REVISTA BOLIVIANA DE QUÍMICA
ISSN 0250-5460 Rev. Bol. Quim. *Paper edition*
ISSN 2078-3949 Rev. boliv. quim. *Electronic edition*
Rollano et Mollinedo RBQ Vol. 34, No.2, pp. 50-55, 2017

BIOINFORMATIC ANALYSIS OF RNA-SEQ WITH A PERSPECTIVE FOR BOLIVIA

ANÁLISIS BIOINFORMÁTICO DE ARN-SEQ CON UNA PERSPECTIVA PARA BOLIVIA

Received 06 13 2016
Accepted 06 27 2017
Published 06 30 2017

Vol. 34, No.2, pp. **50-55**, May./Jun. 2017
34(2), **50-55**, May./Jun. 2017
Bolivian Journal of Chemistry

Short review (Spanish)

Oscar M. Rollano Peñaloza*, Patricia Mollinedo Portugal

Laboratorio de Productos Naturales, Pruebas Biológicas, Instituto de Investigaciones en Productos Naturales IIPN, Ciencias Químicas, Facultad de Ciencias Puras y Naturales FCPN, Universidad Mayor de San Andrés UMSA, P.O. Box 303, Calle Andrés Bello s/n, Ciudad Universitaria Cota Cota, Tel. +59122795878, La Paz, Bolivia, omrollano@umsa.bo, pamollinedo@umsa.bo, pattymollinedo@gmail.com, www.umsa.bo

Keywords: RNA-seq, CnCaBo, Transcriptómica, Nueva generación de Secuenciación masiva, Bioinformática

ABSTRACT

RNA-seq has become the tool of choice for transcriptome studies. Transcriptome provides key information about global changes of an organism under certain conditions at a determined moment. Next-Generation Sequencing technology has made transcriptomic studies available for everyone, increasing the demands of researchers within the bioinformatic field. Therefore, in the present review we are describing the bioinformatics workflow, analysis and applications of RNA-seq. At the same time, we want to present a computing cluster capable to perform bioinformatic analysis placed in Bolivia, known as CnCaBo.

*Corresponding author: omrollano@umsa.bo

RESUMEN

***Spanish title**: Análisis bioinformático de ARN-SEQ con una perspectiva para Bolivia.* RNA-seq se ha convertido en la herramienta más utilizada para el estudio de transcriptomas. El transcriptoma nos provee información clave sobre los cambios globales de un organismo bajos ciertas condiciones en un determinado momento. Las nuevas tecnologías de secuenciación masiva han permitido que la investigación en el área de la transcriptómica esté al alcance de todos, incrementando la demanda de investigadores en el área de la bioinformática. Es por eso que el presente artículo de revisión pretende contribuir al conocimiento del análisis bioinformático de transcriptómica mediante RNA-seq y sus aplicaciones. Al mismo tiempo queremos dar a conocer la implementación de un clúster computacional para realizar análisis bioinformáticos localizado en Bolivia denominado CnCaBo.

INTRODUCCIÓN

La información biológica heredable de todos los seres vivos está codificada en secuencias de DNA segmentadas y localizadas en regiones específicas del cromosoma que cumplen funciones únicas, éstas secuencias se denominan genes. El mecanismo por el cuál estos genes son decodificados y traducidos en unidades estructurales y funcionales se denomina expresión génica. La expresión génica consiste en la transcripción de dichas secuencias de DNA en moléculas de RNA (también conocidas como transcritos), que pueden ser procesados en RNA funcional no codificante o traducidos en proteínas. Es así que las moléculas de DNA, RNA y proteínas constituyen los bloques esenciales para determinar la forma, funcionamiento y diferentes estrategias para la continuidad de la vida que definen a un organismo. La activación de la expresión génica, así como la cantidad de transcritos que se producen, es

un proceso estrictamente regulado que depende de las señales que el organismo reciba. El conjunto total de transcritos en un determinado momento bajo ciertas condiciones fisiológicas es conocido como transcriptoma. Así, un transcriptoma nos brinda la información completa acerca de la expresión génica de una célula, tejido u organismo en diferentes escenarios como ser las etapas del desarrollo, el progreso de una enfermedad y la respuesta a estímulos bióticos o abióticos. La información proveniente de los transcriptomas nos permite identificar nuevos genes, rutas metabólicas y cascadas de señalización representativas de ciertos procesos biológicos, la ciencia encargada de este estudio es la transcriptómica. Los genes expresados de forma significativa pueden ser utilizados como marcadores moleculares para ciertos rasgos de interés. Estos marcadores genéticos pueden ser utilizados para identificar, diferenciar, seleccionar variedades de plantas y generar nuevas variedades con los rasgos deseados [1-3].

