Regulatory Analysis for Exploring Human Disease Progression

Dustin T. Holloway[1], Mark Kon[2], Charles DeLisi[3§]

[1]Molecular Biology Cell Biology and Biochemistry Department, Boston University, 5 Cummington Street, Boston, USA
[2]Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, USA
[3]Bioinformatics and Systems Biology, Boston University, 44 Cummington Street, Boston, USA

[§]Corresponding author

Email addresses:
DTH: dth128@bu.edu
MK: mkon@bu.edu
CD: delisi@bu.edu

Abstract

A crucial goal in post-genomic research is unravelling the regulatory network of transcription factors and their target genes. This is especially important in the human genome where many TFs are involved in disease progression. Similarly to our previous work in the yeast genome, machine learning methods can be applied to known examples of target genes to create decision rules that aid in the identification of new targets. Using targets in publicly available databases as training examples, classifiers have been built and tested for 153 human TFs. These classifiers base their decisions upon the integration of three types of sequence information: composition, conservation, and overrepresentation.

Overall, 33 TF classifiers achieve a precision greater than 60%. Many high confidence predictions are made for all TFs, and all targets are made available for download on our web server and as supplementary information. To highlight the power of this method we briefly discuss the regulator Oct4, a known marker for stem cells. Results for the TF Wt1 are then discussed in detail, showing that many of the predictions share functions with the known targets. Since Wt1 and its targets are intimately involved in development of Wilms' tumor, new models for its action in cancer are proposed. We show that our predictions for Wt1 are statistically enriched for genes which fall into specific chromosomal loci known to be associated with Wilms' tumor. This finding suggests new models for Wilms' tumor progression where dysregulation in these regions is important for disease progression, either through loss of Wt1 or loss of the associated region. Genes in significant loci include several oncogenes and tumor suppressors which are candidates for involvement in cancer progression and may partially explain the observed clinical and biochemical data on this cancer. The anti-apoptotic effects of Wt1 are also discussed along with several new target genes, including *bax* and *pde4b*, which may help mediate this effect. Finally, motif discovery is used to propose a new binding motif for Wt1 which will be useful in later site identification.

SVM-based classifiers provide a comprehensive platform for analysis of regulatory networks, and the results can be used to make new hypothesis for disease progression in humans. Obvious extensions of this work include incorporating new kinds of data, and expanding the analysis to more TFs when binding data becomes available.

Background

Although many factors influence the regulation of genes, the fundamental step of regulatory control is the association of transcription factors (TFs) with their binding sites in DNA. In simpler organisms it is often sufficient to search for these sites in a gene's promoter, which is typically defined as anywhere from 800 to 5000 base pairs upstream of the transcription start site. In higher organisms like humans, TFs may exert regulatory control at a distance of many kilobases from the start site. Complex genomes also show greater incidence of binding sites occurring within 5' UTRs, introns, 3' UTRs, and even far downstream of a gene.

Any TF will have a varying affinity for different nucleotide strings and will thus bind to a repertoire of similar sites in the genome. This site affinity is often described as a motif or preferred pattern of bases. A popular representation of the binding motif is the position specific scoring matrix (PSSM) [1-4], which gives the frequency of observed nucleotide bases at each position of a known motif. However, results produced by scanning DNA with basic PSSM models are often overwhelmed by a high rate of false positive predictions[5]. To improve target prediction, we have previously employed a more sophisticated supervised learning method in *Saccharomyces cerevisiae* which combines many types of genomic data to assist binding site classification[6-8]. We have also developed a method to rank specific genomic features (e.g., presence or conservation of a particular *k*-mer) to select those which are most important for identifying target promoters for a particular TF[7,9]. We now apply these methods, based on the support vector machine (SVM) to produce separate classifiers for 153 TFs in the human genome in an attempt to discover new regulatory interactions important to human disease.

The genomic datasets used include sequence information from promoters (2kb upstream and 5' UTR), introns, and 3' UTRs and take account of 1) sequence composition, 2) sequence conservation in 8 vertebrate genomes, and 3) statistical over-representation. These datasets have high dimensionality (see Methods), often containing thousands of numerical features. During classifier construction SVM recursive feature elimination (SVM-RFE)[10] is used to reduce the feature set to a manageable size. Figure 1 provides a graphical scheme describing classifier construction. Feature ranking as well as feature set and classifier construction are described more completely in the Methods section. Each gene used in the analysis is described by a numerical vector. Each number, or feature, represents one measurement taken in the genome, for example, the number of occurrences of a particular *k*-mer in a gene’s promoter. SVMs efficiently handle high dimensional datasets and have proven effective in a wide range of biological systems [11-17].

Figure 1

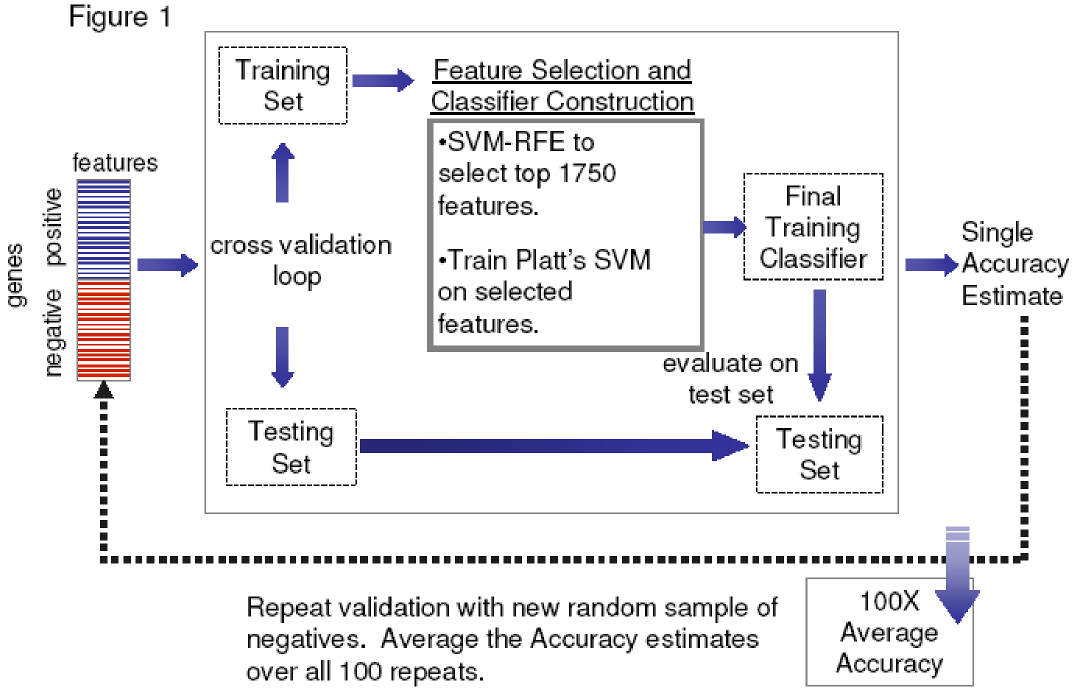
features
genes
positive
negative
Training Set
cross validation loop
Feature Selection and Classifier Construction
•SVM-RFE to select top 1750 features.
•Train Platt's SVM on selected features.
Final Training Classifier
evaluate on test set
Testing Set
Testing Set
Single Accuracy Estimate
Repeat validation with new random sample of negatives. Average the Accuracy estimates over all 100 repeats.
100X Average Accuracy

Figure 1 - SVM Framework

This figure shows the data mining scheme for making TF classifiers. 100 classifiers are constructed for each TF, each using a different random sub-sample of the negative set. A classifier built on the training set is evaluated using leave-one-out cross validation (center, gray box). For every cross-validation split, the top 1750 features are selected using SVM-RFE and the classifier is trained and finally used to classify the test set (left out sample). This process is repeated 100 times, and the accuracy for the procedure is the average of the 100 cross-validation accuracies.

SVMs require the input of positive (known target) genes and negative (non-target) genes to develop a decision rule which can be used to classify new genes as bound or not bound by a TF. Positive examples are curated from several publicly available databases and also ChIP-chip experiments when available (see Methods). The negative set is always chosen randomly from the genome to be of equal size to the positive set. Clearly a random choice of negatives can introduce bias into the classifier since some of the chosen genes may in fact be targets. To resolve this difficulty, each TF classifier is constructed one hundred times, each with a new sampling of negatives. The performance of each classifier is evaluated by cross validation and a final accuracy measurement is then the average accuracy from all one hundred trials. This will average out the fluctuations which occur due to possible errors in negative set selection.

