2016

Universidad del País Vasco

Silvia Fernández Portela

[CARACTERIZACIÓN BIOINFORMÁTICA DE UNIDADES DE TRANSCRIPCIÓN NO ANOTADAS RELACIONADAS CON LA ENFERMEDAD CELÍACA]

RESUMEN

La celiaquía es una enfermedad compleja y crónica que se debe a una reacción de intolerancia a la ingesta del gluten. Sus vías moleculares son poco conocidas aunque han sido muy estudiadas debido a que es multigénica.

Se presenta la secuenciación de ARN como método de identificación de nuevos transcritos por todo el genoma para la compresión de la enfermedad celíaca. Pudiendo extrapolarse a cualquier otra investigación.

Para ello predeciremos transcritos del RNAseq, los identificamos y caracterizamos mediante un análisis bioinformático y validamos los resultados mediante una reacción de la cadena de polimerasa a tiempo real (qPCR).

Se demuestra la eficacia del RNAseq para la identificación de nuevos transcritos no anotados anteriormente para la investigación en enfermedades complejas como el caso de la celiaquía.

De manera que se presenta un nuevo modo de abordar la investigación genética.

Contenido

INTRODUCCIÓN

Enfermedad celíaca

La enfermedad celiaca (EC) es una intolerancia crónica, inmune mediada por el gluten. Es una enfermedad multigénica compleja con componentes genéticos y no genéticos. El gluten es un conjunto de proteínas de pequeño tamaño, contenidas exclusivamente en la harina de los cereales de secano, fundamentalmente el trigo, pero también la cebada, el centeno y la avena, o cualquiera de sus variedades e híbridos

Las vías moleculares que subyacen a la patogénesis causante de la enfermedad celíaca son poco conocidos. Desde hace tiempo que sabe que la celiaquía se desarrolla en personas genéticamente predispuestas. Pero a pesar de que el 40% de la población es portador del factor de riesgo más determinante (los polimorfismos HLA-DQ2 y DQ8 [1]), solo el 1% desarrolla la enfermedad. Es una enfermedad muy compleja en la que numerosos polimorfismos influyen, cada uno con una contribución muy pequeña.

La EC, como la diabetes tipo 1, la artritis reumatoide y esclerosis múltiple, tiene una naturaleza crónica cuando determinadas formas de alelos HLA están sobrerrepresentados entre los pacientes [2]. Comúnmente todos estos trastornos son multifactoriales; los genes HLA y otros genes, junto con los factores ambientales, están implicados en el desarrollo de la enfermedad. La expresión de EC es estrictamente dependiente de la exposición alimentaria al gluten y proteínas de los cereales similares [3]. Los pacientes entran en remisión completa cuando se ponen en una dieta libre de gluten, y vuelven a caer cuando el gluten se introduce de nuevo en la dieta. La EC es a este respecto exclusiva entre las enfermedades asociadas a los genes HLA inflamatorias crónicas en que un factor ambiental crítico ha sido identificado [4].

La EC se presenta comúnmente en la primera infancia con síntomas clásicos que incluyen diarrea crónica, distensión abdominal, y retraso en el desarrollo [5].

El fondo genético desempeña un papel clave en la predisposición a la enfermedad. El haplotipo HLA-DQ2 (DQA1 * 0501 / DQB1 * 0201) se expresa en la mayoría de los pacientes con enfermedad celíaca (90%), mientras que se expresa en un tercio de la población general. En otro 5% de los pacientes con enfermedad celíaca, el haplotipo HLA-DQ8 (DQA1 * 0301 / DQB1 * 0302) se expresa, mientras que casi todo el 5% restante de los pacientes tienen al menos uno de los dos genes que codifican DQ2 (DQB1 * 0201 o DQA1 * 0501). Los haplotipos DQ2 y DQ8 expresados en la superficie de las células presentadoras de antígeno puede unirse activado péptidos del gluten, lo que provocó una respuesta inmune anormal. Los haplotipos DQ2 y DQ8 son necesarios pero no suficientes para el desarrollo de la enfermedad celíaca [6]. Hasta el

momento, se han identificado al menos 39 genes no HLA que confieren una predisposición a la enfermedad, la mayoría de los cuales están involucrados en las respuestas inflamatorias e inmunes [7]. La patogénesis de la enfermedad celíaca consiste en un disparador externo (gluten), cambios en la permeabilidad intestinal, modificación enzimática del gluten, el reconocimiento del HLA, y la respuesta inmune innata y adaptativa a los péptidos del gluten relacionados con antígenos propios (por ejemplo, transglutaminasa), llevando eventualmente a enteropatía celíaca [9, 8]

Uno de estos factores de riesgo añadidos se puede encontrar en el conocido como ADN basura, es decir, el 95% del ADN. Se desea contribuir a desvelar su papel en el control del funcionamiento general del genoma, es decir, regula procesos importantes en nuestro organismo, como la respuesta inmunitaria, y en él se podrían por tanto hallar las causas de enfermedades autoinmunes como la celiaquía.

Proyecto ENCODE

El proyecto de la Enciclopedia del ADN, llamado ENCODE por ENCYCLOPEDIA OF DNA ELEMENTS, ha estudiado con sumo detalle el ADN humano, tanto la parte codificante (genoma), como la no codificante (mal llamada hace una década como "ADN basura") que hasta hace poco creíamos que no tenía función ninguna o poca. El hallazgo más notable de este estudio es que el 80% de todo el ADN contiene elementos vinculados a funciones bioquímicas (es decir, tiene "actividad bioquímica específica"), desterrando la idea de que gran parte del ADN es simplemente "basura" evolutiva. [10]

El ADN humano tiene unos 3.000 millones de bases ("letras" A, G, T, o C), pero solo el 1% contiene unos 21.000 genes que codifican unas 90.000 proteínas. En el ADN entre los genes, el proyecto ENCODE ha descubierto unas 70.000 regiones "promotoras" que se ligan a proteínas para controlar la expresión de los genes. También ha descubierto unas 400.000 regiones "potenciadoras" que regulan (potencian o reducen) la expresión de genes, incluso de genes muy distantes entre sí. Además, ha descubierto 2,9 millones de regiones a las que se ligan proteínas (por ejemplo, factores de transcripción) en los 125 tipos de células estudiados, de las que unos dos tercios se han descubierto en un solo tipo celular y no aparecen en ningún otro tipo. De hecho, solo unas 3,700 de estas regiones son compartidas por todas las células.

ENCODE revela que en el ADN oscuro, conjeturalmente basura, lo que existen son las instrucciones (letra pequeña) que permite a los genes funcionar de manera correcta en unas células y no en otras. Y ha unido la información que teníamos de los genes con este ADN desconocido. Aun así desconocemos qué necesita (cuales son cada una de las instrucciones) cada uno de nuestros genes para funcionar de forma correcta

Lo que nos muestra el proyecto ENCODE es que el ADN y la regulación bioquímica de la célula es mucho más compleja de lo que nunca pudimos imaginar. Una de las cosas que más me interesan sobre el ADN, el estudio detallado de los aspectos dinámicos de la regulación génica, está más allá de los objetivos del proyecto ENCODE y requiere el desarrollo de nuevas tecnológicas. Serán necesarias muchas décadas para que podamos empezar a entender el funcionamiento complejo de cada célula humana a partir de su ADN y su epigenómica.

Identificación de nuevos transcritos

La identificación de transcritos y la cuantificación de la expresión de genes han sido algunas de las actividades básicas dentro la biología molecular desde el descubrimiento de la función del ARN como la clave intermedia entre el genoma y el proteoma. El poder de la secuenciación del ARN reside en el hecho de que los aspectos individuales de descubrimiento y cuantificación se pueden combinar en un ensayo simple de secuenciación de alto rendimiento sencillo llamado secuenciación de ARN (RNA-seq). La secuenciación de ARN, tiene una amplia variedad de aplicaciones.

