International Journal of Medical Sciences & Pharmaceutical Research, Volume1, Issue1, Jan-Jun-2017, pp 25-39,

BIOSTATISTICAL ASSESSMENT OF GENIC POSITIONS AND PROFILES OF SIGNIFICANT TRANSCRIPTION FACTOR BINDING SITES IN SINGLE NUCLEOTIDE POLYMORPHISM SENSITIVITY

BEKA, SYLVIA [1] . RENE TE BOEKHORST [2]. IRINA ABNIZOVA [3] BEKA, NATHAN[4]

[1,2] Department of Computer Science, University of Hertfordshire, Hatfield, Hertfordshire

[3] The Wellcome Trust Sanger Institute, Hinxton, Cambridge

[4] Biocomputation Group, Science and Technology Research Institute, University of Hertfordshire, Hatfield, Hertfordshire

ABSTRACT

Human complex diseases, like Diabetes and Cancer, affect many people worldwide today. Despite existing knowledge, many of these diseases are still not preventable. Complex diseases are known to be caused by a combination of genetic factors, as well as environmental and life style factors. The scope of this investigation covered the genomics of Type 1 Diabetes (T1D). There are 49 human genomic regions that are known to carry markers (disease-associated single nucleotide mutations) for T1D, and these were extensively studied in this work. The aim was to find out in how far this disease may be caused by problems in gene regulation rather than in gene coding. For this, the genetic factors associated with T1D, including the single point mutations and susceptibility regions, were characterised on the basis of their genomic attributes. Furthermore, mutations that occur in binding sites for transcription factors were analysed for change in the conspicuousness of their binding region, caused by allele substitution. This is called SNP (Single nucleotide polymorphism) sensitivity. From this study, it was found that the markers for T1D are mostly non-coding SNPs that occur in introns and non-coding gene transcripts, these are structures known to be involved in gene regulatory activity. It was also discovered that the T1D susceptibility regions contain an abundance of intronic, non-coding transcript and regulatory nucleotides, and that they can be split into three distinct groups on the basis of their structural and functional genomic contents. Finally, using an algorithm designed for this study, thirty-seven SNPs that change the representation of their surrounding region were identified. These regulatory mutations are non-associated T1D-SNPs that are mostly characterised by Cytosine to Thymine (C-T) transition mutations. They were found to be closer in average distance to the disease-associated SNPs than other SNPs in binding sites,

and also to occur frequently in the binding motifs for the USF (Upstream stimulatory factor) protein family which is linked to problems in Type 2 diabetes.

INTRODUCTION

A mutation in a regulatory sequence can affect transcription factor binding and as a consequence, the rate of gene transcription. It may lead to an up-mutation, resulting in increased gene expression, or a down-mutation that does the reverse. Clearly, any study devoted to the genomic aspects of a complex disease should take these mutations seriously. In the case of this research work, it forms the core of the study. Variation in regulatory sequences is common (Garfield et al., 2012) and ever more of these mutations have been detected in binding sites over the years (1,969 in 2005 (Guo and Jamison, 2005), 47,832 in 2008(Kim et al., 2008)) (Zheng et al., 2012). According to statistics compiled by the Human Gene Mutation Database , 1909 regulatory mutations have been identified in more than 700 genes that cause human-inherited disorders.

Although some publications mention a possible association of regulatory SNPs with increased risk of T1D (Gillespie and Owen, 2014), until now no study has been done that takes into account the regulatory T1D SNPs. An important objective of this research is therefore to provide an analysis of all regulatory SNPs and, more specifically, to investigate how they might affect the structure of transcription factor binding motifs.

For this, a "SNP sensitivity test" has been developed based on a previous method by Abnizova et al., (unpublished, 2007). The method , assesses the extent to which a mutant allele in a binding site (from now on referred to as a "TFBS-SNP"), compared to its matching reference allele, distorts the representation of the binding motif in which it occurs. Unlike related methods (Schuab et al, 2012; Laurilla and Lahdesmaki, 2009 Laurilla and Lahdesmaki, 2008)that rely on the correctness of computationally identified functional regulatory sequences, this research work uses those SNPs that occur in experimentally confirmed TFBSs (as given by Ensembl's Genome Browser (Cunningham et al., 2014)). This is done to eliminate the problem of false positives associated with the use of computationally predicted binding sites (Beka, 2012; Struckmann et al., 2011).

AIM OF THIS RESEARCH WORK

The aim of testing for SNP sensitivity is to identify T1D-SNPs in binding sites that cause significant change in the binding signal of their local environment (i.e. the binding motif in which they occur).Testing for SNP sensitivity starts with fitting a Markov model[1](Fink, 2007) for the local environment of a SNP. This is a selected part of the regulatory region surrounding a regulatory SNP. The local environment is made up of 601 base pairs, with 300 bps flanking the SNP on both sides. For each Transcription Factor Binding sites (TFBS-SNP), this should typically encompass the binding site that overlaps the SNP .

METHODOLOGY

Change in Signal Representation

A transcription factor binding site is characterised by a special sequence motif which serves as a binding signal for a specific family of transcription factor proteins. In this work, the binding signal will be taken as the direct neighbourhood of each TFBS-SNP. The direct neighbourhood of the SNP is extracted as a 15bps sequence[2], which includes the SNP at the centre of the sequence and 7bps sequences flanking the SNP on both sides .