This analysis produces several highly accurate classifiers for human TFs, several of which are important to human development and disease. First, new predictions for Oct4 targets are discussed as well as possible implications for diabetes. Finally, Wt1 is examined in detail for its role in cancer development and progression. Many biologically relevant targets are proposed and possible new functions are highlighted, such as involvement in nervous system development, tumor migration, interaction with Wnt signalling, and regulation of new targets responsible for resistance of apoptosis. The predicted targets of Wt1 are also significantly enriched in genes which lie in chromosomal regions known to be associated with progression of Wilms' tumor. This finding is significant since now disease progression can be linked to dysregulation or loss of Wt1's targets in these regions. Motif discovery methods are also used to propose a new binding motif for Wt1 which may be useful in future site identification.

Results and Discussion

Not surprisingly, many TF classifiers show poor performance. Of 153 TFs tested, 33 showed a PPV greater than 0.6. This may be partly due to the fact that our defined promoter region is large and in some cases may be thousands of base pairs long. This size may interfere with the ability of the SVM to identify important regions. Also, the greater complexity and combinatorial regulation occurring in the human genome may not be captured well by single TF classifiers. Finally, most human TFs have few known targets, making it less likely that a classifier will find the correct decision rule. It should be noted that the performance measures are taken with the classifier decision threshold set to 0.5. This is the optimal classifier threshold since it indicates that genes exceeding the threshold have better than a 50% chance of being a true target. For the predictions we discuss below, we only accept genes as targets if they pass a threshold of 0.95 (i.e., 95% likelihood of being positive) on average for all 100 classifiers constructed for a particular TF. Thus, several TFs with lower accuracies may still provide meaningful results at this stringent threshold. Results for all TFs are available as Supplementary Information File 1, and on our web server at http://cagt10.bu.edu/TFSVM/Main%20Frame%20Page.htm. Supplementary File 1 also contains some brief notes on the naming conventions of TFs, and how the classifiers were constructed.

SVM Classifiers Identify Biologically Relevant Targets for Oct4

Regulation by Oct4 is essential in early development, and expression of Oct4 is important for maintaining the pluripotency of embryonic stem cells[18,19]. ChIP-chip analysis of Oct4 and several other regulators revealed Oct4 can act in concert with the TFs Nanog and Sox2[19]. The SVM classifier for Oct4 has an accuracy of 67% and a PPV of 66%.

It has been discovered that Oct4 targets are enriched for transcription factors, with many of these also being important for development[19]. In fact, the known targets in the training set for Oct4 are significantly enriched in the GO term "transcription regulatory activity" (50 genes, p=2.1e-16), and new SVM predictions also show enrichment in this category (111 genes, p=6.7e-34). The known targets and new predictions share many statistically enriched functional terms, including "developmental protein", "homeobox", and "Wnt signalling pathway." For a complete list see Supplementary File 2.

The authors in [19] noted that several targets of Oct4 fall into the Wnt signalling pathway. Indeed, both the known target set and the new predictions are enriched for genes in the Wnt pathway (p = 0.01, p = 0.0014 respectively). Figure 2 shows the Wnt pathway, highlighting SVM predictions alongside previous knowledge. The new targets targets in this pathway fit well with the known biology of Oct4. Other research has indicated that activation of the Wnt pathway can help sustain pluripotency[20]; thus, these results fit with the hypothesis that Oct4 acts to maintain the undifferentiated state by activating the Wnt pathway in stem cells. It is of interest that several of predicted targets of Oct4 appear to be genes that inhibit Wnt signalling, which suggests that Oct4 may repress these genes.

Figure 2

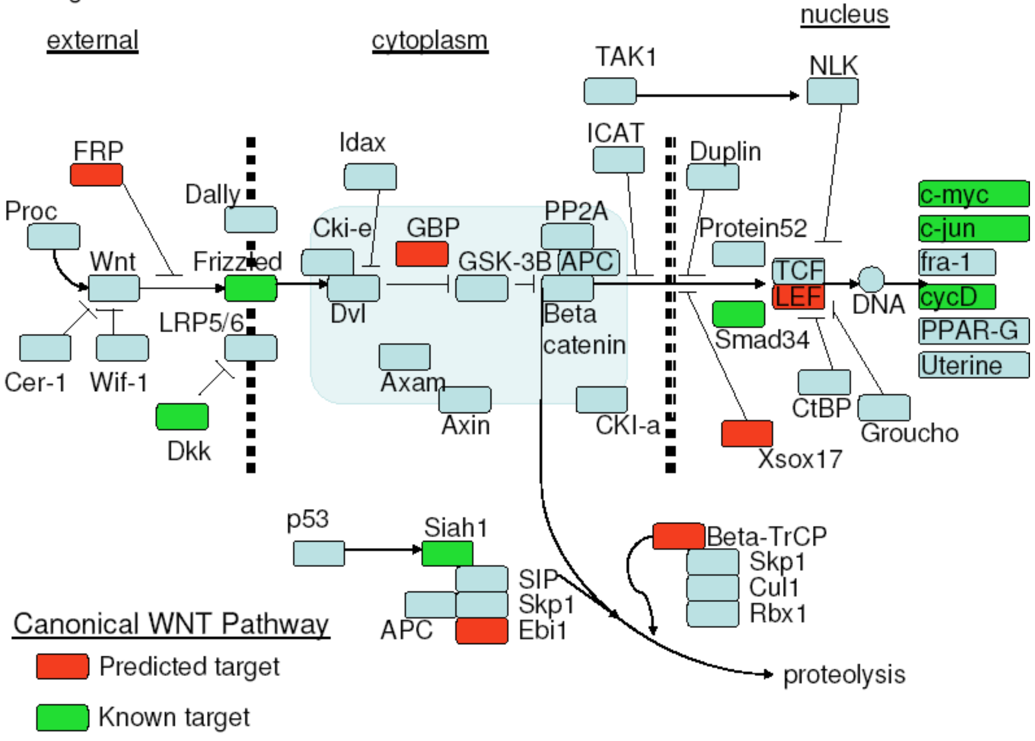

Figure 2 - Oct-4 Targets in the Wnt Signalling Pathway

Known targets of Oct-4 are filled in green and new predictions are filled in red. Known targets and predicted targets are statistically enriched for genes falling in this pathway.

Oct4 regulates several genes involved in Diabetes

Oct4 is known to have high expression in stem cells and reduced expression once cells begin differentiation. Oct4 binds the promoters of several genes important for differentiation, and some of these are factors which can contribute to the onset of diabetes. The known targets of Oct4 are significantly enriched in genes falling into the KEGG pathway Maturity Onset Diabetes of the Young (MODY, p=0.039). Particularly, Oct4 binds the gene PDX1, which causes Type IV MODY when mutated[21]. SVM predicts two new targets falling in this pathway. Most interesting is NeuroD1, which has been shown to cause Type VI MODY when mutated[22]. This evidence hints that Oct4 may play role in diabetes if its mode of regulation is disturbed. Others have hypothesized that disruption of normal transcriptional regulation is the ultimate cause of MODY TypeVI when NeuroD1 is lost[21]. This leaves open the possibility that the disruption of NeuroD1 targets could also be achieved by dysregulation or mutation of Oct4. Further experiments will be needed to explore this possibility.

Regulation by Wt1

The *Wilms tumor 1* (*wt1*) gene codes for an essential transcription factor which plays a role in normal urogenital formation. The *wt1* gene is found to be overexpressed in an assortment of cancers including leukemia[23], lung[24], colon[25], thyroid[26], breast[27], and several others[28-32]. Mutations in *wt1* are also known to result in predisposition to Wilms' Tumor, a renal malignancy accounting for 8% of childhood cancers[33]. Wt1 can either activate or repress target genes, and has a complex role in carcinogenesis acting as both a tumor suppressor[34,35] and an oncogene[36] depending on its context. To further complicate its role, the gene encodes 4 splice variants, each thought to have separate functions and slightly different DNA binding affinities. Regulation by Wt1 is not well defined, and its function may be modulated by post-translational modification or by physical contact with other regulators, including possible dimerization with other proteins or with itself.

The classifier for Wt1 has a prediction accuracy of 68% and a PPV of 75%. 354 new predictions were made for Wt1 at 95% confidence, and these genes show significant enrichment for several KEGG pathways in which there are previously annotated targets. These pathways are Map-kinase (p = 1.1e-3), adherens junction (p = 8.7e-3), and calcium signalling (p = 4.7e-2). Furthermore, the new target set is statistically enriched (p = 1.7e-4, hypergeometric test) in genes showing differential expression in a microarray study[37] of wild-type vs. mutant *wt1* tumors. These data suggest that the classifier for this TF is revealing accurate biological hits.

May Regulate Apoptosis Through Factors Other than Bcl2

As a tumor suppressor, Wt1 is thought to represses several growth factor receptors, and expression of Wt1 has been shown to impede cell growth in a variety of tumors. Inconsistent with this function is the fact that 90% of sporadic Wilm's tumors maintain expression of wildtype Wt1[37-39]. Further investigation has shown that Wt1 provides protection against programmed cell death. This protection is mediated, at least in part, by interaction with *p53* to suppress its apoptotic effects and by direct activation of the anti-apoptotic gene *bcl-2[36]*.

