Universidad de Valencia

Proyecto final de master

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

Autor:
José Carbonell Caballero

Tutor:
José Bermudez Edo

Septiembre 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 cómo 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 cómo 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 secuencia 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 cómo una rama esencial de la biología y gracias a la influencia de otras ramas cómo 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 cómo 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 cómo 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 cómo *dogma central de la biología molecular*. El dogma básicamente describe cómo la porción de ADN correspondiente a un gen se duplica (o transcribe) en forma de ARN mensajero y cómo este sale del núcleo y acaba siendo traducido a una proteína que llevará a cabo la función del gen. A pesar de que con el tiempo se ha demostrado que el dogma central no refleja con exactitud lo que ocurre a nivel celular, ha servido para ilustrar de forma intuitiva el ciclo de vida de los genes y cómo son capaces de llevar a cabo su función en la célula.

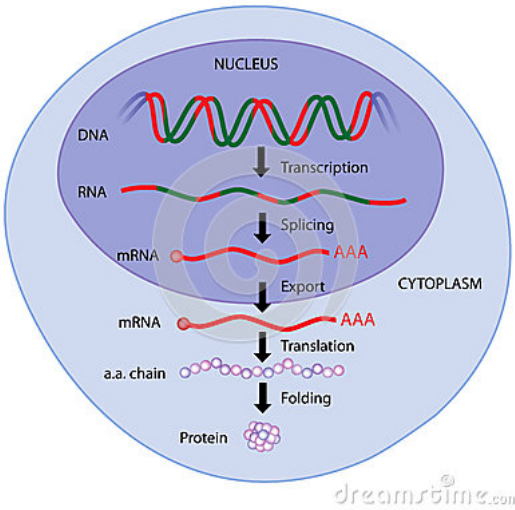

Figura 1.1: Dogma central de la biología molecular

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 fenómenos a escala molecular. Desde el microscopio hasta 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 molecular en los últimos 30 años ha sido la informática. Esta rama ha permitido procesar y almacenar de forma ordenada cantidades enormes de datos en tiempos de cómputo razonables. A la aplicación de la informática a la biología y su unión a otras disciplinas cómo la estadística o las matemáticas, se le denominó biología computacional.

La biología computacional, también denominada bioinformática, es una disciplina en la que, a partir de la combinación de modelos, algoritmos y computadores, es posible abordar la resolución de problemas de tipo biológico. 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 cómo evolucionan y se relacionan las especies cercanas con sus ancestros comunes.

En los últimos años, el volumen de datos aportados por la biología molecular ha crecido de forma considerable, lo que ha generado una gran dependencia a nivel informático que ha obligado 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 el Proyecto Genoma Humano [**?**], encargado de determinación total de

la secuencia de aminoácidos que compone el genoma humano. Esta tarea requirió más de 13 años de trabajo y una inversión superior a 280.000 millones de dólares, aportados mayoritariamente por los gobiernos de Gran Bretaña y Estados Unidos. La identificación de la secuencia del genoma humano y la localización de los genes tuvo una gran relevancia en los ámbitos de la biomedicina y la genética clínica, permitiendo dar luz sobre las bases moleculares de algunas enfermedades hereditarias.

1.2. Enfermedades de origen genético

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 cómo 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 mucho más frecuentes que las monogénicas, y engloban a la mayoría de enfermedades crónicas cómo la hipertensión, la obesidad, o síndromes complejos cómo 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 cómo modelo de herencia de la enfermedad.

Estructura genómica y modelos de 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 idénticas de cada gen, donde una copia vendrá proporcionada por vía paterna y la otra por vía materna durante la meiosis.

*** figura cromosomas en la célula punto de recombinarse

Esta duplicidad génica tiene efectos beneficiosos, tanto sobre el individuo cómo sobre la especie, ya que genera (por combinación) biodiversidad, haciéndonos más robustos a enfermedades. Sin embargo, también tiene efectos funcionales ya que, cada parental dispone de variaciones estructurales y mutaciones propias que generan pequeñas diferencias entre las dos copias génicas que recibe cada descendiente.

Cuando una mutación está presente sólo en una copia, entonces decimos que se presenta en heterocigosis, sin embargo cuando la mutación es compartida por

ambas copias, es decir, ambos parentales disponían de ella, entonces decimos que está presente en homocigosis.

Cuando una enfermedad hereditaria muestra una herencia de tipo recesivo, significa que será necesario disponer de 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 acabará siendo funcional. Sin embargo, cuando la herencia de la enfermedad es de tipo dominante, entonces significa que una sola copia mutada será suficiente para desencadenar todos los cambios celulares. 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 cómo 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 genes que puedan estar implicados en una enfermedad con un origen genético total o parcialmente desconocido. El problema se complica cuando las variaciones genómicas no son responsables en su totalidad del inicio de la enfermedad. En este caso, las condiciones ambientales (cómo los hábitos alimenticios o el clima) explican un porcentaje significativo del riesgo a desarrollar la enfermedad. Existen enfermedades cómo el cáncer de pulmón, donde los hábitos (cómo ser fumador) juegan un papel decisivo en su riesgo y proliferación, y otras enfermedades, principalmente monogénicas, donde presentar ciertas variaciones estructurales asegura una penetrancia casi total, con independencia de los hábitos del individuo.

