Cell

Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity

Outi Hallikas,[1,4] Kimmo Palin,[2,4] Natalia Sinjushina,[3] Reetta Rautiainen,[1] Juha Partanen,[3] Esko Ukkonen,[2,*] and Jussi Taipale[1,*]
[1] Molecular and Cancer Biology Program, Biomedicum Helsinki, PO Box 63
[2] Department of Computer Science, PO Box 68
[3] Developmental Biology Program, Institute of Biotechnology, PO Box 56
FIN-00014 University of Helsinki, Finland
[4] These authors contributed equally to this work.
*Contact: ukkonen@cs.helsinki.fi (E.U.); jussi.taipale@helsinki.fi (J.T.)
DOI 10.1016/j.cell.2005.10.042

SUMMARY

Understanding the regulation of human gene expression requires knowledge of the "second genetic code," which consists of the binding specificities of transcription factors (TFs) and the combinatorial code by which TF binding sites are assembled to form tissue-specific enhancer elements. Using a novel high-throughput method, we determined the DNA binding specificities of GLIs 1–3, Tcf4, and c-Ets1, which mediate transcriptional responses to the Hedgehog (Hh), Wnt, and Ras/MAPK signaling pathways. To identify mammalian enhancer elements regulated by these pathways on a genomic scale, we developed a computational tool, enhancer element locator (EEL). We show that EEL can be used to identify Hh and Wnt target genes and to predict activated TFs based on changes in gene expression. Predictions validated in transgenic mouse embryos revealed the presence of multiple tissue-specific enhancers in mouse c-*Myc* and N-*Myc* genes, which has implications for organ-specific growth control and tumor-type specificity of oncogenes.

INTRODUCTION

Identifying mutations responsible for developmental defects and human diseases has made a major contribution to our understanding of biological processes. However, in part because protein-coding regions of genes present larger targets for mutagenesis than TF binding sequences, many genetic analyses are biased toward detecting mutations that affect the activity of proteins rather than the function of elements that regulate gene expression. Therefore, processes that depend on precise transcriptional control, such as regulation of cell proliferation, are presently relatively poorly understood. Cases where growth appears to be controlled in a tissue-specific manner have proven particularly resistant to genetic dissection. Open questions related to such tissue-specific growth control include regulation of organ size (Conlon and Raff, 1999), the tissue specificity of growth-factor signals, and the tumor-type selectivity of oncogenes.

It is well established that cell proliferation can be induced by oncogenic or growth-factor-activated TFs, such as GLI2 or Tcf4, whose activities are regulated by the Hh and Wnt signaling pathways, respectively (Bienz and Clevers, 2000; Taipale and Beachy, 2001). However, the conserved enhancer or promoter elements through which these TFs regulate the expression of cell-cycle regulatory genes in vivo are generally not known.

Several reasons have made the identification of such elements difficult. First, the information about TF binding specificity is often incomplete, in part due to the difficulty of measuring affinities of large numbers of TFs to DNA using methods such as electrophoretic mobility shift assay (EMSA; Fried and Crothers, 1981) or SELEX (Roulet et al., 2002). Second, the identification of mammalian enhancer elements by computational or experimental methods has proven to be challenging.

Genome-wide in silico analyses of conserved mammalian regulatory sequences have largely concentrated on untranslated regions of mRNAs (Xie et al., 2005) or promoter elements (Suzuki et al., 2004; Xie et al., 2005), ~1–3 kb sequences located immediately upstream of the transcription start site. However, the enhancer elements that control promoter activity are often located quite far from the transcription start site (> > 10 kb). A single promoter can be regulated by one or many relatively short (~1 kb) enhancer modules,

which are activated by binding of multiple TFs. If multiple enhancer modules regulate one promoter, the corresponding gene is expressed in all tissues where one or more of the enhancer elements are active, and thus the expression pattern of a gene reflects the combined activity of all the enhancer modules that are capable of activating its transcription (reviewed in Michelson, 2002).

Whereas in silico methods efficiently identify enhancer modules in *Drosophila* (Michelson, 2002), mammalian enhancer prediction on a genomic scale has not been possible due to the higher complexity of mammalian genomes. Experimentally, individual mammalian enhancer elements regulating a particular gene can be identified by locating genomic sequences that direct tissue-specific expression of marker genes in transgenic embryos (Spitz et al., 2003), followed by progressive deletion of these sequences, often aided by analysis of conservation of the sequences in multiple species. However, this is a difficult and time-consuming process that is not easily adaptable for genome-wide studies.

In this work, we have developed a high-throughput method for TF binding specificity analysis and a novel computational tool, EEL, for the identification of mammalian enhancer elements, which allows genome-wide analysis of human distal enhancer elements. We have further applied these enabling technologies to the identification of target genes of developmental signaling pathways and to the analysis of regulation of two central growth-regulatory genes in mammals, c-*Myc* and N-*Myc*.

RESULTS

Development of High-Throughput Method for Determination of TF Binding Specificities

To allow rapid and accurate analysis of TF binding specificities, we developed a high-throughput method that directly determines the relative affinities of a TF to different DNA sequences. For this purpose, we fused the DNA binding domains of all GLI family TFs (GLI1, 2, and 3 and the *Drosophila* GLI ortholog Ci) to *Renilla reniformis* luciferase. We next expressed the GLI-*Renilla* fusion proteins, incubated them with biotinylated double-stranded oligonucleotide containing the sequence with the highest affinity to GLIs (consensus sequence), and measured the luciferase activity captured on a streptavidin plate. Competing this reaction with different unlabeled oligonucleotides (Figure 1A) allowed determination of the relative affinity of the GLI proteins to all possible single-base substitutions (Liu and Clarke, 2002) to the consensus sequence (Figures 1B and D). The affinities from the binding assay were consistent with results from an EMSA assay (Figure 1C).

We next made similar binding-affinity tables for Tcf4 and c-Ets1 (Figure 1D; see also Table S1 in the Supplemental Data available with this article online), which are regulated by Wnt and Ras/MAPK signaling pathways, respectively. Relative affinities obtained using our assay were consistent with the crystal structure of GLI1 bound to its consensus sequence (Pavletich and Pabo, 1993), published biologically relevant binding sites of GLI and Tcf/LEF TF families, and the known semiquantitative DNA binding preference of c-Ets1 (see Table S1).

Enhancer Element Locator

To identify conserved enhancer elements regulated by GLI and Tcf4, we developed a novel local (Smith and Waterman, 1981) alignment algorithm, enhancer element locator (EEL), that aligns the sequence (i.e., order) of TF binding sites found on two orthologous DNA sequences from two species (Figure 2A). The DNA sequence is not directly used in the alignment because only a fraction of all nucleotides in mammalian genomes code for high-affinity TF binding sites and because multiple DNA sequences can code for the same site. Our approach is thus conceptually similar to aligning peptide sequences instead of the corresponding DNA sequence and should similarly result in increased specificity and sensitivity.

The scoring scheme of EEL takes into account TF binding-site clustering, affinity, and conservation (Figure 2A; see Supplemental Data for details). A negative score is given for increased distance between adjacent conserved TF binding sites, and a positive score is given for conserved TF binding sites on the basis of their total relative affinities. Assessing true affinities of TF binding sites to DNA is difficult because binding of TFs to DNA is often cooperative, and sequences that by themselves bind only weakly to a particular TF can be occupied and biologically relevant in vivo due to increased affinity caused by secondary interactions between TFs or between the TF and other proteins. Because these secondary interactions cannot be modeled using current data, we included correction factors that describe the maximum loss of free energy caused by loss of secondary interactions due to an insertion of sequence between the adjacent TF binding sites. The correction is based on the energy required for twisting and/or compressing the two DNAs of unequal length into structures that would allow similar 3D positions for both pairs of TFs.

An important feature of EEL is that all TFs loaded to the program are treated equally, allowing simultaneous identification of a large number of conserved sites for different TFs. Subsequently, the alignments containing specific TF binding sites can be selected from this general analysis.

We first tested EEL by determining whether it could identify enhancer elements in the best characterized gene that is regulated by multiple enhancers, *Drosophila even-skipped* (*eve*) (Berman et al., 2002; Small et al., 1996). Analyzing *eve* genomic sequences of *D. melanogaster* and *D. pseudoobscura* with EEL using published binding-affinity matrices (Berman et al., 2002) for the five known TFs that regulate *eve*, we could identify all four known enhancer elements that control the segmental expression of *eve* in *Drosophila* embryos (Figures 2B and 2C).

