MITOCHONDRIAL SEQUENCE DIVERSITY AMONG ALLOPLASMIC

AND EUPLASMIC *TRITICUM* SPECIES

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Andrzej Krzysztof Noyszewski

In Partial Fulfillment
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Genomic and Bioinformatics Program

April 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Mitochondrial Sequence Diversity Among Alloplasmic and Euplasmic
Triticum Species

By

Andrzej Krzysztof Noyszewski

The Supervisory Committee certifies that this ***disquisition*** complies

with North Dakota State University's regulations and meets the

accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Shahryar F. Kianian
Chair

Dr. Glenn P. Dorsam

Dr. Richard D. Horsley

Dr. David P. Horvath

Dr. Phil E. McClean

Approved:

04/19/2013	Dr. Richard D. Horsley
Date	Department Chair

ABSTRACT

Four mitochondrial genomes of *Triticum* species were sequenced and annotated: 1) (lo) durum mitochondrial genome, which is an alloplasmic line with *Triticum longissimum* (SS) cytoplasm and *T. turgidum* (AABB) nucleus, 2) *T. longissimum,* 3) *T. turgidum* and 4) *T. tauschii* (DD). Comparison showed major differences in *atp6*, *nad9*, *nad6, rps19*-p, *cob* and *cox2* genes among all four species. Additionally, species-specific ORFs were also identified. A single nucleotide polymorphism search within known genes showed that the alloplasmic line differs from the two parental lines by six nucleotides in the *cox3*, *mttB, rps2*, *rps4* and *rps13* genes. We were able to recognize mitochondrial heteroplasmy based on single nucleotide variation (SNV) and regions of high SNV density within a given species. Structural differences between *T. turgidum*, (lo) durum and *T. longissimum* mitochondrial genomes were observed; however, conserved gene blocks and gene pairs among these species were identified. Three possible recombination events in gene blocks I, V and VI were recognized. We observed differences in the alloplasmic line, compared to its parental lines in: sequence, predicted genes, single nucleotide polymorphism (SNP) and genome structure. These facts support the hypothesis of the accelerated evolution of the mitochondrial genome when transferred into alien nuclear background. We also found that major gene changes recognized here appear to be common among *Triticum* species.

Based on sequence assembly, we report full mitochondrial sequence of *T. turgidum.* We recognized 40 SNP differences compared to the *T. aestivum* mitochondrial genome*,* where 5 SNPs were found in known mitochondrial genes: *rps1*, *rps2*, *cox3* and

ccmFN. The *T. longissimum* and *T. tauschii* share highly similar genomes in structure and content, different only at the level of SNVs. A method to establish phylogenetic relationships based on mitochondrial genome sequence is proposed using differences in reference assembly with a common mitochondrial backbone sequence.

We confirm conservation of the mitochondrial gene content within *Triticum* species. These results create background to explore the role of mitochondrial genes in conditioning nuclear cytoplasmic incompatibility in a wide range of *Triticum* alloplasmic lines and also deepen our understanding of evolutionary relationships that exist in the *Triticum* genus.

ACKNOWLEDGMENTS

<div align="right">Andrzej K. Noyszewski</div>

"Science can purify religion from error and superstition. Religion can purify science from idolatry and false absolutes."

<div align="right">Blessed Pope John Paul II</div>

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

GENERAL INTRODUCTION

This research is divided into two parts that are presented in two papers; each of them consists of an Abstract, Introduction, Results and Discussion. A General Introduction, Literature Review, Objectives, and Materials and Methods are presented before Paper I. The overall Conclusions, a List of References, and Appendixes with additional information concludes this dissertation.

LITERATURE REVIEW

Nuclear–cytoplasmic Interactions

The genetic information which defines a plant's phenotype is primarily stored in the nuclear genome. A secondary source of this information resides in plasmon–that is, in mitochondrial and chloroplast genomes. Contrary to a nuclear genome, organellar genomes are present in multiple copies per cell. In particular, mitochondria are highly conserved in gene content, however variable in structure (Woloszynska, 2010).

It was recognized early in the study of heredity that traits encoded by the nucleus can be altered to make plants agriculturally beneficial. The potential of genetic variation present within species was then utilized. Not much attention was given to genetic variation present in organellar genomes that could also be used for trait improvement. Mitochondria and their genetic content were, for a long time, considered highly conserved among species (Palmer and Hebron, 1988), and were not extensively analyzed for trait improvement. Another reason for this lack of interest was that the exchange of cytoplasm between related plant species typically caused cytoplasmic male sterility (CMS) as the most deleterious and prevalent characteristic of nuclear-mitochondrial incompatibility (Hanson, 1991). However, CMS remains one of the most utilized features connected with mitochondria, particularly in hybrid breeding (Weider *et al*., 2009) or as a preventive method against invasive species propagation (Hodgins *et al*., 2009). The agricultural value of genetic information within organelles has not been fully recognized and remains to be investigated.

Available research on cytoplasmic genetic variation and its utilization in alloplasmic lines of *Triticum* has proven to be valuable for trait improvement and also contributed to the understanding of nuclear-cytoplasmic (NC) interactions in plant species. Improvement of traits such as yield, quality, salt tolerance and disease resistance were observed in the context of NC interactions as a beneficial effect of the *Aegilops crassa* cytoplasm on common wheats (Wu *et al*., 1998 and Liu *et al*., 2002). Investigation of NC interactions by Tsunewaki and coworkers focused on recognition of genotype by plasmon interaction in alloplasmic lines of *Triticum* and *Aegilops*. In their latest work, authors found significant changes in studied traits, indicating a correlation of those changes with substituted cytoplasmic background (Tsunewaki *et al*., 2002). Unfortunately, observed changes are not always favorable, but still allow for a better understanding of the evolution of those genera (Tsunewaki *et al*., 1993, 1996, 2002). The NC interactions have also been studied in many other species. The alloplasmic lines of barley were studied and a number of traits were analyzed such as spike number per plant, spike length, grain number per spike, grain number per plant and grain weight per plant (Goloenko *et al*., 2000). These traits were found to increase in comparison to the original nuclear donor variety, proving that the genetic variability found in the cytoplasm can be beneficial for barley improvement (Goloenko *et al*., 2000). Recent studies on the cytoplasmic effects on seven traits of indica rice indicated significant improvement of grain weight and filled-grain ratio compared to mean values of non-alloplasmic lines (Tao *et al*., 2011). These results support the hypothesis that an opportunity for selection of the appropriate cytoplasmic genomes for a specific nuclear background exists. However, these results cannot be generalized, and each cytoplasmic effect should be

evaluated in the nuclear background of interest and at multiple locations (Tao *et al*., 2011).

Cytoplasmic male sterility is widely understood as an effect of a mitochondrial defect or mutation (Hanson, 1991). In plants, male sterility is recognized by the inability of the plant to produce fertile pollen grains, create anthers, or by other abnormalities of reproductive tissues that result in loss of pollen (Hanson and Bentolila, 2004) and is observed in many plant species (Leon *et al*., 1998). The NC effect on the phenotype can only be recognized in a case of cytoplasm change when the reciprocal crosses are made. The result of hybrid crosses with related species depends on the compatibility between the nuclear donor (transmitted by the pollen of the male parent) and the mitochondria/chloroplast (transmitted by the egg cell of the female parent). Over past years multiple mechanisms were proposed and observations were made to explain cytoplasmic male sterility in plant species. Abnormalities in plant development are believed to be the result of an expression of novel mitochondrial polypeptides, which affect mitochondrial function and pollen development (Chase, 2007), an example being CMS observed in a sunflower. In this instance CMS is caused by a new open reading frame (ORF) that is co-transcribed with the *atpA* gene (Köhler *et al*., 1991). In many cases it was found that the lack of a particular gene being expressed resulted in a deleterious phenotype–as was recognized in maize where two mutations in the the *G* cytoplasm, where the *nad9* and *cox2* mitochondrial genes were found–which altered the cytochorome *c* oxidase activity (Ducos *et al*., 2001). The *atp6* gene played a significant role in contribution to the CMS phenotype and was recognized as one (together with *atp4, cox1* gene) of the major mitochondrial genes which can contribute to this phenotype

(Makaroff *et al*., 1989; Hanson, 1991). The *atp6* gene is a mitochondrial encoded ATP synthase 6 subunit, part of the protein super-complex (Complex V) present in the inner membrane of mitochondria and responsible for the final steps of oxidative phosphorylation. Studies of *atp6* and *atp4* (also called *orf25*) in male-sterile *T. timopheevi,* in comparison to *T. aestivum,* show rearrangements in the upstream region of the *atp6* and *atp4* genes, which may cause the CMS phenotype (Mohr *et al*., 1993). The *atp6* and *cox2* role in the CMS phenotype was studied in *Oryza sativa* L., where rearrangements inside and/or around *atp6* were suspected to be the cause of male sterility in abortive cytoplasm in rice (Hosseini, 2010). A discovery of *atp6* as a possible cause of male sterility was made when it was recognized that *atp6* gene sequence editing restores fertility in male sterile rice (Iwabuchi *et al*., 1993). It was found that the repeat region carries a fertility restorer gene (*Rf-1*) which promotes restoration of male sterility by processing the RNA of an aberrant *atp6* gene in rice (Kazama and Toriyama, 2003) and also explains a possible mechanism of fertility restoration, as a nucleus-guided process. It is important to note that the ATP synthase is essential for normal function of the mitochondrial inner membrane (Paumard *et al*., 2002).

Mitochondria as Organelle

Mitochondria have monophyletic origin and are derived from a bacterial endosymbiotic ancestor, the α-Proteobacteria. Evidence suggests that this event happened only once during the evolution of eukaryotic cells (Gray, 1999). The monophyletic origin of mitochondria would suggest that mitochondria (as organelle and its genome) are highly uniform and conserved among all eukaryotes; this is true, but only as far as general mitochondrial function in eukaryotic organisms. The mitochondrial membrane

structure provides the conditions necessary for ATP production from energy-rich molecules such as lipids, amino acids and carbohydrates (Douce, 1985), and that function is shared among all eukaryotes. Mitochondrial functions are reflected in processes such as development, vigor, reproduction and aging of organisms (Chan, 2006). Detailed characterization of mitochondrial proteins allows proper identification of the process, which can be integrated by mitochondria, expanding the range of already-established functions to areas of metabolism, biosynthesis, oxidative stress regulation, cellular signaling, regulation of transcription, translation and protein complex assembly (reviewed by Millar *et al*., 2005). They are a double-membrane organelle, creating four metabolic compartments first–the outer membrane, second–the volume between the two membranes (inter-membrane space), third–the inner membrane, and fourth-the volume within the inner membrane, known as mitochondrial matrix (Douce, 1985).

Respiratory chain supercomplexes (protein complexes built with multiple proteins) present in plant mitochondria are: complex I (NADH dehydrogenase), II (succinate dehydrogenase), III (cytochrome *c* reductase) and IV (cytochrome *c* oxidase), which are placed in the inner membrane of mitochondria. Additionally, at least five "alternative" pathways were recognized, which also take part in respiratory electron transport (Siedow and Umbach, 1995). Recent studies provided new opportunities to describe a functional role for supercomplexes beyond their primary function, such as dimeric ATP synthase complex, which appears to be responsible for the folding of the inner mitochondrial membrane (Dubkina *et al*., 2006). Mitochondria as organelle constantly change their position within the cell. Studies on meiosis provided evidence that in the cells undergoing cell division, transport processes across the cytoskeleton are

responsible for the proper mitochondrial distribution in daughter cells. Mitochondria can be actively transferred to metabolically more active parts of the cell as, demonstrated in budding yeast (Fehrenbacher *et al*., 2004). Mitochondrial structure is highly dynamic; it continually moves and undergoes fusion and fission processes when structural changes of organelle occur (Okamoto and Shaw, 2005; Logan, 2010). These processes, such as internal changes of the inner membrane organization and general plasticity of mitochondria, have been connected to numerous human disorders and particularly apoptosis (Zick *et al*., 2009). Mitochondria are maternally inherited, with some exceptions of paternal or bi-parental inheritance, depending on species and even particular variety (Weihe *et al*., 2009). In *Triticum,* maternal inheritance is considered the main source of mitochondrial transfer; however, paternal leakage of mitochondria has been observed, where fragments of mitochondrial DNA (mtDNA) were maintained in progeny (Hattori *et al*., 2002).

Mitochondrial Genomics

Mitochondria carry their own genome, which is independent from the nuclear genome. However, during evolution mitochondria adapted themselves to new pre-eukaryotic hosts by transferring genetic material into the nuclear genome (Andersson and Kurland, 1998; Lough *et al*., 2008). Due to that transfer, the majority of mitochondrial encoded genes can be found in the nuclear genome (Adams and Palmer, 2003; Herrmann, 2003). Recent studies of *Mesostigma viride,* considered the earliest ancestor of all green plants, showed that the mitochondrial genome was highly compact at the early stages of green plant evolution. The sequenced genome of the *Mesostigma viride* has only 42,424 bp, with high gene density (86.6% coding sequences), and contains 65 genes. These

genes can also be found in other known mitochondrial genomes (Turmel *et al*., 2002).

This finding provides evidence for the functional conservation of mitochondria in plant

species, and is also a valuable addition to our knowledge of mitochondrial genome

evolution, especially when contrasted with the diversity of mitochondrial genomes found

among animal and plant kingdoms. Plants have the largest known mitochondria, with

genome size ranging from 222 to 773 kbp, looking at the 6 sequenced genomes present in

National Center for Biotechnology Information (2012). The large size of plant

mitochondrial genomes is especially apparent when it is compared to the animal

mitochondria genomes, which range from 16 to 20 kbp (Boore, 1999). Based on 1,000

sequenced animal mtDNAs, we can see a high conservation of gene content in the animal

kingdom with only 13 protein coding genes (Gissi *et al*., 2008). Variation in the genome

size can be observed not only between different species, but within the species. In the

Cucurbitaceae family, the known genome sizes vary from 390 kbp in *Citrullus lanatus* to

2.9 Mbp in *Cucumis melo* (Alverson *et al.,* 2010). The *T. aestivum* mitochondrial genome

encodes only 39 protein coding genes, 34 RNA coding genes, and is 452,528 bp in size as

a circular molecule (Ogihara *et al*., 2005).

As mentioned in the first paragraph, mitochondria carry only a fraction of the

information required to properly build mitochondria as organelle. In *Arabidopsis* it was

predicted, based on proteomic and various machine learning techniques, that

approximately 4,000 proteins can build the mitochondrial structure with its outer and

inner membranes (Heazlewood *et al*., 2004). These proteins are transported from the

nucleus, to the cytosol and finally to the mitochondria. They carry mitochondria targeting

signals (MTSs) located in most cases on the N-terminal protein coding region (Sirrenberg

et al., 1998), but also at the C-terminal end (Heijne *et al*., 1986). This information has been used to identify possible mitochondria-targeted proteins (Schneider and Fecher, 2004).

Another characteristic feature of the mitochondrial genome is that the number of mtDNA copies can differ between organism cell types. On average, in *Arabidopsis thaliana,* there are 59 mtDNA copies per cell; however, generative cells have on average only 0.083 copies of mtDNA (Wang *et al*., 2010). Mitochondria can gain back its genetic material by fusion and fission processes, which occur constantly within cells (Logan, 2010). Thus, it is possible that certain mitochondria will not carry any copy of DNA within itself, but will still provide its functions to the cell.

