Preprint version

Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding

Kirsti Laurila[1]*, Harri Lähdesmäki[1,2]

[1]Department of Signal Processing, Tampere University of Technology, P.O. Box 527, FI-33101 Tampere, Finland
[2]Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland

Email: Kirsti Laurila- kirsti.laurila@tut.fi, Harri Lähdesmäki - harri.lahdesmaki@tut.fi
*Corresponding author

Abstract

Detailed knowledge of the mechanisms of transcriptional regulation is essential in understanding the gene expression in its entirety. Transcription is regulated, among other things, by transcription factors that bind to DNA and can enhance or repress the transcription process. If a transcription factor fails to bind to DNA or binds to a wrong DNA region that can cause severe effects to the gene expression, to the cell and even to the individual. The problems in transcription factor binding can be caused by alterations in DNA structure which often occurs when parts of the DNA strands are mutated. An increasing number of the identified disease-related mutations occur in gene regulatory sequences. These regulatory mutations can disrupt transcription factor binding sites or create new ones. We have studied effects of mutations on transcription factor binding affinity computationally. We have compared our results with experimentally verified cases where a mutation in a gene regulatory region either creates a new transcription factor binding site or deletes a previously existing one. We have investigated the statistical properties of the changes on transcription factor binding affinity according to the mutation type. Our analysis shows that the probability of a loss of a transcription factor binding site and a creation of a new one varies remarkably by the mutation type. Our results demonstrate that computational analysis provides valuable information about the effect of mutations on transcription factor binding sites. The analysis results also give a useful test set for in vitro studies of regulatory mutation effects.

Keywords: transcription factors, binding affinity, regulatory mutation

Introduction

One of the most important mechanisms of gene expression regulation happens at transcription level via binding of transcription factors (TFs) to DNA sequences. Understanding the gene expression process is important since this increases our knowledge of how organisms develop

and how different diseases are caused. If these mechanisms are understood, it is easier to develop new drugs and other treatments for different diseases. Studying mutations that alter the TF binding can help us to understand the complex regulation system and yield useful information for both computational and experimental research topics. Millions of single nucleotide polymorphisms (SNPs) are identified in the human genome. The majority of these SNPs are neutral but some of them are linked to hereditary diseases. Most of these disease-causing mutations alter the protein sequence, but a set of mutations are identified to occur in gene regulatory sequences. These mutations may cause a significant change in individual's phenotype by increasing or decreasing gene expression levels. Some examples of this have been verified experimentally. The expression level of the gene for a 91-kD glycoprotein component of the phagocyte oxidase (gp91-*phox*) is decreased because of promoter region mutations that are associated with X-linked chronic granulomatous disease [1]. Moreover, with Alzheimer disease patients, abnormally high expression levels of the amyloid precursor protein (APP) were measured in vitro when studying three point mutations in the APP promoter region [2]. Although all the mechanisms of gene expression regulation are not known, some cases are identified where the mutations in the promoter regions cause binding of a wrong transcription factor, and this in turn have an effect on transcription levels. For instance, it has been shown experimentally that a point mutation T→C at 77 nucleotides upstream of the transcription starting site (TSS) of the δ-globin gene (HBG) changes the binding affinity of the TF GATA-1 and this also alters the expression level of the gene (see Fig. 1a). This mutation is associated with a hereditary disease δ-thalassemia [3]. Another example is described in [4] where a point mutation at 292 nucleotides upstream of the TSS of the reticulocyte-type 15-lipoxygenase-1 (ALOX15) gene causes a new transcription factor binding site (TFBS) for the TF SPI1 (see Fig. 1b). This again causes three-fold expression levels compared with wild type gene expression. ALOX15 has a role in the development of asthma and some other diseases.
Mutation effect on TF binding has been studied computationally in [5], where authors have used the change of a score computed based on position specific scoring matrices (PSSMs) to infer if the binding affinity of some TFs change. This can be problematic since a single nucleotide change usually causes a small change in the score and more importantly, one cannot directly say whether this change of the score is significant or not. This fact was found when the scores of mutations that are known to affect TF binding were compared with the scores of background substitutions [5].

Figure 1. Examples of point mutations that affect TF binding a) A mutation in the promoter of δ-globin gene destroys the binding site of TF GATA-1. b) A new TF binding site for TF SPI1 is caused by a mutation in the promoter of ALOX15 gene.

In this paper, we use a similar but more fundamental approach as in [5] to analyze regulatory mutations and how they affect TF binding. Instead of analyzing the raw PSSM scores, we use the P-values to compare the wild type and mutated cases and to get the results of different genes and TFs comparable. We also use a larger set of regulatory mutations than in [5] and the mutations used in our analysis are disease-related so most of them are assumed to change the gene expression in some way, even though all are maybe not affecting the TF binding. We also distinguish the cases where a mutation causes the TF binding affinity to get weaker or stronger. A preliminary version of this work has been reported in our previous conference article [6].

As a novel aspect, we study the mutation effects on TF binding by dividing the mutation data into classes based on their type. This analysis is of great importance since it has been found that contacts between TFs and purines are important in TF-DNA complexes especially for some TFs [7]. Further, different dinucleotide steps vary in their ability to form kinks and bends of DNA and the bending of DNA has an effect on TF binding [7-10]. If the mutation changes the DNA structure in the position where TF is to bind, the binding affinity can change remarkably.

Methods

Datasets

Experimentally verified mutations dataset

We made a literature search for known mutations affecting TF binding. We collected 21 experimentally proven mutations. We used the same dataset as in [5] but we excluded those cases where the TFBS could not clearly be measured by PSSMs as a result of spacer molecules in TFBS or other similar reasons. The rest of the mutations included in our dataset come from articles [3, 11, 12] and rSNP DB [13]. These experimentally verified mutations were used to set a threshold for a relevant change in TF binding affinity.

Disease-related mutations dataset

The disease-related mutations we used were the regulatory mutations from Human Gene Mutation Database (HGMD) [14]. The regulatory mutations dataset was filtered to contain only those mutations that occur upstream from transcription or translation starting sites. Altogether we used 474 mutations in 256 genes.

Scores for binding affinity changes

PSSMs are a widely used method in predicting TFBSs and we apply them in our analysis as well [15, 16]. PSSMs were collected from Transfac (Release 10.3) [17] and Jaspar [18]. Only those matrices that have been built (at least partially) using human sequences were used. After this selection, we had 496 matrices for 343 different TFs. We added a small pseudo count (0.005) to all elements in PSSM to prevent zero probabilities.

The score for TF binding to the *n*-length DNA sequence x_1^n was computed by

$$S(x_1^n) = \frac{P_{TF}(x_1^n)}{P_{BG}(x_1^n)},$$

where $P_{TF}(x_1^n)$ is the probability computed by PSSM and $P_{BG}(x_1^n)$ is the background probability. $P_{TF}(x_1^n)$ for a particular PSSM M is computed with formula

$$P_{TF}(x_1^n) = \prod_{i=1}^{n} P_M(x_i, i),$$

where $P_M(x_i, i)$ is the probability of nucleotide x_i in the column i of the matrix M.

As a background model, we used a third order Markov model whose parameters were computed from the promoter sequences of all human genes. As a promoter sequence, we considered commonly used 5000 bases upstream from the start of the first (according to 5' end) annotated mRNA sequence of the gene (altogether we had 23 784 promoter sequences). However, if some promoter sequences overlapped, the overlapping part was used only once. The promoter sequences we used were collected from annotated sequence files (gbk-files) of human

chromosomes. These files were downloaded from the ftp-site of National Center for Biotechnology Information.

We computed the scores for the wild type and the mutated sequences of our regulatory mutations data set. Since the location of mutation in putative binding sites is not known, we computed the scores for all locations within a PSSM. In view of the fact that the score distributions were very different for each PSSM, we did not compare the scores but computed the P-values for each putative binding position. To compute the P-value, we evaluated the null distribution for binding scores for each PSSM separately and thus we did not need to assume any particular distribution for binding scores. The distributions were obtained by computing the scores for each position of each promoter sequence for every human gene and the P-values were estimated based on this score distribution. To measure the change of the binding affinity of each TF we computed the change in P-value between wild type and mutated sequence (the wild type sequence - the mutated sequence). This P-value change was considered as a score to measure the change in TF's binding affinity.

The mutation position in a TFBS was considered as the mutation position in a PSSM. Even though this might not be always the case since the PSSM may contain more or less bases than the actual binding site, this was the only way to measure the position. The positions were normalized according to matrix width so that the leftmost base got the position -1 and the rightmost the position 1. In analysis of mutation positions relative to TSS this information was known for 57 % of mutations. Only these cases were used in this analysis part.

