PERFECTOS-APE
Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation

Ilya E. Vorontsov[1], Ivan V. Kulakovskiy[1,2], Grigory Khimulya[1], Daria D. Nikolaeva[3]
and Vsevolod J. Makeev[1,2,4]

[1]*Department of Computational Systems Biology, Vavilov Institute of General Genetics, Gubkina str. 3, Moscow, Russia*
[2]*Laboratory of Bioinformatics and Systems Biology, Engelhardt Institute of Molecular Biology,*
Vavilova str. 32, Moscow, Russia
[3]*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*
[4]*Department of Medical and Biological Physics, Moscow Institute of Physics and Technology, Moscow Region, Russia*

Keywords: Single Nucleotide Polymorphism, SNP, Single Nucleotide Variant, SNV, P-value, Transcription Factor Binding Site, TFBS, Position Weight Matrix, PWM, PSSM, Transcriptional Regulation.

Abstract: Single nucleotide polymorphisms (SNPs) and variants (SNVs) are often found in regulatory regions of human genome. Nucleotide substitutions in promoter and enhancer regions may affect transcription factor (TF) binding and alter gene expression regulation. Nowadays binding patterns are known for hundreds of human TFs. Thus one can assess possible functional effects of allele variations or mutations in TF binding sites using sequence analysis.

We present PERFECTOS-APE, the software to PrEdict Regulatory Functional Effect of SNPs by Approximate P-value Estimation. Using a predefined collection of position weight matrices (PWMs) representing TF binding patterns, PERFECTOS-APE identifies transcription factors whose binding sites can be significantly affected by given nucleotide substitutions. PERFECTOS-APE supports both classic PWMs under the position independency assumption, and dinucleotide PWMs accounting for the dinucleotide composition and correlations between nucleotides in adjacent positions within binding sites.

PERFECTOS-APE uses dynamic programming to calculate PWM score distribution and convert the scores to P-values with an optional binary search mode using a precomputed P-value list to speed-up the computations. Software is written in Java and is freely available as standalone program and online tool: http://opera.autosome.ru/perfectosape/.

We have tested our algorithm on several disease associated SNVs as well as on a set of cancer somatic mutations occurring in intronic regions of the human genome.

1 INTRODUCTION

Single nucleotide variants (SNVs) are the most studied variations of the human genome. Modern genome-wide association studies link different SNV alleles with different phenotypes, including disease susceptibility. High-throughput technologies become cheaper and push forward sequencing of personal genomes and detection of individual genome variants. However, proper interpretation of the sequencing data remains a challenge. Most of the detected SNVs are located outside of protein coding regions including a special class of SNVs found in gene regulatory regions. Among those an important subclass is formed by promoter and enhancer SNVs, which do not alter protein sequence or structure but possibly affect gene expression. Such SNVs may affect transcription through alterations in transcription factor binding sites.

During the past 10 years a number of tools (Macintyre et al., 2010; Manke et al., 2010; Barenboim and Manke, 2013; Riva, 2012; Teng et al., 2012; Andersen et al., 2008; Khurana et al., 2013) were developed for computational analysis of regulatory SNVs. Basic algorithms were simply predicting TFBS overlapping an SNV position (Ponomarenko et al., 2001). More sophisticated tools compare predicted affinitites of TFs binding to binding sites for different homolo-

E. Vorontsov I., V. Kulakovskiy I., Khimulya G., D. Nikolaeva D. and J. Makeev V..
PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation.
DOI: 10.5220/0005189301020108
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 102-108
ISBN: 978-989-758-070-3

gous alleles (Manke et al., 2010) and estimate statistical significance of the difference (Macintyre et al., 2010). All existing tools use the well-studied but basic model of transcription factor binding sites, the position weight matrix. Nowadays high-throughput data such as ChIP-Seq allows producing advanced models (Kulakovskiy et al., 2013a; Mathelier and Wasserman, 2013), since the volume of data makes it possible to train models with more parameters without the risk of overfitting (Levitsky et al., 2014; Mathelier and Wasserman, 2013). Here we present a novel software, PERFECTOS-APE, to predict how different alleles of SNVs or SNPs may alter affinity of transcription factor binding sites modelled by basic and advanced approaches.

