Int J Mol Epidemiol Genet 2016;7(3):126-136
www.ijmeg.org /ISSN:1948-1756/IJMEG0025402

Original Article

Methodology for single nucleotide polymorphism selection in promoter regions for clinical use. An example of its applicability

Herlander Marques[1,2*], José Freitas[3*], Rui Medeiros[4*], Adhemar Longatto-Filho[1,5*]

[1]Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal; ICVS/3B's-PT Government Associate Laboratory, Braga/Guimarães, Portugal; [2]Department of Oncology, Hospital de Braga, Braga, Portugal; [3]Nova Medical School, New University of Lisbon, Lisbon, Portugal; [4]Molecular Oncology Group & Virology LB-CI, Portuguese Institute of Oncology, Porto, Portugal; ICBAS, Abel Salazar Institute for the Biomedical Sciences, University of Porto, Porto, Portugal; CEBIMED, Faculty of Health Sciences of Fernando Pessoa University, Porto, Portugal; PCC, Research Department-Portuguese League Against Cancer (NRNorte), Porto, Portugal; [5]Laboratory of Medical Investigation (LIM) 14, Faculty of Medicine, University of São Paulo, São Paulo, Brazil. [*]Equal contributors.

Received February 1, 2016; Accepted September 1, 2016; Epub September 30, 2016; Published October 15, 2016

Abstract: Genetic variability in humans can explain many differences in disease risk factors. Polymorphism-related studies focus mainly on the single nucleotide polymorphisms (SNPs) of coding regions of the genes. SNPs on DNA binding motifs of the promoter region have been less explored. On a recent study of SNPs in patients with non-Hodgkin lymphomas we faced the problem of SNP selection from promoter regions and developed a practical methodology for clinical studies. The process consists in identifying SNPs in the coding and promoter regions of the antigen-processing system using the 'dbSNP' database. With the 'HapMap' program, we select SNPs with frequencies >20% in Caucasian populations. For coding regions, we sought biologically and clinically relevant SNPs described in the literature. For the promoter regions, we determined their chromosomal location on 'QiagenSABioscience' site database. The nucleotide sequence of ancestral and variant alleles is available in the 'dbSNP'. These sequences were used in 'Promoter TESS' to determine binding differences of transcription factors. Each sequence may have affinity to different TFs. Thus, SNP selection on the promoter regions was based in the differences on TF binding pattern between the old and the new allele. The potential clinical relevance of the new TFs was also evaluated before the final selection. With this approach, we found that almost half of the relevant SNP fall within the promoter region. In conclusion, we were able to develop a methodology of oriented selection of promoter regions of human genes, comparing the TF with affinity to the ancestral allele with the TF to a variant allele. We selected those SNPs that change the TF's affinity to a pattern with functional significance.

Keywords: Genetic polymorphism, SNP, DNA binding motifs, promoter region, coding region, dbSNP, HapMap, Promoter TESS

Introduction

The Human Genome Project, an international undertaking involving many research labs in several countries, published the complete sequence of the human genome in 2003. The complexity of genetic expression is not only due to gene sequence, but also to extensive interactions between DNA, RNA and proteins, and to the transcriptional regulation and polymorphic variations in human genes, such as SNP (single nucleotide polymorphism). To be considered a polymorphism, the SNP must be present in at least 1% of the population. SNPs are not considered mutations because often they do not change the phenotype significantly, however they can be responsible for diseases when the new amino acid causes even a small alteration in protein function.

The genetic polymorphic variations are currently a major concern because they have been associated to multiple human diseases and can also explain many of individual differences

in behaviour, drug reaction and other biological processes.

Although the main target of polymorphism research in humans has been the coding regions of genes, the National Human Genome Research Institute and its branch ENCODE reported recently that 20% of the non-coding DNA is functional and involved in gene regulation. It is relevant that 90% of sequence alterations are located outside the coding region and maybe associated with diseases [1].

The problem of SNP selection led the authors to study genetic variants associated with the risk of Non-Hodgkin Lymphoma (NHL). These cancer studies of SNP polymorphisms have been done in genes of multiple systems, such as metabolic pathways, biotransformation enzymes, DNA repair genes, folate metabolism, interleukins and immune function proteins. Many of them have been associated with risk or prognosis of the disease [2-8]. However, individual variability in proteins of immune response as antigen processing and presentation may be implicated but were never studied. For that purpose the authors tried to develop a more accurate method to identify SNPs with functional importance in the genes of those immune response proteins.

In the context of malignant diseases, the immunological response depends on class-I Major Histocompatibility Complex (MHC) and on an elaborate enzymatic system required for antigen presentation, including ubiquitin-proteasome (UB) system and transporter associated with antigen processing (TAP).

When identifying SNP in our target proteins, it was easy to locate and select those found in the gene coding area, by consulting the relevant biomedical literature. However, the search and selection of SNPs on the promoter region and on other regulating components raised very complex questions that led us to develop our own methods described below. The problem in studying SNPs in these regions is that they do not change the protein structure because they are not coding regions, but they do control transcription. Transcriptional regulation involves proteins and TFs that bind to short regulatory sequences, or motifs, in the promoter region, also known as transcription factors binding site (TFBS). A SNP in these regions can change the affinity for the usual TFs and in certain situations lead to the introduction of new ones. Therefore, the question becomes "which are now the new TFs for a SNP in the promoter region of interest and which TFs have lost their affinity"?

This is even more complex given that a TF binding to DNA motifs does not occur according to a digital pattern of all or nothing, but has an analogical, probabilistic pattern [9, 10]. The presence of a SNP in a promoter region can reduce, but not eliminate, the affinity of certain TFs to an ideal nucleotide sequence. The probabilistic pattern of ligation protein-DNA predicts, for example, the affinity of a TF to different TFBS with an analogue nucleotide sequence (promiscuity) and also, the existence of different TFs with affinity for the same sequence (sequence degeneration).

Therefore, the goal of this study was to develop a method to select SNP in the coding and essentially the promoter regions of genes. The study of genes involved in antigen processing and presentation were the reason to develop the new methodology and its practical application.

Methods

Computer tools and databases used to select SNPs in coding and promoting regions of the genes

The method presented here for SNP search and selection was developed and subsequently applied to a cohort of patients with NHL. We had a dual goal: develop a methodology for the selection of clinical relevant SNPs in coding and promoter sequences of any gene and test the practical applicability of the method previously developed to study the genes that regulate the proteins involved in antigen processing and presentation.

Protein selection

We used *PubMed* and *Google* to find relevant literature and identify proteins of the antigen processing systems and chose the ones that had a confirmed relevant role on human biology, according to the study published. A useful site was www.genecards.org, part of the Weizmann Institute of Science and LifeMap that describes protein function and gene's location

in the genome such as chromosome, chromosomal band, nucleotide sequence and number of nucleotide base on topography of chromosome DNA. This site also reports the protein expression in different human tissues either, normal or neoplastic and gene expression by RNA sequencing obtained from Illumina Body Map or SAGE (Serial Analysis of Gene Expression). The site directed us to other web pages such as 'HapMap' and 'QIAGEN' which describe the phylogenetic and evolutionary relationship of genes, description of SNPs, their nucleotide sequences and frequency in the populations studied [11].

