Supplemental Data

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

Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Esko Ukkonen, and Jussi Taipale

Captions for Supplemental Tables

Table S1. Measured Binding Profiles (Data for Figure 1D)

Table S2. Frequencies of Occurrence of the 107 TF Binding Sites Used in the Genome-wide Alignment (data for Figure 3B)

Table S3. EEL: Predicted Enhancer Modules with Two or More GLI Sites in a Single Predicted Module

Affinity score is from human to mouse alignment (affinity of the weaker species is shown).

Table S4. Predicted Enhancer Modules with Two or More TCF4 Sites in a Single Predicted Module

Affinity score is from human to mouse alignment (affinity in the weaker species is shown).

Table S5. Predicted Enhancer Modules with Both GLI and TCF4 Sites (at Least One Site Each in Single Predicted Module)

Affinity score is from human to mouse alignment (affinity of the weaker species is shown).

Table S6. Predicted Enhancer Modules with at Least One Conserved GLI Site in Mouse and *Tetraodon nigroviridis*

Table S7. Enhancer Modules Predicted Using DNA Alignment with Two or More Conserved GLI Sites in a Single Module

Compare to Table S2 (predictions made using EEL).

Table S8. Validation of Enhancer Elements with Predicted GLI Binding Sites

Predicted enhancer modules with two or more GLI sites in a single predicted module. Affinity is from human to mouse alignment (affinity of the weaker species is shown). Predicted enhancer elements in the analysis of *Myc* genes.

Table S9. Overview and Statistical Analysis of In Situ Hybridization Results for Predicted HH/GLI Targets and Predicted WNT/TCF4 Targets

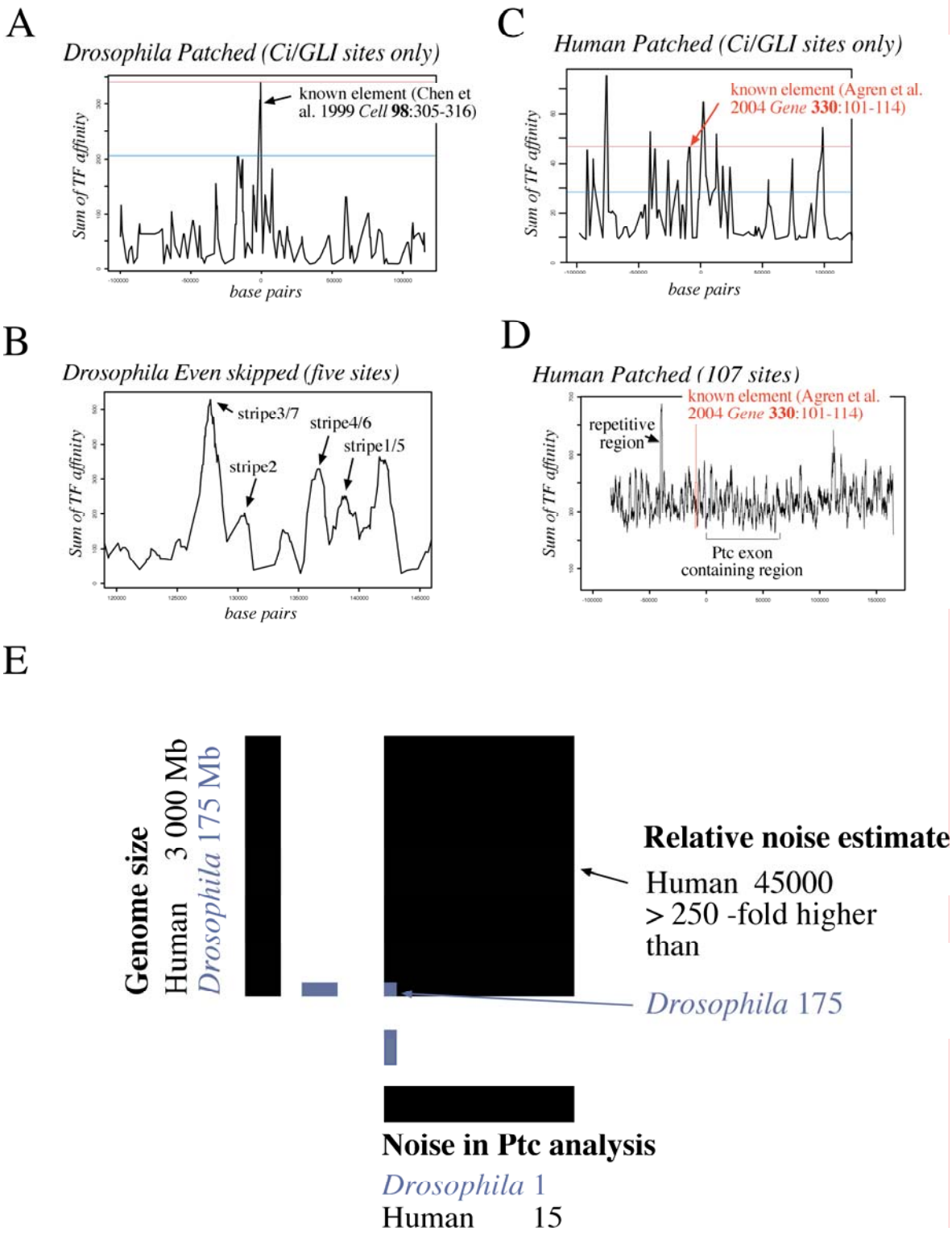

Figure S1. Comparison of Prediction of Enhancer Elements by TF Binding-Site Clustering Method in *Drosophila* and Human

