Proyecto final Genómica Computacional:

"Identificación de módulos y patrones funcionales mediante mutaciones en glioblastomas "

Elaborado por:
- Mier Fenogilo Sofía
- Pérez Romero Natalia Abigail
- Reyes Tapia Valeria
- Torres Aduna Sebastian Ulises

Introducción

Hoy en día la tecnología nos ha abierto muchas puertas, en particular en la medicina y el estudio del ser humano. Con los avances que se han dado ahora podemos, entre otras cosas, estudiar con detalle el genoma del ser humano y hacer análisis sobre su estructura y mutaciones. Poder analizar las mutaciones es especialmente útil e importante hoy en día, dado que de estas se pueden derivar diferentes tipos de cáncer.

Día con día, los proyectos que analizan, identifican y trabajan con tumores han empezado a generar un gran volumen de datos sobre aberraciones genómicas, epigenómicas y de expresión génica, lo que transforma nuestro concepto sobre la biología del cáncer, así como la posible revelación de nuevos biomarcadores, objetivos farmacológicos y la forma en la que posiblemente cambiará el desarrollo de nuevas terapias contra el cáncer.

La manera en la que analizamos los tumores cambió hace poco tiempo, no más de 15 años, pues pasamos de estudiarlos como imágenes a hacer análisis computacional y estadístico (1). Tras este cambio de paradigma hemos sido capaces de obtener mucha información genética y epigenética sobre cómo se comporta el **cáncer**, así como entender el efecto que pueden tener algunos medicamentos y su velocidad de reacción (2).

En el artículo "Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors" publicado por Christopher Miller, Stephen Settle y

Erik Sulman se detallan las diversas técnicas que se han realizado en tumores así como las muestras obtenidas sólo con aberraciones genómicas, entre ellas, mutaciones puntuales y alteraciones del número de copias, se comenta que en análisis anteriores han descubierto en una colección de muestras la alteración de genes que comprenden un módulo funciona y se han percatado que se encuentra en el mismo tumor, una explicación a esto, es que es que existen relaciones funcionales entre los genes (principalmente para los patrones RME). La relevancia principal de este tema es encontrar **aberraciones somáticas** en ciertos tipos de tumores, lo cual nos puede llevar a entender los caminos que derivan en su avance (3).

En este estudio nos enfocaremos en patrones RME (recurrent and mutually exclusive aberrations, es decir, los patrones de las aberraciones recurrentes y mutuamente excluyentes). Emplearemos la metodología empleada por Miller et al. en el artículo *Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors* (2) para encontrar los patrones RME y posteriormente ver cuáles son los **módulos funcionales** para los tumores. La ventaja de este tipo de acercamiento es que nos da suficiente información como para identificar módulos funcionales sin tener información previa más allá de los patrones RME.

Lo que nos interesa ahora es ver si podemos aplicar esta misma metodología a melanomas y de ahí obtener los caminos que nos llevan a tumores y encontrar nuevos módulos funcionales. El artículo que empleamos como base (2) utiliza glioblastomas para verificar el método. El dataset con la información de dichos tumores lo obtienen del portal TCGA (The Cancer Genome Atlas Program). De este mismo sitio sacaremos los datasets que vamos a emplear en nuestro análisis.

Objetivos

Lo que buscamos es recrear una red de módulos funcionales con un método ya existente que permite identificar patrones de mutación recurrentes y mutuamente excluyentes. Con esto vamos a poder filtrar los módulos funcionales que afectan más en el cáncer de piel (melanomas) de forma computacional. A la vez, podremos obtener resultados de forma rápida y sencilla.