2 METHODS

The core idea of our method is similar to that of (Macintyre et al., 2010; Manke et al., 2010): the algorithm estimates the statistical significance (the P-value) of predicted TF binding sites overlapping an SNV. Then it checks if TFBS binding P-values calculated for different homologous alleles differ enough. Extremely small or large ratios of P-values indicate the cases where the binding site exists only for one of the two alleles.

2.1 TFBS Models and P-value Estimation

We use basic and dinucleotide position weight matrices (PWMs) as TFBS models. PWM quantitatively describes which nucleotides are preferred at which position. Classic PWM M is a matrix $4 \times k$ which defines a score function on k-mers $\alpha_1 \dots \alpha_k$ upon nucleotide alphabet $V = \{A, C, G, T\}$:

$$score(M, \alpha_1 \dots \alpha_k) = \sum_{i=1}^{k} M(\alpha_i, i) \qquad (1)$$

Similarly to PWM, the dinucleotide position weight matrix (diPWM) D is a matrix $16 \times (k-1)$ where each element provides score for a pair of consequent nucleotides, where adjacent nucleotide pairs overlap:

$$score(D, \alpha_1 \dots \alpha_k) = \sum_{i=1 \dots k-1} D(\alpha_i \alpha_{i+1}, i) \qquad (2)$$

Dinucleotide PWM (diPWM) takes into account dependent contributions of adjacent nucleotides.

Scoring function with a score threshold defines the set of recognized words scoring no less than the threshold. Higher scores correspond to better TF-DNA recognition, but it's not trivial to define a unified scale for different PWMs. To this end a P-value of a given word for a given PWM is defined as the probability that a random word scores not less than the given threshold. In other words, the P-value corresponds to the area under the right tail of the PWM score distribution. For example, if a random model generates any word with equal probability, than the P-value is the normalized cumulative count of words with scores passing the threshold.

Let's assume the words are generated by an *i.i.d.* random model with p_α frequencies of individual nucleotides for a PWM model or by a Markov(1) model for a diPWM.

Given a PWM score threshold t, P-value for this PWM (denoted M) is:

$$pvalue = \sum_{\substack{w \in V^k \\ score(M,w) \geq t}} P(w) \qquad (3)$$

Here $P(w)$ is the probability of a word w to be generated by the background model (either *i.i.d.* or Markov(1)).

To convert PWM scores to P-values we utilize a simplified dynamic programming approach (ScoreDistribution algorithm) originally presented by H.Touzet and J.Varré (Touzet et al., 2007).

The idea is to discretize PWM elements and produce overall score distribution in a form of a hash (unlike the original approach here we use a predefined discretization level). This allows finding approximate P-value for a given score in reasonable time.

Each key in the hash is a score and the corresponding stored value is a probability to obtain this score. This hash can be constructed through recalculation of the score distribution gradually increasing the length of words:

$$H_0(S) = [S = 0] \cdot 1 \qquad (4)$$

$$H_{l+1}(S') = \sum_{\alpha \in V} \sum_{S: M(\alpha, l+1) + S = S'} H_l(S) p_\alpha \qquad (5)$$

For a dinucleotide position weight matrix (diPWM) the score distribution can be obtained by a similar method. For a dinucleotide model D (score, last-letter) pairs can be used as keys of the hash:

$$H_1(S, \alpha) = [S = 0] \cdot \sum_{\gamma \in V} p_{\gamma\alpha} \qquad (6)$$

$$H_{l+1}(S', \alpha_{l+1}) = \sum_{\alpha_l \in V} \sum_{S: D(\alpha_l \alpha_{l+1}, l) + S = S'} H_l(S, \alpha_l) p_{\alpha_{l+1}|\alpha_l} \qquad (7)$$

Here $p_{\alpha\beta}$ is the background probability of the dinucleotide $\alpha\beta$ and $p_{\beta|\alpha}$ is the background probability of the nucleotide β under condition that the previous letter was α:

$$p_{\beta|\alpha} = \frac{p_{\alpha\beta}}{\sum_{\gamma} p_{\alpha\gamma}} \tag{8}$$

At the last step score distributions of words ending with different nucleotides are summed to obtain the final score distribution:

$$H_k(S) = \sum_{\alpha \in V} H_k(S, \alpha) \tag{9}$$

2.2 Speeding-up P-value Estimation

In practice it is convenient to analyze multiple SNVs. Hundreds of thousands or even millions candidate variants can be subjected to analysis thanks to the personal genomics data.

