PALACKÝ UNIVERSITY OLOMOUC

Faculty of Science

Department of Biochemistry

Summary of the Ph.D. thesis

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

Mgr. Filip Zavadil Kokáš

P1416 Biochemistry

Olomouc 2019

This Ph.D. thesis has been completed within the framework of the Ph.D. study program Biochemie P1416 guaranteed by the Department of Biochemistry, Faculty of Science, Palacký University Olomouc in the period September 2014 – November 2018.

Aspirant: **Mgr. Filip Zavadil Kokáš**
Department of Molecular biology, Centre of the Region Hana for Biotechnological and Agriculture Research, Faculty of Science, Palacký University Olomouc

Supervisor: **Dr. Véronique Bergougnoux, Ph.D.**
Department of Molecular biology, Centre of the Region Hana for Biotechnological and Agriculture Research, Faculty of Science, Palacký University Olomouc

This thesis summary has been sent out on ………

The oral defense will take place on ………………… at the Faculty of Science, Šlechtitelů 27, Olomouc.

The Ph.D. thesis is available in the Library of the Biological Centre, Šlechtitelů 27, Olomouc.

CONTENT

1.0 Summary

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. Down-stream 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.

2.0 Souhrn

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.

3.0 Aims of the work

1. Study of the transcriptome of transgenic barley (*Hordeum vulgare* L. cv. Golden Promise) with altered cytokinin content exposed to drought stress
2. Development of a robust software to provide reliable functional annotation of novel sequences and gene ontology analysis
3. Compare different assembly strategies to obtain the best reference transcriptome of the hexaploid bread wheat
4. Study of the transcriptome of four bread wheat (*Triticum aestivum* L.) inbred lines with various root systems

4.0. Transcriptomic analysis of different wheat inbred lines with different root system

4.1 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).

Recently, several barley transgenic lines overexpressing the *cytokinin dehydrogenase 1* (*CKX1*) gene from *Arabidopsis*, an enzyme of cytokinin catabolism, targeted to various subcellular compartments were prepared. 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.

4.2. Methods

The transcriptome analysis of the upper, vegetative part of barley was done on plants cultivated in shallow trays 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.1).

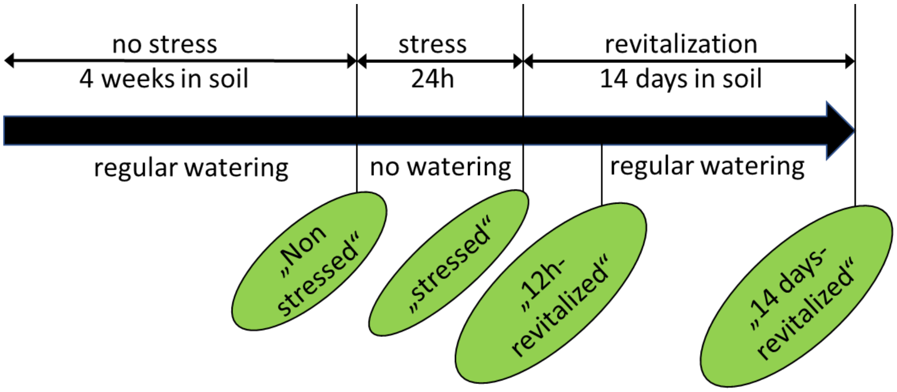

Figure 1: Experimental design used to study the root shoot transcriptome of barley plant grown under drought stress.

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 (Kim *et al.*, 2013) and quantified by using HTSeq v.0.6.0 (Anders *et al.*, 2015). The tests for differential gene expression were performed using the DESeq2 package (Love *et al.*, 2014). GO annotation of the reference genome was improved using the Blast2GO (v.3.0) program (Conesa *et al.*, 2005).

4.3. Results

4.3.1. 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.

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). 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 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.

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.

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).

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

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.

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). 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.

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.