Dentro del estudio de enfermedades hereditarias, es de especial interés el estudio de enfermedades raras. Las enfermedades raras, o también conocidas cómo 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 y 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.

Panorama mundial /penetrancia/ ***

Repositorios de polimorfismos y mutaciones

1000 genomas: mutaciones en población sana ***

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 facilita la iden-

tificación 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 constituye una plantilla sana con la que comparar y definir la normalidad no es algo sencillo. Esta tarea importante recae sobre un consorcio internacional [2] encargado de construir y almacenar la versión oficial del genoma humano, actualizándolo de forma periódica.

La secuenciación de genomas o porciones de ADN ha sido llevada a cabo por medio de diferentes procedimientos a lo largo de los últimos 50 años. Los métodos implementados han permitido determinar con fiabilidad la secuencia de nucleótidos correspondiente a un fragmento de material genético. Todos los métodos han tenido tasas de error cuantificables y parámetros críticos cómo su complejidad, facilidad de ejecución, velocidad de secuenciación o coste económico, los cuales han sido claves para su evolución y adopción por parte de los investigadores.

Uno de los principales handicaps producidos por el desarrollo de tecnología heterogénea es que cada método de secuenciación ha venido acompañado de su propia base científica y de numerosas herramientas y métodos estadísticos de análisis específicos, lo que ha requerido un reaprendizaje de las 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. Cabe destacar, que a pesar de mostrar una cronología clara, todos los métodos son actualmente vigentes y es habitual verlos coexistir en proyectos genómicos, donde cada tecnología es usada en su ámbito adecuado.

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 han recibido un premio nobel dos veces en su vida, fue capaz de demostrar que las proteínas tienen una 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, que se convirtió 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 fue galardonado con 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, ha sido necesario evolucionar a formas más rápidas y automáticas de secuenciación que dejaran atrás algunas limitaciones técnicas cómo la secuenciación de largas cadenas de ADN. Con una perspectiva sistémica, se ha evolucionado hacia métodos de secuenciación que son capaces de obtener información simultanea de miles de genes.

Una de las tecnologías más populares surgidas en la última década han sido los chips de ADN, los cuales, gracias su buena relación coste/prestaciones se extendieron rápidamente a la gran mayoría de estudios biomédicos con una componente genómica.

Un chip de ADN se compone de una matriz de sondas (o pocillos), donde cada elemento es capaz de medir información concreta acerca de un gen. Los chips más empleados han sido los de expresión génica. Estos, recogen en cada pocillo una porción de la secuencia complementaria de cada gen, a partir de la cual son capaces de obtener una medida proporcional a la abundancia de ARN mensajeros y por tanto de la expresión de los genes.
*** figura microarrays

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, que actuan sobre marcadores genómicos asociados a enfermedades.

Si bien, los chips de ADN no han constituido una tecnología capaz de obtener la secuencia precisa de nucleótidos de cada gen, si han servido de forma eficiente y barata para determinación algunas características esenciales relacionadas con la dinámicas 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, algo que hace 15 años habría sido dificil de imaginar. 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 los aparatos son capaces de secuenciar. Actualmente, los

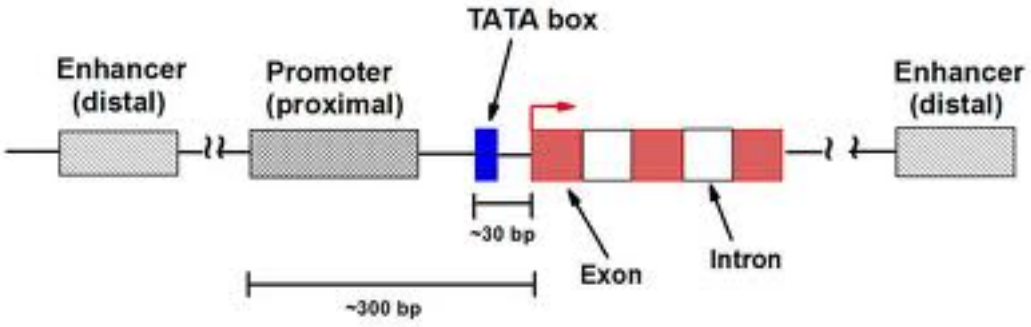

Figura 1.2: Estructura de un gen

secuenciadores son capaces de obtener fragmentos de entre 50 y 500 nucleótidos, lo cual, comparado con técnicas clásicas cómo la secuenciación por Sanger, resulta muy pobre. Además, los fragmentos de ADN secuenciados tienen que volver a ser reposicionados sobre el ADN para poder reconstruir el genoma completo de la muestra y los secuenciadores no aportan ningún tipo de información a este respecto. Por esta razón, es necesario realizar un paso de computado adicional denominado *mapeo* donde cada fragmento secuenciado se compara con un genoma humano de referencia. Algunos parámetros, cómo la longitud del fragmento, serán claves para la fiabilidad del mapeo, ya que, a menor tamaño, más probable será encontrar la misma secuencia de nucleótidos en varios puntos del genoma.