Cuando un genoma de referencia está disponible, el análisis de RNA-seq normalmente implica el mapeo de las lecturas en el genoma de referencia o transcriptoma para inferir que las transcripciones que se expresan. El mapeo únicamente para el transcriptoma de referencia de una especie conocida impide el descubrimiento de nuevo transcritos no anotados y se centra en un análisis puramente basado en la cuantificación.

Método

Este proyecto se centra en esas transcripciones que son detectadas por el RNAseq y no han sido anotadas todavía.

La identificación de nuevos transcritos utilizando el sistema corto lecturas que ofrece la tecnología Illumina es una de las tareas más difíciles de la ARNseq [11]. Existen varios métodos, como por ejemplo, Cufflinks [12], que incorporan anotaciones existentes mediante su inclusión en la lista de posibles isoformas y nuevos transcritos.

HIPÓTESIS

Ya que el análisis RNAseq sirve para detectar la expresión diferencial de transcritos por todo el genoma, y detecta regiones genómicas conocidas que se expresan, también podría ser posible identificar regiones genómicas no anotadas expresadas significativamente.

Estas regiones darían nueva información útil para comprender cuándo y por qué se expresa la enfermedad celíaca.

OBJETIVO

Detectar, identificar, caracterizar y evaluar nuevos transcritos que participan en los pacientes con celiaquía.

Detectamos los conocidos y los nuevos transcritos con la secuenciación de ARN. Posteriormente elegimos los nuevos transcritos solamente, los caracterizamos bioinformáticamente siguiendo el esquema explicado en método y los validaremos con una PCR a tiempo real para saber si los datos proporcionados por el RNAseq en cuanto a nuevos transcritos son significativos.

MÉTODOS

Pacientes y biopsias

La Enfermedad Celíaca (EC) fue diagnosticada de acuerdo a los criterios en vigor en el momento de la contratación con la Sociedad Europea de Pediatría de la Hepatología del Aparato digestivo y Nutrición, incluyendo la determinados por anticuerpos de anti-gliadina, anti-endomisio y anti-transglutaminasa, así como una biopsia confirmatoria del intestino delgado. El estudio fue aprobado por las Juntas Institutionales (Hospital Cruces de la Universidad del País Vasco código CEIC-E09/10 y los Comités Éticos y de Ensayos Clínicos de código PI2013072) y se realizaron análisis después de obtener el consentimiento informado de todos los sujetos o sus padres. Se obtuvieron muestras de biopsia de duodeno distal de cada paciente durante la endoscopia diagnóstico de rutina.

El conjunto de la muestra para el análisis de validación consistió en 16 niños de CD al momento del diagnóstico (en una dieta que contiene gluten, con anticuerpos asociados a la EC, atrofia de las vellosidades intestinales y la hiperplasia de las criptas), y los mismos pacientes en remisión después de haber sido tratados con dieta sin gluten durante> 2 años (asintomática, anticuerpo epitelio negativo y normalizado intestinal en ese momento), más 15 muestras de tejidos de individuos no celíacos que no sufren de inflamación en el momento de la endoscopia se utilizaron como controles.

El cambio, para la secuenciación de ARN se utilizan cuatro pacientes celíacos y cuatro controles (Tabla 1).

Tabla 1. Los pacientes y biopsias para el RNAseq y para los análisis de qPCR. *Active CD*: los niños al momento del diagnóstico (en una dieta que contiene gluten, con anticuerpos CD-asociado, atrofia de las vellosidades intestinales y la hiperplasia de las criptas). *Treated CD*: misma pacientes en remisión después de haber sido tratado con dieta sin gluten durante> 2 años (asintomática, anticuerpo epitelio intestinal negativo y normalizado). *Control*: muestras de tejido de los individuos no celíacos no sufren de inflamación en el momento de la endoscopia, que se utiliza como controles.

	Active CD	Treated CD	Control
RNA_Seq	4		4
Validation	16	16	15

Análisis de la expresión diferencial

Cuando un genoma de referencia está disponible, el análisis de RNA-seq normalmente implica el mapeo de las lecturas en el genoma de referencia o el transcriptoma, para deducir los transcritos que se expresan significativamente. El mapeo de únicamente del transcriptoma de referencia de una especie conocida impide el descubrimiento de nuevos transcritos no anotados y centra el análisis en la cuantificación. Una opción elemental es si la identificación y cuantificación de transcritos se lleva a cabo de forma secuencial o simultánea. En este caso será de forma secuencial.

Alineación

Cuando una secuencia de referencia está disponible, son posibles dos opciones: el mapeo para el genoma o mapeo para el transcriptoma anotado (figura 1a, b; Cuadro 3.). Independientemente de si se utiliza un genoma o transcriptoma de referencia, las lecturas se puede asignar de forma única (pueden ser asignadas a una sola posición en la referencia) o podrían ser lecturas de asignación múltiple (*multireads*). Las multireads genómicas se deben principalmente a secuencias repetitivas o dominios compartidos de genes parálogos. Cuando la referencia es el transcriptoma, las lecturas múltiples surgen aún más a menudo debido a una lectura que se habría sido asignada de forma única en el genoma se asignaría igualmente bien a todas las isoformas de genes en el transcriptoma que comparten el exón. En cualquiera de los casos - mapeo del genoma o del transcriptoma – la identificación y cuantificación de transcritos se convierte en un reto importante para los genes expresados alternativamente.

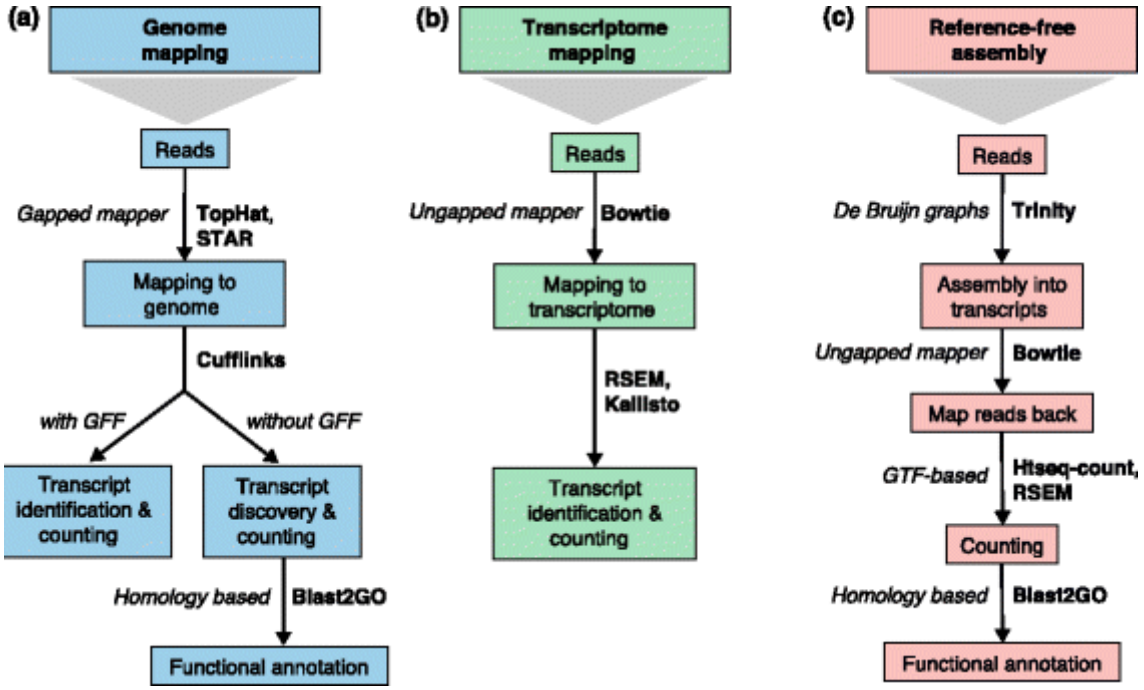


Figura 1. Estrategias de mapeo e identificación de transcritos. *Tres estrategias básicas para el análisis regular RNAseq.* ***A****. Un genoma anotado está disponible y las lecturas se han mapeado en el genoma con un asignador con huecos. El posterior (novel) descubrimiento de transcritos y la cuantificación pueden proceder con o sin un archivo de anotación. Los nuevos transcritos son entonces anotados funcionalmente.* ***b*** *Si no se necesita ningún descubrimiento de nuevos transcritos, las lecturas se pueden asignar al transcriptoma de referencia utilizando un alineador sin huecos. La identificación y la cuantificación de transcritos pueden ocurrir simultáneamente.* ***c*** *Cuando no está disponible un genoma, las lecturas necesitan ser montadas en transcritos. Para la cuantificación, las lecturas se asignan de nuevo al transcriptoma referencia como en (b), seguido de la anotación funcional de los nuevos transcritos como en (a). Los programas representativos que se pueden utilizar en cada etapa de análisis se indican en negrita. Abreviaturas: GFF General Feature Format, GTF gene transfer format, RSEM RNA-Seq by Expectation Maximization*

