Max-Planck-Institut für Molekulare Pflanzenphysiologie

AG „Genomics and Transcript Profiling“, Gruppenleiter PD Dr. Dirk K. Hincha

Evaluation and application of omics approaches to characterize molecular responses to abiotic stresses in plants

Kumulative Dissertation

zur Erlangung des akademischen Grades

"doctor rerum naturalium" (Dr. rer. nat.)

in der Wissenschaftsdisziplin "Bioinformatik"

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät

der Universität Potsdam

von

Stephanie Schaarschmidt

Potsdam, 26. August 2020

Reviewers:

Apl. Prof. Dr. Dirk Walther
Max-Planck-Institut für Molekulare Pflanzenphysiologie
Potsdam, Germany

Prof. Dr. Richard Mott
University College London
London, Great Britian

Prof. Dr. Nils Stein
Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung
Gatersleben, Germany

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-50963
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-509630

Zusammenfassung

Aufgrund des globalen Klimawandels ist die Gewährleistung der Ernährungssicherheit für eine wachsende Weltbevölkerung eine große Herausforderung. Insbesondere abiotische Stressoren wirken sich negativ auf Ernteerträge aus. Um klimaangepasste Nutzpflanzen zu entwickeln, ist ein umfassendes Verständnis molekularer Veränderungen in der Reaktion auf unterschiedlich starke Umweltbelastungen erforderlich. Hochdurchsatz- oder "Omics"-Technologien können dazu beitragen, Schlüsselregulatoren und Wege abiotischer Stressreaktionen zu identifizieren. Zusätzlich zur Gewinnung von Omics-Daten müssen auch Programme und statistische Analysen entwickelt und evaluiert werden, um zuverlässige biologische Ergebnisse zu erhalten.

Ich habe diese Problemstellung in drei verschiedenen Studien behandelt und dafür zwei Omics-Technologien benutzt. In der ersten Studie wurden Transkript-Daten von den beiden polymorphen *Arabidopsis thaliana* Akzessionen Col-0 und N14 verwendet, um sieben Programme hinsichtlich ihrer Fähigkeit zur Positionierung und Quantifizierung von Illumina RNA Sequenz-Fragmenten („Reads“) zu evaluieren. Zwischen 92% und 99% der Reads konnten an die Referenzsequenz positioniert werden und die ermittelten Verteilungen waren hoch korreliert für alle Programme. Bei der Durchführung einer differentiellen Genexpressionsanalyse zwischen Pflanzen, die bei 20 °C oder 4 °C (Kälteakklimatisierung) exponiert wurden, ergab sich eine große paarweise Überlappung zwischen den Programmen. In der zweiten Studie habe ich die Transkriptome von zehn verschiedenen *Oryza sativa* (Reis) Kultivaren sequenziert. Dafür wurde die PacBio Isoform Sequenzierungstechnologie benutzt. Die *de novo* Referenztranskriptome hatten zwischen 38.900 bis 54.500 hoch qualitative Isoformen pro Sorte. Die Isoformen wurden kollabiert, um die Sequenzredundanz zu verringern und danach evaluiert z.B. hinsichtlich des Vollständigkeitsgrades (BUSCO), der Transkriptlänge und der Anzahl einzigartiger Transkripte pro Genloci. Für die hitze- und trockenheitstolerante Sorte N22 wurden ca. 650 einzigartige und neue Transkripte identifiziert, von denen 56 signifikant unterschiedlich in sich entwickelnden Samen unter kombiniertem Trocken- und Hitzestress exprimiert wurden. In der letzten Studie habe ich die Veränderungen in Metabolitprofilen von acht Reissorten gemessen und analysiert, die dem Stress hoher Nachttemperaturen (HNT) ausgesetzt waren und während der Trocken- und Regenzeit im Feld auf den Philippinen angebaut wurden. Es wurden jahreszeitlich bedingte Veränderungen im Metabolitspiegel sowie für agronomische Parameter identifiziert und mögliche Stoffwechselwege, die einen Ertragsrückgang unter HNT-Bedingungen verursachen, vorgeschlagen.

Zusammenfassend konnte ich zeigen, dass der Vergleich der RNA-seq Programme den Pflanzenwissenschaftler*innen helfen kann, sich für das richtige Werkzeug für ihre Daten zu entscheiden. Die *de novo* Transkriptom-Rekonstruktion von Reissorten ohne Genomsequenz bietet einen gezielten, kosteneffizienten Ansatz zur Identifizierung neuer Gene, die durch verschiedene Stressbedingungen reguliert werden unabhängig vom Organismus. Mit dem Metabolomik-Ansatz für HNT-Stress in Reis habe ich stress- und jahreszeitenspezifische Metabolite identifiziert, die in Zukunft als molekulare Marker für die Verbesserung von Nutzpflanzen verwendet werden könnten.

Summary

Due to global climate change providing food security for an increasing world population is a big challenge. Especially abiotic stressors have a strong negative effect on crop yield. To develop climate-adapted crops a comprehensive understanding of molecular alterations in the response of varying levels of environmental stresses is required. High throughput or 'omics' technologies can help to identify key-regulators and pathways of abiotic stress responses. In addition to obtain omics data also tools and statistical analyses need to be designed and evaluated to get reliable biological results.

To address these issues, I have conducted three different studies covering two omics technologies. In the first study, I used transcriptomic data from the two polymorphic *Arabidopsis thaliana* accessions, namely Col-0 and N14, to evaluate seven computational tools for their ability to map and quantify Illumina single-end reads. Between 92% and 99% of the reads were mapped against the reference sequence. The raw count distributions obtained from the different tools were highly correlated. Performing a differential gene expression analysis between plants exposed to 20 °C or 4°C (cold acclimation), a large pairwise overlap between the mappers was obtained. In the second study, I obtained transcript data from ten different *Oryza sativa* (rice) cultivars by PacBio Isoform sequencing that can capture full-length transcripts. *De novo* reference transcriptomes were reconstructed resulting in 38,900 to 54,500 high-quality isoforms per cultivar. Isoforms were collapsed to reduce sequence redundancy and evaluated, e.g. for protein completeness level (BUSCO), transcript length, and number of unique transcripts per gene loci. For the heat and drought tolerant *aus* cultivar N22, I identified around 650 unique and novel transcripts of which 56 were significantly differentially expressed in developing seeds during combined drought and heat stress. In the last study, I measured and analyzed the changes in metabolite profiles of eight rice cultivars exposed to high night temperature (HNT) stress and grown during the dry and wet season on the field in the Philippines. Season-specific changes in metabolite levels, as well as for agronomic parameters, were identified and metabolic pathways causing a yield decline at HNT conditions suggested.

In conclusion, the comparison of mapper performances can help plant scientists to decide on the right tool for their data. The *de novo* reconstruction of rice cultivars without a genome sequence provides a targeted, cost-efficient approach to identify novel genes responding to stress conditions for any organism. With the metabolomics approach for HNT stress in rice, I identified stress and season-specific metabolites which might be used as molecular markers for crop improvement in the future.

Table of contents

List of figures

List of tables

Abbreviations

AlaAT	Alanine aminotransferase
bp	base pairs
BUSCO	Benchmarking universal single-copy orthologs
cDNA	Complementary deoxyribonucleic acid
Col-0	Columbia-0
DGE	Differential gene expression
DS	Dry season
HNT	High-night temperature
LRS	Long-read sequencing
Mb	Mega bases
mGWAS	Metabolome-based genome-wide association study
mQTL	Metabolome quantitative trait loci
PacBio IsoSeq	Pacific Bioscience isoform sequencing
qRT-PCR	quantitative Reverse-Transcription PCR
RNA-seq	Ribonucleic acid sequencing
SMRT	Single-molecule real-time
WS	Wet season
ZMW	Zero-mode waveguide

1 Introduction

1.1 Analyzing the transcriptome to understand genomic functions

With the development of high-throughput or 'omics' technologies, biological research has been revolutionized. Until today, these technologies are constantly evolving for different molecular fields such as genomics, transcriptomics, metabolomics, or proteomics. Genomics appeared as the first omics discipline, focusing on the analysis of entire genomes (Hasin et al. 2017). Already in the early 2000s, it was successfully applied to sequence the first complete human genome (Collins et al. 2003) and provided a framework for mapping and studying specific genetic variants contributing to both Mendelian and complex diseases (Hasin et al. 2017). At this time, also the first plant genomes were sequenced, such as those of *Arabidopsis thaliana* (The Arabidopsis Genome 2000) or *Oryza sativa* (Goff et al. 2002), revealing new gene families, gene losses and duplications.

To understand genomic functions, the analysis of the transcriptome plays an essential role. The transcriptome is defined as the complete set of transcripts in a cell, tissue, or organism, and it can be quantified at specific developmental stages or under different physiological conditions (Vailati-Riboni et al. 2017). Next to differential gene expression (DGE) analysis, an approach to quantify changing gene expression levels under different conditions, key aspects of transcriptomics include cataloging all species of transcripts of a cell, tissue or organism and to determine the transcriptional structure of genes, splicing patterns and other post-transcriptional modifications (Wang et al. 2009). Several technologies have been developed to study these aspects including hybridization- or sequence-based approaches such as microarrays or quantitative Reverse-Transcription PCR (qRT-PCR), and next-generation sequencing methods broadly termed RNA sequencing (RNA-seq). Hybridization-based methods typically involve fluorescently labeled complementary DNA (cDNA) to probe custom-made microarrays that carry cDNAs or oligonucleotides corresponding to target genes. They have a high throughput (Jaluria et al. 2007) but are limited to existing knowledge about the genome sequence. Also, they are prone to technical problems, such as high background noise, cross-hybridization among closely related genes, and the difficulty to compare expression levels across different experiments (Okoniewski & Miller 2006). In contrast to microarray-methods, sequencing-based technologies determine the cDNA sequences directly. While qRT-PCR is highly accurate, it only allows the parallel analysis of a limited number of genes (Nolan et al. 2006), RNA-seq provides an unbiased study of all transcripts at the same time.

RNA-seq allows the analysis of the whole transcriptome without knowledge of the genome sequence (Stark et al. 2019). It normally starts with a population of RNA which is converted into a library of cDNA fragments with adapters attached to one or both ends. Those fragments are amplified by PCR and then sequenced from one (single-end) or both ends (paired-end) normally with a read depth of 10-30 million reads per sample on a high-throughput platform such as provided by Illumina. Depending on the sequencing technology these fragments are between 30 and 400 base pairs (bp) long (Wang et al. 2009).

Based on these sequences, different follow-up analyses can be performed such as DGE analysis, which is still one of the primary applications of RNA-seq. To perform a DGE analysis after sequencing, the final steps are computational. It begins with an alignment or an assembling of the reads to a transcriptome or genome sequence, quantifying the reads that overlap the reference sequence, filtering, and normalization between the samples and a statistical approach to identify significant changes in the expression levels of individual genes and/or between sample groups (Stark et al. 2019).

1.1.1 Computational challenges of Next Generation Sequencing

Over the years, different tools and algorithms were developed to address the alignment of reads to a reference sequence and to perform statistical modeling to obtain biologically relevant information on gene and/or transcript level. But big data also result in big challenges. On the one hand, these high-throughput technologies generate enormous amounts of data that need to be stored and for which improved computational capacities are needed for the acquisition and processing of large data files. On the other hand, implemented tools need to address computational performance to analyze the data in a reasonable amount of time and deliver statistical reliability for large datasets (Arbona et al. 2013).

One of the most computationally intensive steps for RNA-seq data processing is the alignment or mapping of the short fragmented sequences (reads) against a reference sequence. For organisms with an available genome reference, RNA-seq reads are aligned against the reference genome and converted into genomic positions. For organisms without a reference genome, a *de novo* transcriptome assembly is needed and RNA-seq reads are mapped back against this transcriptome sequence (Li & Li 2018). RNA-seq reads can be divided into two groups: reads with full-length alignments to the genome and reads that span exon-exon junctions (Dimon et al. 2010). Full-length alignment of reads against a genome sequence has mostly relied on the Burrows-Wheeler Transform algorithm yielding significant improvements in speed and accuracy for tools such as BWA (Li & Durbin 2009) or bowtie (Langmead et al. 2009). More

difficult is the alignment of reads that bridge exon-exon junctions since they by definition form intron-gapped alignments to the genome, with a very short flanking sequence (Dimon et al. 2010).

Common tools that can perform a gapped alignment of RNA-seq reads against a genome reference are for example "hierarchical indexing for spliced alignment of transcripts 2" (HISAT2) (Kim et al. 2019) or "spliced transcripts alignment to a reference" (STAR) (Dobin et al. 2013). HISAT2 can align both DNA and RNA sequences using a graph Ferragina Manzini index and utilizes not only a global index to represent the genome sequence but also small indices, which collectively cover the reference genome and its variants. This allows a search on local genomic regions, which is especially useful for RNA-seq reads spanning multiple exons, and it provides a much faster lookup due to the local index's small size (Kim et al. 2019). STAR instead is based on a sequential maximum mappable seed search in uncompressed suffix arrays followed by a seed clustering and stitching procedure (Dobin et al. 2013). However, when no reference genome is available, RNA-seq reads can be directly mapped against a transcriptome reference. In the past few years, several tools were developed for specific mapping against transcriptomic reference sequences such as kallisto (Bray et al. 2016) that is based on pseudo-alignments or salmon (Patro et al. 2017), which utilizes a quasi-mapping approach. In general, these tools do not perform the classical alignment against the reference anymore. Instead, they calculate abundances for each read with extremely fast run-times (Patro et al. 2017, Bray et al. 2016).

1.1.2 *Arabidopsis thaliana* as a model organism

Most of the computational tools and their default parameters are optimized for the human genome (Kim et al. 2019, Dobin et al. 2013, Langmead et al. 2009, Li & Durbin 2009). Nevertheless, they are also extensively used in plant research such as for *Arabidopsis thaliana* (Herranz et al. 2019, Hofmann et al. 2019, Zhang et al. 2017b). *A. thaliana,* a small plant in the Brassicaceae family, has been established as a plant model organism in the last decades for several reasons such as a small size that limited the requirement for growth facilities, or seed production through self-pollination (Koornneef & Meinke 2010). Additionally, it has a short reproductive cycle including seed germination, the formation of a rosette plant, bolting of the main stem, flowering, and maturation of the first seeds which is completed in six weeks (Gichner et al. 1995). Arabidopsis also has a broad natural distribution throughout Europe, Asia, and North America resulting in local accessions differing both genetically and phenotypically. It has been shown previously that these differences may constitute environmental adaptions such as a different freezing tolerance (Zuther et al. 2012). Arabidopsis was also the subject of

the first plant genome sequencing project (The Arabidopsis Genome 2000) because of its small genome size for a higher plant of roughly 130 megabases (Mb) distributed over five chromosomes (Michael & Jackson 2013). The many resources available for the Arabidopsis experimental system and data from the last decades can be transferred and adapted to other plant species such as tomato (Mysore et al. 2001). Moreover, research with Arabidopsis has demonstrated the important role that analysis of plant genomes can play in understanding the basic principles of biology relevant to a variety of species, including humans (Meinke et al. 1998).

1.2 The new generation of sequencing

Currently, the Illumina short-read sequencing technology is the most commonly used approach to study the transcriptome and has generated more than 95% of the published RNA-seq datasets (Stark et al. 2019) available in the Sequence Read Archive (SRA) (Leinonen et al. 2011). However, this technique is also limited especially when it comes to correctly identify and quantify multiple isoforms that are expressed from a gene. Additionally, the analysis of complex genomic loci, repetitive elements, multiple mapped reads, or variant phasing (haplotyping) is difficult to perform resulting in inefficient transcriptome characterization (Kraft & Kurth 2019). Several of these limitations can be overcome by third-generation sequencing or long-read sequencing (LRS) methods such as those developed by Pacific Biosciences (PacBio) (Rhoads & Au 2015) or Oxford Nanopore (Deamer et al. 2016).

1.2.1 Long-read sequencing (LRS) technology

Oxford Nanopore LRS utilizes a tiny protein pore that is embedded in an electrically resistant polymer membrane (Figure 1a). By setting a voltage across the membrane, an ionic current is passed through this nanopore. When DNA or RNA passes through the pore via a helicase, a characteristic change in the current occurs which provides information on the respective nucleotide in the nanopore (Deamer et al. 2016). In contrast, PacBio single-molecule real-time (SMRT) sequencing uses a single DNA polymerase that is immobilized at the bottom of a well, the so-called zero-mode waveguide (ZMW) (Figure 1b). The ZMWs are small enough to allow real-time recording of individual emitted fluorescence signals when labeled nucleotides are progressively incorporated by the polymerase during the replication process (Rhoads & Au 2015). Additionally, circular DNAs serve as a sequencing template and allow multiple sequencing rounds of the same cDNA that is later used to improve sequence accuracy by creating consensus sequences (Kraft & Kurth 2019).

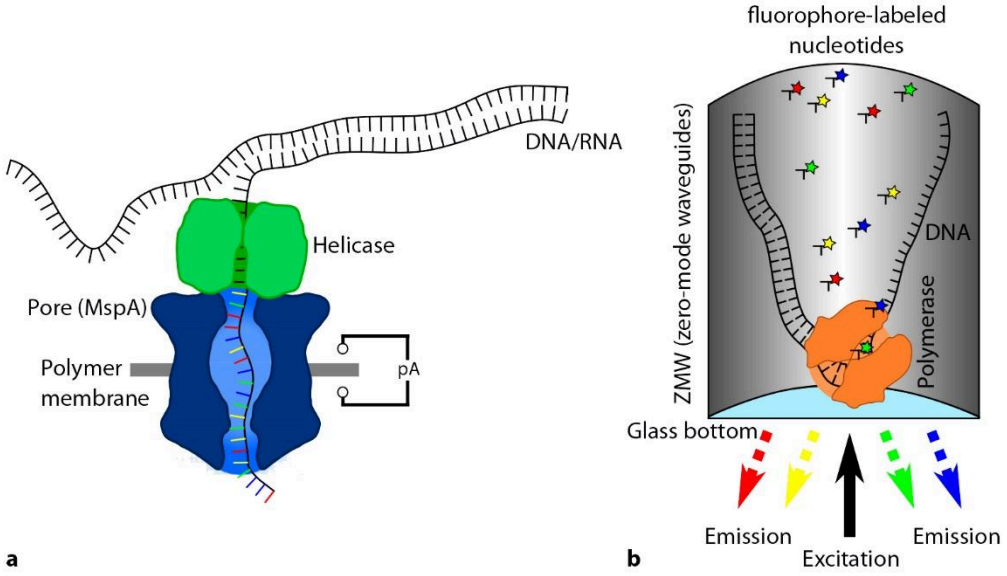


Figure 1: An overview of nanopore and single-molecule real-time (SMRT) sequencing. Oxford Nanopore sequencing **(a)** is based on a biological protein pore (e.g. *Mycobacterium smegmatis* porin A, MspA) that is fixed in a polymer membrane. The DNA is unzipped by a helicase and nucleotides inside the pore disrupt the ion flow through the channel. The changes in the current are recorded and converted to DNA/RNA sequences. Each flow cell can have between 50 (Flongle) and 3000 (PromethION) pores to sequence in parallel. For PacBio SMRT sequencing **(b)** individual molecules are loaded into a sequencing chip (SMRT cell), where they bind to a polymerase immobilized at the bottom of a zero-mode waveguide (ZMW). As each of the fluorescently labeled nucleotides is incorporated into the growing strand, the emitted light is detected and translated into a DNA/RNA sequence. Each SMRT cell can have between 150.000 (RSII) and 8 million (Sequel 2) ZMWs. Image adapted from Kraft and Kurth (2019).

LRS also has its weaknesses regarding library preparation, sequencing error rate, and bioinformatic analysis. For PacBio sequencing, there is a preference for shorter molecules due to the diffusion loading of the sample on the SMRT cell, which might negatively affect sequencing runs (Ardui et al. 2018). Additionally, the sequence template may not bind to the polymerase or may be too short to sequence, leading to a reduced overall output (Kraft & Kurth 2019). But at least loading biases can be addressed for example by using size selection to remove short molecules (Ardui et al. 2018). However, for Nanopore sequencing very large DNA molecules tend to block the pores (Kraft & Kurth 2019). Another major challenge is the high error rate. SMRT sequencing has error rates between 13-15%, where errors are distributed randomly across the reads (Rhoads & Au 2015). This randomness allows the creation of highly accurate consensus sequences by applying the circular consensus chemistry that permits to sequence the same molecule multiple times (Eid et al. 2009). Finally, to handle the long-reads new bioinformatic tools still need to be adapted and/or developed such as for alignment (Li

2018, Križanović et al. 2017, Chaisson & Tesler 2012) and assembly (Koren et al. 2017, Li et al. 2017a, Vaser et al. 2017). Many PacBio specific tools and pipelines are openly available, including demultiplexing, creating circular consensus sequences, *de novo* assemblies, or epigenetic analyses (https://www.pacb.com/support/software-downloads/, accessed on 30.06.2020).

1.2.2 Applications in biology

Besides providing full-length transcripts, LRS offers a range of different applications to study complex biological questions. In contrast to short-read sequencing (e.g. Illumina), these technologies enable unambiguous mapping of reads, for example in regions of high homology, low complexity, or in pseudogenes. Additionally, phasing of alleles (haplotypes) is possible without knowledge of parental single nucleotide polymorphisms (SNPs) and allows to distinguish whether genetic variants occur on the same allele or the opposite strand. The identification of structural variations is one of the biggest advantages, including the detection of balanced chromosomal rearrangements (Kraft & Kurth 2019).

However, these technologies also open up new opportunities in plant research and breeding. Natural plant populations have an extensive genetic variation underpinning phenotypic traits through evolutionary adaption (Henderson & Salt 2017) and human domestication (Doebley et al. 2006). Understanding these genetic polymorphisms and how different genotypes adapted to a changing environment is particularly relevant in times of global climate change and the resulting alteration in temperature, water availability, and other stressors (Henderson & Salt 2017). As a first step, high-quality reference genomes and transcriptomes are necessary. LRS is a useful tool to sequence large and complex plant genomes without the need for expensive or labor-intensive work such as sequencing overlapping BAC clones (Dong et al. 2016). Recently, several high-quality plant genomes were published such as for *Oryza sativa* (430 Mb) (Du et al. 2017), *Chenopodium quinoa* (1500 Mb) (Jarvis et al. 2017), *Triticum aestivum* (1500 Mb) (Zimin et al. 2017), *Zea mays* (2100 Mb) (Jiao et al. 2017) or *Helianthus annuus* (3600 Mb) (Badouin et al. 2017). Combined with LRS-generated transcriptomes and RNA-seq approaches, candidate genes can be identified and introduced via transgenic approaches such as CRISPR/Cas (Zhang et al. 2019b) or introgression by crossing (Ellstrand et al. 2013) into breeding populations to generate climate-adapted cultivars.

1.2.3 Using long-read sequencing to study natural variation in rice

With the fast increase in the world's population and predicted growth to about 9 billion in 2050 (FAOSTAT 2020), there is a corresponding demand for improved food production and food

security. Global climate change is a constant and severe threat due to increased abiotic stresses which negatively affect the yield of all crops (IPCC 2014, FAO 2009). Abiotic stressors are for example UV-B, high light intensities, flooding, drought, heat, cold, or salinity (Raza et al. 2019).

For improved food security in the future, it is necessary to produce new climate-smart crop cultivars, for example for rice (*Oryza sativa*), resistant to these abiotic stressors (Wheeler & von Braun 2013). Rice is a staple crop for more than half of the world's population and studies showed a significant yield decrease due to increased temperatures (Xu et al. 2020b, Jagadish et al. 2015), drought (Yang et al. 2019, Lawas et al. 2018) or salinity (Chinnusamy et al. 2005). However, rice has a wide natural variation, and cultivars exist which have a natural stress tolerance such as cultivars from the *Oryza sativa* ssp. *aus* (Lawas et al. 2018, Jagadish et al. 2008) or *Oryza* wild species (Bierschenk et al. 2020).

In the past, most of the studies in rice were based on the *japonica* cultivar Nipponbare (Goff et al. 2002) due to the lack of proper genome assemblies from different *Oryza sativa* subspecies. For instance, the sequences obtained in the 3,000 Rice Genomes Project (RGP 2014) were mapped against the Nipponbare genome, excluding all sequences that could not be mapped to this reference. This may have led to the loss of genetic information that is specific to the non-*japonica* subspecies. To understand and study these different rice genotypes, several high-quality genomes were published using LRS technology, for example for *Oryza sativa indica* (IR8, Shuhui498, Zhenshan97, Minghui63) and *aus* subspecies (N22) (Stein et al. 2018, Du et al. 2017, Zhang et al. 2016), although the degree of completeness and annotation remains variable. Nevertheless, these new genome sequences can be exploited as a useful genetic resource to generate climate-adapted rice cultivars.

1.3 Metabolomics as a direct connection between genotype and phenotype

Various molecular elements such as DNA, RNA, proteins, and metabolites, as well as environmental factors, define the phenotype of an organism. While gene and protein expression describe the potential of plants to respond to different conditions and environments, metabolites represent the integration of these regulatory aspects with the environment (Arbona et al. 2013). Metabolites constitute the biological endpoint of metabolism, considering the time difference and regulatory processes that occur between gene expression and its physiological manifestation and are therefore bridging the gap between phenotype and genotype (Hall 2006). The quantitative and qualitative profiling of metabolites in an organism is called 'metabolomics' (Dunn et al. 2013) and has emerged as a widely adopted technology in plant

science (Kumar et al. 2017). Over the past years, plant metabolomics has developed steadily (Nakabayashi & Saito 2020, Sawada et al. 2008, Lisec et al. 2006) and revealed a better understanding of the plant metabolome including crop species under abiotic stresses (Matich et al. 2019, Dawid & Hille 2018, Arbona et al. 2013, Krasensky & Jonak 2012, Obata & Fernie 2012). Additionally, integrated 'omics' data-centered on metabolomics revealed novel pathways associated with stress tolerance (Perez de Souza et al. 2019, Nakabayashi & Saito 2015) and were applied for the dissection of complex traits in plants (Fang & Luo 2019, Chen et al. 2016, Luo 2015). Furthermore, metabolomics can be used to predict metabolite markers for stress tolerance in plants that can be used for marker-assisted selection in breeding programs (Lawas et al. 2019, Sprenger et al. 2018, Degenkolbe et al. 2013).

In rice, different metabolomic studies were performed to assess the impact of different abiotic stresses such as drought (Casartelli et al. 2018, Degenkolbe et al. 2013), heat (Yamakawa & Hakata 2010), high night temperature (Schaarschmidt et al. 2020, Glaubitz et al. 2017, Glaubitz et al. 2015) or combined heat and drought (Lawas et al. 2019, Li et al. 2015).

1.3.1 High night temperature stress influences the yield of rice

During the last decades, the global surface temperature has increased by an average of 0.38°C, and until 2100 an increase of 3.7°C has been predicted (IPCC 2014). The temperature increase develops asymmetrically, causing a reduction in the diurnal temperature range which is defined as the difference in daily maximum and minimum temperature (Davy et al. 2017, Vose et al. 2005, Easterling et al. 1997) that leads to 'high night temperatures' (HNT). Crop species like rice are negatively affected by HNT such as shown by reduced yield and grain quality, including increased chalk formation or an altered grain growth dynamic (Shi et al. 2017, Shi et al. 2013, Mohammed & Tarpley 2011, Nagarajan et al. 2010, Peng et al. 2004). Studies have also shown a possible natural variation in HNT tolerance among various rice cultivars based on grain yield, yield-related parameters (Bahuguna et al. 2017, Shi et al. 2017, Shi et al. 2013, Zhang et al. 2013) or phenotypes in the vegetative stage (Glaubitz et al. 2014). Additionally, higher rates of respiration in leaves and panicles were reported. Photosynthetic activity was not affected or decreased, and a reduction of nitrogen and carbohydrate translocation after flowering was observed, with negative effects on grain yield, especially in sensitive cultivars (Bahuguna et al. 2017, Glaubitz et al. 2014, Liang et al. 2013, Mohammed et al. 2013, Shi et al. 2013).

1.3.2 Molecular knowledge of high night temperature stress in rice

However, despite the increasing knowledge of the physiological and agronomic responses to HNT, only little is known of the molecular responses of rice (Table 1). In summary, these

studies revealed tissue-specific effects especially on the tricarboxylic acid (TCA) cycle, amino acid biosynthesis, and starch metabolism in rice upon HNT stress (Dhatt et al. 2019, Glaubitz et al. 2017, Glaubitz et al. 2015). Only a few studies were performed in the field under natural conditions and comparing for example the seasonal influence on the metabolome.

Table 1: Molecular studies on the effects of HNT stress in rice published during the last 10 years.

Area	Reference
Transcriptome	*Glaubitz et al. 2017* , Liao et al. 2015
Proteome/Lipidome	Shi et al. 2017, **Shi et al. 2013**, Li et al. 2011
Metabolome	**Schaarschmidt et al. 2020**, Dhatt et al. 2019, *Glaubitz et al. 2017*, **Bahuguna et al. 2017**, **Sharma et al. 2017**, Glaubitz et al. 2015
bold for studies in the field, *italic* for studies using more than one omics approach	

1.4 Aims of the thesis

With the predicted changes in the global climate and their effects on global food security, analyzing the wide natural variation of plants is essential to ensure food security. For that, different omics approaches need to be utilized to identify on a large-scale candidate genes for crop improvement and biological markers. But also bioinformatic tools and pipelines need to be improved, developed, and finally evaluated to obtain reliable results for further applications such as breeding. Especially analyzing the transcriptome is an untargeted, time- and cost-efficient approach to identify such candidate genes that are regulated under different stress conditions in different genotypes.

I aimed to evaluate seven bioinformatic alignment tools by mapping Illumina single-end RNA-seq reads against the reference genome or transcriptome, based on experimentally generated data from *Arabidopsis thaliana.* Among those seven tools, I aimed to compare in detail key parameters such as mapping rate, raw count distribution, and the positions on the reference genome of the mapped single-end reads. Also, the influence on further downstream analysis procedures such as DGE should be analyzed. To extend this methodological study, I also wanted to expand the analysis of the transcriptome to the more complex transcriptomes of rice (*Oryza sativa*), using a third-generation sequencing technique: PacBio Isoform sequencing (IsoSeq). Here, I aimed to identify cultivar-specific transcripts among ten rice cultivars that mainly lacked a proper genome assembly. In addition, these cultivars were selected, because they showed different tolerance levels to abiotic stressors such as HNT or combined heat and

drought stress to identify possible uncharacterized stress-responsive genes in the tolerant cultivars. Also, I wanted to explore and evaluate different data processing pipelines for this relatively new sequencing technique. As another aspect of the omics investigations, I aimed to analyze the metabolic responses of rice panicles and flag leaves from rice cultivars that were partly used in the PacBio study and collected during the exposure to HNT stress. As an additional factor, the two experiments were grown in the field once during the dry season and once during the wet season to identify possible differences between the metabolic profiles.

In summary, the work presented here intended to expand the knowledge of existing RNA-seq mapping tools and to evaluate it with experimental data from a higher plant. Also, I wanted to show that long-read sequencing can indeed be used for targeted sequencing for organisms without a proper genome reference and also taking into account the broad natural variation in the plant kingdom. Finally, metabolite studies were performed which can help to identify potential biological markers for rice breeding under HNT and can also be combined with transcriptomic studies.

2 Contributions

Paper 1: Evaluation of Seven Different RNA-seq Alignment Tools Based on Experimental Data from the Model Plant *Arabidopsis thaliana*

Study concept were designed by Dirk Hincha and Ellen Zuther. Samples for sequencing were collected by Ellen Zuther. Data analysis and figure generation were performed by me and supervised by Axel Fischer and Ellen Zuther. The first draft of the manuscript was written by me and Dirk Hincha. All authors contributed to the final version of the manuscript.

Paper 2: Season Affects Yield and Metabolic Profiles of Rice (*Oryza sativ*a) under High Night Temperature Stress in the Field

The project idea was developed by SV Krishna Jagadish, Dirk Hincha, and Ellen Zuther. Experiments at the International Rice Research Institute (Philippines) were performed by Lovely Lawas, Xia Li, and Ulrike Glaubitz. Metabolite measurements and annotation were performed by Alexander Erban and Joachim Kopka. Data curation, analysis and generated figures were performed by me and supervised by Ellen Zuther and Dirk Hincha. First draft of the manuscript was written by me, Ellen Zuther and Dirk Hincha. All authors contributed to the final version of the manuscript.

Paper 3: *De novo* Reconstruction of Transcriptomes of ten *Oryza sativa* Cultivars using PacBio Single-Molecule Real-Time Sequencing

Project idea was developed by Ellen Zuther and Dirk Hincha. Experimental design was conceptualized by Ellen Zuther, Dirk Hincha, and Krishna Jagadish. RNA-seq samples for N22 were generated by Lovely Lawas. IR64 root and shoot RNA were provided by Julia Bailey-Serres, Rejbana Alam and Endang Septinigsih. PacBio IsoSeq data was generated by Bruno Hüttel. Data analysis, methodical evaluation, and selected software were chosen and performed by me and supervised by Axel Fischer. The first draft of the manuscript was written by me, Axel Fischer and Dirk Hincha.

_____

Apl. Prof. Dr. Dirk Walther

Potsdam, 24.08.2020

2.1 Paper 1: Evaluation of Seven Different RNA-seq Alignment Tools Based on Experimental Data from the Model Plant *Arabidopsis thaliana*

Stephanie Schaarschmidt, Axel Fischer, Ellen Zuther and Dirk K. Hincha

Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany

Journal: International Journal of Molecular Sciences

Received: 30 January 2020

Accepted: 29 February 2020

Published: 3 March 2020

doi: 10.3390/ijms21051720

International Journal of
Molecular Sciences

MDPI

Article

Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant *Arabidopsis thaliana*

Stephanie Schaarschmidt, Axel Fischer, Ellen Zuther and Dirk K. Hincha *

Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany; schaarschmidt@mpimp-golm.mpg.de (S.S.); afischer@mpimp-golm.mpg.de (A.F.); zuther@mpimp-golm.mpg.de (E.Z.)
* Correspondence: hincha@mpimp-golm.mpg.de; Tel.: +49-331-5678253

Received: 30 January 2020; Accepted: 29 February 2020; Published: 3 March 2020

Abstract: Quantification of gene expression is crucial to connect genome sequences with phenotypic and physiological data. RNA-Sequencing (RNA-Seq) has taken a prominent role in the study of transcriptomic reactions of plants to various environmental and genetic perturbations. However, comparative tests of different tools for RNA-Seq read mapping and quantification have been mainly performed on data from animals or humans, which necessarily neglect, for example, the large genetic variability among natural accessions within plant species. Here, we compared seven computational tools for their ability to map and quantify Illumina single-end reads from the *Arabidopsis thaliana* accessions Columbia-0 (Col-0) and N14. Between 92.4% and 99.5% of all reads were mapped to the reference genome or transcriptome and the raw count distributions obtained from the different mappers were highly correlated. Using the software DESeq2 to determine differential gene expression (DGE) between plants exposed to 20 °C or 4 °C from these read counts showed a large pairwise overlap between the mappers. Interestingly, when the commercial CLC software was used with its own DGE module instead of DESeq2, strongly diverging results were obtained. All tested mappers provided highly similar results for mapping Illumina reads of two polymorphic Arabidopsis accessions to the reference genome or transcriptome and for the determination of DGE when the same software was used for processing.

Keywords: *Arabidopsis thaliana*; differential gene expression; natural genetic variation; read mapping tools; RNA-Seq

1. Introduction

Since the completion of the human genome project in 2003 [1], sequencing technologies have developed extraordinarily fast. The resulting data have revealed the astonishing complexity of genome architecture and transcriptome composition. In this context, transcript identification and the quantification of gene expression play crucial roles in connecting genomic information with phenotypic and biochemical measurements. These two key aspects of transcriptomics can be combined in a single high-throughput sequencing assay called RNA-Sequencing (RNA-Seq). This approach allows detailed transcript profiling including the identification of splicing-induced isoforms, nucleotide variation and post-transcriptional base modification [2].

While comparative studies of diverse read aligners have been performed using data with a corresponding reference genome or transcriptome [3–7] or *de novo* assembly [8–10], only little evaluation is available of the performance of read mappers for data generated from genotypes within a species showing sequence polymorphisms. In this study, the algorithmically different mappers

bwa, CLC Genomics Workbench, HISAT2, kallisto, RSEM, salmon and STAR were used to map experimentally generated RNA-Seq data from the two natural accessions Columbia-0 (Col-0) and N14 of the higher plant *Arabidopsis thaliana* and to quantify the transcripts.

Bwa (Burrows–Wheeler-Alignment) was developed for mapping short DNA sequences against a reference genome and was extended for RNA-Seq data analysis. For indexing, the algorithm constructs a suffix array and Burrows–Wheeler-Transformation (BWT), and subsequently matches the sequences using a backward search [11]. STAR (Spliced Transcripts Alignment to a Reference) is a specialized tool for RNA-Seq reads that uses a seed-extension search based on compressed suffix arrays [12] and can detect splice-junctions. HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2) is also a splice-aware aligner using a graph-based alignment approach (graph Ferragina Manzini index) that can align DNA and RNA sequences [13]. RSEM (RNA-Seq by Expectation Maximization) is a software package that quantifies transcript abundances. It can employ different pre-defined mappers such as bowtie2 and based on the generated alignments utilizes a maximum likelihood abundance estimation, the expectation-maximization algorithm, as the statistical model to quantify transcripts [14]. By contrast, salmon and kallisto are tools which do not perform a classical alignment of individual bases, but instead implement new strategies for RNA-Seq quantification. Salmon is based on the concept of quasi-mapping. It uses a suffix array that is BWT-indexed and searched by an FMD algorithm, allowing the discovery of shared substrings of any length between a read and the complete set of transcripts. Mismatches are handled with chains of maximally exact matches [15]. The concept of kallisto is based on *pseudo-alignments*. Pseudo-alignments define a relationship between a read and a set of compatible transcripts. This relationship is computed based on "mapping" the *k*-mers to paths in a transcript De Bruijn graph. As the pseudo-alignments are generated, equivalence classes are computed and used for the relative isoform quantification [16]. CLC read mapping utilizes an approach described by Mortazavi et al. [3] and is the only commercial tool with a graphical user interface included in our study.

Here, we compare the performance of these seven RNA-Seq mappers in the analysis of experimentally generated transcriptome data covering more than 30,000 *Arabidopsis thaliana* genes. The analysis compares alignment accuracy and quantification to enable comprehensive biological interpretation. For the RNA-Seq experiment, RNA was isolated from the higher plant *Arabidopsis thaliana* and the performance of each software was tested on 150 bp single-end reads from the two natural accessions Col-0 and N14 [17]. Mappability, raw count expression, overall similarity of the count distribution and differential gene expression (DGE) were analyzed to compare the mappers. The two splice-aware aligners HISAT2 and STAR were compared for accuracy by mapping the reads against the reference genome without an annotation. Additionally, an in silico approach to characterize the correctness of the mappers was performed (see Figure 1 for a schematic description of the analysis workflow).

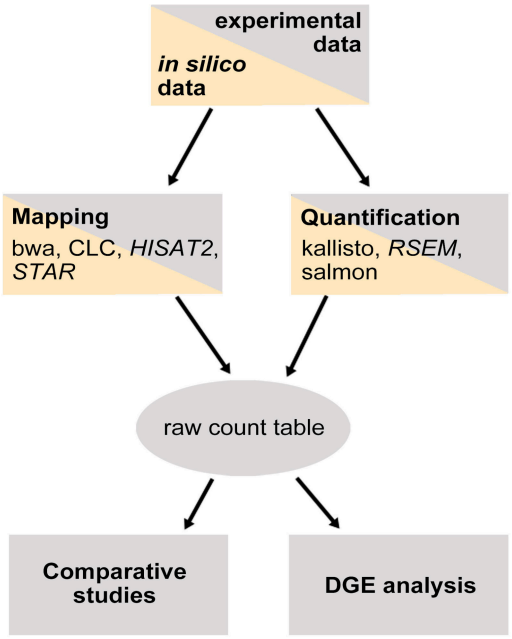


Figure 1. Analysis workflow. Light gray represents all steps performed for experimental data, light orange for analysis of in silico generated data analyzed with HISAT2, RSEM and STAR.

2. Results

2.1. Mapping Statistics

After pre-processing, the resulting dataset contained 36 samples [17] with a sequencing data size ranging from about 21 to almost 33 × 10^6 reads (Table A1). In general, a high fraction of the total reads was mapped for both accessions. The mapping for Col-0 was slightly better than for N14 (Figure 2) with mapped reads between 95.9% (bwa) and 99.5% (STAR). For N14 between 92.4% (bwa) and 98.1% (STAR) of the reads were mapped against the respective reference sequence of Col-0 (Table A2).

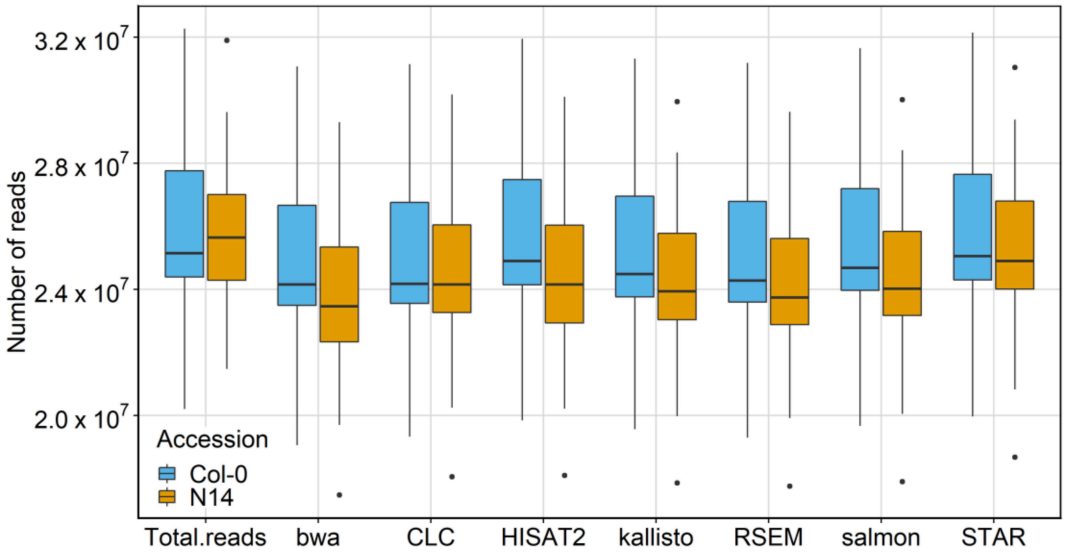


Figure 2. Mapper comparison based on mappability. Number of mapped reads against the Col-0 reference sequence for all seven mappers and each accession separately. The analysis included RNA-Seq data from 36 biological samples. Outliers for N14 were in each case sample V for minimum, sample AF for maximum (see Table A3 for sample information).

2.2. Raw Count Distribution for Individual Samples

Raw count distributions between the mappers were investigated for both accessions. The unfiltered expression values for each mapper were plotted against each other and correlations computed. The results for one control sample of Col-0 (sample A) and N14 (sample B) are shown as an example (Figure 3). For Col-0 (Figure 3a), high correlation coefficients between 0.977 (STAR vs. CLC) and 0.997 (kallisto vs. salmon) were determined. For N14 (Figure 3b) the correlation coefficients ranged from 0.978 (CLC vs. HISAT2) to 0.996 (kallisto vs. salmon). Regarding the STAR and HISAT2 comparisons with all other mappers, a higher variance was observed in the direction of STAR and HISAT2 for lowly expressed genes.

