

Capítulo I

Recursos genéticos

Variación somaclonal en menta (*Mentha* L.)

Luis Felipe López Hernández y Andrés Javier Cortés Vera

Mensajes clave sobre la variación somaclonal en menta:

- La variación somaclonal es todo aquel polimorfismo genético que no se da por mutación en las células germinales sino por mutación en la línea somática.
- La menta es una hierba fuertemente perfumada de la familia Lamiaceae, que se cultiva principalmente por propagación clonal, lo que la hace una especie valiosa para el estudio de variantes somaclonales y sus consecuencias fenotípicas.
- Las introducciones recientes de pocas especies de menta en América del Sur, seguidas de una propagación clonal desenfrenada, hacen que esta región sea ideal para estudiar el alcance de la diversidad genética somaclonal.
- Este capítulo caracteriza la diversidad somaclonal codificante de la menta en el norte de los Andes, para identificar si habían surgido variaciones somaclonales a pesar de introducciones relativamente recientes.
- Un total de 29 materiales colectados en granjas de exportación de menta, cultivada por propagación clonal en Antioquia (Colombia), han sido genotipificados vía RNAseq.
- La captura de 2.033 loci en 912 transcritos a profundidad de 20X fue efectuada mediante la combinación de protocolos GATK4 + Trinity.
- A partir de los 2.033 loci se realizó un análisis de distancia por UPGMA con 10.000 *bootstrap* y los algoritmos de agrupamiento de machine learning: K-means y AGNES, y se identificaron tres grupos genéticos para *M. spicata* y un grupo genético para *M. piperita*.
- Lo anterior muestra que el origen de la *M. spicata* que se cultiva en la región del Oriente antioqueño para exportación y consumo local tiene posiblemente dos ancestrías que se propagaron clonalmente por la zona.
- Uno de los orígenes de *M. spicata* presentó mayor estructura poblacional que el otro, lo que sugiere que posiblemente existan presiones ambientales o antropológicas que intervinieron en la fragmentación de este grupo genético en la región.

La tasa, el alcance y la arquitectura de las mutaciones *de novo*, así como las variaciones somaclonales, han sido durante mucho tiempo preguntas intrigantes en la evolución molecular. Las características genómicas y las restricciones cromosómicas son determinantes en el conocimiento de estos niveles ocultos de variación genética. Los reordenamientos de segregación inusuales y los altos niveles de ploidía también se consideran potenciadores de la diversidad de nucleótidos en escalas de tiempo evolutivo.

Sin embargo, las variaciones somaclonales en periodos de tiempo más superficiales siguen siendo esquivas (Plomion et al., 2018), debido en parte a su supuesta rareza y a la dificultad para identificarlas. Además, a menudo se supone que son neutrales y su valor adaptativo potencial rara vez se ha probado.

El uso de información transcriptómica hace posible a una mayor resolución capturar la variación somaclonal en variantes alélicas. Estudios comparativos como el de Zhao et al. (2019) sugieren que el sistema más eficaz de reconstruir la matriz de variantes alélicas, a partir de la secuenciación de RNA, es la integración de los algoritmos Trinity como ensamblador *de novo* y GATK (*Genome Analysis Toolkit*) como protocolo de llamado de SNPs. La ejecución del protocolo integrado Trinity + GATK recuperó SNPs con una precisión del 100 % en casos de melocotón y mandarina, y fue la primera vez en ser probado para hierbas perfumadas como la menta.

Los cultivos mantenidos por propagación clonal constituyen campos de juego experimentales únicos para apuntar a variantes somaclonales y explorar sus consecuencias fenotípicas. En particular, las especies que se cultivan por sus órganos no reproductivos tienen graves interrupciones en sus sistemas de floración y fructificación, lo que hace que la clonalidad sea una estrategia de propagación obligatoria. La menta, una hierba fuertemente perfumada de la familia Lamiaceae, es un ejemplo de esto. Varias especies se han utilizado durante siglos con fines medicinales y gastronómicos, incluyendo alrededor de 30 especies y especies híbridas que se distribuyen alrededor del globo (Vining et al., 2020). Más allá de su uso como hierbas, especias y para necesidades farmacéuticas, la destilación de aceites esenciales de la menta comercial (mentol) es hoy en día un producto económico mundial. Sin embargo, el potencial bioeconómico de la menta sigue siendo enorme, debido a los metabolitos secundarios y los usos novedosos que aún no se han descubierto.

