METHODS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Quantifying the Effect of Sequence Variation on Regulatory Interactions

Thomas Manke,[1*†] Matthias Heinig,[1,2†] and Martin Vingron[1]

[1]*Max Planck Institute for Molecular Genetics, Berlin, Germany;* [2]*Max Delbrück Centrum for Molecular Medicine, Berlin, Germany*

Communicated by A. Jamie Cuticchia

Received 14 October 2009; accepted revised manuscript 12 January 2010.
Published online 2 February 2010 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21209

ABSTRACT: The increasing amount of sequence data provides new opportunities and challenges to derive mechanistic models that can link sequence variations to phenotypic diversity. Here we introduce a new computational framework to suggest possible consequences of sequence variations on regulatory networks. Our method, called sTRAP (strap.molgen.mpg.de), analyses variations in the DNA sequence and predicts quantitative changes to the binding strength of any transcription factor for which there is a binding model. We have tested the method against a set of known associations between SNPs and their regulatory consequences. Our predictions are robust with respect to different parameters and model assumptions. Importantly we set an objective and quantifiable benchmark against which future improvements can be compared. Given the good performance of our method, we developed a publicly available tool that can serve as an important starting point for routine analysis of disease-associated sequence regions.
Hum Mutat 31:477–483, 2010. © 2010 Wiley-Liss, Inc.

KEY WORDS: regulatory SNP; transcription factor; protein–DNA interaction; genome-wide association studies; eQTL

Introduction

Genetic polymorphisms constitute the basis of phenotypic diversity. An increasing amount of sequence data points to substantial sequence variation among individual organisms, such as single nucleotide polymorphisms (SNPs) and structural variations. Ongoing genotyping and resequencing projects are likely to produce ever more data, that will challenge our molecular understanding about the functional consequences of such variations. The most studied form of sequence variations are SNPs which, in some cases, cause distinct phenotypes, disease, or increased disease susceptibility [Altshuler et al., 2008]. Occasionally, SNPs affect directly the function of a protein, as exemplified by sickle cell disease, which can be linked to a mutation in the coding region of α-globin [Ingram, 1956].

As the result of technological advances, there are now a number of systematic efforts underway to map human variations [Frazer et al., 2007] and the genetic basis for many other common diseases in genome-wide association studies [Burton et al., 2007; Kruglyak, 2008; Schadt, 2009]. Unsurprisingly, most variations have been observed in noncoding regions, where they might alter regulatory interactions. Their functional consequences, however, are more difficult to predict and validate, because the regulatory code is much more complex and flexible than the genetic code. For example, the misregulation of α-globin is known to cause α thalassemia; a reduction in functional hemoglobin [Higgs et al., 1989]. Only recent experiments have provided first insights into possible molecular mechanisms, namely, the creation of a novel Gata1 binding site and other hallmarks of regulatory control in the upstream region of α-globin [De Gobbi et al., 2006].

The regulatory effects of sequence variations can be measured systematically at the level of gene expression data. The transcript level of each individual gene is treated as a quantitative trait, and can subsequently be used for genetic mapping. This idea has given rise to eQTL studies as first proposed in Jansen and Nap [2001] and recently reviewed in Cookson et al. [2009] and Rockman [2008]. These studies have identified a large number of *cis*-regulated eQTL genes [Brem and Kruglyak, 2005; Brem et al., 2002; Hubner et al., 2005; Schadt et al., 2003], providing first links between sequence and function.

However, both GWAS and eQTL studies have some limitations. First, the SNPs found to be associated with a certain trait are not necessarily the causative variation, but rather provide a lead to a larger sequence region. To increase the resolution, additional time-consuming sequencing efforts have to be undertaken, which can only be done for a subset of candidate genes. This will generally yield additional SNPs that might be in linkage disequilibrium with the lead SNP. Even if a causative SNP has been identified, these studies only provide certain associations, but no hypothesis about the actual mechanisms involved.

In principle, computational studies can be used to prioritize SNPs and to generate hypotheses about the regulatory mechanisms involved. Some earlier work on regulatory SNPs considered binding sites of specific transcription factors (TFs) and studied their overlap with comprehensive collections of SNPs [Ameur et al., 2009; Chorley et al., 2008]. Other groups have aimed to generate larger collections of SNP–TF associations using binding site predictions [Kim et al., 2008; Ponomarenko et al., 2003; Stepanova et al., 2006]. Focusing on the functional role of SNPs, the computational approach has also been used to assess the overall correlation of functional SNPs with sequence features, such as predicted binding sites [GuhaThakurta et al., 2006]. In a similar spirit, Andersen et al. [2008] have attempted to utilize evolutionary sequence conservation to assess the functionality of noncoding sequence variations. These authors noted that binding site predictions do not improve the detection of functional SNPs. The main caveat for all these efforts is that experimental

†The first two authors contributed equally to this work.

