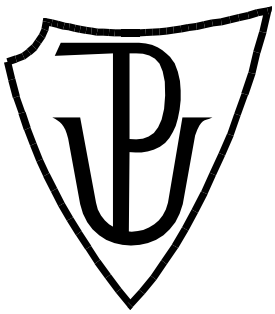
PALACKÝ UNIVERSITY OLOMOUC

Faculty of Science

Department of Biochemistry



Transcriptomic analysis of barley (*Hordeum vulgare* L.) and wheat (*Triticum aestivum* L.): tool for crop improvement.

Ph.D. Thesis

Autor: **Mgr. Filip Zavadil Kokáš**

Study program: P1416 Biochemistry

Supervisor: Dr.Véronique Bergougnoux, Ph.D.

Olomouc 2019

DECLARATION

I hereby declare that the submitted Ph.D. thesis is based on my own research carried out at the Department of Molecular Biology, Centre of the Region Hana for Biotechnological and Agriculture Research, Faculty of Science, Palacký University, Olomouc in the period September 2014 – November 2018. The thesis has been written by me with the use of literature cited in the References.

Most of the data presented in this thesis were included in the following publications:

Vojta, P., **Kokáš, F.**, Husičková, A., Grúz, J., Bergougnoux, V., Marchetti, C.F., Jiskrová, E., Ježilová, E., Mik, V., Ikeda, Y., Galuszka, P. (2016). Whole transcriptome analysis of transgenic barley with altered cytokinin homeostasis and increased tolerance to drought stress. *New Biotechnology*, 33, 676-691. https://doi.org/10.1016/j.nbt.2016.01.010

Kokáš, F., Vojta, P., Galuszka, P. (2016). Dataset for transcriptional response of barley (*Hordeum vulgare* L.) exposed to drought and subsequent re-watering. *Data in Brief*, 8, 334-341. https://doi.org/10.1016/j.dib.2016.05.051

Kokáš, F.Z., Bergougnoux, V., Čudejková, M.M. (2019). SATrans: New free available software for annotation of transcriptome and functional analysis of differentially expressed genes. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 26*(2), 117-123. https://doi.org/10.1089/cmb.2018.0149

In Olomouc 15. 04. 2019 ……..……...

Mgr. Filip Zavadil Kokáš

ACKNOWLEDGEMENTS

I wish to express my thanks

to my advisors Dr. Véronique Bergougnoux Ph.D., Doc. Petr Galuszka and Ph.D., Maria Majeská Čudejková Ph.D., and also Doc. Lenka Luhová Ph.D. and Mgr. Petra Hloušková for their valuable time, fruitful discussions, appreciated advices and for their help with my study problems.

to CESNET for the access to the computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042).

BIBLIOGRAPHICAL IDENTIFICATION

Author's name: Filip Zavadil Kokáš

Title: Mgr.

Type of thesis: Ph.D.

Department: Department of Molecular Biology, CRH UPOL Olomouc

Supervisor: Dr. Véronique Bergougnoux, Ph.D.

The year of presentation: 2019

Abstract

The thesis is focused on differential transcriptomics analysis of barley (*Hordeum vulgare*) and wheat (*Triticum aestivum*), two economically important cereals. It is also described the development of a new bioinformatics software allowing the annotation of novel transcripts and the functional analysis of differentially expressed genes. The theoretical part deals with the description of the main technologies of RNA-sequencing, design of the RNA-sequencing experiment and down-stream bioinformatics analysis.

The experimental part is divided into three chapters. The first chapter aimed to understand the differential transcriptomics of transgenic barley lines overexpressing the *Arabidopsis cytokinin dehydrogenase 1* gene under the control of the mild root specific promotor of maize β-glycosidase. The results showed transgenic lines are more tolerant to drought than the wild-type plants, mainly due to the alteration of their root architecture and a stronger lignification of root tissue.

The second chapter is focused on the development of SATrans, new freely available software for the annotation of transcriptome and the functional analysis of differentially expressed genes. The software was developed with Perl and MySQL as programing languages and has been tested on a test data set. It provided a fast and robust functional annotation of novel sequences and performed advanced gene ontology analysis of the differentially expressed genes.

The last chapter covered different approaches to analyze RNAseq data generated for four new inbred lines of hexaploid wheat with different root architecture. Three different assembly approaches (*ab initio*, *de novo* and combined) were used and evaluated for their

ability to provide the best reference wheat transcriptome for downstream analyses. The combined approach, i.e. coupling *ab initio* and *de novo* transcriptome assemblies, was evaluated as the best tool to generate a root-specific reference transcriptome. Downstream bioinformatics analyses were performed in order to highlight the genes whose differential expression might be related to various root architecture observed between for the four-wheat new inbred lines. The data showed that few biological processes were affected, including the transmembrane transport of the phytohormone auxin and the hydrolysis of its conjugates. Nevertheless, our hypothesis will have to be supported by qPCR analysis and strong physiological experiments.

Keywords: transcriptomics, bioinformatics, software development, functional annotation, wheat, barley
Number of pages: 168
Number of appendices: 21
Language: English

BIBLIOGRAFICKÁ IDENTIFIKACE

Jméno autora: Filip Zavadil Kokáš

Titul: Mgr.

Typ práce: Disertační

Pracoviště: Oddělení molekulární biologie, CRH UPOL Olomouc

Vedoucí práce: Dr. Véronique Bergougnoux, Ph.D.

Rok obhajoby práce: 2019

Abstrakt

Tato práce je zaměřena na diferenciální analýzu transkriptomu u ječmene (*Hordeum vulgare*) a pšenice (*Triticum aestivum*) jako dvou významných zemědělských plodin. Práce rovněž popisuje vývoj nového bioinformatického softwaru vyvíjeného za účelem zlepšení funkční anotace nových sekvencí a následné analýze diferenciálně exprimovaných genů poskytnutých v rámci transkriptomické studie. Teoretická část práce je soustředěna na popis současných technologii používaných pro RNA-sekvenování, design experimentu a následnou bioinformatickou analýzu s využitím rozličných bioinformatických nástrojů.

Experimentální část práce je rozdělena na tři kapitoly. První kapitola se zabývá diferenciální analýzou transkriptomu u transgenních linii ječmene s nadprůměrně exprimovaným genem *cytokinin dehydrogenázou 1* z *Arabidopsis thaliana*, pod kořenově specifickým promotorem β-glykosidázy z kukuřice. Výsledky ukázaly transgenní linie jako více tolerantní vůči suchu, ve srovnání s rostlinami divokého typu, a to v důsledku odlišné architektury kořenového systému a silnější lignifikace kořenové tkáně.

Druhá kapitola se zabývá vývojem a popisem volně dostupného programu SATrans, vytvořeném za účelem lepší a robustnější funkční anotace sekvencí, a rovněž pro analýzu diferenciálně exprimovaných genů. Program byl implementován programovacími jazyky Perl a MySQL, následně byl testován na datovém setu a srovnán s ostatním volně dostupnými programy. Program poskytuje rychlou a robustní funkční anotaci nových sekvencí a poskytuje pokročilou analýzu genové ontologie pro diferenciálně exprimované geny.

Poslední kapitola praktické části se zabývá odlišnými přístupy pro výstavbu kořenově specifické reference za účelem mapování krátkých sekvencí získaných v průběhu sekvenačního experimentu. Na vstupní data (čtyři genotypy pšenice s odlišnou architekturou kořenového systému) byly aplikovány tři různé přístupy pro vytvoření reference (*de novo* přístup, *ab initio* přístup, kombinovaný přístup). Kombinovaný přístup byl na základě sledovaných charakteristik vyhodnocen jako nejlepší a výsledná reference byla použita pro následnou diferenciální analýzu transkriptomu mezi genotypy s odlišnou architekturou kořenového systému. Výsledky ukázaly několik ovlivněných biologických procesů, mezi které náleží zejména transmembránový transport auxinů a hydrolýza jejich konjugátů. Potvrzení výsledků získaných touto analýzou by ovšem mělo být provedeno s pomocí qPCR a kontrolovanými biologickými experimenty.

Klíčová slova: transkriptomika, bioinformatika, vývoj softwaru, funkční anotace, ječmen, pšenice
Počet stran: 168
Počet příloh: 21
Jazyk: anglický

CONTENT

INTRODUCTION AND AIMS OF DISSERTATION

Sustainability in agriculture is challenged by the increase of human population, environmental changes, limited land availability, water shortage and growing demand for biofuel production. For long, traditional plant breeding was used to improve the agronomical value of crop species, consisting in crossing two well-performing parents and selecting new superior genotypes that combine the parental qualities. Despite its importance, breeding has limitations. First, it requires that two plants can sexually mate with each other, leading to the narrowing of the gene pool from which cultivars are drawn. Second, up to 10 to 12 years are necessary to keep by backcrossing only the desired traits. In opposite, the modern genetic engineering is the directed addition of a foreign gene or genes to the genome of an organism. In this regard, exploring and exploiting the plant genomes for determining the function of important genes involved in agronomically relevant traits (responses to biotic or abiotic stress, yield or plant development) is one of the durable strategies to bring sustainability to crops. During the last decade, the decreasing cost of DNA sequencing and the development of sequencing technology provides feasible applications such as whole-genome re-sequencing for variation analysis, RNA sequencing (RNA-seq) for transcriptome and non-coding RNAome analysis, quantitative detection of epigenomic dynamics, and Chip-seq analysis for DNA/protein interactions. In addition to these approaches, which focus on transcriptional regulatory networks, other approaches have been developed, including interactome analysis for networks formed by protein/protein interactions, hormone analysis for phytohormone-mediated cellular signaling, and metabolome analysis for metabolic systems. Bioinformatics that might stand for “Biological Informatics” is at the junction between biological sciences and computer science. Very often the term of “computational biology” is now reported. Bioinformatics has been crucial in every aspect of omics-based research to manage various types of genome-scale data sets effectively and extract valuable knowledge.

The overall objectives of the present PhD thesis were to apply bioinformatics to two cereals – barley (*Hordeum vulgare* L.) and bread wheat (*Triticum aestivum* L.) – in order to get insights in molecular mechanisms regulating i) cytokinin-mediated tolerance to drought and ii) root development. For this purpose, the work was divided into four main sections:

i) Study of the transcriptome of transgenic barley (*Hordeum vulgare* L. cv. Golden Promise) with altered cytokinin content exposed to drought stress

ii) Development of a robust software to provide reliable functional annotation of novel sequences and gene ontology analysis

iii) Compare different assembly strategies to obtain the best reference transcriptome of the hexaploid bread wheat

iv) Study of the transcriptome of four bread wheat (*Triticum aestivum* L.) inbred lines with various root systems

LITERATURE OVERVIEW

The transcriptome is characterized by a complete set of genes (transcripts) expressed in a cell and by their abundance which are both specific for a given physiological conditions or a developmental stage of the organism. Understanding the transcriptome is necessary for interpreting the functional elements of the genome and for uncovering the molecular constituent of cells and tissues. Study of transcriptome of an organism using specific technologies is called transcriptomics (Wang *et al.*, 2009a; Yang & Kim, 2015). *In silico* analysis must be provided using mathematical and statistical techniques for correct interpretation of results, which were obtained by transcriptomics. Bioinformatics provides powerful tools for this purpose as an interdisciplinary field that develops methods and software's for understanding biological data (Lesk, 2013).

2.1 Methods of nucleic acid sequencing

The history of nucleic acid sequencing began in the 1970s with the development of the first-generation sequencing methods. In 1977, the bacteriophage ΦX 174 was the first genome to be sequenced (Sanger *et al.*, 1977) and since then many researchers from around the world have invested a great deal of financial resources to improve technologies that facilitate DNA sequencing (Heather & Chain, 2016). Second-generation sequencing (SGS) technologies, released in the first decade of the twenty-first century, reduced run times and costs and increased throughput (about hundreds of Gbp per run; Goodwin *et al.*, 2016; Schadt *et al.*, 2010). SGS platforms produces millions of short DNA sequence reads which are generally in the range of 25 to 700 bp in length (Unamba *et al.*, 2015).

During the last decade, third-generation sequencing (TGS) technologies (also known as long-read sequencing) was developed. Currently under active development, these methods work by reading the nucleotide sequences at the single molecule level, negating the requirement for DNA amplification, characteristic of the previous methods. The third-generation methods can therefore sequence long molecules of nucleic acids in contrast to earlier methods that require breaking long strands of DNA into small segments before inferring nucleotide sequences (Morey *et al.*, 2013).

Fourth-generation sequencing technology combines traditional imaging analysis techniques with the SGS technologies to offer new opportunities for sequencing nucleic acids such as removing the optical detection (Feng *et al.*, 2015; Ke *et al.*, 2016).

2.1.1 Second generation of sequencing technology

Over the past few years, second-generation sequencers were reported. Representative members of the second-generation DNA sequencers, such as 454 Genome Sequencers (Margulies *et al.*, 2005), Illumina (Liu *et al.*, 2012), HelioScope, and Sequencing by Oligonucleotide Ligation and Detection (SOLiD; Shendure *et al.*, 2005), have gradually replace the first-generation ones, reducing the cost of sequencing per one megabasis.

There are a few key factors in the definition and evaluation of sequencing technology platform, such as read length, throughput, read accuracy, read depth (number of times each base is sequenced in independent events) and cost per base (Morey *et al.*, 2013). The main weakness of those sequencing platforms is that the read length is not as long as with the first-generation sequencing techniques, mainly due to the progressive decline in efficiency of the sequencing chemistry during the run. The result of this process is an asynchronous read elongation for any given amplicon clone and a correspondingly ambiguous fluorescent signal (Shendure & Ji, 2008). Another weakness of SGS is the use of polymerase chain reaction (PCR) which can bring another bias (Acinas *et al.*, 2005). Nevertheless, one of the key advantages is the capacity of SGS to perform highly multiplexed reactions (Morey *et al.*, 2013).

Irrespective of the used technology, there are several conjoint steps in DNA sequencing. First, the DNA sample is fragmented to a target size, depending on the read length, the sequencing platform and the chemistry used. The resulting fragment size generally has a Poisson distribution and can be optionally refined via size selection. The fragments of nucleic acids are subsequently amplified by various enrichment strategies. PCR has been the most widely used technique for sequencing sample enrichment. The main weakness of target enrichment by PCR is that coverage uniformity can be lower in comparison with other techniques such as hybridisation capture or DNA circularisation (Morey *et al.*, 2013).

Enrichment by hybridization is based on the information from microarray research and can be applied to the enrichment of samples for SGS. DNA is hybridised with probes and nonspecific fragments are washed out. Two methods are currently available, namely in-solution hybridization and array hybridization. The crucial advantage of these methods is the ability to capture a larger number of fragments and to provide more homogeneous coverage (Hodges *et al.*, 2007; Okou *et al.*, 2007). An alternative method, known as DNA circularisation, is based on the use of selection probes or molecular inversion probes (Porreca *et al.*, 2007; Dahl *et al.*, 2005).

After library preparation, sequencing is performed in combination with clonal amplification of the signal generated in order to enable detection of the sample during the sequencing process. Clonal amplification aims to produce several copies of each DNA library fragment which will be separated in clusters to produce unambiguous, monoclonal signal (Mardis, 2008; Shendure & Ji, 2008). The amplification technique depends on the sequencing platform which is used for sequencing process. There are a few platforms which are commercially available and are viewed as relevant methods. All SGS platforms use the same approach for signal detection, which is performed by Charge-Coupled-Device (CCD) camera in combination with microscope, computer and storage system (Morey *et al.*, 2013).

2.1.1.1 454 Pyrosequencing

Pyrosequencing as the first major successful commercial SGS method was licensed in 2005 by 454 Life Sciences in collaboration with Roche Diagnostics (Ansorge, 2009). This method is also reported as one of the sequencing-by-synthesis methods. DNA fragments are enriched by specific adapters, immobilized onto beads and amplified via emulsion PCR (emPCR) which is also used in SOLiD technology (Morey *et al.*, 2013).

Amplification by emPCR is based on the principle where DNA fragments with adapters (templates) are isolated in independent aqueous microreactors which are surrounded in oil phase (Fig. 1). emPCR is usually carried out with reaction primers immobilized on the surface of beads which serve to localise the clonally amplified fragments of each reaction on the surface of one bead (Dressman *et al.*, 2003; Nakano *et al.*, 2003).

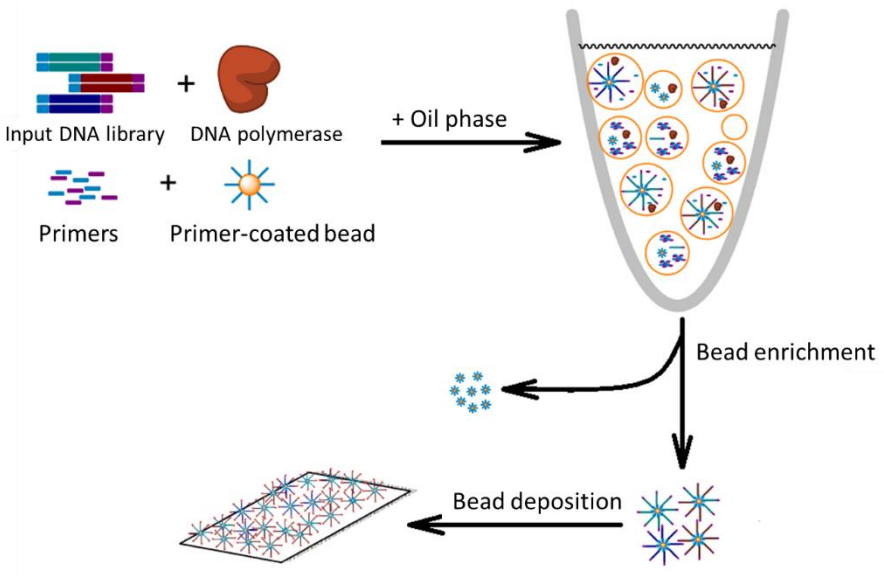


Figure 1: Scheme representing the different steps of the emPCR (adapted from Morey *et al.*, 2013).

After emPCR amplification, the emulsion with beads is broken down by organic solvents and the beads are isolated into the aqueous phase by extraction. Each bead bears multiple copies of a single, unique fragment. The efficiency of emPCR depends on the amount of DNA added to the emPCR reaction. A large amount of DNA can lead to the generation of polyclonal beads, i.e. beads covered by amplicons generated by different DNA fragments. Beads without amplified DNA fragment can be excluded by physical separation of amplified and non-amplified beads (Fig. 1). Beads with amplified DNA fragments are deposited onto a PicoTiterPlate containing millions of micro-wells whose diameter allow the inclusion of only one bead (Shao *et al.*, 2011; Margulies *et al.*, 2005).

After beads deposition, other components are added for sequencing namely DNA-polymerase, ATP sulfurylase, luciferase and apyrase (Fig. 2). Then adenosine-5´-phosphosulphate (APS) as a future substrate for ATP sulfurylase is added to the reaction. After added one from four deoxy-nucleotide triphosphate (dNTP), the synthesis of new DNA strand is catalysed by DNA polymerase and generates one molecule of pyrophosphate (PPi) per base added. ATP sulfurylase convert PPi into ATP by conversion of APS molecule (Fig. 2). The ATP is then used to produce visible light by the luciferase in the presence of luciferin. The light is detected by CCD camera and the light intensity is proportional to the amount of PPi (Nyrén, 1987).

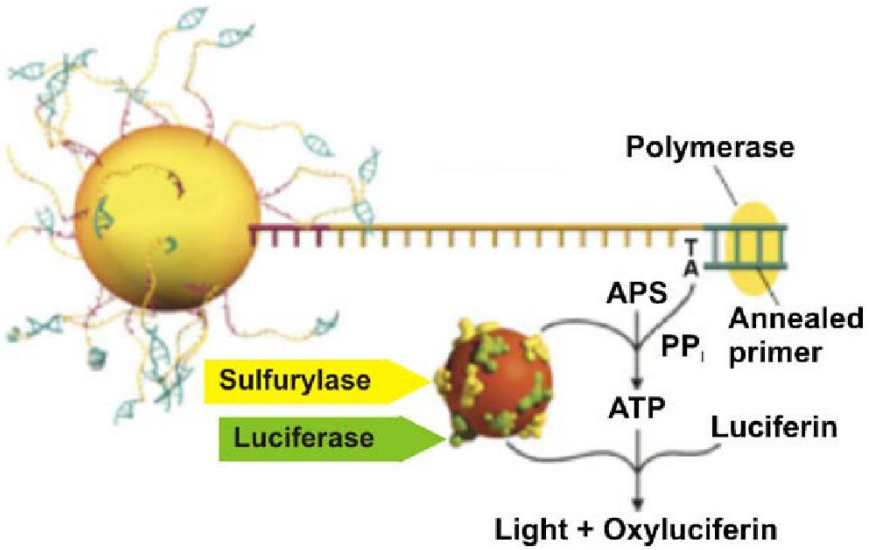


Figure 2: Principle of the pyrosequencing approach. The integration of a nucleotide into the DNA strand generates a molecule of PPi that is converted by the sulfurylase into ATP in the presence of APS. ATP is used by luciferase to convert luciferin into oxyluciferin. This reaction generates light that is measured by CCD camera. APS: adenosine-5´-phosphosulphate; ATP: adenosine-triphosphate; dNTP: deoxy-nucleotide triphosphate; PPi: pyrophosphate; (from Ansorge, 2009).

Consequently, the intensity of the light corresponds with the number of bases which were added. However, noise which is produced allows detecting only four or five identical nucleotides. The apyrase is then used to degrade the non-incorporated nucleotides (Ronaghi *et al.*, 1998).

The data output generated by 454 pyrosequencing is relatively low in comparison to other approaches. Nevertheless, this is compensated by the length of reads which can be obtained (around 400-500 bases). Relatively long reads are of interest to align complex, such as GC-rich regions and repeated regions, with the use of low computational resources, especially in genome *de novo* sequencing (Margulies *et al.*, 2005; Mardis, 2008).

2.1.1.2 Illumina sequencing-by-synthesis

Similarly, to the 454 pyrosequencing, the Illumina method is also a sequencing-by-synthesis sequencing method. Nevertheless, the amplification is performed by bridge-PCR, and not by emPCR as described above. This technology was initially developed by S. Balasubramanian and D. Klenerman from the University of Cambridge, who created

the Solexa Company in 1998. This later was purchased by Illumina in 2006-2007 (Morey *et al.*, 2013).

Illumina platforms (MiSeq, NextSeq 500 and HiSeq series) based on this technology currently dominate the market and are optimized for a variety of throughputs. MiSeq and HiSeq are currently the most established Illumina platforms (Reuter *et al.*, 2015).

The bridge-PCR used by Illumina for amplification is also listed as a two-dimensional PCR which allows the clonal amplification of a large number of DNA fragments in parallel on the surface of a flow-cell (Fig. 3; Adessi *et al.*, 2000).

The oligonucleotides immobilised on the surface of the flow-cell are linked by their 5′-end and complementary to the adaptors which were added to the DNA fragments during the library construction. Initially, double-stranded-DNA (dsDNA) fragments are denatured into individual single-stranded-DNA (ssDNA) molecules and hybridized to the oligonucleotides on the flow-cell surface. After bridge-PCR, the resulting dsDNA is covalently linked to the surface.

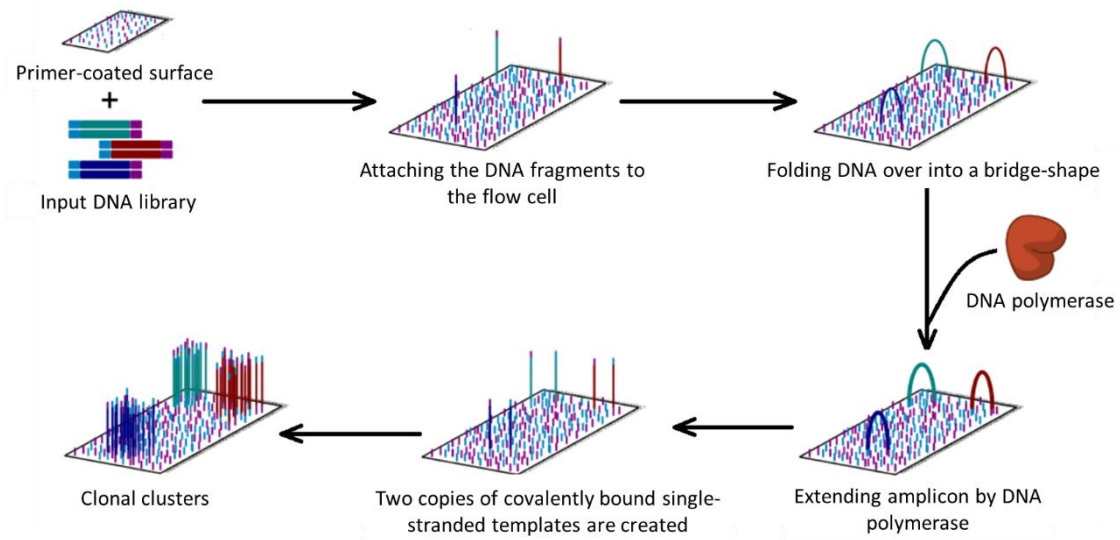


Figure 3: Scheme representing the different steps of the bridge-PCR (adapted from Morey *et al.*, 2013). The DNA fragments contain adapters whose sequences are complementary to the sequence of the primers coated on the surface of the flow-cell. DNA fragments are denatured and hybridize to the flow-cell surface. Oligonucleotides of the flow-cell are linked to the surface by their 5′-end, leaving the 3′-end free for the polymerase. The resulting double-stranded-DNA is covalently attached to the flow-cell. This double-stranded DNA is then denatured and the single strand bends to hybridize to adjacent primers, thus forming a bridge. Polymerases form a double-stranded bridge. After denaturation, two copies of covalently bound single-stranded templates are obtained. This cyclical process is repeated several times, producing clusters of clonal copies of each initial fragment. No primers are required in the reaction solution and clusters are spatially separated.

This dsDNA is then denatured, and a bridge is formed by ssDNA bending towards and hybridization with nearby primers. The newly closed amplicon is extended by DNA polymerase and after denaturation two copies of covalently bound ssDNA are obtained.

The separation between clusters which were synthetized by bridge-PCR strictly depends on the initial quantity of DNA provided for the reaction. High initial concentration of DNA can lead to close localised clusters, leading to interference between them, whereas too small amounts of DNA can lead to a lower yield of bridge-PCR (Morey *et al.*, 2013). After denaturation, the sequencing step can be performed.

Illumina´s technology is based on the polymerase-catalysed addition of reversible terminator fluorescently labelled bases, which are added in parallel during reaction (Fig. 4). Once a fluorescently labelled dNTP is added, addition of subsequent bases is prevented because the fluorophore occupies the 3´-hydroxyl position. It means, that only one base can be added in one cycle (Turcatti *et al.*, 2008).

After base incorporation, residual dNTPs and DNA polymerase are washed out. Fluorophores are excited with appropriate lasers and imaging is ensured by CCD camera. The 3´-position blocking is then chemically removed, and the sequencing cycle is repeated (Heather & Chain, 2016).

Illumina technology provides maximum output approximately 600 Gb per flow-cell with 6 billion reads length 2x100 bases on a paired-end library. Researchers can use the barcodes to distinguish different samples that have been pooled into a single library (Morey *et al.*, 2013). Error rates for Illumina machines are below 1%, and the most common type of error is substitution (Dohm *et al.*, 2008).

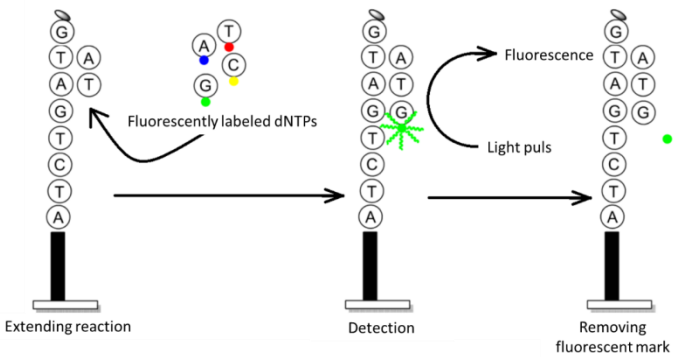


Figure 4: Scheme showing the different steps of the Illumina sequencing methods (adapted from Morey *et al.*, 2013).

2.1.1.3 Sequencing-by-ligation: SOLiD

The method of sequencing-by-ligation was commercialized by Applied Biosystems in 2008. In SOLiD, an emPCR is performed to provide clonal amplification. DNA library is used as an initial material and approach for amplification is similar like in pyrosequencing method. Beads, which contains amplified fragments, are then immobilized to a glass surface in a random pattern and sequencing-by-ligation can be performed (Fig. 5).

Process of sequencing is started by attaching universal primer to one of the library adaptors. Fluorescently labelled nucleotide octamers with two strictly defined bases are used for ligase reaction. The remaining bases are degenerated. After ligation, the three final bases are cleaved so the length of the octamer is reduced to five nucleotides.

After ligation-cleaving step, an imaging step is performed to detect fluorescence and ligation cycle can be repeated. In ligation cycle, the first two nucleotides of each group of five are examined (Morey *et al.*, 2013).

After many ligation cycles, the synthetized nucleotide chain is cleaved together with universal primer. A new universal primer with different length replaces the old one and the same ligation cycle as listed above can be performed. This process is repeated five times with five different universal primers. Every base on each clonally amplified DNA fragment is examined twice in independent ligation. Approach is called as two-base encoding and increase sequencing accuracy.

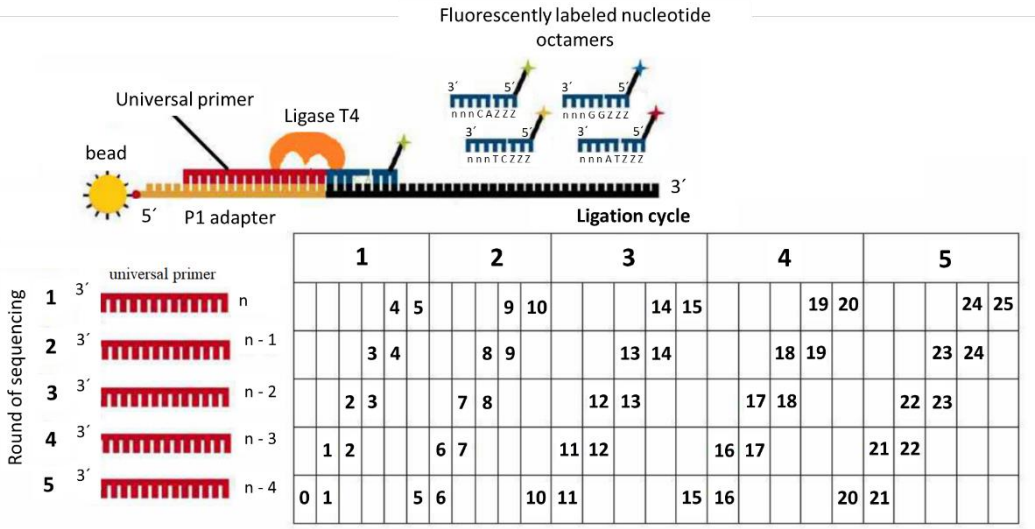


Figure 5: Display the process of SOLiD sequencing process (adapted from Morey *et al.*, 2013).

The commercially available SOLiD platforms run in 7 up to 14 days, depending on the chemistry. The maximum output is 300 Gb per run with approximately 5 billion reads provided with the read length 35-75 bp for paired-end library (Morey *et al.*, 2013).

2.1.1.4 Other sequencing platforms

DNA nanoball sequencing is another notable technology listed as member of SGS technologies that has been commercialized since 2009 by Complete Genomics. This technology is based on the formation of large numbers of compact DNA nanoballs (Heather & Chain, 2016; Morey *et al.*, 2013).

The source DNA is fragmented (size range 400-500 bp) and universal double stranded adaptors (Ad1L and Ad1R; Fig. 6) are attached to both ends. The fragments are then amplified and circularised. The resulting amplicon is exposed to restriction enzyme which cuts 13 bp internally from Ad1R, generating linear product. A second set of universal adaptors (Ad2L and Ad2R) is attached, and the resulting product is amplified and circularized again. The product is processed by restriction enzyme, this time cutting 13 bp internally from Ad2L. A third cycle of adaptor (Ad3L and Ad3R) ligation, PCR amplification and circularisation is repeated. Then, DNA is cutted at 26 bp internally to Ad3L, and 26 bp from Ad2R by the EcoP15 restriction enzyme. Fragments of approximately 350 bp are eliminated and replaced by the last adaptor Ad4. PCR is performed and the product circularized. Then, the circular replication is performed by Φ29 DNA polymerase (Drmanac *et al.*, 2010; Morey *et al.*, 2013).

Adaptors in amplified sequence represent palindromic sequences which are combined with complementary fragments and forms nanoball of DNA sequences. The sequencing of the complete circular template is performed by oligonucleotides which are complementary to the added adaptors.

Ligase T4 and a pool of 10-mer DNA sequences, with some defined nucleotides which correlated with the fluorophore attached to that probe, are added to the reaction. A probe will only bind to the complementary DNA. Then non-bound probes are washed away, and the fluorescence is detected by an imaging system. After detection, probe is removed, and the process is repeated with another pool of 10-mers. A maximum of 70 bases can be sequenced by this technology (Porreca, 2010).

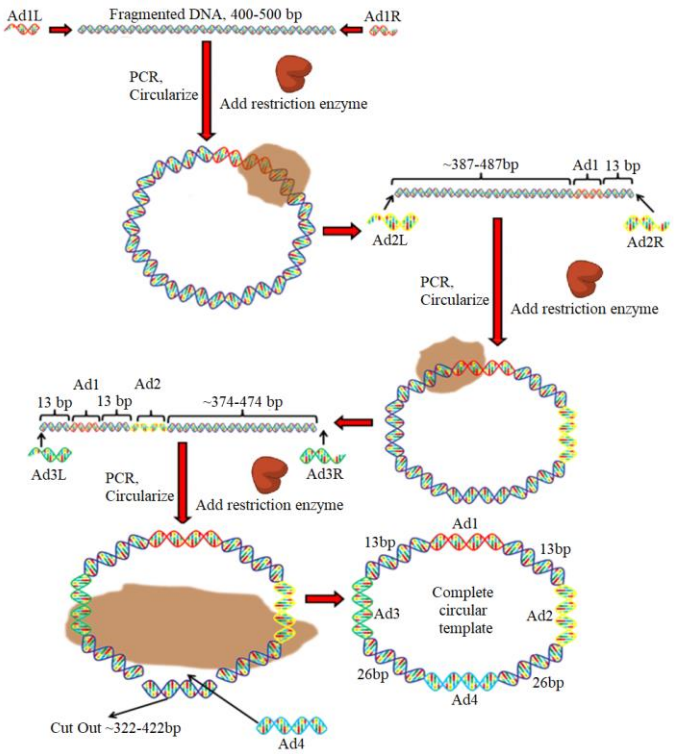


Figure 6: Display process of Nanoball sequencing (adapted from Wikimedia Foundation, 2010).

The main advantage of this technology in contrast to other sequencing approaches is the use of high-density arrays which is provided by nanoball dispersion. Disadvantage of this technology is the production of reads with short length (Drmanac *et al.*, 2010; Porreca, 2010).

Semiconductor sequencing is another technology of second generation and can be listed as a variant of pyrosequencing. The technology was commercialized by Ion Torrent Systems Inc. in 2010 in the form of the benchtop Ion PGM sequencer. The clonal amplification of the library is performed by emPCR similarly to pyrosequencing. The amplified fragments on beads are deposited onto a surface made of micro-wells containing a sensor sensitive to pH changes. In contrast to pyrosequencing, this technology is based on the detection of hydrogen ions generated after dNTP addition, which is performed by a semiconductor chip in contrast to pyrosequencing technology where PPi is measured (Quail *et al.*, 2012; Morey *et al.*, 2013).

Initially, unmodified dNTP is added into the microwells. The reaction by polymerase causes the release of a hydrogen ion, which leads to pH change and detection by the sensor. Ion release and voltage signal depend on the number of added bases. The crucial advantage of this technology is that Ion Torrent avoids optical scanning to distinguish nucleotides which allows the attachment of more than one base in the case of repetitive regions. Nevertheless, the disadvantage lies in the fact that the generated signal does not increase completely linearly, depending on the number of bases added, which can lead to bias (Rothberg *et al.*, 2011). Insertions and deletions form the most common errors types (Liu *et al.*, 2012). Homopolymer in fragment of DNA which is longer than 6 bp lead to increased error rate, because subsequent voltage changes for homopolymers are not perfectly scaled (Rothberg *et al.*, 2011).