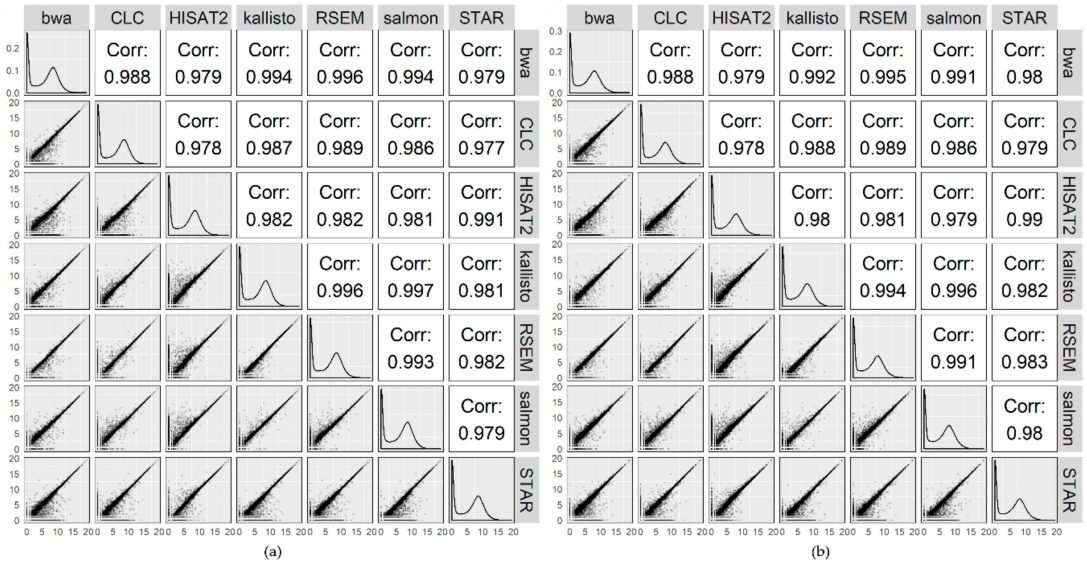


Figure 3. Raw counts of mapped reads determined by each mapper plotted against each other. Results are shown for sample A of Col-0 (**a**) and sample B of N14 (**b**) which both were obtained from plants grown under control conditions at 20 °C (see Table A3 for sample information). Lower triangle represents scatterplots of $\log_2$(counts + 1) transformed, unfiltered raw counts for each mapper plotted against each other. The diagonal histograms show the density of the raw count distribution for each mapper. The upper triangle displays the correlation coefficients.

2.3. Overall Comparison of the Mappers

For a more quantitative comparison, the raw counts generated by each mapper from all samples were compared against each other employing the R_v coefficient to quantify similarity. The raw count tables generated by the seven mappers have a high similarity indicated by R_v values close to 1 (Figure 4). Salmon and kallisto showed the highest similarity (R_v = 0.9999). CLC mapped slightly differently compared to bwa, HISAT2, kallisto, RSEM and salmon. However, it should be stressed that the raw count tables of all mappers were very similar; with 0.9804 as the lowest R_v value (CLC vs. HISAT2).

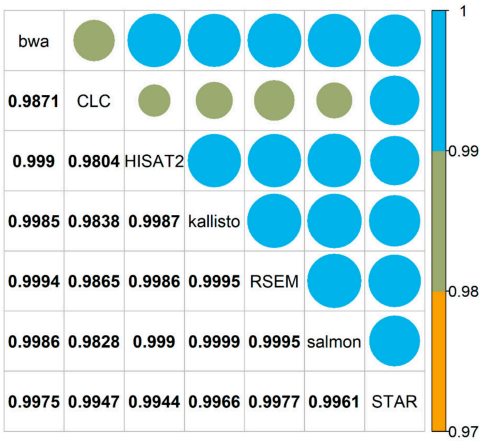


Figure 4. Mapper comparison based on raw count distributions. Graphical representation of the computed R_V values based on the correlation matrices of the unfiltered raw count tables generated by all mappers for all samples from both accessions. Values close to 1 indicate high similarity. The color and shape scales were adjusted to visualize the small differences between the R_V coefficients.

To investigate the effect of mapper choice on further statistical analysis, differentially expressed genes between control and cold acclimated conditions were determined [17]. In the read mapping steps, the aligners bwa, salmon and kallisto, using the transcriptomic reference, identified 32,243 expressed genes and thus 1,359 genes less than the other mappers with 33,602 genes each. This difference is due to the presence of non-coding RNAs such as transfer RNAs (tRNA) and micro RNAs (miRNA) in the genomic reference, which are absent from the transcriptomic reference that is based on poly-adenylated mRNAs. Prior to DGE analysis, transcript raw count tables were filtered to remove lowly expressed genes with less than five counts over all 36 samples, resulting in 23,903 (CLC) to 25,144 (RSEM) genes (Table 1). While this cut-off is admittedly arbitrary, most genes are removed with a cut-off of 1 read count (around 20%), while additional increases from 2 to 10 counts only reduce the number of genes by 2–0.3% per additional count, making the exact cut-off rather uncritical.

Table 1. Number of expressed genes identified in all samples before and after filtering out lowly expressed genes.

	Bwa	CLC	HISAT2	Kallisto	RSEM	Salmon	STAR
Before filtering	32,243	33,602	33,602	32,243	33,602	32,243	33,602
After filtering	24,197	23,903	24,840	24,810	25,144	24,574	24,515

The percentage of overlapping DGE (control vs. cold acclimated) identified by each pair of mappers was analyzed in both directions using *DESeq2* [18] in all cases and was plotted in an asymmetric matrix. For Col-0 (Figure 5a) kallisto and salmon yielded a large overlap of DGE of 98% (kallisto vs. salmon) and 97.7% (salmon vs. kallisto). For N14 (Figure 5b) slightly smaller overlaps were detected, but also here salmon and kallisto (97.6% and 96.4%) yielded the largest overlap. On the other hand, for both Col-0 and N14 the lowest overlaps were detected for bwa and STAR (93.4% and 92.1%, respectively). In general, a smaller overlap of DGE between 92% and 94% was identified for the comparisons of STAR and HISAT2 with the remaining five mappers.

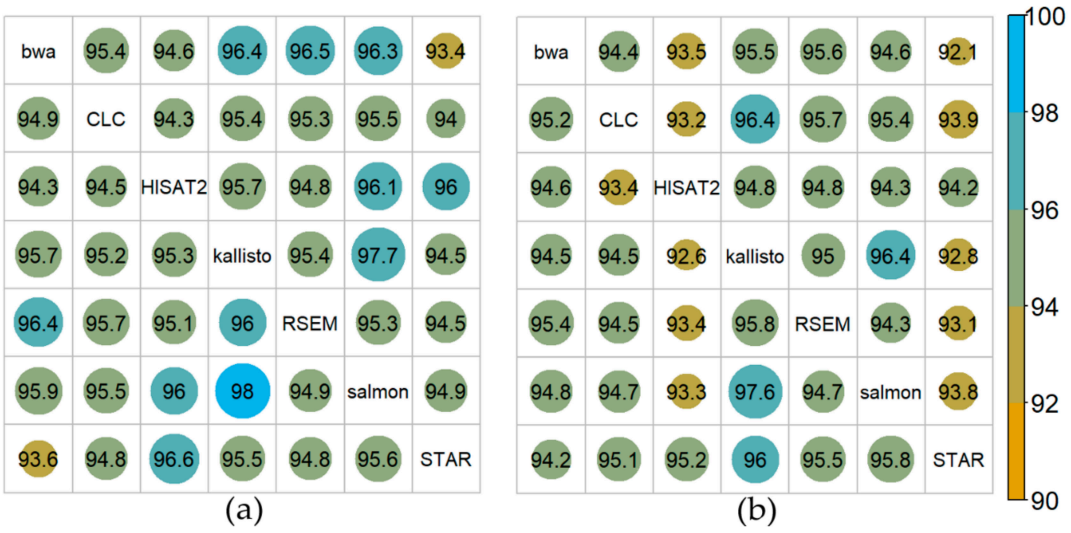


Figure 5. Overlap of significantly differentially expressed genes among the different mappers for cold acclimated vs control plants. Overlap in % for Col-0 (**a**) and N14 (**b**). DGE was determined at FDR $p < 0.1$ and an absolute $\log_2$FC > 1 using the R-package *DESeq2*. Overlap of differentially expressed genes among each pair of mappers is represented in an asymmetric matrix.

DGE analysis [19,20] was additionally performed directly in the CLC software instead of using *DESeq2*. Using the standard significance levels for these two software packages (FDR < 0.1 and FDR < 0.05 for *DESeq2* and CLC, respectively) this resulted in a much higher number of significantly differentially expressed genes for the two exemplary comparisons, detailed under Methods, compared to the *DESeq2* analysis (Table 2). Also, there was only a limited overlap between the results of the two methods.

Table 2. DGE analysis using the CLC software.

		DESeq2	CLC			
Comparison	**Accession**		**Baggerly**	**Overlap DESeq2**	**EDGE**	**Overlap DESeq2**
C28P3/C28	Col-0	2014	3034	1013	2921	1006
	N14	2101	3414	1061	3311	1052
C28P3L7T3/C35P3	Col-0	1	98	0	86	0
	N14	1	168	0	259	0

Differential gene expression was calculated with *DESeq2* (FDR < 0.1, abs (log2FC > 1), based on STAR alignments and two CLC approaches after Baggerly and EDGE (FDR < 0.05, abs (log2FC > 1)).

All mappers have different options to perform RNA-Seq quantification (Table 3). While most mappers can only use either a genome or a transcriptome reference, CLC, HISAT2 and STAR are able to use both types of reference sequences to align transcripts. Depending on the downstream analysis, it is essential which output each mapper provides. The classical alignment-based mappers bwa, CLC, HISAT2, RSEM and STAR provide an alignment output of the reads against the references, whereas salmon and kallisto only provide read quantifications. Nevertheless, kallisto offers a "pseudo-alignment" which can generate alignment files and salmon provides an option to re-quantify RNA-Seq reads using previously generated alignments against the transcriptome as obtained, for example, from STAR. Five out of the seven mappers generate transcript count tables. Only for HISAT2 and bwa additional tools have to be employed for this purpose.

Table 3. Comparison of selected key features of the used mappers. Features indicated by X are included in the specified mapper.

	Bwa	CLC	HISAT2	Kallisto	RSEM	Salmon	STAR
Reference							
Genome		X	X				X
Transcriptome	X	X	X	X	X	X	X
Needs annotation	X	X		X	X	X	
Specifications							
Alignment-based	X	X	X		X		X
Pseudo-alignment				X		X	
Expression values		X		X	X	X	X
Splice aware		X	X				X
Commercial software		X					

For a more detailed investigation of the comparability of the outputs of different mappers, three of the seven mappers were analyzed in detail regarding read position on the reference sequence. The overlap of reads from one sample, which were mapped by HISAT2, bowtie2/RSEM and STAR, was determined and the positions of the mapped reads on the reference genome were compared. For Col-0 around 11.2×10^6 (Figure 6a) of around 24.9×10^6 mapped reads and for N14 around 10.5×10^6 reads (Figure 7a) of around 22.0×10^6 mapped reads were located on the same genomic position by all three mappers. For both accessions, bowtie2/RSEM showed a high number of reads mapping to a different position compared to HISAT2 and STAR. The number of reads with a unique position was between 20.4-fold and 10.9-fold higher for bowtie2/RSEM than for the other two mappers. Hence, the differences in read positions were determined, showing that most of these reads had a position that differed by one base pair. This is the result of soft clipping of the first or last base pair that is performed by HISAT2 and STAR. After adding the base pair back to the reads in HISAT2 and STAR, the overlap with RSEM increased to 20.8×10^6 reads for Col-0 (Figure 6b) and to 17.9×10^6 reads for N14 (Figure 7b). However, RSEM still produced between 18.4-fold and 3.8-fold more uniquely positioned reads than HISAT2 and STAR that cannot be explained by soft clipping.

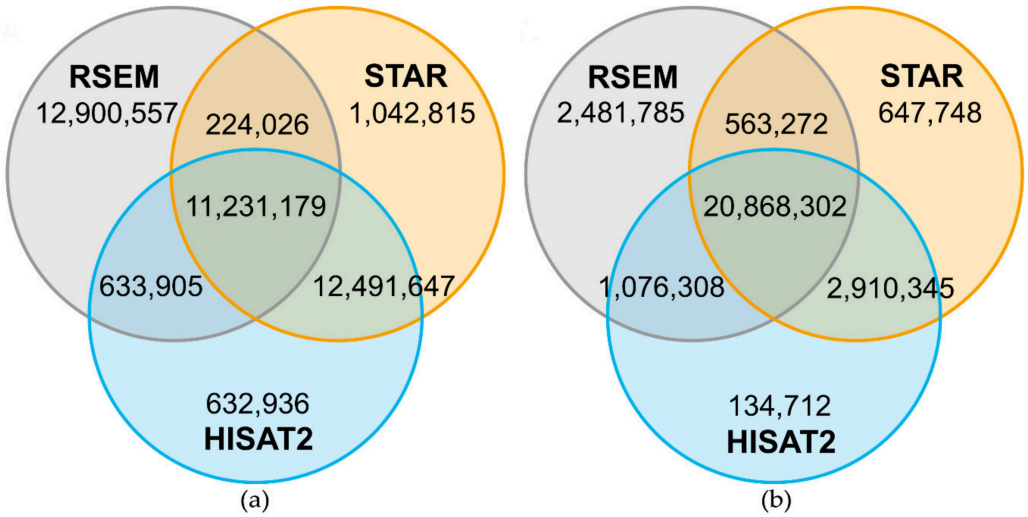


Figure 6. Number of reads mapping on the same genomic position comparing HISAT2, RSEM and STAR for Col-0. Venn diagrams are based on 24,989,667 reads mapped by all three mappers and represent the overlap of mapped reads on the same genomic position for sample A (see Table A3 for sample information). A high number of the uniquely mapped reads in RSEM was based on soft-clipping by one bp performed by HISAT2 and STAR (**a**). The reads in HISAT2 and STAR were corrected by adding the soft-clipped bp back and the overlap with RSEM increased strongly (**b**).

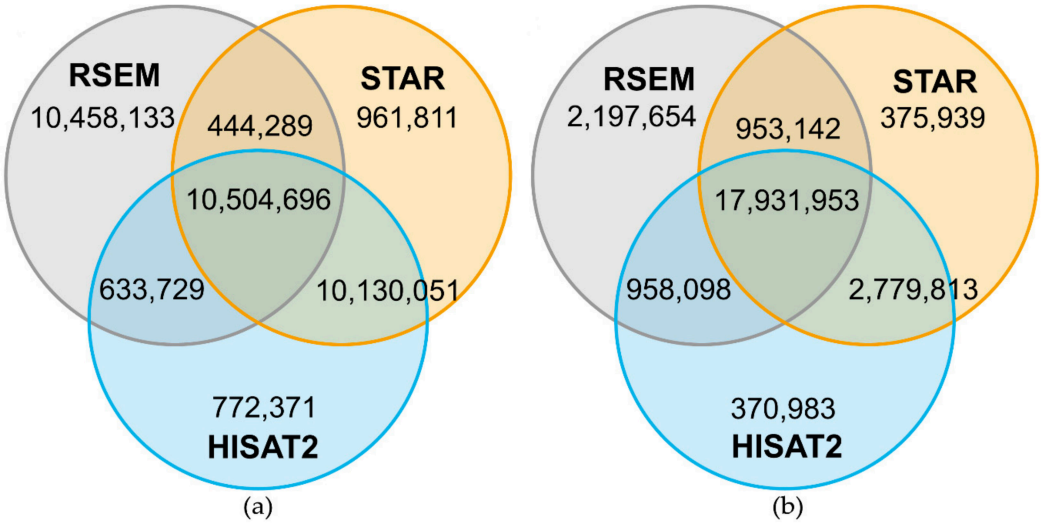


Figure 7. Number of reads mapping on the same genomic position comparing HISAT2, RSEM and STAR for N14. Venn diagrams are based on 22,040,847 reads mapped by all three mappers and represent the overlap of mapped reads on the same genomic position for sample B (see Table A3 for sample information). A high number of the uniquely mapped reads in RSEM was based on soft-clipping by one bp performed by HISAT2 and STAR (**a**). The reads in HISAT2 and STAR were corrected by adding the soft-clipped bp back and the overlap with RSEM increased strongly (**b**).

Additionally, the two splice-aware aligners HISAT2 and STAR were tested for accuracy. Reads of all 36 biological samples were mapped against the reference genome sequence without annotation and reads on exons were determined with featureCounts (Table 4). For Col-0, 93% (STAR) and 94% (HISAT2), and for N14 around 91% (both mappers) of the primary alignments were mapped to known exons. A small fraction of reads were not assigned to the annotated exons due to no mapping, multimapping (i.e., mapping to more than one location) or mapping to intergenic regions.

Table 4. Fraction of reads mapped to known exons for HISAT2 and STAR.

	HISAT2		**STAR**	
	Col-0	**N14**	**Col-0**	**N14**
Assigned to exon	94.34	90.70	93.05	90.72
Unmapped	1.10	5.16	0.50	1.99
Multimapped	4.01	3.61	5.93	6.77
No Feature (intergenic)	0.55	0.53	0.51	0.53

To test accuracy of HISAT2 and STAR, reads of the 36 biological samples were mapped against the reference genome without including an annotation. More than 90% of reads were mapped for both accessions and mappers to known exons while a small fraction was either unmapped, multimapped or mapped to intergenic positions.

2.4. Mapping of in Silico Generated Reads

To investigate whether mappers placed the mapped reads in the correct positions on the reference genome, the alignment results for in silico generated Col-0 RNA-Seq reads were analyzed using HISAT2, bowtie2/RSEM and STAR. All three mappers correctly positioned a high percentage (almost 99%) of the reads on the respective reference sequence (Table 5) for the primary alignments. Almost all remaining reads were mapped to the correct gene, but to a different transcript. Furthermore, only 0.001 to 0.03% of the reads were not mapped against the reference sequence for all mappers. A small number of reads mapped to intergenic regions for STAR and HISAT2 while for bowtie2/RSEM all reads were mapped on known genes. This derives from the fact that the used mapper bowtie2 is a splice unaware aligner that only maps against the transcriptome which was extracted from the genome reference. For the secondary alignments of HISAT2 and STAR, which only constituted 3.2% (STAR) and 3.8% (HISAT2) of the total alignments, 41.5% (HISAT2) and 36.9% (STAR) of the reads were correctly

aligned. The majority of the secondary alignments, 55% for HISAT2 and 59% for STAR, mapped the reads to wrong positions, mostly to wrong (unrelated) or paralogous genes. For bowtie2/RSEM, almost 43% of these reads were mapped multiple times. Nearly 96% of these reads were mapped to the wrong gene.

Table 5. Mapping of the in silico-generated Col-0 transcriptome using HISAT2, RSEM and STAR.

	HISAT2	in %	RSEM	in %	STAR	in %
Primary						
Mapped on right transcript	57,981,570	98.7	58,072,536	98.9	58,000,379	98.8
Mapped on wrong transcript	689,541	1.2	658,699	1.1	668,909	1.1
Unmapped	18,022	0.031	773	0.001	19,526	0.033
Mapped not on known exon	42,875	0.073	0	0.0	43,194	0.1
total reads	58,732,008	100	58,732,008	100	58,732,008	100
Secondary						
Mapped on right transcript	962,756	41.5	1,788,234	4.1	727,039	36.9
Mapped on wrong transcript	1,280,622	55.1	42,112,759	95.9	1,164,065	59.1
mapped on different gene	825,766	64.5	38,112,265	90.5	842,864	72.4
mapped on paralog gene	454,178	35.5	3,957,169	9.4	320,812	27.6
mapped on different isoform	678	0.1	43,325	0.1	389	0.0
Mapped not on exon	79,118	3.4	0	0.0	77,647	3.9
total reads	2,322,496	100	43,900,993	100	1,968,751	100

For a better overview, the alignments were split into primary and secondary alignments. If a read maps multiple times against the reference, one mapping is defined as primary (underlying criteria depend on the mapper), while the other mappings are classified as secondary alignments.

3. Discussion

RNA-Seq data from the *Arabidopsis thaliana* accessions Col-0 and N14 were mapped with five alignment-based and two pseudo-alignment tools. For Col-0, high mappability of the 150 bp single-end Illumina reads to the Col-0 reference genome or transcriptome was found for all seven alignment tools, ranging from 95.9% (bwa) to 99.5% (STAR). A slightly smaller fraction of the reads obtained from N14 was mapped to the same references, ranging from 92.4% to 98.1%. The high quality of the reference sequences may contribute to the high fraction of mapped reads. For both accessions, bwa had the lowest performance and STAR the highest, although it should be stressed that differences in mappability for any sample between the mapping tools ranged only from 1% to 4%. Comparable performance of different mapping tools has been found in previous studies using either simulated reads or RNA-Seq reads obtained from various non-plant organisms [21–25]. On the other hand, another report showed that seed-extended approaches used by STAR performed better than e.g., exon-first approaches, when mapping reads from genetically polymorphic species [26].

Considering the two accessions separately, the high number of mapped reads for Col-0 is in agreement with the fact that the Col-0 reference sequences were used for mapping. However, a small number of reads was not mapped, potentially due to sequencing errors or to polymorphisms between the publicly available genome sequence and the genome of the Col-0 population used in our experiments. In this context it has to be kept in mind that the Col-0 populations used in various laboratories around the world have been separated for many generations and have very likely accumulated different mutations over time [27]. The generally lower percentage of mapped reads for N14 can be explained by natural variation between the accessions [28,29].

In addition to the percentage of mapped reads, the correctness of the mapping of reads to the reference genome or transcriptome is also of crucial importance to obtain reliable biological information from an RNA-Seq experiment. We found that HISAT2 and STAR had a high overlap of reads mapping to the same position in the reference sequence. The differences in read positions between bowtie2/RSEM and HISAT2/STAR originated to a large part from the soft-clipping, mostly of the first base of the

reads, by both aligners. Soft-clipping can be turned off in both tools and that largely eliminates the observed differences. However, STAR has a higher tolerance for more soft-clipped and mismatched bases compared to HISAT2, which leads to a higher mapping rate for STAR and more unmapped reads for HISAT2 [24]. Also, in our analysis, STAR showed the highest fraction of mapped reads for both accessions among all compared mapping tools.

Our analysis of an in silico generated RNA-Seq data set also indicated that differences in the mapping quality between the three mappers are most likely due to their different ability to deal with mismatches. About 99% of the primary aligned reads were correctly positioned and the mappers showed the same performance when synthetic reads without any mismatches between read and reference sequences were used. This indicates that mapper performance may also depend on other factors, such as the complexity of the genome, read length and read quality [22]. The high fraction of correctly mapped reads may in part be due to the comparatively small genome of Arabidopsis with roughly 130 megabases and a low content of repetitive DNA sequences [30,31]. Regarding the secondary alignments, RSEM showed a high number of multimapped reads. The mapping for RSEM was performed with the mapper bowtie2 which searches for distinct, valid alignments for each read. As long as no upper limit is defined, bowtie2 will continue to look for all alignments that are as good or better for one read [32]. If the same read maps multiple times with the same quality string, the primary alignment is chosen randomly. The quantification algorithm of RSEM also depends on a high number of multi-mapped reads.

From a biological point of view, the quantification of gene expression is the most important part of an RNA-Seq experiment as researchers are mostly interested in the identification of differentially expressed genes, either between conditions or between genotypes. Correct mapping, as discussed above, is important to identify the correct genes as being differentially expressed. However, determining the correct read count numbers is of at least equal importance [33]. We have addressed this issue on two levels by comparing raw counts for the different genes or transcripts among the mapping tools and by comparing differentially expressed genes between plants grown under ambient and cold conditions identified by the different tools.

To investigate the results obtained by the different tools on the basis of raw counts, raw count numbers for each gene/transcript of a single sample from Col-0 and N14 each, generated by the different mappers, were plotted against each other. In general, high similarities among the mappers were observed, indicated by correlation coefficients close to 1. Similarly, when the raw counts were compared between mappers for all 36 biological samples generated in this study, R_v values close to 1 indicated a good correspondence in the expression levels computed by all seven software tools.

To analyse the effects of the mapping tools on the DGE analysis, we compared expression levels of control plants grown at ambient temperature with expression levels of plants that were exposed to 4 °C for three days (cold acclimation; compare [17] for a detailed description). Significantly differentially expressed genes were in all cases identified using the *DESeq2* tool. The results showed that the raw counts generated by the different mappers resulted in clear differences in the number of significantly differentially expressed genes, with an overlap between mappers from 98.0% between kallisto and salmon in Col-0, and 92.1% between bwa and STAR in N14. The small sample size (three samples per condition and accession) may of course contribute to the uncertainty in identifying differentially expressed genes unambiguously [34]. However, this sample size is currently the standard in biological experiments and therefore our results give a realistic impression of what the user can expect from the performance of these tools.

Finally, the results from *DESeq2* and from the DGE-pipeline of CLC were compared. Interestingly, CLC identified about 50% more differentially expressed genes than *DESeq2*. Since the same alignments for downstream analysis were used in both cases, this difference cannot originate from differences in the mapping and raw count generation. Therefore, the normalization (to one million counts) as well as the statistical tests used by CLC must have led to these differences. In a transcriptome analysis of mouse tissues, different DGE tools such as *DESeq2* and CLC were compared,

resulting in a better performance for *DESeq2* compared to both CLC approaches [35]. The results were experimentally validated by qRT-PCR for 18 differentially expressed genes. For the CLC Baggerly approach large differences to qRT-PCR results were shown. The CLC EDGE approach yielded results that were more similar to the expression changes found by qRT-PCR and those detected by *DESeq2*. However, in our analysis, the CLC approaches yielded results that were largely different from those obtained by *DESeq2*.

4. Materials and Methods

4.1. Experimental Dataset

RNA samples of the *Arabidopsis thaliana* accessions Col-0 and N14 were used for RNA-Seq as described in detail recently [17]. Plant material was collected from three independent biological experiments resulting in a total of 36 samples. Samples were taken after 28 days of growth at 20 °C, after an additional three days of cold acclimation at 4 °C, after a subsequent seven day period at 20 °C and after a final three days at 4 °C. Additionally, samples from developmental control plants were taken after 35 days at 20 °C and a subsequent three days of cold acclimation at 4 °C (Details of all samples are given in Table A3). Library preparation and sequencing were performed by the Max-Planck Genome Centre Cologne, Germany (https://mpgc.mpipz.mpg.de/home/). Libraries were constructed with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs) including polyA enrichment. Illumina HiSeq 3000 technology was used for sequencing and yielded 150 base pair (bp) long single end reads. RNA-Seq raw counts are available at GEO [36] under the accession number GSE112225. A detailed biological analysis of the RNA-Seq data has been presented recently [17].

4.2. Mapping

Quality control of the raw reads and adapter trimming have been described previously [17]. The genomic FASTA sequence, cDNA and GTF annotation files of *Arabidopsis thaliana* Col-0 were downloaded from EnsemblPlants [37], version TAIR10, release 31 [38]. For read mapping bwa, CLC Genomics Workbench, HISAT2, kallisto, RSEM, salmon and STAR were used, employing pre-defined default parameters as far as possible (Table 6). Bwa aln was used for higher sensitivity and resulting sai files were converted into alignment files with bwa sampe. For kallisto and salmon it was necessary to set parameters for single-end data, define the estimated average read length as well as its estimated standard deviation. As index mode for salmon, –type quasi and a stranded library type were chosen. For expression quantification kallisto and salmon were run in quant mode. For STAR, 1-pass mode was used and additional parameters were defined to sort the alignments, to limit multi-mapping and to keep unmapped reads in the alignments as well as generating the gene count output. HISAT2 was run with default parameters, for index generation annotation was included (Table 6). All tools are freely available except the CLC Genomics Workbench which is a commercial tool that requires purchase of a license. For the mappings without annotation, HISAT2 was run with default parameters and without inserting the annotation into index generation. STAR was run in the 2-pass mode. To determine the reads mapping on exons, *featureCounts* v2.0.0 [39] (–primary -T 10 -f -O -F GTF -t exon -g gene_id) was used. Expression values were natively generated by five of the seven mappers. For bwa, samtools idxstat and for HISAT2, *featureCounts* v. 2.0.0 [39] were used to determine raw counts. For mapping statistics and further analysis of the alignment files, samtools v1.3 [40] was employed.

Table 6. Overview of the seven mappers used in this study.

Mapper	Version	Parameters	Reference
bwa aln	0.7.13	Default	Li and Durbin (2009) [11]
CLC	9	Default	Qiagen, Hilden, Germany [41]
kallisto quant	0.42.5	–single, -l 150 and -s 25	Bray et al. (2016) [16]
HISAT2	2.1.0	Default	Kim et al. (2019) [19]
RSEM	1.2.30	–bowtie2, –fragment-length-mean 150 & –fragment-length-sd 25	Li and Dewey (2011) [14]
salmon quant	0.6.0	–type quasi, -k 31 –fldMean 150, –fldSD 25 and -l SF	Patro et al. (2017) [15]
STAR	2.5.2a	–outSAMtype BAM SortedByCoordinate –outFilterMultimapNmax 20 –alignSJDBoverhangMin 8 –outSAMunmapped Within –quantMode TranscriptomeSAM GeneCounts	Dobin et al. (2012) [12]

4.3. Comparison Based on Expression Values

For the comparison of the expression values (raw counts), samples A for Col-0 and B for N14 (grown under 20 °C control conditions; see Table A3) were chosen as an example. Raw counts were $\log_2$(counts + 1) transformed and results visualized with the R-package *ggplot2* [42]. For an overall comparison the R_v coefficient [43] based on correlation matrices of the unfiltered raw count tables of samples A and B over all mappers was calculated using the R-package *FactoMineR* [44]. Spearman correlation was used for correlation analysis and the significance of the results was tested as described [45]. The results were visualized employing the R-package *corrplot* [46].

4.4. Differential Gene Expression

Prior to the differential gene expression (DGE) analysis, estimated read counts provided by RSEM, kallisto and salmon were rounded to obtain integer values. The resulting count tables for all mappers were filtered to discard lowly expressed genes by keeping only those with a sum greater than five counts per gene for all 36 samples. The DGE analysis was performed using the R-Package *DESeq2* [18] including the normalization step. For CLC, alignment files were extracted and processed in the same way as for the other six mappers. Data was loaded with the function DESeqDataSetFromMatrix. Additional parameters for DGE were used as follows: test = "Wald", fitType="local" and including a batch effect correction in the design formula. For determining differentially expressed genes, a threshold p-value < 0.1 after false-discovery rate correction [47] and an absolute $\log_2$ fold change > 1 were used. Results of the comparison control vs. cold acclimation (Table A3) for Col-0 (samples A, M, Y vs. C, O, AA) and N14 (samples B, N, Z vs. D, P, AB) were investigated in detail.

Additionally, the built-in CLC workbench plugin for DGE was tested based on the mappings generated by CLC. Data was normalized "By totals" to a value of 1,000,000. Normalized data was used for determination of differentially expressed genes using the "Empirical analysis of DGE" [19] and "Baggerly's test on proportions" [20] with multiple testing correction of the generated p-values [47]. Next to the control vs. cold acclimation comparisons described above, the cold acclimated developmental controls (samples I, U, AG for Col-0 and J, V, AH for N14) were compared to the second cold stress treatment (samples K, W, AI for Col-0 and L, X, AJ for N14; Table 1). The numbers of significantly differentially expressed genes (FDR $p < 0.05$, abs($\log_2$ fold change) > 1) were compared with the results obtained by *DESeq2* based on the STAR alignments.

4.5. Mapping of in Silico Generated Reads

To investigate the mapping quality of the tools, reads were generated in silico using the *A. thaliana* transcriptome (TAIR10) and applying a sliding window approach (window size: 150 bp, shift: 1 bp) resulting in approximately 58×10^6 in silico reads. Reads were mapped with HISAT2 (using the same parameters as above), RSEM and STAR (without –outFilterMultimapNmax and –alignSJDBoverhangMin). For identification, the in silico reads contained the transcript name and the position of the read on the transcript as identifiers. Additionally, the GTF annotation file was reduced to the exon entries and the overlap with the resulting alignment files of HISAT2, RSEM and STAR was determined with bedtools [48]. Furthermore, transcript IDs were compared between alignment entry and GTF entry to identify correctly mapped reads.

5. Conclusions

All tested mappers provided highly comparable results for mapping Illumina reads from the genetically distinct Arabidopsis accessions Col-0 and N14 to the Col-0 reference genome or transcriptome. The same was true for the determination of DGE when *DESeq2* was used for processing. We conclude that all seven mappers can be equally used for RNA-Seq data analysis in Arabidopsis, even with different accessions. The only caveat is that using the CLC software for the identification of DGE yielded strongly varying results. Further research will be needed to establish whether read mapping to more complex genomes with larger non-coding regions or higher ploidy levels would pose additional challenges that may reveal larger differences between the mappers.

Author Contributions: Formal analysis, S.S.; Funding acquisition, D.K.H.; Methodology, A.F.; Project administration, D.K.H.; Software, A.F.; Supervision, A.F. and E.Z.; Visualization, S.S.; Writing—original draft, S.S. and D.K.H.; Writing—review & editing, A.F. and E.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was in part supported by a grant from the German Research Foundation (DFG) through Collaborative Research Center 973, Project A3 to DKH. The funders had no role in the design of the study and collection, analysis, and interpretation of the data and in writing the manuscript.

Acknowledgments: We thank the Max-Planck Genome Centre Cologne (http://mpgc.mpipz.mpg.de/home/) for RNA-Seq sequencing, Jessica Alpers for RNA extraction and Dirk Walther for critical reading of the manuscript and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

RNA-Seq	RNA-sequencing
DGE	Differential Gene Expression
BWT	Burrows–Wheeler-Transformation

Appendix A

Table A1. Number of reads for raw and pre-processed data.

Sample	Number of Reads Raw Data	Number of Reads Pre-Processed Data
A	26,551,078	25,965,205
B	24,160,253	23,723,408
C	24,987,211	24,631,398
D	24,679,891	24,314,564
E	32,902,966	32,265,838
F	25,343,870	24,962,434
G	25,633,391	25,255,295
H	24,767,056	24,276,316

Table A1. *Cont.*

Sample	Number of Reads Raw Data	Number of Reads Pre-Processed Data
I	22,434,138	22,074,152
J	27,102,013	26,738,311
K	29,909,220	29,473,355
L	30,039,895	29,625,213
M	25,373,173	25,045,811
N	27,401,911	27,059,316
O	32,172,339	31,758,225
P	26,713,325	26,326,809
Q	28,367,001	27,941,198
R	21,784,606	21,476,277
S	23,466,191	23,142,088
T	25,002,989	24,642,826
U	25,470,737	25,137,081
V	32,322,582	31,890,842
W	31,880,034	31,451,153
X	28,614,863	28,223,380
Y	25,396,753	24,312,026
Z	25,402,962	24,351,761
AA	21,934,477	21,095,112
AB	29,068,271	27,924,700
AC	28,363,133	27,205,327
AD	27,538,807	26,446,048
AE	21,048,979	20,198,121
AF	22,915,893	21,786,356
AG	26,195,089	25,161,103
AH	23,710,160	22,348,705
AI	25,915,840	24,936,936
AJ	27,904,776	26,835,785

Pre-processed raw data was filtered for a minimum read length of 80 base pairs and Illumina adapters were removed.

Table A2. Number of mapped reads for each mapper and sample.

Sample	bwa	CLC	HISAT2	kallisto	RSEM	salmon	STAR
A	24,990,288	25,070,332	25,727,064	25,202,788	25,068,400	25,488,500	25,877,150
B	22,235,860	22,831,185	22,831,427	22,489,984	22,450,834	22,625,100	23,535,895
C	23,568,631	23,650,969	24,398,527	23,911,331	23,729,822	24,096,400	24,545,823
D	22,665,145	23,292,374	23,392,011	23,182,011	23,001,552	23,294,400	24,114,936
E	31,067,889	31,136,183	31,948,360	31,315,586	31,186,635	31,651,600	32,144,692
F	22,079,186	23,828,249	22,529,274	23,226,055	22,975,469	23,289,400	24,362,368
G	24,360,053	24,368,639	25,003,451	24,630,743	24,435,931	24,818,000	25,152,392
H	22,607,768	23,256,060	23,230,510	22,983,234	22,847,497	23,135,800	23,972,434
I	20,887,128	20,897,575	21,647,744	21,094,905	21,052,741	21,301,500	21,759,724
J	25,002,889	25,748,821	25,729,980	25,530,361	25,258,258	25,626,800	26,525,228
K	28,251,892	28,394,902	29,083,561	28,728,018	28,398,031	28,924,400	29,340,134
L	27,691,133	28,565,611	28,456,640	28,333,833	27,965,330	28,411,700	29,380,081
M	24,027,754	24,158,404	24,771,967	24,370,150	24,159,388	24,539,200	24,947,419
N	25,448,518	26,128,347	26,116,046	25,859,912	25,708,968	25,908,500	26,872,750
O	30,483,322	30,538,741	31,426,082	30,970,549	30,650,488	31,145,900	31,631,436
P	23,748,275	24,471,940	24,562,932	24,318,422	24,070,551	24,406,800	25,332,412
Q	26,863,089	26,968,681	27,679,891	27,157,401	26,977,076	27,405,000	27,843,106
R	19,700,000	20,245,101	20,218,359	19,970,836	19,918,383	20,052,800	20,826,196

Table A2. *Cont.*

Sample	bwa	CLC	HISAT2	kallisto	RSEM	salmon	STAR
S	22,171,458	22,274,280	22,902,423	22,444,868	22,280,936	22,624,300	23,035,647
T	23,165,182	23,815,751	23,748,284	23,543,789	23,398,523	23,638,100	24,456,937
U	24,145,575	24,192,319	24,905,595	24,499,411	24,279,099	24,667,800	25,057,610
V	29,305,198	30,181,899	30,105,423	29,951,050	29,635,355	30,012,700	31,037,834
W	30,171,991	30,240,229	31,135,272	30,619,320	30,314,724	30,820,900	31,321,391
X	26,417,781	27,215,579	27,146,639	26,999,068	26,701,971	27,089,600	28,004,846
Y	23,467,493	23,523,457	24,062,637	23,713,957	23,548,791	23,915,800	24,211,437
Z	22,939,890	23,531,637	23,488,307	23,241,258	23,186,262	23,333,700	24,169,828
AA	20,347,062	20,333,841	20,891,798	20,594,460	20,425,031	20,742,500	21,011,777
AB	26,183,324	26,842,810	26,903,033	26,663,997	26,539,902	26,769,800	27,709,817
AC	26,065,885	26,102,795	26,890,847	26,358,644	26,235,209	26,562,400	27,054,414
AD	24,904,560	25,532,006	25,483,657	25,267,366	25,201,022	25,348,100	26,234,545
AE	19,055,414	19,320,692	19,842,597	19,566,392	19,295,597	19,670,300	19,967,391
AF	17,469,949	18,053,590	18,090,815	17,854,059	17,755,997	17,898,700	18,672,118
AG	24,163,365	24,161,812	24,876,952	24,468,971	24,283,108	24,682,500	25,047,179
AH	20,953,174	21,498,746	21,436,520	21,310,760	21,215,866	21,379,400	22,109,670
AI	23,823,058	23,916,429	24,617,944	24,223,766	23,973,245	24,383,700	24,792,766
AJ	25,023,005	25,804,958	25,767,495	25,481,098	25,325,866	25,563,800	26,594,060
Col-0 %	95.9	96.2	98.9	97.2	96.4	97.9	99.5
N14 %	92.4	95.2	94.9	94.2	93.6	94.6	98.1
Total %	**94.1**	**95.7**	**96.9**	**95.7**	**95.0**	**96.3**	**98.8**

Tools are sorted alphabetically by name. Total describes the fraction of mapped reads for both accessions Col-0 and N14.

Table A3. Sample list with sample name, condition (Cond.) and accession (Acc.).

Experiment 1			Experiment 2			Experiment 3	
Cond.	Acc.	Sample	Cond.	Acc.	Sample	Cond.	Acc.
C28	Col-0	M	C28	Col-0	Y	C28	Col-0
C28	N14	N	C28	N14	Z	C28	N14
C28P3	Col-0	O	C28P3	Col-0	AA	C28P3	Col-0
C28P3	N14	P	C28P3	N14	AB	C28P3	N14
C35	Col-0	Q	C35	Col-0	AC	C35	Col-0
C35	N14	R	C35	N14	AD	C35	N14
C28P3L7	Col-0	S	C28P3L7	Col-0	AE	C28P3L7	Col-0
C28P3L7	N14	T	C28P3L7	N14	AF	C28P3L7	N14
C35P3	Col-0	U	C35P3	Col-0	AG	C35P3	Col-0
C35P3	N14	V	C35P3	N14	AH	C35P3	N14
C28P3L7T3	Col-0	W	C28P3L7T3	Col-0	AI	C28P3L7T3	Col-0
C28P3L7T3	N14	X	C28P3L7T3	N14	AJ	C28P3L7T3	N14

Samples were taken from three independent biological experiments. C28/C35: Control plants after 28 days or 35 days of growth at 20 °C; C28P3/C35P3: plants after an additional 3 days of cold treatment at 4 °C; C28P3L7: cold treated plants after a further 7 days at 20 °C; C28P3L7T3: plants after an additional 3 days at 4 °C.

