Detección de mutaciones en TFBSs mediante tecnología difusa

J. A. Morente-Molinera[†], C. Cano[†], M. Cuadros[†], J. M. Martín[§], A. Blanco[†]
[†] Departamento de Ciencias de la Computación e I.A. Universidad de Granada.
[§] Departamento de Tecnologías de la Información, Universidad de Huelva.
E-mail: jamoren@correo.ugr.es, {ccano, marta, armando}@decsai.ugr.es, jmmartin@dti.uhu.es

***Abstract*—Las tecnologías de secuenciación masiva están produciendo un enorme volumen de datos genómicos cuyo análisis requiere la aplicación y desarrollo de nuevas técnicas computacionales capaces de tratar la dimensionalidad y complejidad de estos datos.**

Los polimorfismos de nucleótido simple (Single Nucleotide Polymorphisms -SNPs-) son una de las formas más comunes de variación genética, y miles de estas variaciones han sido descritas en la literatura especializada como marcadores de enfermedades. Identificar SNPs y anotar su papel funcional en el organismo humano resulta, por lo tanto, de una gran relevancia. Esta tarea es habitualmente desempeñada por expertos humanos, que leen la literatura biomédica y extraen el conocimiento relevante para anotar, en bases de datos especializadas, la información sobre cada SNP. Existen herramientas software que asisten a los expertos en este proceso, pero ninguna de estas herramientas estudia la asociación de SNPs a regiones reguladoras como los Sitios de Unión de Factores de Transcripción (Transcription Factor Binding Sites -TFBSs-).

Este artículo presenta una metodología para ayudar a los expertos humanos en la anotación de SNPs en secuencias de genoma completo, incluyendo no sólo genes sino también TFBSs. Nuestra principal aportación es el uso de una medida novedosa, CSintuit, basada en tecnología difusa y conjuntos intuicionistas, para efectuar comparaciones más precisas entre secuencias de ADN y detectar TFBSs afectados por SNPs.

***Index Terms*—bioinformática, regulación genética, conjuntos intuicionistas, tecnología difusa, detección y anotación de SNPs**

I. Introducción

Actualmente, las tecnologías de secuenciación de ADN generan una gran cantidad de datos que hacen necesario el desarrollo de herramientas computacionales que nos permitan analizar y extraer conocimiento relevante de los mismos [1]. Este conocimiento ayuda a los científicos a entender mejor cómo funcionan los organismos vivos y a averiguar cómo se producen las enfermedades, de forma que puedan desarrollarse métodos adecuados para prevenirlas.

Estudios recientes han demostrado que los polimorfismos de nucleótido simple (SNP) tienen un papel muy importante en los desórdenes sistémicos. Los SNPs son mutaciones en la cadena de ADN qué únicamente afectan a un nucleótido, es decir, consisten en la sustitución de una base por otra en la cadena de ADN. Dependiendo de en qué lugar se produzcan, esta mutación puede alterar drásticamente un proceso de regulación celular o no presentar ningún efecto. De hecho, un genoma humano puede llegar a tener más de tres millones de SNPs, de los cuales se conocen, al menos, 9000 cambios en aminoácidos que hacen que varíe la proteína generada [23]. Por tanto, anotar los efectos que dichos SNPs pueden producir es una tarea fundamental para entender los mecanismos de regulación y asignar relevancia clínica a estas mutaciones [2].

Tradicionalmente, se pensaba que los SNPs sólo afectaban a la regulación de los organismos cuando éstos estaban localizados en las regiones de los genes que codifican proteínas (exones). En efecto, los cambios producidos en exones son habitualmente trasladados a la secuencia de aminoácidos de la proteína, lo que a su vez puede modificar las propiedades y estructura terciaria de la misma [3].

Sin embargo, la regulación celular es un proceso complejo en el que intervienen multitud de mecanismos. La transcripción de un gen es uno de los procesos clave en la regulación celular. Este proceso comienza cuando una o varias proteínas, llamadas factores de transcripción (TF), se unen a la cadena de DNA situada cerca del gen en unos lugares denominados Sitios de Unión de Factores de Transcripción (Transcription Factor Binding Site -TFBS-). Estas interacciones proteína-ADN juegan un papel crucial en la regulación de la transcripción, ya que activan e inhiben la maquinaria de transcripción de la célula. Por tanto, estudiar dichas interacciones y encontrar los TFBSs asociados a los TFs en las cadenas de ADN se ha convertido en un problema que está atrayendo la atención de la comunidad científica [4], [18], [15], [16]. Además, existen bases de datos como Jaspar [11] y TRANSFAC [12] que contienen TFs y sus motivos de unión conocidos.