The representation of the direct neighbourhood (which is assumed to include the binding signal surrounding the SNP) is calculated as a standard residual value$\boldsymbol{SR}$. The value of $\boldsymbol{SR}$ should be calculated on thebasis of the established Markov order of the local environment of the SNP.In order to do this a sequence of steps are taken.

The direct neighbourhood is decomposed into sub-strings of "k-mer words" (i.e. a sequence of $\boldsymbol{k}$ nucleotides) using a single step sliding window method(Fi). In this study, $\boldsymbol{k}$ = 3. This value has been chosen because the Markov order of dependency for regulatory sequences could only be established up to $\boldsymbol{m}$ = 2 (i.e. trinucleotide dependency) The sliding window process generates thirteen trimers

per direct neighbourhood. Subsequently, the expected frequency of "trimers" is derived from the best fitting Markov model of the local environment (Thijs et al., 2001, This et al., 2002).

The expected frequencies of each *i*-th trimer ($\boldsymbol{E_i}$) are compared to the corresponding observed frequencies ($\boldsymbol{O_i}$) by converting them to standard residuals$(\boldsymbol{SR_i = (O_i - E_i)/\sqrt{E_i}})$. The$\boldsymbol{SR}$s indicate if a word is significantly over-represented or under-represented (for$\boldsymbol{SR} > 2.00$ or $\boldsymbol{SR} < -2.00$, respectively). $\boldsymbol{SR}$s are obtained for both alleles of the SNP, i.e. for the same sequence containing the mutant allele and the

reference allele. Finally, the difference between scores for both allelic sequences are tested for statistical significance.

Statistical Significance of D_{max}

The change in over/under-representation of trimers in SNP neighbourhoods is captured by the difference between the standard residuals of the *i*-th word in the neighbourhood of the reference allele and that of the matching mutant allele (ΔSR_i).The biological interpretation of this is that such a change may lead to increased or decreased binding affinity of a transcription factor. Thirteen ΔSR_iscores that are generated for each neighbourhood and are subsequently converted to absolute values. The location of their maximum (D_{max}) indicates the region in the neighbourhood where the highest change in over- or under-representation between the reference and the mutant allele sequence occurs (Figure 1). D_{max}values areobtained for all the TFBS-SNPs, as well as the REG-SNPs and NON-REG-SNPs.

A large D_{max}suggests that SR_r is much greater than SR_m or vice versa (where ***r*** = reference allele and ***m*** = mutant allele. This implies that by switching to the mutant allele of the SNP, a core nucleotide in the motif has been affected, consequently causing substantial change in motif representation. Figure 2 illustrates an example of the change in motif representation caused by switching the alleles of SNP rs200372524, $SR_m$is less than SR_rindicating a decrease in representation. The opposite is for SNP rs3130456, SR_r is less than SR_mindicating that the mutant allele of the SNP causes an increase in motif representation.

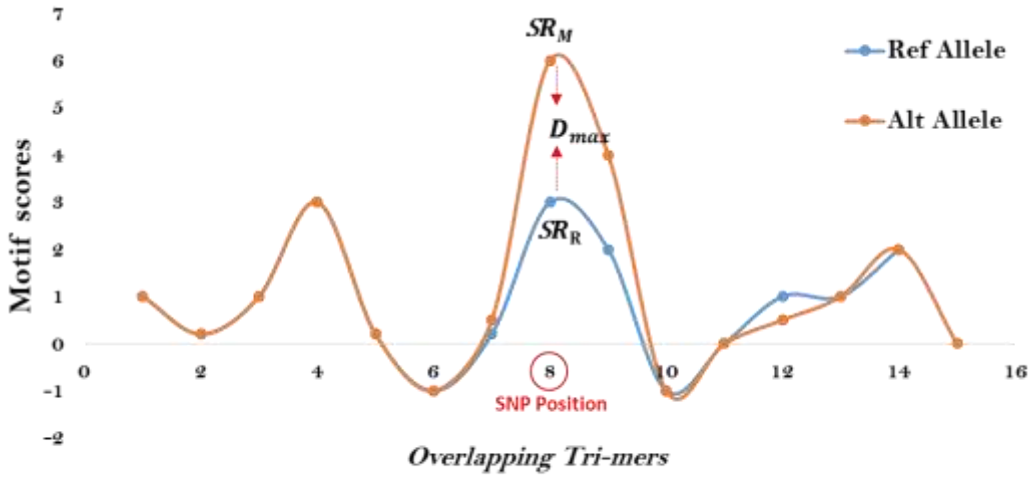

Figure 1. Location of D_{max} , the largest change in over- or under-representation between the reference and the mutant allele sequence