Several newly predicted targets are genes known to be anti-apoptotic or otherwise regulate cell death (Supplementary File 3). One notable prediction is that Wt1 binds the promoter of *bax*, a pro-apoptotic gene whose protein product binds to *bcl2* and disrupts its repression of apoptosis. A possible hypothesis is that Wt1 acts in a dual fashion, activating the expression of *bcl2* while repressing *bax*. Also interesting is the predicted target *pde4b*, which can augment apoptosis when inactivated[40]. The possibility that Wt1 activates *pde4b* suggests that loss of Wt1, and hence downregulation of *pde4b*, contributes to the sensitivity to apoptosis observed in *wt1* mutants.

Plays an Important Role in Migration and is Directly Linked to the Wnt Pathway

Recent evidence indicates that Wt1 is involved in cellular migration[41]. Currently no known targets of the TF are directly involved in this process. Functional grouping of our target predictions reveals a group of 67 genes which are annotated to cellular adhesion, cytoskeleton, or cell motility (Supplementary File 4). This group includes many cadherin and contactin genes known to be involved in adhesion and migration. Notably, this set also contains *wasf1*, *irsp53*, *afadin*, and *arhgap6*, which are all closely related to actin polymerization and associated with adherens junctions and cell migration. Also of interest are *nectin* and *α-catenin*, core components of the adherens junction. Regulation of these genes by Wt1 may play an important role in the modulation of cellular adhesion and migration in cancer.

The complex role of Wt1 requires that different genetic changes must take place in wildtype-*wt1* vs. mutant-*wt1* tumors. Tumors expressing normal *wt1* have increased resistance to cell death and respond poorly to treatment with chemotherapeutic agents that act by induction of apoptosis. Tumors with *wt1* mutations may become sensitized to apoptosis and thus tend to accumulate compensatory mutations which activate cellular growth and proliferation. In a study examining a group of *wt1*-mutant tumors, it was discovered that 75% also contained mutations in the *β-catenin* gene[37], a known oncogene and crucial component of the Wnt-signalling pathway. The Wnt pathway influences cell growth, development, migration, and adhesion, and is known to take part in carcinogenesis. It is also a pathway often disregulated in cancer, containing several oncogenes and tumor supressors. Thus popular hypotheses suggest that *wt1*-mutant tumors become sensitive to apoptosis and thus require a compensating mutation in *beta-catenin* which constitutively activates the Wnt pathway. The malignancy may adapt to the loss of *wt1* by accumulating other mutations as well

Near the plasma membrane β-catenin links cadherins in adherens junctions to α-catenin. Cancerous cells undergoing metastasis progress through what is called the Epithelial-Mesenchymal Transition (EMT), a hallmark of which is a dissociation of the E-cadherin/ β-catenin/α-catenin complex. This would result in loss of adherens junctions and increased cell mobility. The disruption would release β-catenin, allowing it to translocate to the nucleus where it cooperates with the TCF/LEF complex to activate targets of the Wnt pathway. Clearly, disruption of adherens junctions and increased Wnt signalling are important for progression in some cancers. Wt1 may affect this process in two major ways. First, the prediction that Wt1 regulates *α-catenin* is intriguing, implying that Wt1 could directly disrupt adherens junctions by repressing *α-catenin*. This may also activate Wnt signalling by freeing β-catenin from adherens junctions, and allowing it to translocate to the nucleus where it participates in Wnt target activation. Second, Wt1 may contribute to the activation of Wnt signalling by controlling a key Wnt regulator, TCF, which appears in our list of predicted targets. The view that Wt1 enhances Wnt signalling is supported by expression experiments showing reduction in Wnt4 expression in *wt1* knockout cells[42].

Alternatively, Wt1 may act in the opposite way to repress Wnt signalling which is consistent with its role as a tumor suppressor. This view is supported by the fact that several (but not all) Wnt targets are upregulated in *wt1*-mutant as opposed to *wt1*-wildtype tumors. Current research has shown that mutations of β-catenin in Wilms tumors is essentially confined to cases where *wt1* is also mutated. Mutations hyperactivating Wnt signalling may then be the primary reason for the observed upregulation of Wnt targets in *wt1*-mutants. There is also some evidence that Wt1 may actually bind the promoter of *β-catenin* itself, for which the SVM model assigns a probability of 0.7. Closer inspection of the *β-catenin* promoter reveals 11 matches to the Wt1 consensus site within 600bp of the *β-catenin* transcriptional start site (Figure 3). The true functional relationship between Wt1 and the Wnt pathway will have to be elucidated through further experiments, but there is strong evidence that Wt1 is intimately involved with the Wnt pathway and likely exerts significant regulatory control on Wnt mediators and targets.

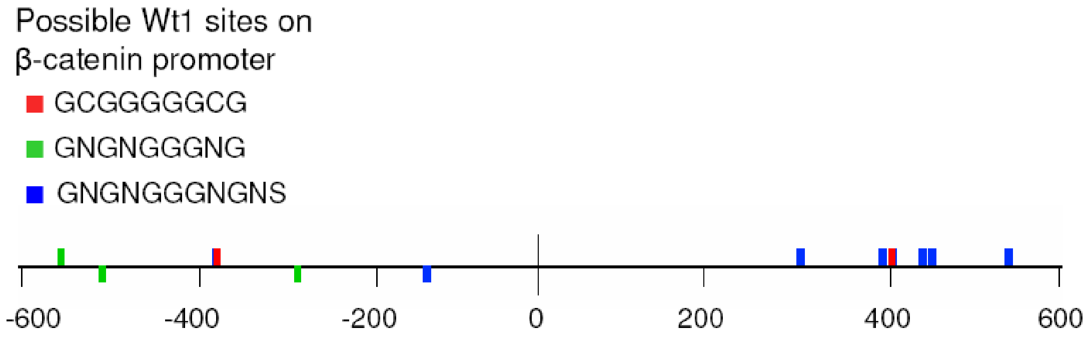

Figure 3 - Possible binding sites for Wt1 near the *β-catenin* gene

This figure shows the region spanning 1200bp centered on the *β-catenin* transcriptional start site. Potential Wt1 binding sites are highlighted as follows: red-GCGGGGGCG[68], green-GNGNGGGNG[69], blue-GNGNGGGNGNS[70].

Role of Wt1 in Nervous Tissue Development and Disease

The set of combined targets (known and newly predicted) for Wt1 is significantly enriched in several annotation categories related to the nervous system and neuron growth: transmission of nerve impulse (p = 0.0069), synaptic transmission (p=0.013), and neurotransmitter receptor (p=0.058). Many genes are annotated to similar categories but do not show statistical significance (Supplementary File 5). These may still be important since they all relate to development or function of the nervous system. Observations have been made of neuronal differentiation markers in Wilms' Tumor[43], demonstrating that some mechanism in these tumors is activating nerve cell signature genes. Wt1 has been shown to be required for normal development of the neurons in retinal[44] and olfactory[45] tissues. Furthermore, analysis in the developing mammalian embryo has shown presence of Wt1 in brain, tongue, and retinal tissues[46]. Surprisingly, one highly significant predicted target for Wt1 is the *tas1r1* gene, which is a receptor responsible for detecting sweet compounds. This implies that, aside from its proven roles in eye and olfactory development, Wt1 is also involved in taste sensation. Also along these lines are the potential new targets *eya1* and *eya4*, which are member of a gene family known to be involved in kidney, eye, and ear development.

Another supporting prediction is the *mtmr2* gene which, when mutated, can cause Charcot-Marie-Tooth Diseases type 4B[47]. This is a demyelinating disease of the nervous system which causes sensory and motor defects. It is interesting that one of the chromosomal loci implicated in Charcot-Marie-Tooth Disease is 11p15[48], also known to be involved in Wilms' Tumor. Finally, 48 high confidence targets can be annotated as being either voltage gated ion channels, integral to the plasma membrane, or being part of a neurotrophic ligand/receptor interaction (Supplementary File 5). Taken together, this is compelling evidence that Wt1 is thoroughly involved in the nervous system, and may play a role in diseases affecting nerve tissue. This view appears to agree with symptoms observed in the clinic. Patients with WAGR syndrome, which causes predisposition to Wilms' Tumor, show mental retardation and aniridia, a defect of the iris. Also, there are reported cases of deafness and mental retardation accompanying Denys-Drash syndrome[49], which also predisposes patients to Wilms' Tumor.

Some Predicted Targets of Wt1 fall into disease associated chromosomal loci

Wilms' tumors have largely been studied in individuals genetically predisposed to tumor formation due to genetic abnormality (as opposed to sporadic Wilms' Tumor). The syndromes resulting from these abnormalities and the associated chromosomal changes are listed in Table 1. It has become clear that deletions, duplications, and other abnormalities at chromosomal loci 11p13[50,51], and 11p15[52] are involved in predisposition to Wilms' Tumor. Other studies have also implicated loss of heterozygosity in regions 16q, 1p, and 22q as correlating with poor outcome in Wilms' Tumor patients[53,54]. One study used a range of probes to determine that loss of heterozygosity(LOH) in the specific region near 11p15.5 is associated with Wilms' Tumor[55].