References

1. Collins, F.S.; Morgan, M.; Patrinos, A. The Human Genome Project: Lessons from large-scale biology. *Science* **2003**, *300*, 286–290. [CrossRef] [PubMed]
2. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef] [PubMed]
3. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Meth.* **2008**, *5*, 621–628. [CrossRef] [PubMed]

4. Dillies, M.-A.; Rau, A.; Aubert, J.; Hennequet-Antier, C.; Jeanmougin, M.; Servant, N.; Keime, C.; Marot, G.; Castel, D.; Estelle, J.; et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **2012**, *14*, 671–683. [CrossRef] [PubMed]
5. Rapaport, F.; Khanin, R.; Liang, Y.; Pirun, M.; Krek, A.; Zumbo, P.; Mason, C.E.; Socci, N.D.; Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **2013**, *14*, R95. [CrossRef]
6. Benjamin, A.M.; Nichols, M.; Burke, T.W.; Ginsburg, G.S.; Lucas, J.E. Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genom.* **2014**, *15*, 570. [CrossRef]
7. Lin, Y.; Golovnina, K.; Chen, Z.X.; Lee, H.N.; Negron, Y.L.; Sultana, H.; Oliver, B.; Harbison, S.T. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genom.* **2016**, *17*, 28. [CrossRef]
8. Amin, S.; Prentis, P.J.; Gilding, E.K.; Pavasovic, A. Assembly and annotation of a non-model gastropod (*Nerita melanotragus*) transcriptome: A comparison of De novo assemblers. *BMC Res. Notes* **2014**, *7*, 488. [CrossRef]
9. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szczesniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X.; et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 13. [CrossRef]
10. Rana, S.B.; Zadlock, F.J.I.V.; Zhang, Z.; Murphy, W.R.; Bentivegna, C.S. Comparison of de novo transcriptome assemblers and k-mer strategies using the killifish, *Fundulus heteroclitus*. *PLoS ONE* **2016**, *11*, e0153104. [CrossRef]
11. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows—Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
12. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2012**, *29*, 15–21. [CrossRef] [PubMed]
13. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [CrossRef] [PubMed]
14. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
15. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Meth.* **2017**, *14*, 417. [CrossRef]
16. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [CrossRef]
17. Zuther, E.; Schaarschmidt, S.; Fischer, A.; Erban, A.; Pagter, M.; Mubeen, U.; Giavalisco, P.; Kopka, J.; Sprenger, H.; Hincha, D.K. Molecular signatures associated with increased freezing tolerance due to low temperature memory in Arabidopsis. *Plant Cell Environ.* **2019**, *42*, 854–873.
18. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
19. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef]
20. Baggerly, K.A.; Deng, L.; Morris, J.S.; Aldaz, C.M. Differential expression in SAGE: Accounting for normal between-library variation. *Bioinformatics* **2003**, *19*, 1477–1483. [CrossRef]
21. Baruzzo, G.; Hayer, K.E.; Kim, E.J.; Di Camillo, B.; FitzGerald, G.A.; Grant, G.R. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Meth.* **2016**, *14*, 135. [CrossRef] [PubMed]
22. Everaert, C.; Luypaert, M.; Maag, J.L.V.; Cheng, Q.X.; Dinger, M.E.; Hellemans, J.; Mestdagh, P. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.* **2017**, *7*, 1559. [CrossRef] [PubMed]
23. Jin, H.; Wan, Y.-W.; Liu, Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinform.* **2017**, *18* (Suppl. 4), 117. [CrossRef]
24. Sahraeian, S.M.E.; Mohiyuddin, M.; Sebra, R.; Tilgner, H.; Afshar, P.T.; Au, K.F.; Bani Asadi, N.; Gerstein, M.B.; Wong, W.H.; Snyder, M.P.; et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* **2017**, *8*, 59. [CrossRef] [PubMed]
25. Teng, M.; Love, M.I.; Davis, C.A.; Djebali, S.; Dobin, A.; Graveley, B.R.; Li, S.; Mason, C.E.; Olson, S.; Pervouchine, D.; et al. Erratum to: A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **2016**, *17*, 203. [CrossRef] [PubMed]

26. Garber, M.; Grabherr, M.G.; Guttman, M.; Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Meth.* **2011**, *8*, 469–477. [CrossRef] [PubMed]
27. Ossowski, S.; Schneeberger, K.; Lucas-Lledo, J.I.; Warthmann, N.; Clark, R.M.; Shaw, R.G.; Weigel, D.; Lynch, M. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **2010**, *327*, 92–94. [CrossRef]
28. Atwell, S.; Huang, Y.S.; Vilhjálmsson, B.J.; Willems, G.; Horton, M.; Li, Y.; Meng, D.; Platt, A.; Tarone, A.M.; Hu, T.T.; et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **2010**, *465*, 627. [CrossRef]
29. Hancock, A.M.; Brachi, B.; Faure, N.; Horton, M.W.; Jarymowycz, L.B.; Sperone, F.G.; Toomajian, C.; Roux, F.; Bergelson, J. Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **2011**, *334*, 83–86. [CrossRef]
30. Meinke, D.W.; Cherry, J.M.; Dean, C.; Rounsley, S.D.; Koornneef, M. *Arabidopsis thaliana*: A model plant for genome analysis. *Science* **1998**, *282*, 662–682. [CrossRef]
31. Mayer, K.; Schüller, C.; Wambutt, R.; Murphy, G.; Volckaert, G.; Pohl, T.; Düsterhöft, A.; Stiekema, W.; Entian, K.D.; Terryn, N.; et al. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **1999**, *402*, 769. [CrossRef] [PubMed]
32. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Meth.* **2015**, *12*, 357–360. [CrossRef] [PubMed]
33. Fonseca, N.A.; Marioni, J.; Brazma, A. RNA-Seq gene profiling—A systematic empirical comparison. *PLoS ONE* **2014**, *9*, e107026. [CrossRef] [PubMed]
34. Soneson, C.; Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **2013**, *14*, 91. [CrossRef] [PubMed]
35. Kumar, P.K.; Hoang, T.V.; Robinson, M.L.; Tsonis, P.A.; Liang, C. CADBURE: A generic tool to evaluate the performance of spliced aligners on RNA-Seq data. *Sci. Rep.* **2015**, *5*, 13443. [CrossRef] [PubMed]
36. Edgar, R.; Domrachev, M.; AE, L. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 2074. [CrossRef] [PubMed]
37. EnsemblPlants Arabidopsis Thaliana Assembly and Gene Annotation. Available online: http://plants.ensembl.org/info/website/ftp/index.html (accessed on 5 June 2016).
38. Berardini, T.Z.; Reiser, L.; Li, D.; Mezheritsky, Y.; Muller, R.; Strait, E.; Huala, E. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **2015**, *53*, 474–485. [CrossRef]
39. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef]
40. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
41. Qiagen CLC Genomics Workbench. Available online: https://www.qiagenbioinformatics.com/ (accessed on 25 February 2019).
42. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
43. Kazi-Aoual, F.; Hitier, S.; Sabatier, R.; Lebreton, J.-D. Refined approximations to permutations tests for multivariate inference. *Comput. Stat. Data Anal.* **1995**, *20*, 643–656. [CrossRef]
44. Lê, S.; Josse, J.; Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **2008**, *25*, 1–18. [CrossRef]
45. Josse, J.; Husson, F.; Pagés, J. Testing the significance of the R_V coefficient. *Comput. Stat. Data Anal.* **2007**, *53*, 82–91. [CrossRef]
46. Wei, T.; Simko, V. R Package "Corrplot": Visualization of a Correlation Matrix. Available online: https://github.com/taiyun/corrplot (accessed on 3 July 2019).
47. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
48. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef] [PubMed]

2.2 Paper 2: Season Affects Yield and Metabolic Profiles of Rice (*Oryza sativa*) under High Night Temperature Stress in the Field

Stephanie Schaarschmidt[1], Lovely Mae F. Lawas[1,2,y], Ulrike Glaubitz[1], Xia Li[1,3], Alexander Erban[1], Joachim Kopka[1], S. V. Krishna Jagadish[2,4], Dirk K. Hincha[1] and Ellen Zuther[1]

[1]Max-Planck-Institute of Molecular Plant Physiology, 14476 Potsdam, Germany;

[2]International Rice Research Institute, Metro Manila 1301, Philippines

[3]Institute of Crop Science, Chinese Academy of Agricultural Science, Beijing 100081, China

[4]Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA

[y]Present address: Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA.

Journal: International Journal of Molecular Sciences

Received: 6 March 2020

Accepted: 29 April 2020

Published: 30 April 2020

doi: 10.3390/ijms21093187

International Journal of
Molecular Sciences

MDPI

Article

Season Affects Yield and Metabolic Profiles of Rice (*Oryza sativa*) under High Night Temperature Stress in the Field

Stephanie Schaarschmidt [1], Lovely Mae F. Lawas [1,2,†], Ulrike Glaubitz [1], Xia Li [1,3], Alexander Erban [1], Joachim Kopka [1], S. V. Krishna Jagadish [2,4], Dirk K. Hincha [1] and Ellen Zuther [1,*]

1 Max-Planck-Institute of Molecular Plant Physiology, 14476 Potsdam, Germany; schaarschmidt@mpimp-golm.mpg.de (S.S.); lfl0008@auburn.edu (L.M.F.L.); glaubitz@mpimp-golm.mpg.de (U.G.); rainbowleelx@hotmail.com (X.L.); Erban@mpimp-golm.mpg.de (A.E.); Kopka@mpimp-golm.mpg.de (J.K.); Hincha@mpimp-golm.mpg.de (D.K.H.)
2 International Rice Research Institute, Metro Manila 1301, Philippines; kjagadish@ksu.edu
3 Institute of Crop Science, Chinese Academy of Agricultural Science, Beijing 100081, China
4 Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA
* Correspondence: zuther@mpimp-golm.mpg.de
† Present address: Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA.

Received: 6 March 2020; Accepted: 29 April 2020; Published: 30 April 2020

Abstract: Rice (*Oryza sativa*) is the main food source for more than 3.5 billion people in the world. Global climate change is having a strong negative effect on rice production. One of the climatic factors impacting rice yield is asymmetric warming, i.e., the stronger increase in nighttime as compared to daytime temperatures. Little is known of the metabolic responses of rice to high night temperature (HNT) in the field. Eight rice cultivars with contrasting HNT sensitivity were grown in the field during the wet (WS) and dry season (DS) in the Philippines. Plant height, 1000-grain weight and harvest index were influenced by HNT in both seasons, while total grain yield was only consistently reduced in the WS. Metabolite composition was analysed by gas chromatography-mass spectrometry (GC-MS). HNT effects were more pronounced in panicles than in flag leaves. A decreased abundance of sugar phosphates and sucrose, and a higher abundance of monosaccharides in panicles indicated impaired glycolysis and higher respiration-driven carbon losses in response to HNT in the WS. Higher amounts of alanine and cyano-alanine in panicles grown in the DS compared to in those grown in the WS point to an improved N-assimilation and more effective detoxification of cyanide, contributing to the smaller impact of HNT on grain yield in the DS.

Keywords: high night temperature; rice; grain yield; wet season; dry season; metabolomics

1. Introduction

Rice is a staple food for more than half of the world's population and the demand is steadily increasing with the growing human population [1]. Climate change is a significant limiting factor for enhancing food production, because increasing abiotic and biotic stresses negatively affect the yield of all major crops [2–4]. During the past century, the global surface temperature has increased by an average of 0.85 °C, and a further increase of up to 3.7 °C has been predicted by 2100 [3]. This temperature increase develops asymmetrically, with a faster rise in daily minimum compared to daily maximum temperatures [5–9], leading to "high night temperature" (HNT) conditions. Asymmetric warming causes a reduction in the temperature difference between daily maximum and minimum

temperatures, i.e., the diurnal temperature range (DTR), with a negative influence on both wild and crop plant species [10]. In particular, the main rice-growing countries in Asia, including China [11], the Philippines [12,13] and India [14,15], are affected.

Several studies have reported a strong decrease in yield and grain quality, such as increased chalk formation, and altered grain growth dynamics in rice under HNT [16–21]. HNT can have a stronger impact on grain weight than high day temperatures in rice and wheat [22–24]. Field studies at the International Rice Research Institute (IRRI) in the Philippines showed that rice grain yield was reduced by 10% per 1 °C increase in night temperatures during the dry season (DS), whereas the effect of increasing day temperatures was not significant within the investigated time period [12].

Differences in HNT sensitivity among various rice cultivars based on grain yield [25–27], yield-related parameters, or phenotypes in the vegetative stage [28] have been reported, indicating natural variation in HNT tolerance. In addition, HNT reduces the starch content in panicles and negatively affects grain yield and quality (chalk and amylose content) in the sensitive cultivars Gharib and IR64, but not in the tolerant cultivar N22 [29].

Different factors may cause HNT sensitivity. Physiological effects reported under HNT include higher rates of respiration in leaves [28,30,31] and panicles [29], whereas photosynthesis is not affected [28] or may be decreased as well [32]. A reduction in nitrogen and carbohydrate translocation after flowering as a possible cause of yield reduction in HNT sensitive cultivars was also discussed [25]. Reduced grain weight and quality may be caused by lower sink strength due to lower cell wall invertase and sucrose synthase activity in sensitive cultivars, accompanied by higher sugar accumulation in the rachis [29].

Despite the increasing knowledge of the physiological responses to HNT, only little is known about the metabolomic responses of rice under these conditions. The metabolic status is important for growth, development and stress tolerance, and additionally influences important traits such as flavor, biomass, yield and nutritional quality [33–35]. Therefore, the assessment of the metabolomic status of wild and crop species can help to evaluate natural variation [33]. Additionally, the metabolome integrates molecular and environmental effects as endpoints of biological processes [36]. Moreover, metabolites constitute potential markers for the selection of tolerant crop genotypes in breeding programs. Several studies investigated metabolic changes in rice in response to abiotic stress conditions, such as salinity [37–41], osmotic stress [42], drought [43–47], heat [44,48], and combined drought and heat stress conditions [49,50].

In a corresponding study on rice under HNT conditions, sucrose and pyruvate/oxaloacetate-derived amino acids were shown to accumulate while sugar phosphates and organic acids involved in glycolysis/gluconeogenesis and the tricarboxylic acid (TCA) cycle decreased in developing caryopses [48]. A dysregulation of central metabolism and an increase in polyamine biosynthesis was described for sensitive cultivars, whereas existing metabolic pre-adaptation under control conditions was found for tolerant cultivars [51,52]. Furthermore, in sensitive cultivars, 4-amino butanoic acid (GABA) signaling—and in tolerant cultivars, the jasmonate precursor *myo*-inositol—were linked to the HNT responses [52]. A metabolomics study investigating early seed development and the early grain-filling stage in six rice cultivars reported a sugar accumulation peak seven days after flowering and 19 significantly different metabolites under HNT compared to under control conditions, with a special focus on the generally higher abundance of sugars and sugar alcohols under HNT [53].

The goal of this study was to investigate the seasonal effects of HNT responses by assessing the metabolic responses to HNT stress in flag leaves and panicles during the DS and wet season (WS) in contrasting rice cultivars under field conditions. Previous studies of the comparison of HNT's effects during the WS and DS were limited to agronomic traits [13,14,20,26,54], while the influence of HNT on the rice metabolome has not been reported yet. The present study sheds new light on the responses of rice to an important climatic stress factor that may severely limit grain yield and quality, and therefore the global food supply.

2. Results

Two field experiments were performed at the IRRI in the Philippines during the WS and DS with eight rice cultivars (Table 1). These cultivars comprised the *indica* and *japonica* subspecies and included HNT tolerant, intermediate and sensitive cultivars, as determined during the vegetative growth stage from a study under controlled environmental conditions [28].

Table 1. Experimental set-up for high night temperature (HNT) field experiments for eight contrasting *Oryza sativa* cultivars. Mean temperatures and relative humidity (RH) are given from the beginning of HNT treatment till the sampling time, when panicles reached 50% flowering.

	Experiment 1		**Experiment 2**	
Season	Wet		Dry	
Conditions	Control	HNT	Control	HNT
Cultivars	CT9993-5-10-1M IR123 IR62266-42-6-2 IR64 IR72 M202 Moroberekan Taipei309			
T_{day} (°C)	27.7		26.1	
T_{night} (°C)	22.2	27.6	22.2	27.8
RH (%)	98.3	89.4	96.4	80.9
Sampling time	Panicle at 50% flowering			
Samples	Flag leaves, panicles			

The WS experiment was performed for 84 to 104 days from transplanting till maturity, and the DS experiment, for 87 to 118 days, depending on the staggered sowing (Figure A1). Samples for metabolite analysis were taken at 59 to 78 days after transplanting in the WS, and in the DS, between 58 and 88 days. During the day (6 a.m.–6 p.m.), plants were exposed to ambient conditions, with an average temperature of 27.7 °C during the WS and 26.1 °C during the DS. The mean daytime temperature ranged from 25.4 to 29.7 °C during the WS, and from 21.8 to 30.9 °C during the DS, with maximum daily temperatures from 26.5 °C to 34.7 °C during the WS and from 24.3 °C to 36.1 °C during the DS (Figure A2e).

During the night, plots were covered by tents, and the temperature was kept constant by air conditioners, set to 22 °C for the control and 28 °C for HNT conditions. The average temperatures measured in the tents were 27.64 °C (± 0.77 °C) and 27.82 °C (± 1.07 °C) under HNT conditions and 22.24 °C (± 0.99 °C) and 22.25 °C (± 0.46 °C) under control conditions during the WS and DS, respectively (Figure 1A,B). The corresponding ambient night temperatures outside the tents are shown in Figure A2f. As the average day temperatures for both seasons were very similar, the night temperature difference of around 5 °C is the main temperature factor driving the physiological and metabolic changes in all cultivars.

Average radiation was about 22% lower in the WS than in the DS and sunshine duration in the WS reached only 45% of the values measured in the DS (Figure A2A,B). Daily rainfall in the WS was recorded between 85 and 0 mm, while it was approximately zero in the DS (Figure A2C. Accordingly, average relative air humidity was lower in the DS with values between 73% and 95%, compared to those between 76% and 98% in the WS (Figure A2D).

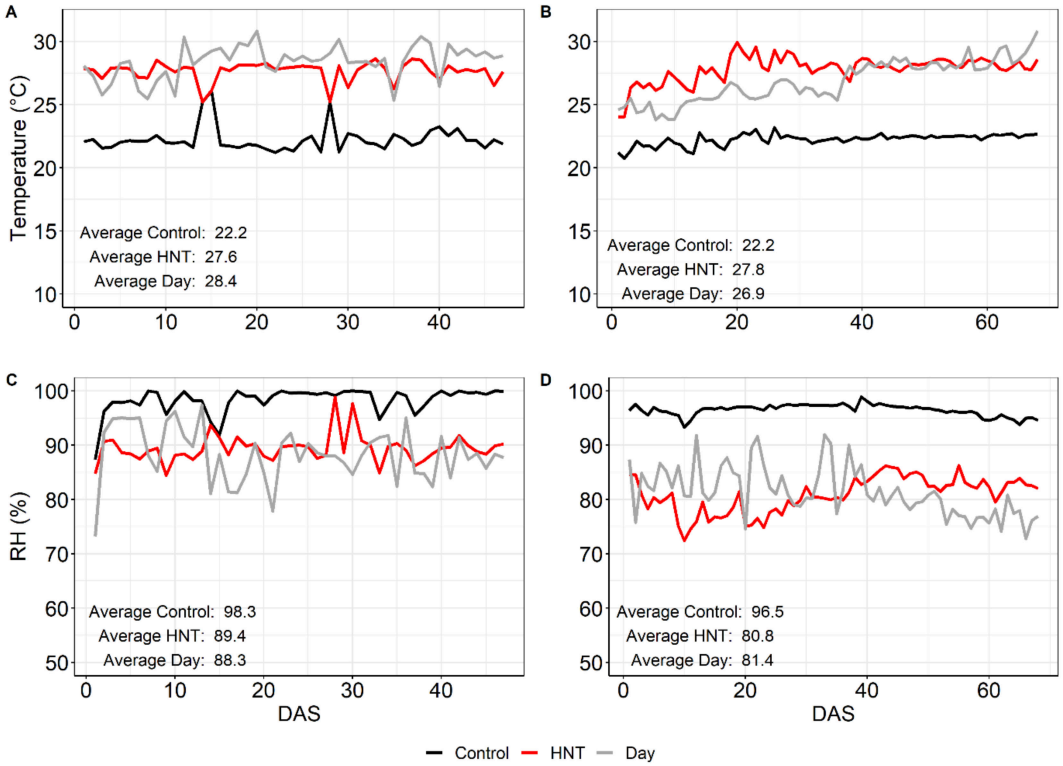


Figure 1. Average temperature (**A**,**B**) and relative humidity (RH) (**C**,**D**) during the night (6 p.m.–6 a.m.) in the wet season (WS) (**A**,**C**) and dry season (DS) (**B**,**D**) under control and HNT conditions, measured till the end of sampling at 50% flowering. For comparison, day temperature and humidity are included (grey lines). Measurements, which were done every 30 min, were averaged. DAS—Days after stress; WS—wet season; DS—dry season.

2.1. Influence of HNT on Agronomic Parameters

For all agronomic parameters, a significant genotype effect was found in both seasons when comparing samples from plants grown under HNT with control conditions (Table 2). Furthermore, a significant seasonal effect was recorded for almost all agronomic parameters. The influence of HNT conditions on the growth response was recorded as differences in plant height. No significant treatment effect but a significant Genotype x Treatment (GxT) effect of HNT on plant height was found over all cultivars for both seasons (Table 2). On average, plant height was slightly lower in the DS compared to in the WS, but cultivar-specific patterns were conserved (Figure 2). In both seasons, plant height was significantly ($p < 0.05$) increased under HNT in three cultivars (IR123, IR64 and IR72), while it was decreased in Moroberekan. IR62266-42-6-2 and M202 showed reduced plant height only in the WS, and Taipei309, only in the DS.

Total grain yield under control conditions was significantly lower in the WS, with a maximum yield among all cultivars of about 617 g·m^{-2} compared to that in the DS of 762 g·m^{-2} (Figure 3A,B). A significant effect of HNT treatment on the grain yield of all eight cultivars compared to control was only detectable in the WS (Table 2), where yield reduction varied between 23% in M202 and 4% in IR123 (Figure A3A). In the DS, yield was only reduced between 8% and 3% in four cultivars, while it was slightly increased (1%–5%) in the other four (Figure A3B). No correlation was found between the yield reduction in our experiments in the WS or DS and the HNT sensitivity rank of the same cultivars in the vegetative stage under controlled environmental conditions determined for the same cultivars in a previous study [28] (not shown).

Table 2. Analysis of variance (ANOVA) on selected agronomic parameters. Sixty plants for the DS and 24 plants for the WS were considered for plant height, tiller number and panicle number. For the remaining parameters, two replicates pooled from twelve plants each were considered for the WS and five replicates pooled from twelve plants each were considered for the DS. Spikelets/panicle represents the number of spikelets per panicle. The seed set was calculated as follows: seed set (%) = filled grains/(filled+half-filled+unfilled grains) × 100. Harvest index was calculated as percentage of dry weight of filled grains relative to total above-ground biomass. The significance of the influence of genotype (G), HNT-treatment (T), season (S) or the interaction between two influences (GxT or TxS) on differences between HNT and control conditions across all eight cultivars is indicated by asterisks: 0.001 < ***; 0.001 > ** < 0.01; 0.01 > * < 0.05. ns—not significant. Original data for plant height, tiller number, panicle number and all yield components for the WS (2011) and the DS (2014) are available in Table S1.

Parameter	WS	WS	WS	DS	DS	DS	Both Seasons		
	G	T	GxT	G	T	GxT	T	S	TxS
Plant Height (cm)	***	ns	**	***	ns.	*	ns	ns	ns
Biomass g/m^2	***	*	ns	***	ns	ns	ns	*	ns
Straw (g)	***	ns	*	***	ns	ns	ns	ns	ns
Rachis (g)	***	ns	ns	***	ns	ns	ns	*	ns
Tiller No	***	ns	ns	***	ns	ns	ns	**	ns
Panicle No	***	ns	ns	***	ns	ns	ns	**	ns
Panicle/m^2	***	ns	ns	***	ns	ns	ns	**	ns
Spikelet/m^2	***	*	ns	***	ns	ns	ns	***	ns
Spikelets/Panicle	***	**	ns	***	ns	ns	ns	***	ns
Seed set (%)	***	ns	ns	***	ns	ns	**	*	ns
Grain yield (g/m^2)	***	***	ns	***	ns	ns	ns	**	ns
1000 grain weight (g)	***	***	***	***	***	ns	ns	***	ns
Harvest Index	***	***	ns	***	**	ns	ns	***	ns

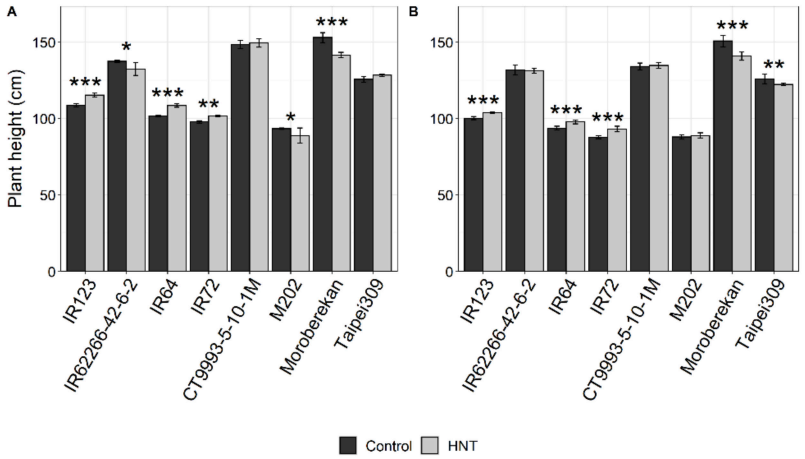


Figure 2. Plant height of the investigated rice cultivars under control and HNT conditions in the WS (**A**) and DS (**B**). Bars for the WS represent means ± SEM of 24 plants per condition, and bars for the DS, those of 60 plants per condition. Cultivars are sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8). Significance levels are indicated by asterisks: 0.001 < ***; 0.001 > ** < 0.01; 0.01 > * < 0.05.

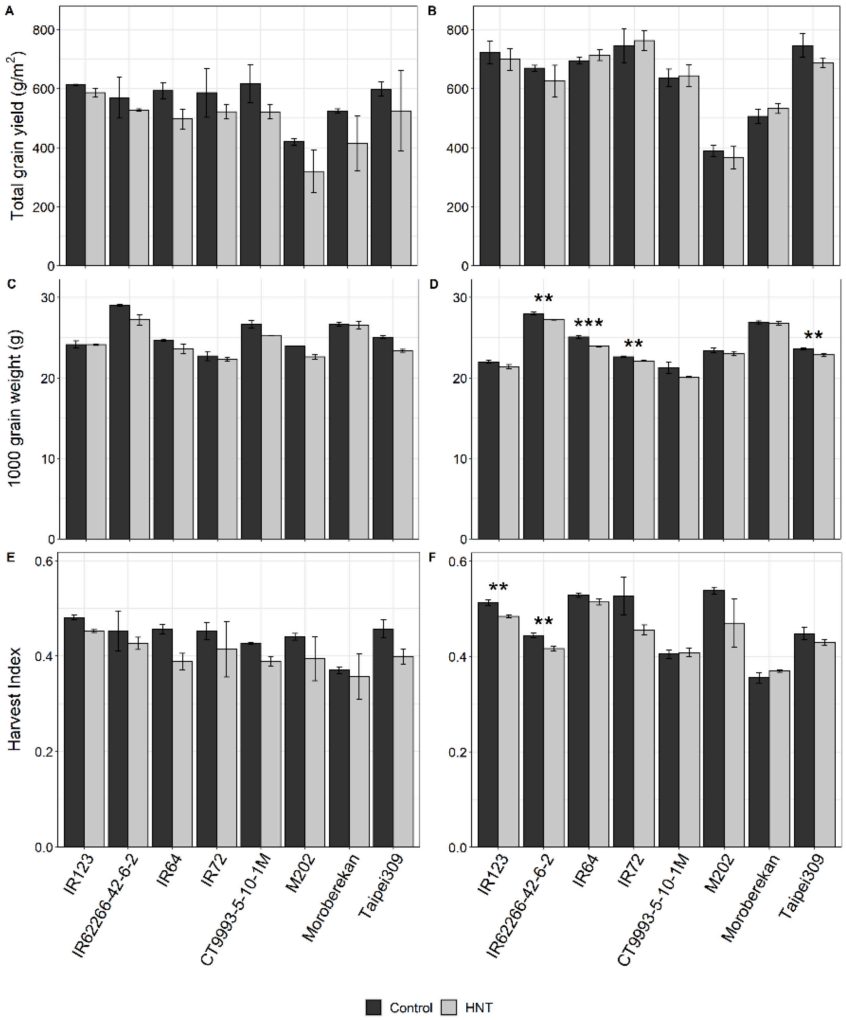


Figure 3. Grain yield (**A**,**B**), 1000-grain weight (**C**,**D**) and harvest Index (**E**,**F**) of eight rice cultivars under control and HNT conditions in the WS (**A**,**C**,**E**) and DS (**B**,**D**,**F**). For the WS, bars represent the means and error bars, the range of two replicates generated from 12 plants, each. For the DS, the bars represent the means ± SEM of five replicates generated from 12 plants, each. Cultivars are sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8). Significance levels were only calculated for the DS due to an insufficient replicate number in the WS and are indicated by asterisks: 0.001 < ***; 0.001 > ** < 0.01; 0.01 > * < 0.05.

A significant negative HNT treatment effect was also found for the 1000-grain weight in both growth seasons (Table 2), with the highest reductions in the WS of about 1.7 and 1.8 g for Taipei309 and IR62266-42-6-2, respectively (Figure 3C,D, Table 2).

The harvest index was significantly affected by HNT across all cultivars in both seasons (Table 2) and showed an overall reduction, except for Moroberekan in the WS and DS and CT9993-5-10-1M only in the DS (Figure 3E,F, Table 2). Furthermore, a significant treatment effect was determined for biomass, spikelets per m^2 and spikelets per panicle only in the WS, but not in the DS (Table 2). For cultivar-specific changes in these parameters, see Figure A4.

2.2. HNT's Effects on the Metabolome Are More Pronounced in Panicles Than in Flag Leaves

Profiling of hydrophilic small metabolites was performed by gas chromatography-mass spectrometry (GC-MS) on flag leaves and panicles of all eight cultivars grown in both seasons. Since it has been shown previously that the metabolite profiles of rice flag leaves and panicles differ widely, making meaningful direct comparisons impossible [49], we treated the data from the two

organs separately. After the pre-processing of both data sets, a total of 76 metabolites for flag leaves and 69 for panicles were determined that were detected in both seasons. Principal Component Analysis (PCA) indicated that metabolite profiles of flag leaves were not strongly affected by HNT conditions in either the WS or DS (Figure 4A,B).

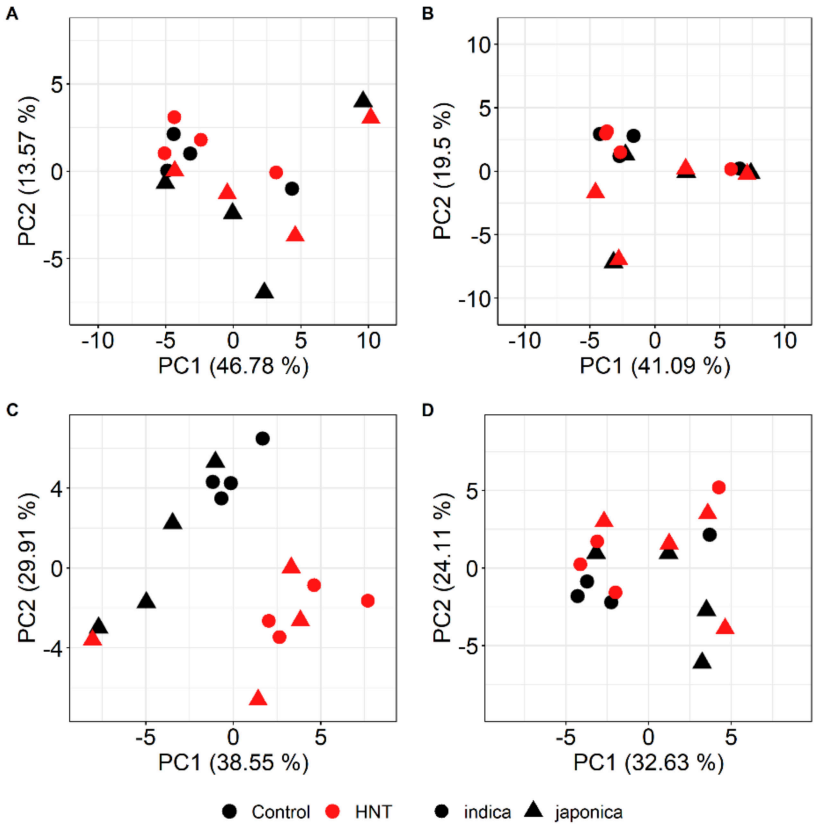


Figure 4. Score plots of the first two Principal Components (PC1 and PC2) from the Principal Component Analysis (PCA) of the metabolite profiles of rice flag leaves (**A**, **B**) and panicles (**C**, **D**) of the eight investigated rice cultivars under control and HNT conditions in the WS (**A**, **C**) and DS (**B**, **D**). For flag leaves, means of the median-normalized and $\log_2$-transformed mass spectral intensities of 76 metabolites, and for panicles, those of 69 metabolites, were used. Numbers in parentheses indicate the fractions of the total variance explained by the respective PCs.

Instead, a separation between cultivars belonging to the subspecies *indica* and *japonica* was visible for both seasons and treatments. By contrast, a clear separation along PC1, explaining 38.55% of the total variance in the data set, between samples from plants grown under control or HNT conditions was observed for panicles collected in the WS (Figure 4C). The single outlier represents the *japonica* cultivar Moroberekan under HNT conditions. For the DS experiment, samples from panicles under different night temperature conditions were separated by PC2, explaining 24.11% of the variance, while PC1 separated the subspecies, explaining 32.63% of the total variance (Figure 4D).

The metabolite composition already varied under control conditions between the two growth seasons in both flag leaves and panicles (Figure 5). Of the 76 metabolites identified in flag leaves, 48 (63%) showed a significantly different content in at least three cultivars in this analysis, while of the 69 metabolites in panicles, 28 (41%) differed between seasons. Only eight of these metabolites (malic acid, A159003, A221004, cis-4-hydroxycinnamic acid, trans-4-hydroxycinnamic acid, fructose-6-phosphate, glyceric acid-3-phosphate and raffinose) were identical in both organs, indicating highly organ-specific metabolic reactions to seasonal variations in rice. In addition, there was variation among the cultivars, which was, however, largely independent of the subspecies that the cultivars belong to.

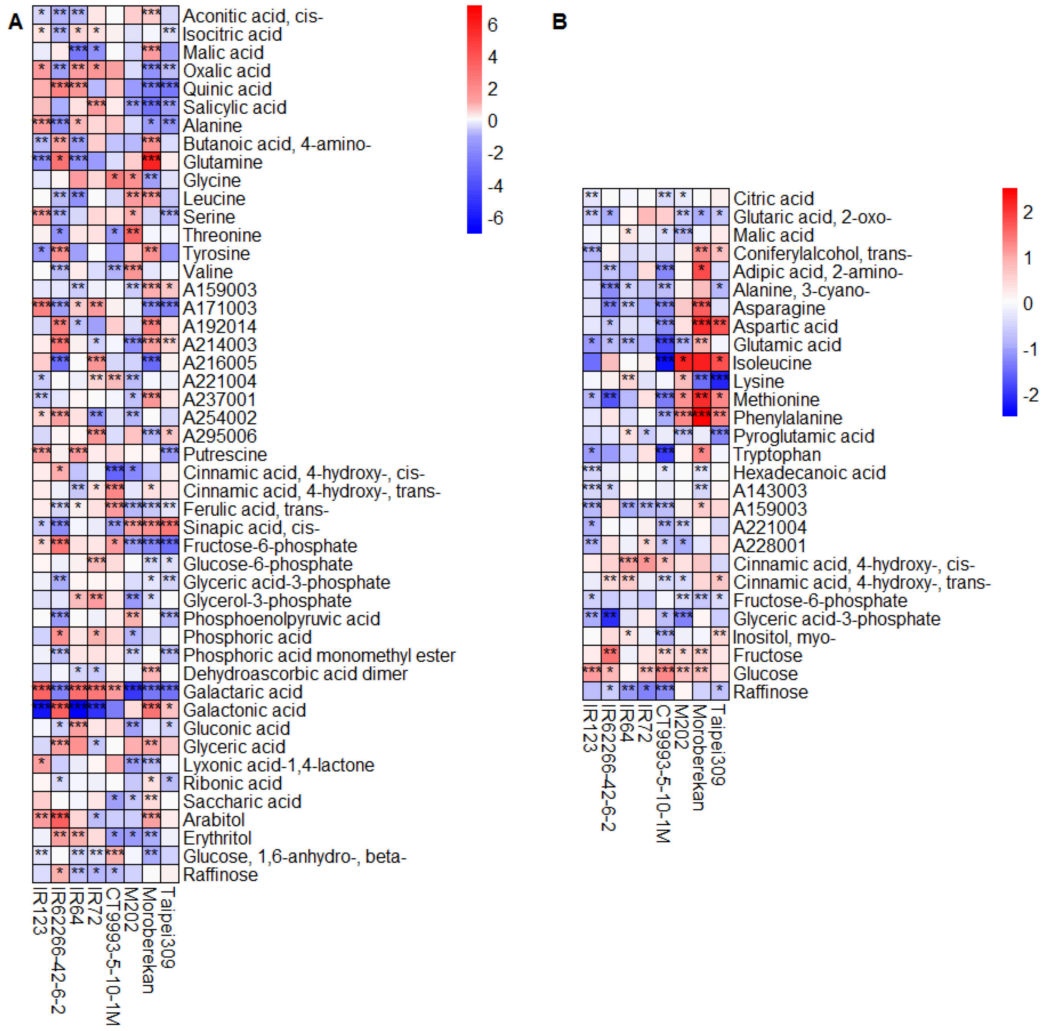


Figure 5. Heat maps showing the $\log_2$ fold changes in metabolite levels under control conditions in the DS compared to the WS for flag leaves (**A**) and panicles (**B**). Only metabolites with a significant change in at least three out of the eight cultivars are displayed. The level of significance is indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$) and the $\log_2$ fold change is represented by the indicated color code. Blue indicates a lower metabolite level in the DS compared to the WS, and red, a higher level. Metabolites are listed alphabetically within the metabolite classes (compare Supplementary Table S2). Cultivars are sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8).

Under HNT conditions, only three metabolites in flag leaves were significantly changed relative to control values in at least three cultivars in the WS, compared to 17 metabolites that were so in the DS (Figure 6). Only erythritol was significantly affected by HNT in both growth seasons. However, while it was increased or unchanged in the DS, it showed a cultivar-specific increase (strongest in Taipei309) or decrease (strongest in IR72) in the WS. In the DS, all metabolites were either reduced/unchanged or increased/unchanged across all cultivars, except for fructose, which was significantly increased in Taipei309 and CT9993-5-10-1M, and significantly decreased in IR64. In addition, while most metabolites showed significant changes in only three or four cultivars, glucose-6-phosphate was significantly reduced under HNT conditions in seven out of the eight cultivars (Figure 6B).

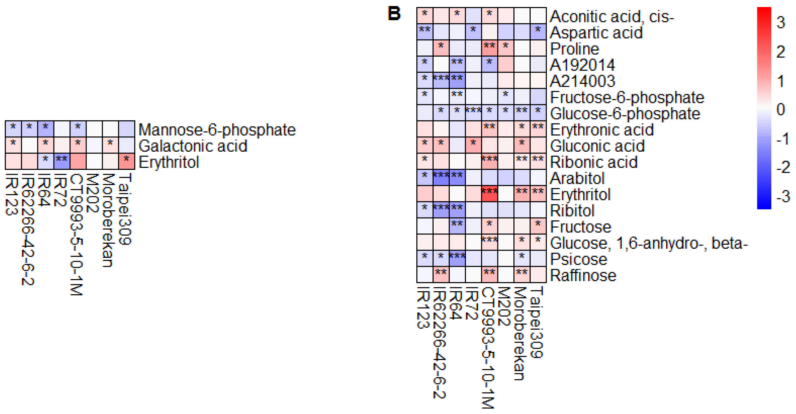


Figure 6. Heat maps showing the $\log_2$ fold changes in metabolite pool sizes in flag leaves under HNT compared to control conditions for the WS (**A**) and DS (**B**). Only metabolites with a significant change in at least three out of the eight cultivars are displayed. The level of significance is indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), and the $\log_2$ fold change is represented by the indicated color code. Blue indicates a lower metabolite level under HNT compared to under control conditions, and red, a higher level. Cultivars were sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8).

In panicles, metabolite changes caused by HNT conditions were more pronounced than in leaves, with higher $\log_2$ fold changes and a larger number of significantly changed metabolites—25 during the WS and 12 during the DS. In addition to the larger number of metabolites that were significantly affected by HNT in the WS than in the DS, opposite to what we observed in flag leaves (Figure 6), changes were generally also larger in the WS than in the DS in panicles (Figure 7).

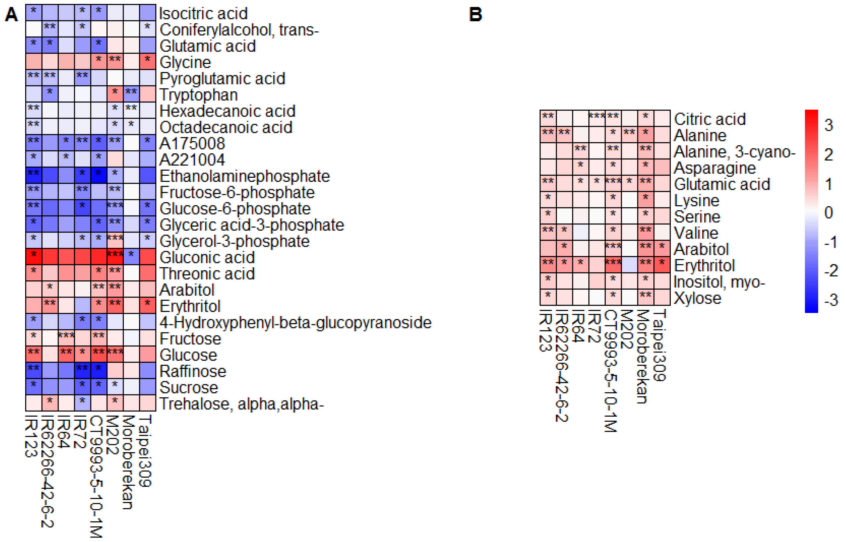


Figure 7. Heat maps showing the $\log_2$ fold changes in metabolite pool sizes in panicles under HNT compared to control conditions for the WS (**A**) and DS (**B**). Only metabolites with a significant change in at least three out of the eight cultivars are displayed. The level of significance is indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), and the $\log_2$ fold change is represented by the indicated color code. Blue indicates a lower metabolite level under HNT compared to control conditions, and red, a higher level. Cultivars were sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8).

A comparison of the significantly changed metabolites in at least three of the eight cultivars in the DS with those in the WS revealed an overlap of glutamic acid, arabitol and erythritol (Figure 7,

Figure A5). Glutamic acid content was mainly reduced in the WS but increased in the DS, while the polyols arabitol and erythritol were mainly increased by HNT in both seasons. There was very little overlap in the metabolites significantly affected by HNT between flag leaves and panicles, with only erythritol affected in the WS and arabitol and erythritol, in the DS. Interestingly, arabitol showed an opposite behavior in response to HNT in the two organs, with decreased levels in flag leaves and increased levels in panicles.

In the WS, the levels of organic acids; amino acids (except glycine); the phosphorylated intermediates fructose-6 phosphate, glucose-6-phosphate, glyceric acid-3-phosphate and glycerol-3-phosphate; and the sugars raffinose and sucrose were in general reduced during HNT in panicles compared to under control conditions. On the other hand, glycine, gluconic acid, threonic acid, arabitol, erythritol, and fructose and glucose were increased (Figure 7A). In a direct comparison of these significantly changed metabolites in the WS with the metabolite levels in the DS, no reduction of any of these metabolites could be observed in the DS (Figure A6). In the DS, all 12 of the significantly influenced metabolites (mainly amino acids, arabitol, erythritol, citric acid, glutamic acid and xylose) were increased under HNT conditions (Figure 7B).

Alanine and 3-cyano alanine were among the metabolites that were significantly changed under HNT conditions in panicles in the DS, but not in the WS. Alanine is a major storage amino acid under stress conditions [55], and the activity of the alanine biosynthetic enzyme alanine aminotransferase (AlaAT) can influence rice yield [56]. In the WS, the activity of AlaAT was generally reduced under HNT to values of 62% to 96% (except for Moroberekan) compared to under control conditions, which was significant at $p < 0.05$ for IR123 and IR72 (Figure 8A). By contrast, AlaAT activity in the DS reached values of 77% to 137% higher under HNT in comparison to under control conditions and was increased in five out of the eight cultivars, although none of the differences were statistically significant (Figure 8B).

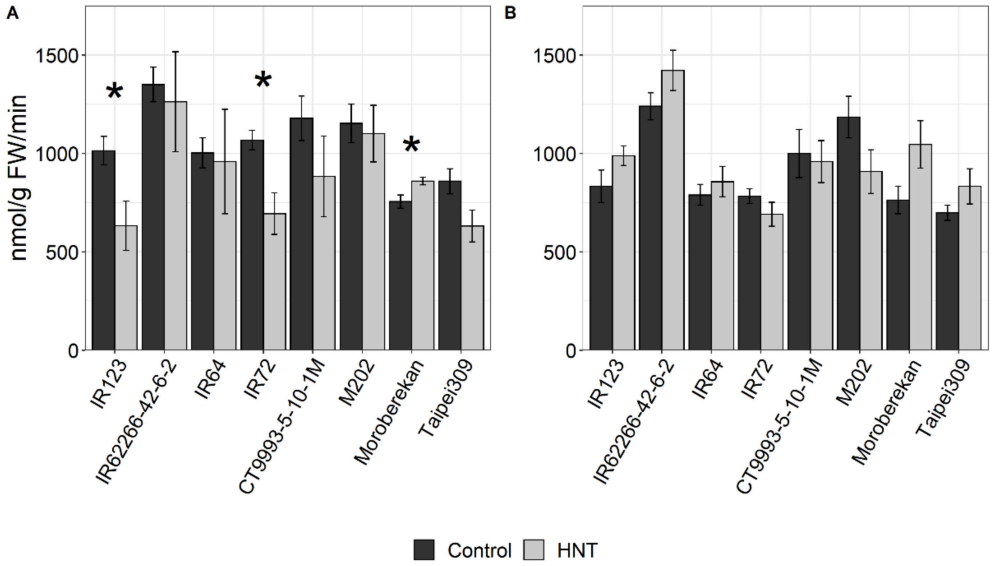


Figure 8. Activity of the enzyme alanine aminotransferase (AlaAT) in panicles of the indicated rice cultivars under control and HNT conditions for the WS (**A**) and DS (**B**). Values are averages of three replicates per cultivar and condition, with four exceptions with two replicates. The level of significance is indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). Cultivars were sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8).