Genome-wide Prediction of Mammalian Enhancer Elements

To adapt EEL for the more complex vertebrate genomes, we optimized the parameters of the penalty function by a greedy hill-climbing procedure (see Supplemental Data) using 107 binding profiles obtained by combining our own analyses

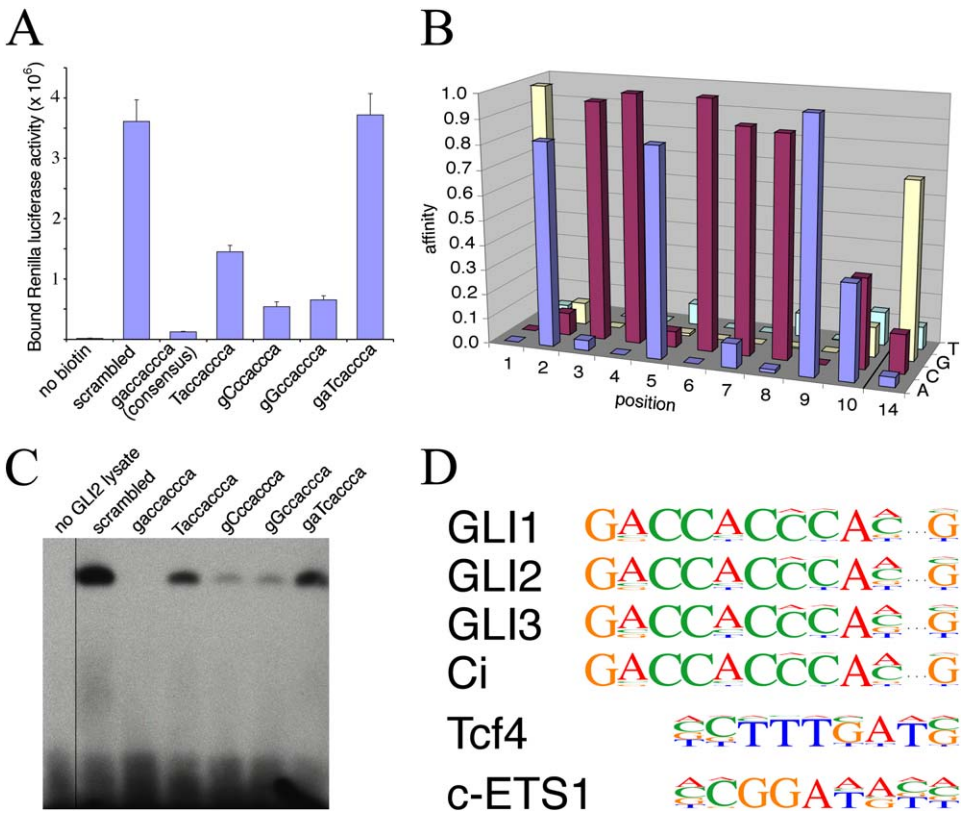

Figure 1. A High-Throughput Method for Measurement of TF Binding Specificity

(A) GLI2-zinc-finger *Renilla* luciferase fusion protein was incubated with competitor oligonucleotides indicated in the absence or presence of a biotinylated oligonucleotide containing the GLI consensus binding sequence. Bound GLI2 was measured as *Renilla* luciferase activity. Error bars represent one standard error (n = 8).

(B) Complete binding profile of GLI2. Bases 1–10 and 14, which contact the GLI protein, were analyzed (see Experimental Procedures).

(C) Verification of the results of the DNA binding assay by EMSA.

(D) Binding profiles of GLIs 1–3, Ci, Tcf4, and c-Ets1 described by differentially sized letters. The height of a letter at a particular position is directly proportional to the effect of that nucleotide substitution on the binding affinity (relative to consensus) of the indicated TF.

with high-quality TF binding profiles available in the literature and in the JASPAR2 database (see Tables S1 and S2 and Sandelin et al., 2004). Optimization resulted in relatively large penalties for differences in distance and angle, consistent with the initial hypothesis on the importance of the secondary interactions for TF binding.

We next tested EEL on a classic example of a distal enhancer in mammals, the −20 kb enhancer of MyoD (Goldhamer et al., 1995). The highest scoring *cis*-module resulting from the alignment of 50 kb mouse and human MyoD sequences was located to another gene 3′ of MyoD, the second was the −20 kb distal enhancer of MyoD, and the third was in the MyoD coding region (Figure 2D). These results indicate that EEL can also identify mammalian distal enhancers.

To predict enhancers genome-wide, we performed an EEL alignment of all 20,173 homologous human-mouse gene pairs (17,429 human genes with their 20,173 mouse orthologs from the ENSEMBL database). The aligned sequences included the genomic sequences from first to last exon and 100 kb of flanking sequence in both directions. The results were placed in a relational database containing information about the aligned regions, predicted enhancer modules, and conserved TF binding sites (Figure 3A). This database was subsequently used to determine the frequency of conserved binding sites for all of the 107 TFs used in the alignment (Figure 3B, top panel).

Identification of Activated TFs Based on Changes in Gene Expression

To test whether the genome-wide data could be used to determine which TFs are activated in an experiment based on gene-expression data, we also determined frequencies of all 107 TF sites in predicted enhancer elements of 13 genes whose expression is induced in the colon of mice after inactivation of the adenomatous polyposis coli (APC) tumor suppressor (Sansom et al., 2004). The second most overrepresented TF site in the flanking regions of these genes was Tcf4 (Figure 3B), which is known to be activated by the loss of

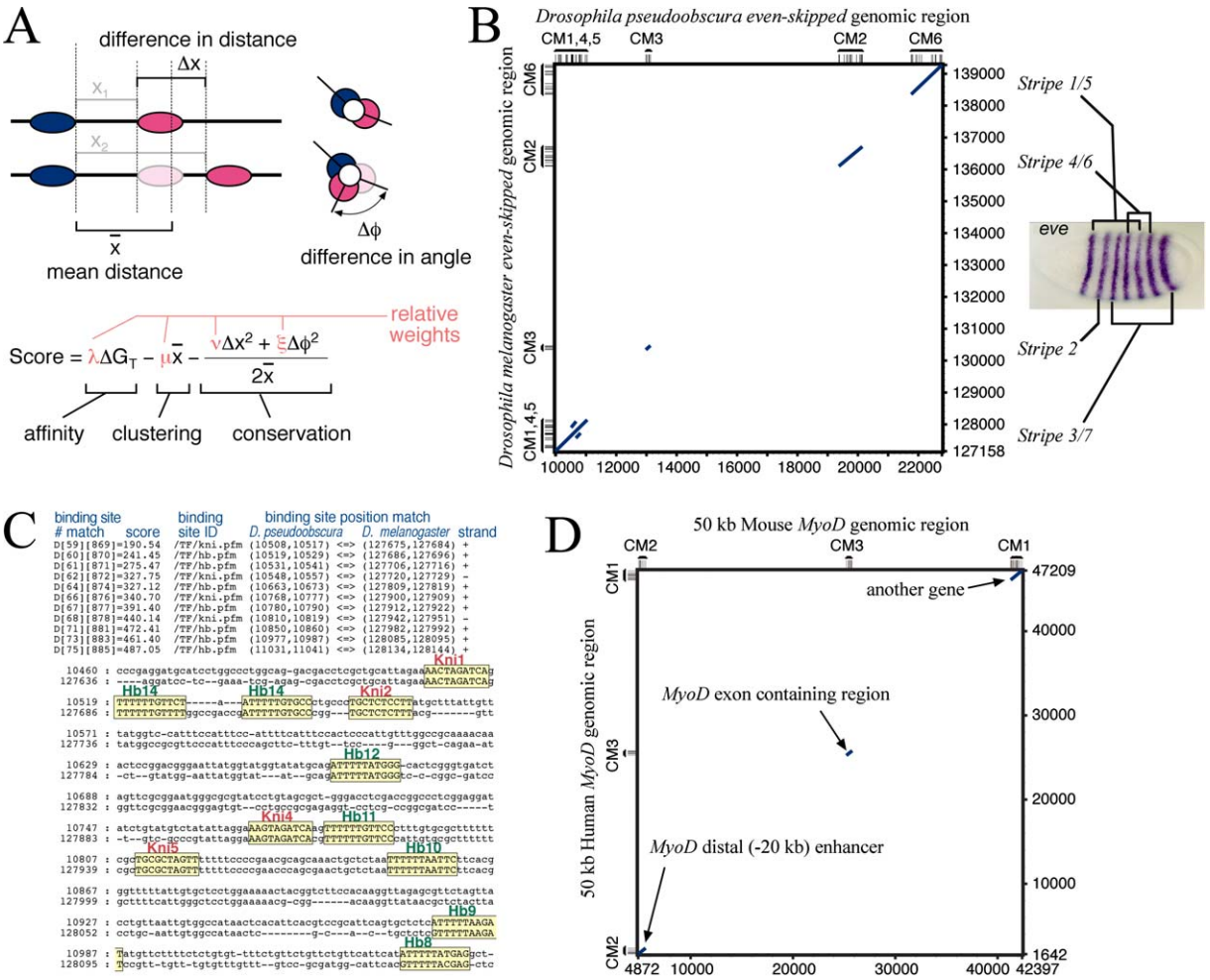

Figure 2. EEL, a Novel Local Alignment Tool that Aligns Two TF Binding-Site Sequences

(A) EEL scoring function. Top: schematic representation of two TFs (blue and red ovals) bound to DNA of unequal length from two different species. Side view (top left) indicates mean distance ($\bar{x}$) and difference in distance (Δx), and front view (top right) indicates difference in angle ($\Delta\phi$) of the two factors bound to DNA (open circle). Position weight matrix scores for TFs were used as a proxy for binding affinity in calculation of ΔG_T, the sum of TF affinities to sites in both species. Bottom: the score function. See Supplemental Data for details.