The *wt1* gene itself is located in 11p13, and naturally explains why disruption of this region contributes to tumor formation. Potentially important genes in the other chromosomal regions have been postulated, including *igf2* and *p57* in 11p15. The true factors causing predisposition to Wilms' Tumor at these regions remains unknown, and it is hypothesized that regulatory targets of Wt1 in these regions might contribute to disease. One suggestion is that the region 11p15 may contain tumor suppressor activity since allelic loss in this region correlates with tumor formation. Clearly a more defined set of regulatory targets for Wt1 would greatly improve the understanding of this gene's role in normal tissue as well as in carcinogenesis. Figure 4 depicts the genetic changes which may lead to tumor formation by the syndromatic or sporadic pathways.

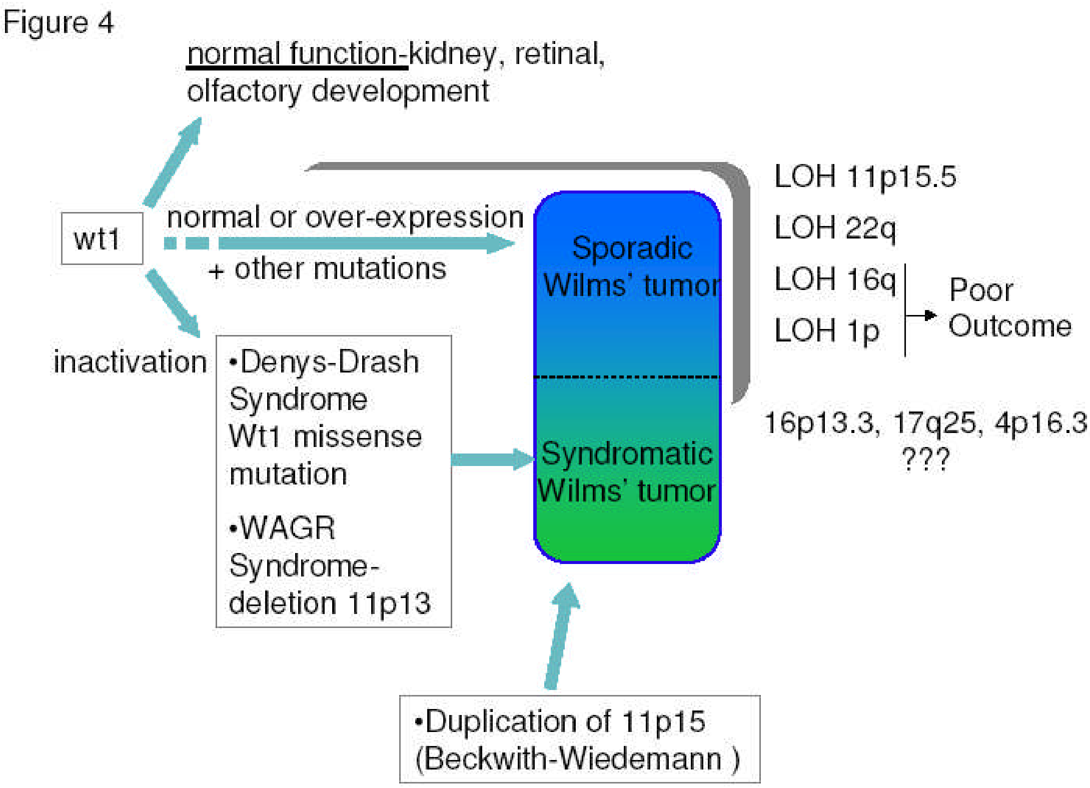
Figure 4
normal function-kidney, retinal, olfactory development
wt1
normal or over-expression
+ other mutations
inactivation
•Denys-Drash Syndrome Wt1 missense mutation
•WAGR Syndrome-deletion 11p13
Sporadic Wilms' tumor
Syndromatic Wilms' tumor
LOH 11p15.5
LOH 22q
LOH 16q
LOH 1p
Poor Outcome
16p13.3, 17q25, 4p16.3
???
•Duplication of 11p15 (Beckwith-Wiedemann)

Figure 4 - Pathways to Wilms' Tumor

Genetic changes leading to Wilms' Tumor. Cancer occurs through the sporadic or the syndromatic pathway. Few cases of tumor occurring in the syndromatic pathway also exhibit loss of heterozygosity(LOH) or loss of imprinting(LOI) at other loci whereas sporadic cases of Wilms tumor regularly exhibit LOH and LOI. The gray bar indicates that the LOH events may occur anywhere along the development of the sporadic cancer. Most sporadic cases (but not all) have a wild-type or overexpressed *wt1* gene. It is possible that LOH, LOI, and other genetic changes in sporadic tumors compensate for the presence of wt1. LOH at regions 16q and 1p correlate with poor prognosis. Other regions often showing LOH are listed. Regions 16p13.3, 17q25, and 4p16.3 are statistically enriched for predicted targets of Wt1 but their involvement in tumor formation is unknown.

Strikingly, examining the predicted targets of Wt1 shows that these genes occur more frequently than expected by chance in cytobands 11p15.5 (p = 6.3e-5), and 1p36.3 (p = 6.3e-4). Three of the new targets for Wt1 in 11p15.5 are possible tumor suppressors: Rnh1[56], Igf2as[57], and CD151[58]. *If in fact Wt1 normally activates these genes it could explain why inactivation of Wt1 or loss of genes in 11p15.5 contributes to cancer formation since in both cases expression of these tumor suppressors would be abolished*. Also in 11p15.5 is *mucdhl*, a cadherin like protein. Loss of *mucdhl* could contribute to loss of cell adhesion by disruption of adherens junctions. This could be another significant step toward tumor migration and metastasis.

Although 16q and 22q, which correlate with poor prognosis, have no statistical enrichment, predicted targets do lie in these regions. Examination of the genes in these regions allows the assembly of a model of Wilms' tumor which explains some of the observed clinical behaviour. There are predicted target genes with known tumor suppressor activity in the regions 16q and 1p which could explain why loss of these regions correspond to poor clinical outcome (*cbfa2t3*[59] in 16q, and *eno1*[60] in 1p). Also lying in 1p is the predicted target *pde4b* which, as mentioned earlier, can augment apoptosis when inactivated[40].

Other chromosomal regions with strong enrichment include 16p13.3 (p = 4.3e-6, most significantly enriched location), 17q25 (p=1.7e-5) and 4p16.3 (p=4.3e-3). These regions contain several new predictions which may be relevant to tumor formation. At 16p13.3 new targets include *kremen2* and *tsc2*, both of which are thought to be tumor supressors. At 17q25 lies the predicted target *fasn*. Inhibition of *fasn* can cause apoptosis[61] and also sensitizes cancer cells to treatment by chemotherapy[62]. Activation of *fasn* could provide another mechanism by which Wt1 supports resistance to apoptosis. At the 4p16.3 locus the gene *fgfr3* is likely to be a target and is known to be important for cancer progression[63,64].

Recent analysis has shown that there are far fewer LOH events in tumors containing *wt1* mutation, suggesting that regions shown to undergo LOH harbor genes regulated by Wt1 or downstream effectors[65]. This supports the idea that without inactivation of *wt1*, tumor cells must undergo alternative mutations which selectively activate or inactivate its targets. Then, once the tumor suppressor effects of Wt1 are abrogated, the cancerous cells are free to benefit from the apoptosis resistance conferred by active Wt1.

Not all predictions make perfect sense, and highlight the complexity of Wt1 regulation and its role in tumor formation. For example, *igf2* (11p15.5) is a proposed target of Wt1 and is a potent growth factor proposed to be important for cancer progression. Igf2 is upregulated in many Wilms' tumors due to either duplication of the 11p15 paternal allele (often in concert with LOH on the maternal allele) or removal of silencing on the maternal allele (loss of imprinting)[66]. Whereas some evidence shows that loss of genes at 11p15 is crucial for cancer formation, other results demonstrate that abnormal activation of some genes is also important. Since dysregulation of genes at 11p15 is due to epigenetic changes as well as genetic mutations, it is difficult to predict the implications of regulation by Wt1 without direct experimentation. Indeed, the specific genomic changes may vary between tumors, specifically between sporadic and syndromatic cases of Wilms' tumor.

Finally, Wt1 is predicted to regulate the transcription factor Pou6f2. This factor has been suggested to be a tumor suppressor, and mutations in Pou6f2 confer a predisposition to Wilms tumor[67]. Repression or activation of pou6f2 by Wt1 could have a profound effect on carcinogenesis. Loss of activation through mutation of Wt1 could enhance cancer progression in -mutant individuals. Alternatively, increased repression of pou6f2 through overexpression of Wt1 could intensify malignancy in -wildtype tumor patients. More studies will be necessary to uncover the details of the interplay between these two factors.