Presentación del Trabajo

El estudio y trabajo realizado para nuestro proyecto será que estos patrones RME puedan usarse para identificar grupos de genes que están funcionalmente relacionados. ¿Cómo lo vamos a realizar? Más adelante les explicaremos paso por paso Los patrones RME corresponden a módulos del tipo "OR" donde anular la función de un miembro en cada módulo es suficiente para que falle (fenotipo tumoral), estos patrones son suficientemente informativos para permitir el descubrimiento de módulos funcionales relacionados con el cáncer sin utilizar ninguna información previa. Una vez realizadas las funciones con los patrones, se diseña un algoritmo para la detección precisa y computacionalmente eficiente de estos módulos (de esta manera probamos nuestra hipótesis de los patrones RME y su descubrimiento de los módulos). Se utiliza el algoritmo Winnow para la construcción de redes y con el método de significación algorítmica, eliminamos la necesidad de costosas pruebas de permutación. Más adelante, mostramos que este algoritmo permite el uso de conjuntos de datos muy grandes y evaluamos los tipos de módulos que se pueden descubrir utilizando datos que se generarán en grandes proyectos de caracterización de tumores. Validamos nuestro método aplicándolo a un conjunto de datos y el algoritmo identifica módulos conocidos de las vías principales de GBM, amplía estos módulos con nuevos miembros y descubre nuevos módulos que pueden informar estudios futuros. Los datos se obtuvieron del Portal de datos del Atlas del genoma del cáncer (http://tcga-data.nci.nih.gov/). En la figura 1, se muestra el esquema de cómo es el proceso para nuestro trabajo

Metodología

Cómo nos estamos basando en el proceso que se realizó en un artículo, explicaremos las etapas que tuvieron que hacer para llegar a los resultados. De igual manera daremos la metodología de nuestro proceso. A grandes rasgos, lo que haremos será crear una matriz de mutación que pueda soportar polimorfismos de un solo nucleótido (SNP por su nombre en inglés, Single Nucleotide Polymorphisms) y alteraciones de número de copia (CNA por el nombre en inglés, Copy-Number Alterations). Después, con un algoritmo Winnow analizamos la matriz de mutaciones, lo cual nos da un “puntaje” para cada par de genes por exclusividad, que están indicados por los puntajes de las aristas de una gráfica. Finalmente, buscamos en la gráfica módulos de hasta un tamaño determinado y calculamos la significancia algorítmica para cada módulo potencial, pero solo nos quedamos con los que

tengan mayor significancia. Veamos todos estos pasos desglosados para ver cuáles son los que tomaremos de Miller et al. y cuáles haremos nosotros.

Crear una matriz de mutación

Para empezar, los autores crearon un algoritmo capaz de analizar información diferente, es decir, el algoritmo creado puede usar información mutacional dispar, entre estos tenemos polimorfismos de un solo nucleótido, alteraciones del número de copias y modificaciones epigenéticas. Todos estos tipos de información diferentes los convirtieron en una matriz de mutación bidimensional.

Para empezar, tomaremos datos de la página de The Cancer Genome Atlas Data Portal. Utilizaremos información sobre melanomas. Los autores utilizaron datos de este instituto, donde filtraron los datos que tenían para únicamente mostrar mutaciones no sinónimas y válidas.

En este paso se filtran para obtener solo las mutaciones válidas y no sinónimas, los umbrales de relación logarítmica para la amplificación y la eliminación se establecieron en 1,5 desviaciones estándar de la intensidad media de la sonda. Estos se cruzan con picos de cambio recurrente en el número de copias identificados por el algoritmo, luego se eliminaron las variantes del número de copias y se seleccionaron los genes impulsores, de esta forma nos arroja una matriz mutaciona

La matriz de mutación se construyó de tal manera que cada gen de cada muestra fuera comparado contra los SNPs y los CNA, tal que si la muestra i tenía una mutación en el gen j, la entrada x_{ij} era igual a 1, mientras que si no tenía dicha mutación, esa entrada era igual a 0. Construiremos la matriz para muestras de datos de melanomas posteriormente.