Mutation classes

Nucleotides can be divided into two bases containing classes in three different ways. The divisions are those for purines (denoted by R, consists of bases A and G) and pyrimidines (denoted by Y, bases C and T), those for bases containing amino-group (denoted by M, bases A and C) and those with keto-group (denoted by K, bases G and T), and division into bases forming strong base pairs in double stranded DNA (denoted by S, bases C and G) and those forming weak base pairs (denoted by W, bases A and T). For each division, there are four different mutation types, i.e. mutations R→R (where A mutates to G or G to A), R→Y, Y→R and Y→Y. If the dinucleotide steps are considered, then eight different mutation classes occur whether the mutation is in the first or in the second nucleotide, for mutations in the first nucleotide (underlined) the classes are $\underline{R}$R→$\underline{R}$R, $\underline{R}$R→$\underline{Y}$R, $\underline{R}$Y→$\underline{R}$Y, $\underline{R}$Y→$\underline{Y}$Y, $\underline{Y}$R→$\underline{R}$R, $\underline{Y}$R→$\underline{Y}$R, $\underline{Y}$Y→$\underline{R}$Y and $\underline{Y}$Y→$\underline{Y}$Y and similar mutation types are, when the mutation occurs in the second nucleotide. We divided mutations into these classes. Each mutation class was studied separately.

Statistical testing

The two-sample Kolmogorov-Smirnov test [19] was performed for two different testing strategies. First, the distributions of the relevant P-value changes of each mutation class were compared to the distribution of all relevant changes. The testing was performed at significance level α=0.01 and Bonferroni correction was used. Second, for each mutation class and for the set of all relevant mutations, we also tested whether the mutation type had an equal effect in increasing and decreasing the binding affinity. This was done by comparing the left and right sides of the distribution of P-value changes. The left side (negative P-value changes) of the distribution was mirrored around the origin by taking the absolute value of negative P-value changes and by re-computing the distribution, after which we can compare the two sides of the

distribution in the standard way. For each mutation class, the equality of the distributions of the P-value changes with the distribution of the P-value changes of all relevant mutations was also tested with the two sample permutation test using the procedure described in [20]. The absolute differences between the distributions of the P-value changes of the mutation classes and the distribution of the P-value changes of all relevant mutations were computed by subtracting the weighted means of each distribution from each other.

Results and Discussion

Experimentally verified mutations data set

We evaluated the effect of the experimentally verified mutations on TF binding by P-values derived from PSSM scores. The data set contained 21 mutations. The list of mutations, their scores for binding affinity changes (i.e. difference in P-value) and P-values for wild types are in Tab. 1. For each mutation, all big changes (P-value change over 0.2) are shown in Tab 1. If the mutation did not show any notable change in binding affinity score, the scores and P-values are not shown. Over two thirds of the verified mutations showed a notable change in P-value between the wild type and mutated sequence. However, the P-values of the sequences which have stronger affinity to TFs were quite high in some cases, i.e. the binding site would not be deemed as statistically significant. Nevertheless, even weaker binding sites can be important, since it has been recently shown that models which include weak binding sites predict the expression patterns better than those models from which the weak binding sites are excluded [21].

For the experimentally verified mutations, big P-value changes are found for several PSSMs of a single TF. For example for the mutation in hemoglobin gamma G (HBG2) promoter, the P-values corresponding to 4 out of 7 PSSMs for TF SP1 showed a difference in binding affinity. However, one of the matrices showed the change in two different matrix positions, which reflects the fact that all of the matrices are not very specific to the binding site.

Six mutations in the data set did not show change in binding affinity. This is most likely caused because of the inaccuracy of PSSMs. The TFBS can be formed of several DNA pieces or the mutation can disrupt the binding via other mechanisms than directly affecting the structure of the TFBS. For example, we did not include in our experimentally verified mutations data set the mutation C→T in the promoter region (position -677 relative to TSS) of the FLT1 gene since in the middle of the binding site of the TF p53 there is a spacer molecule that could not be modeled via ordinary PSSMs [22]. Even though we left similar cases of our data set out, the binding mechanism is not exactly known for every TF and thus analysis using PSSMs may fail in predicting the mutation effect.

Table 1. Experimentally verified mutations and their effect on TF binding.[1]

[1] P-values are presented only for those PSSMs that show relevant changes. For mutation in SP1 gene, the position relative to TSS was not reported in reference, in addition, the mutation is not identified with any disease but the transcription of the SP1 gene is increased. MP=mutation position relative to TSS, MW=matrix width, POM= mutation position on matrix, wt=wild type, Δ P-value=(P-value of wt) -(p-value of mutated sequence), ref=reference.

gene symbol	muta-tion	MP	TF	MW	PO	effect on binding	ΔP-value	P-value of wt	disease	ref
AFP	C→A	-55	HNF-1	21	12	increase	0.349	0.626	hereditary persistence of alpha fetoprotein	[23]
AFP	G→A	-119	HNF-1	15	5	increase	0.319	0.611	"	[24]
AFP	G→A	-119	HNF-1	21	15	increase	0.298	0.446	"	[24]
AFP	G→A	-119	HNF-1	21	9	increase	0.212	0.277	"	[24]
AGTRL1	G→A	-154	SP1			decrease			risk of brain infarction	[25]
ALOX	A→G	-292	SPI1	6	2	increase	0.356	0.592	(anti)inflammatory effects	[4]
AGT	A→C	-20	ER1	19	14	decrease	-0.422	0.488	hypertension	[26]
AGT	A→C	-20	ER1	11	2	decrease	-0.316	0.530	"	[26]
CETP	C→A	-629	SP1			increase			high high density lipo-protein cholesterol levels	[27]
F7	C→G	-94	SP1			decrease			F7 deficiency	[28]
FECH	G→C	-250	SP1	10	8	decrease	-0.373	0.280	erythropoietic porphyria	[29]
FECH	G→C	-250	SP1	10	4	decrease	-0.274	0.203	"	[29]
FECH	G→C	-250	SP1	13	4	decrease	-0.250	0.250	"	[29]
FECH	G→C	-250	SP1	13	5	decrease	-0.361	0.228	"	[29]
FECH	G→C	-250	SP1	13	4	decrease	-0.347	0.434	"	[29]
FECH	G→C	-250	SP1	13	3	decrease	-0.306	0.434	"	[29]
Gp1bβ	C→G	-133	GATA1			decrease			Bernard-Soulier Syndrome	[30]
HBD	T→C	-77	GATA1	13	12	decrease	-0.386	0.553	δ-thalassemia	[3]
HBG2	C→G	-202	SP1	10	4	increase	0.274	0.540	hereditary persistence of fetal hemoglobin	[11]
HBG2	C→G	-202	SP1	10	5	increase	0.402	0.702	"	[11]
HBG2	C→G	-202	SP1	13	6	increase	0.658	0.861	"	[11]
HBG2	C→G	-202	SP1	10	4	increase	0.373	0.653	"	[11]
HBG2	C→G	-202	SP1	10	4	increase	0.206	0.420	"	[11]
ITGA2	C→T	-52	SP1	10	9	decrease	-0.235	0.178	diminished expression of the integrin on platelets	[31]
ITGA2	C→T	-52	SP1	10	7	decrease	-0.363	0.217	"	[31]
ITGA2	C→T	-52	SP1	10	4	decrease	-0.284	0.296	"	[31]
ITGA2	C→T	-52	SP1	10	3	decrease	-0.408	0.217	"	[31]
ITGA2	C→T	-52	SP1	10	4	decrease	-0.842	0.107	"	[31]
LIPC	C→T	-480	USF	14	5	decrease	-0.430	0.210	low hepatic lipase activity	[32]
LIPC	C→T	-480	USF	14	5	decrease	-0.281	0.060	"	[32]
LIPC	C→T	-480	USF	10	3	decrease	-0.326	0.085	"	[32]
LIPC	C→T	-480	USF	12	2	decrease	-0.269	0.256	"	[32]
NFKBIL1	T→A	-62	USF1	8	5	decrease	-0.386	0.420	rheumatoid arthisis	[33]
PROC	T→C	-14	HNF-1	15	7	decrease	-0.216	0.265	protein C deficiency	[34]
PTGS2	G→A	-1195	MYB			increase			risk of esophageal squa-mous cell carcinoma	[35]
SP1	A→C	?	NFY	11	10	decrease	-0.276	0.150	?	[36]
Tcof1	C→T	-346	YY1			decrease			Treacher Collins syndrome	[37]
TNF	G→A	-376	Oct-1	13	6	increase	0.303	0.492	cerebral malaria	[38]
TNF	G→A	-376	Oct-1	11	9	increase	0.414	0.787	"	[38]
TNF	G→A	-376	Oct-1	11	3	increase	0.340	0.665	"	[38]
UROS	C→A	-90	CP2	18	13	decrease	-0.207	0.143	congenital erythro-poietic porphyria	[12]
UROS	C→A	-90	CP2	11	11	decrease	-0.274	0.164	"	[12]
UROS	T→C	-70	GATA1	14	8	decrease	-0.317	0.085	"	[12]
UROS	T→C	-70	GATA1	13	7	decrease	-0.206	0.038	"	[12]