The last member of the mainstream of sequencing technologies is the Polony sequencing that has been at Harvard Medical School. Here again, the amplification is ensured by emPCR. The process of sequencing starts with the addition of universal primers complementary to one of the library adaptors. The ligase attaches a base-specific fluorescently-labelled nonamer nucleotide in the position which is complementary (Morey *et al.*, 2013). The nonamer nucleotides are degenerate, with the exception of one, which represents the investigated position in nonamer. The nonamer is fluorescently labelled and fluorescence is emitted and recorded by light imaging system after each ligation cycle. Then, the universal primer and nonamer are separated from the DNA and new sequencing cycle is initiated similarly like in SOLiD technology. The position of the defined base is changed each time, so the investigated nucleotide is also different. The main advantage of this technology is that it provides a very flexible technique with variable application. Nevertheless, it provides only small amount of usable data which causes limitation in performance of the technique (Shendure *et al.*, 2005; Mitra *et al.*, 2003).

2.1.2 Third generation of sequencing technology

During the last decade several alternative approaches to improve second-generation technologies were developed as well as novel sequencing approaches, including single-molecule real-time sequencing (SMRT), fluorescence resonance energy transfer (FRET)

and transmission electron microscope (TEM). Those technologies are listed as third generation of sequencing technology.

The crucial advantage of TGS technologies is their ability to sequence single molecules of DNA in contrast to the SGS technologies. Moreover, the sequencing reaction is not paused for washing and imaging after incorporation of each base, so time and consumption of reagents are reduced (Morey *et al.*, 2013).

SMRT method was developed by Pacific Biosciences and is currently the most widely used TGS technology. This technology is based on the detection of natural DNA synthesis by a single DNA polymerase (Eid *et al.*, 2009). SMRT use nanophotonic visualisation chamber that contains thousands of zero-mode waveguides (ZMWs; Fig. 7). ZMWs are tiny holes of approximately 45 nm diameter in a 100 nm thick aluminium film, which provide fluorescence detection; the volume is around 20 zeptoliters. ZMWs use light properties that pass-through holes with a diameter less than their wavelength. This phenomenon causes exclusive illumination of the very bottom of the wells and allows visualizing the fluorescence emitted by fluorophore (Levene *et al.*, 2003; Foquet *et al.*, 2008).

SMRT chip contains approximately 75 000 ZMWs, each with a molecule of DNA polymerase immobilized at its bottom. The polymerase used for sequencing is a modified version of Φ29 DNA polymerase, exhibiting reduced 3-5 exonuclease activity, while preserving the original properties Φ29 such as processivity of several hundred kilobases and error rate of 10^{-5} (Buermans & den Dunnen, 2014; Morey *et al.*, 2013).

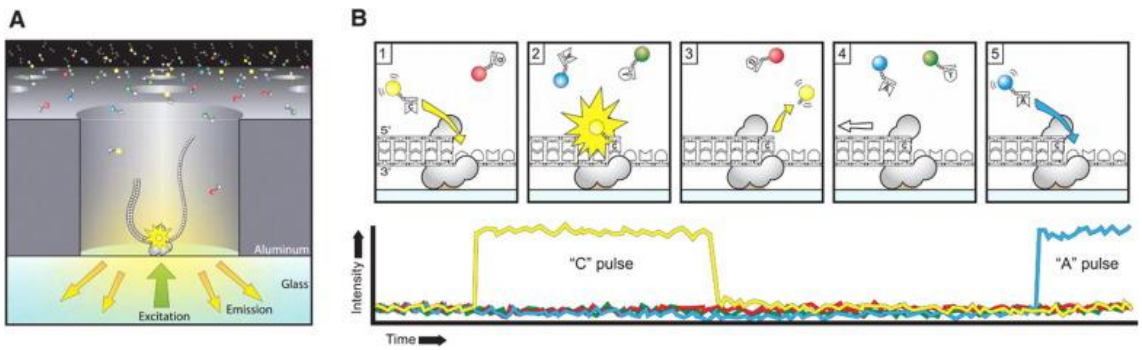


Figure 7: Schematic of SMRT. A. Nanophotonic visualization chamber. **B.** Sequencing process where fluorescently labelled nucleotides are added (1), detection by excitation is provided (2), removing dye-linker-pyrophosphate product (3), translocation polymerase to the next position (4), association the next nucleotide with the template in the active site of the polymerase and initiating the next fluorescence pulse (5) (from Rhoads & Au, 2015).

Template preparation involves ligation of single-stranded adapters at the ends of the fragmented DNA, generating a capped template (Reuter *et al.*, 2015). Template preparation does not require any amplification steps, and the prepared library molecule connected with single-stranded adapter forms the sequencing template (Buermans & den Dunnen, 2014).

SMRT technology uses phospholinked nucleotides, which is a crucial difference in comparison to Illumina technology where different fluorescent label for each nucleotide is used. Fluorescently labelled nucleotides are added simultaneously to the reaction and if fluorescently labelled phosphate is cleaved by DNA polymerase, a base-specific fluorescence is produced (Fig. 7). After extension, the fluorescently linked phosphate diffuses outside the illumination zone, and the next extension can be provided. Base incorporation is detected as a specific-color flash at a rate of multiple bases per second (Margulies *et al.*, 2005; Buermans & den Dunnen, 2014).

The polymerization occurs continuously, and the DNA sequence can be sequenced in real-time. SMRT technology is also capable of producing incredibly long reads, up to 10 kbp, which provide easy way for *de novo* assembly. Read accuracy is reported to be around 83% but can be improved to more than 99% with 15x coverage and corresponding bioinformatic software. Indels are dominated in these sequencing errors and are distributed randomly (Reuter *et al.*, 2015; Heather & Chain, 2016; Morey *et al.*, 2013).

Other technologies, which can be listed as TGS does not have such share on the market in comparison to the SMRT technology and are still under active development. First such technology is the Real-time DNA sequencing using FRET. This technology uses DNA polymerase with an attached fluorophore which is brought into close proximity together with phospholinked nucleotide. Mutual interaction causes emission of a FRET signal and fluorophore linked on the nucleotide is released. Fluorophore can be detected by photodetection.

The last technology which will be listed here is a direct imaging of DNA using TEM. This technology is developing by ZS Genetics and Halcyon Molecular. The technology is based on the direct imaging and chemical detection of atoms to identify nucleotides by annular dark-field imaging. Companies planned to initiate an early-access centre in 2017 for product and application development (Morey *et al.*, 2013).

2.1.3 Fourth generation of sequencing technology

After the development of third generation of DNA sequencing technology, has become nanopore-based DNA sequencing technology one of the most advanced sequencing technology of third generation (Morey *et al.*, 2013). However, some publications listed this technology as a member of fourth generation (Feng *et al.*, 2015).

This technology uses single-molecule strategy that is based on the tunnelling of a DNA molecule or its component bases through a nanopore that separates two compartments (Morey *et al.*, 2013; Wang *et al.*, 2015). Library preparation includes fragmentation of DNA and ligation of adapters. Similarly, like third generation sequencing technology the library preparation does not include PCR amplification (Reuter *et al.*, 2015).

The nanopores are embedded in a biological membrane or formed in solid-state film (Feng *et al.*, 2015). Sequencing flow-cell comprises hundreds of independent micro-wells, each containing a bilayer perforated by nanopores. When DNA is loaded onto one side and a voltage applied across the bilayer, DNA chain can be mobilized through the nanopore. Characteristic changes are measured in current that are induced as the bases are threaded through the nanopore (Reuter *et al.*, 2015; Schadt *et al.*, 2010).

Nanopore technologies can be divided into two categories according to the type of nanopore that is used (biological and solid-state nanopores). The advantage of the biological nanopores is the precise definition and highly-reproducible size and structure. Biological nanopores can be also modified easily with modern techniques of molecular biology. This type of nanopores are generally inserted into a planar lipid bilayer or other polymer films (Feng *et al.*, 2015).

α-Hemolysin is the first and most commonly used biological nanopore with external dimension of 10 nm (Fig. 8). Structure of the channel is formed by a 232.4 kDa transmembrane channel, consisting of 3.6 nm diameter cap and 2.6 nm diameter transmembrane β-barrel (Song *et al.*, 1996). Nanopore structure of α-hemolysin can remain functionally stable at temperatures close to 100°C within a pH range 2-12 (Kang *et al.*, 2005).

Mycobacterium smegmatis porin A is another example of biological nanopore, which is constructed as octamer channel of 1.2 nm diameter at the minimal point (Fig. 8).

Moreover, porin A is very robust and keeps the channel active under extreme experimental conditions with a pH ranging from 0 to 14 and temperature around 100°C (Feng *et al.*, 2015).

Bacteriophage Φ29 (Fig. 8), the last biological nanopore which is listed here has a 12-subunit gp10 connector, six copies of ATP-binding DNA packaging RNA and an ATPase protein, gp16 which provides chemical energy required for DNA translocation. Φ29 channel has a large diameter in comparison to the previously listed nanopores. This allows for the measurement of larger molecules, such as dsDNA, complexes of DNA, and proteins and also provides more flexibility for biochemical modifications (Feng *et al.*, 2015).

Biological nanopores already showed few promising experimental results for ssDNA sequencing and have many advantages. Nevertheless, biological nanopores and supported lipid membranes has affection to the fragility. To remove this limitation, various synthetic nanopores have been fabricated using different methods and can be used as nanopore-based sequencing. Moreover, solid-state nanopores have many other superior advantages, such as chemical, thermal, and mechanical stability and can work under a wide variety of experimental conditions (Venkatesan *et al.*, 2009; Feng *et al.*, 2015).

The already published solid-state nanopores were prepared in various materials such as silicon nitride (Si_3N_4), silicon dioxide (SiO_2), aluminium oxide (Al_2O_3) and graphene. Al_2O_3 films have high chemical stability, improved electrical performance and lower noise during DNA translocation in compared to other solid-state films (Si_3N_4 and SiO_2; Venkatesan *et al.*, 2009; Feng *et al.*, 2015).

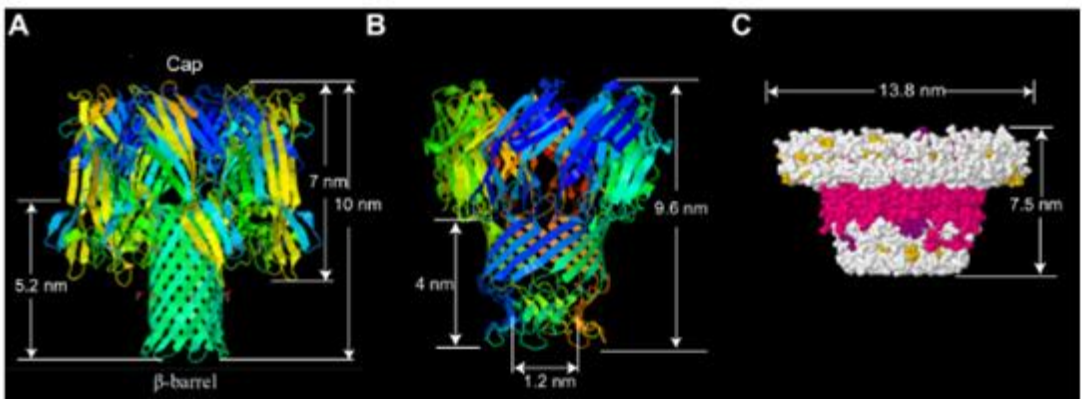


Figure 8: Side views of three biological nanopores. A. Heptameric α-hemolysin toxin from *Staphylococcus aureus*. **B**. Octameric MspA porin from *Mycobacterium smegmatis*. **C.** Dodecamer connector channel from bacteriophage Φ29 DNA packaging motor (from Feng *et al.*, 2015).

The crucial advantages of nanopore-based sequencing include long reads (10^4-10^6 bases), high throughput and low material requirement (Feng *et al.*, 2015). Moreover, the sample preparation is inexpensive due to the use of unmodified DNA and the employment of a nanopore sensors that were described above (Morey *et al.*, 2013). Nanopore DNA sequencing technology is still under rapid development because it has a very high error rate (over 90%), with most reported are insertion, deletion and substitution (Jain *et al.*, 2015; Mikheyev & Tin, 2014). Several bottlenecks hold the further development such as high translocation velocity of DNA molecule passing through the nanopore or large channel length leading to a poor spatial resolution in separating adjacent bases (Liu *et al.*, 2016).

In recent years, Oxford Nanopore Technologies focused onto the development and commercialization of nanopore-based sequencing method. The first commercially available device is the MinION, a USB-powered portable sequencer, which was released in early access trial 2014 and designed for general application of DNA sequencing (Wang *et al.*, 2015; Heather & Chain, 2016). The second device for nanopore-based sequencing is GridION, firstly presented in 2012 and can be extended with additional cartridges contains the nanopores. The GridION was designed for flexible run times ranging with respect to the experimental data requirements (Feng *et al.*, 2015; Heather & Chain, 2016).

2.2 Design of RNA-sequencing experiment

Beside microarray analyses, RNA-seq, also called whole-transcriptome shotgun sequencing, has become one of the standard methods for transcriptome analysis of model organisms and a realistic option for non-model organisms. However, this technology and the bioinformatics tools required for downstream analysis are constantly evolving (Conesa *et al.*, 2016; Wolf, 2013). Until the advent of RNA-seq, gene expression studies were restricted to small-scale quantitative polymerase chain reaction of genes or microarrays as a standard tool for gene expression quantification (Vera *et al.*, 2008). These methods are the basic tools of transcriptomics and are generally in good agreement concerning gene expression quantification (Nookaew *et al.*, 2012).

In comparison to microarray technology, RNA-seq has several major advantages. First, RNA-seq eliminates the need for prior species-specific sequence information.

Second, reads overlapping heterozygous single-nucleotide polymorphisms (SNPs) can be mapped to paternal and maternal chromosomes, providing quantification of allele-specific expression (Pastinen, 2010). Third, RNA-seq does not have an upper limit for quantification, which means that it has a large dynamic range of expression levels. By contrast, microarrays lack sensitivity for genes which are expressed at low or very high levels, and accordingly have a smaller dynamic range (Wang *et al.*, 2009a). This makes the RNA-seq an attractive method for characterizing global changes in transcriptome, which is the reason why over the past years this method has rapidly replaced microarray technology (Fonseca *et al.*, 2014).

The key aims of transcriptomics are to catalogue all species of transcripts (messenger RNA, non-coding RNA and small RNA), to determine the transcriptional structure of genes (Wang *et al.*, 2009a) and to quantify the changing expression levels of each transcript in samples (Wang *et al.*, 2008a). Discovery and quantification of the transcripts can be combined in a single RNA-seq experiment (Conesa *et al.*, 2016). Investigators might be also interested only in messenger RNA (mRNA), microRNA levels or in allele specific expression (Degner *et al.*, 2009; Conesa *et al.*, 2016). RNA-seq may also be used to investigate many different phenomena, such as SNPs (Craig *et al.*, 2008) and alternative splicing detection (Wang *et al.*, 2008a), and to detect changes in the expression of transcripts isoforms from the same gene (Jiang & Wong, 2009) and fusion of transcripts (Maher *et al.*, 2009).

RNA-seq can be coupled with different types of biochemical methods to analyse previously listed phenomena. The integration of RNA-seq data with DNA sequencing can be used for several purposes such as SNPs detection or RNA-editing analysis (Conesa *et al.*, 2016). The combination of RNA-seq and proteomic is rather controversial because the two measurements show generally low correlation (de Sousa Abreu *et al.*, 2009; Vogel & Marcotte, 2012). However, integration of these two technologies can be used to identify novel isoforms of genes (Conesa *et al.*, 2016).

Generally, integration of different datasets is quite complicated because each data type is analysed separately with its own specific algorithms that provide results in different formats. Tools providing format conversions and the extraction of relevant results can be

helpful here. Examples of such software packages include Anduril (Ovaska *et al.*, 2010), Galaxy (Goecks *et al.*, 2010) and Chipster (Kallio *et al.*, 2011).

The decisive prerequisite for a successful RNA-seq study is that the generated data have the potential to provide an answer to a given biological question. There is no optimal pipeline or analysis scenario for different types of RNA-seq applications. However, RNA-seq workflow can be generally divided into five distinct steps, namely (1) sample preparation and RNA isolation, (2) library preparation, (3) sequencing by RNA-seq technologies, (4) bioinformatics analysis of raw data, and (5) interpreting of results (Yang & Kim, 2015; Wolf, 2013). This workflow (Fig. 9) provides many variables which influence the result of RNA-seq experiments.

The number of replicates provides a crucial design factor for RNA-seq (Conesa *et al.*, 2016; Fang & Cui, 2011). Depending on both the amount of technical variability in the RNA-seq procedure and the biological variability of the system under study, this factor should be included in an RNA-seq design. For any inferential analysis, it is necessary to make biological replicates, with three replicates being the minimum (Conesa *et al.*, 2016).

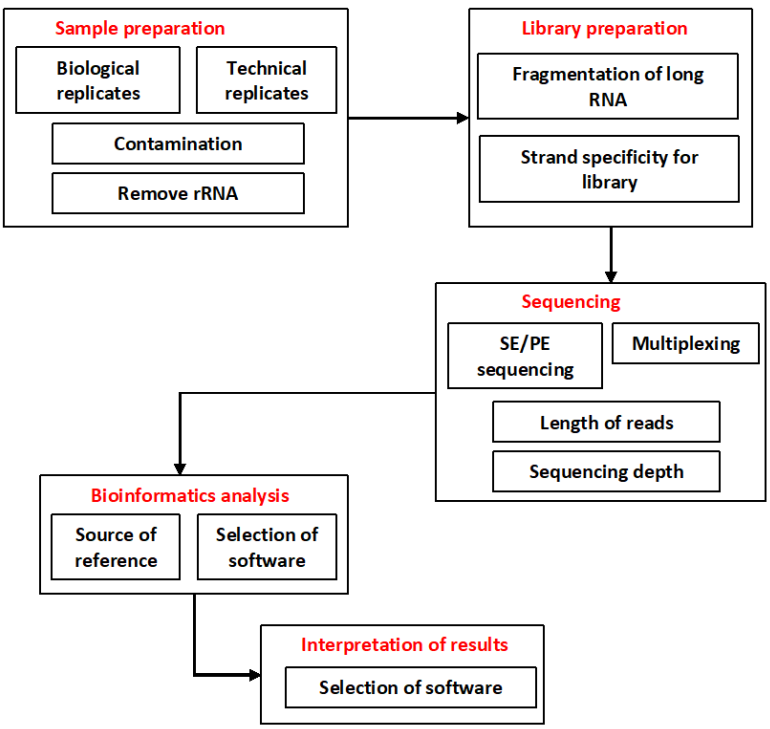


Figure 9: Diagram of RNA-seq experiments showing main steps (red color) and crucial factors which must be considered during the analysis. SE, single-end; PE, paired-end.

The use of replicates brings variation within the target population and can simultaneously counteract random technical variations as a part of independent sample preparation. The choice of biological and technical replicates is governed by time, financial and technical resources, and variability in population of interest (Conesa *et al.*, 2016, Auer *et al.*, 2012).

Arguably, instantaneous shock-freezing tissue by liquid nitrogen is still the most reliable method for sample collection to prevent fragmentation and possible loss of RNA due to RNase activity. Commercially available buffers or other home-made solutions can protect RNA at room temperature (De Wit *et al.*, 2012; Wolf, 2013).

The additional important aspect of the experimental design is the RNA-extraction protocol used to remove the highly abundant ribosomal RNA (rRNA) which normally constitutes over 90% of total RNA in the cell. mRNA extraction is provided by a wide variety of commercially available kits. Special attention should thus be paid to obviate any form of contamination. To exclude contamination, treatment with high-quality DNase is recommended. The appraisal of RNA integrity (e.g. by micro-capillary electrophoresis) is a critical step for validating RNA isolation (Wolf, 2013; Conesa *et al.*, 2016).

Prior to library preparation, rRNA needs to be removed. Poly-adenylated mRNA in eukaryotes can be enriched by capture on oligo-dT-coated magnetic beads, but this approach typically requires a relatively high proportion of mRNA with minimal degradation. If the sequence is known, an alternative way is direct elimination of rRNA from the transcript pool. For bacterial samples in which mRNA is not polyadenylated, the only usable alternative is ribosomal depletion (Künstner *et al.*, 2010; Conesa *et al.*, 2016).

Most sequencing platforms require conversion RNA to cDNA before sequencing. The enzymatic reaction of the reverse transcriptase can be primed by random hexamer primers or by the hybridization of an oligo-dT primer on to the poly-A tail of the mRNA template. Library preparation for RNA-seq is platform specific (Wolf, 2013).

Small molecules RNAs can be directly sequenced after adaptor ligation, whereas larger RNA molecules must be randomly decomposed into short reads to be compatible with most deep-sequencing technologies. Common fragmentation methods include cDNA fragmentation (DNase I treatment or sonication) and RNA fragmentation (RNA

hydrolysis or nebulization). Each of these methods creates a different bias in the outcome (Wang *et al.*, 2009a, Wolf, 2013).

Another important factor to consider is whether to generate strand-preserving libraries. Furthermore, sequencing can include single-end (SE) or paired-end (PE) reads. PE reads are preferable for *de novo* transcript discovery or isoform expression analysis and for characterization of poorly annotated transcriptomes, whereas SE reads are used for healthy annotated transcriptomes (Katz *et al.*, 2010; Garber *et al.*, 2011). Similarly, while longer reads improve mappability and transcript identification, short SE reads are normally sufficient for studies of gene expression levels in well-annotated organisms (Garber *et al.*, 2011; Łabaj *et al.*, 2011).

Each molecule, with or without amplification, is then sequenced to obtain short sequences from one end (SE sequencing) or both ends (pair-end sequencing). The length of reads depends on the used RNA-seq technology (Wang *et al.*, 2009a). Sequencing depth and multiplexing are important options for sequencing.

Sequencing depth represents the number of sequenced reads or bases represented in single sequencing experiment. If the sample is sequenced to a deeper level, more transcripts will be detected. Optimal sequencing depth depends on the complexity of the targeted transcriptome and on the aims of the experiments (Mortazavi *et al.*, 2008; Tarazona *et al.*, 2011). Sequencing depth is generally chosen based on an estimation of total number of bases in transcriptome and the expected dynamic range of transcripts abundances. Greater sequencing depth increases sensitivity to detect smaller changes in relative expression and the number of observed transcripts. However, this does not exclude the possibility of detecting smaller changes in transcriptome as only the result of tolerated fluctuations in transcript abundance (Robles *et al.*, 2012; Fang & Cui, 2011).

Multiplexing plays an important role in RNA-seq experiment. It allows the sequencing of multiple samples in a single sequencing lane or reaction, and consequently permits the reduction in sequencing costs per sample. This approach uses indexing tags, "barcodes" or short ($\leq$ 20 bp) stretches of sequence that are ligated to the start of sample sequence fragments during the library preparation step. Barcodes are different between sample libraries and allow pooling for sequencing followed by allocation of reads back to individual samples after sequencing by analysis of the sequenced barcode (Porreca *et al.*,

2007; Smith *et al.*, 2010). Randomizing samples across the lanes on the flow cell in a completely randomized fashion in combination with the smart management effectively averages out any effect that lanes might have on the gene counts and provides the crucial factor to obtain error-free data (Auer *et al.*, 2012; Conesa *et al.*, 2016).

Processing millions of short reads generated during sequencing by bioinformatics analysis order to obtain results requires computational resources and considerable bioinformatics skills. This process generally comprises aligning the short reads generated to the reference genome or transcriptome. Differences in the choice alignment/quantification tools can have a major effect upon the measurement of expression levels (Fonseca *et al.*, 2014).

The last step in a standard RNA-seq workflow is frequently characterization of the molecular functions or pathways in which differentially expressed genes (DEGs) are involved as well as a biological interpretation of obtained results (Conesa *et al.*, 2016).

2.3 Bioinformatics analysis of RNA-seq experiment

RNA-seq as a one of the high-throughput techniques for transcriptomic studies has a several bioinformatics challenges, including the development of efficient methods to store, retrieve and process of data. Commercial technologies especially second-generation RNA-seq share the limitation that read lengths are much shorter than even the transcript. This limitation is overcoming by over-sampling the target transcriptome with short reads from random positions.

Typical bioinformatics analysis of transcriptomic data contains the following steps. First, the quality check of the reads must be performed and mapping the reads to the reference genome/transcriptome (*ab initio* approach) or assembling reads into contigs before aligning them to the reference sequence to reveal transcription structure (*de novo* approach). Second, the aligned reads are used as an input for methods to identify expressed genes and isoforms, to detecting possible SNPs or to getting results for replying question of interest. Third, the obtained results must be interpreted by visualization. The crucial part of this process is defining the semantic of the genes or transcripts which can be performed by functional annotation process with assistance by specific bioinformatics tools (Wang *et al.*, 2009a; Garber *et al.*, 2011).

2.3.1 Quality control and alignment of reads

Quality control for the raw reads forms the crucial step of bioinformatics analysis and involves the analysis of adaptors presence, GC content, overrepresented k-mers and duplicated reads to detect sequencing errors, contaminations or PCR artefacts. Values for GC content and overrepresented k-mers are strictly experiment- and organism-specific and should be homogeneous for samples in the same RNA-seq experiment (Conesa *et al.*, 2016; Yang & Kim, 2015). Several bioinformatics tools have been developed for quality evaluation. One of the most popular bioinformatics tools is FastQC (Andrews, 2010) specific to Illumina sequencing platform, in contrast to NGSQC (Dai *et al.*, 2010) which can be applied to any sequencing platform. As well, software HTQC (Yang *et al.*, 2013) can be used to assess the quality of raw reads, providing assessment of the overall and per base quality of reads in each sample.

Raw-reads are produced by sequencers and formatted in FASTQ which is utilized by majority of sequencing platforms. Reads (single-letter base calls) plus base quality value for each base call (Ewing & Green, 1998) are encoded based on ASCII (American Standard Code for International Interchange) characters which each ASCII character corresponds to a specific quality score (Góngora-Castillo & Buell, 2013). Low quality score indicates possible sequencing errors.

Read trimming (adapter trimming or quality trimming) can improve the accuracy before the aligning of RNA-seq reads. Adapter trimming involves removal of the adapter sequence by excluding specific sequence which was used during library preparation. Quality trimming is based on the fact that, read quality is significantly decreased towards the 3′ ends of reads and if quality is to low, bases should be removed to improve mappability. Consequently, the ends of reads can be removed by quality trimming (Conesa *et al.*, 2016; Yang & Kim, 2015). Software tools such as Trimmomatic (Bolger *et al.*, 2014), FASTX-Toolkit (FASTX-Toolkit, 2015) and FLEXBAR (Dodt *et al.*, 2012) can be used for both types of trimming and discard low-quality reads.

Alignment of reads to either a reference transcriptome or genome is a second step in bioinformatics analysis of sequencing reads. Genome sequence coverage level is generally uniform across the genome while transcriptome coverage levels are dependent

on gene expression level which must be considered. It forms the crucial difference between alignment of reads to the reference genome and reference transcriptome. Likewise, ambiguities in the genome assembly are generally attributed to repetitive sequences and occasionally ally homologous sequences in compare to variation of transcript isoforms, which contributes to complexity of transcriptome. Single base differences are not problematic for alignment of reads, because most mapping algorithms consider predicting of one or two base difference. On the other side, detecting large differences requires better reference genome/transcriptome and deeper sequencing coverage (Miller *et al.*, 2010; Wang *et al.*, 2009a).

The reference for alignment of reads is represented by assembly genome or transcriptome. Assembly is closely related with the alignment of reads and can be defined as a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. Assembly groups reads into contigs and contigs into supercontigs/scaffolds which may have a topology like a simple path or network. Contigs provide a multiple sequence alignment of reads and the consensus sequence which is represented by reference genome/transcriptome. The supercontigs sometimes called metacontigs, define the order, orientation of the contigs, size and orientation possible gaps between contigs (Miller *et al.*, 2010).

Alignment of reads and as well as assembly of reference uses the similar principles (aligning fragments of nucleic acids) and are under influence of few problems. Significant portion of sequence reads match multiple locations in the genome forms a multimapping problem which can be solved by assign these multi-matched reads by proportionally assigning them based on the number of reads mapped to their neighbouring unique sequence or by excluding reads from alignment (Mortazavi *et al.*, 2008; Wang *et al.*, 2009a). Alternatively, multimapping problem can be reduced by use a PE sequencing strategy. In this strategy reads are determined from both ends of a DNA fragment (Campbell *et al.*, 2008; Wang *et al.*, 2009a). Alignment of reads is made more difficult because software must tolerate imperfect sequence alignment which can lead to false positive alignment. Consequently, false positive alignment can induce chimeric assemblies which represents unreal transcripts (Martin & Wang, 2011).

Also, the short-reads are source of complications such as error rate inherent in short reads, lack of consideration of sequence quality scores in most transcript assembly algorithms and limitations in definitive quality assessment methods for the assembly. Consequence of this is influencing the quality of assemblies (Martin & Wang, 2011; Góngora-Castillo & Buell, 2013).

Some of sequencing reads also contain span exon junction or poly(A) ends and these types of reads cannot be analysed in the same way. Poly(A) tails contain in reads can be simply identified by a presence of multiple adenine or thymine at the end of some reads. Exon-exon junctions which are covered by reads can be identified by the presence of a specific context sequence and can be confirmed by low coverage of intronic sequences, which are removed during alternative splicing. Effectivity of identify alternative splicing isoforms is increased if the pairing information from PE reads is used as well as higher coverage of sequences (Wang *et al.*, 2009a; Góngora-Castillo & Buell, 2013).

Values which represents quality of transcriptome assembly can be used for evaluation of alignment of RNA-seq reads. Percentage of mapped reads forms an important mapping quality parameter and serves as a global indicator of the overall sequencing accuracy and of the presence contaminating fragments of DNA. If reference transcriptome is used for alignment, slightly lower total mapping percentages are expected because reads coming from unannotated transcript are lost (Conesa *et al.*, 2016). Quality of alignment can be examined also by statistics including maximum length of contig, average length, combined total length and N50 value. Value N50 represents length of the smallest contig in the set that contains the longest contigs whose combined length represents at least 50% of the assembly (Miller *et al.*, 2010). Quality of alignment can be examined by bioinformatics tools include Picard (Picard, 2009), RSeQC (Wang *et al.*, 2012) and Qualimap (García-Alcalde *et al.*, 2012).

Regardless of whether a reference genome or transcriptome is used, reads can be mapped uniquely or as multi-mapped reads. When the reference is the transcriptome, multi-mapping is increased because reads could be mapped on all gene isoforms in the transcriptome that share the exon in contrast to reference genome. If a reference is transcriptome, unspliced aligners should be used for accurate read mapping. These aligners are limited to the identification of already known exon and junction. (Martin &

Wang, 2011; Yang & Kim, 2015). Unspliced aligners do not allow large gaps and fall into two main categories, “seed method” and “Burrows-Wheeler transform methods” (Fig. 10).

“Seed methods” are based on finding matches for short subsequence’s, called “seeds”, if that at least one seed in a read will have perfect match with the reference. Each seed is used to obtain limited candidate region and more sensitive algorithms (such as Smith-Waterman algorithm) can be applied to extend seeds to full alignments (Garber *et al.*, 2011). There are several programs which fall into group of seed methods such as Mapping and Assembly with Quality (MAQ; Li *et al.*, 2008) and Stampy (Lunter & Goodson, 2011).

In contrast, the second group of methods which contains software’s such as Burrow-Wheeler Aligner (BWA; Li & Durbin, 2009) and Bowtie (Langmead *et al.*, 2009) is based on the “Burrows-Wheeler transformation” (BWT). These software’s compact the genome into an indexed data structure. However, the performance of these methods is significantly decreased if the mismatches are allowed. BWT methods provides faster performance than seed methods with only small differences in specificity of alignment. Beyond, seed methods are more sensitive if a transcriptome of distant species is used as a reference (Garber *et al.*, 2011).

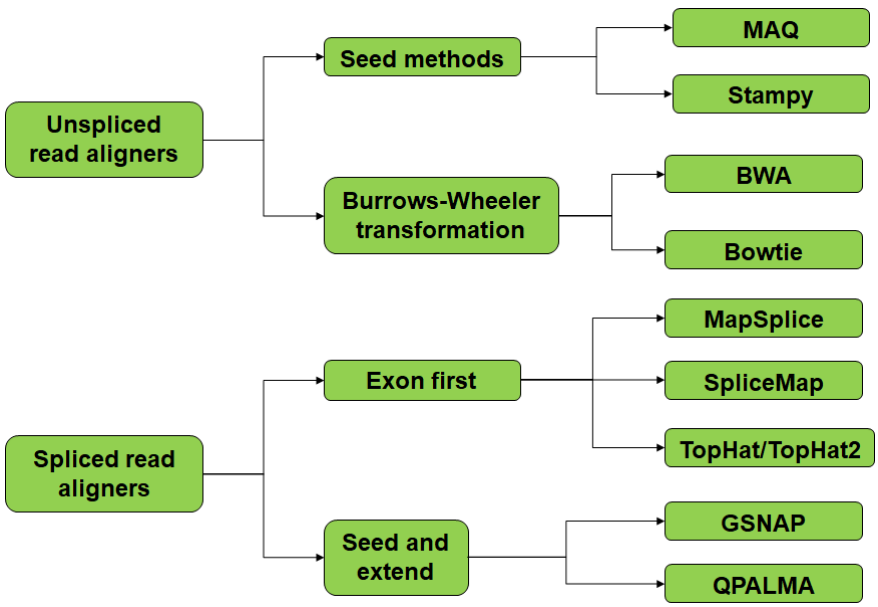


Figure 10: Scheme showing the division of the different groups of aligners and examples.

If the genome is a reference, should be employed spliced aligners (Fig. 10), because in this case reads aligned at exon-exon junctions will be split into two fragments. Spliced aligners allow a wide range of gaps and also increase probability of identifying novel transcripts, which are generated by alternative splicing events (Yang & Kim, 2015; Martin & Wang, 2011). Spliced aligners fall into two main categories called as "exon first" and "seed and extend".

"Exon first" methods are based on the twostep process. First, reads are continuously mapped to the genome by unspliced read aligner. Second, unmapped reads are split into short segments and independently aligned to the reference. The genomic regions which contain mapped reads are than searched for possible spliced connections (Garber *et al.*, 2011). Various aligners based on the exon first methods have been developed, including tools such as TopHat2 (Kim *et al.*, 2013), MapSplice (Wang *et al.*, 2010) and SpliceMap (Au *et al.*, 2010).

Alternatively, "seed and extend" methods break reads into short seeds, which are placed onto the genome to localize the alignment. Regions which contains mapped reads are then examined by more sensitive algorithms such as Smith-Waterman algorithm or iterative extension. Initial seeds are merging to determine splice events for the reads. Seed and extend methods examine spliced and unspliced reads together, which yielding the best placement of each read (Martin & Wang, 2011; Garber *et al.*, 2011). Few aligners based on the seed and extend method can be listed such as Genomic Short-read Nucleotide Alignment Program (GSNAP; Wu & Nacu, 2010) and Optimal Spliced Alignments of Short Sequence Reads (QPALMA; De Bona *et al.*, 2008).

"Exon first" methods are generally optimized to align reads with few errors, are faster than "seed and extend" methods and require fewer computational resources. Aligners are usually different in implementation of aligning reads across introns. "Exon first" methods also create database of all possible combinations of splicing junction a genomic locus in contrast to "seed and extend" methods, which shift the gaps in the local alignment to match known splice sites (Martin & Wang, 2011).

If reference genome/transcriptome is unknown or only partially known, transcriptome reconstruction should be provided and this forms a difficult computational task for few reasons such as a wide range of gene expression, diversity of alternative splice forms

generated from each locus and presence incompletely spliced precursors mRNA. Moreover, plant genome provides source of diversity by whole or segmental genome duplication and polyploidy (Garber *et al.*, 2011; Góngora-Castillo & Buell, 2013).

Transcriptome reconstruction can be defined as the identification of all transcripts expressed in a specimen. The product of reconstruction is transcriptome assembly.

The choice of transcriptome assembly strategy forms a crucial point for the subsequent alignment of reads and depends on many factors, including the existence or completeness of a reference, computer resources and the type of data set generated (length of reads, number of reads; Martin & Wang, 2011; Conesa *et al.*, 2016).

There are several strategies to reconstruct the transcriptome, and they can be divided into three classes, namely genome-guided reconstruction (*ab initio*), genome independent reconstruction (*de novo*) and combined strategy. Genome-independent methods are the necessary choice for organisms without a reference sequence, whereas genome-guided methods are used for annotating organisms with a reference genome. Generally, genome-independent methods require considerable computational resources whereas genome-guided methods are less demanding (Yang & Kim, 2015; Garber *et al.*, 2011).

