De novo transcriptome analysis of peduncle necking in cut *Rosa hybrida* cultivar 'H30'

B. Lear[1,a], A. Marchbank[3], N. Kent[3], K. Tansey[2], R. Andrews[2], P. Devlin[1], H. Rogers[3], A. Stead[1]

[1] School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, UK, TW20 0EX
[2] School of Medicine, Henry Wellcome Building, University Hospital of Wales, Cardiff, UK, CF14 4XW
[3] School of Biosciences, Sir Martin Evans Building, Museum Avenue, Cardiff, UK, CF10 3AX

Abstract

Bent-neck or 'necking' is a phenomenon often seen in cut roses, whereby the flower head droops due to a bending of the peduncle. Necking is thought to be caused by a blockage of the xylem, either due to an air embolism or an accumulation of contamination, limiting water uptake to the flower head and causing the stem to bend. Despite the use of biocides to reduce contamination levels, necking remains an issue for the cut flower sector and roses continue to die prematurely. As the occurrence of necking varies widely within and between cultivars, a transcriptome analysis of cut *Rosa hybrida* cultivar 'H30' has been carried out to investigate the potential molecular mechanisms involved. Peduncle samples of three stages of necking (straight, $<90°$ and $\geq90°$) were sequenced using next generation sequencing to produce over 100 million reads per stage and 203,565 contigs following Trinity *de novo* assembly. Differential expression analysis revealed nearly 2,000 significant transcripts (p adjust <0.05); providing a new resource for further analysis into the process of necking.

Keywords: *Rosa hybrida, de novo, transcriptomics, bent neck, necking*

INTRODUCTION

Roses are one of the UKs most popular cut flowers in terms of sales per stem and have been of significant cultural importance for centuries (Bendahmane et al., 2013; Hanks, 2015). However, bent-neck or 'necking' is a phenomenon often seen in cut rose flowers, whereby the flower head droops due to a bending of the peduncle. Necking is thought to be caused by a blockage in the xylem vessels; either due to an air embolism or an accumulation of contamination at, or within, the stem end. This limits the movement of water to the flower head and causes the stem to bend (Bleeksma and van Doorn, 2003). As this phenomenon often occurs within the first few days of vase life, it is a major cause of customer dissatisfaction within the cut flower industry. Although this association between bacterial contamination and necking has been widely studied, it is currently unknown if there are any molecular mechanisms involved in the process of necking within the peduncle.

De novo approaches allow for transcriptomic studies to be carried on non-model organisms as they do not require a reference genome. *De novo* assembly programmes such as Trinity use a multi-step process with de Bruijn graph analysis to build full length transcripts from a sample of raw sequenced reads (Grabherr et al., 2011). Koning-Boucoiran et al. (2015) used Trinity to successfully develop a *de novo* rose transcriptome and SNP array using *Rosa hybrida* petals, whole flowers and young leaves from 12 garden rose cultivars and stressed leaves from *Rosa multiflora*. Another extensive *de novo* transcriptomic study has also been carried out for *Rosa chinensis* focusing on mixture 13 different rose tissues under varying abiotic and biotic stress conditions (Dubois et al., 2012). However, no studies have currently been carried out on peduncle tissue undergoing necking.

[a]Email: Bianca.Lear.2015@live.rhul.ac.uk

Roses belong to the Rosaceae family, a large family including many important crop plants such as apple, peach and strawberry. An increase in Rosaceae genomic research over the past 14 years has meant that there are currently genomes available for seven genera, all of which are available on the 'Genome Database for Rosaceae' (GDR), an integrated web resource for Rosaceae genomics and genetic research (Jung et al., 2004; Jung et al., 2014). The diploid woodland strawberry (*Fragaria vesca*) genome was the closest related species available for rose studies (Shulaev et al., 2011), however the *Rosa chinensis* cv. 'Old Blush' genome has recently been published (Raymond et al., 2018). This diploid rose genome will therefore be viewed as a new analysis resource for the tetraploid *Rosa hybrida*, in relation to the *Fragaria vesca* and *Arabidopsis thaliana* TAIR10 (Lamesch et al., 2011) genomes.

