Identificación y priorización de genes candidatos en enfermedades hereditarias: un estudio basado en redes génicas

José Carbonell Caballero

1 de septiembre de 2012

Resumen

Esto será el abstract...

Índice general

1 Introducción

1.1. Biología molecular y computacional

1.1.1. Los inicios de la biología moculecular

La biología molecular surge como una disciplina propia a raíz del descubrimiento de la doble hélice de ADN por parte de los investigadores James D. Watson y Francis Crick en 1953 [1], descubrimiento por el cual fueron galardonados, junto a Maurice Wilkins, con el premio nobel de medicina en 1962. Este hallazgo, que pasó desapercibido en un primer momento, fue el inicio de una serie de trabajos que permitieron describir como el ADN codifica en su interior las instrucciones necesarias para el funcionamiento de las células. Se descubrió que los aminoácidos se codificaban en grupos de 3 nucleótidos y que una secuencias de estos, formaba una proteína, la cual es la encargada directa de ejecutar las instrucciones contenidas en una determinada porción del ADN. A partir de ese momento, la biología molecular crece como una rama esencial de la biología y gracias a la influencia de otras ramas como la bioquímica, surge con un carácter marcadamente cuantitativo, dentro de una ciencia más acostumbrada a la descripción que a la medición.

En términos generales, la biología molecular se define como la parte de la biología encargada de estudiar los procesos celulares que ocurren a escala molecular. Estos procesos abarcan todos los mecanismos necesarios para que las células puedan funcionar correctamente, llevando a cabo todas sus funciones vitales con normalidad. Se trata de mecanismo generales que ocurren en todos los organismo conocidos, aunque por supuesto, con diferencias específicas en función del tipo celular.

La biología molecular nos permite comprender como funcionan las células y que mecanismos son esenciales para preservar la vida. Es de especial interés la composición de las diferentes formas de material genético (ARN o ADN) y los elementos moleculares que intervienen en su síntesis y regulación. 5 años después del descubrimiento de la doble hélice, Francis Crick formula lo que se conoce como dogma central de la biología molecular,. El dogma básicamente describe como el ADN de un gen contenido en el nucleo celular se trascribe a un ARN mensajero que sale del núcleo, y que posteriormente será traducido a una proteína que llevará a cabo la función del gen. A pesar de que con el tiempo el dogma central se ha desmitificado, ha servido para ilustrar de forma intuitiva como los genes son capaces de llevar a cabo su función en la célula.

1.1.2. Biología computacional

Desde los descubrimientos de Watson y Crick, la biología molecular ha crecido de la mano de un sinfín de tecnologías que le han permitido abordar de forma cuantitativa el estudio de los fenómenos moleculares. Desde el microscopio a los modernos secuenciadores, todos los descubrimientos han estado asentados sobre una base tecnológica muy fuerte, de la que han carecido otras áreas de la biología.

Una de las ramas tecnológicas que más ha evolucionado y aportado a la biología computacional en los últimos 30 años, ha sido la informática. Esta rama ha permitido procesar y almacenar de forma ordenada cantidades enormes de datos biológicos en tiempos de cómputo razonables. A la aplicación de la informática a la biología y su unión a otras disciplinas como la estadística o las matemáticas, se le denominó biología computacional.

La biología computacional, también denominada en ocasiones como bioinformática, es una disciplina en la que, a partir de la unión de modelos, algoritmos y computadores, es posible abordar la resolución de problemas biológicos. Los problemas más característicos afrontados en el pasado mediante la biología computacional han estado relacionados con el alineamiento de secuencias, lo que ha permitido entre otras cosas construir árboles filogenéticos que describen como evolucionan y se relacionan las distintas especies con sus ancestros comunes.

En los últimos años, el volumen de datos aportados por la biología ha crecido de forma considerable, lo que ha generado una dependencia informática, obligando a la incorporación de procedimientos y herramientas informáticas que permiten automatizar ciertos procesos casi cotidianos.

Entre los logros más importantes de la biología molecular y computacional destaca la secuenciación del primer genoma humano, tarea que fue resuelta en más de 13 años y en la que se invirtieron más de 280.000 millones de dólares por parte de algunos gobiernos y entidades privadas. La identificación de la secuencia del genoma humano y la localización de los genes ha tenido una gran relevancia en los ámbitos de la biomedicina y la genética clínica ya que aporta un visión genómica global que facilitó la interpretación de las bases moleculares de algunas enfermedades hereditarias.

1.2. Biología computacional en el estudio de enfermedades hereditarias

Las enfermedades de origen genético, son síndromes donde un error en la maquinaria de síntesis de proteínas provoca un comportamiento anómalo de las células y como consecuencia, la aparición de una enfermedad. Típicamente, su origen viene determinado por una mutación o variación estructural nociva en uno o varios genes, provocando un daño en cadena.

Cuando la enfermedad está causada por un único gen, decimos que es monogénica o mendeliana (en honor a Gregor Mendel). Ejemplos de enfermedades monogénicas son la fibrosis quística (causada por el gen *CFTR*), la enfermedad Huntington (causada por el gen HTT), o la Hemofilia de tipo A (causada por el gen F8). Por el contrario, cuando la enfermedad está causada por la combinación

de varios genes mutados, entonces decimos que es multigénica. Las enfermedades multigénicas son mucha más frecuentes que las monogénicas, y engloban la mayoría de las enfermedades crónicas como la hipertensión, la obesidad, o síndromes complejos como el alzheimer o la esquizofrénia.

Cuando la enfermedad es de carácter familiar, entonces decimos que es hereditaria. En este caso, la mutación o mutaciones nocivas se segregan de padres a hijos. Si un individuo acaba desarrollando una enfermedad de la que es portador vendrá determinado por lo que conocemos como modelo de herencia de la enfermedad.

Los humanos disponemos un genóma redundante compuesto por 23 pares de cromosomas, lo que significa a efectos prácticos (y a excepción de los cromosomas sexuales) que disponemos de dos copias de cada gen, donde una copia vendrá proporcionada por vía paterna y la otra por vía materna durante la meiosis. Esta duplicidad tiene efectos beneficiosos sobre el individuo y la especie, ya que genera (por combinación) biodiversidad y nos hace más robustos a enfermedades. Sin embargo, puede tener efectos funcionales ya que, aunque ambas copias deberían a ser prácticamente iguales, al provenir de parentales distintos cada copia puede disponer de mutaciones o alteraciones estructurales propias. Cuando una mutación está sólo en una copia, entonces decimos que se presenta en heterocigosis, sin embargo cuando la mutación es compartida por ambas, es decir, ambos parentales disponían de ella, entonces decimos que está en homocigosis. Cuando una enfermedad hereditaria muestra una herencia de tipo recesivo, significa que será necesario tener una mutación nociva en ambas copias para acabar desarrollando la enfermedad, lo que generalmente estará asociado a una pérdida de función del gen, ya que ninguna de sus proteínas seran funcionales. Sin embargo, cuando la herencia de la enfermedad es de tipo dominante, entonces significa que una sola copia mutada será suficiente para producir la enfermedad. En este caso, será irrelevante disponer de una copia intacta del gen, ya que generalmente la proteína mutada dispondrá de una nueva función que resultará nociva para la célula.