(B) EEL analysis (left) using the five known TFs that regulate *eve* (Hunchback, Caudal, Knirps, Bicoid, and Kruppel) identifies all four enhancers driving striped expression of *Drosophila eve* (right). Blue diagonal lines indicate aligned regions, and black lines on the x and y axes represent the conserved TF binding sites that constitute the *cis*-modules (CM). Number after the CM indicates its rank based on its EEL score.

(C) Text display of EEL alignment of part of the *eve* Stripe 3/7 enhancer (CM1 from [B]). *D. pseudoobscura* and *D. melanogaster* sequences are on top and bottom lines, respectively. EEL aligns the DNA sequences between the conserved TF sites for clarity; the DNA alignment does not contribute to the EEL score. Yellow boxes indicate conserved binding sites of Hunchback (Hb) or Knirps (Kni), which regulate this *cis*-module (Small et al., 1996).

(D) A distal −20 kb enhancer element in the mouse and human *MyoD* genes is identified by EEL analysis.

APC (Bienz and Clevers, 2000). With an even higher confidence value, a pair of Tcf4 sites were identified as the most overrepresented pair of the same binding sites in individual enhancer elements of the APC target genes (Figure 3B; p = 0.00083; 92% confidence after correction for multiple hypothesis testing). We also performed similar analyses for all possible TF site pairs, identifying Tcf4+Tcf4 as the second most overrepresented pair (Figure 3C; Table S2). These results validate the biological relevance of our TF binding-specificity assay and indicate that genome-wide EEL results can be used to identify activated TFs on the basis of expression-profiling data.

In Silico Identification of Hh/GLI Target Genes

To further validate EEL through unbiased genome-wide analysis of its predictions and to test the feasibility of identifying conserved target genes of developmental signaling pathways in silico, we performed pairwise genome-wide EEL alignments of human genes to orthologous rat, chicken, and pufferfish genes (Figure 4A). Similarly to in the human-to-

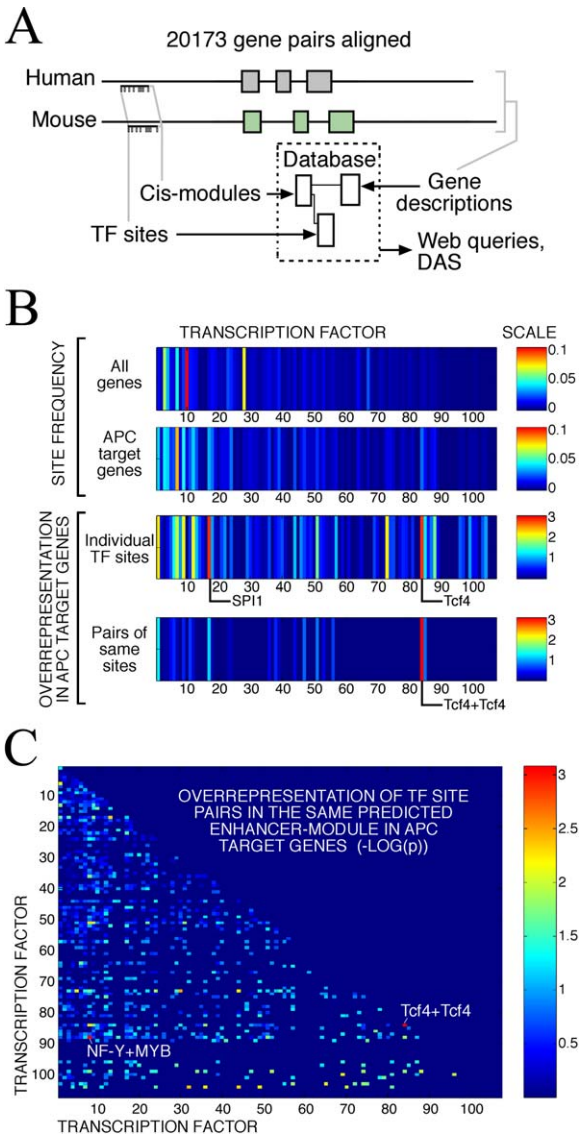

Figure 3. Genome-wide EEL Analysis of Enhancer Elements in Mammals

(A) Schematic description of the alignment procedure.

(B) Analysis of overrepresentation of TF binding sites in genes regulated by the APC tumor suppressor. The 107 vertical colored lines in the left panels represent the different TFs used in the analysis. The colors represent values on a color scale (corresponding right panels) indicating site frequency (top two panels) or overrepresentation of the TFs ($-\log_{10}(p)$) in APC target genes (bottom two panels).

(C) Overrepresentation of pairs of any two TF binding sites in the same predicted *cis*-module in the APC target genes. The two pairs of TFs having the lowest p values are also indicated. See Table S2 for identity of the 107 TFs.

mouse analysis, all the 107 TF sites were included in these alignments.

Since the presence of two binding sites for the same factor resulted in the highest confidence values in the analysis described above (Figure 3B), we selected elements that contained a minimum of two GLI binding sites of combined relative affinity score of 25 or more, were shorter than 2000 bp, and had an EEL alignment score higher than 500. As few clusters of multiple conserved GLI sites are found in the genome, this high an EEL score requires the presence of multiple conserved sites for other TFs in the predicted element.

Using mouse-to-human alignment alone, a total of 42 elements met these criteria. Two out of three ($p = 8.8 \times 10^{-5}$) in vivo-validated direct GLI targets (Table S3; Figure 4B, red typeface; see Supplemental Data for standards of evidence) contained such an element in their aligned regions. These were the two known marker genes for GLI activity, *GLI1* and *PTCH1* (Ingham and McMahon, 2001; Taipale and Beachy, 2001), which are induced by Hh ligands in all tissues examined. The predicted enhancer of *PTCH1* was the same one that was identified previously (Agren et al., 2004), containing one high-affinity and one lower-affinity GLI site, an arrangement which is potentially important in graded responses to Hh. Of the 42 elements, 7 were also conserved in the human-to-rat alignment (Figure 4B, blue and red diamonds). Only one element with two GLI sites was also conserved in chicken (Figure 4B, red diamond), and none was conserved in pufferfish.

To further validate the predictions, we analyzed the expression pattern of a subset of the predicted genes and enhancers. During early embryogenesis, Sonic hedgehog (Shh) expressed by the notochord and floor plate (see Figure 4C) is important for the patterning of the ventral neural tube and the sclerotome and epaxial myotome of the somites (Chiang et al., 1996; Wijgerde et al., 2002). Shh is also characteristically expressed in the endoderm (Figure 4C) and posterior margin of the developing limb buds and at later stages in whisker (E12.5) and hair (E14.5) follicles. Target genes of Shh are expressed in some cases in most or all responding cells (e.g., PTCH1; Figure 4D) but more commonly are restricted to particular Shh-responsive tissues at specific developmental stages (e.g., Tbx2 and FoxF1; Figures 4E and 4F), consistent with the ability of Hh proteins to induce diverse cellular responses during development (for review, see Ingham and McMahon, 2001).

We next analyzed the expression pattern of predicted Shh target genes that were located close to 16 conserved enhancer elements having high GLI affinity scores (see Table S3 for details). Ten genes were expressed at the analyzed stage in a relatively restricted pattern. Five expression patterns were consistent with regulation by Shh: Three corresponded to previously known Shh targets (PTCH1, Tbx2, and FoxF1) and two to genes whose regulation by Shh has not been reported. Of these, GPC3 was expressed in the sclerotome of the somites (Figure 4H) and SOX13 in the ventral neural tube (Figure 4G; see Table S9 for overview).