Despite functional conservation, it is characteristic for the mitochondrial genomes within cells to be highly diverse in structure and nucleotide polymorphism, which is recognized as heteroplasmy. Heteroplasmy is defined as the coexistence of divergent mitochondrial genotypes (mitotypes) in the same cell. This phenomenon has been connected with paternal leakage of mitochondria as observed in *Silene vulgaris* (Bentley *et al*., 2010). It was found that mitotypes can coexist in a single cell (Figure 1), however
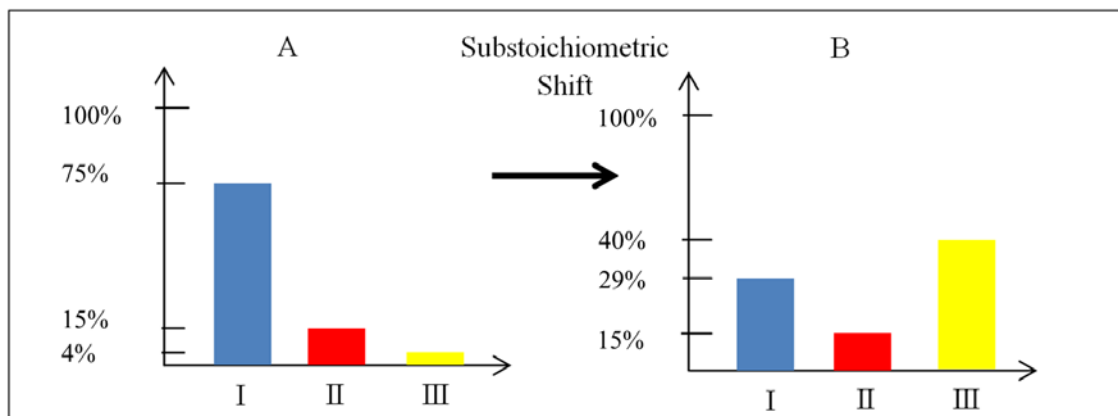
Figure 1. An ideogram representing differences in the mitochondrial composition of two cells. Cells are designated as A and B. Numbers: I, II, III represent different mitotypes present in cell. Change in proportion between mitotypes is recognized as a substoichiometric shift.

the proportion between mitotypes can differ between plants, tissues, and developmental stages (Arrieta-Montiel *et al*., 2001). The coexistence of mitotypes and substoichiometric shift can explain the emergence of CMS as demonstrated in *Brassica napus*. It was found that *pol* male sterile, and *nap* fertile mitotypes coexist in *B. napus* cultivars (Chen *et al*., 2011). Stoichiometric shifting is considered the mechanism for the generation of mitochondrial gene diversity. Studies on the maize *atpA* gene, where four genomic rearrangements were identified (Small *et al*., 1987), provide the background to connect the observed gene diversity and the shift between different mitotypes in the cell. Studies of *Beta vulgaris* where CMS genomes belong to a single sterile lineage showed that CMS lines underwent accelerated rates of nucleotide substitution and genome rearrangement (Darracq *et al*., 2011). Mitochondrial genomes are rich in large repeats (over 1,000 bp), which take part in reciprocal recombination and are responsible for the multipartite organization of the genome (Sugiyama *et al*., 2005). Another common feature, recognized as the main reason for the appearance of new mitotypes, are short repeats, which are very common in mitochondrial genomes (Woloszynska *et al*., 2010). The connection between CMS phenotypes of maize and changes in the genome size was found to be connected with large (0.5-120 kbp) duplications (Allen *et al*., 2007). Mitochondrial genome recombination is controlled by the nucleus. The nuclear gene *msh1* (*mutS* homologue) was found to be a suppressor of recombination at the short repeated sequence regions (size range, from 108 bp to 556 bp) in *Arabidopsis*. By its action, this gene can contribute to mitochondrial evolution, creating lineage-specific patterns of mitochondrial genetic variation in higher plants (Arrieta-Montiel *et al*., 2009). Studies on yeast revealed that the *mgm101* protein is required for mitochondrial repeat-

mediated homologous recombination (Mbantenkhu *et al*., 2011), which additionally confirms the role of nuclear control on mitochondrial recombination.

In the natural population of *Silene vulgaris*, the level of heteroplasmy was quantified using the *atpA* and *cox1* genes. It was concluded that in one population, 26% of the diversity occurred within individuals (Welch *et al*., 2006). Studies on the common bean show that mitotypes are inherited and transmitted from one generation to another (Arrieta-Montiel *et al*., 2001). In conclusion, the heteroplasmy itself is a natural state of mitochondrial genomes, but can be influenced by many factors, such as the alloplasmic condition. Multiple variants of mitochondria can be an evolutionary mechanism of cytoplasm adaptation, since mitochondria behave in a cell as a population.

Mitochondrial Genome Evolution

Mitochondrial genome evolution depends on four different aspects: 1) mutation rate, 2) intra-genomic recombination, 3) transfer of genes between organelle genomes (mainly nuclei and mitochondria), and 4) heteroplasmy.

In general, plant mitochondria accumulate substitutions 3 to 4 times slower than chloroplast and 100 times slower than animal mitochondria (Palmer and Herbon, 1988; Wolfe and Sharp, 1987). Studies on flowering plants demonstrate an exception from the observed slow mutation rate of the plant mitochondrial genome, indicating that the mutation rate is dependent on factors, such as production and detoxification of oxygen free radicals or the efficacy of mtDNA replication and/or repair (Cho *et al*., 2004). The mutation rate of the mitochondrial genome can be variable, as shown in studies of the *Geraniaceae* family where mutation-creation mechanisms in mtDNA were studied (Parkinson *et al*., 2005). Seven mitochondrial genes were compared across taxa of

Pelagronium hortorum, and results indicated exceptionally rapid mitochondrial evolution (Parkinson *et al*., 2005). These multiple mutation rate estimates of mitochondrial DNA give rise to a hypothesis that there can be multiple mechanisms of mutation induction/control.

It is recognized that mitochondria recombine frequently, creating new configurations of the genome and constantly changing structure, which as a consequence is accompanied by gene shuffling. There are multiple factors that can alter homologous recombination such as: cytoplasm substitution (Dieterich *et al*., 2003; Liu *et al*., 2002), tissue culture (Kanazowa *et al*., 1994; Vitart *et al*., 1992), and cellular replication. As mentioned before, homologous recombination is guided by large repeats (larger than 1,000 bp), which are present multiple times in a mitochondrial genome (Sugiyama *et al*., 2005). Recombination can also occur through short repeats (six to several hundred base pairs), which are considered to be "hot spots" for recombination and are present in high numbers in mtDNA (Scotti *et al*., 2004). There is strong connection between the recombination of mitochondrial genomes and the nuclear coded genes which guide those processes. Biochemical activity similar to the bacterial *RecA* gene was observed during the recombination of soybean mitochondria, where recombination structures were observed (using Western blot analysis, 2-D agarose gel electrophoresis), providing the first physical evidence for recombination and connection with nuclear guidance of that process (Manchekar *et al*., 2006).

As mentioned before, the mitochondrial genome has one characteristic feature: transfer of its genes (protein and tRNA coding genes) to the nucleus, which has occurred during the course of evolution (Martin and Herrmann, 1998). Gene content in eukaryotic

mitochondria varies greatly, with the protein coding genes ranging from 3 to 67 and tRNA from 0 to 27 (reviewed by Adams and Palmer, 2003). A highly variable rate and pattern of mitochondrial gene loss and transfer to the nucleus was observed in studies on flowering plants (Adams *et al.*, 2002). This variability of gene content (Gray, 1999), and the single mitochondria origin in all plant species (Turmel *et al.*, 2002), creates a picture where nuclear-mitochondrial gene transfer has to be one of the major evolutionary processes which guides mitochondrial evolution. In addition to the process of gene transfer, differences in genes that are being transferred to the nucleus are observed. Functional gene transfer takes place when the transferred gene can be expressed in the nuclear background. Recent evidence of that process was observed in *Triticum* and *Oryza,* a relatively closely related grass species (Sandoval *et al.*, 2004). Comparison of these species indicated that maize does not have *rps14* and *rpl5* mitochondrial genes, however wheat has the functional *rpl5* gene, but the *rps14* pseudogene is located in the nucleus (Sandoval *et al.*, 2004).

It was demonstrated that heteroplasmy could be of evolutionary importance with relation to the adaptation of species to changing environmental conditions. Studies on insecticide resistance of *Tetranychus urticae* (arthropod pest)–where the resistance is encoded by the mitochondrial genome–provided that evidence. Mitotypes carrying resistance were favored by selection at a frequency of 60% or more (Leeuwen *et al.*, 2008). Heteroplasmy allows species to create various mitotypes and if necessary stoichiometric shifting can occur, creating changes in proportion of a mitotypes (Arrieta-Montiel *et al.*, 2001).

Mitochondria are a flexible organelle defined by constant fusion and fission processes, and by their high mobility within the cell which allows adaptation of the physiological needs of the whole organism. Looking at the mitochondrial genome, we also see a high level of recombination activity, accompanied by gene transfer. Existence of mitotypes creates a population of mitochondrial genomes within a cell, providing the means of adaptation to environmental conditions. This can be contrasted with relatively high conservation of gene content; however, this fact cannot be generalized above kingdom level. The simplest way to describe mitochondria and their function, but also with relation to the evolutionary potential of its genome, would be: adaptable.

OBJECTIVES

In our studies we compared four mitochondrial genomes from the *Triticum* genus: *T. turgidum*, an alloplasmic durum line with a *Triticum longissimum* cytoplasm [hereafter referred to as (lo) durum], *T. longissimum* and *T. tauschii*. Our research focused on two aspects of cytoplasm evolution in the *Triticum* genus. First, a possible connection between phenotypic abnormalities observed in the alloplasmic line and its relation to the cytoplasm (Paper I); and second, evolution of *T. turgidum* mitochondria by comparison to *T. aestivum* and *T. tauschii* mitochondrial genomes (Paper II). In both papers we characterize gene content, looking at major gene changes and polymorphism within gene space and across the genome, as well as compare with sequenced genomes globally, gene space organization and overall synteny. We hypothesize possible causes for phenotypic changes observed in our model and evolution of mitochondria in tetrapolid and hexaplid wheat genomes.

MATERIALS AND METHODS

Plant Material

Four *Triticum* species: *T. turgidum* (line 56-I)*, T. longissimum* (line WCB18)*,* alloplasmic durum line [(lo) durum] and *T. tauschii* (line WCB623) were used for sequencing of the mitochondrial genomes. The (lo) durum wheat is a male sterile alloplasmic line where the cytoplasm is derived from *T. longissimum* and the nucleus is from *T. turgidum*. The scs^{ti} locus on chromosome 1A restores viability in this alloplasmic line and was derived from *T. timophevii* (Maan, 1992a). This line is maintained by crossing with normal durum [(d) – –; 56-1], which does not have a copy of the *scs* gene. The (lo) durum line was chosen to investigate the differences and changes that occur in the mitochondrial genome in the alloplasmic condition. Evolution of the *Triticum* species was studied by sequencing both parents of *T. aestivum*, *T. tauschii*–considered to be the DD paternal nuclear genome donor–and *T. turgidum,* the maternal AABB nuclear genome donor of *T. aestivum* and, as a consequence, the cytoplasm donor for that species (Feldman, 2001).

Mitochondrial DNA Isolation and Amplification

The base for the mitochondria isolation procedure was the protocol developed by Triboush *et al*. in 1998. Plants were grown in the greenhouse in dark conditions for two to three weeks. Trays with soil mix were placed in the greenhouse, in closed cardboard boxes. Only leaf tissue was harvested and collected in aluminum foil on ice. Before the DNA isolation, tissue was rinsed with 4^oC distilled water and transferred to cold room (4^oC). All subsequent steps were performed in the cold room to prevent organelle, and as consequence mtDNA degradation. Tissue homogenization was performed using a hand

blender with STE buffer (400 mM sucrose, 50 mM Tris pH 7.8, 20 mM EDTA-Na$_2$, 0.2% bovine serum albumin, 0.2% β-mercaptoethanol), as described in the original procedure, until the majority of green tissue was separated from cellulose fibers. Homogenized tissue was filtered using nylon woven mesh with 50 microns opening (Nitex). Tubes containing homogenized tissue were transferred into ice and centrifuged in a Sorval RC6 Plus, Thermo Scientific, 3,000 g – 10' (cell/tissue debris pelleting), 16,000 g – 15' (chloroplast pelleting) and 26,000 g – 45' (mitochondria pelleting). Pelleted mitochondria from the last centrifugation were treated by DNase to digest nuclear and chloroplast DNA (Triboush *et al*., 1998). Mitochondrial DNA was isolated immediately after organelle purification using a Mammalian Genomic DNA Miniprep Kit, Sigma-Aldrich Co, St. Louis, USA. We used the "Cultured Cell Preparation" procedure supplied with the kit, starting from 200 μl of RS solution. Quantitative real time PCR, using 7900HT Fast Real-Time PCR System, Applied Biosystems™, was performed to test for the proportion of mitochondrial, chloroplast and nuclear DNA. Three gene-specific primers with TaKaRa SYBR® Premix Ex Taq™ II (Perfect Real Time), Takara Bio, Madison, USA were used for the PCR reaction. Gene-specific primers for chloroplast (gene *psb60*), mitochondrial (gene *nad3*) and nuclear genomes (retrojunction nuclear marker) are listed in Table C1. No amplification of nuclear DNA was reported for any sample in mitochondrial DNA preparation (nuclear DNA was used as positive control from total DNA purification). For sequencing the whole genome, mitochondrial DNA was amplified following the manufacturer's protocol for the Illustra™ GenomiPhi™ V2 DNA Amplification Kit, GE Healthcare, Fairfield, USA. After amplification, aliquots of DNA were separated on 1% agarose gels to confirm the amplification (Figure D1).

Amplified DNA was acetone precipitated (Sambrook *et al*., 1989) and quantified using the Nanodrop spectrophotometer, Thermo Fisher Scientific Inc, Waltham, USA. Approximately 20 ug of each mtDNA sample, dissolved in water was delivered for sequencing.

Mitochondrial Genome Sequencing

Shot gun, one-end sequencing by 454 Pyrosequencing^{TM} was performed in the laboratory of Dr. Yong Q. Gu, by Dr. Naxin Hao, USDA-ARS, Albany, CA, USA. An additional mtDNA concentration check was done using a Quant-iT™ PicoGreen® dsDNA Assay Kit, Invitrogen™, USA. The mitochondrial DNA was fragmented by nebulization and fragments from 500 to 700 bp were selected for library preparation (Figure D2). Sequencing was done using a 454 GS FLX Titanium sequencer, with GS Titanium General Library Prep Kit, GS Titanium SV emPCR kit, GS Titanium LV emPCR kit, GS Titanium sequence kit XLR70 and GS Titanium PicoTiterPlate kit 70x75 according to manufacturer protocols.

Mitochondrial Genome Assembly

Genome assembly was completed using the software MIRA v. 3.xx (Chevreux *et al*., 2004) on a PC/UNIX platform (for its full description, please see Appendix B). Mapping assembly for each genome was performed with the *T. aestivum* (Gene Bank, NC_007579.1) mitochondrial genome as a backbone. As negative assembly control we used *Tripsacum dactyloides* (Gene Bank, DQ984517). The *de novo* assemblies were also performed for each *Triticum* species. The final assembly of the *T. tauschii* mitochondrial genome was used as a backbone sequence for the *T. longissimum* genome assembly. To assemble the (lo) durum genome, we used *de novo* and mapping assembly at the same

time, with the *T. aestivum* genome as a backbone sequence, to partially guide the assembly within common regions of both genomes. The GAP4 software (Bonfield *et al*., 1995) was used to manually edit/join contigs obtained directly from assembly software to create final consensus sequences and also to visually evaluate assembly. Joining was performed only for contigs with an overlap greater than 100 bp and less than about 1% dissimilarity in overlapping region. Artemis: DNA Sequence Viewer and Annotation Tool software was used to create pictures of the mitochondrial genome assembly (Carver *et al*., 2008). Visualization of the mapping genome assembly was done using Tablet software (Milne *et al*., 2010). Pictures were processed using IrfanView v.4.x (author: Irfan Skiljan).

To perform *de novo* genome assembly with one end sequencing data we relied only on the nucleotide sequence of the read. Current software was not able to perform assembly of the mitochondrial genome (completely) with one end data without extensive manual trimming, joining and large contig editing. Additionally, due to the presence of long and short repeated sequences present multiple times in mitochondrial genomes and heteroplasmy, single contigs for each genome could not be created only by algorithms implemented in the software alone. To overcome these difficulties, genome assembly was performed by extension of each contig separately (one by one, large contig against all read data set) to create overlaps and to find sequence-spanning contigs (Figure 2).