SNP selection in coding regions

Information about SNPs obtained from any data source can be used directly for further studies regarding the coding regions of the genes. We considered the results from previous biomedical studies published in English. Preferentially, but not exclusively, we selected SNPs that have a frequency of 20% or higher among Caucasian populations. This information can be obtained in 'HapMap', provided by the NIH, at www.hapmap.ncbi.nlm.nih.gov. For SNPs with lower frequencies, we selected those particularly prevalent in the Iberian population, information obtained from "1000 Genomes" in the same site of 'dbSNP' and 'HapMap'.

Application of this simple direct selection can be much less useful for promoter regions and thus a new methodological strategy was required.

SNP selection in promoter regions

The search for SNPs in the promoter regions of the proteins of interest turned out to be complex and required several steps: First, we determined the gene location in the chromosome and its promoter using Qiagen SABioscience database at www.sabioscience.com/chipqpcrsearch.phap?app=TFBS. This database uses TRANSFAC and JASPER recorded information to provide chromosomal location, TFs, and their binding sites. Then using the chromosomal interval containing the gene promoter, we identified the SNPs available in the 'HapMap' database.

'HapMap' allows human chromosome visualization and permits successive amplifications from chromosomal bands to nucleotide sequence as well as the SNPs identified along the chromosomal locations. Furthermore, 'HapMap' provides the frequency of SNPs in several populations groups already studied including African sub ethnicities, Native Americans, Europeans and Chinese Han ethnicity. Unlike the coding regions, we restricted the study to the SNPs of promoter regions whose frequency on Caucasian populations registered on 'HapMap' were >20%.

For identifying the DNA binding motif sequence where SNPs occurred and its chromosomal location on the target protein gene, we used the powerful database 'dbSNP' that provides SNPs from diverse species including *Homo sapiens*. They are presented in the middle of a nucleotide sequence with approximately 20 adjacent nucleotides, half of which toward 3' and the other half toward 5', of the truncated base. For each SNP the ancestral allele and the new variant sequence were registered.

Afterwards, using the nucleotide sequence identified above at 'dbSNP' we compared TF settings. The affinity of TF for ancestral sequences of TFBS or for the variant TFBS induced by SNPs in the promoter regions can be obtained using 'Promoter TESS' at www.cbil.upenn.edu/cgi-bin/tess?RQ=WELCOME.

Finally, part of the information about TFs role in infection, inflammation, immunity, neoplasia, and aging was obtained through 'Promoter TESS', but was later expanded and analysed using the information provided by 'GeneCards' and in the literature published on 'PubMed' and Google.

Practical application of selected SNPs

DNA collection, genotyping by PCR amplification, clinical registry data and statistical analysis will be performed as the next step of this project during the practical application of the methodology to our target proteins.

Results

Method development for searching polymorphisms and its function in the coding and promoter regions of the genes

In this study, we used the genes of proteins of antigen processing and presentation systems. One of our objectives was to find SNPs with a

Table 1. SNPs from coding and promoter regions of selected genes from antigen processing/presentation systems

Gene	Gene location-pr	SNPs-cr	SNPs-pr
PSMA6	chr14: 35,741,574-35,771,574	rs12878391	rs1755784
		rs1048990	rs10139973
			rs17553775
			rs1766136
			rs7148603
			rs1766135
			rs2787423
			rs1766143
			rs1766145
PSMA7	chr20: 60,708,474-60,738,474	rs3746651	
		rs2281740	
		rs2057169	
PSMB4	chr1: 151,352,041-151,382,041	rs2296840	rs11205209
		rs4603	rs310133
PSMB8	chr6: 32,801,816-32,832,712		rs28772340
			rs2858892
			rs2859112
			rs13199787
			rs7773407
			rs6457644
			rs11758312
			rs9276490
			rs6918223
			rs7770024
PSMD7	chr16: 74,310,681-74,340,681	rs17336700	
PSMD9	chr12: 122,306,646-122,336,646	rs1043307	rs4759415
		rs74421874	
		rs3825172	
		rs14259	
UBQLN2	chrX: 56,570,072-56,600,072	rs12344615	
		rs11140213	
		rs2781003	
		rs2780995	
		rs944947	
		rs2781002	
Hsp70	chr5: 132,367,662-132,397,662	rs14355	
		rs398606	
Bag1	chr9: 33,254,761-33,284,761	rs706118	
PSMD5	chr9: 123,595,506-123,625,206	rs10760117	rs10985387
		rs10739575	rs10818593
			rs4641136
			rs3802488
			rs13299463
			rs4307413
B2M	chr: 15:44,983,685-45,013,685	rs2255235	rs16958856
			rs4349090

potential role in the reconnaissance and immunological response to the Herpes virus family on NHL patients and in NHL oncogenesis.

Protein selection

For clinical application of our methodology, we selected 22 proteins from the hundreds of proteins that integrate the antigen processing UB system, TAP protein, ER conjugation, lysosomal alternative pathway, and MHC system.

First, we only considered proteins that have a role in antigen processing and presentation and, secondly, we only included those that have demonstrated influence in inflammation, antigen processing mechanisms, infectious diseases, cancer, and aging.

From these 22 proteins, 28 SNPs in coding region and 26 of promoter region were selected for further study. SNP selection was based on the following criteria.

Selecting SNPs in coding regions

We used 'dbSNP' database to determine SNPs in the coding regions. Because there are hundreds of SNPs (since 65 to 1785) identified for each protein's gene, we chose only those that are reported in the literature and those that play a possible role in aging and infectious, inflammatory, immunological/autoimmune or neoplastic disorders. By these criteria we selected 74 SNPs (**Tables 1** and **2**).

Using the 'HapMap' following the methodology described

			rs16958871
			rs6493247
TAP1	chr6: 32,811,748-32,841,748	rs6457684	rs5019296
		rs4148882	rs13199787
		rs4148879	rs6457644
		rs2127679	rs7773407
			rs11758312
			rs9276490
			rs6918223
			rs7770024
TAP2	chr6: 32,796,547-32,826,547	rs17583244	rs11758312
		rs10484565	rs7773407
		rs9380326	rs6457644
		rs3819721	rs13199787
		rs3819714	rs2859112
		rs2857104	rs2395237
		rs2228396	rs9461799
		rs1800454	rs9469240
UBA52	Chr19: 18,662,614-18,692,614	rs3209501	rs10419226
		rs6554	rs4808844
			rs7256986
CUL5	Chr11: 107,859,408-107,889,408	rs7104942	rs11212672
			rs12361570
ERAP1	chr5: 96,133,892-96,163,892	rs28366066	rs28096
		rs17482078	rs1057569
			rs1065407
			rs149078
			rs27042
			rs469783
			rs469758
			rs26510

Table 2. SNPs in coding and promoter regions of selected genes from antigen processing/presentation systems (cont.)