In this case P-value estimation can be accelerated using precomputed score distributions for each TFBS model in a given collection. The PWM score distribution is discrete; thus, it's possible to store the P-value for each achievable PWM score. The P-value is a monotonically decreasing function depending on scores. Thus the list of (score, P-value) pairs can be sorted with respect to both the P-value and the score. Given the word score t, *binary search* can be utilized to find a pair of consequent scores t_1, t_2 such that $t_1 \leq t \leq t_2$ and thus the P-value p associated with the score t lies in range $p_1 \geq p \geq p_2$.

A value between p_1 and p_2 can be taken as an estimate of the true P-value; we use the geometric mean $p \approx \sqrt{p_1 p_2}$.

Since in practice we don't need a precise P-value, but a reasonable approximation, we select admissible P-value relative error δ to further improve computation time. From a given score distribution we take not the whole list of (score, P-value) pairs but only a subset such that P-values differ in $(1+\delta)$ times: $p_0, p_0/(1+\delta), p_0/(1+\delta)^2, \ldots$ for any p_0. With the help of this list we can estimate P-value with a given relative error δ. We use the fixed discretization level for PWMs at the distribution estimation step. We call the whole procedure Approximate P-value Estimation (APE).

To measure possible functional effect of allele variants we compute fold change of P-values corresponding to the binding sites with these allele variants (see the next section for details). In turn, the fold change relative error will be not greater than $\frac{1+\delta}{1-\delta}$.

2.3 Predicting Functional Effect of SNVs

To predict possible functional effect of a given SNV, PERFECTOS-APE (1) scans the region overlapping SNV position with (di)PWM model predicting binding sites for given nucleotide variants; (2) selects the best (di)PWM prediction for each allele; (3) computes P-values for the binding sites detected for each allele variant. If any of P-values is small enough (i.e. there is an allele corresponding to the putative binding site) and P-values differ significantly (i.e. the ratio passes a fold change threshold t_{FC}) we state that the SNV may play a regulatory role through disruption or emergence of the transcription factor binding site.

3 ALGORITHM COMPLEXITY

Given a motif of length k, a scanned region around SNV should contain no less than $2k-1$ nucleotides (the SNV position with the flanking genomic regions). The first stage of the algorithm calculates scores for each site position in both strand orientations relative to the given DNA sequence. It takes $O(k^2)$ operations to compute the scores ($2k$ score estimations with each score estimation of $O(k)$ complexity).

The second stage computes P-values either by calculating score distribution using dynamic programming or via binary search in a precomputed list.

Dynamic programming involves k steps. At each step the algorithm updates the hash H of the score distribution (score to probability mapping). For a PWM each update has $O(|H||V|)$ complexity. For a diPWM each update consists of $O(|H||V|^2)$ operations and $k-1$ updates are to be done. Here $|H|$ is a number of elements in the hash and $|V|$ is an alphabet size which is equal to 4. Hash size can be roughly estimated as $|H| \leq (max\_score - min\_score) \cdot discretization\_rate$ for a motif model discretized as:

$$M_{discretized}(\alpha, i) = \lceil M(\alpha, i) \cdot discretization\_rate \rceil \tag{10}$$

Binary search has only $O(\log_2 |S|)$ complexity where $|S|$ is the size of the (score; P-value) list. By default we calculate this list sampling the score distributions in several points such that P-value differs by no more than a given relative error δ for adjacent pairs. The list of P-values is bounded by the minimal achievable P-value $minAchievablePvalue$ and 1. Thus

$$|S| = -\frac{\ln(minAchievablePvalue)}{\ln(1+precision)} \tag{11}$$

It's also possible to make threshold to P-value conversion in O(1) by expanding a P-value list into an array indexed by discretized thresholds. Such an algorithm consumes a bit more memory and works notably faster than binary search. However, practical profit in total computation time could be limited due to the score computation bottleneck. In turn, there is room for score computation improvements using advanced methods (Korhonen et al., 2009).