MATERIALS AND METHODS

Plant material

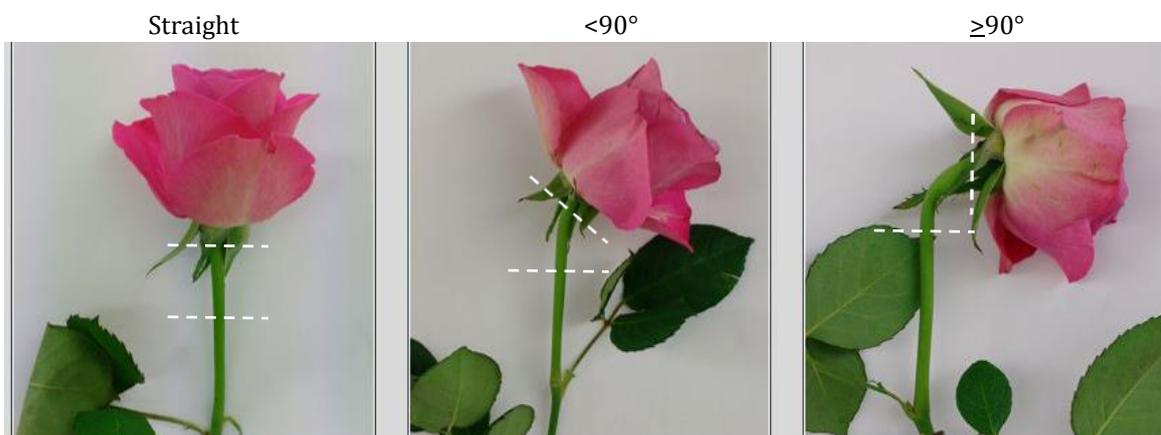| Straight	<90°	≥90°

Figure 1. Stages of necking for RNA sequencing
Three stages of necking were sampled, straight (control), <90° and ≥90°. For each stage of necking, 2cm sections were cut just below the flower head as shown by the dashed white lines.

Rosa hybrida cv. H30 stems were grown and harvested in Naivasha, Kenya and transported to the UK via air freight. After a 24 hour rehydration period, stems were re-cut to 47cm and leaves on the lower 25cm of stem removed. Stems were arranged in bunches of 10 and placed in vases containing a sachet of commercial flower food, made up to 1 litre with tap water. Vases were held at a constant 21°C for the duration of vase life, with a 12 hour light cycle (15-20 µmoles^{-2} sec^{-1}). For each stage of necking (straight, <90° and ≥90°), 2cm stem sections were cut using a sterile razor blade and flash frozen in liquid nitrogen (Figure 1). Samples were taken within 5-8 days of vase life, as the stages appeared, and stored at -80°C for RNA extraction. Vase life and sampling for replicate 1 was carried out in November, with replicates 2 and 3 repeated in October of the following year.

RNA extraction and sequencing

Samples were ground to a fine powder in liquid nitrogen using an autoclaved pestle and mortar. For each stage, three stem sections were ground to produce one pooled sample. This was repeated for each of the three replicates to form nine samples. Total RNA was extracted from each pooled sample using the extraction method devised by Jazi et al. (2015), adapted for use with 1.5ml microcentrifuge tubes. Genomic DNA contamination was removed from the RNA samples using the RapidOut DNA removal kit (Thermo Scientific) following the standard protocol. All

[a]Email: Bianca.Lear.2015@live.rhul.ac.uk

samples were quality tested using a Qubit fluorometer and then sequenced using an Illumina NovaSeq 5000 to produce paired-end reads for each sample.

De novo transcriptome workflow

Bioinformatic analysis was carried out using the Advanced Research Computing at Cardiff (ARCCA), a supercomputer running Unix. This was accessed remotely using MobaXterm Personal Edition v10.2 (https://mobaxterm.mobatek.net/), with Unix scripts written and edited using Notepad++ v7.3.3 (https://notepad-plus-plus.org/). Files were uploaded to the server and managed using the FTP software FileZilla Client v3.34.0 (https://filezilla-project.org/).

The quality of the raw paired-end reads was assessed using the quality control tool FastQC v0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Forward and reverse reads from each sample were then assembled into a reference transcriptome using the Trinity software package v2.3.2 with default settings (Grabherr et al., 2011). Timmomatic v0.35 was run as part of the Trinity pipeline to remove low quality reads and bases from the data (Bolger et al., 2014). The *de novo* transcriptome was annotated by running a blastx alignment against *Fragaria vesca* (Hawaii_1.0), *Arabidopsis thaliana* (TAIR10) and *Rosa chinensis* (Old Blush-v2). Alignments were performed using Blast+ 2.2.29 (Altschul et al., 1997) with an e-value cut off of 10^{-5}.

