Tarea 6.1

José Miguel Amaro Estrada

Instituto de Ecología, UNAM

Posgrado en Ciencias Biológicas, UNAM

**Introducción a la bioinformática e investigación reproducible para análisis genómicos, semestre 2020-2**

Tarea 6.3

Recalibración y llamado de variantes

**MÉTODO**

Se realizó una secuenciación masiva para pacientes de los que se sospechaba una mutación patogénica. Esta se realizó con un equipo MiSeq de Illumina. La secuenciación incluyó un panel de 28 genes, los cuales se enlistan en la tabla 1. El total de todas las regiones blanco representan 91,120 bp.

Tabla 1. Lista de genes incluidos en la secuenciación de las muestras

|  |  |
| --- | --- |
| *Gen (abr)* | Gen (nombre extendido) |
| *MPL* | MPL proto-oncogene, thrombopoietin receptor |
| *SF3B1* | splicing factor 3b subunit 1 |
| *PDGFRA* | platelet derived growth factor receptor alpha |
| *KIT* | KIT proto-oncogene, receptor tyrosine kinase |
| *PDGFRB* | platelet derived growth factor receptor beta |
| *IKZF1* | IKAROS family zinc finger 1 |
| *BRAF* | B-Raf proto-oncogene, serine/threonine kinase |
| *EZH2* | enhancer of zeste 2 polycomb repressive complex 2 subunit |
| *IL7* | interleukin 7 |
| *JAK2* | Janus kinase 2 |
| *PAX5* | paired box 5 |
| *ABL1* | ABL proto-oncogene 1, non-receptor tyrosine kinase |
| *KLLN* | killin, p53 regulated DNA replication inhibitor |
| *PTEN* | phosphatase and tensin homolog |
| *WT1* | WT1 transcription factor |
| *KMT2A* | lysine methyltransferase 2ª |
| *CBL* | Cbl proto-oncogene |
| *KRAS* | KRAS proto-oncogene |
| *FLT3* | fms related receptor tyrosine kinase 3 |
| *BRCA2* | BRCA2 DNA repair associated |
| *RB1* | RB transcriptional corepressor 1 |
| *TP53* | tumor protein p53 |
| *BRCA1* | BRCA1 DNA repair associated |
| *CALR* | Calreticulin |
| *JAK3* | Janus kinase 3 |
| *CEBPA-DT* | CEBPA divergent transcript |
| *CRLF2* | cytokine receptor like factor 2 |
| *P2RY8* | P2Y receptor family member 8 |

**Calibración de la calidad**

Se utilizó el archivo bam de la muestra 19, el cual contiene el alineamiento generado con las lecturas filtradas, con respecto al genoma de referencia h19. En primer lugar, se generó una tabla de calibración utilizando como referencia de sitios conocidos la base de datos dbSNP\_hg19. Posteriormente se aplicó esta calibración al alineamiento bam inicial para obtener un alineamiento recalibrado. Para este paso se utilizó el programa BaseRecalibrator y Print Reads contenido en GATK.

**Llamado de variantes**

Se utilizó HaplotypeCaller para llamar las variantes del alineamiento re-calibrado. Posteriormente se utilizó SelectVariants para filtrar las variantes bajo dos criterios. Primero, solo aquellas variantes con una profundidad mayor que 10. Segundo, solo aquellas variantes con una calidad superior a 30. Estos filtros se aplicaron de manera separada par dos tipos de variantes: SNP’s e indels. Una vez filtrados ambos tipos de variantes, se fusionaron los dos archivos VCF para tener todas las variantes en un solo archivo. Para efectos de este ejercicio, solamente se realizó el llamado de variantes del cromosoma 19.

Finalmente, se realizó una anotación de los SNP, para lo cual se utilizó el programa snpEff para anotar los tipos de variantes recuperadas y predecir sus efectos, tales como un cambio en el sentido de la transcripción. Adicionalmente, se usó el servidor SNPnNexus para realizar una anotación adicional (

**RESULTADOS**

El archivo VCF inicial, con el llamado de variantes previo al filtrado, contuvo 9 variantes para el cromosoma 19, de las cuales una fue filtrada durante el llamado inicial, es decir, se llamó un total de 8 variantes. Después de aplicar ambos filtros, se retuvieron estas mismas variantes, ya que estas variantes tuvieron calidades de 99 y profundidades superiores a 100 (tabla 2). Todas las variantes fueron de tipo SNP, y no se observó ningún indel.