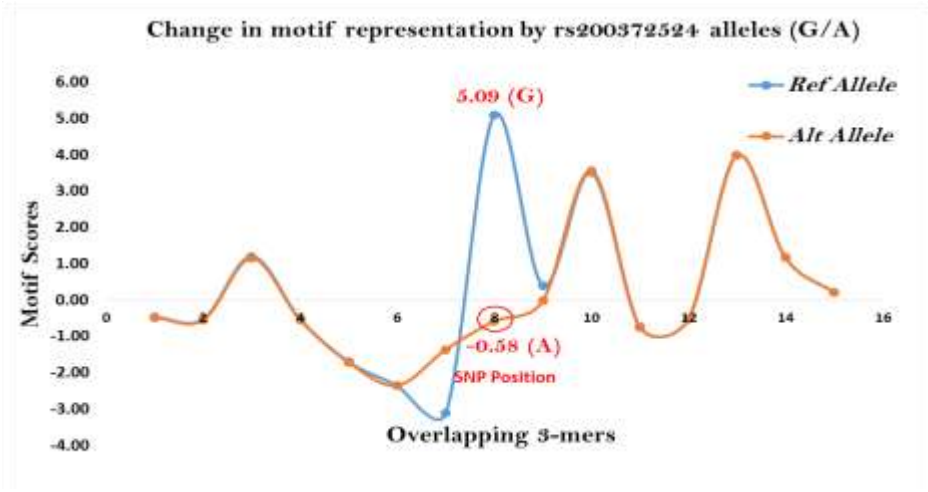

Figure 2. A decrease in motif representation caused by the substitution of alleles of SNP rs200372524 in its local environment, $SR_R > SR_M$

The statistical significance of each D_{max} score was determined so as to assess in how far the change in representation is due to chance. To do this, each Deoxylnitribuneclic acid (DNA) sequence was reshuffled 5000 times to yield random permutations of the same sequence. The D_{max} score was obtained for each permuted sequence following the same procedure used to obtain the original D_{max} score. The original D_{max} is considered to be significant when it is larger than 4750 (95%) of the D_{max} scores of the

permuted sequences. This corresponds to an empirical p-value cut off of 0.05.Those SNPs resulting in a p< 0.05, test positive for SNP sensitivity and are associated withsubstantial change in the trimers making up their direct neighbourhood when alleles are substituted. For those cases in which the Markov order of the sequence could not be established, a SNP was considered significant if at least any of the three D_{max} values (computed for m = 0, 1, and 2) are significant. SNPs that test positive for SNP sensitivity (from now on referred to as significant TFBS-SNPs and distinguished as such from non-significant TFBS-SNPs) are those with mutant alleles that have the potential to distort the recognition of the binding motif. They will be selected as candidate functional mutations with the propensity to disturb transcriptional regulation.

RESULT AND DISCUSSION

Identity and Location of significant TFBS SNPs

In this study, 37 out of 92 TFBS-SNPs were found to test positive for SNP-sensitivity. The names and alleles of these significant SNPs, the susceptibility regions in which they occur, the degree of sensitivity (D_{max}), and the values of features i) and iii) are shown in Table 1.

Identity of significant TFBS-SNPs: With respect to the identity of single nucleotide substitution, two types of mutations are generally distinguished. Transitions (TI) are SNPs of which the reference and mutant allele are of the same nucleotide class, i.e. both are either a pyrimidine (C, T) or a purine (G, A). Hence, transitions are C-T and G-A SNPs (and the reverses T-C and A-G). Transversions (TV) are SNPs in which a purine is substituted by a pyrimidine (i.e. C-G, G-C, A-T and T-A).

Table 1 The Names and features of significant TFBS-SNPs, including p-values and details of nearby disease associated SNPs. D_{max} values are absolute.

SNP ID	Alleles	T1D Region	Nearby associated SNP	Distance to associated SNP (Bps)	Region Size	$\|D_{Max}\|$
rs138680304	C/T	2p23.3	rs478222	91600	468897	4.6429
rs114096282	C/T	2p23.3	rs478222	13274	468897	6.9658
rs117640654	G/A	2p23.3	rs478222	75075	468897	3.0526
rs377664089	G/T	3p21.31	rs333	124049	599694	6.0142
rs34638008	C/T	3p21.31	rs333	131238	599694	3.6256
rs140935015	T/C	MHC	rs9268645	1350064	3808585	5.5389
rs140000554	T/A	MHC	rs9268645	1996519	3808585	3.4372
rs151190212	C/G	MHC	rs9268645	621897	3808585	7.0912
rs2267646	G/T	MHC	rs9268645	561683	3808585	4.7346
rs3134944	C/T	MHC	rs9268645	229660	3808585	4.1205
rs35131721	C/T	MHC	rs9268645	831989	3808585	4.1995
rs7741418	C/T	MHC	rs9268645	2312571	3808585	4.1217
rs3130288	C/A	MHC	rs9268645	280303	3808585	3.9943
rs116431137	A/G	MHC	rs9268645	2594487	3808585	3.7497
rs56245106	T/C	MHC	rs9268645	201542	3808585	3.9326
rs201033718	G/C	MHC	rs9268645	449305	3808585	3.6274
rs6921948	A/C	MHC	rs9268645	1205047	3808585	3.4668
rs9262142	G/A	MHC	rs9268645	1726278	3808585	0.0221
rs8192582	C/T	MHC	rs9268645	212692	3808585	3.4722
rs8192581	C/T	MHC	rs9268645	212640	3808585	2.8728
rs13206219	G/T	MHC	rs9268645	201543	3808585	3.3090
rs78180266	C/T	7p12.2	rs10272724	58315	299719	3.8173
rs188548927	C/T	7p12.2	rs10272724	58315	299719	4.5298
rs182785851	G/A	7p15.2	rs7804356	245105	544327	4.1407
rs184649955	C/T	12q13.2	rs2292239	39710	446498	4.0808
rs141305257	C/G	16p11.2	rs4788084	47776	730672	4.3615
rs7203793	C/G	16p13.3	rs12708716	91596	449453	3.7016
rs371243647	C/T	16p13.3	rs12927355	61513	449453	3.4342
rs139221703	G/A	16p13.3	rs12927356	60111	449453	3.4101
rs187731105	G/A	16p13.3	rs12927357	59878	449453	3.4749
rs191450302	C/A	16q23.1	rs8056814	265264	304790	3.4001
rs200372524	G/A	19p13.2	rs2304256	14159	237839	5.6772
rs201991101	C/T	19p13.2	rs2304256	13749	237839	3.2587
rs371391397	C/A	19p13.2	rs2304256	90052	237839	4.0387
rs372996186	G/C	19p13.2	rs2304256	70838	237839	4.9207
rs201432982	C/T	19p13.2	rs2304256	54719	237839	2.6915
rs141193051	G/C	19p13.2	rs2304256	27736	237839	3.3260