*Correspondence to: Thomas Manke, Max Planck Institute for Molecular Genetics Computational Biology, Ihnestr. 73, Berlin, 14195 Germany. E-mail: manke@molgen.mpg.de

binding data is still sparse, whereas computational binding site predictions tend to be very unspecific and rely on an arbitrary threshold. Moreover, they do not address the more direct question that is often asked by geneticists: Given a specific SNP, the binding of *which* transcription factor is most affected by the sequence variation? The answer to such a question is complicated by the large number of known TFs and the large rate of false positives, which is commonly associated with threshold-based binding site predictions.

Here we do not aim to predict functional or causative SNPs, but we assume that functionality has already been established by other means, such as promoter assays. Instead, we aim to identify those transcription factors, whose predicted binding affinity is most strongly affected by a given sequence variation. This goal is more challenging than previous efforts as its success hinges on the correct identification of a transcription factor from a large set of many possible ones. Our approach makes predictions about individual SNPs rather than the overall sequence properties of classes of SNPs. We believe that this specificity brings our efforts in line with what geneticists would want to learn about selected SNPs.

To this end we extended an earlier framework for transcription factor affinity predictions, TRAP [Roider et al., 2007], and combined it with a statistical approach to normalize the binding affinities for different transcription factors. Our new method, called sTRAP, can predict sequence-induced changes in the binding affinity of a transcription factor. Importantly, and owing to our statistical framework, we are able to compare these changes for a comprehensive set of transcription factors. We validated the approach against a set of known SNP–TF associations and find that sTRAP correctly predicts a large fraction of those associations at a small rate of false predictions.

Finally, we provide the software and a simple Web interface (http://strap.molgen.mpg.de) that calculates affinity changes for any user-specified pair of sequences. This will help geneticists to rapidly assess likely effects of sequence variation on regulatory interactions.

Data and Methods

An overview of our method is given in Figure 1. In the following we provide the details for the individual steps.

SNP Data

Although there is massive data on sequence variation from large-scale mapping efforts [Frazer et al., 2007; Sherry et al., 2001], the regulatory potential of SNPs is badly documented and only occasionally reported. Here we study 20 known associations of regulatory SNPs with transcription factors, which were collected by Andersen et al. [2008]. These comprise SNPs that are naturally occurring or were generated by targeted mutagenesis. Moreover, for those SNPs the binding of selected transcription factors was shown to be affected. For each of these SNPs we retrieve a flanking region of 60, 100, 500, and 1,000 bp.

Binding Models

An increasing number of genome-wide in vivo and in vitro studies of protein–DNA interactions [Harbison et al., 2004; Mukherjee et al., 2004] aim to provide a comprehensive compendium of binding models for transcription factors under different conditions and in various species. For the purpose of this work we use a preliminary compendium of binding models, as available from the TRANSFAC database, version 12.1 [Matys et al., 2003]. We use information on 202 vertebrate transcription factors, which is encoded by 554 position specific weight matrices. In earlier work we showed how this information can be used to predict the sequence-specific binding affinities of a transcription factor using a biophysical framework [Roider et al., 2007]. For the local binding affinities at sequence position l we use

$$a_1(R_0, \lambda) = \frac{R_0 e^{-E_L(\lambda)}}{1 + R_0 e^{-E_l(\lambda)}} \quad (1)$$

where $R_0(W) = 0.6W - 6$ and $\lambda = 0.7$ are two parameters that were fitted in Roider et al. [2007], and W denotes the width of the motif. The local affinity predictions can be utilized in two different ways.

First we consider, for each SNP and every motif matrix, the W pairs of local affinities that are changed when comparing the reference sequence with its variation. This is illustrated on the left side of Figure 2, where the matrix model for Gata1 (width $W = 13$ bp) is scanned over a regulatory SNP in the α-globin promoter, which causes α thalassemia. Such a scan results in 13 pairs of predicted affinities that differ between the reference sequence and its variation. The alignment shown in Figure 2 is determined by the shift with the most deleterious effect on the binding affinity of Gata1.

Such a local comparison may suggest large effects on the predicted binding affinity, even if the flanking sequence contains additional binding sites that may buffer the effect of a sequence variation. Therefore, we also employed a second strategy in which we calculate the overall affinity, A, of a transcription factor to a longer sequence region of size L (i.e., SNP+flanking region). This can be obtained by summing the local binding affinities, a_l, over all accessible sites. This second approach has the added benefit that, for each SNP and every transcription factor, we only need to compare the two overall affinities from the reference sequence and its variation.