After choice of assembly strategy, the other challenge is the choice of assembly program to use. A large number of assemblers were developed for using in assembly of particular organism, sequencing platform and for perform better on a similar data set. The many assemblers are consequently determined for specific sequencing platform and their output, is a set of unassembled or partially assembled reads (Martin & Wang, 2011).

Genome-guided methods fall into two categories which are “exon-identification” and “genome-guided assembly” approach. Both groups are based on mapping all reads to the reference genome and then assemble overlapping reads into transcripts (Garber *et al.*, 2011).

“Exon identification” methods such as G.mor.se (Denoeud *et al.*, 2008) defined exons as coverage islands and then use spliced reads to define exon boundaries. These methods cannot identify full-length structures of lowly expressed and alternatively spliced genes which is essential disadvantage (Garber *et al.*, 2011).

“Genome guided assemblers” use spliced reads directly to reconstruct the transcripts and could incorporate already existing annotation by adding them to the list of possible

isoforms. Several software's can be listed such as Cufflinks (Trapnell *et al.*, 2010), Scripture (Guttman *et al.*, 2010), StringTie (Pertea *et al.*, 2015), iReckon (Mezlini *et al.*, 2013) and SLIDE (Li *et al.*, 2011). Cufflinks constructs an overlap graph from reads aligned to locus and then transverses this graph to obtain minimum set of transcripts explaining intron junction whereas Scripture, creates a splice graph which containing each base of a chromosome and adds join sites between bases as edges. Consequently, Cufflinks provide conservative approach to reconstruct transcripts in contrast to Scripture which produces a larger set of possible transcripts per locus. Both of software's requires a similar hardware resources (Martin & Wang, 2011; Garber *et al.*, 2011).

Genome-guided methods have a several advantages. First, relatively low computational resources are necessary for launch insomuch as they reduced the complex assembly problem to a many individual problems. Second, current methods provide high sensitivity and can detect new isoforms with low abundance. On the other hand, there are also few drawbacks such as depend of effectivity on the quality of the reference genome, or possible exclude of splice reads, during the assembly process (Martin & Wang, 2011). Software Augustus (Stanke *et al.*, 2006) can be used for incorporate RNA-seq data to better annotate protein-coding transcripts, however his achievement is worse for non-coding transcripts (Engström *et al.*, 2013).

Alternative for *ab initio* approach with reference is genome independent reconstruction which uses specific algorithms to directly build consensus transcripts from short reads. Partition reads into disjoin components, which represents all possible isoforms of a gene forms a central challenge for this approach. Genome-independent strategy is conceptually simple, however there are two major complications termed distinguishing sequencing errors from variation and finding the balance between graph complexity and sensitivity.

A commonly strategy use short subsequence's, termed k-mers for build a de Brujin graph and overlaps of k-1 bases between these k-mers forms graph of all possible sequences. Set of sub-sequences of reads (k-mers) are connected by edges. Then, paths are traversed in the graph and PE can eliminate false branch points introduced by k-mers, shared by different transcripts. Each remaining path is then reported as a separate transcript (Garber *et al.*, 2011).

The choice of the k-mer length is essential for the assembly process. Smaller value of k parameter leads to a larger number of overlapping nodes (consequently higher hardware requirements) and to the more complex graph, whereas larger values of k reduce a complexity of graph and leads to a simpler graph structure and provides better performance on highly expressed transcripts. Optimal k-mer value depends on the read length, sequencing depth, read error rate and complexity of transcriptome (Góngora-Castillo & Buell, 2013; Garber *et al.*, 2011).

Similarly, the high number of reads increases the hardware requirements and therefore reduce the number of reads is highly recommended for samples which are deeply sequenced (Haas *et al.*, 2013). One strategy for obtain less set of reads is to sub-sample the read pool by randomly selecting reads from all samples thereby approximating the representation of transcripts in the original RNA samples. However, low abundant transcripts may not be represented in result transcriptome assembly. For comparative analysis, it is recommended to combine all reads for studied samples to obtain a consolidated set of contigs (Garber *et al.*, 2011; Góngora-Castillo & Buell, 2013).

Also, length of reads is well important for genome independent methods because longer reads provide more accurate and robust assemblies and simpler construction of the de Bruijn graph. Moreover, the use of PE reads will significantly improve the assembly as the assembly algorithms can use the directionality of the reads in contrast to the SE reads (Góngora-Castillo & Buell, 2013).

A handful of assemblers may be used for genome independent assembly, such as Trinity (Haas *et al.*, 2013), Oases (Schulz *et al.*, 2012), transABySS (Robertson *et al.*, 2010), Rnnotator (Martin *et al.*, 2010) and Multiple-k (Surget-Groba & Montoya-Burgos, 2010).

A crucial advantage of the *de novo* assembly is that does not depend on the correct alignment of reads to known splice sites and provide alternative to prediction of novel splicing isoforms. On the other hand, these methods have several disadvantages. First, genome independent assembly requires a significantly higher sequencing depth for full-length transcriptome assembly in comparison to reference-guided methods. Second, *de novo* transcriptome assemblers are highly sensitive to sequencing errors and to presence of chimeric reads in the input data set (Cocquet *et al.*, 2006; Martin & Wang, 2011).

Genome-guided and genome independent strategies can be combined to create better reference transcriptome. This strategy can take advantage of the high sensitivity of reference-based assemblers and the possibility of detecting novel isoforms of *de novo* assemblers (Martin & Wang, 2011).

The combined approach should start by assembling the reads by using reference genome. Then the unmapped reads are assembled by *de novo* approach. This align-then-assemble approach is also popular method for filtering out unwanted sequences (detection of pathogen). If the reference genome is incomplete or from closely related species, *de novo* assembly should be performed first and should be followed by alignment contigs to reference. Incomplete transcripts can be merged and can form possibly full-length transcripts by approach assemble-than-align (Martin & Wang, 2011). Combined approach was already used on transcriptome of mosquito and catfish (Crawford *et al.*, 2010; Surget-Groba & Montoya-Burgos, 2010).

Beyond transcriptome reconstruction is also crucial to check the global quality of the RNA-seq dataset by checking on the reproducibility which should be generally high among of technical replicates whereas reproducibility of biological replicates depends on the heterogeneity of the experimental system. Principal component analysis (PCA) serves as a valuable tool for this purpose and biological replicates will cluster together if the same experimental conditions for them are fulfilled (Conesa *et al.*, 2016).

2.3.2 Qualitative and quantitative analysis of transcriptome

Quantitative analysis of transcriptome starts with extraction the numbers of reads that's are uniquely mapped on each contig or gene from reference. These numbers serve as an instrumentality for quantification of expression, which can be compared among tissues, ecotypes or between control and treatment (De Wit *et al.*, 2012; Yang & Kim, 2015). Obtained numbers of reads are not satisfactory to compare expression among samples, because the values are affected by sequencing biases, total number of reads and length of transcript. Few normalization methods were developed for removing impact of transcript length and total number of reads. The widely used methods are RPKM (reads per kilobase exon model per million reads; Mortazavi *et al.*, 2008) and FPKM (fragments per kilobase

of exon model per million mapped reads; Fig. 11). The values for FPKM and RPKM are equivalent in case of SE reads (Conesa *et al.*, 2016).

Numerous methods have been development for expression quantification, and they can be divided into two groups based on the target levels: quantification of the gene and quantification of the alternative splicing forms (isoforms). Expression of the gene is defined as the sum of the expression of all its isoforms (De Wit *et al.*, 2012; Yang & Kim, 2015).

Great part of genes has multiple isoforms and many of which share exons so some reads cannot be assigned uniquely to the transcript, which reduces number of reads which can be used for quantification of isoforms. Therefore, calculating expression of isoforms can be computationally challenging especially for genes with high number of isoforms. Methods for isoform level quantification fall into three groups depending on the requirement of aligned reads and type of reference (Yang & Kim, 2015).

The first group includes the RSEM (Li & Dewey, 2011) and requires transcriptome as reference and sequencing reads for alignment. RSEM is software able to quantify transcripts and this process can be divided into two steps. Firstly, reference transcripts sequences are generated from FASTA file with genome sequence and gene transfer format file (GTF; UCSC, 2006).

Then, a Bowtie alignment program (Langmead *et al.*, 2009) or another user-provided aligner is used for alignment of reads to the reference created in previous step. RSEM then use Expectation-Maximization algorithm for investigate gene-level and isoform level expression (Yang & Kim, 2015).

The second group of methods for quantification of isoforms also requires alignment of sequencing reads but works with genome as a reference in compare with the first group of methods.

$$a)\ RPKM = \frac{total\ exon\ reads}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

$$b)\ FPKM = \frac{total\ fragments}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

Figure 11: Formula for computing of RPKM (a) and FPKM(b).

As a software in the second group can be listed Cufflinks (Trapnell *et al.*, 2010) and StringTie (Pertea *et al.*, 2015). Cufflinks represents widely used software and require mapped reads in bam format as input from TopHat2 (popular spliced aligner). Abundance examination is performed by estimating the maximum likelihood abundance and is export as numbers under FPKM normalization in case of PE or RPKM normalization in case of SE reads respectively. The algorithm is based on the transcript coverage, with respect to the fragment length distribution and was designed to take advantages of PE reads (Yang & Kim, 2015). StringTie similarly like Cufflinks also use alignment results from available aligners for isoforms abundance estimation and can perform transcriptome reconstruction parallel by building a flow network for the path of the best coverage. The results of isoform expression are exported in RPKM for SE reads and FPKM for PE reads (Yang & Kim, 2015).

The last group of methods consists the alignment free methods and Sailfish (Patro *et al.*, 2014) can be listed as a representative software that rely on k-mer counting in reads without of their alignment (Yang & Kim, 2015; Patro *et al.*, 2014). Sailfish initially build the index from a set of reference transcripts and length of k-mers and then computes the isoform´s abundance. The results are reported in RPKM (Yang & Kim, 2015).

In contrast to quantification of isoform abundance, quantification of gene expression is conceptually simpler and most widely used solution is to aggregate counts of mapped reads with software such as HTSeq-count (Anders *et al.*, 2015) or featureCounts (Liao *et al.*, 2014). The GTF file must be included as one of the inputs. A few software packages and pipelines have been developed for analysis of differential gene expression, including edgeR (Robinson *et al.*, 2010), DESeq2 (Love *et al.*, 2014), NOIseq (Tarazona *et al.*, 2015) or SAMseq (Li & Tibshirami, 2013). Methods, such as edgeR use aggregated reads numbers and perform their normalization as well as a differential expression analysis. Similarly, like edgeR, DESeq2 uses the negative binomial distribution and works with own normalization approach (Anders & Huber, 2010; Love *et al.*, 2014). In contrast with these software's, NOIseq (Tarazona *et al.*, 2011) and SAMseq (Li & Tibshirami, 2013) are based on the nonparametric approach and have minimal assumptions about the data and estimate the null distribution for differential analysis (Conesa *et al.*, 2016).

Qualitative analysis of transcriptome includes detection of SNPs and identification novel alternative gene-splicing forms. Qualitative analysis of alternative splicing is strictly connected with reconstruction of transcriptome, which were fully described in previous subchapter. The programs which are used for transcriptome reconstruction can be also applied for detection of novel gene-splicing forms.

SNPs can be listed as one of the fundamental types of genetic variation, and with the growing popularity of next-generation sequencing they are becoming the widely used genetic markers. If mapped read mismatched to a reference at a single base, it can be a mistake of sequencing or a polymorphism. The distribution of SNP in reads on the same position in reference determines the homozygote or heterozygote for that SNP (De Wit *et al.,* 2012). Crucial problem is separation true polymorphisms from sequencing or alignment artefacts and discard data with low quality. Few software packages can be listed as suitable for work with sequencing data in connection with SNPs detection. The typical example is SAMtools (Li *et al.,* 2009), Picard (Picard, 2009) or GATK (McKenna *et al.,* 2010).

In conclude, it is important to depict the results from qualitative and quantitative analysis together with semantic information about function and localization of gene products. The typical example where this is crucial is that entire metabolic pathway can be differentially expressed without a single gene in the pathway being significant. This purpose is realized by annotation of transcripts and visualization of results together with differential expression information (Ogata *et al.,* 1999; Ashburner *et al.,* 2000).

2.3.3 Annotation of novel transcripts

Annotation is term used to describe two processes namely structural annotation and functional annotation. The first is the process of identifying genes/transcripts and their intron-exon structure in reference genome/transcriptome, whereas the functional annotation is the process of attaching information about function of gene/transcript. This subchapter focuses on the functional annotation which forms one of the major challenges of current bioinformatics (Yandell & Ence, 2012).

Functional annotation of genes which are differentially expressed is frequently used for biological interpretation of results by enrichment analysis tools. An extensive number

of software's have been development for this purpose such as DAVID (Huang *et al.*, 2007), BINGO (Maere *et al.*, 2005) or Onto-express (Khatri *et al.*, 2002). These tools generally using statistical methods such as Fisher´s exact test (Fisher, 1935) or χ^2-test for calculate the enrichment *p*-value of pathway for a user defined list of interesting genes. The first test is used for pathways with small numbers of genes whereas χ^2-test is adequate if number of genes is greater than five (Hong *et al.*, 2014).

Functional annotation of transcripts from non-model species can be also used for comparative genomic study and generating valuable information about species-specific genes or genetic diversity (Van Bel *et al.*, 2013).

Transcripts for functional annotation are obtained during genome or transcriptomic sequencing process. There are a few possibilities for determining function of transcript. One way is to compare transcript sequence with known genes in selected database and assume that homology tells something about function of gene or protein as a gene product. Other possibility for annotation of transcript is determining protein function based on the structure similarity or non-homology prediction (De Wit *et al.*, 2012; Lee *et al.*, 2007).

Homology search depends on existence of bioinformatics databases as a source of annotation information. Databases can be divided into three types: largescale public repositories, community-specific databases and project-specific databases (Rhee *et al.*, 2006).

The first group includes GenBank (Wheeler *et al.*, 2005), InterPro (Finn *et al.*, 2017) or Protein Data Bank (Deshpande *et al.*, 2005). The second group typically involves databases which are focused on studying model organisms or clade-oriented comparative databases. They consist information curated with high standards and exact form specification and typical example is metabolism and protein modification. The last group of databases is habitually created for project data management and their structure, focus and duration depend on the project and funding period (Rhee *et al.*, 2006).

The databases are mostly organized as relational databases where data are represented as entities, attributes (properties of the entities) and relationships between the entities which can be represented as Entity-Relationship diagram. Entities and attributes become tables and columns of the database which are filled up by data. Relational databases provide powerful approach for storing large quantities of data but representation of

complex relationships such as signal transduction pathways can be source of problems. Programming language MySQL (Widenius & Axmark, 2002) is widely used for implementation of databases due to simple syntax that requires minimal programming knowledge (Rhee *et al.*, 2006).

InterPro can be listed as typical protein database. Database integrates 13-member databases which are described on Tab. 1 (Sangrador-Vegas *et al.*, 2016) that each provide models to help classify proteins. Proteins can be classified based on structure, sequence or function.

There are also other sources of functional annotation such as MetaCyc, Gene Ontology (GO), KEGG, COG (Tatusov *et al.*, 2003) and its eukaryotic extension KOG (Lee *et al.*, 2007).

Table 1: Description of member databases which are included in InterPro (Finn *et al.*, 2017). Signatures describes a specific subscription which is characteristics for a group of proteins.

Name of database	Description	Number of signatures
CATH-Gene3D	Database of protein families and domain architectures in complete genomes.	6 119
HAMAP	Protein annotation database of well-conserved proteins families or subfamilies.	2 246
PANTHER	Large collection of protein families that are subdivided into functionally related subfamilies, using human expertise.	90 742
Pfam	Large collection of multiple sequence alignments and hidden Markov models focused on many common protein and families.	16 712
PIRSF	Protein classification system reflecting the evolutionary relationship of full-length proteins and domains.	3 285
PRINTS	Database focused on individual protein motifs which forms protein fingerprints.	2 106
ProDom	Domain database focusing on homologous domains.	1 894
PROSITE	Database of protein families and domains consisting biologically significant sites, patterns and profiles.	2 519
SMART	Database focusing on identification and annotation of genetically mobile domains and the analysis of domain architectures	1 312
SUPERFAMILY	Collection of structural domain superfamilies' for detect homologues.	2 019
TIGR-FAMs	Collection of protein families. Protein are divided based on their specific molecular function.	4 488
CDD	Database for annotation of protein consisting annotated multiple sequence alignment models for ancient domains and full-length proteins.	12 805
SFLD	Structure-Function Linkage Database providing hierarchical classification of enzymes in relation to specific structure of sequence.	303

Ontology is determined as a set of vocabulary terms whose meaning and relations with other terms are explicitly stated and GO together with other ontologies such as Sequence Ontology (SO) project (Eilbeck *et al.*, 2005) and the Plant Ontology (PO) project (Consortium PO, 2002) are typical open biological ontologies (Rhee *et al.*, 2006) and are still under active development. Ontologies besides of functional annotation can be also used to test the robustness of semantic similarity searching methods (Lord *et al.*, 2003) and to study adaptive evolution (Aris-Brosou, 2005).

The SO project is focused on define all the terms needed to describe features on a nucleotide sequence, whereas the PO project aims to develop shared vocabularies to define anatomical structures for flowering plants to depict gene expression patterns and plant phenotypes (Rhee *et al.*, 2006).

GO forms the part of the added information provided by InterPro and is the one of the more comprehensive ontologies within the bioinformatics community. It is a rapidly growing connection of about 16 000 phrases, representing terms for the biological domains of "biological process" (BP), "molecular function" (MF) and "sub-cellular component" (CC; Lord *et al.*, 2003; Rhee *et al.*, 2006). MF describes activity on the molecular level whereas BP defines broader functions that are carried out by assemblies of MF. CC describes the compartment or compartments of a cell in which location of gene product is predicted or experimentally confirmed (Lee *et al.*, 2007).

The GO terms should have definition and be placed within a structure of relationships where the most important are the 'is-a' relationship between parent and child and the 'part-of' relationship between part and whole. The structure of GO can be viewed as directed acyclic graph, part of which is shown on Fig. 12. Children´s terms can have multiple parents, as well as multiple children based on the relationship 'is-a'. Each GO term is alphanumerical code that correspond with the metabolic roles known for a gene (De Wit *et al.*, 2012; Lord *et al.*, 2003).

Since the GO terms are not a static and are regularly updated, so also the InterPro is never complete stable as a database (Sangrador-Vegas *et al.*, 2016). In the last decade, number of software tools for visualising, editing and analysing ontologies was developed such as KEGG Automatic Annotation Server (KAAS; Moriya *et al.*, 2007), Blast2GO

(Conesa *et al.*, 2005), TRAPID (Van Bel *et al.*, 2013) and T-ACE (Philipp *et al.*, 2012). These software's can be used for model- as well as for non-model organisms.

TRAPID is a web based and high-throughput analysis pipeline, that used predefined reference database. Analysis of sequence in this software include automatic identification of coding sequence, which is follow by assigning of sequence to multi-species gene families. Finally, TRAPID generating the functional annotations after performing transcript quality control (Van Bel *et al.*, 2013). On the other hand, software such as Blast2GO can work with public databases such as "nt" or "nr" (NCBI Resource Coordinators, 2014) database. As well as TRAPID provides automatic functional annotation of input sequences (Conesa *et al.*, 2005).

A useful source for annotating enzymes is "Enzyme Commission" (EC) numbers which are assigned to the chemical reactions they catalyse. They comprise a hierarchical set of four numbers describes enzyme class, the type of bond or group that is acted on and specification of the catalysed reaction and its substrates. If the two enzymes catalyse the same reaction, can be described with the same EC number. Pairwise sequence identity more than 40% is generally used as a confident threshold to transfer the first three digits of an EC number, whereas for transfer all four digits of an EC number, the sequence identity higher than 60% is required (Chen *et al.*, 2012; Tian & Skolnick, 2003; Rost, 2002).

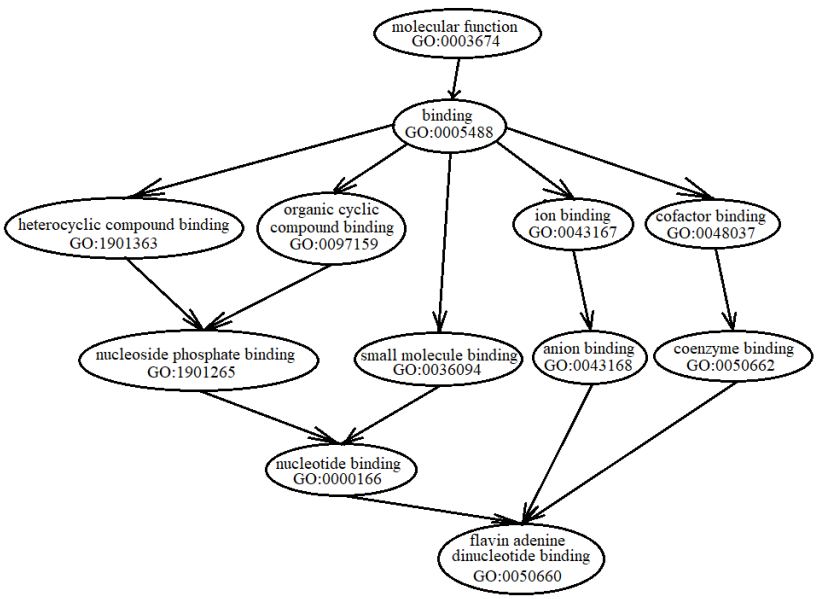


Figure 12: Directed acyclic graph with gene ontology (GO) terms (adapted from Lord *et al.*, 2003).

As the volume of data in databases has continually increased, the number sophisticated of computational methods has been developed for determination inheritance of functional annotation through homology which forms a popular approach to function prediction by sequence alignment (Lee *et al.*, 2007).

Comparing sequences provides a foundation for many bioinformatics tools and is based on sequence similarity between two strings of text. For pair-wise sequence alignment, FASTA and BLAST (Altschul *et al.*, 1990) provide popular tools and the expectation value is used for assessing the confidence level of an alignment result. The BLAST software has several utilities for comparing sequences such as blastn for nucleotide sequences or blastp for protein sequences (Rhee *et al.*, 2006).

The success of this process can be increased by identifying specific residues that discriminate between functions. In some proteins, specific residues can be predicted as a ligand and protein binding sites and these functionally active residues are most likely to have been conserved over evolution (Lee *et al.*, 2007).

The data stored in database can be also used for phylogenetic analysis in comparative genomics. The analysis is typically started by aligning the related proteins using tools such as ClustalW (Thompson *et al.*, 1994).

If the homology-based approach for determining function of gene product cannot be used, the non-homology-based methods forms a good alternative. Approach uses subcellular localization and other protein features such as membrane association and post-translational modifications. The ProtFun (Jensen *et al.*, 2002) is software to predict function of protein based on the post-translational modifications, protein-sorting signals and basic properties of protein sequence such as amino-acid composition, isoelectric point and length (Lee *et al.*, 2007).

For proteins which have crystallographic structure already defined, structure-based prediction can be used for function prediction but currently the three-dimensional structures of <1% of sequences have been experimentally solved. Nevertheless, structured data can serve as an alternative approach to detect proteins with similar function (Lee *et al.*, 2007).

CHAPTER 3.0

TRANSCRIPTOMIC ANALYSIS OF BARLEY TRANSGENIC LINES WITH ALTERED CYTOKININ STATUS AND APPARENT TOLERANCE TO DROUGHT

Vojta, P., **Kokáš, F.**, Husičková, A., Grúz, J., Bergougnoux, V., Marchetti, C.F., Jiskrová, E., Ježilová, E., Mik, V., Ikeda, Y., Galuszka, P. (2016). Whole transcriptome analysis of transgenic barley with altered cytokinin homeostasis and increased tolerance to drought stress. *New Biotechnol. 33*, 676-691. https://doi.org/10.1016/j.nbt.2016.01.010

Kokáš, F., Vojta, P., Galuszka, P. (2016). Dataset for transcriptional response of barley (*Hordeum vulgare* L.) exposed to drought and subsequent re-watering. *Data in Brief 8*, 334-341. https://doi.org/10.1016/j.dib.2016.05.051

3.1 Introduction

Cytokinins (CKs) are plant hormones which together with auxins mainly influence plant morphology. Their role in other physiological processes, such as senescence and nutrient remobilization, is very well described (Zalabák *et al.*, 2013). Evidence from studies mostly conducted with the plant model *Arabidopsis* suggest also an important role of CKs in the regulation of responses to environmental stresses (Ha *et al.*, 2012). Indeed, CK-deficient *Arabidopsis* plants exhibited a strong stress-tolerant phenotype associated with increased cell membrane integrity and abscisic acid (ABA) hypersensitivity (Nishiyama *et al.*, 2011). In this context, ABA mainly induces stomatal closure to prevent water losses under conditions of water limitation. Moreover, loss-of-function mutants of CK receptors and proteins involved in the CK-signalling pathway are strongly tolerant to drought and salt stress; their tolerance was related to the accumulation of many stress-inducible genes (Nishiyama *et al.*, 2013; Tran *et al.*, 2007). Similarly, rice seedlings with knock-down proteins of the CK transduction pathway have been observed to be tolerant to osmotic stress but hypersensitive to salt stress (Sun *et al.*, 2014). To summarize, functional CK receptors negatively control osmotic stress responses and thus confirm that reduced CK status is a prerequisite for better drought tolerance.

Increased drought tolerance or avoidance by stress-induced CK accumulation has been proven in several plant species (Zhang *et al.*, 2010; Ghanem *et al.*, 2011; Merewitz *et al.*, 2011; Peleg *et al.*, 2011; Kuppu *et al.*, 2013). A transgenic approach has exploited expression of the CK biosynthetic *isopentenyl transferase* (*IPT*) gene under a stress- and maturation-inducible promoter. Under drought-stress conditions, the transgenic plants maintained high photosynthetic activity in contrast to control plants due to the direct CK effect on delaying leaf senescence. The acquired drought tolerance was also related to maintenance of nitrate acquisition from soil (Reguera *et al.*, 2013). Thus, stress-induced CK synthesis in these transgenic plants promoted sink strengthening through the maintenance and coordination of N and C assimilation during water stress.

In abiotic stress responses, CKs can act together with other phytohormones. Auxin's role in drought tolerance has been demonstrated when increased activity of auxin conjugating enzyme, which reduces auxin maxima in leaves, led to the accumulation of late-embryogenesis abundant proteins responsible for the switch from plant growth to

stress adaptation (Zhang *et al.*, 2009). Auxin is able to induce the expression of genes encoding enzymes participating in biosynthesis of such stress-related hormones as ethylene (Tsuchisaka & Theologis, 2004); vice versa, ethylene promotes local auxin biosynthesis and consequently reduces root cell elongation (Růžička *et al.*, 2007). As CKs are known to affect production of both auxin and ethylene, a coordinated regulation of hormonal biosynthetic pathways could be crucial for plants' adaptation to abiotic stresses (Peleg & Blumwald, 2011). Plants with stress-induced CK production showed up-regulation of brassinosteroid synthesis and signalling genes (Peleg *et al.*, 2011; Rivero *et al.*, 2010). Brassinosteroids (BRs) act synergistically with gibberellins (GAs), due to common components in their signalling pathways (Wang *et al.*, 2009b). Transgenic *Arabidopsis* seedlings constitutively overexpressing GA-responsive genes exhibited improved tolerance to various abiotic stresses; stress tolerance was accompanied by biosynthesis of salicylic acid (Alonso-Ramírez *et al.*, 2009), another plant hormone mainly implicated in stress responses.

Plants exposed to drought stress show an alteration of CK content. Hormonal analysis of wild-type (WT) maize leaves subjected to drought showed a gradual decline in CK and GA contents during stress (Wang *et al.*, 2008b). A comprehensive study on maize seedlings exposed to salt and osmotic stress also demonstrated rapid decline in some CK forms due to enforced CK catabolism. During acclimatization, however, accelerated CK metabolism led to a moderate increase in active CK forms (Vyroubalová *et al*., 2009). Higher accumulation of all CK forms was also determined in tobacco plants exposed to severe drought stress (Havlová *et al.*, 2008). Hence, accumulation of active CKs among other processes might contribute to the mechanisms by which plants overcome stress status and avoid growth inhibition. Regarding stress signalling, CKs do not, due to the slow response of their biosynthetic genes to stress induction, have a direct function similar to ABA, which directly affects stomatal closure (Vyroubalová *et al*., 2009).

Maintenance of high photosynthetic capacity is an important prerequisite for preserving crop yield under adverse environmental conditions. Although increasing CK content by senescence regulated expression of a CK biosynthetic gene is an efficient tool for prolonging leaf photosynthetic activity (Gan & Amasino, 1995), engineered wheat plants with senescence-regulated CK production showed no differences in yield-related

parameters (Sýkorová *et al.*, 2008). Shoot growth inhibition and promotion of root growth have been regarded as advantageous for crop stability under stressful conditions and constitute an integral part of plant stress tolerance (Sharp & LeNoble, 2002). Accordingly, plants with reduced shoot-to-root ratios as a consequence of CK deficiency showed greater tolerance to or avoidance of drought stress (Werner *et al.*, 2010; Macková *et al.*, 2013). Hence, down- and up-regulation of CK contents *in planta* can have a synergistically positive effect on enhanced tolerance to water deficit: in the first case, through alteration of plant morphology to a root architecture that is better adapted to withstand water deprivation, and, in the second case, through activation of photosynthetic processes and source–sink relations.

Recently, we prepared several barley transgenic lines overexpressing the *cytokinin dehydrogenase 1* (*CKX1*) gene from *Arabidopsis*, an enzyme of cytokinin catabolism, targeted to various subcellular compartments. Transgenic barley exhibited greater tolerance to or avoidance of drought stress that most probably was due to higher lignification and changes in root morphology (Pospíšilová *et al.*, 2016). While focusing primarily on post-stress revitalization, the in-depth transcriptomic analysis of the transgenic barley lines aimed to clarify and describe in detail all processes that enable CK-deficient barley plants to cope better with drought. In addition, this study provides comparative transcriptomic study between observed time points for WT plans, which were exposed to drought stress.

3.2. Material and methods

3.2.1. Plant material and cultivation

The spring barley cv. Golden Promise (WT) and transgenic lines expressing *AtCKX1* (NM_129714.3) under the control of the maize root-specific β-glucosidase (DQ333310.1) promoter were used in the study (Pospíšilová *et al.*, 2016). Transgenic and WT plants were grown in an environmental chamber with a photoperiod of 15°C/16 hours light and 12°C/8 hours darkness. The light source was a combination of mercury tungsten lamps and sodium lamps providing an intensity of 300 $\mu mol.m^{-2}.s^{-1}$. Plants were grown in a 2:1 mixture of soil and perlite (Perlit Ltd., Czech Republic). Soil composition was

1:1 professional substrate (peat type) for growing plants (Rašelina Soběslav, Czech Republic) and a muck-type arable soil from the Olomouc Region (Czech Republic).

3.2.2. Application of drought stress

For transcriptomic analysis of the root system exposed to drought stress, plants were grown hydroponically in a modified Hoagland solution (Vlamis & Williams, 1962) under control conditions as described above. The experiment was performed in the following arrangement: 2/3 of each vessel was filled with two transgenic genotypes and 1/3 with WT plants. In total, 27 plants from each genotype were cultivated together in three vessels. Three plants were pooled per biological replicate. The roots of "**non-stressed plants**" were harvested before stress induction. Drought stress was induced on 4-week-old plants by pouring the solution out of the growth vessel.

Plants were kept for 24 hours without solution. The roots of "**stressed plants**" were harvested before reapplying the nutritive solution. After 24h of stress, the vessel was filled-in back with the nutritive solution (Fig. 13).

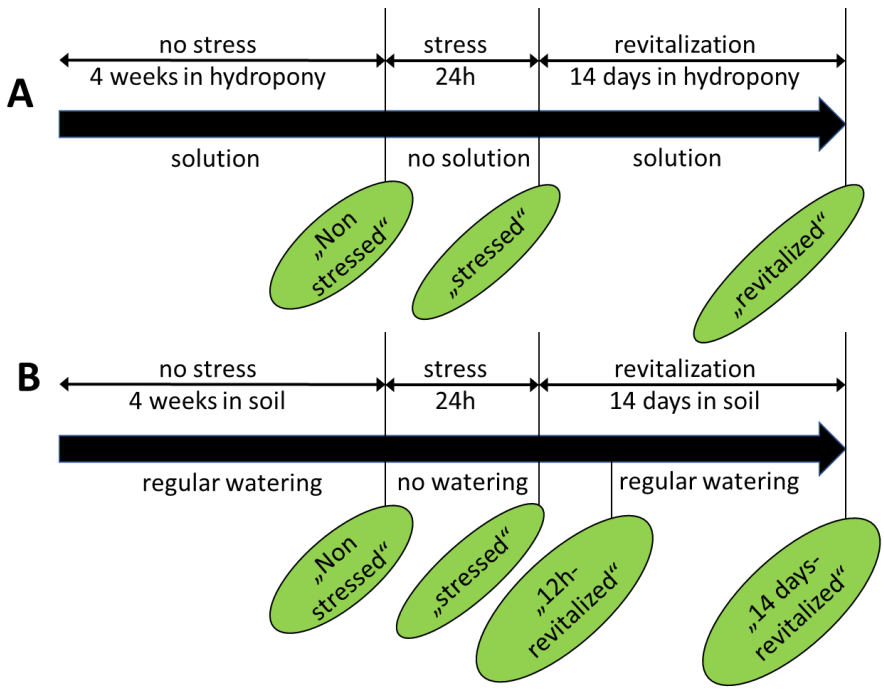


Figure 13: Experimental design used to study the root (A) and shoot (B) transcriptome of barley plant grown under drought stress.

Plants were further grown in non-stress conditions for 14 days. At the end of the experiment, the roots of "**revitalized plants**" were harvested (Fig.13).

The transcriptome analysis of the upper, vegetative part of barley was done on plants cultivated in shallow trays (30 cm x 20 cm x 5 cm) filled with soil and daily watered. Drought stress was applied to 4-week-old plants by cessation of watering for 4 days; thereafter watering was resumed daily. Samples were collected 12 hours after the last watering, on the 4th day of the stress application, at 12 hours after re-watering, and after 14 days of revitalization (Fig.13).

3.2.3. Isolation of RNA and RNA-seq analysis

Isolation of total RNA was performed using an RNAqueous Kit (Life Technologies, USA). Isolated RNA was then treated using a TURBO DNA-free Kit (Life Technologies) and purified using magnetic beads (Agencourt RNA-CLEAN XP, Beckman Coulter, USA).

Working with 2.5 μg of total RNA from each sample, extracted as described above, Illumina® TruSeq® Stranded mRNA Sample Preparation Kit (Illumina, USA) was used for cDNA library preparation. Library concentration was assessed using a Kapa Library Quantification Kit (Kapa Biosystems, USA) and all libraries were pooled to a final 8 pM concentration for cluster generation and sequencing. The clusters were generated using an Illumina® TruSeq® SR Cluster Kit v3 cBot HS and sequenced on a HiSeq SR Flow Cell v3 with a HiSeq 2500 Sequencing System. Two independent libraries were prepared for each genotype at each time point from two biological replicates (3 pooled plants in each).

The reads generated by sequencing were mapped to the reference genome of *Hordeum vulgare* v.25 (Cunningham *et al.*, 2015) using the TopHat2 v.2.0.12 splice-read mapper (Kim *et al.*, 2013) with default parameters. The reads mapped to the transcripts annotated in the reference genome were quantified by using HTSeq v.0.6.0 (Anders *et al.*, 2015) with respect to the stranded library. The tests for differential gene expression were performed using the DESeq2 package (Love *et al.*, 2014) implemented in R (R Development Core Team, 2008). Technical replicates were merged into one technical replicate to obtain higher coverage of the reference transcriptome.

GO annotation of the reference genome was improved using the Blast2GO (v.3.0) program (Conesa *et al.*, 2005), "nt" database (NCBI Resource Coordinators, 2014), the ncbi-blast+ (v.2.2.28) program (Camacho *et al.*, 2009), as well as the InterPro (Finn *et al.*, 2017) and PGSB (Spannagl *et al.*, 2016) databases. Gene description from the National Center for Biotechnology Information database (NCBI; NCBI Resource Coordinators, 2014) were mined using the BLAST module from program Blast2GO with parameters blastn and e-value $\leq 1.10^{-5}$. The annotation from PGSB and UniProtKB database was extracted with cut-off e-value $\leq 1.10^{-5}$.