Debido a la importancia de los TFs y TFBSs en la regulación de la transcripción, las herramientas de anotación de genomas deberían ser capaces de permitir el análisis de aquellos SNPs localizados en TFBSs. Con ello, se podrían estudiar más profundamente los procesos de regulación y descubrir la asociación de SNPs con diversos fenotipos de enfermedades. Sin embargo, las herramientas de anotación disponibles actualmente se centran únicamente en anotar SNPs en genes y exones [2], [19], [20].

En este artículo se propone una nueva herramienta que permite anotar secuencias genómicas pertenecientes a cualquier zona del genoma. La anotación se realiza usando información tanto de genes como de TFBSs conocidos. El procedimiento se describe a continuación. Primero, se alinean las secuencias obtenidas mediante las tecnologías de secuenciación de ADN contra un genoma de referencia. Después, se buscan los SNPs en el alineamiento generado. Una vez identificados los SNPs

y sus secuencias contexto (secuencias formadas por las bases que rodean al SNP), éstos son buscados en diferentes bases de datos de secuencias pertenecientes a genes, SNPs y TFBSs para recuperar la información asociada a la localización del SNP. Una de las claves de este proceso es que la búsqueda en bases de datos de TFBSs se realiza utilizando la medida CSintuit, basada en conjuntos intuicionistas, que nos permite analizar las secuencias con y sin SNP.

El artículo está organizado como sigue. La sección Métodos describe los diferentes componentes de la herramienta y los pasos que se llevan a cabo para realizar el análisis. Esta sección incluye una descripción del proceso de búsqueda de SNPs y la anotación de dichos SNPs mediante consultas en bases de datos especializadas utilizando la medida CSintuit para comparar secuencias de ADN. Por último, la sección Resultados muestra un ejemplo de anotación de SNPs en un cromosoma humano y un resumen de los resultados encontrados.

II. MÉTODOS

La metodología propuesta consiste en la creación de un *pipeline* que consta de dos pasos: la búsqueda de SNPs y la anotación de los mismos. El primer paso aborda el alineamiento de secuencias contra el genoma de referencia y analiza dicho alineamiento para detectar las variaciones (SNPs). La anotación de SNPs consiste en identificar si los SNPs encontrados forman parte de un gen o de un TFBS para poder asignarles un papel funcional a los mismos. En este proceso proponemos el uso de CSintuit, una medida de similitud basada en conjuntos intuicionistas que permite comparar cadenas de ADN de forma más precisa.

A. Búsqueda de SNPs

La fase de búsqueda de SNPs tiene dos partes, la fase de mapeo, en donde las secuencias son alineadas a un genoma de referencia, y la fase de búsqueda de SNPs en la que, utilizando un método bayesiano, cada posición del alineamiento es analizada con el objetivo de determinar si dicha posición contiene un SNP. Los detalles sobre estos métodos bayesianos pueden consultarse en otros trabajos [6], [7]. Para el alineamiento utilizamos el algoritmo de Burrows-Wheeler (BWA) [8].

B. Anotación de SNPs

Tras identificar SNPs en el genoma analizado, éstos deben ser anotados para identificar su funcionalidad y, en su caso, su relevancia clínica. Para ello, los SNPs encontrados son buscados en la base de datos dbSNP[9], utilizando el algoritmo BLAST [21], para determinar si el SNP identificado ya ha sido documentado en trabajos previos. Si el SNP no se encuentra en dbSNP, se realiza una búsqueda en la base de datos Genbank [10] para determinar si el SNP está localizado en un gen.