Descubrimiento de nuevos transcritos

La identificación de nuevos transcritos utilizando el sistema de lecturas cortas que ofrece la tecnología *Illumina* es una de las tareas más difíciles del RNAseq. Las lecturas cortas raramente se extienden a través de varias uniones de empalme y por lo tanto hacen que sea difícil para inferir directamente en transcritos de gran longitud. Además, es difícil identificar los sitios de inicio y fin de transcripción [13], y herramientas como la ARENA [14] que incorpora más información como los extremos 5 'de CAGEo RAMPAGE suelen tener una mayor oportunidad de anotar correctamente las principales isoformas expresadas. En cualquier caso, las lecturas PE ayudan a reconstruir una mayor cobertura de los transcritos expresados menos significativamente, y las repeticiones del experimento son esenciales para resolver los llamados de falsos positivos (es decir, artefactos de mapeo o contaminaciones). Existen varios métodos, tales como Cufflinks [15], iReckon [16], SLIDE [17] y StringTie [18], que incorporan anotaciones existentes mediante su inclusión en la lista de posibles isoformas. Montebello [19] descubrime isoformas y cuantifica los transcritos usando un algoritmo de Monte Carlo basado en la probabilidad de aumentar el rendimiento. Las herramientas de investigación genética, como Augustus [20] pueden incorporar datos de RNA-seq para anotar mejor los transcritos que codifican proteínas, pero nos da un peor rendimiento en los transcritos no codificantes [21]. En general, la precisión en la reconstrucción de transcritos con lecturas cortas es difícil, y muchos métodos suelen mostrar desacuerdo sustancial [21].

Por lo cual, para en análisis RNAseq, primero se ha usado Sickle [22] para quitar las lecturas de baja calidad, se ha usado la opción -n 13 y el resto de opciones con los valores por defecto. Después, se ha seguido el protocolo Tuxedo [23]. Brevemente, las secuencias leídas se mapearon contra el genoma humano de referencia (Hg38) usando Tophat [24] y usando GENCODE 24 [25] como el transcritpoma de referencia. Algunos parámetros se han convertido para adaptarse a sus datos (-g 40 --library-type=fr-firststrand) y para el resto de opciones se utilizaron los valores de los parámetros predeterminados. Se ha usado Cufflinks [26] para el montaje del transcriptoma usando la búsqueda de transcritos *de novo* y GENCODE 24 [25] como referencia, para calcular el valor FPKM para cada transcrito, para examinar las diferencias entre sanos y enfermos, y conseguir valores normalizados.

Reconstrucción de transcritos *de novo*

Cuando un genoma de referencia no está disponible o está incompleto, las lectura de RNA-seq puede ser ensamblada "de novo" (Fig. 1c) en un transcriptoma utilizando paquetes como SOAPdenovo-Trans. Oasis, Trans-Abyss o Trinidad.

En general, se prefiere la secuenciación específica de PE de cadena larga y las lecturas largas porque son llevan más información. Aunque es imposible de montar transcripciones poco expresadas que carecen de suficiente cobertura para un conjunto fiable, demasiadas lecturas también pueden ser problemáticos porque conducen a un aumento del potencial del montaje y de los tiempos de ejecución. Por lo tanto, se recomienda la reducción del número de lecturas para muestras secuenciadas profundamente. Para los análisis comparativos entre las muestras, es recomendable combinar todas las lecturas de múltiples en una sola entrada con el fin de obtener un conjunto consolidado de *contigs* (transcripciones).

Ya sea con una referencia o mediante de novo, la reconstrucción completa de transcriptomas utilizando la tecnología de lectura corta Illumina sigue siendo un problema difícil, y en muchos casos de novo da como resultado un montaje en decenas o cientos de contigs que representan transcripciones fragmentadas.

Caracterización bioinformática

La caracterización bioinformática se ha hecho usando bases de datos para el uso abierto y público: UCSC Genome, Blast, Blastx y Conserved Domain Database (CDD). Se ha seguido el esquema de la figura 3.

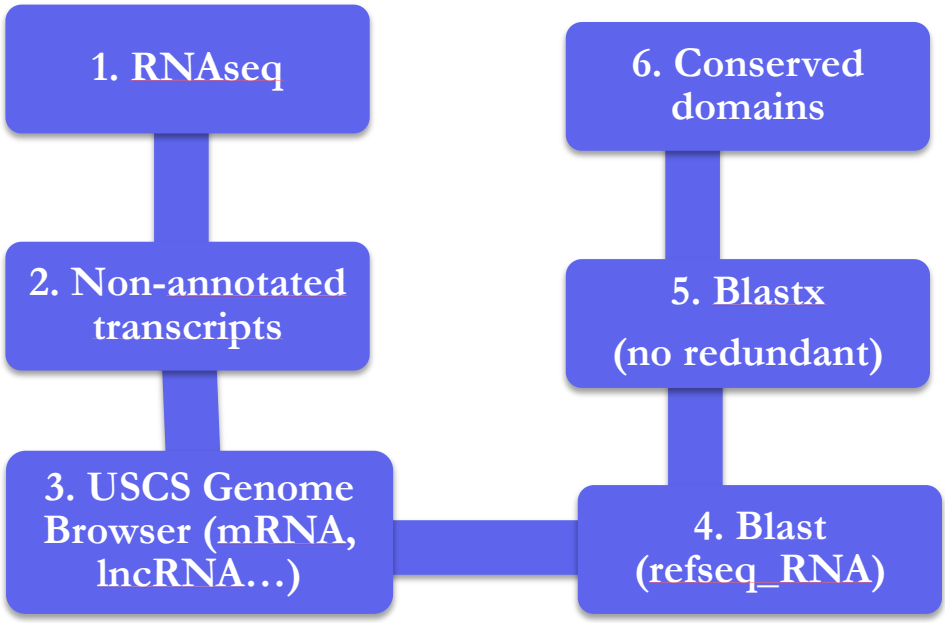


Figura* 2.** ***Workflow de la caracterización bioinformática*. El resultado se define como significante si Query Cover y Identity >90%*

USCS Genome Browser

Primero, se ha obtenido la representación gráfica de la región cromosómica en gracias al sitio de Bioinformática UCSC del genoma humano HG38 (https://genome.ucsc.edu/). Nos fijamos en los ARNm en GenBank, todos los modelos de genes humanos de v22 Gencode, genes codificantes y no codificantes de proteínas en los genes RefSeq.

a. ARNm de GenBank:

Muestra alineaciones entre los ARNm humanos en GenBank y el genoma. Los ARNm humanos GenBank han sido alineados en contra del genoma utilizando el programa Blat. Cuando un solo mRNA se está alineado en varios lugares, se muestra la alineación que tiene la identidad de base más alta. Sólo se mantienen las alineaciones que tienen un nivel de identidad de base dentro de 0,5% de la mejor y la identidad de base al menos 96% con la secuencia genómica.

b. GENCODE v22

Está compuesto por todos los modelos de gen en GENCODE v22. Ello incluye los genes RNA codificantes y no codificantes de proteínas.

c. Genes RefSeq

Los genes RefSeq muestras los genes (codificantes o no) extraídos de la colección NCBI RNA como referencia (RefSeq). Los datos se actualizan semanalmente.