Since many chromosomal regions have been observed to undergo allele loss, duplication, or other mutation in Wilms' tumor, we have compiled a list of known targets and significant predictions which fall into several important regions (Supplementary File 6).

A New Binding Motif for Wt1

Discovery of a binding site for Wt1 has proven difficult since each isoform of the regulator may bind to slightly different sequences in DNA. Dimerization with other proteins and post-translational modifications may also alter the binding affinity in undetermined ways. Several consensus sites for Wt1 have nevertheless been proposed (GCGGGGGCG [68], GNGNGGGNG[69], GNGNGGGNGNS[70], and GCGTGGGAGT[71]); unfortunately, showing that Wt1 binds to a site *in vitro* has proven to be a poor predictor of binding and regulatory action *in vivo[33]*. The four related consensus sites reported in the literature can be seen in Figure 5A. Our classification based approach has yielded a set of 353 high confidence targets to add to the set of 14 genes known to be bound. This provides a rich group from which to perform motif discovery.

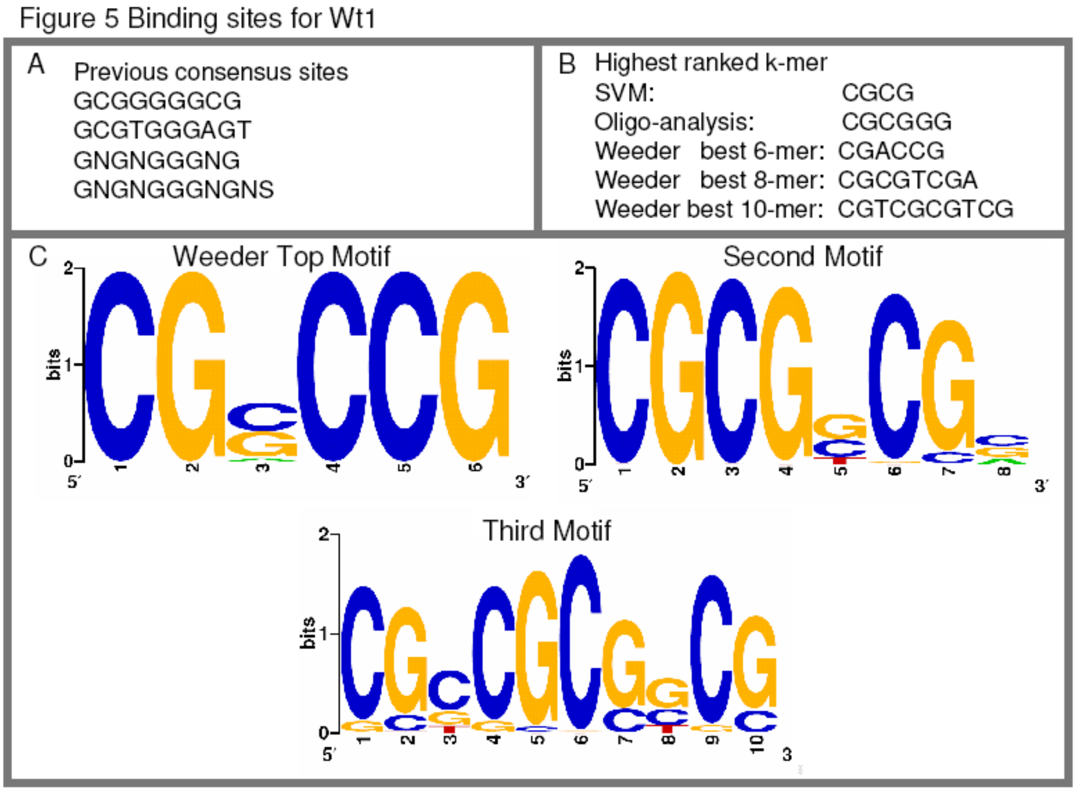

Figure 5 – Motif Discovery on Wt1 Targets

5A lists the proposed consensus binding sites for Wt1 from the literature sources mentioned in the text. 5B shows the top ranked *k*-mer from each motif discovery method, including the best *k*-mer ranked by the

SVM model. 5C shows the top 3 PSSMs created by the Weeder algorithm. Motif discovery was performed on all known and newly predicted targets of Wt1.

A first approach (see Methods) comes from an SVM procedure which iteratively ranks each feature used by the classifier to determine those that are most useful in distinguishing the known targets (positives) form the non-targets (negatives). This method has been applied successfully to the *S. cerevisiae* genome to yield nucleotide strings which matched well with the known affinities of the TFs. In this case it produces a ranking of *k*-mers based on information in the training set alone, so that new predictions do not contribute to the *k*-mer ranking. Two other methods have been applied to the entire set of predictions and known targets. The first of these is oligo-analysis[72,73], which scores each *k*-mer (up to $k = 6$) by its over-representation in promoters of the gene set (see Methods). The second is an algorithm called Weeder[74-76] which implements an efficient search to score and rank all *k*-mers of length 6, 8, and 10, while also allowing mismatches. Weeder was one of the best performing motif discovery algorithms in a recent comparison[77].

Figure 5B shows the top scoring *k*-mers from all methods. The results are uniform in that the discovered sites are GC-rich. The 4-mer ranked highest by SVM (CGCG) is also present in the result given by oligo-analysis and in the best 8 and 10-mers found by Weeder. The Weeder algorithm offers a further advantage since it automatically clusters the most similar of the significant *k*-mers (of any length) and combines them into consensus site and creates a position weight matrix (PWM) based on the occurrences of the consensus in the gene set. Figure 5C shows the top 3 PWMs reported by Weeder. A scan of the known target promoters of Wt1 with the best PWM shows that all but 1 contains a perfect match to this matrix. Binding by Wt1 is complex, and these motifs may describe only one possible binding mode of the regulator. Although experimentation is required to validate any predictions, these motifs may aid investigators in future site identification or binding affinity studies with Wt1. Supplementary File 7 contains the raw outputs from Weeder, oligo-analysis, and results of scanning previously proposed consensus sites against the promoters of predicted Wt1 targets.

Materials and Method

SVM training and validation

SVM is a classification technique developed by Vapnik [78]. Binary labelled training examples (e.g., 0, 1 or negative, positive) are feature vectors **x** corresponding to individual genes, with each vector composed of genome scale sequence measurements (see below). The measurements are the attributes, or features, of the data. Positives are known targets of the TF while negatives are a randomly selected set of genes. The SVM algorithm performs an optimization to find a maximal-margin hyperplane separating the two classes of training data. Maximal-margin refers to the fact that the separator is selected to be as distant as possible from the training points, achieving maximum separation between classes. We have successfully applied machine learning to regulatory analysis before[6-8], and an in-depth tutorial on SVM training is available on our web site.

Training an SVM involves setting a parameter *C*, which adjusts tolerance for misclassifications. The classifier for the Myc transcription factor was used as the prototype for parameter selection. Five-fold cross validation was used to measure the performance of several values of *C*, and the value resulting in lowest classifier error was chosen for subsequent use in all classifiers. Tested values include: [2^{-7} ,2^{-5} ,2^{-3} ,2^{-1} ,1, 1.5, 2, 2^2,2^3,2^4,2^5,2^6]. The value 2^{-7} was reported by the SPIDER machine learning package[79] as having the best performance.

Choosing negatives for classifier construction is difficult since there is no defined set of genes known not to be targets. For every TF, a set of negatives is chosen randomly to be equal in size to the positive set. 100 classifiers are made in this way using a different randomly selected negative sets. All 100 classifiers are tested using leave-one-out cross validation (LOOCV), and the final performance measurements (accuracy, PPV, etc) are averaged over all trials. The full scheme is sketched in Figure XX, and closely follows that reported in [9]. A short outline appears below.

For an example TF A:

1. Assemble positive set (denote size as *n*). Sample *n* genes randomly to construct the negative set.
2. Spit the data for LOOCV.
3. Use SVM-RFE to rank all features in the training set.
4. Construct SVM classifier on best 1750 features. Save full feature ranking.
5. Classify left out example.
6. Repeat steps 2-5 to complete LOOCV. Save all feature rankings.
7. Calculate performance statistics (Accuracy, PPV, etc.)
8. Repeat steps 1-7 100 times.
9. Calculate final performance statistics (i.e., mean Accuracy, mean PPV, etc.).

A new gene can be classified by applying all 100 classifiers for TF A to the feature vector for that gene. Each classification produces a posterior probability (see below), and the mean of all 100 probabilities is calculated. Typically, if $P > 0.5$, a gene is classified as a positive. In this paper we increase the cutoff to $P \geq 0.95$ to select only the highest quality targets for each TF. Feature rankings on each training set are saved and used to calculate the final ranks of each feature (see below).