Comprender como funcionan los distintos modelos de enfermedad resulta absolutamente fundamental, ya que nos permite entender las bases moleculares de las enfermedades, y por consiguiente, avanzar en su prevención y cura. Además, nos permite ser mucho más eficientes a la hora de buscar que genes pueden estar implicados en una enfermedad con un origen genético total o parcialmente desconocido. El problema se complica cuando las variaciones genómicas no son responsables totalmente de la enfermedad. Entonces decimos que las condiciones ambientales (cómo los hábitos alimenticiones o el clima) explican un porcentaje significativo del riesgo a padecer la enfermedad. Existen enfermedades como el cáncer de pulmón, donde el ambiente (ser fumador) puede resultar totalmente decisivo en la proliferación de la enfermedad, y otras como las enfermedades monogénicas, donde presentar ciertas variaciones estructurales asegura una penetrancia casi total.

Biología computacional en la búsqueda de enfermedades***

Es conocido que no todas las mutaciones en genes van a producir los mismos efectos. Hay mutaciones que ni siquiera provocan un cambio de aminoácido en la secuencia proteíca, y otras que directamente impiden la síntesis de la proteina.

Mutaciones sinónimas/no sinónimas, No todas las mutaciones causan los mismos síntomas o la misma intensidad de la enfermedad ***

1000 genomas: mutaciones en población sana ***

Panorama mundial /penetrancia/ ***

Dentro del estudio de enfermedades hereditarias, es de especial interés el estudio de enfermedades raras. Las enfermedades raras, o también conocidas como huérfanas, son aquellas que tienen una baja incidencia en la población (inferior a 5 de cada 10.000 individuos), y por su condición de baja prevalencia, son sometidas a menores inversiones por parte de capital público o privado. Existen más de 7000 enfermedades raras descritas por la OMS y se estima que afectan al 7 % de la población mundial (sólo en España afectan a más de 3 millones de personas). Este tipo de síndromes se benefician claramente de las estrategías de secuenciación de genoma completo, ya que de otra forma, serían necesarios costosos estudios previos que focalizaran el origen de la enfermedad sobre un grupo reducido de genes candidatos.

1.3. Secuenciación de ADN

1.3.1. La secuenciación del genoma

La secuenciación del ADN nos permite conocer el orden específico de los nucleótidos en el genoma de un individuo, lo que posteriormente proporciona la identificación y localización de los genes. Comparando la secuencia obtenida con una secuencia de referencia es posible determinar aquellas diferencias o variaciones estructurales que presenta un individuo y que podrían ser susceptibles de haber causado o causar una enfermedad en el futuro. La construcción de un genoma de referencia con el que comparar ya supone de por sí un reto tecnológico importante que ha sido llevado a cabo en la última década. Además, el genoma de referencia debería constituir una plantilla sana con la que comparar, y definir la normalidad no es algo sencillo. Actualmente existe un consorcio internacional [2] encargado de construir y almacenar la versión oficial del genoma humano, actualizándolo de forma periódica.

A lo largo de los últimos 50 años han surgido multitud de método de secuenciación que han permitido conocer con fiabilidad la secuencia de nucleótidos de la que se compone un fragmento de material genético. Todos los métodos han tenido tasas de error cuantificables y parámetros como su complejidad, facilidad de ejecución, velocidad de secuenciación o coste económico han sido claves para su evolución y adopción por parte de los investigadores. Cabe tener en cuenta que todos los métodos de secuenciación han venido acompañados de herramientas y métodos estadísticos de análisis propios para su análisis, y que cada nueva generación de secuenciadores ha requerido un reaprendizaje de las nuevas herramientas de trabajo por parte de los investigadores. Este hecho ha tenido especial incidencia en los últimos años ya que cada tecnología ha dado lugar a una batería de herramientas informáticas de análisis y de algoritmos encargados de resolver problemas específicos de cada tecnología.

A continuación, se enumeran tres de las metodologías de secuenciación más empleados por la biología molecular.

1.3.2. Secuenciación por Sanger

El método de secuenciación más empleado en las últimas décadas ha sido el denominado método Sanger, en honor a su creador Frederick Sanger. Sanger, el cual es una de las 4 personas que ha recibido 2 veces un premio nobel en su vida, fue capaz de demostrar que las proteínas tienen estructura específica y que esta es fundamental para su función. Para ello, consiguió en 1955 determinar la secuencia de aminoácidos de la insulina y desarrolló un método con el que obtuvo un perfil específico de su estructura. Este trabajo le permitió obtener su primer nobel de química en 1958. Años más tarde (en 1975) desarrolla formalmente su método de secuenciación [3], el cual, a partir de dideoxinucleótidos, un gel de agarosa y la aplicación de electroforesis fue capaz de obtener un patrón de bandas a partir del cual es posible deducir la secuencia subyacente de nucleótidos. Con este método Sanger secuenció al bacteriófago A4, convirtiéndose en el primer organismo cuyo genoma fue secuenciado de forma completa. Los trabajos de Sanger fueron fundamentales en la consecución de proyecto genoma humano y en otros ambiciosos proyectos de secuenciación posteriores, gracias a lo cual obtuvo su segundo nobel de química en 1980.

1.3.3. Chips de ADN

A pesar de los importantes avances obtenidos mediante la secuenciación por Sanger, han sido necesario evolucionar a formas más automáticas de secuenciación que dejaran atrás algunas limitaciones técnicas como la secuenciación de largas cadenas de ADN. Con una perspectiva más de biología de sistemas, se ha evolucionado hacia métodos de secuenciación que son capaces de obtener información simultanea de miles de genes. Una de las tecnologías surgidas en la última década han sido los chips de ADN, los cuales, gracias su bajo coste y altas prestaciones, se han extendido a la gran mayoría de estudios de secuenciación importantes. Se trata de matrices de sondas (o pocillos), donde cada elemento es capaz de medir información concreta acerca un gen. Los chips más populares han sido los de expresión génica. Estos, compuestos por pocillos que contiene la secuencia complementaria de cada gen, son capaces de medir la abundancia de ARN mensajeros y por tanto de la expresión de cada gen. Otros chips ampliamente extendidos, han sido los chips de genotipado. En este caso, el chip se encarga de testar la presencia de una cantidad enorme de polimorfismo en la secuencia de los genes. Si bien, no ha sido una tecnología a partir de la cual obtener la secuencia precisa de nucleótidos de cada gen, si han servido de forma eficiente y barata en la determinación de las características básicas de los genes.