3.3. Results and discussion

CK-mediated tolerance to drought stress can be acquired through two approaches. The first is based on increasing endogenous CK, promoting consequently plant acclimatization and survival rate and minimizes yield losses. This was realized by stress-inducible overexpression of the adenylate isopentenyl transferase, the step-limiting enzyme of the CK biosynthesis (Rivero *et al.*, 2007; Zhang *et al.*, 2010; Ghanem *et al.*, 2011; Merewitz *et al.*, 2011; Peleg *et al.*, 2011). The second approach tends to decrease endogenous CK content in roots by accumulation of CK-degradation enzymes, thus resulting in modified root morphology or enhanced root biomass (Werner *et al.*, 2001; Macková *et al.*, 2013). Degradation of CKs is catalysed by a family of cytokinin dehydrogenases (CKX; Galuszka *et al.*, 2001). CKX isoenzymes vary in their localization to different subcellular compartments (Schmülling *et al.*, 2003; Werner *et al.*, 2003) or in the temporal pattern of their expression (Mrízová *et al.*, 2013). Substrate specificity and localization of individual CKXs are best characterized in *Arabidopsis* which contains two vacuolar CKXs (*AtCKX1* and AtCKX3), one cytosolic (*A*tCKX7), and four probably apoplastic CKXs (Werner *et al.*, 2003). Secreted (apoplastic) CKX forms prefer free CK bases as substrate, whereas vacuolar CKXs prefer CK nucleotides, the primary products of the *de novo* biosynthesis (Kowalska *et al.*, 2010). Because the various CKX forms have different biological functions, a combination of desired enzyme activity, signal sequence, and a suitable tissue- or organ-specific promoter can selectively affect the plant phenotype.

Table 2: Characteristic of the libraries generated by RNA-seq. Read were aligned on the barley reference genome v.26 (Cunningham *et al.*, 2015).

	Sample (library)	Number of reads	Number of aligned reads	Mapping rate
Aerial part (plant grown in soil)	WT before stress 1	84 517 106	78 995 347	93.5%
	WT before stress 2	63 270 045	59 274 446	93.7%
	vAtCKX1 before stress 1	75 279 326	70 471 879	93.6%
	vAtCKX1 before stress 2	71 354 832	66 706 602	93.5%
	WT stressed 1	77 801 469	70 486 367	90.6%
	WT stressed 2	46 104 470	43 271 563	93.9%
	vAtCKX1 stressed 1	71 985 223	68 031 123	94.5%
	vAtCKX1 stressed 2	73 507 253	69 533 881	94.6%
	WT 12 hours revitalization 1	51 142 352	48 101 143	94.1%
	WT 12 hours revitalization 2	65 959 107	62 277 842	94.4%
	vAtCKX1 12 hours revitalization 1	69 436 337	64 832 308	93.4%
	vAtCKX1 12 hours revitalization 2	71 082 657	66 986 453	94.2%
	WT 14 days revitalization 1	66 446 892	62 470 573	94.0%
	WT 14 days revitalization 2	68 521 784	64 326 131	93.9%
	vAtCKX1 14 days revitalization1	59 187 683	54 355 201	91.8%
	vAtCKX1 14 days revitalization 2	71 287 568	66 620 894	93.5%
Roots (plant grown hydroponically)	WT stressed 1	45 007 428	37 581 752	83.5%
	WT stressed 2	200 564 800	168 454 522	84.0%
	vAtCKX1 stressed 1	152 220 463	121 895 816	80.1%
	vAtCKX1 stressed 2	78 643 428	65 854 656	83.7%
	cAtCKX1 stressed 1	75 960 243	64 038 286	84.3%
	cAtCKX1 stressed 2	63 499 681	52 087 213	82.0%
	WT 14 days revitalization 1	138 937 849	104 864 042	75.5%
	WT 14 days revitalization 2	90 088 238	65 540 634	72.8%
	vAtCKX1 14 days revitalization 1	64 437 018	48 016 522	74.5%
	vAtCKX1 14 days revitalization 2	104 760 752	77 622 260	74.1%
	cAtCKX1 14 days revitalization 1	84 791 919	66 888 893	78.9%
	cAtCKX1 14 days revitalization 2	124 010 156	99 545 337	80.3%

In the present study, in order to identify the genes regulated by CKs in response to drought stress, two lines of *AtCKX1*-overexpressing transgenic barley were selected for detailed transcriptomic analysis: one targeting the CKX1 protein to the vacuoles (*vAtCKX1*), whereas for the second line, *AtCKX1* was engineered to have a predicted cytosolic localization (*cAtCKX1*). In addition, non-transgenic WT plants were also analysed at different time points.

To study differences in transcriptomes, 28 single-read libraries were prepared and sequenced on an Illumina Hi-Seq 2500 system. The basic features of all sequenced libraries are summarized in the Table 2.

3.3.1. Improvement of the functional annotation of barley transcriptome

In order to improve the raw reference genome available at the time of the study, the functional annotation of the predicted genes was determined by Blast2GO version 3.0 program (Conesa *et al.*, 2005). Using this strategy, out of the 26 072 predicted genes in *Hordeum vulgare* genome, we obtained a functional annotation for 17 885 genes. The GO analysis assigned a total number of 70 719 GO terms to up to 20 991 predicted genes.

Out of these 40.87%, 42.12% and 17.01% were related to BP, MF and CC categories, respectively (Fig. 14). In BP category (on level 2), the major subcategories of GO terms were “metabolic process” (10 036 terms), “cellular process” (9 205 terms) and “single-organism process” (7 047 terms). Under MF, the top three categories (on level 2) were “binding” (11 328 terms), “catalytic activity” (8 594 terms) and “transporter activity” (937 terms). With respect to CC, “cell” (6 266 terms), “organelle” (4 725 terms) and “membrane” (3 268 terms) were the dominant groups. The number of GO terms assigned to one predicted sequence varied from 1 to 35.

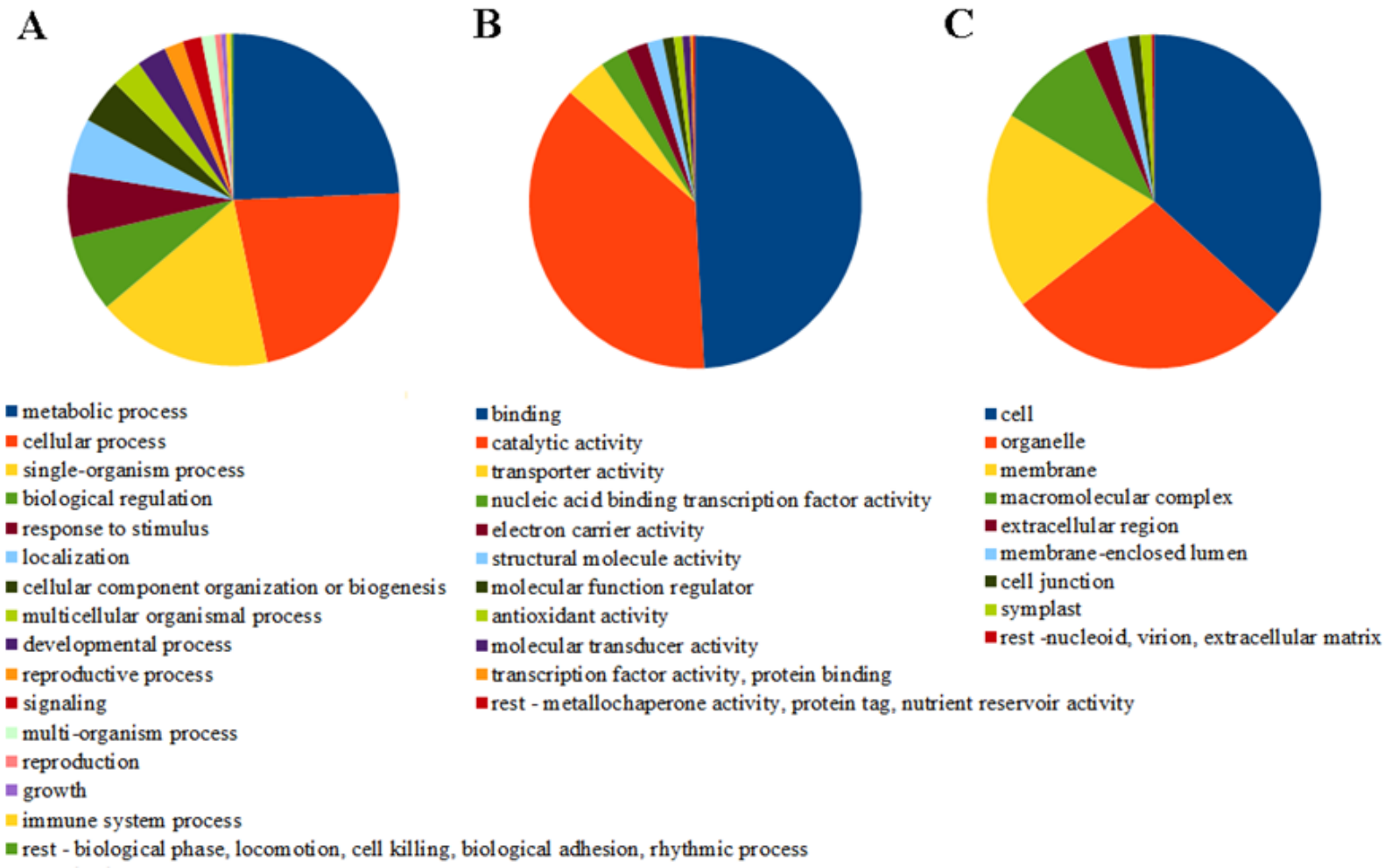


Figure 14: Gene ontology annotation of the whole barley transcriptome (*Hordeum vulgare* v.25; Cunningham *et al.*, 2015). (A) Biological processes; **(B)**, Molecular function; **(C)** Cellular component.

3.3.2. Effect of cytokinin deficiency on the aerial part of *vAtCKX1* plants under optimal conditions

The mild expression of *AtCKX1* under the control of β-glucosidase promoter had a positive effect on root system development whereas the aerial part was not substantially affected. The height of *vAtCKX1* plants was slightly reduced (Fig.15a, d), while *cAtCKX1* plants exhibited no visible changes in their aerial part during the first 4 weeks of development.

Differential expression examination revealed that approximately 400 genes were significantly affected in the 6-week-old aerial part in contrast to more than 2 400 genes affected in the roots of hydroponically cultivated *vAtCKX1* plants (Pospíšilová *et al.*, 2016).

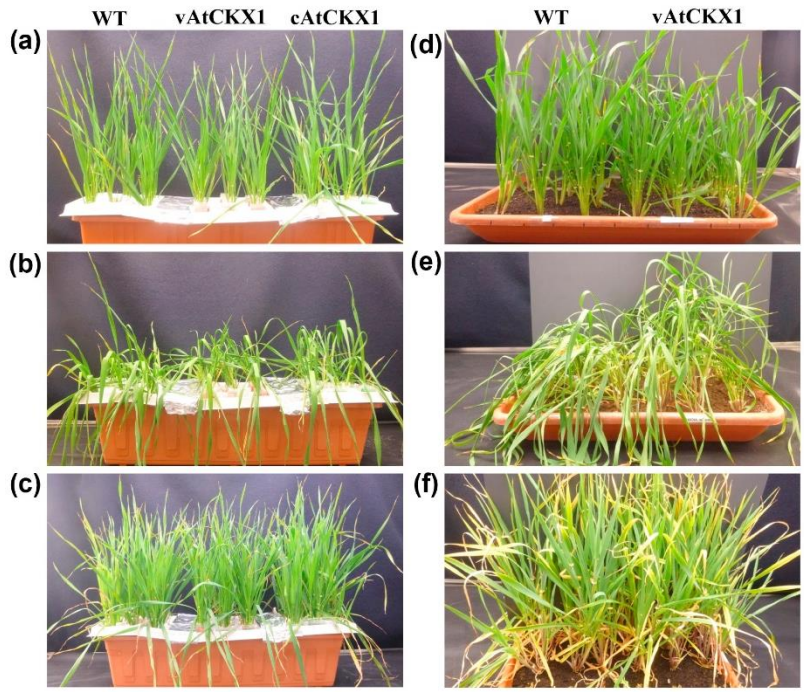


Figure 15: Photographs of transgenic (*vAtCKX1*, *cAtCKX1*) and WT barley plants cultivated in hydroponic system (left) or in shallow soil (right). (a, d), optimally watered 4-week-old plants; **(b, e)**, plants suffering from severe drought stress; **(c, f)**, regenerated plants 2 weeks after the application of drought stress.

In order to understand those mechanisms only regulated by the altered hormonal status during leaf development, we performed an in-depth transcriptomic analysis of *vAtCKX1* plants of the same age but cultivated in the soil. In contrast to hydroponically cultivated plants, approximately four times more genes were found to be affected by altered CK content.

Of the total 26 067 annotated genes, 988 and 609 genes were significantly down- and up-regulated, respectively, in the leaves of the *vAtCKX1* line grown in normal conditions compared to WT (adjusted *p*-value ≤ 0.01; Supplemental table 1). GO terms (level 6) of the most significantly affected genes in the leaves of plants grown both in hydroponic culture and in soil were compared. The 15 most common affected GO terms are summarized in the table 3.

The four most negatively affected (down-regulated) processes in leaves of *vAtCKX1* plants were linked to photosynthesis (ferredoxin-NAD(P) reductase activity, thylakoid membrane organization, establishment of plastid localization and phenylalanine ammonia-lyase activity), indicating that the photosynthetic apparatus and photosynthesis are most probably affected in transgenic plants. The effect was much more pronounced in plants cultivated in the shallow soil.

Three of four putative genes coding for prephenate/arogenate dehydratase, an enzyme participating in the final steps of the aromatic amino acid pathway that produces tyrosine and phenylalanine (Rippert & Matringe, 2002), were up-regulated in the leaves of *vAtCKX1* plants. Phenylalanine is the primary substrate for the phenylpropanoid pathway that gives rise to lignin, flavonoids, and anthocyanins. Accordingly, the most up-regulated GO terms were GO:0009963 (Positive regulation of flavonoid biosynthetic process) and GO:0009718 (Anthocyanin-containing compound biosynthetic process). The up-regulation of genes involved in the synthesis of phenylalanine suggests that the production of phenylalanine might be stimulated in the transgenic *vAtCKX1* line and might serve as a pool for the synthesis of flavonoids and anthocyanins in the leaves, where they contribute to protection mechanisms against various stresses. However, four of eight genes encoding phenylalanine ammonia lyases, whose activity is considered a key switch between the phenylpropanoid pathway and primary aromatic amino acid metabolism (Cass *et al.*, 2015), were significantly down-regulated.

Table 3: The most affected gene ontology (GO) terms in the upper part of *vAtCKX1* plants cultivated hydroponically or in soil compared to wildtype plants. Percentages are shown of differentially expressed genes (adjusted *p*-value ≤ 0.05) at GO level 6 and higher from total number of genes with the same GO number. Genes affected in both culture conditions are in bold. Genes in several GO terms are not listed because the term parsed to several other child terms.

GO number	*Category	GO term	Total #	% of affected genes		Accession of affected genes in format MLOC_XXXXX
				Hydrop.	Soil	
UP-REGULATED						
GO:0004664	MF	prephenate dehydratase activity	4	50.0	25.0	23316, 65725, 56414
GO:0008131	MF	primary amine oxidase activity	6	16.7	50.0	**70980**, 4986, 17390
GO:0016165	MF	linoleate 13S-lipoxygenase activity	16	31.3	18.8	**64972**, 54031, **55029**, 71275, 51884, 69572
GO:0005544	MF	calcium-dependent phospholipid binding	10	20.0	20.0	54932, 40592, 55134, 15770
GO:0004834	MF	tryptophan synthase activity	5	20.0	20.0	59863, 61188
GO:0009963	BP	positive regulation of flavonoid biosynthetic process	9	11.1	22.2	81070, 54366, 19988
GO:0004034	MF	aldose 1-epimerase activity	7	14.3	14.3	**5638**
GO:0005337	MF	nucleoside transmembrane transporter activity	7	14.3	14.3	**55464**
GO:0009718	BP	anthocyanin-containing compound biosynthetic process	11	18.2	9.1	61512, 65788, 64248
GO:0009407	BP	toxin catabolic process	27	7.4	18.5	**17760**, 73593, **68101**, 57709, 72489
GO:0047262	MF	polygalacturonate 4-α-galacturonosyltransferase activity	8	12.5	12.5	11661, 57229
GO:0031418	MF	L-ascorbic acid binding	13	15.4	7.7	**77814**, 64248
GO:0004806	MF	triglyceride lipase activity	41	9.8	12.2	**17298, 18031**, **80878**, **80586**, 58940
GO:0006569	BP	tryptophan catabolic process	14	14.3	7.1	12847, 57323, 69262
GO:0015996	BP	chlorophyll catabolic process	33	3.3	15.2	80455, 34851, 55009, 21175, **64277**
DOWN-REGULATED						
GO:0008937	MF	ferredoxin-NAD(P) reductase activity	4	25.0	50.0	7761, 53537, 40355
GO:0010027	BP	thylakoid membrane organization	106	2.8	61.3	**58382**; not listed

Table 3: The most affected gene ontology (GO) terms in the upper part of *vAtCKX1* plants cultivated hydroponically or in soil compared to wildtype plants. (*continued*)

GO:0051667	BP	establishment of plastid localization	63	4.8	55.6	Not listed
GO:0045548	MF	phenylalanine ammonia-lyase activity	7	28.6	28.6	79728, 62322, 477, 4684
GO:0009658	BP	chloroplast organization	137	2.2	54.0	Not listed
GO:0019682	BP	glyceraldehyde-3-phosphate metabolic process	202	1.5	51.5	Not listed
GO:0006720	BP	isoprenoid metabolic process	237	1.3	46.0	Not listed
GO:0010310	BP	regulation of hydrogen peroxide metabolic process	15	6.7	26.7	33774, 1518, 15501, 65632, 1340
GO:0030855	BP	epithelial cell differentiation	7	14.3	14.3	38181, 54366
GO:0016054	BP	organic acid catabolic process	153	3.9	14.4	Not listed
GO:0008544	BP	epidermis development	9	11.1	11.1	38181,54366
GO:0009699	BP	phenylpropanoid biosynthetic process	22	9.1	9.1	4684, 477, 57736, 79728
GO:0070726	BP	cell wall assembly	11	9.1	9.1	52864, 67760
GO:0032870	BP	cellular response to hormone stimulus	205	8.8	5.7	Not listed
GO:0007166	BP	cell surface receptor signaling pathway	30	3.3	10.0	63541, 17680, 44275, 72162

*MF: molecular function, BP: biological process

Nevertheless, all four genes, MLOC_4684, MLOC_62322, MLOC_79728 and MLOC_477, were found to be up-regulated (by 3.6-, 3.0-, 2.7-, and 1.9-fold, respectively) in the roots of *vAtCKX1* plants (Pospíšilová *et al.*, 2016). Hence, and as much as the aromatic amino acid metabolism seemed not to be affected in *vAtCKX1* transgenic roots, surplus phenylalanine might be translocated from leaves to roots, where it can supply lignin production and deposition.

The third most enriched process in *vAtCKX1* leaves was linked to the activity of lipoxygenases, which enzymes participate in the release of volatile compounds, including jasmonates (JAs), from intracellular lipids (Feussner & Wasternack, 2002). These compounds are usually released during plant defence against various pathogens. As the result is based on two independent experiments in which two biological replicates were sequenced and compared to the respective WT plants, it is not very likely that the observed lipoxygenase activation was merely a response to an undetected biotic stressor.

In addition to plant defences, JAs participate in several developmental processes such as trichome formation and leaf senescence (Wasternack, 2014).

Interestingly, JA-dependent formation of trichomes is accompanied by the production of secondary compounds such as flavonoids, anthocyanins, and terpenoids (Wasternack, 2014; Tian *et al.*, 2012). Although there is not enough evidence to indicate cross-talk between JAs and CKs, it is predicted that their interaction might be antagonistic inasmuch as JAs strongly inhibit the CK-induced callus growth (Ueda & Kato, 1982). Nevertheless, the interplay of both phytohormone groups probably depends not only on the CK:JA ratio but also on other hormones (O'Brien & Benková, 2013). In total, 12 of 55 and 8 of 39 genes categorized as GO:0009753 'Response to jasmonic acid' and GO:0010026 'Trichome differentiation,' respectively, were found to be significantly upregulated in the roots of two independent *vAtCKX1* lines (Pospíšilová *et al.*, 2016). Hence, predicted JA production in the upper part of transgenic plants might affect mainly roots and their fine architecture. Volatile methyl JA can be translocated as a rapid chemical signal from shoot to root and function there as a gene expression inducer (Baldwin *et al.*, 1994). Nevertheless, enforced JA production directly in roots is also feasible inasmuch as GO:0016165 'Linoleate 13S-lipoxygenase activity' is among enriched terms in transgenic roots during revitalization (see below).

3.3.3. Whole transcriptome response of *vAtCKX1* plants during revitalization after drought stress

In addition to optimal conditions, three other kinds of sequencing libraries were generated from the upper part of *vAtCKX1* and WT plants cultivated in the shallow soil: the first from plants exposed for 4 days to drought (Fig.15e), the second from material 12 hours after re-watering, and the third from leaves having undergone revitalization for 14 days (Fig. 15f). Unsurprisingly, only five genes were significantly altered (adjusted *p*-value ≤ 0.05) between stressed transgenic and WT leaves, thereby indicating that transcriptomes of both genotypes were strongly affected by the water deficit (Supplemental table 1). Twelve hours after re-watering, 10 and 9 genes were significantly up- and down-regulated, respectively, between *vAtCKX1* and WT leaves (Supplemental table 1). Additionally, 5 of these 19 genes were not altered between the libraries made from

optimally cultivated versus stressed leaves of either transgenic or WT leaves. Two putative F-box-like proteins (MLOC_75620, MLOC_43997), which have been shown to play an essential role in multiple phytohormone-signalling pathways (Moon *et al.*, 2004), and one receptor-like protein kinase (MLOC_17138) were detected among them. The protein MLOC_17138 contains in addition to the kinase domain, a pfam01657 domain associated with a role in salt stress response and antifungal activity.

Transgenic plants overexpressing *AtCKX1* exhibit better growth parameters (e.g., biomass production and yield) when encountering drought stress (Pospíšilová *et al.*, 2016). To understand processes attributed to the beneficial growth of *vAtCKX1* plants, a comparative transcriptomic analysis was carried out examining transgenic versus WT leaves 2 weeks after revitalization from stress. Of the total 26 067 barley genes, 301 and 31 genes were significantly up- and down-regulated, respectively, in revitalized *vAtCKX1* leaves in contrast to WT (Supplemental table 1). The enriched GO terms in up-regulated genes of *vAtCKX1* are summarized in the table 4.

Products of many genes up-regulated by *vAtCKX1* participate as structural proteins or enzymes of the photosynthetic apparatus, which indicated influence of photosynthetic parameters. Interestingly, the most activated genes comprised those encoded by the barley chloroplast genome (indicated by the prefix EPlHVUG). In total, 14 of 112 translatable chloroplast genes were 2- to 3-fold up-regulated with high significance (adjusted *p*-value ≤ 0.05; Supplemental table 1). Chloroplasts are a known target of CK action. Indeed, exogenously applied CK is able directly to activate the expression of several chloroplast-encoded genes in detached barley leaves which accumulated also the stress hormone ABA (Zubo *et al.*, 2008). Because it is not yet clear whether CK acts directly on chloroplast transcription, we can only speculate that the increase in chloroplast transcripts observed in revitalized *vAtCKX1* transgenic plants relays an accumulation of CK in leaves upon water stress. Our hypothesis is supported by the strong activation of endogenous *IPT* genes in *vAtCKX1* leaves at several developmental time points as a consequence of CK depletion (Pospíšilová *et al.*, 2016). Hence, increased local maxima of CKs, produced by IPT activity localized in chloroplasts, might trigger similar machinery as was described in CK-treated detached leaves to activate the chloroplast genome.

Table 4: The most enriched GO terms in up-regulated genes (adjusted *p*-value ≤ 0.05) in the aerial part of *vAtCKX1* plants collected two-weeks after re-watering. Percentage of differentially expressed genes at GO level 6 and higher from total number of genes with the same GO number is shown.

GO number	*Category	GO term	Total #	% of affected genes	Accession of affected genes in format MLOC_XXXXX or EPlHVUG000000XXXX
GO:0016984	MF	ribulose-bisphosphate carboxylase activity	5	40.0	EPlHVUG00000010074, MLOC_21811
GO:0009765	BP	photosynthesis, light harvesting	32	34.4	Not listed
GO:0030076	CC	light-harvesting complex	8	25.0	EPlHVUG00000010021, MLOC_57061
GO:0009718	BP	anthocyanin-containing compound biosynthetic process	10	20.0	MLOC_5324, 19814
GO:0016165	MF	linoleate 13S-lipoxygenase activity	16	18.8	MLOC_37378, 51884, 71948
GO:0045259	CC	proton-transporting ATP synthase complex	27	14.8	EPlHVUG00000010007, 10016, 10047, MLOC_26730
GO:0009767	BP	photosynthetic electron transport chain	56	14.3	EPlHVUG00000010010, 10072, 10065, 10026, 10021, MLOC_52515, 22512, 39436
GO:0046271	BP	phenylpropanoid catabolic process	20	10.0	MLOC_15203, 61189
GO:0052716	MF	hydroquinone: oxygen oxidoreductase activity	20	10.0	MLOC_15203, 61189
GO:0009579	CC	thylakoid	331	9.9	Not listed
GO:0016597	MF	amino acid binding	31	9.7	MLOC_62844, 19879, 80634
GO:0004499	MF	N, N-dimethylaniline monooxygenase activity	22	9.1	MLOC_11897, 11896
GO:0009637	BP	response to blue light	48	8.3	MLOC_43394, 22512, 11312, 52515
GO:0055082	BP	cellular chemical homeostasis	47	6.4	MLOC_22808, 65878, 69460
GO:0034754	BP	cellular hormone metabolic process	38	5.3	MLOC_6666, 73942

* MF: molecular function, BP: biological process, CC: sub-cellular component.

The analysis of endogenous CK content in chloroplasts under these conditions would provide support to our hypothesis. It is noteworthy that none of the chloroplast-encoded genes were down-regulated in *vAtCKX1* plants cultivated under optimal conditions when as many nucleus-encoded genes participating in photosynthesis were down-regulated compared to those in WT plants (Tab. 3).

Among other interesting genes significantly up-regulated in revitalized *vAtCKX1* leaves were these encoding four putative aquaporins (MLOC_56278, MLOC_71237, MLOC_552, MLOC_22808), which are channel proteins facilitating the transport of water through plasma and intracellular membranes. The increased expression of several genes encoding barley aquaporins had already been observed in plants exposed to salinity stress (Wei *et al.*, 2007). It has been hypothesized that an increase in water channel activity would facilitate maintenance or recovery of growth during or after the stress period.

Two genes classified under the GO term 'phenylpropanoid catabolic process' encode putative laccases – aromatic compound:oxygen oxidoreductase (Tab. 4), which might participate in lignin degradation or its polymerization.

Furthermore, genes significantly affected between non-stressed and revitalized leaves were evaluated separately for *vAtCKX1* and WT plants (Tab. 5). Those from the transgenic plants that did not overlap with WT plants were further compared with genes differentially regulated between the two genotypes (i.e., the 301 up- and 31 down-regulated genes). Only seven up-regulated genes remained as being not developmentally dependent but genotype dependent. None of the down-regulated genes meet both criteria (Tab. 5).

Table 5: List of genes significantly up-regulated in *vAtCKX1* leaves 14 days after re-watering (adjusted *p*-value ≤ 0.05). Genes considered were not developmentally dependent but also significantly up-regulated between revitalized and non-stress leaves of *vAtCKX1* genotype but not in WT (adjusted *p*-value ≤ 0.001).

Gene number	Gene annotation	Mean expression (*R2W)	Fold change	
			vAtCKX1 (*R2W) vs WT (*R2W)	*vAtCKX1* (*R2W) vs *vAtCKX1* (†NS)
MLOC_8529	Nematode-resistance protein	1 471	4.28	2.21
MLOC_14310	GDSL esterase/lipase	1 163	2.51	2.62
MLOC_74636	tolB protein (WD40-like Beta Propeller)	185	2.38	3.60
MLOC_30661	Putative isoflavone 2'-hydroxylase	78	2.31	6.66
MLOC_70609	unknown protein located in chloroplast stroma	104	2.30	4.30
MLOC_74367	Peroxiredoxin (Thioredoxin-like fold)	2 443	2.23	2.66
MLOC_80571	Chalcone isomerase	598	2.19	2.86

*R2W: 2-week-long revitalization; †NS: non-stressed.

Two genes encoding putative enzymes of the flavonoid biosynthesis pathway (chalcone isomerase and isoflavone 2'-hydroxylase) were found to be up-regulated in *vAtCKX1* plants. These two genes combine with two other genes found to be up-regulated in revitalized *vAtCKX1* leaves and participating in the regulation of flavonoid metabolism (chalcone isomerase: MLOC_5324; zinc-finger (B-box) protein: MLOC_19814), indicating an enhanced production of isoflavonoids and anthocyanins. Recently, an unambiguous positive effect of flavonoid and anthocyanin production in improving tolerance to drought stress has been shown (Nakabayashi *et al.*, 2014). Due to their antioxidative activity, the over-accumulation of flavonoids mitigates the negative effect of reactive oxygen species released under stress conditions.

Peroxiredoxin (MLOC_74367) belongs to a family of cysteine-dependent peroxidases which also participate in detoxification of plant cells by scavenging reactive oxygen species (Cha *et al.*, 2015). An orthologue of peroxiredoxin has been found among another 25 over-accumulated proteins in wheat seedlings of a cultivar that is drought-stress tolerant in comparison to a drought-sensitive one (Cheng *et al.*, 2015). MLOC_14310 belongs to a large family of GDSL-type esterase/lipases with hydrolytic activity toward triacylglycerols. Members of this family are involved in plant development, morphogenesis, secondary metabolite synthesis, and defence responses, and some members are activated by JAs. The closest rice orthologue of MLOC_14310 (LOC_Os01g46080) was found to be activated by desiccation stress in rice leaves (Jiang *et al.*, 2012). Moreover, pepper GDSL lipase caused higher susceptibility to pathogens but increased tolerance to osmotic stress when overexpressed in *Arabidopsis* (Hong *et al.*, 2008).

Taken together, the differential expression study in *vAtCKX1* and WT leaves before and after the stress period reveal several genes whose increased expression initiated by the CK imbalance may lead to better drought tolerance and/or faster growth after rewatering.

3.3.4. Response of *vAtCKX1* and *cAtCKX1* roots during stress and revitalization

Due to the impossibility to collect root tissues from soil-grown plants without causing mechanical stress, transcriptome of the root system was studied from plants grown hydroponically. Twelve RNA-seq libraries were generated from *vAtCKX1*, *cAtCKX1*, and WT roots collected at two time points: during the severe drought stress (Fig. 15b) and 2 weeks after revitalization (Fig.15c). Similarly, as in the aerial part, stress induced a strong response at the transcriptome level (Supplemental table 2). Between transgenic plants and WT plants, only a few genes were deregulated during the stress (Supplemental table 2). Just seven genes were significantly up-regulated in both *vAtCKX1* and *cAtCKX1* genotypes, including, for example, putative nicotianamine synthase (MLOC_71596) and 4-coumarate CoA ligase (MLOC_18901) involved in lignification. Fifty-seven genes were significantly down-regulated. The most strongly down-regulated gene in both lines was a putative F-box-like protein (MLOC_75620; 12.6- and 5.3-fold), which was found also among the most strongly down-regulated genes in the early and late phases of leaf revitalization.

The unambiguous and strong depletion of MLOC_75620 transcripts in all transgenic samples indicates that this F-box protein might play a crucial role in regulating responses in *CKX*-overexpressing plants via cross-talk with other hormones. F-box proteins represent one of the largest superfamilies in plants, that is involved in the process of ubiquitination and protein degradation. To date, only a limited number of F-box proteins have been functionally characterized. Most of them are involved in regulating hormone signalling pathways, where they degrade repressors or activators of auxin, GA, ethylene, and JA response (Moon *et al.*, 2004).

The analysis of DEGs in revitalized 6-week-old roots (Supplemental table 2) revealed that a gene encoding one of the cytokinin receptors (*HvHK3*; MLOC_44452) was significantly down-regulated in both transgenic genotypes. The gene was also down-regulated in 6-week-old *vAtCKX1* and *cAtCKX1* roots cultivated under optimal conditions (Pospíšilová *et al.*, 2016), thus leading to the conclusion that cessation of CK perception via HvHK3 is a developmental consequence rather than a response to stress. The

overexpression of the *Arabidopsis CKX1* gene into the barley genome led to a hormonal imbalance that is counterbalanced by a timely and finely regulation of endogenous CKX and IPTs (Pospíšilová *et al.*, 2016). Plant *IPT* genes are generally very weakly expressed, and the enzyme's activity is regulated by farnesylation (Galichet *et al.*, 2008). Significant up-regulation of two abundant endogenous CKX enzymes (HvCKX4 and HvCKX5) was observed in both independent experiments with 6-week-old *AtCKX1*-overexpressing plants, while other *HvCKXs* were down-regulated or unchanged in comparison to WT plants. Although it is difficult to estimate the final CK homeostasis, one might expect a local minimum in CK content that leads to cessation in *HvHK3* transcription and CK perception through this receptor.

The altered homeostasis of CKs in *AtCKX1*-overexpressing barley, which influences root system architecture, might anticipate changes in auxin levels, transport, and perception. Several groups of auxin response genes exist in plant genomes that react sensitively to auxin imbalance. Auxin early response genes are divided into two categories: *Aux/IAA* and *SAUR* (small auxin up RNA); all of them regulate plant physiology by modulating the interaction with auxin response elements of other transcription factors, such as the major group of auxin response factors (ARFs). Auxin action is influenced by its polar transport that is mediated by auxin efflux carriers (*PINs*). Its homeostasis is also regulated by *GH3* (*Gretchen Hagen 3*) genes encoding mainly auxin-amino acid synthetases that form the inactive auxin-amino acid conjugates (McSteen, 2010). The dataset comparing *AtCKX1* vs. WT indicated that transgenic roots contain elevated levels of auxin, inasmuch as 3 of 12 predicted and expressed *PIN* transporters and 3 of 7 *GH3* genes were significantly up-regulated in *cAtCKX1* roots (Supplemental table 3). Analysis of *vAtCKX1* roots showed a similar but slightly weaker tendency.

Many of the 48 significantly down-regulated genes in the *Arabidopsis ahk2/ahk3* double knock-out with higher stress tolerance are auxin early responsive genes (*SAUR* and *Aux/IAA*). The dwarf phenotype of *ahk2/ahk3* plants was attributed to the observed down-regulation of auxin response (Tran *et al.*, 2007). Because *vAtCKX1* and *cAtCKX1* plants were not substantially affected in their aerial part but had positively altered root system morphology, a different way of auxin response might be expected. In contrast to

ARF and *GH3* genes, several *SAUR* and *Aux/IAA* genes exhibited lower expression in transgenic plants than in WT plants. Nevertheless, no straightforward comparison can be made between the auxin response of *Arabidopsis ahk2/ahk3* mutant and *AtCKX1*-overexpressing barley (Supplemental table 3 and 4).

3.3.5. Whole transcriptome analysis of wild-type barley plants during stress and revitalization.

In order to determine the molecular functions that can be affected in barley plants grown under drought, we evaluated genes whose expression is changed in roots and in aerial part of barley during the stress and/or revitalization. For this purpose, differential expression was determined using the DESeq2 package implemented in R (R Development Core Team, 2008), and the GO annotation was obtained with Blast2GO (Conesa *et al.*, 2005). The level six was considered as the most meaningful for our analysis.