Tradicionalmente el método elegido para el estudio de transcriptomas ha sido los chips de DNA (microarrays), que constan de fragmentos de DNA unidos a una superficie sólida, dónde cada fragmento representa un gen o región de un genoma de referencia, que al entrar en contacto con una secuencia complementaria de DNA copia del RNA (cDNA) proveniente de la muestra se hibridizan y generan una señal fluorescente que es cuantificable de forma relativa mediante la intensidad de la señal fluorescente bajo diferentes condiciones. Los chips de DNA tienen ciertas limitaciones, como ser la baja sensibilidad para detectar genes expresados en un bajo número de copias y la dependencia de un genoma de referencia anotado [4-6].

Éstas limitaciones desaparecieron cuando se introdujo la tecnología del RNA-seq para el estudio de transcriptomas que además generó nuevos descubrimientos como el empalme alternativo (splicing alternativo) en diferentes tejidos [7, 8]. La investigación transcriptómica con RNA-seq se ha facilitado bastante debido a que ya existen más de 50 plantas con su genoma completamente secuenciado [9] y se estima que sobrepasen los 100 genomas para el 2018. Además, este método permite la investigación de especies relativamente desconocidas, incrementando el conocimiento sobre las funciones de genes desconocidos y el descubrimiento de nuevos genes [10]. El RNA-seq es una herramienta que consta de la secuenciación masiva del RNA global de un organismo, que normalmente genera varios millones de secuencias (en inglés: reads) por muestra. El método se puede resumir como la extracción del RNA total de una muestra, la captura y purificación del RNA deseado (usualmente mRNA), la conversión a cDNA y finalmente la secuenciación masiva.

La secuenciación varía mucho dependiendo de la tecnología empleada, por lo cual existen varias investigaciones y publicaciones enfocadas en la descripción de la química de secuenciación de las diferentes compañías [11-14]. La diferencia más marcada entre los métodos de secuenciación masiva depende de si es para una sola molécula o varias moléculas amplificadas previamente por PCR (Reacción en Cadena de la Polimerasa). La secuenciación masiva con amplificación previa es la tecnología más utilizada, en la cual el cDNA es fragmentado, ligado a adaptadores específicos y amplificado, posteriormente se procede a un armado de librerías para su secuenciación masiva. Ésta tecnología es conocida como la nueva generación de secuenciación (NGS: Next-Generation Sequencing) que es ofrecida por compañías como SOLiD, IonTorrent e Illumina. Illumina es la tecnología más utilizada que prácticamente ha dominado el mercado [11, 14, 15]. Por otra parte, la secuenciación de una sola molécula es conocida como la Tercera Generación de Secuenciación (TGS) que consta en la ligación del cDNA a adaptadores, dónde éste puede ser o no fragmentado y se procede directamente a secuenciar [16]. Las tecnologías disponibles para secuenciación de una sola molécula en el mercado son Pacific Biosciences, Oxford Nanopore y Helicos Biosciences [20] comercializada bajo la marca Seqll. Las ventajas de la secuenciación de una sola molécula (TGS) comparada con la tecnología NGS son la identificación de transcritos pobremente amplificados por PCR, la eliminación del complejo proceso de la preparación de librerías de cDNA [17] y la gran longitud de las secuencias obtenidas [18,19]. Entre las desventajas de las tecnologías TGS está la alta tasa de error que se producen en los extremos de las moléculas y el menor volumen de información (throughput) generado por cada corrida.

Los equipos para realizar secuenciación masiva todavía tienen costos muy altos, por lo tanto, una mejor alternativa para investigaciones que cuentan con presupuestos limitados es el envío de muestras a centros de secuenciación NGS. Actualmente existen varios centros que se dedican exclusivamente a realizar secuenciación NGS (ej. Sequentia Biotech-España, IGA technology services - Italia, Beijing Genomics Institute - China, CATG - Argentina, Omega Bioservices - USA), los cuáles realizan secuenciación masiva de alta calidad a partir de muestras de tejido, DNA o RNA. La información genómica generada por estos centros puede ser descargada desde cualquier parte del mundo para proceder a su análisis bioinformático. Todo el software necesario es de distribución libre y está en desarrollo permanente, además existe una gran comunidad bioinformática constantemente activa que aporta al aprendizaje y a la resolución de problemas.