4 RESULTS & DISCUSSION

4.1 Software Implementation

Our implementation of PERFECTOS-APE can be used to test a set of SNVs against a given collection of TFBS models. PERFECTOS-APE is implemented in Java and supports both mononucleotide and dinucleotide PWMs. Console and web-based versions are available at the PERFECTOS-APE website http://opera.autosome.ru/perfectosape/. Current implementation is restricted to SNVs with two alternative alleles. PERFECTOS-APE web interface allows testing a set of SNVs against publicly available motif collections: HOCOMOCO (Kulakovskiy et al., 2013b), JASPAR (Portales-Casamar et al., 2009), HT-SELEX (Jolma et al., 2010), SwissRegulon (Pachkov et al., 2007), and HOMER (Heinz et al., 2010).

We performed a basic benchmark with 100 SNPs and the HOCOMOCO collection of 426 mononucleotide PWMs on a common laptop with a Core i3 CPU. Precalculation with default settings took $\sim$ 150 Mb RAM and $\sim$ 30 seconds; the precalculation step took $\sim$ 0.5 hour for a comparable amount of dinucleotide PWMs. In the binary search mode processing of 100 SNPs took $\sim$ 5 min/$\sim$ 0.5 min and $\sim$ 90Mb/$\sim$ 200Mb RAM for basic/binary search mode respectively. For dinucleotide motifs the running time was $\sim$ 3 hours in basic mode and the same $\sim$ 0.5 min as for the mononucleotide case was necessary in binary search mode. Thus, PERFECTOS-APE allows large-scale SNP analysis.

4.2 Case Studies on Real Data

To test whether PERFECTOS-APE is able to produce meaningful predictions we analysed several disease-associated SNPs.

For instance, rSNP rs1314913 (Orr et al., 2012) is located in an intronic region of RAD51B gene that is significantly associated with breast cancer. It was proposed (Orr et al., 2012) that the minor allele negatively affects the binding site of the AP-1 complex. PERFECTOS-APE reproduces this prediction (fig.1): it shows 40× to 280× fold changes for binding site P-values.

We also checked breast cancer-associated rSNPs, which currently have no functional annotation (see fig.2):

- rs3112612 (Fletcher et al., 2011) possibly damages the E4F1 binding site, and
- rs4784227 (Long et al., 2010) possibly damages the HIC1 binding site.

Both those SNPs are located in the upstream region of the TOX3 gene, thus possibly affecting not the protein product directly, but its expression regulation.

To demonstrate PERFECTOS-APE performance on genomic scale, we've taken data on somatic mutations in 21 breast cancers (Nik-Zainal et al., 2012) selecting only those in intronic or promoter regions (76594 SNVs). It is known (Ostrow et al., 2014) that cancer somatic mutations escape the pressure of stabilizing selection responsible for maintaining adequate tissue specific regulation of gene expression, but fall under strong positive selection responsible for fast cell division and survival in the tumor environment. One of the sources of selection is disruption/emergence of a TFBS binding site for the mutated allele. Two prevalent types of breast cancer somatic mutations, CpG and TpC with nucleotide substitutions at C, were considered independently.

Using 426 HOCOMOCO PWMs we predicted TFBS overlapping the set of SNVs and passing 0.0005 P-value threshold for either the reference or the mutated allele. Using PERFECTOS-APE we calculated the number of cases when a somatic mutation caused a significant negative TFBS affinity change for the mutant allele (with a P-value cutoff at 5). To generate control data we shufflled 30bp sequences flanking the SNVs (preserving 1bp mutation context) and repeated TFBS prediction. This procedure was executed 8 times aggregating the results obtained for the same TF in the same mutation context, i.e. only CpG- and TpC-context variants were considered for shuffled sequences.

To evaluate whether somatic mutations are likely to occur in the TFBS favoring context we have constructed 2x2 contingency tables for the number of TFBS with a high/low affinity change caused by mutations (as table columns) and genuine/shuffled SNV context (as table rows). Fisher's exact test was used to estimate the significance of the association between the significant change of TFBS affinity and the intact SNV context. Since we tested all 426 HOCOMOCO models for each mutation site, we calculated Holm's multiple testing correction for the number of TFBS models.