The first thing one may notice in Table 1 is that the majority of significant TFBS-SNPs are C-T mutations, which reflects the dominance of transitions (see Table 2). This is not surprising, because transitions in general are more common than transversions. Furthermore, Laurilla and Lahdesmaki (2009) report that C-T transitions are among the most effective SNPs in terms of weakening transcription factor binding, the nucleotide substitution types are inventoried for significant (S) and non-significant (NS) TFBS-SNPs. Although NS-TFBS-SNPs appear to have more G-A, A-G mutations and less C-T

transitions than S-TFBS-SNPs, there is no significant difference between the two categories (NS, S) concerning mutant composition (χ^2 = 5.16, df = 5, p = 0.40).

Table 2. Counts of TFBS-SNP nucleotide substitution types

		TI			TI Total				TV			TV Total	Grand Total
T1D Susc Region	**A/G**	**G/A**	**C/T**	**T/C**		**A/C**	**C/A**	**C/G**	**G/C**	**G/T**	**T/A**		
12q13.2			1		1								1
16p11.2								1				1	1
16p13.3		2	1		3			1				1	4
16q23.1							1					1	1
19p13.2		1	2		3		1		2			3	6
2p23.3		1	2		3								3
3p21.31			1		1					1		1	2
7p12.2			2		2								2
7p15.2		1			1								1
MHC	1	1	5	2	9	1	1	1	1	2	1	7	16
Grand Total	1	6		2	23	1	3	3	3	3	1	14	37

Location of significant TFBS-SNPs in susceptibility regions: Another striking feature seen in Table 1 is that most of the significant TFBS-SNPs are found in the MHC the Human Leucocyte Antigen (HLA) region, which has been shown to have the associated strongest with T1D . However, this is simply a consequence of its large size; the MHC is made up of about ten times more nucleotides than the other regions. on the basis of its genomic components the MHC is placed in the second cluster, which indeed includes susceptibility regions with the highest SNP density. However, those regions do not have the highest density of significant TFBS -NPs. These happens to be in cluster 3 (CL 3), the one characterised by an abundance of non-coding nucleotides including regulatory DNA (Table 3). This cluster is also composed of regions with on average the highest number of both disease-associated T1D-SNPs and markers for other autoimmune diseases (i.e. “pleiotropic” regions).

Table 3 . Counts of Disease-associated T1D-SNPs and TFBS-SNPs in clusters of T1D susceptibility regions.

Cluster Name	CL1	CL2	CL3
Number of Regions (N)	12	24	10
Counts			
Associated SNPs	19	41	19
TFBS-SNPs	12	43	32
Significant TFBS-SNPs	4	18	12
Normalised values			
Associated SNPs	1.58	1.70	1.90
TFBS-SNPs	1.00	1.79	3.20
Significant TFBS-SNPS	0.33	0.75	1.20

Genic positions and –profiles of significant TFBS SNPs: As, a single SNP can affect more than one gene and intersect multiple transcripts. To see in how far this holds for SNPs that significantly change the motif structure of binding sites, the number of genes and transcripts intersected by each TFBS-SNP was counted. Non-significant TFBS-SNPs (N = 55) appear to affect more transcripts than significant TFBS-SNPs (N = 37) (p = 3.74E-06) (Figure 3). This might be due to a possible relationship between gene size and the number of transcripts that gene can produce (large genes contain more SNPs and more transcripts).

The characterised the TFBS-SNPs by their genic-profiles. Recall that a genic profile comprises the name of each unique type of the genic position in which a SNP occurs. The identification of genic-profiles typical for significant TFBS-SNPs is illustrated in Figure 4. Profiles that are more typical for either significant- or non-significant TFBS -NPs differ more strongly in their proportions, and are therefore farther away from the diagonal line in Figure 4 (i.e. those profiles that constitute identical proportions of significant- and non-significant SNPs fall along the diagonal).

