2022 Lleida 27·1 junio·juny julio·juliol Cataluña Catalunya

8º CONGRESO FORESTAL ESPAÑOL

La **Ciencia forestal** y su contribución a los **Objetivos de Desarrollo Sostenible**

8CFE - 1041

Edita: Sociedad Española de Ciencias Forestales
Cataluña | Catalunya · 27 junio | juny - 1 julio | juliol 2022
ISBN 978-84-941695-6-4
© Sociedad Española de Ciencias Forestales

Organiza

Sociedad Española de Ciencias Forestales SECF

Optimización de los flujos de trabajo bioinformáticos para el análisis de expresión diferencial a partir de RNA-seq en *Pinus radiata*

HURTADO GONZÁLEZ, M.[1], MORA-MARQUEZ, F.[3] MARINO BILBAO, D.[2], SOTO DE VIANA, A.[3], LÓPEZ DE HEREDIA LARREA, U.[3] Y GOICOECHEA, P.G.[1]

[1] Departamento de Ciencias Forestales. NEIKER-BRTA, Instituto Vasco de Investigación y Desarrollo Agrario, Campus Agroalimentario de Arkaute, Crtra N-104 km 355, 01192 Arkaute, Alava, España.
[2] Departamento de Biología Vegetal y Ecología. Facultad de Ciencia y Tecnología. Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV-EHU). Barrio Sarriena s/n 48940 Leioa, Bizkaia.
[3] Departamento de Sistemas y Recursos Naturales. ETSI de Montes, Forestal y del Medio Natural. Universidad Politécnica de Madrid. Ciudad Universitaria s/n 28040, Madrid, Spain.

Resumen

Este estudio pretende optimizar el flujo de trabajo bioinformático para analizar la expresión génica diferencial, mediante RNA-seq, cuando no existe un genoma de referencia de la especie de interés. Para ello hemos analizado la expresión diferencial del cambium en árboles jóvenes y adultos de *Pinus radiata* utilizando el genoma de *Pinus taeda* como referencia. El diseño estadístico incluye tres árboles adultos y tres jóvenes, analizados en verano e invierno. Tras la extracción de ARN, preparación de librerías para RNAseq y secuenciación masiva en la plataforma Illumina□ se analizó la influencia de: (1) diferentes tasas de alineamiento con HISAT2, (2) tablas de cobertura (StringTie) en comparación a las cuentas observadas (HTSeq-count) y (3) análisis de expresión diferencial mediante los dos softwares predominantes (DESeq2 y edgeR). Tras observar una gran variabilidad en los resultados de las alternativas propuestas, se realizó una simulación con el objetivo de obtener un "gold standard" para analizar las tasas de positivos verdaderos/falsos positivos. Los resultados de las simulaciones muestran diferencias relevantes de los flujos de trabajo tanto en número de genes como en la tasa de positivos verdaderos/falsos positivos. Esto ofrece la posibilidad de seleccionar el flujo de trabajo más adecuado para la finalidad de cada estudio.

Palabras clave

Alineamiento, ensamblaje, expresión diferencial, transcriptómica, mejora genética forestal.

1. Introducción

La secuenciación masiva de ARN (RNA-seq) se ha convertido en una herramienta revolucionaria para los estudios de transcriptómica (Wang et al. 2009) y en concreto para el análisis de expresión diferencial (ED). La cantidad de datos obtenidos en el proceso de secuenciación es tan enorme, que los análisis bioinformáticos se han convertido en uno de los cuellos de botella del proceso (Wang et al. 2009La disponibilidad de un genoma debidamente ensamblado y desarrollado, i.e. como el de especies modelo, facilita el análisis bioinformático y permite obtener datos de mayor calidad. Sin embargo, cuando la especie de estudio no tiene un genoma disponible, la bioinformática y los análisis de ED de genes se vuelven un reto. Esta situación se puede afrontar con dos estrategias diferentes: (1) el ensamblaje *de novo* y (2) el uso de la mejor referencia genómica cercana en un flujo de trabajo guiado por genoma de referencia.

El ensamblaje *de novo* es una herramienta poderosa para especies que carecen de genoma de referencia. Sin embargo, este tipo de ensamblajes dependen de dos asunciones implícitas: (1) que el transcriptoma ensamblado es una representación imparcial, aunque incompleta, del verdadero transcriptoma subyacente expresado, y (2) que las estimaciones de expresión del ensamblaje son buenas, aunque ruidosas, y están cercanas a la abundancia relativa de los transcritos expresados. Sin embargo, la evidencia indirecta sugiere que estas suposiciones con frecuencia no se cumplen

(Fredman et al. 2020). Esto podría indicar que el ensamblaje con un genoma de referencia cercano puede ser la mejor alternativa a la hora de realizar un análisis de ED.

Por otro lado, el análisis bioinformático enfocado a especies modelo ha contado con importantes avances, muchos de ellos desarrollados durante el proyecto de los 1000 genomas (The 1000 Genomes Project Consortium 2015). Sin embargo, estos flujos de trabajo optimizados para especies modelo podrían no ser tan efectivos cuando la especie de interés no tiene unos recursos bioinformáticos de calidad disponibles, o las referencias genómicas están altamente fragmentadas e incompletas.

Este contexto es relativamente común en la genética forestal, ya que gran parte de las especies forestales carecen de un genoma secuenciado, o en caso de tenerlo éstos suelen estar fragmentados y son de baja calidad. Cuando se usa una especie filogenéticamente cercana, existe un riesgo alto de que las lecturas no alineen como debieran a la referencia debido a discrepancias entre los genomas de ambas especies, especialmente en genomas ampliamente fragmentados.

El pino radiata es una especie de gran interés económico (Forrest 1973). Por ello, muchas de las investigaciones realizadas sobre ella han sido llevadas a cabo por compañías privadas que no han hecho públicos los datos. Este es el caso de la secuenciación de su genoma, que disponiéndose de un primer borrador, los datos permanecen privados (https://www.scionresearch.com/about-us/about-scion/corporate-publications/scion-connections/past-issues-list/scion-connections-issue-26,-december-2017/radiata-pine-genome-draft-assembly-completed).