Disease-related mutations data set

We computed the change in P-value for each mutation in the disease-related mutations dataset for each TF. This P-value change was considered as a score to measure the change in TF's binding affinity. The distribution of the changes is shown in Fig. 2a. Based on the experimentally verified cases we considered the change to be relevant if the P-value change (absolute value) was over 0.3 or the change was over 0.2 and P-value of either the wild type or the mutated sequence was under 0.3. Approximately 11% of the changes exceeded these boundaries. We tried different thresholds, too, but they did not result in any notable changes in the characteristics of the set of relevant changes and hence the conclusions were similar. The set of experimentally verified mutations is relatively small and that prevents us from inferring more conservative thresholds without losing too many verified cases. However, the experimentally verified mutation set was significantly enriched in the set of relevant mutations, when testing was performed with hypergeometric test (P-value 9.261e-10). Further, current knowledge does not allow us to discriminate true and false changes more carefully (see e.g. [5]). This choice of thresholds, however, results in a set of predicted binding changes that is enriched for true binding affinity changes. Consequently, despite some false positives, our analysis results provide insights into true mutation effects. The analysis provides a list of testable hypothesis, ordered according to the significance of mutation effect, which can be readily tested in a laboratory to verify the real mutation effect in vitro. Besides, if a particular TF is known to regulate some gene and our analysis gives a large absolute P-value change for the affinity of that TF due to mutation, this provides a strong evidence for the mutation effect and this should be taken into account when studying the disease mechanisms at the molecular level.

Figure 2. Distributions of the p-value changes. a) All changes. b) Only changes that are considered relevant.

The distribution of the P-value changes that are considered relevant and are further studied here can be seen in Fig. 2 b). It can be seen that the left side of the bimodal distribution has somewhat larger area than the right side i.e. the mutations seem to cause more often the loss of TF binding affinity than create a new TFBS. The statistical analyses of the significance of the results are in the end of the Results and Discussion section.

We analyzed whether the mutation position in a TFBS affects the change in TF binding affinity. In Fig. 3 are the distributions of the mutation positions in PSSMs. The positions are normalized since the widths of the PSSMs varied from 8 to 30 nucleotides. As expected, the positions are uniformly distributed in all changes (Fig. 3a). For the set of the relevant changes in TF binding affinity the mutation occurs more likely in the central parts of TFBS (Fig. 3b). This result is expected since in PSSMs the nucleotides in the middle of the binding sites are more conserved than those in the sides. However, a part of the effect may be caused because of the quality of PSSMs, too. Besides, in the widest matrices, the side nucleotides may not present the TFBS at all, but the areas near the binding site that are important only by providing right binding environment. We also separately analyzed the relevant changes that either increased or lowered the TF binding affinity but the distributions of mutation positions in PSSMs were similar to all relevant changes.

Next, we analyzed the effect of mutation position according to TSS. We used the same division as in [5] to get comparable result. The distributions of the relevant P-value changes for different mutation positions relative to TSS can be seen in Fig. 4. In the plots there is also the distribution

of all P-value changes that exceeded the thresholds, as a reference distribution. It can be seen that there are only slight variations in the distributions. This was the case when all P-value changes were studied, too. Our analysis confirms the result of [5] that the position of regulatory mutation relative to TSS does not correlate with the strength of mutation effect. This is natural since as a result of DNA looping, a TF that bounds to thousands of base pairs upstream from the target gene can attach the TSS [39]. Our results can also reflect the fact that most genes have several TSSs.

Figure 3. Distributions of the normalized mutation positions in PSSMs. a) All changes. b) Only changes that are considered relevant.

Figure 4. Distributions of the p-value changes according to mutation position in promoter. a) 0-100 bases upstream from TSS b) 0-500 bases upstream from TSS c) 500-2000 bases upstream from TSS d) 2000-10000 bases upstream from TSS e) over 10000 bases upstream from TSS.

Effects of different mutation classes

We divided the mutations by the three common ways to distinguish the nucleotides and analyzed the effects of different mutation types separately. Fig. 4 shows how mutations in purines/pyrimidines affect the TF binding in relevant binding affinity changes. Distribution of all relevant changes is again included as a reference distribution. It can be seen that when mutation changes the type of the base from pyrimidine to purine (Fig. 5c), then the TF binding affinity is more likely to be increased than decreased (that is, a new binding site for TF is more likely to be formed than existing one is disrupted). Otherwise, if the mutation does not change the type of base the effect is contrary (Fig. 5a and 5d). For mutation from purine to pyrimidine the probability of forming a new binding site is the same as that of disrupting an existing one (Fig. 5b).
For base division into strong and weak pairing bases the phenomenon is of the same kind as with division into purines and pyrimidines. However, for mutation from base with weak pairing to a similar base there is no difference whether mutation more often causes weaker or stronger binding of TF. In division into bases with amino groups and keto groups one cannot see similar effects than for other divisions but if the mutation changes the base type the binding affinity more often gets weaker than stronger. Nonetheless, differences exist only for mutations K→K and K→M and they were smaller than for other divisions. For mutations divided by nucleotides the strongest effects were for mutations A→C (stronger affinity for TF binding), C→T (weaker affinity for TF binding) and G→A (weaker affinity for TF binding). The results indicate that type of the mutation is important. Formation of a new binding site naturally requires bigger changes in the structure of DNA than the loss of an existing one. So, by substituting a base of different size or different number of hydrogen bonds in double stranded DNA, the binding affinity can often be increased remarkably. Further, if there is a mutation in already existing TFBS, even smaller changes in DNA can cause a loss of the binding site.

Figure 5. Distributions of the p-value changes in different mutation types. a) R→R b) R→Y c) Y→R d) Y→Y.

We also computed the distributions of the P-value changes for each mutation type for different dinucleotide steps (16 dinucleotide classes for each base division). The distributions for diffeerent classes varied remarkably. In Fig. 6 are the distributions of the P-value changes for dinucleotide division purine-pyrimidine when a mutation is in the second nucleotide. Fig. 6 also shows the distribution of all P-value changes that exceeded the thresholds, as a reference distribution. As in the analysis of single nucleotide steps, one can find interesting features on distributions of dinucleotide steps, too. For mutation R<u>R</u>→R<u>R</u> (Fig. 6a, mutated nucleotide underlined), the probability of generating a new TFBS is approximately as probable as a disruption of an existing binding site. This was also the case for the mutation type YY→YY (mutation in either nucleotide). For mutations R<u>R</u>→R<u>Y</u> (Fig. 6b) and Y<u>Y</u>→Y<u>R</u> (Fig. 6g), <u>R</u>R→<u>Y</u>R and <u>Y</u>Y→<u>R</u>Y the mutation considerably more often caused a new binding site than disrupted an existing one. The rest of the mutations caused more likely the removal of an old binding site than making a new one as can be seen for example in Fig. 6c.

Figure 6. Distributions of the p-value changes in different dinucleotide mutation types. Mutated nucleotide underlined. a) R<u>R</u>→R<u>R</u> b) R<u>R</u>→R<u>Y</u> c)R<u>Y</u>→R<u>R</u> d) R<u>Y</u>→R<u>Y</u> e) Y<u>R</u>→Y<u>R</u> f) Y<u>R</u>→Y<u>Y</u> g) Y<u>Y</u>→Y<u>R</u> h) Y<u>Y</u>→Y<u>Y</u>.

The above results suggest that purine-pyrimidine and pyrimidine-purine dinucleotides play an important role in TF binding. It has been previously shown that pyrimidine-purine steps are flexible allowing the DNA strands to form sharp kinks [9]. This is important for TF binding that usually bends the DNA or TF binds to a bent DNA. Some TFs also recognize particular bends or kinks in DNA or flexible DNA regions. Different TFs are known to bind to bent DNA sites and in these cases dinucleotides play a critical role [7]. Nevertheless, such flexibility is not shown to occur with all purine-pyrimidine steps, even though an RY step GC, for example, can also form more conformations than the AA and TT steps [10]. This can affect the phenomena seen in our analysis. As one property of the DNA flexibility, the different dinucleotide steps have different effect on the width of the major and minor grooves [10]. If there exists a mutation that changes particularly the width of the major groove, this can cause a strong effect on TF binding since it is known that contacts between bases in major groove and amino-acids in TF-DNA complex are especially important [7].