Los ARN RefSeq han sido alineados contra el genoma humano utilizando Blat. Aquellos con una alineación de menos del 15% se descartaron. Cuando un solo RNA está alineado en múltiples lugares, se identifica la alineación que tiene la identidad de base más alta. Sólo se mantuvieron alineaciones que tienen un nivel de identidad de base dentro de 0,1% de la mejor y la identidad de base al menos 96% con la secuencia genómica.

d. Todas las transcripciones GENCODE incluyendo versiones anteriores

El objetivo del proyecto GENCODE (Harrow et al., 2006) es producir un conjunto de anotaciones muy precisas de las características de genes basados en la evidencia en el genoma de referencia humano. Esto incluye la identificación de todos los loci asociados de codificación de proteína con variantes alternativas de empalme, genes no codificantes con las evidencia de transcripción en las bases de datos públicas (NCBI / EMBL / DDBJ) y pseudogenes.

Para muchos estudios de investigación, tales como análisis comparativos o evolutivos, o para el diseño experimental y la interpretación de los resultados, es necesario un conjunto de alta calidad de las estructuras de los genes.

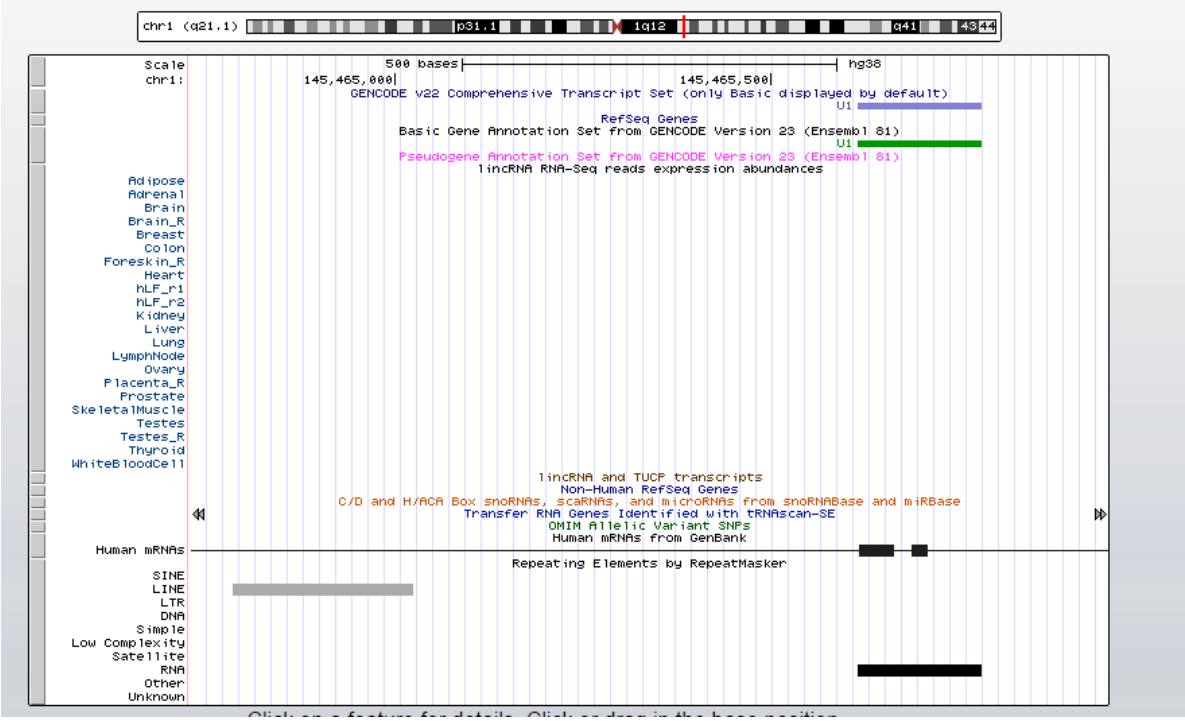


Ilustración 1. *Representación gráfica de chr1:145464729-145465949*

Blastn

Los transcritos que no han sido identificados en el anterior paso pasan al siguiente nivel, el examen mediante Blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Figura 3.

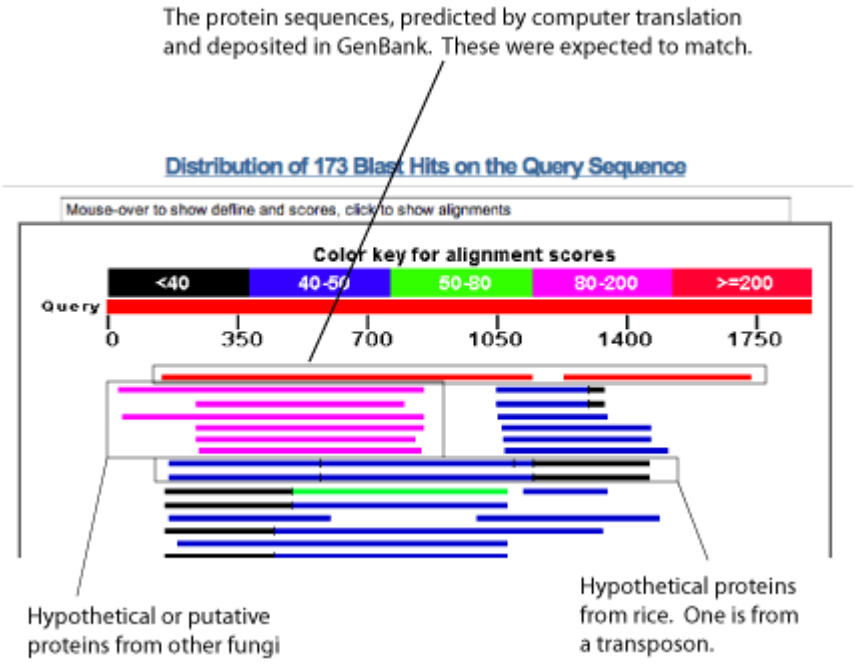


Figura 3. *Ejemplo de Blast. From scienceblogs.com*

Blastn compara la secuencia nucleótida contra la base de datos (RefSeq_RNA). Busca la mayor similitud por todo el genoma usando Megablast (optimización para secuencias altamente similares)

Decimos que el resultado es significante, y que el transcrito está caracterizado si la similitud es mayor del 90% (QV%, I%>90).

Blastx

Los transcritos que no han sido identificados en el anterior paso pasan al siguiente nivel, el examen mediante Blastx (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

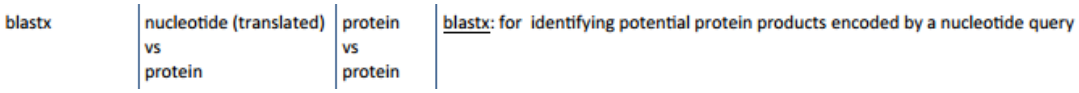


Figura 4. *Ejemplo de Blast. From scienceblogs.com*

Blastx traduce la secuencia y la compara con una base de datos de proteínas no redundantes (Figura 4)

Decimos que el resultado es significante, y que el transcrito está caracterizado si la similitud es mayor del 90% (QV%, I%>90).

Consideraciones al usar BLAST

Se debe recordar que el programa es heurístico y por lo tanto puede que no encuentre la solución óptima. En la actualidad, el abuso y la pobre interpretación de los resultados de BLAST han llevado a múltiples errores de anotación. Una cosa a tener en cuenta al usar BLAST es que cuanta más evidencia externa se pueda obtener para corroborar un alineamiento (fisiológica, filogenética, genética, etc.) es mejor.

El programa de BLAST NO garantiza que las secuencias que alinea sean homólogas y mucho menos que tengan la misma función, simplemente provee posibles candidatos.

La puntuación del BLAST depende del largo de la secuencia, una secuencia muy corta tendrá una puntuación menor que una grande simplemente por la cantidad de caracteres que tiene.

El e-valor depende del tamaño de la base de datos. Para bases de datos muy pequeñas, e-valores altos son más significativos que para bases de datos muy grandes. Para la base de datos no

redundante (NR) de NCBI por lo general e-valores de 0.01 o menos son considerados como significativos, pero esto puede depender de la secuencia que se esté analizando.