Distribution of Binding Affinities

In general, the predicted affinities for different transcription factors are not comparable because they can have very different specificities. Hence, it is important to also model, for each factor, the distribution of binding affinities over genomic sequences. With the help of such a distribution one can normalize the affinities; in other words, one can assign a p-value to each affinity. Here we follow the statistical framework developed in Manke et al. [2008]. There we showed that a simple parameterization effectively describes the distribution of affinities, A, for most transcription factors:

$$\log A \sim P(x|a,b,c) = \exp\left(-\left[1 + a\frac{x-c}{b}\right]^{-1/a}\right) \quad (2)$$

The three parameters of this Generalized Extreme Value (GEV) distribution also depend on the length of the sequence region, and their value was determined in Manke et al. [2008] for all 554 binding models of vertebrate transcription factors from TRANSFAC [Matys et al., 2003]. Although the parameters of the distribution in Equation (2) are based on a sequence model of known human promoters, in some situations it might be preferable to use a different background model that captures the specific sequence properties in the vicinity of the SNP more accurately. In those instances our implementation offers the possibility to reestimate the empirical distribution of binding affinities (and their p-values). Given a sufficiently large sequence region around the SNP, we estimate the p-values based on the rank

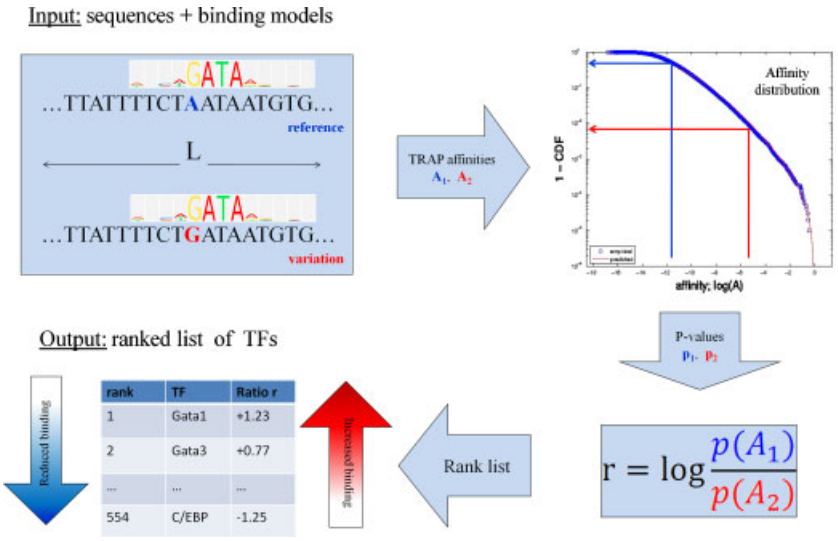

Figure 1. Overview of the sTRAP method. Using sequence data and a comprehensive set of transcription factor binding models as input, we predict the binding affinities of all transcription factors (TF) to the reference sequence and its variation. These affinities are then normalized with the help of the affinity distribution from Manke et al. [2008]. This normalization step ensures that affinities and affinity changes are comparable for different factors. The log-ratio of the two *p*-values is used to rank all TFs according to their change in binding affinity.