One very insightful result is that mutations for opposite directions caused clear contrary effects in three cases: R<u>R</u>→R<u>Y</u> and R<u>Y</u>→R<u>R</u> (Fig. 6b and 6c), Y<u>Y</u>→Y<u>R</u> and Y<u>R</u>→Y<u>Y</u> (Fig. 6g and 6f) and <u>R</u>R→<u>Y</u>R and <u>Y</u>R→<u>R</u>R. For the fourth case (<u>Y</u>Y→<u>R</u>Y and <u>R</u>Y→<u>Y</u>Y) an effect of the same kind can be seen but for the mutation <u>R</u>Y→<u>Y</u>Y the difference was smaller. This analysis indicates the great importance of the surrounding nucleotides of the mutation and the mutation type.

The division for strong and weak bond forming bases in dinucleotide steps resulted in similar but weaker effects than the purine-pyrimidine division. The effects were visible only for mutations S<u>S</u>→S<u>W</u>, W<u>S</u>→W<u>S</u>, <u>S</u>S→<u>W</u>S, <u>W</u>S→<u>W</u>S and <u>W</u>W→<u>W</u>W where the binding affinity was more likely to become lower than stronger and for mutations <u>W</u>S→<u>S</u>S, S<u>W</u>→S<u>S</u>, SW→SW (mutation in either nucleotide) and S<u>S</u>→S<u>S</u> where binding affinity was more probable to strengthen than to get smaller. Thus, the effect of contrary mutations did not occur in all pairs but can be seen only in pairs S<u>S</u>→S<u>W</u> and S<u>W</u>→S<u>S</u> and <u>S</u>S→<u>W</u>S and <u>W</u>S→<u>S</u>S. The mutations in the dinucleotide WW and when a nucleotide was mutated to WW showed only little or no difference in the affinity changes between a new binding site and a loss of an existing one which suggests that this

kind of mutation does not have any specific effects on TF binding. However, it seems clear that change in the number of hydrogen bonds from one to two consecutive strong base pairs has a clear effect on TF binding. For example, importance of hydrogen bonds is demonstrated in a case where arginine in zinc fingers 1 and 2 in transcription factor EGR1 (Early Growth Response Protein 1) site binds to the keto-oxygen of guanine and to the nitrogen of the imidatsole ring of guanine [7]. One could deduce that changing this nucleotide to another keto-oxygen containing base thymine would be destructive to the binding site since the other hydrogen bond with nitrogen could not be formed.

For division into bases with keto and amino groups, the big effects were in mutations KK→MK, KM→MM, MM→MM and slight effects on mutation KM→KM and mutations happening in dinucleotide MM, where affinity is more likely to get stronger. Further, affinity is more likely to get weaker in mutation types KM→KK, KK→KM and if mutation happened in dinucleotide MK. The other mutation classes showed only minor or no differences in binding affinity changes.

Statistical significance of mutation class differences

We performed the statistical tests to see if the effects of different mutation classes are real. We tested the cases with two different testing strategies: by the two-sample Kolmogorov-Smirnov test and by the two-sample permutation test and the results for these were similar. Most of the mutation classes appeared to have some kind of effect on distribution, since only in 2 of 12 single nucleotide mutation cases and in 5 of 48 dinucleotide cases the Kolmogorov-Smirnov test did not reject the null hypothesis that the distribution of changes in the mutation class was the same as the distribution of all changes (with significance level 0.01 and after Bonferroni correction). The cases where the mutation did not affect the distribution were M→K, W→W, KM→KM, MM→KM, WW→WS, WW→WW and SW→SW. When using permutation tests and comparing the means no additional mutation classes showed any statistically significant change in the distribution of scores. The distributions of changes for different mutation positions were all similar to the distribution of all relevant changes with risk level 0.01 when compared with the Kolmogorov-Smirnov test.

We also tested by the Kolmogorov-Smirnov test whether the absolute values of the two sides of the bimodal distributions were the same. We first compared the two sides of the distribution of all relevant changes. The hypothesis that the absolute value of left side of the distribution equals to the right side could be rejected with P-value 0 (with Matlab's computing accuracy). When testing the similarity of the sides of distributions of different mutation classes with risk level 0.01, the sides were the same (i.e. no statistically significant difference was found) for mutation classes R→Y, M→M, W→W, RR→RR, YY→YY, YY→YY, KK→KK, MK→MK, MM→MK, KK→KK, SS→SS and WW→SW.

After two testing procedures, we see that for 9/12 mutation classes, when mutation is considered in one nucleotide, and for 34/48 dinucleotide mutation classes, the effect of mutation was statistically significant. Even though so many mutation classes show consistent effects on transcription factor binding affinity, some of the influences are remarkably stronger compared to the others. We computed the differences between distributions of P-value changes and the biggest values (over 0.2) are shown in Tab. 2. These results confirm our findings of the effects of mutations discussed in previous section. For example, the distribution of changes in mutation class RR→RY differ remarkably from the distribution of all changes (See Tab. 2).

Table 2 – The biggest absolute differences between p-value change distributions of different mutation classes and the set of all relevant changes. [2]

mutation	difference
R**R**→R**Y**	0.302
Y**R**→Y**R**	0.214
Y**Y**→Y**R**	0.259
YY→**R**Y	0.320
KK→**M**K	0.257
KM→**M**M	0.255
MK→**M**K	0.213
S**S**→S**S**	0.237
S**S**→S**W**	0.207
S**W**→S**S**	0.299
S**W**→S**W**	0.233
SW→**S**W	0.285
WS→**S**S	0.257

Conclusion

Although accurate binding site prediction is difficult in general, our results demonstrate that computational analysis can provide valuable information about the effect of mutations on transcription factor binding sites. Our tests also offer a useful test set for in vitro studies of regulatory mutation effects.

We have shown that regulatory mutations can change the TF binding affinity remarkably. This does not originate only from a single nucleotide mutation but also the type of the surrounding nucleotides which should be taken into account when studying the effects of a new point mutation on gene expression regulation. We would like to acknowledge that all regulatory mutations in the HGMD are not verified as causative and to affect TF binding mechanisms. Our results, however, are not likely to be affected by these "false positives" as we look for general trends (differences in histograms) over mutation types and positions. Further, additional data sets, that will become available in the future, will be useful to refine our findings.

PSSMs are a widely used method in modeling TF binding. A problem with PSSMs is, however, the number of false positives in predicting TFBSs, which concerns our analysis as well. Depending on both the chosen threshold (p-value cut-off) and the specificity of PSSM, predictions can report TFBSs in every 500-5000 bases. It is estimated that only about 0.1% of these TFBSs may be functional even though many can be bound by TF in vitro studies [40]. As our studies with experimentally verified TFBSs and the mutations affecting them showed, the PSSM modeling does not always assign extremely small P-values to TFBSs. This can be a result of the structure of PSSMs which does not have any correlation between different bases. This is not very realistic since the structure of TFBS is built by many successive nucleotides that are not independent of each other. The quality of PSSMs affects also the application of computational method for hypothesis generation as variability in binding predictions depends directly on the

[2] Mutated nucleotide is underlined. Only those differences that exceed 0.2 are tabulated.

sequence specificity models. Fortunately, recently developed experimental techniques, such as protein binding microarrays, make it possible to measure TF binding specificities in high-throughput manner and will thereby improve computational analysis of regulatory mutations as well [41]. Our studies have also shown that the dinucleotides in TFBSs affect the binding significantly. This is most likely caused by the ability of DNA strands to bend. If the mutation changes DNA structure considerably, then the mutation effects may be stronger than if the mutation has a smaller effect on DNA structure. This is natural since the structure of TFBS has a vital role in DNA recognition [7]. Since different DNA-binding domains of TFs have different binding mechanisms and demands for DNA bending it could be more appropriate to study each TF family separately.

In the future it is important to incorporate additional knowledge into TF binding prediction. For example, models that combine the nucleosome positions or chromatin immunoprecipitation on chip (ChIP-chip) data are shown to improve TF motif discovery [42, 43]. Other additional data sources, such as DNase hypersensitive sites or protein-protein interactions, can also be incorporated into computational analysis. One possible method for modeling TF binding by combining different data sources is a Bayesian method presented in [44]. Studying mutation effects on models of this kind integrated with several data sources can provide additional insights, since the mutation can disrupt the TF binding not only by occurring directly in the binding site but also causing another molecule to change its binding. For example, if the mutation disrupts the binding of a TF that interacts with another TF the mutation can prevent the binding of both TFs. In addition, the TF binding differs in different states of the cell depending on the TFs present and their concentrations which can change the strength of mutation effect. The effects of mutations on TF binding could also be studied in a more detailed manner by dividing TFs into different classes as it is known that different TF protein families have different DNA-binding characteristics [45]. Also, some additional knowledge about mutations could be used. For example predictors such as presented in [46] could be used for this purpose to filter the data set to those regulatory mutations that are most likely to be functional.