To test whether the predicted sequences functioned as enhancer elements, we assessed their ability to direct LacZ marker-gene expression to specific tissues of transgenic mouse embryos. Three of four highest scoring (see Table S8) predicted Shh-regulated enhancer elements analyzed directed LacZ expression into tissues that are specified

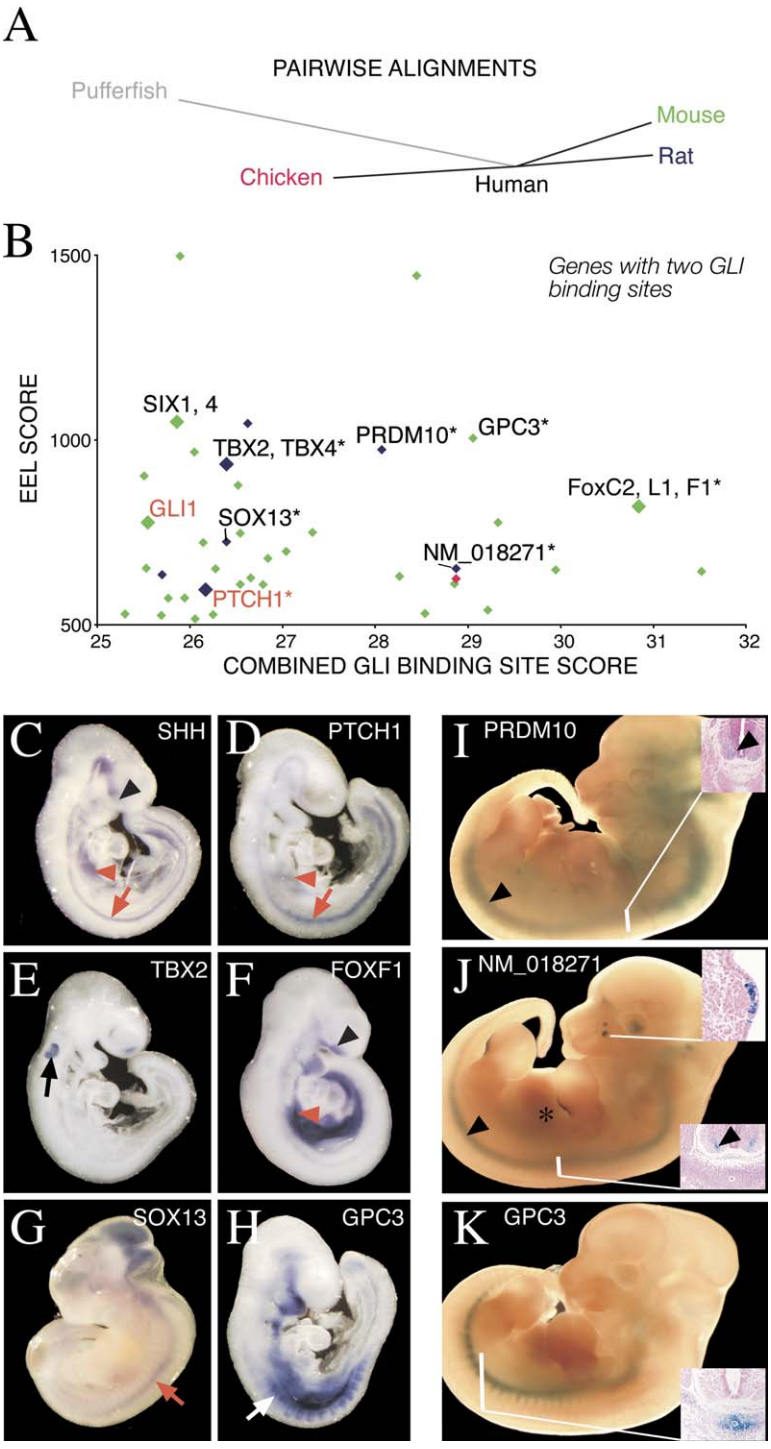

Figure 4. Identification of Hh Target Genes

(A) Description of the pairwise genome-wide alignments (star alignment) performed.

(B) Plot of predicted GLI-regulated enhancer elements. Coloring of diamonds indicates conservation in alignments of human to mouse (green), rat (blue), chicken (red), and pufferfish (gray). Large diamonds represent genes that have been reported to be induced by Hh; direct targets validated in vivo are in red typeface. Genes indicated with an asterisk are analyzed in (D)–(K).

(C–H) Expression pattern of Shh (C) and known (PTCH1 [D], TBX2 [E], and FOXF1 [F]) and predicted (SOX13 [G] and GPC3 [H]) Hh target genes. Black and red arrowheads indicate nasal process and gut, respectively, and arrows indicate ventral neural tube (red), ventral otic vesicle (black), and sclerotome (white). All embryos are analyzed by in situ hybridization at E9.5, except for SOX13 (E11.5).

(I–K) Analysis of predicted Hh-regulated enhancer elements in E12.5 mouse embryos. PRDM10 enhancer drives expression in ventral neural tube ([I], arrowhead). NM_018271 enhancer (J) directs LacZ expression into whisker follicles (inset), ventral neural tube (arrowhead), and posterior aspect of limbs (asterisk). GPC3 enhancer (K) directs LacZ expression into sclerotome-derived tissue of the vertebral cartilage primordia (inset) at E12.5. Sectioning plane is also indicated.

by Shh. Enhancer from PRDM10 drove expression in ventral neural tube (6 of 6 LacZ-positive embryos; Figure 4I); NM_018271 drove expression in ventral neural tube, whisker follicles, and posterior aspect of limb buds (5 of 6 embryos; Figure 4J); and enhancer from glypican3 (GPC3) specifically directed marker-gene expression into sclerotomally derived tissues (vertebral cartilage primordia, 6 of 6 embryos; Figure 4K; at E12.5, GPC3 is expressed in a similar pattern [Pellegrini et al., 1998]). Although all of the enhancer elements tested (Figures 4I–4K) had two GLI binding sites with similar affinity (Table S3), the specific tissues into which they directed expression were different, indicating that the other TF binding sites in the enhancer modules critically restrict expression to particular tissues.

