Reference leaftTranscriptomes for potato cultivars: Desiree and PW363

Maja Zagoršcak[1]✉, Marko Petek[1], Mohamed Zouine[2], Kristina Gruden[1]

[1]National Institute of Biology - Department of Biotechnology and Systems Biology, Ljubljana, Slovenia
[2]Ecole Nationale Supérieure Agronomique de Toulouse, Toulouse, France

The rapid development of modern life science technologies, such as Next Generation Sequencing (NGS) techniques, allows the generation of biological data with increasing speed and precision. Most potato cultivars are highly heterozygous tetraploids with high genetic variability while being susceptible to pathogens, pests and inbreeding depression. To bypass polyploidy related sequencing problems, *Potato Genome Sequencing Consortium* (PGSC, 2011) sequenced a double monoploid derived from *S. tuberosum* group Phureja.

In order to avoid problems, such as discriminating between paralogous genes, divergence and expression bias between the reference genome and potato cultivars, and to identify traits that are not present in initially sequenced genotype, RNA-sequencing for cv. Desiree and cv. PW363 leaves was conducted on Illumina NGS platform. In house generated raw reads were complemented with data already deposited in the NCBI database and stNIB-v1 *S. tuberosum* gene groups, which included two genome assemblies and two EST datasets (Ramšak *et al.*, 2014). The preliminary transcriptomes were assembled using both hybrid and de novo assembly approaches. The hybrid approach, combining genome-guided and de novo RNA-Seq assembly, was implemented using the pipeline available from CLC Genomics Workbench 7.0.3[1]. De novo assembly was performed using Trinity (Grabherr *et al.*, 2011).

The resulting two sets of preliminary transcriptomes were then merged using the CD-HIT clustering algorithm (Limin *et al.*, 2012) and merged with existing gene models with BLAST against stNIB-v1. The presumed novel clusters were annotated using SwissProt, PGSC DM v3.4 superscaffolds and NCBI-nt databases. Initial potato pangenome containing 35609 genes was expanded with 24999 potential new transcripts, and will serve to further expand knowledge on the potato pathogen interactions.

References

Limin F, Beifang N, Zhengwei Z, Sitao W, Weizhong L (2012) CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* **28**(23), 3150-3152. http://dx.doi.org/10.1093/bioinformatics/bts565

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* **29**(7), 644-652. http://dx.doi.org/10.1038/nbt.1883

Ramšak Ž, Baebler Š, Rotter A, Korbar M, Mozetič I, *et al.* (2014) GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic Acids Research* **42**(D1), D1167-D1175. http://dx.doi.org/10.1093/nar/gkt1056

PGSC: The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195. http://dx.doi.org/10.1038/nature10158

---

1   www.clcbio.com