Classifying new targets and prediction significance

As described in [80] and applied in [9] the SVM can produce a probabilistic output. This is a class conditional probability of the form *P*(target | SVM output), where distance from the gene to the hyperplane classifier. We refer to this output simply as the true positive probability and denote it using the upper-case P (e.g., $P \geq 0.95$), while other statistical tests which output p-values are denoted in lower-case (e.g., $p \leq 0.01$). The probability is calculated by fitting a sigmoid function to the SVM output using 3-fold cross validation. Thus, genes lying at a greater distance from the hyperplane on the positive side will have higher probabilities (i.e., more likely to be positive). This form of probabilistic output makes sense as one would expect genes falling deep into the positive region to be more likely to be targets. New genes are classified using the average probability assigned by all 100 classifiers for a given TF. An average posterior probability greater then 0.5 is generally considered to yield a positive, but only genes with P ≥ 0.95 are accepted in this study as high confidence targets.

Genomic feature selection and ranking

As demonstrated in the yeast genome [9], the SVM algorithm can be used to select and rank features. One main output of the SVM procedure is the vector **w**, which contains the learned weights of each data feature. The **w** vector is calculated directly as

$$\mathbf{w} = \sum_{i=1}^{s} \alpha_i y_i \mathbf{x}_i ,$$

where y_i is the class label of the training example $\mathbf{x}_i$, which is the feature vector for gene *i*. Here *s* is the number of support vectors, or genes which lie directly on the classifier margin. These are essentially the examples which are closest to the separator, and thus the only ones necessary to define its placement. The numbers α_i are values of the Lagrange multipliers used during the optimization process[78,81,82].

Features with larger **w** components are more useful in distinguishing between the positives and negatives. The SVM recursive-feature-elimination (SVM-RFE) algorithm uses the **w** vector to iteratively select important features[10]. The original algorithm begins by training an SVM and discarding the attributes which have smallest weights [10]. The process is repeated until a set number of features remains. In this study, half of the features are removed during each iteration until 2050 features remain. They are then eliminated individually until 1750 are left. As indicated in the Discussion, the target of 1750 is determined by exploring the effect of feature selection on the prototype TF-classifier for Myc.

Since ranking is performed on each training set during a LOOCV, and because 100 classifiers are cross validated for each TF, many feature rankings are accumulated for each TF. Lastly, a count is taken of the number of times each feature appears in the top 40 of any ranking. The final rank is made by sorting the features according to the frequency of their appearance as a “top 40 feature.” Genes high on this list are consistently ranking high over all cross validation trials and all choices of negative set, making them reliable in that they are robust to changes in the training set.

Feature Datasets

Several regulatory sequence regions were extracted for 18660 human genes from the UCSC genome browser[83,84]. These regions consist of: 1) 2kb of sequence upstream of the transcription start site plus the 5’UTR, 2) all introns, 3) 3’UTR. Three different types of feature measurements were taken on this sequence data.

1. *k*-mers —The distribution of all *k*-mers in a gene’s regulatory regions may be used to predict whether it is bound or not-bound by a TF. Feature vectors are formed by enumerating all possible strings of nucleotides of length 4, 5, and 6. The number of occurrences of each string is counted in a gene’s promoter region, and this string of counts is the feature vector for the gene. For each gene, the counts for 4-mers, 5-mers, and 6-mers are normalized separately to mean 0 and standard deviation 1. This is separate from the feature normalization which occurs prior to SVM training. *k*-mer

counts are performed separately and summed for each regulatory region mentioned above.

2. *k*-mer Overrepresentation —This method calculates the significance of occurrences of each *k*-mer in the a gene's regulatory regions. This method is the same as that reported in our previous work[9] and follows the equations set out by RSA tools[72,73]. Here, the background sequence set is all gene promoters (2kb upstream), 5'UTRs, introns, and 3'UTRs. *E*-values were used and calculated according to

$$Evalue = -\log_{10}\left(pvalue \times D\right),$$

where *D* is the number of *k*-mers in the analysis. Higher *E*-values correspond to more significant *k*-mers.

3. Conserved *k*-mers—This method for constructing a *k*-mer conservation matrix is based on output generated by the PhastCons algorithm[85,86] and follows the procedure outlined in our work in the yeast genome[9]. Introns and 3'UTRs are now also included for the human genome. Essentially, *k*-mers are counted in gene regulatory regions as in data method 1, but each k-mer instance is weighted according to its level of conservation in a multiple alignment of sequences from human and seven other vertebrate genomes (chimp, dog, mouse, rat, chicken, zebrafish, fugu). Genomic alignments and PhastCons scores may be downloaded from the UCSC genome browser website[83,84].

If a *k*-mer is not conserved, it receives a count of 1 just as it would in simple *k*-mer counting.

As in[9], the weighting metric is:

$$\frac{1}{1-\beta P_c}$$

where P_c is the average PhastCons score for a particular *k*-mer. β is an adjustable parameter which controls how much the conservation of a *k*-mer increases its count. In this study we choose β = 0.75, so that an element with a maximum conservation of 1 has a count of 4.

Functional Analysis

Statistical enrichment of gene sets for particular gene functions was calculated using the Functional Annotation Tool in DAVID 2006[87]. Enrichment for functions was calculated using default background sets provided in DAVID. DAVID uses the Fisher Exact test to measure functional enrichment in annotation categories from numerous public databases (e.g., KEGG pathways, GO terms, Spir keywords, etc). Enrichment for chromosomal locations was found using DAVID but searching only for enriched cytobands. Genes were also clustered according to functional similarity using the Functional Annotation Clustering tool in DAVID.

Positive Binding Targets

Known binding sites for human TFs were parsed from several public databases in January 2006. The databases used are Oregano[88], TRDD[89], Transfac[90], Ensembl[91], and the Eukaryotic Promoter Database[92]. Many binding sites were also manually curated from literature sources. Several large-scale experimental binding studies were also examined to identify binding sites[19,93-97]. In all cases, binding sites found outside of the sequence region studied (i.e., 2kb upstream, 5' UTR, introns, and 3' UTR) were excluded. Lists of literature curated binding sites with Pub-med references and a spreadsheet of binding interactions parsed from the above databases can be downloaded in Supplementary File 1

Motif Discovery

Motif Discovery was performed on Wt1 known targets plus new predictions. Sequence data for each gene went to 1kb upstream and 0.5kb downstream of transcriptional start. The sequence data was downloaded from the human promoter extraction database at Cold Spring Harbor Laboratory[98]. Motif discovery was performed with Weeder[75] and Oligo-analysis[72] available at the RSA-tools website[73]. The full raw output from Weeder and Oligo-analysis along with the accompanying fasta files is available as Supplementary File 7. Matching of consensus strings to promoter regions was performed using RSAtools.

Table 1 Syndromes causing predisposition to Wilms' Tumor

Syndrome	Occurrence of Wilms tumor	Chromosomal abnormality	Ref.
WAGR	98% by age 6	Deletion at 11p13	OMIM: **#194072**
Beckwith-Wiedemann	96% by age 8	Duplication of paternal 11p15. May result in increased gene expression(IGF2) or inactivation(p57).	OMIM: **#130650**
Denys-Drash	96% by age 5	Missense mutation in WT1 (11p13 locus) causing dominant negative phenotype.	OMIM: **#194080**

Table 1

This table highlights the syndromes causing predisposition to Wilms' Tumor development, and the genetic changes associated with the syndrome. These include WAGR[50] Denys-Drash[52], and Beckwith-Wiedemann[51] syndromes.

Supplementary Information

Supplementary File 1

This file contains several sub-folders. The folder "Classifier Results" provides a file containing the probabilities for all possible associations of TFs with genes in humans. It also contains a list of classifiers and their associated performance measures. The folder “Literature_curated_targets” contains the known TF-target interactions taken from databases and the literature. Any interactions manually curated from primary literature are listed, and the Pubmed ID of the article used is given. All files are annotated so as to be self explainatory or have an accompanying Readme file.

Supplementary File 2

This file contains two excel spreadsheets providing the functional annotations of known targets and predicted targets of Oct4 respectively. These are annotations as provided by the DAVID system at NIH and include the statistical significance of each functional category.

Supplementary File 3

Using both known and newly predicted targets, this file contains a list of genes which relate to apoptosis as given by the DAVID functional analysis tools. The genes appear several times in various, similar annotation categories which are related to cell death pathways

Supplementary File 4

Using just the newly predicted targets, this file contains a list of genes which relate to cellular adhesion, cytoskeleton, or motility as given by the DAVID functional analysis tools.

Supplementary File 5

Using both known and newly predicted targets, this file contains a list of genes which are annotated to terms by DAVID which are somehow related to the nervous system. Three main categories are present (represented by folders) which each contain several functional terms and the genes annotated to them. The three main categories are “Neuron related”, “Sensory perception”, and “Voltage gated channels and membrane receptors.”

Supplementary File 6

Using both known and newly predicted targets, this file contains a list of genes and the chromosomal cytobands to which they are mapped. P-values generated by DAVID are also given to show statistical enrichment.

