**HOW CAN WE ANALYZE CANCER?**

El cáncer está causado por alteraciones en el genoma y evoluciona como un proceso Darwiniano. En este ambiente, cada uno tiene una fitness y los mejor adaptados son los que sobreviven y pasan a la siguiente generación.

Durante la vida acumulamos mutaciones por causas endógenas o exógenas. Sucede en cada célula de forma independiente y a veces alguna aumenta la capacidad de proliferación de las células. Esto es lo que llaman drivers y se puede entender tanto a nivel de gen como de mutación. El cáncer se da por una acumulación sucesiva de drivers. Podemos aplicar este conocimiento para desarrollar tratamientos, a la hora de la prevención, diagnosis, tratamiento o prognosis (habrá una mejor o peor expectativa según la mutación).

Para identificar drivers, necesitamos secuenciar los tumores y comparar el DNA del tumor con el de referencia. Donde no concuerdan es donde está la mutación. También necesitamos una muestra de sangre del paciente para saber cuáles mutaciones son específicas de la persona y no del tumor.

Los tumores tienen de cientos a cientos de miles de mutaciones. La variabilidad dentro de un mismo tipo de cáncer es muy grande y solamente una fracción muy pequeña de estas mutaciones son drivers. Se calcula que se necesitan de 4 a 8 drivers para hacer un tumor maligno.

Tenemos que secuenciar a diferentes cohortes de pacientes de cáncer (con un tipo de tumor o varios). Entonces cogemos las mutaciones somáticas (específicas del tumor) para identificarlas y encontrar los patrones que deja la selección positiva en el genoma:

* **Recurrencia:** un gen se predice como driver si tiene más mutaciones de las esperadas por azar en todo el gen.
* **Impacto funcional:** acumulación de mutaciones que tienen un impacto funcional alto: que truncan la proteína, que afectan al centro activo...
* **Clustering:** tener mutaciones acumuladas en una región concreta de la secuencia. Por ejemplo, en las kinasas estos clusters tienden a aparecer en zonas con más impacto funcional (centro activo). Se puede entender a nivel de secuencia nucleotídica o de proteína. Se encuentran mutaciones en una región porque estas están teniendo un impacto selectivo positivo.

Necesitamos diferentes señales de selección positiva para tener una lista completa de drivers, ya que no todos los genes tienen las mismas señales de selección positiva. Por ejemplo, genes supresores como p53 no tienen tanto clustering porque lo que importa es truncar la proteína, entonces se encuentran menos mutaciones pero con alto impacto funcional. Por eso se necesitan diferentes métodos basados en diferentes señales de selección positiva.

Las mutaciones se acumulan dependiendo de muchos factores (regiones que se replican más tarde, regiones que están más cerradas en la cromatina…) y es muy difícil para una misma muestra tener todos estos datos. Por suerte, las tasas de mutación también dependen de la secuencia y sabemos que se acumulan teniendo en cuenta el contexto de cada nucleótido. Es decir, depende de los nucleótidos que están upstream y downstream. Se puede calcular la frecuencia de que una posición concreta del genoma mute teniendo en cuenta la frecuencia de mutación en este contexto en la cohorte en la que están.

**IntOGen**

Lo que hemos explicado hasta ahora es la pipeline de IntOGen. Los datos que se han utilizado corresponden a 27.000 muestras de diferentes tipos de cáncer, como mama, colon, próstata...

Se han juntado 6 métodos diferentes, dos son de recurrencia, tres de clustering y uno de impacto funcional. Cada uno propone una lista de drivers con un p-value y después se tienen que combinar. Se les da diferentes pesos al valor de cada resultado según lo bien que detectan unos genes que sabemos que están involucrados en cáncer. Por ejemplo, si el que mejor detecta es OncodriveFML, este es el que va a tener más peso al final. Al combinar los resultados vemos que la cohorte funciona mejor que utilizando los métodos individuales, por tanto son complementarios.

Una cosa es la cohorte y el otro es el tipo de cáncer. Con esta pipeline se han podido descubrir nuevos genes driver. Algunos que eran conocidos en un tipo de cáncer, se ha encontrado en otro tipo de cáncer. Esto es importante porque hay fármacos para un tipo de cáncer que puede que se puedan aplicar para otro tipo, también.

**OncodriveCLUSTL**

El algoritmo sólo es capaz de analizar variantes nucleotídicas de regiones codificantes o no codificantes. Primero coge la secuencia, mapea las mutaciones y busca donde están los clusters. Mira cuántas mutaciones hay, como de separadas están y les da un valor. Como más mutaciones y más juntas más alto es el score. Después se suman los scores de cada cluster y tenemos el score general de la secuencia.

Entonces tenemos que comparar con el clustering por azar, que viene dado por las frecuencias cada tipo de mutación en la cohorte. Por cada mutación hay una probabilidad de mutar y randomizamos la mutación para que caiga en un sitio aleatorio. Esto se repite 1000 veces, para cada una se mira cómo se hacen los clusters y se les da un score. Por último, miramos si nuestro cluster es significativo o no.

**QQplot:** esperamos que los valores observados se distribuyan como lo esperado menos los que tengan señales de clustering y sean significativos. En negrita están los genes dentro de la lista CGC (los que sabemos que están involucrados en cáncer).

Para ver si éste método funciona mejor que otros se comparó cómo funcionaba este método y el antiguo del laboratorio. Se cogieron los dos y los aplicaron sobre la misma cohorte para ver cómo de bien se detectan CGC. A más alto es el área bajo de la curva, mejor detecta los genes. Hay el método 1 y 2 y la lista de genes priorizada por cada cohorte, miramos en el CGC i si esta el primer gen pondremos un 100%, si esta este pero el segundo no un 50%, y lo hacemos para el top 40, entonces se hace una línea y lo vemos.

Para complementar podemos hacer clustering en 3D donde se tiene en cuenta la estructura de la proteína. Se trata de ver cuántas mutaciones se detectan y si ocurren en clusters 3D. Se vuelven a correr las cohortes con los dos métodos y vemos que genes detectan ambos. En azul están los que solamente detecta hotmaps, en blanco son los que detectan los dos y en rojo los que no pueden detectar. Éstos últimos serían los casos en que no se conoce la estructura o la proteína está truncada porque no hay proteína.

Podemos ver que así por clustering se pueden detectar algunos que los otros no pueden. Por eso es importante tener diferentes métodos para detectarlo.

**Cancer genome interpreter**

Podemos interpretar genomas específicos con esta información, llega la muestra del paciente y necesitamos saber qué mutaciones tienen drivers y si existe alguna terapia que se pueda usar para este paciente con estas mutaciones. Se tienen que mapear las mutaciones, dónde están; en qué genes y si están catalogados como drivers. Una vez sabemos que las mutaciones están en genes driver, se mira si se pueden tratar (tratamiento in silico). También puedes encontrar alguna terapia que se usa en otro cáncer pero donde hay la misma mutación que en el paciente y por lo tanto podamos usarla.