SNP Annotation from Next Generation Sequencing Data

J. A. Morente-Molinera*, J. M. Martín+, C. Cano*, M. Cuadros*, A. Blanco*

*Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain
{jamoren, ccano, marta, armando}@decsai.ugr.es

+Information Technology
University of Huelva
Huelva, Spain
jmmartin@dti.uhu.es

***Abstract*— Massive sequencing technologies are producing an increasing amount of whole genome data, which need to be explored and analyzed. New computational tools are thus required to deal with the dimensionality and complexity of these data. Single Nucleotide Polymorphisms (SNPs) are the most common human genome variation and can be involved in disease conditions. Identifying SNPs and annotating its functional and clinical role in whole human genomes is a challenging task, which requires expert curation. There are several software tools that assist researchers in the SNP calling and SNP annotation processes. However, these tools do not focus on the association of SNPs to regulatory regions such as Transcription Factor Binding Sites (TFBSs).**

This paper proposes a methodology to assist the annotation of SNPs in whole genome sequences, including not only genes but also known TFBSs. Our main contribution is that we use an intuitionistic-based similarity measure (SCintuit [1]), based on fuzzy technology and intuitionistic sets, to perform accurate comparisons between DNA sequences and identify TFBSs affected by a SNP.

Keywords-Bioinformatics; SNP Annotation; TFBS; Sequence similarity measures; intuitionistic fuzzy sets

I. Introduction

Next-generation sequencing technologies are currently generating a huge amount of data, requiring new computational tools to be developed in order to analyze these data and extract relevant knowledge from them [2]. This knowledge may shed light on the understanding of the way living organisms work, towards the ultimate goal of unveiling the mechanisms that regulate system disorders in order to prevent them.

SNPs are mutations in the DNA chain that only affects one nucleotide, i.e. they consist on the substitution of one nucleotide with another. Recent studies have demonstrated the active role of single nucleotide polymorphisms (SNPs) in the genesis of many system disorders. Depending on the affected loci, this alteration in the DNA sequence may either dramatically alter a regulatory process or do not affect the cell regulation a bit. Furthermore, a human genome may present more than 4 million DNA variants compared to a reference genome, including more than 3 million SNPs and around 9000 non-synonymous amino acid changes [3]. Therefore, the annotation of the functional and clinical effects of the SNPs in the human genome is of great relevance to understand the underlying regulatory mechanisms and assign clinical relevance [4].

Traditionally, it was thought that SNPs only affect the regulation of an organism when they were located in protein-coding regions of the genes (exons). The rationale behind this is that changes in the DNA chain may be translated into changes in the aminoacid sequence of a protein, which, in turn, may alter its ternary structure or its properties [5]. However, cell regulation is a complex process involving many regulatory mechanisms.

The transcription of a gene is one of the key processes in cell regulation. This process is regulated by one or several proteins, called transcription factors (TFs). The TFs bind to DNA sequences called transcription factor binding sites (TFBS) near the gene. TFBSs are usually located close to the transcription start site (TSS) of the gene and upstream from it. In some cases, TFBSs may also be found downstream the TSS or even within exons. These interactions between DNA and proteins play a crucial role in regulating the transcription of genes by either activating or inhibiting the transcriptional machinery of the cell. Several measures have been proposed for comparing DNA sequences in order to characterize these interactions and find TFBSs associated to TFs in DNA sequences [1][6][7][8]. Among these measures, SCintuit has been shown to outperform the other approaches in terms of predicting quality of TFs, without compromising sensitivity [1].
Existing knowledge about well-characterized TFs and binding motifs is stored in databases such as TRANSFAC [9] and Jaspar [10].

Given the crucial role of TFs and TFBSs in transcriptional regulation, current whole-genome annotation tools should be able to cope with the annotation of SNPs to TFBSs. This would allow elucidating affected regulatory processes, which, in turn, may be associated with a disease phenotype. However, most of the available annotation tools are usually

978-1-4577-1675-1

restricted to the annotation of SNPs at genes or exons [4] [11] [12].

A scientist accomplishing a SNP calling and annotation process typically faces the following steps. First, he uses some software tools such as SAMTOOLS [13], VARiD [14] or ACCUSA [15] to carry out the SNP calling process. For SNPs located in gene regulatory regions or intragenic regions, further research is needed in order to identify whether the SNPs is located in a TFBS. In this case, a tedious search against TFBS databases is carried out, since there are no tools assisting the scientist in the automatic annotation of SNPs in TFBSs.