Transcriptome of barley roots under drought stress was conducted on 4-week-old plants grown in hydroponic culture and deprived of nutritive solution for 24h. We identified a total of 11 158 DEGs, of which 5 721 and 5 437 genes were significantly up- and down- regulated, respectively (adjusted p-value ≤ 0.05; Supplemental table 5). Genes that accumulated in the roots of barley plants undergoing severe drought stress belonged to GO term categories (Supplemental table 6) related to response to stress such as: "cellular response to water stimulus", "hyperosmotic salinity response", "heat acclimation", "ion homeostasis", "toxin catabolic process", "cellular lipid catabolic process", "response to steroid hormone", "response to hydrogen peroxide" or "phenylpropanoid biosynthetic process" (Tab. 6). The importance of the phenylpropanoid biosynthetic pathway has been already discussed, suggesting that the synthesis of phenylalanine, the precursor of lignin, flavonoids, and anthocyanins might be stimulated in the roots of barley plants exposed to severe drought. All three compounds have been characterized to play a role in response to stress. A detailed analysis of the accumulation of these three metabolites could have support our conclusions. Genes related to the GO terms "toxin catabolic process", "heat acclimation" and "response to hydrogen peroxide" encompass genes encoding proteins such as glutathione transferase (MLOC_68101), glutathione reductase (MLOC_32914; MLOC_61193) and several heat shock proteins,

which were reported as drought-responsive proteins or proteins involved in reactive oxygen species production (Wang *et al.*, 2016).

Table 6: List of the GO terms related to "Biological Processes" (level 6) the most affected in the roots of barley plants grown under stress conditions (adjusted *p*-value ≤ 0.05). Four-week-old barley plants were subjected for 24h to severe drought stress (removal of the nutritive solution from the vessel).

GO number	GO term	*Total #	†% of affected genes
	UP-REGULATED		
GO:0071462	cellular response to water stimulus	11	63.64%
GO:0009407	toxin catabolic process	33	60.61%
GO:0072348	sulphur compound transport	10	60.00%
GO:1902644	tertiary alcohol metabolic process	22	59.09%
GO:0044242	cellular lipid catabolic process	97	57.73%
GO:0033015	tetrapyrrole catabolic process	35	57.14%
GO:0042538	hyperosmotic salinity response	20	55.00%
GO:0010286	heat acclimation	26	53.85%
GO:0046164	alcohol catabolic process	10	50.00%
GO:0046434	organophosphate catabolic process	12	50.00%
GO:0048545	response to steroid hormone	20	50.00%
GO:0050801	ion homeostasis	80	48.75%
GO:0055082	cellular chemical homeostasis	52	48.08%
GO:0042542	response to hydrogen peroxide	62	46.77%
GO:0009699	phenylpropanoid biosynthetic process	28	46.43%
	DOWN-REGULATED		
GO:0010089	xylem development	13	69.23%
GO:0071103	DNA conformation change	104	59.62%
GO:0070726	cell wall assembly	11	54.55%
GO:0048544	recognition of pollen	91	50.55%
GO:0006915	apoptotic process	296	50.34%
GO:0051129	negative regulation of cellular component organization	10	50.00%
GO:0001666	response to hypoxia	16	50.00%
GO:0009664	plant-type cell wall organization	76	48.68%
GO:0046271	phenylpropanoid catabolic process	21	47.62%
GO:0007166	cell surface receptor signalling pathway	42	47.62%
GO:0009834	plant-type secondary cell wall	19	47.37%
GO:0015851	Biogenesis nucleobase transport	13	46.15%
GO:0006002	fructose 6-phosphate metabolic process	11	45.45%
GO:0042886	amide transport	65	44.62%
GO:0000910	cytokinesis	106	43.40%

* Total #: number of genes in the GO category showing significantly changes in expression.

† % of affected genes: % of genes in the GO category with altered expression.

Interestingly, severe drought induced the accumulation of genes belonging to the GO term “response to steroid hormone”, suggesting that BRs could participate in the establishment of the response to drought in barley roots. BRs are involved in a huge array of physiological responses including, but not limited to, growth, seed germination, rhizogenesis, senescence, and resistance to plants against various abiotic stresses (Vardhini & Anjum, 2015). How BRs can confer tolerance or resistance to drought is a complex process. Indeed, whereas the overexpression of *BRL3*, a vascular-enriched member of the BR receptor family, confer drought stress tolerance in *Arabidopsis* without affecting the overall plant growth, loss-of-function mutations in the ubiquitously expressed *BRI1* receptor leads to drought resistance (Fàbregas *et al.*, 2018). BRs regulate various components of the antioxidant defence system (Vardhini & Anjum, 2015). Moreover, recent advances in phytohormone research suggest that BRs effect of might rely on its interplay with other hormones, notably abscisic acid that is the probably the most important hormones regulating responses to stress (Ahammed *et al*., 2015).

Genes whose expression was down-regulated in the roots of barley plants grown under severe stress belong to GO terms categories (Supplemental table 6) such as " xylem development”, “cell wall assembly”, “apoptotic process”, “plant-type cell wall organization”, “plant-type secondary cell wall” and “cytokinesis” (Tab. 6). This was probably related to the inhibition of root growth during stress as a protection against exposure to stress conditions (Janiak *et al.*, 2018).

The analysis of the transcriptome of the aerial parts revealed that a total of 2 628 and 2 400 genes were differentially up- and down-regulated, respectively, upon stress (adjusted p-value ≤ 0.05; Supplemental table 7). Genes that were the most up-regulated in the leaves of barley plants exposed to severe drought were notably annotated as belonging to GO terms “hyperosmotic salinity response” and “response to water deprivation” (Tab. 7; Supplemental table 8). Up-regulation of genes from these categories is in direct reaction to the drought stress (Sinha *et al.*, 2017). The most significant down-regulated process in barley leaves during severe drought stress was “photosynthesis, light harvesting”, which is in direct relation with water deficiency (Janiak *et al.*, 2018).

In order to get insights concerning molecular mechanisms involved in re-vitalization, we further determined the transcriptome of leaves of stressed barley plants 12 h after re-

watering. Among the 6 580 genes differentially regulated genes (Supplemental table 7; adjusted *p*-value ≤ 0.05), 3 690 genes were up-regulated, and 2 890 genes were down-regulated compared to the non-stressed plants.

Table 7: List of the GO terms related to "Biological Processes" (level 6) the most affected in the aerial part of barley plants grown under stress conditions (adjusted *p*-value ≤ 0.05). Four-week-old barley plants grown in the soil were subjected for 4 days to severe drought stress (no watering).

GO number	GO term	*Total #	†% of affected genes
	UP-REGULATED		
GO:0042538	hyperosmotic salinity response	20	50.00%
GO:0009962	regulation of flavonoid biosynthetic process	11	36.36%
GO:0010647	positive regulation of cell communication	15	33.33%
GO:0006026	aminoglycan catabolic process	18	33.33%
GO:0046348	amino sugar catabolic process	18	33.33%
GO:1901071	glucosamine-containing compound metabolic process	18	33.33%
GO:0060548	negative regulation of cell death	29	31.03%
GO:0010583	response to cyclopentenone	14	28.57%
GO:0046271	phenylpropanoid catabolic process	21	28.57%
GO:0033015	tetrapyrrole catabolic process	35	28.57%
GO:0009414	response to water deprivation	68	27.94%
GO:1902644	tertiary alcohol metabolic process	22	27.27%
GO:0043067	regulation of programmed cell death	63	25.40%
GO:0006662	glycerol ether metabolic process	32	25.00%
GO:0009407	toxin catabolic process	33	24.24%
	DOWN-REGULATED		
GO:0009765	photosynthesis, light harvesting	32	87.50%
GO:0019750	chloroplast localization	67	71.64%
GO:0051667	establishment of plastid localization	67	71.64%
GO:0009668	plastid membrane organization	123	70.73%
GO:0009658	chloroplast organization	146	67.12%
GO:0016226	iron-sulfur cluster assembly	70	62.86%
GO:0019682	glyceraldehyde-3-phosphate	216	62.04%
GO:0051156	glucose 6-phosphate	121	61.16%
GO:0033014	tetrapyrrole biosynthetic process	102	59.80%
GO:0042727	flavin-containing compound biosynthetic process	12	58.33%
GO:0010374	stomatal complex development	69	53.62%
GO:0009767	photosynthetic electron transport chain	57	50.88%
GO:0006720	isoprenoid metabolic process	255	49.02%
GO:0006778	porphyrin-containing compound metabolic process	138	47.83%
GO:0016143	S-glycoside metabolic process	60	46.67%

* Total #: number of genes in the GO category showing significantly changes in expression.

† % of affected genes: % of genes in the GO category with altered expression.

The up-regulated genes were involved in affected GO terms (Tab. 8; Supplemental table 8) according to Blast2GO analysis. In contrast to the stress conditions few new GO categories were observed, mainly “ribosomal large subunit biogenesis” (14 involved genes), “karyogamy” (25 genes) and “pyrimidine-contain compound biosynthetic process” (97 genes involved). Those changes indicated that the plants accelerating the specific processes of synthesis and cell division as a part of the revitalization process. The GO categories, which contained down-regulated genes (Table 8; Supplemental table 8) consistently showed down-regulation of “photosynthesis, light harvesting”. This process is probably reactivated after longer time period.

Table 8: List of the GO terms related to "Biological Processes" (level 6) the most affected in the aerial part of barley plants 12h after re-watering (adjusted *p*-value ≤ 0.05). Four-week-old barley plants grown in the soil were subjected for 4 days to severe drought stress (no watering), then re-watered to normal condition. Transcriptome was analyzed 12h after the re-watering.

GO number	GO term	[*]Total #	[†]% of affected genes
	DOWN-REGULATED		
GO:0009765	photosynthesis, light harvesting	32	81.25%
GO:0071462	cellular response to water stimulus	11	63.64%
GO:0051156	glucose 6-phosphate metabolic process	121	53.72%
GO:0009767	photosynthetic electron transport chain	57	52.63%
GO:0019750	chloroplast localization	67	50.75%
GO:0051667	establishment of plastid localization	67	50.75%
GO:0072525	pyridine-containing compound biosynthetic process	20	50.00%
GO:0009637	response to blue light	49	48.98%
GO:0019682	glyceraldehyde-3-phosphate metabolic process	216	46.30%
GO:0010109	regulation of photosynthesis	13	46.15%
GO:0043085	positive regulation of catalytic activity	79	45.57%
GO:0016143	S-glycoside metabolic process	60	45.00%
GO:0006778	porphyrin-containing compound metabolic process	138	44.93%
GO:0009668	plastid membrane organization	123	44.72%
GO:0033014	tetrapyrrole biosynthetic process	102	43.14%
	UP-REGULATED		
GO:0042273	ribosomal large subunit biogenesis	14	71.43%
GO:0000741	karyogamy	25	64.00%
GO:0072528	pyrimidine-containing compound biosynthetic process	97	59.79%
GO:0000085	mitotic G2 phase	23	56.52%
GO:0043572	plastid fission	11	54.55%
GO:0006518	peptide metabolic process	498	49.20%
GO:0043604	amide biosynthetic process	512	48.44%
GO:0007292	female gamete generation	69	46.38%
GO:0007006	mitochondrial membrane organization	11	45.45%
GO:0051604	protein maturation	47	44.68%
GO:0006026	aminoglycan catabolic process	18	44.44%
GO:0046348	amino sugar catabolic process	18	44.44%
GO:1901071	glucosamine-containing compound metabolic process	18	44.44%
GO:0051169	nuclear transport	125	44.00%
GO:0009553	embryo sac development	110	43.64%

[*] Total #: number of genes in the GO category showing significantly changes in expression.
[†] % of affected genes: % of genes in the GO category with altered expression.

CHAPTER 4.0

SATRANS: A TOOL DESIGN FOR FAST FUNCTIONAL ANNOTATION OF RNA-SEQ DATA SETS

Kokáš, F.Z., Bergougnoux, V., Čudejková, M.M. (2019). SATrans: New free available software for annotation of transcriptome and functional analysis of differentially expressed genes. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 26*(2), 117-123. https://doi.org/10.1089/cmb.2018.0149

4.1. Introduction

Massive parallel sequencing, such as of RNA, opens up great possibilities for transcriptomic studies to measure gene expression changes within an entire transcriptome despite having no previous knowledge of the sequences (Wang *et al.*, 2009a). Rapid technological advances have made this technique very cost effective and widely used. Thus, the amount of RNA-seq data in the NCBI Short Read Archive database increases every day (NCBI Resource Coordinators, 2014). Together with this increasing RNA-seq data, the number of sequences still to be characterized also grows. Particularly if the genome of a given organism of interest is not already sequenced and annotated, thousands of *de novo* reconstructed transcripts with unknown function are produced (Geniza & Jaiswal, 2017). Although several tools promising efficient functional analysis of these nucleotide sequences already have been developed, rapid and effective functional characterization of such a large number of sequences remains a challenging task. For example, TRAPID (Van Bel *et al.*, 2013) and Transcriptator (Tripathi *et al.*, 2015) are quite comprehensive and freely available transcriptome analysis tools. Both tools are dedicated to functionally annotate transcript sequences by sequence similarity search, which is performed by RAPSearch2 (Zhao *et al.*, 2012) or BLAST (Altschul *et al.*, 1990), respectively, in the selected protein databases. Both tools provide GO (Ashburner *et al.*, 2000) term assignment, TRAPID moreover performs a phylogenetic analysis of gene families and Transcriptator identifies non-coding RNAs (Van Bel *et al.*, 2013; Tripathi *et al.*, 2015). Both TRAPID and Transcriptator are web interface-based applications, which makes them user-friendly but also gives rise to certain limitations. For example, the number of sequences that can be analyzed in a single project is limited and there is a "wait in a queue" for computing capacity. Mercator, another web interface-based application suited for processing of large scale of data sets, –is a pipeline specifically designed to functionally annotate plant "omics" data (Lohse *et al.*, 2014). This pipeline computes functional annotations of protein or nucleotide input sequences using the MapMan BIN ontology. It combines BLAST-based and protein domain-based search to compute "BIN" assignments, the system of functional categories different from GO (Lohse *et al.*, 2014). Among the desktop application is a very popular Blast2GO (Conesa *et al.*, 2005). The freeware BASIC version of this software performs functional annotation of nucleic acid

or protein sequences using BLAST and InterProScan search (Jones *et al.*, 2014), and the current version of the charged PRO version integrates the analysis of differential expression of count data arising from RNA-seq technology.

We present here SATrans (Supplementary material 1), a freeware desktop application providing ORF (Open Reading Frame) prediction and sequence similarity BLAST-based search not only in the protein databases but also in the nucleotide databases. Furthermore, it offers the possibility to choose the type of BLAST search to be performed (blastn, blastp, blastx, tblastn, tblastx). Obviously, each annotation tool is helping to understand biological meaning of transcriptomic data, however, each is designed for a certain purpose as well as SATrans. The goal of SATrans is not to replace any of the tools mentioned above, but to increase the effectivity of the annotation process and to provide a functional analysis of differential gene expression data generated by RNA-seq, helping to functionally and biologically interpret such large datasets. Up to now, SATrans is the only freely available software for functional annotation of nucleotide/amino acid sequences, which allows the analysis of differential gene expression data with respect to the biological function, process and cellular localization based on the GO annotation.

In the following sections, we will explain the main functionalities of the software and provide a typical example use case to illustrate the applicability of SATrans in transcriptomic studies.

4.2. Methods

4.2.1 Software description and functionality

SATrans is written in Perl (Wall *et al.*, 2000), as the main programming language, and MySQL (Widenius & Axmark, 2002), as the language for communicating between the software and the database for enduring storage of the data. The software is primarily designed for GNU/Linux, and the current version is designed for a command line. The software is divided into three Perl modules: blast.pm, ipr_scan.pm, and datab.pm. It operates through eight modes: create, update, repair, analysis, delete, show, export, and import. The module blast.pm provides the BLAST function to search for homologous sequences in the database of interest and the communication between the user's PC and

the remote NCBI server providing the public databases (NCBI Resource Coordinators, 2014). The module also allows a local BLAST against a local FASTA-formatted database. This, however, requires the installation of the BLAST standalone (Camacho *et al.*, 2009) and is high memory consuming. The Ipr_scan.pm module consists of the functions used for the InterProScan search (Jones *et al.*, 2014). InterProScan allows scanning of the input sequences for matches against the InterPro protein signature databases. It is executed by HTTP protocol and service on the website iprscan5. InterPro integrates 11 protein family databases: HAMAP (Pedruzzi *et al.*, 2015), PANTHER (Mi *et al.*, 2016), Pfam (Finn *et al.*, 2016), PRINTS (Attwood *et al.*, 2012), ProDom (Bru *et al.*, 2005), PROSITE Patterns and PROSITE Profiles (Sigrist *et al.*, 2013), SMART (Letunic *et al.*, 2015), TIGRFAMs (Haft *et al.*, 2013), SUPERFAMILY (Oates *et al.*, 2015), and CATCH-Gene3D (Lam *et al.*, 2016). It is one of the main sources of GO annotation (Sangrador-Vegas *et al.*, 2016). Before InterProScan is launched, the best ORF in the nucleotide sequence is determined. The algorithm for the best ORF search depends on ORF length. The longest ORF is selected as the best, translated to a protein, then used by the InterProScan, which searches all the InterPro databases (Finn *et al.*, 2017) to collect such information as GO terms or, conserved domains. Functions stored in the datab.pm module enable communication between the SATrans software and the MySQL database. These include transport of the BLAST and InterProScan results into the database, and other database-associated operations such as “update,” “select” and “delete.” The module also contains functions for export of the results (“export”), creation (“create”), and quality control (“show”) of the MySQL database. Furthermore, the module contains functions for the analysis (“analysis”) of DESeq2 data (Love *et al.*, 2014), which are provided as an input by the user.

The software is designed to perform the following tasks: (1) functional annotation of nucleotide or amino acid sequences, and (2) GO enrichment analysis of DEGs. A simplified scheme of the analytical process is shown in Figure 16.

(1) The process of functional annotation is carried out in the “create” mode and requires providing as an input a multi-FASTA file containing sequences to be analyzed (for example, an entire transcriptome). For this purpose, SATrans uses two external services, BLAST (NCBI Resource Coordinators, 2014) and

InterProScan (Finn *et al.*, 2017). The two can be run in two parallel procedures. Both services (BLAST and InterProScan) produce XML output, which is stored in a MySQL database and can be used as an input for further analysis. The results of the functional annotation can be exported using the "export" mode. Output files are provided separately for the BLAST and InterProScan results as well as for the GO annotation.

(2) The functional annotation stored in the MySQL database can be paired with information about the differential gene expression coming from a separate analysis performed by the DESeq2 program (Love *et al.*, 2014). The DESeq2 results in csv format serve as the input file, and the analysis is carried out in "analysis" mode. At this step, the basic statistics about the significantly DEGs are calculated with respect to their GO terms. For example, the ratio is provided for affected/all genes belonging to the selected GO term. Moreover, the significance of the differential expression of each GO term is calculated using Fisher's exact test (Fisher, 1935; for details see the SATrans Manual). Other analyses provided are the mean number as well as the minimum and maximum values of log2FoldChange (depending on whether genes are

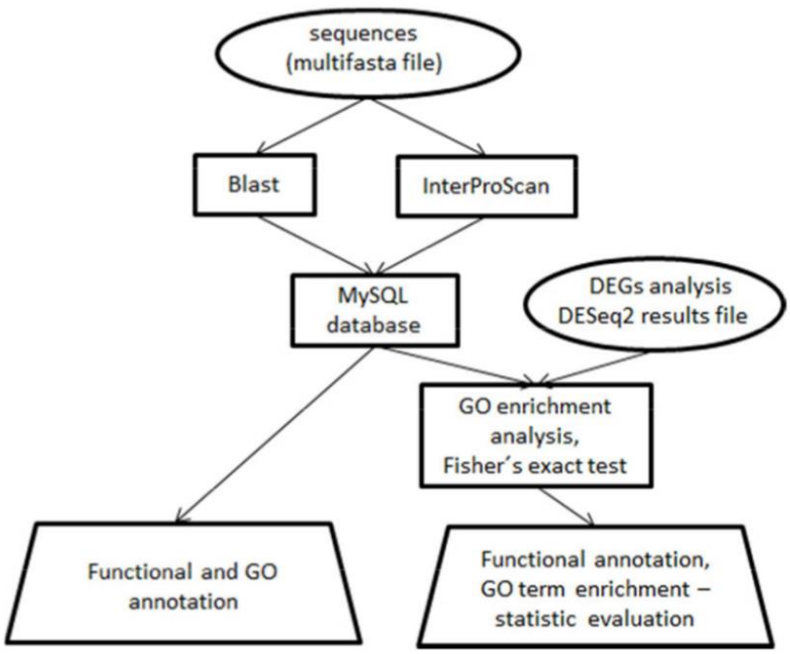


Figure 16: Simplified scheme of SATrans annotation and analysis process. Ellipses, input files; rectangles, data storage and analysis processes; trapezium, output files; DEGs, differentially expressed genes; GO, Gene Ontology.

upregulated or downregulated) for the selected GO term. The results of the analysis are summarized in the main output file, which is the GO analysis output file. This mode generates another two output files: Histogram and Annotation file. These provide a user-defined (by cutoff values) overview of the analyzed data, consisting of log2FoldChange (log2FC) values (originally calculated by DESeq2) for each GO term or functional annotation (results of BLAST and InterProScan) for each gene from the input DESeq2 file.

(3) The process of functional annotation is carried out in the "create" mode and requires providing as an input a multi-FASTA file containing sequences to be analyzed (for example, an entire transcriptome). For this purpose, SATrans uses two external services, BLAST (Altschul *et al.*, 1990) and InterProScan (Jones *et al.*, 2014). The two can be run in two parallel procedures. Both services (BLAST and InterProScan) produce XML output, which is stored in a MySQL database and can be used as an input for further analysis. The results of the functional annotation can be exported using the "export" mode. Output files are provided separately for the BLAST and InterProScan results as well as for the GO annotation.

4.2.2 Error handling

SATrans handles any server errors occurring during the sequence search by stopping the process for ten seconds and then retrying with the next query sequence. A new search for the query sequence with error will be repeated after all query sequences are searched by BLAST and InterProScan. Maximal count of the repeated searches is five. More than five errors per ten searches cause a further 60 seconds pause before trying a retrieval operation again. If the number of failures is more than 60 per 100 sequences the program is stopped permanently. All errors are reported in the error.log file and the terminal window. Stopping the program by the user at any time will not affect results which have been already stored in the database. Stopped analysis can be resumed by the "repair" mode using the input fasta file.

4.3. Results and discussion

4.3.1 Case Study

To demonstrate the SATrans function and the usage in the real biological studies we have chosen the study made by Čudejková *et al*. (2016). In this study, the authors performed qualitative and quantitative transcriptome analysis of several strains of ergot fungus *Claviceps purpurea* (*Cp*). For the purpose of the current case study, we have chosen to functionally re-analyze results of the differential gene expression between the mycelial and sclerotial tissues of *Cp* strain 20.1, summarized in the Supplemental table 9. The whole *Cp* transcriptome (protein coding sequences) was downloaded from the http://fungi.ensembl.org/Claviceps_purpurea_20_1/Info/Index (Cunningham *et al.*, 2015) and annotated using the "create" mode with default parameters. Annotation results are summarized in the stat.log file and the detailed results (Supplementary material 2) have been exported as text files using the "export" mode. These results were subsequently used for the functional analysis of DEGs. The input csv file for the "analysis" mode (Supplemental table 10) have been provided to the software and the options log2fold change, cutoff value, *p*-adjusted cutoff value and format of output file remained default. The detailed results of "analysis" mode have been saved as the text files (Supplementary material 3) and the summary of the results is provided again in the stat.log file. *Claviceps* genome contains 8 825 predicted genes. SATrans functionally annotated 7 595 sequences, whereas 1 230 sequences remained unannotated. Among the SATrans annotated sequences, 3 930 have NCBI "nt" database BLAST annotation, 7 433 have InterProScan annotation, and 4 625 have GO annotation comprising of 119 885 GO terms. From the total number 629 DEGs, 548 DEGs were functionally annotated by SATrans, of these 257 have NCBI "nt" database BLAST annotation, 537 have InterProScan annotation, and 311 have GO annotation, whereas 81 DEGs remained unannotated. When we compare the SATrans annotation with the original annotation of DEGs published in Čudejková *et al*. (2016) (407 functionally annotated DEGs), we obtained a clear improvement of the annotation by 22%. The GO analysis of the DEGs, provided in the output GO analysis files, revealed that regulation of 95 GO terms was significantly changed between the examined tissues. When analyzed separately downregulated and upregulated genes, 69

and 72 GO terms were significantly down- and upregulated, respectively. Among the downregulated processes were mostly processes associated with membrane transport, especially ion transport. Whereas, upregulated processes were principally associated with secondary metabolism – toxin biosynthesis. These results are in great compliance with the metabolic changes coupled with sclerotial differentiation (Čudejková *et al.*, 2016).

4.3.2 Comparison of SATrans with Blast2GO, TRAPID and MERCATOR

We have compared the features of different annotation tools: two desktop tools (SATrans and Blast2GO) and two web-based tools (TRAPID and MERCATOR), which are summarized in Tab. 9.

In the case of Blast2GO, only the BASIC freeware version was used, since other tools are also freeware. The comparison of the features shows that SATrans is the only freeware annotation tool providing DEGs GO analysis integration. Comparing to web-based tools, it provides sequence similarity search in any database of interest, not only in the default ones, what is very likely the reason why it provides the better functional annotation results (Tab. 10).

Table 9: Feature comparison of the transcript analysis freeware tools.

Features	*SATrans	*Blast2GO – BASIC	*TRAPID	MERCATOR
Sequence similarity search	NCBI BLAST, local BLAST	NCBI BLAST, local BLAST	RAPSeqrch2	BLAST, RPS-Blast
†ORF finding	yes	no	yes	yes
Reference database	any database of interest	any database of interest	PLAZA 2.5	TAIR 10, TIGR5 rice proteins
Functional annotation	GO, InterProScan	GO, InterProScan, Enzyme codes, KEGG	GO, InterProScan/PFAM	InterProScan
‡DEGs GO analysis integration	yes	no	no	no
Availability	freeware - desktop application	freeware - desktop application	freeware - web tool	freeware - web tool

* GO: gene ontology; † ORF: open reading frame; ‡ DEGs: differentially expressed gene

Table 10: Comparison of the results of the annotation of 5 000 sequences performed by different annotation tools. The dataset – 5 000 sequences were randomly selected from the barley transcriptome stored in the Ensembl database (Cunningham *et al.*, 2015).

Search	SATrans	Blast2GO - BASIC	TRAPID	MERCATOR
Sequence similarity	4 531	4 531	4 349	3 886
***GO annotation**	2 359	2 359	2 128	-
InterProScan	3 953	3 953	3 687	2 880

* GO: gene ontology

We conducted a series of benchmarks to assess both runtime and number of annotated genes for selected annotation tools. As a representative dataset, we used 5 000 transcripts randomly selected from the barley transcriptome stored in the Ensembl database. The tools were launched with default parameters and the results are summarized in Table 10 and Table 11. Comparing the quality of the annotation (number of annotated sequences), SATrans provided the same results as Blast2GO-BASIC, which were obviously better than TRAPID or MERCATOR (Tab. 10).

When analysing runtime, SATrans needed much longer time to annotate the same number of sequences comparing to web-based tools TRAPID and MERCATOR (Tab. 11). However, when comparing to the Blast2GO-BASIC, SATrans performed its job much faster than Blast2GO-BASIC, annotating 5 000 sequences in 2 131 minutes comparing to 10 890 minutes required by Blast2GO-BASIC. The greater efficiency of SATrans compared to Blast2GO is very likely caused by parallel running of BLAST and InterProScan and better control of deadlock, SATrans not allowing sequence similarity search running over 15 minutes.

Table 11: Comparison of computational time for different annotation tools. Time is measured in minutes. Dataset represents randomly selected transcripts from barley transcriptome stored in the Ensembl database.

Dataset (number of sequences)	SATrans	Blast2GO - BASIC	TRAPID	MERCATOR
50	27	102	5	6
500	227	621	10	13
5 000	2 131	10 890	35	79

Considering the runtime and the results quality performance, SATrans seems to be the best tool to annotate a large number of the sequences; however, in contrast to other analyzed tools it requires the basic knowledge of a Linux operating system and does not provide a graphical user interface at the current version.

4.3.3 Future development of software

Architecture of the SATrans software enables an easy application of upgrades or even adding the new modules. Future releases of the software might provide the greater use of data from InterProScan for analysis of the impact of DEGs in the frame of whole transcriptome and adding a graphical user interface together with on-line interactive help system. Possibility to work with future versions of Perl and MySQL database system is also expected.

CHAPTER 5.0

TRANSCRIPTOMIC ANALYSIS OF DIFFERENT WHEAT INBRED LINES WITH DIFFERENT ROOT SYSTEM

5.1 Introduction

The domestication of wheat (*Triticum* spp.) around 10 000 years ago marked a dramatic turn in the development and evolution of human civilization, as it enabled the transition from a hunter-gatherer and nomadic pastoral society to a more sedentary agrarian one (Gegas *et al.*, 2010). Wheat counts between the first crops that have been domesticated in the Fertile Crescent. With a production of 771 718 579 tonnes in 2017, it represents the third most important crop, grown on more than 218 million ha worldwide (FAOstat, 2017). Wheat provides one-fifth of the calories consumed in human diet. The common, present-day wheat cultivars fall into two groups: the tetraploid durum wheat, *T. durum* Desf. (2n = 28, BBAA) and the allohexaploid bread wheat, *T. aestivum* L. (2n = 42, AABBDD; Özkan *et al.*, 2011; Li *et al.*, 2013; Duan *et al*., 2012). In 2018, the sequencing of the bread wheat (Chinese Spring) genome revealed a genome of 14.5 Gbp with 97% of the sequences assigned and ordered along the 21 chromosomes of the 3 subgenomes (A, B and D). A total of 107 891 high-confidence gene models have been predicted and annotated in the wheat genome (IWGSC, 2018). Nevertheless, up to 85% of the wheat genome is represented by highly repetitive DNA (Clavijo *et al.*, 2017; IWGSC, 2018).

With the evergrowing worldwide population and increasing needs for food resources, the agriculture faces the challenge to increase the production of crops while the arable surface is decreasing, and the environmental conditions become more extremes. Drought is one of the important environmental disorders that affect plant development, and more particularly crop productivity. To respond to drought, plants have evolved different strategies: 1) drought avoidance that is the ability to maintain higher tissue water content and 2) drought tolerance that is the ability to endure low tissue water content through adaptive traits (Basu *et al.*, 2016). Avoidance and tolerance are often referred to as drought resistance (DR). Adaptive traits of DR include maintenance of cell turgor by osmotic adjustment, protoplasmic resistance, reduction of transpiration, and optimization of water uptake, through modification of rooting. Roots are an important organ ensuring not only anchorage, but also water and nutrient absorption, and are the site of interaction with soil microorganisms which regulate crop productivity (Hochholdinger *et al.*, 2004). Due to their hidden status, only minor attention has been given to roots and mechanisms

controlling their development. In cereals, the root system is mainly composed of post-embryonically developed roots that emerge from the crown, i.e. the junction between the roots and the shoot. It has been assumed that the root architecture can predict the ability of the plant to withstand stress. In this regard, it is considered that a shallow root system is adapted for acquiring nutrients present in the higher layer of soil, whereas a deep root system is required for foraging lower soil layers where the water accumulates (Li *et al.*, 2016; Zhao *et al.*, 2017; El Hassouni *et al.*, 2018).

The present study was initiated to characterize new bread wheat bred lines towards their ability to withstand drought stress. Four genotypes were selected among a large collection based on the architecture of their root system: two genotypes presented an overall long root system, whereas two others were characterized by short but abundant roots. To provide bases of genetic engineering through molecular farming, omics approaches, especially RNA-seq, are powerful tools for i) discovering mechanisms that control responses to biotic and abiotic stresses, and be later targeted by genome editing (Jia *et al.*, 2017; Duan *et al.*, 2012), ii) discovering new genes (Hahn *et al.*, 2009), iii) detecting alternative splice variants and getting insights concerning gene regulation (Anders *et al.*, 2012), or iv) detecting polymorphism (Van Bellghem *et al.*, 2012). RNA-seq has proven to be beneficial and cost-efficient for transcriptomics studies both in model and non-model plant species (Goyal *et al.*, 2016). Therefore, RNA-seq analysis was used to understand the molecular basis of the difference of the root architecture of the four wheat genotypes. For the current study, the reference genome was the genome of the Chinese spring bread wheat that was still incompletely annotated on the time of our research.

As already mentioned, a good reference genome/transcriptome is crucial for each RNA-seq analysis. When a reference genome is available and well-characterized, one can use the *ab initio* strategy, i.e. mapping short reads on the reference genome (Li *et al*., 2013). In opposite, when the reference genome is not available or partially annotated, one might consider other approaches, such as the *de novo* approach or the combined approach, taking advance of the already available reference genome and building new genes. There are tremendous challenges to *de novo* assembly because of handling the millions to billions of shorts reads generated during the sequencing (He *et al.*, 2015). Furthermore, it

is necessary to reconstruct full-length transcripts from these short reads which can be hardly reachable in the case of plants with large genome containing huge number of isoforms and/or a high proportion of repetitive sequences (Duan *et al*., 2012). The read length is crucial for the *de novo* assembly. Therefore, the Roche 454 pyrosequencing, generating longer reads than the Illumina technology, was used for some non-model organisms; nevertheless, short reads produced by Illumina are much more economical (Duan *et al.,* 2012). Moreover, it has been described that *de novo* transcriptome assembly software's are very sensitive to sequencing errors and can hardly distinguish isoforms with high degree of homology (Li *et al*., 2013). Interestingly when considering more than one genotype, reads from various genotypes can be combined before assembly or after assembly, into potential transcripts (He *et al.,* 2015). Plethora of software's has been developed for *de novo* assembly of transcriptomes, such as Trinity (Haas *et al.,* 2013), Trans-ABySS (Simpson *et al.*, 2009), Oases (Schulz *et al*., 2012) or SOAPdenovo2 (Luo *et al.*, 2012). They were successfully applied to assemble transcriptomes in many organisms (Li *et al*., 2013); nevertheless, Trinity was evaluated as highly efficient in few transcriptome assembly studies (Li *et al*., 2013; Duan *et al.*, 2012).

For organisms with partially known or annotated reference genomes, a combined assembly strategy can be used to analyse transcriptome data (Zhao *et al*., 2011). Reads are firstly mapped to the known reference; the non-mapped reads are extracted from the raw data and used in *de novo* assembly strategy to obtain potential genes that are not present or represented in the reference genome (Duan *et al.*, 2012). Previous study indicates, that a combination of *ab initio* and *de novo* strategy would be another suitable approach for the reconstruction of wheat reference (Li *et al*., 2013).

Therefore, in this study, we systematically compared the performance of *ab initio, de novo* and combined strategies in assembling transcriptomes of the four wheat genotypes. To reconstruct an accurate and nearly complete reference, several factors affecting read assembly were monitored (N50, max length of transcripts etc.). Based on our results, we provided guidelines for the selection of optimal assembly strategy. This study also provided differential expression analysis and GO analysis between groups of genotypes with different root system.

5.2 Material and methods

5.2.1 Plant materials

Grains of four bread wheat inbred lines (W501, W509, W527 and W533) were obtained from RAGT Czech s.r.o. (for commercial reason, the pedigree of the different lines cannot be provided). These lines were produced in the frame of a selection program dealing with efficient use of inputs and high resistance to stresses. The 4 lines could be separated into two groups: i) the group I (W501 and W509) presents a deep root system composed of a limited amount of post-embryonically developed adventitious roots, and ii) the group II (W527 and W533) is characterized by a shallow root system composed by numerous short adventitious roots (Fig. 17).

Plants were grown in hydroponic culture with half-strength Hoagland solution in culture chamber with a long-photoperiod (16h-day/8h-night) and a temperature regime of 21°C-day/18°C-night. For RNA-seq analysis, the full root system of 7-week-old plants were considered. Each genotype was analysed into three independent biological replicates.

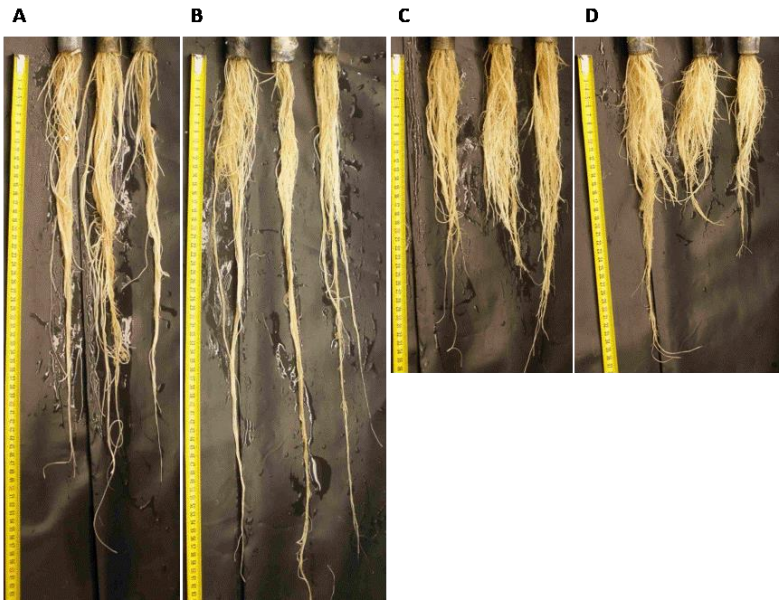


Figure 17: Photographs showing the root system of the four bread wheat lines used in the study. The genotypes W501 **(A)** and W509 **(B)** can be considered as long-type root, whereas W527 **(C)** and W533 **(D)** are short-type roots with apparent high number of adventitious roots. Plants were grown for 7 weeks in a ½ Hoagland solution in culture chamber under control conditions (16h-light/8h-dark; 21°C-day/18°C-night).