1.3.4. Ultrasecuenciadores o técnicas de secuenciación masiva

Los avances de la última década han permitido el desarrollo de técnicas de secuenciación másiva a un coste muy bajo. Se trata de tecnologías que permiten procesar y secuenciar simultáneamente millones de fragmentos de ADN. Son las llamadas tecnologías de ultrasecuenciación o secuenciación másiva, y son capaces de secuenciar un genoma completo en aproximadamente una semana, con un coste inferior a 20.000 dolares. Lógicamente, estas técnicas han revolucionado

la genómica computacional, dirigiendo los estudios hacia una perspectiva más genómica que génica.

A pesar de sus numerosas prestaciones, los técnicas de secuenciación másiva también muestran algunas limitaciones importantes. La más destacada reside en la longitud máxima que son capaces de secuenciar. Actualmente, los secuenciadores son capaces de obtener fragmentos de entre 50 a 500 nucleótidos, lo cual, comparado con técnicas clásicas como la secuenciación por Sanger, resulta muy pobre. Por esta razón los secuenciadores precisan un paso de computado adicional denominado mapeo. Este paso básicamente se encarga de volver a posicionar (o localizar) cada fragmento secuenciado sobre genoma de referencia. Algunos parámetros como la longitud del fragmento serán claves para la fiabilidad del mapeo, ya que, teniendo en cuenta que la secuencia del genóma dispone de sólo 4 nucleótidos diferentes, a menor tamaño, más probable será encontrar la misma porción en varios puntos del genoma. Otra limitación importante reside en la tasa de error a la hora de ir obtenido la secuencia de nucleótidos, la cual es superior a otras técnicas como la secuenciación por Sanger. Por esta razón, es habitual que después de la fase de análisis estadístico, se recurra al método Sanger para validar las mutaciones obtenidas mediante los ultrasecuenciadores.

A pesar de todo, los ultrasecuenciadores constituye la tecnología más prometedora y, gracias a su enfoque genómico, resultando muy eficientes en el estudio de aquellas enfermedades donde los genes causantes de la enfermedad son desconocidos, y un enfoque más centrado en genes individuales sería totalmente infructuoso.

1.4. Identificación y priorización de genes candidatos

1.4.1. Indentificación de genes candidatos: estudio caso/control

Los genes constituyen únicamente el 5 % del genoma, el resto, antaño considerado como ADN basura, tiene una función todavía desconocida, que en algunas ocasiones estará relacionada con aspectos más estructurales que funcionales en la molécula de ADN. A su vez, los genes tienen en su interior cierto grado de estructura. Poseen porciones que serán codificadas directamente a proteínas, llamadas exones o regiones codificantes, pero también poseen otras regiones de control relacionadas con su regulación, transcripción y otras regiones denominadas intrones, que flanquean a las regiones codificantes.

Debido a la estructura y funcionamiento del genoma, es mucho más probable encontrar una mutación causante de enfermedad dentro de una zona codificante, que en zonas intergénicas, u otras zonas interiores al gen, donde un cambio de nucleótido tendría consecuencias mucho menores. Este factor produce que en la práctica no se realice una secuenciación orientada a genoma completo, por el contrario, se recurre a kits especiales de secuenciación que cubren únicamente las regiones codificantes de los genes. Esta reducción en el área de búsqueda tiene ciertas limitaciones, ya que en algunas ocasiones, la variación estructural nociva no aparecerá en la zona codificante, sino en zonas de control relacionadas con la regulación del gen. Sin embargo, la porción de ADN estudiado se reduce

a sólo un 1 % del total, lo cual aporta grandes beneficios para toda la fase de procesamiento, almacenamiento y análisis que ocurre después de la secuenciación, lo cual permite un aprovechamiento máximo de los recursos económicos disponibles dentro de un proyecto de investigación.

El protocolo de selección de genes candidatos se realiza a partir de un estudio caso/control. Como es habitual, el grupo de casos está formado por una serie de individuos que comparten el mismo síndrome, en este caso, de origen genético. El grupo de casos se acompaña de un grupo de individuos control que nos permite filtrar todas aquellas mutaciones potencialmente nocivas que por estar presentes en población sana no deberían estar relacionadas con el síndrome. Una vez realizada la secuenciación y el resto de pasos computacionales necesarios para la obtención de todas las mutaciones propias de cada individuo del estudio, se realiza una selección de mutaciones candidatas que, en general, presenten una prevalencia significativamente mayor en los casos que en los controles. Si consideramos que los genes responsables de la enfermedad no tiene necesariamente que estar compartidos por el 100 % de casos, o permitimos que un porcentaje minoritario de controles pueda disponer de algunas de las mutaciones nocivas (como suele ocurrir en enfermedades multigénicas), la selección de mutaciones o genes candidatos se realizará habitualmente por medio de un test estadístico de proporciones, como el test exacto de Fisher. Si por el contrario, la selección de mutaciones se realiza de forma más estricta, únicamente se escogerán aquellas mutaciones o genes presentes en el 100 % de casos, sin prevalencia alguna en los controles, lo que hace irrelevante el uso de un test estadístico.

Una vez se realiza el primer filtro en función de casos y controles, se realiza una segunda selección donde quedan eliminadas aquellas mutaciones que producen cambios sinónimos (cambio de nucleótido que no conlleva un cambio de aminoácido) o que por estar en una zonas no codificantes, sea dificil asegurar su malignidad. Finalmente, el grupo de genes que han pasado todos los filtros serán considerados como el conjunto de genes candidatos a la enfermedad.

Debido a limitaciones relacionadas con el tamaño muestral y otro tipo de sesgos surgidos durante todo el proceso, el gen, o genes causantes de la enfermedad serán seleccionando junto a un grupo de genes aleatorios que cumplirán los mismos requisitos y que por tanto no podrán ser filtrados ni separados de estos. De esta forma, será tarea de estrategias posteriores encontrar argumentos biológicos o técnicos que permitan reducir el número de candidatos, o priorizar convenientemente a los que quedan. Esta es la razón principal que obliga a disponer de una fase de priorización (tanto manual como automática) de los genes candidatos, ya que los mecanismos de validación que se ejecutaran al final del estudio, incorporan necesariamente trabajos de laboratorio costosos tanto en tiempo como en gastos económicos. El tamaño del conjunto inicial de genes candidatos será principalmente función del tamaño muestral y de otros parámetros de calidad relacionados con la secuenciación, aunque otros aspectos relacionados con la fase computacional posterior también podrían tener gran influencia.

1.4.2. Reducción y priorización de genes candidatos

Modelos animales***
Selección de genes candidatos***

priorización basada en funciones biológicas***
comparación con genes conocidos***

1.4.3. Validación del conjunto final de genes candidatos

***modelos animales
***grupo de test

1.5. Metodologías de priorización en ausencia de genes conocidos

2 Material y métodos

2.1. Situación inicial: genes candidatos

La metodología propuesta en el presente trabajo trata de describir como priorizar un conjunto de genes candidatos obtenidos a partir de un típico estudio caso/control. El contexto de aplicación serán las enfermedades hereditarias y se va a plantear una metodología compatible tanto con un estudio de tipo familiar, donde se dispone de 1 o más familias distintas con el mismo síndrome, pero no necesariamente con la misma mutación, como con un estudio de tipo más general, donde los individuos disponibles no tienen parentesco alguno.