Construir una red de genes con el algoritmo Winnow

Como indicamos anteriormente, vamos a emplear un algoritmo Winnow para filtrar los datos que tenemos en la matriz de mutación. Analizaremos sólo los genes que cumplan con cierto grado de recurrencia y posteriormente analizar el grado de exclusividad entre pares de genes, que se calcula como sigue:

$$\frac{\#\ de\ muestras\ donde\ exactamente\ un\ gen\ en\ el\ par\ está\ mutado}{\#\ de\ muestras\ donde\ al\ menos\ uno\ de\ los\ genes\ en\ el\ par\ está\ mutado}$$

De esta manera los autores, y ahora también nosotros, podemos crear una red donde cada nodo es un gen y el peso de las aristas que unen a los nodos es el grado de exclusividad entre dicho par de genes.

Claramente, este método de analizar la exclusividad entre genes se vuelve muy pesado conforme tenemos más genes. Aquí es donde entra el algoritmo Winnow, que nos ayuda a detectar señales de exclusividad entre genes, incluso cuando la red sea muy compleja y con muchas conexiones irrelevantes. Así, con la velocidad y certeza que nos provee el algoritmo, podemos filtrar las puntuaciones de salida antes de generar la red. Con esto, finalmente, construimos una red más pequeña y de mejor calidad que la que habríamos creado si únicamente consideramos la exclusividad entre pares de genes

Para entrenar el algoritmo, los autores tomaron un gen como clasificador y el resto de la matriz de mutación como datos para entrenarlo. Se llevaron a cabo múltiples ejecuciones del algoritmo Winnow: en la primera, se voltearon los bits de la matriz, de tal forma que calcularon qué tan bien cada aberración en el clasificador es predictivo de no-aberraciones en cada gen de la matriz. En la segunda ejecución, los bits que se voltearon fueron los del clasificador, lo cual les permitió calcular qué tan bien cada no-aberración del clasificador calculaba las aberraciones de cada gen en la matriz. Así obtuvieron pesos para dar a los vértices de la gráfica. Por último, eliminaron los pesos más bajos (para cada gen clasificador tomaron el segundo peso más alto y mantuvieron todas las aristas que tenían un peso mayor o igual al seleccionado).

Dado que el rango de pesos para cada ejecución estuvo determinado por la rapidez con la que Winnow encuentra un clasificador óptimo, no utilizamos un valor de umbral absoluto al eliminar las aristas. En cambio, para cada gen clasificador, tomamos el segundo peso más alto y conservamos todas las aristas con una puntuación mayor o igual a ese valor.

Nota: La puntuación de exclusividad está definida como el número de muestras en las que exactamente uno del par está mutado dividido por el número de muestras en las que al menos uno del par está mutado.

Nota: Si solo hiciéramos la red sin el algoritmo de Winnow, la desventaja es que las redes rápidamente se vuelven demasiado grandes y densamente conectadas, lo que genera un enorme problema a la hora de identificar subredes de manera efectiva

Identificando módulos candidatos

Para esta etapa lo que hicimos fue emplear cada nodo de la red "filtrada" que se creó con un algoritmo voraz para encontrar los módulos RME y así poder evaluar todos los conectados y tienen un tamaño menor que cierto valor elegido. Solo tomaremos los módulos que tienen significancia algorítmica. La significancia algorítmica nos da un límite superior en la probabilidad de que la similaridad entre secuencias haya ocurrido por casualidad (4).

Posteriormente, tomamos los módulos RME que fueron buenos candidatos y los agrupamos por número de genes y ordenados por su valor de significancia. De estos módulos, tomamos el que tiene mayor número de genes y valor de significancia y descartamos todos los otros módulos que tengan los mismos genes. Repetimos esto hasta haber pasado por todos los grupos de genes.

Evaluación de los módulos haciendo una prueba de significancia algorítmica

Para determinar si un módulo tiene un patrón RME significativo podemos emplear modelos probabilísticos y heurísticos. Para facilitar el cálculo, que puede volverse computacionalmente muy pesado, no vamos a emplear ninguno de esos dos métodos, sino una prueba de significancia algorítmica.