Otra limitación importante reside en la tasa de error obtenida durante la determinación de la secuencia de nucleótidos, la cual es superior al resto de técnicas de secuenciación. Por esta razón, es habitual realizar un análisis estadístico a partir de datos de ultrasecuenciación, complementados con una fase de validación posterior realizada por el método *Sanger*.

A pesar de todo, los ultrasecuenciadores constituyen una tecnología más que prometedora que aporta un enfoque genómico muy eficiente en el estudio de aquellas enfermedades donde los genes causantes de la enfermedad son desconocidos

1.4. Identificación de genes candidatos

1.4.1. Regiones de interés

Los genes constituyen únicamente el 5 % del genoma, el resto, antaño considerado cómo 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 dedicadas al control de su expresión, e incluso otras regiones, denominadas intrones, implicadas en la formación de las diferentes isoformas del gen.

Debido a la estructura y funcionamiento del genoma, se considera mucho más probable encontrar una mutación causante de enfermedad dentro de una

zona codificante, ya que esta produciría un cambio directo sobre las proteínas generadas y por consiguiente, un comportamiento anómalo. Esta circunstancia 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, que puede tener implicaciones serias en casos donde la mutación nociva no afecta a una zona codificante sinó a una zona reguladora, permite en la práctica trabajar sólo con el un 1 % del total del genoma, lo cual aporta grandes beneficios para toda la fase de procesamiento, almacenamiento y análisis que ocurre después de la secuenciación, consiguiendo un aprovechamiento máximo de los recursos económicos disponibles dentro de un proyecto de investigación.

1.4.2. Estudio caso/control

El protocolo de selección de genes candidatos se realiza a partir de un estudio caso/control. En este caso, se trata de identificar a aquellos genes mutados que podrían tener una relación directa con la enfermedad. cómo 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, se obtienen todas las mutaciones propias de cada individuo del estudio, descartando a aquellas que no muestren un efecto nocivo probado. Las mutaciones seleccionadas, serán sometidas, generalmente, a un test estadístico de proporciones (cómo el test exacto de Fisher) que identificará a aquellos candidatos de mayor prevalencia en casos frente a controles. Cabe destacar que este proceso puede realizarse a nivel mutacional, donde todos los individuos enfermos compartirán exactamente el mismo cambio nocivo, o a nivel génico, donde lo importante es identificar al gen mutado, por encima de las variaciones estructurales propias de cada individuo enfermo.

Debido a limitaciones relacionadas con el tamaño muestral y otro tipo de sesgos surgidos durante todo el proceso, el gen, o genes causantes serán seleccionados junto a un grupo de genes aleatorios que cumplirán los mismos requisitos y que por tanto no podrán ser filtrados ni separados de estos. Será tarea de estrategias posteriores encontrar argumentos biológicos o técnicos que permitan reducir el número de candidatos y priorizar convenientemente a los que quedan, ya que, los mecanismos de validación posteriores son complejos e incorporan de forma habitual costosos trabajos de laboratorio. El tamaño muestral y otros parámetros de calidad relacionados con la secuenciación condicionarán claramente la talla del conjunto de genes candidatos, aunque otros aspectos relacionados con la fase computacional también podrían tener gran influencia.

1.4.3. Reducción y priorización de genes candidatos

La identificación de los genes implicados en el origen de una enfermedad hereditaria constituye una tarea ardua, con unos requerimientos temporales y

económicos considerables, en la que es habitual enfrentarse al estudio detallado de cientos de genes candidatos. En este contexto, es necesario contar con protocolos de priorización (manuales o automáticos) que permitan reducir la lista de candidatos iniciales o priorizar convenientemente a los seleccionados en función de criterios biológicos.

El proceso de priorización establece un ranking de genes que estima la relevancia de cada candidato sobre los procesos biológicos que se presuponen clave en la enfermedad de estudio, siendo los mejor priorizados aquellos que serán empleados en los análisis posteriores. Los criterios empleados hasta la fecha para determinar la implicación de un gen sobre una enfermedad son muy amplios. En general, se comienza haciendo uso de la estructura de los individuos del estudio caso/control, para posteriormente, recopilar propiedades inherentes a los genes que describen su relevancia global, y en consecuencia, permite modular las probabilidades a priori de cada candidato propuesto.