Transcript quantification was performed using RSEM v1.3.0 (Li and Dewey, 2011). Reads were first mapped against the *de novo* reference transcriptome using Bowtie v1.0.0 (Langmead et al., 2009) following default parameters, before 'isoform' level counts were produced for each sample. The count tables were then uploaded onto the Galaxy web platform, available at usegalaxy.org (Afgan et al., 2016), for differential expression analysis using the edgeR tool v3.20.7.2 (Robinson et al., 2009). A likelihood ratio test was used in line with traditional edgeR pipelines, and a Benjamini and Hochberg (1995) p-value adjustment method was applied at a threshold of 0.05 to reduce the false discovery rate. Default trimmed mean of M (TMM) normalisation and filtering parameters were also applied to the data. Contrasts were made between each of the stages to produce three lists of statistically up and down regulated transcripts ($\geq$90 vs. Straight, <90 vs. Straight and $\geq$90 vs. <90).

Gene Ontology analysis

Functional annotation for the *Rosa chinensis* genome have been completed using InterProScan to provide gene ontology (GO) terms and are publicly available through GDR (Jung et al., 2004). Using this resource, the differentially expressed transcripts lists produced through edgeR analysis were matched with their GO terms through the Join two files tool v1.1.1 on Galaxy. GO terms were then grouped using a simplified version of the TAIR GO-slim categories available through the TAIR website (Lamesch et al., 2011).

RESULTS

Millions of good quality paired-end reads were produced through the RNA extraction and sequencing of rose peduncle tissue for each of the three stages of necking. Replicate 1 samples were sequenced at a greater read depth, prior to replicates 2 and 3, as seen by the higher read counts for replicate 1 samples (Table 1). The ARCCA supercomputer enabled this large amount of sequencing data to be assembled into a Trinity transcriptome of 203, 565 contigs with unique identifiers. Through a blastx analysis, a total of 67.3% of these contigs were aligned to either rose, arabidopsis or strawberry (data not shown). As expected the rose genome provided the largest

ᵃEmail: Bianca.Lear.2015@live.rhul.ac.uk

amount of hits to the *de novo* assembled transcripts, with 25% and 11% more hits than arabidopsis and strawberry respectively (Figure 2).

Table 1. Total number of paired-end reads in millions per sample

Stage	Replicate 1	Replicate 2	Replicate 3
Straight	62.5	23.5	27.0
<90	73.5	27.7	25.0
≥90	51.6	22.6	26.1

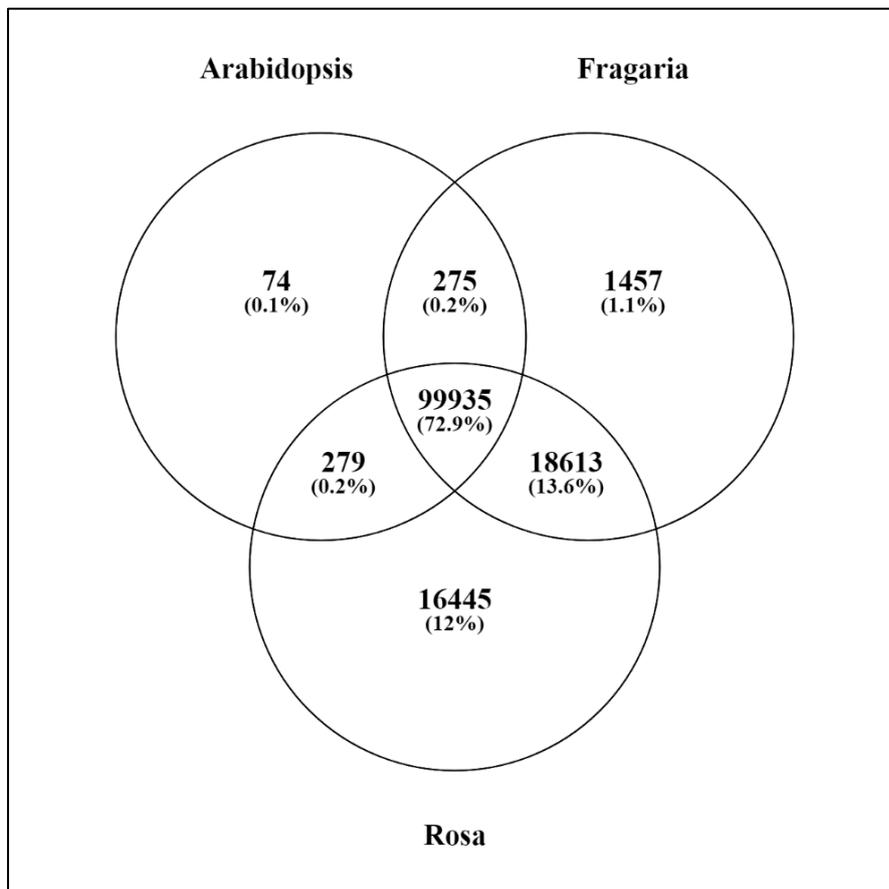*Values are shown to 3 significant figures.*