Anotación de SNPs en TFBSs: Para determinar si un SNP forma parte de un TFBS, utilizamos las bases de datos TRANSFAC [11] y Jaspar [12]. Estas bases de datos contienen una lista de secuencias motivo conocidas a las que cada factor de transcripción tiende a unirse, junto con una matriz PSSM (position-specific scoring matrix) que proporciona información acerca de las frecuencias de aparición de cada base en cada posición de la secuencia consenso. Determinar si una secuencia con un SNP está asociada a un TFBS no resulta trivial, y la comparación de cadenas de ADN continúa siendo un problema abierto que atrae la atención de la comunidad científica [4], [15], [16].

Mientras que la mayoría de las medidas actuales utilizan únicamente la matriz PSSM, CSintuit [4] hace uso de la información que porta esta matriz para asignar valores de similitud en función del grado de conservación de las bases de las distintas posiciones de la secuencia. De este modo, nuestro enfoque utiliza CSintuit [4] para obtener un valor de similitud entre una secuencia de consulta y un motivo TFBS.

CSintuit representa cada combinación de dos posiciones en la secuencia como un conjunto intuicionista. La definición de conjunto intuicionista está basada en la teoría de conjuntos difusos [13]. La teoría de conjuntos difusos utiliza una función de pertenencia para determinar el grado de pertenencia de un elemento al conjunto. La teoría de conjuntos intuicionistas añade el concepto de función de *no pertenencia* [14]. Sea X el universo de discurso. Un conjunto difuso intuicionista A en X se define como:

$$A = \{(x, \mu_A(x), \nu_A(x)) : x \in X\}, \tag{1}$$

donde $\mu_A, \nu_A : X \to [0, 1]$ denota la función de pertenencia y no pertenencia de A, respectivamente, de forma que se satisface: $0 \leq \mu_A + \nu_A \leq 1$ para todo $x \in X$. Por lo tanto, el grado de incertidumbre de x en A es $\pi_A(x) = 1 - \mu_A - \nu_A$.

El universo de discurso usado por CSintuit es $B \times B$ es decir, el conjunto de las 16 combinaciones posibles de dos bases (AA, AC, ..., TT).

La función de pertenencia se calcula usando la siguiente expresión:

$$\begin{aligned}\mu_{I^M_{i,j}}(b_1, b_2) =& P(b_1, b_2, i, j) + \\ &+ (1 - P(b_1, b_2, i, j))\frac{P(b_1, i) + P(b_2, j)}{2},\end{aligned} \tag{2}$$

donde $\mu_{I^M_{i,j}}$ representa el conjunto intuicionista de las columnas i y j de la representación del TFBS ya comentada a la que llamaremos motivo, $P(b_1, b_2, i, j)$ es la probabilidad de que las bases b_1 y b_2 aparezcan en la posición i y j en una de las secuencias del motivo, $\mu_{I^M_{i,j}}(b_1, b_2)$ está en el rango $0 \leq \mu_{I^M_{i,j}}(b_1, b_2) \leq 1$, y $P(b_1, i)$ es la probabilidad de que b_1 aparezca en la posición i de una secuencia, información que puede obtenerse de la matriz PSSM.

La función de no pertenencia puede calcularse mediante la siguiente expresión:

$$\nu_{I^M_{i,j}}(b_1, b_2) = \left(\frac{IC^{b_1}_i + IC^{b_2}_j}{2}\right)(1 - \mu_{I^M_{i,j}}(b_1, b_2)), \tag{3}$$

donde $IC^b_p = \frac{2+P(b,p)\log_2(P(b,p))}{2}$ es el contenido de información normalizada de la base b en la posición p, $\nu_{I^M_{i,j}}(b_1, b_2)$ está en el rango $0 \leq \nu_{I^M_{i,j}}(b_1, b_2) \leq 1$, y se cumple que $\mu_{I^M_{i,j}}(b_1, b_2) + \nu_{I^M_{i,j}}(b_1, b_2) \leq 1$.

El valor de la medida para una cadena de ADN de dos bases se calcula a partir de la función de pertenencia y no pertenencia mediante la siguiente expresión:

$$SC^{i,j}_{intuit}(b_1, b_2) = \mu_{I^M_{i,j}}(b_1, b_2)(\max(\nu_{I^M_{i,j}}) - \nu_{I^M_{i,j}}(b_1, b_2)), \quad (4)$$

donde $\max(\nu_{I^M_{i,j}})$ es el máximo grado de no pertenencia encontrado para las posiciones i y j considerando todas las combinaciones posibles de $b_1, b_2 \in B^2$.