To obtain insight into the potential function of particular metabolites in HNT tolerance in the field and to identify possible candidate marker metabolites for HNT tolerance, we performed correlation analyses between the grain yield reduction in eight cultivars under HNT compared to under control conditions and the change in relative metabolite pool sizes ($\log_2$ fold change) under HNT in the WS,

wherein HNT significantly affected grain yield. While we only identified one significant correlation for metabolites detected in flag leaves (ribitol), we found seven such correlations among panicle metabolites (Figure 9). In addition to one yet unidentified compound, the others comprised four amino acids (including 3-cyano alanine), pyroglutamic acid (representing the sum of pyroglutamate, glutamine and glutamate pools) and fructose-6-phosphate. All eight metabolites showed positive correlations, i.e., a larger change in metabolite pool size indicates a smaller yield loss.

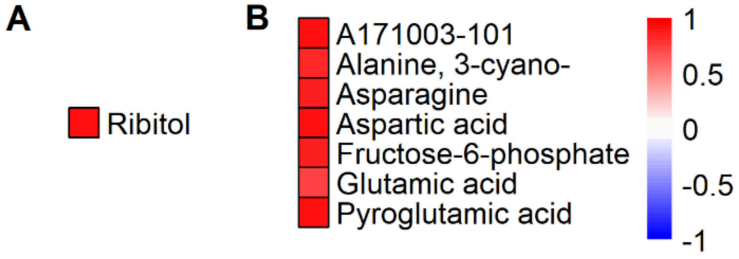


Figure 9. Metabolites with significant correlations (Spearman's rank correlation, $p < 0.05$) between total grain yield reduction under HNT in the WS and the corresponding changes in metabolite contents ($\log_2$ fold change) in flag leaves (**A**) and in panicles (**B**). Red color indicates positive correlations. Metabolites are sorted alphabetically.

3. Discussion

The response of agronomic parameters and metabolic patterns to HNT have been analyzed for eight rice cultivars with different HNT tolerance under field conditions at the IRRI in two different seasons. A comparison of the weather data for both seasons and the respective agronomic parameters identified a slightly longer time to maturity in the DS than in the WS as an important difference. During the DS, plants were exposed to higher radiation intensity and sunshine duration, but lower rainfall and relative humidity compared to in the WS. Similar differences for radiation and sunshine have been reported for a comparison of the DS and the WS from 2005 to 2009 at the IRRI [13]. Furthermore, temperature data for the two growth seasons largely agree between our study and two earlier reports for the same location [13,20], indicating that the plants in our study were exposed to normal climatic conditions without any extreme weather events.

Under control conditions, total grain yield was higher for most cultivars in the DS than in the WS, in agreement with published data [13]. Under HNT conditions, no clear changes in grain yield were observed during the DS, while it was reduced to different degrees in all cultivars in the WS. Under controlled environmental conditions, a yield reduction caused by HNT was previously observed for the cultivars IR62266-42-6-2 and CT9993-5-10-1M, while IR123 showed no change, and IR72 even showed an increased grain yield [28]. During the WS, under our field conditions, IR62266-42-6-2 and CT9993-5-10-1M also showed clear yield reductions of about 22% and 12%, respectively. However, IR123 and IR72 behaved differently under field than under climate chamber conditions, with yield reductions of 16% and 11%, respectively, emphasizing the need for field experiments to determine the effects of stress treatments on rice yield.

A similar influence of the growing season on yield reduction under HNT was previously reported for the *indica* cultivar Gharib and six tropical hybrid cultivars [20]. Additionally, for the tolerant *aus* cultivar N22, a significantly lower yield under HNT was only recorded in the WS. This yield variation was mainly attributed to a reduced grain weight and number of spikelets per m^2, parameters also with significant negative treatment effects during the WS in the present study. Grain yield was reduced in both seasons by around 11% under HNT except for tolerant cultivars in four consecutive years, again partially attributable to a decrease in grain weight [26]. Other authors also highlighted the combination of decreased grain weight, spikelet number per panicle, and biomass production together with a decreased seed set as important for the decline in grain yield under HNT [57]. In general, a reduction in grain weight under HNT conditions was demonstrated for field-grown rice when exposed

to HNT stress from panicle initiation to maturity [18,25–27,58]. In agreement with this, we also found a negative HNT effect on the 1000-grain weight in both seasons under similar stress conditions as used in the previous studies.

Grain yield is influenced by carbon and nitrogen supply to the grain, which are affected by HNT [59]. Temperature-sensitive respiration, known to be increased under HNT (e.g., [28]), might have resulted in increased respiratory carbon loss, previously described to be important during the ripening period [60]. Dark respiration was also considered by other reports to be the main factor affecting biomass and yield under HNT conditions [12,20,25,61] and might be responsible for a decline in assimilation supply to developing grains [57].

This hypothesis is in agreement with the metabolite data obtained during the WS. We found a lower abundance of sucrose and the intermediates of glycolysis, such as glucose-6-phosphate, fructose-6-phosphate, glyceric acid-3-phosphate and glycerol-3-phosphate, whereas the monosaccharides glucose and fructose were increased in panicles. A similar decrease in sugar phosphates, but not in sucrose, was also reported for developing rice caryopses exposed to HNT during the milky stage [48]. Likewise, we also found a significant correlation between the magnitude of the changes in the fructose-6-phosphate content of the panicles under HNT conditions and the yield reduction in the WS. This emphasizes the importance of glycolysis for HNT tolerance in rice.

Glycolysis generates biosynthetic intermediates for respiration. Therefore, a high turnover of glycolysis, as indicated by reduced levels of intermediates, could be expected as respiration is highly increased under HNT. In addition, the products of glycolysis also feed into the TCA cycle, which was shown to be dysregulated in leaves under HNT conditions in climate chamber experiments [51,62]. On the other hand, no significant differences in the metabolites associated with the TCA cycle were found in the developing seeds of different rice cultivars under HNT [53]. Likewise, our data did not provide evidence for an altered TCA cycle under HNT conditions in either panicles or flag leaves.

Interestingly, the effects on glycolysis that we found in panicles in the WS were not observed in either flag leaves in our present study or previously in leaves of the vegetative stage [52]. Apparently, photosynthesis, which is unimpaired under HNT conditions, results in largely unaltered carbohydrate pools in leaves [28,52]. It is therefore reasonable to assume that the carbohydrate supply to the panicles is limiting for grain yield under HNT conditions. Lower sink capacity [26], possibly related to a reduction in the activity of enzymes involved in starch synthesis, has been discussed as a reason for the reduction in grain weight under HNT [63], which we have also observed. A further possibility is an impaired import of sucrose into the panicles under HNT conditions, as has been shown in rice under heat stress [44]. Further experiments will be necessary to test these hypotheses.

The larger reduction in grain yield in the WS compared to in the DS may nevertheless, at least in part, be related to carbohydrate availability. One factor may be faster development during a slightly shorter growing period in the WS, caused by higher daytime T_{min}, preventing the accumulation of sufficient biomass, as shown previously in simulation models [54]. In addition, irradiance levels in the WS were much lower than in the DS, resulting in lower photosynthesis rates [64]. This may have led to a lower overall carbon supply for grain filling [20], leading to lower yield in the WS than the DS under control conditions and a more pronounced effect of HNT on yield in the WS [25] that was mitigated by the higher carbohydrate supply in the DS.

The amino acid alanine was among the significantly increased metabolites in panicles under HNT in the DS but not in the WS. Similarly, alanine was also increased under HNT during early seed development and in the early grain-filling stage in six rice cultivars [53] and in wheat spikes [65]. Alanine is synthesized by the enzyme AlaAT, which catalyzes the reversible synthesis of alanine and 2-oxoglutarate from pyruvate and glutamic acid [66]. It is therefore considered an intercellular nitrogen and carbon shuttle involved in both carbon fixation and nitrogen metabolism [67]. AlaAT is localized in various plant organs and is active in developing rice seeds [68]. The activity of AlaAT is increased in developing rice seeds under heat stress [48], and we observed a moderate increase under HNT conditions in the DS and a moderate decrease in the WS. While the overexpression of *AlaAT* from

barley in rice or canola results in increased nitrogen uptake efficiency and a higher biomass and seed yield compared to in wild type plants [56,66,69–71], a rice mutant of *AlaAT1* exhibits decreased kernel weight [69]. The higher AlaAT activity in the DS may have led to increased nitrogen uptake and assimilation, as described for plants overexpressing *AlaAT* [56], while reduced activity in the WS may have had the opposite effect.

Another metabolite that was significantly increased in response to HNT in the DS, but not in the WS, specifically in panicles, was 3-cyano alanine. This compound is generated by the enzyme 3-cyano alanine synthase (EC 4.4.1.9) during the detoxification of cyanide, which is generated as a by-product of ethylene biosynthesis [72], when the precursor 1-aminocyclopropane-1-carboxylic acid (ACC) is converted into ethylene and hydrogen cyanide (HCN) by the activity of the enzyme ACC synthase [73]. The resulting 3-cyano alanine is then enzymatically converted to asparagine [74], which was also increased under HNT in the DS, indicating a functional detoxification process. Ethylene is a volatile plant hormone that is important for plant growth and development, and various biotic and abiotic stress responses [75]. HCN, on the other hand, is toxic to cells and therefore needs to be efficiently removed [74]. The lower amounts of 3-cyano alanine and asparagine in the panicles collected in the WS might point to a less efficient detoxification of HCN. This is in agreement with the finding that the magnitude of the reduction of both 3-cyano alanine and asparagine in panicles in the WS is significantly correlated with the reduction in grain yield in the WS observed across the eight cultivars. This may indicate that HCN toxicity plays an important role in the HNT sensitivity of panicles. Additionally, however, HCN may play a direct regulatory role in gene expression in low, non-toxic concentrations [76]. Whether this has any impact on HNT tolerance is currently not known.

Two polyols, arabitol and erythritol, were significantly increased in the flag leaves and panicles of almost all cultivars under HNT in both seasons. Both metabolites were also increased under HNT in the vegetative leaves of 12 rice cultivars, including the eight in the present study, in climate chamber experiments [51]. Polyols generally function as compatible solutes and antioxidants under abiotic and biotic stress conditions [77]. Furthermore, arabitol accumulates in flowering spikelets and developing seeds under combined drought and heat stress in the tolerant *aus* cultivar N22 and has a higher content in N22 compared to in sensitive cultivars in flag leaves in the field under control conditions [49]. Similarly, erythritol is accumulated in flowering spikelets and flag leaves under the same conditions, while it is decreased in developing seeds under combined drought and heat stress. Increased levels of erythritol were also found under drought conditions in Arabidopsis [78,79] and in flag leaves of 292 rice accessions [80]. In fact, arabitol and erythritol were both identified as potential metabolic markers for combined drought and heat tolerance [49], and erythritol content under control conditions was the best predictor of drought-induced yield loss in rice [80]. In the present study, however, no correlation between changes in arabitol or erythritol levels and grain yield under HNT was found. The accumulation of these sugar alcohols may therefore be an unspecific response to HNT stress.

4. Materials and Methods

4.1. Plant Material, Cultivation and HNT Stress Treatment

Eight *Oryza sativa* ssp. *indica* (IR123, IR62266-42-6-2, IR64 and IR72) and *japonica* (CT9993-5-10-1M, M202, Moroberekan and Taipei309) cultivars with different HNT tolerance in the vegetative stage under controlled environmental conditions [28] were used (Table 1). IR72, Taipei309 and Moroberekan were characterized as HNT tolerant; IR64, IR123 and CT9993-5-10-1M showed intermediate tolerance; and M202 and IR62266-42-6-2 were sensitive to HNT under these conditions [28]. The seeds for all cultivars were produced at the IRRI. The experiments were carried out during the WS and DS at the IRRI (14°11'N, 121°15'E, 21 MASL) in the Philippines. The seeds were pre-germinated in water after incubation at 50 °C for 3 d to break dormancy and were then sown in seeding trays. Fourteen-day old seedlings were transplanted to the field to a spacing of 0.2 × 0.2 m. The WS experiment was started in June 2011, with four seedlings per hill and each cultivar (42–48 hills) randomly assigned to

two replicate plots per treatment. Phosphorus (15 kg·ha^{-1} p as single superphosphate), potassium (20 kg·ha^{-1} K as KCl), and zinc (2.5 kg·ha^{-1} Zn as zinc sulfate heptahydrate) were applied to all plots as a basal fertilizer a day before transplanting. Nitrogen (N as urea) was incorporated in four splits (30 kg·ha^{-1} as basal, 20 kg·ha^{-1} at mid-tillering, 30 kg·ha^{-1} at panicle initiation (PI), and 20 kg·ha^{-1} just before heading). For the DS experiment, seedlings were transplanted in a staggered approach with one batch in December 2013 and two batches in January 2014. The stagger sowing was based on the phenology data from the first experiment. Each cultivar was randomly assigned to five replicate plots per treatment with one seedling per hill and a total of 28–40 hills per plot. Basal fertilizer (30 kg·ha^{-1} P as single superphosphate, 40 kg·ha^{-1} K as KCl, and 5 kg·ha^{-1} Zn as zinc sulfate heptahydrate) was applied one day before transplanting. N fertilizer as urea was applied in four splits (45 kg·ha^{-1} as basal, 30 kg·ha^{-1} at mid-tillering, 45 kg·ha^{-1} at PI, and 30 kg·ha^{-1} just before heading).

During the day (6 a.m–6 p.m.), plants were exposed to ambient conditions (compare Figure A2 and Table 1). Overnight (6 p.m.–6 a.m.), plants were exposed to the temperature treatments in manually-covered tents with temperature-control devices as described previously [25]. Air conditioners were programmed to maintain the temperature setting at control (22 °C) or HNT (28 °C). Temperature and relative humidity were monitored by sensors connected to data loggers (HOBO, Onset Computer Corporation, Bourne, MA, USA). Temperature treatments started at the panicle initiation stage and lasted until physiological maturity (Figure A1). During the flowering stage, panicles that had flowered for at least 50% were identified and tagged. These were then collected, together with the corresponding flag leaves, the next morning just before the tents were opened (~4 a.m.–6 a.m.). All samples were collected in liquid nitrogen and stored at −80 °C until use.

4.2. Weather Data

Weather data (radiation, sunshine duration, rainfall, relative humidity, maximum temperature (T_{max}) and minimum temperature (T_{min})) recorded by the IRRI wetland agrometeorological station were obtained from the IRRI Climate Unit.

4.3. Growth Analysis, Grain Yield and Yield Components

Twelve hills from each replicate plot were harvested at physiological maturity for the determination of plant height, tiller number, panicle number, and straw and rachis weight and processed for the analysis of yield components [81]. Sixty plants for the DS and 24 plants for the WS were considered for plant height, tiller number and panicle number. For the remaining parameters, two replicates pooled from twelve plants each were considered for the WS, and five replicates pooled from twelve plants each were considered for the DS. The number of panicles per hill was counted for the calculation of panicles per m^2. Afterwards, plants were separated into straw and panicles and panicles were manually threshed. Filled and unfilled grains were submerged in water and separated with a seed blower. Filled, half-filled and empty grains were counted to obtain spikelets per m^2, spikelets per panicle, seed set and 1000-grain weight. Total above ground biomass was determined from the dry weight of straw; rachis; and filled, half-filled and empty grains after drying at 70 °C until constant weight. The harvest index was calculated as the percentage of the dry weight of filled grains relative to the total above ground biomass. Plants from central areas of two m^2 from each plot (two for the WS and five for the DS, per condition and cultivar) were also harvested for the determination of grain yield. Grain weight data were adjusted to a standard moisture content of 0.14 g H_2O g^{-1}.

4.4. Metabolite Profiling and Data Processing

A fraction enriched in small polar metabolites was prepared from 120 mg of fresh weight of snap-frozen and ground flag leaves or panicles from five biological replicates per cultivar and condition and analyzed by gas chromatography coupled to electron impact ionization-time of flight-mass spectrometry (GC/EI-TOF-MS) as described in [82]. Chromatograms were acquired and baseline corrected by the ChromaTOF software (LECO Instrumente GmbH, Mönchengladbach,

Germany). TagFinder [83], the NIST08 software, (http://chemdata.nist.gov/dokuwiki/doku.php?id=start) (U.S. Department of Commerce, Gaithersburg, USA, MD) and the mass spectral and retention time index reference collection of the Golm Metabolome Database [84,85] were used for the manually supervised annotation of metabolites. Mass spectral intensities were normalized to fresh weight and $^{13}C_6$-sorbitol (Sigma-Aldrich, Taufkirchen, Germany) as internal standard. The normalized data are available in Table S2.

Data pre-processing was done separately for both organs and included the omission of metabolites with more than 75% missing values and a missing value imputation for the remaining metabolites with half the minimum amount of the respective mass spectral intensity. Furthermore, contaminations were identified using hierarchical clustering and correlation matrices with a set of known contaminating compounds and removed. A batch effect correction of different measurements of the whole data set was performed using an ANOVA tool [86]. The intensities of each metabolite were divided by the median intensity across all measurements and $\log_2$-transformed to approximate a normal distribution. All presented metabolite data thus represent relative metabolite abundance measures. Outliers were detected with the function *grubbs.test* included in the R-package *outliers* [87] using a threshold of $p < 0.0001$. Finally, 132 metabolite intensities were detected for panicles and 161 metabolite intensities were detected for flag leaves for the DS, and 195 metabolites were detected for both tissues for the WS. For further analysis, the overlap of metabolites per tissue was determined, showing 69 metabolites for panicles and 76 metabolites for flag leaves.

To enable direct comparison, overlapping metabolites for each tissue between both experiments were determined, resulting in 69 metabolites for panicles and 76 for flag leaves.

4.5. Enzyme Activity

The activity of alanine aminotransferase (AlaAT, E.C.2.6.1.2) was measured according to a published method [88]. Ground panicle material (20 mg) was used from three biological replicates per cultivar and condition. In four cases (IR72, IR62266-42-6-2—C, HNT, Moroberekan—C, Moroberekan—HNT), only two replicates were available.

4.6. Statistical Analysis

PCA was perfomed with the R-package *pcaMethods* [89]. For the data processing and visualization, *R v3.4.2* [90] and *R-Studio v1.1.383* [91] were used including the following packages: *ggplot2* [92], *grid* [93], *gridExtra* [94] and *reshape2* [95].

Changes in metabolite content were investigated by calculating the $\log_2$ fold change between the averages of metabolite levels under control conditions in the DS compared to in the WS, or under HNT compared to under control conditions. Unpaired, two-sided t-tests were performed over all replicates, comparing control and HNT conditions to determine the statistical significance of the observed changes. For agronomic data, t-tests were applied for the DS. For the WS, only two replicates were available for most parameters and t-tests were only applied for plant height, tiller number and panicle number. To test the significance of the influence of genotype (G), treatment (T) and GxT interactions across all cultivars, a 2-way ANOVA design was used.

The statistical significance of differences in enzyme activity between control and HNT treatments were evaluated by an unpaired two-sided *t*-test, performed in RStudio [91].

Correlations between total grain yield reduction under HNT in the WS and the corresponding changes in metabolite content ($\log_2$ fold change) were done in R with the package *cor.test* using Spearman Rank Correlation with $p < 0.05$.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/21/9/3187/s1.

Author Contributions: Conceptualization, S.V.K.J., D.K.H., E.Z.; Methodology, L.M.F.L., X.L., A.E., J.K.; Formal analysis, S.S., U.G., X.L., A.E.; Data curation, S.S., A.E., J.K.; Writing, S.S.; D.K.H., E.Z.; Review and editing, all authors, Funding acquisition, S.V.K.J., E.Z., D.K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Federal Ministry for Economic Cooperation and Development through Contracts No. 81141844 and 81206686.

Acknowledgments: We thank Ines Fehrle for her excellent technical assistance with the GC-MS measurements and Jessica Alpers for her excellent support with the enzyme activity measurements.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

ACC	1-aminocyclopropane-1-carboxylic acid
AlaAT	Alanine aminotransferase
DAS	Days after stress
DAT	Days after transplanting
DS	Dry season
DTR	Diurnal temperature range
FC	Fold change
G	Genotype
GABA	4-amino butanoic acid
GC-MS	Gas chromatography – Mass spectrometry
HCN	Hydrogen cyanide
HNT	High night temperature
IRRI	International Rice Research Institute
PC	Principal Component
PCA	Principal Component Analysis
RH	Relative humidity
S	Season
T	Treatment
TCA	Tricarboxylic acid
WS	Wet season

Appendix A

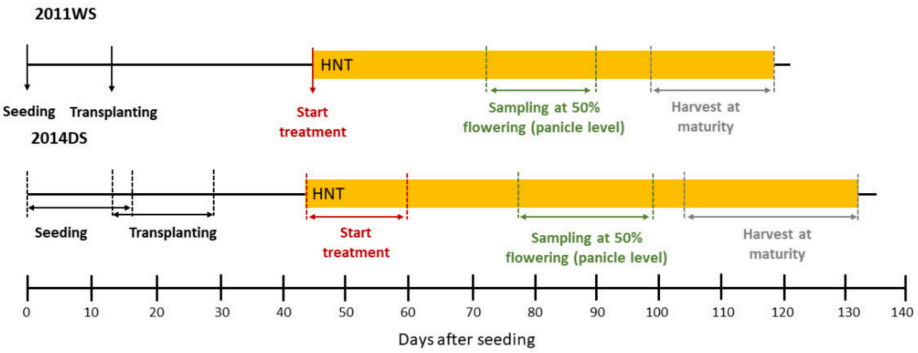


Figure A1. Experimental set-up for the DS and WS experiment. WS—wet season, DS—dry season.

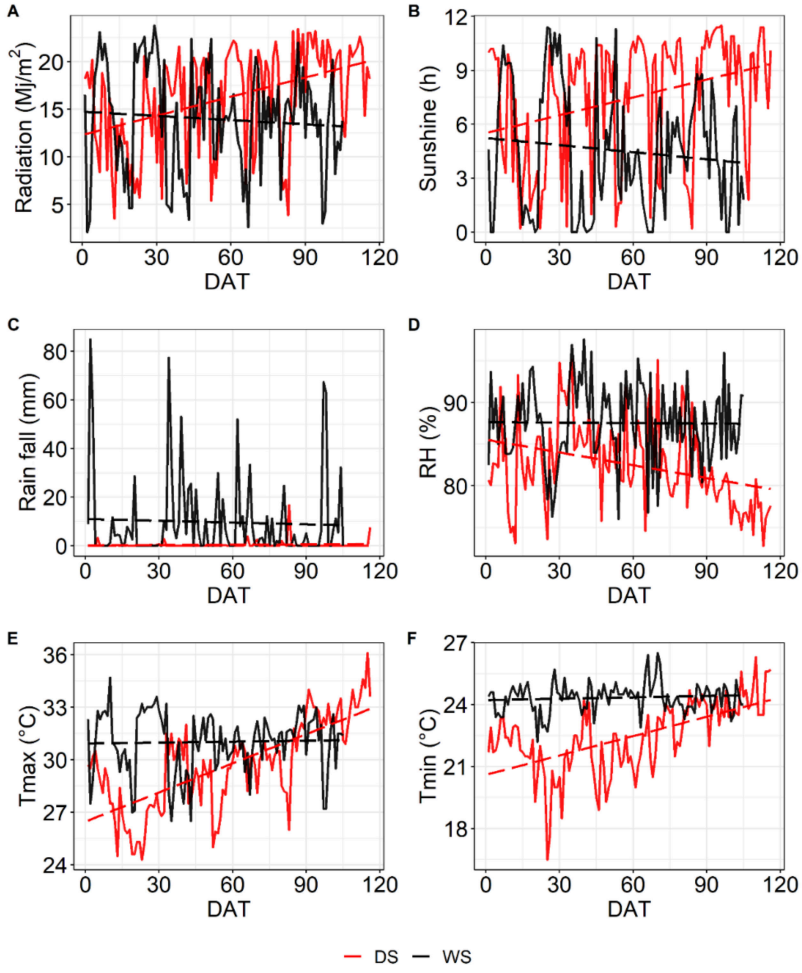


Figure A2. Weather data for the DS and WS experiment measured at the IRRI weather station as average values per day. Radiation (**A**), sunshine duration (**B**), rainfall (**C**), relative humidity (**D**), maximal temperature T_{max} (**E**), minimal temperature T_{min} (**F**). Broken lines represent a trend line for the respective data set. DAT—days after transplanting. Average values for all weather parameters were significantly different ($p < 0.05$) between WS and DS.

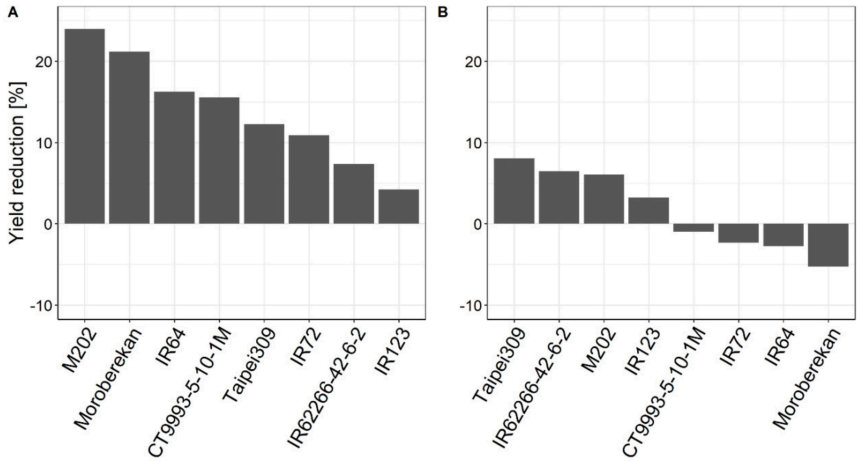


Figure A3. Yield reduction under HNT in the WS (**A**) and DS (**B**). Cultivars are sorted from highest to lowest yield reduction.

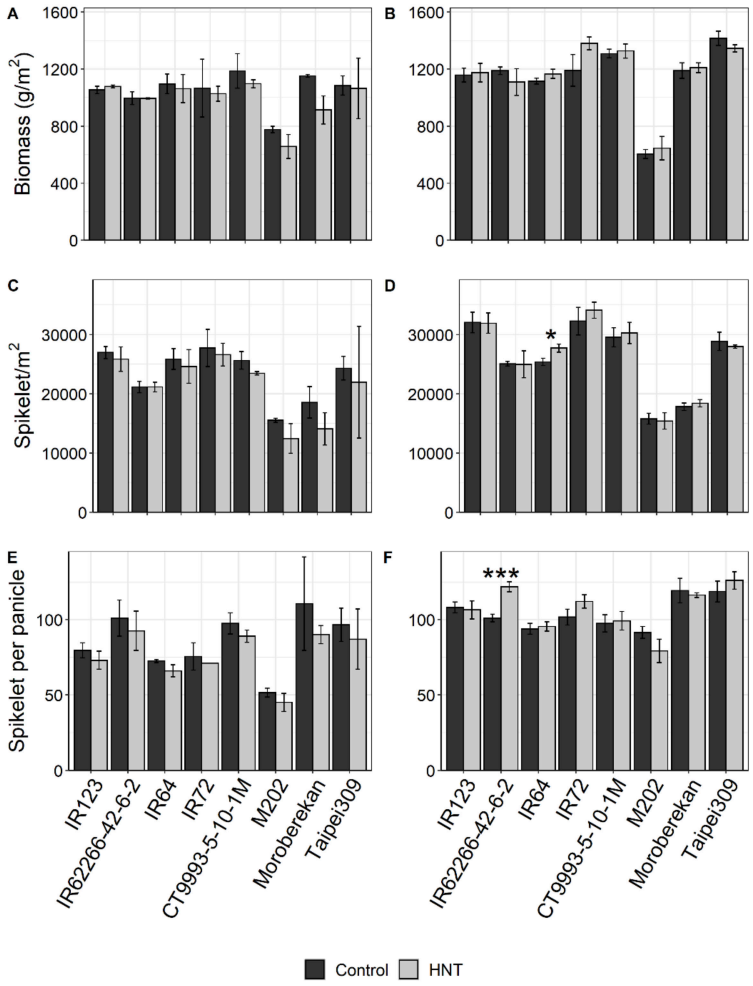


Figure A4. Biomass (**A**, **B**), spikelets per m^2 (**C**, **D**), and spikelets per panicle (**E**, **F**) of eight rice cultivars in response to HNT stress for the WS (**A**, **C**, **E**) and DS (**B**, **D**, **F**). For the WS, variance is displayed as range between means of two replicates with 12 plants each; for the DS, the standard error of the mean of five replicates with 12 plants each is shown. Cultivars are sorted alphabetically within the respective *O. sativa* subspecies *indica* (1-4) and *japonica* (5–8). Significance levels were only calculated for the DS due to the insufficient replicate number in the WS and are indicated by asterisks: 0.001 < ***; 0.001 > ** < 0.01; 0.01 > * < 0.05.

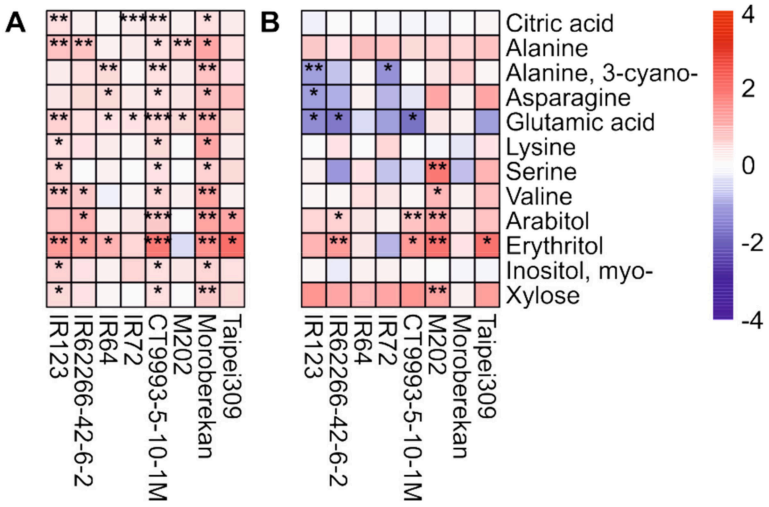


Figure A5. Log_2 fold changes in significantly changed metabolite pools under HNT compared to under control conditions in panicles for the DS (**A**). For comparison, the same metabolites are shown for the WS (**B**) independent of a significant change. For the DS, only metabolites that showed a significant change in at least three out of eight cultivars are displayed in (**A**). The level of significance is indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), and the log_2 fold difference is indicated by the color code. Blue indicates a lower metabolite level compared to under the control condition, and red, a higher level. Cultivars were sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8).

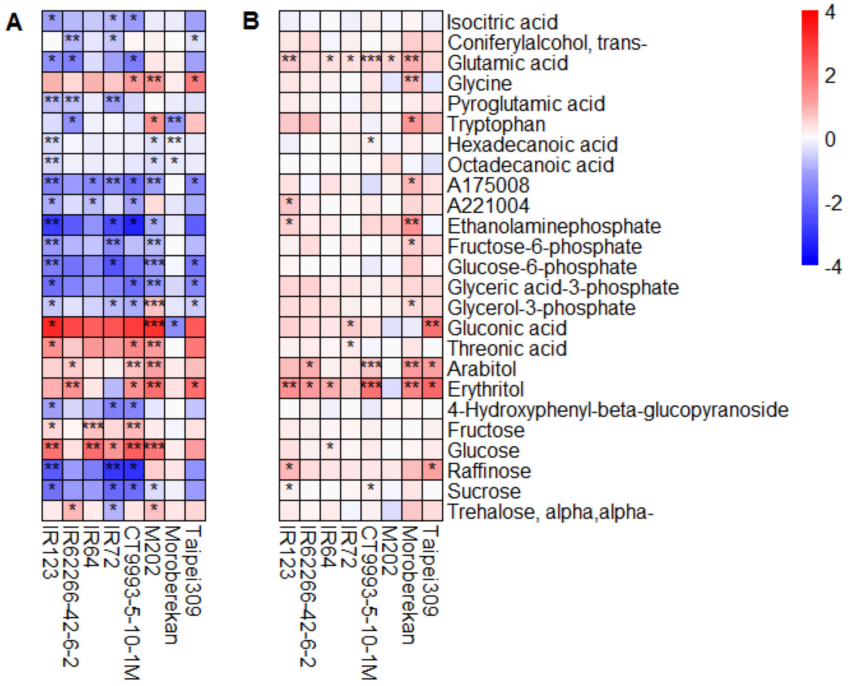


Figure A6. Log_2 fold changes in significantly changed metabolite pools under HNT compared to under control conditions in panicles for the WS (**A**). For comparison, the same metabolites are shown for the DS (**B**) independent of a significant change. For the WS (**A**), only metabolites that showed a significant change in at least three out of eight cultivars are displayed. The level of significance is indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), and the log_2 fold difference is indicated by the color code. Blue indicates a lower metabolite level compared to under the control condition, and red, a higher level. Cultivars were sorted alphabetically within the respective *O. sativa* subspecies *indica* (1–4) and *japonica* (5–8).

References

1. GRISP. *Global Rice Science Partnership*; International Rice Research Institute: Los Baños, Philippines, 2013.
2. FAO. FAOSTAT Database 2009. Available online: http://faostat.fao.org/ (accessed on 23 October 2019).
3. IPCC. *AR5 Climate Change 2014: Impacts, Adaptation, and Vulnerability*; Cambridge Univ. Press: Cambridge, UK, 2014.
4. FAO. FAOSTAT Database 2014. Available online: http://faostat.fao.org/ (accessed on 2 November 2019).
5. Easterling, D.R.; Horton, B.; Jones, P.D.; Peterson, T.C.; Karl, T.R.; Parker, D.E.; Salinger, M.J.; Razuvayev, V.; Plummer, N.; Jamason, P.; et al. Maximum and minimum temperature trends for the globe. *Science* **1997**, *277*, 364–367. [CrossRef]
6. Donat, M.G.; Alexander, L.V. The shifting probability distribution of global daytime and night-time temperatures. *Geophys. Res. Lett.* **2012**, *39*, L14707. [CrossRef]
7. Vose, R.S.; Easterling, D.R.; Gleason, B. Maximum and minimum temperature trends for the globe: An update through 2004. *Geophys. Res. Lett.* **2005**, *32*, L23822. [CrossRef]
8. Sillmann, J.; Kharin, V.V.; Zhang, X.; Zwiers, F.W.; Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J. Geophys. Res. Atmos.* **2013**, *118*, 1716–1733. [CrossRef]
9. Davy, R.; Esau, I.; Chernokulsky, A.; Outten, S.; Zilitinkevich, S. Diurnal asymmetry to the observed global warming. *Int. J. Climatol.* **2017**, *37*, 79–93. [CrossRef]
10. Peng, S.; Piao, S.; Ciais, P.; Myneni, R.B.; Chen, A.; Chevallier, F.; Dolman, A.J.; Janssens, I.A.; Peñuelas, J.; Zhang, G.; et al. Asymmetric effects of daytime and night-time warming on Northern Hemisphere vegetation. *Nature* **2013**, *501*, 88–92. [CrossRef]
11. Zhou, Y.; Ren, G. Change in extreme temperature event frequency over mainland China, 1961–2008. *Clim. Res.* **2011**, *50*, 125–139. [CrossRef]
12. Peng, S.; Huang, J.; Sheehy, J.E.; Laza, R.C.; Visperas, R.M.; Zhong, X.; Centeno, G.S.; Khush, G.S.; Cassman, K.G. Rice yields decline with higher night temperature from global warming. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9971–9975. [CrossRef]
13. Zhao, X.; Fitzgerald, M. Climate change: Implications for the yield of edible rice. *PLoS ONE* **2013**, *8*, e66218. [CrossRef]
14. Rao, B.B.; Chowdary, S.P.; Sandeep, V.M.; Rao, V.U.M.; Venkateswarlu, B. Rising minimum temperature trends over India in recent decades: Implications for agricultural production. *Glob. Planet. Chang.* **2014**, *117*, 1–8. [CrossRef]
15. Padma Kumari, B.; Londhe, A.L.; Daniel, S.; Jadhav, D.B. Observational evidence of solar dimming: Offsetting surface warming over India. *Geophys. Res. Lett.* **2007**, *34*, L21810. [CrossRef]
16. Ambardekar, A.A.; Siebenmorgen, T.J.; Counce, P.A.; Lanning, S.B.; Mauromoustakos, A. Impact of field-scale nighttime air temperatures during kernel development on rice milling quality. *Field Crops Res.* **2011**, *122*, 179–185. [CrossRef]
17. Mohammed, A.R.; Tarpley, L. Effects of night temperature, spikelet position and salicylic acid on yield and yield-related parameters of rice (*Oryza sativa* L.) plants. *J. Agron. Crop Sci.* **2011**, *197*, 40–49. [CrossRef]
18. Nagarajan, S.; Jagadish, S.V.K.; Prasad, A.S.H.; Thomar, A.K.; Anand, A.; Pal, M.; Agarwal, P.K. Local climate affects growth, yield and grain quality of aromatic and non-aromatic rice in northwestern India. *Agric. Ecosyst. Environ.* **2010**, *138*, 274–281. [CrossRef]
19. Shi, W.; Yin, X.; Struik, P.C.; Solis, C.; Xie, F.; Schmidt, R.C.; Huang, M.; Zou, Y.; Ye, C.; Jagadish, S.V.K. High day- and night-time temperatures affect grain growth dynamics in contrasting rice genotypes. *J. Exp. Bot.* **2017**, *68*, 5233–5245. [CrossRef] [PubMed]
20. Shi, W.; Yin, X.; Struik, P.C.; Xie, F.; Schmidt, R.C.; Jagadish, K.S.V. Grain yield and quality responses of tropical hybrid rice to high night-time temperature. *Field Crops Res.* **2016**, *190*, 18–25. [CrossRef]
21. Lanning, S.B.; Siebenmorgen, T.J.; Counce, P.A.; Ambardekar, A.A.; Mauromoustakos, A. Extreme nighttime air temperatures in 2010 impact rice chalkiness and milling quality. *Field Crops Res.* **2011**, *124*, 132–136. [CrossRef]
22. Morita, S.; Yonemaru, J.; Takanashi, J. Grain growth and endosperm cell size under high night temperatures in rice *(Oryza sativa* L.). *Ann. Bot.* **2005**, *95*, 695–701. [CrossRef]

23. Rehmani, M.I.A.; Wei, G.; Hussain, N.; Ding, C.; Li, G.; Liu, Z.; Wang, S.; Ding, Y. Yield and quality responses of two *indica* rice hybrids to post-anthesis asymmetric day and night open-field warming in lower reaches of Yangtze River delta. *Field Crops Res.* **2014**, *156*, 231–241. [CrossRef]
24. Lobell, D.B.; Ortiz-Monasterio, J.I.; Asner, G.P.; Matson, P.A.; Naylor, R.L.; Falcon, W.P. Analysis of wheat yield and climatic trends in Mexico. *Field Crops Res.* **2005**, *94*, 250–256. [CrossRef]
25. Shi, W.; Muthurajan, R.; Rahman, H.; Selvam, J.; Peng, S.; Zou, Y.; Jagadish, K.S. Source-sink dynamics and proteomic reprogramming under elevated night temperature and their impact on rice yield and grain quality. *New Phytol.* **2013**, *197*, 825–837. [CrossRef] [PubMed]
26. Zhang, Y.; Tang, Q.; Peng, S.; Zou, Y.; Chen, S.; Shi, W.; Qin, J.; Laza, M.R.C. Effects of high night temperature on yield and agronomic traits of irrigated rice under field chamber system condition. *Aust. J. Crop Sci.* **2013**, *7*, 7–13.
27. Yang, Z.; Zhang, Z.; Zhang, T.; Fahad, S.; Cui, K.; Nie, L.; Peng, S.; Huang, J. The effect of season-long temperature increases on rice cultivars grown in the central and Southern regions of China. *Front. Plant Sci.* **2017**, *8*, 1908. [CrossRef] [PubMed]
28. Glaubitz, U.; Li, X.; Köhl, K.I.; van Dongen, J.T.; Hincha, D.K.; Zuther, E. Differential physiological responses of different rice (*Oryza sativa*) cultivars to elevated night temperature during vegetative growth. *Funct. Plant Biol.* **2014**, *41*, 437–448. [CrossRef]
29. Bahuguna, R.N.; Solis, C.A.; Shi, W.; Jagadish, K.S.V. Post-flowering night respiration and altered sink activity account for high night temperature-induced grain yield and quality loss in rice (*Oryza sativa* L.). *Physiol. Plant.* **2017**, *159*, 59–73. [CrossRef] [PubMed]
30. Liang, J.; Xia, J.; Liu, L.; Wan, S. Global patterns of the responses of leaf-level photosynthesis and respiration in terrestrial plants to experimental warming. *J. Plant Ecol.* **2013**, *6*, 437–447. [CrossRef]
31. Mohammed, R.; Cothren, J.T.; Tarpley, L. High night temperature and abscisic acid affect rice productivity through altered photosynthesis, respiration and spikelet fertility. *Crop Sci.* **2013**, *53*, 2603–2612. [CrossRef]
32. Dong, W.; Chen, J.; Wang, L.; Tian, Y.; Zhang, B.; Lai, Y.; Meng, Y.; Qian, C.; Guo, J. Impacts of nighttime post-anthesis warming on rice productivity and grain quality in East China. *Crop J.* **2014**, *2*, 63–69. [CrossRef]
33. Fernie, A.R.; Schauer, N. Metabolomics-assisted breeding: A viable option for crop improvement? *Trends Genet.* **2009**, *25*, 39–48. [CrossRef]
34. Oikawa, A.; Matsuda, F.; Kusano, M.; Okazaki, Y.; Saito, K. Rice Metabolomics. *Rice* **2008**, *1*, 63–71. [CrossRef]
35. Nadella, K.D.; Marla, S.S.; Kumar, P.A. Metabolomics in agriculture. *Omics A J. Integr. Biol.* **2012**, *16*, 149–159. [CrossRef] [PubMed]
36. Krumsiek, J.; Bartel, J.; Theis, F.J. Computational approaches for systems metabolomics. *Curr. Opin. Biotechnol.* **2016**, *39*, 198–206. [CrossRef] [PubMed]
37. Zuther, E.; Koehl, K.; Kopka, J. Comparative metabolome analysis of the salt response in breeding cultivars of rice. In *Advances in Molecular Breeding toward Drought and Salt Tolerant Crops*; Jenks, M.A., Hasegawa, P.M., Jain, S.M., Eds.; Springer: Dordrecht, The Netherlands, 2007; pp. 285–315. [CrossRef]
38. Liu, D.; Ford, K.L.; Roessner, U.; Natera, S.; Cassin, A.M.; Patterson, J.H.; Bacic, A. Rice suspension cultured cells are evaluated as a model system to study salt responsive networks in plants using a combined proteomic and metabolomic profiling approach. *Proteomics* **2013**, *13*, 2046–2062. [CrossRef] [PubMed]
39. Zhao, X.; Wang, W.; Zhang, F.; Deng, J.; Li, Z.; Fu, B. Comparative metabolite profiling of two rice genotypes with contrasting salt stress tolerance at the seedling stage. *PLoS ONE* **2014**, *9*, e108020. [CrossRef] [PubMed]
40. Nam, M.H.; Bang, E.; Kwon, T.Y.; Kim, Y.; Kim, E.H.; Cho, K.; Park, W.J.; Kim, B.G.; Yoon, I.S. Metabolite profiling of diverse rice germplasm and identification of conserved metabolic markers of rice roots in response to long-term mild salinity stress. *Int. J. Mol. Sci.* **2015**, *16*, 21959–21974. [CrossRef]
41. Ma, N.L.; Che Lah, W.A.; Abd Kadir, N.; Mustaqim, M.; Rahmat, Z.; Ahmad, A.; Lam, S.D.; Ismail, M.R. Susceptibility and tolerance of rice crop to salt threat: Physiological and metabolic inspections. *PLoS ONE* **2018**, *13*, e0192732. [CrossRef]
42. Baldoni, E.; Mattana, M.; Locatelli, F.; Consonni, R.; Cagliani, L.R.; Picchi, V.; Abbruscato, P.; Genga, A. Analysis of transcript and metabolite levels in Italian rice (*Oryza sativa* L.) cultivars subjected to osmotic stress or benzothiadiazole treatment. *Plant Physiol. Biochem.* **2013**, *70*, 492–503. [CrossRef]
43. Degenkolbe, T.; Do, P.T.; Kopka, J.; Zuther, E.; Hincha, D.K.; Köhl, K.I. Identification of drought tolerance markers in a diverse population of rice cultivars by expression and metabolite profiling. *PLoS ONE* **2013**, *8*, e63637. [CrossRef]

