Universidad San Jorge

Facultad de Ciencias de la Salud

Grado en Bioinformática

Proyecto Final

Alteraciones metabólicas y mutaciones: descubriendo biomarcadores en esquizofrenia

Autora: Andrea Roque Sola

Directores: Francisco José Roig Molina, Francesc Josep Montoro Salvador

Universidad San Jorge, 15 de junio de 2025

Este trabajo constituye parte de mi candidatura para la obtención del título de Graduado o Graduada en Bioinformática por la Universidad San Jorge y no ha sido entregado previamente (o simultáneamente) para la obtención de cualquier otro título.

Este documento es el resultado de mi propio trabajo, excepto donde de otra manera esté indicado y referido.

Doy mi consentimiento para que se archive este trabajo en la biblioteca universitaria de Universidad San Jorge, donde se puede facilitar su consulta.

Fecha y Firma

Zaragoza, a 15 de junio de 2025.

Fdo: Dª **Andrea Roque Sola**
DNI.: **73613376J**

Dedicatoria y agradecimiento

A mi madre, por su apoyo y confianza incondicional, su esfuerzo sin medida, y por enseñarme que rendirse no es una opción. Este logro es más tuyo que mío.

A mi padre, que partió al comenzar este camino, pero que me sigue acompañando en cada paso.

A Molly, por enseñarme lo que es la amistad y sentarse a mi lado día tras día, en silencio pero siempre presente. Te fuiste poco antes del final, pero este logro también lleva tu nombre.

A Emma, por enseñarme el valor de las cosas pequeñas, y hacer más ligero el camino con esas cuatro patitas esperándome al volver a casa.

A Gisela, Pepe, y Fer, por estar en las buenas, pero más aún en las malas.

A mi abuela, por su confianza y amor constante. Gracias.

A Pedro, por no haber dejado de confiar en mí ni un solo día, y haber sabido sostenerme incluso cuando ni yo sabía cómo hacerlo.

A Fran, por enseñarme a confiar en mí y a disfrutar del proceso. Gracias por estar ahí con una paciencia y una dedicación infinitas.

Y, por último, a todo el profesorado del Grado en Bioinformática y del Grado en Farmacia, por ser una fuente constante de inspiración. Ha sido un privilegio aprender de los mejores.

Índice de Contenido

Índice de Ilustraciones

Índice de Tablas

Resumen

La esquizofrenia es un trastorno psiquiátrico crónico y complejo que generalmente se diagnostica en fases avanzadas, cuando los pacientes ya han experimentado brotes psicóticos significativos. Este diagnóstico tardío limita las oportunidades de intervención temprana y aumenta la carga clínica y social de la enfermedad. Actualmente, no existen biomarcadores específicos que permitan la identificación precoz de la esquizofrenia, lo que resalta la necesidad de enfoques moleculares innovadores.

Este estudio está centrado en identificar posibles biomarcadores tempranos de esquizofrenia mediante análisis transcriptómico. Para ello, se analizaron secuencias de RNA extraídas de sangre periférica de pacientes diagnosticados con esquizofrenia y de individuos sanos como grupo control. El procesamiento de los datos se llevó a cabo mediante tres *pipelines* bioinformáticas específicas: *HISAT-GATK/PICARD-SnpEff* para el llamado de variantes, *HISAT-HTSeq count-EdgeR* para el análisis de expresión diferencial de genes, y *HISAT-StringTie* para el ensamblaje de transcriptomas y la cuantificación de la expresión génica.

El análisis transcriptómico permitió identificar una expresión diferencial significativa en 39 genes al comparar las muestras de pacientes con esquizofrenia con las de los controles sanos. Además, se detectaron 16 variantes genéticas exclusivas del grupo de pacientes, ausentes en el grupo control.

Estos resultados sugieren que la expresión diferencial de estos genes y la presencia de variantes genéticas específicas podrían estar asociadas con la predisposición a la esquizofrenia, proporcionando información valiosa para el desarrollo de estrategias diagnósticas tempranas.

Palabras clave: Biomarcadores, cuantificación de la expresión, expresión diferencial, esquizofrenia, llamado de variantes, transcriptómica.

Abstract

Schizophrenia is a chronic and complex psychiatric disorder that is generally diagnosed in advanced stages, when patients have already experienced significant psychotic episodes. This late diagnosis limits opportunities for early intervention and increases the clinical and social burden of the disease. Currently, there are no specific biomarkers that allow for the early identification of schizophrenia, highlighting the need for innovative molecular approaches.

This study is focused on identifying possible early biomarkers of schizophrenia by transcriptomic analysis. For this purpose, RNA sequences extracted from peripheral blood of patients diagnosed with schizophrenia and from healthy individuals as a control group were analyzed. Data processing was performed using three specific bioinformatics pipelines: *HISAT-GATK/PICARD-SnpEff* for variant calling, *HISAT-HTSeq count-EdgeR* for differential gene expression analysis, and *HISAT-StringTie* for transcriptome assembly and gene expression quantification.

Transcriptomic analysis identified significant differential expression in 39 genes when comparing samples from schizophrenia patients with those from healthy controls. In addition, 16 genetic variants unique to the patient group, absent in the control group, were detected.

These results suggest that differential expression of these genes and the presence of specific genetic variants could be associated with predisposition to schizophrenia, providing valuable information for the development of early diagnostic strategies.

Keywords: Biomarkers, expression quantification, differential expression, schizophrenia, variant calling, transcriptomics.

1. Introducción

La esquizofrenia es un trastorno psiquiátrico grave que afecta aproximadamente al 1% de la población mundial y forma parte de las principales causas de discapacidad a nivel global (1). Se caracteriza por una combinación de síntomas positivos (alucinaciones y delirios), negativos (apatía y anhedonia), y cognitivos (deterioro de la memoria de trabajo y dificultades en la toma de decisiones)(2).

A pesar de los avances en el conocimiento de la enfermedad, su etiología exacta sigue siendo desconocida, lo que dificulta el desarrollo de estrategias terapéuticas y diagnósticas más eficaces. No obstante, se considera que su origen involucra una combinación de predisposición genética, factores ambientales, y elementos neurobiológicos, lo que conduce a déficits cognitivos y síntomas psicóticos con diversos grados de manifestaciones positivas, negativas y desorganización. Generalmente, el trastorno se presenta en la adultez temprana, con una edad pico de aparición de 20,5 años y una edad mediana de inicio de 25 años (3).

El diagnóstico de la esquizofrenia sigue dependiendo principalmente de criterios clínicos establecidos en el *Diagnostic Statistical Manual of Mental Disorders* (DSM-V) y en la *International Classification of Diseases* (ICD-11) (4). Sin embargo, la ausencia de biomarcadores objetivos ha impulsado numerosas investigaciones en busca de perfiles moleculares que permitan una detección más temprana y precisa.

El Programa Internacional de Seguridad Química, dirigido por la Organización Mundial de la Salud (OMS) define un biomarcador como "cualquier sustancia, estructura o proceso que pueda medirse en el organismo o en sus productos y que influya o prediga la incidencia de una enfermedad o su resultado" (5).

Los biomarcadores relacionados con la esquizofrenia abarcan alteraciones bioquímicas detectables provocadas por el estrés (como el incremento de la carga alostérica), problemas en el funcionamiento de las mitocondrias, inflamación del sistema nervioso, y desequilibrios por estrés oxidativo y nitrosativo, además de trastornos en el ritmo circadiano (6).

Desde un punto de vista genético, la esquizofrenia es altamente hereditaria, con estimaciones que indican que aproximadamente el 80% del riesgo de padecer la enfermedad es atribuible a factores genéticos. Los estudios de asociación de todo el genoma (*Genome-Wide Association Studies*, GWAS) han identificado múltiples *loci* asociados significativamente con la enfermedad, los cuales contienen genes involucrados en la neurotransmisión, la función sináptica, y la regulación del desarrollo neuronal (7).

Sin embargo, a pesar de estos avances, los efectos individuales de las variantes identificadas suelen ser pequeños y no explican por completo la compleja etiología de la esquizofrenia. Esta

limitación ha impulsado la búsqueda de variantes genéticas con un mayor impacto funcional mediante enfoques más profundos.

En este sentido, se han descubierto deleciones en la región 22q11.2 y mutaciones en genes como *SETD1A*, que desempeña un papel esencial en la mitosis y la proliferación celular (8), lo que ha llevado a una exploración más profunda mediante enfoques como la secuenciación del exoma y el genoma completo, los cuales han revelado mutaciones raras y estructurales con un impacto funcional más pronunciado.

Además, los estudios epigenéticos han mostrado que factores ambientales pueden modular la expresión génica a través de mecanismos como la metilación del ADN y la modificación de histonas. De la misma forma, el papel como biomarcador diagnóstico del RNA no codificante, incluido el micro-RNA (miRNA), ha sido ampliamente estudiado, así como la aceleración de la edad epigenética y acortamiento de los telómeros (9).

Dado que la esquizofrenia es un trastorno complejo con alteraciones a múltiples niveles biológicos, el análisis transcriptómico, y en particular la secuenciación de RNA (RNA-seq), se ha consolidado como una herramienta clave para caracterizar los cambios en la expresión génica asociados con la enfermedad (10). Esta aproximación permite examinar la actividad génica en tejidos relevantes, como el cerebro o la sangre, lo que facilita la identificación de rutas metabólicas alteradas y permite la identificación de biomarcadores implicados en la fisiopatología del trastorno.

Por ejemplo, la activación crónica del eje hipotálamo-hipófisis-adrenal (HHA) se refleja en la expresión diferencial de genes relacionados con el cortisol y la respuesta al estrés, por lo que cambios en genes como *NR3C1* (receptor de glucocorticoides) pueden influir en la susceptibilidad de la esquizofrenia, aunque esto sólo se ha conseguido relacionar en mujeres (11). Asimismo, el análisis transcriptómico permite detectar disfunciones mitocondriales mediante la identificación de patrones anómalos de expresión en genes involucrados en la cadena de transporte de electrones y producción de ATP, como *NDUFS7* o *COX4I1*, proceso relacionado con la etiología de la esquizofrenia (12). Del mismo modo, se ha observado que genes reguladores del ritmo circadiano, como *CLOCK*, *ARNTL* o *PER2*, pueden presentar expresión diferencial en pacientes con esquizofrenia, lo que explica las alteraciones del sueño y los desajustes hormonales característicos del trastorno (13).

Para facilitar el acceso y análisis de datos transcriptómicos por parte de la comunidad científica, repositorios como el *Sequence Read Archive* (SRA) de GenBank (14) proporcionan almacenamiento y acceso a secuencias de RNA de diversos estudios. SRA es una base de datos pública que contiene secuencias obtenidas mediante RNA-seq, lo que permite la validación y el análisis de datos en distintos contextos. A través de SRA, se pueden explorar perfiles de expresión

génica asociados con la esquizofrenia y comparar datos entre diferentes estudios, contribuyendo al entendimiento de los mecanismos moleculares subyacentes a la enfermedad.

La expresión diferencial de genes en la esquizofrenia no solo permite comprender los mecanismos biológicos subyacentes, sino que también aporta herramientas para el diagnóstico y el desarrollo de tratamientos personalizados. Comparando perfiles transcriptómicos entre individuos con esquizofrenia y controles sanos, se pueden identificar biomarcadores específicos de la enfermedad.

Uno de los enfoques más utilizados es la identificación de firmas génicas en sangre periférica, lo que facilita el desarrollo de pruebas diagnósticas menos invasivas. Por ejemplo, la expresión alterada de genes proinflamatorios como *IL6*, *TNF-α,* y *CCL2* sugiere una posible disfunción inmunológica en la esquizofrenia, lo que respalda la hipótesis neuroinflamatoria de la enfermedad (15). Asimismo, cambios en la expresión de genes relacionados con la integridad de la barrera hematoencefálica, como *OCLN* o *CLDN5*, pueden indicar una disrupción en la comunicación neurovascular (16).

El análisis de la expresión diferencial también permite subclasificar a los pacientes según su perfil molecular. Se ha visto que ciertos subgrupos de pacientes con esquizofrenia presentan una regulación anómala en genes sinápticos como *SYN1* y *DLG4*, lo que sugiere una afectación en la plasticidad neuronal y la transmisión sináptica (17). Esta estratificación podría ayudar en la selección de tratamientos dirigidos a corregir déficits específicos.

En términos de intervención terapéutica, la expresión diferencial puede orientar el uso de fármacos según la firma transcriptómica del paciente. Se ha demostrado que algunos antipsicóticos pueden modular la expresión génica, normalizando en parte los patrones alterados en la esquizofrenia (18). Sin embargo, además de los cambios en la expresión génica, la variabilidad en la secuencia del DNA de cada individuo también juega un papel clave en la susceptibilidad y evolución de la enfermedad.

Esta heterogeneidad clínica y biológica de la esquizofrenia representa un reto importante para el desarrollo de terapias eficaces, pero también una oportunidad para la identificación de subtipos moleculares mediante técnicas transcriptómicas, lo que podría facilitar una medicina más personalizada y basada en perfiles biológicos específicos.

Una de las estrategias más utilizadas para estudiar estas variaciones genéticas es la identificación de polimorfismos de un solo nucleótido (SNPs, por sus siglas en inglés) mediante técnicas de *variant calling*, que generalmente se aplican a datos genómicos, aunque también pueden utilizarse con datos transcriptómicos (19). Los SNPs son cambios en una sola base del ADN que pueden afectar la función génica y están ampliamente distribuidos en el genoma humano. En el contexto de la esquizofrenia, el análisis de SNPs permite no solo identificar

variantes asociadas con la presencia de la enfermedad, sino también con su gravedad, respuesta a tratamientos y progresión clínica.

El *variant calling* a partir de datos transcriptómicos (RNA *variant calling*) permite detectar variantes genéticas expresadas en el RNA, lo que significa que estas mutaciones tienen un impacto funcional directo en la célula. A diferencia del análisis de variantes a nivel del DNA genómico, el RNA *variant calling* ofrece la ventaja de identificar mutaciones que afectan genes activos en tejidos específicos, como el cerebro o la sangre, lo que resulta especialmente útil en enfermedades neuropsiquiátricas como la esquizofrenia.

Entre los genes más estudiados mediante *variant calling* en transcriptómica se encuentran *DISC1* y *CACNA1C*, ambos implicados en funciones clave del sistema nervioso. En el caso de *DISC1*, se han descrito múltiples variantes funcionales, incluyendo mutaciones como Q31L, L100P, D453G y R264Q, asociadas con fenotipos psiquiátricos en humanos y modelos animales (20). Por otro lado, *CACNA1C*, gen implicado en la regulación del calcio intracelular y la excitabilidad neuronal, presenta variantes como rs2283274 y rs2239061, que se relacionan con disfunción autonómica cardíaca en pacientes con esquizofrenia (21). La detección de variantes expresadas en estos genes permite comprender mejor su impacto funcional y su contribución a la patología de la esquizofrenia.

El uso de técnicas de RNA-seq en combinación con *variant calling* representa un enfoque poderoso para identificar biomarcadores genéticos que podrían servir para la estratificación de pacientes y el diseño de estrategias terapéuticas personalizadas. Así, la integración del análisis de SNPs con la transcriptómica permite avanzar en la comprensión de los mecanismos moleculares subyacentes a la esquizofrenia y su variabilidad clínica (22).

En este sentido, para abordar de forma integral el análisis transcriptómico en esquizofrenia, se han implementado diferentes *pipelines* bioinformáticas que permiten tanto la cuantificación de la expresión génica como la identificación de variantes genéticas expresadas. Un ejemplo sería el análisis de la expresión diferencial basada en la secuencia de herramientas *HISAT– StringTie–Ballgown*. Esta *pipeline* permite obtener perfiles transcriptómicos detallados a partir de datos de RNA-seq. Las lecturas se alinean al genoma de referencia utilizando *HISAT*, un alineador eficiente y sensible para lecturas genómicas y transcriptómicas (19). Los archivos de alineamiento generados se procesan con *Samtools*, y la cuantificación génica se realiza con *HTSeq-count*. A continuación, *StringTie* permite ensamblar transcritos y estimar su abundancia, mientras que *Ballgown* se encarga del análisis estadístico y la visualización de los datos de expresión diferencial (23).

