistematic.caltech.edu/~alim/cispaper

Motif similarity score

We define the motif similarity score of two PSFMs A and B as:

$$\mathrm{MSS}(A, B) = \mathrm{Max}(\Sigma\ \mathrm{PearsonCorr}(A_i,B_i),\Sigma\ \mathrm{PearsonCorr}(A_i,\mathrm{revB}_i))/\mathrm{length}(A)$$

where the index i represents the corresponding columns in the PSFMs, revB is the reverse complement PSFM of B, and PearsonCorr is the Pearson Correlation. The MSS of two motifs ranges between 0 and 1.0.

Genome-wide *cis*-regulatory cohort identification

We used the Cistematic Locate experiment object class to map every instance of our motifs in human, mouse, and dog with either the consensus or the PSFM. The consensus score for a candidate window m of length L was calculated as:

$$\sum_i f_i(m_i)$$

where f_i is the frequency of the nucleotide at position m_i in the i^{th} column of the PSFM.

The best possible score for each PSFM was calculated and all matches that scored higher than the best score times a predetermined threshold (see Results and Fig. 3) were accepted as matches. We have found that this particular scoring function performs as well as the traditional log-likelihood scoring (data not shown), allows us to use PSFMs without resorting to pseudocounts or Dirchlet distributions to account for unseen valid nucleotides, and that the threshold can be intuitively related to the number of mismatches of the site to the consensus of the PSFM (about 5% per major mismatch in the case of NRSE2).

One or more genes were identified for every match as members of the *cis*-regulatory cohort using the criteria that the match instance is (1) within the gene model or (2) within a 10-kb radius of either the 3′ or 5′ gene model boundaries. The relative location of the motif to each neighboring gene was noted as upstream, 5′-UTR, coding sequence, intron, 3′-UTR, or downstream. Results from each genome were saved to a separate file to serve as inputs for the ensuing steps of the analysis.

We used the following annotations from NCBI or UCSC along with the corresponding genomic sequences from UCSC: human (NCBI Build 35), mouse (NCBI Build 35), dog (NCBI Build 2), and *Tetraodon* (geneid, UCSC tetNig1).

Orthology matching

Genes from each genome that were flagged as neighbors in our genome-wide search were cross-matched using the Cistematic orthology database, which is built from a combination of NCBI's HomoloGene (version 41.2) supplemented with precomputed best reciprocal BLAST searches for additional genomes that are not yet included into HomoloGene. Cistematic considers a motif occurrence in genes in a genome (human) conserved if the orthologous gene was present in the genome-wide search results for one or more of the other genomes (here mouse and dog) or if it was present in another paralog in the original genome.

cisMatcher and motif refinement

Cistematic can identify motif conservation arbitrarily far from a gene using the cisMatcher algorithm, as outlined in Supplemental Figure S2 and described below. The objective is to be able to specify gene proximity with flexibility that will bring in all flanking sequence to the next gene, for example, whether that distance is several hundred kilobases in a gene desert or only 1 kb in a gene-dense neighborhood. Cistematic genome-wide results were purged of matches that were marked as occurring within repeats; operationally, these are any of the partially or completely lowercase matches in the genomic sequences from UCSC, which are soft-RepeatMasked. Remaining matches were used to retrieve 65-bp sequences with 22 bp upstream from the motif, the motif itself, and the remainder downstream of the motif, which were saved in one file in fasta format per genome. The resulting sequence files from mouse and dog were used to build a BLAST database, which was then searched using the human sequences. For each human match, the best match with an e-value less than 0.01 with length longer than 25 bp and similarity >85% in each of mouse and dog were imported into a custom sqlite database. A query was used to retrieve the best mouse or dog match for each human sequence that was available. For each human match with a conserved match in another genome, the nearest human gene within 200 kb was mapped using a radius that was expanded in 1-kb increments; in cases where more than one gene are within the same radius, the one with the lower starting numerical coordinate on the pseudomolecule was picked. Matches were annotated as occurring upstream, in the 5′-UTR or 3′-UTR, in the coding sequence, introns, or downstream relative to their nearest gene. Matches within coding sequences were optionally filtered where indicated. Matching sequences and their corresponding gene and relative locations are sorted in decreasing order of similarity and length.