44. Li, X.; Lawas, L.M.F.; Malo, R.; Glaubitz, U.; Erban, A.; Mauleon, R.; Heuer, S.; Zuther, E.; Kopka, J.; Hincha, D.K.; et al. Metabolic and transcriptomic signatures of rice floral organs reveal sugar starvation as a factor in reproductive failure under heat and drought stress. *Plant Cell Environ.* **2015**, *38*, 2171–2192. [CrossRef]
45. Shu, L.; Lou, Q.; Ma, C.; Ding, W.; Zhou, J.; Wu, J.; Feng, F.; Lu, X.; Luo, L.; Xu, G. Genetic, proteomic and metabolic analysis of the regulation of energy storage in rice seedlings in response to drought. *Proteomics* **2011**, *11*, 4122–4138. [CrossRef]
46. Nam, K.H.; Shin, H.J.; Pack, I.S.; Park, J.H.; Kim, H.B.; Kim, C.G. Metabolomic changes in grains of well-watered and drought-stressed transgenic rice. *J. Sci. Food Agric.* **2016**, *96*, 807–814. [CrossRef] [PubMed]
47. Barnaby, J.Y.; Rohila, J.S.; Henry, C.G.; Sicher, R.C.; Reddy, V.R.; McClung, A.M. Physiological and metabolic responses of rice to reduced soil moisture: Relationship of water stress tolerance and grain production. *Int. J. Mol. Sci.* **2019**, *20*, 1846. [CrossRef] [PubMed]
48. Yamakawa, H.; Hakata, M. Atlas of rice grain filling-related metabolism under high temperature: Joint analysis of metabolome and transcriptome demonstrated inhibition of starch accumulation and induction of amino acid accumulation. *Plant Cell Physiol.* **2010**, *51*, 795–809. [CrossRef] [PubMed]
49. Lawas, L.M.F.; Li, X.; Erban, A.; Kopka, J.; Jagadish, S.V.K.; Zuther, E.; Hincha, D.K. Metabolic responses of rice cultivars with different tolerance to combined drought and heat stress under field conditions. *Gigascience* **2019**, *8*, giz050. [CrossRef]
50. Lawas, L.M.F.; Erban, A.; Kopka, J.; Jagadish, S.V.K.; Zuther, E.; Hincha, D.K. Metabolic responses of rice source and sink organs during recovery from combined drought and heat stress in the field. *Gigascience* **2019**, *8*, giz102. [CrossRef]
51. Glaubitz, U.; Erban, A.; Kopka, J.; Hincha, D.K.; Zuther, E. High night temperature strongly impacts TCA cycle, amino acid and polyamine biosynthetic pathways in rice in a sensitivity-dependent manner. *J. Exp. Bot.* **2015**, *66*, 6385–6397. [CrossRef]
52. Glaubitz, U.; Li, X.; Schaedel, S.; Erban, A.; Sulpice, R.; Kopka, J.; Hincha, D.K.; Zuther, E. Integrated analysis of rice transcriptomic and metabolomic responses to elevated night temperatures identifies sensitivity- and tolerance-related profiles. *Plant Cell Environ.* **2017**, *40*, 121–137. [CrossRef]
53. Dhatt, B.K.; Abshire, N.; Paul, P.; Hasanthika, K.; Sandhu, J.; Zhang, Q.; Obata, T.; Walia, H. Metabolic dynamics of developing rice seeds under high night-time temperature stress. *Front. Plant Sci.* **2019**, *10*, 1443. [CrossRef]
54. Van Oort, P.A.J.; Zwart, S.J. Impacts of climate change on rice production in Africa and causes of simulated yield changes. *Glob. Chang. Biol.* **2018**, *24*, 1029–1045. [CrossRef]
55. Good, A.G.; Muench, D.G. Long-term anaerobic metabolism in root tissue (Metabolic products of pyruvate metabolism). *Plant Physiol.* **1993**, *101*, 1163–1168. [CrossRef]
56. Shrawat, A.K.; Carroll, R.T.; DePauw, M.; Taylor, G.J.; Good, A.G. Genetic engineering of improved nitrogen use efficiency in rice by the tissue-specific expression of alanine aminotransferase. *Plant Biotechnol. J.* **2008**, *6*, 722–732. [CrossRef] [PubMed]
57. Xiong, D.; Ling, X.; Huang, J.; Peng, S. Meta-analysis and dose-response analysis of high temperature effects on rice yield and quality. *Environ. Exp. Bot.* **2017**, *141*, 1–9. [CrossRef]
58. Jagadish, S.V.; Murty, M.V.; Quick, W.P. Rice responses to rising temperatures—challenges, perspectives and future directions. *Plant Cell Environ.* **2015**, *38*, 1686–1698. [CrossRef]
59. Mohammed, A.R.; Tarpley, L. Effects of high night temperature and spikelet position on yield-related parameters of rice (*Oryza sativa* L.) plants. *Eur. J. Agron.* **2010**, *33*, 117–123. [CrossRef]
60. Kanno, K.; Makino, A. Increased grain yield and biomass allocation in rice under cool night temperature. *Soil Sci. Plant Nutr.* **2010**, *56*, 412–417. [CrossRef]
61. Coast, O.; Ellis, R.H.; Murdoch, A.J.; Quiñones, C.; Jagadish, K.S.V. High night temperature induces contrasting responses for spikelet fertility, spikelet tissue temperature, flowering characteristics and grain quality in rice. *Funct. Plant Biol.* **2015**, *42*, 149–161. [CrossRef]
62. Liao, J.-L.; Zhou, H.-W.; Peng, Q.; Zhong, P.-A.; Zhang, H.-Y.; He, C.; Huang, Y.-J. Transcriptome changes in rice (*Oryza sativa* L.) in response to high night temperature stress at the early milky stage. *Bmc Genom.* **2015**, *16*, 18. [CrossRef]

63. Dong, W.; Tian, Y.; Zhang, B.; Chen, J.; Zhang, W. Effects of asymmetric warming on grain quality and related key enzymes activities for *japonica* rice (Nanjing 44) under FATI facility. *Acta Agron. Sin.* **2011**, *37*, 832–841. [CrossRef]
64. Chen, Y.; Murchie, E.H.; Hubbart, S.; Horton, P.; Peng, S. Effects of season-dependent irradiance levels and nitrogen-deficiency on photosynthesis and photoinhibition in field-grown rice (*Oryza sativa*). *Physiol. Plant.* **2003**, *117*, 343–351. [CrossRef]
65. Impa, S.M.; Vennapusa, A.R.; Bheemanahalli, R.; Sabela, D.; Boyle, D.; Walia, H.; Jagadish, S.V.K. High night temperature induced changes in grain starch metabolism alters starch, protein, and lipid accumulation in winter wheat. *Plant Cell Environ.* **2020**, *43*, 431–447. [CrossRef]
66. Good, A.G.; Johnson, S.J.; De Pauw, M.; Carroll, R.T.; Savidov, N.; Vidmar, J.; Lu, Z.; Taylor, G.; Stroeher, V. Engineering nitrogen use efficiency with alanine aminotransferase. *Can. J. Bot.* **2007**, *85*, 252–262. [CrossRef]
67. Beatty, P.H.; Shrawat, A.K.; Carroll, R.T.; Zhu, T.; Good, A.G. Transcriptome analysis of nitrogen-efficient rice over-expressing alanine aminotransferase. *Plant Biotechnol. J.* **2009**, *7*, 562–576. [CrossRef] [PubMed]
68. Kikuchi, H.; Hirose, S.; Toki, S.; Akama, K.; Takaiwa, F. Molecular characterization of a gene for alanine aminotransferase from rice (*Oryza sativa*). *Plant Mol. Biol.* **1999**, *39*, 149–159. [CrossRef] [PubMed]
69. Zhong, M.; Liu, X.; Liu, F.; Ren, Y.; Wang, Y.; Zhu, J.; Teng, X.; Duan, E.; Wang, F.; Zhang, H.; et al. FLOURY ENDOSPERM12 encoding alanine aminotransferase 1 regulates carbon and nitrogen metabolism in rice. *J. Plant Biol.* **2019**, *62*, 61–73. [CrossRef]
70. Beatty, P.H.; Carroll, R.T.; Shrawat, A.K.; Guevara, D.; Good, A.G. Physiological analysis of nitrogen-efficient rice overexpressing alanine aminotransferase under different N regimes. *Botany* **2013**, *91*, 866–883. [CrossRef]
71. Selvaraj, M.G.; Valencia, M.O.; Ogawa, S.; Lu, Y.; Wu, L.; Downs, C.; Skinner, W.; Lu, Z.; Kridl, J.C.; Ishitani, M.; et al. Development and field performance of nitrogen use efficient rice lines for Africa. *Plant Biotechnol. J.* **2017**, *15*, 775–787. [CrossRef]
72. Lai, K.W.; Yau, C.P.; Tse, Y.C.; Jiang, L.; Yip, W.K. Heterologous expression analyses of rice OsCAS in Arabidopsis and in yeast provide evidence for its roles in cyanide detoxification rather than in cysteine synthesis in vivo. *J. Exp. Bot.* **2009**, *60*, 993–1008. [CrossRef]
73. Lim, P.O.; Kim, H.J.; Nam, H.G. Leaf senescence. *Annu. Rev. Plant Biol.* **2007**, *58*, 115–136. [CrossRef]
74. Siegień, I.; Bogatek, R. Cyanide action in plants—from toxic to regulatory. *Acta Physiol. Plant.* **2006**, *28*, 483–497. [CrossRef]
75. Helliwell, E.E.; Wang, Q.; Yang, Y. Ethylene biosynthesis and signaling is required for rice immune response and basal resistance against *Magnaporthe oryzae* infection. *Mol. Plant Microbe Interact.* **2016**, *29*, 831–843. [CrossRef]
76. Yu, L.; Liu, Y.; Xu, F. Comparative transcriptome analysis reveals significant differences in the regulation of gene expression between hydrogen cyanide- and ethylene-treated *Arabidopsis thaliana*. *BMC Plant Biol.* **2019**, *19*, 92. [CrossRef] [PubMed]
77. Tian, L.; Liu, L.; Yin, Y.; Huang, M.; Chen, Y.; Xu, X.; Wu, P.; Li, M.; Wu, G.; Jiang, H.; et al. Heterogeneity in the expression and subcellular localization of POLYOL/MONOSACCHARIDE TRANSPORTER genes in *Lotus japonicus*. *PLoS ONE* **2017**, *12*, e0185269. [CrossRef] [PubMed]
78. Pires, M.V.; Pereira Junior, A.A.; Medeiros, D.B.; Daloso, D.M.; Pham, P.A.; Barros, K.A.; Engqvist, M.K.; Florian, A.; Krahnert, I.; Maurino, V.G.; et al. The influence of alternative pathways of respiration that utilize branched-chain amino acids following water shortage in Arabidopsis. *Plant Cell Environ.* **2016**, *39*, 1304–1319. [CrossRef] [PubMed]
79. Fabregas, N.; Fernie, A.R. The metabolic response to drought. *J. Exp. Bot.* **2019**, *70*, 1077–1085. [CrossRef] [PubMed]
80. Melandri, G.; AbdElgawad, H.; Riewe, D.; Hageman, J.A.; Asard, H.; Beemster, G.T.S.; Kadam, N.; Jagadish, K.; Altmann, T.; Ruyter-Spira, C.; et al. Biomarkers for grain yield stability in rice under drought stress. *J. Exp. Bot.* **2020**, *71*, 669–683. [CrossRef]
81. Yoshida, S.; Forn, D.A.; Cock, J.H.; Gomez, K.A. *Laboratory Manual for Physiological Studies of Rice*; International Rice Research Institute: Los Banos, Philippines, 1976.
82. Dethloff, F.; Erban, A.; Orf, I.; Alpers, J.; Fehrle, I.; Beine-Golovchuk, O.; Schmidt, S.; Schwachtje, J.; Kopka, J. Profiling methods to identify cold-regulated primary metabolites using gas chromatography coupled to mass spectrometry. *Methods Mol. Biol.* **2014**, *1166*, 171–197. [CrossRef] [PubMed]

83. Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* **2008**, *24*, 732–737. [CrossRef]
84. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; et al. GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics* **2005**, *21*, 1635–1638. [CrossRef]
85. Hummel, J.; Strehmel, N.; Selbig, J.; Walther, D.; Kopka, J. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics* **2010**, *6*, 322–333. [CrossRef]
86. Lisec, J.; Romisch-Margl, L.; Nikoloski, Z.; Piepho, H.P.; Giavalisco, P.; Selbig, J.; Gierl, A.; Willmitzer, L. Corn hybrids display lower metabolite variability and complex metabolite inheritance patterns. *Plant J.* **2011**, *68*, 326–336. [CrossRef]
87. Komsta, L. Outliers: Tests for Outliers. R Package Version 0.14. 2011. Available online: https://cran.r-project.org/web/packages/outliers/outliers.pdf (accessed on 16 August 2019).
88. Gibon, Y.; Blaesing, O.E.; Hannemann, J.; Carillo, P.; Höhne, M.; Hendriks, J.H.M.; Palacios, N.; Cross, J.; Selbig, J.; Stitt, M. A robot-based platform to measure multiple enzyme activities in Arabidopsis using a set of cycling assays: Comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell* **2004**, *16*, 3304–3325. [CrossRef] [PubMed]
89. Stacklies, W.; Redestig, H.; Scholz, M.; Walther, D.; Selbig, J. pcaMethods—a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **2007**, *23*, 1164–1167. [CrossRef] [PubMed]
90. RCore, T. *R: A Language and Environment for Statistical Computing, 3.4.2*; R foundation for statistical computing: Vienna, Austria, 2017.
91. RStudio, T. *RStudio: Integrated Development for R*; RStudio: Boston, MA, USA, 2016.
92. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2009.
93. Murrell, P. *R Graphics*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2005.
94. Auguie, B. Gridextra: Miscellaneous Functions for "Grid" Graphics. R Package Version 2.3. Available online: http://CRAN.R-project.org/package=gridExtra (accessed on 16 August 2019).
95. Wickham, H. Reshaping Data with the reshape Package *J. Stat. Softw.* **2007**, *21*, 1–20.

2.3 Paper 3: *De novo* Reconstruction of Transcriptomes of ten *Oryza sativa* Cultivars using PacBio Single-Molecule Real-Time Sequencing

Stephanie Schaarschmidt[1], Axel Fischer[1], Lovely Mae F. Lawas[1,6], Rejbana Alam[2], Endang M. Septiningsih[3,7], Julia Bailey-Serres[2], S. V. Krishna Jagadish[3,4], Bruno Huettel[5], Ellen Zuther[1], Dirk K. Hincha[1]*

[1]Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany

[2]Center for Plant Cell Biology, Department of Botany and Plant Sciences, University of California Riverside, Riverside, CA 92521, USA

[3]International Rice Research Institute, DAPO Box 7777, Metro Manila, Philippines

[4]Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA

[5]Max-Planck Genome Centre Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

[6]Present address: Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA

[7]Present address: Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843, USA

*****Corresponding Author:

Dirk K. Hincha, Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany

Published in October 2020. DOI: https://doi.org/10.3390/ijms21218148

Abstract

Background: Reduced yield under increasing environmental stresses due to climate change poses severe challenges for crop improvement. The wide natural variation present in rice is an important source of genes to facilitate stress tolerance breeding. However, the identification of candidate genes from transcriptome (RNA-seq) studies is hampered by the lack of high-quality genome assemblies for the most stress tolerant cultivars. A more targeted and therefore more cost-effective solution could be the reconstruction of transcriptomes to provide templates to map short reads from Illumina RNA-seq experiments.

Results: We sequenced transcriptomes of ten rice cultivars of the subspecies *aus*, *indica* and *japonica* on the PacBio Sequel I platform. RNA was isolated from different organs of plants grown under control and abiotic stress conditions in climate chambers, net-houses and in the field. *De novo* reference transcriptomes were reconstructed resulting in approximately 37,500 to 54,600 plant-specific high-quality isoforms per cultivar. Isoforms were collapsed to reduce sequence redundancy and collapsed transcriptomes were evaluated e.g. for protein completeness level (BUSCO), transcript length, and number of unique gene loci and transcripts. About 40% of all identified transcripts were novel isoforms compared to the Nipponbare reference transcriptome. For 17% and 28% of transcripts no homologous sequence or a conserved protein domain, respectively, could be identified. About 830 *aus*-specific transcripts were determined, of which 56 were significantly differentially expressed in developing seeds of the drought and heat tolerant *aus* cultivar N22 when well-watered plants were compared to plants subjected to combined drought and heat stress in the field.

Conclusions: The newly generated rice transcriptomes are useful to identify candidate genes for stress tolerance breeding that are not present in the reference transcriptome/genome. In addition, our approach provides a general, cost-effective alternative to genome sequencing for the identification of candidate genes in highly stress tolerant "exotic" genotypes.

Keywords: abiotic stress, dehydrins, natural genetic variation, PacBio Sequel, RNA-seq, SMRT sequencing, transcriptome sequencing, rice (*Oryza sativa*)

Background

Global climate change is causing an increase in the severity and frequency of abiotic stress conditions such as heat, drought and high night temperatures that all have a strong negative impact on crop yield [1-5]. In combination with the increasing world population, plant breeders face the challenging task to develop new cultivars that produce higher yield, with enhanced quality, and accompanied by reduced environmental footprints [6]. Rice (*Oryza sativa*) is the main source of calories for more than half of the world's population, especially for the poorest in Asia [7]. As an important reservoir for genes that may be used for crop improvement, the wide natural genetic diversity within the species and its wild relatives, which is preserved in more than 230,000 rice germplasm accessions maintained in gene banks worldwide [8], is an invaluable resource.

While almost 80% of rice cultivation in the world is based on *indica* varieties [9], the current gold standard genome assembly and annotation is derived from the *japonica* cultivar Nipponbare [10]. Due to the lack of proper genomic assemblies, studies of cultivars from different *Oryza sativa* subspecies have largely been based on this reference genome. For instance, the sequences obtained in the 3,000 Rice Genomes Project [11] were mapped against the Nipponbare genome, excluding all sequences that could not be mapped to this reference [12]. This may have led to the loss of genetic information that is specific to the non-*japonica* subspecies. However, more recently the genomes of cultivars belonging to additional *Oryza sativa* subspecies have been sequenced, such as *indica* (e.g. the cultivars Shuhui498 (R498 genome; [13], Zhenshan 97, and Minghui 63 [14]), or *aus* (e.g. Kasalath [15], N22 [6]) cultivars, although the degree of completeness and annotation remains variable.

In particular, the *aus* subspecies (addressed as a subpopulation within the *indica* subspecies [16]) has been a valuable source of genes underlying traits for disease resistance [17], tolerance to phosphate starvation [18], submergence [19], deep water [20], anaerobic germination [21, 22] and drought [23]. For example, the phosphate-starvation tolerance gene *OsPSTOL1*, the deepwater escape genes *OsSNORKEL1/2* and the submergence tolerance gene *OsSUB1A* were identified in the genomes of *aus* cultivars. Significantly, these genes are absent in the genome sequence of the *japonica* reference cultivar Nipponbare.

During the last years, RNA sequencing (in particular Illumina-based short-read RNA-seq) has emerged as a powerful tool for analyzing transcriptomes to identify genes that show differential expression between unstressed control and various environmental stress conditions. However, the determination of transcript levels from RNA-seq data requires reference genome or transcriptome sequences for read mapping and annotation. In rice, the identification of

differentially expressed genes and transcript isoforms is determined by the reference genome [24]. Obviously, the expression data of any gene that is not represented in the reference genome/transcriptome will be lost from the analysis. This could be particularly relevant when investigating stress-tolerant exotic cultivars, land races or wild rice species, as they may contain tolerance genes not present in the reference cultivar Nipponbare. This would then severely limit the possibility to identify novel candidate genes that can support crop improvement programs.

An obvious solution to this problem would be the sequencing, assembly, and annotation of the required genomes. However, this is still comparatively expensive and time-consuming. Here, we have explored a more targeted approach of sequencing and reconstructing partial transcriptomes of rice cultivars from three different subspecies that can be used as references to map RNA-seq reads from abiotic stress experiments. For this purpose, we have used Pacific Bioscience (PacBio) Single-Molecule Real-Time (SMRT) long-read sequencing technology isoform sequencing (Iso-Seq), belonging to a new generation of sequencing methods that provide full-length transcript sequences with high throughput [25]. It thus offers the ability to sequence transcriptomes without the need for an assembly based on an existing reference genome. This approach has been successfully applied to explore and extend existing plant transcriptomes and annotations for example in sorghum [26], wheat [27, 28], sugarcane [29], wild cotton [30], different panicoid grass species [31], and alfalfa [32].

Data description

We selected ten rice cultivars of the subspecies *aus* (Dular, N22), *indica* (Anjali, IR6226-42-6-2, IR64, IR72) and *japonica* (CT9993-5-10-1M, M202, Moroberekan, Nipponbare) for this study that we have used in previous stress experiments [33-37 and unpublished obervations]. RNA was isolated from different organs and tissues of plants grown under various control and stress conditions in climate chambers, net-houses, and in the field (Table 1 and Additional file 1). It should be stressed that we did not aim to obtain (near) complete transcriptomes, but rather to assemble targeted partial transcriptomes with relevance to the RNA-seq analysis of these stress treatments. Pooled RNA samples were sequenced on the PacBio Sequel I platform. The raw data have been deposited at the NCBI's Sequence Read Archive (SRA) [38] under the BioProject number PRJNA640670 and are freely available. Based on the PacBio isoform data, *de novo* reference transcriptomes were reconstructed resulting in approximately 37,500 to 54,600 plant-specific high-quality isoforms per cultivar. High-quality isoforms were collapsed to reduce redundancy in the sequences using the tools `TAMA`, `cDNA cupcake`, and `cogent`. The collapsed transcriptomes from all three approaches

can be found in the additional folders 1-3. Reconstructed transcriptomes were evaluated regarding their protein completeness level (BUSCO), length, and GC content of the transcripts, as well as the number of unique gene loci. Finally, 834 *aus*-specific transcripts were identified. Among these, we identified 56 transcripts in an Illumina RNA-seq data set (raw reads have been deposited in the NCBI's Gene Expression Omnibus (GEO) [39] under the accession number GSE153030 and are freely available) that were significantly changed in abundance in developing seeds of the *aus* cultivar N22 under combined drought and heat stress in the field [35]. As an example, we provide a more detailed analysis of one of the encoded proteins. Our data indicate that reconstructing targeted partial transcriptomes can indeed aid in the analysis of RNA-seq data and allows identification of rice subspecies-specific candidate genes for stress tolerance traits.

Table 1. Sampling for PacBio isoform sequencing. RNA of ten *Oryza sativa* cultivars from different organs and conditions was extracted and pooled for each cultivar (FL - flag leaves, LE - leaves, PA - panicles, FS - flowering spikelets, DS - developing seeds, SH - sheaths, RO - roots, SO - shoots, PP - pollinated pistils, AN - anthers). Seed database accession numbers (IRTP/IRGC/IRIS ID No.) from the International Rice Research Institute (IRRI) are shown. Plants were grown in climate chambers (CC), net-houses (NH), and/or in the field (F). Cultivars were sorted alphabetically within the subspecies (Subsp.) *aus, indica,* and *japonica.* See Additional file 1 for a more detailed description of all samples used for RNA isolation.

Cultivar	**Subsp.**	**ID No.**	**Organ**										**Set-up**		
			FL	**LE**	**PA**	**FS**	**DS**	**SH**	**RO**	**SO**	**PP**	**AN**	**CC**	**F**	**NH**
Dular	*aus*	IRGC 636	X			X	X							X	
N22		IRTP 3911	X			X	X							X	
Anjali	*indica*	IRTP 23206	X			X	X							X	
IR62266-42-6-2		IRGC 117597	X	X	X	X		X					X	X	
IR64		IRTP 12158	X	X	X				X	X			X	X	X
IR72		IRTP 14747	X	X	X	X		X					X	X	
CT9993-5-10-1M	*japonica*	IRIS 71-1229921	X	X	X	X		X					X	X	
M202		IRGC77142	X	X	X	X		X					X	X	
Moroberekan		IRGC12048	X	X	X	X					X	X	X	X	
Nipponbare		IRGC12731	X	X	X								X	X	

Analysis

De novo reconstruction of transcriptomes

Pooled samples representing mRNAs from ten different *Oryza sativa* cultivars were collected from various tissues and treatments (see Table 1 and Additional file 1) and were sequenced on two or three SMRT cells per cultivar (Table2). In total, between 15.49 and 24.51 gigabases (GB) of sequences were obtained for the different cultivars. Sequence raw data was processed with the software `IsoSeq3` using the steps `ccs` and `lima`, resulting in between 460,340 and 736,747 full-length non-chimeric reads (FLNC, containing 5' primer, 3' primer and poly(A) tail) for the combined SMRT cells per cultivar. After the `IsoSeq3 cluster` and `polish` steps, between 37,951 and 54,684 high-quality (HQ), as well as between 1,233 and 2,170 low-quality (LQ) sequences were obtained. Possible sequence contaminations by non-plant organisms were identified by alignment against the NCBI nucleotide database using `blastn` [40] ($E \leq 1e^{-10}$). Isoforms without a significant hit were aligned against the NCBI protein database using `blastx` [40] ($E \leq 1e^{-10}$). All sequences that showed no significant similarity to sequences from the *Viridiplantae* (green plants) family were removed, resulting in between 37,535 and 54,594 HQ full-length transcripts for further analysis (Table 2).

Table 2. Overview of results from PacBio full-length isoform sequencing from ten *Oryza sativa* cultivars. Identified high (HQ) and low quality (LQ) isoforms were analysed for non-plant contamination using blast. Contaminating sequences (not in the group of *Viridiplantae*) were removed (HQ after filt.). PB - number of PacBio SMRT cells, GB - total number of sequenced basepairs in gigabases, FLNC - full-length non-chimeric reads. Cultivars were sorted alphabetically within the subspecies (Subsp.) *aus, indica,* and *japonica*.

Cultivar	Subsp.	PB	GB	FLNC	HQ	LQ	HQ after filt.
Dular	*aus*	2	18.46	460,340	42,252	1,960	41,396
N22		3	24.17	736,747	54,572	1,807	52,333
Anjali	*indica*	2	15.49	481,094	40,208	1,732	39,438
IR62266-42-6-2		2	22.48	649,085	50,569	1,659	50,510
IR64		2	21.97	622,881	49,633	1,279	49,327
IR72		2	20.31	554,872	44,176	2,170	44,049
CT9993-5-10-1M	*japonica*	2	20.81	620,595	48,537	1,465	48,401
M202		2	24.07	656,740	48,836	1,501	48,676
Moroberekan		2	24.51	675,251	54,684	1,721	54,594
Nipponbare		3	15.65	544,792	37,951	1,233	37,535

It has been shown for the previous PacBio sequencing platform (RSII) that correcting long reads using corresponding RNA-seq data could lead to an increased number of HQ

sequences [26, 28, 29, 32]. This was necessary because of a relatively high rate of LQ sequences with insertions and deletions (InDels). However, the newer PacBio Sequel platform produces a higher sequencing output compared to the RSII, including a higher number of HQ and a lower number of LQ sequences [41] which we have also seen in our own data when comparing it to previous RSII studies [26, 42]. To evaluate, whether InDels could be a problem in our data set, we mapped all uncorrected HQ transcripts with `minimap2` against the genome sequences of the corresponding subspecies. The number of InDels was extracted from the cigar string of the alignment files (Additional file 2). The analysis indicated that the uncorrected sequences showed only a small fraction of InDels (between 0.08% and 0.14%). Because of this low frequency of InDels and the low number of LQ sequences (Table2), further data analysis was performed without error correction and excluding LQ transcripts.

Collapsing redundant isoforms

During library preparation, 5' RNA degradation products can be formed and are subsequently sequenced. These degraded products have the same exonic structure but lack some 5' sequence information and hence yield redundant isoforms that are not associated with technical bias or biological context. To tackle the problem, three different approaches to collapse redundant isoform models were tested, namely `cogent`, `cDNA cupcake`, and `TAMA`. While `cDNA cupcake` and `TAMA` perform collapsing based on a reference genome sequence, `cogent` can be used without a reference sequence. Instead, it reconstructs a coding genome based on the PacBio sequences and maps the same sequences back to the reconstructed genome. Based on this mapping, it then collapses the redundant isoforms using the `cDNA cupcake` algorithm. For `TAMA` and `cDNA cupcake`, transcripts were mapped against the respective *Oryza sativa* subspecies genome sequences using `minimap2`. Only a small number of transcripts were not mapped by these approaches (Table 3). With `cogent`, a much larger number of transcripts (5,441 to 7,979) could not be mapped back against the respective reconstructed coding genomes. In general, all three tools reduced the number of isoforms strongly, by 47.6% (`cDNA cupcake`, Nipponbare) to 68.3% (`cogent`, Dular) after collapsing.

Table 3. Number of isoform models after collapsing with TAMA, cDNA cupcake, and cogent. #Tr. - number of filtered, high-quality isoforms used for collapsing. Cultivars were sorted alphabetically within the subspecies (Subsp.) *aus, indica,* and *japonica*.

			Reference-based				Reference-free	
Cultivar	**Subsp.**	**Reference**	**# Tr.**	**TAMA**	**cDNA cupcake**	**Unmapped**	**cogent**	**Unmapped**
Dular	*aus*	N22	41,396	13,995	18,239	313	13,107	7,340
N22			52,333	18,787	23,954	149	19,026	6,603
Anjali	*indica*	S498	39,438	14,371	18,170	178	13,237	6,476
IR62266-42-6-2			50,510	18,926	23,803	220	18,773	6,913
IR64			49,327	19,064	23,435	1,911	17,874	7,979
IR72			44,049	15,954	20,646	143	15,251	7,426
CT9993-5-10-1M	*japonica*	Nipponbare	48,401	18,789	23,415	223	18,359	6,611
M202			48,676	18,925	23,670	240	18,091	6,695
Moroberekan			54,594	20,604	26,009	268	20,378	7,358
Nipponbare			37,535	16,584	19,674	42	14,345	5,441

Uncollapsed (Figure 1, A) and collapsed (Figure 1, B) isoforms were evaluated by a BUSCO assessment against a set of 430 highly conserved ortholog proteins in plants and shown here for HQ transcripts collapsed with TAMA. Because of the incomplete sampling, between 54% and 27% of the essential proteins were missing, while in the reference transcriptome of Nipponbare (IRGSP) only six essential proteins were missing. The tissue localization of the missing proteins was checked exemplary in the InterPro database [43]. This only provided information on a small fraction of the proteins but those were mostly expressed in roots, flowers, stems and seedlings, or expressed during a specific developmental stage (Additional file 3). Due to our pooling of several RNA samples before library construction, we would also expect to miss rare transcripts due to a dilution effect.

For all cultivars, between 3% and 7% of all identified proteins were fragmented before collapsing. This fraction decreased to 2% to 5% after collapsing (Figure 1). Similarly, the number of complete and duplicated transcripts was reduced in favor of single-copy proteins. While for the uncollapsed isoforms, around 19% (Dular) up to 40% (CT9993-5-10-1M) of the proteins were complete and duplicated, this fraction decreased after collapsing to approximately 8% (Anjali) and 18% (IR64) with a corresponding increase of complete and single-copy proteins. For the IRGSP Nipponbare reference transcriptome the majority of transcripts encoded complete and single-copy proteins. Similar results were obtained for cDNA cupcake (Additional file 4, panel B). For cogent (Additional file 4, panel A) more than 50% of the BUSCO proteins were missing, most likely due to not mapping back to the reconstructed genome.

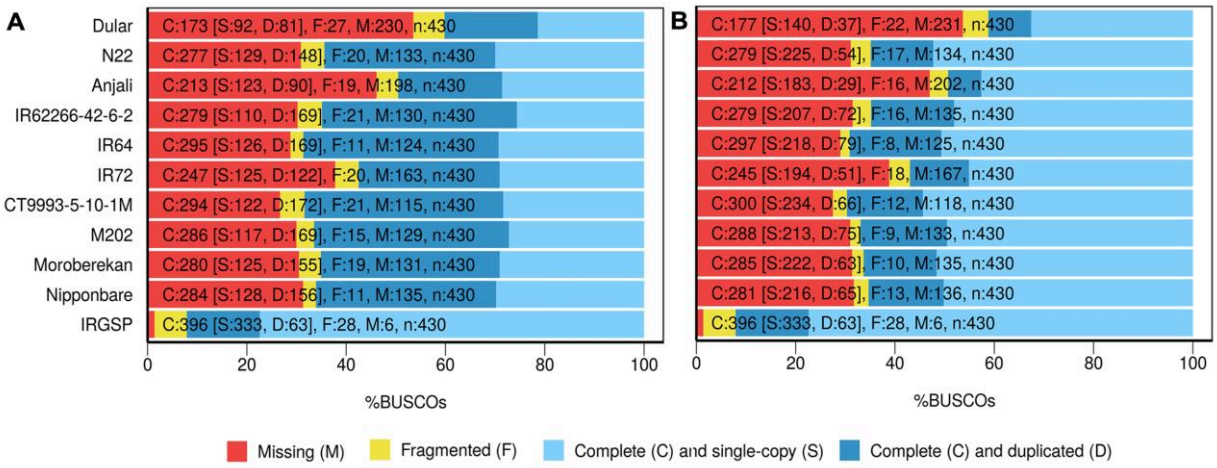


Figure 1. BUSCO assessment analysis of uncollapsed (A) and collapsed (B) transcripts. Results of collapsed transcripts obtained by TAMA are shown. Corresponding results obtained by cDNA cupcake and cogent are shown in Additional file 4. Cultivars were sorted alphabetically within the subspecies *aus*, *indica*, and *japonica*. IRGSP indicates the Nipponbare reference transcriptome.

Through collapsing, the median transcript length increased for all cultivars and for all three methods, as shown for `TAMA` in Additional file 5. The length distribution and median length of the transcripts from each cultivar were more similar to the Nipponbare reference transcriptome after collapsing. Additionally, the number of isoforms per gene locus was determined for all three collapsing methods (Figure 2). `TAMA` yielded the highest fraction of unique isoform models per gene locus, with around 75% for each cultivar. `cDNA cupcake` resulted in around 60%, whereas `cogent`, the reference-free approach, collapsed around 50% of the HQ isoforms into unique isoform models. The relative number of isoforms per gene locus was also determined for the Nipponbare reference transcriptome (IRGSP) resulting in 85% unique isoform models per gene locus.

The three *O. sativa* subspecies *aus, indica*, and *japonica* differ in their genomic sequences and cultivars from the same subspecies are more closely related in their genome sequences [16]. To evaluate genetic distances among our candidate cultivars and to compare the effect of collapsing by different tools, a phylogenetic study was performed. Single nucleotide polymorphisms (SNPs) were called in the collapsed transcriptome datasets based on the IRGSP Nipponbare genome reference and phylogenetic trees were drawn based on an analysis with `SNPhylo` (Figure 3). `SNPhylo` extracts high-quality and representative SNPs for the analysis and resulted in around 30,000 SNPs for `cDNA cupcake`, 23,200 SNPs for `cogent` and around 16,000 SNPs for `TAMA`. For all three approaches, the cultivars of the same subspecies clustered together. The trees constructed from the `cogent` (Figure 3, A) and `TAMA` (Figure 3, C) analyses were more similar to each other than to the tree obtained after collapsing

with `cDNA cupcake` (Figure 3, B). By all three approaches, the *aus* cultivars were clearly separated from the *indica* and *japonica* cultivars. However, the separation between cultivars of the *indica* and *japonica* subspecies was less clear for `cogent` and `TAMA` than for `cDNA cupcake`.

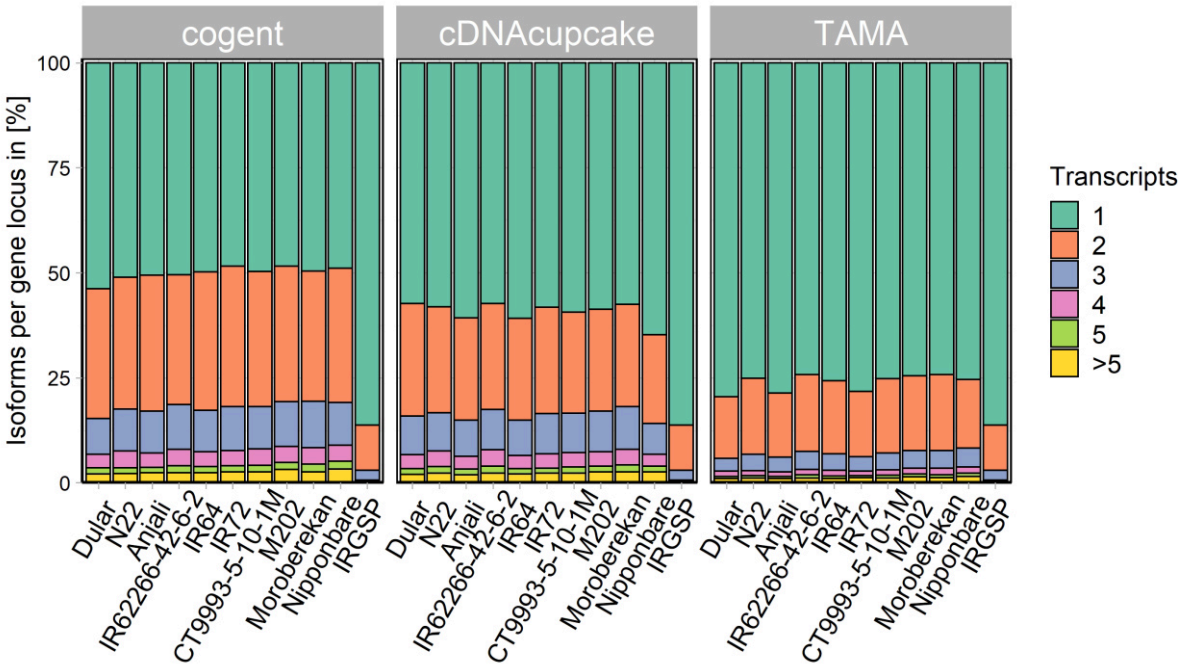


Figure 2. Fraction of isoforms per gene locus for the ten *Oryza sativa* cultivars and the Nipponbare reference transcriptome (IRGSP). Cultivars were sorted alphabetically within the subspecies *aus*, *indica*, and *japonica*.

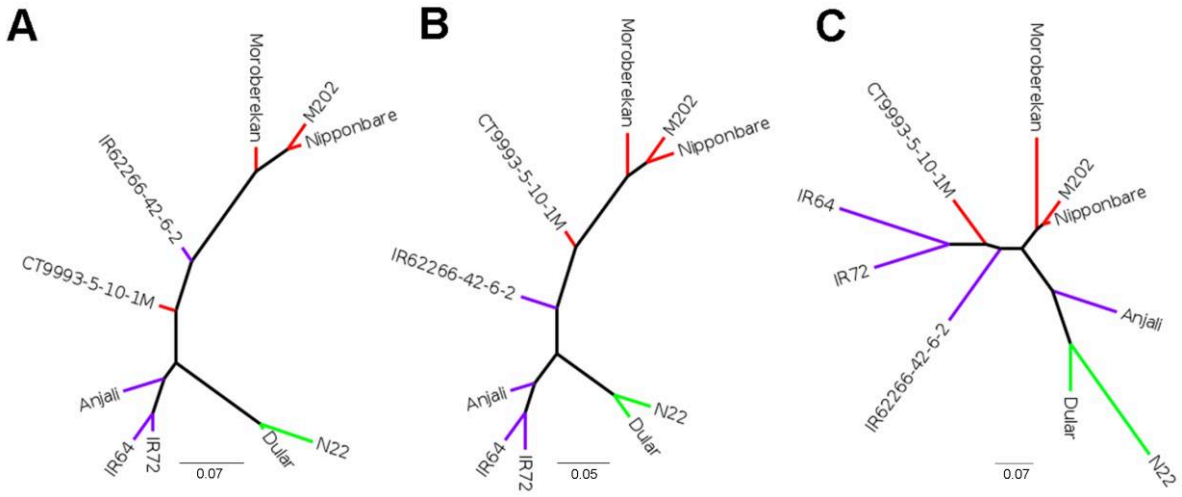


Figure 3. Phylogenetic trees constructed with SNPhylo. Trees are based on SNPs from the transcriptomes of ten *Oryza sativa* cultivars from the subspecies *aus*, *indica* and *japonica* after collapsing redundant transcripts with cogent (A), cDNA cupcake (B) and TAMA (C). Red – *japonica*, purple – *indica*, green – *aus*.

Evaluation of reconstructed transcriptomes

For further biological analysis, collapsed HQ transcripts obtained with `TAMA` were used. Because `TAMA` only collapses transcripts mapped against the reference genome, unmapped transcripts were collapsed additionally with `cogent`. The combined data for each cultivar resulted in 10,511 (Dular) to 15,011 (IR64) reconstructed gene loci as well as between 14,255 (Dular) and 20,803 (Moroberekan) unique isoform models (Table4). Compared to the Nipponbare transcriptome reference (IRGSP), around one third of the gene loci and about half of the transcript models were reconstructed. The average number of transcripts per gene locus was about 1.4 to 1.5 for each cultivar, which was slightly higher than for the reference transcriptome with 1.2. The median transcript length ranged from 986bp (Dular) to 1,394bp (Nipponbare) and was similar to the Nipponbare reference of 1,385bp. The average GC content was between 50.87% (Dular) and 52.76% (IR64), again similar to the reference GC content of 51.24%.

Table 4. Summary of reconstructed transcriptomes including the Nipponbare reference transcriptome (IRGSP). #GL - Number of gene loci, #TR - Number of transcripts, #TR/GL - average number of transcripts per gene locus, Total #bp - total number of bp of all transcripts, Min - shortest transcript in bp, Max - longest transcript in bp, Median - median length of transcripts in bp, GC - content of the nucleotides G and C in %. Cultivars were sorted alphabetically within the subspecies *aus, indica,* and *japonica*.

Cultivar	Sub-species	# GL	# TR	# TR/ GL	Total # bp	Min [bp]	Max [bp]	Median [bp]	GC [%]
Dular	*aus*	10,511	14,255	1.4	15,447,641	56	4,551	986	50.87
N22		13,343	18,913	1.4	26,290,969	62	5,911	1,295	52.26
Anjali	*indica*	10,616	14,499	1.4	17,717,403	75	4,216	1,156	51.99
IR62266-42-6-2		13,227	19,093	1.4	26,791,848	51	7,190	1,314	51.37
IR64		15,011	20,672	1.4	28,663,408	56	6,919	1,299	52.76
IR72		11,647	16,081	1.4	19,678,018	53	5,475	1,149	51.16
CT9993-5-10-1M	*japonica*	13,354	18,963	1.4	26,757,988	55	5,752	1,318	51.97
M202		13,143	19,105	1.5	26,258,012	59	6,644	1,287	51.74
Moroberekan		14,324	20,803	1.5	28,446,682	57	7,072	1,278	51.80
Nipponbare		11,366	16,622	1.5	24,760,098	75	6,035	1,394	52.60
IRGSP	*japonica*	38,866	45,660	1.2	69,184,066	30	16,029	1,385	51.24

The *de novo* reconstructed transcriptomes of the ten *O. sativa* cultivars were compared with the existing Nipponbare reference annotation using `gffcompare`. This tool reports transcripts that fully match, partially match or do not match a reference transcript. A full match transcript has an exact intron-exon-chain matching ("Annotated") to the reference annotation, whereas partially matched transcripts share at least one splice junction with the reference

transcript or show intron retention ("Novel isoform", "Retrained intron"). Additionally, `gffcompare` also reports isoforms on the antisense strand ("Novel antisense") compared to the reference, fully contained exon-chains within a reference intron ("Novel intronic") and on intergenic ("Novel intergenic") regions as well as intron matches on the opposite strand, exonic overlap on the opposite strand, and others ("Novel other"). About 60% of our reconstructed transcripts were fully matched to a known reference transcript of Nipponbare, while around 40% were reported in a broader sense as novel (Additional file 6).