Cuando afrontamos un estudio de tipo familiar, los individuos seleccionados, tanto casos como controles, se distribuyen a lo largo de las distintas familias. Si se ha realizado un diseño experimental en condiciones, cada nucleo familiar, dispondrá tanto de sujetos caso, como de sujetos control propios de la familia. La ventaja proporcionada por las estructuras familiares es que a priori se espera que todos los individuos enfermos pertenecientes a la misma familia compartan exactamente el mismo origen, es decir, la misma mutación dañina (heredada de padres a hijos), lo cual sería mucho más difícil de afirmar en el caso de individuos independientes. Esta situación permite reducir drásticamente el número de candidatos finales con sólo realizar una simple intersección entre las mutaciones de los individuos enfermos. Por otro lado, los familiares sanos incluidos en el estudio dispondrán de una capacidad de filtrado mucho mayor, ya que al compartir con sus familiares enfermos una porción de genoma mayor de lo esperable con un control externo, será posible eliminar una cantidad mayor de mutaciones no relacionadas con la enfermedad. Además, dicha potencia aumentará con el grado de parentesco, se calcula que un hermano sano (el cual compartirá el 50 % de su genoma con el hermano enfermo) puede tener una capacidad de filtrado equivalente a cientos de controles externos. La figura xxx muestras las diferencias básicas entre el protocolo inicial de selección de candidatos para un estudio familiar frente a uno más general. Al final, en cualquiera de los dos casos, se dispone de una lista de genes candidatos para cada grupo de individuos independientes. Si el estudio es familiar, dispondremos de una lista de candidatos para cada familia, mientras que en el caso general, cada individuo aportará su lista de genes.

2.2. Metodología general

2.2.1. Esquema general del método

La metodología propuesta trata de puntuar a cada gen candidato en función del grado de importancia global sobre el total de familias o individuos estudiados. El proceso se podría describir como una especie de intersección entre los candidatos aportados por cada grupo independiente, donde los genes que aparezcan en un mayor número de grupos serán en general los que tienen una probabilidad mayor de ser responsables de la enfermedad de estudio y por tanto, los mejor priorizados.

Una forma sencilla de abordar el problema cuando esperamos que todos los individuos del estudio tengan mutado el mismo gen, sería mediante la intersección directa de las listas de candidatos proporcionadas por cada familia. En un escenario sencillo, esta sería probablemente la metodología más adecuada, ya que, salvo por un error de secuenciación o procesamiento posterior, el gen de la enfermedad estará presente en todas las listas de candidatos obtenidas. Además, al aumentar el número de familias en el estudio se estará aumentando de forma exponencial la fiabilidad del resultado, ya que, se reduce la probabilidad de encontrar un gen que por azar esté presente también en todas las familias. Sin embargo, cuando disponemos de familias o individuos heterogéneos, con el mismo síndrome, pero con genes mutados diferentes, la metodología de intersección directa no sería valida. El problema se agrava al considerar que los errores del protocolo podrían provocar que el gen de la enfermedad no fuera correctamente secuenciado y que por tanto, no apareciera en la lista de candidatos iniciales. En este caso, es necesario aplicar metodologías más sofisticadas que permitan detectar cuando un gen, aun no habiendo sido seleccionado directamente por una familia, presenta interacciones descritas con los genes que sí han sido seleccionados.

Se calcula que más de un 15 % de enfermedades mendelianas disponen de más de un gen distinto con capacidad para producir la enfermedad, este porcentaje aumentaría mucho en el caso de las enfermedades complejas. Esto se debe a que, aun teniendo una función claramente definida, generalmente los genes actuan de forma colaborativa para llevar a cabo los distintos procesos biológicos que ocurren en la célula. Los genes se coordinan en rutas metabólicas, rutas de señalización u otro tipo de procesos biológicos donde el grado de coordinación entre las distintas moléculas participantes es muy alto. De esta forma, si un síndrome está causado por el mal funcionamiento de un determinado proceso biológico en la célula, es probable que mutando cualquiera de sus genes importantes, se acabe produciendo el mismo síndrome. Esto nos obliga a adaptar las metodologías de trabajo, y poder recoger así la heterogeneidad presente en un set de individuos seleccionados para un estudio.

El proceso de priorización comienza con el reclutamiento de todas las listas de genes candidatos aportadas por los distintos grupos independientes del estudio. Seguidamente, se construye el conjunto global de candidatos a partir de la unión de todos los genes seleccionados por las familias. A continuación, se procede a realizar el cómputo del peso, o estadístico de priorización, para cada uno de los genes incluidos en el conjunto global de candidatos. Por último, una vez obtenidos los pesos de cada uno de los genes candidatos, se realiza un

ordenamiento de la lista de genes en función del estadístico calculado. De esta forma, los genes que estén en lo alto de la lista serán los mejor priorizados, y por tanto, los primeros a validar.

El cómputo del estadístico de priorización para un determinado gen candidato, se realiza básicamente a partir de la suma del peso que tiene el gen en cada una de las familias del estudio. Cuanto mayor sea el peso del gen en las distintas familias, o mayor sea el número de familias en las que el gen esta presente, mejor será su ponderación. Para calcular la relación entre un gen candidato y una familia de estudio, se realiza una evaluación en dos partes: en primer lugar se proporciona un peso inicial en función de si el gen está presente en la lista de candidatos aportada por la familia (equivalente a la que se realizaría con el método de intersección directa) y en un segundo lugar, se añade un segundo término proporcional a la cantidad de interacciones descritas entre el gen candidato y los genes seleccionados por la familia. Esta metodología permite recoger de forma más eficiente el grado de proximidad entre un gen y una familia, lo que a la postre proporciona una priorización más precisa, permitiendo incluso ponderar de forma satisfactoria, genes que no han sido directamente seleccionados por la familia como candidatos.

2.2.2. Estimación del grado de interacción entre genes candidatos y familias

La parte complicada del proceso reside en como calcular y ponderar el grado de interacción entre un gen candidato y el conjunto de genes aportados por una familia. Para ello, en primer lugar, es necesario disponer de una base de datos que recoja el total de interacciones descritas entre los genes. En este caso, se hará uso de diferentes interactomas. Se trata de ficheros que recogen todas las interacciones descritas entre cada par de genes, a partir de los cuales es posible reconstruir fácilmente la red de interacción global de todas las proteínas o genes. Esta red de interacción, donde los nodos son los genes y las aristas sus interacciones, permite, además de recuperar los genes vecinos que interaccionan directamente con un determinado gen, reconstruir totalmente los caminos o secuencias de genes que podríamos emplear para llegar desde un nodo (o gen) de la red a otro. El conjunto total de caminos existente entre cada par de genes de la red nos va a permitir calcular una medida de distancia que va a ser directamente empleada para estimar el grado de interacción entre ellos. Intuitivamente, una distancia pequeña o un número grande de caminos posibles describirá una interacción fuerte entre dos genes, mientras que un número elevado de intermediarios o un número pequeño de caminos posibles describirán una interacción pobre entre los mismos. La medida de distancia va a ser una herramienta fundamental a la hora de evaluar el grado de interacción entre un gen candidato, y cualquiera de los genes aportados por una familia. Más adelante se describirán diferentes formas de computarlo.