(A) TF clustering analysis identifies the element in *Drosophila* Ptc locus that is regulates Ptc in response to Hh. Graph indicates the sum of relative affinities of Ci/GLI binding sites within a moving window. Red and blue horizontal lines indicate height of the peaks corresponding to the known enhancer and first background cluster, respectively.

(B) Similar analysis using five TF sites (Kruppel, Bicoid, Hunchback, Knirps and Caudal) detects peaks corresponding to known enhancers of the *Drosophila* even-skipped gene.

(C) Analysis similar to that in (A) fails to identify the element that is responsible for GLI regulation of mammalian Ptc. Red horizontal line indicates the height of the peak corresponding to the known enhancer. Note that five peaks are higher than the peak corresponding to the known enhancer, and that a total of 15 peaks are higher than the blue line placed at the relative position of first background peak in the *Drosophila* Ptc analysis.

(D) Inclusion of 107 TF sites to the human Ptc analysis results in complete loss of peaks, except at a repetitive region indicated.

(E) Order of magnitude estimate of noise in genome-wide analysis of enhancer elements in *Drosophila* and human. Ptc represents the only currently known gene that is regulated by either Wnt or Hh in both *Drosophila* and humans and whose regulatory element is known in both species. Note that although human has 2-3 times more genes than *Drosophila*, the increased noise due to decreased clustering of TF sites (x-axis) and increased genome size (y-axis) makes enhancer prediction in mammals more difficult than in *Drosophila*. Although this analysis is based on a case study, Ptc represents the only case of a known target gene of either Hh or Wnt that is conserved between *Drosophila* and mammals, and whose regulatory elements are known in both organisms.

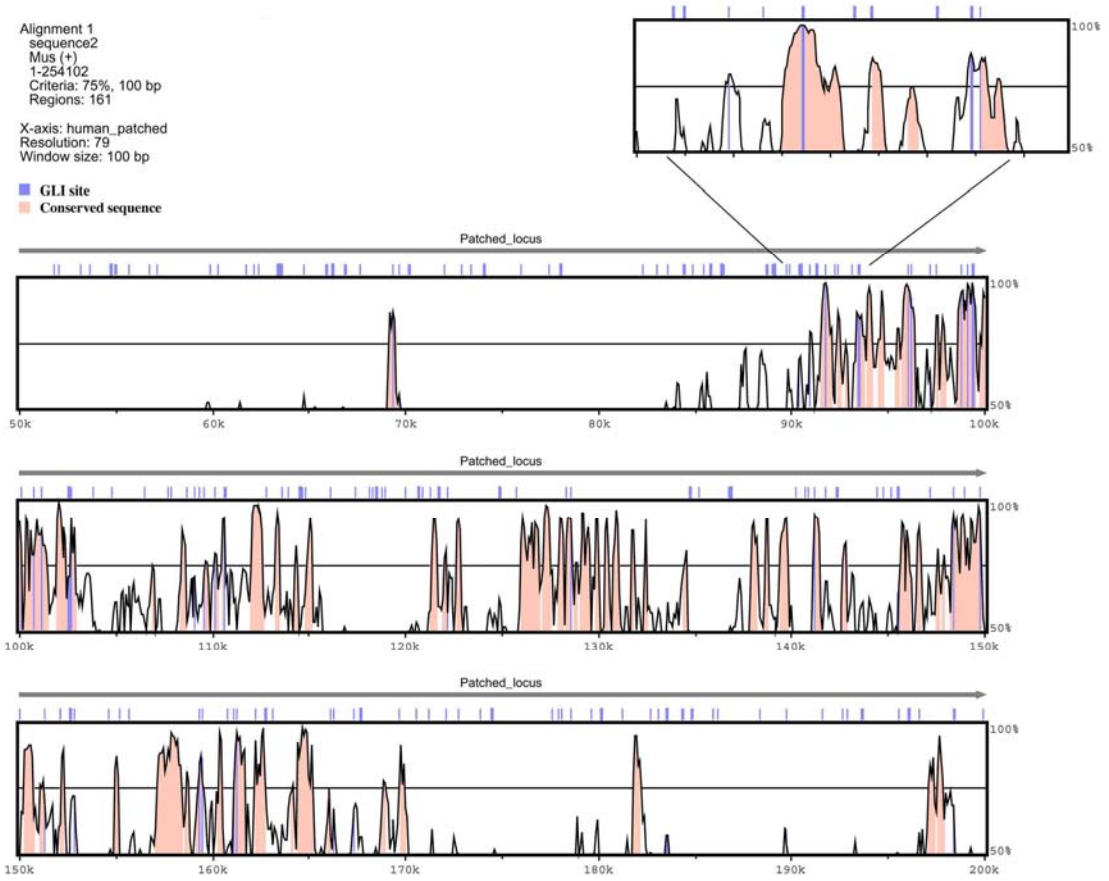

Figure S2. Identification of Mammalian Enhancers Using DNA Alignment and Mapping of TF Binding Sites