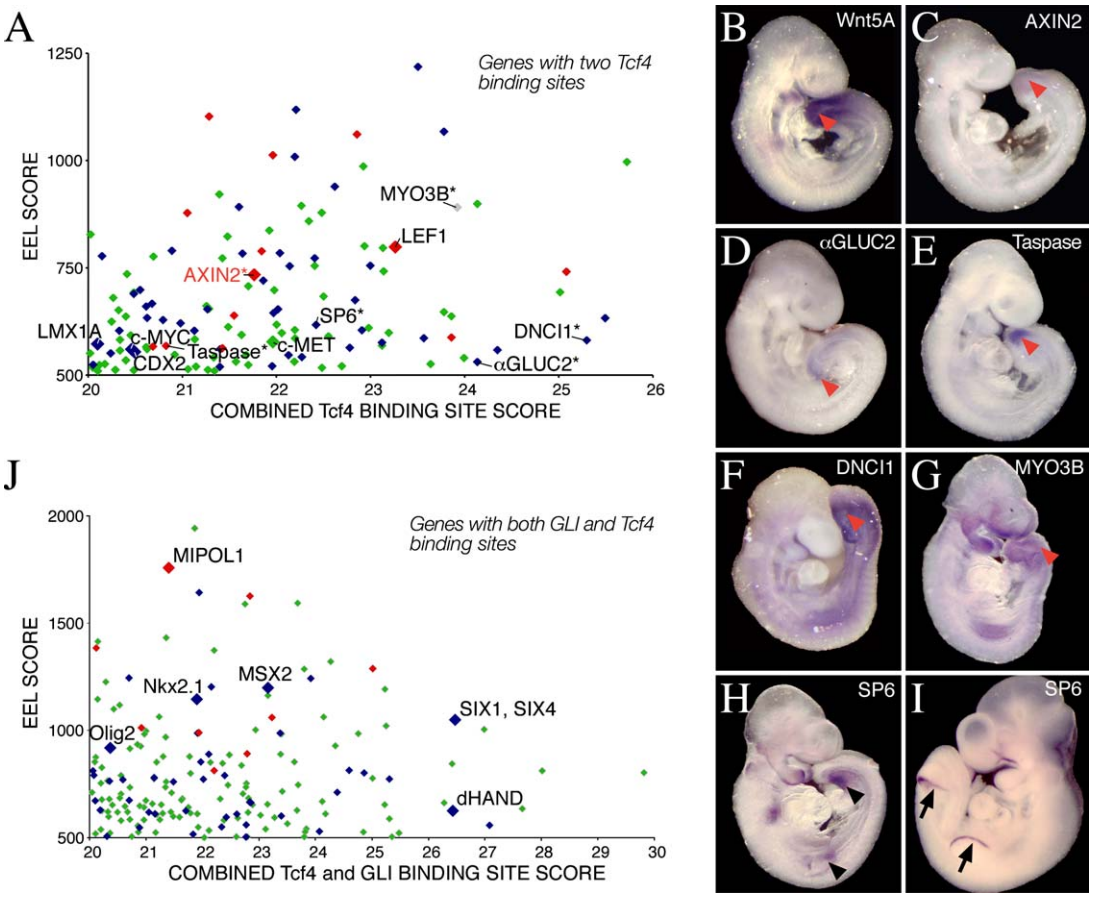

Figure 5. Identification of Wnt Target Genes

(A) Plot of predicted Tcf4-regulated enhancer elements. Coloring of diamonds indicates conservation in alignments of human to mouse (green), rat (blue), chicken (red), and pufferfish (gray). Large diamonds represent genes that have been reported to be induced by Wnt; direct targets validated in vivo are in red typeface. Expression of genes indicated with an asterisk is analyzed in (B)–(I).

(B–I) Expression pattern of Wnt5A (B) and one known (AXIN2 [light stain], [C]) and five predicted Tcf4 target genes. Tail bud (red arrowhead), limb buds (black arrowheads), and AER (black arrows) are also indicated. Embryos are E9.5 (B–H) and E10.5 (I).

(J) Plot of predicted enhancer elements containing both GLI and Tcf4 sites.

Identification of Wnt/Tcf4 Target Genes

Similar analysis of Tcf4-regulated genes based on the human-to-mouse alignment identified 132 predicted enhancer elements. One of these elements was in aligned regions of AXIN2, one of the three known direct Tcf/LEF target genes that have been validated by enhancer analysis in vivo in transgenic mice (Figure 5A, red typeface; Table S4). A total of six elements (4.5%) were located close to other reported Tcf4-inducible target genes, including *LEF-1*, *LMX1A*, c-*Met*, *CDX2*, and c-*Myc* (Figure 5A; Table S4). Elements located close to reported Tcf4 target genes were further enriched among the predictions if only elements conserved also in rat (8.5%; 5 of 59) or both rat and chick (14%; 2 of 14) were considered (Figure 5A).

We next assessed whether the predicted genes were expressed in a pattern consistent with Wnt regulation. At E9.5, Wnt3, Wnt5A (Figure 5B), Wnt5B, and many Wnt target genes (e.g., AXIN2; Figure 5C) are expressed in the tail bud, a structure whose formation depends on Wnt signals (Huelsken et al., 2000). Wnts are also required for the formation of the apical ectodermal ridge (AER) of the developing limbs (Barrow et al., 2003). Twelve predicted Wnt target genes analyzed that were located close to 25 conserved enhancer elements with high Tcf4 affinity scores (Table S4) were expressed at E9.5 in a specific pattern. Expression patterns of five genes that had previously not been characterized as Wnt targets were clearly consistent with Wnt regulation. Of these, four were expressed in the tail bud (Figures 5D–5G) and one in the AER (Figures 5H and 5I). Four additional genes, including a known Wnt target (LEF1), had somewhat more general expression patterns with markedly elevated expression in the tail (Figure S3).

It is interesting to note that several genes that are known to be induced by Shh and/or Wnt also contained enhancer elements having both a conserved GLI and Tcf4 sites, raising the possibility that these genes may be involved in integration

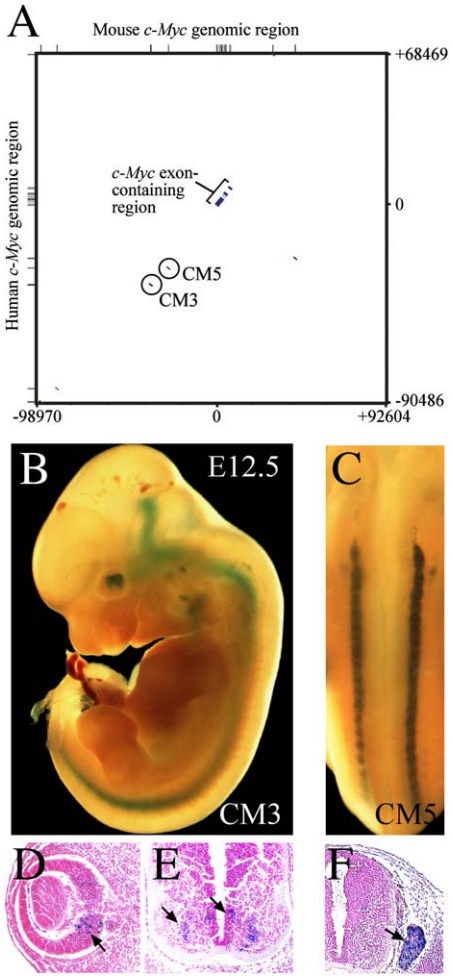

Figure 6. Analysis of Enhancer Elements in c-*Myc*

(A) EEL predicts two enhancer elements (CM3 and CM5) in human c-*Myc* that contain conserved Tcf4 binding sites (circled).

(B–F) Analysis of the predicted enhancers in E12.5 mouse embryos. c-Myc-CM3 enhancer directs LacZ expression to the ventral neural tube ([B] and [E], arrows) and to the eye ([D], arrow). c-Myc-CM5 drives expression in dorsal root ganglia ([C] and [F], arrow).

of Hh and Wnt signals during development (Figure 5J; Table S5).

Analysis of Regulation of c-*Myc* and N-*Myc*

Finally, we tested the practical utility of EEL in dissecting biological problems by applying it to the analysis of organ-specific growth control. For this purpose, we analyzed whether the expression of a central family of growth-regulatory genes, the *Myc* genes, is under the control of multiple tissue-specific enhancer elements.

We first analyzed the predicted Tcf4-regulated enhancer on the c-*Myc* gene. c-*Myc* is a known target of the Wnt pathway in colorectal cancer (He et al., 1998), and its expression in vivo appears to depend on distal elements that have not been identified (Lavenu et al., 1994). EEL predicted several conserved enhancer elements for the c-*Myc* locus, two of which (CM3 and CM5) contained conserved Tcf4 sites (Figure 6A). CM3 directed marker-gene expression into the ventral aspect of the neural tube and in the eye (4 of 5 LacZ$^+$ embryos; Figures 6B, 6D, and 6E). CM5, in turn, drove expression in dorsal root and trigeminal ganglia (7 of 8 embryos; Figures 6C and 6F; compare to Schmid et al., 1989).

The Hh-GLI pathway induces the expression of N-*Myc*, which encodes a protein that functions similarly to c-Myc. Induction of N-*Myc* is critical for Shh-induced cell proliferation of cerebellar granule neuron progenitors (CGNPs; Kenney et al., 2003; Oliver et al., 2003), and its expression depends on distal elements that have not been identified (Charron et al., 2002). Multiple predicted enhancer modules were identified in the N-*Myc* locus (Figure 7A), two of which (CM5 and CM7) contained GLI binding sites conserved in human, chimpanzee, and rat. At E12.5, the predicted enhancer located in the second intron (CM7) drove expression specifically in the maxillary arch derivatives, including mouth (Figure 7B), and in the developing tooth buds (6 of 7 embryos; Figure 7E). The pattern of expression in the tooth placode is localized to regions where Shh is specifically expressed and acts as a mitogen (Cobourne et al., 2001) to induce localized epithelial thickenings that invaginate to form the tooth bud. Also, the distal +65 kb enhancer (CM5) drove expression in a tissue-specific manner in the forebrain (Figures 7C and 7F; thalamus and roof of neopallial cortex) and in dorsal aspect of the neural tube (4 of 5 embryos; Figure 7G). Although LacZ is present also in postmitotic neurons, probably due to stability of the protein, the position of the LacZ expression along the dorsoventral axis of the neural tube is consistent with the known expression domain of N-*Myc* RNA (Kenney et al., 2003). Consistent with a role of this enhancer also in mediating growth responses to Shh, CM5 drove expression at postnatal day 3 (PN3), specifically in the CGNPs of the external granule cell layer of the cerebellum (4 of 6 LacZ-positive mice; Figures 7D and 7H). Two additional N-*Myc*-derived sequences tested that contained a conserved GLI site (or sites), one in the coding region of N-*Myc* and the other at +48.5 kb (GLI site not conserved in chimpanzee), did not drive expression in a tissue-specific manner at E12.5. These sequences either do not represent enhancers or function at a different developmental stage.