La selección de genes candidatos comienza generalmente por la delimitación de aquellas regiones del genoma que podrían contener al gen o genes causantes. Esta tarea se ha llevado a cabo por medio de estudios que relacionan zonas cromosómicas con rangos fenotípicos, o en casos familiares, a partir de la intersección entre las zonas mutadas de uno o varios individuos enfermos con sus parentales también afectados.

Desgraciadamente, las técnicas de posicionamiento no consiguen centrar convenientemente el problema sobre una reducida lista de regiones que todavía contenga al gen de la enfermedad. Por el contrario, muestran algunas limitaciones claras, especialmente en síndromes de origen regulatorio, donde es dificil delimitar las regiones cromosómicas afectadas. Además, no tienen en cuenta el plegamiento tridimensional de los cromosomas que provoca que dos genes cercanos en el espacio, aparezcan alejados al tratar la secuencia de ADN cómo un fragmento lineal. A pesar de todo, constituyen un buen punto de inicio, y permiten filtrar algunas regiones candidatas que sólo aportarían ruido al análisis.

El uso de información funcional acerca de los genes permite establecer de forma eficiente criterios cuantitativos y cualitativos de *culpabilidad* sobre los candidatos. Las bases de datos con información biológica describen cómo los genes intervienen en rutas metabólicas y de señalización [9], qué funciones biológicas desempeñan en la célula y en qué compartimentos [8], que mutaciones han sido descritas previamente [6] y en qué enfermedades han sido implicadas [**?**], así cómo otro tipo de anotaciones biológicas que pueden tener gran relevancia a la hora de obtener un perfil específico de cada gen.

La información cubre generalmente las siguientes áreas:

- *Redes de interacción proteina-proteina:*
  Interacciones físicas entre las proteínas generadas por los genes.
- *Anotaciones funcionales:*
  Funciones biológicas, compartimentos celulares y otros roles asignados a las proteinas.
- *Rutas metabólicas y de señalización:*
  Diagrama de las diferentes rutas biológicas en las células y su implicación en algunas enfermedades.

- *Asociación a enfermedades:*
  Asociaciones descritas en la literatura de tipo gen/enfermedad o mutación/enfermedad.
- *Expresión:*
  Expresión de los genes por tejido o momento del desarrollo.
- *Conservación:*
  Grado de conservación de la secuencia de nucleótidos entre diferentes especies cercanas.
- *Redes de regulación:*
  Relaciones entre genes reguladores y regulados.
- *Minería de textos:*
  Frecuencia de aparición de un gen en publicaciones científicas, junto a otros términos relativos a enfermedades, procesos biológicos u otras entidades de interés.

El concepto de priorización computacional, tal y cómo hoy lo conocemos, fue introducido por primera vez en 2002 por Perez-Iratxeta [**?**]. A partir de ese momento, multitud de métodos y algoritmos, en su mayoría de uso libre, han sido desarrollados para llevar a cabo tal tarea [**?**, **?**, **?**]. Los métodos propuestos consiguen clasificar y priorizar de forma automática grandes cantidades de genes, lo que los hace muy adecuados en estudios de genoma completo, especialmente en el estudio de aquellas enfermedades genéticas de origen desconocido y el número de candidatos es grande.

En general, la inmensa mayoría de métodos propuestos tratan de recuperar la información disponible sobre aquellos genes que ya se han descrito en la enfermedad, con el fin de establecer un criterio de priorización sobre los nuevos candidatos. El mecanismo habitual consiste en obtener un vector de características sobre los genes de referencia, que posteriormente será empleado para medir el grado de similitud entre cada candidato y los genes conocidos, siendo el gen más similar el de mejor priorización. Esta metodología reposa sobre la idea básica de que aquellos genes de enfermedad que todavía no han sido descritos, deberían tener características funcionales, similares a los genes existentes.

1.4.4. Validación del conjunto final de genes candidatos

***modelos animales
***grupo de test

1.5. Metodologías de priorización en ausencia de genes conocidos

La metodología habitual presenta algunas limitaciones. En primer lugar, es necesario disponer de una lista de genes previamente descritos para la enfermedad de estudio, algo que no siempre ocurre. En estos casos, aquellos procesos

biológicos relevantes podrían aportar una lista de genes cercanos con los que establecer una referencia secundaria. Por otro lado, la mayor parte de enfermedades mendelianas, dispone de muy pocos genes descritos, lo que a nivel estadístico, supone un problema importante durante la fase de construcción del perfil.

2 Material y métodos

2.1. Obtención de genes candidatos

La metodología propuesta en el presente trabajo trata de describir cómo 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 (aunque no necesariamente con la misma mutación), cómo con un estudio de tipo más general, donde los individuos disponibles no muestran parentesco alguno.