Gene Finding, ORF Prediction, and Genome Annotation

Gene sequences and names were annotated based on the *T. aestivum* mitochondrial genome (Ogihara *et al*., 2005). The NCBI database was searched for plant

19
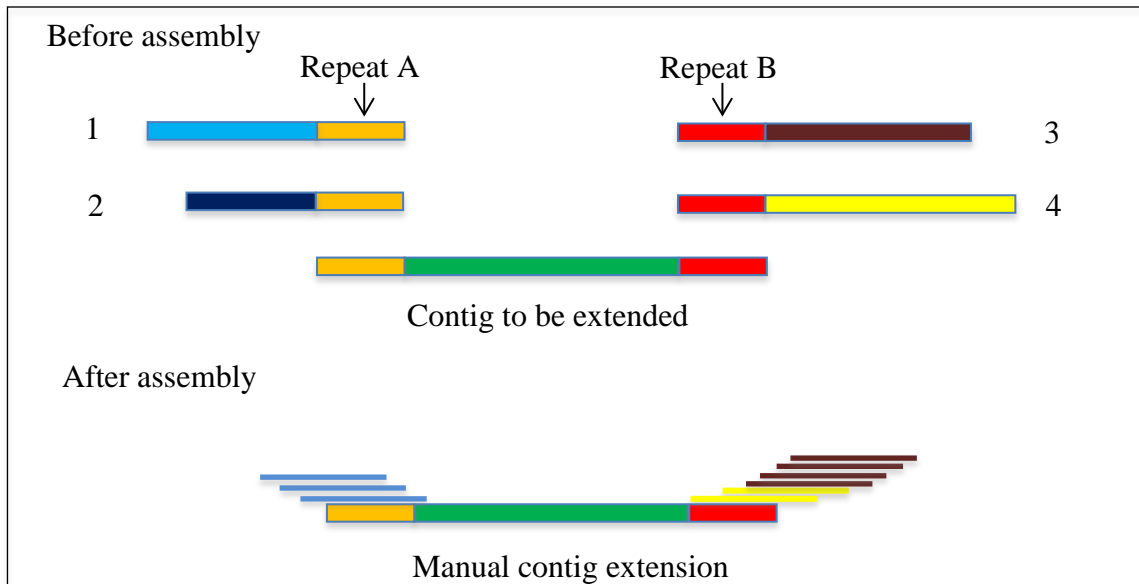
Figure 2. Ideogram of manual assembly process in case of long repeats. Multiple repeats in the mitochondrial genome hindered the assembly process and extension of contigs. Figure demonstrates manual extension process of contigs which are ended by repeated sequences (marked orange and red). Manual selection of one of the variants, in this example there are four possible contigs (numbers 1, 2, 3, 4), which can be joined in repeated region and sequence can be further extended. Subsequently, genome assembly was performed, where contigs from *de novo* assembly were used to perform an additional round of mapping assembly.

mitochondrial genome sequences. An open reading frame search was performed using NCBI ORF Finder, and a standard translation table. The ORFs were named starting from 0, as the first one identified in submitted sequence, with subsequently increasing numbers. Only 300 bp or larger ORFs are reported. The ORF comparison between the three mitochondrial genomes was performed and unique ORFs were reported for each species with comparison to the other two. We used the BLAST algorithm to compare the ORFs. Only ORFs which were 95 % or less of the query size were considered as unique and these ORFs are reported and described.

Mitochondrial Genome Polymorphism

Deep sequencing, coverage above 10x, allowed us to perform polymorphism analysis within each mitochondrial genome. Assembly provided us with consensus sequences for each genome. Those consensus sequences were used further as backbone sequences, where sequencing reads from the same species were mapped against them. We used raw assembly files to search for polymorphism for each sequenced position. Raw assembly files were obtained from GAP4 software. The SNP was reported only when the number of different nucleotides, in comparison to other nucleotides at that position, was equal to or greater than 10% of position coverage. SNP search between mitochondrial genomes was performed. Even though there were a high number of rearrangements, many regions were highly similar in sequence. Because of the high number of rearrangements between genomes, we used BLAST to find positions of similar genomic sequences, with a cutoff of more than 95% identity. A particular region was searched for SNP differences using an algorithm developed for that search: each position of the nucleotide sequence was compared to find nucleotide differences reported. An error rate for a particular SNP search was established comparing 17,046 bp of gene space (which are: *atp1, atp9, rps13, cox3, matR, rps1, ccmFN, cob, rpl16, rps3, ccmFC, orf359, nad4, orf240, ccmB, atp4, nad2, orf173, nad9 orf349, rps2*) between sequenced species and the *T. aestivum* reference, using *de novo* assembly as the quality check.

Confirmation of Genome and Gene Structure

A new allelic form of the *atp6* gene was found in the assembly of (lo) durum, *T. longissimum* and *T. tauschii* when compared to the *T. turgidum.* Presence/absence and sequence of *atp*6 was confirmed by PCR using atp6LO, atp6TU set of primers (Table

C1), which were designed to differentiate between two alleles of that gene. These markers have common forward primer sequence placed in the core region of the gene. Reverse primer was designed from a non-conserved region of the *atp6* gene. This design allowed us proper differentiation between the two gene alleles. Also, PCR was performed to confirm synteny and specific breakages between *T. turgidum* and (lo) durum and to confirm the proper contig structure. PCR primer design was made using OLIGO Primer Analysis Software v.7.x (Rychlik, 2007), Primer3 v. 0.4.0, and Primer-BLAST (Rozen and Skaletsky, 2000).

Final genome assembly was used to explain overall structural differences between genomes. Assignment of gene positions into each genome allowed us to create a gene synteny map showing only the gene relationship between mitochondrial genomes in sequenced species. Double ACT v2, Health Protection Agency, together with ACT: Artemis Comparison Tool, Sanger Institute, was used to demonstrate syntenic differences and gene positions across sequenced mitochondrial genomes.

PAPER I. ACCELERATED MITOCHONDRIAL GENOME EVOLUTION OF TRITICUM ALLOPLASMIC LINE

Abstract

To investigate the level of changes in the mitochondrial genome as a consequence of the alloplasmic condition, three mitochondrial genomes of the *Triticum-Aegilops* species were sequenced: 1) a durum alloplasmic line with *T. longissimum* cytoplasm designated as (lo) durum, 2) *T. turgidum*, durum wheat and 3) *T. longissimum.*

The mitochondrial genomes obtained for *T. turgidum* was 451,925 bp and the sequence was similar to the previously sequenced *T. aestivum* cv. 'Chinese Spring' with few alterations. The estimated sizes of the (lo) durum and *T. longissimum* genomes were about 432,606 bp and 399,074 bp, respectively. A high coverage of 61-133x for all three genomes enabled observation of heteroplasmy within each mitochondrial genome. Genome rearrangements and structural changes were also observed through *de novo* assembly of the mtDNAs studied. The mitochondrial genome structure in the alloplasmic line was not only distant from *T. turgidum,* but also different from its maternal parent *T. longissimum*. The alloplasmic durum and *T. longissimum* carry the same versions of *atp6*, *nad6*, *rps19*-p, *cob and cox2-2,* which are different from *T. turgidum*. Evidence of paternal leakage was also observed by analyzing *nad9* and o*rf359* among all three lines. An open reading frame search found 184, 191, and 163 ORFs in *T.turgidum*, (lo) durum and *T. longissimum,* respectively. 27 were specific to the (lo) durum line.

Several heteroplasmic regions were observed within genes and also intergenic regions in all the three genomes. The amount of rearrangements and nucleotide changes

in mtDNA of the alloplasmic line that occurred in less than half a century was surprising in contrast to the changes found between *T. turgidum* and *T. aestivum* genomes as a result of 10,000 years of evolution. The changes in genes from paternal leakage were high, emphasizing other active mechanisms of heteroplasmy such as recombination and mutation. In the newly-formed ORFs, differences in gene sequences and copy numbers, heteroplasmy, and substoichiometric changes show the potential of alloplasmic condition to accelerate the evolution toward forming new mitochondrial genomes.

Introduction

The primary cause of change and the most prevalent phenotype connected with nuclear-cytoplasmic incompatibility is cytoplasmic male sterility (CMS), recognized in more than 150 plant species (Leon *et al*., 1998). In the case of the alloplasmic (lo) durum line (normal *Triticum turgidum* nucleus and alien *T. longissimum* [designated as (lo)] cytoplasm), presence of the *scs* gene (species cytoplasm specific), and *Vi* (vitality) gene pairs are required to restore proper communication between the alien cytoplasm and the nucleus (Maan, 1992b). The *scs^{ti}* gene is a nuclear gene derived from *T. timopheevii* (Maan, 1992a). Incompatibility between the cytoplasm and nucleus observed in this model results in seed and plant developmental abnormalities where three major phenotypic changes are observed: CMS, weak vs. normal plant development and shriveled vs. full plump seeds (Figure 3). Cytoplasmic male sterility was investigated primarily for use in hybrid wheat production. A sterility-fertility restoration system was developed using *T. timopheevi* cytoplasm and *Rf* genes (Maan *et al*., 1984). A class of genes called *species cytoplasm specific* gene, present on group 1 chromosomes of the *Triticum* species control developmental and sexual reproductive traits in alloplasmic lines
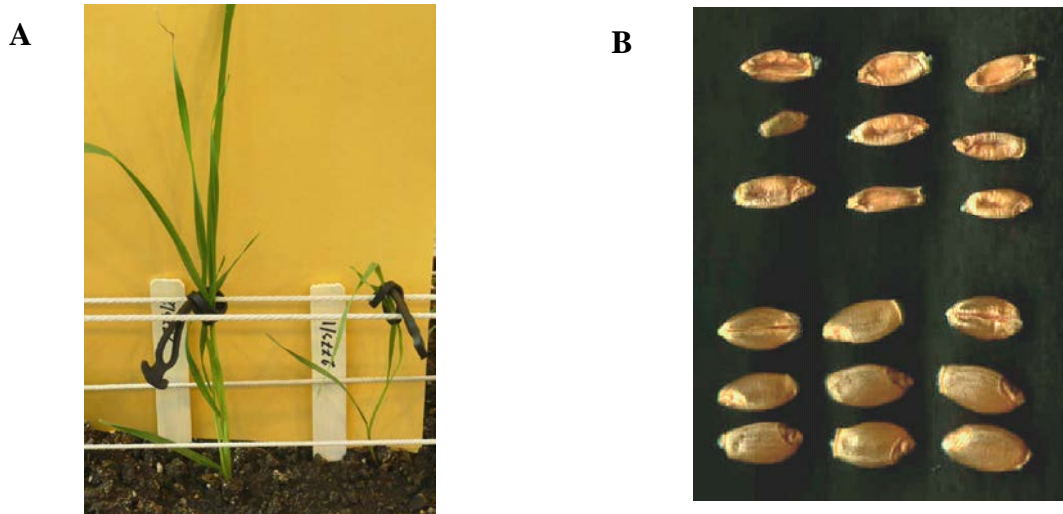
Figure 3. Comparison of the NC incompatibility and action of the scs^{ti} gene. **(A)** The photograph shows [(lo) scs^d scs^{ti}] vigorous plants (on the left) and [(lo) scs^d _] weak plant (on the right). **(B)** Shriveled, inviable seeds (top) without *scs* show an incompatibility reaction between the nucleus and cytoplasm. Plump, viable seeds with *scs* (bottom) indicate compatibility between the nucleus and cytoplasm.

(Maan, 1991). Nuclear *scs* genes can be identified by their interactions in other species of interest. The transfer of the nucleus from *T. turgidum* to the *T. longissimum* cytoplasm was initially unsuccessful. The alloplasmic 'Selkirk' hexaploid line was used in a cross euplasmic durum wheat as male, which carried 1D or telomeric 1DL chromosome. The 1DL chromosome, carrying scs^{ae}, was removed by crossing to *T. timopheevi*, and repeated backcrossing to *T. turgidum* plants. This method produced (lo) durum, the alloplasmic line that carries scs^{ti} on 1A, and can be backcrossed to normal durum, resulting in a 1:1 ratio of plump, viable seeds and shriveled, inviable seeds (Kalavacharla *et al.*, 2005).

It is known that common wheat, *T. aestivum* L. (2n = 6x = 42), is a hexaploid with an AABBDD genome (Feldman, 2001) and is much more flexible in adaptation to new environments than durum wheat; *T. turgidum* L. var. *durum* (2n = 4x = 28) is a tetraploid with genome AABB (Feldman, 2001). The polyploid condition of wheat is one of the

major reasons for its wide adaptation around the globe (Dubcovsky and Dvorak, 2007). However, due to differences in the ploidy level, durum wheat is more sensitive to cytoplasmic substitution than common wheat and is partially or completely incompatible with *T. longissimum* (S. & M.) Bowden, [S^1S^1 $2n = 2x = 14$] (Feldman, 2001).

Mitochondria and their functions were connected with NC compatibility in wheat species (Wu *et al*., 1998). Multiple studies have identified a connection between mitochondrial encoded genes and the CMS phenotype (Köhler *et al*., 1991, Mohr *et al*., 1993, Hosseini, 2010). At present there are three mitochondrial genomic sequences available in *Triticum* species: one from *T. aestivum* cv. 'Chinese Spring' (Ogihara *et al*., 2005), *T. aestivum* cv. 'Chinese Yumai' (Cui *et al*., 2009)–both with a size of 452,528 bp–and a third genome from male sterile *Triticum* line with K-type cytoplasm from *Aegilops kotschyi* and genome size 647,559 bp (Liu *et al*., 2011). These studies provide valuable background information on mitochondrial genome structure and diversity, which can serve for further functional studies of observed phenotypic changes (Soltani *et al*., 2011).

Results

Genome assembly

The read lengths were as expected for this sequencing technology (Figure D3), with the average sequencing read length of 449 bp, indicating a good quality of the data obtained. The purification method used (Triboush *et al*., 1998) provided mitochondrial DNA free from nuclear DNA contamination. Based on a sequencing read count 3 to 9% of overall reads were chloroplast DNA contaminations.

When assembling the mitochondrial genomes, a majority of the sequencing reads were used (66-79% of total reads) and assembled to create the mitochondrial genome sequences. In case of *T. longissimum,* only 32% of the reads were used, since bacterial contamination was recognized. The remaining 12-28% of reads were not assembled and not included in the final assembly, but they created additional contigs. Using BLAST searches against the assembled mitochondrial genomes indicated that those additional contigs are valid mitochondrial sequences.

The size of the *T. turgidum* mitochondrial genome was 451,925 bp and the sequence was almost identical to that previously described for *T. aestivum* cv. 'Chinese Spring' (Ogihara *et al.*, 2005). The estimated size of the (lo) durum and *T. longissimum* genomes are about 432,606 bp and 399,074 bp respectively. Overall coverage of the sequenced genomes is 61-133x. This high-quality data is suitable for full description of known mitochondrial gene space, open reading frames search and comparison, and single nucleotide polymorphisms found within and across the sequenced genomes.