We have found that original "genuine" mutations were the subject of either emergence or disruption of TF binding sites significantly more often than the shuffled random sequences.

In addition to Fisher's Test Holm-corrected P-values passing 0.05 threshold, we selected TFs for which the rate of TFBS disruption in real somatic mutations was no less than 1.3 times greater than for the synthetic data. The resulting list of TFs contained several well-known oncogenes, such as HIF1A, SP2, MYC for CpG context mutations and HLF, C/EBP-family, RUNX2 and others for TpC context mutations.

It is also notable, that a simpler approach comparing overall counts of TFBS predictions overlapping real somatic and ”shuffled” mutations was also showing high significance (FDR-corrected Fisher’s exact test P-value < 0.05) for the most of those TFs.

4.3 Comparison to Existing Tools

Existing tools for regulatory SNP annotation are mostly provided as web services rather than stand-alone command-line applications; the notable exceptions are sTRAP (implemented as R-package) and FunSeq (a C++ based command line tool). Web-only realization restricts applicability to large-scale data sets.

Another limitation comes from predefined sets of transcription factor binding motifs collections with most of the web-tools providing only two built-in collections: TRANSFAC and JASPAR. The choice of an appropriate motif collection plays an important role in finding effects of regulatory substitutions. It is desirable to have an option to utilize a user-defined motif collection since the general purpose databases are incomplete and motifs for new TFs keep emerging.

P-value ratio is a common approximation for binding affinity change and is used by almost all tools except RAVEN (Andersen et al., 2008), which calculates score difference. PERFECTOS-APE uses an approach for P-value calculation very similar to that used in is-rSNP (Macintyre et al., 2010); it allows exact calculations of P-value with a given precision. This differs our approach from RegSNP (Teng et al., 2012) and sTRAP (Manke et al., 2010) which rely on approximate techniques.

Most of tools use PWM motif models whereas PERFECTOS-APE supports both PWM and diPWM. The other tool that employs motif models more complex than PWMs is rSNP-MAPPER, which uses Hidden Markov Models.

RAVEN, ChroMoS (Barenboim and Manke, 2013), RegSNP and FunSeq (Khurana et al., 2013) are complex integrative tools which provide annotation compiling information from different data sources, while is-rSNP, sTRAP, rSNP-Mapper (Riva, 2012) and PERFECTOS-APE perform solely sequence analysis.

PERFECTOS-APE was intentionally designed as a stand-alone tool accepting raw sequences (not gene names or dbSNP identifiers) and making no preliminary filtering of SNPs according to their location, functionality etc. (unlike integrative tools).

PERFECTOS-APE does not perform any multiple testing correction, which is left up to downstream analysis if necessary.

We tested several tools on a verified SNP (rs1314913) and gathered ranks of AP-1 complex in the resulting predictions (see table 1). Our observations show that all methods behave similarly, but the results depend on the motif collection. For instance, is-rSNP fails to recognize rs1314913 as a regulatory SNP using JASPAR motif collection. For other tools ranks of transcription factors vary significantly depending on selected collection.

4.4 Discussion

PERFECTOS-APE can be improved in many ways, including direct support for multiallelic SNVs and more efficient procedure to scan sequences for binding sites. Also an effective parallelization is possible since each SNV versus PWM test can be run as a separate task or thread. Dinucleotide PWMs have shown better TFBS recognition than classic PWMs (Kulakovskiy et al., 2013a; Levitsky et al., 2014), yet the classic PWMs are much more common and are available for a wider range of transcription factors. As more data on TF binding becomes available and advanced models are constructed for a wider range of TFs, PERFECTOS-APE can be efficiently used for analysis of regulatory SNVs and SNPs.

ACKNOWLEDGMENTS

The authors thank the School for Molecular and Theoretical Biology of the Dynasty Foundation, and personally Anya Piotrovskaya and Fyodor Kondrashov. This work was supported by the Dynasty Foundation Fellowship; by Russian Scientific Foundation [14-14-01140]; by the Program on Molecular and Cell Biology from the Presidium of RAS.