Este flujo de trabajo ha demostrado ser especialmente útil para identificar genes y transcritos diferencialmente expresados, lo que permite detectar biomarcadores transcriptómicos y vías

moleculares alteradas. Además, ha sido validada y ampliamente empleada en estudios transcriptómicos debido a su alta precisión tanto en el ensamblaje de transcritos como en la cuantificación de su expresión, incluso en genes con múltiples isoformas, aspecto clave para reflejar la complejidad del transcriptoma en enfermedades multifactoriales como la esquizofrenia (23).

De la misma forma, la combinación *TOPHAT–GATK/PICARD*, también permite la identificación de SNPs y pequeñas inserciones o deleciones (indels) expresadas en los tejidos analizados. El flujo de trabajo incluye el alineamiento de las lecturas con *TOPHAT*, una herramienta especializada en detectar sitios de empalme. Luego, los archivos de alineamiento son analizados con *PICARD* para la marcación de duplicados y la preparación de los datos. Finalmente, *GATK* se encarga de la detección de variantes, permitiendo identificar mutaciones que afectan genes funcionalmente activos (24). Esta aproximación resulta particularmente relevante en el estudio de enfermedades neuropsiquiátricas como la esquizofrenia, ya que las variantes expresadas pueden tener un impacto directo sobre la función celular y contribuir a la comprensión de la fisiopatología de la enfermedad, así como a la estratificación de pacientes y al diseño de tratamientos personalizados.

Esta última estrategia está basada en las buenas prácticas recomendadas por el equipo de desarrollo del *Genome Analysis Toolkit* (GATK), lo que garantiza la obtención de llamadas de variantes de alta confianza y calidad (24), especialmente importantes en estudios de biomarcadores funcionales expresados en tejidos relevantes como el cerebro.

Si bien la transcriptómica de célula única (scRNA-seq) ha revolucionado el estudio de la heterogeneidad celular al permitir la identificación de subpoblaciones celulares específicas, su aplicación en estudios de esquizofrenia presenta limitaciones importantes. En primer lugar, el acceso a muestras cerebrales frescas humanas, necesarias para una separación celular eficaz, es restringido y generalmente se basa en tejido *post mortem*, lo que puede afectar la calidad del RNA y la interpretación de los resultados (25). En segundo lugar, la preparación celular para scRNA-seq puede inducir sesgos técnicos debido al estrés mecánico o enzimático durante la disociación celular, lo cual puede alterar los perfiles de expresión génica (26). Por último, el coste elevado y la complejidad bioinformática de esta técnica pueden limitar su aplicabilidad en estudios de mayor escala.

Por tanto, el análisis transcriptómico de datos de RNA-seq en muestras de pacientes con esquizofrenia, podrían ser muy útiles para identificar firmas génicas y variantes expresadas que puedan actuar como biomarcadores diagnósticos o predictivos, y contribuir a una mejor comprensión de los mecanismos moleculares de la enfermedad.

La estructura de este estudio sigue el flujo natural de análisis de los datos de RNA-Seq. En primer lugar, se realiza una fase de control de calidad y preprocesamiento de los datos, incluyendo filtrado con *PRINSEQ* y evaluación con *FastQC*. A continuación, las lecturas son alineadas al genoma de referencia mediante *HISAT2*, punto de partida para el análisis de la expresión diferencial con *HTSeq-count* y *EdgeR*, de *StringTie* para la cuantificación de transcritos, y de *GATK-Picard* y *SnpEff* para el análisis de variantes. Esta organización metodológica permite explorar distintos niveles de información molecular a partir de los mismos datos, ofreciendo una visión más completa de los procesos biológicos implicados en la esquizofrenia.

2. Antecedentes

A pesar de los avances significativos en la comprensión genética y molecular de la esquizofrenia, persisten múltiples desafíos que limitan la traducción clínica de estos hallazgos. En primer lugar, a pesar de que con *GWAS* se han identificado más de 100 *loci* asociados al riesgo, estos explican sólo una fracción de la heredabilidad estimada del trastorno, y muchas de las variantes descubiertas se ubican en regiones no codificantes del genoma, lo que dificulta su interpretación funcional (27). Además, la mayoría de estos estudios no tienen en cuenta el contexto tisular ni el estado transcripcional de las variantes, dejando abierta la pregunta sobre su impacto real a nivel celular.

Frente a esta limitación, la transcriptómica ha ofrecido un camino alternativo para estudiar las consecuencias funcionales de las alteraciones genéticas. Sin embargo, los estudios transcriptómicos realizados hasta la fecha en esquizofrenia presentan una notable heterogeneidad metodológica y clínica, lo que dificulta la comparación entre cohortes y la replicabilidad de los hallazgos. Diferencias en los tipos de muestra (tejido cerebral *post mortem* vs. sangre periférica), protocolos de extracción, plataformas tecnológicas y *pipelines* bioinformáticas pueden introducir variabilidad técnica significativa (28). Además, las propias características clínicas de la esquizofrenia (presentación heterogénea, evolución variable y respuesta diferencial a tratamientos) añaden una capa de complejidad biológica que los enfoques tradicionales han tenido dificultades en capturar (29).

Por otro lado, aunque el RNA-seq ha sido ampliamente utilizado para cuantificar la expresión génica, muchos estudios transcriptómicos en esquizofrenia no integran de forma sistemática el análisis de variantes expresadas, a pesar de que estas pueden tener implicaciones funcionales directas en la célula. La combinación de expresión génica y análisis de variantes expresadas permite una aproximación más completa, capaz de identificar no solo qué genes están desregulados, sino también qué mutaciones específicas podrían estar modulando esa desregulación (30).

En este contexto, el uso de *pipelines* robustas y estandarizadas como *HISAT–StringTie–Ballgown* (23) para expresión diferencial y *TopHat–GATK/PICARD* (24) para detección de variantes, representa una estrategia metodológica consolidada, ya que permite integrar información cuantitativa y cualitativa del transcriptoma, facilitando la identificación de rutas moleculares alteradas y variantes funcionales expresadas, especialmente en tejidos periféricos como la sangre, que podrían servir como biomarcadores accesibles.

Finalmente, aunque la tecnología de célula única (scRNA-seq) ha revelado con gran resolución la diversidad celular en el cerebro humano (25), su uso en esquizofrenia sigue siendo

limitado por razones técnicas, económicas, y éticas, especialmente debido a la necesidad de muestras cerebrales *post mortem* o de difícil obtención mediante biopsias.

Por tanto, el análisis transcriptómico sigue siendo una herramienta válida y poderosa, especialmente si se desarrolla bajo criterios metodológicos estrictos y se complementa con técnicas de análisis de variantes. Esta estrategia puede ayudar a superar algunas de las barreras que han dificultado la identificación de biomarcadores confiables y avanzar hacia una clasificación molecular más precisa del trastorno.

3. Objetivos

1. Objetivo primario: detectar biomarcadores asociados a la esquizofrenia que contribuyan al avance en el diagnóstico temprano y preciso de la patología.

2. Objetivos secundarios:
   - Detectar SNPs asociados con la esquizofrenia, que puedan ser utilizados como biomarcadores para la diagnosis temprana y prognosis de la enfermedad.
   - Evaluar los patrones de expresión diferencial en pacientes con esquizofrenia en comparación con individuos sanos, para identificar posibles marcadores de expresión que puedan ser usados en el diagnóstico de la enfermedad y entender la evolución de esta.

4. Metodología

4.1 Obtención de secuencias

Las secuencias de RNA utilizadas en este estudio fueron obtenidas a través de SRA (14), con el código de identificación SRP499627. Consisten en lecturas *paired-end* de longitud media de 149 pares de bases y un tamaño medio de fragmento de 9400 pares de bases. Para su análisis, se empleó el genoma referencia GRCh38.p14, utilizado tanto para el mapeo como para la anotación transcriptómica.

Dichas secuencias fueron obtenidas en formato FASTQ, formato estándar utilizado para almacenar secuencias producidas por tecnologías NGS (*Next Generation Sequencing*). Este formato está basado en el antiguo formato FASTA, diferenciándose en que ya no sólo incluye las secuencias, sino que también añade los valores de calidad que indican la fiabilidad de cada nucleótido. Como el formato FASTA, FASTQ empieza con una línea de encabezado, pero en el caso de FASTQ, esta línea se identifica con @, en lugar del símbolo ">" utilizado en FASTA (31).

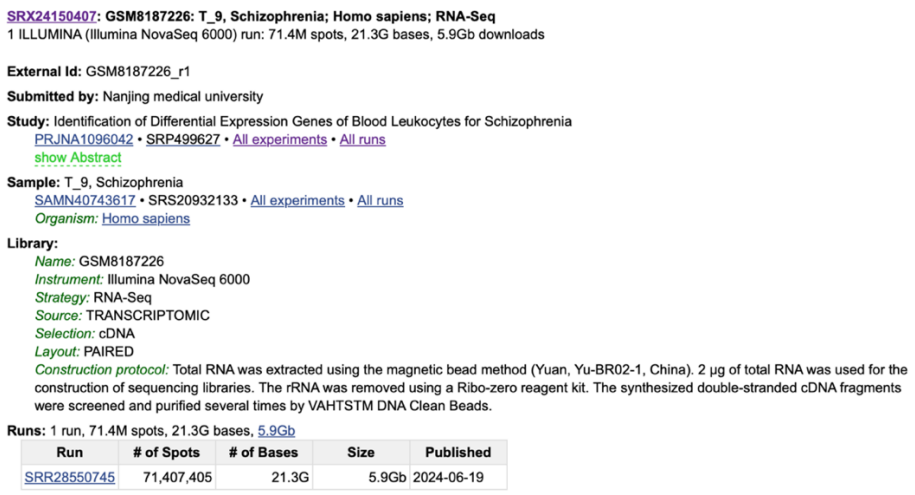
Para obtener cada una de las secuencias utilizadas, los siguientes comandos fueron ejecutados. El ejemplo mostrado es el de la secuencia SRR28550745.

prefetch SRR28550745

fastq-dump --split-files SRR28550745

En total se analizaron 18 secuencias, de las cuales nueve provenían de pacientes con esquizofrenia, identificadas con los códigos SRR28550745 - SRR28550753 (*Ilustración 1*), y nueve correspondientes a individuos control, con los identificadores SRR28550736 - SRR28550744 (*Ilustración 2*).

Para el desarrollo y ejecución de los códigos implementados en este proyecto, se utilizó el servidor Comba de la Universidad San Jorge (USJ).



SRX24150407: GSM8187226: T_9, Schizophrenia; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina NovaSeq 6000) run: 71.4M spots, 21.3G bases, 5.9Gb downloads

External Id: GSM8187226_r1

Submitted by: Nanjing medical university

Study: Identification of Differential Expression Genes of Blood Leukocytes for Schizophrenia
PRJNA1096042 • SRP499627 • All experiments • All runs
show Abstract

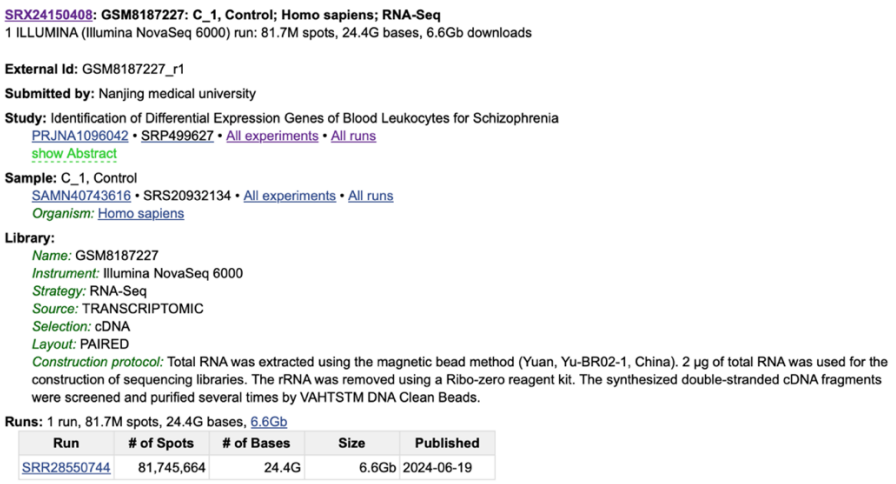
Sample: T_9, Schizophrenia
SAMN40743617 • SRS20932133 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: GSM8187226
Instrument: Illumina NovaSeq 6000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED
Construction protocol: Total RNA was extracted using the magnetic bead method (Yuan, Yu-BR02-1, China). 2 μg of total RNA was used for the construction of sequencing libraries. The rRNA was removed using a Ribo-zero reagent kit. The synthesized double-stranded cDNA fragments were screened and purified several times by VAHTSTM DNA Clean Beads.

Runs: 1 run, 71.4M spots, 21.3G bases, 5.9Gb

Run	# of Spots	# of Bases	Size	Published
SRR28550745	71,407,405	21.3G	5.9Gb	2024-06-19

Ilustración 1. Secuencia SRR28550745 (esquizofrenia)



SRX24150408: GSM8187227: C_1, Control; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina NovaSeq 6000) run: 81.7M spots, 24.4G bases, 6.6Gb downloads

External Id: GSM8187227_r1

Submitted by: Nanjing medical university

Study: Identification of Differential Expression Genes of Blood Leukocytes for Schizophrenia
PRJNA1096042 • SRP499627 • All experiments • All runs
show Abstract

Sample: C_1, Control
SAMN40743616 • SRS20932134 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: GSM8187227
Instrument: Illumina NovaSeq 6000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED
Construction protocol: Total RNA was extracted using the magnetic bead method (Yuan, Yu-BR02-1, China). 2 μg of total RNA was used for the construction of sequencing libraries. The rRNA was removed using a Ribo-zero reagent kit. The synthesized double-stranded cDNA fragments were screened and purified several times by VAHTSTM DNA Clean Beads.

Runs: 1 run, 81.7M spots, 24.4G bases, 6.6Gb

Run	# of Spots	# of Bases	Size	Published
SRR28550744	81,745,664	24.4G	6.6Gb	2024-06-19

Ilustración 2. Secuencia SRR28550744 (control)

4.2 Control de Calidad

El análisis de calidad se llevó a cabo usando *FASTQC*. Este paso es muy importante dado que el conjunto de parámetros usados para el preprocesado depende de la calidad de la lectura. *FASTQC* ofrece un conjunto modular de análisis que proporciona una descripción general de los datos, ayudando a identificar posibles problemas antes de continuar con el análisis.

Las principales características de *FASTQC* incluyen (32):

- Acepta múltiples formatos de archivo, incluyendo BAM, SAM, y todas las variantes de FastQ.
- Rápida evaluación de los datos, resaltando áreas donde podrían existir problemas.
- Resumen de gráficos y tablas para una evaluación fácil y rápida de la calidad de los datos.
- Informes exportables en formato HTML.
- Funcionamiento *offline*, permitiendo generación automatizada de informes, sin requerir la aplicación interactiva.

4.3 Preprocesado

El preprocesamiento de las secuencias resulta fundamental para asegurar la calidad de los datos de secuenciación, incluso en ausencia de secuencias adaptadoras, como ocurre en este estudio. Este paso se llevó a cabo con *Prinseq*, permitiendo filtrar lecturas de baja calidad, recortar bases deficientes y eliminar secuencias cortas que podrían no proporcionar información fiable. Al establecer umbrales de calidad y filtros de longitud, ayuda a eliminar datos ruidosos o erróneos, garantizando que únicamente se utilicen secuencias precisas y de alta calidad para los análisis posteriores.