These results indicate that the expression of the *Myc* genes is controlled by multiple tissue-specific enhancer elements and further demonstrate the utility of EEL in identifying distal enhancers in mammals.

DISCUSSION

Determination of TF Binding Specificities

Information about TF binding-site specificity is often incomplete (i.e., only the site with maximal affinity is known) or biased by the prediction methods used (such as alignment of multiple potential binding sites). To resolve this problem, we developed a novel microwell-plate-based TF binding-specificity assay. The assay has broad utility, as it can be used for multiple classes of TFs, including zinc-finger (GLI), high-mobility-group (Tcf4), and ETS-domain (c-Ets1) DNA

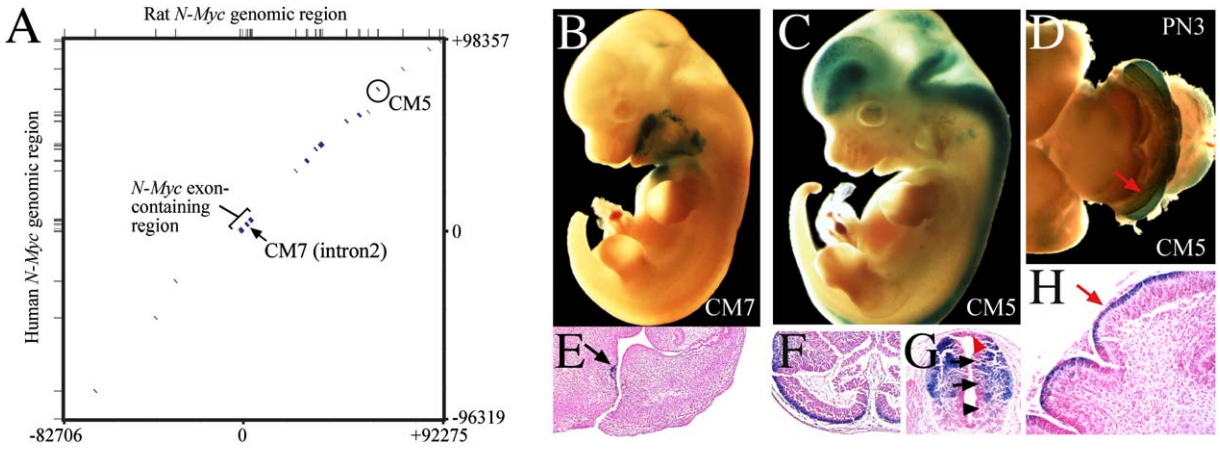

Figure 7. Analysis of Enhancer Elements in N-*Myc*

(A) EEL analysis identifies two GLI binding-site-containing predicted enhancer elements in human N-*Myc* (arrow and circle).

(B–H) At E12.5, N-Myc intronic enhancer (N-Myc-CM7) directs expression to the ventral side of the neck and in the mouth (B), specifically to the developing tooth buds ([E], arrow). N-Myc-CM5 enhancer directs expression to the forebrain ([C] and [F]) and to the dorsal aspect of the neural tube ([G], arrows), excluding ventral neural tube (black arrowhead) and roof plate (red arrowhead). At postnatal day 3 (PN3), N-Myc-CM5 drives expression into the external granule cell layer of the cerebellum ([D] and [H], red arrow).

binding proteins. Therefore, using high-throughput methods similar to those described here, it should be feasible in the near future to determine the binding specificities of the estimated 2000 DNA binding proteins (Tupler et al., 2001) in the human genome.

Because our method depends on prior knowledge of the site with maximal affinity but is capable of directly determining relative affinities, it complements recently described microarray-based high-throughput TF binding-specificity assays (Liu et al., 2005; Mukherjee et al., 2004) that can identify unknown consensus sequences but rely on indirect means to estimate affinity. In particular, the established high-throughput methods rely on computational tools to find DNA binding-site motifs from the sequences included on microarrays. This leads to a bias toward the sequences that are included. More importantly, using alignment-based methods, it is very difficult to determine affinities as opposed to rank of affinities. Because high-affinity sites can become saturated, the concentration of TF used and the threshold for inclusion of sequences into the alignment affects the ''stringency'' of the obtained TF binding-specificity matrix.

Prediction of Mammalian Enhancer Elements

We report here the genome-wide prediction of mammalian enhancer elements and assignment of all publicly available high-quality TF binding sites to these enhancers in human, mouse, rat, chick, and pufferfish. Whereas a large number of studies have identified enhancer elements in *Drosophila*, genome-scale methods have not previously been applied to mammalian enhancer prediction.

In *Drosophila*, enhancer elements can be efficiently identified by using algorithms based on clustering of a limited number of TF binding sites in one (Berman et al., 2002; Markstein et al., 2002; Rajewsky et al., 2002) or multiple (Sinha et al., 2004) species. However, clustering analysis is not as powerful in mammals, as mammalian regulatory elements typically have a very limited number of binding sites for any individual TF (see, for example, Agren et al., 2004; Lickert and Kemler, 2002). For example, clustering analysis clearly identifies the Hh-regulated element in the Ptc gene of *Drosophila* (Figure S1A), but the analysis of corresponding human sequences results in identification of five elements that have higher scores than that corresponding to the known enhancer (Figure S1C). A total of 15 elements are found using a signal-to-noise cutoff that in *Drosophila* would result in the inclusion of the first incorrect module. The decreased clustering of human GLI sites thus decreases signal intensity, and the larger size of the human genome increases noise, making enhancer prediction in mammals more difficult than in *Drosophila* (Figure S1E).

Presumably due to these difficulties, genome-scale studies in mammals have concentrated on promoters or 3′ untranslated regions (e.g., Suzuki et al., 2004; Xie et al., 2005) or multispecies conserved sequences (MCS). The MCS are sequences that are very well conserved in multiple vertebrate species. Although the MCS are much more conserved than the sequences of known enhancers and their role in other processes (e.g., as replication origins or modulators of chromatin structure) has not been carefully studied, some of these elements clearly regulate gene expression. However, none of the 1400 sequences conserved between human and pufferfish (Woolfe et al., 2005) overlaps with known Hh- or Wnt-responsive elements.

Because DNA-based alignment methods treat all nucleotides as equivalent, their use for identification of enhancer modules has important information theoretical limitations (see Supplemental Data), and they cannot reach the sensitivity and specificity of EEL, which only analyzes the information that is relevant for enhancer function (TF binding sites and their relative positions). The power and specificity of EEL is

demonstrated by the fact that even though only between 5% and 20% of all TF binding specificities are currently known and were included in our analysis, we were able to accurately predict mammalian enhancer elements on a genomic scale. Despite the large gap in our knowledge of TF DNA binding specificities, a method using EEL score also outperformed the use of DNA-alignment-based score in identification of known GLI target genes (Table S7). These results suggest that alignment based on TF binding sites will become an even more powerful method of analysis of regulatory elements when more information on TF binding specificities becomes available.

In Silico Identification of Hh and Wnt Target Genes

We also applied the genome-wide data on conserved enhancer modules and binding sites for the identification of target genes of developmental signaling pathways. Identification of target genes of pathways such as Hh and Wnt by expression profiling or chromatin immunoprecipitation is made difficult by the cell-type- and developmental-stage-specific cellular responses to these signals. These context-dependent responses do not affect our in silico analysis based on conserved TF binding sites in genomic sequences.

Seven out of ten predicted enhancer elements that we tested in vivo directed tissue-specific expression at the one developmental stage tested (E12.5 mouse embryo). In addition, a significant fraction (3 of 6) of well-established (see Tables S3 and S4) direct targets of the Hh and Wnt pathways were identified by EEL. Between 5% and 25% of the genes predicted to be regulated by the Hh or Wnt pathways had been previously reported as targets for these pathways. To further validate the predictions, we analyzed expression of the predicted genes in tissues that are specified or induced by Wnt (tail bud, AER) or Hh (ventral neural tube, sclerotome). In these tissues, the Hh or Wnt signals are by definition the most upstream modulators, and, in the genetic sense, all genes specifically expressed in the induced tissues are direct or indirect targets of Hh or Wnt. At the one developmental stage analyzed, 31% and 36% of the GLI- and Tcf4-regulated enhancers predicted, respectively, were located close to genes expressed in a pattern consistent with our predictions (Table S9). Prediction of GLI or Tcf4 target genes significantly enriched also novel genes whose expression patterns were consistent with regulation by Hh or Wnt, respectively ($p < 3.4 \times 10^{-3}$ and $< 3.3 \times 10^{-5}$ for novel and all genes, respectively; see Table S9). As there are other TF binding sites in addition to GLI and Tcf4 in the predicted enhancers, we cannot rule out that they also direct expression independently of Hh and/or Wnt. Thus, further validation of the predictions by targeted mutations of the TF binding sites in the mouse genome are needed to analyze the biological consequences and conclusively determine the directness of the individual predicted regulatory interactions.