El proceso que seguimos para esta prueba es el siguiente:
Sea k el número de muestras, m el número de genes en un módulo y X una matriz de tamaño kxm de valores únicamente binarios que representan presencia o ausencia (si hay presencia, el valor $x_{ij} = 1$; si tenemos ausencia, $x_{ij} = 0$) de aberraciones en el j-ésimo gen de la i-ésima muestra. Lo que hace la prueba de significancia algorítmica es comparar el número de bits que se utilizan para codificar la matriz binaria creada con el algoritmo RME con el

número de bits necesarios para codificar la matriz bajo la hipótesis nula. En otras palabras, el algoritmo RME intenta codificar los datos en menos bits utilizando la premisa de que las mutaciones ocurren con una frecuencia inusualmente alta de forma mutuamente excluyente, a diferencia que con la hipótesis nula, que asume que las aberraciones ocurren de forma independiente en sus frecuencias de fondo.

Si existe un patrón RME el algoritmo RME va a poder codificar la matriz X de forma más concisa. Para minimizar la longitud total de codificación, al algoritmo RME le dan la identidad de los m genes del total de n genes analizados. Así obtenemos un conteo de aberraciones para cada muestra, dado por a_{i0} con $i = 1, \dots, k$ y un conteo de aberraciones para cada gen, dado por b_{0j} con $j = 1, \dots, m$. Con estos datos, el algoritmo RME primero organiza las muestras, después los genes por su conteo de aberraciones (en la parte superior de la matriz pone las muestras con mayor número de alteraciones, y al inicio de cada fila pone los genes con mayor número de aberraciones).

Ahora lo que hace el algoritmo es analizar la matriz ordenada fila por fila, de izquierda a derecha, cuenta cuántas aberraciones hay y calcula la probabilidad de encontrar una aberración en la próxima célula de la matriz y codifica al bit de forma óptima según la probabilidad calculada.

La probabilidad la calcularemos de la siguiente manera y con la siguiente notación:

- $p(x_{ij} = 1)$ es el número de mutaciones sin observar dividido entre el número de posiciones sin observar que quedan en la matriz
- a_{ij} es el número de aberraciones sin observar en el gen en el que nos encontramos
- b_{ij} es el número de aberraciones sin observar en la muestra en la que nos encontramos

Dada esta notación podemos codificar elementos de la matriz X acorde con la siguiente distribución probabilística: si a_{ij} y b_{ij} son ambos mayores a 0, y no tenemos un 1 en esta fila aún, empleamos la siguiente fórmula derivada de Bayes:

$$p_{RME}(x_{ij} = 1) \approx p(x_{ij} = 1 | a_{ij}, b_{ij}) = \frac{p(x_{ij}=1|a_{ij}) \cdot p(x_{ij}=1|b_{ij})}{p(x_{ij}=1) \cdot \left(\frac{p(x_{ij}=0|a_{ij}) \cdot p(x_{ij}=0|b_{ij})}{p(x_{ij}=0)} + \frac{p(x_{ij}=1|a_{ij}) \cdot p(x_{ij}=1|b_{ij})}{p(x_{ij}=1)}\right)}$$

Si lo anterior no es el caso, los autores estimaron que la probabilidad sería casi 0, entonces tenemos que $p_{RME}(x_{ij} = 1) \approx 0$

Por otro lado, la hipótesis nula, que nos da el algoritmo nulo, codifica de forma óptima con el supuesto de que k genes contienen aberraciones con una frecuencia de fondo, que denotaremos $p_{NULL}(1)$, y que las mutaciones ocurren de forma independiente en cada uno de los k genes.

Veamos cómo se calcula la diferencia de codificación en ambos algoritmos (nulo y RME) y cómo se calcula la significancia algorítmica.