En la práctica, se emplearán varios interactomas diferentes, encargados de recoger interacciones de diferente naturaleza. Concretamente, para el presente trabajo han sido empleados los siguientes interactomas:

- Binding: interacción física entre las proteínas generadas por dos genes

- Ptmod: modificaciones post-transcripcionales
- Functional: funciones comunes
- Regulation: relaciones de tipo regulador-regulado
- Text-mining: relaciones entre genes obtenidas a partir de articulos y técnicas de minería de datos

Es importante señalar que el cómputo del grado de interacción descrito con anterioridad se realiza de forma independiente para cada interactoma, por lo que se obtendrán tantos rankings como interactomas hayan sido empleados en el estudio. Así pues, una de las tareas importantes dentro de la metodología propuesta consiste en como unir o ponderar los resultados obtenidos con cada interactoma. Más adelante, se discutirán diferentes formas de llevarlo a cabo.

Conocer a priori cual es el interactoma que mejor recoge las relaciones existentes entre los genes implicados en una enfermedad es una tarea complicada, ya que depende de la naturaleza misma del síndrome. Existen enfermedades (como la xxxx) donde los procesos biológicos sobre los que actúan sus genes implicados se describen de forma casi completa mediante interacciones físicas entre sus proteínas, y sin embargo, existen otras enfermedades de origen regulatorio (como la XXXXX), que serían mejor descritas por el interactoma de regulación.

Por otro lado, es importante indicar que no todos los interactomas describen en realidad interacciones físicas. Sin embargo, representar la relación de cualquier tipo existente entre los genes a partir de una red, proporciona una herramienta de estudio muy potente. Con esta representación es fácil determinar la relación existente entre un determinado gen y sus vecinos indirectos. Si por ejemplo el gen A y el gen B se expresan simultáneamente en algún tejido y el gen B y el gen C se expresan simultáneamente en algún momento del desarrollo, es facil inferir que A y C tienen una estrecha relación y que podrían incluso coexpresar bajo condiciones muy determinadas. De esta forma, podríamos concluir que A y C están a una distancia mucho menor de la que, en promedio, encontrariamos entre dos genes aleatorios.
interactomas incompletos*****

2.2.3. Búsqueda de vecindarios compartidos entre familias

La forma en que se desarrolla todo el proceso de secuenciación y su posterior análisis estadístico provocan que, en general, en un estudio de estas características, el número de falsos positivos sea muy elevado en relación al número positivos esperados. Tanto si se trabaja con familias como con individuos independientes, es conveniente aumentar el tamaño muestral ya que, debido a la metodología de intersección y filtrado empleada, este tiene un impacto directo sobre el tamaño del conjunto de genes candidatos a priorizar, y por tanto, sobre la precisión final del resultado obtenido.

Si se trabaja con una enfermedad mendeliana, al analizar una familia, únicamente deberíamos encontrar un gen responsable, ya que todos los familiares enfermos deberían coincidir en su mutación nociva. El resto de genes seleccionados como candidatos van a ser genes aleatorios obtenidos principalmente a causa de limitaciones derivadas del tamaño muestral máximo que una familia

normal puede ofrecer. Hay que tener en cuenta que, si disponemos de un diseño experimental razonable, el número de genes totales aportados por la familia podrían estar en torno a un rango de 20 a 200 genes. Esto significa que, en el mejor de lo casos, estaremos introduciendo aproximadamente un 95 % de falsos positivos.

La intersección posterior realizada entre familias, debería eliminar de forma considerable la mayor parte de genes aleatorios propios de cada familia, de forma equivalente a cómo se reduce el ruido al promediar varias adquisiciones de la misma señal. En la práctica, tal y como se ha discutido anteriormente, ni todas las familias deben disponer del mismo gen mutado, ni todos los genes responsables del síndrome en cada familia podrán llegar a la fase de selección. Esta es la razón por la que se aplica un proceso de priorización posterior, y también la que justifica parte de la metodología planteada. Si consideramos el caso extremo en el cual disponemos de N familias con el mismo síndrome pero cada una con un gen causante distinto, deberíamos ajustar la metodología para que fuera capaz de detectar entre los candidatos un grupo de genes con un grado de interacción por encima de lo esperado. Si dichas interacciones pudieran estar descritas en un interactoma, significa que el cluster de genes causante del síndrome, debería reflejarse como un vecindario de la red con una densidad de genes candidatos también por encima de lo normal.

El hecho de que en general, la mayoría de genes aportados por las familias sean de carácter aleatorio permite sugerir que el grado de interacción medio debería ser equivalente al obtenido en promedio para un grupo de genes escogidos de forma aleatoria. De esta forma, el vecindario que contiene al grupo de genes causantes, debería ser identificado al evaluar el grado de interacción de cada uno de ellos con respecto a sus vecinos. Por supuesto, las premisas planteadas son sensibles a la talla final del conjunto de genes candidatos, ya que a mayor número de genes aleatorios, más probable es que surgan otros vecindarios altamente conectados por azar.

Por otro lado, es problable que la región de la red donde figuran los genes a identificar no esté descrita de forma completa en el interactoma empleado, lo que provocaría un descenso en la conectividad media del cluster y por tanto una subestimación del estadístico de priorización para los genes causantes de la enfermedad.

2.3. Estimación de parámetros

2.3.1. Medidas de distancia

Tal y como se ha descrito anteriormente, la metodología propuesta trata de mejorar el peso asignado a cada gen candidato incorporando un término matemático que recoge el grado de interacción entre el gen y cada una de las familias incluidas en el estudio. El grado de interacción reposa directamente sobre la medida de distancia calculada entre el gen de estudio y cada uno de los candidatos familiares. En la práctica, estimar la distancia entre dos genes dentro de una red de interacción no es algo trivial, ya que, debido a que los interactomas son redes altamente conectadas, lo normal es disponer de más de un camino distinto para llegar de un punto a otro de la red.

La función de distancia toma como entrada al conjunto formado por los N caminos disponibles entre ambos nodos. El caso es especialmente complicado cuando se dispone de una gran cantidad de caminos, con longitudes muy distintas. Para este trabajo, se han implementado y validado 3 medidas distintas de distancia que recogen varios enfoques distintos a la hora de cuantificar la distancia entre dos nodos. Las medidas empleadas se describen a continuación:

- Shortest path (camino más corto): La distancia entre dos nodos viene definida por la longitud del camino más corto entre ellos. Se trata de una medida que simplifica enormemente el cómputo de distancia, pero que deja de lado algunas características propias de la topología formada por la subred existente entre los dos genes de estudio.
- Integrated distance (distancia integrada): La distancia entre dos nodos se calcula empleando el total de caminos existentes entre ellos. La medida de distancia final vendrá determinada tanto por la longitud de los caminos existentes como por el número de caminos disponibles. Se trata de una medida mucho más compleja y costosa computacionalmente que la técnica del Shortest path, pero consigue recoger de forma más eficiente la relación entre la topología de la red formada por ambos genes.
- Random walk: La medidad de distancia viene determinada por la probabilidad de paso, tomando como origen un gen y como destino el siguiente. Se trata de la adaptación del algoritmo random walk para el trabajo con redes.