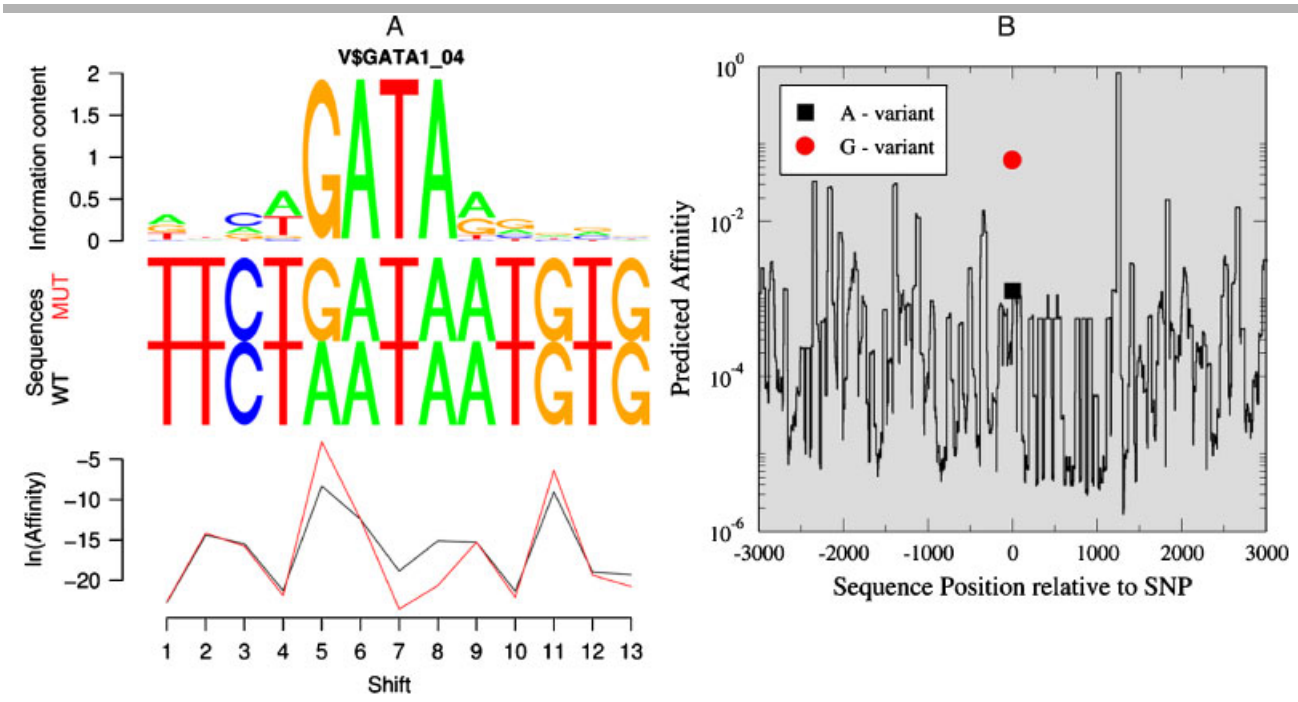

Figure 2. The left figure (**A**) illustrates how an SNP may cause changes to the local binding affinity of a transcription factor. This example is for a regulatory SNP in the α-globin promoter region and transcription factor Gata1, whose sequence logo is shown at the top. Shifting this particular motif (width $W = 13$ bp) across the SNP-region, gives rise to 13 differential pairs of binding affinities. Those are plotted at the bottom on a natural logarithmic scale. The alignment shown on the top corresponds to a shift of 5 bp with respect to the SNP. The right figure (**B**) shows a regional approach, where the affinity is calculated for a larger window ($L = 60$ bp) which was shifted across a $\pm$3,000-bp region around the SNP. The affinity of Gata1 is strongly affected as evident from a large shift at the position of the SNP. Notice that the affinities shown are not yet normalized, but the variance observed in the surrounding sequence helps to quantify this change. For this article we also utilized a parameterization of the affinity distribution from [Manke et al., 2008].

statistics obtained from all affinities in the neighbourhood of the SNP (see right-hand side of Fig. 2 for an example). Here we implemented this option to test the robustness of our predictions, but in general, the local background model may be better suited for species with GC-content very different from humans.

Prediction of Differential Binding Affinities

Here we describe a simple ranking scheme to compare different transcription factors with respect to the changes induced by sequence variations. For a given transcription factor, *X*, and a pair

of sequences, S1 and S2, we calculate the *p*-values, $p_X(S1)$ and $p_X(S2)$, as described in the previous section. In the following we consider S1 as the reference sequence and S2 its variation and define a log-ratio

$$r_X = \log_{10}\left(\frac{p_X(S1)}{p_X(S2)}\right) \quad (3)$$

Large positive values denote cases where the factor X increases its binding affinity, whereas for large negative values the binding affinity is decreased with respect to the reference sequence $S1$. Importantly, the ratios for different transcription factors, X, are directly comparable, because they are based on *p*-values, rather than absolute affinities. Conceptually we aim to detect differential binding affinities that are changed in one or the other sequence, and the score of Equation (3) provides a corresponding quantitative ranking. Notice that if $p(S1)/p(S2)$ follows a uniform ratio distribution, then $r = 1$ corresponds to $p = 0.05$. In this work, however, we focus on the ranking of scores r from different transcription factors and assess the performance of our predictions based on such a ranking.