Codificar la matriz binaria de aberraciones

Definiremos una variable, d', que es donde vamos a almacenar la longitud de codificación Examinamos la matriz de aberraciones fila por fila, de izquierda a derecha. El valor de d' va a incrementar conforme avanzamos en la matriz. Así, tenemos las siguientes reglas:

- Si $x_{ij} = 1$ entonces

  $d' \leftarrow d' + (-log(p_{NULL}(1)) + log(1 - p_{RME}(1))$

- En cualquier otro caso, entonces

  $d' \leftarrow d' + (-log(p_{NULL}(0)) + log(1 - p_{RME}(1))$

En ambas reglas log denota el logaritmo binario

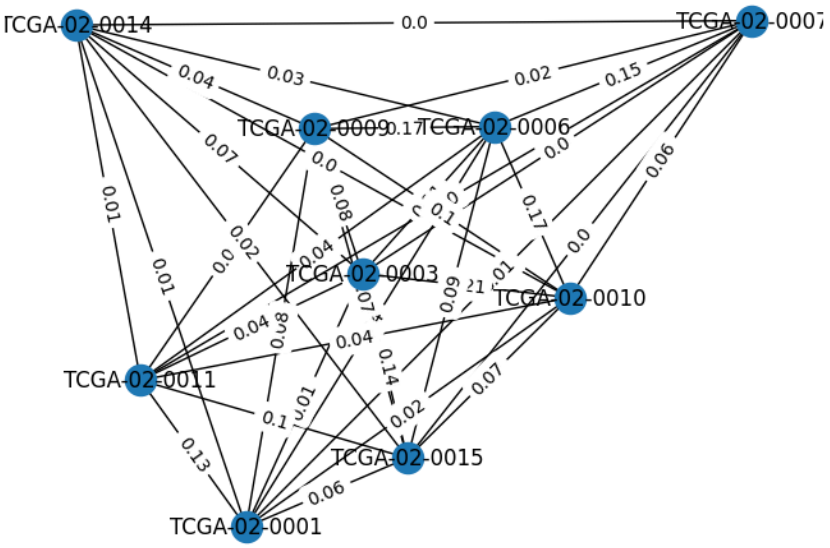
Para calcular la significancia usamos la siguiente fórmula:

- $d = d' - m\,log(n) - k\,log^{*}(m) - m\,log^{*}(k)$

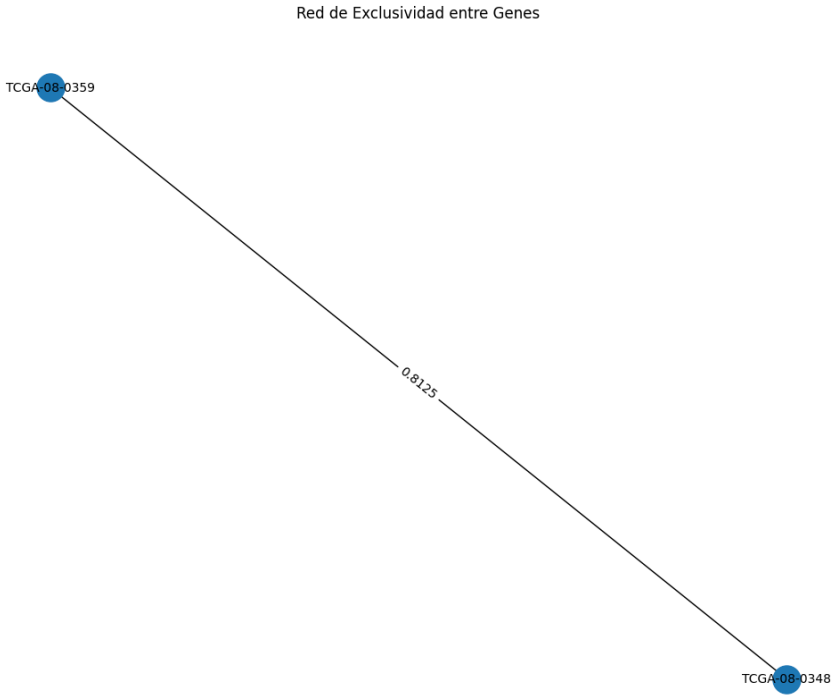
Nota: El enfoque de modelos probabilísticos o de heurísticas generalmente requerirán el establecimiento de valores de significancia extremadamente bajos (corrección previa a Bonferroni), lo que lleva a muchos ciclos de pruebas de permutación. Para eliminar este cuello de botella, se usa prueba de significancia algorítmica mucho menos exigente