Major gene differences (*atp6*, *rps19*-p, *cob*, *nad9*, *nad6*)

We used the *T. aestivum* mitochondrial genome annotated in Ogihara *et al.*, 2005 as a reference to describe our results, and found five genes with major nucleotide changes in (lo) durum and *T. longissimum*: *atp6*, *rps19*-p, *cob, nad9,* and *nad6* genes. A major difference was recognized in the *atp6* gene between *T. turgidum* (1,161 bp) and (lo) durum (predicted gene size 1,248 bp; Figure 4). The mitochondrial *atp6* gene encodes an ATP synthase F0 subunit 6. The *apt6* gene found in *T. longissimum* and (lo) durum (designation: *atp6*-L) has an altered pre-sequence in comparison to *T. turgidum* (designation: *apt6*-T). However, that difference is not reflected in the surrounding region,
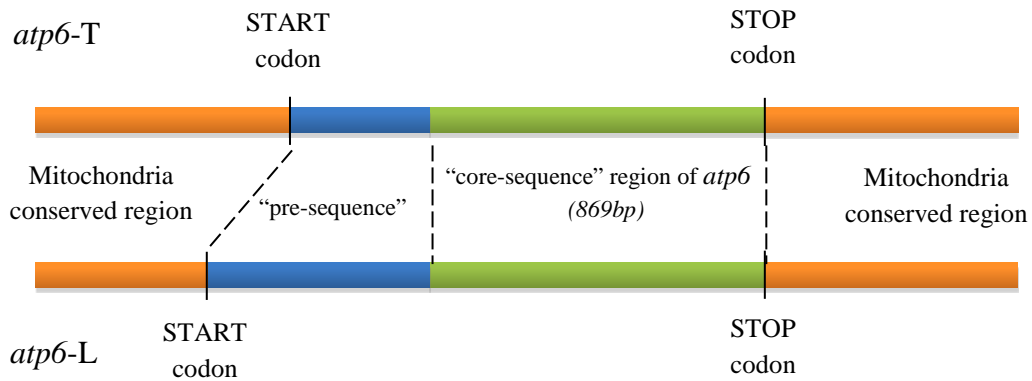
Figure 4. Graphical representation of the *atp6* gene structure and its comparison between sequenced species. Orange bars symbolize mitochondrial conserved regions; blue bars represent a pre-sequence portion of the *atp6* gene. The new allele (*atp6*-L) of that gene has a longer pre-sequence (found in (lo) durum and *T. longissimum*) than the *atp6* gene from *T. turgidum* (*atp6*-T). The core sequence is conserved, except for 6 SNPs.

which is conserved among all three species. In addition, they share a highly similar (4 SNPs and 2 dinucleotide changes) core region of 869 bp (sequence details in Figure D4). A BLAST search of the NCBI nucleotide sequence showed that the pre-sequence of *atp6* from *T. longissimum* and (lo) durum was not present in any previously sequenced mitochondrial genome, with the exception of the newly sequenced mitochondria of a male sterile *T. aestivum* line (Liu *et al*., 2011), where a highly similar sequence to *atp6*-L was found (Figure D5). In addition to the characteristic pre-sequence for the female parent, (lo) durum has a number of SNPs in the core region when compared to the *T. turgidum* gene. These changes are C/A^{520}, (A/T^{543}, T/A^{544}), (A/T^{580}, T/A^{581}), A/C^{595}, T/G^{639}, including SNP T/G^{560} were characteristic only for (lo) durum (Figure D4). We confirmed the allelic differences in the *atp6* gene by amplification of the *atp6*-L and *atp6*-T gene alleles from the mitochondria and total DNA of each species (Figure 5). No *atp6*-T amplification product was found in *T. longissimum* and (lo) durum mitochondrial

28
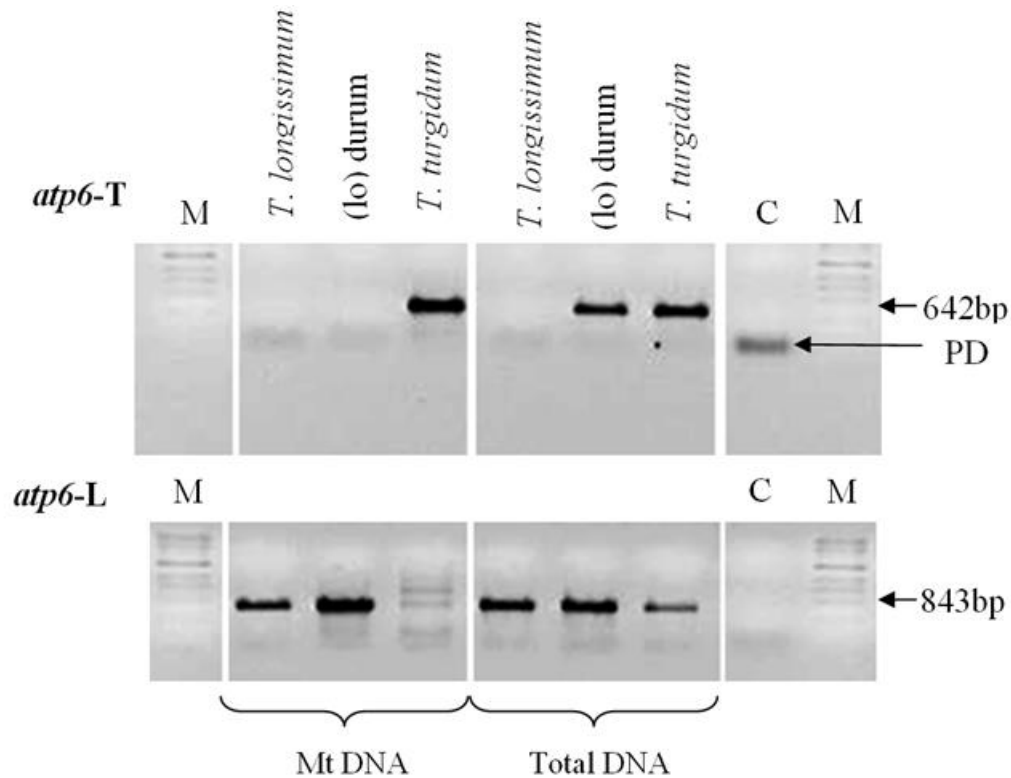
Figure 5. Differentiation of the *atp6* alleles using polymerase chain reaction (PCR). Amplification used two sets of forward primers to differentiate between alleles (Table C1). *Atp6*-T amplification of *atp6* allele from *T. turgidum* and *atp6*-L amplification of *atp6* from *T. longissimum* and (lo) durum line. DNA was separated on a 1% agarose gel. M–indicates bp size marker, C–indicates control PCR, PD–primers dimers

DNA, but it was present in the *T. turgidum* mitochondrial DNA sample. The PCR results confirmed data obtained from assembly. However when PCR was performed based on the total DNA, no product was observed in the case of *T. longissimum,* but for (lo) durum, and *T. turgidum* the proper allele was amplified. Amplification of the *atp6*-L allele (mitochondria DNA only) confirmed its presence in *T. longissimum* and (lo) durum. The *T. turgidum* amplification results were inconclusive. The *atp6*-L allele is present in all three species when total DNA was used in amplification, indicating a possible nuclear genome copy. Another major gene difference was observed in *nad9*

(gene length 864 bp). A four-nucleotide deletion (TGTG[134-137]) was present in *nad9* in *T. longissimum.* The alloplasmic durum line doesn't share this deletion. However, a characteristic dinucleotide difference, TG/CA[134-135] in relation to *T. turgidum,* was observed in (lo) durum (Figure 6). The four-nucleotide deletion found at position 135 creates a frame shift in the *T. longissimum* gene and generates a STOP codon at position 158 (TAG[158-160]). This creates a major alteration in protein length compared to *T. turgidum* (Figure 6). Beside those changes, three additional SNPs were observed in this region before the premature STOP codon, but no other polymorphisms were found across the rest of the gene. There is one SNP characteristic only for (lo) durum, A/C[118], and two SNPs, C/A[122] and A/T[125], found in both *T. longissimum* and (lo) durum. Each of the SNPs creates an amino-acid change N/H[40], S/Y[41], K/I[42], respectively; and the dinucleotide change results in a V/A[45] substitution in the *nad9* protein. For the full gene sequence refer to Figure D6 in Appendix D.

```
                       111                                       150
                       |--------|---------|---------|---------|
    nad9 T.turgidum    GATACATAACTCTAAAAAGTGTGTGTGCAGTTTTGGATCT
    nad9 (lo) durum    .......C...A..T.........CA...............
    nad9 T. longissimum .........A..T.........----..............

                       151                                       190
                       |--------|---------|---------|---------|
    nad9 T. turgidum   TTATTGGTAGCCAGTCTTTCACTTCTGCCTCTCCACTCCC
    nad9 (lo) durum    .......................................
    nad9 T.longissimum .......___.............................
```

STOP (only *T. longissimum*)
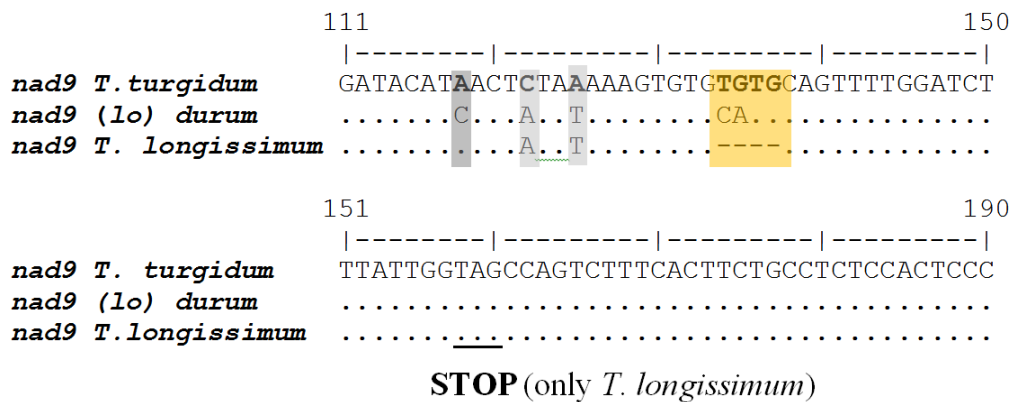
Figure 6. Multi-alignment results for *nad9* gene, STOP codon region. Above picture only gives the sequence from 111 bp to 190 bp. Three SNPs were recognized (light gray boxes) in comparison to *T. turgidum*, one of them (dark gray box) found only in (lo) durum line. Four-nucleotide deletion in *T. longissimum* creates a STOP codon; dinucleotide change in (lo) durum was also recognized (orange box).

The *nad6* (744 bp) mitochondrial gene encodes subunit 6 of Complex I, the NADH-ubiquinone oxidoreductase, which is built from at least 30 different subunits, that are mainly encoded by the nucleus. Multi-alignment of the *nad6* gene alleles (Figure 7) shows that a truncated version of *nad6* is present in the *T. longissimum* and (lo) durum lines, but not in *T. turgidum*. The conserved region begins from the ATG start codon and continues until nucleotide 703. No polymorphisms were found between the alloplasmic line and its parents within the conserved region of the gene. The full sequence of the gene is provided in Appendix D Figure D7.

A nine-nucleotide deletion (AAAGGTTGG[122-130]) in *rps19*-p pseudo gene was recognized in the *T. turgidum* mitochondrial sample in relation to (lo) durum and *T. longissimum* without allelic difference between the two species (Figure 8).

We identified two genes, *cob* and *cox2* (exon 2), from *T. turgidum* that, when compared with (lo) durum and *T. longissimum*, were shorter by 10 nucleotides at the 3'-end. No nucleotide polymorphism between the alloplasmic line and *T. longissimum* were found within region common for all species.

```
                     ....|....| ....|....| ....|....| ....|....|
                     681        690        700        710        720
nad6 T. turgidum     CTTCTCCTCC GGAGCGGGGG GACCG---CC TGACAATTAC
nad6 (lo) durum      CTTCTCCTCC AAAGCGGAGG GAGGGGAATC TCAGGATTCC
nad6 T. longissimum  CTTCTCCTCC AAAGCGGAGG GAGGGGAATC TCAGGATTCC

                                                Highly polymorphic region
```
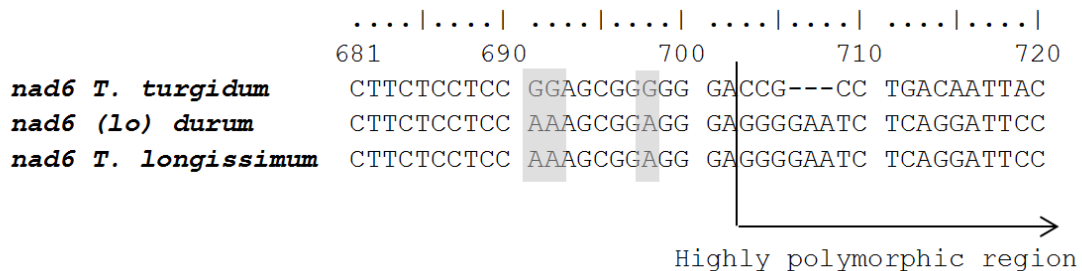
Figure 7. Sequence alignment showing differences in the *nad6* gene as found in *T. turgidum*, (lo) durum and *T. longissimum*. In the above picture only sequences from 681 bp to 730 bp are shown. Overall three SNP's were recognized and two dinucleotide changes (light gray boxes) in comparison to *T. turgidum*. From position 703 bp there is the start of a highly polymorphic region for the gene.

31

```
       ....|....| ....|....| ....|....| ....|....|
                110        120        130        140
rps19-p T. turgidum     TTCGTTGTAA GATCACTGAA GG-------- -TCATAAATT
rps19-p (lo) durum      TTCGTTGTAA GATCACTGAA GGAAAGGTTG GTCATAAATT
rps19-p T. longissimum  TTCGTTGTAA GATCACTGAA GGAAAGGTTG GTCATAAATT
```

Figure 8. The result of *rps19*-p sequence multi-alignment. In the above picture only sequences from 100 bp to 140 bp are shown. It was found that *T. turgidum* has 9 bp deletion in relation to the (lo) durum and *T. longissimum* alleles, where both of these species share the same allele. For full gene alignment see Appendix D, Figure D8.

Open reading frame search and comparison

Previously, several ORFs (ORF designations: 359, 240, 173, and 349) have been identified in *T. aestivum.* The sequenced species were checked for presence and potential polymorphism in these ORFs. ORFs 173 and 240 did not show any polymorphism among sequenced species. We found that *orf349* in the alloplasmic line and *T. longissimum* carried a dinucleotide variation (TT/AG[68]) when compared to *T. turgidum. Orf359* was present in *T. turgidum* at 100% similarity and coverage, but was absent in *T. longissimum*. Only a portion of that gene, from 73 bp to 1,080 bp was found in (lo) durum. This ORF is significantly different in the alloplasmic line; where we found 48 SNPs, three dinucleotide changes, one trinucleotide change, and one three-nucleotide insertion in relation to *T. turgidum* copy of the gene. This appears to be the most polymorphic region found in sequenced genomes when compared to *T. turgidum* (Figure D9).

The ORF search was performed using the NCBI ORF finder. In *T.turgidum*, (lo) durum and *T. longissimum,* respectively*,* we identified 184 ORFs (total of 81,978 bp), 191 ORFs (total of 90,237 bp), 163 ORFs (total of 78,651 bp) that were larger than 300 bp. These ORFs cover 18.1%, 20.8% and 19.7% of the total genome size of each species, respectively. The alloplasmic line has the largest number of ORFs compared to its parents

32

and the highest percentage of ORF per genome size. To determine if there are species-specific ORFs, a comparison of the ORFs was performed. There were 27 ORFs in (lo) durum which are not present in *T. turgidum* and *T. longissimum*, 15 ORFs in *T. longissimum* not present in both *T. turgidum,* and (lo) durum, and 26 ORFs in *T. turgidum* not present in *T. longissimum* and (lo) durum.

The ORFs found to be unique to the alloplasmic line were further described (Table 1). Based on Mega BLAST search against the NCBI database the ORFs 0, 1, 3, 50, 61, 74, 97, 108, 159, 177 were not present, possibly making them characteristic for this line only.