Cuando afrontamos un estudio de tipo familiar, los individuos seleccionados, tanto casos cómo 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, cómo de sujetos control propios. La ventaja proporcionada por las estructuras familiares es que a priori se espera que todos los individuos enfermos pertenecientes a la misma familia compartan la misma alteración genómica, 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 aproximadamente el 50 % de su genoma con su hermano enfermo) puede tener una capacidad de filtrado equivalente a cientos de controles externos.

Finalmente, se dispone de una lista de genes candidatos por 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 puntua a cada gen candidato en función de su relevancia global sobre el total de familias o individuos estudiados. Básicamente,

se trata de computar la frecuencia de aparición de cada candidato sobre las listas de genes independientes y complementar este valor con un término proporcional a la cantidad de interacciones descritas entre el candidato y dichos genes.

El término de interacción trata de incrementar el peso de un gen sobre una familia cuando este no ha sido explicitamente seleccionado cómo un candidato, de forma que, en caso de existir evidencias claras de interacción con uno o varios de sus genes, se pueda afirmar que el gen también resulta importante para la familia .

Aproximadamente un 15 % de enfermedades mendelianas disponen de más de un gen distinto con capacidad para producir la misma enfermedad, este porcentaje aumentaría mucho en el caso de las enfermedades complejas, donde un mismo gen mutado combinado con otros genes distintos, puede jugar un rol importante en varios síndromes.

La biología de sistemas demuestra que los genes actuan de forma colaborativa para llevar a cabo las distintas tareas esenciales para la célula. Los genes se comunican en procesos biológicos donde el grado de coordinación entre las distintas moléculas participantes es muy alto. Esto implica que si un síndrome está causado por el mal funcionamiento de una ruta metabólica, es probable que alterando cualquiera de sus genes importantes, se acabe produciendo el mismo resultado para el individuo.
*** imagen pathway

Hay que señalar que, a menudo, las células disponen de mecanismos naturales de compensación que les permiten sobrevivir pese al mal funcionamiento de alguno de sus genes importantes. Desgraciadamente, estos mecanismos no siempre consiguen evitar la aparición de anomalias graves que posteriormente provocarán la aparición de una enfermedad.

El proceso de priorización comienza con el reclutamiento de todas las listas de genes candidatos aportadas por cada familia.

$$C = [c_1, c_2, ..., c_i, .., c_n] \tag{2.1}$$

donde C se corresponde con el vector de listas de candidatos, n con el número de familias (o grupos independientes) del estudio y c_i con la lista de candidatos aportada por la familia *iesima*.

A continuación, se procede a construir el set global de candidatos G formado a partir de la unión de todas las listas independientes

$$G = unico(\cup \forall c_i, i \exists C) \tag{2.2}$$

Seguidamente, se procede a calcular el estadístico de priorización para cada gen contenido en el set global de candidatos:

$$\rho_i = \Phi(g_i), \forall g_i, i \exists G \tag{2.3}$$

Por último, se ordena la lista global de candidatos en función del estadístico de priorización computado.

$$R = orden(\rho) \tag{2.4}$$

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.

$$\rho_i = \sum_{j=1}^{n} \Phi(g_i, c_j) \tag{2.5}$$

donde $\Phi(g_i, c_j)$ se corresponde con el peso del gen i sobre la familia j

Cuanto mayor sea el peso del gen en las distintas familias, o mayor sea el número de familias en las que el gen está 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 y en 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. Concrétamente:

$$\rho_{i,x} = \sum_{j=1}^{n} \alpha_j * \gamma_{ij} + \delta(i,j) * (1 - \gamma_{ij}) \tag{2.6}$$

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.

2.2.2. Estimación del grado de interacción entre genes

La parte más compleja del proceso consiste en cómo estimar el grado de interacción entre un gen candidato y el conjunto de genes aportados por una familia. Para ello, 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 (o redes de interacción génica). Se trata de bancos de datos 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.

La red de interacción, donde los nodos son los genes y las aristas sus interacciones, se define cómo:

$$I = (V, E) \tag{2.7}$$

donde I se corresponde con el grafo general, V con el conjunto total de genes, y E con sus interacciones descritas.

La red permite recuperar los vecinos directos de un determinado gen, pero también reconstruir totalmente los caminos o secuencias de genes que podríamos emplear para llegar desde un nodo (o gen) de la red a otro:

$$P_{i,j} = [P_{i,j,0}, P_{i,j,1}, ..., P_{i,j,t}] \tag{2.8}$$

donde $P_{i,j}$ se corresponde con el conjunto total de t caminos entre el nodo V_i y el nodo V_j.

Asímismo, cada camino se compone del conjunto de interacciones entre ambos genes.

$$P_{i,j,\alpha} = [E_{i,0}, E_{0,1}, E_{1,2}, ..., E_{k-1,k}, E_{k,j}], V_i \in V, V_j \in V \tag{2.9}$$

done $P_{i,j,\alpha}$ se corresponde con un posible camino del nodo V_i al nodo V_j compuesto por las interacciones entre ambos nodos y los k intermediarios necesarios.