A pesar del área de distribución nativa cosmopolita de la menta, el único subcontinente donde tuvieron que introducirse como parte del intercambio colombino fue América del Sur. Las introducciones recientes de pocas especies de menta, seguidas de una propagación clonal desenfrenada, hacen de esta

región un escenario ideal para estudiar el alcance de la diversidad genética somaclonal. Por lo tanto, se caracteriza la diversidad somaclonal codificante de la menta en el norte de los Andes, para abordar las siguientes preguntas: ¿Habían surgido variaciones somaclonales a pesar de introducciones relativamente recientes? ¿En qué medida albergan potencialmente consecuencias fenotípicas?

Herramientas para el estudio de la variación somaclonal

A partir del recurso genético de menta en el noroeste de los Andes, se desarrolló una investigación para el estudio de la variación somaclonal con 38 materiales colectados en 14 fincas productoras en la región del oriente del departamento de Antioquia (Colombia), entre septiembre y diciembre de 2019, y bajo condiciones protegidas y de libre exposición. De los 38 materiales colectados, 36 correspondían a *Mentha spicata* y 2 a *Mentha piperita*.

Personal especializado de la Universidad Católica de Oriente (UCO) y productores de cada predio identificaron los materiales en cada especie. Las muestras colectadas consistieron en tallos con raíces, las cuales fueron entregadas a la UCO para su reproducción clonal, con el fin de disponer de suficiente material endurecido para su procesamiento por secuenciación de ARN (ARNseq), en conjunto con la colección de referencia de la UCO para las especies mencionadas.

Estudio de la diversidad transcriptómica de la menta

Con el fin de recuperar la mayor cantidad y calidad de RNA de la menta colectada, se realizó un experimento comparativo de protocolos de extracción. Se efectuaron seis protocolos diferentes de extracción de RNA, en el laboratorio de genética molecular del Centro de Investigación Tibaitatá de AGROSAVIA (Mosquera, Colombia), en dos muestras aleatorias de *M. piperita* y *M. spicata*, además de las referencias proporcionadas por la UCO para las mismas especies. Los protocolos evaluados fueron: 1) casero Agrosavia; 2) casero Agrosavia, modificado en volúmenes y tiempo; 3) casero Agrosavia con variaciones para especies con altos contenidos de polisacáridos y polifenoles; 4) casero Agrosavia con Trizol; 5) kit comercial Qiagen RNeasy Plant Mini Kit, y 6) kit comercial Qiagen RNeasy Plant Mini Kit con Trizol. El ARN genómico se extrajo a 29 individuos a partir de las hojas de las muestras almacenadas a -80 ºC, mediante el kit de extracción comercial de ARN Qiagen RNeasy Plant Mini Kit. La cuantificación de ADN extraído se realizó por método de espectrofotometría con el equipo Nanodrop® 2000 (Thermo Fischer Scientific, Estados Unidos), y por método fluorimétrico con el fluorómetro Qubit® dsDNA HS (Life Technologies, Suecia). La construcción de librerías se realizó a partir

del kit Sure Select Strand-Specific RNA® para secuenciación multiplexada por Illumina®. Las librerías se cuantificaron por medio del método fluorimétrico en el fluorómetro Qubit® dsDNA HS. La concentración y los tamaños de los fragmentos de las librerías de cDNA se evaluaron en el equipo TapeStation 4200® (Agilent Technologies, Estados Unidos) y el kit High Sensitivity D1000. Las secuencias de ADN fueron obtenidas mediante el secuenciador Hiseq de Illumina 2500 (Macrogen, Corea del Sur) en un solo sentido (*single-end*).

Con el fin de depurar los datos de la secuenciación de RNAseq, se corrió una rutina automatizada usando el programa Trimmomatic (Bolger et al., 2014) con los parámetros principales: ILLUMINACLIP:TruSeq3-SE:2:30:10, ventana SLIDINGWINDOW:4:20 y MINLEN:50. Posteriormente, se realizó un análisis de la calidad de los archivos *fastq* mediante el programa FastQC (Andrews, 2010), a través de la codificación 1,9 de Illumina. Estudios comparativos como el de Zhao et al. (2019) sugieren que el sistema más eficaz para reconstruir la matriz de variantes genómicas informativas (SNPs), a partir de RNAseq, es la integración de los algoritmos Trinity + GATK. De este modo, se obtuvo el transcriptoma *de novo* de las 29 muestras de menta por medio del programa Trinity (Grabherr et al., 2011) en la plataforma en el proyecto Galaxy versión 2.9.1 (Afgan et al., 2018).