5.2.2 RNA extraction, library construction and Illumina sequencing

For each line, roots were collected from plant grown in hydropony for seven weeks after germination and immediately frozen in liquid nitrogen. Samples were ground in liquid nitrogen using a mortar and pestle. RNAqueous Total RNA Isolation Kit (Ambion, Life Technologies) was used for total RNA extraction according to the manual and additional DNase treatment was done with 2U TURBO DNase (Ambion, Life Technologies) twice for 30 min at 37°C.

PE sequencing was considered for the purpose of the current student. For this purpose, libraries were prepared from 2.5 µg of total RNA according to instructions of the Illumina TruSeq1 Stranded mRNA Sample Preparation Kit (Illumina). Library concentration was assessed with the Kapa Library Quantification Kit (Kapa Biosystem) and all libraries were pooled to the final 8 pM concentration for cluster generation and sequencing. The clusters were generated using an Illumina TruSeq1 PE Cluster Kit v3cBot HS and sequenced on HiSeq PE Flow Cell v3 with a HiSeq 2500 Sequencing System. Three independent biological replicates were sequenced for each sample; each biological sample represents a pool of the root system of 3 different plants grown in the same conditions.

5.2.3 RNA-seq data pre-processing and Quality Control

The simple quality control of paired end reads was performed with FASTQC v.0.11.5 (Andrews, 2010). Within this quality control, the following parameters were checked: per base sequence quality, per sequence quality scores, per base sequence content and k-mer content. RNA-seq data were based on the quality control pre-processed before aligning to the reference wheat genome or being used for *de novo* assembly. PE raw reads from all samples were trimmed using the FASTX-Toolkit version 0.0.14 (FASTX-Toolkit, 2015) without primer/adaptor filtering, with parameter “first base to keep” set on 12.

5.2.4 *Ab initio* approach

Genome sequence and gene annotations for wheat were downloaded from the EnsemblPlants release 40 (Cunningham *et al.*, 2015) as well as the existing gene annotation (GTF file), containing 110 790 annotated genes.

Short reads obtained during sequencing were aligned to the reference genome by TopHat2 v2.1.1 (Kim *et al.*, 2013), with default parameters. FeatureCounts v1.4.6 (Liao *et al.*, 2014) was used to quantify the number of reads aligned to the wheat reference genome. The command-line parameters were: "-F GTF -t exon -p -s 2 -T 10". Stringtie v1.3.1c, a fast and highly efficient assembler of RNA-seq alignments into potential transcripts, (Pertea *et al*., 2015) was used to improve the annotation of the wheat reference transcriptome; default parameters were considered. For this purpose, we used first the bam files generated after short read mapping onto the reference genome for each sample separately; second, the merged mode was used to obtain a consensus GTF file. Finally, the sequences listed in the consensus GTF file were used as reference to map reads and FeatureCounts v1.4.6 (Liao *et al.*, 2014) was applied again.

Finally, to evaluate the effectiveness of the *ab initio* approach, the following parameters were taken into account: number of unique reads, number of reads mapped on multiple sites, number of unassigned reads for ambiguity, and number of unassigned reads for presence in junk DNA.

The analysis was performed on a 64-bit Linux system (Ubuntu 12.04) with 503G physical memory and 24 CPUs.

5.2.5 *De novo* approach

De novo assembly and removing redundancy

To obtain the best possible *de novo* reference, two sub-strategies were employed: "Single Genotypes" (SG) and "All Merged Genotypes" (AMG). Trinity v2.0.6 (Haas *et al.*, 2013) was used to assemble individual transcriptome of the four genotypes (SG sub-strategy; Fig. 18), as well as to generate the "merged" transcriptome, containing the information of the four genotypes together (AMG sub-strategy; Fig. 18). The command-line parameters for Trinity v2.0.6 were: "--seq_type fq -max_memory 40G -left

Reads_R1.fastq -right Reads_R2.fastq -SS_lib_type RF -CPU 10 -min_contig_length 300 --normalize_reads -KMER_SIZE 32". For the SG sub-strategy, two fastq files (R1 and R2 corresponding to left and right sequencing, respectively) were generated for each genotype, representing a total of 8 fastq files. For AMG sub-strategy, two fastq files were obtained, R1 and R2 corresponding to left and right sequencing, respectively.

Subsequently, CD-HIT-EST v4.6.1 (Fu *et al.*, 2012) was used to remove the redundancy of the new assembled contigs. For this purpose, a transcript was marked as redundant when covered by other transcripts with an identity higher than 80%. Thus, only non-redundant transcripts were used to count the basic assembly statistics. Transcripts obtained from SG sub-strategy were collected and used as an input for the CD-HIT-EST to remove redundancy of collected transcripts (SGM sub-strategy). The detailed pipelines of AMG, SG and SGM sub-strategy is shown in diagram (Fig. 18).

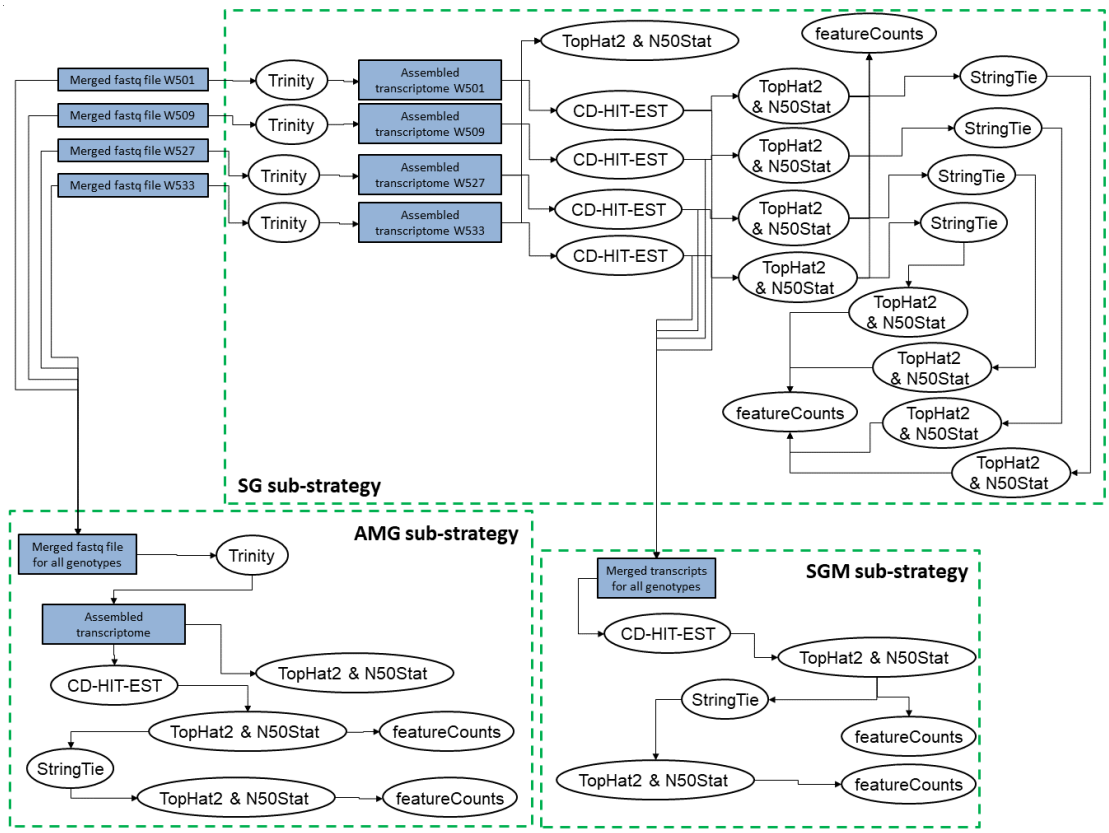


Figure 18: Diagram showing the different sub-strategies used for *de novo* assembly of the wheat reference transcriptome. The different software's used are indicated. The blue rectangle indicates the data file generated after use of a specific software. For this study, four genotypes were used (W501, W509, W527 and W533).

Basic assembly statistics were employed to evaluate the performance of the different assembly strategies. For this purpose, the N50 value, the number of contigs longer than 1 kbp, the average length of contigs, the total number of contigs, the total number of bases, the GC content and the maximum length of contig were used as criteria. These metrics were calculated by N50Stat.pl tool from NGSQCToolkit (Patel, 2012).

Further redundancy removal was performed by Stringtie v1.3.1c (Pertea *et al*., 2015) for all three sub-strategies, with default parameters and by approach which is listed before. Basic assembly statistics were re-examined as described previously.

Again, most of the analysis was performed on a 64-bit Linux system (Ubuntu 12.04) with 503G physical memory and 24 CPUs, except for CD-HIT-EST v4.6.1 analysis, which was calculated by MetaCentrum (METAVO, 2017).

Read mapping

TopHat2 version 2.1.1 (Kim *et al*., 2013) was used to get extended assembly statistics, i.e. the number of reads that could be mapped back to transcripts. Default parameters were set to map back all input short reads to the reconstructed transcripts. This procedure was performed for each *de novo* strategy following the reference assembly by Trinity (except of SGM sub-strategy), redundancy removal by CD-HIT-EST and subsequent data processing with StringTie.

Quantification of unique reads and comparative study

FeatureCounts v1.4.6 (Liao *et al.*, 2014) was employed to determine the number of reads aligned to the reference genome for each *de novo* strategy. The command-line was: "-F GTF -t exon -p -s 2 -T 10". To evaluate the different *de novo* sub-strategies, the same parameters as these previously described for assessment of the *ab initio* approach were used.

To determine the similarity between the different assembled transcriptomes (after redundancy removal) of the four wheat lines (SG sub-strategy), the VennBLAST, an integrated software that combines a fast-parallelized BLAST filtering utility with whole-transcriptomic alignment comparison, was used (Zahavi *et al*., 2015). For this purpose, the following Blast command-line was applied on results of paired comparisons between

genotypes, which were obtained by the ncbi-blast+ (v.2.2.28; Camacho *et al*., 2009): "-p blastn -m 8 -b 1 -e 0.000001 -a 10".

5.2.6 Combined approach

To complement the results from *ab initio* analysis, unassigned reads were used as an input for Trinity (v2.0.6; Haas *et al.*, 2013) using the following parameters: "--seq_type fq -max_memory 40G -left Reads_R1.fastq -right Reads_R2.fastq -SS_lib_type RF -CPU 10 -min_contig_length 300 -normalize_reads -KMER_SIZE 23". Afterwards, all contigs were searched with ncbi-blast+ tool (v.2.2.28; Camacho *et al*., 2009) against the best *de novo* reference and the genome reference (*ab initio* approach). Blast command-line parameters were: "-p blastn -m 8 -b 1 -e 0.000001 -a 10". The contigs, for which blastn found a hit against the best *de novo* reference, but not against genome reference, were submitted to InterProScan (Jones *et al*., 2014). Then, contigs for which an IPRscan hit was found were added to the newly created combined wheat reference.

Subsequently, short reads from sequencing were aligned to the combined reference by TopHat2 (v2.1.1; Kim *et al*., 2013). The results were subjected to StringTie (v1.3.1c; Pertea *et al*., 2015) analysis, with the same setup as previously described. Again, the quantification of reads aligned to the reference genome was performed with featureCounts (v1.4.6; Liao *et al*., 2014). The Command-line parameters were: "-F GTF -t exon -p -s 2 -T 10".

Finally, gene differential expression was analysed by DESeq2 (Love *et al*., 2014) between the first group (W527 and W533; short root system) and the second group (W501 and W509; long root system). A gene was determined as significantly differentially expressed when adjusted *p*-value ≤ 0.01 and log2FC $\geq |2|$. Positive log2FC values indicated genes that were more abundant in the first group (W527/W533) compared to the second group (W501/W509); whereas negative log2FC values indicated genes that were less abundant in the first group (W527/W533) compared to the second group (W501/W509).

5.2.7 Annotation and GO analysis

The SATrans (v1.3; Kokáš *et al*., 2019) was used to functionally annotate the transcripts of the wheat combined reference obtained in the frame of the present study. The program was launched with default parameters. Transcripts from combined reference were used as an input in "create" mode. GO analysis of DEGs from the combined reference was performed by SATrans (v1.3) in "analysis" mode.

5.3 Results and discussion

5.3.1 Sequencing of samples

The root transcriptome of four genotypes of common wheat with contrasted root system architecture (RSA) was assessed in order to get insights concerning the molecular regulation of the wheat RSA. For this purpose, cDNA libraries were constructed and prepared according manufacturer's instruction (Illumina) for PE sequencing. The full root system of seven-week old plants grown in hydropony as previously described was harvested. Each genotype was represented by three independent biological replicates, and each replicate encompassed the root system of 3 plants grown in the same condition.

High throughput sequencing was performed on Illumina HiSeq platform and a total number of 1 454.2 million of 91 bp PE reads, representing 157.6 Gb of raw data were generated. After low-quality trimming of the 5'-end of each read, the same number of trimmed reads was used for downstream analysis (Tab. 12).

The GC content gives important indication of the stability of DNA. The average content of samples was 50.83%, in range with GC levels of monocots coding sequences (Kuhl *et al*., 2004). The GC content of raw reads did not show up substantially changes after read trimming. The length of raw reads was shortened from 91 bp to 80 bp.

Table 12: Statistics of trimmed reads. For this study, four genotypes were used (W501, W509, W527 and W533).

*Sample	Total reads	†GC content [%]
W501_A	118 215 047	51(51)
W501_B	152 126 685	51(51)
W501_C	197 091 190	50(50)
W509_A	171 684 759	50(50)
W509_B	154 768 440	52(51)
W509_C	83 681 849	51(51)
W527_A	121 900 138	52(52)
W527_B	83 313 983	51(51)
W527_C	64 007 073	51(50)
W527_A	79 425 504	51(51)
W527_B	124 557 305	51(51)
W527_C	103 457 690	51(51)

* Letters A, B, C indicates the different biological replicates.
† The numbers in the parentheses indicated GC content before trimming.

5.3.2 Different strategies for construction of the wheat reference genome

Because wheat has three sub-genomes (A, B and D) and because the reference genome was still incomplete on the time of the study, several strategies were used (*ab initio*, *de novo* and combined) to obtain the best wheat reference genome. Each strategy was evaluated for its suitability such purpose and showed specific restrictions.

The great limitation of *de novo* assembly is the misassembly of a large number of isoforms and diverse alleles. We chose two "sub-strategies" of *de novo* assembly, which use single genotypes (SG sub-strategy) and all genotypes (AMG sub-strategy) to generate assemblies. Trinity was used in both *de novo* assembly sub-strategies as an assembler, for his high efficiency in recovering full-length transcripts and correctly spliced isoforms, as described in the literature (Haas *et al.*, 2013; Duan *et al.*, 2012). To obtain consensual reference for all genotypes, which were used in this study, results from SG sub-strategy were merged and examined (SGM sub-strategy). The pipelines for sub-strategies of *de novo* strategy are shown in the figure 18.

The genome reference obtained from IWGSC (IWGSC, 2018) was used during the *ab initio* strategy as a reference, but inaccurate annotation significantly reduced his quality.

The combined strategy was designed based on the mapping-first approach. Reads were firstly mapped to the reference genome (from IGWSC), and unmapped reads were used to assemble new potential transcripts. Subsequently, transcripts with blast hit (obtained by IPRscan) were added to the reference genome.

In this study, we designed an optimal pipeline to build the wheat reference that fits best for downstream analysis of DEGs. Several steps were processed in the frame of the different strategies, including assembly with Trinity, reduction of redundancy and alignment of reads. The effect of each step was recorded in detail.

5.3.3 Statistics of assemblies from *de novo* strategy

For each *de novo* sub-strategy, the same options were used. To determine the relation of *de novo* assembly strategies, the size distribution of assembled contigs was compared between AMG and SG sub-strategy, including N50 value, number of longer than 1 000 bp assembled contigs, average contig size, etc...

The preliminary assessment of the two *de novo* assembly sub-strategies showed, that the AMG sub-strategy is more powerful than the SG one. Indeed, in contrast to the SG sub-strategy, the AMG sub-strategy assembly allowed to obtain longer contigs and in higher number. Also, the SG sub-strategy produced smaller transcriptomes, which were on average 1.98 times smaller than the transcriptomes generated with the AMG sub-strategy (Tab. 13).

Nevertheless, a large assembly includes transcripts with a wide range of expression levels but contains many redundant contigs. Therefore, it could be worse than a small assembly, which contains only unique transcripts. Redundancy constitutes an important problem for downstream analysis. In our study, we could observe that before deduplication (clustering related contigs to remove redundancy), assemblies produced by *de novo* sub-strategies had a much greater number of contigs and total number of bases.

To increase the performance of the SG sub-strategy, the assemblies after CD-HIT were combined into a single file (SGM sub-strategy) and the approach for redundancy reduction was applied similarly to SG and AMG sub-strategies.

Table 13: Statistics of each assembly step. For this study, four genotypes were used (W501, W509, W527 and W533).

Characteristics	*SG				†AMG	‡SGM
	W501	W509	W527	W533		
Initial assembly						
Number of contigs	628 498	630 943	430 086	500 674	1 126 997	-
Total bases	674 654 353	679 921 892	426 435 063	534 544 391	1 155 233 183	-
Number of contigs (≥ 1 kbp)	238 769	241 111	148 775	191 376	391 890	-
Max contig length	15 347	15 470	15 416	15 241	18 933	-
Average contig length	1 073	1 077	992	1 068	1 025	-
N50	1 471	1 470	1 321	1 443	1 382	-
GC content [%]	48.86	48.28	49.25	48.56	47.75	-
After CD-HIT						
Number of contigs	305 349	306 548	209 414	240 909	571 185	530 749
Total bases	290 243 525	295 938 117	185 234 499	229 058 031	532 534 850	503 408 811
Number of contigs (≥ 1 kbp)	92 933	96 012	56 952	74 390	167 917	157 650
Max contig length	15 347	15 470	15 416	15 241	18 933	15 470
Average contig length	951	965	885	951	932	949
N50	1 275	1 294	1 141	1 263	1 208	1 285
GC content [%]	48.15	47.02	47.62	47.10	46.63	46.80
After StringTie						
Number of contigs	81 338	87 189	73 732	79 712	130 488	122 469
Total bases	116 392 915	125 608 411	97 313 075	111 695 200	180 187 069	179 817 206
Number of contigs (≥ 1 kbp)	47 659	51 510	39 626	46 155	72 068	72 024
Max contig length	15 347	15 470	15 416	15 241	18 933	15 470
Average contig length	1 431	1 441	1 320	1 401	1 381	1 468
N50	1 820	1 819	1 672	1 759	1 772	1 922
GC content [%]	49.09	48.36	49.17	48.69	47.72	48.39

* SG: single genotype; † AMG: all merged genotypes; ‡ SGM: single genotypes merged

Contigs, that overlapped with a minimum length of 50 bp and with a minimum identity of 80%, were collapsed into single contig using cd-hit-est, therefore removing redundancy (Fu *et al.*, 2012; Duan *et al*., 2012).

After redundancy removal, both the number of contigs and the total number of bases were drastically lowered in comparison to the initial assembly. A large proportion of contigs (50% of total number) had been merged (Tab. 13). Merging of single genotype assemblies introduced more redundancy. After redundancy reduction, the N50 of SGM sub-strategy slightly increased by approximately 3.3% compared to the SG assemblies.

A major concern of removing redundancy is the mis-assembly of highly similar transcripts. After redundancy removal, the overall contig numbers were still very high (ranging from 209 414 to 571 185), even when considering the various transcript isoforms (Tab. 13). The reason might be sequencing errors, and potential sequence variation among individual genotypes. Thus, we implemented a StringTie launch to select sequences with high coverage of sequencing reads and decrease the number of contigs without loss of mappable reads.

The StringTie was applied on each *de novo* sub-strategy to filtrate poorly covered isoforms which might have been generated by misassembling. This step caused that, number of contigs was significantly decreased. Moreover, this step significantly increased the N50 and average contig length for all tested *de novo* sub-strategies. After StringTie launch, for *de novo* assembly sub-strategies, removal of roughly 70% contig number and 57% of total length was observed.

When considering maximum contig lengths and number of contigs, the AMG sub-strategy showed a higher performance than other sub-strategies (Tab. 13). This was observed at any points of the analysis (initial preassembly, after CD-HIT, after StringTie). Therefore, we concluded that AMG sub-strategy was the best *de novo* approach to compare wheat genotypes apparently distinctly related from the Chinese spring wheat reference genome.

5.3.4 Assessment of mapping, redundancy and similarity of *de novo* assemblies

To determine whether the transcriptomes of the four different wheat genotypes were similar, the four individual transcriptomes obtained by the SG sub-strategy was compared by BLAST as shown by Venn diagram (Fig. 19).

The similarity of transcriptomes was determined by paired comparison taking into consideration the following threshold values: a minimum of 80% length overlap between the aligned sequences and a minimum identity of 80%.

Venn diagram showed, that the W509 assembly had the greatest number of unique contigs, whereas W533 had fewest. Moreover, W509 had the largest overlaps with W501, indicating that they are probably very closely related. This could be related to the fact that they belong to the same "root phenotype" group. Only 18 662 contigs were common to all four genotypes. This supported our strategy to build a *de novo* consensual wheat reference transcriptome for downstream analysis of differential gene expression between genotypes. Otherwise, it would lead to the rapidly decrease of the mappable reads, and potentially erroneous analyse.

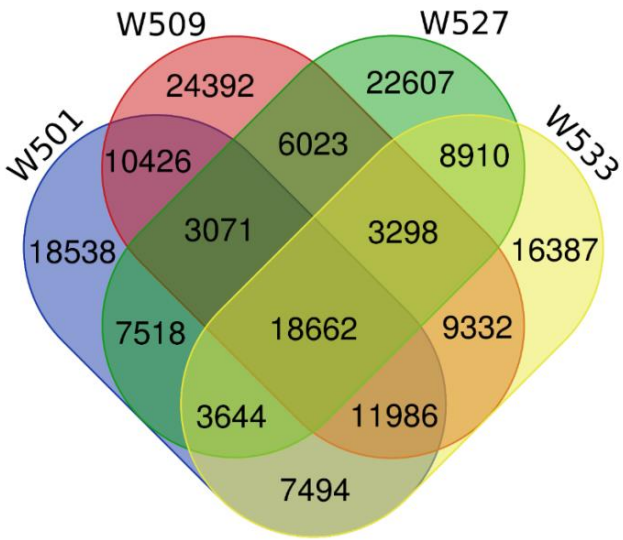


Figure 19: Venn diagram shows the similarity between reference transcriptomes obtained by *de novo* SG sub-strategy. For this study, four genotypes were used (W501, W509, W527 and W533).

To evaluate assembly redundancy, reads from all samples, were mapped to the *de novo* assemblies (Tab. 14). The proportion of reads mapping to more than one location was used as an indicator of redundancy.

On average, after the process for redundancy removal with CD-HIT, SG assemblies were significantly less redundant than AMG assembly (Tab. 14). The same trend was observed after contig filtration with StringTie.

Merging individual assemblies from SG sub-strategy slightly decreased mappable reads and significantly increased the number of reads which had more than one hit. The most redundant assembly was SGM, for which 14.57% of the total input reads matched more than one hit after deduplication and contig filtering.

The use of StringTie for read filtration had a significant effect on the number of contigs in assemblies, but with only slight loss of the mappable reads. Except of SGM sub-strategy, the proportion of reads mapped to more than one location decreased about 1.12% in average, and the uniquely mapped reads were reduced by about 1.56%. Small share of the reads, which were mapped to the multiple locations was expected, because the hexaploidy nature of the wheat genome. The comparison of reads mapped on multiple location and uniquely mapped reads showed, that SG sub-strategy better performed than other *de novo* sub-strategies.

Table 14: Statistics of reads mapped to assemblies/references. For this study, four genotypes were used (W501, W509, W527 and W533).

Characteristic	*SG				†AMG	‡SGM
	W501	W509	W527	W533		
Initial assembly/reference						
Uniquely aligned [%]	30.00	28.24	28.41	29.95	27.93	-
Multimapping [%]	46.73	41.43	45.37	45.74	45.96	-
After CD-HIT						
Uniquely aligned [%]	48.70	44.83	48.08	48.81	47.31	44.42
Multimapping [%]	8.10	6.66	7.53	7.39	8.15	17.45
After StringTie						
Uniquely aligned [%]	49.90	42.81	47.15	47.32	45.74	36.26
Multimapping [%]	6.63	5.65	6.53	6.40	7.02	14.57

* SG: single genotype; † AMG: all merged genotypes; ‡ SGM: single genotypes merged

Nevertheless, for downstream analysis of DEGs, the availability of a consensus reference for all genotypes is essential. Based on this fact, the results of the SG sub-strategy cannot be used. Among other *de novo* sub-strategies, AMG sub-strategy performed better than the SGM one, and only slightly increase the proportion of reads mapped to more than one location. The SGM sub-strategy, i.e. merging assembly from individual genotypes, had the following shortcomings: significantly less proportion of unique mapped reads and slightly increased redundancy.

5.3.5 Assessment quality of the wheat reference genome

To evaluate the suitability of strategies to build the reference, and performance of StringTie, we performed summarization of reads by FeatureCounts. The purpose of this evaluation was to determine which strategy has the best performance to create a new wheat reference transcriptome, and potential for downstream analysis of differential expression. Reads from each genotype were mapped to the assembly/reference and each assembly/reference had a different number of contigs/transcripts which were annotated (Tab. 15).

With the exception of SGM sub-strategy, the proportion of reads which were uniquely aligned to the exon regions were equivalent for the all *de novo* sub-strategies, also after deduplication with StringTie. The same observation was done for reads, that do not overlap any exon (Unassigned_NoFeatures), and the reads which overlap two or more genes (Unassigned_Ambiguity).

The AMG sub-strategy showed better performance over the SGM sub-strategy and formed the best consensus *de novo* wheat transcriptome reference for downstream analysis. When the *ab initio* strategy was used, as well as the combined strategy after StringTie, the proportion of reads mapped to the exon regions was 49.94% and 58.93%, respectively. The use of StringTie caused that, the proportion of uniquely aligned reads to the exon region, was significantly increased for all strategies, mostly for *de novo* strategy (17.29% in average). Moreover, this step significantly decreased the number of reads, that do not overlap any exon (Unassigned_NoFeatures) whatever the considered strategy. This was caused by improving annotation (for all strategies) and decreasing number of sequences (in case of *de novo* strategy).

Table 15: Statistics of reads aligned in region of mRNA. For this study, four genotypes were used (W501, W509, W527 and W533).

Characteristic	*De novo* strategy						*Ab initio* strategy	Combined strategy
	*SG				†AMG	‡SGM		
	W501	W509	W527	W533				
After CD-HIT								
Number of contigs/genes	305 349	306 548	209 414	240 909	571 185	530 749	110 790	117 826
Unique reads [%]	27.93	28.42	27.89	28.36	28.95	21.65	47.27	51.90
Unassigned (No feature) [%]	6.37	6.16	8.11	6.05	7.75	4.71	15.52	10.18
Unassigned (Ambiguity) [%]	29.12	28.46	29.26	30.89	27.52	31.35	0.69	1.50
After StringTie								
Number of contigs/genes	81 338	87 189	73 732	79 712	130 488	122 469	69 981	83 300
Unique reads [%]	43.89	45.32	46.39	47.45	44.95	37.52	49.94	58.93
Unassigned (No feature) [%]	1.94	2.04	1.98	1.90	2.19	1.57	6.60	7.16
Unassigned (Ambiguity) [%]	1.41	1.18	1.74	1.75	1.83	1.61	1.41	1.60

* SG, single genotype; † AMG, all merged genotypes; ‡ SGM, single genotypes merged

In conclusion, we found that a combined strategy is the best strategy to obtain a wheat reference transcriptome when considering not related genotypes. This was supported by the highest proportion of uniquely aligned reads to the exon regions (Tab. 15) and the low number of reads mapping to multiple places, especially when comparing with the *ab initio* strategy (Supplemental table 11).

5.3.6 Functional annotation of putative genes

A final comprehensive wheat root reference transcriptome was generated using a combined strategy. A total of 83 300 sequences were identified (Tab. 15) as putative genes expressed in the roots of 7-week old wheat plants grown in hydropony in controlled conditions. This reference transcriptome will be used for differential expression between genotypes with distinct RSA.

Still at that point, the putative genes were not annotated. Therefore, the obtained sequences were compared against the NCBI Nucleotide database ("nt"; NCBI Resource Coordinators, 2014) and InterPro (Finn *et al.*, 2017) database with a cut-off e-value $\leq 1.10^{-3}$ using SATrans. BLAST alignment to "nt" database showed that 74 531 (89.47%) putative genes aligned to "nt" database while the remaining 8 761 (10.53%) did not show homology to any known sequence in the database.

Among the aligned query sequences, 73.06% had an e-value of less than 1.10^{-26} and showed very strong homology (more than 80%) and coverage (more than 90%) with the sequence in the database. The remaining 26.94% query sequences had an e-value ranging from 1.10^{-3} to 1.10^{-26}.

To study the sequence conservation of our reference compared to other plant species, we analysed the species distribution of the sequence dataset by aligning sequences against the "nt" database. Approximately 87.98% of total query sequences matched with sequences from five top-hit species, i.e., *Hordeum vulgare* (49.14%), *Triticum aestivum* (14.20%), *Brachypodium distachyon* (13.52%), *Oryza sativa* (7.26%) and *Aegilops tauschii* (3.86%) all of which are members of the *Poaceae* family. The top ten-hit species based on "nt" annotation are shown in Fig. 20.

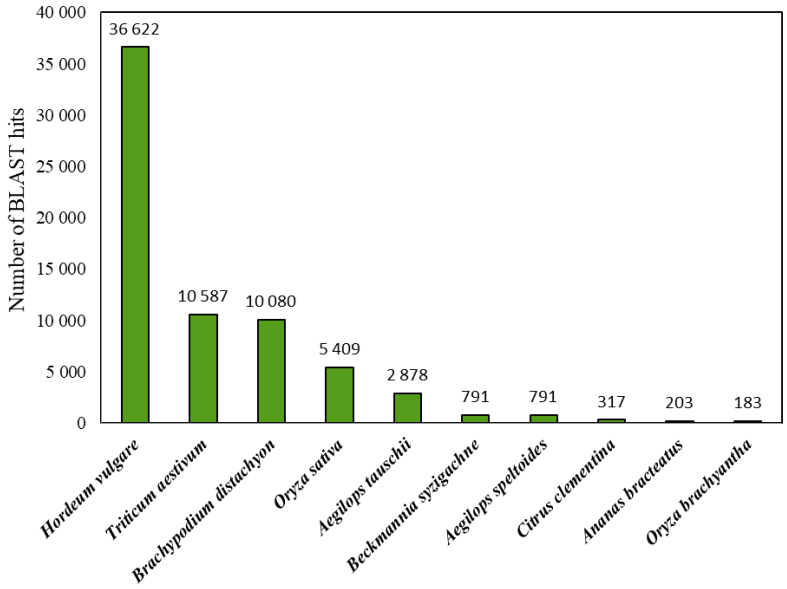


Figure 20: Species distribution of BLAST hits for functional annotation of putative genes.

Sequence homology based on GO classification using SATrans tool revealed that out of the assembled sequences 29 775 were annotated in the three main GO categories, including 39 functional groups (Supplemental table 12). A total of 720 488 GO assignment were obtained. From total count of annotated genes, 26 731 had record in the largest category – MF, followed by BP (17 662 of putative genes) and CC (7 060 of putative genes) at level 2.

5.3.7 Analysis of differentially expressed genes between genotypes

In order to identify the putative molecular mechanism that could regulate RSA in wheat, we performed the differential analysis of genes expressed in the roots of 7-week old plants grown in hydropony. For the initial purpose of the study, four genotypes were used (W501, W509, W527 and W533). Nevertheless, for simplest analysis and because the pedigree (as the resulting correlation) was unknown, it was decided to group the genotypes based on their RSA: short (W527 and W533) and long (W501 and W509). The differential expression analysis was investigated by DESeq2 and normalized RPKMs values were subjected to PCA, in order to control the quality of independent biological replicates. The PCA demonstrated clustering of the biological replicates and of the different genotypes (Fig. 21).

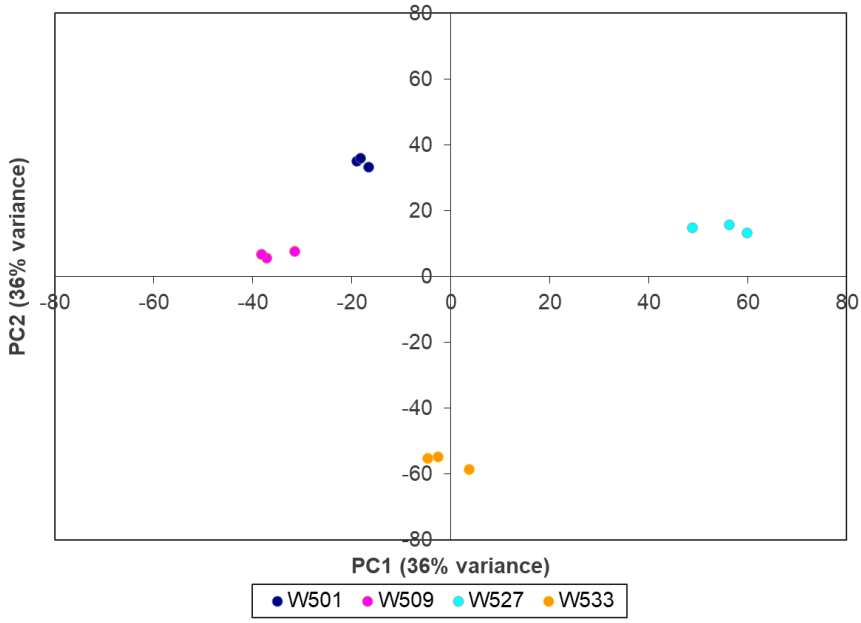


Figure 21: The PCA analysis for replicates from genotypes of wheat. For this study, four genotypes were used (W501, W509, W527 and W533).

The comparison between genotypes with a short (W527, W533) and a long root system (W501, W509) under defined conditions identified a total of 3 942 DEGs (adjusted *p*-value ≤ 0.01, log2FC ≥ |2|). Of them, 2 193 were up-regulated (i.e. more expressed in "short roots") and 1 749 were down-regulated (i.e. less expressed in "short roots" or "more expressed in "long roots"; Supplemental table 13).

To further look into the functional categories of genes differentially expressed between genotypes of wheat with short and long root system, GO analysis was performed using SATrans (Supplemental table 14 for down-regulated genes and 15 for up-regulated genes). The level six was selected as the most meaningful for our analysis.

The most significantly GO-terms (GO level 6; *p*-value ≤ 0.05; number of DEGs ≥ 10) of the up-regulated genes in genotypes with short roots (W527 and W533) are shown in Tab. 16. In the BP category, the majority of the up-regulated genes was related to "cellular protein metabolic process" (89 genes), "phosphate-containing compound metabolic process" (65 genes) and "gene expression" (64 genes). Under the MF, the top three categories were "ribonucleoside binding" (130 genes), "kinase activity" (57 genes) and "phosphotransferase activity, alcohol group as acceptor" (52 genes). This suggested that these functions were more active in the short root system.

Table 16: The most affected GO terms for genes which were up-regulated in the "short-roots" genotypes (W527 and W533) compared to "long-roots" genotypes (W501 and W509), at the GO level 6 (*p*-value ≤ 0.05; number of DEGs ≥ 10).