Supplementary File 7

This file contains the results of running the Weeder algorithm on 1) the set of known and newly predicted ($P \geq 0.95$) targets of Wt1, and 2) the known targets of Wt1. Sequence regions used are as defined in Methods. The file also contains the results of Oligo-analysis. Also included is the matching results after scanning the literature derived consensus sites for Wt1 against the full set of Wt1 targets (predicted and known).

Author contributions.

DH coded the required software in Matlab and Perl, conceived of many of the design implementations, and wrote this article. All authors made contributions to this manuscript and the experimental design. CD initially conceived and motivated this work. All authors read and approved the final manuscript

Competing interests. The authors have declared that no competing interests exist.

References

1. Stormo GD (2000) DNA Binding Sites: Representation and Discovery. Bioinformatics 16: 16-23.
2. Workman CT, Stormo GD (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput: 467-478.
3. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. Journal of Molecular Biology 188: 415-431.
4. Schneider T, Stephens R (1990) Sequence logos: a new way to display consensus sequences. Nucl Acids Res 18: 6097-6100.
5. Fickett JW (1996) Coordinate Positioning of MEF2 and Myogenin Binding Sites. Gene 172: 19-32.
6. Holloway D, Kon M, DeLisi C (2005) Integrating genomic data to predict transcription factor binding. Proc of the Workshop on Genome Informatics 16: 83-94.
7. Holloway D, Kon M, DeLisi C (2006) Machine Learning for Predicting Targets of Transcription Factors in Yeast: in press. Systems and Synthetic Biology.
8. Holloway D, Kon M, DeLisi C (2006) Machine Learning Methods for Transcription Data Integration. IBM Journal of Research and Development on Systems Biology 50.
9. Holloway D, Kon M, DeLisi C (2006) Building Transcription Factor Classifiers and Discovering Relevant Biological Features: submitted. PLOS Computational Biology.
10. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46: 389-422.
11. Jaakola T, Diekhans M, Haussler D (1999) Using the Fisher kernel method to detect remote protein homologies. Proc Int Conf INtell Syst Mol Biol: 149-158.
12. Hua (2001) A novel method of protein secondary structure prediction with high segment overlap measure:support vector machine approach. Journal of Molecular Biology 308: 397-407.
13. Hua., Sun. (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 18: 721-728.
14. Zien A, Ratsch G, Mika S, Scholkopf B, Lengauer T, et al. (2000) Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 16: 799-807.

15. Wang M, Yang J, Chou K-C (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28: 395-402.
16. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, et al. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16: 906-914.
17. Pavlidis P, Noble WS (2001) Gene Functional Classification from Heterogeneous Data. RECOMB Conference Proceedings: 249-255.
18. Chambers I (2004) The molecular basis of pluripotency in mouse embryonic stem cells. Cloning And Stem Cells 6: 386-391.
19. Boyer LA, Tong IL, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.
20. Sato N, Meijer L, Skaltsounis L, Greengard P, Brivanlou AH (2004) Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. Nature Medicine 10: 55-63.
21. Fajans SS, Bell GI, Polonsky KS (2001) Molecular Mechanisms and Clinical Pathophysiology of Maturity-Onset Diabetes of the Young. N Engl J Med 345: 971-980.
22. Malecki MT, Jhala US, Antonellis A, Fields L, Doria A, et al. (1999) Mutations in NEUROD1 are associated with the development of type 2 diabetes mellitus. 23: 323-328.
23. Inoue K, Sugiyama H, Ogawa H, Nakagawa M, Yamagami T, et al. (1994) WT1 as a new prognostic factor and a new marker for the detection of minimal residual disease in acute leukemia. Blood 84: 3071-3079.
24. Yusuke Oji SM, Hajime Maeda, Seiji Hayashi, Hiroya Tamaki, Shin-Ichi Nakatsuka, Masayuki Yao, Eigo Takahashi, Yoko Nakano, Hirohisa Hirabayashi, Yasushi Shintani, Yoshihiro Oka, Akihiro Tsuboi, Naoki Hosen, Momotaro Asada, Tatsuya Fujioka, Masaki Murakami, Keisuke Kanato, Mari Motomura, Eui Ho Kim, Manabu Kawakami, Kazuhiro Ikegame, Hiroyasu Ogawa, Katsuyuki Aozasa, Ichiro Kawase, Haruo Sugiyama, (2002) Overexpression of the Wilms' tumor gene <I>WT1</I> in <I>de novo</I> lung cancers. International Journal of Cancer 100: 297-303.
25. Oji Y, Yamamoto H, Nomura M, Nakano Y, Ikeba A, et al. (2003) Overexpression of the Wilms' tumor gene WT1 in colorectal adenocarcinoma. Cancer Science 94: 712-717.
26. Oji Y, Miyoshi Y, Koga S, Nakano Y, Ando A, et al. (2003) Overexpression of the Wilms' tumor gene WT1 in primary thyroid cancer. Cancer Science 94: 606-611.
27. Loeb DM, Evron E, Patel CB, Sharma PM, Niranjan B, et al. (2001) Wilms' Tumor Suppressor Gene (WT1) Is Expressed in Primary Breast Tumors Despite Tumor-specific Promoter Methylation. Cancer Res 61: 921-925.
28. Oji Y, Yano M, Nakano Y, Abeno S, Nakatsuka S-i, et al. (2004) Overexpression of the Wilms' tumor gene WT1 in esophageal cancer. Anticancer Research 24: 3103-3108.

29. Oji Y, Nakamori S, Fujikawa M, Nakatsuka S-i, Yokota A, et al. (2004) Overexpression of the Wilms' tumor gene WT1 in pancreatic ductal adenocarcinoma. Cancer Science 95: 583-587.
30. Oji Y, Inohara H, Nakazawa M, Nakano Y, Akahani S, et al. (2003) Overexpression of the Wilms' tumor gene WT1 in head and neck squamous cell carcinoma. Cancer Science 94: 523-529.
31. Ueda T, Oji Y, Naka N, Nakano Y, Takahashi E, et al. (2003) Overexpression of the Wilms' tumor gene WT1 in human bone and soft-tissue sarcomas. Cancer Science 94: 271-276.
32. Oji Y, Suzuki T, Nakano Y, Maruno M, Nakatsuka S-i, et al. (2004) Overexpression of the Wilms' tumor gene WT1 in primary astrocytic tumors. Cancer Science 95: 822-827.
33. Lee BS, Haber D (2001) Wilms Tumor and the WT1 Gene. Experimental Cell Research 264: 74-79.
34. Luo X, Reddy J, Yeyati P, Idris A, Hosono S, et al. (1995) The tumor suppressor gene WT1 inhibits ras-mediated transformation. Oncogene 11: 743-750.
35. Haber D, Park S, Maheswaran S, Englert C, Re G, et al. (1993) WT1-mediated growth suppression of Wilms tumor cells expressing a WT1 splicing variant. Science 262: 2057-2059.
36. Mayo M, Wang C, Drouin S, Madrid L, Marshall A, et al. (1999) WT1 modulates apoptosis by transcriptionally upregulating the bcl-2 proto-oncogene. EMBO 18: 3990-4003.
37. Li C-M, Kim CE, Margolin AA, Guo M, Zhu J, et al. (2004) CTNNB1 Mutations and Overexpression of Wnt/{beta}-Catenin Target Genes in WT1-Mutant Wilms' Tumors. Am J Pathol 165: 1943-1953.
38. Coppes M, Liefers G, Paul P, Yeger H, Williams B (1993) Homozygous Somatic WT1 Point Mutations in Sporadic Unilateral Wilms Tumor. PNAS 90: 1416-1419.
39. Little M, Wells C (1997) A clinical overview of WT1 gene mutations. Human Mutation 9: 209-225.
40. Moon E, Lee R, Near R, Weintraub L, Wolda S, et al. (2002) Inhibition of PDE3B Augments PDE4 Inhibitor-induced Apoptosis in a Subset of Patients with Chronic Lymphocytic Leukemia. Clin Cancer Res 8: 589-595.
41. Jomgeow T, Oji Y, Tsuji N, Ikeda Y, Ito K, et al. (2006) Wilms' tumor gene WT1 17AA(-)/KTS(-) isoform induces morphological changes and promotes cell migration and invasion in vitro. Cancer Science 97: 259-270.
42. Sim E, Smith A, Szilagi E, Rae F, Ioannou P, et al. (2002) Wnt-4 regulation by the Wilms' tumour suppressor gene, WT1. Oncogene 21: 2948-2960.
43. Hussong J, Perkins S, Huff V, McDonald M, Pysher T, et al. (2000) Familial Wilms' Tumor with Neural Elements: Characterization by Histology, Immunohistochemistry, and Genetic Analysis. Pediatric and Developmental Pathology 3: 561-567.
44. Wagner K-D, Wagner N, Vidal VP, Schley G, Wilhelm D, et al. (2002) The Wilms' tumor gene Wt1 is required for normal development of the retina. EMBO 21: 1398-1405.