Figure 2. Number of overlapping homologous genes
Comparison of blastx alignments against Arabidopsis *(TAIR10),* Fragaria vesca *and* Rosa chinensis*. (%) percentage of total identified. E value cut off of 10^{-5}.*

Just under 2,000 differentially expressed transcripts were identified using edgeR at a robust p adjusted threshold of <0.05 (Table 2). The ≥90 vs Straight contrast provided the largest amount of statistically up and down regulated transcripts, with 1,103 in total; over double that produced by the <90 vs Straight and the >90 vs <90 contrasts which produced 473 and 384 respectively. Across all contrast groups more statistically downregulated than upregulated transcripts were discovered (Table 2).

[a]Email: Bianca.Lear.2015@live.rhul.ac.uk

Protein binding and protein metabolism show the highest amount of differentially expressed genes for the molecular function and biological process gene ontology categories. Similar trends can be seen in most of the GO sub categories for up and down regulation between the two contrasts (>90vStr and <90vStr). (Figure 3)

Table 2. Differential expression counts

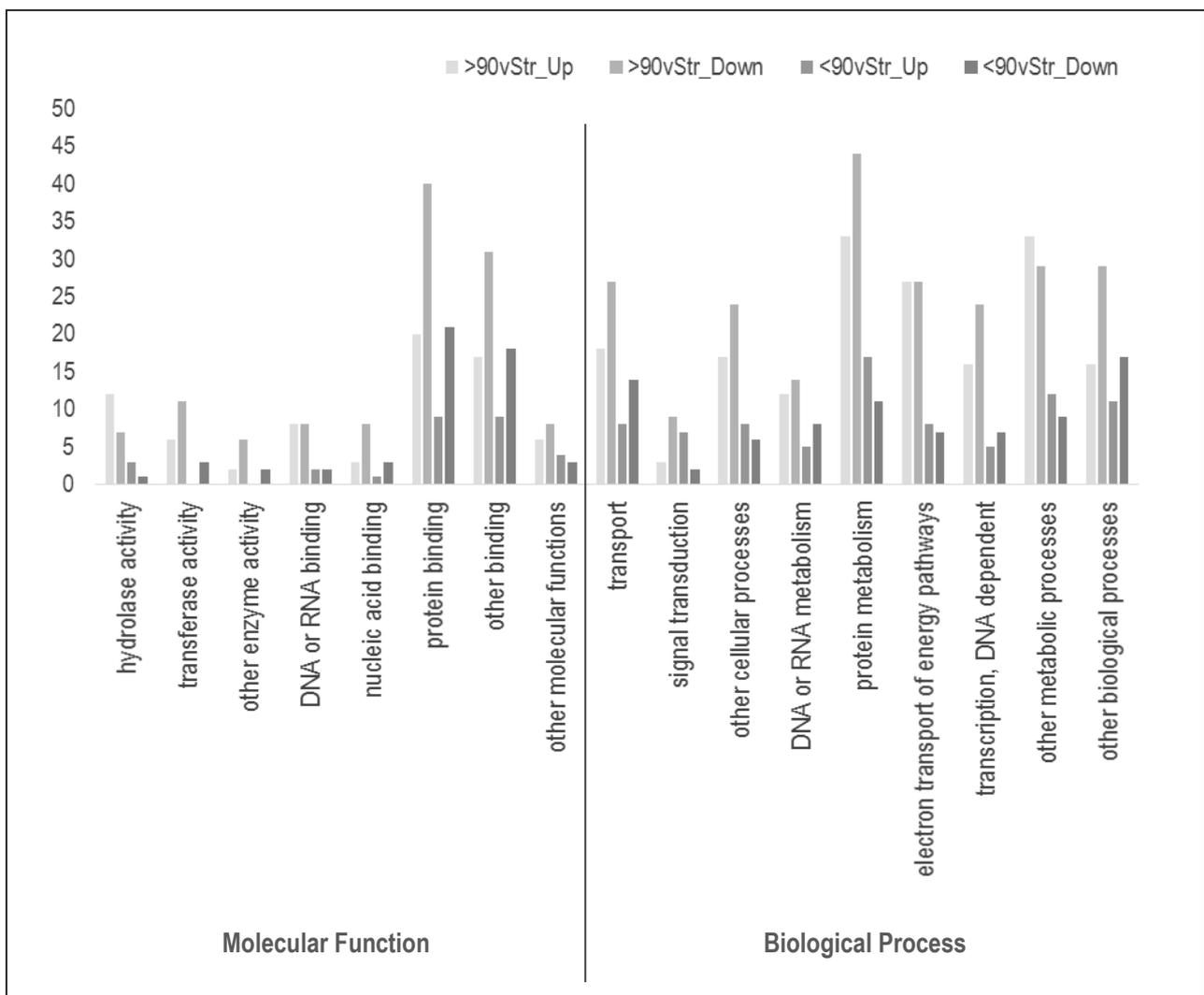|	<90 vs Straight	>90 vs Straight	>90 vs <90
Up regulated	220	481	184
Down Regulated	253	622	200