Table 1. Mega BLAST results of unique ORFs for the alloplasmic line. Comparison of ORF recognized in alloplasmic durum line (lo) durum (column ORF#) against the alloplasmic line of *T. aestivum* (Liu *et al*. 2011) and normal *T. aestivum* (Ogihara *et al*., 2005).

| ORF# | Alloplasmic, CMS *T. aestivum* line (GU985444.1) | | Normal *T. aestivum* (EU534409.) | |
	Coverage (%)	Identity (%)	Coverage (%)	Identity (%)
14	14	100	14	97
26*	100	97	100	97
63	100	100	-	-
65	100	99	-	-
84	-	-	31	100
112	100	99	-	-
113[&]	85	99	91	99
114	100	100	10	98
122*	100	97	100	98
138	22	100	-	-
146	100	98	66	98

* - ORFs with 100% of coverage between all three species, & - ORF characteristic only for (lo) durum line

The BLAST search results showed that ORFs 63, 65, and 112 are only present in the male sterile, alloplasmic line of *T. aestivum* and are highly similar (~ 99% similarity) to the ORFs present in the alloplasmic durum line. Complete *orf146* is present in

alloplasmic *T. aestivum* (Liu *et al*., 2011)*,* but only a portion of this ORF exists in a normal common wheat line (Ogihara *et al*., 2005). Additionally, in male fertile and male sterile lines of *T. aestivum, orf14* and *orf113*, are different from that found in alloplasmic durum. Based on its composition, the (lo) durum *orf113* (396 bp) is composed of fragments from other known functional mitochondrial genes *rps2*, *cox1*, *nad4-2* and *rps19*-p genes (Figure 9).
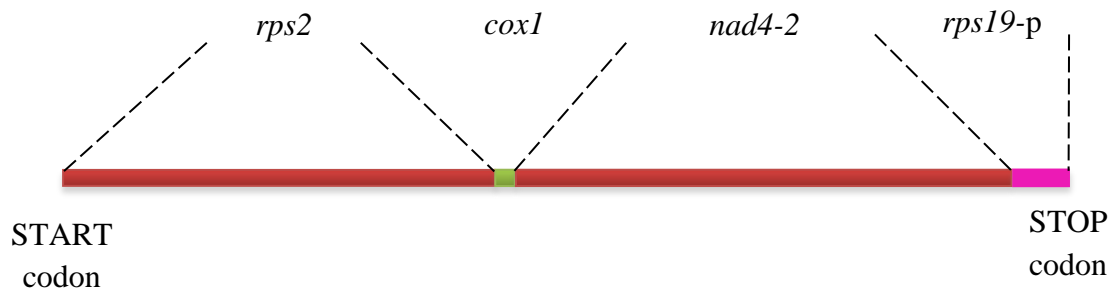
Figure 9. Structure of *orf113* characteristic for the alloplasmic durum line. Database search showed similarity to *rps2*, *cox1*, *nad4-2*, *rps19*-p mitochondrial genes. Shown are the relative fragment size and location of each gene found in the ORF.

Polymorphism within genes, genomes and between genomes

The 39 protein coding genes, three rRNA and 17 tRNA genes characterized in the *T. aestivum* mitochondrial genome were identified in the three sequenced species. Besides previously described major gene differences, the majority of the genes are highly conserved between mitochondria in the *Triticum* species (Table 2).

We report here single nucleotide variations for a total of 11 loci: 5 ribosomal proteins, 4 electron transport coding genes, one transfer RNA and one other protein (Table C4 and C5). Those present in *T. longissimum* and the alloplasmic (lo) durum line are altered relative to *T. turgidum* and represent variations in each species' mitochondrial genome. Multiple variants of a given locus within a species were observed for genes *nad3*, *atp8*, and *rps1* in *T. turgidum.* These may not represent copy number variations in

the master genome assembly, but most likely are the result of heteroplasmy within this species. Here we report only the most abundant allele(s) of particular genes. The total number of nucleotide variations observed in the alloplasmic (lo) durum line was higher than any of the other species examined. When nucleotide variations from multiple alleles present in *T. turgidum* were omitted, we identified in *T. longissimum* and the alloplasmic lines 10 SNPs and 14 SNPs respectively. Six SNPs–*cox3*-exon 2 (GA/TC^{687}), *mttB* (T/G^{41}), *rps2* (A/T^{233}), *rps4* (C/A^{495}), and *rps13* (A/C^{170})–were identified as unique to the alloplasmic line, and differentiate it from both parents.

Table 2. Groups of genes present in mitochondrial genome without polymorphism. Species: *T. turgidum*, (lo) durum and *T. longissimum*.

Complex I
nad1-1, nad1-2, nad1-3, nad1-4, nad1-5
nad2-1, nad2-2, nad2-3, nad2-4, nad2-5
nad4-1, nad4-2, nad4-3, nad4-4
nad4L
nad5-1, nad5-2, nad5-3, nad5-4
nad7-1, nad7-2, nad7-3, nad7-4, nad7-5
Complex IV
cox1, cox2-2
Complex V
apt4, atp9
Complex c biogenesis
ccmB, ccmFC
Other proteins
matR
Ribosomal proteins
rps3-1, rps3-2, rps3-3, rps3-4
rps7, rps12, rpl16
tRNA
Asn tRNA, Asp tRNA, Cys tRNA, Gln-1 tRNA, Gln-II tRNA, Gln-III tRNA*, Glu tRNA, Lys-1 tRNA, Lys-2 tRNA, Lys-3 tRNA*, Met-1 tRNA, Met-2 tRNA, Met -3 tRNA, Met- IV tRNA, Phe tRNA, Pro -1 tRNA, Pro - 2 tRNA, Ser -1 tRNA, Ser -2 tRNA, Ser -3 tRNA, Trp tRNA, Tyr tRNA
rRNA
rrn5-1, rrn5-2$^{\#}$, rrn5-3$^{\#}$, rrn18-1, rrn18-2$^{\#}$, rrn18-3$^{\#}$, rrn26-1, rrn26-2, rrn26-3$^{\$}$

* Copies of genes not present in final the genome assembly, but present in multiple copies in raw assembly data
$^{\#}$ Copies of genes not present in final assembly in *T. longissimum*
$^{\$}$ Only part of the gene was found in final the assembly of (lo) durum. Arabic and roman numerals designate alleles of the same genes

Our data shows that the ribosomal coding genes are the most variable when compared to other mitochondrial protein coding genes. Three out of four SNPs observed between (lo) durum and *T. longissimum* were found in the ribosomal protein coding genes *rps2*, *rps4* and *rps13* (Table C4).

We were able to identify the nucleotide, and the associated amino-acid polymorphisms, that are characteristic for each studied species. In *T. longissimum* the *cox3* gene has an amino acid change Q/H at position 244. In *T. turgidum* we identified changes in the *rps1* gene not present in the alloplasmic line and the cytoplasm donor. Alloplasmic durum has characteristic amino acid changes in *mttB*, *rps13*, and *rps2* (Table 3).

Table 3. Summary of amino acid changes for the sequenced species.

Gene	*T. turgidum*	*T. longissimum*	(lo) durum
cox3*	$Q^{229} + M^{230}$	$H^{229} + L^{330}$	$Q^{229} + M^{230}$
rps1[#]	T^{133}	I^{133}	I^{133}
mttB[$]	L^{16}	L^{16}	W^{16}
rps13[$]	E^{57}	E^{57}	A^{57}
rps2[$]	Y^{78}	Y^{78}	F^{78}

*genes where *T. longissimum* differs from (lo) durum and *T. turgidum*
[#] genes where (lo) durum and *T. longissimum* differ from *T. turgidum*
[$] genes where (lo) durum differs from both its parents

Analysis of nucleotide polymorphisms and their corresponding amino acids indicated *cox3*, *rps13*, *mttB* genes have possible amino acid changes that could result in a change of protein function. We compared those genes within recognized amino acid changes among closely related monocots (*Zea perennis*, *Oryza sativa* cv. 'Indica') and using *Arabidopsis thaliana* as an out group (if available). It was found that observed amino acid changes exist among selected species in genes *cox3* and *mttB*, with an exception for the *rps13* gene, where change A/E^{57} was unique for the (lo) durum line only. Among other *Triticeae* species examined in the *mttB* gene we found amino acid

change characteristic only for (lo) durum, however present in other non-wheat species

(Figure 10).

cox3

```
                    70        80          240        250
           --------+---------+--   ------+---  --  ----+-
A. thaliana  SLGLIFILYTMFVWWRDVLRE  IICGIRQYLG HL  TKEHHV
Z. perennis  SLGLIFLLYTMFVWWRDVLRE  IICGIRQYLG HL  TKEHHV
O. sativa Indica SLGLIFLLYTMFVWWRDVLRE  IICGIRQYLG HL  TKEHHV
T. longissimum SLGLIFLLYTMFVWWRDVLRE IICGIRQYLG HL  TKEHHV
T. aestivum  SLGLIFILYTMFVWWRDVLRE  IICGIRQYLG QM  TKEHHV
T. turgidum  SLGLIFILYTMFVWWRDVLRE  IICGIRQYLG QM  TKEHHV
(lo) durum   SLGLIFLLYTMFVWWRDVLRE  IICGIRQYLG QM  TKEHHV
Consensus    SLGLIFlLYTMFVWWRDVLRE  IICGIRQYLG hl  TKEHHV
```

rps13

```
                    60
              -- - --+--
Z. perennis   QI E QMIAQ
O. sativa Indica QI E QMIAQ
T. aestivum   QI E QMIAQ
T. turgidum   QI E QMIAQ
T. longissimum QI E QMIAQ
(lo) durum    QI A QMIAQ
Consensus     QI e QMIAQ
```

mttB

```
            1        10          20
            +--------+----- - ---+-
A. thaliana  NPSLLALNYLYISYE F NFAS
Z. perennis  MPKMHLSFELLI--E W NFAP
O. sativa Indica MPKMHLSFELLI--E W NFAP
T. aestivum  MPQIHFSFELLI--F L NFAP
T. turgidum  MPQIHFSFELLI--F L NFAP
T. longissimum MPQIHFSFELLI--F L NFAP
(lo) durum   MPQIHFSFELLI--F W NFAP
Consensus    mP..h.sfeLlI..e . NFAP
```

Figure 10. Amino acid multi-alignment of *cox3, rps13, mttB* gene fragments. Grey boxes indicate amino acid changes. Only segments of genes have been shown where polymorphism was present.

In the final assembly there were three copies of *rrn18* in the *T. turgidum* and in (lo) durum genomes, while just one copy was found in *T. longissimum,* possibly due to incomplete gene assembly, but not actual lack of gene copy. No polymorphism within this gene was found among these three lines. Two copies *rrn26* were found in *T. turgidum,* two copies in *T. longissimum* (with one copy having the T/C^{743}SNP), and one full copy in (lo) durum.

Mitochondrial genomes are known to carry polymorphism within each cell and tissue type (Arrieta-Montiel *et al*., 2001). Deep sequencing (61-133x) allowed description of nucleotide variation *within* each mitochondrial genome. We were able to confirm the existence of heteroplasmy, based on the existence of SNVs within mitochondrial genomes, in each studied species. We identified 244, 218, 148 and 196 SNVs in *T. turgidum*, (lo) durum, and *T. longissimum* (contig A and contig B respectively). During the SNV search we identified multiple regions of higher SNV density within a given species when compared to regions with only single nucleotide changes (Figure 11), which we called "SNV blocks". These regions were unevenly distributed throughout the mitochondrial genomes examined. In *T. turgidum* we found 22 regions of high SNV density, containing a total of 151 SNVs, in (lo) durum's 15 regions, with 155 SNVs (Figure 12) and in *T. longissimum's* 14 regions, with 99 SNPs in contig I and 13 regions, and 147 SNVs in contig II. No specific pattern of SNV blocks was recognized. The number of SNVs outside of those SNV blocks was 93, 63, 98 in *T. turgidum*, (lo) durum and *T. longissimum,* respectively (Table C3).

A polymorphism search between final consensus sequences was also performed. We identified 349 and 740 SNVs in the (lo) durum line between *T. longissimum* and *T. turgidum*, respectively, and 731 SNVs between *T. longissimum* and *T. turgidum*.
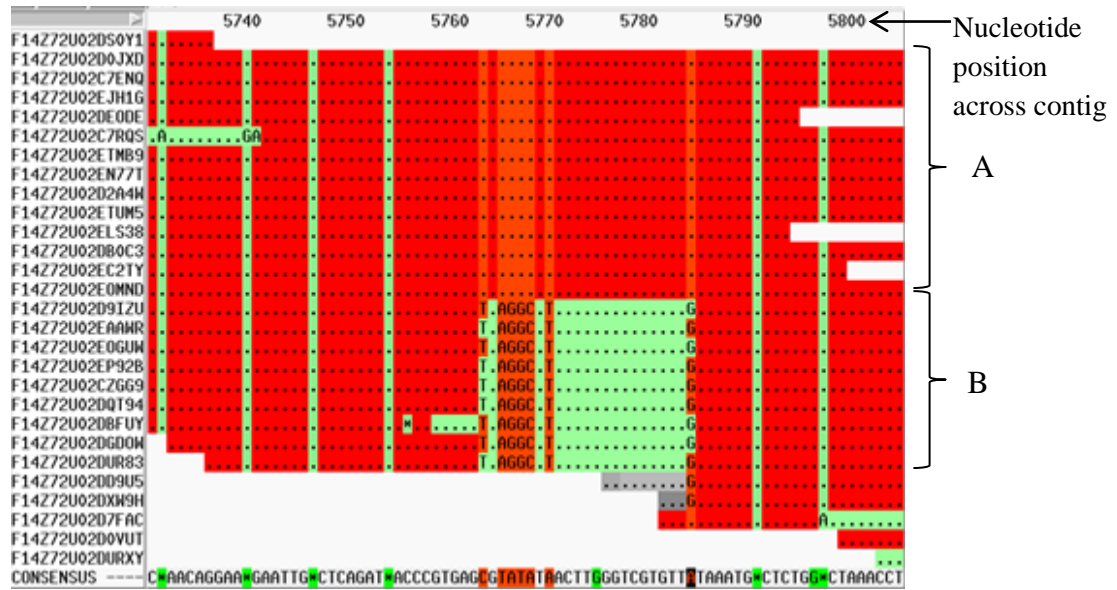
Figure 11. A segment of genome from the (lo) durum raw assembly (GAP4 contig assembly visualization). Region detected by single nucleotide polymorphism search algorithm. Picture demonstrates raw data reads assembled to create final consensus sequence. This kind of analysis shows hidden data, present in raw assembly data, which demonstrate regions of heteroplasmy and/or regions involved in recombination of mitochondrial genomes. In this case we see two mitotypes differentiated by region:
A – CGTATATAACTTGGGTCGTGTTA and B – TGAGGCTTACTTGGGTCGT GTTG. F14Z72U02xxxxx – designation of each sequencing read.
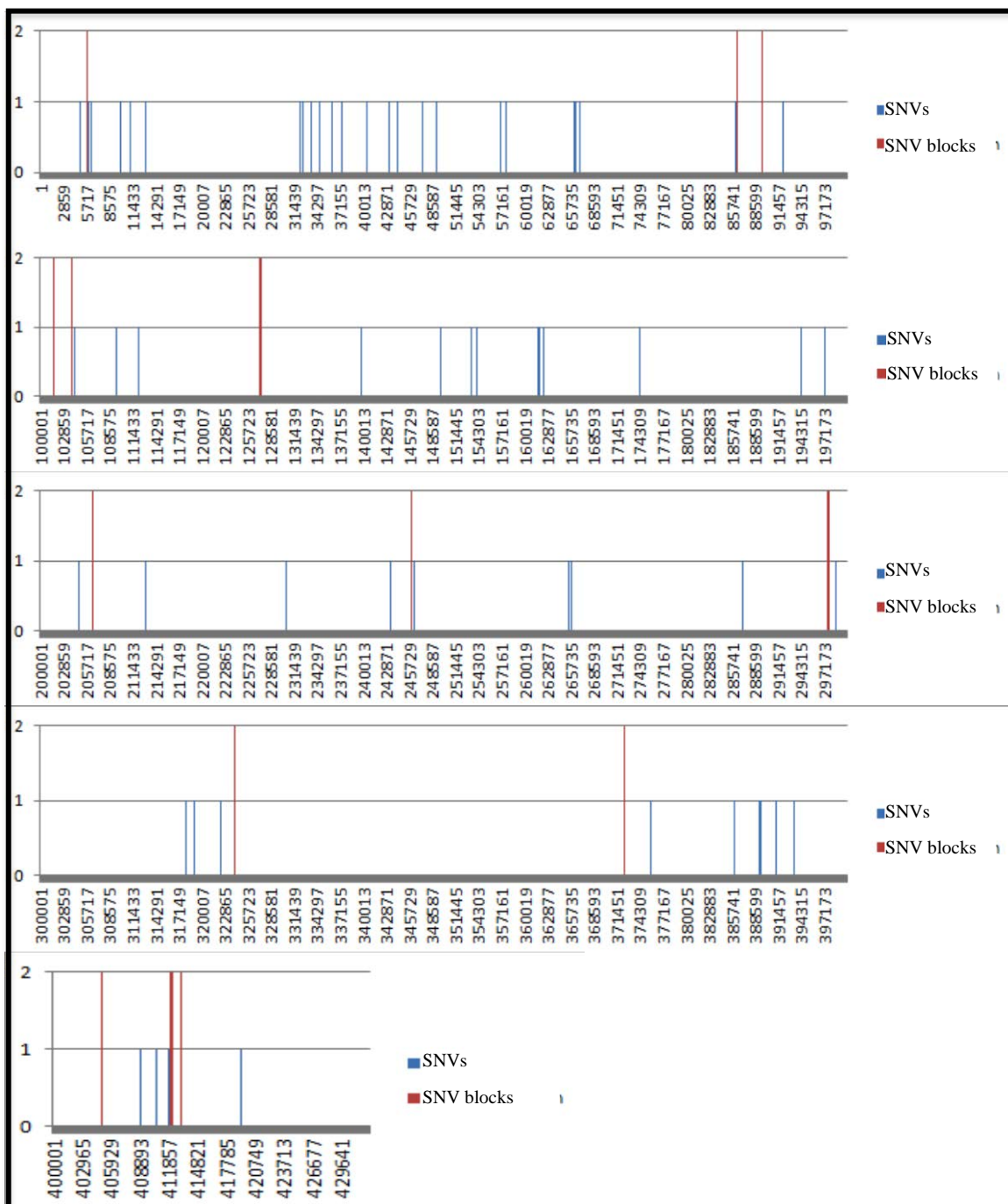
Figure 12. SNV distribution across (lo) durum mitochondrial genome. Blue lines indicates SNV positions, red line indicate the position of SNV blocks. We found 216 SNVs across the genome, out of which 63 were SNVs and the remaining 155 created 15 SNV blocks. The overall SNV density is 1 SNV/7,091 bp.