This paper proposes a methodology to help the annotation of SNPs in whole genome sequences. Our focus relies not only on genes but also on known TFBSs in the sequences. Therefore, after aligning the reads to a reference genome and calling the SNPs, identified SNPs and context sequences are searched against different databases, with sequences and functional information associated to known genes, SNPs and TFBSs, to retrieve relevant functional information associated to the SNP. A key step in our approach is that sequence matching is performed using an Artificial Intelligence approach based on fuzzy technology and intuitionistic sets.

This paper is organized as follows. Section Methods describes the different components of the proposed methodology and the steps to carry out the analysis. This section comprises both the implemented SNP calling pipeline and the method for annotating SNPs in TFBSs. Finally, the Experiments section describes a case of study on the annotation of SNPs on a human chromosome, a comparison of different measures and a summary of the results found.

II. METHODS

The proposed methodology implements a pipeline with two main steps: SNP calling and SNP annotation. The SNP calling maps the reads generated by the sequencing platform to a reference genome and compares the alignments, detecting SNPs. The SNP annotation process consists of identifying if a SNP affects a known gene or TFBS for assigning a functional role to it.

A. SNP Calling

We start from a reference genome and a set of short sequences, called reads, obtained by Next Generation Sequencing (NGS) technologies. The SNP calling has two parts. First, a mapping step in which the reads are aligned to the reference genome. Second, the so-called SNP calling phase in which, using the Bayesian method included in SAMTOOLS, each position of the genome alignment is analyzed in order to determine whether that position can be a SNP candidate position. A SNP position is not only determined by searching for differences between the reference and the reads. It is also important to analyze the error rates in order to discard base changes that are produced by sequencing or mapping errors (among others). To do this, a Bayesian network is built with all this information and the posterior probabilities of having a SNP are computed for each position in the genome. Depending on the final result, each position is classified as a SNP or not. The maximum read depth to call a SNP is set to 100 and the genotypes are set to be computed at variants sites. More details about Bayesian methods for calling SNPs can be found in the SAMTOOLS description [13] [16]. For the alignment, the burrows-wheeler algorithm (BWA) [17] has been used.

B. SNP Annotation in Genes

The sequences in which the SNPs have been found are annotated using the variation effect predictor of the Ensembl database [18].

This process uses a local copy of Ensembl to query the database to reduce the reply time and be able to analyze a huge amount of SNPs at the same time. As a result of this step, SNPs affecting a gene are identified and annotated with the information in Ensembl (gene, transcripts, exon/intron, etc.).

C. SNP Annotation in TFBSs

In our methodology, the TRANSFAC [9] and Jaspar [10] databases are queried in order to determine if a SNP is located in a TFBS. These databases contain a list of motif sequences where each known transcription factor tends to bind, and a position-specific scoring matrix (PSSM) that provides information about the frequencies of each base in each position of these sequences.

Determining whether a given sequence is similar to a motif sequence for a TFBS is not a trivial problem and many similarity measures have recently been proposed [1] [6] [7] [8]. Our approach uses SCintuit [1] to obtain a similarity score for a query sequence and a TFBS motif. SCintuit was shown to outperform other similarity measures such as SCindep, SCdep [7] and SCmat [8].

While most of the measures only use the PSSMs, SCintuit also uses the sequences list, since PSSMs do not provide information about which sequence carries each base. Therefore, SCintuit is able to assign better scores to query sequences that are similar to the majority of the sequences in the TFBS motif and penalize the sequences that only show good correlation for some of the positions.

SCintuit works by representing each pairwise combination of the sequences positions as an intuitionistic set. An intuitionistic set is a set based on the Fuzzy Set Theory [19]. The Fuzzy Set Theory uses a membership function to determine the membership degree of an item to the set. The intuitionistic theory modifies the fuzzy sets theory by adding the concept of a non-membership function that helps to classify inconclusive data [20]. Let X be the universe of discourse. An intuitionistic fuzzy set A in X is an object having the form:

$$A = \{(x, \mu_A(x), \nu_A(x)) : x \in X\} \qquad (2.1)$$

where $\mu_A, \nu_A : X \rightarrow [0,1]$ denote the membership function and non-membership function respectively of A, satisfying $0 \leq \mu_A + \nu_A \leq 1$ for every $x \in X$. Therefore, the degree of uncertainty of x to A is $\pi_A(x) = 1 - \mu_A - \nu_A$.