El supertranscrito es útil al ser una referencia para mapear lecturas e identificar polimorfismos alélicos en el contexto de ensamblaje *de novo* sin genoma de referencia. Un supertranscrito se construye colapsando regiones de secuencias únicas y comunes entre isoformas de empalme en una sola secuencia lineal (Davidson et al., 2017). De este modo, en la misma plataforma Galaxy versión 2.9.1 (Afgan et al., 2018), se obtuvo el supertranscrito mediante Trinity para ser usado como referencia en el protocolo de llamado de variantes de GATK4.

Con el fin de identificar polimorfismos alélicos entre las 29 muestras de menta, se construyó un script para automatizar el proceso de llamado de variantes con la función *Haplotype Caller* del protocolo GATK4 (McKenna et al. 2010) con BWA-MEM. Los estadísticos del mapeo se obtuvieron a partir de la función *flagstat* (Li et al., 2009) en la plataforma Samtools 1,9, en el proyecto Galaxy versión 2.0.3 (Afgan et al., 2018). La matriz de SNPs se filtró usando el programa Tassel 5.2.78 (Bradbury et al., 2007), a una profundidad de 20X y porcentaje de datos perdidos por loci de 96 %.

A partir de la identificación de polimorfismos alélicos consolidados en una matriz de SNPs, se realizó un análisis de reducción dimensional lineal por componentes principales moleculares PCA con la función *glPca* del paquete R *adegenet 2.1.4* (Jombart, 2008). Posteriormente, con el fin de encontrar una subestructura en los conjuntos de *Mentha spicata* y *Mentha piperita*, se realizó un agrupamiento particional y jerárquico a partir de las variables

reducidas. Paralelamente, se corrió un proceso de validación de clustering por medio de la función *optCluster* del paquete R *optCluster* (Sekula et al., 2017), una versión optimizada de la función *clValid* (Brock et al. 2008). Los algoritmos comparados en *optCluster* fueron los algoritmos particionales: K-means (MacQueen, 1967; Lloyd, 1982) y k-medoids (Kaufman & Rousseeuw, 2009); los algoritmos jerárquicos: AGNES (AGglomerative NESting) y DIANA (Divisive analysis) (Kaufman & Rousseeuw, 2009). Los métodos de validación usados en *optCluster* fueron algoritmo genético y entropía cruzada. También se ejecutó un algoritmo usado ampliamente en ciencia de datos para validación de clustering llamado *NbClust* (Charrad et al. 2014). Por otro lado, se reconstruyó un dendograma UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) con muestreo aleatorio *bootstrap* de 10.000 réplicas con la función *aboot* del paquete R *poppr* (Kamvar et al., 2014).

Recomendaciones para optimizar perfiles de ARN

Los 38 materiales colectados y las referencias de *M. piperita* y *M. spicata* se llevaron al invernadero de la UCO para aclimatación homogénea en las mismas presiones ambientales de humedad y temperatura por un periodo de nueve meses. La Unidad de Sanidad Vegetal reportó que todos los materiales se encontraban en estado óptimo de conservación y propagados clonalmente para disponer de suficiente material foliar para la preparación de librerías.

De los seis protocolos realizados a las muestras de menta, el kit comercial Qiagen® RNeasy Plant Mini Kit permitió obtener una mejor calidad de concentración y relaciones de absorbancia A260/A280 y A260/230. Posteriormente, se realizó la extracción de RNA mediante kit comercial a los 29 materiales en estado óptimo de conservación, donde se encontró una concentración media de ARN (ug/uL) de 922,0206897 (IC: 236,5158607), usando Nanodrop®, concentración media Qubit® (ug/uL) de 94,79172414 (IC: 8,775199689), relación media A260/280 de 2,11034483 (IC: 0,016035813) y relación media A260/230 de 2,014827586 (IC: 0,228618311).

Las 29 librerías genéticas construidas para secuenciación presentaron una concentración media Qubit® (ug/uL) de 18,77241379 (IC: 5,542889311), un tamaño medio de fragmento (bp) de 282,6551724 (IC: 3,099013281) y una cuantificación media TapeStation® de 99,96206897 nM (IC: 29,16482462). Los electroferogramas para cada material sugirieron la distribución de los fragmentos con picos definidos, sin presencia de contaminantes. La información genética de todas las muestras se componía de secuencias con score mayor a 30 usando la codificación 1,9 de Illumina, sin presencia de adaptadores, y con altos porcentajes de duplicación y desviación de GC, lo que era de esperarse en datos RNAseq.