Para poder realizar comparaciones entre los distintos valores obtenidos es necesario realizar un proceso de normalización:

$$NSC^{i,j}_{intuit}(b_1, b_2) = \frac{SC^{i,j}_{intuit}(b_1, b_2) - \min(SC^{i,j}_{intuit}(b_1, b_2))}{\max(SC^{i,j}_{intuit}(b_1, b_2)) - \min(SC^{i,j}_{intuit}(b_1, b_2))}. \quad (5)$$

donde $\min(SC^{i,j}_{intuit})$ y $\max(SC^{i,j}_{intuit})$ son los valores mínimos y máximos obtenidos, respectivamente, en la posición (i, j) del motivo.

Por último, para una cadena de ADN S de tamaño n, la medida CSintuit puede calcularse del siguiente modo:

$$SC_{intuit} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} NSC^{i,j}_{intuit}(S_i, S_j). \quad (6)$$

Gracias a la precisión de CSintuit en la comparación de cadenas de ADN, resulta muy conveniente utilizar esta medida para estudiar si un SNP afecta de forma crítica a un TFBS. Para esto, se realizan dos búsquedas, una para la cadena de ADN mutada y otra para la cadena original. Si una de las dos cadenas es similar a un motivo y la otra no, podemos concluir que el SNP está silenciando o activando el TFBS correspondiente. En el caso de que la cadena mutada sea la que se corresponda con un TFBS, podríamos concluir que el SNP ha incrementado significativamente la afinidad de unión del TF correspondiente a la cadena donde aparece el SNP. Si el TFBS se encuentra únicamente en la secuencia original, podríamos argumentar que la aparición del SNP en esta secuencia está reduciendo la afinidad de unión a la misma del TF. En ambos casos llegamos a la conclusión de que el proceso de transcripción está siendo alterado por la ocurrencia del SNP.

III. Resultados

La metodología propuesta ha sido aplicada al análisis del genoma número 96 del proyecto de los 1000 genomas [17]. Las secuencias pertenecientes al sujeto han sido alineadas al genoma de referencia humano más reciente, la versión g1k_v37. El proceso de identificación de SNPs se ejecutó sobre todo el genoma, obteniendo un total de 2598644 SNPs. Este resultado es acorde al de trabajos previos, que establecen un número de SNPs cercano a 3 millones al comparar dos genomas humanos [23]. Para un análisis más en profundidad, elegimos los primeros 37990 SNPs localizados en el cromosoma 15, debido a su gran variabilidad, y procedemos con la fase de anotación de los mismos. En particular, centramos nuestra atención en la anotación de SNPs en TFBSs.

Para ello, la base de datos TRANSFAC es consultada para identificar SNPs localizados en TFBSs. Esta consulta se ejecuta utilizando CSintuit para identificar secuencias de TFBSs similares a las secuencias en las que se detectaron los SNPs en el cromosoma estudiado. Tras realizar la búsqueda encontramos un total de 9942 resultados (26% de los 3790 SNPs del cromosoma 15), considerando un umbral inferior para el valor de similitud de 0.9. La Tabla I muestra algunos resultados obtenidos. Algunos de los motivos incluyen TFs tales como E2F1, DBP, LBP1, LEF1, TBP y USF2. Los factores de transcripción localizados en la dirección 3' (DBP, LBP1, LEF1, TBP, and USF2) se unen cerca del lugar de iniciación de la transcripción para estimularla o reprimirla.

Para cada SNP realizamos dos búsquedas, una para la secuencia contexto con la mutación y otra sin ella. De entre todos los resultados nos centramos en aquellos que muestran un valor de similitud alto (superior a 0.9) para una secuencia y bajo (inferior a 0.8) para otra, ya que en estos casos la presencia o ausencia del SNP está alterando significativamente la afinidad de unión del TF al TFBS (ver Tabla II). Entre los resultados encontrados caben destacar algunos TFBSs que se encontraron en las cadenas con mutación y no se encontraron en las cadenas sin mutación (como los TFBSs de los genes TBP ,LBP1, LEF1, IRF-1, GATA, GR, DBP y E2F) y casos en los que los TFBSs se encontraron en las cadenas de referencia, y no en las cadenas mutadas (como TFBSs para los genes TEF-1, USF2, HNF4, CREB, PEA3 y ETF). Además, en algunos casos se encontró un TFBS en la cadena mutada y otro distinto en la cadena no mutada, por ejemplo, en la Tabla II se muestra una mutación en la región (20076473 - 20076553) de la secuencia negativa del cromosoma 15 y el TFBS asociado al gen HNF4 cambia pasando a convertirse en el del gen AP-1.