The universe of discourse used by SCintuit is $B \times B$, in other words, the set of all 16 possible combinations of bases for two given positions (AA, AC, …, TT).

The membership function is calculated using the following expression:

$$\mu_{I^M_{i,j}}(b_1,b_2) = P(b_1,b_2,i,j) + (1 - P(b_1,b_2,i,j))\frac{P(b_1,i)+P(b_2,j)}{2} \qquad (2.2)$$

where $I^M_{i,j}$ represents the intuitionistic set for the columns i and j of the TFBS representation discussed above that is called motif, $P(b_1, b_2, i, j)$ represents the probability that the bases b_1 and b_2 appears in the position i and j on a sequence of the motif, $\mu_{I^M_{i,j}}(b_1,b_2)$ is in the range $0 \leq \mu_{I^M_{i,j}}(b_1,b_2) \leq 1$ and $P(b_1, i)$ represents the probability that b_1 appears in the position i of a sequence, this last piece of information is located in the PSSM.

The non-membership function is calculated using the following expression:

$$\nu_{I^M_{i,j}}(b_1,b_2) = \left(\frac{IC_i^{b_1} + IC_j^{b_2}}{2}\right)(1 - \mu_{I^M_{i,j}}) \qquad (2.3)$$

where $IC_p^b = \dfrac{2 + P(b,p)\log_2(P(b,p))}{2}$ is the normalized information content of base b in position p, $\nu_{I^M_{i,j}}(b_1,b_2)$ is in the range $0 \leq \nu_{I^M_{i,j}}(b_1,b_2) \leq 1$ and $\nu_{I^M_{i,j}}(b_1,b_2) + \mu_{I^M_{i,j}}(b_1,b_2) \leq 1$.

The score for a two bases DNA string is calculated using the membership and non-membership functions resolving the next expression:

$$SC^{i,j}_{\text{intuit}}(b_1,b_2) = \mu_{I^M_{i,j}}(b_1,b_2)(\max(\nu_{I^M_{i,j}}) - \nu_{I^M_{i,j}}(b_1,b_2)) \qquad (2.4)$$

where $\max(\nu_{I^M_{i,j}})$ is the maximum degree of non-membership found in the pair of positions i and j considering all the combinations of $b_1, b_2 \in B^2$.

In order to be able to compare scores, a normalization step needs to be performed:

$$NSC^{i,j}_{\text{intuit}}(b_1,b_2) = \frac{SC^{i,j}_{\text{intuit}}(b_1,b_2) - \min(SC^{i,j}_{\text{intuit}})}{\max(SC^{i,j}_{\text{intuit}}) - \min(SC^{i,j}_{\text{intuit}})} \qquad (2.5)$$

where $\min(SC^{i,j}_{\text{intuit}})$ and $\max(SC^{i,j}_{\text{intuit}})$ are the minimum and maximum scores respectively in the position (i, j) of the motif.

Finally, for a given DNA sequence S of length n the SCintuit can be computed as:

$$SC_{\text{intuit}} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} NSC^{i,j}_{\text{intuit}}(S_i, S_j) \qquad (2.6)$$

Two searches against the TFBS database are needed: one for the mutated DNA sequence and other for the original sequence. If one of the two sequences matches a TFBS from the database and the other one does not, we can conclude that a SNP that is silencing/activating a TFBS has been potentially found. In case that only the mutated chain matched a TFBS, we can hypothesize that the presence of the SNP has significantly increased the binding affinity of the selected region to the associated TF. In case that only the original sequence matched a TFBS, we can argue that the presence of the SNP has reduced the binding affinity of the associated TF to the given sequence. In both cases, the same conclusion is lead: the transcriptional regulation may be affected by the presence or absence of SNPs located in potential TFBSs. This, in turn, may alter the regulation of the genes controlled by the associated TFs.