Ci/GLI binding sites (blue vertical lines) are mapped onto a global alignment (using the VISTA tool at http://genome.lbl.gov/vista/index.shtml) of 150 kb segment of human and mouse Ptch loci. Highly conserved regions are indicated by red color. Inset (top right) shows a magnification of the region containing the known element (wide conserved peak in the center with two GLI sites).

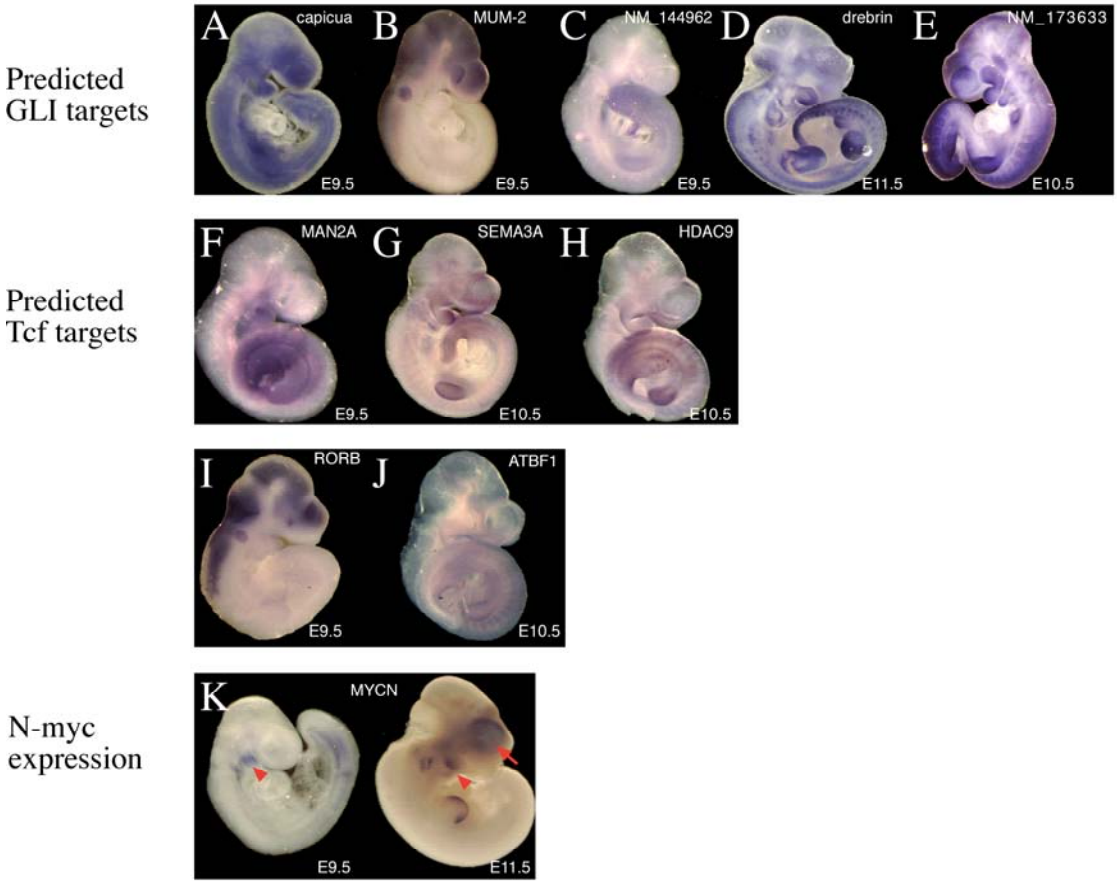

Figure S3. Expression Patterns of Predicted Hh/GLI and Wnt/Tcf Target Genes and *N-myc* Analyzed by Whole-Mount In Situ Hybridization

(A-E) expression pattern of predicted Hh targets not scored as consistent with Hh regulation. The obtained expression pattern of capicua (A) is example of a pattern scored as "B" (blue, general or non-specific expression) in the analyses described in Tables S3, S4 and S9. Expression patterns of indicated GLI target genes which displayed reasonably specific expression pattern but were not scored consistent with Hh regulation are shown in B-E. Expression pattern of NM_144962 (C) was scored as a weak Wnt-pattern. Consistently, in addition to two conserved GLI sites, this locus also contains a conserved Tcf4 site (see Table S5). Expression of one of the genes scored as specific (APLP2) was weakly detected in the otic vesicle, the pattern is not shown.

(F-H) Expression patterns of the indicated predicted Wnt/Tcf target genes that were scored as weak Wnt-patterns ("+") due to enhanced expression in the tail.