p adjusted value <0.05

Figure 3. Gene ontology categories of the differentially expressed genes
Gene ontology comparison for >90 vs Straight and <90 vs Straight up and down differentially expressed genes.

[a]Email: Bianca.Lear.2015@live.rhul.ac.uk

DISCUSSION AND CONCLUSIONS

Rosa hybrida peduncles at three stages of necking were successfully sequenced to produce millions of high quality reads (Table 1). The freely available web resource usegalaxy.org (Galaxy) and the ARCCA supercomputer were used in combination to process the high memory workloads of the *de novo* pipeline (Afgan et al., 2016). Unlike supercomputers such as the ARCCA, Galaxy does not require knowledge of unix script but instead has a preloaded set of 'tools' within a user-friendly interface and is therefore highly accessible and reproducible. However, using unix command can give you greater control over the set up and parameters, and larger jobs such as the Trinity *de novo* assembly exceeded Galaxy runtime limits and were therefore completed through the ARCCA.

The blastx alignment of the *de novo* reference transcriptome to the *Fragaria vesca* genome provided hits for 59.1% of the contigs (data not shown; e value cut off 10^{-5}). This is in line with the transcriptome study by Koning-Boucoiran et al. (2015) who found that 60.7% of their rose *de novo* transcripts mapped to strawberry genes, and therefore provides validation for the success of the de novo assembly produced in this study. Alignment of the reference transcriptome to the *Rosa chinensis* genome however provided hits for 66.5% of the contigs. This presents a great new resource for *Rosa* research and allowed for as many as 16, 724 more contigs to be identified (Figure 2).

As a similar amount of differentially expressed genes can be seen for the <90 vs straight and the $\geq$90 vs <90 contrasts (Table 2) and comparable trends can be seen between the $\geq$90 vs straight and <90 and straight contrast in figure 3; it can be determined that the <90 stage dataset represents a successful mid-point in the necking process. This is important as early stage necking roses can be hard to categorise, and due to the destructive sampling process it was unknown if those selected for this stage would go on to progress into full necking. This stage is also of particular importance as if there are molecular processes underlying necking, this early stage rather than late stage necking is most likely to include genes causative to necking rather than those present as result of the necking process. Differences between this early stage necking (<90) and later stage necking, full necking (>90), as shown in the $\geq$90vs<90 contrast, will therefore be of interest and require further analysis.

Gene ontology and pathway analysis programmes such as DAVID, AgriGO, KEGG and Mapman can be used for the next stages of analysis to easily identify significant pathways and patterns within the dataset (Kanehisa and Goto, 2000; Du et al., 2010; Huang et al., 2009; Thimm et al., 2004). However, these are not yet compatible with the new rose genome identifiers in their default setup. Carrying out a blastx alignment against a widely used model genome such as *Arabidopsis thaliana* or older genomes such as *Fragaria vesca* therefore currently remains necessary for using these tools in further study of the differential gene lists.