45. Wagner N, Wagner K-D, Hammes A, Kirschner KM, Vidal VP, et al. (2005) A splice variant of the Wilms' tumour suppressor Wt1 is required for normal development of the olfactory system. Development 132: 1327-1336.
46. Armstrong J, Pritchard-Jones K, Bickmore W, Hastie N, Bard J (1993) The expression of the Wilms' tumour gene, WT1, in the developing mammalian embryo. Mechanisms of Development 40: 85-97.
47. Bolino A, Muglia M, Conforti FL, LeGuern E, Salih MAM, et al. (2000) Charcot-Marie-Tooth type 4B is caused by mutations in the gene encoding myotubularin-related protein-2. 25: 17-19.
48. Othmane KB, Johnson E, Menold M, Graham FL, Hamida MB, et al. (1999) Identification of a New Locus for Autosomal Recessive Charcot-Marie-Tooth Disease with Focally Folded Myelin on Chromosome 11p15. Genomics 62: 344-349.
49. Jadresic L, Leake J, Gordon I, Dillon M, Grant D, et al. (1990) Clinicopathologic review of twelve children with nephropathy, Wilms tumor, and genital abnormalities (Drash syndrome). Journal of Pediatrics 117: 717-125.
50. Heyningen V, Bickmore W, Seawright A, Fletcher J, Maule J, et al. (1990) Role for the Wilms Tumor Gene in Genital Development. PNAS 87: 5383-5386.
51. Elliott M, Maher E (1994) Beckwith-Wiedemann syndrome. Journal of Medical Genetics 31: 560-564.
52. Meuller R (1994) The Denys-Drash syndrome. Journal of Medical Genetics 31: 471-477.
53. Klamt B, Schulze M, Thäte C, Mares J, Goetz P, et al. (1998) Allele loss in Wilms tumors of chromosome arms 11q, 16q, and 22q correlates with clinicopathological parameters. Genes, Chromosomes and Cancer 22: 287-294.
54. Grundy PE, Telzerow P, Breslow N, Moksness J, Huff V, et al. (1994) Loss of heterozygosity for chromosomes 16q and p1 in Wilms' tumors predicts an adverse outcome. Cancer Research 54: 2331-2331.
55. Mannens M, Slater R, Heyting C, Bliek J, de Kraker J, et al. (1988) Molecular nature of genetic changes resulting in loss of heterozygosity of chromosome 11 in Wilms' tumours. Human Genetics 81: 41-48.
56. Fu P, Chen J, Tian Y, Watkins T, Cui X, et al. (2004) Anti-tumor effect of hematopoietic cells carrying the gene of ribonuclease inhibitor. 12: 268-275.
57. Yang J, Chen W, Liu Z, Luo Y, Liu W (2003) Effects of insulin-like growth factors-IR and -IIR antisense gene transfection on the biological behaviors of SMMC-7721 human hepatoma cells. Journal of Gastroenterology and Hepatology 18.
58. Saur G, Kurzeder C, Grundmann R, Kreienberg R, Zeillinger R, et al. (2003) Expression of tetraspanin adaptor proteins below defined threshold values is associated with in vitro invasiveness of mammary carcinoma cells. Oncology Reports 10.
59. Kochetkova M, McKenzie OLD, Bais AJ, Martin JM, Secker GA, et al. (2002) CBFA2T3 (MTG16) Is a Putative Breast Tumor Suppressor Gene from the Breast Cancer Loss of Heterozygosity Region at 16q24.3. Cancer Res 62: 4599-4604.
60. Ejeskar K, Krona C, Caren H, Zaibak F, Li L, et al. (2005) Introduction of in vitro transcribed ENO1 mRNA into neuroblastoma cells induces cell death. BMC Cancer 5: 161.

61. De Schrijver E, Brusselmans K, Heyns W, Verhoeven G, Swinnen JV (2003) RNA Interference-mediated Silencing of the Fatty Acid Synthase Gene Attenuates Growth and Induces Morphological Changes and Apoptosis of LNCaP Prostate Cancer Cells. Cancer Res 63: 3799-3804.
62. Menendez J, Colomer R, Lupu R (2004) Inhibition of tumor-associated fatty acid synthase activity enhances vinorelbine (Navelbine)-induced cytotoxicity and apoptotic cell death in human breast cancer cells. Oncology Reports 12: 411-422.
63. Logie A, Dunois-Larde C, Rosty C, Levrel O, Blanche M, et al. (2005) Activating mutations of the tyrosine kinase receptor FGFR3 are associated with benign skin tumors in mice and humans. Hum Mol Genet 14: 1153-1160.
64. van Oers JMM, Lurkin I, van Exsel AJA, Nijsen Y, van Rhijn BWG, et al. (2005) A Simple and Fast Method for the Simultaneous Detection of Nine Fibroblast Growth Factor Receptor 3 Mutations in Bladder Cancer and Voided Urine. Clin Cancer Res 11: 7743-7748.
65. Ruteshouser C, Hendrickson BW, Colella S, Krahe R, Pinto L, et al. (2005) Genome-wide loss of heterozygosity analysis of WT1-wild-type and WT1-mutant Wilms tumors. Genes, Chromosomes and Cancer 43: 172-180.
66. Satoh Y, Nakadate H, Nakagawachi T, Higashimoto K, Joh K, et al. (2006) Genetic and epigenetic alterations on the short arm of chromosome 11 are involved in a majority of sporadic Wilms' tumours. British Journal of Cancer 95: 541-547.
67. Perotti D, De Vecchi G, Testi MA, Lualdi E, Modena P, et al. (2004) Germline mutations of the POU6F2 gene in Wilms tumors with loss of heterozygosity on chromosome 7p14. Human Mutation 24: 400-407.
68. Rauscher F, Morris J, Tournay O, Cook D, Curran T (1990) Binding of the Wilms' tumor locus zinc finger protein to the EGR-1 consensus sequence. Science 250: 1259-1262.
69. Fraizer G, Wu Y, Hewitt S, Maity T, Ton C, et al. (1994) Transcriptional regulation of the human Wilms' tumor gene (WT1). Cell type-specific enhancer and promiscuous promoter. J Biol Chem 269: 8892-8900.
70. Hewitt SM, Fraizer GC, Wu Y-J, Rauscher FJ, III, Saunders GF (1996) Differential Function of Wilms' Tumor Gene WT1 Splice Isoforms in Transcriptional Regulation. J Biol Chem 271: 8588-8592.
71. Nakagama H, Heinrich G, Pelletier J, Housman D (1995) Sequence and structural requirements for high-affinity DNA binding by the WT1 gene product. Mol Cell Biol 15: 1489-1498.
72. van Helden J, Collado-Vides J (1998) Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies. Journal of Molecular Biology 281: 827-842.
73. van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Research 31: 3593-3596.
74. Pavesi G, Mauri G, Pesole G (2001) An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics 17: S207-214.
75. Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucl Acids Res 32: W199-203.

76. Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, et al. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucl Acids Res 34: W566-570.
77. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology 23: 137-144.
78. Vapnik V (1998) Statistical Learning Theory. Text:The Nature of Statistical Learning Theory.
79. Weston J, Elisseeff A, Bakir G, Sinz F, et al SPIDER: object oriented machine learning library: [http://www.kyb.tuebingen.mpg.de/bs/people/spider/].
80. Platt JC (1999) Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers, MIT Press.
81. Tan PN, Steinbach M, Kumar V (2006) Introduction to Data Mining; Harutunian K, editor: Pearson Addison Wesley.
82. Sholkopf B, Smola AJ (2002) Learning with Kernels. MIT Press.
83. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucl Acids Res 32: D493-496.
84. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucl Acids Res 31: 51-54.
85. Siepel A, Haussler D (2004) Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis. Journal of Computational Biology 11: 413-428.
86. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.
87. Dennis GJ, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4.
88. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, et al. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics 22: 637-640.
89. Kolchanov N, et al. (2002) Transcription Regulatory Regions Database (TRDD): its status in 2002. Nucl Acids Res 30: 312-317.
90. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl Acids Res 34: D108-110.
91. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucl Acids Res 34: D556-561.
92. Schmid CD, Perier R, Praz V, Bucher P (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucl Acids Res 34: D82-85.
93. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, et al. (2004) Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. Science 303: 1378-1381.

94. Zhang X, Odom DT, Koo S-H, Conkright MD, Canettieri G, et al. (2005) Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. PNAS 102: 4459-4464.
95. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. (2004) Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs. Cell 116: 499-509.
96. Kim J, Bhinge A, Morgan X, Iyer V (2005) Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. Nature Methods 2: 47-53.
97. Wei C-L, Wu Q, Vega VB, Chiu KP, Ng P, et al. (2006) A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome. Cell 124: 207-219.
98. Xuan Z, Zhao F, Wang J, Chen G, Zhang M (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biology 6: R72.