Functional annotation

To get insight into the biological context of the reconstructed transcripts, functional annotation was performed. Open reading frames (ORFs) were predicted using `TransDecoder` (Figure 4), including `blast` and `PFAM` searches, indicating the presence of approximately 60% to 70% complete ORFs (including start and stop codon). Between 26% and 38% 5´ partial ORFs (only start codon), and low percentages of 3´ partial (only stop codon) and internal (neither start nor stop codon) ORFs were additionally identified.

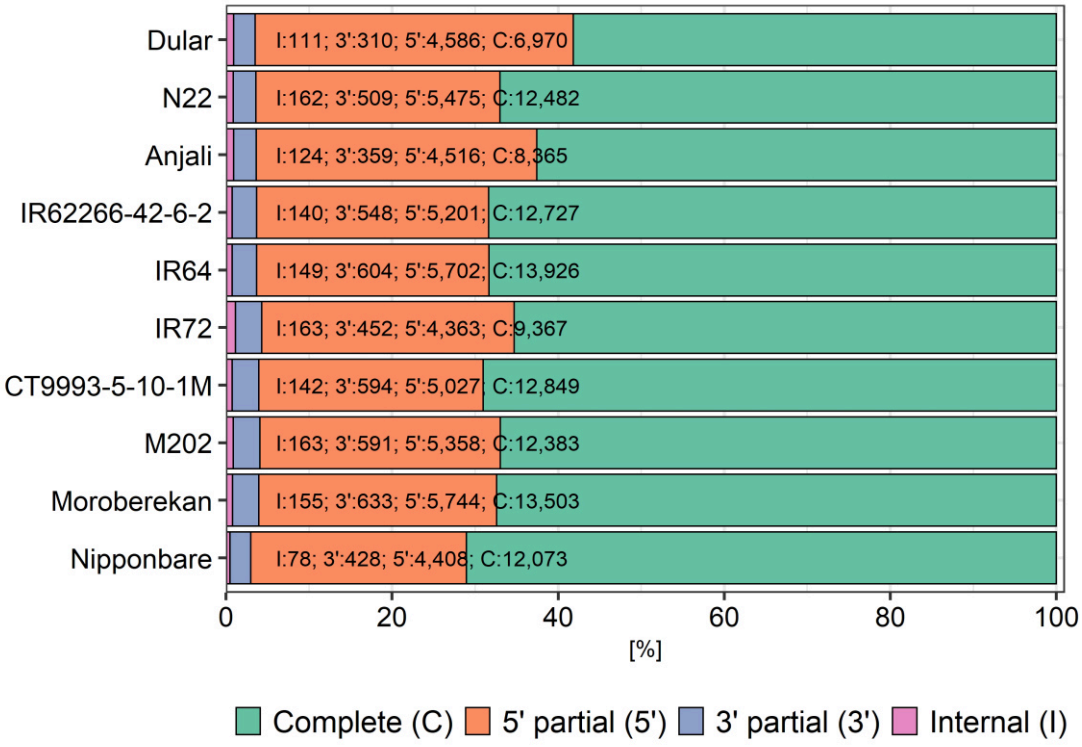


Figure 4. Fraction of predicted open reading frames (ORFs) using TransDecoder. Complete ORFs include start and stop codon, 5' partial/3' partial ORFs contain only the start or the stop codon, respectively, and internal ORFs contain neither start nor stop codon. Numbers represent the number of transcripts for each category per cultivar. Cultivars were sorted alphabetically within the subspecies *aus, indica*, and *japonica*.

Functional annotation was performed with `Trinotate` and `Mercator4`. `Mercator4` was developed specifically for plants and uses a simple hierarchical tree structure of terms referred to as "bins" that describe biological concepts [44]. Major biological processes such as photosynthesis are represented by top-level bins and each offspring bin describes a more narrowly focused subprocess, ending at the single-protein level for each parent bin. Currently, the ontology comprises 27 functional top-level categories representing a diverse range of biological processes in plants. The number of annotated sequences in each `Mercator` bin for the cultivars N22, IR64 and Nipponbare, as representative cultivars for each subspecies, were compared with all known genes for *O. sativa* in the `Mercator` ontology (Figure 5). The relative distribution is similar among the three cultivars and also to the reference. However, the `Mercator` ontology has over 28,000 known *O. sativa* genes (Additional file 7) that have not been assigned to a functional bin and hence, between approximately 8,000 and 10,000 transcripts were also not assigned to functional bins for the three cultivars.

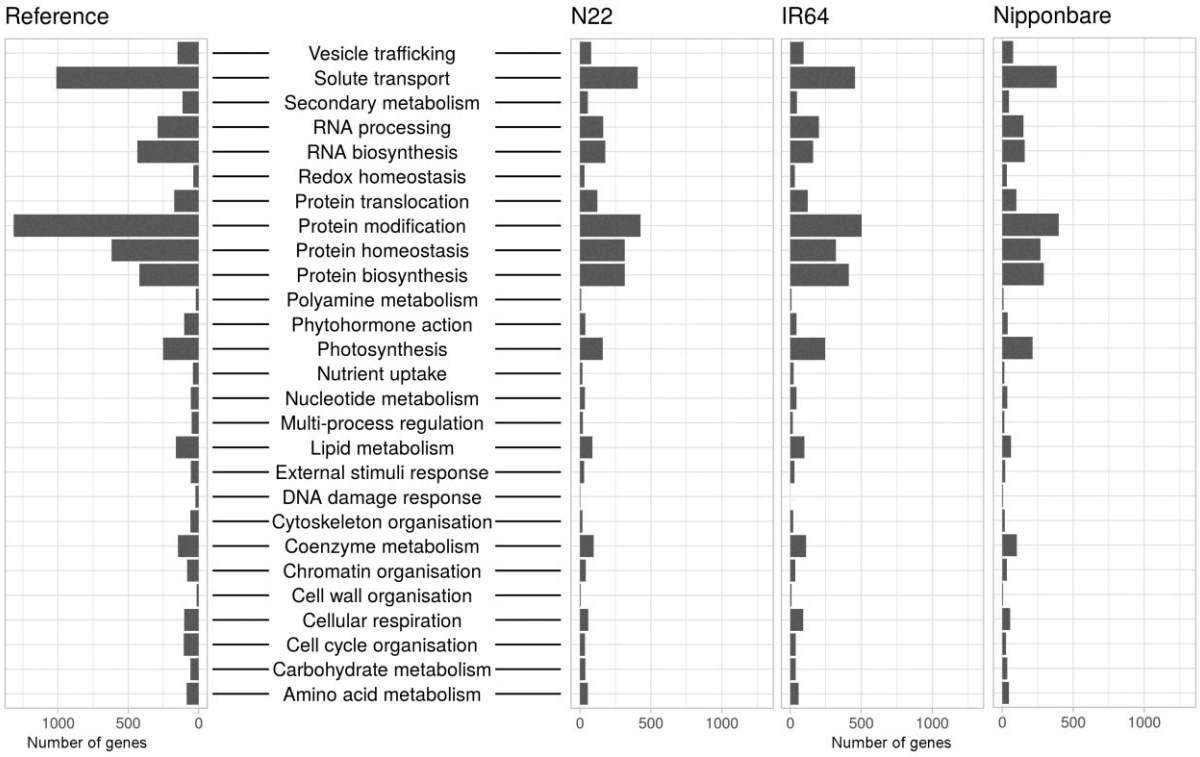


Figure 5. Classification of transcripts into functional bins. Transcripts of N22 (*aus*), IR64 (*indica*) and Nipponbare (*japonica*) were classified into functional bins using Mercator. The bins "not assigned.annotated" and "not assigned.not annotated" are not included.

The complete results of the functional annotation using the `TransDecoder-Trinotate` pipeline and `Mercator` are shown for each cultivar (Additional file 8). The fraction of sequences with at least one significant hit are summarized in Table 5. For

Mercator, blastx, blastp and PFAM retrieved between approximately 60% and 75% significant hits for annotations. For GO terms, only around 38-48% of the transcripts of each cultivar were connected to a functional annotation. Finally, between about 17% and 28% of the transcripts could not be functionally annotated. Because the Swiss-Prot database was used for annotation, which only includes manually curated proteins, data of *Oryza* wild species were mainly not represented. To investigate, whether unannotated transcripts were derived from wild ancestors of *Oryza sativa*, cDNA sequences of all available *Oryza* wild species were downloaded from EnsemblPlants and compiled as a blast database. Unannotated transcripts were searched against it and between 82% and 92% of these transcripts were highly similar to cDNA sequences of *Oryza* wild species.

Table 5. Fraction of transcripts (%) for which at least one significant annotation was found by Mercator or the TransDecoder-Trinotate pipeline (blastx, blastp, PFAM or GO). Also shown is the percentage of transcripts for which no annotation was reported. All unannotated transcripts (No annotation) were additionally compared with an *Oryza* wild species cDNA database using blast. The fraction of unannotated transcripts with a highly similar sequence to an *Oryza* wild species cDNA is shown (Homologs WS). Cultivars were sorted alphabetically within the subspecies (Subsp.) *aus, indica,* and *japonica*.

Cultivar	Subsp.	Mercator	blastx	blastp	PFAM	GO	No annotation	Homologs WS
Dular	*aus*	61.60	65.17	59.57	59.81	37.98	27.60	90.54
N22		68.40	72.05	68.43	70.01	45.52	19.24	91.33
Anjali	*indica*	65.77	69.46	65.43	66.90	43.06	22.03	89.82
IR62266-46-6-2		68.08	71.53	67.16	68.64	44.85	20.19	91.19
IR64		67.78	71.27	67.37	69.55	45.31	20.23	82.03
IR72		63.55	67.20	62.26	63.78	41.22	24.96	88.54
CT9993-5-10-1M	*japonica*	68.57	71.80	67.58	69.24	45.01	19.62	92.43
M202		67.78	71.08	66.69	67.97	44.44	20.68	90.71
Moroberekan		65.72	69.03	64.66	66.85	43.42	22.37	91.68
Nipponbare		71.25	74.35	70.26	72.16	47.59	16.81	91.31

Common and specific transcripts among cultivars

To investigate the question of cultivar-specific transcripts, the transcriptome of one cultivar of each subspecies (N22, IR64, Nipponbare) was used as a blast database and the sequences of the remaining nine cultivars were searched against it. The most highly significant hit for each database entry of each cultivar was selected and the common overlap with all other cultivars was determined (Figure 6). For N22 (Figure 6, A) around 18,000 transcripts were included in the database, of which about 9,000 were highly similar to transcripts from the other

nine cultivars. 652 transcripts were unique to N22 and over 184 transcripts were only found in the *aus* cultivars N22 and Dular. The *aus* specific transcripts were extracted, including their annotations (Additional file 9). For the *indica* cultivar IR64 (Figure 6, B) and the *japonica* cultivar Nipponbare (Figure 6, C) the search space included approximately 15,000 and 20,000 transcripts each, resulting also in around 9,000 common transcripts over all cultivars. While for IR64 2,426 cultivar-specific transcripts were identified, only 349 were determined for Nipponbare (Additional file 9).

Differential gene expression analysis for aus specific transcripts

The *aus* cultivar N22 is particularly tolerant to combined drought and heat stress [35]. We therefore asked whether any of the identified *aus* specific transcripts were regulated under these conditions. A differential gene expression (DGE) analysis was performed for N22 plants grown in the field under control and combined drought and heat stress. RNA-seq was performed using RNA isolated from developing seeds and the resulting Illumina reads were mapped against the *de novo* reconstructed N22 transcriptome. After identifying significantly differentially expressed genes with DESeq2 (FDR $p < 0.1$, absolute $\log_2$ fold-change ≥ 1), 56 *aus* specific genes were extracted (Additional file 10). As determined by a `blast` search, about 46% of these genes had *Arabidopsis thaliana* homologs, 27% lacked any annotation, 11% each were either only described by a PFAM domain or were homologous to sequences in other plant species, while 5% had known homologs in *Oryza*.

As an example, we describe the gene B12288, which was significantly up regulated during combined heat and drought stress (Additional file 10). It has homologous genes in both *japonica* and *indica* cultivars annotated as *RAB21*. The gene is induced by drought and the corresponding protein belongs to the dehydrin family of Late Embryogenesis Abundant (LEA) proteins. Evolutionary relationships with other *Oryza* dehydrins [45] were investigated by multiple sequence alignment and visualized as a tree (Additional file 11). The N22 gene product was closely related to four other dehydrins in wild rice species and *Oryza sativa* ssp. *japonica.* It showed 89.5% sequence coverage and 86.0% sequence identity compared to the *japonica* protein (Figure 7) including the highly conserved repeat regions characteristic of dehydrins [46, 47]. The N22 protein was more similar to the proteins from *Oryza* wild species than to the *japonica* protein (see Figure 7 and Additional file 11).

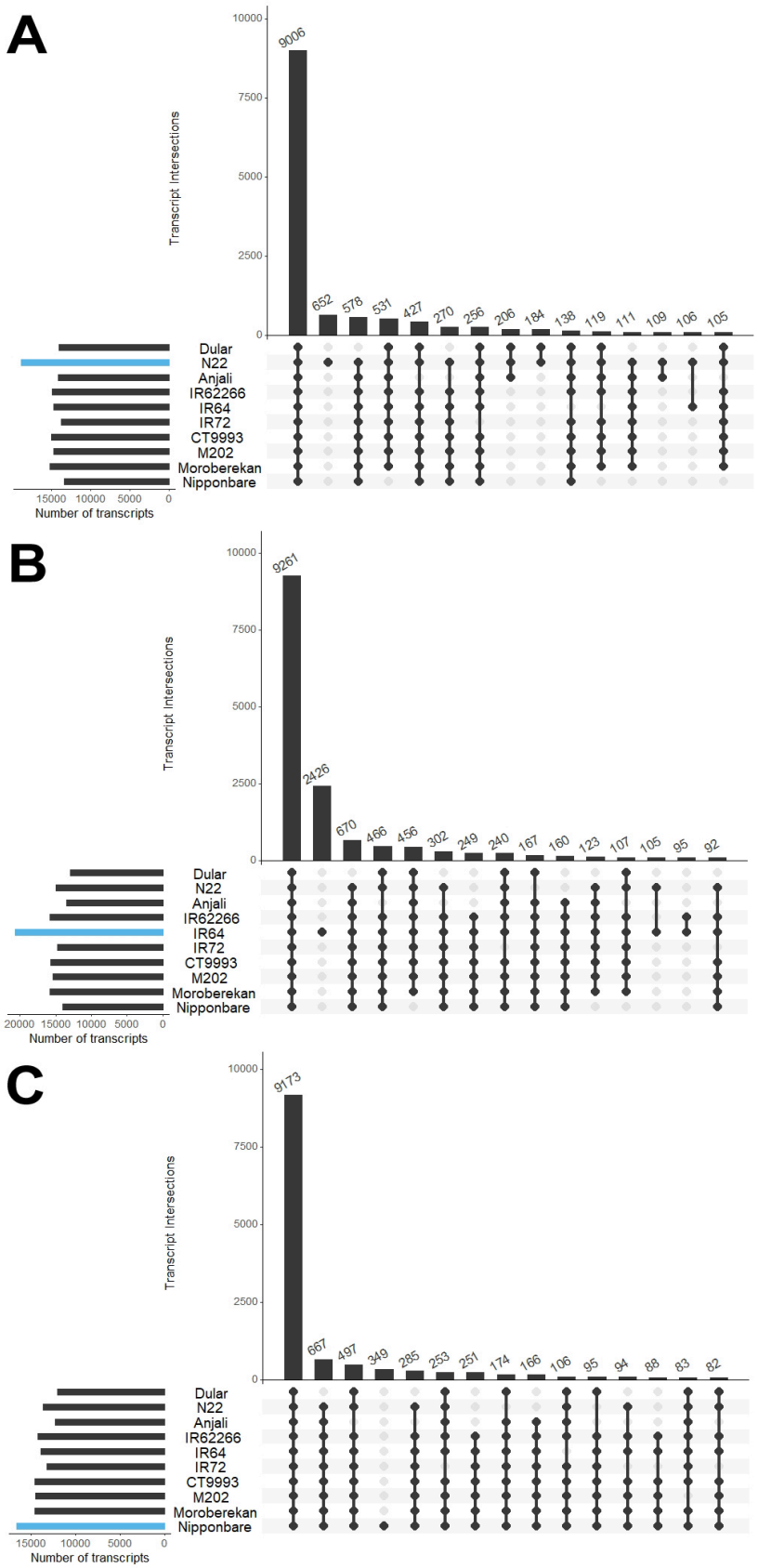


Figure 6. Identified common and specific transcripts over all cultivars.

Sequence similarities were identified by a blastn search using a representative cultivar of each subspecies as a database. The best hit for each database entry was selected based on the cultivars N22 (A), IR64 (B), and Nipponbare (C). The 15 largest categories were visualized in an UpSet plot. The barplots on the left of the cultivar names represent the size of the datasets, with the blue bars indicating the size of the search space. Dots and vertical lines indicate the cultivars included in the overlap. Barplots in the top panels represent the number of transcripts in the respective comparison. Cultivars were sorted alphabetically within the subspecies *aus*, *indica*, and *japonica*.

Figure 7. Multiple sequence alignment of five *Oryza* RAB21 dehydrin proteins. Os – *Oryza sativa* ssp. japonica, OGLUM - *Oryza glumaepatula*, ORUF – *Oryza rufipogon*, ONIVA – *Oryza nivara*, B12288_1_PacBio – RAB21 protein from the *Oryza sativa* ssp. *aus* cultivar N22. The encoding transcript was identified as *aus*-specific in our analysis. Color theme is "identity" by MVIEW [88].

Discussion

Sequencing performance

Between 15.7 and 24.5 GB of cDNA were sequenced for each cultivar on two or three SMRT cells resulting in 460,340 up to 736,747 full-length non-chimeric (FLNC) reads. Using the `IsoSeq3` protocol, between 38,000 and 54,700 high-quality (HQ) transcripts for each cultivar were obtained before filtering out contaminants. Among the ten cultivars, the sequencing output was similar. In previous studies of plant transcriptomes, more SMRT cells were used, but resulted in a similar output of FLNC reads and HQ transcripts. For example, for the wheat cultivar Xiaoyan 81 [42] around 197,800 FLNC reads were obtained on the RSII platform and processed into 91,800 HQ reads based on eight SMRT cells. With the newer PacBio Sequel platform that we also used in our study, around 650,000 FLNC reads were obtained using five SMRT cells analyzing the transcriptome of the wild cotton species *Gossypium australe* [30] but in this case an older chemistry and software were used. Therefore, it is difficult to directly compare the sequencing output from different studies. However, our results indicate that two to three SMRT cells are sufficient to obtain useful IsoSeq data with the currently available technology.

The PacBio technology has a relatively high sequencing error-rate, but these errors are distributed randomly among the sequence [25]. Since sequencing is performed on circularized cDNA molecules, several sequencing passes can be generated for a given cDNA, carrying errors in different random locations. The PacBio `IsoSeq3` tool is then generating a consensus sequence based on the multiply sequenced cDNA template to eliminate these errors. However, even after the correction, InDels and SNPs may still occur. In a study on sorghum [26] using the older RSII technology, HQ reads were mapped against a reference genome sequence and a per-nucleotide error rate of 2.34% was observed. This made a correction using corresponding

RNA-seq data necessary. Using the Sequel technology, we found a per-nucleotide error rate between 0.08 and 0.14% for the uncorrected HQ reads, based on mapping against the respective subspecies reference genome sequences. This low error rate made further correction unnecessary.

Collapsing redundant transcripts and transcriptome quality assessment

During library preparation, degradation products can be formed and are subsequently sequenced. These shorter transcripts lack some of the 5' sequence but are otherwise identical to the full-length transcripts, resulting in large numbers of redundant transcripts. This effect can be reduced experimentally using specific 5' end capturing library preparation methods, or it can be partly compensated computationally by the use of collapsing software. We compared the utility of the tools `cogent`, `cDNA cupcake` and `TAMA` to reduce the number of redundant transcripts. `Cogent` does not need a reference genome sequence to collapse redundant isoforms and was successfully applied to transcriptomes from organisms without an available genome reference [30, 48, 49]. `cDNA cupcake` and `TAMA`, on the other hand, need a reference genome sequence and have been more commonly used [50-53]. In our study, the number of transcripts after collapsing decreased by up to 68%, indicating the necessity to reduce redundancy and thereby improve data quality. While `TAMA` and `cogent` resulted in similar numbers of collapsed transcripts, the numbers were slightly higher after processing with `cDNA cupcake`. `Cogent` left more transcripts unmapped, compared to the other tools. This may be due to the generation of transcript orphans, i.e. putative single-isoform transcripts that were not incorporated into the reconstructed transcriptomes.

Transcriptome quality improvement after collapsing was shown by the `BUSCO` assessment, where the number of encoded complete and single-copy proteins increased by approximately 20% to between about 35% and 55% of all proteins included in `BUSCO`, while for the reference transcriptome this was about 75%. However, as expected, only about 70% of all `BUSCO` proteins were covered by our partial transcriptomes. For comparison, PacBio sequencing of the sugarcane transcriptome [29] using a pooled RNA sample derived from leaf, internode and root tissues at different developmental stages collected from 22 varieties resulted in a coverage of 90% of the `BUSCO` proteins. However, since no collapsing was performed, this study found 66% complete but duplicated `BUSCO` proteins.

Collapsing transcripts with `TAMA` resulted in the highest fraction of one isoform models per gene locus and the average number of isoforms per locus in our different transcriptomes was very similar to the Nipponbare reference transcriptome. This is, however, not always the

case. A PacBio IsoSeq study in maize [54] identified an average of 6.56 isoforms per gene locus using the `cDNA cupcake` pipeline, more than twice the number found for the reference genome annotation with an average of 2.84 transcripts per gene locus. `Cogent` and `cDNA cupcake` yielded lower fractions of one isoform models per gene locus in our study. Since there are, to the best of our knowledge, no other direct comparisons of the three collapsing tools available, it cannot be judged whether the tools may perform differently with different data sets or different reference transcriptomes. Obviously, only `cogent` could be used in cases where no reference genome sequence is available.

Around 70% of the transcripts covered a complete ORF in most of the cultivars. Only Dular and Anjali showed a smaller fraction of complete ORFs. The differences among the cultivars are due to a different fraction of 5´truncated ORFs. In these cases, either the collapsing tool (`TAMA`) has not worked sufficiently, or no full-length ORFs were sequenced for these particular transcripts. Either way, it seems that a certain fraction of incomplete ORFs cannot be avoided, given the methodology we employed in our study. A PacBio IsoSeq study of the chicken transcriptome compared brain and embryo RNA libraries, where both libraries were normalized to reduce over-represented transcripts, but only for the embryo library a 5' cap selection was performed [55]. Here, the number of transcripts dropped by 60% for the brain data and by 21% for the embryo data after collapsing with an older version of `cDNA cupcake`, indicating lower transcript redundancy for the capped library. However, it remains to be tested in detail, whether other library preparation methods would yield better results, perhaps in combination with the collapsing approach.

Common transcripts and differential gene expression analysis

Even for the well-annotated Nipponbare transcriptome, around 17% of the transcripts that we found did not have a functional description and are therefore considered to be novel isoforms. Similarly, for the remaining cultivars, between 19% and 28% of the transcripts could not be assigned with a functional description. This is supported by the identification of a large fraction of potential novel isoform models by the `gffcompare` tool compared with the Nipponbare reference transcriptome. However, `gffcompare` also reports isoforms as "novel" models which share at least one splice junction with the reference transcript and differ in the remaining splice junctions for multiple-exon transcripts. This criterion can be weak for example where exon-exon boundaries are shifted due to sequencing errors [56].

Since all ten cultivars that we analyzed belong to the same species, they should have a large fraction of common transcripts that may be identified by a `blast` search. We therefore used

the transcriptome of one cultivar from each subspecies to generate a database for `blast` searches of the other nine transcriptomes. With this approach, we were able to identify common, cultivar- and subspecies-specific transcripts within our datasets. It must be stressed, however, that the lack of a transcript in the transcriptome of a particular cultivar may have two reasons. It could indeed be absent from the transcriptome and genome of this cultivar, or it could be missing from the transcriptome of this cultivar relative to one of the databases because of differences in sampling, such as different tissues or growth conditions.

Our analysis indicated, as expected, that the largest fraction of the transcripts identified in N22 (47.6%), IR64 (44.8%) and Nipponbare (55.2%) were common to all transcriptomes. Using the *aus* cultivar N22 as the database yielded 652 N22-specific and an additional 184 *aus* specific transcripts, resulting in a total of 836 transcripts (4.4% of the total N22 transcripts) that were only found in the *aus* cultivars. Interestingly, we also identified 160 transcripts in IR64 and 166 in Nipponbare that were not present in either of the *aus* transcriptomes, while neither the IR64 nor the Nipponbare transcriptomes contained any transcripts that were specific for the respective subspecies. The Nipponbare transcriptome only contained a very small fraction (2.1%) of cultivar-specific transcripts. This was very different in the IR64 transcriptome with over 2,426 unique transcripts, comprising 11.7% of the transcriptome. We attribute this high fraction of cultivar specific IR64 transcripts to the fact that only in this case roots were included in the analysis and submergence and salt stress were applied. In all other cultivars, only above-ground tissues were sampled, and treatments involved exclusively high night temperatures, heat, and drought stress.

Aus cultivars are known to be more stress tolerant than *indica* or *japonica* cultivars and contain genes, such as the phosphate starvation tolerance gene *OsPSTOL1* [18], the submergence tolerance gene *OsSUB1A* [19], and the deepwater escape genes *OsSNORKEL1/2* [20] that are absent in the Nipponbare reference genome. To test whether our transcriptome sequencing approach might aid in the identification of such *aus* specific stress-related genes, we performed a differential gene expression analysis by Illumina-based RNA-seq. The samples from developing seeds were obtained from N22 plants grown under control and combined drought and heat stress in the field [35]. More than 50 significantly differentially expressed genes were identified as unique to the *aus* subspecies transcriptomes. Over 45% of the gene products were annotated as homologous to an *Arabidopsis thaliana* gene, such as the gene B12989 annotated as encoding a RALF precursor polypeptide, which may regulate plant stress responses, growth, and development in Arabidopsis and tobacco [57].

We characterized one of the significantly induced genes in more detail. The gene B12288 is annotated as *RAB21*. This gene has homologs in different *Oryza sativa* subspecies and in various wild species of *Oryza*. It belongs to the dehydrin family of LEA proteins and high levels of expression of *RAB21* have been found in mature seeds, as well as in vegetative tissues under salt and drought stress, and after treatment of rice seedlings with the plant stress hormone abscisic acid [58]. The drought and heat induced *RAB21* gene we identified in N22 was more closely related to *RAB21* isoforms from wild rice species than to the homolog from Nipponbare. The sequence differences are not large but may nevertheless be functionally significant. It has been shown with *in-vitro* assays that some dehydrins are able to protect enzymes from inactivation under heat stress [59, 60], indicating a possible function of RAB21 under combined drought and heat stress conditions that led to transcriptional upregulation. It is still unclear which structural characteristics determine the ability of a dehydrin to act as an enzyme stabilizer under heat stress and therefore, the functional significance of the sequence differences between RAB21 from Nipponbare and N22 cannot be evaluated. However, it has recently been shown that changes in only four amino acids in the LEA protein COR15A from Arabidopsis significantly increased the stabilizing effect of this protein for membranes during freezing [61]. It is therefore conceivable that the minor differences in amino acid sequence between the RAB21 proteins from different subspecies and wild rice species may have significant functional effects. Obviously, further experimental work will be necessary to test this hypothesis.

Potential implications

The central question of our study was whether targeted partial transcriptomes obtained by PacBio IsoSeq may be useful for the down-stream RNA-seq analysis in rice cultivars from subspecies such as *aus*, which are not well represented by the Nipponbare reference genome sequence. Our analysis has shown that for all 10 cultivars that were investigated, cultivar-specific transcripts could be identified. In addition, a number of *aus* subspecies (i.e. N22+Dular) specific transcripts were identified. These results strongly suggest that this approach will be useful for future analysis of RNA-seq datasets. The transcriptomes that we have reconstructed here will be directly available for the research community. In addition, the general approach should also be useful for many other plant species for which no high-quality genome assemblies are available, as it represents a much cheaper and computationally less challenging alternative when the aim is the targeted analysis of RNA-seq data. In principle, the approach should also be applicable to species outside of the plant kingdom.

The example of the *RAB21* gene shows that the identification of novel homologs of genes that have already been annotated in the reference genome and that are most likely introgressions from wild relatives, can yield interesting information. In the case of the RAB21 proteins from Nipponbare, N22 and several wild relatives of *Oryza sativa*, our analysis suggests obvious mutational studies that could be performed to understand the potential functional significance of the relatively minor amino acid sequence differences among these proteins. Other interesting candidates have been identified that could be functionally characterized using available functional genomics tools to improve the environmental stress tolerance of rice in an effort to generate climate change resilient cultivars through targeted molecular breeding.

Methods

Plant material

Different tissues of ten *Oryza sativa* ssp. *japonica, indica,* and *aus* cultivars were used for RNA isolation. Plants were grown under combined drought and heat stress in the field at IRRI (Dular, N22, Anjali) [35], under heat and combined drought and heat stress under controlled climate conditions at IRRI (N22, Moroberekan) [34], under shoot submergence root salinity, and combined shoot submergence and root salinity in net-houses at IRRI (IR64) [37 and unpublished observations], under high night temperature stress under controlled climate conditions at MPI Potsdam (IR62266-42-6-2, IR64, IR72, CT9993-5-10-1M, M202, Moroberekan, Nipponbare) [33] and under high night temperature stress in the field at IRRI (IR62266-42-6-2, IR64, IR72, CT9993-5-10-1M, M202, Moroberekan) [36]. Samples were obtained from plants grown under both stress and control conditions (see Additional file 1 for a complete list of all samples). An overview of cultivars, tissues and growth environments is given in Table 1. The selection of cultivars was based on their different sensitivity to high night temperature [33], heat, drought, or combined heat and drought stress [34, 35].

RNA extraction and sequencing

Total RNA was isolated from homogenized frozen material from all samples listed in Additional file 1 using Trizol-based methods [62, 63]. RNA was quantified spectrophotometrically (NanoDrop Technologies, Wilmington, DE, USA) and genomic DNA contamination was removed by DNase treatment (Rapid Out DNA Removal Kit, Thermo Scientific). Absence of genomic DNA was verified by qRT-PCR using a primer pair amplifying an intron sequence [64]. Final RNA quality and integrity were assayed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). For each cultivar, RNA isolated from all organs and treatments was pooled to generate one sample per cultivar. PacBio library

preparation and sequencing were performed at the Max-Planck Genome Center Cologne, Germany. cDNA was synthesized and amplified according to the Pacific Biosciences´s protocol using the SMARTer PCR cDNA Synthesis kit (Clontech) and amplification by the KAPA HIFI PCR Kit (Kapa Biosystems). The cDNAs were not size-selected and PacBio libraries were prepared with the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences) and sequenced on the PacBio Sequel I with Sequel DNA polymerase and binding kit and sequencing chemistry version 2.1 for 600 min. Each library was sequenced on two or three SMRT cells to achieve sufficient coverage.

For RNA-seq analysis, RNA was isolated from developing seeds of the *aus* cultivar N22. Plants were grown in the field in 2013 under either well-watered control conditions or under combined drought and heat stress [35] and RNA was extracted using Ribospin Seed/Fruit and Riboclear *plus!* (GeneAll Biotechnology, Songpa-gu, Republic of Korea) following the manufacturer´s instructions. Three biological replicates were generated for each condition (control/stress). Quantification of RNA and quality controls were performed as described above. Library preparation and sequencing were performed at the Max-Planck Genome Centre Cologne. Libraries were prepared with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs) and sequenced using Illumina HiSeq 3000 technology generating approximately 30 million 150 base pair-long single-end reads per sample.

De novo transcriptome reconstruction

To generate full-length isoforms, the software `IsoSeq3` v3.0 (PacBio) included in `smrtlink` v5.1 was used to perform the following four steps: `consensus` (ccs 3.0.0), `lima` (lima 1.0.0), `cluster` (sierra 0.7.1) and `polish` (tango 0.7.1). Raw data processing for each library was performed on combined data from two or three SMRT cells (using the `smrtlink` command `create`) with default parameters:

```
ccs $in.subreads.bam $out.bam --noPolish --minPasses=1
lima $in.xml primer.fasta $out.demux.ccs.bam --isoseq --no-pbi --dump-clips
isoseq3 cluster $in.demux.ccs.bam $out.unpolished.bam
isoseq3 polish $in.unpolished.bam $out.polished.bam
```

As final output high-quality (HQ) and low-quality (LQ) isoforms were obtained. Only HQ isoforms were used for subsequent analysis. To identify contaminations, HQ isoforms of all cultivars were aligned against the NCBI nucleotide database (downloaded: 24.07.2018) with `blastn` v2.3.0 [40] ($E \leq 1e^{-10}$). Isoforms without a hit were aligned against the NCBI protein database (downloaded: 24.07.2018) using `blastx` v2.3.0 [40] ($E \leq 1e^{-10}$). All isoforms without

a significant hit for the family *Viridiplantae* (green plants) were defined as contaminations and removed.

Genome references

For insertion and deletion (InDel) determination, collapsing and mapping, three *O. sativa* genome references from the subspecies *aus* (N22) [6], *indica* (Shuhui498 (R498 genome)) [13] and *japonica* (Nipponbare, IRGSPv1.0.44) [10] were used.

InDel analysis

HQ isoforms of each cultivar were mapped against the subspecies-specific reference genomes using `minimap2`, v2.17-r941 [65] with the parameters `--ax splice`, `--uf --C5` and `--secondary=no`. Insertions and deletions were determined by extracting the cigar string from the alignment files in bam format [66].

Collapsing redundant isoforms

For the removal of redundant PacBio isoforms, three tools were tested, namely Transcriptome Annotation by Modular Algorithms (`TAMA`) [51], `cDNA cupcake` [67] and COding GENome reconstruction Tool (`cogent` v3.9) [68] followed by the `cDNA cupcake` collapse pipeline. For further descriptions, we will refer to the latter only as `cogent`. `TAMA` and `cDNA cupcake` use a reference genome to collapse PacBio isoforms, while cogent employs a reference-free approach, where it reconstructs gene loci based on PacBio isoforms creating its own "coding genome". Afterwards, `cDNA cupcake` is employed to collapse the isoforms based on the created reference. For `TAMA`, the following parameters were used: `-x no_cap,-e longest_ends,-a 100,-z 100,-m 30` and `-d merge_dup`. `cDNA cupcake` and `cogent` were run with default parameters following the descriptions on the corresponding websites [69, 70]. For both reference-based approaches, the respective reference genome of the appropriate subspecies was used and HQ isoforms were mapped with `minimap2` v2.17-r941 [65]. For all downstream analysis, collapsed transcript models obtained by `TAMA` were used. While `cogent` and `cDNA cupcake` provide the PacBio transcripts after collapsing, `TAMA` generates a bed file with the coordinates of the collapsed transcripts and sequences were extracted from the corresponding genome sequence of each subspecies for the ten cultivars using `bedtools` v2.27.0 [71] `getfasta`. Additionally, remaining unmapped transcripts were collapsed with `cogent` and added to the final datasets (Additional folder 4).

BUSCO analysis

A set of 430 *Viridiplantae* conserved ortholog proteins was used in `BUSCO` v3.0.2 (Benchmarking Universal Single-Copy Orthologs) [72] to assess the completeness of the

conserved content of the *de novo* reconstructed transcriptomes using the BUSCO transcriptome mode.

Phylogenetic analysis

For phylogenetic analysis, SNPs of the collapsed transcripts from TAMA, cDNA cupcake and cogent were used for analyses with SNPhylo [73]. Collapsed transcripts of all cultivars obtained by cogent and cDNA cupcake were mapped against the IRGSP Nipponbare reference genome and SNPs were called utilizing the bcftools v1.9 pipeline [74]. For TAMA, HQ transcripts of all cultivars were collapsed based on the Nipponbare reference genome and the generated variant file was used to determine SNPs. Entries were filtered for the "M" type and defined as alternative alleles. The respective reference alleles were extracted with bedtools v2.27.0 from the reference genome. A simple SNP file was generated and used as input for SNPhylo. Phylogenetic trees were visualized with Figtree [75].

Comparison of reconstructed transcriptomes

HQ collapsed sequences were classified and compared with the existing IRGSP Nipponbare annotation using gffcompare v0.11.2 [76]. The classifications defined by gffcompare were generalized into annotated (classes "=" and "c"), novel isoform (classes "j" and "k"), retrained intron (classes "m" + "n"), novel antisense (class "x"), novel intronic/intergenic (classes "i" and "u") and novel others (classes "o", "y", "e", "s" and "p").

Functional annotation

ORFs were predicted with TransDecoder v5.5.0 [77]. The candidate protein coding regions were extracted by transDecoder.LongOrfs with a minimum length of 100 amino acids. Resulting ORFs were characterized according to similarities to known proteins by a blastp v2.3.0 search [40] ($E \leq 1e^{-5}$) of the comprehensive Swiss-Prot protein database [78] (downloaded 09 Sep 2019) and for conserved protein domains using Hmmer v3.2.1 [79] based on the Pfam database [80] (downloaded 18 Sep 2019). Finally, likely coding regions were reported by the transDecoder.Predict module including all peptides with blast or domain hits. Additionally, HQ collapsed transcripts of all ten cultivars were searched against the Swiss-Prot database using blastx v2.3.0 ($E \leq 1e^{-10}$). All results (blastp, blastx and Pfam) were parsed by Trinotate v3.2.0 [81], stored in a SQLite relational database and then reported as a tab-delimited transcript annotation summary file. Additional Gene Ontology (GO) information was extracted by Trinotate based on the Swiss-Prot database entries. Mercator v4.2 [44] was used as an additional functional annotation pipeline. HQ collapsed

nucleotide sequences were submitted online [82] and resulting tables were downloaded. `Trinotate` and `Mercator` tables were merged to one table per cultivar (Additional file 8). For a detailed comparison with existing *Oryza sativa* bins, results were also compared to the rice MSU7 annotation on the `Mercator` website and saved. All transcripts without any annotation for `Mercator` or the `TransDecoder-Trinotate` pipeline were extracted and a `blastn` search (min. identity 85%, $E \leq 1e^{-10}$) performed against all available cDNA files of *Oryza* wild species obtained from EnsemblPlants [83].

Determination of common overlap

Common overlap of transcripts among the cultivars was determined using `blastn` v2.3.0 [40] with stricter thresholds than before ($E \leq 1e^{-10}$; min. identity 95%). The transcriptome data of the cultivars N22, IR64, and Nipponbare were transformed into blast databases and the transcripts of the remaining nine cultivars were searched against these databases. Results were filtered for the best hit for each database entry, the common overlap determined and visualized using the R package `UpSetR` [84].

Differential gene expression analysis

RNA-seq data for the *aus* cultivar N22 were mapped against the reconstructed PacBio N22 transcriptome using `kallisto` v0.45 [85]. Based on the mappings, a differential gene expression analysis was performed using the R-package `DESeq2` v1.26.0 [86]. *Aus*-Specific differentially expressed genes were extracted, and transcript annotations merged on gene level. A selected candidate gene (B12288) investigated in more detail. Based on the annotation, the product of B12288 is a dehydrin and hence, a multiple sequence alignment was performed with rice specific dehydrin sequences [45] using `Clustal Omega` [87]. The resulting phylogenetic tree was visualized using `Figtree` [75]. Protein sequences were downloaded from www.uniprot.org. The multiple sequence alignment of four closely related protein sequences to the candidate protein B12288 was visualized with `MView` [87].

Graphical visualization

If not mentioned otherwise, the R packages `ggplot2` [88], `ggpubr` [89], `gridExtra` [90] and `reshape2` [91] were used for graphical visualization of the results.

Availability of supporting data and materials

PacBio raw data are available in the NCBI's SRA database under the accession number PRJNA640670. Collapsed and filtered HQ sequences and functional annotation of all ten cultivars will be publicly available soon. RNA-seq data are available at GEO [39] under the accession number GSE153030.

Availability of source code and requirements

Project name: PacBio-IsoSeq-Workflow-for-Rice

Project home page: GitHub (https://github.com/steffi778/PacBio-IsoSeq-Workflow-for-Rice)

Operating system: Ubuntu 18.04

Programming language: R, bash

License: GNU General Public License

Abbreviations

bp – Basepairs; BUSCO – Benchmarking Universal Single-Copy Orthologs; FLNC – Full-Length Non-Chimeric; GB – Gigabases; HNT - High Night Temperature; HQ – High Quality; InDel – Insertion/deletion; IRGSP – International Rice Genome Sequencing Project; IsoSeq – Isoform sequencing; LQ – Low Quality; ORF – Open Reading Frame; RNA-seq – RNA sequencing; SMRT – Single-Molecule, Real-Time; SNP – Single Nucleotide Polymorphism

Acknowledgments

We would like to acknowledge the Max-Planck Genome Center Cologne for sequencing and Rod Wing (University of Arizona, Tucson) for providing the updated N22 genome assembly. We thank Ulrike Glaubitz and Xia Li for the contribution of rice samples, and Jessica Alpers for excellent technical support.

Funding

DKH gratefully acknowledges an exceptional grant from the Max-Planck Society to cover the cost of PacBio and Illumina sequencing. This research was in part funded by the German Federal Ministry for Economic Cooperation and Development, through Contracts No. 81141844 and 81206686 (to SVKJ and DKH) and the U.S. National Institute of Food and Agriculture–Agriculture and Food Research Initiative (2011−04015 J.B.-S. and 2017-67013-26194 to J.B.-S. and E.S.). LMFL gratefully acknowledges a PhD fellowship from the University of Potsdam and R.A acknowledges a Monsanto Beachell-Borlaug International PhD Scholarship at the University of California Riverside. The funding bodies had no role in study design, data collection, analysis or interpretation, or in writing the manuscript.

Authors' contributions

DKH conceived the project. SVKJ and LMFL organized the field experiments. RA, EMS and JB-S planned and performed the net-house experiments and provided RNA from the samples. EZ organized the high night temperature experiments in climate chambers. LMFL provided the RNA-seq data. BH supervised all sequencing and provided technical input for the experimental

planning. SS performed the data analysis under supervision of AF, with contributions from EZ and DKH. SS, AF and DKH wrote the manuscript with edits from all co-authors.

Competing interests

The authors declare that they have no financial or non-financial competing interests.

Additional files

Additional file 1 (XLS). List of samples used for RNA isolation and PacBio sequencing

HNT - high night temperature.

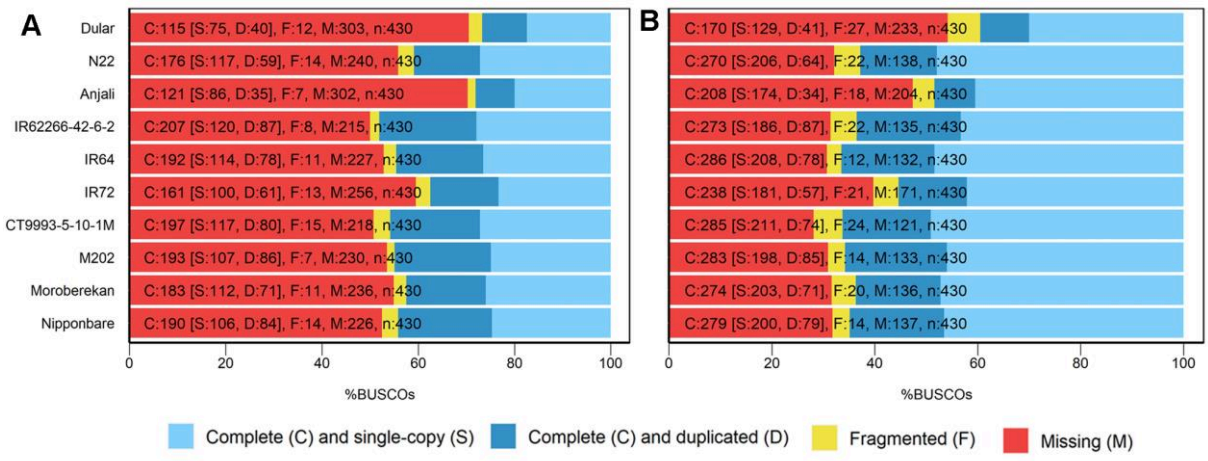
Additional file 2 (XLS). Results from InDel analysis

The total number of InDels (Insertions/deletions) per transcriptome is shown, along with the fraction of InDels in each transcriptome as per-nucleotide error rate in %.