5.0 SATrans: a tool design for fast functional annotation of RNA-seq data sets

5.1 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.*, 2009). 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.

We present here SATrans, 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. 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.

5.2 Description of the software

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 divided into three Perl modules: blast.pm, ipr_scan.pm, and datab.pm. 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 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. 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. Functions stored in the datab.pm module enable communication between the SATrans software and the MySQL database. 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 2.

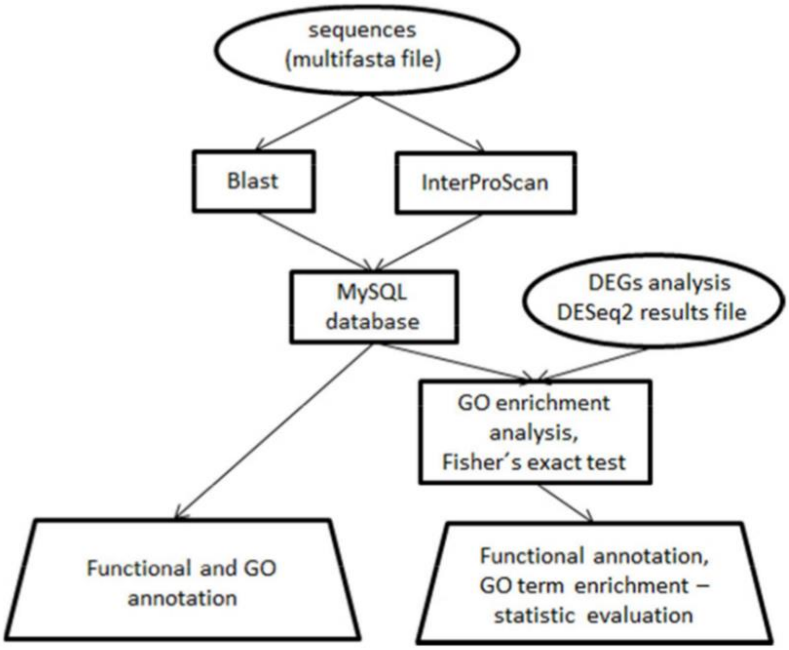

Figure 2: Simplified scheme of SATrans annotation and analysis process. Ellipses, input files; rectangles, data storage and analysis processes; trapezium, output files; DEGs, differentially expressed genes; GO, Gene Ontology.

5.3 Results

We have compared the features of different annotation tools: two desktop tools (SATrans and Blast2GO) and two web-based tools (TRAPID and MERCATOR). 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 (Cunningham *et al.*, 2015). The tools were launched with default parameters and the results are summarized in Table 1 and Table 2. 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. 1).

Table 1: 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

When analysing runtime, SATrans needed much longer time to annotate the same number of sequences comparing to web-based tools TRAPID and MERCATOR (Tab. 1). 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 (Tab. 2).

Table 2: 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

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.

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.

6.0 Transcriptomic analysis of different wheat inbred lines with different root system

6.1 Introduction

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).

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 (W501, W509), whereas two others were characterized by short but abundant roots (W527, W533).

6.2 Methods

6.2.1 *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. 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.

6.2.2 *De novo* approach

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. 3), as well as to generate the “merged” transcriptome, containing the information of the four genotypes together (AMG sub-strategy; Fig. 3).

Subsequently, CD-HIT-EST v4.6.1 (Fu *et al.*, 2012) was used to remove the redundancy of the new assembled contigs. 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. 3).