Sin embargo, la información bioinformática en bases de datos para los países hispanos y en idioma español es escasa, por lo cual en el siguiente artículo describimos el procedimiento y análisis bioinformático para la investigación transcriptómica mediante RNA-seq.

ANÁLISIS BIOINFORMÁTICO PARA RNA-SEQ

El análisis bioinformático para RNA-seq, empieza con la filtración de secuencias con calidad aceptable para alinear fidedignamente al genoma o transcriptoma de referencia. Las secuencias alineadas deben ser cuantificadas, normalizadas y sometidas a un análisis estadístico para conocer los genes que han sido expresados significativamente. Finalmente, se identifican los roles y funciones de los genes expresados.

Control de calidad

Las secuencias obtenidas (reads) pueden tener errores de tipo instrumental o fallas técnicas, por lo cual es necesario evaluar la calidad de las mismas. El software más utilizado para estas tareas es FastQC que permite obtener parámetros de calidad por nucleótido secuenciado. En caso de que la calidad de las secuencias fuera muy baja, existen diversas herramientas que ayudan a eliminar ciertas secuencias o a recortar el extremo de las secuencias que tiene una baja calidad con software como Prinseq ó Trimmomatic.

Mapeo y visualización

El mapeo genómico consiste en alinear las secuencias (reads) obtenidas mediante NGS con el genoma (o transcriptoma) de referencia. Mapeando las secuencias obtenemos información de su localización en el genoma, la cual puede ser utilizada para identificar y cuantificar los genes expresados, conocer las variantes alélicas y además descubrir nuevos genes. En especies poco estudiadas puede ser que el genoma de referencia aún no esté disponible, en estos casos es posible alinear a un transcriptoma ensamblado *de novo* [21]. La primera consideración para elegir un software de alineamiento es el splicing alternativo. En organismos que no tienen intrones es mejor utilizar alineadores contiguos como Bowtie2 [22] ó BWA [23] que fueron diseñados para alinear DNA. Sin embargo, cuando se tienen genomas que poseen intrones es mejor utilizar alineadores como Tophat2 [24], Hisat2 [25], STAR [26], GSNAP [27], SOAP2 [28] o Kallisto [29].

Manipulación de secuencias alineadas y detección de polimorfismos de un solo nucleótido (SNPs)

La alineación provee grandes cantidades de información que requieren ser procesadas de distintas maneras para su análisis posterior. Los millones de secuencias alineadas pueden ser ordenadas de varias maneras y pueden proveernos valiosa información como inserciones y deleciones, las herramientas más utilizadas son Samtools [30] y Picard. Esta información puede ser utilizada por programas como Bedtools para detectar SNPs, que a su vez se pueden utilizar para detectar todo tipo de mutaciones y variantes en el genoma analizado automáticamente.

Visualización del alineamiento

La visualización de un genoma alineado es altamente recomendado porque nos permite comprobar los datos provenientes del análisis automático. Además de identificar inserciones y deleciones, SNPs, nuevas variantes de splicing alternativo y otras diferencias con el genoma de referencia de forma manual. Existen diferentes tipos de software con diferentes ventajas pero uno de los más recomendados es el IGV (Integrative Genomics Viewer) [31].

Conteo de secuencias y normalización

Este conteo nos permite identificar cuántas secuencias han sido alineadas a una parte o al genoma completo. La cantidad de secuencias (RNA-seq) alineadas a un gen especifico nos determinará su nivel de expresión génica. Uno de los programas más utilizados es el Htseq [32]. Para realizar un análisis apropiado de expresión génica mediante RNA-seq, lo usual es una normalización debido a los diversos tamaños de los genes y a la variación de la profundidad (depth) en la secuenciación de cada muestra. Diferentes unidades se han propuesto como RPKM (Secuencias (reads) por Kilobase de transcrito por Millón de secuencias mapeadas) y FPKM (Fragmentos de transcrito por Millón de secuencias mapeadas), sin embargo dos unidades son las más utilizadas: Transcritos por Millón (TPM) y Medias de los valores *M* ajustados (TMM) debido a la menor variación obtenida en la normalización de los genes de referencia. Si se utiliza un transcriptoma *de novo* para el alineado, el software recomendado es RSEM [33] que también facilita valores de normalización.