(I and J) Expression pattern of RORB (I) and ATBF1 (J), predicted Wnt/Tcf target genes which were not scored as Wnt-patterns. Also NM_006360 was classified to this category (not shown, expression detected in the head and in the branchial arches). Because responses to Hh and Wnt ligands are highly dependent on cell type, relatively few direct target genes for these pathways were known. However, because of this cell-type specific response, it is not

practically possible to determine which genes are not regulated by Hh, as this would require analysis of all cell types during all developmental stages.

(K) *N-myc* is expressed in mouse embryos at E9.5 (left) in branchial arches (arrowhead) and caudally (excluding the tailbud). At E11.5 (right) expression is observed in the apical ectodermal ridge, branchial arches (arrowhead), forebrain (arrow) and in the tail, excluding the posterior-most part (not visible).

Supplemental Methods for "Genome-wide prediction of mammalian enhancer elements based on high-throughput analysis of transcription factor binding affinity"

Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Jussi Taipale, Esko Ukkonen

EEL Algorithm

Alignment approach to enhancer element prediction

We want to predict putative enhancer elements (more generally, cis–regulatory modules) in DNA. Roughly, such an element is a segment of DNA that contains several predicted binding sites of transcription factors (TFs), clustered relatively closely together. Among such segments, the most plausible are the ones whose *pattern* of binding sites of TFs is conserved. Therefore we look at the DNA in the surroundings of orthologous genes to find conserved patterns of binding sites.

Given two such orthologous DNA sequences, our algorithm first finds all binding sites of the TFs. Given the positional weight matrices representing the binding affinity distribution of each TF, the putative binding sites with associated affinity scores can be computed by standard methods [Stormo & Fields, 1998, Stormo, 2000]. The next step is to find conserved clusters of sites. We do this by using alignment techniques. Here the main issue is to devise a suitable scoring function to measure the quality

of the alignment, i.e., the degree of conservation. Once we have a scoring function, the rest is just to find highly-scoring local alignments. The binding sites in the local alignments constitute the conserved patterns. The putative enhancer elements proposed by the algorithm consist of the DNA segments that correspond to these local alignments.

The idea behind the aligment scoring model is that adjacent TFs are very likely to cooperate with each other. This cooperation takes the form of interactions of TFs with each other or with other proteins via a surface that is formed by multiple TFs. Any increase in length of DNA between two TFs that bind near to each other is expected to change this interaction surface, and thus result in changes in avidity of other proteins of the transcription machinery. Thus, the surface created by two TFs that are close to each other is better conserved if there is no change in their distance from each other. The farther the TFs are apart from each other, the easier it is for the DNA to bend to compensate for the differences in TF position, and accommodate the secondary protein-protein interactions. Therefore the penalty becomes smaller with increase in average distance. The correction factor is relative, and penalizes for lack of conservation of distance and angle, and therefore is not affected the presence of additional structures such as nucleosomes. Bending of DNA containing enhancer-elements to the promoter can occur over very long distances ($>>$ 10 kb), and therefore our alignment score is not affected in any way by the distance of TFs from transcription start site(s).

Finding putative binding sites

Let G and G' denote the two orthologous DNA sequences, and let $\{M_1, \ldots, M_k\}$ be the positional weight matrices for the TFs available. The positional weight matrices M_i are all matched to the DNA sequences G and G', and the affinity of the transcription factor to each subsequence g of the DNA sequences is estimated from the positional weight matrix [Stormo & Fields, 1998, Stormo, 2000] by

$$W(g) = \log_2 \frac{\mathrm{P}_{M_i}(g)}{\mathrm{P}_0(g)} \qquad (1)$$

where $\mathrm{P}_{M_i}(g)$ is the likelihood of sequence g being generated by nucleotide distribution represented by M_i, and $\mathrm{P}_0(g)$ is the likelihood of g being generated by the background nucleotide distribution. We assumed independent, identically and uniformly distributed background genome.

In this way we obtain two sequences of putative transcription factor binding sites, sequence $S = (s_1, \ldots, s_L)$ for sites from G and $S' = (s'_1, \ldots, s'_{L'})$ for sites from G'. Each element s_j of sequence S represents a binding site of G as a quadruplet $s_j = (f_j, p_j, q_j, W_j)$ where f_j names the transcription factor, p_j is the start and q_j is the end position of the site on the DNA G, and W_j is the binding affinity of that site estimated by (1) using the weight matrix of the TF f_j [Stormo & Fields, 1998]. The elements $s'_j = (f'_j, p'_j, q'_j, W'_j)$ of sequence S' have analogous structure. The sites are given in sequences S and S' in the increasing order of their start positions. As the number of putative sites may be very high, we only take the sites whose affinity W_j is above some given threshold. We used threshold value 9.

Next we proceed to locally aligning the two sequences of putative binding sites. It should be noted that in this phase the underlying DNA sequences are no longer used and all information is transported via the sequences S and S'.