Human matches that were picked up by cisMatcher as well as their corresponding mouse and/or dog matches are then used to calculate the refined PSFM, which can then be used again by Cistematic to repeat the analysis.

Gene Ontology analysis

Cistematic can flag particular Gene Ontology terms as enriched or depleted, at a statistically significant level, for any set of genes. Cistematic tabulates Gene Ontology (GO) terms associated with a gene cohort using its own GO annotations, provided for mammalian genomes from NCBI's loc2go data set. *P*-values are calculated for every GO term using the hypergeometric. We apply a stringent protocol for significance in which the Bonferroni correction is applied to account for multiple hypotheses testing, where each GO term in the genome represents a hypothesis. We report as significantly enriched or depleted GO terms that (1) are still significant following the Bonferroni correction, and that (2) contain more than 15 genes in the genome. Note that we only show the GO terms (rows) in our GO summary figures that have at least one statistically significant enrichment or depletion in one cohort (column) included in each figure.

To test for the robustness of our analysis, we also generated 100 motifs, where we scrambled the order of the columns in NRSF and repeated the entire analysis pipeline in Figure 1 and asked whether we recovered any enriched GO terms.

Expression analysis

The GNF expression data set was pre-processed by discarding all entries with NCBI gene ID's that are missing or that are not found in the latest NCBI human annotations. If more than one expression pattern for the same gene ID was available, only the first one was kept. For the remaining genes, tissue replicates were averaged and each gene was median centered.

Confusion matrices were done as by Hart el al. (2005) with the following modifications to accommodate genes present in only one cohort. Outersects were defined as the relative complement of each cluster i of set A with respect to set B, i.e.,

$$A_i \setminus B = \{x \mid x \in A_i, x \notin B\}$$

Cell culture conditions

Culture conditions were as follows: Jurkat cells were grown in Advanced RPMI 1640 (GIBCO Invitrogen Cell Culture) supplemented with 15% fetal bovine serum, 100 U/mL of penicillin-streptomycin, and 1× Glutamax (GIBCO Invitrogen Cell Culture) at 37°C with 5% CO_2.

Chromatin immunoprecipitation

This protocol was adapted from the laboratory of Peggy Farnham (http://mcardle.oncology.wisc.edu/farnham/protocols). We cross-linked the Jurkat cells by adding formaldehyde to a final concentration of 1% for 10 min. Cross-linking was stopped by adding glycine to a final concentration of 0.125 M. Then, we collected 2×10^7 cells per IP and washed once with 1× phosphate-buffered saline (PBS). We resuspended the cells in lysis buffer (5 mM 1,4-piperazine-bis-[ethanesulphonic acid], at pH 8.0, 85 mM KCl, 0.5% NP-40, Protease Inhibitor Cocktail [Roche]) and centrifuged to collect the crude nuclear preparation.

We resuspended the crude nuclear preparation in RIPA buffer (1× PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate [SDS], Protease Inhibitor Cocktail) and sonicated at power output 5–6 with the Sonics Vibra-Cell VC130 (Sonics) four times for 30 sec each on ice to produce an average DNA fragment size of 500 bp. We centrifuged the chromatin solution at 4°C for 15 min at 20,000 rcf. Sonicated chromatin was incubated with NRSF mouse monoclonal antibody (12C11; Chen et al. 1998) coupled to sheep anti-mouse IgG magnetic beads (Dynabeads M-280, Invitrogen). After bead pelleting, the supernatant was retained as mock IP DNA for use in quantitative PCR. The magnetic beads were washed five times with wash buffer (100 mM Tris, 500 mM LiCl, 1% NP-40, 1% Deoxycholate) and washed once with TE (10 mM Tris at pH 8.0, 1 mM EDTA). After washing, the bound DNA was eluted by heating the beads to 65°C in elution buffer (0.1 M $NaHCO_3$ and 1% SDS). The eluted DNA and mockIP DNA were incubated at 65°C for 12 h more to reverse the cross-links. Then, we extracted with phenolchloroform and back extracted the organic phase once. We concentrated the DNA in the aqueous phase using the QIAquick PCR Purification Kit (Qiagen), substituting 3 vol of Qiagen Buffer PM for 5 vol of Qiagen Buffer PB.