Tabla 2. Variantes retenidas después de los dos filtros aplicados. Las ocho variantes iniciales pasaron todos los filtros.

|  |  |  |
| --- | --- | --- |
| Variante | Profundidad | Calidad |
| rs1049481 | 128 | 99 |
| rs2302600 | 156 | 99 |
| rs2302601 | 153 | 99 |
| rs2302603 | 121 | 99 |
| rs1122385 | 113 | 99 |
| rs3212741 | 200 | 99 |
| rs3212733 | 59 | 99 |
| rs3212730 | 105 | 99 |
| rs34529039 | filtrada | filtrada |

De las ocho variantes llamadas para el cromosoma 19, 7 cayeron dentro la región correspondiente al gen JAK3, sin embargo, todas estas variantes corresponden a regiones intrónicas del gen, mientras que la variante restante, corresponde al exón 9 del gen CALR.

Captura de pantalla de un celular

Descripción generada automáticamente

Figura 1. Dos de las siete variantes mapeadas en el gen JAK3, correspondientes a una región intrónica..

Según la anotación hecha con snpEff, la mayoría de las variantes son de tipo intrónicas, mientras que el 19% corresponde a mutaciones río arriba y el 14%, río abajo. Solo el 4% de las variantes corresponden a exones (tabla 3).

Tabla 3. Número de efectos por tipo y región de las variantes

|  |  |  |
| --- | --- | --- |
| Tipo | Conteo | Porcentaje |
| Río abajo | 9 | 14.51% |
| Exón | 3 | 4.83% |
| Intrón | 35 | 56.45% |
| Transcrito | 1 | 1.61% |
| Río arriba | 12 | 19.35% |
| UTR 3 prime | 2 | 3.22% |

**CONCLUSIONES**

De acuerdo con los resultados, 7 SNP corresponden al gen *JAK3,* el cual pertenece a una familia de tirosin-cinasas, involucradas en la transducción de señales, principalmente en células del sistema inmune, y para las cuales, se existen mutaciones asociadas con el desarrollo de la enfermedad autoinmune SCID (NCBI, 2020). En la figura 1, se muestran dos de estas mutaciones que ocurren cercanas una de la otra. Sin embargo, al caer todos estos SNP en regiones intrónicas, es poco probable que estos sean causantes directos de alguna patología, aunque podrían estar asociados a algún SNP en una región codificante del gen que sí sea patogénico. Por ejemplo, la anotación con SNPNexus reveló asociaciones de dos de estos SNP con esclerosis lateral amiotrófica.

Por otro lado, el SNP correspondiente al gen CALR sí corresponde a un exón, por lo tanto, puede implicar algún cambio funcional. Según la anotación de snpEff, algunos de los SNP causan un cambio en codón, alguno de los cuales puede corresponder con esta variante. Algunas mutaciones en este gen están relacionadas con afecciones mieloproliferativas, y están asociadas con efectos en genes de la familia JAK, lo cual podría asociar estas variantes con las encontradas en JAK3 (Nangalia et al., 2019).

Debido al tipo de SNP’s encontrados, en posible que el paciente exhiba síntomas relacionados con alguna patología neorodegenerativa o de cáncer asociado al sistema nervioso. Sin embargo, haría falta evaluar los SNP’s en las demás regiones blanco, ya que en este estudio solo se consideraron aquellas correspondientes al cromosoma 19.

**FUENTES**

Nangalia J, Massie CE, Baxter EJ, Nice FL, Gundem G, Wedge DC, Avezov E, Li J, Kollmann K, Kent DG, Aziz A, Godfrey AL, Hinton J, Martincorena I, Van Loo P, Jones AV, Guglielmelli P, Tarpey P, Harding HP, Fitzpatrick JD, Goudie CT, Ortmann CA, Loughran SJ, Raine K, Jones DR, Butler AP, Teague JW, O'Meara S, McLaren S, Bianchi M, Silber Y, Dimitropoulou D, Bloxham D, Mudie L, Maddison M, Robinson B, Keohane C, Maclean C, Hill K, Orchard K, Tauro S, Du MQ, Greaves M, Bowen D, Huntly BJ, Harrison CN, Cross NC, Ron D, Vannucchi AM, Papaemmanuil E, Campbell PJ, Green AR (Dec 2013). "Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2". The New England Journal of Medicine. 369 (25): 2391–405. doi:10.1056/NEJMoa1312542