Because of the cell-type-specific response to Hh and Wnt, it is not feasible to determine which genes are not regulated by these pathways, as this would require analysis of all cell types during all developmental stages. As not all target genes are expected to be regulated at the developmental stage analyzed here, the fraction of the predicted enhancers that are targets of Shh or Wnt is likely to be higher than the 31%–36% estimated above. Furthermore, because the EEL approach is general and simultaneously identifies conserved sites for a large number of TFs, similar analyses can be performed to identify target genes for any TF whose DNA binding specificity is known.

Conservation of Enhancer Elements

Increasing the number of species analyzed from two to three appeared to increase the quality of the EEL predictions (Tables S3 and S4). However, at the same time, the total number of predicted modules was decreased. This effect could be partially alleviated by requiring that only one GLI or Tcf4 site be conserved (Table S6). One factor explaining the decrease is the cumulative effect of incorrect or missing sequences or annotation in the present draft genomes. Thus, improvement of the quality of genomic sequences and development of multiple-alignment programs using several mammalian sequences are also expected to further improve the EEL method. However, inclusion of multiple species may not be beneficial in all cases, as evolutionary changes in the function or expression pattern of genes are also expected to contribute to the decreased conservation of enhancer elements. On the other hand, EEL makes it possible to study such regulatory evolution on a genomic scale. Furthermore, EEL can also be applied to prediction of regulatory single-nucleotide polymorphisms (SNPs), which are believed to be a major factor contributing to differences in gene expression in the human population.

Organ-Specific Growth Control

Our results suggest that the expression of the *Myc* genes is controlled by multiple independent tissue-specific enhancer modules. This allows growth to be regulated specifically in distinct tissues and organs, contributing to the understanding of the hitherto poorly understood mechanisms of organ-specific growth control.

We also find here that, instead of being regulated by a single element that is responsive to Shh in all tissues, N-*Myc* appears to be regulated by at least two distinct tissue-specific enhancer elements containing conserved GLI binding sites. These enhancers drive expression in the tooth bud and the external granule cell layer of the cerebellum, tissues where Shh is known to regulate growth and induce N-*Myc* expression (Cobourne et al., 2001; Kenney et al., 2003; Oliver et al., 2003). Thus, it is likely that tissue-specific TFs restrict the ability of Shh to induce N-*Myc* to particular tissues.

Specificity of Oncogenes

Tissue-specific regulation of the *Myc* genes is also relevant to the problem of tumor-type selectivity of oncogenes. Despite heterogeneity in genotype, all cancer cells share common phenotypic characteristics, such as unrestricted growth (Hanahan and Weinberg, 2000). Most if not all human malignancies express one or more of the *Myc* genes (Pelengaris et al., 2002a). This expression is induced by oncogenes acting upstream of *Myc* (He et al., 1998; Kenney et al., 2003;

Oliver et al., 2003), suggesting that the *Myc* genes serve as intermediaries through which multiple oncogenes regulate cell growth. Tissue specificity of enhancers in the *Myc* genes suggests that, in addition to a TF induced by an oncogene, an enhancer element requires tissue-specific cooperating factors to induce *Myc* transcription. A particular oncogenic mutation that results, for example, in the induction of the Hh pathway would be predicted to induce N-*Myc* expression and cause tumors only in tissues where the presence of these collaborating factors would allow activation of enhancer elements such as CM5 and CM7 (see Figures 7B and 7C). This would explain why mutations activating the Hh pathway are only observed in some tumor types and could thus provide a general mechanism explaining the tumor-type selectivity of oncogenes. In addition, because continued expression of *Myc* is required for tumorigenesis (Pelengaris et al., 2002b), these collaborating factors will also represent potential targets for chemotherapeutic drugs.

EXPERIMENTAL PROCEDURES

Constructs

Coding regions of GLI1–3 zinc-finger domains, Tcf4 lacking 30 NH_2-terminal amino acids, and full-length c-Ets1 were amplified by PCR using *Pfu* polymerase (Stratagene) and cloned into pGEN expression vector (Taipale et al., 2000) as N-terminal fusions to *Renilla* luciferase.

For generation of the *lacZ* reporter constructs, 1–2 kb genomic sequences carrying the predicted enhancer element were amplified by PCR and cloned into pTKPD (Goldhamer et al., 1995), which contains a TK minimal promoter followed by an *E. coli lacZ* gene with SV40 T nuclear localization signal. The genomic sequences included 200–250 bp flanks (see Table S8) that were not contained in the EEL alignments. The flanks were included because not all TF binding specificities are known and, consequently, the EEL alignment may start too late or terminate prematurely. All constructs were sequence verified.

Cell Culture and Transfections

Drosophila S2 cells were cultured in *Drosophila*-SFM (Invitrogen) with 10% fetal bovine serum (FBS) and antibiotics. Human 293T cells were cultured in RPMI medium supplemented with 10% FBS. *Drosophila* proteins were expressed in S2 cells transiently transfected with Effectene (Qiagen) according to manufacturer's instructions. Mammalian proteins were expressed in 293T cells transiently transfected using FuGENE 6 (Roche) essentially as described (Taipale et al., 2000). Cell extracts were collected 48 hr after transfection.

TF Binding Assay and EMSA Analysis

Binding was performed in 100 μl of binding buffer (140 mM KCl, 5 mM NaCl, 1 mM K_2HPO_4, 2 mM $MgSO_4$, 20 mM HEPES [pH 7.05], 100 μM EGTA, 1 μM $ZnSO_4$) supplemented with 0.2% TX-100, 1% milk powder, and 5 μg/ml poly(dI-dC) (Amersham). One picomole of biotinylated consensus double-stranded DNA oligonucleotide was competed with non-biotinylated competitor DNA in 30-fold molar excess. TF-*Renilla* luciferase fusion protein lysate containing 2.5×10^6 relative light units was added into the DNA mixture and incubated for 2 hr at RT. Subsequently, the mixture was added onto streptavidin-coated plates (ABgene), incubated for 2 hr at RT, and washed with binding buffer, and the amount of TF bound to the plate was measured using a luminometer (BMG FluoStar) and the *Renilla* luciferase assay (Promega). Relative affinity ($K_{dsample}/K_{dconsensus}$) was calculated from the light units obtained using the following equation derived from the law of mass action: $[(L_{scrambled}/L_{consensus}) - (L_{sample}/L_{consensus})]/[(L_{sample}/L_{consensus}) - 1] \times [(L_{scrambled}/L_{consensus}) - 1]$. Oligonucleotides used are described in Supplemental Data.

EMSA was performed according to manufacturer's instructions (Pierce LightShift Kit). Briefly, the TF-*Renilla* luciferase fusion protein lysate was incubated for 1 hr with biotinylated DNA probe and nonbiotinylated competitor DNA oligonucleotide (50-fold molar excess) in binding buffer supplemented with 0.2% TX-100, 1% milk powder, and 25 ng/μl poly(dI-dC). The resulting complexes were resolved in a 5% nondenaturing PAGE-gel, transferred onto membrane, and detected using streptavidin-HRP conjugate and a chemiluminescent substrate.

In Silico Methods

Computational methods are described in the Supplemental Data. The open source EEL program is available under GNU general public license at http://www.cs.helsinki.fi/u/kpalin/EEL/.

In Situ Hybridization and Transgenic Analyses

Whole-mount in situ hybridization was performed essentially as described (Henrique et al., 1995). Probes were generated using PCR (primers described in Tables S3 and S4). For enhancer analysis, enhancer-module-minimal promoter *lacZ* constructs were liberated from vector sequences, and TG embryos were produced by pronuclear injection of FVB/N one-cell-stage embryos. LacZ staining was performed essentially as described (Nagy et al., 2003). At least four LacZ-positive F_0 embryos were analyzed for each construct. All constructs resulted in LacZ expression in 50% or more of the TG-positive embryos (genotyping PCR primers AAGCGGTGAAGTGCCTCTGG and GGGGAGCGTCACACTGAGGT). To rule out ectopic expression due to differences in TG integration sites, only consistent expression patterns are indicated (75% or more of LacZ+ embryos expressed LacZ in these tissues). Embryos were photographed under dark-field illumination.