Quantitative PCR

We used Primer3 software to design primers by inputting 500 bp of upstream genomic sequence and 500 bp downstream of each predicted NRSE. Each primer pair was required to flank the NRSE. We performed real-time PCR to quantitate the absolute amount of enriched DNA for each NRSE (amplicon size range between 60 and 217 bp, average size of 79 bp). Each reaction contained 3.5 mM $MgCl_2$, 0.125 mM dNTPs, 0.5 uM forward primer, 0.5 uM reverse primer, 0.1× Sybr Green (Molecular Probes Invitrogen Detection Technologies), 1 U Stoffel fragment (Applied Biosystems), and template DNA in a final volume of 20 uL. For each amplicon, we measured a standard curve of 50 ng, 5 ng, 500 pg, and 50 pg mock IP DNA in addition to our replicate ChIP DNA samples. We measured product accumulation for 40 cycles on the Bio-Rad Icycler and calculated the threshold cycle for each dilution of the standard curve. We then performed a linear regression to fit the threshold cycle from our ChIP DNA sample to this standard curve and divided that result by the amplicon size to measure the absolute number of genomic equivalents of that NRSE in the pool of ChIP DNA. We measured the levels of five random nongenic, nonconserved regions in each ChIP DNA preparation to normalize for any variation in absolute quantities of DNA in each prep.

Acknowledgments

We thank Drs. Erich Schwarz and Brian Williams for thoughtful comments on the manuscript; Dr. Ken McCue for advice on statistical analysis; members of the Wold and Myers groups, Sarah Aerni, and Profs. David Bartel and Paul Sternberg for helpful discussions of the project. This work was supported by a NIH/NSF Southern California Bioinformatics Summer Institute Fellowship and a NIH Training Grant Fellowship to A.M., a Hubert Shaw and Sandra Lui Stanford Graduate Fellowship to E.C.L.T., a Stanford Genome Training Program traineeship to S.T.G. (NHGRI grant T32 HG00044), NIH grant U01 HG003162 to R.M.M., and grants from DOE BER, NASA, and NIHGMS to B.W.

References

Abderrahmani, A., Steinmann, M., Plaisance, V., Niederhauser, G., Haefliger, J.A., Mooser, V., Bonny, C., Nicod, P., and Waeber, G. 2001. The transcriptional repressor REST determines the cell-specific expression of the human MAPK8IP1 gene encoding IB1 (JIP-1). *Mol. Cell. Biol.* **21:** 7256–7267.

Andres, M.E., Burger, C., Peral-Rubio, M.J., Battaglioli, E., Anderson, M.E., Grimes, J., Dallman, J., Ballas, N., and Mandel, G. 1999. CoREST: A functional corepressor required for regulation of neural-specific gene expression. *Proc. Natl. Acad. Sci.* **96:** 9873–9878.

Atouf, F., Czernichow, P., and Scharfmann, R. 1997. Expression of neuronal traits in pancreatic β cells—Implication of neuron-restrictive silencing factor/repressor element silencing transcription factor, a neuron-restrictive silencer. *J. Biol. Chem.* **272:** 1929–1934.

Ballas, N., Grunseich, C., Lu, D.D., Speh, J.C., and Mandel, G. 2005. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell* **121:** 645–657.

Boffelli, D., Weer, C.V., Weng, L., Lewis, K.D., Shoukry, M.I., Pachter, L., Keys, D.N., and Rubin, E.M. 2004. Intraspecies sequence comparisons for annotating genomes. *Genome Res.* **14:** 2406–2411.

Bruce, A.W., Donaldson, I.J., Wood, I.C., Yerbury, S.A., Sadowski, M.I., Chapman, M., Göttgens, B., and Buckley, N.J. 2004. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc. Natl. Acad. Sci.* **101:** 10458–10463.

Chen, Z.F., Paquette, A.J., and Anderson, D.J. 1998. NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat. Genet.* **20:** 136–142.

Chong, J.A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J.J., Zheng, Y., Boutros, M.C., Altshuller, Y.M., Frohman, M.A., Kraner, S.D., and Mandel, G. 1995. REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80:** 949–957.