As noted above, transcriptional regulation can occur also by a number of different mechanisms, including e.g. chromatin modifications. Neither the currently available data nor the analysis carried out in this study can account for these additional regulation mechanisms. Understanding the role of mutations on other transcription mechanisms will be an important future direction.

Acknowledgements

Financial support from Tampere Graduate School in Information Science and Engineering (TISE) is gratefully acknowledged. Work was also supported by the Academy of Finland, project nos 213462 (Finnish Programme for Centres of Excellence in Research 2006-2011), 106030, and 124615 and the Finnish Foundation for Technology Promotion.

References

1. Newburger, P. E., Skalnik, D. G., Hopkins, P.J., Eklund, A. A. and Curnutte, J. T. (1994). Mutations in the promoter region of the gene for gp91-phox in X-linked chronic granulomatous disease with decreased expression of cytochrome b558. J Clin Invest. **94**, 1205-1211.

2. Theuns, J., Brouwers, N., Engelborghs, S., Sleegers, K., Bogaerts, V., Corsmit, E., de Pooter, T., van Duijn. C. M., de Deyn, P. P. and van Broeckhoven, C. (2006). Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. Am J Hum Genet. **26**. 936-946.

3. Matsuda, M., Sakamoto, N. and Fukunaki, Y. (1992). δ-Thalassemia Caused by Disruption of the Site for an Erythroid-Specific Transcription Factor, GATA-1, in the δ-Globin Gene Promoter. Blood. **80**. 1347-1351.
4. Wittwer, J., Marti-Juan, J. and Hersberg, M. (2006). Functional Polymorphism in ALOX15 Results in Increased Allele-Spesific Transcription in Macrophages Through Binding of the Transcription Factor SPI1. Hum Mutat. **27**. 78-87.
5. 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 Comput Biol. **4**. e5.
6. Laurila, K. and Lähdesmäki, H. (2008). Effects of Disease-Related Mutations on Transcription Factor Binding. In Proceedings of the Fifth TICSP Workshop on Computational Systems Biology (WCSB 2008). 89-92.
7. Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: Structural Families and Principles of DNA recognition. Annu Rev Biochem. **61**. 1053-1095.
8. Harrington, R. E. (1992). DNA curving and bending in protein-DNA recognition. Mol Microbiol. **6**. 2549-2555.
9. Suzuki, M., Loakes, D. and Yagi, N.(1996). DNA conformation and its changes upon binding transcription factors. Adv Biophys. **32**. 53-72.
10. Travers, A. A. (2004). The structural basis of DNA flexibity. Philos Transact A Math Phys Eng Sci. **15**. 1423-1438.
11. Collins, F. S., jr Stoeckert, C. J., Serjeant, G. R., Forger, B. G. and Weissman, S. M. (1984). G gamma beta+ hereditary persistence of fetal hemoglobin: cosmid cloning and identification of a specific mutation 5' to the G gamma gene. Proc Natl Acad Sci USA. **81**. 4894-4898.
12. Solis, C., Aizencan, G. I., Astrin, K. H., Bishop, D. F. and Desnick, R. J. (2001). Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. J Clin Invest. **107**. 753-762.
13. Ponomarenko, J. V., Orlova, G. V., Ponomarenko, M. P., Lavryushev, S. V. and Merkulova, T. I. (2000). rSNP Guide: a database documenting influence of substitutions in regulatory gene regions onto their interaction with nuclear proteins and predicting protein binding sites, damaged or appeared de novo due to these substitutions. In Proceedings of BGRS'2000. 69-72.
14. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel. J. A., Thomas, N. S., Abeysinghe, S., Krawczak, M. and Cooper, D.N. (2003). The Human Gene Mutation Database (HGMD®): 2003 Update. Hum Mutat. **21**. 577-581.
15. Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics, **16**.1416-1423.
16. Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. **12**. 505-519.
17. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O.V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, P., Reuter, I., Rotert, S., Saxel, H., Scheer,. M., Thiele, S. and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. **31**. 374-378.
18. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. and Lenhard, B. (2004). JASPAR: an open access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. **32**. D95-97.
19. Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. Ann Math Statist. **28**. 823-838.
20. Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Boca Raton: Chapman & Hall/CRC.
21. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. Nature. **451**. 535-540.

22. Menedez, D., Krysiak, O., Inga, A., Krysiak, B., Resnick, M. A. and Schönfelder, G. (2006). A SNP in the flt-1 promoter integrates the VEGF system into the p53 transcriptional network. Proc Natl Acad Sci USA. **103**. 1406-1411.
23. Alj, Y., Georgiakaki, M., Savouret, J. F., Mal, F., Attali, P., Pelletier, G., Fourre, C., Milgrom, E., Buffet, C., Guiochon-Mantel, A. and Perlemuter,G. (2004). Hereditary persistence of alpha-fetoprotein is due to both proximal and distal hepatocyte nuclear factor-1 site mutations. Gastroenterology. **126**. 308-317.
24. McVey, J. H., Michaelides, K., Hansen, L. P., Ferguson-Smith, M., Tilghman, S., Krumlauf, R. and Tuddenham, E. G. (1993). A G→A substitution in an HNF I binding site in the human alpha-fetoprotein gene is associated with hereditary persistence of alpha-fetoprotein (HPAFP). Hum Mol Genet. **2**. 379-384.
25. Hata, J., Matsuda, K., Ninomiya, T., Yonemoto, K., Matsushita, T., Ohnishi, Y., Saito, S., Kitazono, T., Ibayashi, S., Iida, M., Kiyohara, Y., Nakamura, Y. and Kubo, M. (2007). Functional SNP in a Sp1-binding site of AGTRL1 gene is associated with susceptibility to brain infarction. Hum Mol Gen. **16**. 630-639.
26. Ishigami, T., Umemura, S., Tamura, K., Nyui, N., Kihara, M, Yabana, M., Watanabe, Y., Sumida, Y., Nagahara, T., Ochiai, H. and Ishii, M. (1997). Essential hypertension and 5' upstream core promoter region of human angiotensinogen gene. Hypertension. **30**. 1325-1330.
27. Dachet, C., Poirier, O., Cambien, F., Chapman, J. and Rouis, M. (2000). New functional Polymorphism CETP/-629, in Cholesteryl Ester Transfer Protein(CETP) Gene Related to CETP Mass and High Density Lipoprotein Cholesterol Levels: role of Sp1/Sp3 in Transcriptional Regulation. Arterioscler Thromb Vasc Biol. **20**. 507-515.
28. Carew, J. A., Pollak, E. S., High, K. A. and Bauer, K. A. (1998). Severe Factor VII Deficiency Due to a Mutation Disrupting an SP1 Binding Site in the Factor VII Promoter. Blood. **92**. 1639-1645.
29. Pierro, E. D., Moriondo, V. and Cappellini, M. D. (2004). Human gene mutations. Gene symbol: FECH. Disease: Porphyria, erythropoietic. Hum Genet. **114**. 221.
30. Ludlow, L. B., Schick, B. P., Budarf, M. L., Driscoll, D. A., Zackai, E. H., Cohen, A. and Konkle, B. A. (1996). Identification of a Mutation in a GATA Binding Site of the Platelet glycoprotein Ibβ Promoter Resulting in the Bernard-Soulier Syndrome. J Biol Chem. **271**. 22076-22080.
31. Jacquelin, B., Tarantino, M. D., Kritzik, M., Rozenshteyn, D., Koziol, J. A., Nurden, A. T. and Kunicki, T. J. (2001). Allele-dependent transcriptional regulation of the human integrin alpha2 gene. Blood. **15**. 1721-1726.
32. Jansen, H., Verhoeven, A. J., Weeks, L., Kastelein, J. J., Halley, D. J., van den Ouweland, A., Jukema, J. W., Seidell, J. C. and Birkenhäger, J. C. (1997). Common C-to-T substitution at position -480 of the hepatic lipase promoter associated with a lowered lipase activity in coronary artery disease patients. Arterioscler Thromb Vasc Biol. **17**. 2837-2842.
33. Okamoto, K., Makino, S., Yoshikawa, Y., Takaki, A., Nagatsuka, Y., Ota, M., Tamiya, G., Kimura, A., Bahram, S. and Inoko, H. (2003). Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. Am J Hum Genet. **72**. 303-312.
34. Berg, L. P., Scopes, D. A., Alhaq, A., Kakkar, V. V. and Cooper, D. N. (1994). Disruption of a binding site for hepatocyte nuclear factor 1 in the protein C gene promoter is associated with hereditary thrombophilia. Hum Mol Genet. **3**. 2147-2152.
35. Zhanglow, X., Miaolow, X., Tanlow, W., Ning, B., Liu, Z., Hong, Y., Songlow ,W., Guolow, Y., Zhanglow, X., Shen, Y., Qiang, B., Kadlubar, F. F. and Linlow, D. (2005) Identification of Functional Genetic Variants in Cyclooxygenase-2 and Their Association With Risk of Esophageal Cancer. Gastroenterology. **129**. 556-576.
36. Nicolas, M., Noe, V. and Ciudad, C. J. (2003). Transcriptional regulation of the human Sp1 gene promoter by the specificity protein (Sp) family members nuclear factor Y (NF-Y) and E2F. Biochem J. **371**. 265-275.