Gene	Gene location-pr	SNPs-cr	SNPs-pr
HLA-A	chr6: 29,890,331-29,920,331	rs9260109	rs2523769
		rs9260105	rs1077432
		rs9260090	rs1318083
		rs9260102	rs1610678
		rs926100	rs1611165
		rs2735113	rs1610682
		rs2230954	rs407238
			rs2735003
HLA-B	chr6: 31,314,989-31,344,989	rs4997052	rs3868082
		rs2596501	rs3132496
		rs2523608	rs28480108
		rs1140412	rs3134766
			rs9264179
			rs9264219

above for SNP selection for specific populations, we chose those that exist in at least 20% of Caucasian populations (this cut-off was also adequate for the relatively small size of our population sample) [12, 13]. The type of population and SNP frequency selection results in 56 SNPs.

Of these SNPs, some were not compatible during pairing reactions and allele-specific extension oligonucleotides from PCR amplification in iPlexSequenom, MALDI-TOF. With this technical restriction, 28 of 56 SNPs could be selected (**Table 3**).

Promoter regions

We applied the step-by-step strategy, previously described, to obtaining 26 SNPs.

We started by determining their exact promoter region using the Qiagen SABioscience's database. For example the TAP1 gene promoter region has about 30 Kb and is located between the nucleotides from db 32,811,748 and db 32,841,748 of chromosome 6 (**Figure 1**).

Then, we accessed 'HapMap' database and searched SNPs located in the promoter region that was previously determined in the 'QIAGEN' database. We retrieved 440 SNPs from the promoter regions of the 22 proteins chosen. Then, we selected several SNPs whose minor allele frequency was above 20% [12, 13]. We observed approximately 10-30 SNPs per promoter, 0-14 of which fell within the criteria set out above. With these criteria, we selected 253 SNPs for further study.

			rs3130427
			rs1793891
			rs2524119
			rs2844626
			rs2853961
			rs2248902
			rs2524099
			rs1049281
HLA-C	chr6: 31,229,855-31,259,855	rs13207315	rs1128175
		rs13191343	rs885948
		rs7773175	rs3094188
		rs2395471	rs887466
		rs2249742	rs3131018
		rs2074488	rs1265155
			rs9501066
HLA-E	chr6: 30,437,271-30,467,271	rs1264457	rs3132628
		rs1264459	rs3132626
			rs3132622
			rs3094623
			rs3130133
			rs6936943
			rs3130139
			rs3130144
			rs1012411
			rs2022082
			rs2844746
			rs3132644
			rs3130362
			rs2844745
HLA-F	chr6: 29,671,117-29,701,117	rs17875380	
		rs9258170	
		rs2072895	
		rs1362126	
HLA-G	chr6: 29,774,756-29,804,756	rs12722477	rs7776082
		rs9380142	rs9258122
		rs2735022	rs3094727
		rs1736936	rs2394660
		rs1736935	rs3131863
		rs1632949	rs1476572
		rs1632947	rs1610586
		rs1632933	rs1610594
		rs915668	rs1611356
		rs1710	rs1611381

Gene location-pr: gene location on promoter region; SNPs-cr: Selected SNPs of coding region; SNPs-pr: Selected SNPs of promoter region.

We identified the nucleotide sequence of ancestral and variant alleles of each SNP with the 'dbSNP' database and record this information for using in the next steps.

To understand which SNPs could potentially have greater impact on which TF binding to the promoter region, we used 'Promoter TESS' for functional analysis. With this software, we compared the TFs which bind to the sequence of the ancestral allele with TFs that bind to the sequence of the variant allele (that has the SNP). Because the usual length of TFBS is 6-12 bp [14], we searched 'Promoter TESS' using DNA sequences that included the SNP and 10 nucleotides both the 5' and 3' directions. Thus, we identified 114 SNPs (from the previous 253) that could significantly change the TFs that bind to the variant promoter region (**Tables 1** and **2**).

Finally, we decided to search the literature for the role of these "new" TFs and select the SNPs for those which were related to aging and infectious, inflammatory, immunological/autoimmune and neoplastic disorders. Adopting this approach, we finally restricted our study to 87 SNPs located in the promoter regions (see the Supporting Information; Supplementary File 1 shows the difference between the ancestral and variant nucleotide sequence and corresponding TFs with affinity to them).

Similar to SNP selection from coding regions, some SNPs were not compatible in the iPlexSequenom MALDI-TOF platform, resulting in some technical, but not methodological, restriction to 26 SNPs (**Table 3**).

Considerations in selection criteria

The primary goal of this work was to try to understand the change in TF affinity induced by SNPs on DNA binding motifs of promoter regions. With this method it became possible to choose TFs and SNPs in an oriented manner, permitting selection of those with a potential functional role.

Figure 2 provides a summary of all steps.

Table 3. The final results in accordance with the methodology

Coding region	Promoter region
rs17583244	rs4641136
rs7104942	rs11758312
rs2781003	rs16958871
rs12344615	rs1611356
rs10739575	rs7256986
rs3825172	rs17553775
rs706118	rs1766135
rs2228396	rs26510
rs12878391	rs4808844
rs2296840	rs2844745
rs9380326	rs28096
rs4148879	rs9461799
rs6924102	rs10139973
rs3209501	rs6457644
rs2281740	rs2787423
rs10484565	rs10419226
rs28366066	rs27042
rs2780995	rs7776082
rs3819721	rs7148603
rs6457684	rs469758
rs17875380	rs11205209
rs17336700	rs3132622
rs10760117	rs1766136
rs1632933	rs13199787
rs4148882	rs5019296
rs398606	rs1611381
rs1264457	
rs1800454	
Total: 54 SNPs	

Discussion

This study describes methods for SNP search and selection that can be applied to any human gene. This is an empirical method to select functionally significant SNPs. The authors could not find in the medical literature a previous description of a practical approach to select SNPs in promoter regions based on their predictable functionality. The present method emphasizes the importance of promoter region, where the effects of the SNPs can be more complex than those of SNPs in the coding regions, even though its functional effect has a smaller impact. It compares the TFs with affinity to the TFBS of the conserved ancestral sequence with the TFs that have affinity to the variant allele. Using this approach we were able to identify SNPs responsible for generating important differences in TFs pattern to a TFBS.

The authors identified and selected the SNPs that induced TFs active in NHL. Along the way, the applicability of the method was tested. For example, we studied the genes of the protein of several enzymatic systems involved in processing and presentation of antigenic determinant.

The complexity found in transcription control and its multiple effects has been remarkable. Among several difficulties, we are faced with the promiscuity of the TFs. This phenomenon is possible because TFs can tolerate small differences in nucleotide sequences, although with different affinities. Another manifestation of tolerance in protein-DNA binding proprieties is the possibility of a DNA binding motif (TFBS), having affinity to multiple TFs albeit with different binding energies. This last effect is called “degenerate consensus sequence”.