Decimos que el resultado es significante, y que el transcrito está caracterizado si la similitud es mayor del 90% (QV%, I%>90).

Dominios conservados

Los dominios conservados (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) informan sobre la función de las proteínas. (Figura 5)

La base de datos de dominios conservados (CDD) es un recurso de anotación de proteínas que consiste en una colección de varios modelos de alineación de secuencias bien anotadas para los dominios ancestrales y proteínas de longitud completa. Estos están disponibles como matrices de puntuación posición específica (PSSMs) para una rapida identificación de los dominios conservados en las secuencias de proteínas a través de RPS-BLAST. El contenido de CDD incluye dominios de NCBI, que utilizan la información de estructura 3D para definir explícitamente los límites del dominio y proporciona una visión de las relaciones entre secuencia / estructura / función de la proteina, así como los modelos de dominio importados de una serie de bases de datos de fuente externa (Pfam, SMART, COG, PRK, TIGRFAM).

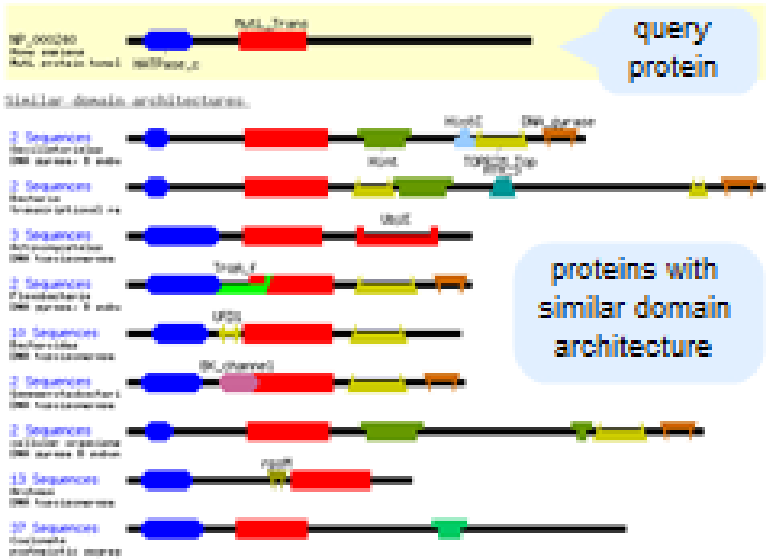


***Figura 5**. Dominios conservados. Extraida de [27]*

Buscamos en la base de datos CDD para prever la familia proteica a la que pertenece y ver otras proteínas con ese dominio a la que se pueden parecer además, en su función. Las proteínas de

las mismas familias por ejemplo todas tienen un dominio conservado, un dominio igual, así puedes saber a qué familia o a que se parece)

QPCR

La reacción en cadena de la polimerasa, cuyas iniciales en inglés son PCR ("polymerase chain reaction"), es una técnica que fue desarrollada por Kary Mullis a mediados de los años 80. Con esta metodología se pueden producir en el laboratorio múltiples copias de un fragmento de ADN específico, incluso en presencia de millones de otras moléculas de ADN.

La PCR se hizo siguiendo el siguiente esquema de la Figura 4

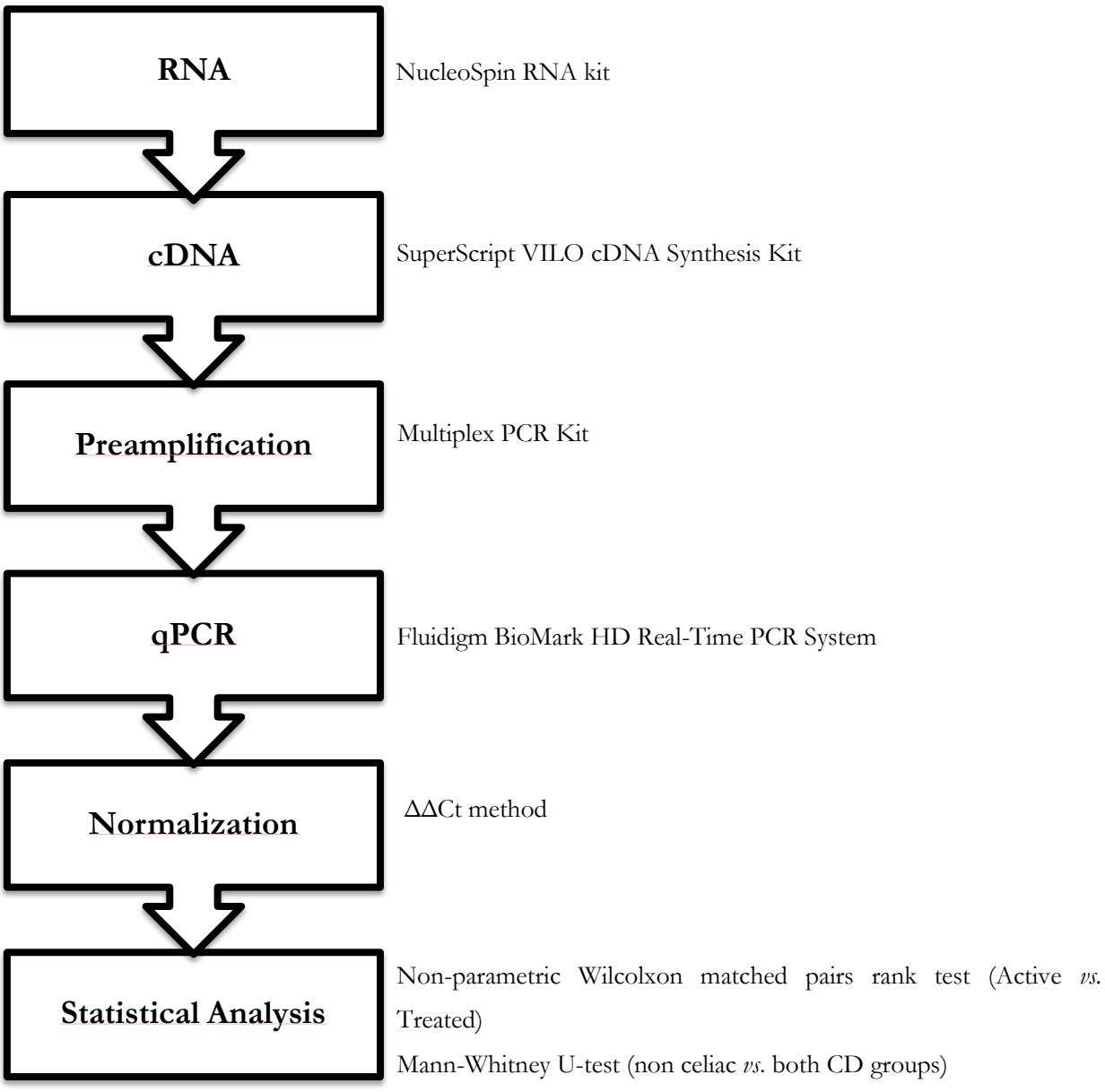


Figure 6. *Esquema de trabajo de la qPCR experimental. Las diferencias se asumen como significativas cuando* $p < 0.05$.

La clave en la PCR cuantitativa es la posibilidad de detectar en tiempo real la amplificación de nuestro genoma de interés. Para llevar a cabo esta detección existen varios métodos pero casi todos basados en la utilización de otro fragmento de ADN (sonda) complementario a una parte intermedia del ADN que queremos amplificar. Esta sonda lleva adherida una molécula fluorescente y otra molécula que inhibe esta fluorescencia ("quencher"), de tal forma que sólo cuando la sonda es desplazada de su sitio por acción de la ADN polimerasa la molécula fluorescente se libera de la acción del "quencher" y emite fluorescencia al ser iluminada con un láser. La cuantificación de la fluorescencia emitida durante cada ciclo de la PCR será proporcional a la cantidad de ADN que se está amplificando. En general para que sea válida esta técnica requiere realizar en paralelo una curva patrón en las mismas condiciones para conocer la cantidad total de ADN que se está amplificando.