Each sequence containing a SNP is compared with all the motifs stored in these databases using a sliding window whose size depends on the length of the consensus sequence of the motif. Only the TFBSs matching the query sequence with a similarity value above a given threshold are stored.

III. EXPERIMENTS

We have applied the proposed methodology on the genome ID 96 from the 1000 genome project [21]. The SNP calling was performed on the whole genome. The reads were mapped to the most recent human reference genome (g1k_v37). This SNP calling provided us with 2598644 SNPs (satisfying our expectations of around 3 million SNPs [3]).

To carry out a detailed SNP annotation, we chose the chromosome 15 due to its known high variability. Particularly, the first 37990 SNPs found in that chromosome were studied.

A. SNP Annotation in Genes

The set of SNPs located in the chromosome 15 are searched using the Ensembl database [18]. Five examples of genes having some of these SNPs in them are showed in Table 1.

TABLE I.

SNP id	Gene symbol	Description
15-20165227	VSIG7	V-set and immunoglobulin domain containing 7.
15-20590186	HERC2P3	Hect domain and RLD2 pseudogene 3.
15-22087625	POTEB	POTE ankyrin domain family, member B.
15-22363754	OR4M2	Olfactory receptors interact with odorant molecules in the nose, to initiate neuronal response that triggers the perception of smell.
15-22510355	MIR1268	MicroRNA 1268. MicroRNAs are short non-coding RNAs that are involved in post-transcriptional regulation of gene expression.

a. Example of some genes containing SNPs in their exons. SNP id indicates the chromosome and the position of the SNP.

B. P-Match and SCintuit Comparison

A comparison of SCintuit with the official TRANSFAC web page search method called P-Match [22] has been done in order to check which one fits the SNP annotation in TFBSs better. P-match has three different forms of filtering the obtained results: minimize false positives, minimize false negatives and minimizes sum of squared errors. Depending of the filter, the output of the method varies.

P-match and SCintuit provide similar results when the motif is very well conserved. However, when the motif includes poorly-conserved positions, SCintuit outperforms P-match, since it discriminates between the relative importance of poorly-conserved positions and well-conserved positions. An illustrative example is shown in Table 2 and Figure 1.

In the first example of Table 2, SCintuit found a high similarity of the sequence with the M00278 motif, while P-match does not find any hits (using the sums of errors filter). We can see that the sequence found matches positions 3, 4, 5 and 6 of the motif (the best conserved). Positions 8 and 9 have a G predomination that also matches the sequence.

In the second example, SCintuit and P-match return a different motif matching the same query sequence. Figure 1 shows the sequence logos for both motifs. A visual inspection of these sequence logos confirms that M00117 seems to be more similar to the query sequence than M00109. Although positions 4, 5, 10 and 11 are well-conserved in both motifs, motif M00109 presents a very well-conserved G in position 8, while the query sequence has an A. Furthermore, the motif M00117 shows a preference for G in position 6 and C in position 7 while M00109 shows no preference for those positions.

The other P-Match filters do not show much improvement. The minimum false positives filter does not find any results for the two examples. The minimum false negatives filter finds the match M00278 for the first example. However, it finds the match M00119 for the second example, which is not as similar to the query sequence as M00109.

TABLE II.

DNA sequence searched	SCintuit Motif ID	Motif binding sequences	P-match Motif ID	Motif binding sequences
CAGAT AAGG	M00278	GCGATAAGG ACGATAAGG GCGATACGG CCGATAGCC GCGATAGTC TTGATAAGC CCGATAGCG GCGATACGA CAGATAAAC TAGATAACG TAGATAGTG CTGATACCG NAGATAGGC TAGATACGC CTGATATGA CAGATAGTG CCGATAGGC ACGATACGG CCGATATGG CAGATAACG ACGATAGCG CTGATACCG AAGATACCG AAGATAACA GTGATAATT CTGATATCG CAGATAACC CAGATAGGG GAGATAGGG TAGATAAGG CAGATAGGG	No results	
TGGTT GCACA ACTC	M00117	GTCTTGCGCAAAAT TATGTGCGCACCTG AGGTTGGGAAATCC ATGTTGCACAAGAG GTATTGAGAAATCT GAGTTGCGCCACTA ATGTTGCTTAAATG TGATTACATCACTG TGTTTGCACAACGT CTCTTGCTTCATAC ACATTGCCCAACTC CTCTTAGGAAATGC TGATTGCAAAACTA TCATTTCCTGTGCA AGGTTACACAACTG ATCTTGCATAATCC GGATTGCTCCATTA	M00109	AGATTGTGCAATGT ACATTGTGCAATCT TGATTTTGTAATGG AGATTGAGCAATCT GTCTTAAGCAAAGC GTATTAGGACATGT ATGTTGAGTAAGAT GTGTGAAGCAAGAG GAATTACGAAATGG AAGTTGTGCAATGG TAGTGGCGCAAACT CAGTGATGTAATCA AAGTTGAGAAATTT GGCTGAGGAAATAC ACAAGTTGCAACAT GGATTTGGAAAGTT GTTTTGTGAAATCG ACATAATGAAAAGA GGGTGTGGAAAGTC TGGTTTTGCAAGAG ATCTGTGGTAAGCA