En este estudio hemos llevado a cabo un análisis de ED a partir del RNAseq de *P. radiata* mediante el flujo de trabajo de StringTie (Pertea et al. 2016). Nuestro análisis incluye 3 tipos de variantes/modificaciones en las fases más importantes del proceso, con el objetivo de optimizarlo para análisis en los que no se dispone de un genoma de referencia. Además, se presenta una propuesta de benchmarking mediante lecturas simuladas y un 'gold standard' para comprobar los efectos de dichas modificaciones en la tasa de positivos verdaderos (PV) y falsos positivos (FP).

2. Objetivos

El objetivo principal de este trabajo es optimizar el flujo de trabajo bioinformático para analizar la ED mediante RNA-seq cuando no existe un genoma de referencia de la especie de interés. Además, se pretende valorar dichas modificaciones mediante un benchmarking con lecturas simuladas y un "gold standard".

3. Metodología

3.1 Material vegetal, secuenciación y filtrado de lecturas

En este estudio utilizamos tejido cambial de 3 árboles jóvenes (< 12 años) y de 3 árboles adultos (> 20 años) de *P. radiata* de diferentes plantaciones del País Vasco (España). Todos los árboles fueron muestreados durante verano e invierno, con la ayuda de un sacabocados en los árboles jóvenes, mientras que la corteza gruesa de los árboles adultos se desprendió con una hacha pequeña. Inmediatamente el cámbium se recogió con un bisturí y las muestras de tejido se cortaron en trozos pequeños y se sumergieron en RNAlater™ donde se almacenaron a -20 ºC hasta la extracción del ARN.

Para la extracción de ARN, el exceso de RNAlater™ se secó con toallitas de papel y seguidamente se homogenizaron 100 mg de biomasa con la ayuda de nitrógeno líquido y un mortero. La extracción se realizó siguiendo el protocolo de LiCl (Le Provost et al. 2007). La contaminación de

ADN se eliminó mediante el tratamiento de las muestras con DNAsa I (Zymo Research) y su posterior purificación con RNA Clean & Concentrator™-25 (Zymo Research) siguiendo las instrucciones del fabricante. El ARN se conservó después de la extracción a -20 ºC. Los análisis del número de integridad del ARN (RIN, por sus siglas en inglés RNA integrity number), construcción de librerías con TrueSeq Stranded Total RNA y la secuenciación NGS con lecturas pareadas (siguiendo el protocolo de Illumina), se llevaron a cabo por los servicios de RNAseq de Diagenode. Las lecturas se filtraron para eliminar secuencias del adaptador, así como las de baja calidad (< Q30) con el software Cutadapt v2.10 (Martin 2011). Por último, se llevaron a cabo controles visuales de calidad con el software FASTQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

3.2 Flujo de trabajo de StringTie

El análisis de ED del tejido cambial de árboles adultos y jóvenes se llevó a cabo modificando 3 fases críticas del flujo de trabajo de StringTie (Pertea et al. 2016), mostrado en la Figura 1. Para este análisis se emplearon las 6 librerías de invierno, así como el genoma de *P. taeda* v.2_0.1 (Zimin et al. 2017) localizado en la base de datos pública TreeGenes (Falk et al. 2018 y Wegrzyn et al. 2019). Las 6 librerías de verano se descartaron para los análisis de DE debido a fuertes efectos de lote y a las grandes diferencias de cobertura. El alineamiento de las lecturas al genoma se realizó con HISAT2 v2.2 (Kim et al. 2019) usando (1) los parámetros por defecto (excepto el uso de la opción –dta) y (2) parámetros modificados que permiten incrementar la tasa de alineamiento (cambio del coeficiente A del parámetro –score-min de A=-0,2 a A=-1). Los alineamientos resultantes (ficheros BAM) se procesaron con SAMtools v1.9 (Danecek et al. 2021) para eliminar duplicados y extraer los alineamientos concordantes. El ensamblaje individual de las librerías se llevo a cabo mediante StringTie v2.1.2 (Kovaka et al. 2019) a partir del GTF de referencia del genoma y los alineamientos. En este paso se permitió el ensamblaje de nuevos transcritos no incluidos en el GTF de referencia. Después se fusionaron los GTFs individuales (opción 'merge' de StringTie), y este GTF unido se utilizó junto con los ficheros de alineamiento para generar (1) las tablas de cobertura con StringTie y (2) las tablas de cuentas observadas con (HTseq-count v0.12.4 (Anders et al. 2014). Unos y otras se emplearon posteriormente para analizar la ED con (1) DESeq2 y con (2) edgeR (Robinson et al. 2010, McCarthy et al. 2012 y Chen et al. 2016). Ambos paquetes de análisis de ED utilizan una distribución binomial negativa para modelizar los resultados observados, pero mantienen varias diferencias. Entre ellas podemos destacar: (1) la fase de normalización, en la que DESeq2 utiliza la mediana de ratios de cuentas observadas mientras que edgeR utiliza la llamada Media recortada de valores M (TMM, de las siglas en inglés Trimmed Mean of M Values); (2) los estadísticos utilizados, ya que mientras que DESeq2 utiliza el modelo linear generalizado (GLM) para experimentos de factores únicos (edgeR recomienda utilizars tests exactos.