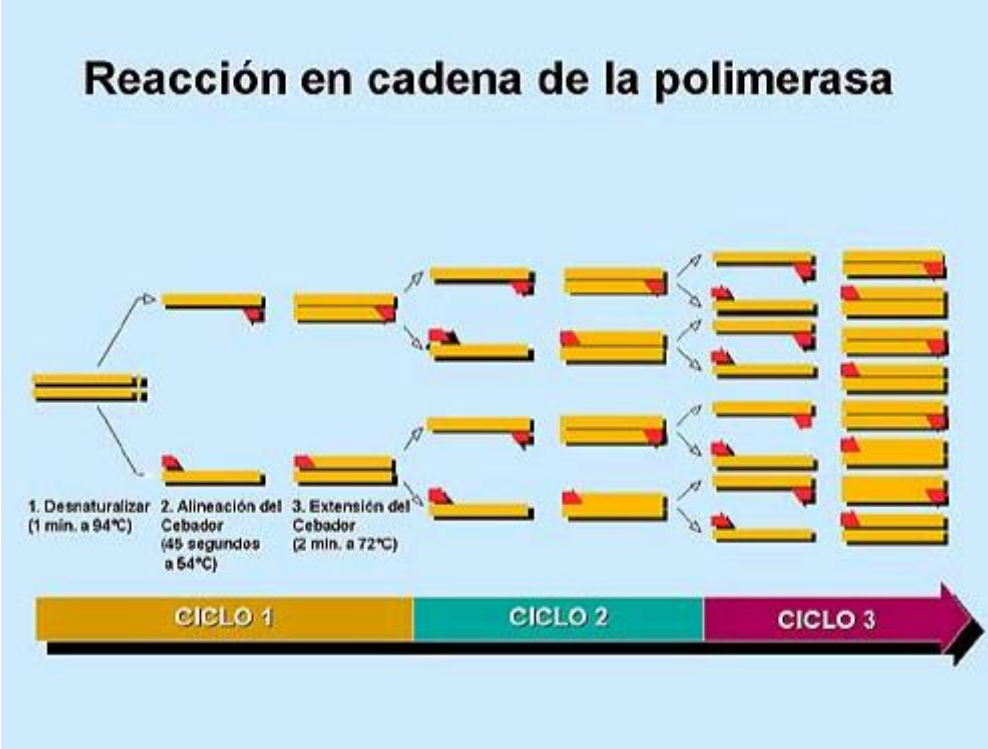


Figura 6. *Ciclos qPCR.*

Extracción de ARN

Todo el ARN fue extraído de pequeñas biopsias usando el kit NucleoSpin microRNA ((Macherey-Nagel, Düren, Germany) siguiendo las instrucciones del fabricante. Después de la purificación de las columnas, el ARN fue eluido en un volumen final de 20µl, cuantificado usando el espectrómetro Nanodrop (ND-1000, Thermo Scientific, Waltham, MA, USA) y diluido en una concentración final de 8ng/µl. Las muestras fueron almacenadas en el congelador a una temperatura de -70°C hasta el momento de su uso.

Expresión génica y síntesis del cADN.

El ARN fue convertido en **cADN** usando el kit SuperScript VILO cDNA Synthesis (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA) empezando desde 500ng del total de RNA. Para este experimento se añadieron 5X VILO Reaction Mix 4µl y 10X SuperScript 2ul a cada muestra con un volumen final de 20ul.

Los tubos con el contenido fueron incubados a una temperatura de 25°C durante 10 minutos, a 42°C durante 60 minutos y 5 minutos a 85°C. El cADN fue diluido en 1:10 para el experimento de qPCR.

PCR a tiempo real: sistema de array dinámico de Fluidigm Biomark

Amplificación específica (STA: Specific Target Amplification)

Las muestras fueron pre amplificadas usando el kit Multiplex PCR Kit (QIAGEN) para incrementar la concentración del target. Los assays TaqMan Gene Expression (20X) se combinaron y se diluyeron usando el buffer low EDTA TE (10 mM Tris, pH 8.0, 0.1mM EDTA) para una concentración final de 0.2X. El volumen final para cada reacción de pre amplificación fue de 5µl (2.5µl 2x QIAGEN Multiplex PCR Master Mix, 1.25µl assay pool (0.2X) y 1.25 cDNA). Las muestras fueron amplificadas en las siguientes condiciones:: 15 minutos a una temperatura de 95ºC y después 14 ciclos a 95ºC durante 15 segundos y a 60ºC durante 4 minutos.

qPCR

El análisis de expresión génica fue hecho usando la PCR a tiempo real de Fluidigm BioMar HD en un formato de arrays dinámicos 96.96 IFC (Integrated Fluidic Circuit), que permiten la preparación de 9216 reacciones (96 muestras por 48 assays en duplicado) al mismo tiempo. El protocolo usado en este experimento ha sido el sistema de trabajo Fluidigm 96.96 Fast Real-Time PCR.

Para la reacción Quanta PerfeCTa™ qPCR Fast Mix, se ha usado el kit low ROX (Quanta BioSciences Inc., Gaithersburg, MD, USA) junto con los reagentes de Fluidigm (2X Assay Loading Reagent y 20X GE Sample Loading Reagent). El array fue puesto en el controlador NanoFlex IFC y puesto en marcha.

El protocolo de PCR usado ha sido: 95ºC durante 1 minuto y 35 ciclos de 95ºC durante 5 segundos y 60ºC durante 20 segundos.

Análisis de datos

Obteniendo resultados de la PCR a tiempo real

El software usado para el análisis ha sido Fluidigm Real-Time PCR versión 3.1.3. Este programa analiza el Ct (Cycle Threshold), el número de ciclos en los cuales la fluorescencia supera el umbral durante la pase de amplificación exponencial. Como cada ensayo se hizo por duplicado, se han obtenido dos valores CT para cada gen. El valor medio para esos valores Ct fue calculado usando el programa Microsoft Excel 2010.

Normalización y cálculo de la expresión relativa de genes: método ΔΔCt

El gen RPLPO se cuantificó y se utilizó como control endógeno para el ARN de entrada (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA). Cada expresión del gen diana se normalizó con el gen RPLPO (valor ΔCt). Después, los valores ΔCt fueron normalizados con la media de los valores de la muestra control (valor ΔΔCt) y finalmente, se aplicó el algoritmo ΔΔCt para obtener las expresiones relativas: RQ: $2^{-\Delta\Delta Ct}$.

Los cálculos se hicieron usando el programa Microsoft Excel 2010.

Análisis estadístico

Los análisis estadísticos se llevaron a cabo para la evaluación de la expresión diferencial entre los grupos (active CD vs. treated CD, active CD vs. controls y treated CD vs. controls). Para examinar los celiacos activos y los celiacos tratados se usa el test no paramétrico de Wilcoxon que tiene en cuenta que ambos grupos son los mismos pacientes. Para comparar los controles con los otros dos grupos de celiacos, se ha usado el test Mann–Whitney U-test. Los análisis fueron llevados a cabo usando el programa Prism 5.0 (GraphPad Software Inc., La Jolla, CA, USA).

Las diferencias son significativas cuando p (valor p) < 0.05.

RESULTADOS Y DISCUSIÓN

Caracterización bioinformática

Del experimento de RNAseq se extraen 276 unidades de transcripción no anotadas y referidas a cuatro grupos diferentes:

- **135** Transcritos no anotados significativamente expresados o subexpresados.
- **33** Transcritos no anotados que se apagan (o su expresión es casi igual a cero) en celíacos y se mantienen encendidos en controles
- **13** Transcritos no anotados que se encienden en celiacos y permanecen apagados en controles (o su expresión es casi igual a cero).
- **95** Transcritos no anotados que se coexpresan significativamente.

La clasificación bioinformática resultó muy exitosa, teniendo solamente un 5% de regiones no anotadas que no se asemejan a nada significativamente.