La Tabla II muestra algunos otros casos interesantes. Por ejemplo, las dos últimas filas presentan un caso que muestra que, para la misma posición del genoma, pueden encontrarse dos TFBSs distintos, uno para cada dirección de la cadena de ADN. Dichos TFBS no tienen por qué estar relacionados: en la dirección negativa hay un TFBS que regula el gen HNF4 y en la dirección positiva uno que regula el gen ETF. Además, un mismo TFBS puede aparecer en diferentes zonas del genoma tal y como se aprecia en la Tabla II respecto al TFBS M01033.

IV. Conclusión

Anotar SNPs en secuencias de genoma competo es una tarea compleja y costosa que requiere nuevas herramientas computacionales. En este trabajo se ha presentado una metodología que propone el uso de una medida de similitud basada en conjuntos intuicionistas para realizar comparaciones precisas de cadenas de ADN y poder asignar SNPs a TFBSs. Del mismo modo, se ha mostrado un procedimiento que permite detectar si la afinidad de la unión del TF sobre un TFBS se ve significativamente alterada por la presencia de un SNP. Este tipo de técnicas tienen una creciente demanda en el ámbito del análisis de datos genómicos generados por las nuevas tecnologías de secuenciación, y su perfeccionamiento permitirá ahondar en el estudio de los procesos de regulación genética y de la base genética de las enfermedades.

AGRADECIMIENTOS

Este trabajo se ha llevado a cabo como parte del proyecto P08-TIC-4299 J. A., Sevilla, TIN2009 13489 DGICT, Madrid, GREYB-PYR2010-02 (CC) y GREYB-PYR05 (MC). Los datos utilizados para la experimentación han sido proporcionados por el 1000 Genomes Project Consortium y Wellcome Trust Sanger Institute en ftp://ftp-trace.ncbi.nih.gov/1000genomes/ y ftp://ftp.sanger.ac.uk/pub/1000genomes/tk2/main_project_reference/.

REFERENCES

[1] L. W. Hillier et al., "Whole-genome sequencing and variant discovery in C. elegans", Nature Methods, vol. 5, pp. 183–188, 2008.
[2] M. Sana et al., "GAMES identifies and annotates mutations in next-generation sequencing projects", Bioinformatics, vol. 27, pp. 9-13, October 2010.
[3] J. Setubal y J. Meidanis, "introduction to computational molecular biology", PWS Publishing Company, Boston, 1997.
[4] F. Garcia-Alcaide et al., "An intuitionist approach to scoring DNA sequences against transcription factor binding site motifs", BMC Bioinformatics, vol. 11:551, 2010.
[5] J. E. Stajich et al., "The Bioperl Toolkit: Perl modules for the live science", Genome Research, vol. 12, pp. 1611-1618, 2002.
[6] SAMTOOLS webpage, http://samtools.sourceforge.net/
[7] H. Li et al., "Mapping short DNA sequencing reads and calling variants using mapping quality scores", Genome Research, vol. 18, pp. 1851-1858, August 2008.
[8] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform", Bioinformatics, vol, 25, pps. 1754-1760. May 2009.
[9] S. T. Sherry et al., "dbSNP: the NCBI database of genetic variation", Nucleic Acids Research, vol. 29, pp. 308-311. 2001.
[10] D. A. Benson et al., "GenBank". Nucleic Acids Research, vol. 36 pp. D25-D30, December 2007.
[11] V. Matys et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles", Nucleic Acids Research, vol. 31, pp. 374-378. 2003.
[12] A. Sandelin et al., "JAPAR: an open-access database for eukaryotic transcription binding sites profiles", Nucleic Acids Research, vol. 32, pp. D91-D94. 2004.
[13] L. Zadeh. "Fuzzy Sets", Information and Control, vol. 8, pp. 338-353. 1965.
[14] K. Atanassov. "Intuitionistic Fuzzy Sets: theory and applications", Physica-Verlag, Heidelberg. New York 1999.
[15] A. Tomovic and E. Oakeley. "Position dependencies in transcription factor binding sites", BMC Bioinformatics, vol. 23, pp. 933. 2007.
[16] F. Zare-Mirakabad et al. "New scoring schema for finding motifs in DNA sequences", BMC Bioinformatics, vol. 10, pp. 94. 2009.
[17] N. Siva. "1000 Genomes Project", Nature Biotechnology, vol. 26, pp. 256. 2008.
[18] F. Garcia et al., "FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral", BMC Bioinformatics, vol. 10:224. 2009.
[19] D. Ge et al., "WGAViewer: Software for Genomic Annotation of Whole Genome Association Studies", Genome Research, vol. 18, pp. 640-643. 2008.
[20] 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.
[21] M. Johnson et al. "NCBI BLAST: a better web interface", Nucleic Acids Research, vol. 36(supp. 2), pp. W5-W9. April 2008.
[22] D.S. Chekmenev et al. "PMatch: transcriptcion binding site search by combining pattern and weight matrices", Nucleic Acids Research, vol. 33(supp. 2), pp. W432-W437. March 2005.
[23] Z. Zhao, Y-X Fu, D. Heweet and E. Boerwinkle. "Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution", Science, Vol. 312, pp. 207-213.2003.