3.3 Ensamblaje del transcriptoma y generación de lecturas simuladas

Dada la gran variabilidad de resultados entre las modificaciones realizadas (ver resultados) al flujo de trabajo de StringTie, se recurrió a simulaciones de lecturas controlando los valores de TPMs (Wagner 2012), que funcionarían como "gold standard y a la comparación de los PV y FP obtenidos por los diferentes flujos de trabajo de análisis de expresión diferencial. Para la simulación de lecturas partimos de un transcriptoma ensamblado *de novo* con Trinity v2.12.0 (Grabherr et al. 2011) a partir de las 12 librerías generadas por RNA-seq. Tras filtrar el pseudo-trancriptoma por tamaño (transcritos entre 300 y 10000 bp) con NGShelper v0.56 (https://github.com/GGFHF/NGShelper) se realizó un alineamiento al genoma de referencia de *P. taeda* con GMAP-GSNAP versión 2021-05-27 (Wu and Watanabe 2005), lo que nos permitió separar los transcritos en 3 grupos: los que no mapearon (path 0), los que mapearon una única vez (path 1) y los que mapearon multiples veces (path n).

El transcriptoma, junto con las librerías de invierno fueron necesarias para simular lecturas mediante RSEM v1.3.1 (Li and Dewey 2011). En la tabla 1 se muestran los pasos a seguir durante el proceso de simulación y benchmarking en los que además de los programas mencionados se

emplearon R versión 4.0.5 (https://www.r-project.org/) y GffCompare v0.12.2 (Pertea, G. y Pertea, M. 2020). Se simularon 6 librerías con $2x10^7$ lecturas cada una, en las que 100 transcritos seleccionados aleatoriamente estaban sobrexpresados (50 sobreexpresados en adultos y 50 en jóvenes). Todos los transcritos sobreexpresados pertenecían al path 1, para evitar ambigüedades a la hora de realizar el conteo de PV y FP. Los niveles de sobrexpresión se dividieron en 5 grupos (x4, x6, x8, x10 y x12). para los transcritos de relleno. Cada librería simulada se completó con 99900 transcritos de relleno, expresados con un valor base de 9,96 TPMs.

3.4 Control de calidad del transcriptoma y librerías

Con la finalidad de evaluar la calidad del del pseudo-transcriptoma ensamblado se emplearon los programas BUSCO v4.0.6 (Manni et al. 2021) y RNA-QUAST v.5.0.2 (Mikheenko et al. 2018). El primero da información sobre la presencia de genes muy conservados en los organismos vivos (BUSCOs), de manera que se espera que un alto porcentaje de estos esté presente en un transcriptoma de alta calidad. RNA-QUAST genera estadísticas en base a la longitud de los transcritos. Por otro lado, la calidad de las librerías para estudios de ED se analizó en términos de correlación entre réplicas, análisis llevado a cabo con los scripts complementarios incluidos con Trinity. También se extrajeron datos de cobertura de lecturas con SAMtools (SAMtools stats) a partir de los ficheros de alineamiento (BAM) generados previamente con HISAT2.

4. Resultados

4.1 Controles de calidad

El control de calidad llevado a cabo por BUSCO para el pseudo-transcriptoma mostró que se llegaron a ensamblar una gran cantidad de genes conservados, también llamados BUSCOs, (Tabla 2). Las estadísticas de RNA-QUAST (Tabla 3) muestran valores generales aceptables si bien el n50 no es muy alto (1632).

Las librerías de invierno muestran una elevada correlación entre réplicas (Figura 2), de lo que se deduce que son apropiadas para el análisis ED. Las estadísticas de SAMtools muestran coberturas de lecturas similares cuando el alineamiento se llevó a cabo con los parámetros por defecto de HISAT2 o modificados (Figura 3).

4.2 Flujo de trabajo con librerías reales

Las tasas de alineamiento de las librerías reales con HISAT2 se incrementaron sensiblemente (alrededor del 20 %) notoriamente cuando se utilizó el parámetro más permisivo (–score-min A=-1), en lugar de los parámetros por defecto (–score-min A=-0,2). Esto ocurrió tanto en el ratio general de alineamiento (OAR, por sus siglas en inglés overall alignment rate) que incluye el alineamiento de todas las lecturas tratadas por individual en lugar de parejas, como para los alineamientos concordantes. Los alineamientos concordantes son aquellos que se alinean respetando las diferencias en la posición genómica que caben de esperar para cada pareja.

Por otro lado, en los resultados de ED (Tabla 5) también se pudo observar un notable aumento en el número de genes expresados diferencialmente (DEGs) cuando los parámetros de HISAT2 fueron modificados para un alineamiento más permisivo. La comparación entre StringTie y HTSeq-count muestra que las tablas de cobertura generada por StringTie producen un mayor número de DEGs que las matrices de cuentas observadas generadas por HTSeq-count. La comparación entre DESeq2 y edgeR, ambos utilizados con las cuentas observadas por HTSeq-count, mostró un mayor número de DEGs con DESeq2.

Finalmente, la comparación entre los genes expresados diferencialmente en los diferentes flujos de trabajo se realizó mediante diagramas de Venn (nótese que únicamente se realizan comparaciones dentro del mismo modo de HISAT2 (por defecto o modificado), ya que diferentes alineamientos producen diferentes ensamblados, y, por lo tanto, diferentes identidades (nombres) de los genes. En los diagramas se observó un comportamiento similar para las 3 alternativas analizadas (StringTie/DESeq2, HTSeq-count/DESeq2 y HTSeq-count/edgeR) al comparar los alineamientos de HISAT2 con parámetros por defecto y modificados (Figura 4). A destacar la mayor cantidad de genes exclusivos para StringTie, que se podría esperar debido a las diferencias en las cantidades totales de DEGs.