Expresión génica diferencial

El análisis de expresión génica diferencial (DGE: por sus siglas en inglés) se refiere a la identificación global de genes (transcriptoma) que se expresan significativamente en ciertas condiciones comparadas a un control experimental. DGE puede ser realizado para observar expresión génica entre diversos tejidos de un mismo organismo y así detectar splicing alternativo [7], para conocer los cambios globales en diferentes estadios del desarrollo, organismos expuestos a diferentes condiciones químicas o biológicas, individuos sanos comparados con enfermos, y otros. Para realizar éste análisis existen diferentes enfoques que vienen incluidos en software que usualmente está basado en el lenguaje de programación gratuito R [34]. Los dos programas más utilizados son edgeR [35] y DESeq que tienen una base de modelos lineares basado en un programa originalmente diseñado para estudiar Microarrays (Limma). Sin embargo, existen otras herramientas disponibles: Cufflinks [36], ALEXA-seq [37]. Las grandes cantidades de análisis múltiples que se realizan requieren de análisis estadísticos avanzados para descubrir las tasas de descubrimientos falsos (False Discovery Rate: FDR por sus siglas en inglés) [38].

Anotación genómica

La anotación genómica en RNA-seq se refiere al proceso de identificación de los genes expresados, y a conocer la información disponible sobre su función en los genomas de referencia y bases de datos públicas.

La ventaja de trabajar con genomas conocidos es que una buena parte de la anotación genómica se realiza junto a la secuenciación del genoma. Mientras que para genomas poco estudiados el proceso inicia realizando un alineamiento de las secuencias de genes que se identificaron como significativamente expresadas con software como BLAST [39]. Luego se procede a identificar las familias a las cuales los genes y sus productos pueden pertenecer mediante software como HMMER, InterProScan [40] y SignalP. Cuando se trabaja con transcriptomas *de novo*, es recomendado resumir ésta información con Trinotate.

Análisis de ontología génica (GO)

Grandes cantidad de genes requieren ser agrupados para conocer su significado biológico. Es así que el análisis de ontología génica (GO por sus siglas en inglés) nos permite agrupar genes por proceso biológico, función molecular y componente celular. Además también se puede analizar rutas metabólicas (KEGG) [41] y otras agrupaciones más complejas (PANTHER) [42]. Herramientas como AgriGO nos permiten visualizar mediante arboles jerárquicos los términos GO significativamente expresados.

A través de todos estos pasos, el análisis bioinformático de RNA-seq nos permite generar listas de todos los genes expresados y sus funciones para cada organismo y condición estudiada, nos permite identificar las inducciones o represiones en la expresión de los genes estudiados de manera cuantitativa y nos da la posibilidad de identificar SNPs en las secuencias de estos. Finalmente, la agrupación de genes expresados nos da pautas acerca de patrones de co-expresión, cascadas de señalización y rutas metabólicas activadas o reprimidas durante la condición fisiológica dada.

De todas maneras, los recursos computacionales para estos análisis aún requieren grandes cantidades de memoria RAM y espacio en disco duro. Por ejemplo, la base de datos de secuencias en investigación contra el Cáncer por si misma supera los 2 petabytes [43]. A su vez, estos análisis también necesitan un número considerable de núcleos que puedan funcionar constantemente, en algunos casos por días enteros. Por lo cual los análisis son generalmente realizados en clústeres computacionales regionales o en computación en la nube (Cloud Computing) mediante servicios web como el servicio web de Amazon (AWS) y otros servicios web de libre acceso (ej. Chipster [44] y Galaxy [45]). En lo que a nuestro conocimiento respecta, en Bolivia no existen clústeres computacionales para investigación bioinformática. Es por eso que hemos visto la necesidad de implementar un clúster computacional para investigación bioinformática de libre acceso denominado CnCaBo. El objetivo de CnCaBo s proveer los recursos computacionales tanto en almacenamiento como en procesamiento para cumplir con las demandas que requieran los investigadores en genómica y bioinformática de toda Bolivia. CnCaBo cuenta con todo el software y hardware necesario para realizar RNA-seq, además de una página web que incluye guías bioinformáticas y a la fecha se encuentra en línea y funcionando.