El ensamblaje del transcriptoma sin genoma de referencia mediante Trinity, a partir de todos los *fastqc* depurados, se compone de 509.754 transcritos con una longitud media de 557,9 pb, mínima de 178 pb, máxima de 12.186 pb y porcentaje GC de 43,2 %. El transcriptoma presenta isoformas de empalme que podrían incrementar la tasa de falsos positivos en el llamado de variantes alélicas. Así, a partir del transcriptoma y los *fastqc* depurados, se construyó un supertranscrito que colapsa regiones de secuencias únicas y comunes entre isoformas de empalme en una sola secuencia lineal. El supertranscrito se compone de 352.512 transcritos con una longitud media de 472,2 pb, mínima de 201 pb, máxima de 15.765pb y porcentaje GC de 43,1 %.

Variación somaclonal revela grupos crípticos en menta

A partir de la información transcriptómica de los 29 materiales de menta, se obtuvo una matriz de variantes alélicas del supertranscrito construido a partir del mismo set de trabajo, usando la combinación de protocolos GATK + Trinity, dedicados a datos RNAseq para llamado de SNPs.

A partir de la información de variantes alélicas, también se hizo un análisis de reducción dimensional lineal por componentes principales como entrada a algoritmos de agrupamiento (por ejemplo, k-means, k-medoids y jerárquicos), así como análisis de distancia para visualización por dendograma. Todos los algoritmos muestran ampliamente la diferencia entre materiales de *M. piperita* y *M. spicata*, lo que comprueba el uso de *M. piperita* como control para la exploración de la diversidad genética de *M. spicata*. El análisis de búsqueda de número óptimo de clúster y el dendograma por UPGMA sugieren la presencia de tres posibles grupos de *M. spicata* y uno de *M. piperita*.

Siguiendo el protocolo GATK4, se desarrolló un *script* para automatizar el proceso para su reproducción. De los 29 materiales, se tuvo una media global de 15.588.705 (IC: 1.368.899) de lecturas, de las cuales se mapearon un total de 15.136.482 (IC: 1.403.825), lo que sugiere un porcentaje de mapeo del 96,81 %. Se presentó un alto contenido de lecturas duplicadas en el mapeo, con un promedio de 11.257.168 (IC: 1.166.846). Al ser removidos en el protocolo GATK4, se obtuvo un promedio de 3.879.314 (IC: 293.552) lecturas sin duplicados para el conjunto de 29 materiales. Finalmente, se obtuvo una matriz de variantes alélicas de 2.033 loci en 912 transcritos para las 29 muestras, a profundidad de 20X y porcentaje de datos perdidos por loci de 96 %.

Con el fin de explorar la variabilidad genética de la menta que actualmente se produce en el Oriente antioqueño, se realizó un análisis de agrupamiento vía PCA y agrupamiento por distancias genéticas mediante UPGMA.

A partir de los 2.033 loci distribuidos en 912 transcritos, se hizo una reducción dimensional de componentes principales, donde el primer componente explica el 58,552 % de la varianza, el segundo componente, un 22,274 % y el tercero, un 11,283 %. A partir de los tres primeros componentes principales (92,109 %), se corrió el análisis de validación de clustering, donde los algoritmos *NbClust* y *optCluster,* validados por algoritmo genético y entropía cruzada, sugirieron un total de cuatro clústeres.

Por otro lado, los mejores algoritmos de reconstrucción del agrupamiento fueron K-means y AGNES, los cuales generaron la misma organización de muestras. Finalmente, el enfoque UPGMA muestra cuatro grupos diferenciados por *bootstrap*, tal como se dispone en la figura 1.1. En los cuatro grupos genéticos obtenidos por enfoques de algoritmos de agrupamiento particional y por distancia genética, se mostró que el grupo *M. piperita* está ampliamente diferenciado del resto. Los grupos restantes son todos *M. spicata*, que se dividen en tres grupos (A, B y C), donde A y B tiene muy poca divergencia frente al primer componente y están distantes frente a C, por el segundo.