GO Level	†Ontology source	GO id	Description of GO	Number of GO annotated sequences	*DEGs number	Percent [%]	*p*-value
6	BP	GO:0008299	isoprenoid biosynthetic process	86	12	13.95	0.0000
6	BP	GO:0006720	isoprenoid metabolic process	87	12	13.79	0.0000
6	BP	GO:0031326	regulation of cellular biosynthetic process	1 351	24	1.78	0.0339
6	BP	GO:0010468	regulation of gene expression	1 379	24	1.74	0.0286
6	BP	GO:0010556	regulation of macromolecule biosynthetic process	1 348	24	1.78	0.0338
6	BP	GO:0019219	regulation of nucleobase-containing compound metabolic process	1 324	24	1.81	0.0484
6	BP	GO:0019438	aromatic compound biosynthetic process	1 909	25	1.31	0.0001
6	BP	GO:0018130	heterocycle biosynthetic process	1 937	25	1.29	0.0000
6	BP	GO:0034654	nucleobase-containing compound biosynthetic process	1 763	25	1.42	0.0004
6	BP	GO:1901362	organic cyclic compound biosynthetic process	1988	25	1.26	0.0000
6	BP	GO:0016070	RNA metabolic process	2 347	26	1.11	0.0000
6	BP	GO:0043603	cellular amide metabolic process	946	38	4.02	0.0151
6	BP	GO:0006518	peptide metabolic process	935	38	4.06	0.0141
6	BP	GO:0055085	transmembrane transport	1 068	41	3.84	0.0279
6	BP	GO:0090304	nucleic acid metabolic process	2 890	44	1.52	0.0000
6	BP	GO:0043412	macromolecule modification	3 537	51	1.44	0.0000
6	BP	GO:0036211	protein modification process	3 374	51	1.51	0.0000
6	BP	GO:0010467	gene expression	3 022	64	2.12	0.0451
6	BP	GO:0006796	phosphate-containing compound metabolic process	3 360	65	1.93	0.0046
6	BP	GO:0044267	cellular protein metabolic process	4 449	89	2.00	0.0027
6	CC	GO:0005634	nucleus	1 352	19	1.41	0.0022
6	CC	GO:0005840	ribosome	532	37	6.95	0.0000

Table 16: The most affected GO terms for genes which were up-regulated in the "short-roots" genotypes (W527 and W533) compared to "long-roots" genotypes (W501 and W509), at the GO level 6 (*p*-value ≤ 0.05; number of DEGs ≥ 10). (*continued***)**

6	MF	GO:0003723	RNA binding	862	11	1.28	0.0077
6	MF	GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	345	16	4.64	0.0423
6	MF	GO:0020037	heme binding	768	39	5.08	0.0002
6	MF	GO:0016773	phosphotransferase activity, alcohol group as acceptor	2 776	52	1.87	0.0060
6	MF	GO:0016301	kinase activity	2 868	57	1.99	0.0160
6	MF	GO:0032549	ribonucleoside binding	5 977	130	2.18	0.0101

* DEGs: differentially expressed genes.
† MF: molecular function, BP: biological process, CC: sub-cellular component.

With respect to CC, “ribosome” (39 genes) and “nucleus” (19 genes) were the dominant groups. Interestingly, in the transcriptome of wheat genotypes with short roots, genes related to “isoprenoid biosynthetic process” were highly represented. Genes in this category included the deoxyxylulose-5-phosphate (DXP) synthase (DXS: MSTRG.11778, MSTRG.11779, MSTRG.16127, MSTRG.2126, MSTRG.4093 and MSTRG.19223), the DXP-reductoisomerase (DXR: MSTRG.25866), the 4-diphosphocytidyl-2C-methyl-D-erythritol kinase (CMK: MSTRG.29720 and MSTRG.32099), and the 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (HMB-PP synthase or HDS: MSTRG.55623, MSTRG.62022 and MSTRG.57977). Their log2FC expression varied from 2.19 to 4.13. All these enzymes are part of the non-mevalonate pathway of isoprenoid precursor biosynthesis (Hunter, 2007) that leads to the synthesis of compounds such as GAs, carotenoids and ABA, chlorophylls, tocopherols, plastoquinone’s and phylloquinones (Rodríguez-Concepción & Boronat, 2002). Both GAs and ABA are known to have a function in root development (Garay-Arroyo *et al.*, 2012). Nevertheless, because the genes described above are too high in any of the pathway, it would be too much speculative to formulate hypothesis related to the importance of GAs or ABA in the development of roots in the “short-root” type genotype of wheat under examination.

Another affected BP was related to “transmembrane transport”. The few genes in this category were manually annotated as genes involved in the transport of plant hormones across biological membranes (Tab. 17). The development of the root system is controlled by plant hormones, especially auxins and cytokinins (Aloni *et al.*, 2006). Two putative *PIN-FORMED* (*PIN*) genes and five *PIN-LIKES* (*PILS*) genes were listed as significantly up-regulated in short-roots genotypes (W527/W533) in comparisons with long-roots genotypes (W501/W509). These genes encode the transporters which ensures the efflux of auxin across the plasma membrane and its release from the endoplasmic reticulum, respectively (Feraru *et al.*, 2012; Talboys *et al.*, 2014). It can support the assumption that the differential root system can be affected by stronger transport of auxins, which can generate local auxin maxima during development of the root system.

Table 17: Selected putative genes which are involved in metabolism of plant hormones.

Name of genes	Description of annotation hit	log2FoldChange	Functional group
MSTRG.59389	Auxin-responsive protein SAUR36	2.30	Auxin response factor
MSTRG.43645	Protein PIN-LIKES 3	3.74	Auxin transport – PILS proteins
MSTRG.44906	Protein PIN-LIKES 7	2.43	Auxin transport – PILS proteins
MSTRG.48538	Protein PIN-LIKES 7	3.00	Auxin transport – PILS proteins
MSTRG.48589	Protein PIN-LIKES 3	2.10	Auxin transport – PILS proteins
MSTRG.53839	Protein PIN-LIKES 3	2.15	Auxin transport – PILS proteins
MSTRG.70818	Probable auxin efflux carrier protein	3.48	Auxin transport – PIN proteins
MSTRG.73507	Probable auxin efflux carrier protein	2.70	Auxin transport – PIN proteins
MSTRG.23683	Cytokinin oxidase/dehydrogenase (*CKX4*) gene	2.14	Cytokinines – CKX
MSTRG.30927	Cytokinin oxidase/dehydrogenase (*CKX4*) gene	2.86	Cytokinines – CKX
MSTRG.13174	Probable indole-3-acetic acid-amido synthetase GH3.8	2.66	IAA amido synthetase
MSTRG.17957	Probable indole-3-acetic acid-amido synthetase GH3.8	2.35	IAA amido synthetase
MSTRG.22517	Probable indole-3-acetic acid-amido synthetase GH3.8	3.15	IAA amido synthetase
MSTRG.23806	IAA-amino acid hydrolase ILR1-like 2	2.64	IAA-amino acid hydrolase
MSTRG.26776	IAA-amino acid hydrolase ILR1-like 2	2.70	IAA-amino acid hydrolase
MSTRG.32390	IAA-amino acid hydrolase ILR1-like 2	2.41	IAA-amino acid hydrolase
MSTRG.76037	IAA-amino acid hydrolase ILR1-like 8	-2.77	IAA-amino acid hydrolase

The local auxin maxima are crucial for establishing and maintaining root primordium, and consequently root branching. Moreover, elevated steady-state auxin concentration in elongating root cells has been shown to promote the elongation of those cells (Feraru *et al.*, 2012; Pacheco-Villalobos *et al.*, 2016).

The differential regulation of genes related to auxin biosynthetic pathway can be another indicator of changes of endogenous auxin status. Interestingly, none of the genes related to auxin synthesis were differentially expressed, except of those related to IAA-amino acid conjugate metabolism (3 genes coding for IAA amido synthetase) and hydrolysis (4 genes encoding IAA-amino acid hydrolase) that occurs in the endoplasmatic reticulum (Ludwig-Müller, 2011; Ostrowski *et al.*, 2014). In summary, the local auxin maxima can also be regulated by processes of biosynthesis of conjugates IAA, and their hydrolysis.

In contrast to auxin, CKs have an antagonistic effect on lateral root branching (Aloni *et al.*, 2006). In agreement with the hypothesis that auxin might dominate in the root system of genotypes with short root system (W527 and W533), two genes involved in reversible degradation of cytokinins (MSTRG.23683; MSTRG.73507) were identified as differentially up-regulated, suggesting a higher degradation of active CKs in the short-root type genotypes (W527 and W533). Nevertheless, once again our hypothesis based on the analysis of RNA-seq data require further investigations such as the validation of gene expression by classical qPCR and the determination of endogenous auxins and CKs content in the root of the 4 wheat genotypes under investigation.

The enriched GO terms in down-regulated genes (GO level 6; p-value ≤ 0.05; number of DEGs ≥ 10) in genotypes with short root system are summarized in Tab. 18. In BP category, the majority of up-regulated genes are involved in “cellular protein metabolic process” (64 genes), “macromolecule modification” (59 genes) and “cellular nitrogen compound biosynthetic process” (32 genes). Under “molecular-function”, the top three categories were “purine ribonucleoside triphosphate binding” (80 genes), “heme binding” (67 genes) and “hydrolase activity, acting on acid anhydrites, in phosphorus-containing anhydrides” (13 genes).

Table 18: The most affected GO terms for genes which were up-regulated, at the GO level 6 (*p*-value ≤ 0.05; number of DEGs ≥ 10).

GO Level	†Ontology source	GO id	Description of GO	Number of GO annotated sequences	*DEGs number	Percent [%]	*p*-value
6	BP	GO:1901362	organic cyclic compound biosynthetic process	1 988	12	0.60	0.00000
6	BP	GO:0018130	heterocycle biosynthetic process	1 937	12	0.62	0.00000
6	BP	GO:0019438	aromatic compound biosynthetic process	1 909	12	0.63	0.00000
6	BP	GO:0010467	gene expression	3 022	13	0.43	0.00000
6	BP	GO:0090304	nucleic acid metabolic process	2 890	13	0.45	0.00000
6	BP	GO:0044106	cellular amine metabolic process	72	14	19.44	0.00000
6	BP	GO:0009308	amine metabolic process	81	14	17.28	0.00000
6	BP	GO:0016053	organic acid biosynthetic process	347	14	4.03	0.02301
6	BP	GO:0009059	macromolecule biosynthetic process	2 929	19	0.65	0.00000
6	BP	GO:0034645	cellular macromolecule biosynthetic process	2 929	19	0.65	0.00000
6	BP	GO:0044271	cellular nitrogen compound biosynthetic process	2 868	32	1.12	0.00004
6	BP	GO:0043412	macromolecule modification	3 537	59	1.67	0.04965
6	BP	GO:0044267	cellular protein metabolic process	4 449	64	1.44	0.00063
6	MF	GO:0030410	nicotianamine synthase activity	21	10	47.62	0.00000
6	MF	GO:0003677	DNA binding	1 838	13	0.71	0.00000
6	MF	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	1 201	13	1.08	0.00855
6	MF	GO:0020037	heme binding	768	67	8.72	0.00000
6	MF	GO:0035639	purine ribonucleoside triphosphate binding	4 862	80	1.65	0.01402

* DEGs: differentially expressed genes.
† MF: molecular function, BP: biological process, CC: sub-cellular component.

We could observe that genes related to “phenylpropanoid catabolic process” and “lignin catabolic process” were down-regulated in the roots of genotypes with short-roots (Supplemental table 14). Metabolism of lignin play an important role in the growth and development of plants (Fan *et al.*, 2006; Liu *et al.*, 2018).

In addition, lignin metabolism can be actively involved in response to various environmental stresses. Plants with higher level of lignin metabolites, exhibit better growth parameters (such as yield), when encountering drought stress (Liu *et al.*, 2018; Pospíšilová *et al.*, 2016). Nevertheless, further deep phenotyping of the different genotypes would be required to make a precise hypothesis concerning lignin accumulation in the roots and the tolerance to drought of the different wheat genotypes under consideration in the present study.

As other “molecular-function” which was enriched, the “nicotianamine synthase activity” is listed. Ten genes, coding putative nicotianamine synthase (MSTRG.26900; MSTRG.42305; MSTRG.55609; MSTRG.56827; MSTRG.56828; MSTRG.58708; MSTRG.60450; MSTRG.63015; MSTRG.63552; MSTRG.63617), were significantly up-regulated (log2FC in range from -2.29 to -5.31). Nicotianamine serves as metal chelator and is synthetized by trimerization of three molecules of S-adenosyl-L-methionine by nicotianamine synthase (Zhou *et al*., 2013). The enzyme is reported, as the candidate for improving the Fe-deficiency tolerance (Nozoye, 2018), and drought stress tolerance (Zhang & Zheng, 2008).

In conclusion, to verify all results which were obtained by RNA-seq, the biological experiments should be done and possible changes in gene expression and levels of metabolites which are indicated here should be observed by different techniques (for example qPCR).

CONCLUSIONS

The high throughput RNA-seq represents a crucial approach to study transcriptome, and consequently understand plant phenotype. In the first chapter of the current dissertation, we present a detailed review dealing with RNA-seq methods, design of RNA-seq experiment and the bioinformatic tools available for downstream analysis of the generated data. RNA-seq and bioinformatics tools have been used to infer the molecular regulation during drought tolerance and root development in barley (*Hordeum vulgare* L.) and wheat (*Triticum aestivum* L.), respectively.

The second chapter described transgenic barley lines with altered endogenous CK content overexpressing the *CKX1* gene from *Arabidopsis* under the control of the mild root-specific promotor of the maize β-glucosidase, targeted to various subcellular compartments. When submitted to water stress, the transgenic lines were more tolerant to drought than the WT plants, mainly due to the alteration of their root architecture and a stronger lignification of root tissue. The RNA-seq study that has been carried out enabled a comprehensive inspection of the molecular regulations occurring in the tissue of plants with CK imbalance, as well as in WT plants, not only during drought stress but also during revitalization. For instance, the up-regulation of four genes encoding aquaporin might contribute to the fact that all transgenic lines were able to increase water potential faster than WT plants. In addition, the process of leaf revitalization is accompanied by the up-regulation of genes encoding proteins involved in photosynthesis, and especially those of chloroplastic origin. This aspect led to faster regeneration of transgenic plants which was observed as higher biomass accumulation. Altered CK status noticeably affected the secondary metabolism derived from phenylalanine and led to the accumulation of intermediates of the phenylpropanoid pathway in the roots.

The third chapter of the dissertation describes SATrans, a novel bioinformatics tool which was developed to contribute to understand and biologically interpret the RNA-seq data. The software is primarily focused on transcriptome research to provide fast and reliable functional annotation of nucleotide/amino acid sequences. The other crucial function of the software is functional analysis differential gene expression at the whole transcriptome level. SATrans is highly robust and requires only the basic knowledge of a

Linux operating system and provides outputs in a user-friendly environment. In addition, the SATrans is a freeware that might be easily upgraded in the future and extended by new modules, thereby giving it great potential for additional, future tasks.

The fourth chapter describes transcriptomic analysis of wheat inbred lines with different root architecture. The hexaploid nature of the wheat inbred lines made the transcriptomic analysis very difficult. Several approaches were considered: mapping toward the available reference genome, creating a new reference transcriptome by *de novo* assembly, or building a new reference genome combining reference wheat transcriptome and *de novo* assembly that can be considered as pan-transcriptomic. In our conditions, the combined approach was evaluated as the best option for building highly quality reference transcriptome prior read mapping. The down-stream analysis aimed to unravel molecular mechanisms that could be responsible for the difference in root system (short vs. long). Our data showed that few biological processes were affected. For instance, the biological process related to "biosynthesis of isoprenoids" was up-regulated in the two genotypes with short root system, showing an apparent increase tolerance to stress. Accumulation of isoprenoids, such as lignin, can serve as an essential advantage for plants which are exposed to drought stress. Other processes such as transmembrane transport of auxins, hydrolysis of its conjugates or its degradation were also found to be differentially regulated between genotypes with short and long roots. Therefore, it can be hypothesized that the different architecture of the root system observed between the 4 genotypes might be related to different hormonal (auxins, but also cytokinins) content. However, biological experiments are required to further validate our hypothesis based on RNA-seq analysis. One might consider experimenting the tolerance to stresses (drought) of the different genotypes, in combination with monitoring plant hormones and other compounds.

LIST OF FIGURES AND TABLES

List of figures

List of tables

ABBREVIATIONS

ABA	abscisic acid
AMG	all merged genotypes
APS	adenosine-5′-phosphosulphate
ARF	auxin response factor
ASCII	American standard code for international interchange
ATP	adenosine-triphosphate
Aux	auxin
BLAST	basic local alignment search tool
bp	base pairs
BP	biological process
BR	brassinosteroid
BWT	Burrows-Wheeler transformation
CC	sub-cellular component
CCD	charge-coupled-device
cDNA	coding DNA
CK	cytokinin
CKX	cytokinin dehydrogenase
CKX1	cytokinin dehydrogenase 1
CMK	4-diphosphocytidyl-2C-methyl-D-erythritol kinase
Cp	*Claviceps purpurea*
DEG	differentially expressed gene
dNTP	deoxy-nucleotide triphosphate
DR	drought resistance
dsDNA	double-stranded-DNA
DXR	deoxyxylulose-5-phosphate reductoisomerase
DXS	deoxyxylulose-5-phosphate synthase
EC	enzyme commission
emPCR	emulsion polymerase chain reaction
FPKM	fragments per kilobase of exon model per million mapped reads
FRET	fluorescence resonance energy transfer
GA	gibberellin
GH3	Gretchen Hagen 3
GO	gene ontology
GTF	gene transfer format
HDS	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
IAA	indole-3-acetic acid
IPT	isopentenyl transferase
JA	jasmonate
Log2FC	log2FoldChange
MF	molecular function
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
ORF	open reading frame
PCA	principal component analysis

PCR	polymerase chain reaction
PE	paired-end
PGSB	Plant Genome and Systems Biology
PILS	Pin-Likes
PIN	Pin-Formed
PO	plant ontology
PPi	pyrophosphate
RNA-seq	RNA sequencing
RPKM	reads per kilobase exon model per million reads
rRNA	ribosomal RNA
RSA	root system architecture
SAUR	small auxin up RNA
SE	single-end
SG	single genotypes
SGM	single merged genotypes
SGS	second-generation sequencing
SMRT	single-molecule real-time sequencing
SNP	single-nucleotide polymorphism
SO	sequence ontology
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
ssDNA	single-stranded-DNA
qPCR	quantitative polymerase chain reaction
TEM	transmission electron microscope
TGS	third-generation sequencing
WT	wild-type
ZMW	zero-mode waveguide

REFERENCES

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., & Polz, M. F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, *71*(12), 8966–8969. https://doi.org/10.1128/AEM.71.12.8966-8969.2005

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.-J., Mayer, P., & Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research*, *28*(20), E87.

Ahammed, G.J., Xia, X.-J., Li, X., Shi, K., Yu, J.-Q., & Zhou, Y.-H. (2015). Role of brassinosteroid in plant adaptation to abiotic stresses and its interplay with other hormones. *Current Protein & Peptide Science, 16*(5), 462-473. https://doi.org/10.2174/1389203716666150330141427

Aloni, R., Aloni, E., Langhans, M., & Ullrich, C. I. (2006). Role of cytokinin and auxin in shaping root architecture: regulating vascular differentiation, lateral root initiation, root apical dominance and root gravitropism. *Annals of Botany*, *97*(5), 883–893. https://doi.org/10.1093/aob/mcl027

Alonso-Ramírez, A., Rodríguez, D., Reyes, D., Jiménez, J. A., Nicolás, G., López-Climent, M., … Nicolás, C. (2009). Evidence for a role of gibberellins in salicylic acid-modulated early plant responses to abiotic stress in *Arabidopsis* seeds. *Plant Physiology*, *150*(3), 1335–1344. https://doi.org/10.1104/pp.109.139352

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

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

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, *25*(4), 195–203. https://doi.org/10.1016/j.nbt.2008.12.009

Aris-Brosou, S. (2005). Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Molecular Biology and Evolution*, *22*(2), 200–209. https://doi.org/10.1093/molbev/msi006

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., … Mitchell, A. L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database: The Journal of Biological Databases and Curation*, *2012*, bas019. https://doi.org/10.1093/database/bas019

Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, *38*(14), 4570–4578. https://doi.org/10.1093/nar/gkq211

Auer, P. L., Srivastava, S., & Doerge, R. W. (2012). Differential expression--the next generation and beyond. *Briefings in Functional Genomics*, *11*(1), 57–62. https://doi.org/10.1093/bfgp/elr041

Baldwin, I. T., Schmelz, E. A., & Ohnmeiss, T. E. (1994). Wound-induced changes in root and shoot jasmonic acid pools correlate with induced nicotine synthesis in *Nicotiana sylvestris* spegazzini and comes. *Journal of Chemical Ecology*, *20*(8), 2139–2157. https://doi.org/10.1007/BF02066250

Basu, S., Ramegowda, V., Kumar, A., & Pereira, A. (2016). Plant adaptation to drought stress. *F1000Research*, *5*,1554. https://doi.org/10.12688/f1000research.7678.1

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

Bru, C., Courcelle, E., Carrére, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Research*, *33*(Database issue), D212-D215. https://doi.org/10.1093/nar/gki034

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta*, *1842*(10), 1932-1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., … Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, *40*(6), 722–729. https://doi.org/10.1038/ng.128

Cass, C. L., Peraldi, A., Dowd, P. F., Mottiar, Y., Santoro, N., Karlen, S. D., … Sedbrook, J. C. (2015). Effects of PHENYLALANINE AMMONIA LYASE (PAL) knockdown on cell wall composition, biomass digestibility, and biotic and abiotic stress responses in *Brachypodium*. *Journal of Experimental Botany*, *66*(14), 4317–4335. https://doi.org/10.1093/jxb/erv269

Cha, J.-Y., Barman, D. N., Kim, M. G., & Kim, W.-Y. (2015). Stress defense mechanisms of NADPH-dependent thioredoxin reductases (NTRs) in plants. *Plant Signaling & Behavior*, *10*(5), e1017698. https://doi.org/10.1080/15592324.2015.1017698

Chen, T.W., Gan, R.C., Wu, T.H., Huang, P.J., Lee, C.Y., Chen, Y.Y., … Tang, P. (2012). FastAnnotator—an efficient transcript annotation web tool. *BMC Genomics, 13*(Suppl 7): S9. https://doi.org/10.1186/1471-2164-13-S7-S9

Cheng, Z., Dong, K., Ge, P., Bian, Y., Dong, L., Deng, X., … Yan, Y. (2015). Identification of Leaf Proteins Differentially Accumulated between Wheat Cultivars Distinct in Their Levels of Drought Tolerance. *PLoS ONE*, *10*(5), e0125302. https://doi.org/10.1371/journal.pone.0125302

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., … Clark, M. D. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research*, *27*(5), 885–896. https://doi.org/10.1101/gr.217117.116

Cocquet, J., Chong, A., Zhang, G., & Veitia, R. A. (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics*, *88*(1), 127–131. https://doi.org/10.1016/j.ygeno.2005.12.013

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Consortium PO. (2002). The plant ontology consortium and plant ontologies. *Comparative and Functional Genomics*, *3*(2), 137-142. https://dx.doi.org/10.1002/cfg.154

Craig, D. W., Pearson, J. V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J. J., … Huentelman, M. J. (2008). Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*, *5*(10), 887–893. https://doi.org/10.1038/nmeth.1251

Crawford, J. E., Guelbeogo, W. M., Sanou, A., Traoré, A., Vernick, K. D., Sagnon, N., & Lazzaro, B. P. (2010). *De novo* transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. *PloS ONE*, *5*(12), e14202. https://doi.org/10.1371/journal.pone.0014202

Čudejková, M.M., Vojta, P., Valík, J., & Galuszka, P. (2016). Quantitative and qualitative transcriptome analysis of four industrial strains of *Claviceps purpurea* with respect to ergot alkaloid production. *New Biotechnology*, *33*(5 Pt B), 743–754. https://doi.org/10.1016/j.nbt.2016.01.006

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., … Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research*, *43*(D1), D662-D669. https://doi.org/10.1093/nar/gku1010

Dahl, F., Gullberg, M., Stenberg, J., Landegren, U., & Nilsson, M. (2005). Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Research*, *33*(8), e71. https://doi.org/10.1093/nar/gni070

Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., … Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*, *11 Suppl 4*, S7. https://doi.org/10.1186/1471-2164-11-S4-S7

De Bona, F., Ossowski, S., Schneeberger, K., & Rätsch, G. (2008). Optimal spliced alignments of short sequence reads. *Bioinformatics*, *24*(16), i174-i180. https://doi.org/10.1093/bioinformatics/btn300

de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular BioSystems*, *5*(12), 1512–1526. https://doi.org/10.1039/b908315d

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., … Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, *12*(6), 1058–1067. https://doi.org/10.1111/1755-0998.12003

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–3212. https://doi.org/10.1093/bioinformatics/btp579

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., … Artiguenave, F. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biology*, *9*(12), R175. https://doi.org/10.1186/gb-2008-9-12-r175

Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., … Bourne, P.E. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Research*, *33*, D233–D237. https://dx.doi.org/10.1093%2Fnar%2Fgki057

Dodt, M., Roehr, J. T., Ahmed, R., & Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, *1*(3), 895–905. https://doi.org/10.3390/biology1030895

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105. https://doi.org/10.1093/nar/gkn425

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., & Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 8817–8822. https://doi.org/10.1073/pnas.1133470100

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., … Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, *327*(5961), 78–81. https://doi.org/10.1126/science.1181498

Duan, J., Xia, C., Zhao, G., Jia, J., & Kong, X. (2012). Optimizing *de novo* common wheat transcriptome assembly using short-read RNA-Seq data. *BMC Genomics*, *13*, 392. https://doi.org/10.1186/1471-2164-13-392

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, *6*(5), R44. https://doi.org/10.1186/gb-2005-6-5-r44

El Hassouni, K., Alahmad, S., Belkadi, B., Filali-Maltouf, A., Hickey, L.T., & Bassi, F.M. (2018). Root system architecture and its association with yield under different water regimes in Durum wheat. *Crop Science*, *58*(6), 2331–2346. https://doi.org/10.2135/cropsci2018.01.0076

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., The RGASP Consortium, … Bertone, P. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, *10*(12), 1185–1191. https://doi.org/10.1038/nmeth.2722

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Research*, *8*(3), 186–194.

Fàbregas, N., Lozano-Elena, F., Blasco-Escámez, D., Tohge, T., Martínez-Andújar, C., Albacete, A. … Caño-Delgado, A.I. (2018). Overexpression of the vascular brassinosteroid receptor BRL3 confers drought resistance without penalizing plant growth. *Nature Communications, 9,* 4680. https://doi.org/10.1038/s41467-018-06861-3

Fan, L., Linker, R., Gepstein, S., Tanimoto, E., Yamamoto, R., & Neumann, P.M. (2006). Progressive inhibition by water deficit of cell wall extensibility and growth along the elongation zone of maize roots is related to increased lignin metabolism and progressive stelar accumulation of wall phenolics. *Plant Physiology, 140*(2), 603-612. https://doi.org/10.1104/pp.105.073130

Fang, Z., & Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics*, *12*(3), 280–287. https://doi.org/10.1093/bib/bbr004

FAOstat. (2017). Retrieved from: http://www.fao.org/faostat

FASTX-Toolkit. (2015). Retrieved from http://hannonlab.cshl.edu /fastx_toolkit/

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics*, *13*(1), 4–16. https://doi.org/10.1016/j.gpb.2015.01.009

Feraru, E., Vosolsobě, S., Feraru, M. I., Petrášek, J., & Kleine-Vehn, J. (2012). Evolution and Structural Diversification of PILS Putative Auxin Carriers in Plants. *Frontiers in Plant Science*, *3*, 227. https://doi.org/10.3389/fpls.2012.00227

Feussner, I., & Wasternack, C. (2002). The lipoxygenase pathway. *Annual Review of Plant Biology*, *53*, 275–297. https://doi.org/10.1146/annurev.arplant.53.100301.135248

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279-D285. https://doi.org/10.1093/nar/gkv1344

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., … Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, *45*(D1), D190–D199. https://doi.org/10.1093/nar/gkw1107

Fisher, R.A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*(1), 39–82.

Fonseca, N. A., Marioni, J., & Brazma, A. (2014). RNA-Seq gene profiling--a systematic empirical comparison. *PloS ONE*, *9*(9), e107026. https://doi.org/10.1371/journal.pone.0107026

Foquet, M., Samiee, K. T., Kong, X., Chauduri, B. P., Lundquist, P. M., Turner, S. W., … Roitman, D. B. (2008). Improved fabrication of zero-mode waveguides for single-molecule detection. *Journal of Applied Physics, 103, 034301*. https://doi.org/10.1063/1.2831366

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Galichet, A., Hoyerová, K., Kamínek, M., & Gruissem, W. (2008). Farnesylation directs AtIPT3 subcellular localization and modulates cytokinin biosynthesis in *Arabidopsis*. *Plant Physiology*, *146*(3), 1155–1164. https://doi.org/10.1104/pp.107.107425

Galuszka, P., Frébort, I., Šebela, M., Sauer, P., Jacobsen, S., & Peč, P. (2001). Cytokinin oxidase or dehydrogenase? Mechanism of cytokinin degradation in plants. *European Journal of Biochemistry*, *268*(2), 450–461. https://doi.org/10.1046/j.1432-1033.2001.01910.x

Gan, S., & Amasino, R. M. (1995). Inhibition of leaf senescence by autoregulated production of cytokinin. *Science*, *270*(5244), 1986–1988. https://doi.org/10.1126/science.270.5244.1986

Garay-Arroyo, A., De La Paz Sánchez, M., García-Ponce, B., Azpeitia, E., & Álvarez-Buylla, E.R. (2012). Hormone symphony during root growth and development. *Developmental dynamics, 241*(12), 1867-1885. https://doi.org/10.1002/dvdy.23878

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, *8*(6), 469–477. https://doi.org/10.1038/nmeth.1613

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., … Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Gegas, V.C., Nazari, A., Griffiths, S., Simmonds, J., Fish, L., Orford, S., … Snape, J.W. (2010). A genetic framework for grain size and shape variation in wheat. *Plant Cell*, *22*(4), 1046–1056. https://doi.org/10.1105/tpc.110.074153.

Geniza, M., & Jaiswal, P. (2017). Tools for building *de novo* transcriptome assembly. *Current Plant Biology*, *11–12*, 41–45. https://doi.org/https://doi.org/10.1016/j.cpb.2017.12.004

Ghanem, M. E., Albacete, A., Smigocki, A. C., Frébort, I., Pospišilová, H., Martínez-Andújar, C., … Pérez-Alfocea, F. (2011). Root-synthesized cytokinins improve shoot growth and fruit yield in salinized tomato (*Solanum lycopersicum* L.) plants. *Journal of Experimental Botany*, *62*(1), 125–140. https://doi.org/10.1093/jxb/erq266

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86. https://doi.org/10.1186/gb-2010-11-8-r86

Góngora-Castillo, E., & Buell, C. R. (2013). Bioinformatics challenges in *de novo* transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Natural Product Reports*, *30*(4), 490–500. https://doi.org/10.1039/c3np20099j

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Goyal, E., Amit, S. K., Singh, R. S., Mahato, A. K., Chand, S., & Kanika, K. (2016). Transcriptome profiling of the salt-stress response in *Triticum aestivum* cv. Kharchia Local. *Scientific Reports*, *6*, 27752. https://doi.org/10.1038/srep27752

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., … Regev, A. (2010). *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, *28*(5), 503–510. https://doi.org/10.1038/nbt.1633

Ha, S., Vankova, R., Yamaguchi-Shinozaki, K., Shinozaki, K., & Tran, L.-S. P. (2012). Cytokinins: metabolism and function in plant adaptation to environmental stresses. *Trends in Plant Science*, *17*(3), 172–179. https://doi.org/10.1016/j.tplants.2011.12.005

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Research*, *41*(Database issue), D387-D395. https://doi.org/10.1093/nar/gks1234

Hahn, D. A., Ragland, G. J., Shoemaker, D. D., & Denlinger, D. L. (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics*, *10*, 234. https://doi.org/10.1186/1471-2164-10-234

Havlová, M., Dobrev, P. I., Motyka, V., Štorchová, H., Libus, J., Dobrá, J., … Vankova, R. (2008). The role of cytokinins in responses to water deficit in tobacco plants over-expressing *trans*-zeatin O-glucosyltransferase gene under 35S or SAG12 promoters. *Plant, Cell & Environment*, *31*(3), 341–353. https://doi.org/10.1111/j.1365-3040.2007.01766.x

He, B., Zhao, S., Chen, Y., Cao, Q., Wei, C., Cheng, X., & Zhang, Y. (2015). Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. *BMC Genomics*, *16*, 65. https://doi.org/10.1186/s12864-014-1192-7

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., … McCombie, W. R. (2007). Genome-wide *in situ* exon capture for selective resequencing. *Nature Genetics*, *39*(12), 1522–1527. https://doi.org/10.1038/ng.2007.42

Hochholdinger, F., Park, W.J., Sauer, M., & Woll, K. (2004). From weeds to crops: Genetic analysis of root development in cereals. *Trends in Plant Science*, *9*(1), 42–48. https://doi.org/10.1016/j.tplants.2003.11.003

Hong, J. K., Choi, H. W., Hwang, I. S., Kim, D. S., Kim, N. H., Choi, D. S., … Hwang, B. K. (2008). Function of a novel GDSL-type pepper lipase gene, *CaGLIP1*, in disease susceptibility and abiotic stress tolerance. *Planta*, *227*(3), 539–558. https://doi.org/10.1007/s00425-007-0637-5

Hong, G., Zhang, W., Li, H., Shen, X., & Guo, Z. (2014). Separate enrichment analysis of pathways for up- and downregulated genes. *Journal of the Royal Society Interface*, *11*(92), 20130950. https://doi.org/10.1098/rsif.2013.0950

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., … Lempicki, R. A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, *35*(Web Server issue), W169-W175. https://doi.org/10.1093/nar/gkm415

Hunter, W.N. (2007). The non-mevalonate pathway of isoprenoid precursor biosynthesis. *Journal of Biological Chemistry, 282*, 21573-21577. https://doi.org/10.1074/jbc.R700005200

IWGSC: International Wheat Genome Sequencing Consortium. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), eaar7191. https://doi.org/10.1126/science.aar7191

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, *12*(4), 351–356. https://doi.org/10.1038/nmeth.3290

Janiak, A., Kwasniewski, M., Sowa, M., Gajek, K., Zmuda, K., Kościelniak, J., & Szarejko, I. (2018). No Time to Waste: Transcriptome Study Reveals that Drought Tolerance in Barley May Be Attributed to Stressed-Like Expression Patterns that Exist before the Occurrence of Stress. *Frontiers in Plant Science*, *8*, 2212. https://doi.org/10.3389/fpls.2017.02212

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., … Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology*, *319*(5), 1257–1265. https://doi.org/10.1016/S0022-2836(02)00379-0

Jia, M., Guan, J., Zhai, Z., Geng, S., Zhang, X., Mao, L., & Li, A. (2017). Wheat functional genomics in the era of next generation sequencing: An update. *The Crop Journal, 6*(1), 7-14. https://doi.org/10.1016/j.cj.2017.09.003

Jiang, H., & Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, *25*(8), 1026–1032. https://doi.org/10.1093/bioinformatics/btp113

Jiang, Y., Chen, R., Dong, J., Xu, Z., & Gao, X. (2012). Analysis of GDSL lipase (GLIP) family genes in rice (*Oryza sativa*). *Plant Omics Journal, 5*(4), 351-358.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., … Korpelainen, E. I. (2011). Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, *12*, 507. https://doi.org/10.1186/1471-2164-12-507

Kang, X.-F., Gu, L.-Q., Cheley, S., & Bayley, H. (2005). Single protein pores containing molecular adapters at high temperatures. *Angewandte Chemie International Edition*, *44*(10), 1495–1499. https://doi.org/10.1002/anie.200461885

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009–1015. https://doi.org/10.1038/nmeth.1528

Ke, R., Mignardi, M., Hauling, T., & Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Human Mutation*, *37*(12), 1363–1367. https://doi.org/10.1002/humu.23051

Khatri, P., Draghici, S., Ostermeier, G. C., & Krawetz, S. A. (2002). Profiling gene expression using onto-express. *Genomics*, *79*(2), 266–270. https://doi.org/10.1006/geno.2002.6698

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. https://doi.org/10.1186/gb-2013-14-4-r36

Kokáš, F. Z., Bergougnoux, V., & Čudejková, M. M. (2019). SATrans: New Free Available Software for Annotation of Transcriptome and Functional Analysis of Differentially Expressed Genes. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 26*(2), 117-123. https://doi.org/10.1089/cmb.2018.0149

Kowalska, M., Galuszka, P., Frébortová, J., Šebela, M., Béres, T., Hluska, T., … Frébort, I. (2010). Vacuolar and cytosolic cytokinin dehydrogenases of *Arabidopsis thaliana*, heterologous expression, purification and properties. *Phytochemistry, 71*(17-18), 1970–1978. https://doi.org/10.1016/j.phytochem.2010.08.013

Kuhl, J. C., Cheung, F., Yuan, Q., Martin, W., Zewdie, Y., McCallum, J., … Havey, M. J. (2004). A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders *Asparagales* and *Poales*. *The Plant Cell*, *16*(1), 114–125. https://doi.org/10.1105/tpc.017202

Künstner, A., Wolf, J. B. W., Backström, N., Whitney, O., Balakrishnan, C. N., Day, L., … Ellegren, H. (2010). Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology*, *19 Suppl 1*, 266–276. https://doi.org/10.1111/j.1365-294X.2009.04487.x

Kuppu, S., Mishra, N., Hu, R., Sun, L., Zhu, X., Shen, G., … Zhang, H. (2013). Water-deficit inducible expression of a cytokinin biosynthetic gene IPT improves drought tolerance in cotton. *PloS ONE*, *8*(5), e64190. https://doi.org/10.1371/journal.pone.0064190

Łabaj, P. P., Leparc, G. G., Linggi, B. E., Markillie, L. M., Wiley, H. S., & Kreil, D. P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, *27*(13), i383-i391. https://doi.org/10.1093/bioinformatics/btr247

Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., … Lees, J. G. (2016). Gene3D: expanding the utility of domain assignments. *Nucleic Acids Research*, *44*(D1), D404-D409. https://doi.org/10.1093/nar/gkv1231

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, *8*(12), 995–1005. https://doi.org/10.1038/nrm2281

Lesk, A.M. (2013). Bioinformatics. *Encyclopaedia Britannica*. Retrieved from: https://www.britannica.com/science/bioinformatics

Letunic, I., Doerks, T., & Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research, 43*(Database issue), D257-D260. https://doi.org/10.1093/nar/gku949

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, *299*(5607), 682–686. https://doi.org/10.1126/science.1079700

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/gr.078212.108

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J.,Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

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

Li, J. J., Jiang, C.-R., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 19867–19872. https://doi.org/10.1073/pnas.1113972108

Li, H.-Z., Gao, X., Li, X.-Y., Chen, Q.-J., Dong, J., & Zhao, W.-C. (2013). Evaluation of assembly strategies using RNA-seq data associated with grain development of wheat (*Triticum aestivum* L.). *PloS ONE*, *8*(12), e83530. https://doi.org/10.1371/journal.pone.0083530

Li, J., & Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22*(5), 519–536. https://doi.org/10.1177/0962280211428386

Li, X., Zeng, R., & Liao, H. (2016). Improving crop nutrient efficiency through root architecture modifications. *Journal of Integrative Plant Biology 58*(3), 193–202. https://doi.org/10.1111/jipb.12434

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, *2012*, 251364. https://doi.org/10.1155/2012/251364

Liu, Z., Wang, Y., Deng, Tao., & Chen, Q. (2016). Solid-State Nanopore-Based DNA Sequencing Technology. *Journal of Nanomaterials*, *2016*, 5284786. https://doi.org/http://dx.doi.org/10.1155/2016/5284786

Liu, Q., Luo, L., & Zheng, L. (2018). Lignins: Biosynthesis and Biological Functions in Plants. *International Journal of Molecular Sciences*, *19*(2), 335. https://doi.org/10.3390/ijms19020335

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., … Usadel, B. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell & Environment*, *37*(5), 1250–1258. https://doi.org/10.1111/pce.12231

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, *19*(10), 1275–1283.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Ludwig-Müller, J. (2011). Auxin conjugates: their role for plant development and in the evolution of land plants. *Journal of Experimental Botany*, *62*(6), 1757–1773. https://doi.org/10.1093/jxb/erq412

Lunter, G., & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Macková, H., Hronková, M., Dobrá, J., Turečková, V., Novák, O., Lubovská, Z., … Vanková, R. (2013). Enhanced drought and heat stress tolerance of tobacco plants with ectopically enhanced cytokinin oxidase/dehydrogenase gene expression. *Journal of Experimental Botany*, *64*(10), 2805–2815. https://doi.org/10.1093/jxb/ert131

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, *21*(16), 3448–3449. https://doi.org/10.1093/bioinformatics/bti551

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., … Chinnaiyan, A. M. (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature*, *458*(7234), 97–101. https://doi.org/10.1038/nature07638

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., … Wang, Z. (2010). Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, *11*, 663. https://doi.org/10.1186/1471-2164-11-663

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*(10), 671–682. https://doi.org/10.1038/nrg3068

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … 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://doi.org/10.1101/gr.107524.110

McSteen, P. (2010). Auxin and monocot development. *Cold Spring Harbor Perspectives in Biology*, *2*(3), a001479. https://doi.org/10.1101/cshperspect.a001479

Merewitz, E. B., Gianfagna, T., & Huang, B. (2011). Photosynthesis, water use, and root viability under water stress as affected by expression of *SAG12-ipt* controlling cytokinin synthesis in *Agrostis stolonifera*. *Journal of Experimental Botany*, *62*(1), 383–395. https://doi.org/10.1093/jxb/erq285

METAVO. Metacentrum Virtual Organization. 2017. Retrieved from: https://metavo.metacentrum.cz.