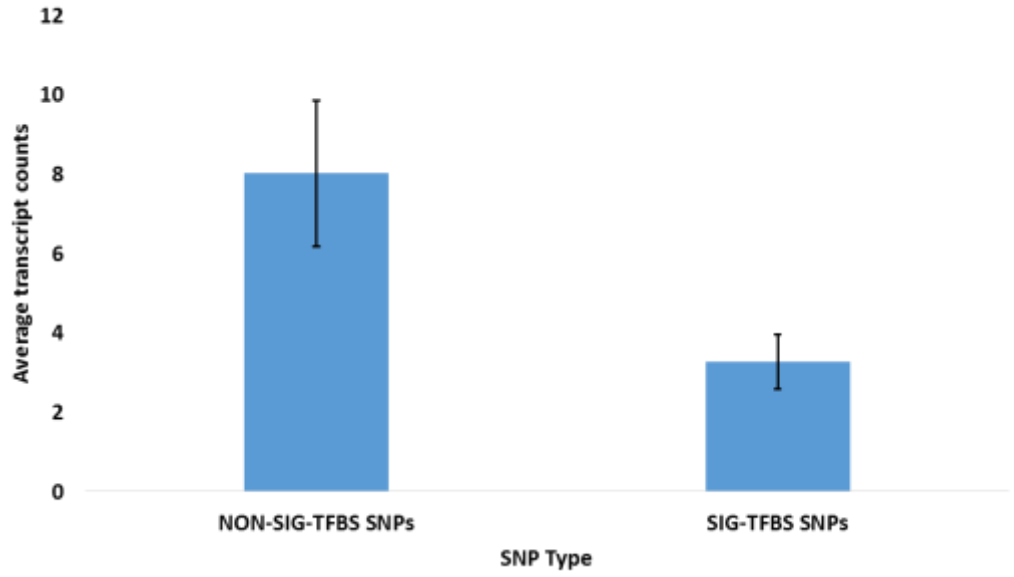

Figure 3 Average number of gene transcripts affected by significantTFBS-SNPs (N= 37) and non-significant TFBS-SNPs N= 55).

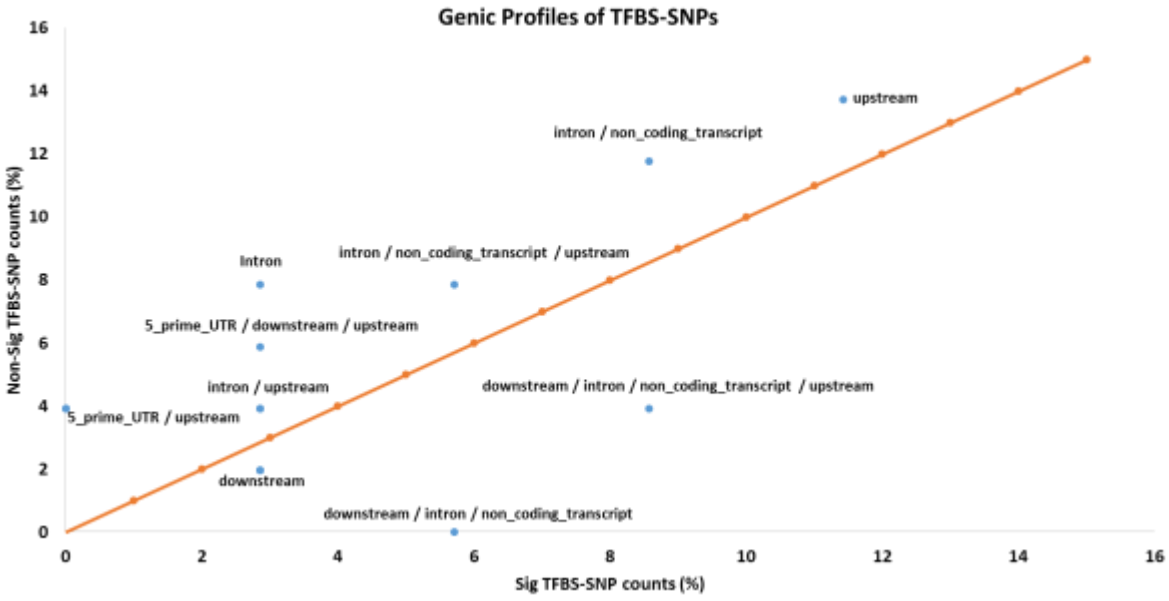

Figure 4 .Isolation of genic-profiles common to the significant and non-significant TFBS-SNPs using a scatter plot

Most of the SNPs are in upstream regions of genes, this is where TFBSs are most likely to be found (Table 4). But significant TFBS-SNPs do affect other genomic parts as well, including introns and non-coding transcripts. Note that the same was found for the disease-associated SNPs. This may suggest that

the disease-associated SNPs and TFBS-SNPs are close to each other. Also, all the components of the genic profiles typical for significant TFBS-SNPs (i.e. upstream, intronic, non-coding transcripts and downstream positions) are parts that are exclusively associated with regulatory activity. In other words, apart from affecting the binding motif in which they occur, some of the significant TFBS-SNPs may have an additional effect on overlaying transcripts.

Table 4 . The most frequent genic-profiles of the significant and non-significant TFBS-SNPs

Genic Profile	SIG TFBS-SNPS		Non-SIG TFBS-SNPs	
	Counts	*(%)*	*Counts*	*(%)*
upstream	4	11.43	7	13.73
intron / non_coding_transcript	3	8.57	6	11.76
intron / non_coding_transcript / upstream	2	5.71	4	7.84
downstream / intron / non_coding_transcript / upstream	3	8.57	2	3.92
intron	1	2.86	4	7.84
5_prime_UTR / downstream / upstream	1	2.86	3	5.88
intron / upstream	1	2.86	2	3.92
missense / non_coding_transcript_exon / non_coding_transcript / upstream	1	2.86	2	3.92
downstream / intron / non_coding_transcript	2	5.71	0	0.00
downstream	1	2.86	1	1.96
downstream / intron / non_coding_transcript_exon / non_coding_transcript	1	2.86	1	1.96
intron / NMD_transcript	1	2.86	1	1.96
intron / NMD_transcript / non_coding_transcript	1	2.86	1	1.96
5_prime_UTR / intron / non_coding_transcript / upstream	0	0.00	2	3.92
5_prime_UTR / upstream	0	0.00	2	3.92
downstream / intron / upstream	0	0.00	2	3.92
Total	35		51	