Alternatively, one may also want to assign a high score if $S1$ or $S2$ or both show strong binding to a transcription factor. In this case one could replace Equation (3) by the minimum of $p_X(S1)$ and $p_X(S2)$

$$r_X = \min\{p_X(S1), p_X(S2)\} \quad (4)$$

The biological rationale for the latter score is to detect strong binding sites that might be affected by the SNP, even if the resolution of simple binding models cannot detect differential affinities. Such effects may be caused by synergistic interactions with other transcription factors, and they are beyond our current framework. Ultimately one would require a model to predict changes in gene expression that could be considerable, even for minor changes in binding affinity. In the absence of any more sophisticated model, our heuristic scores aim for simplicity and avoid overfitting.

Results

sTRAP: A Framework to Rank Affinity Changes

In earlier work we had proposed a quantitative framework for the computational prediction of transcription factor binding sites and determined a simple parameterization for binding affinities [Roider et al., 2007] and their distribution [Manke et al., 2008]. Similar approaches have been studied by a number of other groups [Djordjevic et al., 2003; Foat et al., 2006; Segal et al., 2008; Tanay, 2006].

Here we extend this approach to annotate pairs of sequences with respect to changes in their affinity. Just as we had previously ranked transcription factors with respect to their affinity for a single sequence region, we now rank transcription factors with respect to affinity changes induced by sequence changes. We think of these sequence pairs as being derived from a reference sequence and a possible mutation. In particular, we are interested in scoring the abolishment or the creation of a binding site in one or the other sequence. This notion is made quantitative by the log-ratio of *p*-values (Eq. 3) as introduced in the Methods section. In general, we calculate the binding affinity of transcription factors to longer sequence regions of length L, but a special case is $L = W$ (local approach), where the length is equal to the width, W, of the binding motif. The output of sTRAP consists of a list of transcription factors, which are ranked according to changes induced by the SNP. We reasoned that top-ranking transcription factors are most likely to be affected and provide natural candidates for subsequent analysis. A software package and a Web-based interface is provided that permits an identical analysis for any given sequence variation (http://strap.molgen.mpg.de).

sTRAP Predicts Many Known SNP–TF Associations

For a first test of our method we applied sTRAP to a list of known regulatory sequence variations from humans and their associated transcription factors as collected by Andersen et al. [2008]. The set of SNPs is based on extensive literature search, and includes naturally occurring SNP as well as those generated in mutagenesis experiments; see Method section for more details. We have also verified that the set of transcription factors known to be effected by those SNPs includes only factors for which at least one motif matrix can be found in TRANSFAC.

For each SNP, our method predicts a ranked list of TFs that is compared to the list of TFs known to be affected by the variation. As an input list we took 554 vertebrate transcription factor motifs from the TRANSFAC database [Matys et al., 2003]. A good method would predict known TFs at the top of the list. Indeed, in Figure 3 it is shown that a large fraction of known TFs appear top when ranked according to Equation (3). We also compared this to random expectations where all TF motifs are assigned a random rank. The deviation from random expectations is clearly significant. Notice that the slight deviation of the random set from uniformity apparent in Figure 3 is due to the fact that some TFs have multiple motifs and we always take the best rank.

sTRAP Predictions Are Specific

Although it is encouraging to see that many known associations of regulatory SNPs with their respective transcription factors can be detected, we now investigate more carefully the rate of true positives as a function of false positives. To this end we introduce a variable threshold, θ which defines when the binding of a transcription factor is said to be affected by the SNP, $|r_X| > \theta$. For every threshold, our predictions can be compared to the reference set of known TF–SNP associations. Because only individual transcription factors have been tested experimentally for any given SNP, there is a generic lack of information regarding the effect of the SNP on other factors. For our purposes we make the conservative assumption that all other transcription factors are not affected by the SNP and therefore count predicted associations of experimentally untested factors as false positives. This approach is likely to inflate the estimated rate of false positives. In Figure 4 we plot the rate of true positives against the rate of false positives, when the threshold θ is changed. This curve (the receiver operator characteristic) can be used to compare the performance for different parameterizations and competing models. The performance of our method can be quantified by the area under the curve and it is significantly larger than 0.5, which would be expected from random assignments. This performance is also much better than what can be achieved using simple score differences (area ≈ 0.55). For this comparison we utilized the scoring system developed in Rahmann et al. [2003]. Score differences have previously been used in Andersen et al. [2008] and GuhaThakurta et al. [2006], but with no beneficial effect on their ability to predict functional SNPs. Although Equations (3) and (4) are based on *p*-values, it is important to notice that the above evaluation merely utilizes a ranking of those heuristic scores and does not provide any significance level. For all practical