En las *Figure 3* podemos observar esto en forma gráfica. La mayoría de los transcritos se caracterizan usando USCS genome browser, pero también con todos los demás. En el eje de las ordenadas podemos ver los cuatro grupos en los que hemos dividido nuestros transcritos, en azul aparecen los transcritos que sí hemos caracterizado, en rojo los que dependiendo de nuestro grado de precisión, y en verde los que no tienen nada significativo encontrado.

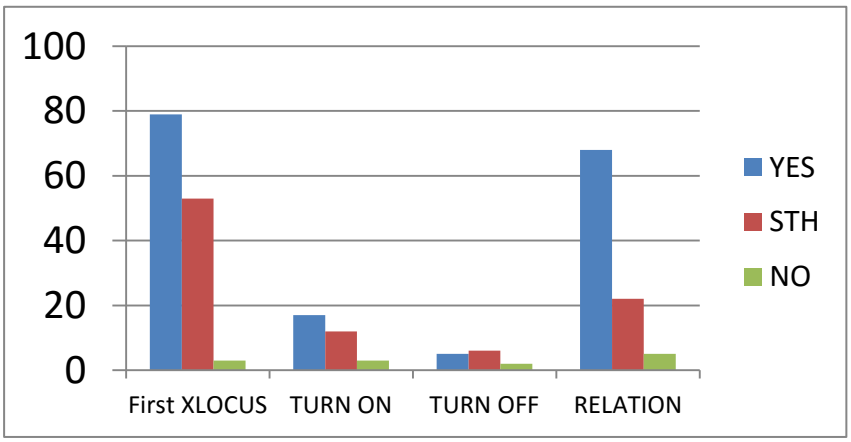


Figura 7. *Caracterización bioinformática.*

En la tabla 2 también podemos ver esta información.

Tabla 2. *Caracterización bioinformática*

	Characterized	Not Clear	No
USCS Genome Browser	79	53	3
Blast	17	12	3
Blastx	5	6	2
CDD	68	22	5

En la figura cuatro podemos ver los transcritos unidos en un solo grupo para subrayar la eficacia de la caracterización bioinformática.

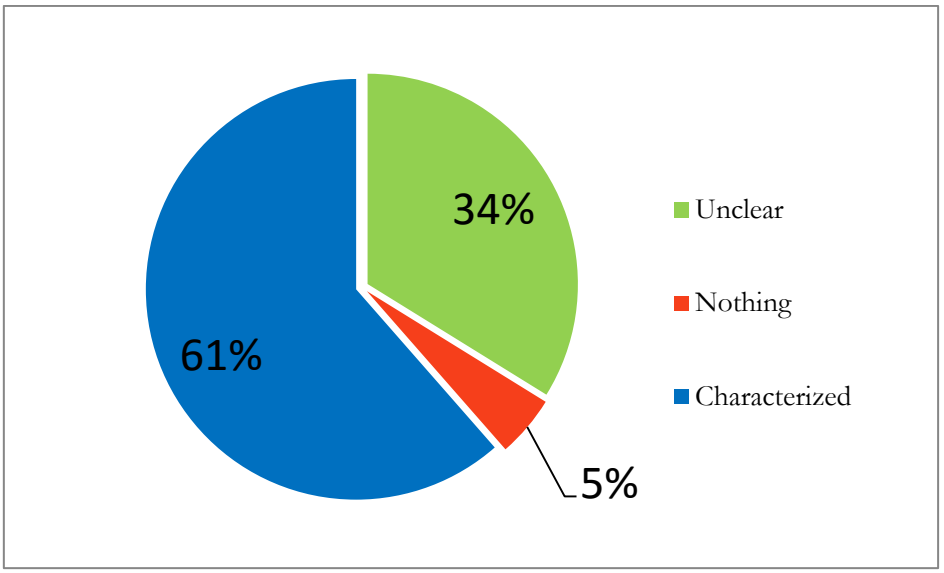


Figura 8. *Caracterización bioinformática*

De cualquier modo, el análisis bioinformático merece un estudio más exhaustivo por que puede pasar que algunos transcritos no anotados estén ya anotados anteriormente. Esto puede pasar por un error en la lectura en el RNA seq.

De la misma manera, se podría realizar una herramienta bioinformática expresa para estos casos, de manera que el análisis de caracterización no tenga que realizarse a mano y pueda tener en cuenta todos los parámetros de manera automática.

Elección de transcritos

Elegimos cuatro transcritos no anotados más interesantes centrándonos en que su número de isometrías era igual a 1 y que tengan solamente un exón.

Debido al interés en los dos grupos que se apagan y encienden dependiendo de si el paciente tiene o no celiaquía, nos hemos quedado con un transcrito no anotado que se enciende en celíacos (**XLOC_022314**) y dos transcritos no anotados que se apagan en celíacos (**XLOC_010878, XLOC_012919**).

Los podemos observar en la tabla X.

***Table 3**. Tránscritos elegidos.*

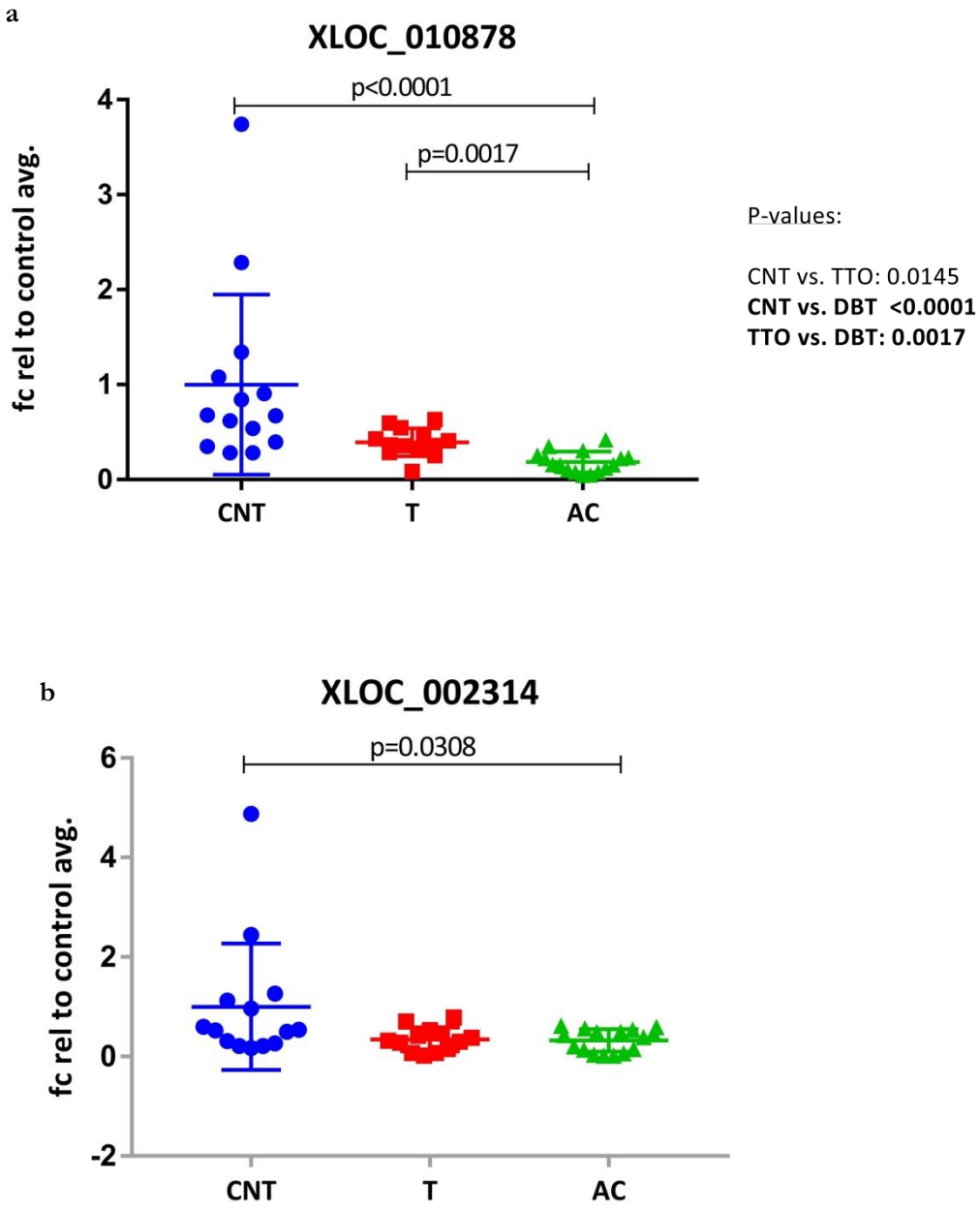
	XLOC_002314	XLOC_012919	XLOC_010878
Localization	chr20:50,278,410-50,279,321	chr16: 5,017,237-5,018,210	chr14:104,082,856-104,083,931
Search	USCS Genome Browser	USCS Genome Browser	Blast
Similar to	LincRNA: LINC01272	mRNA: SEC14L5	Predicted LOC105370691

Dos de los transcritos no anotados elegidos se han caracterizado con USCS Genome Browser, y el tercero con Blast. Esto también lo utilizamos para valorar el método de caracterización informática elegido.