Localisation of significant TFBS-SNPs relative to disease-associated SNPs: Although none of the TFBS-SNPs show up as being statistically associated with T1D in Genome Wide Association Studies (GWAS)c, this should not be taken as proof for a lack of causality. Current research supports rather the opposite view: disease-associated SNPs, instead of being causative, might be no more than just markers for a region of disease association. As such, any other mutation within that region is a putative causal variant. Therefore, many genomic studies nowadays aim to identify other (potentially causal) SNPs that occur in close proximity and linkage with disease-associated SNPs (Schuab et al., 2012). With this in mind, the distance (bps) between the disease-associated SNPs and both the significant and non-significant TFBS-SNPs were compared. The hypothesis is that significant TFBS variants are in closer in proximity to disease-associated SNPs than non-significant TFBS-SNPs.

Indeed, the average distance between significant TFBS-SNPs and nearby disease-associated SNPs (172185 bps) turns out to be less than the average distance between the non-significant TFBS-SNPs and nearby disease-associated SNPs (355876 bps). The relationship was tested by means of a two way

ANOVA in which the possible effect of susceptibility region. The strong effect of susceptibility region is due to a large difference in the average SNP distance in the MHC/HLA (Figure 5). If the HLA is taken out, the effect of region disappears but the difference between groups still remains significant (p = 0.040).

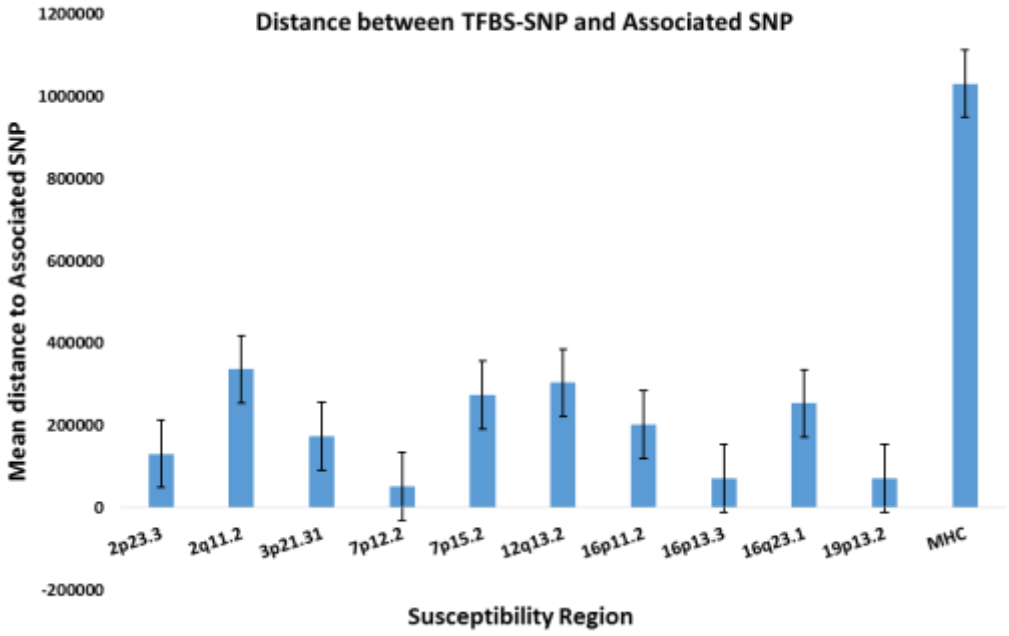

Figure 5 . Plot showing differences in average distances to disease-associated SNPs between susceptibility regions.

Binding motifs in which the significant TFBS-SNP occur

For each TFBS-SNP, the names of binding motifs in which it occurs, and the family of transcription factors that recognise and bind to those motifs, were obtained from the Jasper database (Mathelier et al., 2014) (viathe Ensembl genome browser). This was done to identify families of transcription factor proteins that might distinguish significant TFBS-SNPs from non-significant TFBS-SNPs. These proteins will then be briefly described in relation to T1D.

From the study, the TFBS-SNPs occur in a total of 31 different binding motifs. Eighteen different transcription factor protein families bind to these motifs. Some of the transcription factors, like JunD and USF, bind to more than one type of motif. These are transcription factors that display diverse target specificity and so have more than a single motif model (Mathelieret al., 2014). A scatter plot of the proportion of significant TFBS-SNPs in each type of binding motif against the proportion of non-significant TFBS-SNPs is depicted in Figure 6. The plot highlights the binding motifs more specific to either of both TFBS-SNP categories. The significant TFBS-SNPs occur more frequently, and twice as

much as the non-significant TFBS-SNPs, in binding motifs for the Upstream transcription factor 1 (USF1). The significant TFBS-SNPs also have a high occurrence in binding motifs for the E2F4 transcription factor.