The TF affinity varies according to the 4 nucleotide distribution in a particular sequence and has been defined by entropy-based mathematical models. Each position and type of nucleotide in the TFBS features its own weight according to its binding energy. A position weight matrix (PWM) is commonly used to represent this type of sequence motif [15]. The possible binding sequences can be identified using models and databases, such as MATCH and TESS, that use PWMs registered in JASPAR or TRANSFAC [10, 16-18].

The SNPs in the TFBS can introduce significant alterations in the type of TFs that bind there, eventually inducing changes in the gene transcription. Recent studies by Michal et al led to the development of computer models that can estimate the result of replacing a base at the TFBS [19]. These predicted values for certain SNPs were compared with the published results, and a very good correlation was found. These studies need to be extensively done for all promoters in *Homo Sapiens* genes.

Although it is essential to understand the functionality of a SNP in the regulating sequences, computational models of PWM and the methodology developed in our work, do not solve the issue of the transcriptional control and gene expression. As described below, the TFBS can, and often are redundant in the promoter region.

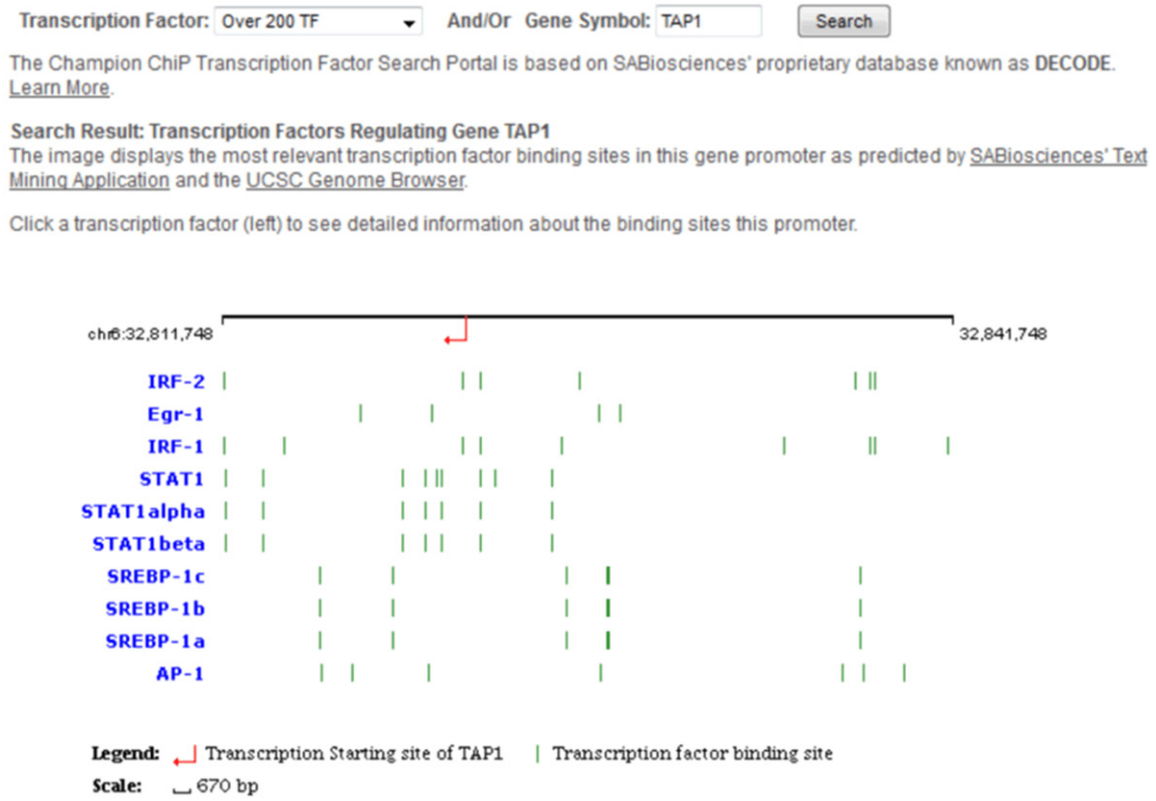

Figure 1. 'QIAGEN' database reveals the promoter region and the TFs that bind to it. In this image, we can see that the promoter region of TAP1 gene is located between nucleotides 32,811,748 and 32,841,748 in chromosome 6. As represented, there are several TFs that can bind to this region and all TFs can bind to different binding sites in the promoter region. These are the concepts of "degenerate consensus sequence" and "promiscuity".

Beyond its redundancy, its location and the effect of the "degenerate consensus sequence" also decrease functional impact of SNPs in regulatory regions. However, several associations between human diseases and these variations in the promoter binding sequences have been described, such as susceptibility for lupus, arthritis, HIV/AIDS, diabetes and heart disease. One of the most interesting findings regarding protection against HIV/AIDS is due to the polymorphism in the cis-regulatory region in the gene *CCR5* that codes for CC chemokine receptor 5, required for the HIV1 virus entrance into the cell [20].

To explore the effect of promiscuity/degenerate consensus sequence, we adopted an empirical approach to the genes of the proteins being studied, comparing the group of TFs that bind to conserved consensus motifs (ancestral allele) in the promoter regions with those that bind alternate sequence, where the SNP appeared (variant allele). Three possible scenarios occurred: 1) new TFs appeared with affinity to the new TFBS; 2) the usual TFs were kept due to the degenerate consensus sequence; 3) an intermediate process such as small changes in the TFs, or the new TFs have low binding affinity for the new TFBS containing the SNP. In practical terms, SNP selection lies in those that induced significant alterations in the TF for the new sequence, that is, the ones that could potentially change the gene transcription.

There is an extensive complexity in the transcriptional control currently known which accounted for some limitations of the study. For example, it did not consider factors that can influence the role of SNPs on the promoter regions, such as the number of repletion of TFBS for the same TF along the promoter region, their location, proximal or distal relative to the initiation site and the synergism between them. We can use the promoter region of TAP1 located between 32,811,746 and 32,841,748