In conclusion, a successful *de novo* transcriptome has been produced for *Rosa hybrida* and provides a huge new resource for studying peduncle necking. However, further analysis of the differentially expressed transcripts is needed and may lead the identification of novel molecular mechanisms linked to this process.

[a]Email: Bianca.Lear.2015@live.rhul.ac.uk

LITERATURE CITED

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*. *44*, 3–10. http://doi.org/10.1093/nar/gkw343

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25, 3389–3402.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*. 22 (10), 2008–2017. http://doi.org/10.1101/gr.133744.111

Bendahmane, M., Duboid, A., Raymond, O., and Bris, M.L. (2013). Genetics and genomics of flower initiation and development in roses. *Journal of Experimental Botany*. 64(4), 847-857. http://doi.org/10.1093/jxb/ers387

Benjamini, Y., and 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), 289-300.

Bleeksma, H.C., and van Doorn, W.G. (2003). Embolism in rose stems as a result of vascular occlusion by bacteria. *Postharvest Biology and Technology*. 29(3), 335-341. http://dx.doi.org/10.1016/S0925-5214(03)00049-8

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

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGo: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*. 38(2), 64-70.

Dubois, A., Carrere, S., Raymond, O., Pouvreau, B., Cottret, L., Roccia, A., Onesto, J., Sakr, S., Atanassova, R., Baudino, S., et al. (2012). Transcriptome database resource and gene expression atlas for the rose. *BMC Genomics*. 13, 638. http://doi.org/10.1186/1471-2164-13-638

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). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*. 29(7), 644–652. http://doi.org/10.1038/nbt.1883

Hanks, G. (2015). A review of production statistics for the cut-flower and foliage sector 2015. *AHDB Horticulture*.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols*. 4(1), 44-57. http://doi:10.1038/nprot.2008.211

Jazi, M., Rajaei, S., and Seyedi, S.M. (2015). Isolation of high quality RNA from pistachio (*Pistacia vera* L.) and other woody plants high in secondary metabolites. *Physiology and Molecular Biology of Plants*. 21(4), 597–603. http://doi.org/10.1007/s12298-015-0319-x

Jung, S., Jesundurai, C., Staton, M., Du, Z., Ficklin, S., Cho, I., Abbott, A., Tomkins, J., and Main, D. (2008). GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *BMC Bioinformatics*. 5(130). https://doi.org/10.1186/1471-2105-5-130

Jung, S., Ficklin, S.P., Lee, T., Cheng, C.H., Blenda, A., Zheng, P., Yu, J., Bombarely, A., Cho, I., Ru., S., et al. (2014). The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Research*. 42, 1237-1244. http://doi.org/10.1093/nar/gkt1012

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 28, 27-30. https://doi.org/10.1093/nar/28.1.27

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*. 12 (4), 357–360. http://doi.org/10.1038/nmeth.3317

Koning-Boucoiran, C.F.S., Esselink, G.D., Vukosavljev, M., van 't Westende, W.P.C., Gitonga, V.W., Krens, F.A., Voorips, R.E., van de Weg, W.E., Schulz, D., Debener, T., Maliepaard, C., et al. (2015). Using RNA-Seq to assemble a rose

ªEmail: Bianca.Lear.2015@live.rhul.ac.uk

transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.). *Frontiers in Plant Science.* 6, 249. http://doi.org/10.3389/fpls.2015.00249

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hermamdez, M., et al. (2011). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research.* 40(1). 1202-1210. https://doi.org/10.1093/nar/gkr1090

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology.* 10. https://doi.org/10.1186/gb-2009-10-3-r25

Li, B., and Dewey, C. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* 12, 323. http://doi.org/10.1186/1471-2105-12-323

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology.* 15 (12). http://doi.org/10.1186/s13059-014-0550-8

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., et al. (2018). The *Rosa* genome provides new insights into the domestication of modern roses. *Nature Genetics.* 50, 772-777. https://doi.org/10.1038/s41588-018-0110-3

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 26 (1), 139-140. http://doi.org/10.1093/bioinformatics/btp616

Shulaev, V., Sargent, D. J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockatis, K., Liston, A., Mane, S.P., et al. (2011). The genome of woodland strawberry (Fragaria vesca). *Nature Genetics.* 43(2), 109–116. http://doi.org/10.1038/ng.740

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal.* 37(6), 914-939. http://doi:10.1111/j.1365-313X.2004.02016.x

ªEmail: Bianca.Lear.2015@live.rhul.ac.uk