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.

$$d_{i,j} = f(P_{i,j}) \tag{2.10}$$

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.

En la práctica, se emplearán varios interactomas distintos, 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 cómo interactomas hayan sido empleados en el estudio. Así pues, una de las tareas importantes dentro de la metodología propuesta consiste en cómo unir o ponderar los resultados obtenidos con cada interactoma para obtener un resultado global.

$$\rho_i = f([\rho_{i,0}, \rho_{i,1}, \rho_{i,m}]) \tag{2.11}$$

donde ρ_i se corresponde con el estadístico de priorización global obtenido para el gen i, a partir del valor indivual obtenido en cada uno de los m interactomas empleados.

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. Algunas enfermedades mendelianas pueden ser descritas de forma casi completa mediante la alteración de un conjunto de interacciones físicas entre sus proteínas, sin embargo, otras enfermedades, principalmente complejas, serían mejor descritas por un interactoma de regulación.

Por otro lado, es importante indicar que no todos los interactomas describen en realidad interacciones directas entre los genes. En la práctica, se emplea el mismo tipo de representación para describir cualquier tipo de relación entre dos nodos, cómo por ejemplo, una evidencia de coexpresión entre dos genes. Esta representación permite evaluar de forma sencilla las relaciones indirectas entre dos genes. 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 escogidos de forma aleatoria.

También es importante resaltar, que la gran mayoría de interacciones existentes entre dos proteinas, son todavía desconocidas. Esto significa que los interactomas disponibles muestran un grado de incompletitud bastante grande, que en la práctica podría suponer un handicap considerable a la hora de realizar la priorización.

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 provoca que, en general, se disponga de un número de falsos positivos muy elevado en relación al número de positivos esperados. Tanto si se trabaja con familias, cómo 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 la talla del conjunto de genes candidatos a priorizar y por tanto, sobre la precisión final del resultado obtenido.

Si se realiza un estudio familiar con una enfermedad mendeliana, únicamente se espera encontrar un gen responsable por familia, ya que todos los individuos enfermos deberían coincidir en su mutación nociva. El resto de genes seleccionados cómo candidatos se corresponden con falsos positivos 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 cada familia estará entre 20 y 200 genes, lo cual significa que, en el mejor de lo casos, estaremos introduciendo aproximadamente un 95 % de falsos positivos.

En la metodología propuesta, la intersección realizada entre familias debería descartar de forma clara 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.

Tanto si se trabaja con una enfermedad mendeliana, o sobre un síndrome complejo, es probable encontrar, en el conjunto total de individuos, más de un gen distinto que de forma directa o indirecta regule los procesos biológicos alterados en el síndrome de estudio. Si las interacciones existentes entre los genes causantes estuvieran correctamente descritas en un interactoma, significa que se acabaría observando un cluster o vecindario en la red con una densidad de genes candidatos por encima de lo normal.

El hecho de que la mayoría de genes aportados por las familias sean de carácter aleatorio permite sugerir que el grado de interacción medio para una red debería ser equivalente al obtenido en promedio para un grupo de genes escogidos de forma aleatoria. Eso significa que el vecindario que contiene al grupo de genes causantes tendrá un valor de priorización medio significativamenter superior a la media y que por tanto debería ser identificable. Sin embargo, en la práctica resulta más complicado ya que si el proceso de selección de los candidatos iniciales no consigue acotar de forma considerable el conjunto de genes a evaluar, surgiran otros vecindarios altamente conectados simplemente por azar.

Otro aspecto a tener en cuenta es que la región de la red donde figuran los genes a identificar podría no estar 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 el grupo de genes causante de la enfermedad.

2.3. Estimación de parámetros

2.3.1. Medidas de distancia

El estadístico de priorización incorpora 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 cómo entrada al conjunto formado por los N caminos disponibles entre ambos nodos (ecuación 2.10). 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 de distancia diferentes que recogen varios enfoques distintos a la hora de cuantificar la proximidad entre dos nodos.

Las medidas empleadas son las siguientes:

Shortest path (camino más corto)

La distancia entre dos nodos viene definida por la longitud del camino más corto entre ellos.

$$SP_{i,j} = min(L(P_{i,j})) \tag{2.12}$$

donde L se corresponde con la función de longitud computada sobre el conjunto de caminos $P_{i,j}$ y $SP_{i,j}$ su valor mínimo.

Se trata de una medida que simplifica enormemente el cómputo de distancias, pero que deja de lado algunas características propias de la topología de la red, cómo el número de caminos existentes entre dos nodos.

Intermediación

Mide la frecuencia en la que el gen i aparece en el total de caminos cortos que parte del gen j

$$IM_{i,j} = \frac{L(i \in SP_j)}{L(SP_j)} \tag{2.13}$$

donde $IM_{i,j}$ se corresponde con la frecuencia de intermediación de i sobre j y SP_j el conjunto total de caminos cortos tomando cómo origen al gen j.