Conservation model and alignment scoring

The scoring function for evaluating an alignment of two sequences of binding sites has two components: some bonus is given for each aligned pair of sites and some penalty is given for distances between two adjacent aligned pairs. We only allow aligning the sites for the same TF but do not give any penalty for the sites that remain unaligned. The bonus score for an aligned pair with affinities W_i and W'_j is just

$$\lambda(W_i + W'_j) \tag{2}$$

where λ is a parameter to be optimized. This quantity is denoted by ΔG_T in the Figure 2A of the paper.

To define the penalty score, let (s, s') and (t, t') be two adjacent pairs in

the alignment, and let the distance (on DNA G) between sites s and t be x basepairs and the distance (on DNA G') between s' and t' be x' basepairs. Then our penalty score which penalizes for loose clustering and for the energy needed to bend and twist the two DNA helices to a common conformation is given as

$$F(x, x') = \mu \frac{x + x'}{2} + \nu \frac{(x - x')^2}{x + x'} + \xi \frac{(\Delta\phi)^2}{x + x'} \tag{3}$$

where μ, ν and ξ are parameters to be optimized separately.

The first term of function (3) penalizes for loose clustering proportionally to the distance between the sites.

The second term models the energy needed to displace a spring with spring constant $\nu/(x + x')$ the amount of $|x - x'|$ i.e. the energy needed to compress the two DNAs to a common length assuming that DNA behaves like a spring whose spring constant is inversely proportional to the length. See [Bryant *et al.*, 2003] and references therein.

The final term similarly models the energy needed to twist the DNA helices so that the two transcription factors can reach a common 3D structure on both DNAs. The twist angle for regular B–DNA having one rotation for every 10.4 basepairs on average becomes $\Delta\phi = (x - x')\frac{2\pi}{10.4} - 2k\pi$ where k is an integer such that $\Delta\phi$ will be the minimum distance to full rotation, i.e., $-\pi \leq \Delta\phi < \pi$.

Dynamic programming for local alignments

We want to find the best local alignments of the binding site sequences S and S' under the above scoring scheme (2,3). The resulting alignment algorithm resembles the traditional dynamic programming for local alignment [Smith & Waterman, 1981]. We use a matrix D of size $L \times L'$ for which $D_{i,0} = D_{0,j} = -\infty$ for all i and j. The cell $D_{i,j}$ holds the score for best local alignment whose last aligned pair is the ith site of S and the jth site of S'. The algorithm finds the best local aligment by evaluating D from the recursion

$$D_{i,j} = \begin{cases} \max\limits_{\substack{0<p_i-q_k<1000 \\ 0<p'_j-q'_l<1000}} \{\lambda w_{ij}, D_{k,l} + \lambda w_{ij} - F(p_i - q_k, p'_j - q'_l)\} & \text{, if } f_i = f'_j \\ -\infty & \text{, otherwise.} \end{cases} \quad (4)$$

where $w_{ij} = (W_i + W'_j)$. Note that we have limited the range of the maximization such that the distance between adjacent aligned pairs must be < 1000.

The highest scoring putative enhancer element corresponds to the best local alignment. The algorithm finds it from the matrix D by backtracking from the cell with highest score along the maximizing path. Additional suboptimal local alignments, or enhancer elements, are traced from the next highest scoring cell from which the backtrack does not overlap with previously backtracked alignments. Repeating this, a desired number of non–overlapping putative enhancers can be reported, in decreasing order of the score.

EEL software

Our software system, called EEL (Enhancer Element Locator), implements the above algorithm. The current software is capable of aligning sequences several megabases long. The software has user friendly graphical interface and runs under several platforms, including Linux, MacOSX and Windows.

The EEL software including the source code but excluding the parameter optimization is available at `http://www.cs.helsinki.fi/u/kpalin/EEL/` under the GNU General Public License.

Application of EEL for Genome Wide Prediction of Mammalian Enhancer Elements

Assignment of TF sites

The sequences were scanned for binding sites after addition of a pseudocount of 0.001 to all cells of the binding site matrices. Binding sites of score 9 or higher were included in the alignments. For Drosophila eve analysis (Figure 2A), only DNA binding matrices for Hunchback, Caudal, Knirps, Bicoid and Kruppel (from [Berman *et al.*, 2002]) were used. The 107 binding matrices that were used for mammalian analyses and parameter optimization included matrices from our own analyses and high-quality transcription factor binding profiles available in the literature [Horvath *et al.*, 1995, Tun *et al.*, 1994] and in the JASPAR2 database [Sandelin *et al.*, 2004] (see Table S1).

Parameter Optimization

The values $\lambda = 2$, $\mu = 0.12$, $\nu = 200$ and $\xi = 200$ were used for the free parameters of the EEL scoring function. The values were selected after utilizing greedy hill climbing optimization as follows:

1. Pick a random set of parameters as current.
2. Compute the goodness of the current parameters.
3. While enough iteration do:
4. Take a small random step in the parameter space.
5. Recompute the goodness for the new parameters.
6. If improvement, set the new parameters to be current.