4.2 Benchmarking con librerías simuladas

Al igual que en las librerías reales, las lecturas simuladas también obtuvieron mayores tasas de alineamiento cuando HISAT2 se utilizó con parámetros modificados (Tabla 6). El benchmarking de las simulaciones, también mostró que los PV aumentaron alrededor de 20% cuando se empleó el modo de alineamiento más permisivo (Figura 5a). Sin embargo, los FP aumentaron entre un 40% y 64% dependiendo de la variante del flujo de trabajo empleada (Figura 5b). Dentro de modos de alineamiento (por defecto o modificado), las diferencias en PV fueron muy pequeñas, permaneciendo ligeramente por debajo la combinación de HTSeq-count /DESeq2. Por el contrario, las diferencias en los FP, fueron muy notables. El uso de StringTieresultó en un mayor número de FP en comparación a la utilización de HTSeq-count. En cuanto a la comparación entre softwares de ED, edgeR, mostró un número ligeramente mayor de falsos positivos que DESeq2. La comparación de PV mostrada en diagramas de Venn (Figura 6) muestra cómo dentro de cada modo de alineamiento, prácticamente la totalidad de genes se comparten entre las variantes del flujo de trabajo. En cuanto a los FP, las alternativas con tasas menores, no producen FP únicos (específicos de la alternativa) si no que son compartidos con el resto; siendo así las que mayor tasa de FP tienen, las únicas que generan FP específicos.

5. Discusión

El uso de un genoma de referencia de una especie cercana cuando la especie de interés no lo posee requiere el ajuste de los flujos de trabajo existentes para lograr unos resultados de mayor calidad. Además, HISAT2 que está optimizado para el genoma humano puede requerir también modificaciones para alineamientos con mega-genomas, sobretodo si estos se encuentran muy fragmentados, como el *P. taeda*. Sin embargo, en ocasiones no es fácil determinar la fiabilidad cuando se realizan este tipo de ajustes. En este estudio se han evaluado los efectos de las modificaciones al popular flujo de trabajo de Stringtie guiado por un genoma de referencia (Pertea et al. 2016), a partir del RNA-seq de tejido cambial de *P. radiata* y el uso de un genoma de referencia altamente fragmentado de *P. taeda*.

A pesar de disponer de un total de 12 librerías, incluyendo lotes de verano e invierno, se decidió utilizar únicamente el lote de invierno para los análisis de ED, debido a que las librerías de verano mostraron un efecto de lote demasiado elevado (datos no mostrados). Tanto el transcriptoma como las 6 librerías restantes mostraron calidades aceptables con los que llevar a cabo el estudio.

En el análisis de ED en las librerías de invierno se apreció una gran variabilidad en los resultados obtenidos para las diferentes modificaciones propuestas al flujo de trabajo (Figura 1). Estas diferencias nos indican el gran efecto que pueden tener el uso de unos u otros programas o parámetros en un determinado flujo de trabajo.

edgeR admite el output de StringTie de las tablas de cobertura para su análisis mediante un script complementario, siempre y cuando se emplee el parámetro -e al generar los GTFs de los

ensamblajes individuales. Sin embargo, dicho parámetro no era de interés en el presente estudio debido a que limita los resultados a los genes anotados en la referencia. Por este motivo y a pesar de que se pretendió comparar las 4 combinaciones posibles entre los programas para generar tablas de cobertura/matrices de expresión y para analizar la ED, ocurrió la incapacidad de analizar el uso de StringTie junto con edgeR. Al momento de la escritura de este trabajo se ha encontrado una vía alternativa a este inconveniente, sin embargo, dado que se encuentra en fase de análisis, no ha sido posible presentar estos resultados.

En cualquier caso, el punto más crítico de todo el flujo de trabajo se encuentra en el alineamiento ya que es dónde mayores diferencias se obtuvieron, no solo en los FP sino también en los PV. En términos generales los resultados de este estudio parecen indicar que el uso de HTseq-count en conjunto con DESeq2 parece ser la configuración más equilibrada obteniendo el menor número de falsos positivos de entre todas las alternativas analizadas, sin afectar prácticamente a los positivos verdaderos obtenidos. Sin embargo, si el objetivo del estudio requiere un mayor número de PV, la optimización del parámetro de permisividad podría ser esencial, mientras que si se prefiere una relación PV/FP más equilibrada el alineamiento con parámetros por defecto parece ser el indicado.

Este estudio da una idea general del comportamiento de diferentes programas y parámetros en un flujo de trabajo, así como una propuesta de benchmarking para testar futuras modificaciones en este y otros flujos de trabajo que tienen como objetivo final el análisis de ED. Sin embargo, la propia simulación tiene margen de mejora. Prueba de ello se encuentra en que no se ha conseguido reproducir la realidad en la relación del número de DEGs conseguidos por DESeq2 y edgeR. Una solución para ello consiste en utilizar niveles de expresión basados en la realidad en lugar de niveles prefijados (artificiales). Esto es algo en lo que ya se ha comenzado a trabajar en el momento de la escritura de este trabajo y está ofreciendo resultados prometedores. Por último, cabe mencionar que las opciones de modificación durante el transcurso de todo el flujo de trabajo son innumerables y las propuestas en este estudio sirven como punto de partida para la adaptación a diferentes casos cuando no existe un genoma de referencia de calidad para la especie a analizar.

6. Conclusiones

Con el presente estudio se ha podido observar que la modificación de los parámetros y programas en un flujo de trabajo pueden dar lugar a resultados muy diversos. En particular los parámetros de alineamiento son cruciales en los positivos verdaderos y falsos positivos y dependiendo del objetivo del estudio el ajuste de interés podría diferir. Por otro lado, el uso de las diferentes combinaciones posibles entre StringTie/HTseq-count y DESeq2/edgeR no tiene gran efecto en el número de positivos verdaderos. Por el contrario, el número de falsos positivos varía notablemente, mostrando que HTseq-count junto con DESeq2 puede ser la combinación con resultados más robustos.

7. Agradecimientos

Este estudio ha sido financiado por NEIKER-BRTA, Instituto Vasco de Investigación y Desarrollo Agrario. M. Hurtado agradece la concesión de una beca de doctorado al organismo de Desarrollo Económico e Infraestructuras del Gobierno Vasco.