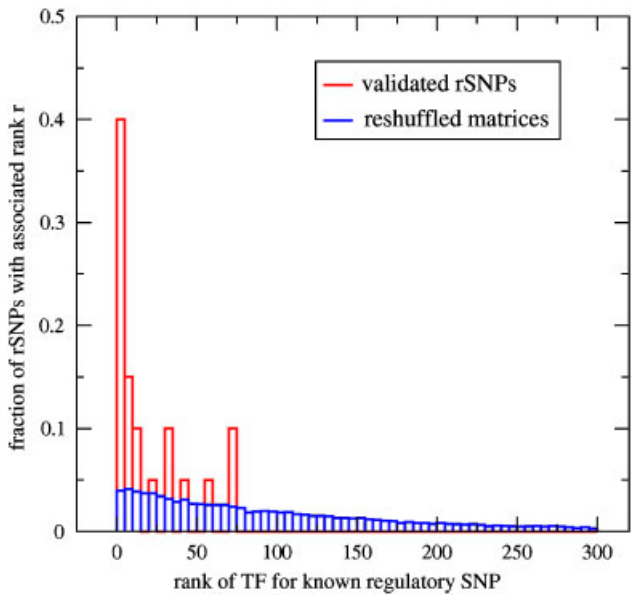

Figure 3. The sTRAP approach successfully predicts many known TF–SNP associations. For each known regulatory SNP and its associated transcription factor we record the rank of the corresponding matrix according to our scoring scheme. This figure shows that many known associations get a significantly high rank according to our scoring scheme (Eq. 3). In blue we show the same histogram for a set of reshuffled matrices. The slight increase toward higher ranks is due to multiple matrices assigned to some factors.

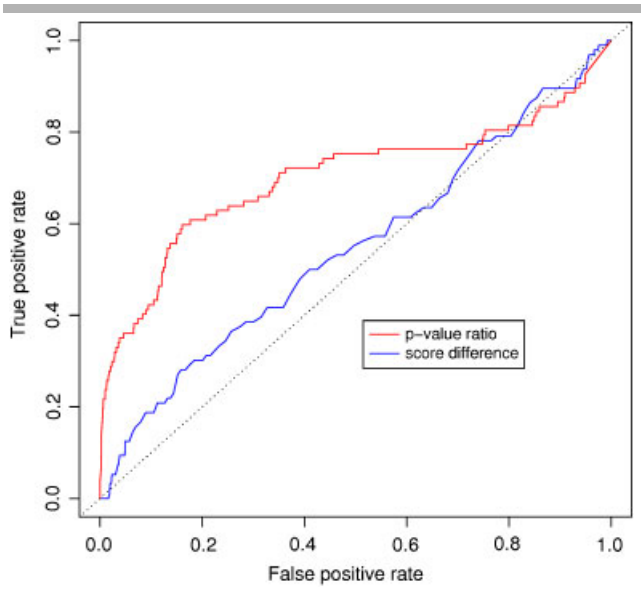

Figure 4. Here we show that the sTRAP approach is specific. The red curve is obtained by varying the threshold r_X in Equation (3), where the *p*-values are estimated from the distribution of Equation (2). A sequence length of $L = 61$ bp around the SNPs was used as in Andersen et al. [2008]. At a rate of 10% false positives we recover 50% of all true positives. The area under the curve is 0.70. For comparison, we also included as blue line the results from a traditional analysis, in which we utilized score differences to rank the transcription factors. Here we used the scoring system introduced in Rahmann et al. [2003]. This approach has a much reduced performance as quantified by the area of 0.55, which is close to random expectation (dotted line).

purposes one could select a specific point from the curve in Figure 4. For example, at a false positive rate of 10% we recover more than half of the known TF–SNP associations, which corresponds to a score threshold of $\theta = 0.21$.

To assess the robustness of our method, we have compared the classification performance of different parameter settings and classifiers. As classifiers we have used different thresholds on the absolute log ratio or the minimum of the two *p*-values, corresponding to the two scenarios where SNPs affect the binding strongest or fall in strong binding sites without drastic changes of affinity. We evaluated different lengths of the sequence regions used to compute the binding affinities, and different background models to determine the distribution of affinities. For the length of the sequence region, L, we used $L \in \{61, 100, 500, 1000\}$ bp to account for the effect of multiple binding sites, or $L = W$, where W is the motif width. The latter corresponds to the local approach described in Section 2. Furthermore, we also evaluated the classification performance for two classes of background models: (1) the GEV parameterization and (2) empirical *p*-values derived from the neighbouring sequence around the SNP of interest. In the latter case we used background sequences of 100,000 bp around the SNP to estimate the *p*-values with sufficient accuracy. Our results are summarized in Table 1, which shows that the method is robust with respect to the choice of classifiers and the length of the sequence region π.