Resultados

- Construimos la red de genes sin aplicar el algoritmo de Winnow. Este análisis puede ayudar a identificar genes que están en rutas mutualmente exclusivas, lo que puede sugerir que las mutaciones en estos genes están proporcionando ventajas similares a las células tumorales y que una mutación en uno puede hacer innecesaria una mutación en el otro.

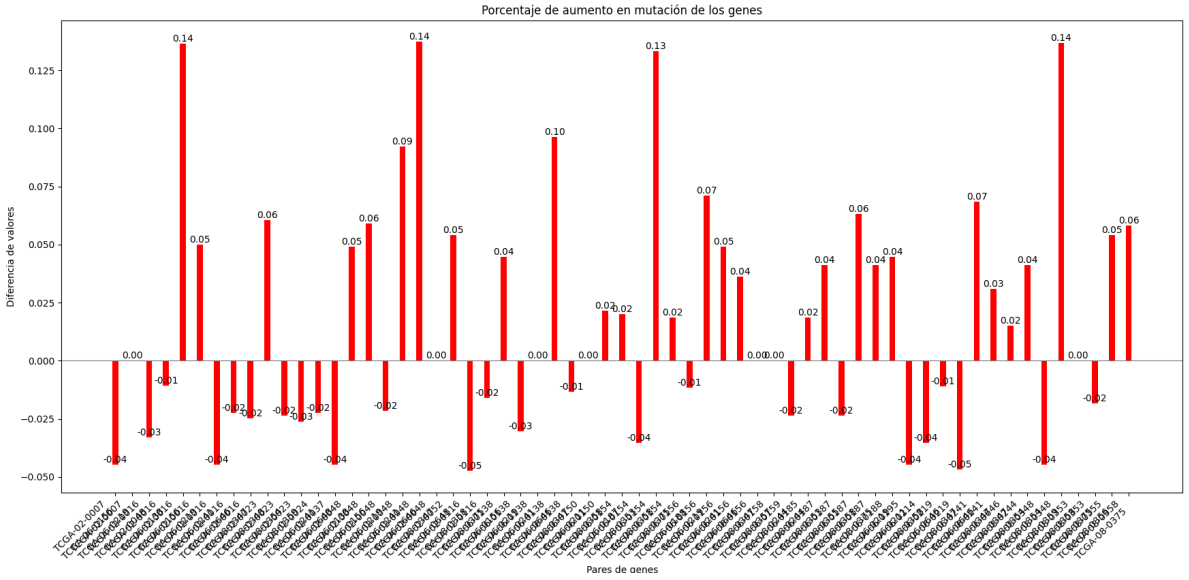


- Utilizamos el algoritmo Winnow para construir una red de genes a partir de los datos de mutaciones obtenidos de melanomas, con el objetivo de identificar patrones de aberraciones recurrentes y mutuamente excluyentes (RME).



-
- Identificamos un número de patrones RME en tumores de melanoma. Estos patrones proporcionaron información significativa acerca de los mecanismos moleculares subyacentes al cáncer. Al identificar estos patrones, pudimos entender más sobre los genes que son importantes para el desarrollo y progresión del cáncer.
- Logramos observar que las mutaciones en los genes de los módulos muy bajo en el crecimiento de su mutación y la supervivencia de las células del melanoma.