8. Bibliografía

ANDERS, S.; PYL, P.T.; HUBER, W.: 2014. HTseq — A Python framework to work with high-throughput sequencing data. Bioinformatics 31(2) 166-169

CHEN, Y.; LUN, A.T.L.; SMYTH, G.K.; 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research 5:1438 1-51

DANECEK, P.; AUTON, A.; ABECASIS, G.; ALBERS, C.A.; BANKS, E.; DEPRISTO, M.A.; HANDSAKER, R.E.; LUNTER, G.; MARTH, G.T.; SHERRY, S.T.; MCVEAN, G.; DURBIN, R.; 1000 Genomes Project Analysis Group; 2011. The variant call format and VCFtools. Bioinformatics 27(15) 2156-2158.

DANECEK, P.; BONFIELD, J.K.; LIDDLE, J.; MARSHALL, J.; OHAN, V.; POLLARD, M.O.; WHITWHAM, A.; KEANE, T.; MCCARTHY, S.A.; DAVIES, R.M.; LI, H.; 2021. Twelve years of SAMtools and BCFtools. GigaScience 10 1-4

FALK, T.; HERNDON, N.; GRAU, E.; BUEHLER, S.; RICHTER, P.; ZAMAN, S.; BAKER, E.M.; RAMNATH, R.; FICKLIN, S.; STATON, M.; FELTUS, F.A.; JUNG, S.; MAIN, D.; WEGRZYN, J.L.; 2018. Growing and cultivating the forest genomics database, TreeGenes. Database 2018 1-11

FREEDMAN, A.H.; CLAMP, M.; SACKTON, T.B. SACKTON; 2020. Error, noise and bias in de novo transcriptome assemblies. Mol Ecol Resour. 21 18–29

FORREST, W.G.; 1973. Biological and Economic Proeduction in Radiata Pine Plantations. Journal of Applied Ecology 10(1) 259-267

GRABHERR, M.G.; HAAS, B.J.; YASSOUR, M.; LEVIN, J.Z.; THOMPSON, D.A.; AMIT, I.; ADICONIS, X.; FAN, L.; RAYCHOWDHURY, R.; ZENG, Q.; CHEN, Z.; MAUCELI, E.; HACOHEN, N.; GNIRKE, A.; RHIND, N.; DI PALMA, F.; BIRREN, B.W.; NUSBAUM, C.; LINDBLAD-TOH, K.; FRIEDMAN, N.; REGEV, A.; 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 29(7) 644-652

KIM, D.; PAGGI, J.; PARK, C.; BENNETT, C.; SALZBERG, S.; 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37 907-915

KOVAKA, S.; ZIMIN, A.V.; PERTEA, G.M., RAZAGHI, R.; SALZBERG, S.L.; PERTEA, M.; 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology 20 Article number 278 1-13

LI, B.; Dewley, C.N.; 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011 12:323

LE PROVOST, G.; HERRERA, R.; PAIVA, J.; CHAUMEIL, P.; SALIN, F.; PLOMION, C.; 2007. A micromethod for high throughput RNA extraction in forest trees. Biol Res 40 291-297
LÓPEZ DE HEREDIA, U.; VÁZQUEZ-POLETTI, J.L.; 2016. RNA-seq analysis in forest tree species: bioinformatic problems and solutions. Tree Genetics & Genomes (2016) 12:30

LOVE, M.I., HUBER, W.; ANDERS, S.; 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology Article number 550 1-21

MANNI, M.; BERKELEY, M.R.; MATHIEU SEPPEY, SIMÃO, F.A.; ZDOBNOV, E.M.; 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution 38(10) 4647–4654
MARTIN, M.; 2011. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. EMBnet.journal 17.1 10-12

MCCARTHY, D.J.; CHEN, Y.; SMYTH, G.K.; 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40(10) 4288-4297

MCKENNA, A.; HANNA, M.; BANKS, E.; SIVACHENKO, A.; CIBULSKIS, K.; KERNYTSKY, A.; GARIMELLA, K.; ALTSHULER, D.; GABRIEL, S.; DALY, M.; DEPRISTO, M.A.; 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data Genome Res. 20(9) 1297-303

MIKHEENKO, A.; PRJIBELSKI, A.; SAVELIEV, V.; ANTIPOV, D.; GUREVICH, A.; 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34(13) 142-150

PERTEA, M.; KIM, D.; PERTEA, G.M.; LEEK, J.; SALZBERG, S.; 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols 11 1650-1667

PERTEA, G.; PERTEA, M.; 2020 GFF Utilities: GffRead and GffCompare. F1000Research 9:304

ROBINSON, M.D.; MCCARTHY, D.J.; SMYTH, G.K.; 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1) 139-140

THE 1000 GENOMES PROJECT CONSORTIUM; 2015. A global reference for human genetic variation. Nature **526** 68–74

WAGNER, G.; 2012. Measurement of mRNA abundance using RNA-Seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131 281–285

WANG, Z.; GERSTEIN, M.;SNYDER, M.; 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 January 10(1) 57–63.

WEGRZYN, J.L.; STATON, M.A.; STREET, N. R.; MAIN, D.; GRAU, E.; HERNDON, N.; BUEHLER, S.; FALK, T.; ZAMAN, S.; RAMNATH, R.; RICHTER, P.; SUN, L.; CONDON, B.; ALMSAEED, A.; CHEN, M.; MANNAPPERUMA, C.; JUNG, S.; FICKLIN, S.; 2019. Cyberinfrastructure to Improve Forest Health and Productivity: The Role of Tree Databases in Connecting Genomes, Phenomes, and the Environment. Frontier in Plant Science 10 Article 813 1-8

WU, T.D.; WATANABE, C.K.; 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005 21 1859-1875

ZIMIN, A.V.; STEVENS, K.A.; CREPEAU, M.W.; PUIU, D.; WEGRZYN, J.L.,; YORKE, J.A.; LANGLEY, C.H.; NEALE, D.B.; SALZBERG, S.L.; 2017. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 6(1) 1-4.

Tabla 1. Descripción del proceso de simulación de lecturas.