Mezlini, A. M., Smith, E. J. M., Fiume, M., Buske, O., Savich, G. L., Shah, S., … Brudno, M. (2013). iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research*, *23*(3), 519–529. https://doi.org/10.1101/gr.142232.112

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, *44*(D1), D336-D342. https://doi.org/10.1093/nar/gkv1194

Mikheyev, A. S., & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, *14*(6), 1097–1102. https://doi.org/10.1111/1755-0998.12324

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Mitra, R. D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik, & Church, G. M. (2003). Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry*, *320*(1), 55–65.

Moon, J., Parry, G., & Estelle, M. (2004). The ubiquitin-proteasome pathway and plant development. *The Plant Cell*, *16*(12), 3181–3195. https://doi.org/10.1105/tpc.104.161220

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., & Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, *110*(1–2), 3–24. https://doi.org/10.1016/j.ymgme.2013.04.024

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(Web Server issue), W182-W185. https://doi.org/10.1093/nar/gkm321

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. https://doi.org/10.1038/nmeth.1226

Mrízová, K., Jiskrová, E., Vyroubalová, Š., Novák, O., Ohnoutková, L., Pospíšilová, H., … Galuszka, P. (2013). Overexpression of cytokinin dehydrogenase genes in barley (*Hordeum vulgare* cv. Golden Promise) fundamentally affects morphology and fertility. *PLoS ONE*, *8*(11), e79029. https://doi.org/10.1371/journal.pone.0079029

Nakabayashi, R., Yonekura-Sakakibara, K., Urano, K., Suzuki, M., Yamada, Y., Nishizawa, T., … Saito, K. (2014). Enhancement of oxidative and drought tolerance in *Arabidopsis* by overaccumulation of antioxidant flavonoids. *The Plant Journal: For Cell and Molecular Biology*, *77*(3), 367–379. https://doi.org/10.1111/tpj.12388

Nakano, M., Komatsu, J., Matsuura, S., Takashima, K., Katsura, S., & Mizuno, A. (2003). Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology*, *102*(2), 117–124.

NCBI Resource Coordinators. (2014). Database resources of the National Center for Biotechnology Information, *Nucleic Acids Research*, *42*(D1), D7–D17. https://doi.org/10.1093/nar/gkt1146

Nishiyama, R., Watanabe, Y., Fujita, Y., Le, D. T., Kojima, M., Werner, T., … Tran, L.-S. P. (2011). Analysis of cytokinin mutants and regulation of cytokinin metabolic genes reveals important regulatory roles of cytokinins in drought, salt and abscisic acid responses, and abscisic acid biosynthesis. *The Plant Cell*, *23*(6), 2169–2183. https://doi.org/10.1105/tpc.111.087395

Nishiyama, R., Watanabe, Y., Leyva-Gonzalez, M. A., Ha, C. V., Fujita, Y., Tanaka, M., … Tran, L.-S. P. (2013). *Arabidopsis* AHP2, AHP3, and AHP5 histidine phosphotransfer proteins function as redundant negative regulators of drought stress response. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(12), 4840–4845. https://doi.org/10.1073/pnas.1302265110

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, *40*(20), 10084–10097. https://doi.org/10.1093/nar/gks804

Nozoye, T. (2018). The *Nicotianamine Synthase* Gene Is a Useful Candidate for Improving the Nutritional Qualities and Fe-Deficiency Tolerance of Various Crops. *Frontiers in Plant Science*, *9*, 340. https://doi.org/10.3389/fpls.2018.00340

Nyrén, P. (1987). Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical Biochemistry*, *167*(2), 235–238.

Oates, M. E., Stahlhacke, J., Vavoulis, D. V, Smithers, B., Rackham, O. J. L., Sardar, A. J., … Gough, J. (2015). The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Research*, *43*(Database issue), D227-D233. https://doi.org/10.1093/nar/gku1041

O'Brien, J. A., & Benková, E. (2013). Cytokinin cross-talking during biotic and abiotic stress responses. *Frontiers in Plant Science*, *4*, 451. https://doi.org/10.3389/fpls.2013.00451

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*(1), 29–34. https://doi.org/10.1093/nar/27.1.29

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., & Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, *4*(11), 907–909. https://doi.org/10.1038/nmeth1109

Ostrowski, M. K., Świdziński, M., Ciarkowska, A., & Jakubowska, A. (2014). IAA-amido synthetase activity and *GH3* expression during development of pea seedlings. *Acta Physiologiae Plantarum*, *36*, 3029–3037. https://doi.org/10.1007/s11738-014-1673-y

Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomäki, V., … Hautaniemi, S. (2010). Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, *2*(9), 65. https://doi.org/10.1186/gm186

Özkan, H., Willcox, G., Graner, A., Salamini, F., & Kilian, B. (2011). Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*). Genetic Resources and Crop Evolution, 58(1), 11–53. https://doi.org/10.1007/s10722-010-9581-5

Pacheco-Villalobos, D., Díaz-Moreno, S. M., van der Schuren, A., Tamaki, T., Kang, Y. H., Gujas, B., … Hardtke, C. S. (2016). The Effects of high steady state auxin levels on root cell elongation in *Brachypodium*. *The Plant cell*, *28*(5), 1009–1024. https://doi.org/10.1105/tpc.15.01057

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, *11*(8), 533–538. https://doi.org/10.1038/nrg2815

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS ONE*, *7*(2), e30619. https://doi.org/10.1371/journal.pone.0030619

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*(5), 462–464. https://doi.org/10.1038/nbt.2862

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., de Castro, E., … Bridge, A. (2015). HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Research*, *43*(Database issue), D1064-D1070. https://doi.org/10.1093/nar/gku1002

Peleg, Z., & Blumwald, E. (2011). Hormone balance and abiotic stress tolerance in crop plants. *Current Opinion in Plant Biology*, *14*(3), 290–295. https://doi.org/10.1016/j.pbi.2011.02.001

Peleg, Z., Reguera, M., Tumimbang, E., Walia, H., & Blumwald, E. (2011). Cytokinin-mediated source/sink modifications improve drought tolerance and increase grain yield in rice under water-stress. *Plant Biotechnology Journal*, *9*(7), 747–758. https://doi.org/10.1111/j.1467-7652.2010.00584.x

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Philipp, E. E. R., Kraemer, L., Mountfort, D., Schilhabel, M., Schreiber, S., & Rosenstiel, P. (2012). The Transcriptome Analysis and Comparison Explorer--T-ACE: a platform-independent, graphical tool to process large RNAseq datasets of non-model organisms. *Bioinformatics*, *28*(6), 777–783. https://doi.org/10.1093/bioinformatics/bts056

Picard. (2009). Retrieved from http://picard.sourceforge.net

Porreca, G. J. (2010). Genome sequencing on nanoballs. *Nature Biotechnology*, *28*(1), 43-44. United States. https://doi.org/10.1038/nbt0110-43

Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., … Shendure, J. (2007). Multiplex amplification of large sets of human exons. *Nature Methods*, *4*(11), 931–936. https://doi.org/10.1038/nmeth1110

Pospíšilová, H., Jiskrová, E., Vojta, P., Mrízová, K., Kokáš, F., Čudejková, M. M., … Galuszka, P. (2016). Transgenic barley overexpressing a cytokinin dehydrogenase gene shows greater tolerance to drought stress. *New Biotechnology*, *33*(5 Pt B), 692–705. https://doi.org/10.1016/j.nbt.2015.12.005

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. https://doi.org/10.1186/1471-2164-13-341

R Development Core Team. (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. ISBN: 3-900051-07-0. http://www.R-project.org

Reguera, M., Peleg, Z., Abdel-Tawab, Y. M., Tumimbang, E. B., Delatorre, C. A., & Blumwald, E. (2013). Stress-induced cytokinin synthesis increases drought tolerance through the coordinated regulation of carbon and nitrogen assimilation in rice. *Plant Physiology*, *163*(4), 1609–1622. https://doi.org/10.1104/pp.113.227702

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Rhee, S.Y., Dickerson, J., & Xu, D. (2006). Bioinformatics and Its Applications in Plant Biology. *Annual Review of Plant Biology, 57*, 335-360. https://doi.org/10.1146/annurev.arplant.56.032604.144103

Rhoads, A., & Au, K.F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics & Bioinformatics, 13*(5), 278-289. https://doi.org/10.1016/j.gpb.2015.08.002

Rippert, P., & Matringe, M. (2002). Molecular and biochemical characterization of an *Arabidopsis thaliana* arogenate dehydrogenase with two highly similar and active protein domains. *Plant Molecular Biology*, *48*(4), 361–368. https://doi.org/10.1023/A:1014018926676

Rivero, R. M., Kojima, M., Gepstein, A., Sakakibara, H., Mittler, R., Gepstein, S., & Blumwald, E. (2007). Delayed leaf senescence induces extreme drought tolerance in a flowering plant. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19631–19636. https://doi.org/10.1073/pnas.0709453104

Rivero, R. M., Gimeno, J., Van Deynze, A., Walia, H., & Blumwald, E. (2010). Enhanced cytokinin synthesis in tobacco plants expressing PSARK::IPT prevents the degradation of photosynthetic protein complexes during drought. *Plant & Cell Physiology*, *51*(11), 1929–1941. https://doi.org/10.1093/pcp/pcq143

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., … Birol, I. (2010). *De novo* assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

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. https://doi.org/10.1093/bioinformatics/btp616

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, *13*, 484. https://doi.org/10.1186/1471-2164-13-484

Rodríguez-Concepción, M., & Boronat, A. (2002). Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics. *Plant Physiology, 130*(3), 1079-1089. https://doi.org/10.1104/pp.007138

Ronaghi, M., Uhlén, M., & Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science, 281*(5375), 363-365. https://doi.org/10.1126/science.281.5375.363

Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology*, *318*(2), 595–608. https://doi.org/10.1016/S0022-2836(02)00016-5

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. https://doi.org/10.1038/nature10242

Růžička, K., Ljung, K., Vanneste, S., Podhorská, R., Beeckman, T., Friml, J., & Benková, E. (2007). Ethylene regulates root growth through effects on auxin biosynthesis and transport-dependent auxin distribution. *The Plant Cell*, *19*(7), 2197–2212. https://doi.org/10.1105/tpc.107.052126

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., … Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695.

Sangrador-Vegas, A., Mitchell, A. L., Chang, H.-Y., Yong, S.-Y., & Finn, R. D. (2016). GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations. *Database: The Journal of Biological Databases and Curation*, *2016*, 1-8. https://doi.org/10.1093/database/baw027

Shao, K., Ding, W., Wang, F., Li, H., Ma, D., & Wang, H. (2011). Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection. *PloS ONE*, *6*(9), e24910. https://doi.org/10.1371/journal.pone.0024910

Sharp, R. E., & LeNoble, M. E. (2002). ABA, ethylene and the control of shoot and root growth under water stress. *Journal of Experimental Botany*, *53*(366), 33–37. https://doi.org/10.1093/jexbot/53.366.33

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., … Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728–1732. https://doi.org/10.1126/science.1117389

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227-R240. https://doi.org/10.1093/hmg/ddq416

Schmülling, T., Werner, T., Riefler, M., Krupkova, E., & Batrina y Manns, I. (2003). Structure and function of cytokinin oxidase/dehydrogenase genes of maize, rice, *Arabidopsis* and other species. *Journal of Plant Research*, *116*(3), 241–252.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., … Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(Database issue), D344-D347. https://doi.org/10.1093/nar/gks1067

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

Sinha, R., Gupta, A., & Senthil-Kumar, M. (2017). Concurrent Drought Stress and Vascular Pathogen Infection Induce Common and Distinct Transcriptomic Responses in Chickpea. *Frontiers in Plant Science*, *8*, 333. https://doi.org/10.3389/fpls.2017.00333

Smith, A. M., Heisler, L. E., St. Onge, R. P., Farias-Hesson, E., Wallace, I. M., Bodeau, J., … Nislow, C. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Research*, *38*(13), e142. https://doi.org/10.1093/nar/gkq368

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H., & Gouaux, J. E. (1996). Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science*, *274*(5294), 1859–1866.

Spannagl, M., Nussbaumer, T., Bader, K.C., Martis, M.M., Seidel, M., Kugler, K.G., … Mayer, K.F.X. (2016). PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Research*, *44*(Database issue), D1141-D1147. https://doi.org/10.1093/nar/gkv1130

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435-W439. https://doi.org/10.1093/nar/gkl200

Sun, L., Zhang, Q., Wu, J., Zhang, L., Jiao, X., Zhang, S., … Sun, Y. (2014). Two rice authentic histidine phosphotransfer proteins, OsAHP1 and OsAHP2, mediate cytokinin signaling and stress responses in rice. *Plant Physiology*, *165*(1), 335–345. https://doi.org/10.1104/pp.113.232629

Surget-Groba, Y., & Montoya-Burgos, J. I. (2010). Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research*, *20*(10), 1432–1440. https://doi.org/10.1101/gr.103846.109

Sýkorová, B., Kurešová, G., Daskalova, S., Trčková, M., Hoyerová, K., Raimanová, I., … Kamínek, M. (2008). Senescence-induced ectopic expression of the *A. tumefaciens* ipt gene in wheat delays leaf senescence, increases cytokinin content, nitrate influx, and nitrate reductase activity, but does not affect grain yield. *Journal of Experimental Botany*, *59*(2), 377–387. https://doi.org/10.1093/jxb/erm319

Talboys, P. J., Healey, J. R., Withers, P. J. A., & Jones, D. L. (2014). Phosphate depletion modulates auxin transport in *Triticum aestivum* leading to altered root branching. *Journal of Experimental Botany*, *65*(17), 5023–5032. https://doi.org/10.1093/jxb/eru284

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, *21*(12), 2213–2223. https://doi.org/10.1101/gr.124321.111

Tarazona, S., Furió-Tarí, P., Turrá, D., Pietro, A.D., Nueda, M.J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21), e140. https://doi.org/10.1093/nar/gkv711

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., … Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, *4*, 41. https://doi.org/10.1186/1471-2105-4-41

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.

Tian, W., & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, *333*(4), 863–882. https://doi.org/10.1016/j.jmb.2003.08.057

Tian, D., Tooker, J., Peiffer, M., Chung, S. H., & Felton, G. W. (2012). Role of trichomes in defense against herbivores: comparison of herbivore response to woolly and hairless trichome mutants in tomato (*Solanum lycopersicum*). *Planta*, *236*(4), 1053–1066. https://doi.org/10.1007/s00425-012-1651-9

Tran, L.-S. P., Urao, T., Qin, F., Maruyama, K., Kakimoto, T., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2007). Functional analysis of AHK1/ATHK1 and cytokinin receptor histidine kinases in response to abscisic acid, drought, and salt stress in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20623–20628. https://doi.org/10.1073/pnas.0706547105

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Tripathi, K.P., Evangelista, D., Zuccaro, A., & Guarracino, M.R. (2015). Transcriptator: An Automated Computational Pipeline to Annotate Assembled Reads and Identify Non Coding RNA. *PloS ONE* *10*(11), e0140268. https://doi.org/10.1371/journal.pone.0140268

Tsuchisaka, A., & Theologis, A. (2004). Unique and overlapping expression patterns among the *Arabidopsis* 1-amino-cyclopropane-1-carboxylate synthase gene family members. *Plant Physiology*, *136*(2), 2982–3000. https://doi.org/10.1104/pp.104.049999

Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A.-P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, *36*(4), e25. https://doi.org/10.1093/nar/gkn021

UCSC Genome Bioinformatics: Frequently Asked Questions: Data File Formats. (2006). Retrieved from https://genome.ucsc.edu/FAQ/ FAQformat.html

Ueda, J., & Kato, J. (1982). Identification of Jasmonic Acid and Abscisic Acid as Senescence-promoting Substances from *Cleyera ochnacea* DC. *Agricultural and Biological Chemistry*, *46*(7), 1975–1976. https://doi.org/10.1271/bbb1961.46.1975

Unamba, C. I. N., Nag, A., & Sharma, R. K. (2015). Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. *Frontiers in Plant Science*, *6*, 1074. https://doi.org/10.3389/fpls.2015.01074

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., & Vandepoele, K. (2013). TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. *Genome Biology*, *14*(12), R134. https://doi.org/10.1186/gb-2013-14-12-r134

Van Belleghem, S. M., Roelofs, D., Van Houdt, J., & Hendrickx, F. (2012). *De novo* transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). *PloS ONE*, *7*(8), e42605. https://doi.org/10.1371/journal.pone.0042605

Vardhini, B.V., & Anjum, N.A. (2015). Brassinosteroids make plant life easier under abiotic stresses mainly by modulating major components of antioxidant defense system. *Frontiers in Environmental Science*, *2*, 67. https://doi.org/10.3389/fenvs.2014.00067

Venkatesan, B. M., Dorvel, B., Yemenicioglu, S., Watkins, N., Petrov, I., & Bashir, R. (2009). Highly Sensitive, Mechanically Stable Nanopore Sensors for DNA Analysis. *Advanced Materials*, *21*(27), 2771-2776.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., & Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, *17*(7), 1636–1647. https://doi.org/10.1111/j.1365-294X.2008.03666.x

Vlamis, J., & Williams, D. E. (1962). Ion Competition in Manganese Uptake by Barley Plants. *Plant Physiology*, *37*(5), 650–655.

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, *13*(4), 227–232. https://doi.org/10.1038/nrg3185

Vyroubalová, S., Vaclavíková, K., Turečková, V., Novák, O., Šmehilová, M., Hluska, T., … Galuszka, P. (2009). Characterization of new maize genes putatively involved in cytokinin metabolism and their expression during osmotic stress in relation to cytokinin levels. *Plant Physiology*, *151*(1), 433–447. https://doi.org/10.1104/pp.109.142489

Wall, L., Christiansen, T., & Orwant, J. (2000). Programming Perl. *Beijing Cambridge, Mass: O'Reilly, Print.* ISBN: 0596000278.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., … Burge, C. B. (2008a). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, C., Yang, A., Yin, H., & Zhang, J. (2008b). Influence of water stress on endogenous hormone contents and cell damage of maize seedlings. *Journal of Integrative Plant Biology*, *50*(4), 427–434. https://doi.org/10.1111/j.1774-7909.2008.00638.x

Wang, Z., Gerstein, M., & Snyder, M. (2009a). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wang, L., Wang, Z., Xu, Y., Joo, S.-H., Kim, S.-K., Xue, Z., … Chong, K. (2009b). *OsGSR1* is involved in crosstalk between gibberellins and brassinosteroids in rice. *The Plant Journal : For Cell and Molecular Biology*, *57*(3), 498–510. https://doi.org/10.1111/j.1365-313X.2008.03707.x

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., … Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, *38*(18), e178. https://doi.org/10.1093/nar/gkq622

Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, *28*(16), 2184–2185. https://doi.org/10.1093/bioinformatics/bts356

Wang, Y., Yang, Q., & Wang, Z. (2015). The evolution of nanopore sequencing. *Frontiers in Genetics*, *5*, 449. https://doi.org/10.3389/fgene.2014.00449

Wang, X., Cai, X., Xu, C., Wang, Q., & Dai, S. (2016). Drought-Responsive Mechanisms in Plant Leaves Revealed by Proteomics. *International Journal of Molecular Sciences*, *17*(10). https://doi.org/10.3390/ijms17101706

Wasternack, C. (2014). Action of jasmonates in plant stress responses and development--applied aspects. *Biotechnology Advances*, *32*(1), 31–39. https://doi.org/10.1016/j.biotechadv.2013.09.009

Wei, W., Alexandersson, E., Golldack, D., Miller, A. J., Kjellbom, P. O., & Fricke, W. (2007). HvPIP1;6, a barley (*Hordeum vulgare* L.) plasma membrane water channel particularly expressed in growing compared with non-growing leaf tissues. *Plant & Cell Physiology*, *48*(8), 1132–1147. https://doi.org/10.1093/pcp/pcm083

Werner T., Motyka, V., Strnad, M., & Schmülling, T. (2001). Regulation of plant growth by cytokinin. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(18), 10487–10492.

Werner, T., Motyka, V., Laucou, V., Smets, R., Van Onckelen, H., & Schmülling, T. (2003). Cytokinin-deficient transgenic *Arabidopsis* plants show multiple developmental alterations indicating opposite functions of cytokinins in the regulation of shoot and root meristem activity. *Plant Cell*, *15*(11), 2532–2550. https://doi.org/10.1105/tpc.014928

Werner, T., Nehnevajova, E., Köllmer, I., Novák, O., Strnad, M., Krämer, U., & Schmülling, T. (2010). Root-specific reduction of cytokinin causes enhanced root growth, drought tolerance, and leaf mineral enrichment in *Arabidopsis* and tobacco. *The Plant Cell*, *22*(12), 3905–3920. https://doi.org/10.1105/tpc.109.072694

Wheeler, D.L., Smith-White, B., Chetvernin, V., Resenchuk, S., Dombrowski, S.M., Pechous, S.W., … Ostell, J. (2005). Plant genome resources at the national center for biotechnology information. *Plant Physiology*, *138*, 1280–1288. https://dx.doi.org/10.1104%2Fpp.104.058842

Widenius, M., & Axmark, D. (2002). MySQL reference manual: documentation from the source. *Beijing Farnham: O'Reilly Community Press. Print.* ISBN: 978-0596002657.

Wikimedia Foundation. (2010). DNA nanoball sequencing. *Academic Dictionaries and Encyclopedias*. Retrieved from: http://enacademic.com/dic.nsf/enwiki/11602930

Wolf, J. B. W. (2013). Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular Ecology Resources*, *13*(4), 559–572. https://doi.org/10.1111/1755-0998.12109

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873–881. https://doi.org/10.1093/bioinformatics/btq057

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., … Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics*, *14*, 33. https://doi.org/10.1186/1471-2105-14-33

Yang, I. S., & Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics & Informatics*, *13*(4), 119–125. https://doi.org/10.5808/GI.2015.13.4.119

Zahavi, T., Stelzer, G., Strauss, L., Salmon, A. Y., & Salmon-Divon, M. (2015). VennBLAST-whole transcriptome comparison and visualization tool. *Genomics*, *105*(3), 131–136. https://doi.org/10.1016/j.ygeno.2014.12.004

Zalabák, D., Pospíšilová, H., Šmehilová, M., Mrízová, K., Frébort, I., & Galuszka, P. (2013). Genetic engineering of cytokinin metabolism: prospective way to improve agricultural traits of crop plants. *Biotechnology Advances*, *31*(1), 97–117. https://doi.org/10.1016/j.biotechadv.2011.12.003

Zhang, Z.-X., & Zheng, Y.-Z. (2008). Overexpression of Nicotianamine Synthase (NAS) Gene Results in Enhanced Drought Tolerance in Perennial Ryegrass. *Biotechnology & Biotechnological Equipment*, *22*(4), 938–941. https://doi.org/10.1080/13102818.2008.10817583

Zhang, S.-W., Li, C.-H., Cao, J., Zhang, Y.-C., Zhang, S.-Q., Xia, Y.-F., … Sun, Y. (2009). Altered architecture and enhanced drought tolerance in rice via the down-regulation of indole-3-acetic acid by TLD1/OsGH3.13 activation. *Plant Physiology*, *151*(4), 1889–1901. https://doi.org/10.1104/pp.109.146803

Zhang, P., Wang, W.-Q., Zhang, G.-L., Kaminek, M., Dobrev, P., Xu, J., & Gruissem, W. (2010). Senescence-inducible expression of isopentenyl transferase extends leaf life, increases drought stress resistance and alters cytokinin metabolism in cassava. *Journal of Integrative Plant Biology*, *52*(7), 653–669. https://doi.org/10.1111/j.1744-7909.2010.00956.x

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., & Hao, P. (2011). Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, *12 Suppl 1*, S2. https://doi.org/10.1186/1471-2105-12-S14-S2

Zhao, Y., Tang, H., & Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, *28*(1), 125–126. https://doi.org/10.1093/bioinformatics/btr595

Zhao, J., Bodner, G., Rewald, B., Leitner, D., Nagel, K.A., & Nakhforoosh, A. (2017). Root architecture simulation improves the inference from seedling root phenotyping towards mature root systems. *Journal of Experimental Botany*, *68*(5), 965–982. https://doi.org/10.1093/jxb/erw494

Zhou, M.-L., Qi, L.-P., Pang, J.-F., Zhang, Q., Lei, Z., Tang, Y.-X., … Wu, Y.-M. (2013). Nicotianamine synthase gene family as central components in heavy metal and phytohormone response in maize. *Functional & Integrative Genomics*, *13*(2), 229–239. https://doi.org/10.1007/s10142-013-0315-6

Zubo, Y. O., Yamburenko, M. V., Selivankina, S. Y., Shakirova, F. M., Avalbaev, A. M., Kudryakova, N. V., … Börner, T. (2008). Cytokinin stimulates chloroplast transcription in detached barley leaves. *Plant Physiology*, *148*(2), 1082–1093. https://doi.org/10.1104/pp.108.122275

SUPPLEMENTAL DATA

List of the electronic supplementary materials

Supplementary material 1: SATrans software – installation zip archive.

Supplementary material 2: Detailed results from annotation sequences obtained from *Claviceps purpurea* (from ENSEMBL; Cunningham *et al.,* 2015) provided via SATrans software in "Create" mode and exported in "Export" mode. Files which are considered in the material are "Sequence_export.fa", "Hits_export.txt", "Annotation_export.txt", "GO_analysis_export.txt", "GO_parse_export.txt", "InterProScan_data_export.txt" and "stat.txt". Format of the files is fully described in the manual of SATrans which is consider in the Supplementary material 1.

Supplementary material 3: Detailed results from "analysis" mode for significantly down/upregulated genes found between mycelium and sclerotia of the strain Gal404 of *Claviceps purpurea* at the adjusted *p*-value < 0.01. The genes, whose expression differed less than 4 times (log2FoldChange < -2 and > 2) were filtered out. Files which are considered in the material are "out_annotation.txt", "out_GO_analysis_all.txt", "out_GO_analysis_downregulated.txt", "stat.txt", "out_GO_analysis_upreg-ulated.txt", "out_Histogram_downregulated.txt" and "out_Histogram_upregulated.txt". Format of the files is fully described in the manual of SATrans which is consider in the Supplementary material 1.

List of the electronic supplemental tables

Supplemental table 1: Comparative transcriptomics of *vAtCKX1* leaves from plants grown in soil under optimal conditions, during drought stress, 12 h after re-watering, and after 14-day revitalization.

Supplemental table 2: Comparative transcriptomics of *vAtCKX1* and *cAtCKX1* roots during drought stress and, after 14-day revitalization. Average expression level (baseMean) and change due to optimal conditions (log2FoldChange) with statistical significance (padj) are presented.

Supplemental table 3: Changes in expression of genes implicated in auxin response, translocation, and deactivation in *vAtCKX1* and *cAtCKX1* roots during 14-day revitalization.

Supplemental table 4: Comparison of significantly up- and down-regulated genes between *ahk2/ahk3 Arabidopsis* double knock-out (Tran *et al.,* 2007) and *vAtCKX1* and *cAtCKX1* overexpressors.

Supplemental table 5: Comparative transcriptomics of *Hordeum vulgare* roots during the drought stress. Average expression level (baseMean) and change due to optimal conditions (log2FoldChange) with statistical significance (padj) are presented.

Supplemental table 6: GO analysis at the level 6 of differentially expressed genes (adjusted *p*-value ≤ 0.05) in *Hordeum vulgare* roots during the drought stress. Data are

categorized according to Biological Processes (BP). Molecular Function (MF) and sub-Cellular Component (CC) and sorted by percentage of affected genes. Total, Number of sequences which are annotated by GO term in whole transcriptome; %, percentage share of #Seqs.

Supplemental table 7: Comparative transcriptomics of *Hordeum vulgare* aerial part during the drought and 12 h after re-watering. Average expression level (baseMean) and change due to optimal conditions (log2FoldChange) with statistical significance (padj) are presented.

Supplemental table 8: GO analysis at the level 6 of differentially expressed genes (adjusted *p*-value ≤ 0.05) in *Hordeum vulgare* aerial part during the drought and 12 h after re-watering. Data are categorized according to Biological Processes (BP). Molecular Function (MF) and sub-Cellular Component (CC) and sorted by percentage of affected genes. Total, Number of sequences which are annotated by GO term in whole transcriptome; %, percentage share of #Seqs.

Supplemental table 9: Tables of significantly down/upregulated genes found between mycelium and sclerotia of the strain Gal404 at the adjusted *p*-value < 0.01. The genes, whose expression differed less than 4 times (log2FoldChange < -2 and > 2) were filtered out. Annotation 1 is original annotation enriched by the protein blast of the reference proteome against the database of all fungal proteins downloaded from UniProtKB. Annotation 2 is enriched annotation acquired by blasting reference proteome against the database of all fungal amino acid associated proteins and all fungal transcriptional factors and regulators downloaded from UniProtKB. COG category was assigned by blasting reference proteome against COG – Clusters of Orthologous Groups database. The first best blast hit with e-value < 10^{-3} was accepted (from Čudejková *et al.*, 2016).

Supplemental table 10: Table of differentially expressed genes extracted from Supplemental table 9 and converted into input format for SATrans (Kokáš *et al.*, 2019). log2FoldChange, change of expression between mycelium and sclerotia.; *p*-value and *p*-adjusted, values which can be obtained from DESeq2. In this case were set to 0 because the gene are already filtered in Supplemental table 9.

Supplemental table 11: Numbers of uniquely aligned reads to the exon regions (Unique), reads that do not overlap any exon (Unassigned_NoFeatures), reads which overlap two or more genes (Unassigned_ambiguity). The numbers were observed before StringTie and after launch of StringTie.

Supplemental table 12: Gene ontology annotation at the level of 3 (GO_Level) for whole wheat transcriptome which was obtained by combined approach. BP, Biological processes; MF, Molecular function; CC, sub-Cellular component.

Supplemental table 13: Comparative transcriptomics of roots between wheat genotypes with short root system (W527, W533) and long root system (W501, W509). Data were obtained from DESeq2 and annotation from "nt" database (column Blast_hit and AC_number; NCBI Resource Coordinators, 2014) and InterPro (column IPR_annotation, GO_description and GO_IDs; Finn *et al.*, 2017) were added.

Supplemental table 14: GO analysis provided via SATrans (Kokáš *et al.*, 2019) of down regulated differentially expressed genes (adjusted *p*-value ≤ 0.01) in wheat genotypes

with short root system (W527, W533) in compare to genotypes with long root system (W501, W509). Data are categorized according to Biological Processes (BP). Molecular Function (MF) and sub-Cellular Component (CC) and sorted GO level (column GO_level).

Supplemental table 15: GO analysis provided via SATrans (Kokáš *et al.*, 2019) of up regulated differentially expressed genes (adjusted *p*-value ≤ 0.01) in wheat genotypes with short root system (W527, W533) in compare to genotypes with long root system (W501, W509). Data are categorized according to Biological Processes (BP). Molecular Function (MF) and sub-Cellular Component (CC) and sorted GO level (column GO_level).

List of the supplemental materials

Supplemental material 1: Research article "Whole transcriptome analysis of transgenic barley with altered cytokinin homeostasis and increased tolerance to drought stress.".

Supplemental material 2: Research article "Dataset for transcriptional response of barley (*Hordeum vulgare* L.) exposed to drought and subsequent re-watering.".

Supplemental material 3: Research article "SATrans: New free available software for annotation of transcriptome and functional analysis of differentially expressed genes.".