Conaco, C., Otto, S., Han, J., and Mandel, G. 2006. Reciprocal actions of REST and a microRNA promote neuronal identity. *Proc. Natl. Acad. Sci.* **103:** 2422–2427.

Dallman, J.E., Allopenna, J., Bassett, A., Travers, A., and Mandel, G. 2004. A conserved role but different partners for the transcriptional corepressor CoREST in fly and mammalian nervous system formation. *J. Neurosci.* **24:** 7186–7193.

Dehal, P., Predki, P., Olsen, A.S., Kobayashi, A., Folta, P., Lucas, S., Land, M., Terry, A., Zhou, C.L.E., Rash, S., et al. 2001. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science* **293:** 104–111.

Fickett, J.W. 1996. Quantitative discrimination of MEF2 sites. *Mol. Cell. Biol.* **16:** 437–441.

Gradwohl, G., Dierich, A., LeMeur, M., and Guillemot, F. 2000. Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl. Acad. Sci.* **97:** 1607–1611.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res.* **32:** D109–D111.

Hamilton, A.T., Huntley, S., Kim, J., Branscomb, E., and Stubbs, L. 2003. Lineage-specific expansion of KRAB zinc-finger transcription factor genes: Implications for the evolution of vertebrate regulatory networks. *Cold Spring Harb. Symp. Quant. Biol.* **68:** 131–140.

Hart, C.E., Sharenbroich, L., Bornstein, B.J., Trout, D., King, B., Mjolsness, E., and Wold, B.J. 2005. A mathematical and computational framework for quantitative comparison and integration of large-scale gene expression data. *Nucleic Acids Res.* **33:** 2580–2594.

Hersh, L.B. and Shimojo, M. 2003. Regulation of cholinergic gene expression by the neuron restrictive silencer factor/repressor element-1 silencing transcription factor. *Life Sci.* **72:** 2021–2028.

Huang, Y.F., Myers, S.J., and Dingledine, R. 1999. Transcriptional repression by REST: Recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat. Neurosci.* **2:** 867–872.

Huang, H.P., Liu, M., El-Hodiri, H.M., Chu, K., Jamrich, M., and Tsai, M.J. 2000. Regulation of the pancreatic islet-specific gene β2 (neuroD) by neurogenin 3. *Mol. Cell. Biol.* **20:** 3292–3307.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. 2004. Human MicroRNA targets. *PLoS Biol.* **2:** 1862–1879.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Kloosterman, W.P., Wienholds, E., de Bruijn, E., Kauppinen, S., and Plasterk, R.H. 2006. In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nat. Methods* **3:** 27–29.

Kosik, K.S. and Krichevsky, A.M. 2005. The elegance of the microRNAs: A neuronal perspective. *Neuron* **47:** 779–782.

Krebs, C.J., Larskins, L.K., Khan, S.M., and Robins, D.M. 2005.