Paso	Proceso	Programa	Descripción
1	**Aprendizaje de parámetros**	RSEM	Se adquiere la información de parámetros de secuenciación de librerías reales para emplear para aplicar a las lecturas simuladas.
2	**Selección de transcritos**	R	Selección aleatoria de transcritos sobrexpresados y de relleno. Los transcritos sobrexpresados pertenecen al path 1 del transcriptoma para evitar ambigüedades.
3	**Edición de TPMs**	R	Edición de las TPM de las tablas de expresión del paso 1 para los transcritos seleccionados con los niveles de expresión establecidos.
4	**Simulación de lecturas**	RSEM	Se generan lecturas con las tablas editadas en el paso 3, parámetros aprendidos en el paso 1 y el transcritpoma ensamblado.
5	**Chequeo del gold standard**	DESeq2	Análisis de expresión diferencial con las tablas de expresión obtenidas en la simulación (valores de TPM finales). Se espera obtener de resultado los genes correspondientes a los transcritos sobrexpresados seleccionados en el paso 2.
6	**Ejecución del flujo de trabajo guiado por referencia (StringTie)**	Explicado en el apartado 3.2	Se ejecuta el flujo de trabajo guiado por referencia con todas las modificaciones a testar empleando las lecturas simuladas y el genoma de *P. taeda* como referencia.
7	**Unión de transcritos de novo y transcritos ensamblados con StringTie**	GffCompare	Es necesaria la búsqueda de equivalencias entre los transcritos del transcriptoma de referencia y los ensamblados con StringTie durante el flujo de trabajo guiado por referencia. Para ello se cruzan los GTFs del mapeo de GMAP del transcritoma al genoma de *P. taeda* y el obtenido en el paso de StringTie.
8	**Conteo de PV/FP**		Se cuentan los PV y FP de los análisis de expresión diferencial respecto a lo que se espera del gold standard.

Tabla 2. Estadísticas BUSCO para la evaluación de calidad del transcriptoma.

	Total	%
BUSCOs completos	1491	92,40 %
BUSCOS completos y únicos	476	29,50 %
BUSCOS completos y duplicados	1015	62,90 %
BUSCOs fragmentados	28	1,70 %
BUSCOs faltantes	95	5,90 %
Total de grupos de BUSCO buscados	1614	

Tabla 3. Estadísticas QUAST para la evaluación de calidad del transcriptoma.

# contigs (>= 0 bp)	1637886
# contigs (>= 1000 bp)	243968
# contigs (>= 5000 bp)	13471
# contigs (>= 10000 bp)	1
# contigs (>= 25000 bp)	0
# contigs (>= 50000 bp)	0
Longitud total (>= 0 bp)	1178253686
Longitud total (>= 1000 bp)	542420676
Longitud total (>= 5000 bp)	86233498
Longitud total (>= 10000 bp)	10000
Longitud total (>= 25000 bp)	0
Longitud total (>= 50000 bp)	0
# contigs	630389
Contig más largo	10000
Longitud total	800989753
GC (%)	38.81
N50	1632
N75	826

L50	131713
L75	308341
# N's por cada 100 kbp	0,00

Tabla 4. Tasas de alineamiento globales (OAR) y de alineamientos concordantes (conc.) de las librerías reales con HISAT2 para parámetros por defecto y parámetros modificados.

Librería	Defecto (OAR)	Defecto (conc.)	Modificado (OAR)	Modificado (conc.)
AWA	68,73 %	37,98 %	87,89 %	59,02 %
AWY	74,48 %	37,32 %	92,63 %	57,80 %
GWA	71,61 %	37,50 %	91,52 %	59,06 %
GWY	71,76 %	36,72 %	91,97 %	58,41 %
VWA	30,44 %	13,80 %	71,15 %	42,49 %
VWY	68,67 %	33,52 %	89,75 %	55,41 %
Media	64,28 % ☐ 16,72 %	32,81 % ☐ 9,25 %	87,48 % ☐ 8,19 %	55,37 % ☐ 6,45 %

Tabla 5. DEGs en librerías reales de invierno para todas las modificaciones al flujo de trabajo realizadas.

	HISAT2 defecto	HISAT2 modificado
StringTie DESeq2	213	639
HTseq-count DESeq2	114	336
HTseq-count edgeR	67	88

Tabla 6. Tasas de alineamiento concordantes de las librerías simuladas con HISAT2 para parámetros por defecto y parámetros modificados.

Librería	HISAT2 defecto	HISAT2 Modificado
AWA_sim	24,77 %	45,52 %
AWY_sim	23,14 %	41,36 %
GWA_sim	23,79 %	43,06 %
GWY_sim	23,32 %	42,11 %
VWA_sim	25,71 %	44,34 %
VWY_sim	22,44 %	40,50 %
Media	23,86 % ☐ 1,19 %	42,81 % ☐ 1,88 %