4.4 Mapeo

El mapeo implica alinear las lecturas contra una secuencia de referencia conocida y construir una secuencia que sea similar pero no necesariamente idéntica a la referencia. Es una etapa fundamental en el análisis de datos de secuenciación de RNA, ya que permite ubicar las lecturas obtenidas mediante secuenciación en una referencia genómica, con el objetivo de identificar las regiones del genoma de donde provienen.

4.5 Cuantificación de la expresión génica

Tras completar el proceso de mapeo de las lecturas mediante *HISAT2*, se procedió a la cuantificación de la expresión génica utilizando *HTSeq-count*, software diseñado para contar la cantidad de lecturas alineadas a cada gen, y constituye una etapa clave en los análisis de RNA-Seq, ya que permite obtener una matriz de expresión génica que será utilizada posteriormente en el análisis diferencial (33).

4.6 Ensamblaje y cuantificación de transcritos

Además de la cuantificación de la expresión génica, se llevó a cabo el ensamblaje y la cuantificación de transcritos utilizando la herramienta *StringTie* (23).

A diferencia de *HTSeq-count*, que se centra en la cuantificación de genes, *StringTie* permite ensamblar transcritos completos a partir de las lecturas mapeadas, facilitando la identificación de isoformas conocidas y potencialmente nuevas, así como la cuantificación de su nivel de expresión.

4.7 Análisis de Expresión Diferencial

Una vez obtenidos los niveles de expresión génica mediante el procesamiento de los archivos BAM con herramientas de cuantificación basadas en conteo, se procedió al análisis de expresión diferencial utilizando el paquete *EdgeR* del entorno R. Esta herramienta está específicamente diseñada para el análisis estadístico de datos de RNA-Seq basados en conteos crudos (*raw counts*) y emplea modelos lineales generalizados basados en la distribución binomial negativa, lo que permite una estimación precisa de la variabilidad biológica entre muestras (34).

4.8 Variant Calling

La llamada de variantes (*Variant Calling*) normalmente se realiza a partir de datos genómicos, pero su aplicación en RNA-Seq representa un valor añadido, ya que permite aprovechar al máximo

la información contenida en los datos transcriptómicos, facilitando la identificación de SNPs y pequeñas inserciones o deleciones expresadas muestras analizadas, lo que amplía el potencial para detectar biomarcadores relevantes en enfermedades como la esquizofrenia. Para llevar a cabo este paso, se siguieron las pautas recomendadas por el *Broad Institute* para datos de RNA-Seq, utilizando las herramientas *Picard* y *GATK* (*Genome Analysis Toolkit*). Dado que el servidor disponible contaba con Java 11 como versión instalada, se optó por emplear la versión 4.1.9.0 de *GATK*, que está compilada específicamente para esa versión de Java. Esta elección permitió resolver de forma eficiente los problemas de compatibilidad y asegurar el correcto funcionamiento del software durante el análisis.

Para la anotación de variantes obtenidas tras el llamado de variantes, se utilizó *SnpEff* (35), una herramienta ampliamente reconocida para la predicción de efectos de variantes genómicas. *SnpEff* permite clasificar las variantes según su impacto biológico y proporciona anotaciones detalladas sobre las consecuencias funcionales de SNPs y pequeñas inserciones/deleciones en regiones codificantes y no codificantes del genoma.

Las variantes se identificaron comparando las secuencias de las muestras con el genoma de referencia, como es habitual en los análisis de *variant calling*. Posteriormente, se realizó una comparación entre los grupos de pacientes y controles para detectar aquellas variantes (SNPs e *indels*) presentes únicamente en los pacientes, con el objetivo de identificar posibles biomarcadores específicos de la enfermedad.

5. Implementación y/o Desarrollo

La implementación del flujo de trabajo seguido en este estudio se resume en la *Ilustración 3*, la cual detalla las principales herramientas utilizadas en cada etapa del análisis, desde el control de calidad inicial hasta el análisis de variantes genéticas.

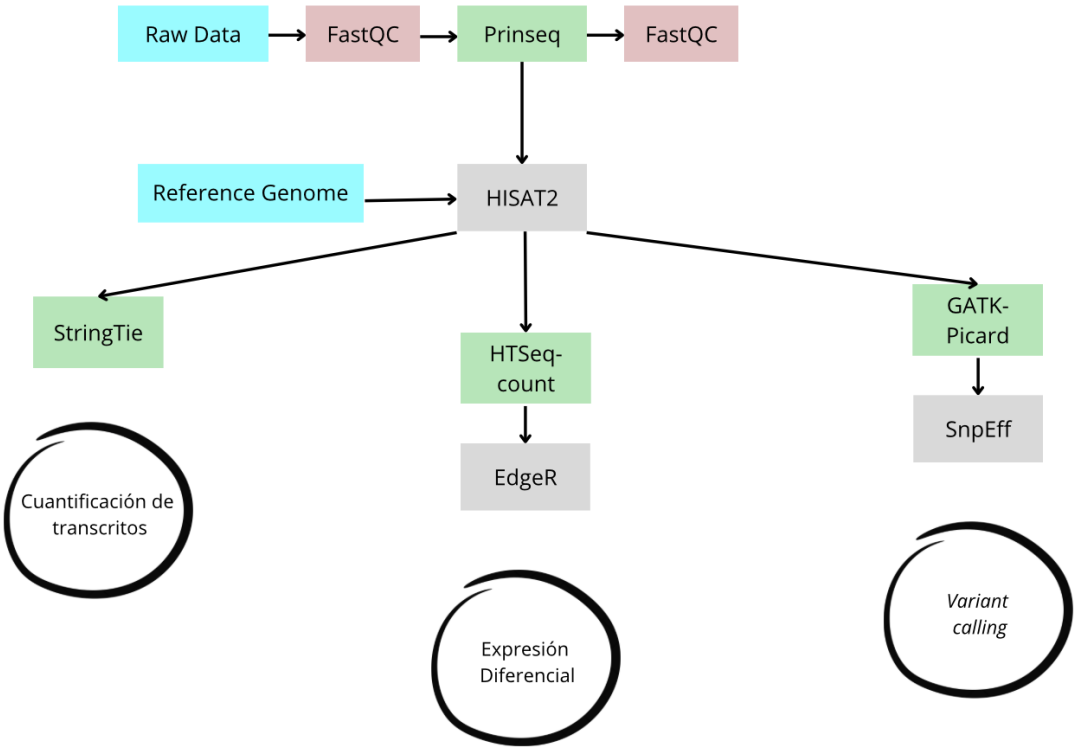


Ilustración 3. Flujo de trabajo implementado

Todas las secuencias obtenidas a través de SRA fueron sometidas a análisis de calidad con *FastQC*. Una vez verificada la calidad y confirmada su adecuación para el análisis, se procedió al preprocesamiento de las lecturas con *Prinseq*. Este paso se aplicó a cada par de secuencias, tanto de los grupos control como de los pacientes con esquizofrenia.

Inicialmente, se intentó realizar el mapeo de las lecturas con *TopHat*, un software ampliamente utilizado para el alineamiento de lecturas de RNA-Seq en genomas eucariotas. *TopHat* está basado en el alineador *Bowtie* y es capaz de identificar de manera eficiente sitios de empalme (*splice junctions*), lo que lo convierte en una herramienta útil para analizar la transcripción en organismos complejos. Sin embargo, durante su ejecución se presentaron errores técnicos que impidieron completar el análisis. En particular, se presentó un conflicto relacionado con la integración de *Picard*, una herramienta requerida en etapas posteriores del flujo de trabajo, que mostraba el error "*Unsupported major.minor version*", al ejecutarse con versiones de Java más recientes que las compatibles con la versión de *TopHat* integrada en el entorno.

Debido a estos inconvenientes, se optó por utilizar *HISAT2*, un alineador de segunda generación que también permite el mapeo de lecturas de RNA-Seq teniendo en cuenta los eventos de *splicing*. *HISAT2* está basado en una arquitectura de indexado más eficiente, lo que le permite realizar alineamientos más rápidos y con menor uso de memoria que *TopHat*. Además, *HISAT2* presenta una mayor compatibilidad con versiones actuales de sistemas operativos y herramientas bioinformáticas, lo que facilitó su implementación exitosa en este estudio.

Para realizar el alineamiento con *HISAT2*, se descargaron archivos de referencia del genoma humano (GRCh38.p14) desde el repositorio de NCBI (genoma, transcriptoma, y GFF).

Para realizar el mapeo con *HISAT2*, se intentó utilizar un índice incluyendo información de *splice sites* y exones derivados del archivo de anotación GFF. Esta estrategia suele recomendarse para mejorar la precisión del alineamiento en datos de RNA-Seq, ya que permite a *HISAT2* anticipar las uniones entre exones.

Sin embargo, durante la ejecución del comando de alineamiento, el programa generó el siguiente error:

Error reading _rstarts[] array: 21704, 23124

Error: Encountered internal HISAT2 exception (#1)

Command: /usr/bin/HISAT2-align-s --wrapper basic-0 -p 10 --dta -x /home/aroque/HISAT/GRCh38_HISAT_index -S /home/aroque/HISAT/SRR28550736.sam -1 /tmp/4140867.inpipe1 -2 /tmp/4140867.inpipe2

(ERR): HISAT2-align exited with value 1

Este error está relacionado con un fallo interno en la lectura del índice y puede deberse a una corrupción del archivo o incompatibilidades entre los archivos de entrada.

Debido a esta limitación, se optó por utilizar un índice genómico estándar, generado únicamente a partir del archivo FASTA del genoma, sin incluir información adicional de *splicing*. Si bien esta opción no proporciona el mismo nivel de sensibilidad en la detección de eventos de empalme, permitió ejecutar *HISAT2* con éxito y obtener alineamientos consistentes.

Una vez verificado el funcionamiento de *HISAT2* con una secuencia, se preparó un *script* para poder ejecutarlo en todas las secuencias del estudio (Anexo 2). Debido a las limitaciones de capacidad del servidor y para optimizar el uso de recursos, el procesamiento se dividió en dos *scripts* independientes: uno para las muestras de controles y otro para las de pacientes, permitiendo así una ejecución más eficiente y ordenada del mapeo de datos en paralelo.

Para llevar a cabo la cuantificación de la expresión génica con *HTSeq-count*, se utilizaron los archivos de alineamiento generados por *HISAT2* en formato SAM, junto con el archivo de

anotación del genoma humano en formato GFF (GCF_000001405.40_GRCh38.p14_genomic.gff.gz), previamente descargado desde NCBI. Es importante destacar que *HTSeq-count* requiere que el archivo de anotación esté correctamente formateado y que el tipo de atributo especificado (por ejemplo, gen o exón) sea coherente con la información que se desea cuantificar.

Durante la cuantificación con *HTSeq-count*, se encontró una dificultad relacionada con el archivo de anotación en formato GTF. Al intentar ejecutar el comando de verificación con la primera secuencia, se generó el siguiente error:

Error processing GFF file (line 56 of file /home/aroque/HISAT/GCF_000001405.40_GRCh38.p14_fixed.gtf):

Feature gene-SEPTIN14P18 does not contain a 'gene_id' attribute

[Exception type: ValueError, raised in features.py:337]

Este error se debía a que algunas anotaciones en el archivo GTF carecían del atributo obligatorio "gene_id", necesario para que HTSeq-count pueda asignar lecturas a genes de forma correcta. En concreto, ciertas líneas contenían únicamente el atributo "gene_name".

Para resolver este inconveniente, se utilizó una instrucción sed que añadió automáticamente un campo "gene_id" igual al "gene_name" en todas aquellas líneas correspondientes a exones que no lo tuvieran:

sed '/exon/ {/gene_id/! s/gene_name "\([^"]\)"/gene_name "\1"; gene_id "\1"/}' /home/aroque/HISAT/GCF_000001405.40_GRCh38.p14_genomic.gtf > /home/aroque/HISAT/GCF_000001405.40_GRCh38.p14_fixed.gtf*

Este comando modificó el archivo de anotación para que todas las entradas tipo exón incluyeran correctamente un campo "gene_id", permitiendo así que *HTSeq-count* pudiera procesar el archivo sin errores y completar la cuantificación. El *script* implementado se muestra en el Anexo 3.

Tras la obtención de los conteos con *HTSeq-count*, se utilizó *EdgeR* con el fin de analizar las diferencias de expresión entre pacientes y controles (Anexo 4). El primer paso consistió en la generación de una matriz de conteo (counts.tsv) unificando los archivos individuales de expresión por muestra generados con *HTSeq*, y se utilizó un archivo de metadatos (pheno_data.csv) para definir los grupos experimentales (control vs. paciente). Posteriormente, la normalización de las bibliotecas mediante el método TMM (*Trimmed Mean of M-values*), permitió aplicar filtros para eliminar genes con baja expresión y se estimaron las dispersiones comunes y específicas por gen.

A continuación, se realizó el test exacto de Fisher modificado, propio de *EdgeR*, para identificar genes diferencialmente expresados entre las condiciones experimentales. Los

resultados se exportaron en un archivo con el listado completo de genes (Ctrl_vs_Disease.tsv) incluyendo sus valores de $\log_2$ *fold change,* p-value, y FDR (*False Discovery Rate)*. Además, se generó un subconjunto de genes con una expresión diferencial significativa (|log2FC| ≥ 2,5), almacenado en el archivo genes_+-2,5FC.tsv.

Paralelo al análisis de la expresión diferencial con *HTSeq-count-EdgeR*, se llevó a cabo la cuantificación de transcritos con *StringTie* (Anexo 5), el cual trabaja directamente sobre los archivos de alineamiento en formato BAM, ordenados por coordenadas genómicas. El objetivo de esta etapa fue obtener una representación más detallada de los transcritos expresados en cada muestra, con el fin de explorar eventos de *splicing* alternativo, isoformas no anotadas, o diferencias en la estructura de los transcritos entre condiciones experimentales (pacientes vs. controles).

Asimismo, partiendo de los archivos BAM generados con *HISAT2*, se llevó a cabo el llamado de variantes (*variant calling*), para lo cual se implementó un flujo de trabajo estructurado en varias etapas consecutivas, permitiendo preparar adecuadamente los archivos de alineamiento, identificar variantes genéticas, y anotar su posible impacto funcional. Las etapas fueron las siguientes:

1. Preparación de archivos BAM

Previo al llamado de variantes, los archivos de alineamiento en formato BAM, generados por *HISAT2* y ordenados por coordenadas, fueron sometidos a un preprocesamiento con *Picard* para garantizar la calidad y compatibilidad de los datos con *GATK*. Dicho preprocesamiento incluyó:

- *AddOrReplaceReadGroups* (*Picard*). Este paso añade información sobre los grupos de lectura (*Read Groups*) a los archivos BAM. Los grupos de lectura son etiquetas necesarias para que *GATK* pueda distinguir entre diferentes muestras o experimentos, facilitando el manejo de metadatos en análisis posteriores. En este caso, se asignó a cada muestra un identificador único (RGID), así como otros campos obligatorios como la librería (RGLB), la plataforma de secuenciación (RGPL), la unidad de plataforma (RGPU) y la muestra (RGSM).

- *MarkDuplicates* (*Picard*). Las lecturas duplicadas, que pueden originarse durante el proceso de amplificación por PCR, fueron marcadas utilizando esta herramienta. La identificación de duplicados es importante para evitar que introduzcan sesgos en el análisis de variantes, ya que pueden aumentar de forma artificial la evidencia de ciertas posiciones del genoma. Los duplicados fueron marcados, no eliminados, y se generó un archivo de métricas que resume la cantidad de duplicados detectados.