37. Masotti, C., Armelin-Correa, L. M., Splendore, A., Lin, C. J., Barbosa, A., Sogayar, M. C. and Passos-Bueno, M, R. (2005). A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY2 DNA-protein interaction. Gene. **359**. 44-52.
38. Knight, J. C., Udalova, I., Hill, A. V., Greenwood, B. M., Peshu, N., Marsh, K. and Kwiatkowski, D. (1999). A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. Nat Genet. **22**. 145-150.
39. de Laat, W. and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. Chromosome Res. **11**. 447-459.
40. Wasserman, W.W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements.Nat Rev Genet. **5**. 276-287.
41. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep III, P. W. and Bulyk, M. L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotech. **24**. 1429-1435.
42. Narlikar, L., Gordan, R. and Hartemink, A. J. (2007). A nucleosome-guided map of transcription factor binding sites in yeast. PLoS Comput Biol. **3**. 2199-2208.
43. Kim, H., Kechris, K. J. and Hunter, L. (2007). Mining Discriminative Distance Context of Transcription Factor Binding Sites on ChIP Enriched Regions. In ISBRA. 338-349.
44. Lähdesmäki, H., Rust, A. G. and Shmulevich, I. (2008). Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources. PLoS ONE. **3**.
45. Suzuki, M. and Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. Proc Natl Acad Sci USA. **91**. 12357-12361.
46. Torkamani, A., Schork, N. J. (2008). Predicting Functional Regulatory Polymorphisms. Bioinformatics. **24**. 1787-1792.

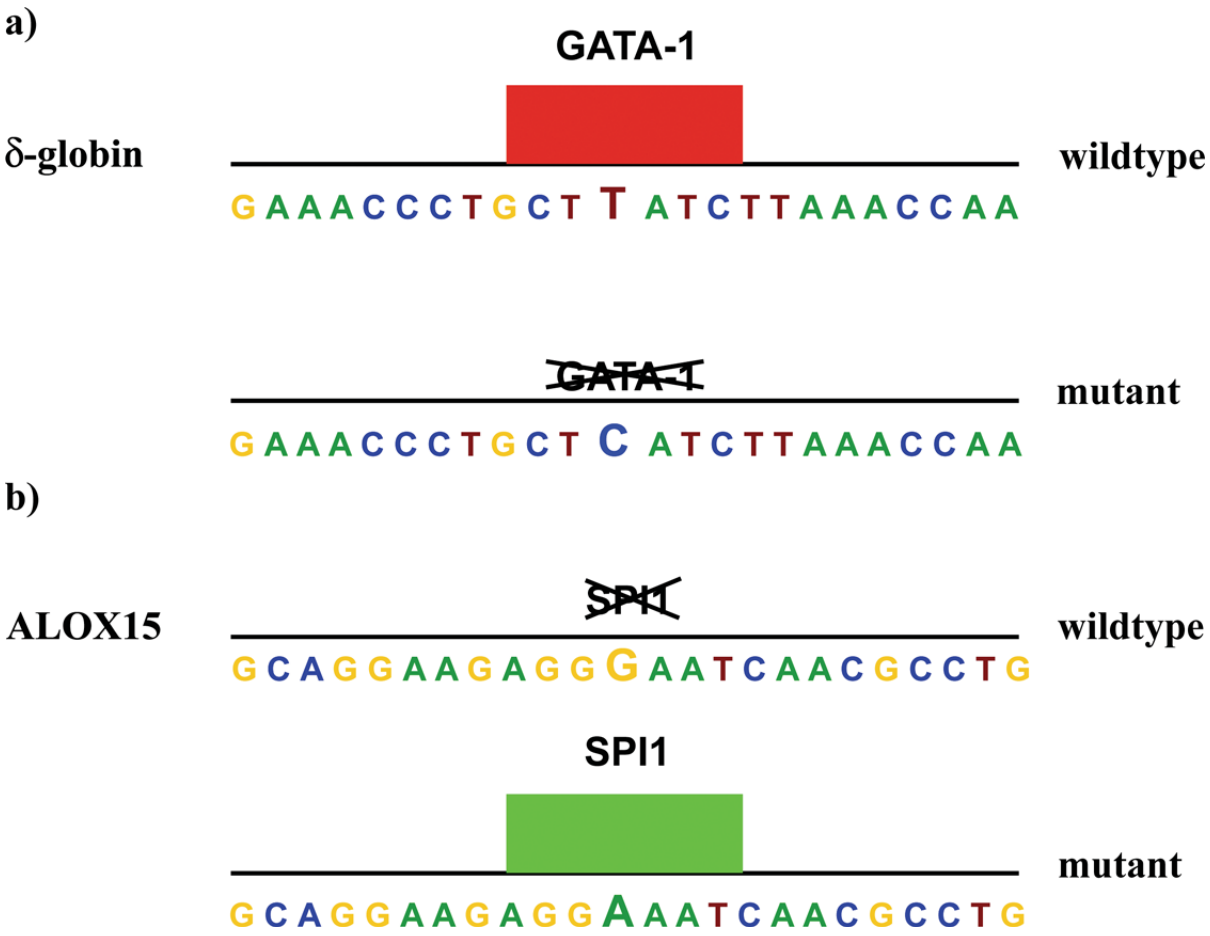
a)
GATA-1
δ-globin
wildtype
GAAACCCTGCT T ATCTTAAACCAA
GATA-1
mutant
GAAACCCTGCT C ATCTTAAACCAA
b)
SPI1
ALOX15
wildtype
GCAGGAAGAGG G AATCAACGCCTG
SPI1
mutant
GCAGGAAGAGG A AATCAACGCCTG

Figure 1.

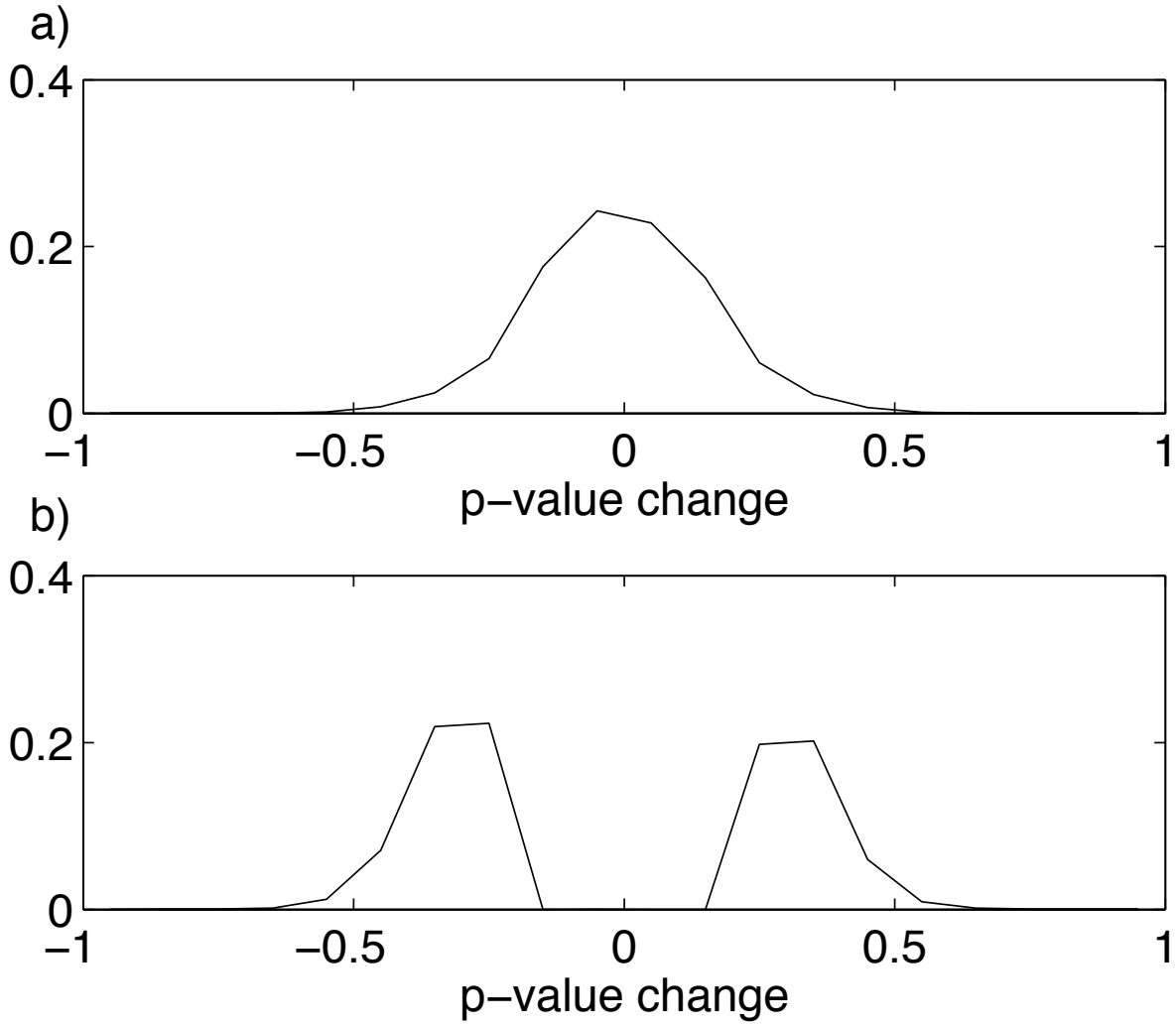
a)
0.4
0.2
0
−1
−0.5
0
0.5
1
p−value change
b)
0.4
0.2
0
−1
−0.5
0
0.5
1
p−value change