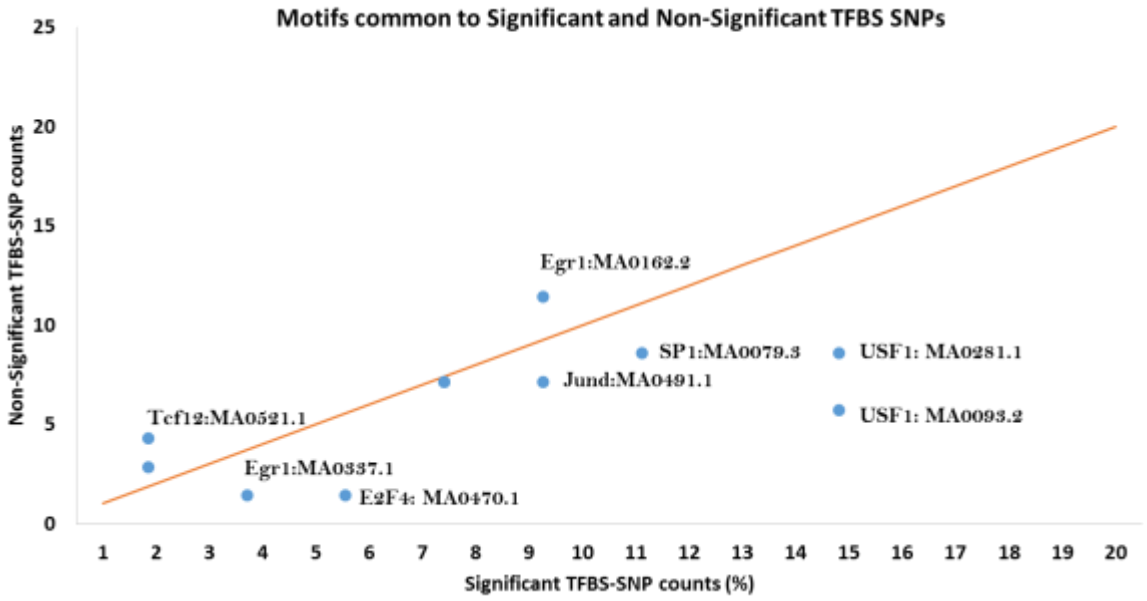

Figure 6 . Scatter plot depicting counts of significant and non-significant SNPs in different binding motif structures

The USF1 protein is a cellular transcription factor (Shieh et al., 1993; Corre and Gallibert, 2006) that is thought to activate transcription through binding enhancer (E)-box motifs. (Corre and Gallibert, 2006).Proteins that bind E-box motifs are said to play a major role in regulating gene transcriptional activity.

The target genes of USF1 include genes that contribute to the regulation of glucose and lipid metabolism. (Fan et al., 2014,). Along with another transcription factor, USF2, the USF1 protein has been found to be important for the regulation of different pancreatic islet genes involved in the control of glucose metabolism (Boonsaen et al., 2007).Already, this protein has been linked to other forms of diabetes. The locus of USF1in humans is associated with increased risk of developing Type 2 diabetes . It is also associated with maturity onset diabetes of the young (MODY) (Bernardo et al, 2008; Qian et al., 1999). More recent studies also link upstream Transcription Factor (USF1 and USF2) to activation of the promoter for the Alx3 gene (Mirasierra et al., 2011). Expression of Alx3 is required for maintaining adequate levels of expression of pancreatic islet genes including insulin; Alx3 loss-of-function in mice

models have shown a progressive decrease in pancreatic islet cell mass and alterations in glucose homoeostasis[3] leading to diabetes.

Variants that affect regulatory functions have been recognised in the aetiology of certain diseases. Some examples include the blood related diseases β-thalassemia andhaemophilia, atherosclerosis , as well as Gilberts syndrome in humans (Bosma et al., 1995). But for T1D, no regulatory SNPs have yet been implicated in the disease mechanism. A SNP that occurs in an experimentally detected binding site, and is closely linked with a disease associated SNP, is more likely to play a biological role in the genome than other SNPs that occur in parts for which there is no particular known function (Schuab et al., 2012). Through this work, It was found that though the associated T1D-SNPs are not regulatory SNPs that may influence transcription factor binding, there are other nearby non-associated SNPs that can influence this process. Thirty-seven of these rare regulatory TFBS-SNPs have been identified by their testing positive for SNP sensitivity. In addition to significantly changing the representation of their local environment, they are significantly closer to disease-associated SNPs than the other TFBS-SNPs. The significant TFBS-SNPs are mostly characterised by C-T transitions, which have previously been shown to cause weaker affinity for transcription factor (TF) binding.

Significant and non-significant TFBS SNPs influence 31 different binding sites for 18 transcription factor families. The binding sites for the USF family of transcription factors are themost affected; these proteins, USF1 and USF2, have been linked to genetic disorders involving the regulation of insulin genes and of the metabolism of glucose. These are typical features of T1D, where insulin is primary auto-antigen[4]. Despite these important findings, further researchis needed to determine whether these SNPs do affect function in vivo. Experimentation can reveal if the recognition and binding of TFs to the binding sites in which the significant TFBS-SNPs occur is altered, and how this in turn disturbs the transcription of target genes.