Genome structure–reference assembly

The *T. aestivum* and *T. turgidum* mitochondrial genomes were completely collinear and nearly identical. There was a region of about 1,200 bp (position in *T. aestivum* mitochondrial genome from 179,037 to 180,129 bp) where there was no coverage from the *T. turgidum* genome reads. Using the same approach, (lo) durum and *T. longissimum* genomes were assembled with *T. aestivum* as the reference mitochondrial genome sequence. We identified 12 and 13 regions (with size >3,000 bp) of no read coverage for *T. lon*gissimum and (lo) durum respectively. However, only 9 of these regions (regions I, II, V, VII, VIII, XII, XIII, XIV and XV) appear to be absent on both species. Additionally, size differences between those regions were observed (region I, V and IX). Regions III, IV, VI, VII, X, XI and XVI could be used to differentiate *T. longissimum* and the alloplasmic line. These could indicate the existence of genome rearrangement, gain or loss of sequence, or level of polymorphism at a particular position preventing proper assembly (Figure 13).
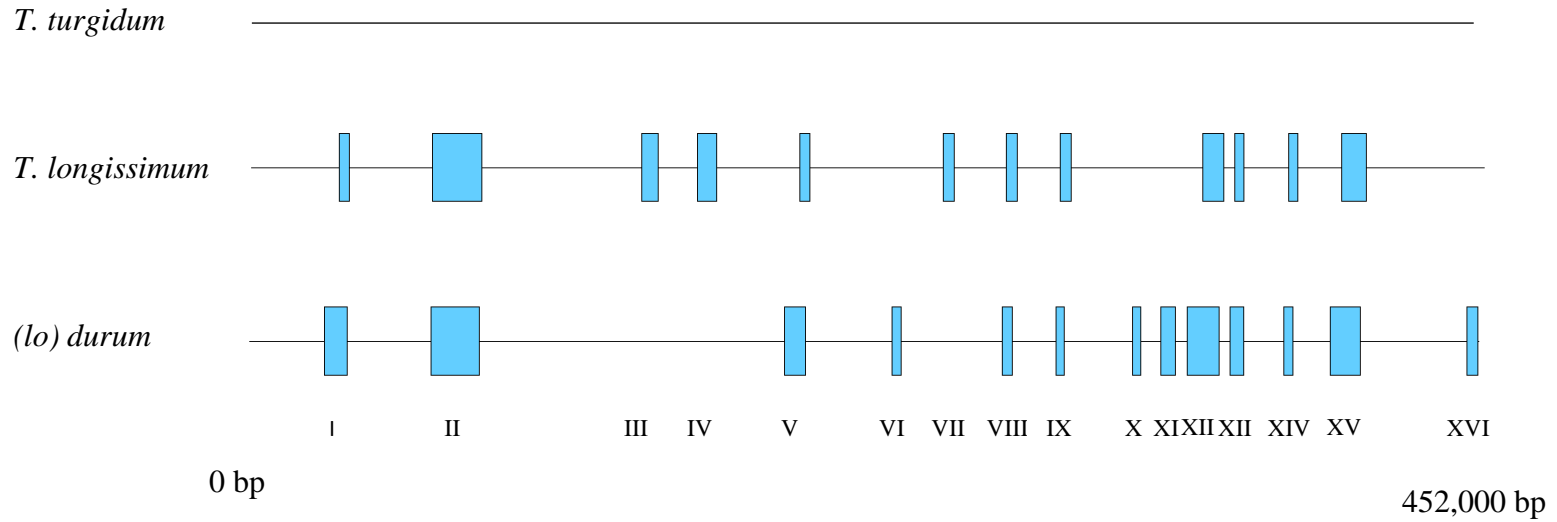
Figure 13. Comparison of *Triticum* mitochondrial genomes based on reference assembly. Blue boxes represent gaps (>3,000 bp) in reference assembly compared to *T. aestivum* (NC_007579.1), *T. turgidum* has a gap of 1,200 bp that is not indicated. Gaps could result from inversions, deletions/insertions or possible sequence differences in the genome.

Genome structure–*de novo* assembly

Assembled genomes indicated overall structural changes that may possibly reflect recombination events within the mitochondrial genomes. To precisely describe the possible recombination events within genomes and differences which could appear in the (lo) durum line as a result of the alloplasmic condition, a detailed gene order description was established. Based on the syntenic relationship between genomes, three gene categories were established: first genes which do not follow any recognizable pattern, second gene pairs that move together, and third, gene blocks that are found together when the three mitochondrial genomes were compared (Figure 14). Referring to these blocks there were 19 recombination/rearrangement events between *T. turgidum* and alloplasmic (lo) durum and 13 between *T. longissimum* and alloplasmic (lo) durum. Four gene pairs were indicated: *nad2-2*, *nad2-1*; *nad1-1*, *mttB*; *rps12*, *nad3*; *atp1*, and *atp9*. Despite the multiple genome rearrangements observed when compared to the *T. turgidum* genome, we were able to distinguish 7 conserved gene blocks between the alloplasmic line and both parents. These blocks account for 44 genes (or gene exons) out of 61 total, providing evidence for the relative stability of the gene order between species. Another interesting aspect related to gene order is that three gene blocks show gene rearrangements within them: blocks I, V and VI, with lengths of 28,642 bp, 20,955 bp and 22,939 bp in *T. turgidum* (Figure 15). Blocks I and VI show the same rearrangement, where genes *atp4* and *cob* were moved from one end to the other. In block V, *T. turgidum* has *orf359* added after the *ccmFC-2* exon and *nad5* exons 3 and 4 are reversed relative to (lo) durum and *T. longissimum*. In addition, *rps7* has been added to the block of *T. longissimum* and (lo) durum. The possible recombination events within gene blocks were only observed

between *T. turgidum* and both *T. longissimum* and (lo) durum. Observed gene blocks consist of about 32% (143,445 bp) of the entire genome.

We also identified differences in the number of multiple genes or gene alleles which were not present in our final assemblies. Gene *ccmB* and *rpl5* (marked red on Figure 14) are not present and genes *atp6* and *atp8* (marked green) are present only in one copy per genome in the (lo) durum assembly. The *rpl5* gene is present in the *T. longissimum* genome, but does not carry *orf359* (both genes marked blue), despite the fact that this genome was the donor of mitochondria to (lo) durum. Also, this genome does not have both copies of *atp6* and *atp8*.
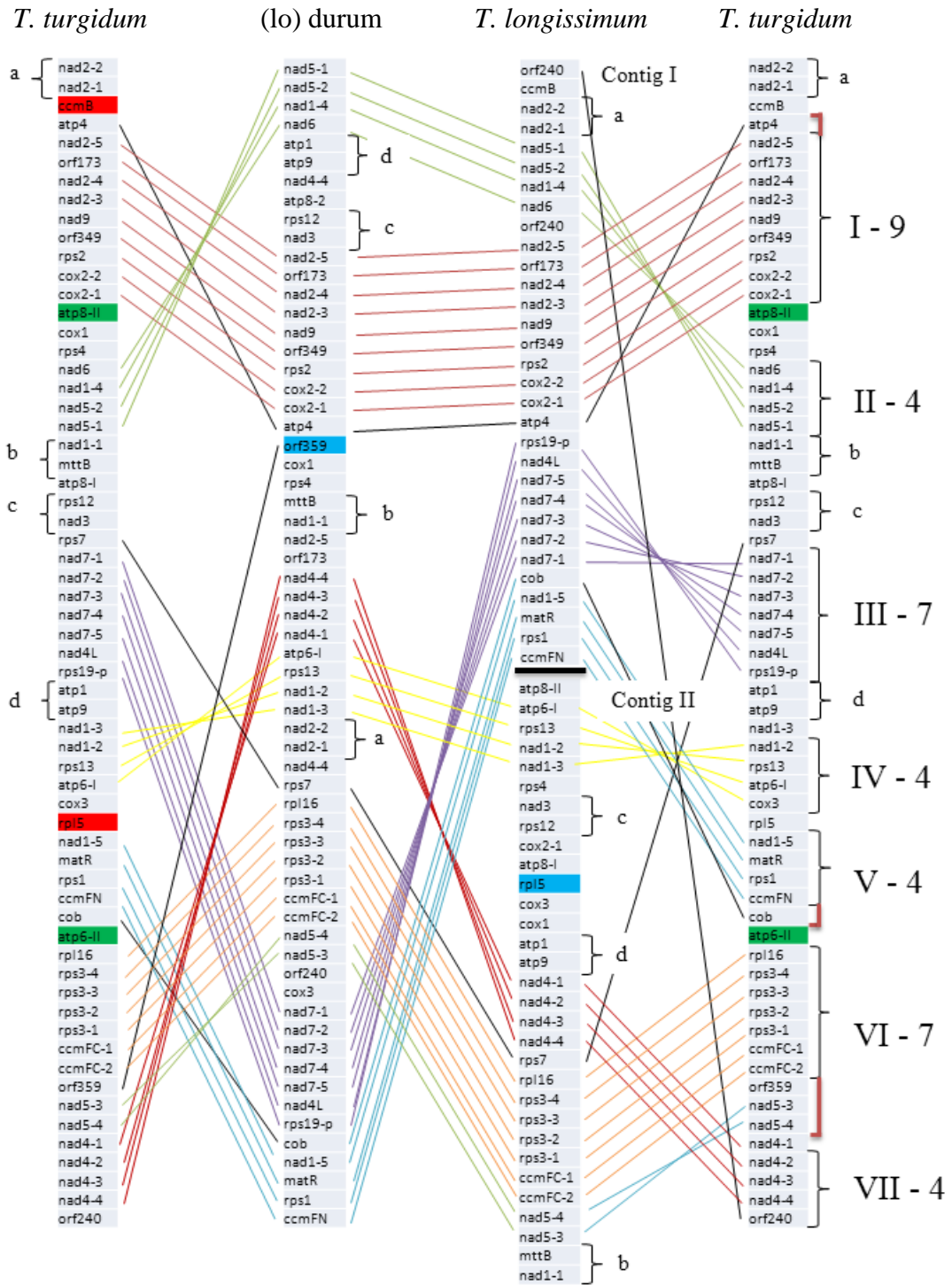
Figure 14. Graphical representation of the synteny observed in *T. turgidum*, (lo) durum and *T. longissimum* based on gene localization. Pairs of genes found together in all lines a, b, c, d and gene blocks (I -VII) are shown. Black line indicates position of end/beginning of two *T. longissimum* contigs. Genes labeled as green are present only in one copy in (lo) durum and *T. longissimum*, blue are missing from either *T. longissimum* or (lo) durum and genes labeled red are not present in our final assembly of (lo) durum line.

Block I

T. turgidum

| atp4 | nad2-5 | ORF 173 | nad2-4 | nad2-3 | orf349 | rps2 | cox2-3 | cox2-1 |

28,642 bp

(lo) durum

| nad2-5 | ORF 173 | nad2-4 | nad2-3 | orf349 | rps2 | cox2-3 | cox2-1 | atp4 |

32,742 bp

T. longissimum

| nad2-5 | ORF 173 | nad2-4 | nad2-3 | orf349 | rps2 | cox2-3 | cox2-1 | atp4 |

32,788 bp

Block V

T. turgidum

| nad1-5 | matR | rps1 | ccmFN | cob |

20, 955 bp

(lo) durum

| cob | nad1-5 | matR | rps1 | ccmFN |

19, 135 bp

T. longissimum

| cob | nad1-5 | matR | rps1 | ccmFN |

19, 234 bp

Block VI

T. turgidum

| rpl16 | rps3-4 | rps3-3 | rps3-2 | rps3-1 | ccmFC-1 | ccmFC-2 | orf359 | nad5-3 | nad5-4 |

22,939 bp

(lo) durum

| rps7 | rpl16 | rps3-4 | rps3-3 | rps3-2 | rps3-1 | ccmFC-1 | ccmFC-2 | nad5-4 | nad5-3 |

30,327 bp

T. longissimum

| rps7 | rpl16 | rps3-4 | rps3-3 | rps3-2 | rps3-1 | ccmFC-1 | ccmFC-2 | nad5-4 | nad5-5 |

29,989 bp

Figure 15. Selected gene blocks and their sizes among three sequenced species. Picture shows possible recombination events within each mitochondrial genome. Colors of the blocks correspond to color on Figure 14. Genes marked as red indicate possible recombination of mitochondrial genome; genes marked as blue are not present in *T. turgidum* in close proximity to gene Block VI.

Discussion

It is an established fact that mitochondrial genomes within a single cell or tissue type represent a highly polymorphic population. Multiple studies were performed to characterize mitochondrial heteroplasmy; however, most of those studies describe it only based on a few genes or ORFs (Hattori *et al.*, 2002; Welch *et al.*, 2006; Leeuwen *et al.*, 2008). In our studies, we utilized the Roche/454 GS FLX sequencing method, widely used in various genome sequencing projects. Deep sequencing (61-133x read coverage) of mitochondrial genomes allowed us to not only describe polymorphism of gene space, ORFs or structural genome differences between genomes, but also to search for nucleotide polymorphism across the whole mitochondrial genome. Next generation sequencing can provide the possibility to describe heteroplasmy and substoichiometric changes of the mitotypes.

Nuclear vs. mitochondrial gene copy

A search for genes identified in the *T. aestivum* mitochondrial genome (Ogihara *et al.* 2005), with the three sequenced genomes in this study, created a picture of stable mitochondrial gene content. All previously characterized genes were present in the alloplasmic line and its parents. Despite gene content conservation, multiple genes with major gene sequence differences were identified. The alloplasmic durum and *T. longissimum* lines share the same differences in genes such as *atp6*, *nad6*, *rps19*-p, *cob* and *cox2-2,* compared to *T. turgidum*. The *atp6* subunit of the F_1F_2–ATP synthase is considered to be a mitochondrial coded gene. Early studies recognized the chimeric structure of *atp6* in *Triticum* species (Bonen and Bird, 1988), and later its potential role was associated with cytoplasmic male sterility (Iwabuchi, 1993), where the proper

47

sequence editing of altered *apt6* restored male fertility. The latest work by Liu *at al*., 2011, where the alloplasmic line of *T. aestivum* with the cytoplasm of *Ae. kotschyi* was sequenced, showed the same differences in *atp6*, *nad6*, *rps19*-p, *cob* and *cox2-2*. However, these differences seem to be common among *Triticum* and *Aegilops* genera rather than possible cause for the CMS condition as reported by Hosseini, 2010 for the *atp6* gene. The alloplasmic line described by Liu *at al*., 2011, expressed CMS phenotype, and no other phenotypic differences were described.