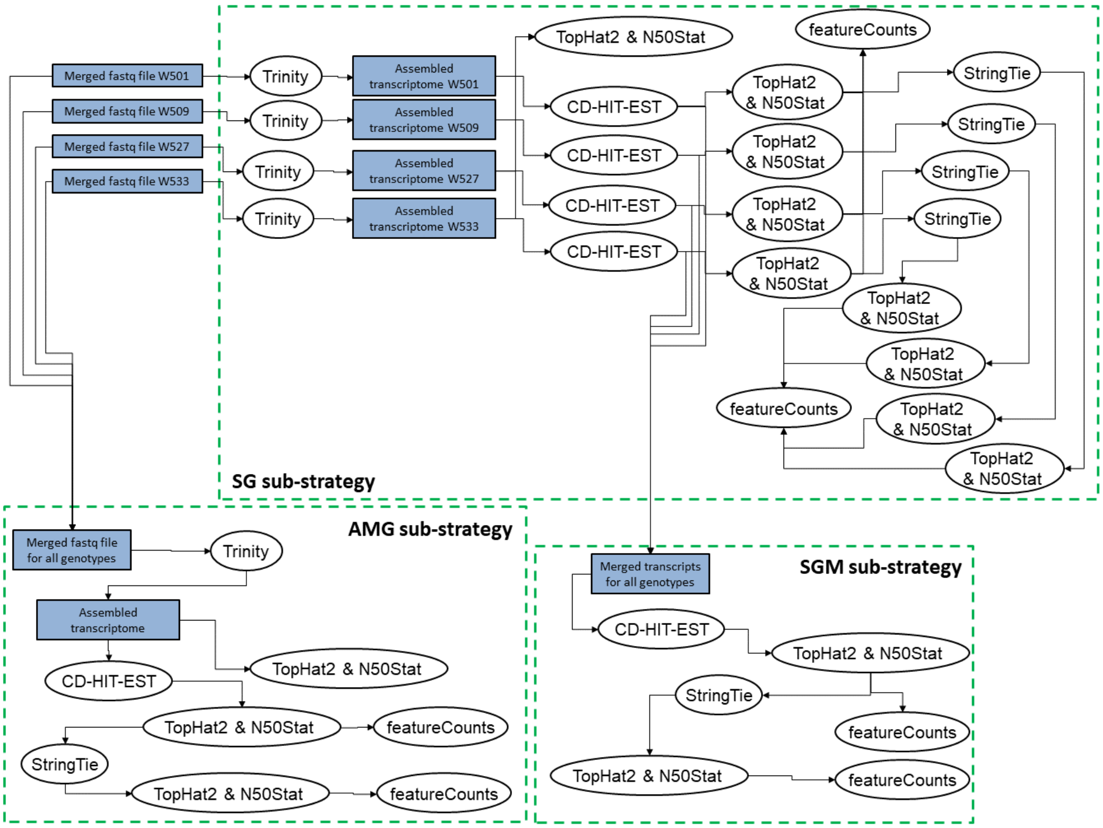

Figure 3: Diagram showing the different sub-strategies used for *de novo* assembly of the wheat reference transcriptome. The different software’s used are indicated. The rectangle indicates the data file generated after use of a specific software. For this study, four genotypes were used (W501, W509, W527 and W533).

Further redundancy removal was performed by Stringtie v1.3.1c (Pertea *et al*., 2015) for all three sub-strategies. 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.

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).

6.2.3 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). The created contigs 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).

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). The SATrans (v1.3; Kokáš *et al*., 2019) was used to functionally annotate the transcripts and for GO analysis of DEGs from the combined reference.

6.3 Results

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.

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.

6.3.1 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. 4).

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.