RECONOCIMIENTOS

Los autores agradecen a Sergio Alvarez Molina por su ayuda en la implementación de CnCaBo, a Enzo Aliaga e Isabel Morales Belpaire por sus comentarios sobre el manuscrito. A la Agencia Sueca de cooperación para el Desarrollo Internacional (SIDA) por el respaldo financiero.

REFERENCIAS

1. Desta, Z.A. & Ortiz, R. **2014**, Genomic selection: genome-wide prediction in plant improvement, *Trends in Plant Science*, *19* (9), 592-601.
2. Varshney, R.K., Terauchi, R. & McCouch, S.R. **2014**, Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding, *PLOS Biology*, *12* (6), e1001883.
3. Jannink, J.-L., Lorenz, A.J. & Iwata, H. **2010**, Genomic selection in plant breeding: from theory to practice, *Briefings in Functional Genomics*, *9* (2), 166-177..
4. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. **2009**, mRNA-Seq whole-transcriptome analysis of a single cell, *Nature Methods*, *6* (5), 377-382.
5. Wang, Z., Gerstein, M. & Snyder, M. **2009**, RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*, *10* (1), 57-63.
6. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. **2014**, Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells, *PLOS ONE*, *9* (1), e78644.
7. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. **2008**, Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*, *5* (7), 621-628.
8. Jain, M. **2012**, Next-generation sequencing technologies for gene expression profiling in plants, *Briefings in Functional Genomics*, *11* (1), 63-70.
9. TüRktaş, M., YüCebilgili KurtoğLu, K., Dorado, G., Zhang, B., Hernandez, P. & üNver, T. **2015**, Sequencing of plant genomes – a review, *Turkish Journal of Agriculture and Forestry*, *39*, 361-376.
10. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. **2011**, Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nature Biotechnology*, *29* (7), 644-652.
11. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. & Vakhlu, J. **2016**, High Throughput Sequencing: An Overview of Sequencing Chemistry, *Indian Journal of Microbiology*, *56* (4), 394-404.
12. Mardis, E.R. **2013**, Next-Generation Sequencing Platforms, *Annual Review of Analytical Chemistry*, *6* (1), 287-303.
13. Metzker, M.L. **2010**, Sequencing technologies — the next generation, *Nature Reviews Genetics*, *11* (1), 31-46.
14. Ozsolak, F. & Milos, P.M. **2011**, RNA sequencing: advances, challenges and opportunities, *Nature Reviews Genetics*, *12* (2), 87-98.
15. Bentley, D.R. & Balasubramanian, S. & Swerdlow, H.P. & Smith, G.P. & Milton, J. & Brown, C.G. & Hall, K.P. & Evers, D.J. & Barnes, C.L. & Bignell, H.R., et al. **2008**, Accurate whole human genome sequencing using reversible terminator chemistry, *Nature*, *456* (7218), 53-59.
16. Schadt, E.E., Turner, S. & Kasarskis, A. **2010**, A window into third-generation sequencing, *Human Molecular Genetics*, *19* (R2), R227-R240.
17. Fu, G.K., Xu, W., Wilhelmy, J., Mindrinos, M.N., Davis, R.W., Xiao, W. & Fodor, S.P.A. **2014**, Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations, *Proceedings of the National Academy of Sciences*, *111* (5), 1891-1896.
18. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. **2009**, Real-Time DNA Sequencing from Single Polymerase Molecules, *Science*, *323* (5910), 133-138.
19. Loose, M., Malla, S. & Stout, M. **2016**, Real-time selective sequencing using nanopore technology, *Nature Methods*, *13* (9), 751-754.
20. Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P. & Causey, M. **2009**, Quantification of the yeast transcriptome by single-molecule sequencing, *Nature Biotechnology*, *27* (7), 652-658.
21. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. **2013**, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nature Protocols*, *8* (8), 1494-1512.
22. Langmead, B. & Salzberg, S.L. **2012**, Fast gapped-read alignment with Bowtie 2, *Nature Methods*, *9* (4), 357-359.
23. Li, H. & Durbin, R. **2009**, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, *25* (14), 1754-1760.
24. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. **2013**, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biology*, *14* (4), R36.
25. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. **2013**, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biology*, *14* (4), R36.
26. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. **2013**, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics (Oxford, England)*, *29* (1), 15-21.
27. Wu, T.D. & Nacu, S. **2010**, Fast and SNP-tolerant detection of complex variants and splicing in short reads, *Bioinformatics*, *26* (7), 873-881.
28. Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. & Wang, J. **2009**, SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*, *25* (15), 1966-1967.
29. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. **2016**, Near-optimal probabilistic RNA-seq quantification, *Nature Biotechnology*, *34* (5), 525-527.
30. Li, H. **2011**, Improving SNP discovery by base alignment quality, *Bioinformatics*, *27* (8), 1157-1158.

31. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. **2013**, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Briefings in bioinformatics*, *14* (2), 178–192.
32. Anders, S., Pyl, P.T. & Huber, W. **2015**, HTSeq--a Python framework to work with high-throughput sequencing data, *Bioinformatics*, *31* (2), 166-169
33. Li, B. & Dewey, C.N. **2011**, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinformatics*, *12* (1), 323.
34. R Core Team. 2016, R: A Language and Environment for Statistical Computing, Journal, (Issue).
35. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. **2010**, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, *26* (1), 139-140.
36. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. & Pachter, L. **2012**, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature Protocols*, *7* (3), 562-578.
37. Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R., Morrissy, A.S., Morin, R.D., Corbett, R., Tang, M.J., Hou, Y.-C., Pugh, T.J., et al. **2010**, Alternative expression analysis by RNA sequencing, *Nature Methods*, *7* (10), 843-847.
38. Benjamini, Y. & Hochberg, Y. **1995**, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, *57* (1), 289-300.
39. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. **2009**, BLAST+: architecture and applications, *BMC Bioinformatics*, *10* (1), 421.
40. Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. **2015**, The InterPro protein families database: the classification resource after 15 years, *Nucleic Acids Research*, *43* (D1), D213-D221.
41. Kanehisa, M. **2004**, The KEGG resource for deciphering the genome, *Nucleic Acids Research*, *32* (90001), 277-280.
42. Mi, H., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. **2013**, Large-scale gene function analysis with the PANTHER classification system, *Nature Protocols*, *8* (8), 1551-1566.
43. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J., et al. **2016**, The real cost of sequencing: scaling computation to keep pace with data generation, *Genome Biology*, *17* (1).
44. Kallio, M.A., Tuimala, J.T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., Koski, M., Käki, J. & Korpelainen, E.I. **2011**, Chipster: user-friendly analysis software for microarray and other high-throughput data, *BMC Genomics*, *12* (1), 507.
45. Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. **2016**, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, *Nucleic Acids Research*, *44* (W1), W3-W10.
46. Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., et al. **2011**, Genome sequence and analysis of the tuber crop potato, *Nature*, *475* (7355), 189-195.
47. Motamayor, J.C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone Iii, D., Cornejo, O., Findley, S., Zheng, P., Utro, F., Royaert, S., et al. **2013**, The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color, *Genome biology*, *14* (6): r53.
48. Argout, X., Salse, J., Aury, J.-M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., et al. **2011**, The genome of Theobroma cacao, *Nature Genetics*, *43* (2), 101-108.
49. Clouse, J.W., Adhikary, D., Page, J.T., Ramaraj, T., Deyholos, M.K., Udall, J.A., Fairbanks, D.J., Jellen, E.N. & Maughan, P.J. **2016**, The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly, *The Plant Genome*, *9* (1), doi: 10.3835/plantgenome2015.07.0062.
50. Yasui, Y., Hirakawa, H., Oikawa, T., Toyoshima, M., Matsuzaki, C., Ueno, M., Mizuno, N., Nagatoshi, Y., Imamura, T., Miyago, M., et al. 2016, Draft genome sequence of an inbred line of Chenopodium quinoa , an allotetraploid crop with great environmental adaptability and outstanding nutritional properties, *DNA Research*, dsw037.
51. Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., Schmöckel, S.M., Li, B., Borm, T.J.A., Ohyanagi, H., Mineta, K., Michell, C.T., Saber, N., et al. 2017, The genome of Chenopodium quinoa, *Nature*, *542* (7641), 307-312.