We measure the goodness of the parameter settings by the sum of relative local alignment scores that are better in homologous than in non–homologous

sequences. To be exact, we aligned 10 randomly picked non-homologous sequence pairs and computed the mean of the score of the best local alignments on them. We then aligned 10 randomly picked homologous sequence pairs and summed the suboptimal alignment scores divided by the mean from the non-orthologous alignments while this ratio exceeded one. For all sequences the exons and tandem repeats were first masked to focus the alignments on the non-coding regions. The hill–climbing search to a local optimum was repeated several times using random initial values of the parameters. The final parameter values for the scoring function were selected as approximate average of the local optima.

Genome wide EEL alignment

For aligment of all human and mouse orthologous genes (as defined by EnsMart 23.1), sequences were downloaded from the ensembl database (releases 23.34e, 23.32c for human and mouse, respectively). A total of 20173 gene pairs were aligned, corresponding to 17429 human genes with the remaining pairs representing genes that have two orthologues in the mouse. In the star-alignment, the following Ensembl assemblies were used: Chick:24.1a, Puffer fish:25.1, Rat:24.3c, Human: 24.34e.

The genome-wide alignment of orthologous gene pairs with 100 kb flanking sequences on both sides of the gene required approximately 2000 CPU hours (approx. 6 min/gene/processor) on a cluster of 64 bit AMD Opteron Linux-servers running at 2 GHz. The results were placed to a MySQL database utilizing Distributed Annotation System (DAS) allowing display of predicted enhancer modules on other DAS-compatible tools, such as the Ensembl ContigView website. The DAS server and the full database are available for downloading at http://www.cs.helsinki.fi/u/kpalin/EEL/.

Single gene alignments and data visualization

The Drosophila even-skipped, mammalian MyoD, N-myc and c-myc genomic sequences were aligned using stand-alone version of EEL designed for single-

gene analyses in two species. 2D representations of the alignments were generated using the program gff2aplot [Abril *et al.*, 2003].

Testing of EEL predictions

To determine which TFs are overrepresented in the APC target genes, we selected all APC targets (c-MYC, CD44, TIAM1, SEMA3C, EPHB3, EPHB2, AXIN2, SOX17, MMP7, LAMININγ2, GPR49, FGF4 and TASR2) mentioned in the text of a single publication [Sansom *et al.*, 2004] to avoid investigator bias. We then determined the 'background' distribution of all 107 analyzed TFs in all predicted enhancer modules in the human to mouse alignment that were shorter than 2000 bp and whose EEL score were higher than 400. Similarly, we determined the number of occurrence of all TF pairs in such predicted enhancers. Probability of occurrence in all genes (p_{ALL}) was calculated by dividing the number of a given TF or TF pair found by the total number of all TFs or TF pairs found, respectively.

Putative enhancer modules were then selected also from the APC target gene alignments according to the criteria described above, and the overrepresentation of a given TF was determined using binomial distribution, with probability being p_{ALL} and number of tests being the total number of TFs or TF pairs in the modules selected from the APC target genes. Correction for multiple hypothesis testing was performed using the equation $p_{multiple} = 1 - (1 - p_{single})^n$, where n is the number of hypotheses (107) and p_{single} is the probability derived from the binomial distribution and $p_{multiple}$ is the corrected and reported p–value.

To estimate false negative rate in target gene identification, we selected genes which had been shown to be directly regulated by GLI and TCF in mouse embryos. To be sure that the comparison set represented true positives, the standard of evidence we used was very strict. Either the directly-regulated enhancer needed to be characterized in transgenic mouse embryos, or the gene needed to be induced in all tissues tested in vivo by factors activating GLI or TCF (Hh and Wnts), and its enhancer characterized and the

regulation shown to be direct in vitro.

There are three such genes for TCF (AXIN2 [Jho *et al.*, 2002], Cdx1 [Lickert & Kemler, 2002] and Brachyury [Yamaguchi *et al.*, 1999], and four for GLI (PTCH1 [Agren *et al.*, 2004, Goodrich *et al.*, 1997]),
GLI1 [Dahmane *et al.*, 1997, Dai *et al.*, 1999, Lee *et al.*, 1997],
HNF-3β [Sasaki *et al.*, 1997] and Myf5), one of which (Myf5) was excluded because of conflicting reports [Gustafsson *et al.*, 2002, Teboul *et al.*, 2003]. Of these, we found 2 out of 3 for GLI, and 1 out of 2 for TCF. Generally similar fraction of genes that have been reported as direct targets of Hh-GLI or Wnt-TCF by usually reliable but less conclusive methods were also identified by EEL. Because of the different experimental methods used caused considerable difficulty of imposing an unbiased and uniform standard of evidence on this set, we did not use it as a basis for the false negative analysis, simply indicated which identified and not identified genes have been reported in the literature as Hh or Wnt targets.

Comparison of EEL with DNA alignment method

The five GLI binding motifs were matched to orthologous sequences from human and mouse as described above. From the list of putative binding sites we obtained list of non–overlapping segments where there was at least 2 sites with affinity greater than 25 within 1000bp region. These segments were found by greedy left to right scan, and 260646 such pairs were found in human sequences.