2. Preparación específica para RNA-Seq

Debido a las características propias de las lecturas de RNA-Seq, que suelen abarcar múltiples exones separados por intrones, fue necesario un paso adicional antes del llamado de variantes:

- *SplitNCigarReads* (*GATK*). Esta herramienta divide las lecturas en las posiciones donde se encuentran intrones (indicadas por los caracteres "N" en los *CIGAR strings* del archivo BAM). Esto es crucial porque las lecturas de RNA-Seq abarcan uniones de exones, y *GATK* requiere un formato específico para poder interpretar correctamente estas regiones durante el llamado de variantes.

3. Llamado de variantes

Finalmente, se ejecutó el proceso de llamado de variantes utilizando:

- *HaplotypeCaller* (*GATK*). Esta herramienta permite identificar variantes en las lecturas alineadas, reconstruyendo haplotipos locales a partir de los datos. Se utilizó el genoma humano GRCh38.p14 como referencia, y el análisis se configuró en modo GVCF (-ERC GVCF), que genera un archivo intermedio en formato GVCF (*Genomic Variant Calling File*). Este formato permite combinar posteriormente las variantes de todas las muestras en un solo análisis conjunto, mejorando la precisión del llamado de variantes.

En este estudio se utilizó el genoma completo de referencia para el llamado de variantes, en lugar de una referencia transcriptómica. Esta elección se fundamenta en las ventajas analíticas que ofrece trabajar a nivel genómico.

Alinearse contra el genoma, y no solo contra una base de datos de transcritos, permite una detección más precisa de variantes que se encuentren no solo en regiones codificantes, sino también en regiones intrónicas, no codificantes, y reguladoras, las cuales pueden ser relevantes para la expresión génica y la patogénesis de enfermedades. Además, herramientas como *GATK* requieren un contexto genómico completo para ejecutar pasos clave como *SplitNCigarReads* y *HaplotypeCaller*, los cuales dependen de la estructura exón-intrón para interpretar adecuadamente las lecturas RNA-Seq, que abarcan empalmes.

Otra ventaja clave del uso del genoma es la compatibilidad con herramientas de anotación funcional como *SnpEff*, que permite clasificar variantes según su efecto y localización genómica (exones, intrones, regiones reguladoras, etc.). Esto sería limitado si solo se utilizara una referencia de transcritos, ya que muchas regiones potencialmente relevantes quedarían fuera del análisis.

En resumen, el uso del genoma de referencia garantiza un análisis más exhaustivo y flexible, especialmente cuando el objetivo incluye la detección e interpretación funcional de variantes genéticas, y no únicamente la cuantificación de expresión génica.

Todos los pasos fueron automatizados mediante un *script* (Anexo 6), diseñado para procesar cada muestra de forma independiente y guardar los resultados en carpetas específicas por muestra.

Después del análisis de los archivos BAM con herramientas de *Picard* y *GATK*, se procedió a ejecutar el módulo *GenotypeGVCFs* con el fin de obtener un archivo VCF final que contenga las variantes genotipadas por muestra (Anexo 7), ya que el archivo generado previamente por *HaplotypeCaller* en formato GVCF aún no representa un conjunto definitivo de variantes, sino que contiene información intermedia, incluyendo probabilidades de variantes y no variantes por posición. Por ello, se aplica *GenotypeGVCFs*, que transforma este GVCF en un archivo VCF estándar, convirtiendo las probabilidades de genotipo en llamadas concretas. Este paso garantiza que las variantes estén correctamente definidas y listas para su posterior análisis funcional, como la anotación con *Variant Effect Predictor* (VEP) o *SnpEff*.

Una vez obtenido el archivo VCF definitivo, se procedió a realizar la anotación funcional de las variantes utilizando *SnpEff* (Anexo 8), una herramienta ampliamente empleada para predecir los efectos de las variantes genéticas. Dado que las secuencias de referencia utilizadas corresponden a *scaffolds* de RefSeq fue necesario crear una base de datos personalizada en *SnpEff*, ya que la base de datos estándar no reconoce adecuadamente estos nombres de cromosomas no convencionales. *SnpEff* espera nombres de cromosomas estándar (como 1, 2, ..., X), por lo que, para asegurar una anotación precisa, se construyó una base adaptada a las características del ensamblaje utilizado.

Esta etapa fue crucial para clasificar las variantes según su posible impacto biológico (por ejemplo, sinónimas, no sinónimas, etc.) y asignarlas correctamente a regiones genómicas específicas como exones, intrones o elementos reguladores.

Posteriormente, para facilitar el manejo y análisis de los datos, se realizó la conversión de los archivos VCF anotados a formato TSV. Esta transformación se llevó a cabo utilizando *SnpSift*, lo que permitió extraer campos relevantes como cromosoma, posición, nucleótido de referencia y nucleótido alternativo, efecto, impacto, entre otros.

Finalmente, se identificaron los SNPs comunes a todos los pacientes mediante la intersección de los conjuntos de variantes no sinónimas de cada individuo de este grupo (Anexo 9). El archivo final snps_unicos_pacientes.tsv contiene estas variantes específicas, que representan potenciales marcadores diferenciales entre los grupos estudiados.

6. Estudio económico

El presente proyecto se desarrolla en la empresa *BIAMICS*, Servicios de Bioinformática S.L., especializada en el análisis bioinformático aplicado a la investigación biomédica, incluyendo el estudio de enfermedades neuropsiquiátricas, como la esquizofrenia. El objetivo principal es analizar datos de RNA-Seq para identificar biomarcadores genéticos y de expresión asociados a esta enfermedad.

El análisis se realiza empleando herramientas bioinformáticas *open-source*, sin necesidad de licencias comerciales. El trabajo es ejecutado por un único bioinformático, encargado de todas las etapas del análisis, desde el preprocesamiento de las secuencias hasta el llamado de variantes, análisis de expresión diferencial, y redacción del informe técnico final.

Aunque este proyecto se desarrolla en un contexto de infraestructura compartida, en un entorno empresarial real es importante tener en cuenta los gastos fijos mensuales de operación, que incluyen el espacio de trabajo y los suministros básicos. En la *Tabla 1* se presenta una estimación proporcional de estos gastos, calculados en base a la estructura de personal de *BIAMICS*, donde actualmente trabajan tres personas a tiempo completo. En este proyecto, uno de los trabajadores dedica 37,5 horas semanales (jornada completa) exclusivamente al análisis durante un mes, lo que implica que el proyecto ocupa aproximadamente un tercio de la capacidad operativa total de la empresa en ese periodo.

Tabla 1. Gastos fijos de la empresa

Concepto	**Frecuencia**	**Precio unitario (€)**	**Total (€)**
Alquiler de oficina	Mensual	600 €	200,00 €
Electricidad	Mensual	100 €	33,33 €
Internet y telefonía	Mensual	50 €	16,67 €
Agua	Mensual	30 €	10,00 €
TOTAL	-	-	**260,00 €**

Fuente: elaborado por la autora.

Además de los costes fijos, el proyecto requiere una serie de recursos directos y específicos para su ejecución. En la *Tabla 2* se detalla una estimación realista de los costes variables asociados.

El coste de personal se ha calculado estimando un total de 120 horas de trabajo dedicadas a las distintas fases del proyecto, a una tarifa de 20 €/hora, correspondiente a un perfil de bioinformático *junior*.

Tabla 2. Gastos específicos del proyecto

Concepto	**Cantidad/Tiempo**	**Precio unitario (€)**	**Total (€)**
Coste de personal	150 h	20 €/h	3000,00 €
Tutorización del investigador	15 h	40 €/h	600,00 €
Amortización del servidor	120 h	-	150,24 €
Almacenamiento de datos	1 TB	50 €	50,00 €
Soporte técnico	-	-	100,00 €
Material fungible informático	-	-	325,00 €
Housing	1 mes	120 €/mes	120,00 €
TOTAL	-	-	**4345,24 €**

Fuente: elaborado por la autora.

Se estima un gasto aproximado de 150,24€ de amortización vinculada al proyecto del servidor (amortización a 5 años). Este gasto se calcula en base al uso efectivo del servidor (120 horas sobre una vida útil de 9.375 h, con un coste estimado de 1,25 €/h).

Se contempla un gasto adicional de 100 € para cubrir posibles gastos administrativos o de soporte técnico relacionado con el servidor.

El material fungible informático hace referencia a los equipos y dispositivos utilizados directamente por el personal para la realización del análisis (*MacBook Air* 2019, ratón, adaptadores, etc.). Si bien estos no se consumen en un solo proyecto, su uso intensivo justifica la imputación proporcional del coste de amortización.

El *housing*, equivalente al uso compartido de infraestructura, energía y climatización, se estima en 120 € para este proyecto.

No existen costes asociados a licencias de software, ya que se utilizan exclusivamente herramientas *open-source* (*FASTQC, Prinseq, HISAT2, HTSeq-count, StringTie, Ballgown, GATK*, entre otras).

Respecto a los potenciales beneficios, a pesar de tratarse de un proyecto puntual, tiene un alto potencial de expansión comercial, ya que la empresa puede ofrecer análisis transcriptómicos

y de variantes para enfermedades neuropsiquiátricas, un área en expansión donde pocas empresas ofrecen estudios específicos.

Este tipo de análisis bioinformático (expresión diferencial y *variant calling*) se cobra en torno a 600 € por muestra en el mercado.

Si la empresa cobrara 600 €/muestra y analizara las 18 muestras (9 controles + 9 pacientes), obtendría 10.800 € de ingreso.

En la *Tabla 3* se muestran los cálculos de beneficio neto y retorno de inversión (ROI) estimados.

Tabla 3. Beneficio y ROI

Concepto	**Total**
Ingresos estimados	10.800,00 €
Costes del proyecto	4.345,24 €
Beneficio neto	6.454,76 €
ROI (%)	148,55 %

Fuente: elaborado por la autora.

El ROI calculado es del 148,55%, lo que refleja una elevada rentabilidad para la empresa en este tipo de proyectos.

Además del rendimiento económico, este proyecto podría permitir a *BIAMICS* posicionarse en el mercado de análisis bioinformático especializado en enfermedades neuropsiquiátricas, un sector en crecimiento y con alta demanda de soluciones personalizadas. Además, podría generar valor científico y comercial, ya que la detección de SNPs y patrones de expresión diferencial asociados a la esquizofrenia pueden dar lugar a la publicación de artículos científicos o presentaciones en congresos, aumentando el prestigio de la empresa y atrayendo nuevos clientes.

7. Resultados

7.1 Control de Calidad

Se obtuvo un informe FASTQC para cada secuencia, en un archivo HTML. Cada archivo contiene información sobre las estadísticas básicas, la calidad de la secuencia por base, las puntuaciones de calidad por secuencia, el contenido de secuencia por base, el contenido de GC (guanina-citosina) por secuencia, el contenido de N por base, la distribución de la longitud de la secuencia, las secuencias duplicadas, las secuencias sobrerrepresentadas, y el contenido de adaptadores.

En la *Ilustración 4* se muestran los resultados del control de calidad de la secuencia SRR28550736, sirviendo como ejemplo representativo del resto. Se puede observar que el contenido de N es nulo en todas las posiciones (gráfico "*Per base N content*"), lo cual indica ausencia de bases indefinidas, un buen indicador de calidad.

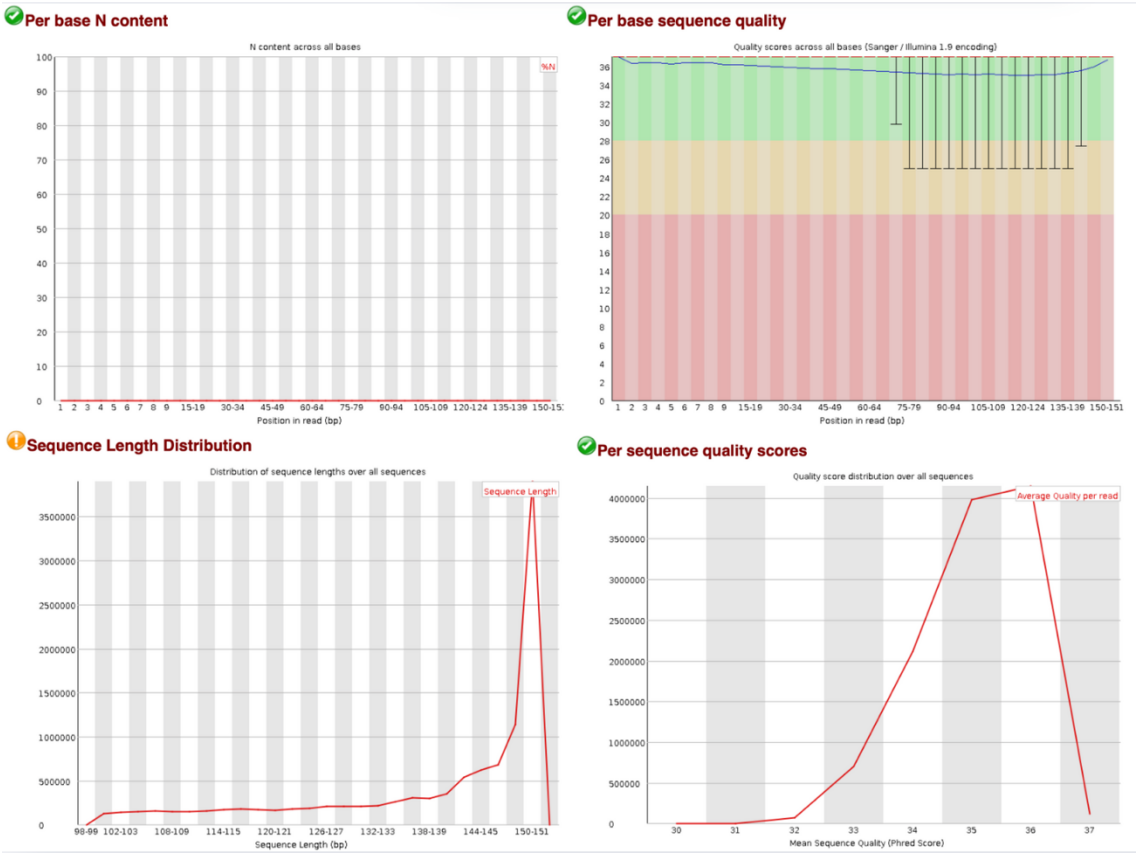


Ilustración 4. FastQC SRR28550736

El gráfico "*Per base sequence quality*" muestra puntuaciones Phred elevadas en la mayor parte de la lectura, con valores medios cercanos a 35-36 en las posiciones iniciales. A partir de la posición ~80 se aprecia un ligero descenso en la calidad y un aumento en la dispersión de los valores, con algunos percentiles por debajo de 30. No obstante, este patrón es común en datos

de secuenciación, especialmente en plataformas *Illumina*, donde es habitual observar una caída progresiva en la calidad hacia el extremo 3' de la lectura. En conjunto, la calidad general de las lecturas es alta y adecuada para continuar con el procesamiento bioinformático. Del mismo modo, el gráfico "*Per sequence quality scores*" evidencia que la mayoría de las lecturas tienen una calidad media por encima de 34.

En cuanto a la longitud de las secuencias ("*Sequence Length Distribution*"), la mayoría de las lecturas tienen una longitud uniforme de 150 pb, lo que indica que no hubo truncamientos ni recortes inesperados.

Además, se generó un informe MultiQC que agrupa y resume todos los resultados obtenidos en los análisis FASTQC individuales. Este informe presenta de forma conjunta y comparativa las métricas de calidad de todas las muestras analizadas, lo cual permite detectar de manera eficiente posibles inconsistencias o tendencias comunes entre ellas. Incluye representaciones gráficas y tablas que facilitan la interpretación de datos como la calidad por base, el contenido de GC, la presencia de secuencias duplicadas o adaptadores, entre otros aspectos relevantes para evaluar la calidad global del conjunto de datos.