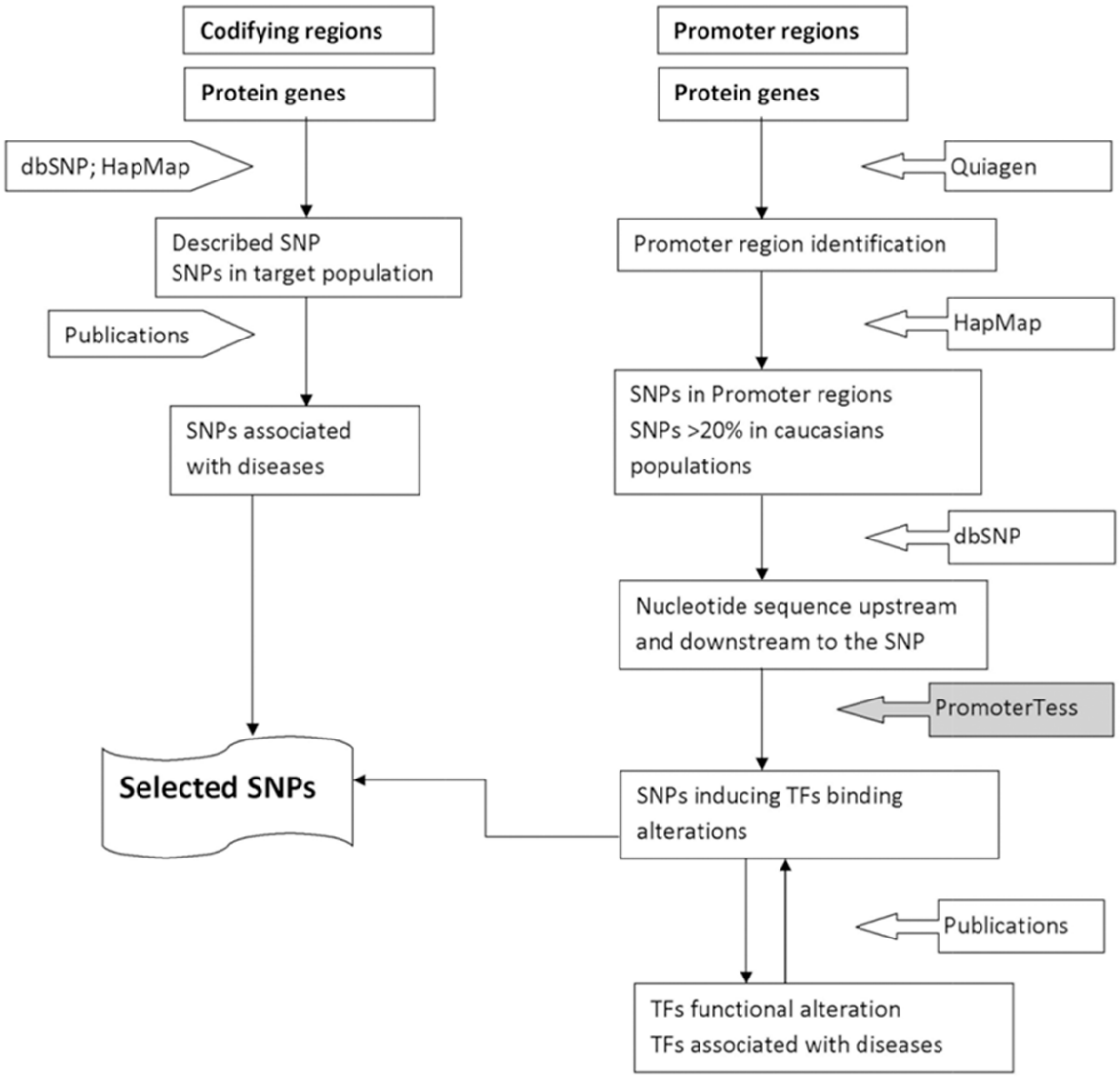

Figure 2. Methodology of SNPs' search in coding and promoter regions of the genes. Flowchart showing the steps used to find SNPs in coding region (left) and promoter region (right) and corresponding computer tools.

in chromosome 6 (chr6) as an example of degenerate consensus sequence and promiscuity shown in **Figure 1**. It has 9 binding sites for the transcription factor IRF-2. This TF has TFBS in >4000 genes: TAP1, GRAP2, CASP8, CIITA, CXCL11, DST, NF1, SPI1, TRIM63, and PRDM1 are the 10 most relevant that are involved in apoptosis, antigenic presentation and many other important processes in oncology and immunology. On the other hand each TFBS can be the binding site for multiples TFs. The TFBS centered in the nucleotide 32,811,824 is the binding site for IRF-1 and IRF-2, whereas the TFBS centered in the nucleotide 32,811,882 is the binding site for STAT1, STAT1α, and STAT1β.

The redundancy of TFBS and the phenomenon of binding entropy (degenerate consensus sequence) are not the only factor that modulates transcriptional decision of TF. For example, TFBS from promoter regions can be present either in repressive or activator places. Even more important can be the TFBS location. Experimental studies, using reporter gene assays, found that time-conserved TFBS appear in the promoter proximal regions, close to the transcription start site and that muta-

tions in these TFBS had an important functional role. However, a considerable number of them were found 1 kb upstream already in the 3'cis-regulatory promoter regions [21].

Although there are complexes mathematical models to determine PWM, few clinical models can predict alterations in the transcription decision with the introduction of an SNP in the TFBS nucleotide sequence. In the present study, the authors selected individual SNPs based on the differences of TFs affinity between the ancestral and the variant allele that potentially could have immediate clinical application.

TF selection should follow the clinician's area of interest. Some groups have attempted to develop methods to identify new entities and prognostic groups of NHL and breast cancer based on tumor gene expression profiling that mandates the study of thousands of genes. This requires very expensive and sophisticated technology not available in most clinical settings. The identification of several dominant TF would allow inferring the expression of a vast group of genes, i.e. genes containing the binding motifs for those TF [22]. We can imagine the replacement of the "tumor genetic profile" for the "tumor transcriptional profile" at a much lower cost.

Using our methodology we are currently studying the selected SNPs in samples of patients with NHL and its role as etiologic risk factors as well as its prognostic value.

Acknowledgements

We thank Professor Fernando Rodrigues from ICVS, UM for insights into genetic transcriptional organization; QIAGEN's application specialist, Jorge Posadas; and HapMap's staff.

Disclosure of conflict of interest

None.

Address correspondence to: José Freitas, NOVA Medical School Faculdade de Ciências Médicas, Campo Mártires da Pátria, 130, 1169-056 Lisboa, Portugal. E-mail: jfrei1992@gmail.com

References

[1] Maher B. ENCODE: The human encyclopaedia. Nature 2012; 489: 46-48.

[2] Han XS, Zheng TZ, Foss FM, Lan Q, Holford TR, Rothman N, Ma S, Zhang YW. Genetic polymorphisms in the metabolic pathway and non-Hodgkin lymphoma survival. Am J Hematol 2010; 85: 51-56.

[3] Sarmanova J, Benesova K, Gut I, Nedelcheva-Kristensen V, Tynková L, Soucek P. Genetic polymorphisms of biotransformation enzymes in patients with Hodgkin's and non-Hodgkin's lymphomas. Hum Mol Gen 2001; 12: 1265-1273.

[4] Liu J, Song B, Wang ZH, Song XR, Shi Y, Zheng JS, Han JX. DNA repair gene *XRCC1* polymorphisms and non-Hodgkin lymphoma risk in a Chinese population. Cancer Genet Cytogenet 2009; 191: 67-72.