En este caso, se considera que cuando un gen i aparece de forma sistemática en los caminos que parten de j hacia el resto de genes, significa que ambos interaccionan en un gran número de procesos celulares y que por tanto, están muy próximos, aunque el camino más corto que los une sea largo.

Random walk

La medidad de distancia viene determinada por la probabilidad de llegar al gen j, partiendo desde el gen i, cuando el *viajero* elige caminos aleatorios.

$$RW_{i,j} = P(j|i) * P(j) \tag{2.14}$$

Se trata de la adaptación del conocido algoritmo de *random walk* para el trabajo con redes. Se trata de un método computacionalmente costoso, pero que tiene en cuenta la estructura total de la red de interacción.

2.3.2. 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, cómo a partir de interactomas nuevos, que describen otro tipo de relaciones no empleadas hasta ese momento. El hecho de

almacenar tantos tipos de interacción cómo sea posible, permite disponer de un criterio más amplio y preciso a la hora de evaluar la relación entre dos genes.

Después del proceso de priorización se dispone para cada gen de tantos valores cómo interactomas hayan sido incluidos en el estudio. De esta forma, el estadístico de priorización del gen i para un interactoma x se define cómo:

$$\rho_{i,x} = \sum_{j=1}^{n} \alpha_j * \gamma_{ij} + \delta(i,j,x) * (1 - \gamma_{ij}) \tag{2.15}$$

donde $\rho_{i,x}$ se corresponde con el valor de priorización para el interactoma x y $\delta(i,j,x)$ cómo el grado de interacción entre el gen i y la familia j en ese mismo interactoma.

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

$$\delta(i,j,x) = f([d(i,G_{j1},x), d(i,G_{j2},x), ..., d(i,G_{jk},x)]) \tag{2.16}$$

donde d es la función de distancia entre el gen i y un gen de la familia j en el interactoma x 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. En este caso, se ha definido el valor de interacción gen/familia cómo la media de aquellas medidas de distancia con un valor superior al cuantil 0.95 de la distribución, es decir, la interacción entre un candidato i y una familia j se computa únicamente a partir de los genes con mayor proximidad.

Por último, se computa el valor de priorización global tal que:

$$\rho_i = f([\rho_{i,1}, \rho_{i,2}, ..., \rho_{i,m}]) \tag{2.17}$$

donde ρ_i se corresopnde con el valor global de priorización para el gen i obtenido a partir de un función f que integra los valores de priorización de los m interactomas.

Desgraciadamente, en la práctica los genes no disponen de información para todos los interactomas disponibles, ya que, en general todos muestran en mayor o menor medida signos evidentes de incompletitud. Debido a esto, es necesario determinar cuando el valor de priorización obtenido a partir de un interactoma aporta información, o por el contrario, provoca una subestimación del peso. 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.

Asímismo, se han empleado dos funciones integradoras distintas: la media y el máximo de los valores de priorización que no son 0. Cabe destacar que en el mejor de los casos, se dispondrá de tantos valores cómo interactomas, lo que puede impedir la aplicación de otras funciones integradoras de mayor complejidad.

2.3.3. 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, la cual permite estimar 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 corregir 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 cómo una única mutación puede ser capaz de inutilizar o provocar el mal funcionamiento de un gen y cómo 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, cómo SIFT [4] o Polyphen [5] que evalúan algunas características esenciales de la secuencia de aminoacidos mutada, y proporcionan un estadístico que describe el 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 condiciones necesarias para producir la enfermedad. 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.

Rol general del gen en la red de interacción

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 cómo 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.

$$Grado_i = L([E_{i,1}, E_{i,2}, ..., E_{i,k}]) \tag{2.18}$$

donde L es la función de longitud sobre el vector de k interacciones que contienen al gen i

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.

$$Cercania_i = \sum_{j=1}^{k} d(c_{i,j})/k \tag{2.19}$$

donde $Cercania_i$ se corresponde con la media de la distancia de todos los k caminos entre el gen i y cualquier otro nodo j

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 la 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 cómo 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.

2.3.4. 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 el 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.

$$c_i = [T_i, R_1, R_2, ..., R_k] \tag{2.20}$$

donde c_i se corresponde con la lista de candidatos aportada por la familia $iesima$, compuesta por un gen de enfermedad T_i y un k genes escogidos de forma aleatoria.