REFERENCES

Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., Wasserman, W. W., and Odeberg, J. (2008). In silico detection of sequence variations modifying transcriptional regulation. *PLoS computational biology*, 4(1):e5.

Barenboim, M. and Manke, T. (2013). Chromos: an integrated web tool for snp classification, prioritization and functional interpretation. *Bioinformatics*, 29(17):2197–2198.

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., Zelenika, D., Gut, I., Heath, S., Palles, C., et al. (2011). Novel breast cancer susceptibility locus at 9q31. 2: results of a genome-wide association study. *Journal of the National Cancer Institute*.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., et al. (2010). Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome research*, 20(6):861–873.

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al. (2013). Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science*, 342(6154):1235587.

Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. (2009). Moods: fast search for position weight matrix matches in dna sequences. *Bioinformatics*, 25(23):3181–3182.

Kulakovskiy, I., Levitsky, V., Oshchepkov, D., Bryzgalov, L., Vorontsov, I., and Makeev, V. (2013a). From binding motifs in chip-seq data to improved models of transcription factor binding sites. *Journal of bioinformatics and computational biology*, 11(01).

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2013b). Hocomoco: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(D1):D195–D202.

Levitsky, V. G., Kulakovskiy, I. V., Ershov, N. I., Oschepkov, D. Y., Makeev, V. J., Hodgman, T., and Merkulova, T. I. (2014). Application of experimentally verified transcription factor binding sites models for computational analysis of chip-seq data. *BMC genomics*, 15(1):80.

Long, J., Cai, Q., Shu, X.-O., Qu, S., Li, C., Zheng, Y., Gu, K., Wang, W., Xiang, Y.-B., Cheng, J., et al. (2010). Identification of a functional genetic variant at 16q12. 1 for breast cancer risk: results from the asia breast cancer consortium. *PLoS genetics*, 6(6):e1001002.

Macintyre, G., Bailey, J., Haviv, I., and Kowalczyk, A. (2010). is-rsnp: a novel technique for in silico regulatory snp detection. *Bioinformatics*, 26(18):i524–i530.

Manke, T., Heinig, M., and Vingron, M. (2010). Quantifying the effect of sequence variation on regulatory interactions. *Human mutation*, 31(4):477–483.

Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9):e1003214.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993.

Orr, N., Lemnrau, A., Cooke, R., Fletcher, O., Tomczyk, K., Jones, M., Johnson, N., Lord, C. J., Mitsopoulos, C., Zvelebil, M., et al. (2012). Genome-wide association study identifies a common variant in rad51b associated with male breast cancer risk. *Nature genetics*, 44(11):1182–1184.

Ostrow, S. L., Barshir, R., DeGregori, J., Yeger-Lotem, E., and Hershberg, R. (2014). Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS genetics*, 10(3):e1004239.

Pachkov, M., Erb, I., Molina, N., and Van Nimwegen, E. (2007). Swissregulon: a database of genome-wide annotations of regulatory sites. *Nucleic acids research*, 35(suppl 1):D127–D131.

Ponomarenko, J. V., Merkulova, T. I., Vasiliev, G. V., Levashova, Z. B., Orlova, G. V., Lavryushev, S. V., Fokin, O. N., Ponomarenko, M. P., Frolov, A. S., and Sarai, A. (2001). rsnp_guide, a database system for analysis of transcription factor binding to target sequences: application to snps and site-directed mutations. *Nucleic acids research*, 29(1):312–316.

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., and Sandelin, A. (2009). Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research*, page gkp950.

Riva, A. (2012). Large-scale computational identification of regulatory snps with rsnp-mapper. *BMC genomics*, 13(Suppl 4):S7.

Teng, M., Ichikawa, S., Padgett, L. R., Wang, Y., Mort, M., Cooper, D. N., Koller, D. L., Foroud, T., Edenberg, H. J., Econs, M. J., et al. (2012). regsnps: a strategy for prioritizing regulatory single nucleotide substitutions. *Bioinformatics*, 28(14):1879–1886.

Touzet, H., Varré, J.-S., et al. (2007). Efficient and accurate p-value computation for position weight matrices. *Algorithms Mol Biol*, 2(1510.1186):1748–7188.