a. Example of a sequence search in TRANSFAC with the SCintuit and the P-Match method using the sum of both errors filter. First column shows the query sequence. Second and third columns, the motif ID and motif sequences found by SCintuit. Fourth and fifth columns, the motif ID and motif sequences found by P-Match.

C. SNP Annotation in TFBS Results

The TRANSFAC database was searched in order to find out if the SNPs are located in TFBSs. We used SCintuit to search for TFBS sequences matching the sequences with SNPs found in chromosome 15. Out of the 37990 SNP sequences, 9942 matching entries from the TRANSFAC database were found (26% of the total, using a similarity score above 0.9) in 19 hours in a server with 15x4 cores at 2’67 GHZ.

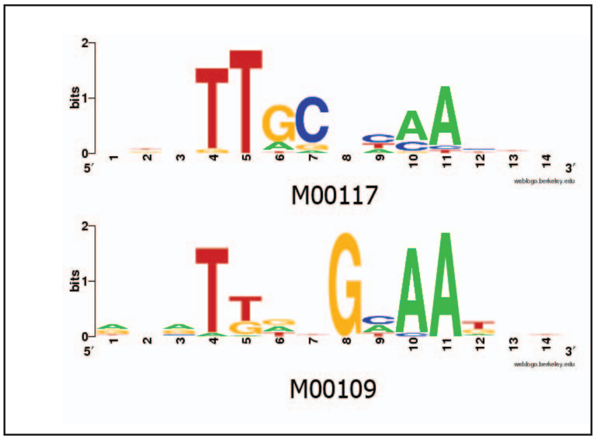

Figure 1. Logo representations of the motifs M00117 and M00109 for the sequence: TGGTTGCACAACTC.

A subset of the results is shown in Table 3. We can see that some of the motifs found in TRANSFAC are associated to the transcription factors E2F1, DBP, LBP1, LEF1, TBP, and USF2 (Table 3). Upstream transcription factors (DBP, LBP1, LEF1, TBP, and USF2) bind somewhere upstream of the initiation site to stimulate or repress transcription.

The TATA-binding protein (TBP) is a transcription factor that binds specifically to a DNA sequence called the TATA box. This DNA sequence is found about 35 base pairs upstream of the transcription start site in some eukaryotic gene promoters.

TABLE III.

Motif ID	Sequence	Gene symbol	Official symbol	Gene Description
M00801	CGTCAC	**CREB**	CREB1	cAMP responsive element binding protein 1
M00624	AGTACAC	**DBP**	DBP	D site of albumin promoter
M00803	CCCGCC	**E2F**	E2F1	E2F transcription factor 1
M00695	GCGGAGA	**ETF**	TEAD2	TEA domain family member 2
M00789	AGACAGG	**GATA**	GATA1	GATA binding protein 1
M00921	AGAACAGA	**GR**	NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
M01033	AGGGCA	**HNF4**	HNF4A	hepatocyte nuclear factor 4, alpha
M00747	TTCACTT	**IRF-1**	IRF1	interferon regulatory factor 1
M00644	CAGCCGC	**LBP1**	LBP1	upstream binding protein 1 (LBP-1a)
M00805	TCAGAG	**LEF1**	LEF1	upstream binding protein 1 (LBP-1a)
M00655	ACATCCG	**PEA3**	ETV4	ets variant 4
M00960	GAGAGGACAT	**PR**	PGR	progesterone receptor
M01131	CTTCGTA	**SOX10**	SOX10	SRY (sex determining region Y)-box 10
M00148	TTTGTTC	**SRY**	SRY	sex determining region Y
M00980	TTTATAG	**TBP**	TBP	TATA box binding protein
M00704	CATTCC	**TEF-1**	USP7	ubiquitin specific peptidase 7 (herpes virus-associated)
M00726	CACGTG	**USF2**	USF2	upstream transcription factor 2
M00924	TGACTCACAGGG	**AP-1**	FOSB	FBJ murine osteosarcoma viral oncogene homolog B