Además, las 3 medidas pueden ser adaptadas tanto para el computo de una distancia gen-gen, como para el cómputo de una distancia gen-lista de genes.

2.3.2. Computo del estadístico de priorización

El estadístico de priorización se computa para cada uno de los genes incluidos en el conjunto global de candidatos y para cada uno de los interactomas incluidos en el estudio. Concretamente, el estadístico de priorización para el gen i y el interactoma x se define como:

$$\rho_{i,x} = \sum_{j=1}^{n} \alpha_j * \gamma_{ij} + \delta(i,j)$$

donde n se corresponde con el número de familias, α_j con el peso inicial asociado a la familia j, γ_{ij} con un factor con valores 0 o 1 en función de si el gen i está seleccionado por la familia j y $\delta(i,j)$ con la función que estima el grado de interacción entre el gen i y el vector de genes seleccionado por la familia j.

A su vez, el término de interacción δ se define como:

$$\delta(i,j) = f([d(i,G_{j1}), d(i,G_{j2}), ..., d(i,G_{jk})])$$

donde d es la función de distancia entre el gen i y un gen de la familia j, y f la función que integra todas las medidas de distancia obtenidas y proporciona finalmente el valor de interacción entre el gen candidato i y la familia j. La función de integración se define de forma genérica porque puede ser substituida por varias funciones distintas, como el máximo o la media del conjunto de valores.

2.3.3. Integración de los estadísticos computados por interactoma

La comunicación entre genes puede producirse a diferentes niveles. Cada tipo de interacción se representa mediante el mismo modelo de red, pero en interactomas separados. Se trata de un sistema de información que puede ser enriquecido y actualizado de forma periódica, tanto con nuevas interacciones descritas en la literatura reciente, como a partir de interactomas nuevos, que describe otro tipo de relaciones no empleados hasta ese momento. El hecho de almacenar tantos tipos de interacción como sea posible, permite disponer de un criterio más amplio y preciso a la hora de evaluar el grado de interacción entre dos genes.

Después del proceso de priorización se dispone para cada gen de tantos valores como interactomas hayan sido incluidos en el estudio. Desgraciadamente, en la práctica los genes no disponen de valores de priorización para todos los interactomas, ya que, en general, todos muestran en mayor o menor medida signos evidentes de incompletitud, especialmente en aquellos interactomas que recogen relaciones de tipo complejo como los procesos de regulación. Debido a esto, es necesario determinar cuando el valor de priorización obtenido a partir de un interactoma puede ser aprovechable. Para el presente trabajo, se ha considerado que un interactoma no debe ser empleado cuando su valor de priorización es igual a 0, es decir, cuando no dispone de interacciones descritas para el gen.

Una vez seleccionados los valores de priorización correspondientes a cada interactoma válido, se computa un valor global de priorización, que permite establecer la relevancia del gen a lo largo de todo el sistema de información. La forma en la que se computa el valor global resulta limitada, ya que en general se dispone de muy pocos valores. Para el presente trabajo se han empleado dos funciones distintas para computar el valor de priorización global: la media y el máximo de los valores de priorización válidos.

2.3.4. Otras ponderaciones complementarias al método

La metodología propuesta comienza en el instante posterior a la selección inicial de candidatos por parte de cada familia. Hasta el momento, cada uno de los genes seleccionados por una familia muestra a priori la misma probabilidad de causar la enfermedad. Sin embargo, hay determinadas estrategias que pueden enriquecer o completar la metodología planteada estableciendo unas probabilidades a priori diferentes para cada gen. Estas estrategias pueden trabajar a partir de los propios datos del estudio, o con información conocida almacenada en bases de datos públicas. Esta información permiten inferir la importancia de cada gen dentro del contexto global de la célula y por tanto ponderar de forma positiva a aquellos genes que por su rol, podrían acarrear consecuencias mucho peores a la célula en caso de mal funcionamiento. Estas medidas pueden llegar a ser muy importantes ya que corrigen el valor de priorización obtenido en presencia de errores de secuenciación o por la falta de información en los interactomas.

A continuación, se describen algunas estrategias posibles.

Evaluación de las mutaciones del gen

La reglas biológicas que rigen el proceso de traducción de un ARN mensajero en una proteína totalmente funcional, describen como una única mutación puede ser capaz de inutilizar o provocar el mal funcionamiento de un gen y como consecuencia, una cascada de anomalías que derive en una enfermedad. No obstante, esto no debería obviar el hecho de que aquellos genes que acumulen un mayor número de mutaciones nocivas, deberían tener una probabilidad mayor de contener a la mutación causante de la enfermedad, por lo que deberían ser a priori mejor ponderados.

Por otro lado, se conoce que determinadas mutaciones en zonas codificantes, aun habiendo provocado un cambio de aminoacido en la secuencia de la proteína, en realidad no producen cambios significativos en su conformación y por tanto en su funcionamiento. En ese sentido, existen en la actualidad algunas herramientas informáticas disponibles, como SIFT [4] o Polyphen [5] que evalúan algunas características de la secuencia de aminoacidos mutada, y son capaces de proporcionan un estadístico proporcional al grado cambio en la proteína.

Otra de las formas de evaluar el potencial efecto de una mutación consiste en determinar si esta ha sido descrita anteriormente en población sana no incluida en el estudio, ya que de ser así, podría no reunir las características necesarias. Para tal efecto, es posible consultar si la mutación ha sido recogida por dbSNP [6], lo cual probaría a priori probaría su inocuidad, o consultar si ha sido descrita en el proyecto de los 1000 genomas [7], y en caso de ser así, con qué frecuencia alélica. Este dato resulta de gran utilidad, ya que, en general, las enfermedades complejas surgen a raíz de una combinación de mutaciones, que de forma individual sí pueden estar presentes en población sana.

Estimación del rol general del gen

El estadístico de priorización obtenido a partir de los interactomas depende totalmente del set de candidatos escogidos, de tal forma que un mismo gen, podría tener valores de ponderación muy diferentes, en función de los genes que le acompañen. Sin embargo, existen algunas medidas generales interesantes acerca del gen que pueden ser computadas de forma determinista a partir de un interactoma. Estas medidas permiten evaluar la importancia del gen en la red en función de parámetros como el número de conexiones. Una de las medidas que mejor describe el rol del gen dentro de la red de interacción lo constituye el concepto de centralidad, el cual trata precisamente de determinar la importancia relativa de un nodo en el contexto global del interactoma.