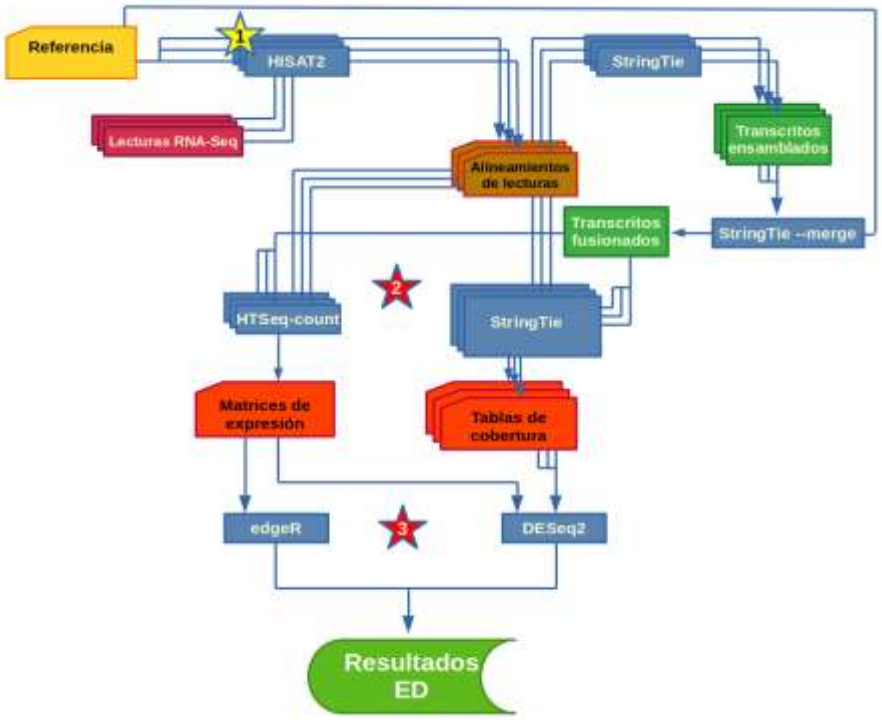

Figura 1. Modificaciones realizadas al flujo de trabajo guiado por genoma de referencia (Pertea et al. 2016). Las 3 modificaciones se muestran con estrellas amarillas si consisten en la modificación de parámetros de programas existentes o en rojo si consiste en nuevos programas. Imagen modificada a partir de la obtenida de https://ccb.jhu.edu/software/StringTie/index.shtml?t=manual.

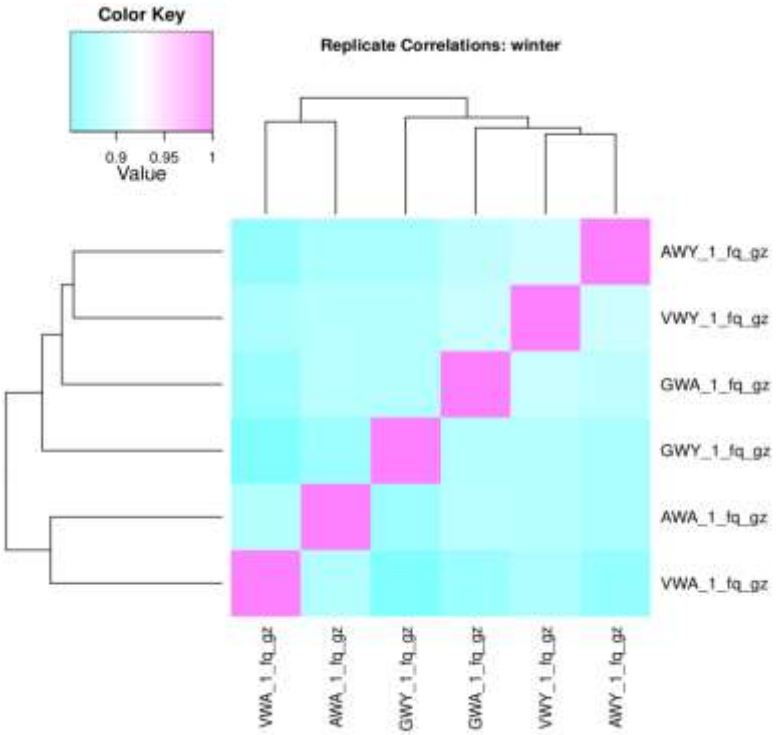

Figura 2. Correlación de réplicas en las librerías de invierno.

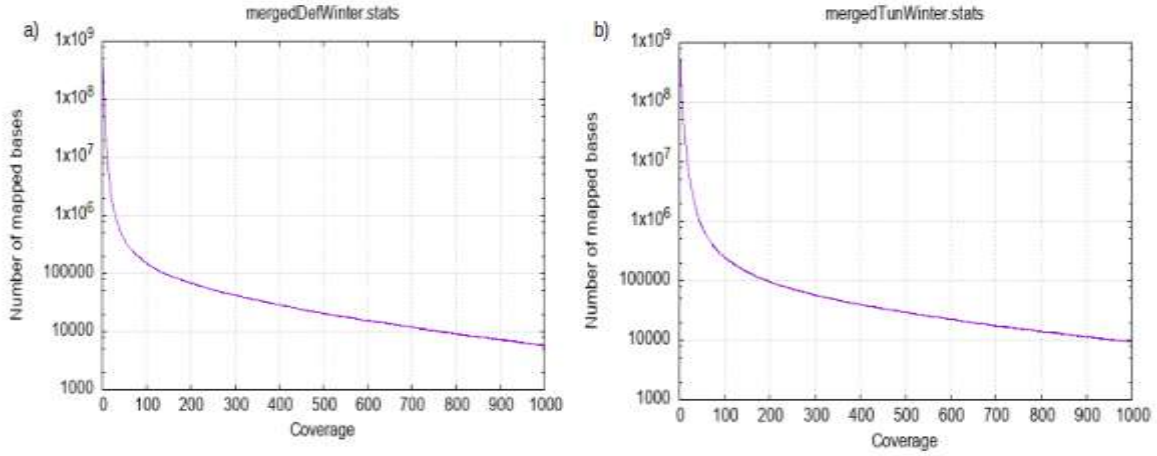

Figura 3. Gráficos de cobertura de los alineamientos de las librerías reales al genoma de P. taeda con HISAT2 con parámetros por defecto (a) y parámetros modificados (b)

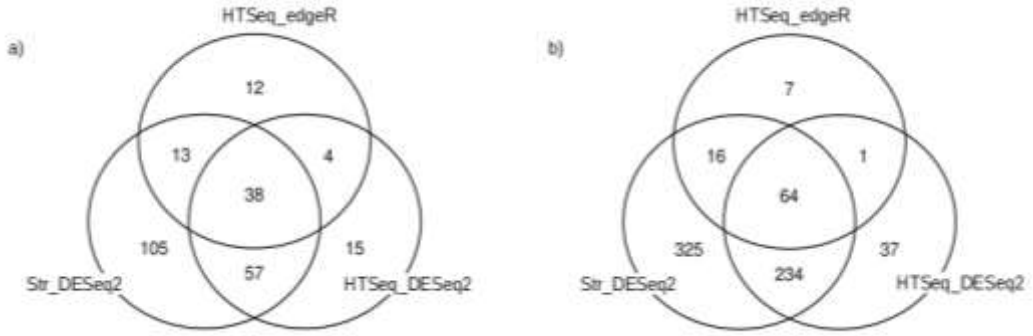

Figura 4. Diagramas de Venn para la comparación de los resultados ED en las diferentes variantes del flujo de trabajo para las librerías reales con los parámetros por defecto de HISAT2 (a) y modificados (b).