Discussion

Current studies from molecular medicine result in many SNPs that are found to be associated with certain diseases, some of which might be causative. However, there is a scarcity of follow-up mechanistic studies to rationalize those predictions in molecular terms.

In this work we developed a novel method to predict *which* transcription factor is likely to be affected by a sequence variation. Our new approach (sTRAP) is based on an earlier biophysical framework (TRAP), and predicts sequence-induced changes in binding affinities. This quantitative approach relies on a proper normalization of binding affinities and permits a robust ranking of the most affected transcription factors. There are a number of advantages of the sTRAP method compared to the traditional annotation of binding sites. First, it allows for the prediction of the most affected transcription factors in a quantitative and threshold-free manner. Other works have utilized a threshold and

Table 1. sTRAP Performs Robustly for Different Choices of Parameters and Distributions

	Ratio		Min	
Length	GEV	Regional	GEV	Regional
W	0.743	–	0.833	–
61	0.702	0.69	0.852	0.87
100	0.713	0.71	0.817	0.85
500	0.736	0.67	0.769	0.81
1,000	0.749	0.62	0.767	0.78

Here we summarize the area under the ROC curve (AUC) as a performance measure of our method. As described in the main text we utilized different length of the flanking region. For the *local* method, the length was set to the variable motif width, W, of the different transcription factors. The specific choice of L = 61 bp was motivated by our analysis of the data from Andersen et al. [2008]. We provide the AUC for two alternative scores defined by equations 3 and 4, respectively. For each method and score we also compared two different background models: the GEV-model from a parametrization of the affinity distributions in human promoters Manke et al. [2008] and a regional background model obtained from the affinity distribution in 100,000 bp around the SNP.

focused on the overlap of predicted binding sites with SNPs, but without ranking the difference in binding strength [Ameur et al., 2009; Kim et al., 2008; Ponomarenko et al., 2003]. In GuhaThakurta et al. [2006] the authors had also used quantitative scores, but without effect on their conclusion regarding the overall correlation of predicted binding sites and functional classes of SNPs. Second, although other works [Ameur et al., 2009; Kim et al., 2008] have annotated only the reference sequence with binding sites, we always consider both the reference sequence and its variation. Therefore, we can account for both a decrease and a possible increase of binding affinity, corresponding to the abolishment or creation of new sites. Third, our method can be adjusted to assess the binding capacity of longer sequence fragments. In this way we can account for buffering effects from neighboring sites that could maintain a high overall binding affinity, even when the SNP in question is disruptive. Fourth, we have introduced a performance test and measure, against which future developments and possible improvements can be tested.

We have shown that the performance of sTRAP is robust against the choices of parameters and different background models used to normalize predicted binding affinities. Different scores were chosen based on different biological premises, and the sequence models were chosen to capture some of the local sequence variability more or less accurately. Alhough there is no theoretical basis for choosing one or the other background model, it is encouraging that all approaches we have tested have comparable performance (Table 1).

We have implemented our method as part of an R-package called tRap and provide a public Web interface, which allows the user to submit pairs of sequences corresponding to the two SNP alleles for analysis (http://strap.molgen.mpg.de). This makes sTRAP a valuable tool for the exploratory data analysis elucidating the mechanisms and possible consequences of regulatory SNPs. In addition to its importance for understanding genetic diseases, our approach also provides clear suggestions for transcription factors that affect gene expression in a human specific manner. Other possible applications include the study of eQTL and species-specific sequence variation. Although our method does not directly assess the functionality of a SNP, it may help to prioritize a list of candidate SNPs if prior knowledge about the role of the implied transcription factors is available.

As with all sequence-based methods, our approach assumes that the binding dynamics of transcription factors to the sequence is rapid and that the equilibrium binding strength is the key parameter to control gene expression. Currently we employed a large but limited set of transcription factor motifs from the TRANSFAC database. There are still many transcription factors with unknown motifs that are beyond our model, and that will give rise to false negatives. Recent experimental advances and high-throughput data, such a protein-binding arrays [Badis et al., 2009; Berger et al., 2006], are likely to alleviate this limitation in the near future and permit an even more comprehensive assessment of the effect of sequence variations. Large-scale binding screens also raise the hope that we will soon be able to model small differences in the binding preferences of structurally similar proteins, as well as binding differences of the same transcription factor in variable cellular conditions. Currently such differences are not easily resolved and contribute to limitations of our approach.