[5] Shen M, Purdue MP, Kricker A, Lan Q, Grulich AE, Vajdic CM, Turner J, Whitby D, Chanock S, Rothman N, Armstrong BK. Polymorphisms in DNA repair genes and risk of non-Hodgkin's lymphoma in New South Wales, Australia. Haematologica 2007; 92: 1180-1185.

[6] Nasr AS, Sami RM, Ibrahim NY. Methylenetetrahydrofolate redutase gene polymorphisms (677C>T and 1298A>C) in Egyptian patients with non-hodgkin lymphoma. J Can Res Ther 2012; 8: 355-360.

[7] Rothman N, Skibola CF, Wang SS, Morgan G, Lan Q, Smith M. Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. Lancet Oncol 2006; 7: 27-38.

[8] Purdue MP, Lan Q, Kricker A, Grulich AE, Vajdic CM, Turner J, Whitby D, Chanock S, Rothman N, Armstrong BK. Armstrong. Polymorphisms in immune function genes and risk of non-Hodgkin lymphoma: findings from the New South Wales non-Hodgkin Lymphoma Study. Carcinogenesis 2007; 3: 704-712.

[9] Lähdesmäki H, Rust AG and Shmulevich I. Probabilistic inference of transcription factor binding from multiple data source. PLoS One 2008; 3: e1820.

[10] Bryne JC, Valen E, Tang MHE, Marstrand T, Winther O, Piedade I. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 2008; 36: D102-D106.

[11] Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D. Gene Cards Version 3: the human gene integrator. Database 2010; 2010: baq020.

[12] Flores C, Maca-Meyer N, González AM, Oefner PJ, Shen P, Pérez JA, Rojas A, Larruga JM, Underhill PA. Reduced genetic structure of the Iberian península revealed by Y-chromosome analysis: implications for population demography. Eur J Hum Genet 2004; 12: 855-863.

[13] Pereira L, Richards M, Goios A. High-resolution mtDNA evidence for the late-glacial resettle-

ment of Europe from an Iberian refugium. Genome Res 2005; 15: 19-24.
[14] Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genoma. Annu Rev Genomics Hum Genet 2006; 7: 29-59.
[15] D'Haeseleer P. What are DNA sequence motifs? Nat Biotechnol 2006; 4: 243-245.
[16] Kel AE, Goßling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH™: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 2003; 13: 3576-3579.
[17] Schug J. Using TESS to predict transcription factor binding sites in DNA sequence. Curr Protoc Bioinformatics 2008; Chapter 2: Unit 2.6.
[18] Wingender E, Dietze P, Karas H, Knüppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 1996; 1: 238-241.
[19] Michal L, Mizrahi-Man O, Pilpel Y. Functional Characterization of Variations on Regulatory Motifs. PLoS Genet 2008; 4: e1000018.
[20] McDermott DH, Zimmerman PA, Guignard F, Kleeberger CA, Leitman SF, Murphy PM. CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS). Lancet 1998; 352: 866-870.
[21] Rockman MV, Wray GA. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 2002; 19: 1991-2004.
[22] Veerla S, Ringner M, Hoglund M. Genome-wide transcription factor binding site/promoter database for the analysis of gene sets and co-occurrence of transcription factor binding motifs. BMC Genomics 2010; 11: 145.

Supplementary File 1. The difference between the ancestral and variant nucleotide sequence (according to the SNP which is in the 2nd column) and corresponding TFs with affinity to them