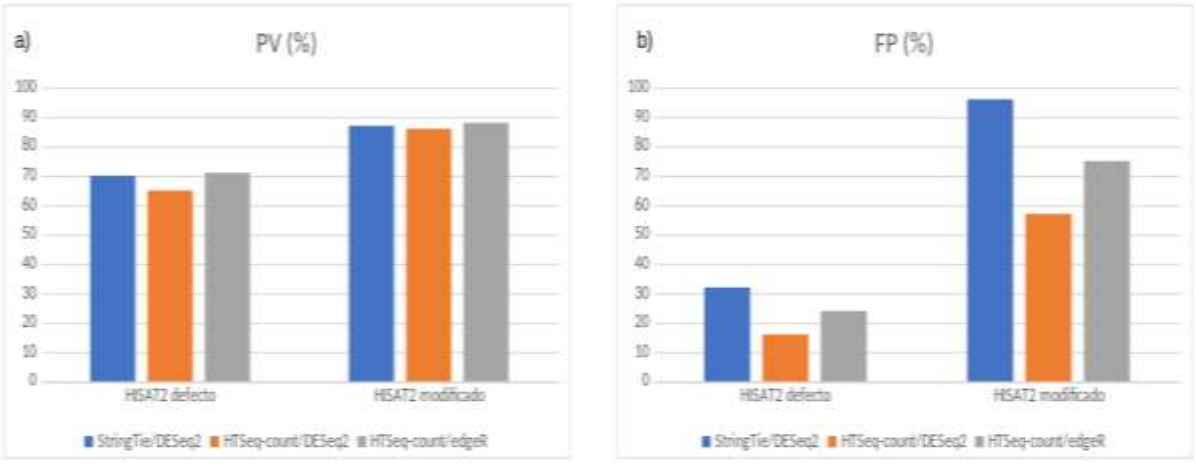

Figura 5. Resultados del análisis de expresión diferencial (DEGs) de las librerías simuladas en todas las variantes del flujo de trabajo mostrados en positivos verdaderos (a) y falsos positivos (b).

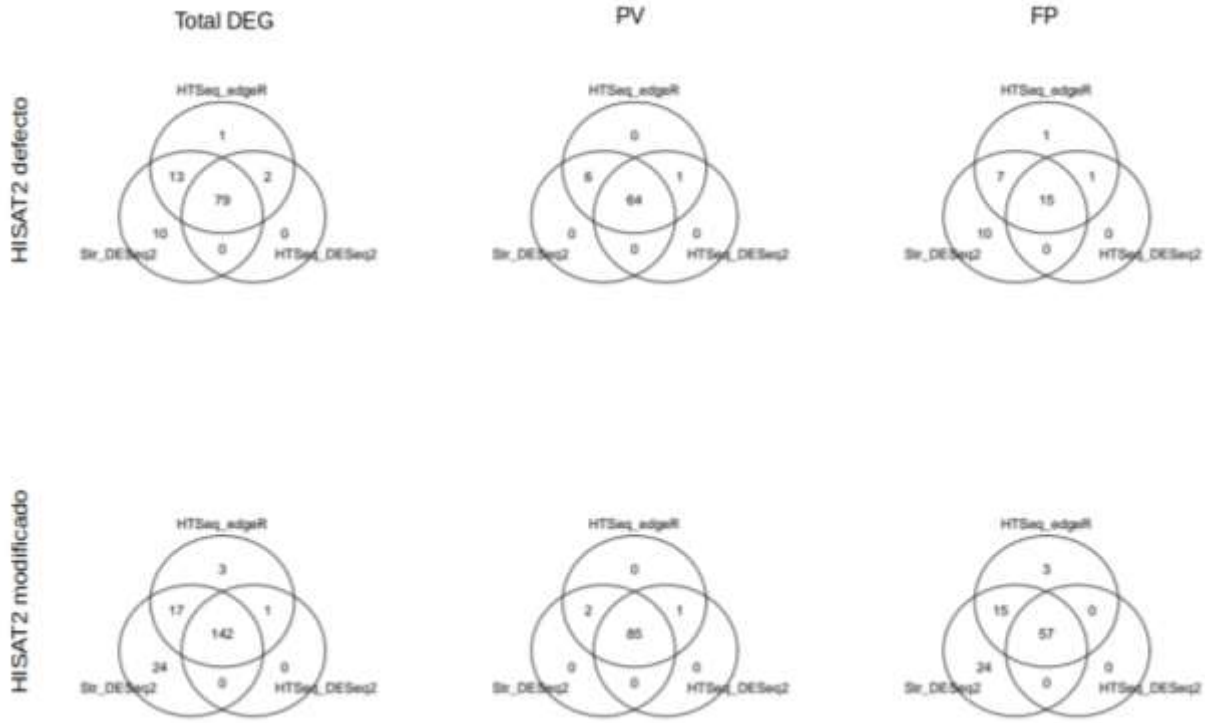

Figura 6. Diagramas de Venn para la comparación de los resultados de ED, PV y FP en las diferentes variantes del flujo de trabajo para las librerías simuladas con los parámetros por defecto de HISAT2 y modificados.