Figure 2.

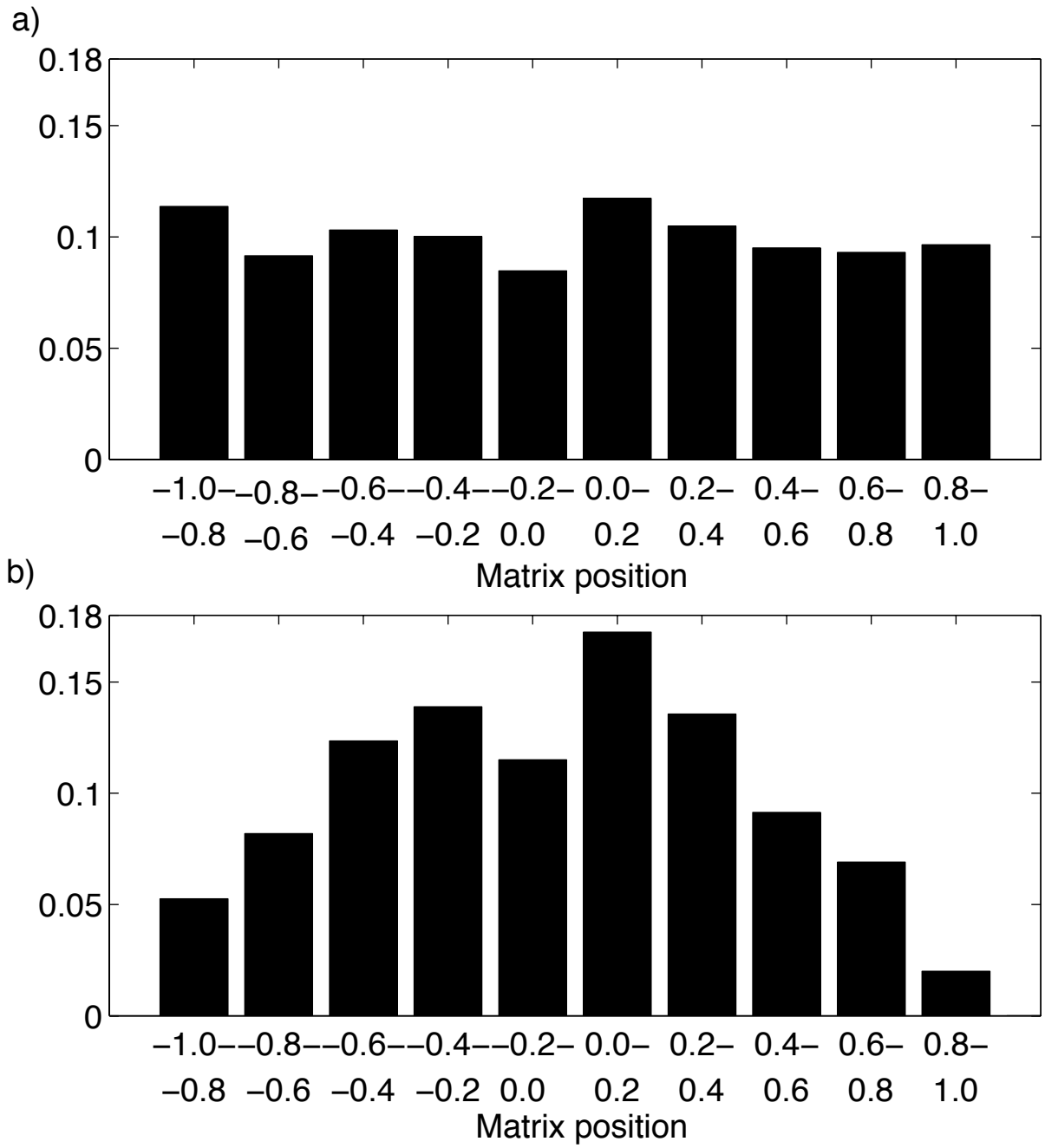
a)
0.18
0.15
0.1
0.05
0
−1.0−0.8
−0.8
−0.8−0.6
−0.6
−0.6−0.4
−0.4
−0.4−0.2
−0.2
−0.2−0.0
0.0
0.0−0.2
0.2
0.2−0.4
0.4
0.4−0.6
0.6
0.6−0.8
0.8
0.8−1.0
1.0
Matrix position
b)
0.18
0.15
0.1
0.05
0
−1.0−0.8
−0.8
−0.8−0.6
−0.6
−0.6−0.4
−0.4
−0.4−0.2
−0.2
−0.2−0.0
0.0
0.0−0.2
0.2
0.2−0.4
0.4
0.4−0.6
0.6
0.6−0.8
0.8
0.8−1.0
1.0
Matrix position

Figure 3.

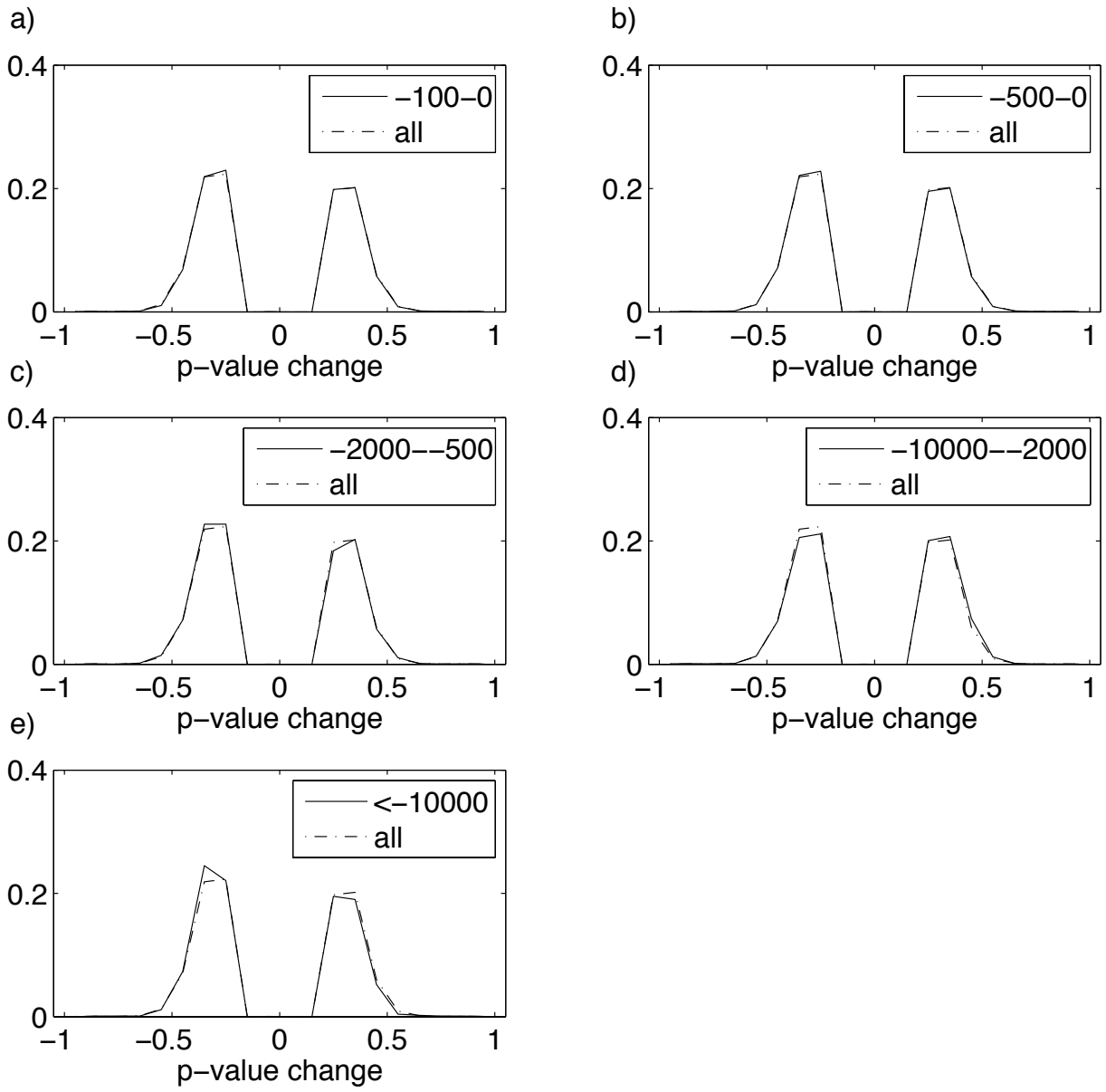
a)
0.4
0.2
0
−1
−0.5
0
0.5
1
p−value change
−100−0
all
b)
−500−0
all
c)
−2000−−500
all
d)
−10000−−2000
all
e)
<−10000
all