En términos generales, las muestras presentaron una buena calidad de secuenciación. Las puntuaciones Phred por posición se mantuvieron elevadas, indicando una alta fiabilidad en la lectura de bases. Asimismo, el contenido GC fue uniforme y acorde con lo esperado para transcriptomas humanos.

Sin embargo, se detectaron algunas advertencias menores en ciertas muestras, principalmente asociadas a:

- Secuencias duplicadas: en algunos casos, el porcentaje de duplicación fue moderado, posiblemente relacionado con la expresión de genes altamente abundantes o regiones de baja complejidad.
- Contenido de bases por posición: se observaron ligeras desviaciones en las primeras posiciones de algunas secuencias, situación común en datos de secuenciación masiva y atribuible a artefactos técnicos de los primeros ciclos.

Estas advertencias no afectan de forma significativa la calidad de los datos. Por tanto, todas las muestras fueron consideradas aptas para continuar con el preprocesamiento y análisis posteriores.

7.2 Preprocesado

Todas las secuencias fueron sometidas a un proceso de filtrado y control de calidad utilizando la herramienta *Prinseq*. Como se muestra en la *Ilustración 5,* los resultados fueron consistentes

entre las distintas muestras. En promedio, se partió de más de 83 millones de secuencias por muestra, con longitudes medias de lectura cercanas a los 149 nucleótidos. A pesar del filtrado, más del 98 % de las secuencias fueron clasificadas como de buena calidad, lo que refleja que la calidad de los datos era elevada incluso antes del preprocesamiento.

```
Input and filter stats:
        Input sequences: 90,344,036
        Input bases: 13,408,269,729
        Input mean length: 148.41
        Good sequences: 88,971,970 (98.48%)
        Good bases: 13,297,487,196
        Good mean length: 149.46
        Bad sequences: 1,372,066 (1.52%)
        Bad bases: 103,360,890
        Bad mean length: 75.33
        Sequences filtered by specified parameters:
        min_len: 1372025
        ns_max_p: 41
Input and filter stats:
        Input sequences: 83,785,371
        Input bases: 12,495,757,214
        Input mean length: 149.14
        Good sequences: 83,068,100 (99.14%)
        Good bases: 12,439,092,736
        Good mean length: 149.75
        Bad sequences: 717,271 (0.86%)
        Bad bases: 53,165,410
        Bad mean length: 74.12
        Sequences filtered by specified parameters:
        min_len: 717271
Input and filter stats:
        Input sequences: 83,785,371
```

Ilustración 5. Salida por pantalla (Prinseq)

Solo un pequeño porcentaje de las secuencias (entre 0.8 % y 1.5 %) fue descartado por presentar baja calidad o no cumplir con los parámetros especificados. El contenido de bases con bajas puntuaciones de calidad también fue mínimo, y las secuencias eliminadas presentaban una longitud media inferior al resto.

En conjunto, estos resultados confirman que las lecturas utilizadas para el análisis presentaban una calidad adecuada, y que el preprocesamiento con *Prinseq* sirvió principalmente para reforzar la homogeneidad y fiabilidad del conjunto de datos.

Posteriormente al filtrado con *Prinseq*, se repitió el análisis con *FastQC* sobre los archivos. El objetivo fue verificar si el preprocesamiento había tenido algún efecto relevante sobre la calidad de las lecturas. Los resultados obtenidos mostraron que las métricas principales se mantuvieron estables y en rangos altos, pero además se observó una leve mejora en las regiones que inicialmente presentaban mayor dispersión en la calidad.

La *Ilustración 6* muestra que se mantuvieron puntuaciones Phred elevadas a lo largo de todas las posiciones de lectura y la distribución de calidad por secuencia continuó mostrando valores altos y concentrados en torno a 35-36. Del mismo modo, la longitud de las secuencias permaneció prácticamente inalterada, con lecturas uniformes de 149–150 pb.

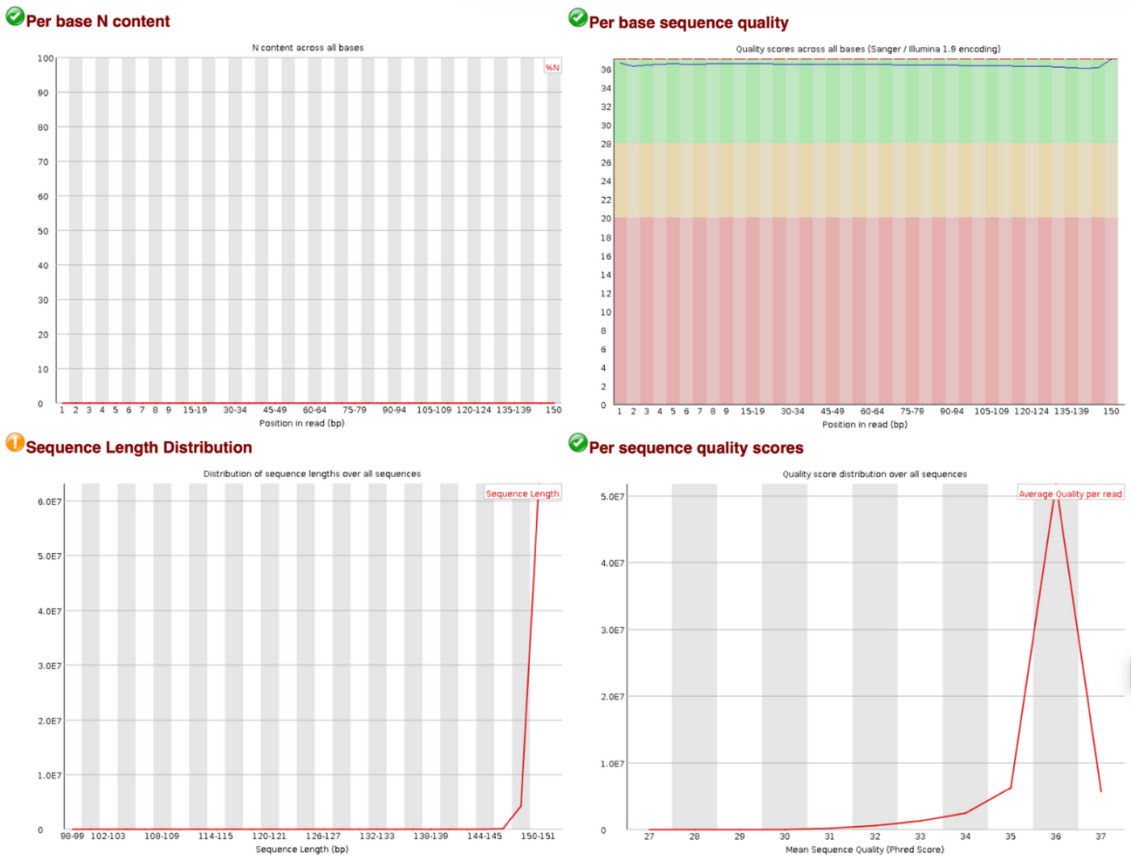


Ilustración 6. FastQC posterior a Prinseq

7.3 Mapeo

El alineamiento de las secuencias preprocesadas se llevó a cabo utilizando la herramienta *HISAT2*, empleando como referencia el genoma humano GRCh38. Los resultados obtenidos fueron altamente consistentes entre las distintas muestras analizadas. La *Ilustración 7* muestra un ejemplo representativo de una de las muestras, el cual refleja las tendencias generales observadas en todo el conjunto de datos.

En este caso, de un total de más de 93 millones de lecturas, el 97.38 % se alineó con éxito al genoma de referencia, lo que indica una tasa de alineamiento global muy alta. Del total de lecturas, el 65.71 % se alineó exactamente una vez de forma concordante, mientras que el 28.58 % se alineó más de una vez, lo cual es esperable en regiones repetidas o con múltiples isoformas. Solo un 5.71 % no logró alinearse concordantemente, y de estos, una fracción menor (15.52 %) presentó un alineamiento discordante.

```
93710232 reads; of these:
  93710232 (100.00%) were paired; of these:
    5351958 (5.71%) aligned concordantly 0 times
    61577009 (65.71%) aligned concordantly exactly 1 time
    26781265 (28.58%) aligned concordantly >1 times
    ----
    5351958 pairs aligned concordantly 0 times; of these:
      830486 (15.52%) aligned discordantly 1 time
    ----
    4521472 pairs aligned 0 times concordantly or discordantly; of these:
      9042944 mates make up the pairs; of these:
        4903250 (54.22%) aligned 0 times
        2828785 (31.28%) aligned exactly 1 time
        1310909 (14.50%) aligned >1 times
97.38% overall alignment rate
```

Ilustración 7. Salida por pantalla (HISAT2)

Estos resultados confirman la buena calidad y compatibilidad de las secuencias con el genoma de referencia, lo que proporciona una base sólida para continuar con los análisis transcriptómicos posteriores.

El proceso de alineamiento con *HISAT2* produjo varios archivos, mostrados en la *Ilustración 8*. Entre los archivos generados, destacan:

- Archivos SAM: generados inicialmente por *HISAT2* para cada muestra, contienen las alineaciones de las lecturas con el genoma de referencia.
- Archivos BAM ordenados: los archivos SAM fueron convertidos y ordenados usando samtools, obteniendo archivos en formato BAM. Estos archivos contienen las alineaciones en un formato comprimido y optimizado, y están nombrados según el identificador SRA de cada muestra.
- Archivos de índice BAM: junto a cada archivo BAM ordenado se generó un archivo de índice con extensión .bai que permite un acceso rápido y eficiente a las regiones específicas de los alineamientos durante las fases posteriores de cuantificación y ensamblaje.
- Índices del genoma: también se encuentran los archivos del índice del genoma de referencia generados por *HISAT2*, con extensiones .ht2. Estos índices son esenciales para permitir el alineamiento rápido y eficiente de las lecturas. Estos archivos constituyen la entrada para las etapas posteriores de cuantificación y ensamblaje de transcritos, asegurando la trazabilidad y la integridad de los datos en cada paso del análisis.

```
aroque@comba:~/HISAT$ ls
exons.txt                                hisat_pipeline_controles.sh  SRR28550743_sorted.bam.bai
GCF_000001405.40_GRCh38.p14_fixed.gtf    hisat_pipeline.sh            SRR28550744_sorted.bam
GCF_000001405.40_GRCh38.p14_genomic.fna  splicesites.txt              SRR28550744_sorted.bam.bai
GCF_000001405.40_GRCh38.p14_genomic.gff  SRR28550725.sam              SRR28550745_sorted.bam
GCF_000001405.40_GRCh38.p14_genomic.gtf  SRR28550725.sorted.bam      SRR28550745_sorted.bam.bai
GCF_000001405.40_GRCh38.p14_rna.fna      SRR28550725.sorted.bam.bai  SRR28550746_sorted.bam
GRCh38_hisat_index.1.ht2                 SRR28550736_sorted.bam      SRR28550746_sorted.bam.bai
GRCh38_hisat_index.2.ht2                 SRR28550736_sorted.bam.bai  SRR28550747_sorted.bam
GRCh38_hisat_index.3.ht2                 SRR28550737_sorted.bam      SRR28550747_sorted.bam.bai
GRCh38_hisat_index.4.ht2                 SRR28550737_sorted.bam.bai  SRR28550748_sorted.bam
GRCh38_hisat_index.7.ht2                 SRR28550738_sorted.bam      SRR28550748_sorted.bam.bai
GRCh38_hisat_index.8.ht2                 SRR28550738_sorted.bam.bai  SRR28550749_sorted.bam
GRCh38_hisat_index.rf                    SRR28550739_sorted.bam      SRR28550749_sorted.bam.bai
GRCh38_hisat_index_simple.1.ht2          SRR28550739_sorted.bam.bai  SRR28550750_sorted.bam
GRCh38_hisat_index_simple.2.ht2          SRR28550740_sorted.bam      SRR28550750_sorted.bam.bai
GRCh38_hisat_index_simple.3.ht2          SRR28550740_sorted.bam.bai  SRR28550751_sorted.bam
GRCh38_hisat_index_simple.4.ht2          SRR28550741_sorted.bam      SRR28550751_sorted.bam.bai
GRCh38_hisat_index_simple.5.ht2          SRR28550741_sorted.bam.bai  SRR28550752_sorted.bam
GRCh38_hisat_index_simple.6.ht2          SRR28550742_sorted.bam      SRR28550752_sorted.bam.bai
GRCh38_hisat_index_simple.7.ht2          SRR28550742_sorted.bam.bai  SRR28550753_sorted.bam
GRCh38_hisat_index_simple.8.ht2          SRR28550743_sorted.bam      SRR28550753_sorted.bam.bai
```

Ilustración 8. Archivos generados con HISAT2

7.4 Cuantificación de la expresión génica

Tras completar el alineamiento de las lecturas mediante *HISAT2*, se procedió a la cuantificación de la expresión génica utilizando la herramienta *HTSeq-count*. Esta herramienta permite contar el número de lecturas alineadas a cada gen, generando así una matriz de conteos que constituye la base para los análisis de expresión diferencial.

Para este proceso, se utilizaron los archivos BAM ordenados generados en la etapa de alineamiento, junto con el archivo de anotación del genoma humano en formato GTF. Dado que el archivo de anotación original presentaba inconsistencias (algunas anotaciones carecían del atributo "gene_id"), se aplicó una corrección previa para garantizar la compatibilidad con *HTSeq* (descrita en la sección de metodología).

El proceso de cuantificación produjo, para cada muestra, un archivo de conteos con extensión .htseq.counts.txt. Estos archivos contienen el número de lecturas asignadas a cada gen.

La *Ilustración 9* muestra los archivos generados, uno por cada muestra analizada, identificados por el código SRA correspondiente. Estos archivos constituyen la entrada para los análisis posteriores de expresión diferencial.

```
aroque@comba:~/HTSEQ$ ls
htseq_pipeline_controles.sh   SRR28550739.htseq.counts.txt  SRR28550745.htseq.counts.txt  SRR28550751.htseq.counts.txt
htseq_pipeline_pacientes.sh   SRR28550740.htseq.counts.txt  SRR28550746.htseq.counts.txt  SRR28550752.htseq.counts.txt
SRR28550725.htseq.counts.txt  SRR28550741.htseq.counts.txt  SRR28550747.htseq.counts.txt  SRR28550753.htseq.counts.txt
SRR28550736.htseq.counts.txt  SRR28550742.htseq.counts.txt  SRR28550748.htseq.counts.txt
SRR28550737.htseq.counts.txt  SRR28550743.htseq.counts.txt  SRR28550749.htseq.counts.txt
SRR28550738.htseq.counts.txt  SRR28550744.htseq.counts.txt  SRR28550750.htseq.counts.txt
```

Ilustración 9. Archivos generados con HTSeq-count

7.5 Ensamblaje y cuantificación de transcritos

Tras el alineamiento de las lecturas con *HISAT2*, se realizó el ensamblaje y la cuantificación de transcritos utilizando *StringTie*, una herramienta que permite reconstruir y cuantificar transcritos completos a partir de lecturas de RNA-seq mapeadas. A diferencia de *HTSeq*, que realiza una cuantificación a nivel de gen, *StringTie* ofrece una resolución más fina al nivel de isoformas, permitiendo identificar variantes transcriptómicas y eventos de *splicing* alternativo.

Para cada muestra, *StringTie* generó un archivo de salida en formato GTF que contiene los transcritos ensamblados junto con sus niveles de expresión. Estos archivos están nombrados según el identificador SRA correspondiente a cada muestra. La *Ilustración 10* muestra todos los archivos GTF generados para las muestras del estudio.

Además de los archivos GTF por muestra, *StringTie* generó una serie de archivos con extensión .ctab:

- e2t.ctab: asocia expresiones con transcritos.
- i2t.ctab: asignaciones internas entre transcritos e isoformas.
- t_data.ctab: contiene los niveles de expresión cuantificados para cada transcrito ensamblado.