Gene	SNPs (promoter region)	Nucleotide region	Transcription factors	Nucleotide region-Wild type	Transcription factors
PSMA6	rs1755784	CAAAGCCTTAGATGTTTTACT	POU2F1/E12	CAAAGCCTTAAATGTTTTACT	POU2F1
	rs10139973	TTATCATGTAGCAAAAGACAT	GATA-2	TTATCATGTAACAAAAGACAT	E4BP4/TCF-4E
	rs17553775	TCAATGGTGAGTTATAGTGAG	AP-1/c-Myc/c-Fos	TCAATGGTGACTTATAGTGAG	--------
	rs1766136	ATTTTCTTTCCCACTGTTTAA	NP-TCII	ATTTTCTTTCACACTGTTTAA	--------
	rs7148603	AGTGACCAAAGTAAGAATCTG	LEF-1	AGTGACCAAAATAAGAATCTG	c-Fos/c-Jun/AP-1
	rs1766135	TGAATTCCCCTTTTCCCCCAA	Pu box binding factor/ NF-Atc/NFAT-1	TGAATTCCCCCTTTCCCCCAA	NP-TCII
	rs2787423	CAGGTAGAGTGAGCTGGAAAG	--------	CAGGTAGAGTAAGCTGGAAAG	NP-TCII
	rs1766143	CCCCCAAAAGGTATGTCCACC	c-Rel/TEF-1	CCCCCAAAAGATATGTCCACC	GATA-3
	rs1766145	ATCAATTTTCTGGTAACTCCA	c-Rel	ATCAATTTTCAGGTAACTCCA	AREB6
PSMB4	rs11205209	AGAAATATAAGATAAGGCAAG	--------	AGAAATATAAAATAAGGCAAG	NC-2
	rs310133	ACCCCTTAACGACAACCATAA	--------	ACCCCTTAACAACAACCATAA	c-Myc
PSMD5	rs10985387	CAATGTGGACGCAGATGCATC	E12/E47/ITF-2	CAATGTGGACACAGATGCATC	E12/E47/ITF-2
	rs10818593	TATACAAAATTGAGATTGCTT	GATA-1	TATACAAAATCGAGATTGCTT	GATA-1/POU2F1
	rs4641136	CAGACTCGAGGGGCGTGCCAT	Sp1	CAGACTCGAGAGGCGTGCCAT	Sp1/p53
	rs3802488	ACTTGAATCCGCCCTCCCCAG	--------	ACTTGAATCCACCCTCCCCAG	Sp1
	rs13299463	AAGAGCTAACGAGAGTGGCCT	c-Myb	AAGAGCTAACAAGAGTGGCCT	--------
	rs4307413	CTTTGTGAAAGCCTGGATTTA	C/EBPbeta	CTTTGTGAAAACCTGGATTTA	--------
PSMD9	rs4759415	AGCCTTGTCATGGCTGAGAAT	--------	AGCCTTGTCACGGCTGAGAAT	--------
PSMB8	rs28772340	CATAAGAGATTACATCCCCAT	AP-2alphaB	CATAAGAGATCACATCCCCAT	c-Myc
	rs2858892	GTAGTTCTTATACAACTGAAG	c-Myc/c-Myb	GTAGTTCTTACACAACTGAAG	Zta
	rs2859112	AGTGAGGCTTGGATGATGCCC	Sp1	AGTGAGGCTTAGATGATGCCC	--------
	rs13199787	TTCATCAATGTATAAAATTAG	NC2	TTCATCAATGCATAAAATTAG	--------
	rs7773407	CTGCAACCTCCAAAACCCTCT	c-Ets2/c-Jun	CTGCAACCTCAAAAACCCTCT	NF-Gma/ELF-1
	rs6457644	ACTGAACCCATGACTTCCCTT	c-Ets2/c-Jun	ACTGAACCCACGACTTCCCTT	c-Ets2
	rs11758312	GATTGGGTTGCTAAGAGAACT	Zta/c-Myc	GATTGGGTTGATAAGAGAACT	c-Myc
	rs9276490	GAATGCAACTGTAAAGAATGT	c-Myc/c-Myb	GAATGCAACTATAAAGAATGT	--------
	rs6918223	AGCTTGTCTGCCTTAATGACA	--------	AGCTTGTCTGACTTAATGACA	c-Jun
	rs7770024	TGCACAGATGGAACTATAACA	--------	TGCACAGATGAAACTATAACA	E12/E47/ITF-2/Tal-1/ Tal1-beta
B2M	rs16958856	TGTTATATTTTCTTCCATGAC	--------	TGTTATATTTCCTTCCATGAC	c-Ets2
	rs4349090	CAATAAACAGGTGTGTGACTG	AREB6	CAATAAACAGCTGTGTGACTG	--------
	rs16958871	GCAATAGTTATGTTGTAAAGT	--------	GCAATAGTTACGTTGTAAAGT	(C)EBPbeta
	rs6493247	AAAAAATCCCGACAAGCTAGG	--------	AAAAAATCCCAACAAGCTAGG	(C)EBPbeta
TAP1	rs5019296	TGAGGCCAGGTGCAGTGGCTC	AREB6/Lmo2	TGAGGCCAGGCGCAGTGGCTC	AP-2alphaA/AP-2alphaB/ LBP-1
	rs13199787	TTCATCAATGTATAAAATTAG	NC2	TTCATCAATGCATAAAATTAG	--------
	rs6457644	ACTGAACCCATGACTTCCCTT	c-Ets2/c-Jun	ACTGAACCCACGACTTCCCTT	c-Ets2
	rs7773407	CTGCAACCTCCAAAACCCTCT	NF-Gma	CTGCAACCTCAAAAACCCTCT	NF-Gma/ELF-1
	rs11758312	GATTGGGTTGCTAAGAGAACT	Zta/c-Myc	GATTGGGTTGATAAGAGAACT	c-Myc
	rs9276490	GAATGCAACTGTAAAGAATGT	c-Myb/c-Myc	GAATGCAACTATAAAGAATGT	--------
	rs6918223	AGCTTGTCTGCCTTAATGACA	--------	AGCTTGTCTGACTTAATGACA	c-Jun
	rs7770024	TGCACAGATGGAACTATAACA	--------	TGCACAGATGAAACTATAACA	E12/E47/ITF-2/Tal-1/ Tal1-beta
TAP2	rs11758312	GATTGGGTTGCTAAGAGAACT	Zta/c-Myc	GATTGGGTTGATAAGAGAACT	c-Myc
	rs7773407	CTGCAACCTCCAAAACCCTCT	c-Ets2/c-Jun	CTGCAACCTCAAAAACCCTCT	NF-GMA/ELF-1
	rs6457644	ACTGAACCCATGACTTCCCTT	NF-Gma	ACTGAACCCACGACTTCCCTT	c-Ets2
	rs13199787	TTCATCAATGTATAAAATTAG	NC2	TTCATCAATGCATAAAATTAG	--------
	rs2859112	AGTGAGGCTTGGATGATGCCC	Sp1	AGTGAGGCTTAGATGATGCCC	--------
	rs2395237	GAAATAATAACGATAAGTTGT	Cart-1/c-Myb	GAAATAATAAAGATAAGTTGT	Cart-1/TCF-1A
	rs9461799	TCCCAATGGGTAACTGATTGC	c-Myb	TCCCAATGGGCAACTGATTGC	c-Myb/c-Myc
	rs9469240	GAGTGTGTAGTGAGATTGTTG	p300/GATA-3	GAGTGTGTAGCGAGATTGTTG	GATA-3
UBA52	rs10419226	AGTCACAAATTACCACAAAGT	--------	AGTCACAAATGACCACAAAGT	PEBP2beta