TABLE I
EJEMPLOS DE GENES QUE SON REGULADOS POR UN TFBS QUE TIENE ALGÚN SNP. "ID MOTIVO" MUESTRA LA ID DE TRANSFAC ASIGNADA AL MOTIVO. LA SECUENCIA MUESTRA LA SECUENCIA CONSENSO DEL MOTIVO CON LA LOCALIZACIÓN DEL SNP RESALTADA.

ID motivo	Secuencia	Gen
M00801	CG̲TCAC	CREB
M00624	AGT̲ACAC	DBP
M00803	CCC̲GCC	E2F
M00695	GCGGAGA̲	ETF
M00789	AGAC̲AGG	GATA
M00921	AGAA̲CAGA	GR
M01033	AG̲GGCA	HNF4
M00747	TTCACT̲T	IRF-1
M00644	CAGCC̲GC	LBP1
M00805	TCAG̲AG	LEF1
M00655	ACATCC̲G	PEA3
M00960	GAGAGGA̲CAT	PR
M01131	CTTC̲GTA	SOX10
M00148	TTTGTTC̲	SRY
M00980	TTTA̲TAG	TBP
M00704	CATTCC̲	TEF-1
M00726	CAC̲GTG	USF2
M00924	TGAC̲TCACAGGG	AP-1

TABLE II
ALGUNAS SECUENCIAS ANALIZADAS EN LAS QUE SE HAN DETECTADO SNPS QUE PUEDEN ALTERAR LA AFINIDAD DE UNIÓN DE UN TFBS DE LA SECUENCIA. SEQ. ID IDENTIFICA LA SECUENCIA CON EL NÚMERO DE CROMOSOMA Y LA REGIÓN QUE ABARCA DENTRO DE ÉSTE. TFBS ID INDICA EL ID DEL TFBS EN TRANSFAC. LAS COLUMNAS CUARTA Y QUINTA MUESTRAN EL VALOR DE SIMILITUD DE CSINTUIT PARA LA SECUENCIA CON SNP Y LA NO MUTADA. LA SEXTA COLUMNA MUESTRA LA DIRECCIÓN DE LA CADENA, ("+" , "-").

Seq. ID	TFBS ID	Gen	Similitud no mutado	Similitud mutado	Dir.
15(20134248-20134328)	M01032	HNF4	0.696426	0.932944	-
15(20076473-20076553)	M00924	AP-1	0.807852	0.921382	-
15(20076473-20076553)	M01033	HNF4	0.967851	0.713052	+
15(20076473-20076553)	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	-
15(20454362-20454442)	M00695	ETF	0.976793	0.759638	+