Expansion and diversification of KRAB zinc-finger genes within a cluster including regulation of sex-limitation 1 and 2. *Genomics* **6:** 752–761.
Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., daPiedade, I., Gunsalus, K.C., Stoffel, M., et al. 2005. Combinatorial microRNA target predictions. *Nat. Genet.* **37:** 495–500.
Kuwahara, K., Saito, Y., Takano, M., Arai, Y., Yasuno, S., Nakagawa, Y., Takahashi, N., Adachi, Y., Takemura, G., Horie, M., et al. 2003. NRSF regulates the fetal cardiac gene program and maintains normal cardiac structure and function. *EMBO J.* **22:** 6310–6321.
Kuwabara, T., Hsieh, J., Nakashima, K., Taira, K., and Gage, F.H. 2004. A small modulatory dsRNA specifies the fate of adult neural stem cells. *Cell* **116:** 779–793.
Lakowski, B., Eimer, S., Gobel, C., Bottcher, A., Wagler, B., and Baumeister, R. 2003. Two suppressors of sel-12 encode C2H2 zinc-finger proteins that regulate presenilin transcription in *Caenorhabditis elegans*. *Development* **130:** 2117–2128.
Lee, J.E., Hollenberg, S.M., Snider, L., Turner, D.L., Lipnick, N., and Weintraub, H. 1995. Conversion of *Xenopus* ectoderm into neurons by NeuroD, a basic helix-loop-helix protein. *Science* **268:** 836–844.
Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773.
Lunyak, V.V., Burgess, R., Prefontaine, G.G., Nelson, C., Sze, S.H., Chenoweth, J., Schwartz, P., Pevzner, P.A., Glass, C., Mandel, G., et al. 2002. Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. *Science* **298:** 1747–1752.
McCormick, M.B., Tamimi, R.M., Snider, L., Asakura, A., Bergstrom, D., and Tapscott, S.J. 1996. neuroD2 and neuroD3: Distinct expression patterns and transcriptional activation potentials within the neuroD gene family. *Mol. Cell. Biol.* **16:** 5792–5800.
Mori, N., Schoenherr, C., Vanderbergh, D.J., and Anderson, D.J. 1992. A common silencer element in the SCG10 and type II Na+ channel gene binds a factor present in non-neuronal cells but not in neuronal cells. *Neuron* **9:** 45–54.
Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X.S., MacDonald, P.E., Pfeffer, B., Tuschl, T., Rajewsky, N., Rorsman, P., et al. 2004. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432:** 226–230.
Schoenherr, C.J. and Anderson, D.J. 1995. The neuron-restrictive silencer factor (Nrsf)—a coordinate repressor of multiple neuron-specific genes. *Science* **267:** 1360–1363.
Schoenherr, C.J., Paquette, A.J., and Anderson, D.J. 1996. Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl. Acad. Sci.* **93:** 9881–9886.
Scholl, T., Stevens, M.B., Mahanta, S., and Strominger, J.L. 1996. A zinc finger protein that represses transcription of the human MHC class II gene, DPA(1,2). *J. Immunol.* **156:** 1448–1457.
Sempere, L.F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol.* **5:** R13.
Shannon, M., Hamilton, A.T., Gordon, L., Branscomb, E., and Stubbs, L. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. *Genome Res.* **13:** 1097–1110.
Sommer, L., Ma, Q., and Anderson, D.J. 1996. Neurogenins, a novel family of atonal-related bHLH transcription factors, are putative mammalian neuronal determination genes that reveal progenitor heterogeneity in the developing CNS and PNS. *Mol. Cell. Neurosci.* **8:** 221–241.
Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101:** 6062–6067.
Thiel, G., Lietz, M., and Cramer, M. 1998. Biological activity and modular structure of RE-1-silencing transcription factor (REST), a repressor of neuronal genes. *J. Biol. Chem.* **273:** 26891–26899.
Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., et al. 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37:** 844–852.
Wagner, S., Hess, M.A., Ormonde-Hanson, P., Malandro, J., Hu, H.P., Chen, M., Kehrer, R., Frodsham, M., Schumacher, C., Beluch, M., et al. 2000. A broad role for the zinc finger protein ZNF202 in human lipid metabolism. *J. Biol. Chem.* **275:** 15685–15690.
Wasserman, W.W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5:** 276–287.
Weber, M.J. 2005. New human and mouse microRNA genes found by homology search. *FEBS J.* **272:** 59–73.
Yeo, M., Lee, S.K., Lee, B., Ruiz, E.C., Pfaff, S.L., and Gill, G.N. 2005. Small CTD phosphatases function in silencing neuronal gene expression. *Science* **307:** 596–600.
Ying, S.Y. and Lin, S.L. 2004. Intron-derived microRNAs—fine tuning of gene functions. *Gene* **342:** 25–28.
Zhang, C., Xuan, Z., Otto, S., Hover, J.R., McCorkle, S.R., Mandel, G., and Zhang, M.Q. 2006. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.* **34:** 2238–2246.
Zhao, Y.U., Sheng, H.Z., Amini, R., Grinberg, A., Lee, E., Huang, S.P., Taira, M., and Westphal, H. 1999. Control of hippocampal morphogenesis and neuronal differentiation by the LIM homeobox gene *Lhx5*. *Science* **284:** 1155–1158.

Received December 5, 2005; accepted in revised form July 19, 2006.