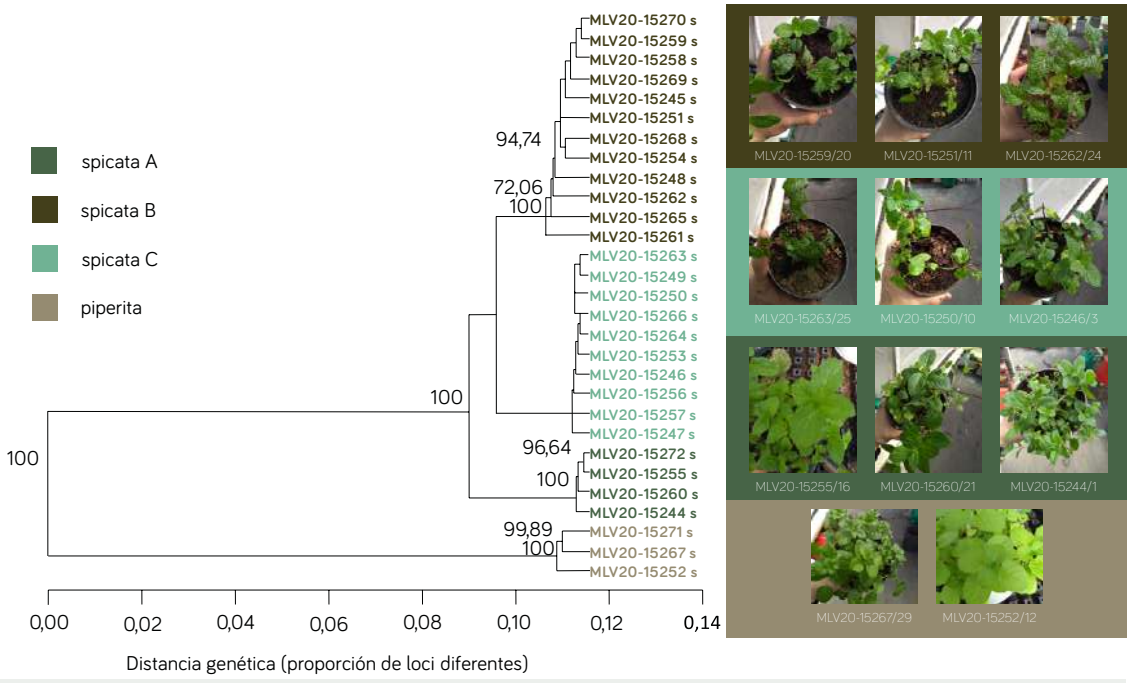


Figura 1.1. Dendograma reconstruido por UPGMA, a partir de los 2.033 loci distribuidos en 912 transcritos, con metodología de validación por remuestreo aleatorio *bootstrap*, con 10.000 iteraciones, basados en la distancia de *Nei*. Se evidencia la presencia de cuatro grupos genéticos validados por *bootstrap* de 100 %.

Fuente: Elaboración propia

Conclusiones

El origen de la menta que se cultiva en la región oriente de Antioquia para exportación y consumo local tiene posiblemente dos orígenes, tal vez debido a su propagación clonal por toda la zona productiva.

Uno de los posibles orígenes de *M. spicata* presentó mayor estructura poblacional que el otro, lo que sugiere presiones ambientales o antropológicas que intervinieron en la fragmentación de este grupo genético.

Referencias

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537-W544. https://doi.org/10.1093/nar/gky379

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2010). FastQC. A quality control tool for high throughput sequence data. *Babraham Bioinformatics*, *370*.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633-2635. https://doi.org/10.1093/bioinformatics/btm308

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, *25*, 1-22. https://www.jstatsoft.org/article/view/v025i04

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'nbclust'. *Journal of Statistical Software*, *61*, 1-36. https://www.jstatsoft.org/article/view/v061i06

Davidson, N. M., Hawkins, A. D., & Oshlack, A. (2017). SuperTranscripts: A data driven reference for analysis and visualisation of transcriptomes. *Genome Biology*, *18*(1), 1-10. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1284-1

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. *Nature Biotechnology*, *29*(7), 644-652.

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403-1405. https://doi.org/10.1093/bioinformatics/btn129

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://peerj.com/articles/281/

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Lloyd, S. P. (1982). Least-Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129-137.

MacQueen, J. B. (1967). *Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

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 Research*, *20*(9), 1297-1303. https://genome.cshlp.org/content/20/9/1297.full.pdf

Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., Faye, S., Francillonne, N., Labadie, K., Le Provost, G., Lesur, I., Bartholomé, J., Faivre-Rampant, P., Kohler, A., Leplé, J. C., Chantret, N., Chen, J., Diévart, A., Alaeitabar, T., Barbe, V., & Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, *4*, 440-452. https://www.nature.com/articles/s41477-018-0172-3

Sekula, M., Datta, S., & Datta, S. (2017). optCluster: An R package for determining the optimal clustering algorithm. *Bioinformation*, *13*(3), 101. https://pubmed.ncbi.nlm.nih.gov/28584451/

Vining, K. J., Hummer, K. E., Bassil, N. V., Lange, B. M., Khoury, C. K., & Carver, D. (2020). Crop wild relatives as germplasm resource for cultivar improvement in mint (*Mentha L.*). *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.01217

Zhao, Y., Wang, K., Wang, W. L., Yin, T. T., Dong, W. Q., & Xu, C. J. (2019). A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics*, *20*(1), 1-10. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5533-4