a. Examples of genes that are regulated by TFBSs that have SNPs on them. The motif ID shows the ID of the TRANSFAC database for that motif. The sequence gives the sequence of the motif with the SNP highlighted. Gene symbol represents the symbol of the gene that appears in the TRANSFAC database, the official symbol is the standard symbol given by Refseq and the description gives some information about the gene.

Two searches were made for each SNP, one for the context sequence without the mutation and other with it. We focus on the matches showing a high similarity value (above 0,9) for one sequence and a low similarity value (below 0,8) for the other, since they represent cases where the presence or absence of the SNP may alter the binding affinity of the TF to the TFBS (see Table 4). High and low similarity threshold values were determined experimentally. Particularly, four different cases can be differentiated:

- If a TFBS is found in the mutated chain and not found in the original chain, it means that the SNP has significantly increased the binding affinity of a TF to the TFBS. The TF binding there might change the transcription process of the gene the TFBS regulate. This is the case of TFBSs of the TBP, LBP1, LEF1, IRF-1, GATA, GR, DBP and E2F genes.
- If a TFBS is found in the reference sequence and not found in the mutated one, it means that the SNP has silenced the TFBS, reducing the binding affinity of the associated TFs. That implies that the gene affected is not going to be regulated by that TFBS so the transcription process is modified. This is the case of TFBSs of the TEF-1, USF2, HNF4, CREB, PEA3 and ETF genes.
- If the same TFBS is found in both cases, we hypothesize that the SNP does not affect the binding affinity of the TF to the TFBS.
- If one TFBS is found in the non-mutated sequence and another TFBS is found in the mutated chain, we may argue that the SNP affected a well-conserved position of the TFBS, increasing the binding affinity of a different TF to that sequence. An example of this type of modification is given in the second example of Table 4: there is a mutation in the negative strand of the region (20076473 - 20076553) of the chromosome 15 and the TFBS of the HNF4 changes into a new TFBS that regulates the AP-1 gen.

TABLE IV.

Sequence ID	TRANSFAC TFBS ID	TRANSFAC Gen Symbol	Similarity not-mutated	Similarity mutated	Strand direction
15(20134248-20134328)	M01032	HNF4	0.696426	0.932944	-
15(20076473-20076553)	M00924	AP-1	0.807852	0.921382	-
	M01033	HNF4	0.967851	0.713052	+
	M00926	AP-1	0.760734	0.926963	+
15(24997660-24997740)	M01033	HNF4	0.97464	0.682869	-
15(20454362-20454442)	M01033	HNF4	0.627531	0.914015	-
	M00695	ETF	0.976793	0.759638	+

a. Some search results examples of TFBSs. The first column indicates the sequence id that is constructed by the chromosome number and the sequence region inside that chromosome. The second column shows the TRANSFAC TFBS id. The third column indicates the gen that is regulated by the TFBS. The fourth and fifth columns show the SCintuit value for the sequence with the SNP and the reference one. The sixth column specify the strand of the sequence that is being examined, "+"means that the TFBS has been found in the positive strand and "–" in the negative strand.

Table 4 shows some other conclusions of interest. First, one SNP may affect two different TFBSs. Furthermore, these TFBS may be located at different strands of the DNA chain. These TFBSs do not need to be related, for example, the last example shows a TFBS in the negative strand that regulates the gene HNF4 and another one in the positive strand that regulates the gene ETF. Furthermore, the same TFBS may appear in different regions of the genome (for example, M01033).

IV. CONCLUSIONS

Annotating SNPs in whole-genome sequences is a complex task that requires new computational tools to be developed. In this paper, a novel methodology that gives us a mechanism to detect accurately if a SNP is located in a TFBS and if the binding affinity of the TFBS is affected by the mutation is proposed. This tool makes heavy use of SCintuit, a similarity measure for DNA sequences which is based on intuitionistic sets.