Figure 4.

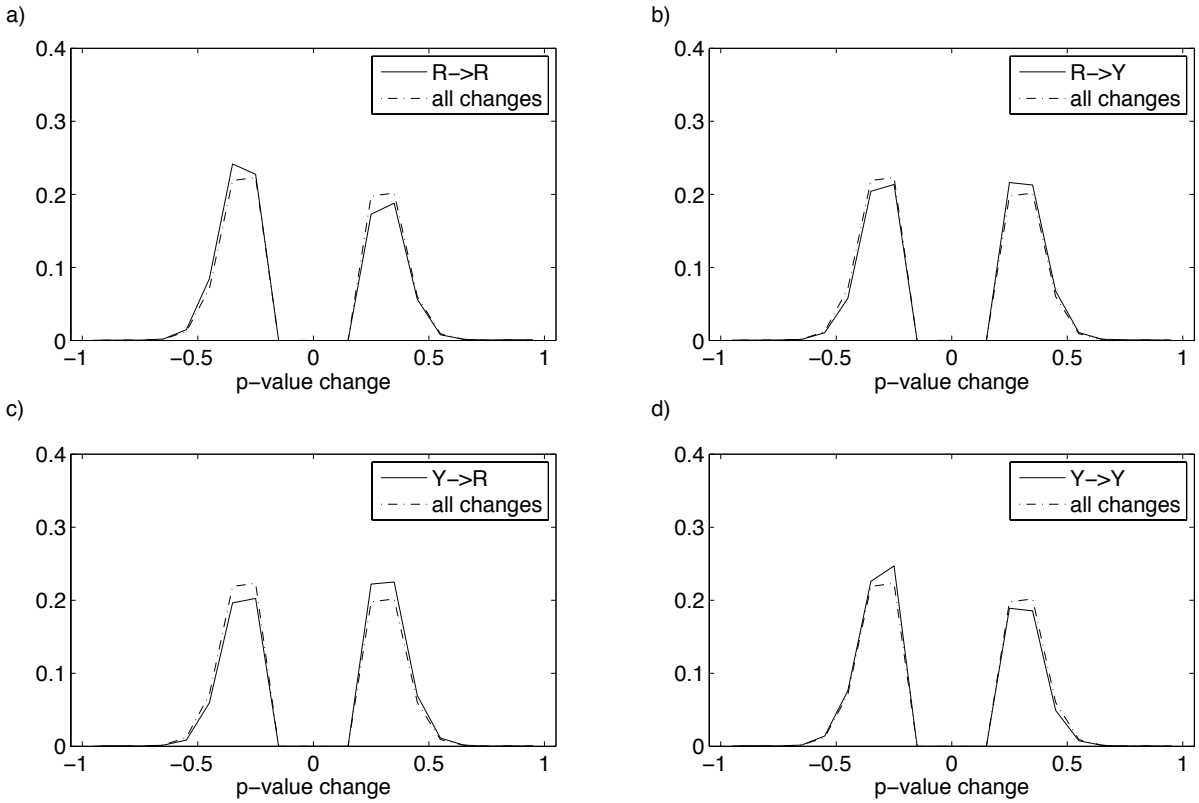
a)
R–>R
all changes
p–value change
b)
R–>Y
all changes
p–value change
c)
Y–>R
all changes
p–value change
d)
Y–>Y
all changes
p–value change

Figure 5.

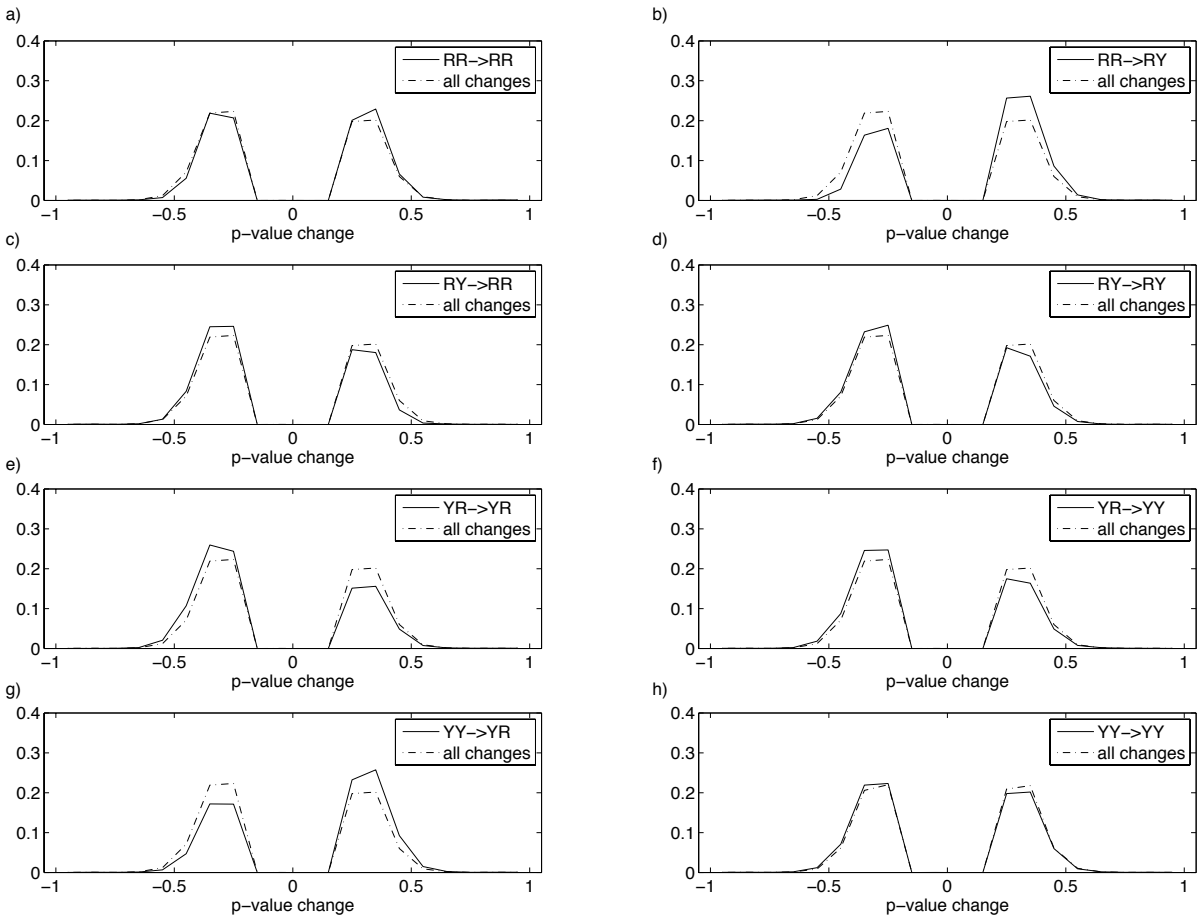
a)
RR–>RR
all changes
b)
RR–>RY
all changes
c)
RY–>RR
all changes
d)
RY–>RY
all changes
e)
YR–>YR
all changes
f)
YR–>YY
all changes
g)
YY–>YR
all changes
h)
YY–>YY
all changes
p–value change

Figure 6.