Selection of SNPs in promoter regions

	rs4808844	CCGGGGCAGAGGGAGGAGCCT	--------	CCGGGGCAGAAGGAGGAGCCT	Sp1
	rs7256986	TGTTTAAAACGGGAGCATAAC	c-Myb	TGTTTAAAACAGGAGCATAAC	--------
CUL5	rs11212672	CTGGCAAACATGAAACAACTT	c-Fos/c-Jun/Fra-1	CTGGCAAACACGAAACAACTT	AP-2alphaA
	rs12361570	AAATTCTTCTACGTCAACTTAG	--------	AAATTCTTCTAAGTCAACTTAG	c-Jun/c-Myb
ERAP1	rs28096	ACTGTATAGCGTCTGGCTTTA	E47	ACTGTATAGCATCTGGCTTTA	E47/E12/ITF-2/Tal-1beta
	rs1057569	GTGGCAGCGGGGCAAGCAAAA	POU2F2/E2F1	GTGGCAGCGGAGCAAGCAAAA	--------
	rs1065407	CTCCCTTGCCCGGTTCCTGTT	ELF-1/c-Ets-1	CTCCCTTGCCAGGTTCCTGTT	Pu.1/Elf-1
	rs149078	AAAAAGCTACGAGACTGTAAC	TCF-1	AAAAAGCTACAAGACTGTAAC	LEF-1/TCF-1
	rs27042	GTTTCATCACTTATTCATTGC	GATA-3	GTTTCATCACCTATTCATTGC	GATA-3/E4BP4
	rs469783	TAATGAGACTTGCCCGATCAT	Sp1/p53	TAATGAGACTCGCCCGATCAT	p53
	rs469758	ATTTGCTCCCTTGCCTGAAGA	--------	ATTTGCTCCCCTGCCTGAAGA	Pax5
	rs26510	TAATCAAGGATCTCAGAAAGT	--------	TAATCAAGGACCTCAGAAAGT	E12
HLA-A	rs2523769	GCCAAAGCAGTATTGTAAACT	TCF-1	GCCAAAGCAGGATTGTAAACT	--------
	rs1077432	CATGTGATGGTTCATTTTCAA	AP-1	CATGTGATGGCTCATTTTCAA	Zta
	rs1318083	TTAAGACTTCTAGTATGTTCC	TEF-1	TTAAGACTTCAAGTATGTTCC	--------
	rs1610678	TAGTTTTTTCGATAACTGGCT	c-Myb	TAGTTTTTTCAATAACTGGCT	--------
	rs1610682	AAATTCTCTTGTTATTTCTTT	--------	AAATTCTCTTATTATTTCTTT	c-Myc
	rs407238	TGGGGAAGATGCTCTCTCACT	c-Ets2	TGGGGAAGATCCTCTCTCACT	--------
	rs2735003	CCATACCAAACCCCTAGGTTC	Sp1/AREB6	CCATACCAAAACCCTAGGTTC	LBP-1
HLA-B	rs3868082	TTGATTTTTATACATTTGGTT	NC2	TTGATTTTTACACATTTGGTT	Zta
	rs3132496	TTTCTAAGAACTGAGTGAATC	c-Myb	TTTCTAAGAAATGAGTGAATC	--------
	rs28480108	GACGGCTGCAGAAGTATCTTC	NF-Gma	GACGGCTGCAAAAGTATCTTC	--------
	rs3134766	ACACCATCACGCCTTACCCCT	SP1/GATA-3	ACACCATCACACCTTACCCCT	AREB6
	rs9264179	GCATTTGAGTTCAGCCAGAGA	--------	GCATTTGAGTCCAGCCAGAGA	LBP-1
	rs9264219	TGAGACTACTTCTGTTTTTGG	--------	TGAGACTACTCCTGTTTTTGG	Sp1
	rs3130427	CTGAACCACAGTGCCCAGATA	AML1/AML3	CTGAACCACAATGCCCAGATA	AML1/AML3/SP1
	rs1793891	AGGGACATGAGGTTCTGCTGC	--------	AGGGACATGAAGTTCTGCTGC	c-Myb
	rs2524119	CAAAGGTCCCGCCTCTTAAAA	Sp1	CAAAGGTCCCACCTCTTAAAA	Sp1/AREB6
	rs2844626	GACCAAGGACTGTACCTGGTA	LBP-1	GACCAAGGACAGTACCTGGTA	--------
	rs2853961	ACTGTTGTTGTGGGAAGTCAA	AML1c/c-Ets-2	ACTGTTGTTGCGGGAAGTCAA	c-Ets-2
	rs2248902	TTCTCCAAGAGGTGAGTGAGA	AREB6	TTCTCCAAGAAGTGAGTGAGA	--------
	rs2524099	GAAACCTGATTGTGTGCTGCA	POU2F1	GAAACCTGATCGTGTGCTGCA	--------
	rs1049281	GTCAATTCCTGGAAGTTGAGA	--------	GTCAATTCCTAGAAGTTGAGA	c-Myb
HLA-C	rs1128175	ATAGCTAGAATGGAAAAAAGA	NF-Atc/TCF-1A/NFAT-1	ATAGCTAGAACGGAAAAAAGA	--------
	rs885948	AGAAGGCAGATAGAGCCACTG	GATA-3	AGAAGGCAGACAGAGCCACTG	TCF-1
	rs3094188	TTTTATGTCTTAGTTGGAAGG	POU2F1/PEA3	TTTTATGTCTGAGTTGGAAGG	PEA3
	rs887466	TCTCCGGAAATACCTGAAAGC	--------	TCTCCGGAAACACCTGAAAGC	AREB6/c-Myc
	rs3131018	GAACCAAGCATAGCTGCAGAA	LEF-1	GAACCAAGCAGAGCTGCAGAA	--------
	rs1265155	CTGTGAGTTGTTGGGGAACCG	AP-2alphaA/AP-2alphaB/(C)EBPbeta	CTGTGAGTTGCTGGGGAACCG	AP-2alphaA/AP-2alphaB
	rs9501066	GTGAAGTGGGGTGGTATCTGA	Sp1	GTGAAGTGGGATGGTATCTGA	Lmo2
HLA-E	rs3132628	GGAATATATAGTTAGTTAAAA	--------	GGAATATATAATTAGTTAAAA	Sp1
	rs3132626	ACTTACCAGGGAAACAACAAC	--------	ACTTACCAGGAAAACAACAAC	NFAT-1/Pu box binding factor
	rs3132622	CAAGCTCTTTGAAAAATAACT	--------	CAAGCTCTTTAAAAAATAACT	TCF-1
	rs3094623	TGCTGATCTATCTGTTCATGT	GATA-3	TGCTGATCTACCTGTTCATGT	--------
	rs3130133	GAGGAGCCAGCTTCTTCTAAA	--------	GAGGAGCCAGATTCTTCTAAA	NF-Gma
	rs6936943	ATCTGGGAAGGAAAAAAAAAA	NF-AT1	ATCTGGGAAGAAAAAAAAAAA	--------
	rs3130139	TTAAAATTTCTGCTTCACATC	--------	TTAAAATTTCCGCTTCACATC	NP-TCII
	rs3130144	TCAAAAATTATAAAATAATTA	--------	TCAAAAATTACAAAATAATTA	E4BP4
	rs1012411	GACAACTATGCCCTGTAGATG	Sp1	GACAACTATGACCTGTAGATG	--------
	rs2022082	TGCAGCTACAGAGGCTCGGGGG	TCF-1	TGCAGCTACAAAGGCTCGGGGG	LEF-1/TCF-1
	rs2844746	TACACAAGGTGAAAAGAGGAC	AREB6	TACACAAGGTAAAAAGAGGAC	--------
	rs3132644	ATTGATAATGGTAATGTTGGC	POU3F2/POU2F1	ATTGATAATGATAATGTTGGC	POU2F1
	rs3130362	TTGCCTATTCGTTTATTAGTT	Pbx-1a	TTGCCTATTCATTTATTAGTT	POU3F2
	rs2844745	ACCTTTATCCGTTAGATAAAA	GATA-3	ACCTTTATCCATTAGATAAAA	--------
HLA-G	rs7776082	TGGCTTTACCGTTTTCCATTC	c-Myb/Pu box binding factor	TGGCTTTACCATTTTCCATTC	POU3F2

Selection of SNPs in promoter regions

rs9258122	TTTGCTGGGGTATAAATGTAA	--------	TTTGCTGGGGCATAAATGTAA	POU3F2
rs3094727	ATTTTCAGGTGTTGAATAGAA	AREB6/c-Myc/E12	ATTTTCAGGTATTGAATAGAA	c-Myb/ELF-1
rs2394660	AATTTCATGTGGCAGCTGTAA	E12/HEB	AATTTCATGTAGCAGCTGTAA	E12
rs3131863	AGAGCAAGAGTGATGGACAGA	Pbx-1a/Pbx-1b/NFAT-1	AGAGCAAGAGCGATGGACAGA	NFAT-1
rs1476572	ATATCTACTGCAGGCCACAGC	GATA-3/HEB	ATATCTACTGAAGGCCACAGC	GATA-3
rs1610586	AGAATGAACTGAGAGATACAC	--------	AGAATGAACTAAGAGATACAC	c-Myc
rs1610594	CAGCCTCATTGTCCATCCTCT	PEA3	CAGCCTCATTATCCATCCTCT	Sp1
rs1611356	TTAGACATGAGTTAGTTGTCC	c-Fos/E12	TTAGACATGACTTAGTTGTCC	--------
rs1611381	GAATGGTAAATCAGCTTATTT	POU3F2	GAATGGTAAACCAGCTTATTT	LBP-1
rs1632957	CCACAAACCTTAGGATTACAG	AREB6	CCACAAACCTCAGGATTACAG	AML1a/E12