Although sTRAP predictions are generically powerful as evidenced by the ROC curve of Figure 4, it will be a challenging task to validate individual predictions and integrate them into a molecular understanding of signaling and gene expression. The overarching goal of this project is to render computational binding predictions more quantitative. Clearly, much work remains to be done. However, there is hope that the theoretical developments will increasingly be driven by technological advances and quantitative data [Mittler et al., 2009], against which the models can be optimized.

Acknowledgments

We would like to thank Andre Franke, Peter Robinson, Norbert Hübner. Financial support from the German National Genome Research Network (NGFN-Plus #01GS0815) is gratefully acknowledged.

References

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. Science 322:881–888.

Ameur A, Rada-Iglesias A, Komorowski J, Wadelius C. 2009. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. Nucleic Acids Res 37:e85.

Andersen MC, Engstrom PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J. 2008. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol 4:e5.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML. 2009. Diversity and complexity in DNA recognition by transcription factors. Science 324:1720–1723.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep3rd PW, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24:1429–1435.

Brem RB, Kruglyak L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci USA 102:1572–1577.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans DM, Leung HT, Marchini JL, Morris AP, Spencer CC, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, StClair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskivina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop TD, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Matthew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop MG, Connell J, Dominiczak A, Marcano CA, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hilder SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DP, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JR, Shields BM, Weedon MN, Hattersley AT, Hitman GA, Walker M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Hill AV, Bradbury LA, Farrar C, Pointon JJ, Wordsworth P, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Gough SC, Seal S; Breast Cancer Susceptibility Collaboration (UK), Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Newport M, Sirugo G, Rockett KA, Bumpstead SJ, Chaney A, Downes K, Ghori MJ, Gwilliam R, Hunt SE, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Widden C, Withers D, Cardin NJ, Davison D, Ferreira T, Pereira-Gale J, Hallgrimsdo'ttir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Brown MA, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AV, Parkes M, Pembrey M, Stratton MR, Mitchell SL, Newby PR, Brand OJ, Carr-Smith J, Pearce SH, McGinnis R, Keniry A, Deloukas P, Reveille JD, Zhou X, Sims AM, Dowling A, Taylor J, Doan T, Davis JC, Savage L, Ward MM, Learch TL, Weisman MH, Brown M. 2007. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39:1329–1337.

Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659:147–157.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. Nat Rev Genet 10:184–194.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR . 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 312:1215–1217.

Djordjevic M, Sengupta AM, Shraiman BI. 2003. A biophysical approach to transcription factor binding site discovery. Genome Res 13:2381–2390.

Foat BC, Morozov AV, Bussemaker HJ. 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics 22:e141–e149.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

GuhaThakurta D, Xie T, Anand M, Edwards SW, Li G, Wang SS, Schadt EE. 2006. Cis-regulatory variations: a study of SNPs around genes showing cis-linkage in segregating mouse populations. BMC Genomics 7:235.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104.

Higgs DR, Vickers MA, Wilkie AO, Pretorius IM, Jarman AP, Weatherall DJ. 1989. A review of the molecular genetics of the human alpha-globin gene cluster. Blood 73:1081–1104.

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A. 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 37:243–253.

Ingram VM. 1956. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 178:792–794.

Jansen RC, Nap JP. 2001. Genetical genomics: the added value from segregation. Trends Genet 17:388–391.

Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. 2008. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinformatics 9(Suppl 1):S2.

Kruglyak L. 2008. The road to genome-wide association studies. Nat Rev Genet 9:314–318.

Manke T, Roider HG, Vingron M. 2008. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. PLoS Comput Biol 4: e1000039.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV and others. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374–378.

Mittler G, Butter F, Mann M. 2009. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Res 19:284–293.

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet 36:1331–1339.

Ponomarenko JV, Merkulova TI, Orlova GV, Fokin ON, Gorshkova EV, Frolov AS, Valuev VP, Ponomarenko MP. 2003. rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. Nucleic Acids Res 31:118–121.

Rahmann S, Muller T, Vingron M. 2003. On the power of profiles for transcription factor binding site detection. Stat Appl Genet Mol Biol 2:Article7.

Rockman MV. 2008. Reverse engineering the genotype–phenotype map with natural genetic variation. Nature 456:738–744.

Roider HG, Kanhere A, Manke T, Vingron M. 2007. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics 23:134–141.

Schadt EE. 2009. Molecular networks as sensors and drivers of common human diseases. Nature 461:218–223.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G and others. 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. Nature 451:535–540.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.

Stepanova M, Tiazhelova T, Skoblov M, Baranova A. 2006. Potential regulatory SNPs in promoters of human genes: a systematic approach. Mol Cell Probes 20: 348–358.

Tanay A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res 16:962–972.