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, cómo 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

Medidas de distancia

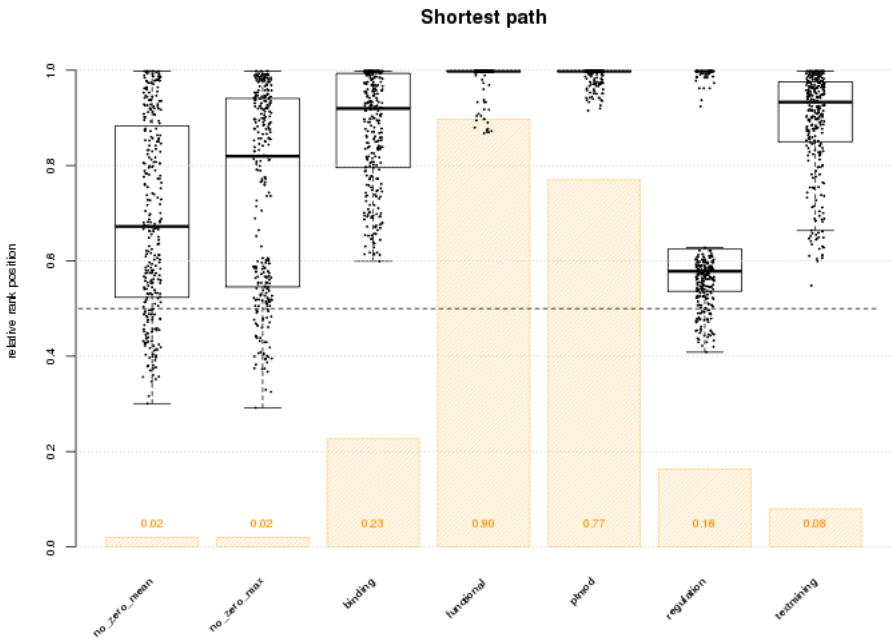

Figura 3.1: Shortest path. Posición relativa en el ranking para los genes de enfermedad en los distintos interactomas

Número de familias

3.2. Caso de uso

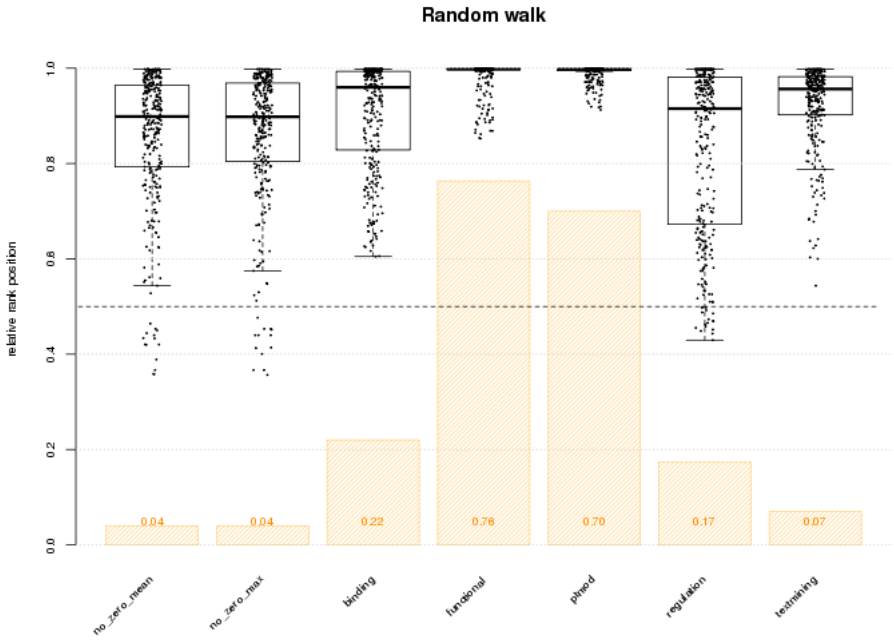

Figura 3.2: Random walk. Posición relativa en el ranking para los genes de enfermedad en los distintos interactomas

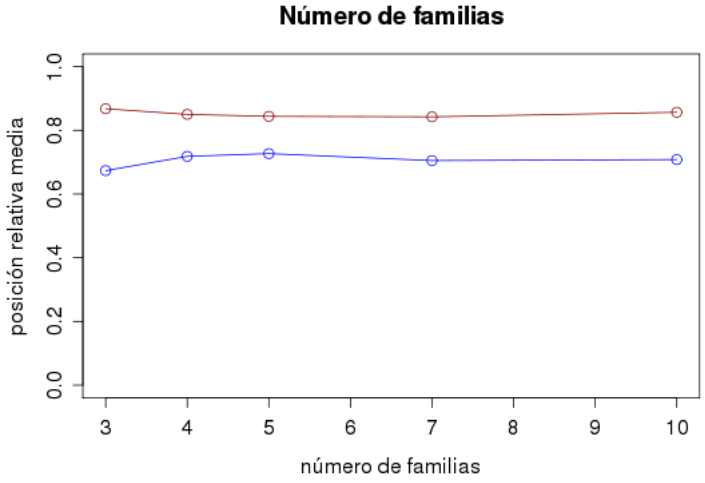

Figura 3.3: Comparación entre métodos para el número de familias medio

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