APPENDIX

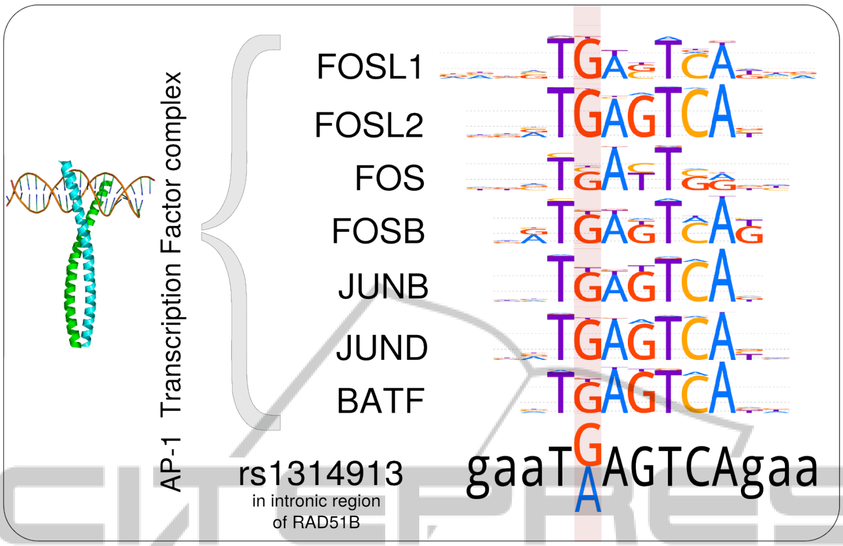

Figure 1: SNP rs1314913 is located in the intronic region of the breast cancer-associated RAD51B gene that is involved in DNA recombination and DNA repair. The T(A) allele is associated with about 1.6 times higher odds of breast cancer in men (Orr et al., 2012). Motif LOGO representations are given for HOCOMOCO PWMs (Kulakovskiy et al., 2013b). PERFECTOS-APE P-value fold changes are 40× to 280×, TFBS P-value threshold was 0.0005

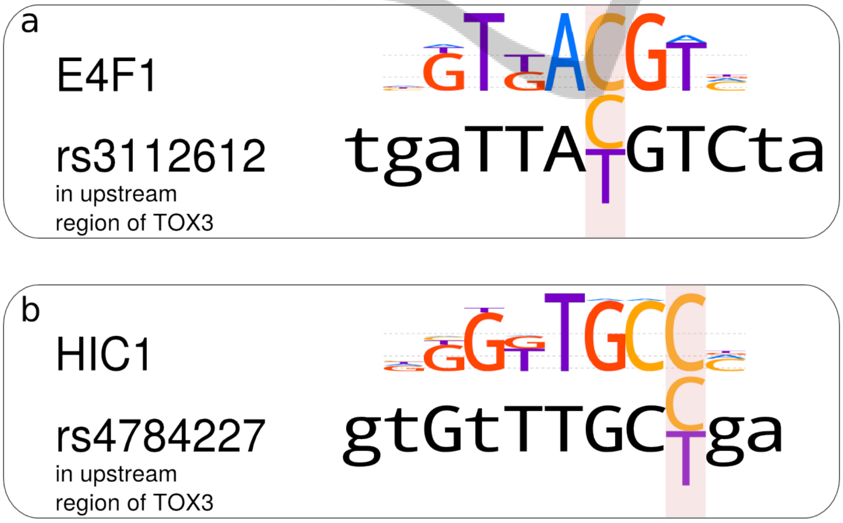

Figure 2: Breast cancer-associated SNPs in the promoter region of the TOX3 gene (fold-change cutoff 5; TFBS P-value threshold 0.001). (a) rs3112612, Risk allele (T) damages putative E4F1 binding site. (b) rs4784227, Risk allele (T) damages putative HIC1 binding site.

Table 1: AP-1 complex ranking in resulting predictions for rs1314913 analyzed by different tools.

PERFECTOS-APE	RegSNP	rSNP-MAPPER	sTRAP	is-rSNP
1-7,13	9,10,13	3,5,6,9	1-6	3,8