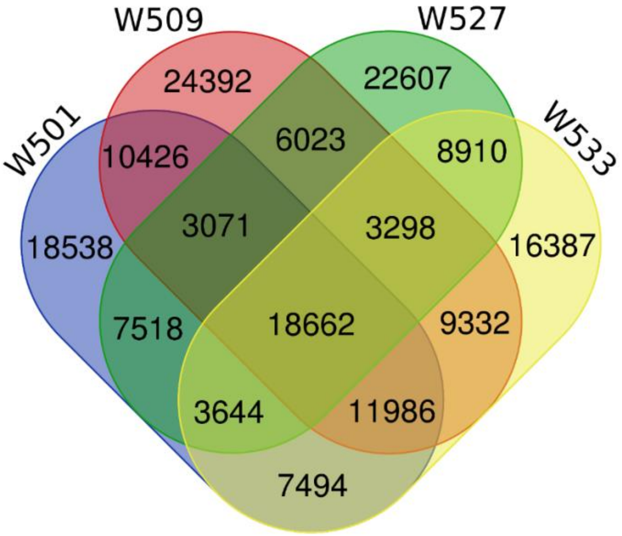

Figure 4: Venn diagram shows the similarity between reference transcriptomes obtained by *de novo* SG sub-strategy. For this study, four genotypes were used (W501, W509, W527 and W533).

To evaluate assembly redundancy, reads from all samples, were mapped to the *de novo* assemblies (Tab. 3). 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. 3). 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. 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.

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.

Table 3: 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

6.3.2 Functional annotation and analysis of differentially expressed genes between genotypes

A final comprehensive wheat root reference transcriptome was generated using a combined strategy. A total of 83 300 sequences were identified as putative genes expressed in the roots of 7-week old wheat plants grown in hydropony in controlled conditions. For functional annotation, the obtained sequences were compared against the NCBI Nucleotide database ("nt"; NCBI Resource Coordinators, 2014) and InterPro (Finn *et al.*, 2017) database using SATrans (Kokáš *et al.*, 2019).

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.

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 (BP, MF, CC). A total of 720 488 GO assignment were obtained.

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").

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 (Kokáš *et al.*, 2019).

The important BP which was affected 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. 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.

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 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).

Table 4: 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

7.0 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.

8.0 Abbreviations

ABA	abscisic acid
AMG	all merged genotypes
BLAST	basic local alignment search tool
bp	base pairs
BP	biological process
CC	sub-cellular component
CK	cytokinin

CKX	cytokinin dehydrogenase
CKX1	cytokinin dehydrogenase 1
DEG	differentially expressed gene
GH3	Gretchen Hagen 3
GO	gene ontology
GTF	gene transfer format
IAA	indole-3-acetic acid
IPT	isopentenyl transferase
JA	jasmonate
Log2FC	log2FoldChange
MF	molecular function
NCBI	National Center for Biotechnology Information
ORF	open reading frame
PILS	Pin-Likes
PIN	Pin-Formed
RNA-seq	RNA sequencing
SAUR	small auxin up RNA
SG	single genotypes
SGM	single merged genotypes
qPCR	quantitative polymerase chain reaction
WT	wild-type

9.0 References

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

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

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

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

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

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

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., 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

10.0 List of authors publications

1. Hluska T., Dobrev P.I., Tarkowská D., Frébortová J., Zalabák D., Kopečný D., Plíhal O., **Kokáš F.**, Briozzo P., Zatloukal M., Motyka V., Galuszka P. (2016) Cytokinin metabolism in maize: Novel evidence of cytokinin abundance, interconversions and formation of a new trans-zeatin metabolic product with a weak anticytokinin activity. Plant Sci. 247, 127-137; doi:10.1016/j.plantsci.2016.03.014.

2. 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

3. **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

4. Pospíšilová H., Jiskrová E., Vojta P., Mrízová K., **Kokáš F**., Majeská Čudejková M., Bergougnoux V., Plíhal O., Klimešová J., Novák O., Dzurová L., Frébort I., Galuszka P. (2016) Transgenic barley overexpressing a cytokinin dehydrogenase gene shows greater tolerance to drought stress. New Biotechnol. 33, 692-705; doi:10.1016/j.nbt.2015.12.005.

5. Frébortová J., Plíhal O., Florová V., **Kokáš F.,** Kubiasová K., Greplová M., Šimura J., Novák O., Frébort I. (2017) Light influences cytokinin biosynthesis and sensing in Nostoc (cyanobacteria). J. Phycol. 53, 703-714; doi: 10.1111/jpy.12538

6. **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