ACKNOWLEDGMENT

This work has been carried out as part of projects P08-TIC-4299 of J. A., Sevilla, TIN2009-13489 of DGICT, Madrid, GREIB-PYR-2010-05 of University of Granada (MC) and GREIB-PYR-2010-02 of University of Granada (CC).
The data used in these experiments were provided by the 1000 Genomes Project Consortium and the Wellcome Trust Sanger Institute and can be obtained from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ and ftp://ftp.sanger.ac.uk/pub/1000genomes/tk2/main_project_reference/.

REFERENCES

[1] F. Garcia, A. Blanco and A. J. Sheperd, "An intuitionist approach to scoring DNA sequences against transcription factor binding site motifs", BMC Bioinformatics, vol. 11:551, 2010.
[2] L. W. Hillier et al., "Whole-genome sequencing and variant discovery in C. elegans", Nature Methods, vol. 5, pp. 183–188, 2008.
[3] Z. Zhao, Y-X. Fu, D. Heweet and E. Boerwinkle. "Investigating single nucleotide polymorphishm (SNP) density in the human genome and its implications for molecular evolution", Science Direct, vol. 312, pp. 207-213. 2003.
[4] M. Sana et al., "GAMES identifies and annotates mutations in next-generation sequencing projects", Bioinformatics, vol. 27, pp. 9-13, October 2010.
[5] J. Setubal and J. Meidanis, "introduction to computational molecular biology", PWS Publishing Company, Boston, 1997.
[6] F. Garcia, F. J. López, C. Cano and A. Blanco, "FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral", BMC Bioinformatics, vol. 10:224. 2009.
[7] A. Tomovic and E. Oakeleyñ "Position dependencies in transcription factor binding sites", BMC Bioinformatics, vol. 23, pp. 933. 2007.
[8] F. Zare-Mirakabad, H. Ahrabian, M. Sadeghi, A. N. Dalini and B. Goliaiei, "New scoring schema for finding motifs in DNA sequences", BMC Bioinformatics, vol. 10, pp. 94. 2009.
[9] V. Matys et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles", Nucleic Acids Research, vol. 31, pp. 374-378. 2003.
[10] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman and B. Lenhard, "JASPAR: an open-access database for eukaryotic transcription binding sites profiles", Nucleic Acids Research, vol. 32, pp. D91-D94. 2004.
[11] D. Ge et al., "WGAViewer: Software for Genomic Annotation of Whole Genome Association Studies", Genome Research, vol. 18, pp. 640-643. 2008.
[12] A. D. Johnson et al., "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap", Bioinformatics, vol. 24, pp. 2938-2939. October 2008.
[13] SAMTOOLS webpage, http://samtools.sourceforge.net/
[14] A. V. Dalca, S. M. Rumble, S. Levy and M. Brudno, "VARiD: A variation detection framework for color-space and letter-space platforms", Bioinformatics, vol. 26(12), pp. i343-i349. 2010.
[15] S. Fröhler and C. Dieterich. "ACCUSA – accurate SNP calling on draft genomes", Bioinformatics, vol. 26, pp. 1364-1365. 2010.
[16] H. Li, J. Ruan and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores", Genome Research, vol. 18, pp. 1851-1858, August 2008.
[17] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform", Bioinformatics, vol, 25, pps. 1754-1760. May 2009.
[18] T. Hubbard et al. "The ensembl genome database project", Nucleic Acids Research, vol. 30, pps. 38-41.2002.
[19] L. Zadeh. "Fuzzy Sets", Information and Control, vol. 8, pp. 338-353. 1965.
[20] K. Atanassov, "Intuitionistic Fuzzy Sets: theory and applications", Physica-Verlag, Heidelberg. New York 1999.
[21] N. Siva. "1000 Genomes Project", Nature Biotechnology, vol. 26, pp. 256. 2008.
[22] D.S. Chekmenev, C. Haid and A. E. Kel, "PMatch: transcriptcion binding site search by combining pattern and weight matrices", Nucleic Acids Research, vol. 33(supp. 2), pp. W432-W437. March 2005.