Estos archivos .ctab permiten realizar análisis posteriores, como la comparación de isoformas entre muestras y el estudio de la expresión diferencial a nivel de transcrito.

```
aroque@comba:~/StringTie$ ls
e2t.ctab     SRR28550736.gtf  SRR28550740.gtf  SRR28550744.gtf  SRR28550748.gtf  SRR28550752.gtf                 t_data.ctab
e_data.ctab  SRR28550737.gtf  SRR28550741.gtf  SRR28550745.gtf  SRR28550749.gtf  SRR28550753.gtf
i2t.ctab     SRR28550738.gtf  SRR28550742.gtf  SRR28550746.gtf  SRR28550750.gtf  stringtie_pipeline_controles.sh
i_data.ctab  SRR28550739.gtf  SRR28550743.gtf  SRR28550747.gtf  SRR28550751.gtf  stringtie_pipeline_pacientes.sh
```

Ilustración 10. Archivos generados con StringTie

7.6 Expresión Diferencial

A partir de la matriz de conteos generada con *HTSeq* para todas las muestras del estudio, se llevó a cabo el análisis de expresión diferencial utilizando el paquete *EdgeR*. Esta herramienta permitió identificar genes cuya expresión difiere significativamente entre las condiciones de estudio (pacientes vs. controles), considerando tanto la magnitud del cambio como su significancia estadística.

Tras aplicar los criterios de filtrado y normalización adecuados, se realizó la estimación de la dispersión biológica y se realizó el test exacto de Fisher modificado. Los resultados obtenidos incluyeron valores de $\log_2$*Fold Change* ($\log_2$FC), valor p, y *False Discovery Rate* (FDR) para cada gen analizado.

La ejecución del paquete *EdgeR* generó dos archivos diferentes. El archivo *Ctrl_vs_Disease.tsv* recoge el listado completo de genes analizados, junto con sus métricas estadísticas. Como criterio de selección de genes diferencialmente expresados, se utilizó un umbral de |log2FC| ≥ 2,5, con el objetivo de centrarse únicamente en aquellos genes que presentan cambios de expresión marcados y potencialmente relevantes desde el punto de vista biológico. Este valor umbral exigente permite filtrar variaciones menores que podrían no tener un impacto funcional claro, priorizando así la identificación de candidatos robustos como posibles biomarcadores o dianas moleculares. La lista final de genes significativos fue almacenada en el archivo genes_+-2,5FC.tsv.

La *Ilustración 11* muestra un gráfico de componentes principales (PCA) generado a partir de los datos normalizados. Este tipo de gráfico permite evaluar visualmente la similitud global entre muestras según su perfil de expresión génica, lo que es especialmente útil para identificar patrones de agrupamiento o separación entre grupos biológicos (como pacientes y controles), así como para detectar posibles muestras atípicas (*outliers*), que puedan influir en los resultados del análisis.

En este caso, se observa una clara separación entre las muestras correspondientes a pacientes y controles, lo que sugiere diferencias en la expresión génica entre ambas condiciones. Cada punto representa una muestra y está etiquetado con su identificador. Los colores distinguen el grupo experimental al que pertenece cada muestra.

Cada eje representa el valor de logFC (*Logarithmic Fold Change*), que indica la variación en la expresión génica entre dos condiciones (control vs. paciente). El número entre paréntesis (13% y 12%) indica el porcentaje de variación explicada por esta dimensión.

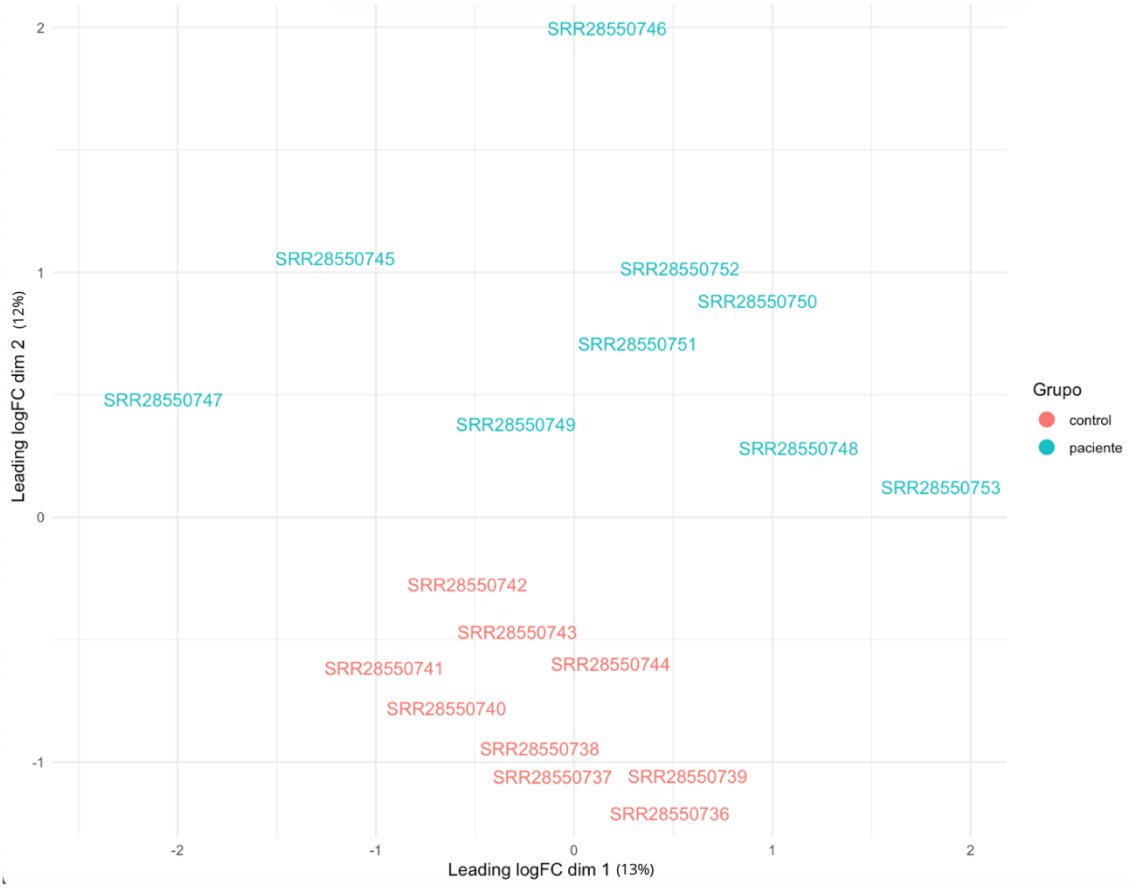


Ilustración 11. PCA

La *Ilustración 12* muestra todos los genes analizados mediante *EdgeR* en la comparación entre pacientes y controles. En este tipo de gráfico, el eje X muestra el logaritmo en base 2 del cambio en la expresión (log_2 FC). Las líneas verticales punteadas marcan los umbrales de $log_2FC = \pm 1$, por lo que los genes situados a la derecha tienen una expresión aumentada en pacientes respecto a controles ($log_2FC > 0$), mientras que los que están hacia la izquierda presentan una expresión disminuida.

Por otro lado, el eje Y representa el $-log_{10}$ del valor p, indicando la significación estadística del cambio de expresión. Cuanto más alto se sitúa un punto en el gráfico, mayor es su significación estadística (menor p-valor).

Los puntos (genes) están coloreados según la magnitud del cambio y su significación estadística. El color rojo representa genes significativamente sobreexpresados en pacientes ($log_2FC > 1$, $p < 0,05$), el color naranja genes con sobreexpresión moderada ($0 < log_2FC < 1$, $p < 0,05$), el azul oscuro genes significativamente subexpresados en pacientes ($log2FC < -1$, $p < 0,05$), el azul claro genes con subexpresión moderada ($-1 < log_2FC < 0$, $p < 0,05$), y el gris genes no significativos ($p \geq 0,05$). Además, en amarillo están representados aquellos genes que son de

especial interés ya que se encuentran incluidos en la lista *genes_+-2,5FC.tsv*, que cumplen |log2FC| ≥ 2,5.

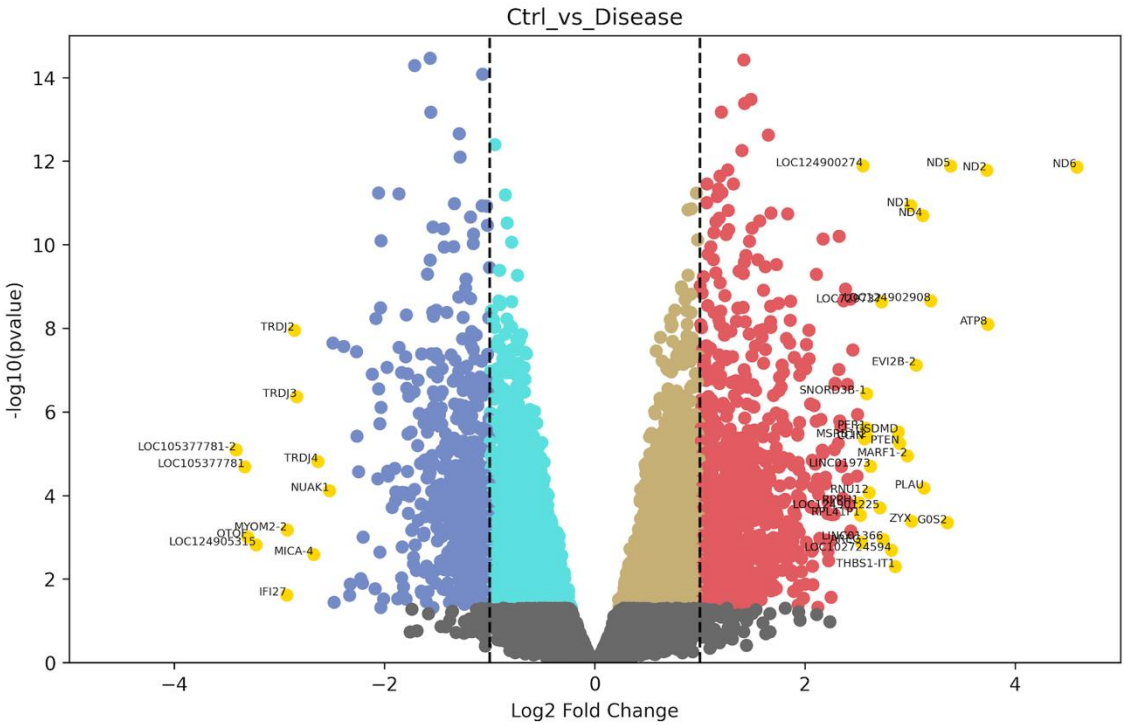


Ilustración 12. Control vs. Esquizofrenia (Genes)

En total, se identificaron 39 genes con un cambio de expresión significativo ($|\log_2 FC| \geq 2,5$), lo que podría indicar una diferencia pronunciada entre las muestras de pacientes y controles. Entre estos genes destacan varios genes mitocondriales, como *ND1, ND2, ND4, ND5, ND6 y ATP8*, implicados en la cadena de transporte de electrones y la producción de energía celular. También se identificaron varios genes relacionados con la inmunidad adaptativa, como *TRDJ2, TRDJ3 y TRDJ4*, que codifican segmentos de la región variable del receptor de células T. Asimismo, se detectaron genes no anotados o putativos como *LOC124900274*, *LOC124902908*, y *LOC729737*, cuya función biológica aún no ha sido completamente caracterizada. También se observó la presencia de *RN7SL3*, un ARN no codificante relacionado con la maquinaria de traducción. Adicionalmente, se detectaron genes con funciones en apoptosis y respuesta inflamatoria, como *GSDMD* (implicado en piroptosis) y *IFI27* (inductor por interferones), así como genes reguladores del ciclo celular y la proliferación como *PTEN, NUAK1, y AREG.*

7.7 GATK

Los resultados del análisis con *GATK* evidencian la correcta ejecución de cada una de las etapas del flujo de trabajo. La *Ilustración 13* muestra los archivos obtenidos para SRR28550736

con el *script* implementado, los cuales son representativos del procesamiento realizado para el resto de las muestras del estudio.

```
[aroque@comba:~/GATK/SRR28550736$ ls
SRR28550736_dedup.bam            SRR28550736_metrics.txt                SRR28550736_split.bai
SRR28550736_dedup.bam.bai        SRR28550736_raw_variants.g.vcf.gz      SRR28550736_split.bam
SRR28550736_filtered.vcf.gz      SRR28550736_raw_variants.g.vcf.gz.tbi
SRR28550736_filtered.vcf.gz.tbi  SRR28550736_RG.bam
```

Ilustración 13. Archivos generados con GATK (SRR28550736)

En primer lugar, se generó un archivo BAM con duplicados marcados (dedup.bam), tras aplicar la herramienta *MarkDuplicates* de *Picard*. Este archivo contiene las lecturas alineadas donde se identificaron y etiquetaron duplicados, con el objetivo de evitar sesgos durante el llamado de variantes. Junto a este archivo se encuentra su correspondiente índice (.bam.bai), necesario para permitir un acceso eficiente durante los pasos posteriores.

Asimismo, se observa la presencia del archivo RG.bam, que incluye la información de los grupos de lectura añadida con *AddOrReplaceReadGroups*. Este paso es indispensable para que *GATK* pueda distinguir entre muestras y asegurar la compatibilidad con sus herramientas.

Posteriormente, se llevó a cabo la etapa de preparación específica para datos de RNA-Seq mediante la herramienta *SplitNCigarReads*, generando el archivo split.bam y su índice. Este archivo contiene las lecturas ajustadas en las regiones de empalme (*splicing*), un paso esencial para que *GATK* interprete correctamente la estructura exón-intrón característica del transcriptoma.

En cuanto al llamado de variantes, se produjo un archivo intermedio comprimido en formato GVCF (raw_variants.g.vcf.gz), el cual incluye tanto variantes como sitios no variantes con sus probabilidades asociadas. Este archivo es el resultado de ejecutar *HaplotypeCaller* en modo GVCF y sirve como base para realizar el genotipado conjunto.

Finalmente, se generó un archivo VCF filtrado (filtered.vcf.gz), que contiene únicamente las variantes de alta calidad tras aplicar criterios de filtrado como la profundidad de lectura, la calidad del mapeo y otras métricas estándar. Este archivo representa el conjunto final de variantes candidatas para análisis funcionales, como la anotación con *SnpEff*. Además, el archivo de métricas generado (metrics.txt) resume la cantidad de duplicados detectados, proporcionando una evaluación adicional de la calidad de los datos procesados.

7.8 Variant Calling

A partir del archivo VCF filtrado generado con *GATK*, y utilizando el genoma creado para *SnpEff*, se obtuvo un archivo en formato VCF con la información sobre el impacto funcional de

las variantes detectadas. Estos archivos anotados contienen información sobre las características genómicas, incluyendo el tipo de variante y el efecto a nivel de genes y proteínas.