En la práctica, la centralidad puede ser computada de muchas maneras. Por ejemplo, se puede hacer uso de los siguientes indicadores:

- *Grado*: Número de conexiones existentes para el nodo. Es la medida más simple para describir la centralidad. A mayor número de conexiones, se le atribuye mayor importancia.
- *Cercanía*: Suma (o ocasiones media) de las distancias existentes entre un nodo y todos aquellos nodos accesibles. Se trata de una medida más compleja donde, a valores más pequeños, mayor cercanía y por tanto, mayor importancia.

- *Intermediación*: Frecuencia con la que un nodo aparece en el camino más corto entre cada para de nodos de la red. A mayor frecuencia, mayor importancia.

Otro de los enfoque actuales más interesantes para estimar la importancia de un nodo en una red lo constituyen los estadísticos empleados por los motores de búsqueda de internet para determinar la importancia de una *web*. El caso más popular lo representa el algoritmo PageRank de Google, el cual, debido a su generalidad, es directamente aplicable para estimar la importancia de un gen dentro de un interactoma.

Información funcional

Actualmente, se dispone de gran cantidad de información biológica relativa a los genes y los procesos biológicos en los que intervienen. Repositorios como Gene Ontology [8], o KEGG [9] ofrecen información estructurada en forma de ontologías acerca de rutas metabólicas, rutas de señalización y otros procesos biológicos descritos en la literatura. Esta información puede ser de gran utilidad, ya que si el investigador responsable del estudio conoce a priori aquellas funciones biológicas en las que el gen de la enfermedad debería estar implicado, se podría llevar a cabo un filtrado drástico de forma directa. El problema de esta metodología es que no se puede sistematizar con facilidad, ya que el proceso requiere de la intervención del investigador para definir las funciones clave, que en ocasiones, estarán descritas de forma diferente según el repositorio de consulta.

2.3.5. Diagrama general del método

Entrada:

G = [g1,g2,...gj] → lista de genes candidatos por familia
I = lista de interactomas
PC = priorizaciones complementarias

Algoritmo:

Construcción del set global de genes candidatos
C = union(G)

Cómputo de estadísticos de priorización
Para todo gen i contenido en el set de candidatos C

Para todo interactoma x

$p_{i,x} = 0$

Para toda familia j

$p_{i,j,x}$ = computo_estadístico (i, j,x)
$p_{i,x} = p_{i,x} + p_{i,j,x}$

fin

Cómputo del estadístico global
p_i = computo_estadístico_global ($p_{i,x}$)

Repriorización con estadísticos complementarios
Para todo estadístico complementario pc contenido en PC
$p_i = p_i * pc(i)$
fin

fin
fin

2.4. Validación

Después de plantear en detalle la metodología de priorización, es necesario realizar una serie de experimentos que permitan evaluar el procedimiento de forma global y la influencia de cada parámetro característico sobre la fiabilidad del resultado. Dicha validación se ha realizado en dos partes, en primer lugar se han empleado simulaciones para evaluar de forma exhaustiva cada parámetro del estudio, y por último, la metodología propuesta se ha aplicado en un caso real donde se conoce el gen causante de la enfermedad.

2.4.1. Simulaciones

Las simulaciones nos permiten evaluar de forma exhaustiva el rendimiento de los parámetros del método, los cuales, a partir de datos reales costarían mucho de calibrar. Con el fin de simular de forma realista un caso de estudio típico, los genes de enfermedad seleccionados para cada simulación han sido extraidos de síndromes reales. Concretamente, se han extraído a partir del repositorio OMIM, mediante el cual se preparó una lista de enfermedades mendelianas y sus correspondientes genes.

Para simular un estudio real, en primer lugar se decide el número de familias que lo componen y el número de genes por familia. A continuación, se escoge al azar una enfermedad de OMIM que contenga un número de genes mayor o igual al de familias. Seguidamente, se le asigna a cada familia un gen de la enfermedad seleccionada. Por último, se añade a cada familia un conjunto de genes aleatorios, componiendo así el set final de genes candidatos. Este conjunto de genes es el equivalente al que habría seleccionado una familia después de haber procesado los individuos que la componen.

Con el esquema de simulación planteado, se ha confeccionado una serie de experimentos destinados a evaluar algunos aspectos críticos de la metodología de priorización. A continuación se describen las diferentes tandas de simulación.

Medidas de distancia

En primer lugar, se ha realizado una tanda de experimentos con el fin de evaluar la eficacia de las distintas medidas de distancia propuestas. Los experimentos planteados son los siguientes:

distancia	repeticiones	familias	genes por familia
SP	100	3	150
ID	100	3	150
RW	100	3	150

Número de familias

El número de familias empleadas en el estudio es un parámetro crítico que va a influir en la fiabilidad de los resultados, ya que a mayor número de familias, menor es la probabilidad de encontrar genes aleatorios en todas las familias. Concretamente, se plantean las siguientes simulaciones.

distancia	repeticiones	familias	genes por familia
SP	100	**3**	100
SP	100	**4**	100
SP	100	**5**	100
SP	100	**7**	100
SP	100	**10**	100
ID	100	**3**	100
ID	100	**4**	100
ID	100	**5**	100
ID	100	**7**	100
ID	100	**10**	100
RW	100	**3**	100
RW	100	**4**	100
RW	100	**5**	100
RW	100	**7**	100
RW	100	**10**	100

Número de genes por familia

Otro de los parámetros importantes a evaluar lo constituye el número de genes por familia, ya que a mayor número de genes, más ruido entrará en el sistema y por tanto, más complicada será la priorización. Los experimentos planteados son los siguientes.

distancia	repeticiones	familias	genes por familia
SP	100	3	**25**
SP	100	3	**50**
SP	100	3	**100**
SP	100	3	**200**
SP	100	3	**500**
ID	100	3	**25**
ID	100	3	**50**
ID	100	3	**100**
ID	100	3	**200**
ID	100	3	**500**
ID	100	3	**25**
ID	100	3	**50**
ID	100	3	**100**
ID	100	3	**200**
ID	100	3	**500**

Solapamiento entre familias

En la práctica, las familias de un estudio pueden coincidir en el gen de la enfermedad. A nivel computaciones, esto provoca que el término asociado a la intersección directa tenga más peso que el término de interacción. Para este fin, se ha empleado otra tanda de experimentos donde se evalúa el grado de solapamiento entre familias:

distancia	repeticiones	familias	genes por familia	genes de enfermedad
SP	100	3	100	**5**
SP	100	3	100	**4**
SP	100	3	100	**3**
SP	100	3	100	**2**
SP	100	3	100	**1**
ID	100	3	100	**5**
ID	100	3	100	**4**
ID	100	3	100	**3**
ID	100	3	100	**2**
ID	100	3	100	**1**
ID	100	3	100	**5**
ID	100	3	100	**4**
ID	100	3	100	**3**
ID	100	3	100	**2**
ID	100	3	100	**1**

Otras ponderaciones complementarias

Por último, se ha considerado importante evaluar el grado de mejora ofrecido por la ponderación del estadístico de priorización con factores relativos al gen. En este caso, se han probado el estadístico de centralidad grado, y el algoritmo PageRank. La siguiente tanda de experimentos se ha diseñado para probar su influencia:

distancia	repeticiones	familias	genes por familia	complementario
SP	100	3	100	**ninguno**
SP	100	3	100	**grado**
SP	100	3	100	**pagerank**
SP	100	3	100	**grado + pagerank**
ID	100	3	100	**ninguno**
ID	100	3	100	**grado**
ID	100	3	100	**pagerank**
ID	100	3	100	**grado + pagerank**
RW	100	3	100	**ninguno**
RW	100	3	100	**grado**
RW	100	3	100	**pagerank**
RW	100	3	100	**grado + pagerank**

2.4.2. Caso de uso

La validación con simulaciones ha sido complementada con un caso real. Se trata de xx individuos correspondientes a 3 familias (figura xx) cuyo con el síndrome xxxx, cuyo gen causante es conocido.