To confirm allelic differences between *atp6* alleles in (lo) durum and parental lines, a PCR check was performed. It seems that the alloplasmic and euplasmic maintainer in our study both have *atp6*-T and *atp6*-L versions of the gene. However, the *atp6*-T version is only present in the (lo) durum nucleus rather than in the mitochondria. Recently it has been shown that two versions of *atp6* are present in a *Ae. crassa*, *T. aestivum* cv. 'Chinese Spring' and the alloplasmic line of 'Chinese Spring' with *Ae. crassa* cytoplasm (Kawaura *at al*., 2011). The *atp6*-CR (crassa) version in 'Chinese Spring' and *atp6*-AE (aestivum) version in *Ae. crassa* was present in less than 10% of the mitochondrial pool in the cell. However, the frequency of *atp6*-AE was 30% in the alloplasmic line, probably due to the paternal leakage. However, the presence of *Aegilops* and *Triticum* versions of *atp6* as sublimones in our study is very weak. There was no obvious amplification for *atp6*-T in (lo) durum or *T. longissimum* mtDNA and *atp6*-L in *T. turgidum* mtDNA. Further studies are necessary to validate nuclear/mitochondrial localization of that gene; however, a strong possibility of nuclear copies of both *atp6* alleles exists.

In plants, transfer of the mitochondrial (in general organelle) genes to the nucleus is well recognized and described (Martin and Herrmann, 1998; Adams *et al*., 2002; Sandoval *et al*., 2004; Leister, 2005). In the case of *Drosophila* it was established that the rate of evolution of the nuclear copy of a mitochondrial gene is much slower (10x) than that present in the mitochondria itself. That observation would allow us to compare the nuclear copy of *atp6* to that of mitochondria and determine the ancestral state.

Paternal linkage and recombination based on *nad9* gene analysis

We also observed major differences in the *nad9* gene, where the alloplasmic line has a different allele of the gene when compared to maternal donor line. Comparison of three *nad9* alleles among species shows that the allele found in the alloplasmic line is more similar to the *T. turgidum* copy of the gene, where the STOP codon is not present. This STOP codon is found in *T. longissimum* (Figure 6) and was also observed in alloplasmic wheat with an *Ae. kotschi* cytoplasm (Liu *et al*., 2011). The (lo) durum copy of *nad9* could be the result of recombination between maternal and paternal mitochondrial genomes, either during cytoplasm transfer or as a result of paternal leakage during subsequent crosses to the male parent. We can assume that the alloplasmic condition is not creating new gene alleles *per se*, since they are present among other *Triticum* species. In wheat the paternal leakage has been investigated in detail in the alloplasmic hexaploid wheat having *Ae. crassa* cytoplasm. It seems the proportion of paternal genes in the alloplasmic line increases by each initiated backcrossing, and after several backcrosses it stays at a constant level (Kawaura *at al*., 2011). The sequence of *nad9* in (lo) durum has similarities to both parents. Therefore, this copy of *nad9* not only shows paternal leakage, but also potential recombination between maternal and paternal

mitochondrial genomes. It is known that nuclear genes control the stoichiometry of alternative mitotypes (Zaegel *et al.*, 2006). Data also shows that CMS can emerge because of stoichiometric shift of particular mitotypes (Chen *at al.*, 2011). These facts allow one to conclude that in the alloplasmic line, the nuclear component responsible for proper communication between mitochondria is missing. In our model compatibility between the nucleus and cytoplasm is partially restored by the *scs* gene; however, the nature of that gene and its role is under investigation. The role of observed sequence differences and expression analysis is necessary to create a connection with observed phenotypic differences.

Substoichiometric shift

In the final assembly of the (lo) durum mitochondrial genome we were not able to find two genes, *rpl5* and *ccmB*. The *rpl5* gene is a ribosomal protein, responsible for rRNA maturation and formation of the 60S ribosomal subunits. Its function could be critical to the alloplasmic line and its survival. When we searched for the *rpl5* gene in raw sequencing data, we found a full gene sequence. In recent publication about the mitochondrial sequence of CMS *Triticum* line, the lack of an *rpl5* gene also was reported (Liu *et al.*, 2011). Our suspicion is that the lack of this gene in final the assemblies of both alloplasmic *Triticum* lines can be related to the proportion of mitochondrial (substoichiometric levels) carrying of the gene (Chen *et al.*, 2011). Final mitochondrial genome assembly reflects the Master Circle; however, mitochondrial genomes within a cell or cell types do not necessarily exist as that one, uniform structure. It is possible that the *rpl5* gene can exist in a separate strand of DNA, in much smaller proportion to other sublimons. A similar situation as for *rpl5* likely exists for the *ccmB* gene. In our

assembly, *ccmB* also was not found in the final assembly, but present in the raw

sequencing data. The *ccmB* gene is involved in cytochrome *c* biogenesis. These

observations create the basis for a hypothesis that a single gene could be "taken out" from

a master circle of predominant mitotype, due to possible environmental or developmental

needs causing stoichiometric shifts in the genome. In humans, heteroplasmy was

considered a rare phenomenon. The advent of new sequencing technologies allowed us to

describe mitochondria as a population of a cell with great precision (Goto *et al*., 2011).

The importance of heteroplasmy was recognized in multiple human studies, where

connection between the heteroplasmic state of a cell to human diseases was made (Elson

and Lightowlers *et al*., 2006; Avital *et al*., 2012). It is of great importance to recognize

the mechanisms which control heteroplasmy. Possible ways of regulating heteroplasmy

would allow us to control the compatibility of hybrids created during the breeding

process giving breeders access to genetic variability, beyond the scope of their programs,

due to NC incompatibility (Tsunewaki *et al*., 1993, 1996, 2002). It was identified that a

nuclear gene could play a role in the regulation of substoichiometric shift (Abdelnoor *et

al*., 2003), giving a chance to create compatibility restoration system similar to CMS

restoration, however targeted at other traits.

Open reading frames (ORFs)

We were able to recognize ORFs characteristic of each line, demonstrating the

possibility that new genes are formed as a result of environmental stress and evolutionary

needs such as the alloplasmic condition. The same ORFs as 63, 65, and 112 from the

alloplasmic durum line were also observed in the alloplasmic hexaploid wheat containing

Ae. kotschyi cytoplasm (Liu *et al*., 2011). The most characteristic ORFs recognized was

chimeric *orf113* (Figure 9) in the (lo) durum line, composed from four fragments of other mitochondrial genes, such as *rps2*, *cox1*, *nad4*-2, *rps19*-p. Appearance of new chimeric ORFs and their expression has been connected with CMS phenotypes. For instance, *orf72* in wild cabbage is built from parts of *atp9* and was expressed in CMS lines, but not normal lines (Shinada *et al.*, 2006).

Gene space polymorphism

We also performed a nucleotide polymorphism search within known mitochondrial genes in sequenced species. The most important fact is a difference between the (lo) durum line and *T. longissimum* where 6 nucleotide changes are reported, with three SNPs found in the *rps2*, *rps4* and *rps13* genes. Nucleotide variation within ribosomal genes were also observed in alloplasmic line of *T. aestivum*, where among others, differences in *rps2* and *rps4* were recognized (Liu *et al.*, 2011). The observation that ribosomal genes could be possible mutation "hot spots" in alloplasmic lines can be critical in understanding CMS in plants. Ribosomal proteins are responsible for protein expression mechanisms and their mutations could adversely affect this fundamental biological process. Also this finding can be utilized to create potential SNP-based markers, to identify other cytoplasms of *Triticum* species. The nucleotide variations in *cox3*, *rps13*, *mttB* genes were compared to the *A. thaliana, Z. perennis, T. aestivum* and *O. sativa* 'Indica' version of the genes. Nucleotide variation found in those genes appear to be conserved among these species (Figure 10). Only in the case of the *rps13* gene from the alloplasmic line, did we found and amino acid change, which was not recognized in other species. These data supports the hypothesis of accelerated evolutionary changes in the alloplasmic line as observed in other cases (Darracq *et al.*, 2011).

Nucleotide polymorphism across mitochondrial genomes

We were able to detect nucleotide polymorphism within each mitochondrial genome by identifying single SNVs and SNV blocks. No specific patterns between species were recognized. The SNV blocks can be identifiers of heteroplasmy in each genome and possibly used to quantify the proportion of particular mitotype. A higher density of SNPs recognized as haplotype blocks, could be possible regions of mitochondrial recombination, as proposed based on analysis of human haplotype blocks (Hey, 2004). Overall, mitochondrial genome polymorphism comparison showed, as expected, that the alloplasmic line is more closely related to the maternal donor *T. longissimum* than to *T. turgidum*. In the (lo) durum line, one particular region of DNA was observed, designated as *orf359* (Ogihara *et al*., 2005), with the highest level of polymorphism compared to other regions of the genome. Gene content and order showed that *orf359* doesn't exist in the cytoplasm donor line, but was completely conserved (sequence and position) among *T. aestivum* (Ogihara *et al*., 2005) and *T. turgidum*. Existence of *orf359* in the alloplasmic durum line is more evidence of paternal leakage, although highly mutated. It would be important to track the changes in this DNA region during the backcrossing process of the alloplasmic line, and establish the existence of this ORF in various *Triticum* lines.

Mitochondrial genome structural differences

By assembling various genomes, we were able to observe the differences in organization between *T. turgidum*, *T. longissimum* and the alloplasmic line (Figure 14). It is evident that the alloplasmic line is the most distant to *T. turgidum* and *T. longissimum*. Genome rearrangements and structural changes were also observed by *de novo* assembly

of the mtDNA of the three lines studied here. In addition those changes were confirmed

by a search of mitochondrial gene synteny. We observed multiple synteny blocks. It is

possible that the conservation of gene order within blocks is important to gene

expression. Studies on wheat mitochondrial gene expression show that *nad6* and *nad1*

(exon d) are co-transcribed (Haouazine *et al.*, 1993). That both genes were found to be

present in Block II, despite high structural differences, further support this hypothesis.

Different orders of genes have not only been reported between species of a same family

due to gene shuffling (Ogihara *et al.*, 2005), but also among different ecotypes of a single

species like *Arabidopsis* (Ullrich *et al.*, 1997). Different genome structure has been

observed, comparing two normal lines as well as the CMS lines in maize (Allen *et al.*,

2007), rice (Bentolila *et al.*, 2012) and wheat (Liu *et al.*, 2011). Structural changes

between the three genomes were expected, but the amount of these changes established in

the short period of time in the alloplasmic line suggests strong evolutionary forces at

play. The number of rearrangements between mitochondrial genomes of (lo) durum and

T. longissimum were much higher than *T. turgidum* and *T. aestivum,* even not considering

the 10,000 years of evolution that separates the later species. It was amazing that the *T.*

longissimum mitochondrion has gone through such a drastic structural change possibly as

a result of the alloplasmic condition. There could be two reasons for these observed

structural genome differences, the first of which is that recombination in the

mitochondrial genome is controlled by the nucleus and not by a random event. There is

evidence which suggests involvement of two nuclear genes in mitochondrial

recombination and repair, identified as *MSH1* and *RECA3* in *Arabidopsis* (Shedge *et al.*,

2007). Second would be the possibility of preferential selection for proper mitotype,

related to substoichiometric shift. In that case, recombination could occur in a random fashion, but only the best form of a mitotype would become prevalent. This phenomenon could be related to "genome shock" (McClintock, 1983), where the mitochondrial genome would be adapting, by developing drastic changes in its structure to survive in a new nuclear background. This accelerated evolution of newly formed mitotypes will lead to different forms than the maternal type in organization and genetic information, as found here and in other species (Allen *et al*., 2007; Bentolila and Stefanov, 2012). Particular variability created by unfavorable conditions of alloplasmic environment could give a possibility to manipulate mitochondrial genome content and create new genetic variability, to be utilized. Considering a number of traits identified, which are affected by mitochondria (Tsunewaki *et al*., 1993, 1996, 2002) and rate of evolution of mitochondrial genes (Parkinson *et al*., 2005; Allen *et al*., 2007; Bentolila and Stefanov, 2012), there is a potential to obtain desired results relatively fast, when compared to nuclear gene mutation rates.

Conclusions

Many important questions on the nature of mitochondria and its role in determining phenotype still remain. Next generation sequencing technologies are powerful tools to study the heteroplasmy and substoichiometric changes in plants especially when working with alloplasmic lines. Studies of structure, recombination and role of particular genome elements found in mtDNA in the context of NC hybrids are necessary. In the past, the connection between observed CMS phenotypes to particular gene change (Mohr *at al*. 1993) and the appearance of new ORF (Köhler *et al*., 1991) was made; however, mtDNA analysis of alloplasmic lines on genomic level, reveals a

more complex view of changes between parent and progeny. The level of rearrangements and nucleotide changes in mtDNA of the alloplasmic line that occurred in less than half a century was surprising when compared to almost no differences between the *T. turgidum* and *T. aestivum* genome even after 10,000 years of separation. Comparative genomic approach provided us with an excellent background for future explanation of the role of particular elements in creating aberrant phenotypes of hybrids. The amount of changes in the genes received from the paternal donor was high, emphasizing mechanisms of heteroplasmy, recombination and mutation. In connection with previous studies on different cytoplasmic backgrounds and its role in developing particular traits (Tsunewaki *et al*., 1993, 1996, 2002), our results should bring identification and understanding of the processes which take place in wheat alloplastic lines to a new light. This can provide a wide range of basic information allowing, the possibility of creating variations for crop improvement.

PAPER II. GENOME CONSERVATION BETWEEN *T. TURGIDUM* AND *T. AESTIVUM*, BUT NOT *T. TAUSCHII*, REVEALS THEIR EVOLUTIONARY RELATIONSHIP

Abstract

To better understand the genetic differences between the ancestral parents of *Triticum aestivum* (AABBDD), we sequenced the mitochondrial genome of durum wheat (*Triticum turgidum* var. *durum* L.; AABB) and *T. tauschii* (DD). Genomes were sequenced using pyrosequencing method. These two species are considered to be female and male parents to *T. aestivum* respectively. Comparison of *T. turgidum* and *T. aestivum* showed 40 SNP differences overall across the mitochondrial genome and 5 SNPs in the *rps1*, *rps2*, *cox3*, *ccmFN* genes. When *T. tauschii* was compared to *T. aestivum*, we found major differences in *atp6*, *nad9*, *nad6, rps19*-p, *cob* and *cox2* genes, and identified 605 SNP differences across both genomes and 25 SNPs within known gene space. Using reference assembly, the structure of *T. turgidum* and *T. aestivum* appear to be nearly identical, with no rearrangement, but the *T. tauschii* genome shows a high number of syntenic differences. The presence of multiple mitotypes in *T. turgidum* was noted. These results confirm the overall high conservation of mitochondrial genomes of *Triticum* species and the ancestral relationship between durum and bread wheat. Hybridization of cytoplasms between these two species did not create additional polymorphism or paternal leakage in hexaploid wheat. Detected heteroplasmy in durum wheat can provide additional information as to the hybridization events during wheat evolution.

Introduction

Bread and pasta wheat are the most important cereal crops in the world (Feldman *et al.*, 1995), next to rice and corn, with about 704, 722 and 883 million tons produced, respectively (http://faostat.fao.org/). The importance of durum wheat comes from unique characteristics and end products, which makes this crop suitable for pasta production (Elias, 1995). There has been an increasing interest in recent years for products derived from durum, which is considered a good low-fat protein and carbohydrate source that is convenient and nutritional (Hu, 2003). Bread wheat is widely used as human food and livestock feed. It provides essential amino acids, minerals, vitamins, and dietary fiber to the human diet (Shewry, 2009). Therefore there is a constant need for improvement of quality, resistance to environmental stresses, pests and diseases, and to provide a product that will match the changing expectations of consumers and growers. Genetic variability has allowed the creation of more than 25,000 types of wheat grown in a wide range of environments (Feldman *et al.*, 1995). Genetic improvement of crops can be extended into analysis of the genetic potential of the cytoplasm, which underlay photosynthesis and cell metabolism.