Although the aetiology of T1D is not fully understood, aberrations in the regulation of certain susceptibility genes, like CTLA-4, PTPN22 and IFIH19 are suspected to contribute to the cause of disease . For T1D to be viewed as a disease that is caused by problems in gene regulation, the classical expectation would be for the disease-associated SNPs to be frequently located in regulatory regions and binding sites. Although this is not the case, this does not rule out the hypothesis that T1D is for a large part due to gene regulatory defects. The disease-associated SNPs occur the least often (7%) in protein coding regions; therefore, T1D cannot be described as a disease that is caused by disruption in protein

coding alone. The findings of this research can be related to two current trends in the study of complex diseases. The first is centred on the function of disease-associated SNPs in susceptibility regions. It is now widely suggested that the disease-associated SNPs may be no more than markers that capture the variation present at a locus associated with disease risk (i.e. the disease susceptibility region). They are unlikely to be the mutations that underlie disease association, but rather are in linkage with other genetic variants, all of which are putatively causal (Marian, 2012; Schuab et al., 2012). This recent turn in complex disease genomics has come about because despite a number of post Genome wide association studies (GWAS), many of the disease associated SNPs are still yet to be implicated as the underlying causal variant in associated complex diseases. .

To conclude, in complex diseases studies, discovering a contributing factor and then characterizing its contribution to the disease is not quite an easy undertaking. This research initially set out to describe the SNPs associated with susceptibility to T1D as variants that cause disease by influencing transcription factor binding. This was not found to be. Instead, nearby non-associated T1D-SNPs were identified as the putative causal regulatory SNPs that could impact binding. The associated SNPs are either likely to influence regulation through other alternative processes or they are markers that have led to the identification of potential causal SNPs.

REFERENCES

1. Abnizova, I., Foco, L., te Boekhorst, R., Bernardinelli, L. (2007). Sequence-oriented epidemiology: Regulatory SNPs associated with disease can be inferred by DNA sequence information directly. [Unpublished Work].

2. Beka, .N (2012) Web interface Development for a Biological Database. Un Published B.Sc Project report. University of Hertfordshire, U.K

3. Bernardo, A. S., Hay, C. W. and Docherty, K. (2008) Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic β cell. *Molecular and Cellular Endocrinology, 294*, 1-9.

4. Boonsaen, T., Rojvirat, P., Surinya, K. H., Wallace, J. C., Jitrapakdee, S. (2007). Transcriptional regulation of the distal promoter of the rat pyruvate carboxylase gene by hepatocyte nuclear factor 3β/Foxa2 and upstream stimulatory factors in insulinoma cells. *Biochemical Journal, 405*(Pt 2), 359–367. doi:10.1042/BJ20070276.

5. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.… Flicek, P. (2014), Ensembl 2015. *Nucleic Acids Research, 42*(22). doi: 10.1093/nar/gku1010.

6. Corre, S. and Galibert, M.D. (2006). USF as a key regulatory element of gene expression. *Médecine Sciences Paris, 22* (1), 62–67. doi:10.1051/medsci/200622162.

7. Fan, Y.-M., Hernesniemi, J., Oksala, N., Levula, M., Raitoharju, E., Collings, A., Hutri-Kähönen, N., Juonala, M., Marniemi, J., Lyytikäinen, L.P.,… Lehtimäki, T. (2014). Upstream Transcription Factor 1 (USF1) allelic variants regulate lipoprotein metabolism in women and USF1 expression in atherosclerotic plaque. *Scientific Reports, 4*, 4650. doi:10.1038/srep04650.

8. Fink, G.A. (2007). *Markov Models for Pattern Recognition: From Theory to Applications*. Springer: New York.

9. Garfield, D., Haygood, R., Nielsen, W., Wray, G. (2012). Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin Strongylocentrotus purpuratus. *Evolution and Development, 14*(2), 152-167. doi: 10.1111/j.1525-142X.2012.00532.x.

10. Guo, Y. and Jamison, D. C. (2005). The distribution of SNPs in human gene regulatory regions. *BMC Genomics, 6*, 140. doi:10.1186/1471-2164-6-140.

11. Gillespie, K. and Owen, K. (2014). IL2RA. *Diapedia*, 2104135143(12). [Online]. Available at: http://dx.doi.org/10.14496/dia.2104135143.12.[Accessed 02 February, 2015].

12. Kim, T. H. and Ren, B. (2006). Genome-Wide Analysis of Protein-DNA Interactions. *Annual Review of Genomics and Human Genetics, 7*, 81-102.

13. 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), pp. 89-92.

14. Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research, 42*, 142-147.

15. Mirasierra, M., Fernández-Pérez, A., Díaz-Prieto, N., Vallejo, M. (2011) Alx3-deficient mice exhibit decreased insulin in beta cells, altered glucose homeostasis and increased apoptosis in pancreatic islets. *Diabetologia, 54*, 403-414.

16 Marian, A.J. (2012). Molecular genetic studies of complex phenotypes. *Translational Research, 159*(2):64-79. doi: 10.1016/j.trsl.2011.08.001.

17 .Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome *Research, 22*(9), 1748–1759. doi:10.1101/gr.136127.111.

18. Struckmann, S., Esch, D., Schöler, H., Fuellen, G. (2011). Visualization and Exploration of Conserved Regulatory Modules Using ReXSpecies 2. *BMC Evolutionary Biology, 11*, 267. doi:10.1186/1471-2148-11-267

19. Shieh, B.H., Sparkes, R.S., Gaynor, R.B., Lusis, A.J. (1993). Localization of the gene-encoding upstream stimulatory factor (USF) to human chromosome 1q22-q23. *Genomics, 16*(1), 266–268. doi:10.1006/geno.1993.1174.

20. Zheng, W., Gianoulis, T.A., Karczewski, K.J., Zhao, H., Snyder, M. (2010). Regulatory variation within and between species. *Annual Review of Genomics and Human Genetics, 12*, 327-346.