Como se muestra en la *Ilustración 14*, junto a cada archivo VCF anotado, se generó un archivo de resumen con el conteo de los diferentes tipos de mutaciones identificadas. Este archivo fue producido mediante el procesamiento de los campos de anotación, permitiendo cuantificar la frecuencia de cada tipo de mutación presente en las muestras.

```
aroque@comba:~/SnpEff/resultados$ ls
mover_resultados.sh.gz                  SRR28550742_annotated_snpeff.vcf.gz        SRR28550748_mutation_counts_snpeff.txt.gz
SRR28550736_annotated_snpeff.vcf.gz     SRR28550742_mutation_counts_snpeff.txt.gz  SRR28550749_annotated_snpeff.vcf.gz
SRR28550736_mutation_counts_snpeff.txt.gz SRR28550743_annotated_snpeff.vcf.gz      SRR28550749_mutation_counts_snpeff.txt.gz
SRR28550737_annotated_snpeff.vcf.gz     SRR28550743_mutation_counts_snpeff.txt.gz  SRR28550750_annotated_snpeff.vcf.gz
SRR28550737_mutation_counts_snpeff.txt.gz SRR28550744_annotated_snpeff.vcf.gz      SRR28550750_mutation_counts_snpeff.txt.gz
SRR28550738_annotated_snpeff.vcf.gz     SRR28550744_mutation_counts_snpeff.txt.gz  SRR28550751_annotated_snpeff.vcf.gz
SRR28550738_mutation_counts_snpeff.txt.gz SRR28550745_annotated_snpeff.vcf.gz      SRR28550751_mutation_counts_snpeff.txt.gz
SRR28550739_annotated_snpeff.vcf.gz     SRR28550745_mutation_counts_snpeff.txt.gz  SRR28550752_annotated_snpeff.vcf.gz
SRR28550739_mutation_counts_snpeff.txt.gz SRR28550746_annotated_snpeff.vcf.gz      SRR28550752_mutation_counts_snpeff.txt.gz
SRR28550740_annotated_snpeff.vcf.gz     SRR28550746_mutation_counts_snpeff.txt.gz  SRR28550753_annotated_snpeff.vcf.gz
SRR28550740_mutation_counts_snpeff.txt.gz SRR28550747_annotated_snpeff.vcf.gz      SRR28550753_mutation_counts_snpeff.txt.gz
SRR28550741_annotated_snpeff.vcf.gz     SRR28550747_mutation_counts_snpeff.txt.gz
SRR28550741_mutation_counts_snpeff.txt.gz SRR28550748_annotated_snpeff.vcf.gz
```

Ilustración 14. Archivos generados con SnpEff

En cuanto al filtrado de variantes, se generaron archivos específicos con variantes no sinónimas (*Ilustración 15*). Esto se realizó descartando aquellas variantes clasificadas como sinónimas, obteniendo así conjuntos más relevantes desde el punto de vista funcional. El resultado final de este paso son archivos _nonsyn.tsv para cada muestra, los cuales contienen únicamente las variantes con potencial efecto funcional significativo.

```
aroque@comba:~/SnpEff$ ls
anotar_36_44.sh             SRR28550736_nonsyn.tsv.gz  SRR28550741_nonsyn.tsv.gz  SRR28550746_nonsyn.tsv.gz  SRR28550751_nonsyn.tsv.gz
anotar_45_53.sh             SRR28550736.tsv.gz         SRR28550741.tsv.gz         SRR28550746.tsv.gz         SRR28550751.tsv.gz
resultados                  SRR28550737_nonsyn.tsv.gz  SRR28550742_nonsyn.tsv.gz  SRR28550747_nonsyn.tsv.gz  SRR28550752_nonsyn.tsv.gz
snpEff                      SRR28550737.tsv.gz         SRR28550742.tsv.gz         SRR28550747.tsv.gz         SRR28550752.tsv.gz
snpEff_v5_0_core.zip        SRR28550738_nonsyn.tsv.gz  SRR28550743_nonsyn.tsv.gz  SRR28550748_nonsyn.tsv.gz  SRR28550753_nonsyn.tsv.gz
snps_pacientes.py           SRR28550738.tsv.gz         SRR28550743.tsv.gz         SRR28550748.tsv.gz         SRR28550753.tsv.gz
snps_pacientes_umbral8.tsv  SRR28550739_nonsyn.tsv.gz  SRR28550744_nonsyn.tsv.gz  SRR28550749_nonsyn.tsv.gz
snps_pacientes_umbral.py    SRR28550739.tsv.gz         SRR28550744.tsv.gz         SRR28550749.tsv.gz
snps_pacientes_umbral.tsv.gz SRR28550740_nonsyn.tsv.gz SRR28550745_nonsyn.tsv.gz  SRR28550750_nonsyn.tsv.gz
snps_unicos_pacientes.tsv.gz SRR28550740.tsv.gz        SRR28550745.tsv.gz         SRR28550750.tsv.gz
```

Ilustración 15. Archivos con variantes no sinónimas (SnpEff)

El análisis de los SNPs e *indels* únicos en pacientes permite establecer diferencias genéticas potencialmente relevantes entre los pacientes y los controles, proporcionando información valiosa para estudios posteriores de asociación genética o análisis funcional.

A pesar de no encontrarse SNPs o *indels* presentes en todos los pacientes y ausentes en controles, se encontraron un total de 9 SNPs y 7 *indels* en 8 de los 9 pacientes, no presentes en los controles. La *Tabla 4* muestra los SNPs e *indels* comunes a los pacientes y ausentes en los controles. Muestra el cromosoma y posición en el que se ubican (coordenadas genómicas), el nombre del gen, el nucleótido/secuencia de referencia y la alteración presente en los pacientes.

Los nombres de los genes fueron obtenidos a través de las coordenadas genómicas utilizando la base de datos *BioMart*. Estas alteraciones representan variantes genéticas que podrían estar asociadas con el fenotipo de la esquizofrenia.

Tabla 4. SNPs e indels exclusivos de pacientes

Cromosoma	Posición	Gen	Referencia	Alteración
NC_000001.11	30743984	*LAPTM5*	T	C
NC_000003.12	127678940	*ABTB1*	A	G
NC_000003.12	167697822	*PDCD10*	C	CTTGTGT
NC_000003.12	167697823	*PDCD10*	A	G
NC_000003.12	167697826	*PDCD10*	A	AC
NC_000003.12	167697830	*PDCD10*	GGC	G
NC_000003.12	167697834	*PDCD10*	TAGA	T
NC_000003.12	167697838	*PDCD10*	G	T
NC_000003.12	167697842	*PDCD10*	TGAG	T
NC_000003.12	167697848	*PDCD10*	T	TTC
NC_000003.12	167697849	*PDCD10*	A	AG
NC_000006.12	11255676	*NEDD9*	T	C
NC_0000010.11	45491375	*MARCHF8*	T	C
NC_0000015.10	34249536	*SLC12A6*	T	C
NC_0000016.10	23915446	*PRKCB*	A	G
NC_0000022.11	35660703	*APOL6*	A	G

Fuente: elaborado por la autora.

8. Discusión

Aunque el mecanismo de acción de los genes identificados es complejo y sus vías metabólicas amplias, lo que dificulta el análisis genético funcional, es fundamental considerarlos en el contexto del trastorno esquizofrénico. Actualmente, la comprensión de la regulación genética en enfermedades psiquiátricas está en desarrollo, y muchos genes identificados aún carecen de una anotación funcional clara. No obstante, el análisis de expresión diferencial permite identificar patrones que podrían estar asociados a mecanismos moleculares específicos.

Al igual que en otros estudios transcriptómicos, uno de los desafíos es la variabilidad interindividual en la expresión génica, que puede depender de factores como la heterogeneidad genética, la influencia ambiental o el estado clínico de los pacientes.

En este estudio, a pesar de que se han identificado 39 genes con cambios significativos en la expresión, la interpretación funcional sigue siendo compleja debido a la falta de estudios previos específicos sobre esquizofrenia.

Dentro de los estos genes diferencialmente expresados, *ND1*, *ND2*, *ND4*, *ND5*, *ND6*, y *ATP8*, codifican subunidades esenciales de los complejos I y V de la cadena de transporte de electrones mitocondrial. Numerosos estudios han implicado la disfunción mitocondrial en la fisiopatología de la esquizofrenia, sugiriendo que alteraciones en estos genes podrían contribuir a déficits energéticos neuronales observados en pacientes (35, 36).

Por otro lado, varios estudios han demostrado que algunos pacientes con esquizofrenia tienen valores anómalos de receptores de células T, esenciales para que los linfocitos T reconozcan y respondan a antígenos específicos, lo que sugiere una posible implicación de la inmunidad adaptativa en la etiología del trastorno. Por este motivo, la expresión diferencial de genes involucrados en la inmunidad adaptativa (*TRDJ2*, *TRDJ3*, y *TRDJ4*) sugiere una regulación defectuosa del sistema inmunitario, lo que contribuiría a los procesos neuroinflamatorios y alteraciones inmunológicas observadas en la esquizofrenia. Esta hipótesis es consistente con estudios previos que han vinculado anomalías inmunológicas con síntomas neuropsiquiátricos, sugiriendo que la modulación de la respuesta inmune podría desempeñar un papel en la patogénesis del trastorno (37, 38).

En este contexto, resulta relevante considerar el papel de *GSDMD*, un ejecutor clave de la piroptosis, una forma de muerte celular inflamatoria. Un estudio reciente ha destacado su papel en la inflamación del sistema nervioso central y su posible contribución a la neuroinflamación observada en la esquizofrenia (39). Esto sugiere que la activación de procesos inflamatorios mediados por *GSDMD* podría estar vinculada a las alteraciones inmunológicas identificadas en pacientes con esquizofrenia, proporcionando un vínculo adicional entre la disfunción inmune y el desarrollo del trastorno.

PTEN es un gen supresor tumoral que regula la proliferación celular y la señalización intracelular. Debido a su influencia en el desarrollo y la función neuronal, mutaciones en *PTEN* se han asociado con trastornos neuropsiquiátricos como el autismo, por lo que podría considerarse su evaluación en trastornos como la esquizofrenia (40).

Las mutaciones de dos miembros de la familia relacionada con la proteína quinasa activada por AMP (AMPK), *NUAK1* y *NUAK2*, han sido asociadas con trastornos del espectro autista (ASD), trastorno por déficit de atención e hiperactividad (ADHD), esquizofrenia, y discapacidad intelectual (ID). *NUAK1* participa en la ramificación axonal y la función mitocondrial en neuronas corticales, por lo que alteraciones en este gen podrían afectar el desarrollo neuronal, contribuyendo a los mecanismos patológicos de la esquizofrenia (41). Por otro lado, *AREG*, un miembro de la familia de factores de crecimiento epidérmico (EGF), actúa como ligando del receptor EGFR y se ha observado que está involucrado en procesos de neurodesarrollo y plasticidad sináptica. La expresión alterada de *AREG* también podría estar relacionada con la esquizofrenia, lo que sugiere que tanto las vías de señalización asociadas a *NUAK1* como a *AREG* pueden influir en el desarrollo de alteraciones neuropsiquiátricas características del trastorno (42).

Además, los genes no anotados y regiones no codificantes (*LOC124900274*, *LOC124902908*, *LOC729737*, *LOC105377781-2*, *LOC105377781*, *LOC124901225*, *LOC124905315*, *LOC102724594*) representan regiones genómicas con funciones aún no completamente caracterizadas. Sin embargo, la creciente evidencia sugiere que los ARN no codificantes desempeñan roles cruciales en la regulación génica y su expresión diferencial podría ser un signo de implicación en la esquizofrenia.

Finalmente, el gen *PER1* es fundamental en la regulación del ritmo circadiano. Alteraciones en su expresión se han asociado con trastornos del sueño en pacientes con esquizofrenia, lo que sugiere una posible implicación en la fisiopatología del trastorno (13, 43).

Es importante considerar que el análisis de variantes genéticas debe integrarse con datos clínicos detallados para establecer relaciones causales. Además, debido a la alta variabilidad observada en algunos genes, sería relevante analizar la influencia de factores adicionales, como el ambiente o cofactores epigenéticos.

Por otro lado, en el análisis realizado con *SnpEff*, se identificaron 16 variantes genéticas exclusivas en pacientes que no se encontraron en los controles. Entre los genes más destacados con inserciones, deleciones, o sustituciones, se encuentran aquellos implicados en procesos celulares esenciales, señalización y respuesta inmune, así como en la regulación de funciones neuronales.

Dentro de estas 16 variantes genéticas, 9 corresponden al gen *PDCD10* (*Programmed Cell Death 10*), también conocido como *CCM3*, el cual codifica una proteína involucrada en la regulación de la apoptosis, la señalización intracelular y la homeostasis vascular. Está implicado en la formación de malformaciones cavernosas cerebrales (CCM) y juega un papel crucial en la integridad estructural de los vasos sanguíneos del cerebro. Además, se ha identificado su participación en vías de señalización que regulan la supervivencia celular y la respuesta al estrés (44).

Esto sugiere que este gen podría ser un punto clave en la cascada de eventos moleculares que contribuyen a la patogénesis de la esquizofrenia. Sin embargo, es fundamental considerar que no todas las variantes necesariamente implican una pérdida de función o un efecto patogénico directo. Para interpretar adecuadamente estos hallazgos, sería importante realizar análisis adicionales, como estudios funcionales para evaluar el impacto de cada mutación, así como el análisis de coexpresión con otros genes relacionados con la apoptosis y el mantenimiento estructural cerebral, y estudios de asociación genética para determinar la frecuencia de estas variantes en poblaciones con y sin esquizofrenia.

Los resultados obtenidos sugieren la importancia de continuar investigando la implicación de los genes identificados en procesos biológicos clave asociados con la esquizofrenia. Integrar análisis funcionales adicionales permitirá establecer con mayor precisión su papel en la fisiopatología de la enfermedad.

9. Limitaciones

Este estudio, centrado en el análisis de datos transcriptómicos de pacientes con esquizofrenia, presenta algunas limitaciones inherentes. En primer lugar, el análisis estuvo limitado a 18 secuencias obtenidas de SRA (9 pacientes y 9 controles), lo que puede afectar la generalización de los resultados a poblaciones más amplias. Además, debido a la complejidad del manejo de grandes volúmenes de datos, el enfoque metodológico utilizado priorizó el análisis de los genes con mayor relevancia estadística, lo que puede haber dejado fuera genes con cambios sutiles pero significativos.

Otra limitación importante es la ausencia de información clínica detallada de los pacientes, lo cual impidió correlacionar directamente los cambios en la expresión génica con parámetros clínicos específicos. Asimismo, algunos genes identificados como diferencialmente expresados no cuentan con información funcional clara, lo que dificulta la interpretación biológica de sus variaciones.

Finalmente, para validar los resultados obtenidos, se requieren estudios adicionales que incluyan un mayor número de muestras y profundicen en la caracterización funcional de los genes destacados. Ampliar el análisis a otras variantes genómicas, así como considerar otros factores ambientales y epigenéticos, podría enriquecer las conclusiones y aportar una visión más completa del impacto genético en la esquizofrenia.

10. Conclusiones

En el presente estudio, se llevaron a cabo diversas etapas de procesamiento y análisis de datos de secuenciación masiva, enfocadas en el análisis transcriptómico de muestras asociadas con esquizofrenia.

El análisis de la expresión diferencial con *EdgeR* reveló 39 genes con un cambio significativo en la expresión ($|\log_2 FC| \geq 2,5$) entre pacientes y controles. Entre estos genes, se identificaron varios relacionados con la cadena de transporte de electrones y la inmunidad adaptativa, lo que sugiere una posible asociación con los mecanismos moleculares implicados en la esquizofrenia.

El análisis de variantes con *GATK* y su posterior anotación con *SnpEff* permitió identificar nueve SNPs y siete *indels* exclusivos en pacientes que no están presentes en los controles. Entre estas variantes genéticas, destacan aquellas ubicadas en genes mitocondriales y relacionados con la inmunidad adaptativa, fortaleciendo la hipótesis de que la activación de vías inflamatorias podría desempeñar un papel importante en el desarrollo de alteraciones neuropsiquiátricas en pacientes con esquizofrenia.

Además, las múltiples variantes genéticas comunes a los pacientes en el gen *PDCD10* resalta la necesidad de explorar su papel en la esquizofrenia de manera más profunda, incluyendo estudios funcionales y análisis de variantes en grupos más extensos.

Los resultados obtenidos en este estudio constituyen una base sólida para el análisis transcriptómico y la identificación de variantes genéticas potencialmente asociadas con la esquizofrenia. Aunque estos hallazgos requieren validación experimental adicional, el enfoque metodológico empleado ha permitido detectar genes y variantes de interés que podrían contribuir significativamente a la comprensión de los mecanismos moleculares subyacentes al trastorno.