These segments of length 1000bp with binding sites in the middle were compared with the alignment tool BLAST (bl2seq) with sensitive word size parameter $W = 7$. All high scoring pairs (HSPs) that obtained e–value less than one were retrieved and the number of sites whose midpoint was included in an HSP was counted. There was in total 14216 regions that were homologous with mouse sequences (a stringency where one false positive is expected i.e. e-value less than 1/260646) and contained two conserved GLI

sites with score > 25.

In supplementary table S7 we provide list of segments that contain at least two conserved sites and for which the best BLAST score is over 164. This cutoff is the greatest score that yields more (112 in total) results than the original EEL approach (110). The DNA alignment method was inferior to EEL in identifying known GLI target genes. Whereas EEL finds two out of three direct, well validated GLI targets, the DNA-based alignment finds only one. EEL is thus clearly better at outperforming random selection ($p = 8.8 \cdot 10^{-5}$; Hypergeometric test, Population 20173, Sample 110, Positive 3, Observed 2) than the clustering + DNA-alignment method ($p = 1.6 \cdot 10^{-2}$; Hypergeometric test, Sample 112, Observed 1). Consistently, out of GLI targets for which there is no evidence for directness or the evidence is not conclusive, EEL identifies 7, the DNA alignment 4. These results are consistent with the fact that only some nucleotides in enhancer elements are used to code for TF binding sites, and some mutations in TF sites are conservative, resulting in little or no change in affinity. DNA alignment algorithms treat all nucleotides as equivalent, thus resulting in increased noise (see Figure S2, Table S7), which necessarily makes signal detection more difficult. Aligning DNA and subsequently identifying TF sites on the aligned sequence is analogous to analyzing conservation at the protein level by first aligning DNA and then identifying conserved codons. This results in a complete loss of the improved sensitivity and specificity of protein sequence -based alignment (see for example [Wernersson & Pedersen, 2003]. Similarly, DNA-alignment -based methods applied to enhancer identification are not expected to perform as well as methods such as EEL which only analyze sequences that code TF binding sites and their relative distances.

Constructs and oligonucleotides

The amino acids included in the TF-Renilla Luciferase fusion proteins are shown in table 1.

Table 1: TF-Renilla Luciferase fusion constructs

Factor	Accession no.	Amino acids
GLI1	NP_005260	226-437 (KREP...GAMK),
GLI2	NP_084657	99-325 (KQEA...SSGL)
GLI3	NP_000159	471-702 (KQEP...KPMT)
Ci	NP_524617	442-668 (KDEP...SDIS)
Tcf4	CAA72166	32-596 (SENS...KSLE)
c-Ets1	CAA32903	1-353 (KAAV...PDAD)

The DNA oligonucleotides for the TF binding assay were from TAG Copenhagen (Denmark). The sequence of the biotinylated and 'consensus' oligonucleotides included the binding sequences for the TFs, flanked on both sides by 5-7 bp of unrelated sequence. Binding sequences used were gaccaccca [Kinzler & Vogelstein, 1990] (GLI1-3 and Ci) [Korinek *et al.*, 1997], cctttgatc (Tcf4) and caggaagtg [Woods *et al.*, 1992] (c-ETS1). Competitor oligonucleotides were similar, but contained single base substitutions to the binding sequence. The scrambled oligonucleotides contained the same bases as the consensus oligonucleotide, with the bases in the binding site in a random order. To generate double-stranded DNA (dsDNA) oligonucleotides, equal molar amounts of the complementary single-stranded oligonucleotides were annealed, and the concentration of the resulting dsDNA was measured using PicoGreen (Molecular Probes).

References

[Abril *et al.*, 2003] Abril, J. F., Guigo, R., & Wiehe, T. (2003). gff2aplot: Plotting sequence comparisons. Bioinformatics *19*(18), 2477–9. 1367-4803 Evaluation Studies Journal Article.

[Agren *et al.*, 2004] Agren, M., Kogerman, P., Kleman, M. I., Wessling, M., & Toftgard, R. (2004). Expression of the ptch1 tumor suppressor gene is

regulated by alternative promoters and a single functional gli-binding site. Gene *330*, 101–14. 0378-1119 Journal Article.

[Berman *et al.*, 2002] Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., & 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 U S A *99*(2), 757–762.

[Bryant *et al.*, 2003] Bryant, Z., Stone, M. D., Gore, J., Smith, S. B., Cozzarelli, N. R., & Bustamante, C. (2003). Structural transitions and elasticity from torque measurements on dna. Nature *424*(6946), 338–341.

[Dahmane *et al.*, 1997] Dahmane, N., Lee, J., Robins, P., Heller, P., & Ruiz i Altaba, A. (1997). Activation of the transcription factor gli1 and the sonic hedgehog signalling pathway in skin tumours. Nature *389*(6653), 876–81. 0028-0836 Journal Article.

[Dai *et al.*, 1999] Dai, P., Akimaru, H., Tanaka, Y., Maekawa, T., Nakafuku, M., & Ishii, S. (1999). Sonic hedgehog-induced activation of the gli1 promoter is mediated by gli3. J Biol Chem *274*(12), 8143–52. 0021-9258 Journal Article.