2.5. Implementación

El método propuesto ha sido implementado en el lenguaje de programación R [10]. Para la gestión de redes se ha empleado el paquete iGraph. Tanto las simulaciones, como el procesamiento de las secuencias del caso de uso han sido ejecutados en un cluster formado por 4 máquinas de 48 Gb y 8 procesadores.

La redacción del presente trabajo ha sido confeccionada bajo el lenguaje Latex [11,12], por medio del editor Texmaker [13], bajo una máquina con sistema operativo Mac OSX.

3 Resultados

3.1. Simulaciones

3.2. Caso de uso

4 Conclusiones

4.1. Validez del método

4.2. Aportación del método en ausencia de genes conocidos

4.3. Limitaciones de la metodología y líneas futuras

medidas de distancia más óptimas***
interactomas incompletos***

5 Agradecimientos

Bibliografía

[1] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[2] Genome reference consortium. `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/`.

[3] F Sanger, S Nicklen, and AR Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of The National Academy of Sciences of The United States Of America*, 74(12):5463–5467, 1977.

[4] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4(7):1073–1081, 2009.

[5] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat. Methods*, 7(4):248–249, Apr 2010.

[6] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, Jan 2001.

[7] D. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. Wilson, R. A. Gibbs, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Wang, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, E. S. Lander, D. L. Altshuler, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, D. R. Bentley, N. Gormley, S. Humphray, Z. Kingsbury, P. Kokko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, H. Lehrach, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, M. Egholm, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, J. Kebbler, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte,

S. Wang, E. R. Mardis, R. K. Wilson, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, R. M. Durbin, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, R. A. Gibbs, D. Wheeler, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, J. Wang, X. Fang, X. Guo, R. Li, Y. Li, R. Luo, S. Tai, H. Wu, H. Zheng, X. Zheng, Y. Zhou, G. Li, J. Wang, H. Yang, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W. P. Lee, W. F. Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, D. L. Altshuler, A. D. Ball, E. Banks, T. Bloom, B. L. Browning, K. Cibulskis, T. J. Fennell, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, D. B. Jaffe, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, E. Shefler, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, C. D. Bustamante, A. G. Clark, A. Boyko, J. Degenhardt, S. Gravel, R. N. Gutenkunst, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, P. Flicek, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stutz, S. Humphray, M. Bauer, R. K. Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, A. Chakravarti, K. Ye, F. M. De La Vega, Y. Fu, F. C. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, S. T. Sherry, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, G. A. McVean, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, B. Desany, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. C. Melton, G. R. Abecasis, Y. Li, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. M. Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zollner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. C. Sahinalp, P. H. Sudmant, E. R. Mardis, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, D. Dooling, G. Weinstock, J. W. Wallis, M. C. Wendl, Q. Zhang, R. M. Durbin, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, D. M. Carter, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Q. Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, J. Stalker, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, Y. Li, R. Luo, G. T. Marth, E. P. Garrison, D. Kural, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, S. A. McCarroll, E. Banks, M. A.

DePristo, R. E. Handsaker, C. Hartl, J. M. Korn, H. Li, J. C. Nemesh, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, J. Degenhardt, M. Kaganovich, L. Clarke, R. E. Smith, X. Zheng-Bradley, J. O. Korbel, S. Humphray, R. K. Cheetham, M. Eberle, S. Kahn, L. Murray, K. Ye, F. M. De La Vega, Y. Fu, H. E. Peckham, Y. A. Sun, M. A. Batzer, M. K. Konkel, J. A. Walker, C. Xiao, Z. Iqbal, B. Desany, T. Blackwell, M. Snyder, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, K. Chen, A. Chinwalla, L. Ding, M. D. McLellan, J. W. Wallis, M. E. Hurles, D. F. Conrad, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, J. Du, F. Grubert, R. Haraksingh, J. Jee, E. Khurana, H. Y. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, R. A. Gibbs, M. Bainbridge, D. Challis, C. Coafra, H. Dinh, C. Kovar, S. Lee, D. Muzny, L. Nazareth, J. Reid, A. Sabo, F. Yu, J. Yu, G. T. Marth, E. P. Garrison, A. Indap, W. F. Leong, A. R. Quinlan, C. Stewart, A. N. Ward, J. Wu, K. Cibulskis, T. J. Fennell, S. B. Gabriel, K. V. Garimella, C. Hartl, E. Shefler, C. L. Sougnez, J. Wilkinson, A. G. Clark, S. Gravel, F. Grubert, L. Clarke, P. Flicek, R. E. Smith, X. Zheng-Bradley, S. T. Sherry, H. M. Khouri, J. E. Paschall, M. F. Shumway, C. Xiao, G. A. McVean, S. J. Katzman, G. R. Abecasis, E. R. Mardis, D. Dooling, L. Fulton, R. Fulton, D. C. Koboldt, R. M. Durbin, S. Balasubramaniam, A. Coffey, T. M. Keane, D. G. MacArthur, A. Palotie, C. Scott, J. Stalker, C. Tyler-Smith, M. B. Gerstein, S. Balasubramanian, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, C. D. Bustamante, N. Gharani, R. A. Gibbs, L. Jorde, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, G. A. McVean, P. N. Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, C. Tyler-Smith, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, F. S. Collins, A. Duncanson, E. D. Green, M. S. Guyer, J. L. Peterson, A. J. Schafer, Y. Xue, and R. A. Cartwright. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[9] M. Kotera, M. Hirakawa, T. Tokimatsu, S. Goto, and M. Kanehisa. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.*, 802:19–39, 2012.

[10] M. Kotera, M. Hirakawa, T. Tokimatsu, S. Goto, and M. Kanehisa. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.*, 802:19–39, 2012.

[11] Latex – a document preparation system. `http://www.latex-project.org/`.

[12] Michel Mittelbach, Frank; Goosens. *The LaTeX Companion (2nd ed.)*. Addison-Wesley, 2004.

[13] Texmaker.

6 Material suplementario

.1. First Appendix