El primer transcrito (**XLOC_022314**) se ha caracterizado usando USCS Genome Browser y es similar a un LincRNA anotado como LINC01271. El segundo transcrito ubicado en el cromosoma 16, según USCS Genome Browser, es similar al mRNA anotado como SEC14L5. Y el tercer transcrito es elegido sobre todo por ser similar según Blast a un LincRNA predicho y anotado como LOC105370691. Demostraríamos su existencia con nuestros resultaos.

Resultados qPCR

Los resultados de la qPCR analizados con Prism 5.0 los podemos observar en la figura 5.



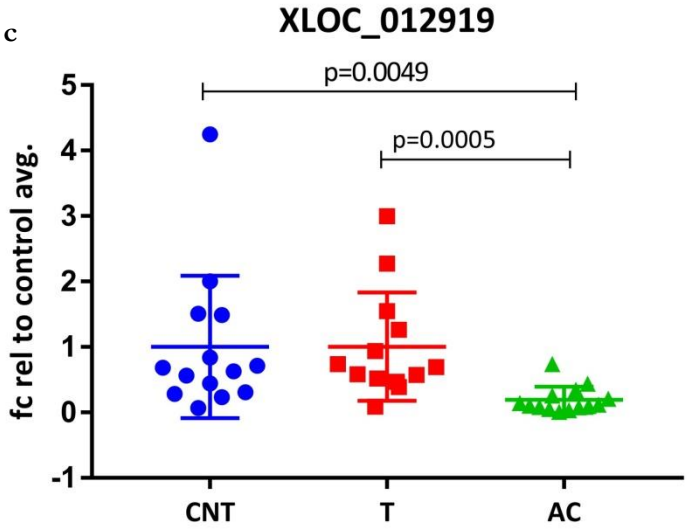


Figura 9*. Resultados de los tres transcritos no anotados en la qPCR. Control (CNT) Tratados (T) y celiacos activos (AC).*

En el primer transcrito (*Figura 9a*), que se apaga en celiacos activos según el análisis RNAseq, según el resultado de la qPCR está altamente expresado en controles comparando con los celiacos activos y tratados. De la misma manera, también son significativas las diferencias entre tratados y activos. Por lo que por un lado parece que tenemos un transcrito que se debe a la ingesta de gluten (por que sube de nuevo cuando dejamos de ingerir gluten) pero también, propio de la enfermedad celíaca, porque aunque quitemos el gluten y la expresión del transcrito sigue siendo significativamente baja comparado con los controles.

En el segundo transcrito *(Figura 9b)*, según el RNA se enciende en celíacos activos. Viendo el resultado de la qPCR sale significativamente lo contrario. Por lo que deberíamos repetir este experimento para sacar una conclusión clara.

El tercer transcrito *(Figura 5c)*, sabíamos que se apagaba en celíacos activos por el análisis RNAseq. Los resultados de la qPCR confirman esto. Estos transcritos se expresa en pacientes sanos, y permanece activado cuando éstos dejan de comer gluten. Y deja de expresarse cuando el celiaco ingiere gluten. Por lo que permanence activado en ausencia de gluten. Este transcrito se debe únicamente a la ingesta de gluten y no propiamente a la enfermedad celiaca *per sé* como el primer transcrito examinado.

En el primer y el tercer caso, la reacción de cadena de polimerasa a tiempo real nos confirma los datos obtenidos del RNAseq.

CONCLUSION

Hemos demostrado la existencia de un transcrito no anotado catalogado de predicted (1) ubicado entre: “chr16:5017237-5018210” y otro ubicado entre “chr14:104082856-104083931”. Además de su significancia para con la celiaquía.

Quedando patente que las regiones no anotadas cambian en los pacientes sanos y enfermos sería muy recomendable crear una línea de estudio que se centre en estas regiones para que nos ayude a entender mejor el funcionamiento de la celiaquía.

De manera que el RNAseq sirve para descubrir nuevos transcritos y el método de caracterización informática es correcto.

REFERENCIAS

[1] Sollid, L. M., & Lie, B. A. (2005). Celiac disease genetics: current concepts and practical applications. Clinical Gastroenterology and Hepatology, 3(9), 843-851.

[2] Thorsby E. 1997. Invited anniversary review: HLA associated diseases. Hum. Immunol. 53:111

[3] Trier JS. 1991. Celiac sprue. N. Engl. J. Med. 325:1709–19

[4] Sollid, L. M. (2000). Molecular basis of celiac disease. Annual review of immunology, 18(1), 53-81.

[5] Schmitz J. 1992. Coelic disease in child- hood. In Coeliac Disease, ed. MN Marsh, pp. 17–48. Oxford: Blackwell

[6] Karell K, Louka AS, Moodie SJ, et al. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum Immunol 2003;64:469-77.

[7] Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 2011;43:1193-201.

[8] Jabri B, Sollid LM. Tissue-mediated control of immunopathology in coeliac disease. Nat Rev Immunol 2009;9:858-70.

[9] Schuppan D, Yunker Y, Barisani D. Celiac disease: from pathogenesis to novel therapies. Gastroenterology 2009;137: 1912-33.

[10] http://francis.naukas.com/2012/09/06/el-proyecto-piloto-encode-dice-adios-al-adn-basura-el-80-del-adn-tiene-funciones-bioquimicas/

[11] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A Survey of Best Practices for RNA-seq Data Analysis.

[12] Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).

[13] Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10:1177–84.

[14] Boley N, Stoiber MH, Booth BW, Wan KH, Hoskins RA, Bickel PJ, et al. Genome-guided transcript assembly by integrative analysis of RNA sequence data. Nat Biotechnol. 2014;32:341–6.

[15] Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27:2325–9.

[16] Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, Shah S, et al. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res. 2013;23:519–29.

[17] Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci U S A. 2011;108:19867–72.

[18] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.

[19] Hiller D, Wong WH. Simultaneous isoform discovery and quantification from RNA-Seq. Stat Biosci. 2013;5:100–18.

[20] Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: AB INITIOprediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.

[21] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013;10:1185–91.

[22] Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle. 2011 (2011).

[23] Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–78 (2012).

[24] Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013)

[25] Harrow, J. et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 22, 1760–1774 (2012)

[26] Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).

[27] Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., ... & Lanczycki, C. J. (2005). CDD: a Conserved Domain Database for protein classification. Nucleic acids research, 33(suppl 1), D192-D196.

AGRADECIMIENTOS

Este trabajo forma parte de los proyectos ISCIII-PI13 / 1201 GVSAN- 2011111034 a JBR, y están aprobados por el Hospital Universitario DE Cruces y los Comités de Ética de Ensayos Clínicos del País Vasco (Códigos CEIC- E09 / 10 y PI2013072). Las biopsias de duodeno distal se obtuvieron mediante endoscopia después de obtener el consentimiento informado de todos los sujetos o sus padres en caso de menores.

Me gustaría agradecer al Dr. José Ramón Bilbao por su apoyo y aliento y el Dr. Koldo García-Etxebarria por su valiosa asistencia técnica.