[Goodrich *et al.*, 1997] Goodrich, L. V., Milenkovic, L., Higgins, K. M., & Scott, M. P. (1997). Altered neural cell fates and medulloblastoma in mouse patched mutants. Science *277*(Aug 22), 1109–1113.

[Gustafsson *et al.*, 2002] Gustafsson, M. K., Pan, H., Pinney, D. F., Liu, Y., Lewandowski, A., Epstein, D. J., & Emerson, C. P., J. (2002). Myf5 is a direct target of long-range shh signaling and gli regulation for muscle specification. Genes Dev *16*(1), 114–26. 0890-9369 Journal Article.

[Horvath *et al.*, 1995] Horvath, C. M., Wen, Z., & Darnell, J. E., J. (1995). A stat protein domain that determines dna sequence recognition suggests

a novel dna-binding domain. Genes Dev *9*(8), 984–94. 0890-9369 Journal Article.

[Jho *et al.*, 2002] Jho, E. H., Zhang, T., Domon, C., Joo, C. K., Freund, J. N., & Costantini, F. (2002). Wnt/beta-catenin/tcf signaling induces the transcription of axin2, a negative regulator of the signaling pathway. Mol Cell Biol *22*(4), 1172–83. 0270-7306 Journal Article.

[Kinzler & Vogelstein, 1990] Kinzler, K. W. & Vogelstein, B. (1990). The gli gene encodes a nuclear protein which binds specific sequences in the human genome. Mol Cell Biol *10*(2), 634–42. 90136577 0270-7306 Journal Article.

[Korinek *et al.*, 1997] Korinek, V., Barker, N., Morin, P. J., van Wichen, D., de Weger, R., Kinzler, K. W., Vogelstein, B., & Clevers, H. (1997). Constitutive transcriptional activation by a beta-catenin-tcf complex in apc-/- colon carcinoma. Science *275*(5307), 1784–7. 0036-8075 Journal Article.

[Lee *et al.*, 1997] Lee, J., Platt, K. A., Censullo, P., & Ruiz i Altaba, A. (1997). Gli1 is a target of sonic hedgehog that induces ventral neural tube development. Development *124*(13), 2537–52. 0950-1991 Journal Article.

[Lickert & Kemler, 2002] Lickert, H. & Kemler, R. (2002). Functional analysis of cis-regulatory elements controlling initiation and maintenance of early cdx1 gene expression in the mouse. Dev Dyn *225*(2), 216–20. 1058-8388 Journal Article.

[Sandelin *et al.*, 2004] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., & Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32 Database issue*, D91–4. 1362-4962 Journal Article.

[Sansom *et al.*, 2004] 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., Clarke, A. R., & Winton, D. J. (2004). Loss of apc in vivo immediately perturbs wnt signaling, differentiation, and migration. Genes Dev *18*(12), 1385–90. 0890-9369 Journal Article.

[Sasaki *et al.*, 1997] Sasaki, H., Hui, C., Nakafuku, M., & Kondoh, H. (1997). A binding site for gli proteins is essential for hnf-3beta floor plate enhancer activity in transgenics and can respond to shh in vitro. Development *124*(7), 1313–22. 97236478 0950-1991 Journal Article.

[Smith & Waterman, 1981] Smith, T. F. & Waterman, M. S. (1981). Identication of common molecular subsequences. Journal of Molecular Biology *147*, 195–197.

[Stormo, 2000] Stormo, G. D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*(1), 16–23.

[Stormo & Fields, 1998] Stormo, G. D. & Fields, D. S. (1998). Specificity, free energy and information content in protein-dna interactions. Trends Biochem Sci *23*(3), 109–113.

[Teboul *et al.*, 2003] Teboul, L., Summerbell, D., & Rigby, P. W. (2003). The initial somitic phase of myf5 expression requires neither shh signaling nor gli regulation. Genes Dev *17*(23), 2870–4. 0890-9369 Journal Article.

[Tun *et al.*, 1994] Tun, T., Hamaguchi, Y., Matsunami, N., Furukawa, T., Honjo, T., & Kawaichi, M. (1994). Recognition sequence of a highly conserved dna binding protein rbp-j kappa. Nucleic Acids Res *22*(6), 965–71. 0305-1048 Journal Article.

[Wernersson & Pedersen, 2003] Wernersson, R. & Pedersen, A. G. (2003). Revtrans: Multiple alignment of coding dna from aligned amino acid sequences. Nucleic Acids Res *31*(13), 3537–9.

[Woods *et al.*, 1992] Woods, D. B., Ghysdael, J., & Owen, M. J. (1992). Identification of nucleotide preferences in dna sequences recognised specifically by c-ets-1 protein. Nucleic Acids Res *20*(4), 699–704.

[Yamaguchi *et al.*, 1999] Yamaguchi, T. P., Takada, S., Yoshikawa, Y., Wu, N., & McMahon, A. P. (1999). T (brachyury) is a direct target of wnt3a during paraxial mesoderm specification. Genes Dev *13*(24), 3185–90. 0890-9369 Journal Article.