Además, los resultados abren nuevas posibilidades para investigaciones futuras centradas en la validación funcional de los genes identificados y en el análisis de su implicación en los procesos patológicos característicos de la esquizofrenia.

En resumen, los hallazgos de este estudio proporcionan una base prometedora para el desarrollo de estrategias diagnósticas y terapéuticas más eficaces, y refuerzan la utilidad del análisis transcriptómico como herramienta clave en el estudio de enfermedades neuropsiquiátricas complejas.

11. Referencias

1. BALU, D. y GOFF, D., 2012. Schizophrenia 2012. 1 de enero de 2012, pp. 80–106.
2. FADEN, J. y CITROME, L., 2023. Schizophrenia: One Name, Many Different Manifestations. Medical Clinics of North America, 107(1), pp. 61–72.
3. CORRELL, C., ARANGO, C., FAGERLUND, B., GALDERISI, S., KAS, M. y LEUCHT, S., 2024. Identification and treatment of individuals with childhood-onset and early-onset schizophrenia. European Neuropsychopharmacology, 82, pp. 57–71.
4. VALLE, R., 2020. Schizophrenia in ICD-11: Comparison of ICD-10 and DSM-5. Revista de Psiquiatría y Salud Mental, 13(2), pp. 95–104.
5. STRIMBU, K. y TAVEL, J., 2010. What are biomarkers? Current Opinion in HIV and AIDS, 5(6), pp. 463–466.
6. FISAR, Z., 2023. Biological hypotheses, risk factors, and biomarkers of schizophrenia. Progress in Neuro-Psychopharmacology & Biological Psychiatry, 120.
7. WAHBEH, M. y AVRAMOPOULOS, D., 2021. Gene-Environment Interactions in Schizophrenia: A Literature Review. Genes, 12(12).
8. TAKATA, A., XU, B., IONITA-LAZA, I., ROOS, J., GOGOS, J. y KARAYIORGOU, M., 2014. Loss-of-Function Variants in Schizophrenia Risk and SETD1A as a Candidate Susceptibility Gene. Neuron, 82(4), pp. 773–780.
9. SRIVASTAVA, A., DADA, O., QIAN, J., AL-CHALABI, N., FATEMI, A., GERRETSEN, P., et al., 2021. Epigenetics of Schizophrenia. Psychiatry Research, 305.
10. WU, Y., ZHANG, C., WANG, L., LI, Y. y XIAO, X., 2023. Genetic Insights of Schizophrenia via Single Cell RNA-Sequencing Analyses. Schizophrenia Bulletin, 49(4), pp. 914–922.
11. LIU, L., WU, J., QING, L., LI, J., YANG, H., JI, A., et al., 2020. DNA Methylation Analysis of the NR3C1 Gene in Patients with Schizophrenia. Journal of Molecular Neuroscience, 70(8), pp. 1177–1185.
12. SCHOONOVER, K., MILLER, N., FISH, K. y LEWIS, D., 2024. Scaling of Smaller Pyramidal Neuron Size and Lower Energy Production in Schizophrenia. Biological Psychiatry, 95(10), pp. S271–S272.
13. VON SCHANTZ, M., LEOCADIO-MIGUEL, M., MCCARTHY, M., PAPIOL, S. y LANDGRAF, D., 2021. Genomic perspectives on the circadian clock hypothesis of psychiatric disorders. Advances in Genetics, 107, pp. 153–191.
14. NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Sequence Read Archive (SRA) [en línea]. Bethesda (MD): NCBI, [s.f.] [consulta: 9 junio 2025]. Disponible en: https://www.ncbi.nlm.nih.gov/sra
15. SMITH, R. y MAES, M., 1995. The macrophage-T-lymphocyte theory of schizophrenia: Additional evidence. Medical Hypotheses, 45(2), pp. 135–141.

16. NAJJAR, S. y PEARLMAN, D., 2015. Neuroinflammation and white matter pathology in schizophrenia: systematic review. Schizophrenia Research, 161(1), pp. 102–112.
17. FROMER, M., POCKLINGTON, A., KAVANAGH, D., WILLIAMS, H., DWYER, S., GORMLEY, P., et al., 2014. De novo mutations in schizophrenia implicate synaptic networks. Nature, 506(7487), pp. 179–186.
18. FATEMI, S. y FOLSOM, T., 2009. The Neurodevelopmental Hypothesis of Schizophrenia, Revisited. Schizophrenia Bulletin, 35(3), pp. 528–548.
19. KIM, D., et al. HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 2015, 12(4), pp. 357–360.
20. CHEN, Y., LIN, C. y LANE, H., 2022. Distinctively lower DISC1 mRNA levels in patients with schizophrenia, especially in those with higher positive, negative, and depressive symptoms. Pharmacology Biochemistry and Behavior, 213.
21. REFISCH, A., KOMATSUZAKI, S., UNGELENK, M., SCHUMANN, A., CHUNG, H., SCHILLING, S., et al., 2022. Analysis of CACNA1C and KCNH2 Risk Variants on Cardiac Autonomic Function in Patients with Schizophrenia. Genes, 13(11).
22. ZHAO, S., et al. A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics, 2019, 20(1), pp. 1–12.
23. PERTEA, M., KIM, D., PERTEA, G., LEEK, J. y SALZBERG, S., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols, 11(9), pp. 1650–1667.
24. VAN DER AUWERA, G.A., CARNEIRO, M.O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., et al., 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics, 43(1110), pp. 11.10.1–11.10.33.
25. LAKE, B., AI, R., KAESER, G., SALATHIA, N., YUNG, Y., LIU, R., et al., 2016. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science, 352(6293), pp. 1586–1590.
26. VAN DEN BRINK, S., SAGE, F., VÉRTESY, A., SPANJAARD, B., PETERSON-MADURO, J., BARON, C., et al., 2017. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nature Methods, 14(10), pp. 935–936.
27. RIPKE, S., NEALE, B., CORVIN, A., WALTERS, J., FARH, K., HOLMANS, P., et al., 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), pp. 421–428.
28. GANDAL, M., HANEY, J., PARIKSHAK, N., LEPPA, V., RAMASWAMI, G., HARTL, C., et al., 2018. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science, 359(6376), pp. 693–697.

29. BIRNBAUM, R. y WEINBERGER, D., 2017. Genetic insights into the neurodevelopmental origins of schizophrenia. Nature Reviews Neuroscience, 18(12), pp. 727–740.
30. CUMMINGS, B., KARCZEWSKI, K., KOSMICKI, J., SEABY, E., WATTS, N., SINGER-BERK, M., et al., 2020. Transcript expression-aware annotation improves rare variant interpretation. Nature, 581(7809), pp. 452–456.
31. COCK, P.J.A., FIELDS, C.J., GOTO, N., HEUER, M.L. y RICE, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 2010, 38(6), pp. 1767–1771.
32. ANDREWS, S. FastQC: a quality control tool for high throughput sequence data [en línea]. 2010 [consulta: 9 junio 2025]. Disponible en: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
33. ANDERS, S., PYL, P.T. y HUBER, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 2015, 31(2), pp. 166–169.
34. ROBINSON, M.D., MCCARTHY, D.J. y SMYTH, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26(1), pp. 139–140.
35. CINGOLANI, P., PLATTS, A., WANG, L.L., COON, M., NGUYEN, T., LÉGÉR, P., et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 2012, 6(2), pp. 80–92.
36. GONÇALVES, V., GIAMBERARDINO, S., CROWLEY, J., VAWTER, M., SAXENA, R., BULIK, C., et al., 2018. Examining the role of common and rare mitochondrial variants in schizophrenia. PLoS ONE, 13(1).
37. KAKIUCHI, C., ISHIWATA, M., KAMETANI, M., NELSON, C., IWAMOTO, K. y KATO, T., 2005. Quantitative analysis of mitochondrial DNA deletions in the brains of patients with bipolar disorder and schizophrenia. International Journal of Neuropsychopharmacology, 8(4), pp. 515–522.
38. LUO, C., PI, X., HU, N., WANG, X., XIAO, Y., LI, S., et al., 2021. Subtypes of schizophrenia identified by multi-omic measures associated with dysregulated immune function. Molecular Psychiatry, 26(11), pp. 6926–6936.
39. DEBNATH, M., 2015. Adaptive Immunity in Schizophrenia: Functional Implications of T Cells in the Etiology, Course and Treatment. Journal of Neuroimmune Pharmacology, 10(4), pp. 610–619.
40. JIA, C., ZHANG, M., WU, X., ZHANG, X., LV, Z., ZHAO, K., et al., 2025. HERV-W Env Induces Neuron Pyroptosis via the NLRP3-CASP1-GSDMD Pathway in Recent-Onset Schizophrenia. International Journal of Molecular Sciences, 26(2).

41. GROSS, C., 2017. Defective phosphoinositide metabolism in autism. Journal of Neuroscience Research, 95(5), pp. 1161–1173.
42. BENNISON, S., LIU, X. y TOYO-OKA, K., 2022. Nuak kinase signaling in development and disease of the central nervous system. Cellular Signalling, 100.
43. CARPENTER, G., 1983. The biochemistry and physiology of the receptor-kinase for epidermal growth factor. Molecular and Cellular Endocrinology, 31(1), pp. 1–19.
44. JOHANSSON, A., OWE-LARSSON, B., HETTA, J. y LUNDKVIST, G., 2016. Altered circadian clock gene expression in patients with schizophrenia. Schizophrenia Research, 174(1–3), pp. 17–23.
45. VALENTINO, M., DEJANA, E. y MALINVERNO, M., 2021. The multifaceted PDCD10/CCM3 gene. Genes & Diseases, 8(6), pp. 798–813.

12. Anexos

Anexo 1. FI-030 Propuesta de Proyecto Final.

Anexo 2. Hisat.sh

Este *script* lleva a cabo la alineación y procesamiento de secuencias utilizando el software *HISAT2* y *Samtools*. Primero, define las rutas a los archivos de índice, directorios de entrada y salida, así como el número de hilos para el procesamiento paralelo. A continuación, realiza un bucle que recorre las muestras. Para cada muestra, realiza el mapeo de lecturas filtradas contra el genoma de referencia usando *HISAT2*, genera archivos de alineación en formato SAM, los convierte a formato BAM, los ordena, los indexa y elimina los archivos intermedios para optimizar el almacenamiento.

Anexo 3. HTSeq.sh

Este *script* cuantifica la expresión génica a partir de los archivos BAM generados con *HISAT2*, utilizando la herramienta *HTSeq-count*. Especifica el directorio de archivos BAM generados previamente, la ubicación del archivo de anotación en formato GTF, y el directorio de salida donde se almacenarán los resultados. Para cada muestra, ejecuta el comando htseq-count, que cuenta el número de lecturas alineadas a cada gen:

-f bam: Indica que el formato de entrada es BAM.

-r pos: Especifica que las lecturas están ordenadas por posición.

-s no: Indica que los datos no son específicos de cadena (*strand*).

-t exon: Utiliza los exones como unidades de conteo.

-i gene_id: Usa el identificador del gen para agrupar los conteos.

Anexo 4. EdgeR.R

Este *script* en R realiza el análisis de expresión diferencial utilizando el paquete EdgeR. Tras verificar la instalación del paquete, carga los datos de cuentas y fenotipos, ordena las muestras y crea un objeto DGEList para el análisis. Filtra los genes de baja expresión usando la función filterByExpr(), que retiene únicamente los genes con un nivel de expresión suficiente, mejorando la potencia estadística. A continuación, estima la dispersión para calcular la variabilidad biológica y realiza una prueba exacta de tipo binomial negativo mediante la función exactTest(), adecuada para comparar dos grupos. Los resultados incluyen genes

diferencialmente expresados con |log2FC| ≥ 2,5 y p-valor < 0,05, guardados en archivos TSV. Además, genera gráficos MDS, BCV y un volcán (*smear plot*) para visualizar diferencias entre condiciones.

Anexo 5. StringTie.sh

Este *script* lleva a cabo el ensamblaje y cuantificación de transcritos a partir de los archivos BAM generados con *HISAT2* utilizando la herramienta *StringTie*. Define el directorio que contiene los archivos BAM generados previamente, el directorio donde se almacenarán los resultados de *StringTie*, la ruta del archivo de anotación en formato GTF y el número de hilos para el procesamiento paralelo. Para cada muestra, ejecuta el comando *stringtie*, que realiza el ensamblaje de transcritos y la cuantificación de la expresión génica:

$BAMDIR/${SAMPLE}_sorted.bam: Archivo BAM ordenado como entrada.

-G $GTF_REF: Archivo GTF de referencia para la anotación génica.

-o $OUTDIR/${SAMPLE}.gtf: Archivo de salida con los ensamblajes en formato GTF.

-p $THREADS: Número de hilos para ejecución paralela.

-e: Modo de estimación de abundancia solo para genes presentes en el archivo de referencia.

-B: Genera archivos adicionales útiles para la fusión posterior de ensamblajes.

Anexo 6. Variant_calling.sh

Este *script* en Bash realiza el proceso de llamado de variantes (*variant calling*) utilizando las herramientas *Picard*, *GATK* y *Samtools*. El objetivo es identificar variantes genéticas a partir de archivos BAM alineados.

Como parámetros iniciales, define las rutas a las herramientas *Picard* y *GATK*, así como al archivo de referencia genómica (GRCh38).

Paso 1: *AddOrReplaceReadGroups* (*Picard*). Añade o reemplaza los grupos de lectura (*Read Groups*) en el archivo BAM, lo que es esencial para el procesamiento posterior con *GATK*:

- RGID: Identificador del grupo de lectura.
- RGLB: Biblioteca de secuenciación (lib1).
- RGPL: Plataforma de secuenciación (Illumina).
- RGPU: Unidad de procesamiento (unit1).
- RGSM: Muestra.

Paso 2: *MarkDuplicates* (*Picard*). Marca duplicados para evitar sesgos en el llamado de variantes. Genera un archivo BAM sin duplicados y un archivo de métricas. Posteriormente, indexa el archivo BAM procesado usando *Samtools*.

Paso 3: *SplitNCigarReads* (*GATK*). Realiza el preprocesamiento necesario para manejar lecturas con *indels* y alineamientos complejos en regiones codificantes.

Paso 4: *HaplotypeCaller* (*GATK*). Realiza el llamado de variantes utilizando el algoritmo *HaplotypeCaller*. Genera un archivo de variantes en formato GVCF.

Anexo 7. Genotype_gcvf.sh

Este *script* verifica la existencia del archivo GVCF y, si está presente, ejecuta el comando GenotypeGVCFs para convertir el GVCF en un archivo VCF filtrado, utilizando el genoma de referencia GRCh38.

Anexo 8. SnpEff.sh

Este *script* realiza la anotación de variantes genéticas utilizando *SnpEff* procesando archivos VCF previamente filtrados. Para cada muestra, ejecuta *SnpEff* para generar un archivo VCF anotado que incluye información sobre el impacto biológico de cada variante. Además, extrae y cuenta los tipos de mutación presentes en el archivo anotado, guardando el resumen en un archivo de texto ordenado.

Anexo 9. Snps_pacientes.py

Este *script* en Python identifica variantes genéticas exclusivas de pacientes en comparación con controles utilizando archivos TSV que contienen variantes no sinónimas. Primero, carga los archivos de ambos grupos y extrae las variantes clave (CHROM, POS, REF, ALT) para cada paciente y control, generando conjuntos de variantes comunes entre pacientes y entre controles. Luego, determina las variantes que están presentes únicamente en pacientes, y no en controles. Finalmente, filtra las variantes únicas a partir del primer archivo de pacientes y guarda los resultados en un archivo "snps_unicos_pacientes.tsv", indicando el número total de variantes exclusivas encontradas.