The genes analyzed in the in situ hybridization and TG validation (Figure 4) experiments were picked in a systematic fashion (see Tables S3, S4, and S8) from a group of genes located close to predicted enhancers with high TCF or GLI affinity scores conserved in human and mouse or human, mouse, and rat.

Supplemental Data

Supplemental Data include three figures, Supplemental Experimental Procedures, Supplemental References, and nine tables and can be found with this article online at http://www.cell.com/cgi/content/full/124/1/47/DC1/.

ACKNOWLEDGMENTS

We thank P.A. Beachy, K. Alitalo, and H. Clevers for constructs; S. Small for the picture of *eve* expression; M. Berg and B. Ranjan for help in coding; and J. Saharinen of Biomedicum Bioinformatics Unit for computing infrastructure. We also thank K. Salonen for production of TG embryos and P. Salven, P. Ojala, M. Bonke, M. Björklund, M. Taipale, and I. Thesleff for critical review of the manuscript. This work was supported by the Center of Excellence in Translational Genome-Scale Biology of the Academy of Finland, the BioSapiens and Regulatory Genomics projects of the EU, Biocentrum Helsinki, University of Helsinki, the Magnus Ehrnrooth and Sigrid Juselius foundations, and Finnish Cancer Research Organizations.

Received: June 14, 2005
Revised: September 21, 2005
Accepted: October 21, 2005
Published: January 12, 2006

REFERENCES

Agren, M., Kogerman, P., Kleman, M.I., Wessling, M., and Toftgard, R. (2004). Expression of the PTCH1 tumor suppressor gene is regulated by alternative promoters and a single functional Gli-binding site. Gene *330*, 101–114.

Barrow, J.R., Thomas, K.R., Boussadia-Zahui, O., Moore, R., Kemler, R., Capecchi, M.R., and McMahon, A.P. (2003). Ectodermal Wnt3/beta-catenin signaling is required for the establishment and maintenance of the apical ectodermal ridge. Genes Dev. *17*, 394–409.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA *99*, 757–762.

Bienz, M., and Clevers, H. (2000). Linking colorectal cancer to Wnt signaling. Cell *103*, 311–320.

Charron, J., Gagnon, J.F., and Cadrin-Girard, J.F. (2002). Identification of N-myc regulatory regions involved in embryonic expression. Pediatr. Res. *51*, 48–56.

Chiang, C., Litingtung, Y., Lee, E., Young, K.E., Corden, J.L., Westphal, H., and Beachy, P.A. (1996). Cyclopia and defective axial patterning in mice lacking *Sonic hedgehog* gene function. Nature *383*, 407–413.

Cobourne, M.T., Hardcastle, Z., and Sharpe, P.T. (2001). Sonic hedgehog regulates epithelial proliferation and cell survival in the developing tooth germ. J. Dent. Res. *80*, 1974–1979.

Conlon, I., and Raff, M. (1999). Size control in animal development. Cell *96*, 235–244.

Fried, M., and Crothers, D.M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res. *9*, 6505–6525.

Goldhamer, D.J., Brunk, B.P., Faerman, A., King, A., Shani, M., and Emerson, C.P., Jr. (1995). Embryonic activation of the myoD gene is regulated by a highly conserved distal control element. Development *121*, 637–649.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

He, T.C., Sparks, A.B., Rago, C., Hermeking, H., Zawel, L., da Costa, L.T., Morin, P.J., Vogelstein, B., and Kinzler, K.W. (1998). Identification of c-MYC as a target of the APC pathway. Science *281*, 1509–1512.

Henrique, D., Adam, J., Myat, A., Chitnis, A., Lewis, J., and Ish-Horowicz, D. (1995). Expression of a Delta homologue in prospective neurons in the chick. Nature *375*, 787–790.

Huelsken, J., Vogel, R., Brinkmann, V., Erdmann, B., Birchmeier, C., and Birchmeier, W. (2000). Requirement for beta-catenin in anterior-posterior axis formation in mice. J. Cell Biol. *148*, 567–578.

Ingham, P.W., and McMahon, A.P. (2001). Hedgehog signaling in animal development: paradigms and principles. Genes Dev. *15*, 3059–3087.

Kenney, A.M., Cole, M.D., and Rowitch, D.H. (2003). Nmyc upregulation by sonic hedgehog signaling promotes proliferation in developing cerebellar granule neuron precursors. Development *130*, 15–28.

Lavenu, A., Pournin, S., Babinet, C., and Morello, D. (1994). The cis-acting elements known to regulate c-myc expression ex vivo are not sufficient for correct transcription in vivo. Oncogene *9*, 527–536.

Lickert, H., and Kemler, R. (2002). Functional analysis of cis-regulatory elements controlling initiation and maintenance of early Cdx1 gene expression in the mouse. Dev. Dyn. *225*, 216–220.

Liu, X., and Clarke, N.D. (2002). Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. J. Mol. Biol. *323*, 1–8.

Liu, X., Noll, D.M., Lieb, J.D., and Clarke, N.D. (2005). DIP-chip: rapid and accurate determination of DNA-binding specificity. Genome Res. *15*, 421–427.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc. Natl. Acad. Sci. USA *99*, 763–768.

Michelson, A.M. (2002). Deciphering genetic regulatory codes: a challenge for functional genomics. Proc. Natl. Acad. Sci. USA *99*, 546–548.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet. *36*, 1331–1339.

Nagy, A., Gertsenstein, M., Vinterstein, K., and Behringer, R. (2003). Manipulating the Mouse Embryo: A Laboratory Manual, Third Edition (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Oliver, T.G., Grasfeder, L.L., Carroll, A.L., Kaiser, C., Gillingham, C.L., Lin, S.M., Wickramasinghe, R., Scott, M.P., and Wechsler-Reya, R.J. (2003). Transcriptional profiling of the Sonic hedgehog response: a critical role for N-myc in proliferation of neuronal precursors. Proc. Natl. Acad. Sci. USA *100*, 7331–7336.

Pavletich, N.P., and Pabo, C.O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science *261*, 1701–1707.

Pelengaris, S., Khan, M., and Evan, G. (2002a). c-MYC: more than just a matter of life and death. Nat. Rev. Cancer *2*, 764–776.

Pelengaris, S., Khan, M., and Evan, G.I. (2002b). Suppression of Myc-induced apoptosis in beta cells exposes multiple oncogenic properties of Myc and triggers carcinogenic progression. Cell *109*, 321–334.

Pellegrini, M., Pilia, G., Pantano, S., Lucchini, F., Uda, M., Fumi, M., Cao, A., Schlessinger, D., and Forabosco, A. (1998). Gpc3 expression correlates with the phenotype of the Simpson-Golabi-Behmel syndrome. Dev. Dyn. *213*, 431–439.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics *3*, 30.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat. Biotechnol. *20*, 831–835.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32*, D91–D94.

Sansom, O.J., Reed, K.R., Hayes, A.J., Ireland, H., Brinkmann, H., Newton, I.P., Batlle, E., Simon-Assmann, P., Clevers, H., Nathke, I.S., et al. (2004). Loss of Apc in vivo immediately perturbs Wnt signaling, differentiation, and migration. Genes Dev. *18*, 1385–1390.

Schmid, P., Schulz, W.A., and Hameister, H. (1989). Dynamic expression pattern of the myc protooncogene in midgestation mouse embryos. Science *243*, 226–229.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics *5*, 129.

Small, S., Blair, A., and Levine, M. (1996). Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. Dev. Biol. *175*, 314–324.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.

Spitz, F., Gonzalez, F., and Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell *113*, 405–417.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K., and Sugano, S. (2004). Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res. *14*, 1711–1718.

Taipale, J., and Beachy, P.A. (2001). The Hedgehog and Wnt signalling pathways in cancer. Nature *411*, 349–354.

Taipale, J., Chen, J.K., Cooper, M.K., Wang, B., Mann, R.K., Milenkovic, L., Scott, M.P., and Beachy, P.A. (2000). Effects of oncogenic mutations

in Smoothened and Patched can be reversed by cyclopamine. Nature *406*, 1005–1009.

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature *409*, 832–833.

Wijgerde, M., McMahon, J.A., Rule, M., and McMahon, A.P. (2002). A direct requirement for Hedgehog signaling for normal specification of all ventral progenitor domains in the presumptive mammalian spinal cord. Genes Dev. *16*, 2849–2864.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. *3*, e7. Published online November 11, 2004. 10.1371/journal.pbio.0030007.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature *434*, 338–345.