- Como podemos ver en la gráfica se muestra el nivel de mutación en cada gen y analizamos que cada gen tiende a mutar con un valor menor del 0.001 respecto a su valor obtenido del puntaje de exclusividad.
- Recordemos que los módulos que encontramos en esta parte, son las relaciones de las posibles mutaciones de cada gen, además de que si existe una relación entre dos genes es muy probable que el segundo gen tienda a mutar a partir del primero.



Discusión

Gracias a este proyecto pudimos conocer uno de los proyectos que analizan, identifican y trabajan con tumores que lidian con un gran volumen de datos sobre aberraciones genómicas, epigenómicas y de expresión génica, lo que transforma nuestro concepto sobre la biología del cáncer.

Conclusión

Con este análisis logramos:

- Descubrir patrones de exclusividad que pueden indicar interacciones genéticas específicas en el cáncer.
- Comparar la efectividad de este método con técnicas previas.
- Comparar los genes que posiblemente muten y las condiciones, así como analizar el cambio que se generan en éstas.
- El algoritmo previo permite analizar diversas formas de cáncer siempre y cuando se presenten en matrices de mutaciones proporcionadas por instituciones de confianza.

Glosario

cáncer: enfermedad en la que las células mutan, razón por la cual crecen de forma descontrolada. Estas células pueden moverse e invadir tejidos diferentes que no son en los que iniciaron

aberraciones somáticas: alteración del ADN que sucede después de la concepción. Se pueden presentar en cualquier célula del cuerpo, a excepción de las células germinativas. Una aberración en uno de los genes puede resultar en el desarrollo de un fenotipo tumorigénico clave, eliminando la presión selectiva para la mutación de los demás.

módulos funcionales: grupo de genes o sus productos que están relacionados por una o más interacciones genéticas o celulares. La función de un módulo es separable de la de otros y sus miembros tienen más relación entre ellos que con miembros de otros módulos. Ejemplos de esto son la co-regulación, co-expresión o afiliación de un complejo de proteínas o camino de señalización de un agregado celular.

tumores: Masa anormal de tejido que aparece cuando las células se multiplican más de lo debido o no se mueren cuando deberían. Los tumores son benignos (no cancerosos) o malignos (cancerosos). Las masas benignas a veces crecen mucho pero no se diseminan y tampoco invaden los tejidos cercanos ni otras partes del cuerpo. Las masas malignas suelen diseminarse o invadir los tejidos cercanos, y también es posible que se diseminen a otras partes del cuerpo a través de la sangre y el sistema linfático. También se llama neoplasia.

glioblastoma: Tipo de tumor del sistema nervioso central de crecimiento rápido que se forma a partir del tejido glial (de sostén) del encéfalo y la médula espinal; tiene células cuyo aspecto es muy diferente al de las células normales.

mutación genética: Una mutación es el cambio al azar en la secuencia de nucleótidos o en la organización del ADN o ARN de un ser vivo que produce una variación en las características de este y que no necesariamente se transmite a la descendencia.

módulo funcional: Se refiere a la manera en la que una mutación afecta la función de un gen o una proteína. Esto incluye la comprensión de cómo una alteración en la secuencia de ADN puede cambiar la estructura y función de los productos génicos.

Bibliografía

1. Navin, N.E. Cancer genomics: one cell at a time. *Genome Biol* **15**, 452 (2014). https://doi.org/10.1186/s13059-014-0452-9
2. Miller, C.A., Settle, S.H., Sulman, E.P. *et al.* Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics* **4**, 34 (2011). https://doi.org/10.1186/1755-8794-4-34
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068 (2008). https://doi.org/10.1038/nature07385
4. Milosavljević A. Discovering sequence similarity by the algorithmic significance method. Proc Int Conf Intell Syst Mol Biol. 1993;1:284-91. PMID: 7584347.
5. *Diccionario de Cáncer del NCI*. Comprehensive Cancer Information - NCI. (n.d.). https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/tumor
6. Bioinformatics Pipeline: DNA-Seq Analysis - GDC Docs. (n.d.). https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/#cnv-from-wgs-file-format