Additional file 3 (XLS). Results of InterPro analysis of proteins missing from our transcriptomes, but present in the BUSCO data base

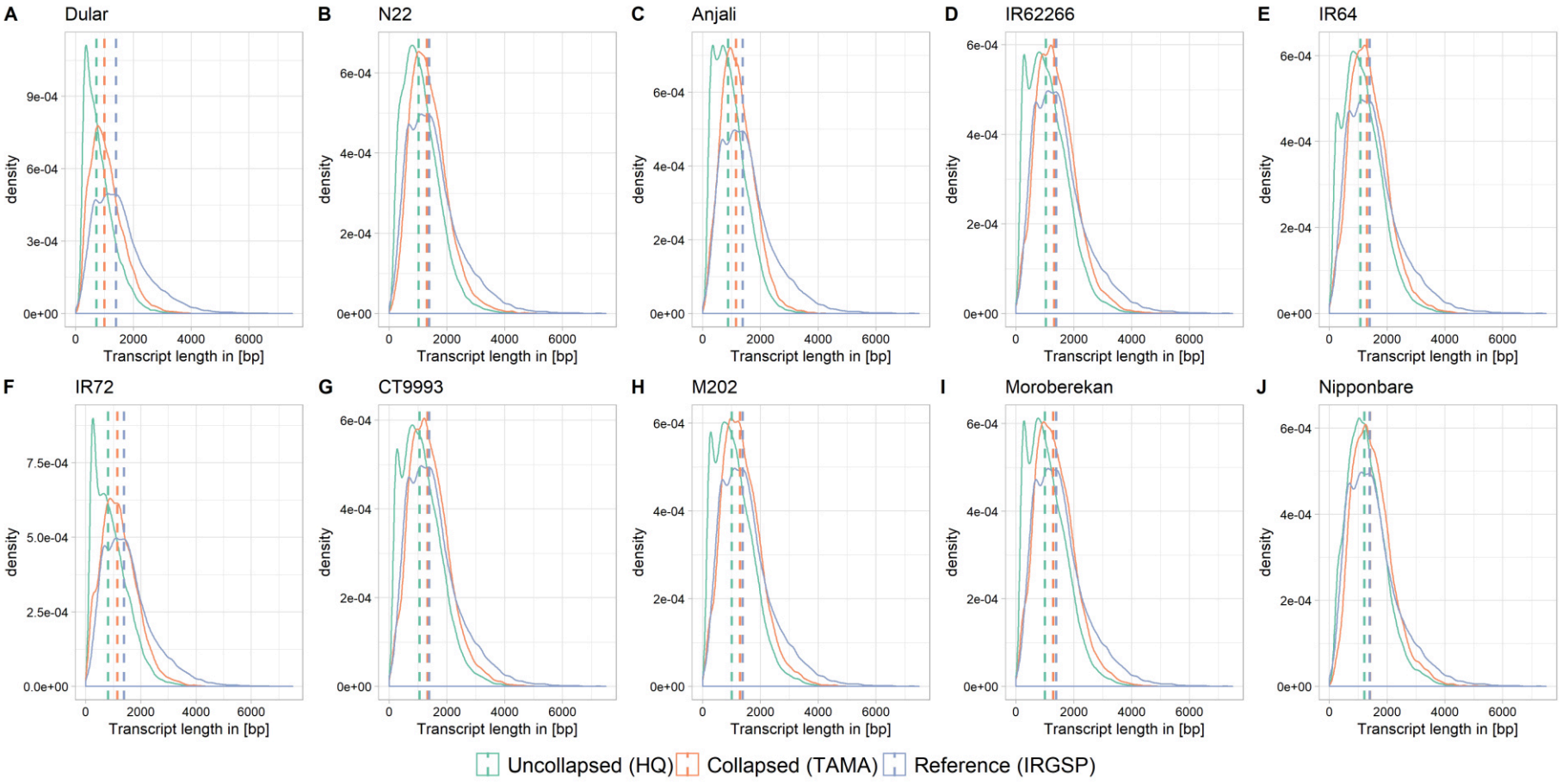
The table provides the BUSCO ID, the InterPro description of the proteins, the InterPro ID and where available information on organ localization and the developmental stage where the protein has been detected.

Additional file 4 (PNG). `BUSCO` assessment results for the collapsing tools `cogent` (A) and `cDNA cupcake` (B)



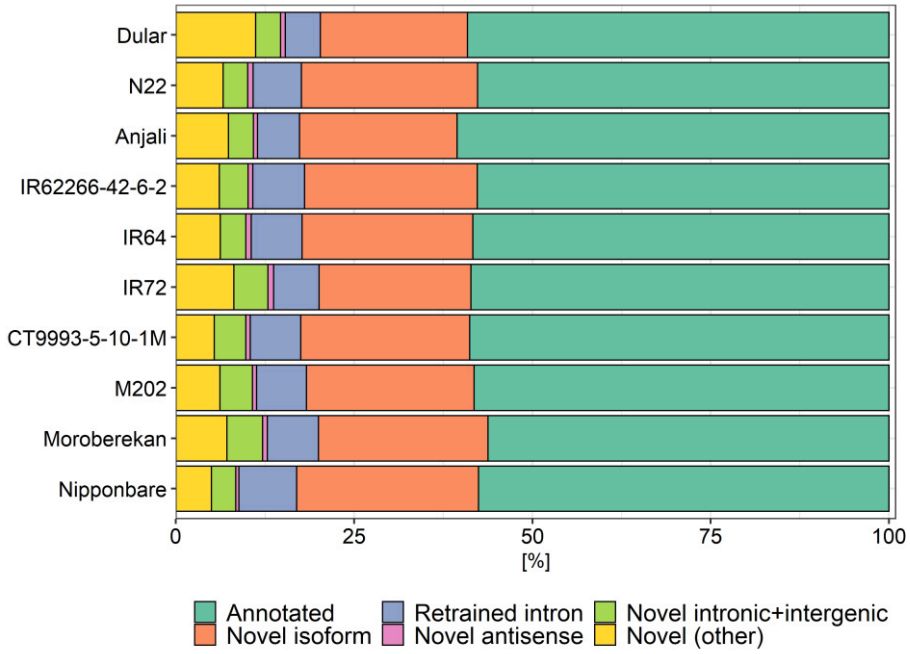
Additional file 5 (PNG). Transcript length distribution for the 10 *Oryza sativa* cultivars

Length distribution of uncollapsed transcripts is indicated in green, length distribution of transcripts after collapsing by `TAMA` is indicated in orange and length distribution of the Nipponbare IRGSP reference transcriptome is indicated in purple. Dashed lines show the median length of transcripts for the respective datasets.


A Dular
B N22
C Anjali
D IR62266
E IR64
F IR72
G CT9993
H M202
I Moroberekan
J Nipponbare
density
Transcript length in [bp]
Uncollapsed (HQ)
Collapsed (TAMA)
Reference (IRGSP)

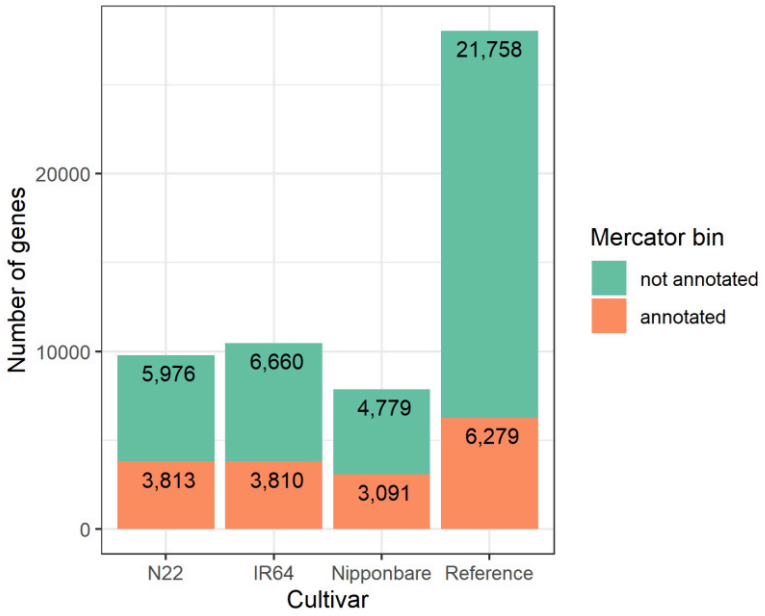
Additional file 6 (PNG). Annotated and novel transcripts as identified by `gffcompare` using the Nipponbare reference annotation

Cultivars were sorted alphabetically within the subspecies *aus*, *indica*, and *japonica*.



Additional file 7 (PNG). Number of genes classified as "annotated" or "not annotated" among the genes not assigned to a functional bin in the `Mercator` ontology for rice

Data are shown for the transcriptomes of N22, IR64, and Nipponbare and the Nipponbare reference transcriptome.



Additional file 8 (XLS). Merged functional annotations using the TransDecoder-Trinotate and Mercator pipeline for all ten cultivars

Additional file 9 (XLS). Identity and annotation of transcripts specific to representative cultivars of the three subspecies

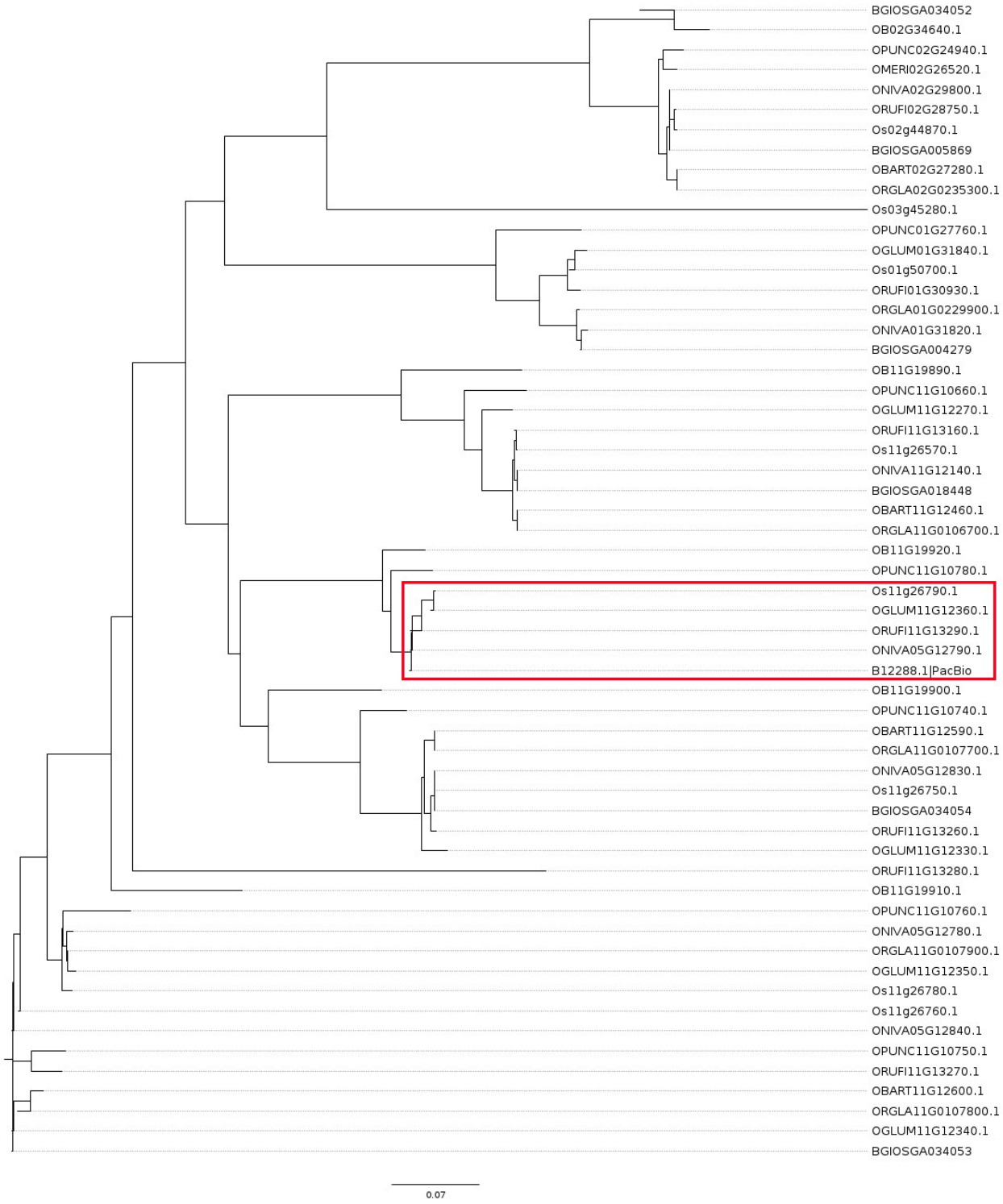
Data are shown for the *aus* subspecies cultivar N22, the *indica* cultivar IR64, and the *japonica* cultivar Nipponbare. In addition to the transcript annotation from `Mercator`, the table also shows annotations retrieved by `BlastX` and `BlastP` searches, Pfam and GO annotations. Further, we provide information about the best-hit `BlastN` analysis gene ID from the Nipponbare IRGSP genome.

Additional file 10 (XLS). Drought and heat regulated genes in N22

Heat and drought regulated genes were determined by Illumina-based RNA-seq in developing seeds of plants of the cultivar N22 grown in the field. Therefore, RNA-seq reads were mapped against the *de novo* reconstructed N22 transcriptome using kallisto. Expression of the transcripts was summarized on gene level and differential gene expression analyzed with DESeq2. Genes identified as *aus*-specific among significantly induced genes were extracted and are listed, including the $\log_2$ fold-change values, FDR p-values, and annotation. Distribution of these *aus*-specific genes among different annotation classes is also presented.

Additional file 11 (PNG). Phylogenetic tree of *Oryza* dehydrin proteins

Dehydrin selection was based on Verma et al. (2017) and includes the following *Oryza* species: BGI – *Oryza sativa* ssp. *indica*, Os – *Oryza sativa* ssp. *japonica*, OB – *Oryza brachyantha*, OPUNC – *Oryza punctata*, OGLUM – *Oryza glumaepatula*, ORUF – *Oryza rufipogon*, ORGLA – *Oryza glaberrima*, ONIVA – *Oryza nivara*, OBART – *Oryza barthii*, OMERI - *Oryza meridionalis*. The red box indicates the proteins used for the sequence alignment shown in Figure 7.


BGIOSGA034052
OB02G34640.1
OPUNC02G24940.1
OMERI02G26520.1
ONIVA02G29800.1
ORUFI02G28750.1
Os02g44870.1
BGIOSGA005869
OBART02G27280.1
ORGLA02G0235300.1
Os03g45280.1
OPUNC01G27760.1
OGLUM01G31840.1
Os01g50700.1
ORUFI01G30930.1
ORGLA01G0229900.1
ONIVA01G31820.1
BGIOSGA004279
OB11G19890.1
OPUNC11G10660.1
OGLUM11G12270.1
ORUFI11G13160.1
Os11g26570.1
ONIVA11G12140.1
BGIOSGA018448
OBART11G12460.1
ORGLA11G0106700.1
OB11G19920.1
OPUNC11G10780.1
Os11g26790.1
OGLUM11G12360.1
ORUFI11G13290.1
ONIVA05G12790.1
B12288.1|PacBio
OB11G19900.1
OPUNC11G10740.1
OBART11G12590.1
ORGLA11G0107700.1
ONIVA05G12830.1
Os11g26750.1
BGIOSGA034054
ORUFI11G13260.1
OGLUM11G12330.1
ORUFI11G13280.1
OB11G19910.1
OPUNC11G10760.1
ONIVA05G12780.1
ORGLA11G0107900.1
OGLUM11G12350.1
Os11g26780.1
Os11g26760.1
ONIVA05G12840.1
OPUNC11G10750.1
ORUFI11G13270.1
OBART11G12600.1
ORGLA11G0107800.1
OGLUM11G12340.1
BGIOSGA034053
0.07

REFERENCES

1. Lamaoui M, Jemo M, Datla R and Bekkaoui F. Heat and Drought Stresses in Crops and Approaches for Their Mitigation. Front. Chem. 2018;6:26. doi:10.3389/fchem.2018.00026.
2. Dawson TP, Perryman AH and Osborne TM. Modelling impacts of climate change on global food security. Clim. Change. 2016;134 3:429-40. doi:10.1007/s10584-014-1277-y.
3. Zhao C, Liu B, Piao S, Wang X, Lobell DB, Huang Y, et al. Temperature increase reduces global yields of major crops in four independent estimates. PNAS. 2017;114 35:9326. doi:10.1073/pnas.1701762114.
4. Iizumi T and Ramankutty N. Changes in yield variability of major crops for 1981–2010 explained by climate change. Environ. Res. Lett. 2016;11 3:034003. doi:10.1088/1748-9326/11/3/034003.
5. Peng S, Huang J, Sheehy JE, Laza RC, Visperas RM, Zhong X, et al. Rice yields decline with higher night temperature from global warming. PNAS. 2004;101 27:9971-5. doi:10.1073/pnas.0403720101.
6. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat. Genet. 2018;50 2:285-96. doi:10.1038/s41588-018-0040-0.
7. FAO: Food systems for better nutrition. http://www.fao.org/publications/sofa/2013/en/ (2013). Rome. Accessed: 28 July 2020.
8. Li J-Y, Wang J and Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience. 2014;3 1 doi:10.1186/2047-217x-3-7.
9. Mahesh HB, Shirke MD, Singh S, Rajamani A, Hittalmani S, Wang GL, et al. Indica rice genome assembly, annotation and mining of blast disease resistance genes. BMC Genomics. 2016;17:242. doi:10.1186/s12864-016-2523-7.
10. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, et al. A draft sequence of the rice genome (*Oryza sativa L.* ssp. *japonica*). Science. 2002;296 5565:92-100.
11. RGP. The 3,000 rice genomes project. GigaScience. 2014;3 1:7. doi:10.1186/2047-217X-3-7.
12. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557 7703:43-9.
13. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, et al. Sequencing and de novo assembly of a near complete *indica* rice genome. Nat. Commun. 2017;8:15324.
14. Zhang J, Chen LL, Xing F, Kudrna DA, Yao W, Copetti D, et al. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. PNAS. 2016;113 35:E5163-E71.
15. Sakai H, Kanamori H, Arai-Kichise Y, Shibata-Hatta M, Ebana K, Oono Y, et al. Construction of pseudomolecule sequences of the aus rice cultivar Kasalath for comparative genomics of Asian cultivated rice. DNA Res. 2014;21 4:397-405. doi: 10.1093/dnares/dsu006.
16. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, et al. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. PNAS. 2009;106 30:12273-8. doi:10.1073/pnas.0900992106.
17. Garris AJ, McCouch SR and Kresovich S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa L.*). Genetics. 2003;165 2:759-69.
18. Gamuyao R, Chin JH, Pariasca-Tanaka J, Pesaresi P, Catausan S, Dalid C, et al. The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. Nature. 2012;488:535. doi:10.1038/nature11346
19. Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R and Heuer S. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature. 2006;442 doi:10.1038/nature04920.
20. Hattori Y, Nagai K, Furukawa S, Song XJ, Kawano R, Sakakibara H, et al. The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. Nature. 2009;460 7258:1026-30. doi:10.1038/nature08258.

21. Baltazar MD, Ignacio JCI, Thomson MJ, Ismail AM, Mendioro MS and Septiningsih EM. QTL mapping for tolerance of anaerobic germination from IR64 and the aus landrace Nanhi using SNP genotyping. Euphytica. 2014;197 2:251-60. doi:10.1007/s10681-014-1064-x.
22. Baltazar MD, Ignacio JCI, Thomson MJ, Ismail AM, Mendioro MS and Septiningsih EM. QTL mapping for tolerance to anaerobic germination in rice from IR64 and the aus landrace Kharsu 80A. Breed. Sci. 2019;69 2:227-33. doi:10.1270/jsbbs.18159.
23. Bernier J, Kumar A, Venuprasad R, Spaner D, Verulkar S, Mandal NP, et al. Characterization of the effect of a QTL for drought resistance in rice, qtl12.1, over a range of environments in the Philippines and eastern India. Euphytica. 2009;166 2:207-17. doi:10.1007/s10681-008-9826-y.
24. Slabaugh E, Desai JS, Sartor RC, Lawas LMF, Jagadish SVK and Doherty CJ. Analysis of differential gene expression and alternative splicing is significantly influenced by choice of reference genome. RNA. 2019;25 6:669-84.
25. Rhoads A and Au KF. PacBio Sequencing and Its Applications. Genom Proteom Bioinf. 2015;13 5:278-89. doi:https://doi.org/10.1016/j.gpb.2015.08.002.
26. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of the sorghum transcriptome using single-molecule long reads. Nat. Commun. 2016;7:11706. doi:10.1038/ncomms11706.
27. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res. 2017;27 5:885-96. doi:10.1101/gr.217117.116.
28. Dong L, Liu H, Zhang J, Yang S, Kong G, Chu JS, et al. Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC Genomics. 2015;16:1039. doi:10.1186/s12864-015-2257-y.
29. Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L, Thirugnanasambandam PP, et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics. 2017;18 1:395. doi:10.1186/s12864-017-3757-8.
30. Feng S, Xu M, Liu F, Cui C and Zhou B. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in *Gossypium australe*. BMC Plant Biol. 2019;19 1:365. doi:10.1186/s12870-019-1968-7.
31. Carvalho DS, Nishimwe AV and Schnable JC. IsoSeq transcriptome assembly of C3 panicoid grasses provides tools to study evolutionary change in the Panicoideae. Plant Direct. 2020;4 2:e00203. doi:10.1002/pld3.203.
32. Chao Y, Yuan J, Li S, Jia S, Han L and Xu L. Analysis of transcripts and splice isoforms in red clover (*Trifolium pratense L.*) by single-molecule long-read sequencing. BMC Plant Biol. 2018;18 1:300. doi:10.1186/s12870-018-1534-8.
33. Glaubitz U, Li X, Köhl KI, van Dongen JT, Hincha DK and Zuther E. Differential physiological responses of different rice (*Oryza sativa*) cultivars to elevated night temperature during vegetative growth. Funct. Plant Biol. 2014;41 4:437.
34. Li X, Lawas LM, Malo R, Glaubitz U, Erban A, Mauleon R, et al. Metabolic and transcriptomic signatures of rice floral organs reveal sugar starvation as a factor in reproductive failure under heat and drought stress. Plant Cell Environ. 2015;38 10:2171-92. doi:10.1111/pce.12545.
35. Lawas LMF, Shi W, Yoshimoto M, Hasegawa T, Hincha DK, Zuther E, et al. Combined drought and heat stress impact during flowering and grain filling in contrasting rice cultivars grown under field conditions. Field Crops Res. 2018;229:66-77. doi:10.1016/j.fcr.2018.09.009.
36. Schaarschmidt S, Lawas LMF, Glaubitz U, Li X, Erban A, Kopka J, et al. Season Affects Yield and Metabolic Profiles of Rice (*Oryza sativa*) under High Night Temperature Stress in the Field. Int J Mol Sci. 2020;21 9 doi:10.3390/ijms21093187.

37. Alam R, Hummel M, Yeung E, Locke AM, Ignacio JCI, Baltazar MD, et al. Flood resilience loci SUBMERGENCE 1 and ANAEROBIC GERMINATION 1 interact in seedlings established underwater. Plant Direct. 2020;4 7 doi:10.1002/pld3.240.
38. Leinonen R, Sugawara H, Shumway M and International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011;39 Database issue:D19-D21. doi:10.1093/nar/gkq1019.
39. Edgar R, Domrachev M and AE L. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30 1:2074.
40. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. doi:10.1186/1471-2105-10-421.
41. Ardui S, Ameur A, Vermeesch JR and Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018;46 5:2159-68. doi:10.1093/nar/gky066.
42. Dong X, Gao Y, Chen W, Wang W, Gong L and Liu X. Spatiotemporal distribution of phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice. Mol Plant. 2015;8 doi:10.1016/j.molp.2014.11.003.
43. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017 - beyond protein family and domain annotations. Nucleic Acids Res. 2017;45 D1:D190-D9. doi:10.1093/nar/gkw1107.
44. Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A, et al. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Mol Plant. 2019;12 6:879-92. doi:10.1016/j.molp.2019.01.003.
45. Verma G, Dhar YV, Srivastava D, Kidwai M, Chauhan PS, Bag SK, et al. Genome-wide analysis of rice dehydrin gene family: Its evolutionary conservedness and expression pattern in response to PEG induced dehydration stress. PLoS One. 2017;12 5:e0176399. doi:10.1371/journal.pone.0176399.
46. Hundertmark M and Hincha DK. LEA (late embryogenesis abundant) proteins and their encoding genes in *Arabidopsis thaliana*. BMC Genomics. 2008;9:118. doi:10.1186/1471-2164-9-118.
47. Graether SP and Boddington KF. Disorder and function: a review of the dehydrin protein family. Frontiers in Plant Science. 2014;5:576. doi:10.3389/fpls.2014.00576.
48. Workman RE, Myrka AM, Wong GW, Tseng E, Welch KC, Jr. and Timp W. Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. GigaScience. 2018;7 3:1-12. doi:10.1093/gigascience/giy009.
49. Li J, Harata-Lee Y, Denton MD, Feng Q, Rathjen JR, Qu Z, et al. Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus membranaceus* reveals transcript variants involved in bioactive compound biosynthesis. Cell Dis. 2017;3:17031. doi:10.1038/celldisc.2017.31.
50. Xie L, Teng K, Tan P, Chao Y, Li Y, Guo W, et al. PacBio single-molecule long-read sequencing shed new light on the transcripts and splice isoforms of the perennial ryegrass. Mol Gen Genom. 2020;295 2:475-89. doi:10.1007/s00438-019-01635-y.
51. Kuo RI, Cheng Y, Smith J, Archibald AL and Burt DW. Illuminating the dark side of the human transcriptome with TAMA Iso-Seq analysis. bioRxiv. 2019; doi:10.1101/780015.
52. Zhang G, Sun M, Wang J, Lei M, Li C, Zhao D, et al. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J. 2019;97 2:296-305. doi:10.1111/tpj.14120.
53. Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, et al. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. New Phytol. 2018;217 1:163-78. doi:10.1111/nph.14762.
54. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016;7:11708. doi:10.1038/ncomms11708.

55. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL and Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017;18 1:323. doi:10.1186/s12864-017-3691-9.
56. Tung LH, Shao M and Kingsford C. Quantifying the Benefit Offered by Transcript Assembly on Single-Molecule Long Reads. bioRxiv. 2019:632703. doi:10.1101/632703.
57. Olsen AN, Mundy J and Skriver K. Peptomics, Identification of Novel Cationic Arabidopsis Peptides with Conserved Sequence Motifs. In Silico Biol. 2002;2:441-51.
58. Mundy J and Chua NH. Abscisic acid and water-stress induce the expression of a novel rice gene. EMBO Journal. 1988;7 8:2279-86.
59. Koubaa S, Bremer A, Hincha DK and Brini F. Structural properties and enzyme stabilization function of the intrinsically disordered LEA_4 protein TdLEA3 from wheat. Sci Rep. 2019;9 1:3720. doi:10.1038/s41598-019-39823-w.
60. Kovacs D, Kalmar E, Torok Z and Tompa P. Chaperone Activity of ERD10 and ERD14, Two Disordered Stress-Related Plant Proteins. Plant Physiol. 2008;147 1:381. doi:10.1104/pp.108.118208.
61. Sowemimo OT, Knox-Brown P, Borcherds W, Rindfleisch T, Thalhammer A and Daughdrill GW. Conserved Glycines Control Disorder and Function in the Cold-Regulated Protein, COR15A. Biomolecules. 2019;9 3 doi:10.3390/biom9030084.
62. Li Z and Trick HN. Rapid method for high-quality RNA isolation from seed endosperm containing high levels of starch. Biotechniques. 2005;38 6:872, 4, 6. doi:10.2144/05386bm05.
63. Chomczynski P and Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nat Protoc. 2006;1 2:581-5. doi:10.1038/nprot.2006.83.
64. Do PT, Degenkolbe T, Erban A, Heyer AG, Kopka J, Kohl KI, et al. Dissecting rice polyamine metabolism under controlled long-term drought stress. PLoS One. 2013;8 4:e60325. doi:10.1371/journal.pone.0060325.
65. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34 18:3094-100. doi:10.1093/bioinformatics/bty191.
66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25 16:2078-9. doi:10.1093/bioinformatics/btp352.
67. Tseng E: cDNA cupcake. https://github.com/Magdoll/cDNA_Cupcake (2019). Accessed 29.11 2019.
68. Tseng E: Cogent. https://github.com/Magdoll/Cogent (2019). Accessed 29.11 2019.
69. Tseng E: Cogent Tutorial. https://github.com/Magdoll/Cogent/wiki/Tutorial%3A-Using-Cogent-to-collapse-redundant-transcripts-in-absence-of-genome (2019). Accessed 29.11 2019.
70. Tseng E: cDNA cupcake Wiki. https://github.com/Magdoll/cDNA_Cupcake/wiki (2019). Accessed 29.11 2019.
71. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26 6:841-2. doi:10.1093/bioinformatics/btq033.
72. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017; doi:10.1093/molbev/msx319.
73. Lee T-H, Guo H, Wang X, Kim C and Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 2014;15 162.
74. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27 21:2987-93. doi:10.1093/bioinformatics/btr509.
75. Rambaut A: FigTree v1.4. http://tree.bio.ed.ac.uk/software/figtree/ (2012).
76. Pertea G and Pertea M. GFF Utilities: GffRead and GffCompare [version 1; peer review: 3 approved]. F1000Research. 2020;9 304 doi:10.12688/f1000research.23297.1.

77. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8 8:1494-512. doi:10.1038/nprot.2013.084.
78. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2018;47 D1:D506-D15. doi:10.1093/nar/gky1049.
79. Eddy SR: Hidden Markov Models - hmmer.org. http://hmmer.org/ (2019). Accessed 19 Nov 2019.
80. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47 D1:D427-D32. doi:10.1093/nar/gky995.
81. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 2017;18 3:762-76. doi:10.1016/j.celrep.2016.12.063.
82. Usadel B: Mercator4 Webtool. https://plabipd.de/portal/mercator4 (2020). Accessed 30 Mar 2020.
83. EnsemblePlants: *Oryza* wildspecies https://plants.ensembl.org/index.html (2020). Accessed 02 Apr 2020.
84. Conway JR, Lex A and Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33 18:2938-40. doi:10.1093/bioinformatics/btx364.
85. Bray NL, Pimentel H, Melsted P and Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34 5:525-7.
86. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15 12:550. doi:10.1186/s13059-014-0550-8.
87. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47 W1:W636-W41. doi:10.1093/nar/gkz268.
88. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
89. Kassambara A: ggpubr R Package: ggplot2-Based Publication Ready Plots. https://github.com/kassambara/ggpubr (2020). Accessed 30 Apr 2020.
90. Auguie B: gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. http://CRAN.R-project.org/package=gridExtra (2017). Accessed 30 Apr 2020.
91. Wickham H. Reshaping Data with the reshape Package. J Stat Softw. 2007;21 12:1-20.

3 Discussion

With the development of next-generation sequencing technologies, the knowledge of the genetic base of diverse plant species increased dramatically in the last years (Michael & Jackson 2013). To identify the genetic regulators that respond to various abiotic stressors, it is essential to study the transcriptome. One of the most common approaches for that purpose is RNA-seq. The first step of RNA-seq data processing is the alignment and quantification of the reads against a reference sequence. A variety of studies exist comparing tools performing these tasks using either a reference genome or a *de novo* transcriptome assembly (Zhang et al. 2017a, Conesa et al. 2016, Benjamin et al. 2014, Dillies et al. 2012) but only little evaluation is available on the performance of read mappers for data generated from genotypes within a species showing sequence polymorphisms. To address this issue, I used experimentally generated RNA-seq data from two *Arabidopsis thaliana* accessions that were mapped and evaluated with five alignment-based and two alignment-free tools (Paper 1).

While *Arabidopsis thaliana* is still a simple model organism, sequencing technologies need also to be utilized in more complex plant systems, for example in rice. Since the successful sequencing of the first rice genome of the cultivar Nipponbare from the *Oryza sativa* ssp. *japonica*, most of the available rice studies were based on this genome/transcriptome and neglected the wide natural variation of rice (Li et al. 2014, Schatz et al. 2014). Therefore, I sequenced the transcriptomes of different organs of ten different *Oryza sativa* cultivars using a third-generation sequencing technology (PacBio IsoSeq) and identified previously unknown cultivar-specific stress-responsive genes (Paper 3).

Finally, a selection of the sequenced rice cultivars was analyzed for their metabolic response to high night temperature (HNT) stress under field conditions (Paper 2). Despite the several advantages of stress experiments in controlled environments (Telfer et al. 2018), it has been shown that plants respond differently to stress under controlled and field conditions (Sprenger et al. 2016, Bahuguna et al. 2015). Hence, more accurate quantification and validation of plant stress-responses need to be obtained in the field (Hartman et al. 2014, Limpens et al. 2012). To the best of my knowledge, this is the first study investigating the primary metabolite profile of different rice cultivars under HNT stress during the wet and dry season.

3.1 Bioinformatic approaches to identify molecular regulators of abiotic stress

For plants, abiotic stresses such as drought, heat, or cold have dramatic effects on growth as well as on crop yield, which can be reduced by as much as 50% (Qin et al. 2011). Many abiotic stresses are complex processes controlled by networks of genetic and environmental factors

that limit traditional breeding approaches and crop improvements (Akpınar et al. 2013). Yield stagnation has been reported over the last decades for major cereals such as maize, rice, or wheat and emphasizes the need to adopt new strategies in agriculture that can ensure global food security (Ray et al. 2012). Bioinformatic tools including software, databases, and web resources have brought major changes in analyzing and interpreting the vast amount of data generated by stress experiments and measured on high-throughput platforms (Kumar & Shanker 2018). In this study, transcriptomic and metabolomic approaches were applied and several bioinformatic tools were evaluated to characterize molecular responses under abiotic stress conditions such as cold (Paper 1), HNT (Paper 2), and heat and drought (Paper 3) in rice or *Arabidopsis thaliana*.

Further applications based on these approaches include quantitative trait loci (QTL) mapping, genome-wide association studies (GWAS), or marker-assisted selection (MAS). Molecular breeding can help to identify tolerant cultivars at a much faster and cheaper rate compared to classical breeding (Shah et al. 2018, Saade et al. 2016). For example, a study in potato identified 20 metabolite and transcript markers for drought stress utilizing a Random Forest machine learning approach. The markers were validated in 16 independent agronomic field trials and demonstrated a stable yield prediction under drought that was largely independent of seasonal and regional agronomic conditions (Sprenger et al. 2018).

3.2 The era of high-throughput sequencing

The analysis of high-throughput data is one of the most challenging tasks today. In transcriptomics, the development of RNA-seq introduced new algorithms and tools to address correct mapping against a reference sequence in a reasonable amount of time (Kim et al. 2015a, Dobin et al. 2013, Li & Durbin 2009), to identify changes in gene expression (Teng et al. 2016, Dillies et al. 2012), or to detect alternative splicing isoforms (Shen et al. 2014, Zhou et al. 2012) and SNPs (Zhao et al. 2019).

3.2.1 Comparison of RNA-seq mapping tools using data from *Arabidopsis thaliana*

After RNA extraction and library preparation in the wet lab, RNA-seq libraries are sequenced to obtain reads. Those reads are mapped to an index of DNA or RNA sequences followed by a quantification step that counts the number of mapped reads to an individual transcript or gene. Most of the tools such as STAR, kallisto, or salmon include an in-built quantification algorithm and hence, can be considered as a wrapper program for both steps (Patro et al. 2017, Bray et al. 2016, Dobin et al. 2013).

Here, five alignment-based and two alignment-free RNA-seq mappers were evaluated using experimentally and *in silico* generated data from the two *Arabidopsis thaliana* accessions Columbia-0 (Col-0) and N14 (Paper 1). It should be pointedout that despite the attempt to assess and compare mapping algorithms intensively in several studies, the evaluation is not straightforward for RNA-seq data (Engström et al. 2013, Korf 2013). Different factors such as sequencing errors, repeats, the complexity of the organism, and genetic variants increase the uncertainty in read mapping and challenge algorithms to find the true genomic source for each read, and thus, to define a 'correct mapping' (Hatem et al. 2013). However, these studies can help to determine the top performers and to select the right tool for data analysis, depending on the biological question, input data, and application (Finotello & Di Camillo 2015).

One key aspect of RNA-seq data analysis is to determine how many reads could be mapped against the reference sequence (mappability). For the Arabidopsis accession Col-0, a high mappability of the 150 bp single-end Illumina reads against the Col-0 reference genome or transcriptome was observed, ranging between 95.9% and 99.5% (Paper 1). A slightly smaller fraction was obtained for the other accession N14, ranging between 92.4% to 98.1%. The high fraction of mapped reads for both accessions may be in part due to the high-quality reference sequence, to the comparatively small genome of Arabidopsis with roughly 130 megabases and to the low content of repetitive DNA sequences (Kim et al. 2015a, Mayer et al. 1999).

Another important aspect next to mappability is the 'correct' mapping of the reads against the reference sequence to obtain reliable biological information. Comparing the alignment-based mapper's STAR, HISAT2, and bowtie2/RSEM a high overlap of reads mapping to the same position at the reference sequence was observed for STAR and HISAT2 for both accessions (Paper 1). The differences in read positions between bowtie2/RSEM and HISAT2/STAR originated to a large part from soft-clipping, mostly of the first base of the reads by both aligners. An additional approach using *in silico* reads from the Col-0 reference sequence revealed that about 99% of the reads were mapped to the correct position by the three mappers and hence showed the same performance when synthetic reads without any mismatches between read and reference sequences were used. Finally, HISAT2 and STAR were also tested by mapping the experimentally generated reads against the reference genome without annotation information resulting in more than 90% of the reads mapped to known exons. These results indicate that mappers most likely mapped the reads back to the right position of the reference sequence. It was reported that alignment-based methods have a high sensitivity in mapping, but longer running times and higher memory costs due to the requirement to align each read accurately (Patro et al. 2017, Bray et al. 2016). Much faster are alignment-free

approaches such as kallisto (Bray et al. 2016) or salmon (Patro et al. 2017) which were also used in our study. Salmon for example only uses unique k-mers that are mapped to the transcriptome to identify the transcripts and leads to an increase in speed, but a decrease in sensitivity (Babarinde et al. 2019).

3.2.2 Quantification and biological analysis based on different mapping algorithms

From a biological point of view, the quantification of gene expression is the most important part of an RNA-seq experiment as researchers are mostly interested in the identification of differentially expressed genes, either between conditions or between genotypes. Correct mapping is essential to determine changes in expression, but read quantification is at least equally important (Fonseca et al. 2014). The number of reads or k-mer per feature depends on the actual expression level, library size, transcript length, GC content, and other parameters (Love et al. 2016, Benjamini & Speed 2012). In this context, we have compared the raw count distribution of all seven tools using one sample of each accession by plotting the results against each other (Paper 1). High similarities among the mappers were observed, indicated by correlation coefficients close to one. Similarly, when the raw counts were compared between mappers for all 36 biological samples, R_v values close to one indicated a good correspondence in the expression levels between all seven tools. In general, read quantification is an essential step and can vary greatly with different algorithms (Fonseca et al. 2014). Alignment-free tools are more suited for transcript-level quantification as they exploit unique splicing patterns to identify unique k-mers (Bray et al. 2016), but they perform poorly with lowly expressed transcripts or short RNAs (Wu et al. 2018) and work better in well-annotated genomes with sufficient transcript annotations (Babarinde et al. 2019). However, in well-annotated organisms, gene-level quantification may be all that is required for many biological questions because the properties of genes are relatively well known, it mostly focus on protein-coding genes (Babarinde et al. 2019), and most highly expressed genes have single dominant isoforms (Ezkurdia et al. 2015).

To analyze the effects of the mapping tools on differential gene expression (DGE) analysis (Paper 1), expression levels of control plants grown at ambient temperature and plants grown for three days at 4°C were compared (cold acclimation, see Zuther et al. 2019) using the R-package *DESeq2* (Love et al. 2014). The results showed that the raw counts generated by the different mappers resulted in clear differences in the number of significantly differentially expressed genes, with an overlap between mappers of 98.0% in Col-0, and 92.1% in N14. The small sample size (three samples per condition and accession) may have contributed to the differences in identifying differentially expressed genes (Soneson & Delorenzi 2013). Previous

studies have demonstrated that the analysis pipeline affects the results and that no single method is likely to perform favorably for all datasets (Rapaport et al. 2013, Seyednasrollah et al. 2013, Soneson & Delorenzi 2013). However, higher detection power was reported with increasing sample sizes (Conesa et al. 2016, Rapaport et al. 2013). Comparisons of different DGE tools for human or mouse data revealed significant differences in differentially expressed genes deriving from normalization, statistical model, and sample size (Rapaport et al. 2013, Seyednasrollah et al. 2013). In contrast, a study in yeast has shown more comparative results among different read aligners and DGE analysis tools (Nookaew et al. 2012). Most of the differences in the results are derived from genetic variation and lowly expressed genes. Like Arabidopsis, yeast is a well-characterized organism, and high-quality genome and annotation data is available that may contribute to mapping a high fraction of reads on the reference sequence and hence, to accurately estimate gene expression levels.

3.2.3 *De novo* transcriptome assembly using short- and long-read technologies

The last years short-read sequencing technologies were dominantly used to study the transcriptome, recently newer technologies have been emerging to sequence longer, full-length transcripts (Deamer et al. 2016, Rhoads & Au 2015). The major advantage of LRS is that transcriptomes can be reconstructed without the need for an assembly or a reference genome. Especially in plants with a large natural variation including tolerance to abiotic stresses, this approach can help to identify cultivar-specific stress-responsive genes (Xu et al. 2020a , Teng et al. 2019). Using LRS technology from PacBio, we have explored a targeted approach of sequencing and reconstructing partial transcriptomes of ten rice cultivars from three different subspecies (Paper 3). These transcriptomes can be used as references to map RNA-seq reads from abiotic stress experiments in stress-tolerant genotypes without a reference genome and were performed exemplary in this study with RNA-seq data obtained from combined heat and drought experiments (see details in Lawas et al. 2018). Large parts of the rice transcriptomes were recovered by using only two or three single-molecule real-time (SMRT) sequencing cells resulting in 38,000 and 54,700 high-quality transcripts without the need for a reference genome or classical short-read *de novo* assembly. However, it is challenging to compare the sequencing output directly with other studies as the technology, chemistry, and software are improved constantly resulting in higher throughput and better data quality (Amarasinghe et al. 2020).

While LRS sequences full-length transcripts, short-read *de novo* transcriptome assemblers reconstruct transcriptomes by the identification of a mutual overlap of the short fragments (Babarinde et al. 2019). Different tools and pipelines exist to perform the assembly either with or without a reference genome but a reference-free assembly of short-reads is normally less

accurate (Steijger et al. 2013). The accurate reconstruction of transcript models with short-reads remain a challenging computational problem (Babarinde et al. 2019), for example, a comparative study including 24 protocols and 14 independent algorithms reported a poor assembly of isoform structures including missing exons and incorrect splice junctions using RNA-seq data based on the human genome (Steijger et al. 2013). Nevertheless, with *de novo* assemblies from short-reads, gene expression can be studied from any species and cell type within a species (Babarinde et al. 2019). However, these *de novo* assemblers are restricted, especially by the assembly of lowly expressed genes and genes with complex splicing patterns (Steijger et al. 2013). LRS technologies can overcome these limitations by sequencing full-length transcripts without the need of a *de novo* assembly. This is interesting for plant breeding, especially for crops with more complex genomes such as the hexaploid wheat (Clavijo et al. 2017). Applications varied by using LRS alone or together with short-reads (Miller et al. 2017) and helped to identify novel transcripts and to update existing annotations in human (Wu & Ben-Yehezkel 2019, Au et al. 2013), animals (Chen et al. 2017) and plants (Feng et al. 2019, Minio et al. 2019, Abdel-Ghany et al. 2016).

3.2.4 Data redundancy and tool development for PacBio isoform sequencing

During the library preparation mRNA degradation products can be formed and are subsequently sequenced. These shorter transcripts lack some of the 5' sequences but are identical to the full-length transcripts resulting in a large number of redundant transcripts. This redundancy can influence follow-up analyses such as alternative splicing studies. In this study, we have compared three collapsing tools to merge redundant transcripts and investigated the influence on data quality, number of unique isoform models per gene locus, and phylogenetic resolution (Paper 3). The first tool, called cogent, does not need a reference sequence to collapse redundant isoforms and was successfully applied to transcriptomes from organisms without an available genome reference (Feng et al. 2019, Workman et al. 2018, Li et al. 2017b). The two other tools, cDNA cupcake and TAMA need a reference genome sequence and have been more commonly used (Xie et al. 2020, Kuo et al. 2019, Zhang et al. 2019a, Wang et al. 2018). In our study, the number of transcripts after collapsing decreased by up to 60% indicating a high redundancy in the datasets. While TAMA and cogent resulted in similar numbers of collapsed transcripts the numbers were slightly higher for cDNA cupcake. Cogent had the highest number of unmapped reads compared to the two other tools. This may be due to the generation of transcript orphans, i.e. putative single-isoform transcripts that were not incorporated into the reconstructed transcriptomes. However, to the best of my knowledge, no studies are available at the moment comparing the performance of these tools in more detail.

Not only data redundancy is a challenge for LRS analyses, but also the development of new bioinformatic tools to process and analyze the data is constantly ongoing. For example, mappers that were developed for short-reads have problems with the high raw sequencing error rate and the length of the reads (Križanović et al. 2017). Križanović et al. (2017) showed that common splice-aware RNA-seq mappers such as HISAT2 and Tophat2 were not able to map long-reads to the reference genome at all. STAR was algorithmically extended for long reads and performed well in this study aligning 96.8% of the Illumina reads, but for the LRS data, the mappability ranged only between 0.1% to 62.2%. The tested data sets had different error rates and STAR seemed to be affected by the increased complexity of the reads. However, new mappers were developed for LRS such as minimap2 which can handle both short and long reads (greater than 1kb at an error rate of 15%) including long insertions and deletions (Li 2018).

3.2.5 Combining short- and long-read technologies to identify molecular regulators for organisms without a reference genome

Compared to Illumina-based short-read sequencing technologies, LRS has a lower sequencing throughput per run, and higher raw error rates, but longer read lengths (Križanović et al. 2017). Fortunately, PacBio and Illumina sequencing are highly complementary to each other (Conesa et al. 2016). Ideally, both technologies are used for studies, for example, RNA-seq reads can be utilized to correct sequencing errors for LRS reads employing tools like LoRDEC (Salmela & Rivals 2014) or proovread (Hackl et al. 2014), to verify splice junctions (Zhang et al. 2019a) or for hybrid *de novo* genome assemblies (Li et al. 2020). Furthermore, DGE can be performed with RNA-seq reads using the generated LRS transcriptome as a reference sequence if no genome sequence is available. In our study, we performed DGE exemplarily for the rice *aus* cultivar N22 (Paper 3). *Aus* cultivars are known to be more stress-tolerant than *indica* or *japonica* cultivars and contain genes, such as the phosphate starvation tolerance gene *OsPSTOL1* (Gamuyao et al. 2012) that are absent in the Nipponbare reference genome. The RNA-seq samples included developing seeds obtained from plants grown under control and combined drought and heat stress in the field (Lawas et al. 2018). More than 50 significantly differentially expressed genes were identified as unique to the *aus* subspecies transcriptomes in our study, most of the gene products were annotated as homologous to an *Arabidopsis thaliana* gene (Paper 3). In more detail, one gene product (*Rab21*) was characterized that has homologs in Nipponbare and different *Oryza* wild relatives and which is probably induced through water deficit (Mundy & Chua 1988). Multiple sequence alignment studies revealed that *Rab21* in N22 was closer related to *Rab21* isoforms from rice wild species than to the homolog from Nipponbare. Other studies have previously successfully utilized this approach to identify stress-

responsive genes in non-model plant organisms such as in ryegrass under cadmium stress (Hu et al. 2020) or heat and drought stress in pearl millet (Sun et al. 2020).

3.3 Utilizing metabolomics to understand molecular responses during abiotic stress

Metabolites are essential for plants regarding growth, development, and defense against climatic alterations or natural predators (Oikawa et al. 2008). Abiotic stress can dramatically affect the plant metabolome, such as heat and drought (Lawas et al. 2019, Das et al. 2017), salt (Siahpoosh et al. 2012), or HNT stress (Dhatt et al. 2019, Glaubitz et al. 2017). Not surprisingly, the field of metabolomics has become an important tool for crop breeding and improvement in the last years (Fernie & Schauer 2009).

3.3.1 Metabolite profiles of rice are affected by season upon high night temperature stress

Next to transcriptomics approaches, we performed metabolite profiling for eight rice cultivars during the wet and dry season upon HNT stress (Paper 2). Here, primary metabolites were measured for flag leaves and panicles using gas-chromatography mass-spectrometry (GCMS). So far, only a few studies exist analyzing the impact of HNT on the molecular level under controlled or field conditions, mostly for sink and source tissues (Table 1). Additionally, we analyzed the differences in the metabolite profiles depending on season. Previous studies have shown that the metabolite profile in plants can differ dramatically depending on the season (Gong et al. 2020, Kim et al. 2015b).

For molecular analyses, research in crop plants needs an agronomic characterization to connect molecular changes to a phenotype or trait. A key parameter here is the total grain yield. For this parameter a reduction was observed for all cultivars during the wet season (WS) upon HNT while for the dry season (DS) no significant yield decrease occurred (Paper 2). Under controlled conditions total grain yield was higher for most cultivars in the DS compared to the WS which was also reported before by other groups (Zhao & Fitzgerald 2013). On the physiological level grain yield is influenced by carbon and nitrogen flux to the grain that is affected by HNT (Mohammed & Tarpley 2011). The carbon loss could be caused by respiration which is known to be increased under HNT (Glaubitz et al. 2014) and may have a strong effect on biomass and yield (Shi et al. 2016, Shi et al. 2013, Peng et al. 2004). Therefore, it might be responsible for a decline in the assimilation supply to developing grains (Xiong et al. 2017). This hypothesis can be supported by the metabolite data obtained during the WS (Paper 2) where glycolysis intermediates such as sugar phosphates and sucrose levels were decreased while the abundance of monosaccharides was increased in panicles. Glycolysis generates biosynthetic intermediates

for respiration and a high turnover indicated by reduced levels of intermediates could be expected.

During HNT stress accumulation of amino acids and regulation of related pathways were reported, such as for the shikimate pathway. Metabolites related to this pathway including tyrosine, shikimic acid, and quinic acid, were significantly affected by HNT treatment in winter wheat leaves (Impa et al. 2019) and rice leaves (Glaubitz et al. 2015) under controlled environmental conditions. Increased accumulation of alanine and phenylalanine was observed for developing rice seeds (Dhatt et al. 2019) and wheat spikes (Impa et al. 2019). During the DS, also higher amounts of alanine in panicles were observed (Paper 2). Alanine is synthesized by the enzyme alanine aminotransferase (*AlaAT*) that catalyzes the reversible synthesis from pyruvate and glutamic acid (Good et al. 2007). The pathway is connected to carbon fixation and nitrogen metabolism. Overexpression of *AlaAT* in rice resulted in increased nitrogen uptake efficiency, higher biomass, and seed yield (Beatty et al. 2009). In our study, *AlaAT* activity showed a moderate increase during the DS that may have led to increased nitrogen assimilation and higher yield (Paper 2).

In both seasons, the polyols arabitol and erythritol were significantly increased in flag leaves and panicles among most of the cultivars upon HNT (Paper 2). The accumulation of polyols in many plant species is a common response to abiotic stress. These metabolites interact with membranes, protein complexes, or enzymes and act as antioxidants (Djilianov et al. 2005). A study in rice analyzing the metabolome under combined heat and drought stress reported an accumulation of arabitol and erythritol in flowering spikelets and developing seeds (Lawas et al. 2019). Both metabolites were identified as potential metabolic markers to predict combined heat and drought tolerance. However, in our studies, no correlations between sugar alcohol levels and grain yield were identified and thus, the accumulation of these metabolites may be an unspecific response to HNT (Paper 2).

Finally, correlation analysis between grain yield reduction and changed metabolite contents between control and HNT conditions for the WS revealed seven significant positive correlations among panicle metabolites. They included one unidentified compound, 3-cyano alanine, asparagine, aspartic acid, glutamic acid, pyroglutamic acid, and fructose-6-phosphate. Asparagine and aspartic acid have been identified in previous studies to be associated with HNT sensitivity under controlled conditions and higher levels were reported for HNT-sensitive rice cultivars (Glaubitz et al. 2017, Glaubitz et al. 2015). Thereby, these two metabolites can be possible molecular markers for further breeding attempts to improve HNT tolerance in rice.

3.3.2 Further applications of metabolomics in plant breeding

Our study has focused on the changes in metabolite levels of different rice cultivars under HNT stress during the DS and the WS. However, metabolomics can be extended to identify molecular markers connected to genetic regions in plants, including major crops (Matsuda et al. 2015, Schauer et al. 2006, Fiehn et al. 2000). More specifically, the integration of metabolomics, linkage mapping studies, and metabolome-based genome-wide association studies (mGWAS) provide comprehensive insight into the extent of natural variation in metabolism and its genetic control in plants (Chaudhary et al. 2019). For example, the adaption of barley to drought and combined heat and drought stress were analyzed by using genotype and metabolite data (Templer et al. 2017). Through the combined analysis of mQTL mapping and mGWAS, three major QTL for metabolites were identified involved in antioxidative defense that co-localize with genes of the corresponding pathways.

3.4 Future directions

The results generated from these studies have revealed novel insights into mapper performance for RNA-seq data generated from *Arabidopsis thaliana* (Paper 1), how to utilize PacBio IsoSeq to identify novel stress-responsive genes for rice cultivars without a genome reference (Paper 3) and into the analysis of metabolite changes in rice under HNT stress and season (Paper 2).

The first study showed that all tested mappers provided highly similar results for mapping Illumina reads of two polymorphic Arabidopsis accessions to a reference sequence. In plants, not only mapping performance regarding polymorphism but also analyzing the performance of mapping tools for polyploid species would be interesting. Only a few studies exist comparing mapping pipelines with data from polyploid species such as for a tetraploid blueberry cultivar (Payá-Milans et al. 2018). However, the genome and transcriptome assembly of polyploid species is still a challenge but LRS can help immensely to reconstruct the necessary high-quality reference sequences (Kyriakidou et al. 2018).

In the second part, PacBio IsoSeq was used to reconstruct the transcriptome of ten rice cultivars from different subspecies (Paper 3). This approach provides a general, cost-effective alternative to whole-genome sequencing for the identification of candidate genes in highly stress-tolerant ‘exotic’ genotypes without an available genome sequence. The example of the *Rab21* gene showed that the identification of novel genes with annotated homologs in other cultivars or species can yield interesting information. In the case of the *Rab21* proteins from Nipponbare, N22, and several wild relatives of *Oryza sativa*, the analysis suggests mutational studies that could be performed to understand the potential functional significance of the relatively minor

amino acid sequence differences among these proteins. Also, the data provide a comprehensive resource for the identification of other interesting candidate genes that could be functionally characterized using available functional genomics tools and to improve the abiotic stress tolerance of rice through targeted molecular breeding.

In the last study, the response of agronomic parameters and metabolic patterns to HNT has been analyzed for eight rice cultivars under field conditions in two different seasons (Paper 2). Possible marker metabolites (asparagine, aspartic acid) were identified that could be used as a starting point for metabolomics-based breeding. These markers will still need further validation involving more cultivars and experiments before an established biomarker could be determined (Zabotina 2013). As metabolomics is considered to complement other 'omics' technologies, transcript-based profiling could give a more holistic overview of the response to HNT stress and could be used as a source for additional transcript-markers. An integrated data analysis using the metabolite and transcriptome data can give novel insights into pathways and regulated genes under field conditions using tools such as MetaboAnalyst (Chong et al. 2019) or MapMan (Schwacke et al. 2019), both previously applied for rice under HNT stress at controlled conditions (Glaubitz et al. 2017). For further breeding programs the identification of genetic regions correlated to HNT stress by using mQTL and/or mGWAS approaches could be interesting and so far, have not been performed for HNT (Xu et al. 2020b).

Clearly, these studies cover only a small part of the huge field of omics technologies. Nevertheless, the insights discovered in the transcriptomic studies can help researchers to choose the right tool and explore new technologies for their research questions. Additionally, the metabolite profiling in rice under HNT is a first step that can help to develop cultivars with higher stress resilience. In the future, 'omics' approaches need to be integrated and connected to understand complex traits in crop plants and for the application of novel insights in molecular breeding to ensure global food security.

References

Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, et al. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun* 7: 11706

Akpınar BA, Lucas SJ, Budak H. 2013. Genomics approaches for crop improvement against abiotic stress. *Sci World J* 2013: 361921

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 21: 30

Arbona V, Manzi M, Ollas C, Gomez-Cadenas A. 2013. Metabolomics as a tool to investigate abiotic stress tolerance in plants. *Int J Mol Sci* 14: 4885-911

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Res* 46: 2159-68

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, et al. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *PNAS* 110: E4821-E30

Babarinde IA, Li Y, Hutchins AP. 2019. Computational methods for mapping, assembly and quantification for coding and non-coding transcripts. *Comput Struct Biotechnol J* 17: 628-37

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, et al. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 546: 148-52

Bahuguna RN, Jha J, Pal M, Shah D, Lawas LMF, et al. 2015. Physiological and biochemical characterization of NERICA-L-44: A novel source of heat tolerance at the vegetative and reproductive stages in rice. *Physiol Plan* 154: 543-59

Bahuguna RN, Solis CA, Shi W, Jagadish KS. 2017. Post-flowering night respiration and altered sink activity account for high night temperature-induced grain yield and quality loss in rice (*Oryza sativa L.*). *Physiol Plant* 159: 59-73

Beatty PH, Shrawat AK, Carroll RT, Zhu T, Good AG. 2009. Transcriptome analysis of nitrogen-efficient rice over-expressing alanine aminotransferase. *Plant Biotechnology Jl* 7: 562-76

Benjamin AM, Nichols M, Burke TW, Ginsburg GS, Lucas JE. 2014. Comparing reference-based RNA-seq mapping methods for non-human primate data. *BMC Genomics* 15: 570

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40: 10-e72

Bierschenk B, Tagele MT, Ali B, Ashrafuzzaman Md, Wu LB, et al. 2020. Evaluation of rice wild relatives as a source of traits for adaptation to iron toxicity and enhanced grain quality. *PLOS ONE* 15: e0223086

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34: 525-27

Casartelli A, Riewe D, Hubberten HM, Altmann T, Hoefgen R, Heuer S. 2018. Exploring traditional *aus*-type rice for metabolites conferring drought tolerance. *Rice* 11: 9

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13: 238

Chaudhary J, Khatri P, Singla P, Kumawat S, Kumari A, et al. 2019. Advances in omics approaches for abiotic stress tolerance in tomato. *Biology* 8: 90

Chen SY, Deng F, Jia X, Li C, Lai SJ. 2017. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci Rep* 7: 7648

Chen W, Wang W, Peng M, Gong L, Gao Y, et al. 2016. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat Commun* 7: 12767

Chinnusamy V, Jagendorf A, Zhu J-K. 2005. Understanding and improving salt tolerance in plants. *Crop Sci* 45: 437-48

Chong J, Wishart DS, Xia J. 2019. Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Curr Protoc Bioinformatics* 68: e86

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, et al. 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 Res* 27: 885-96

Collins FS, Morgan M, Patrinos A. 2003. The Human Genome Project: Lessons from large-scale biology. *Science* 300: 286-90

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol* 17: 13

Das A, Rushton PJ, Rohila JS. 2017. Metabolomic profiling of soybeans (*Glycine max L.*) reveals the importance of sugar and nitrogen metabolism under drought and heat stress. *Plants* 6: 21

Davy R, Esau I, Chernokulsky A, Outten S, Zilitinkevich S. 2017. Diurnal asymmetry to the observed global warming. *Int J Climatol* 37: 79-93

Dawid C, Hille K. 2018. Functional metabolomics—a useful tool to characterize stress-induced metabolome alterations opening new avenues towards tailoring food crop quality. *Agronomy* 8:138

Deamer D, Akeson M, Branton D. 2016. Three decades of nanopore sequencing. *Nat Biotechnology* 34: 518-24

Degenkolbe T, Do PT, Kopka J, Zuther E, Hincha DK, Köhl KI. 2013. Identification of drought tolerance markers in a diverse population of rice cultivars by expression and metabolite profiling. *PLOS ONE* 8: e63637

Dhatt BK, Abshire N, Paul P, Hasanthika K, Sandhu J, et al. 2019. Metabolic dynamics of developing rice seeds under high night-time temperature stress. *Fron Plant Sci* 10: 1443

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, et al. 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinformatics* 14: 671-83

Dimon MT, Sorber K, DeRisi JL. 2010. HMMSplicer: A tool for efficient and sensitive discovery of known and novel splice junctions in RNA-seq data. *PLOS ONE* 5: e13875

Djilianov D, Georgieva T, Moyankova D, Atanassov A, Shinozaki K, et al. 2005. Improved abiotic stress tolerance in plants by accumulation of osmoprotectants—gene transfer approach. *Biotechnol Biotechnol Equip* 19: 63-71

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21

Doebley JF, Gaut BS, Smith BD. 2006. The molecular genetics of crop domestication. *Cell* 127: 1309-21

Dong J, Feng Y, Kumar D, Zhang W, Zhu T, et al. 2016. Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *PNAS* 113: 7949-56

Du H, Yu Y, Ma Y, Gao Q, Cao Y, et al. 2017. Sequencing and *de novo* assembly of a near complete *indica* rice genome. *Nat Comm* 8: 15324

Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, et al. 2013. Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 9: 44-66
Easterling DR, Horton B, Jones PD, Peterson TC, Karl TR, et al. 1997. Maximum and minimum temperature - trends for the globe. *Science* 277: 364-67
Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133-8
Ellstrand NC, Meirmans P, Rong J, Bartsch D, Ghosh A, et al. 2013. Introgression of crop alleles into wild or weedy populations. *Annu Rev Ecol Evol Syst* 44: 325-45
Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Met* 10: 1185-91
Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. 2015. Most highly expressed protein-coding genes have a single dominant isoform. *J Prot Res* 14: 1880-87
Fang C, Luo J. 2019. Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J* 97: 91-100
FAO. 2009. How to feed the world in 2050. *Rome, FAO*
FAOSTAT. 2020. http://www.fao.org/faostat/en/. Accessed 21.07.2020
Feng S, Xu M, Liu F, Cui C, Zhou B. 2019. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in *Gossypium australe*. *BMC Plant Biol* 19: 365
Fernie AR, Schauer N. 2009. Metabolomics-assisted breeding: A viable option for crop improvement? *TIG* 25: 39-48
Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. 2000. Metabolite profiling for plant functional genomics. *Nat Biotechnol* 18: 1157-61
Finotello F, Di Camillo B. 2015. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* 14: 130-42
Fonseca NA, Marioni J, Brazma A. 2014. RNA-seq gene profiling - a systematic empirical comparison. *PLOS ONE* 9: e107026
Gamuyao R, Chin JH, Pariasca-Tanaka J, Pesaresi P, Catausan S, et al. 2012. The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* 488: 535
Gichner T, Meyerowitz E, Somerville C. 1995. Arabidopsis. *Biol Plant* 37: 540
Glaubitz U, Erban A, Kopka J, Hincha DK, Zuther E. 2015. High night temperature strongly impacts TCA cycle, amino acid and polyamine biosynthetic pathways in rice in a sensitivity-dependent manner. *J Exp Bot* 66: 6385-97
Glaubitz U, Li X, Köhl KI, van Dongen JT, Hincha DK, Zuther E. 2014. Differential physiological responses of different rice (*Oryza sativa*) cultivars to elevated night temperature during vegetative growth. *Funct Plant Biol* 41: 437
Glaubitz U, Li X, Schaedel S, Erban A, Sulpice R, et al. 2017. Integrated analysis of rice transcriptomic and metabolomic responses to elevated night temperatures identifies sensitivity- and tolerance-related profiles. *PCE* 40: 121-37
Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. 2002. A draft sequence of the rice genome (*Oryza sativa L.* ssp. *japonica*). *Science* 296: 92-100
Gong AD, Lian SB, Wu NN, Zhou YJ, Zhao SQ, et al. 2020. Integrated transcriptomics and metabolomics analysis of catechins, caffeine and theanine biosynthesis in tea plant (*Camellia sinensis*) over the course of seasons. *BMC Plant Biol* 20: 294
Good AG, Johnson SJ, De Pauw M, Carroll RT, Savidov N, et al. 2007. Engineering nitrogen use efficiency with alanine aminotransferase. *Can J Bot* 85: 252-62
Hackl T, Hedrich R, Schultz J, Forster F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30: 3004-11

Hall RD. 2006. Plant metabolomics: From holistic hope, to hype, to hot topic. *New Phytol* 169: 443
Hartman Y, Hooftman DAP, Uwimana B, Schranz ME, van de Wiel CCM, et al. 2014. Abiotic stress QTL in lettuce crop-wild hybrids: Comparing greenhouse and field experiments. *Ecol Evol* 4: 2395-409
Hasin Y, Seldin M, Lusis A. 2017. Multi-omics approaches to disease. *Genome Biol* 18: 83
Hatem A, Bozdağ D, Toland AE, Çatalyürek Ü V. 2013. Benchmarking short sequence mapping tools. *BMC Bioinformatics* 14: 184
Henderson IR, Salt DE. 2017. Natural genetic variation and hybridization in plants. *J Exp Bot* 68: 5415-17
Herranz R, Vandenbrink JP, Villacampa A, Manzano A, Poehlman WL, et al. 2019. RNA-seq analysis of the response of *Arabidopsis thaliana* to fractional gravity under blue-light stimulation during spaceflight. *Fron Plant Sci* 10: 1529
Hofmann F, Schon MA, Nodine MD. 2019. The embryonic transcriptome of *Arabidopsis thaliana*. *Plant Repro* 32: 77-91
Hu Z, Zhang Y, He Y, Cao Q, Zhang T, et al. 2020. Full-length transcriptome assembly of italian ryegrass root integrated with RNA-seq to identify genes in response to plant cadmium stress. *Int J Mol Sci* 21: 1067
Impa SM, Sunoj VSJ, Krassovskaya I, Bheemanahalli R, Obata T, Jagadish SVK. 2019. Carbon balance and source-sink metabolic changes in winter wheat exposed to high night-time temperature. *PCE* 42: 1233-46
IPCC. 2014. AR5 Climate change 2014: Impacts, adaptation, and vulnerability. Cambridge Univ. Press, Cambridge, UK
Jagadish SVK, Craufurd PQ, Wheeler TR. 2008. Phenotyping parents of mapping populations of rice for heat tolerance during anthesis. *Crop Sci* 48: 1140-46
Jagadish SVK, Murty MVR, Quick WP. 2015. Rice responses to rising temperatures – challenges, perspectives and future directions. *PCE* 38: 1686-98
Jaluria P, Konstantopoulos K, Betenbaugh M, Shiloach J. 2007. A perspective on microarrays: Current applications, pitfalls, and potential uses. *Microb Cell Fact* 6: 4
Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, et al. 2017. The genome of *Chenopodium quinoa*. *Nature* 542: 307-12
Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, et al. 2017. Improved maize reference genome with single-molecule technologies. *Nature* 546: 524-27
Kim D, Langmead B, Salzberg SL. 2015a. HISAT: A fast spliced aligner with low memory requirements. *Nat Met* 12: 357-60
Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37: 907-15
Kim NK, Park HM, Lee J, Ku KM, Lee CH. 2015b. Seasonal variations of metabolome and tyrosinase inhibitory activity of *Lespedeza maximowiczii* during growth periods. *J Agri Food Chem* 63: 8631-39
Koornneef M, Meinke D. 2010. The development of Arabidopsis as a model plant. *Plant J* 61: 909-21
Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 27: 722-36
Korf I. 2013. Genomics: the state of the art in RNA-seq analysis. *Nat Met* 10: 1165-6
Kraft F, Kurth I. 2019. Long-read sequencing in human genetics. *Med Gen* 31: 198-204
Krasensky J, Jonak C. 2012. Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J Exp Bot* 63: 1593-608
Križanović K, Echchiki A, Roux J, Šikić M. 2017. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* 34: 748-54

Kumar R, Bohra A, Pandey AK, Pandey MK, Kumar A. 2017. Metabolomics for plant improvement: Status and prospects. *Front Plant Sci* 8: 1302-02
Kumar S, Shanker A. 2018. Bioinformatics resources for the stress biology of plants. In *Biotic and Abiotic Stress Tolerance in Plants*, ed. S Vats, pp. 367-86. Singapore: Springer Singapore
Kuo RI, Cheng Y, Smith J, Archibald AL, Burt DW. 2019. Illuminating the dark side of the human transcriptome with TAMA Iso-Seq analysis. *bioRxiv*
Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. 2018. Current strategies of polyploid plant genome sequence assembly. *Front Plant Sci* 9: 1660-60
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Gen Biol* 10: R25
Lawas LMF, Li X, Erban A, Kopka J, Jagadish SVK, et al. 2019. Metabolic responses of rice cultivars with different tolerance to combined drought and heat stress under field conditions. *Gigascience* 8: 5
Lawas LMF, Shi W, Yoshimoto M, Hasegawa T, Hincha DK, et al. 2018. Combined drought and heat stress impact during flowering and grain filling in contrasting rice cultivars grown under field conditions. *Field Crops Res* 229: 66-77
Leinonen R, Sugawara H, Shumway M. 2011. The sequence read archive. *Nucleic Acids Res* 39: D19-21
Li C, Lin F, An D, Wang W, Huang R. 2017a. Genome sequencing and assembly by long reads in plants. *Genes (Basel)* 9: 6-1
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094-100
Li H, Chen Z, Hu M, Wang Z, Hua H, et al. 2011. Different effects of night versus day high temperature on rice quality and accumulation profiling of rice grain proteins during grain filling. *Plant Cell Rep* 30: 1641-59
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-60
Li JY, Wang J, Zeigler RS. 2014. The 3,000 rice genomes project: New opportunities and challenges for future rice research. *GigaScience* 3: 8
Li J, Harata-Lee Y, Denton MD, Feng Q, Rathjen JR, et al. 2017b. Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus membranaceus* reveals transcript variants involved in bioactive compound biosynthesis. *Cell Disc* 3: 17031
Li W, Li K, Zhang QJ, Zhu T, Zhang Y, et al. 2020. Improved hybrid *de novo* genome assembly and annotation of African wild rice, *Oryza longistaminata*, from Illumina and PacBio sequencing reads. *Plant Gen* 13: e20001
Li WV, Li JJ. 2018. Modeling and analysis of RNA-seq data: A review from a statistical perspective. *Quant Biol* 6: 195-209
Li X, Lawas LM, Malo R, Glaubitz U, Erban A, et al. 2015. Metabolic and transcriptomic signatures of rice floral organs reveal sugar starvation as a factor in reproductive failure under heat and drought stress. *PCE* 38: 2171-92
Liang J, Xia J, Liu L, Wan S. 2013. Global patterns of the responses of leaf-level photosynthesis and respiration in terrestrial plants to experimental warming. *J Plant Ecol* 6: 437-47
Liao JL, Zhou HW, Peng Q, Zhong PA, Zhang HY, et al. 2015. Transcriptome changes in rice (*Oryza sativa L.*) in response to high night temperature stress at the early milky stage. *BMC Genomics* 16: 18
Limpens J, Granath G, Aerts R, Heijmans MMPD, Sheppard LJ, et al. 2012. Glasshouse vs field experiments: Do they yield ecologically similar results for assessing N impacts on peat mosses? *New Phytol* 195: 408-18

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR. 2006. Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nat Prot* 1: 387-96
Love MI, Hogenesch JB, Irizarry RA. 2016. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat Biotech* 34: 1287-91
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Gen Biol* 15: 550
Luo J. 2015. Metabolite-based genome-wide association studies in plants. *Curr Opinion Plant Biol* 24: 31-38
Matich EK, Chavez Soria NG, Aga DS, Atilla-Gokcumen GE. 2019. Applications of metabolomics in assessing ecological effects of emerging contaminants and pollutants on plants. *Journal of Hazardous Materials* 373: 527-35
Matsuda F, Nakabayashi R, Yang Z, Okazaki Y, Yonemaru J, et al. 2015. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant J* 81: 13-23
Mayer K, Schüller C, Wambutt R, Murphy G, Volckaert G, et al. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402: 769
Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M. 1998. *Arabidopsis thaliana*: A model plant for genome analysis. *Science* 282: 662-82
Michael TP, Jackson S. 2013. The first 50 plant genomes. *Plant Gen* 6: 1-7
Miller JR, Zhou P, Mudge J, Gurtowski J, Lee H, et al. 2017. Hybrid assembly with long and short reads improves discovery of gene family expansions. *BMC Genomics* 18: 541
Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D. 2019. Iso-Seq allows genome-independent transcriptome profiling of grape berry development. *G3 (Bethesda)* 9: 755-67
Mohammed AR, Tarpley L. 2011. Effects of night temperature, spikelet position and salicylic acid on yield and yield-related parameters of rice (*Oryza sativa L.*) plants. *J Agro Crop Sci* 197: 40-49
Mohammed R, Cothren JT, Tarpley L. 2013. High night temperature and abscisic acid affect rice productivity through altered photosynthesis, respiration and spikelet fertility. *Crop Sci* 53: 2603-12
Mundy J, Chua NH. 1988. Abscisic acid and water-stress induce the expression of a novel rice gene. *EMBO J* 7: 2279-86
Mysore KS, Tuori RP, Martin GB. 2001. Arabidopsis genome sequence as a tool for functional genomics in tomato. *Gen Biol* 2: 1
Nagarajan S, Jagadish SVK, Prasad ASH, Thomar AK, Anand A, et al. 2010. Local climate affects growth, yield and grain quality of aromatic and non-aromatic rice in northwestern India. *Agric Ecosyst Environ* 138: 274-81
Nakabayashi R, Saito K. 2015. Integrated metabolomics for abiotic stress responses in plants. *Curr Opin Plant Biol* 24: 10-16
Nakabayashi R, Saito K. 2020. Higher dimensional metabolomics using stable isotope labeling for identifying the missing specialized metabolism in plants. *Curr Opin Plant Biol* 55: 84-92
Nolan T, Hands RE, Bustin SA. 2006. Quantification of mRNA using real-time RT-PCR. *Nat Prot* 1: 1559-82
Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, et al. 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 Res* 40: 10084-97
Obata T, Fernie AR. 2012. The use of metabolomics to dissect plant responses to abiotic stresses. *Cell Mol Life Sci* 69: 3225-43

Oikawa A, Matsuda F, Kusano M, Okazaki Y, Saito K. 2008. Rice metabolomics. *Rice* 1: 63-71
Okoniewski MJ, Miller CJ. 2006. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* 7: 276
Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Met* 14: 417
Payá-Milans M, Olmstead JW, Nunez G, Rinehart TA, Staton M. 2018. Comprehensive evaluation of RNA-seq analysis pipelines in diploid and polyploid species. *GigaScience* 7: giy132
Peng S, Huang J, Sheehy JE, Laza RC, Visperas RM, et al. 2004. Rice yields decline with higher night temperature from global warming. *PNAS* 101: 9971-75
Perez de Souza L, Scossa F, Proost S, Bitocchi E, Papa R, et al. 2019. Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (*Phaseolus vulgaris*) metabolome. *Plant J* 97: 1132-53
Qin F, Shinozaki K, Yamaguchi-Shinozaki K. 2011. Achievements and challenges in understanding plant abiotic stress responses and tolerance. *Plant Cell Physiol* 52: 1569-82
Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, et al. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 14: R95
Ray DK, Ramankutty N, Mueller ND, West PC, Foley JA. 2012. Recent patterns of crop yield growth and stagnation. *Nat Commun* 3: 1293
Raza A, Razzaq A, Mehmood SS, Zou X, Zhang X, et al. 2019. Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. *Plants (Basel)* 8: 34
RGP. 2014. The 3,000 rice genomes project. *GigaScience* 3: 7
Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *GPB* 13: 278-89
Saade S, Maurer A, Shahid M, Oakey H, Schmöckel SM, et al. 2016. Yield-related salinity tolerance traits identified in a nested association mapping (NAM) population of wild barley. *Sci Rep* 6: 32586
Salmela L, Rivals E. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30: 3506
Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, et al. 2008. Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiol* 50: 37-47
Schaarschmidt S, Lawas LMF, Glaubitz U, Li X, Erban A, et al. 2020. Season affects yield and metabolic profiles of rice (*Oryza sativa*) under high night temperature stress in the field. *Int J Mol Sci* 21: 3187
Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, et al. 2014. Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Gen Biol* 15: 506
Schauer N, Semel Y, Roessner U, Gur A, Balbo I, et al. 2006. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat Biotechnol* 24: 447-54
Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, et al. 2019. MapMan4: A refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol Plant* 12: 879-92
Seyednasrollah F, Laiho A, Elo LL. 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinformatics* 16: 59-70
Shah T, Xu J, Zou X, Cheng Y, Nasir M, Zhang X. 2018. Omics approaches for engineering wheat production under abiotic stresses. *Int J Mol Sci* 19: 2390

Sharma N, Yadav A, Khetarpal S, Anand A, Sathee L, et al. 2017. High day–night transition temperature alters nocturnal starch metabolism in rice (*Oryza sativa L.*). *Acta Physiol Plant* 39: 74

Shen S, Park JW, Lu Z-x, Lin L, Henry MD, et al. 2014. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *PNAS* 111: E5593

Shi W, Muthurajan R, Rahman H, Selvam J, Peng S, et al. 2013. Source-sink dynamics and proteomic reprogramming under elevated night temperature and their impact on rice yield and grain quality. *New Phytol* 197: 825-37

Shi W, Yin X, Struik PC, Solis C, Xie F, et al. 2017. High day- and night-time temperatures affect grain growth dynamics in contrasting rice genotypes. *J Exp Bot* 68: 5233-45

Shi W, Yin X, Struik PC, Xie F, Schmidt RC, Jagadish KSV. 2016. Grain yield and quality responses of tropical hybrid rice to high night-time temperature. *Field Crops Re* 190: 18-25

Siahpoosh MR, Sanchez DH, Schlereth A, Scofield GN, Furbank RT, et al. 2012. Modification of *OsSUT1* gene expression modulates the salt response of rice *Oryza sativa* cv. Taipei 309. *Plant Sci* 182: 101-11

Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14: 91

Sprenger H, Erban A, Seddig S, Rudack K, Thalhammer A, et al. 2018. Metabolite and transcript markers for the prediction of potato drought tolerance. *Plant Biotechnol J* 16: 939-50

Sprenger H, Kurowsky C, Horn R, Erban A, Seddig S, et al. 2016. The drought response of potato reference cultivars with contrasting tolerance. *PCE* 39: 2370-89

Stark R, Grzelak M, Hadfield J. 2019. RNA sequencing: The teenage years. *Nat Rev Genet* 20: 631-56

Steijger T, Abril JF, Engström PG, Kokocinski F, Consortium R, et al. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Met* 10: 1177-84

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat Genet* 50: 285-96

Sun M, Huang D, Zhang A, Khan I, Yan H, et al. 2020. Transcriptome analysis of heat stress and drought stress in pearl millet based on Pacbio full-length transcriptome sequencing. *BMC Plant Biol* 20: 323-23

Telfer P, Edwards J, Bennett D, Ganesalingam D, Able J, Kuchel H. 2018. A field and controlled environment evaluation of wheat (*Triticum aestivum*) adaptation to heat stress. *Field Crops Res* 229: 55-65

Templer SE, Ammon A, Pscheidt D, Ciobotea O, Schuy C, et al. 2017. Metabolite profiling of barley flag leaves under drought and combined heat and drought stress reveals metabolic QTLs for metabolites associated with antioxidant defense. *J Exp Bot* 68: 1697-713

Teng K, Teng W, Wen H, Yue Y, Guo W, et al. 2019. PacBio single-molecule long-read sequencing shed new light on the complexity of the *Carex breviculmis* transcriptome. *BMC Genomics* 20: 789

Teng M, Love MI, Davis CA, Djebali S, Dobin A, et al. 2016. Erratum to: A benchmark for RNA-seq quantification pipelines. *Genome Biol* 17: 203

The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815

Vailati-Riboni M, Palombo V, Loor JJ. 2017. What are omics sciences? In *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*, ed. BN Ametaj, pp. 1-7. Cham: Springer International Publishing

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* 27: 737-46
Vose RS, Easterling DR, Gleason B. 2005. Maximum and minimum temperature trends for the globe: An update through 2004. *Geophys Res Letters* 32: L23822
Wang M, Wang P, Liang F, Ye Z, Li J, et al. 2018. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytol* 217: 163-78
Wang Z, Gerstein M, Snyder M. 2009. RNA-seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63
Wheeler T, von Braun J. 2013. Climate change impacts on global food security. *Science* 341: 508
Workman RE, Myrka AM, Wong GW, Tseng E, Welch KC, Jr., Timp W. 2018. Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *GigaScience* 7: 1-12
Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO. 2018. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics* 19: 510
Wu I, Ben-Yehezkel T. 2019. A Single-Molecule Long-Read survey of human transcriptomes using LoopSeq synthetic long read sequencing. *bioRxiv*: 532135
Xie L, Teng K, Tan P, Chao Y, Li Y, et al. 2020. PacBio single-molecule long-read sequencing shed new light on the transcripts and splice isoforms of the perennial ryegrass. *Mol Genet Genomics* 295: 475-89
Xiong D, Ling X, Huang J, Peng S. 2017. Meta-analysis and dose-response analysis of high temperature effects on rice yield and quality. *Environ Experimen Bot* 141: 1-9
Xu J, Fang M, Li Z, Zhang M, Liu X, et al. 2020a. Third-generation sequencing reveals LncRNA-regulated HSP genes in the Populus x Canadensis Moench heat stress response. *Front Gen* 11: 249-49
Xu J, Henry A, Sreenivasulu N. 2020b. Rice yield formation under high day and night temperatures—A prerequisite to ensure future food security. *PCE* 43: 1595-608
Yamakawa H, Hakata M. 2010. Atlas of rice grain filling-related metabolism under high temperature: Joint analysis of metabolome and transcriptome demonstrated inhibition of starch accumulation and induction of amino acid accumulation. *Plant Cell Physiol* 51
Yang X, Wang B, Chen L, Li P, Cao C. 2019. The different influences of drought stress at the flowering stage on rice physiological traits, grain yield, and quality. *Sci Rep* 9: 3742
Zabotina OA. 2013. Metabolite-based biomarkers for plant genetics and breeding. In *Diagnostics in Plant Breeding*, ed. T Lübberstedt, RK Varshney, pp. 281-309. Dordrecht: Springer Netherlands
Zhang C, Zhang B, Lin LL, Zhao S. 2017a. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* 18: 583
Zhang G, Sun M, Wang J, Lei M, Li C, et al. 2019a. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. *Plant J* 97: 296-305
Zhang J, Chen LL, Sun S, Kudrna D, Copetti D, et al. 2016. Building two *indica* rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. *Sci Data* 3: 160076
Zhang R, Calixto CPG, Marquez Y, Venhuizen P, Tzioutziou NA, et al. 2017b. A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res* 45: 5061-73
Zhang Y, Malzahn AA, Sretenovic S, Qi Y. 2019b. The emerging and uncultivated potential of CRISPR technology in plant science. *Nat Plants* 5: 778-94

Zhang Y, Tang Q, Peng S, Zou Y, Chen S, et al. 2013. Effects of high night temperature on yield and agronomic traits of irrigated rice under field chamber system condition. *Aust J Crop Sci* 7: 7-13

Zhao X, Fitzgerald M. 2013. Climate change: Implications for the yield of edible rice. *PLOS ONE* 8: e66218

Zhao Y, Wang K, Wang Wl, Yin TT, Dong WQ, Xu CJ. 2019. A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics* 20: 160

Zhou A, Breese MR, Hao Y, Edenberg HJ, Li L, et al. 2012. Alt Event Finder: A tool for extracting alternative splicing events from RNA-seq data. *BMC Genomics* 13: S10

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience* 6: 1-7

Zuther E, Schaarschmidt S, Fischer A, Erban A, Pagter M, et al. 2019. Molecular signatures associated with increased freezing tolerance due to low temperature memory in Arabidopsis. *PCE* 42: 854-73

Zuther E, Schulz E, Childs LH, Hincha DK. 2012. Clinal variation in the non-acclimated and cold-acclimated freezing tolerance of *Arabidopsis thaliana* accessions. *PCE* 35: 1860-78

Acknowledgements

In the end of the thesis I faced a big loss. My supervisor PD Dr. Dirk K. Hincha died a sudden dead in the month of submission. I hope, he knew how grateful I was for his great supervision, his open-mindedness, his ideas and his support on the scientific as well as personal level over all the years. He was never tired explaining things to me and taking the time to discuss and develop our projects. He was a great mentor and helped me and all his students to be confident but also critical about our science. This work is dedicated to him.

I truely want to thank Dr. Ellen Zuther, who not only co-supervised me all these years by giving me great scientific and personal advice, but also supporting me and our group immensively. She keeps us moving in these really sad times.

I want to thank the whole AG Hincha, former and current members, for the help, the friendly working atmosphere and great support. Over the years working at the MPI as a Master and PhD student, I enjoyed every day being at work with you all.

I would also like to thank apl. Prof. Dr. Dirk Walther, Axel Fischer and the whole AG Bioinformatics. Dirk Walther helped me without hesitation by taking over the supervision and welcomed me in his group for my bioinformatic analyses. Axel Fischer gave immense input for the transcriptomics studies and was always there to support and talk. Great thanks also to our collaborators Dr. Joachim Kopka, Alexander Erban and Ines Fehrle for the GC-MS measurements and help.

In addition, I want to thank my collegues at the MPI-MP for there great support during these difficult times and the Max Planck Society as well as the “Federal Ministry for Economic, Cooperartion, and Development” for funding.

Finally, all these years would have been much more difficult without my friends, partner and family. My family gave me the opportunity to study, to develop and to be confident about myself and thus my research. My friends and my partner gave me the support I needed in the past and now. Thank you to all of you!

Curriculum vitae

This page contains personal information and was removed.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass meine hinsichtlich der früheren Teilnahme an Promotionsverfahren gemachten Angaben richtig sind und, dass die eingereichte Arbeit oder wesentliche Teile derselben in keinem anderen Verfahren zur Erlangung eines akademischen Grades vorgelegt worden sind.

Ich versichere darüber hinaus, dass bei der Anfertigung der Dissertation die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der DFG eingehalten wurden, die Dissertation selbständig und ohne fremde Hilfe, insbesondere für die allgemeine Einleitung und Diskussion, verfasst wurde, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt worden sind und die den benutzten Werken wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht wurden.

Einer Überprüfung der eingereichten Dissertation bzw. der eingereichten Schriften mittels einer Plagiatsprüfungssoftware stimme ich zu.

Potsdam, den _____ Unterschrift _____