Genetic diversity studies and historical perspective

Studies on genetic diversity in the cytoplasm of *Triticum* species date back to 1951, with Kihara's work describing the evidence of genetic diversity between *Triticum* and *Aegilops* (Kihara, 1951). Three main centers interested in studies of the *Triticum* cytoplasm were S.S. Maan at North Dakota State University, USA, I. Panayotov at the Institute for Wheat and Sunflower, Bulgaria and K. Tsunewaki at Kyoto University, Japan. These three centers were conducting extensive studies on nuclear–cytoplasmic

interactions of various *Aegilops* and *Triticum* species. The main goal of these studies was to create a clear picture of the evolutionary relationship within the *Triticeae* tribe based on the cytoplasm and use that knowledge in breeding programs.

Tetraploid and hexaploid wheat evolution

The *Triticum* tribe arouse as a result allopolyploidization and domestication. By definition an allopolyploid is an organism that has more than two sets of haploid chromosomes inherited from different species. The beginning of agriculture is dated to start about 10,000 years ago in the Fertile Crescent, Near East, where the domestication of wheat species was recognized to take place (Smith, 2001) and from there wheat spread to other continents. The polyploid condition of wheat is one of the major reasons for its wide presence around the globe (Dubcovsky and Dvorak, 2007). The common wheat, *T. aestivum* L. ($2n = 6x = 42$), is a hexaploid with genome AABBDD; and *T. turgidum* L. var. *durum* ($2n = 4x = 28$), is a tetraploid with genome AABB (Feldman, 2001). It was recognized that the *T. turgidum* nuclear donors are diploid *T. urartu* (diploid AA, with chromosome number of $2n = 2x = 14$) and the B genome is likely derived from the *Aegilops* genus, where the *Ae. speltoides,* S genome is genetically the closest species to the B genome. *Triticum aestivum* was created by the hybridization event between *T. turgidum* and *T. tauschii* (diploid DD, with chromosome number of $2n = 2x = 14$) also referred to as *Ae. squarrosa* (McFadden and Sears, 1946; Feldman, 2001). *Triticum urartu* and *T. tauschii* are likely monophyletic in origin, but have diverged since and show major phenotypic differences (Kihara, 1951). Different genomes of the *Triticeae* tribe show little similarity, and potential interspecific hybrids show complete sterility and are isolated. Multiple studies on the *Triticum* and *Aegilops* cytoplasm revealed plasmon

differentiation among diploid species (Maan, 1975; Tsunewaki, 1988). It was found that

the cytoplasm from *T. urartu* (AA genome donor) was incompatible with emmer wheat,

the successor of domesticated durum (Tsunewaki *et al*., 1999). *Triticum tauschii* was

recognized as the male parent of *T. aestivum* (McFadden and Sears, 1946). Recognition

of evolutionary relationships within the *Triticeae* tribe (Figure 16) was critical to utilizing

the potential genetic variability remaining in the tribe, facilitating easier nuclear genetic

exchange between various species to accommodate the growing need for novel

variations.

Importance of cytoplasm in wheat breeding

Wild species are the primary source of new genes for cultivated wheat. These

genes can provide breeding programs with the various traits necessary to improve

cultivars for adaptation to a changing environment. Genes improving resistance to abiotic

(Nevo and Chen, 2010) and biotic (Reader and Miller, 1991) stresses are found in that

gene pool, especially in the wild durum progenitor emmer wheat. Crosses with alien

cytoplasm are not always successful. Lack of crossability can be related to nuclear

genome differences, creating various chromosomal aberrations, often resulting in hybrid

sterility (Baum *et al*., 1992). Nuclear-cytoplasmic (NC) incompatibility are an additional

obstacle to wheat hybrid creation (Maan, 1975). One example of NC incompatibility was

a cross between *T. longissimum* and *T. turgidum*, where no viable progeny could be

obtained. Two genes, *scs* (species cytoplasm-specific gene), and *Vi* (vitality gene), were

identified which restored compatibility between the cytoplasm and nucleus in *T. turgidum*

(Maan, 1992a and 1992b). It is possible, that a similar compatibility restoration system

can be utilized in other wide crosses of *Triticum*. Determining connections between the evolution of wild species, domestication events and nuclear–cytoplasmic compatibility is critical to a wider use of wild species. That knowledge creates an opportunity to access a diverse gene pool.
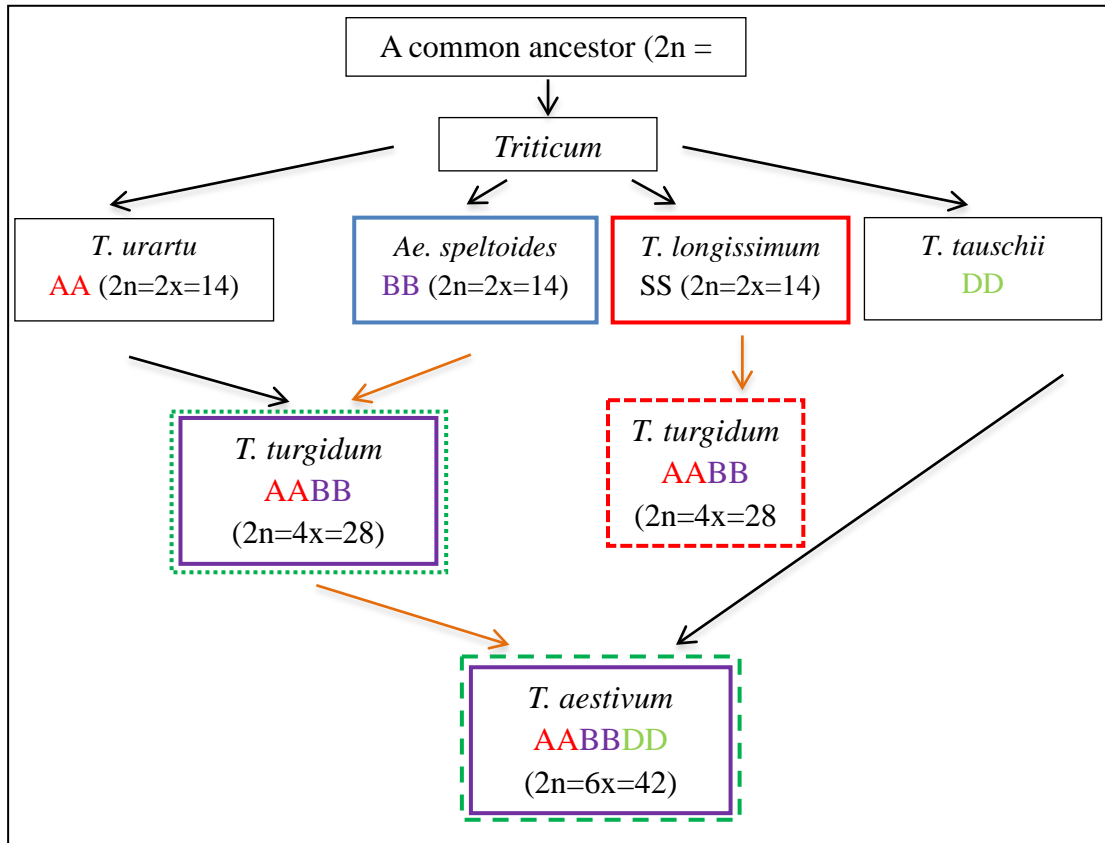
Figure 16. Cytoplasm and genome evolution in *Triticeae* species. *Aegilops speltoides* is the most probable donor of B genome and cytoplasm for *T. turgidum* (Feldman *et al*., 1995). The A genome donor was *T. urartu*, where cytoplasm is not compatible with *T. turgidum* nuclear genome (Tsunewaki *et al*., 1999). *Triticum tauschii* was recognized as the D genome donor for *T. aestivum*. Incompatibility of cytoplasm from *T. longissimum* with *T. turgidum* was recognized by Maan, 1991.

<center>**Results**</center>

Genome assembly

Sequencing of the *T. tauschii* sample provided results as expected. In the final genome assembly, average sequencing read length was 468 bp, indicating good quality data. We were able to create two contigs, with a total consensus length of 399,374 bp and an average coverage of 116x (Table C2). The *Run10Bastien3* and *Run4* assembly runs resulted in 105 and 63 contigs, with 33 and 36x average coverage, respectively (Table C6). We used that data to describe known mitochondrial gene space (Ogihara *et al*., 2005), search for open reading frames and identify single nucleotide polymorphisms within and across the sequenced genomes.

Genome and gene space polymorphism

High conservation of mitochondrial gene content was observed between *T. aestivum*, *T. turgidum* and *T. tauschii*. We identified all previously characterized genes among sequenced species. *Triticum tauschii* showed a multiple differences, based on sequence synteny, gene sequence and polymorphism. Analysis of polymorphism in the gene space between *T. aestivum* and *T. turgidum* a conservation sequence. We identified only 5 SNPs in gene space in *T. turgidum* in relation to *T. aestivum,* where two SNPs are in *rps1* and one SNP in the *rps2, cox3* and *ccmFN* genes respectively. Overall we identified 40 SNPs across the genomes between these two species. When *T. tauschii* was compared to *T. aestivum* we identified 27 SNPs and 2 dinucleotide changes within described gene space. From these 10 SNPs and one dinucleotide, change took place in ribosomal genes *rps13*, *rpl5*, *rps1*, *rps2* and *rps4*, which represent about 37% of polymorphism observed within gene space. Overall, mitochondrial genome

<center>62</center>

polymorphism comparison showed 679 SNPs between *T. tauschii* and *T. aestivum* (in homologous regions of genomes only).

We found differences in SNPs between the final and raw assembly data. In *T. turgidum, matR*, *rps1* and *ccmFN* show additional polymorphism in raw assembly data. In our final assembly we did not identify any SNPs in *matR*, but two SNPs in *rps1* and one SNP in the *ccmFN* gene were observed. However, when we looked at raw assembly data (contigs directly generated by assembler), one SNP was present in *matR,* no SNPs in *rps1* and two SNPs in *ccmFN*. This result indicates that reference assembly can overlook polymorphisms present in the mitochondrial genomes. Because of that, we described SNP polymorphism only based on raw assembly results. In *T. tauschii* assembly similar situation was observed. However, only in the case of *rps1* gene was an additional SNP recognized (Table 4)*.* It is important to note that all SNPs appear to be present in both species, indicating that observed polymorphism is common between species. Observed differences are possible results of heteroplasmy of mitochondrial genomes.

Table 4. Comparison of raw and final assembly alleles of *matR*, *rps1* and *ccmFN* genes relative to *T. aestivum* sequence.

	T. turgidum		*T. tauschii*	
	Final assembly	Raw assembly	Final assembly	Raw assembly
matR	+	T/G	T/G	T/G
rps1	C/A, C/T	+	C/T	C/A, C/T
ccmFN	C/A	C/A, A/T	C/A, A/T	C/A, A/T

+ - indicates identical allele in *T. aestivum*.

Polymorphism in genes with major differences (*atp6*, *nad9*, *nad6*, *cob*, *rps19*-p) and genes with multiple alleles in the cell were not taken into account in the above description. During the SNP search we recognized multiple nucleotide insertion/deletion events in the final assemblies. The majority of them were present in poly A, T, G and C

regions. Because of the presence of these errors we described only nucleotide changes. In *T. turgidum, nad3* and *rps1* show at least two alleles each and *atp8* had four alleles. Two of the *atp8* alleles from *T. turgidum* are present in *T. tauschii.* The same is seen in the case of *rps1* and *nad3*. These results demonstrate that a description of polymorphism based only on consensus data can create an imprecise view of nucleotide variation among different mitochondrial genomes.

Major gene difference between *T. tauschii, T. turgidum* and *T. aestivum*

Further analysis of gene space indicated that: *atp6*, *nad6*, *nad9, rps19*-p*, cob* show sequence differences between *T. tauschii* and both *T. turgidum* and *T. aestivum*. *Atp6* is made from two regions, pre-sequence and core region. This gene has the same sequence in *T. turgidum* and *T. aestivum*; however, in *T. tauschii* the pre-sequence is different. Additionally, we recognized 7 SNPs (A/C^{521}, TA/AT$^{543-544}$, TA/AT$^{580-581}$, C/A^{595}, G/T^{637}) which exist in *T. tauschii* when compared with *T. turgidum* and *T. aestivum* (Figure D4).

In the *nad6* locus we recognized 3 SNPs (G/A^{483}, A/T^{537}, A/G^{697}) and two dinucleotide changes (GA/CC^{630631}, AA/GG$^{691-692}$) in addition to polymorphism within the conserved region. We identified a region of high dissimilarity starting from position 703 in *T. tauschii* (Figure D7).

Another major gene difference was recognized in the *nad9* locus, where four nucleotide deletions in region TGTG$^{134-137}$ create a premature STOP codon at position 158 (TAG$^{158-160}$). As a consequence, a major alteration in protein length is created when compared to *T. turgidum* and *T. aestivum* (Figure D6).

Nine nucleotide differences were recognized in the *rps19*-p locus. We identified two alleles of the gene in *T. turgidum*. The most abundant allele *rps19*-p (1) (Figure D8) was found in the final durum assembly, and another allele, *rps19*-p (2) was found in raw assembly data only. In *T. tauschii* and *T. aestivum* only *rps19*-p (2), was present. Unavailability of raw sequencing data from the *T. aestivum* mitochondrial genome did not allow us to confirm the presence of additional *rps19*-p allele in that species. The *cob* gene is 10 nucleotide shorter sequence in *T. tauschii* at the 3'-end of the gene.

We were able to identify parts of the *atp6*, *nad6*, *nad9* genes in *T. tauschii* and in *T. turgidum* raw data assembly. That information, together with SNP data, strongly suggests the presence of at least two mitotypes in *T. turgidum*. However, low quality of sequence data for *T. tauschii* indicate the need for further confirmation of results.

Four ORFs were reported in the *T. aestivum* sequence *orf359*, *349*, *240* and *173*. The *orf359* did not show any polymorphism between *T. turgidum* and *T. aestivum*, but was missing from *T. tauschii* assembly data. The *orf240, 173* were 100% identical in the three species, but *orf173* was truncated in the final genome assembly of *T. tauschii*. We identified one SNP (G/T^{744}) in *orf349* of *T. turgidum* and four SNPs (TT/AG^{68}, G/T^{221} and G/T^{744}) in *T. tauschii*.

Ribosomal and transfer RNA

In *T. turgidum* tRNA genes we did not identify any polymorphism in comparison to *T. aestivum.* Also there was no polymorphism in *rrn18* (three copies in the genome of *T. turgidum*), *rrn26* (two copies), *rrn26*-p and *rrn5* (three copies each). In *T. tauschii* we found polymorphism in His-RNA at T/A^{10} and T/G^{42}. We also identified copy number differences in Gln-tRNA, Lys-tRNA (two copies out of three), Met-tRNA (three copies

65

out of four), and single copies of *rrn18*, *rrn26, rrn5* compared to the three copies in the *T. turgidum* and *T. aestivum* genome. Two copies of *rrn26*-p were identified in our final *T. tauschii* assembly.

Structural differences between genomes

The sequence of *T. turgidum* is nearly identical to that of previously reported *T. aestivum*, with the exception of a region of about 1,200 bp (approximate position in *T. aestivum* mitochondrial genome is from 179,037 to 180,129 bp) where there was no coverage from the *T. turgidum* genome reads. We were able to obtain one linear molecule of the *T. turgidum* mitochondrial genome, joining the "ends" of the *T. aestivum* mitochondrial sequence. Comparing *T. tauschii* and *T. aestivum,* we found multiple synteny breakages. Looking at gene organization between genomes, we identified 7 and 12 breakages in contig I and contig II of *T. tauschii*. We identified 6 gene blocks, which appear to be in common between all three species, in addition to 5 gene pairs and 7 genes not connected to gene blocks (Figure 17).
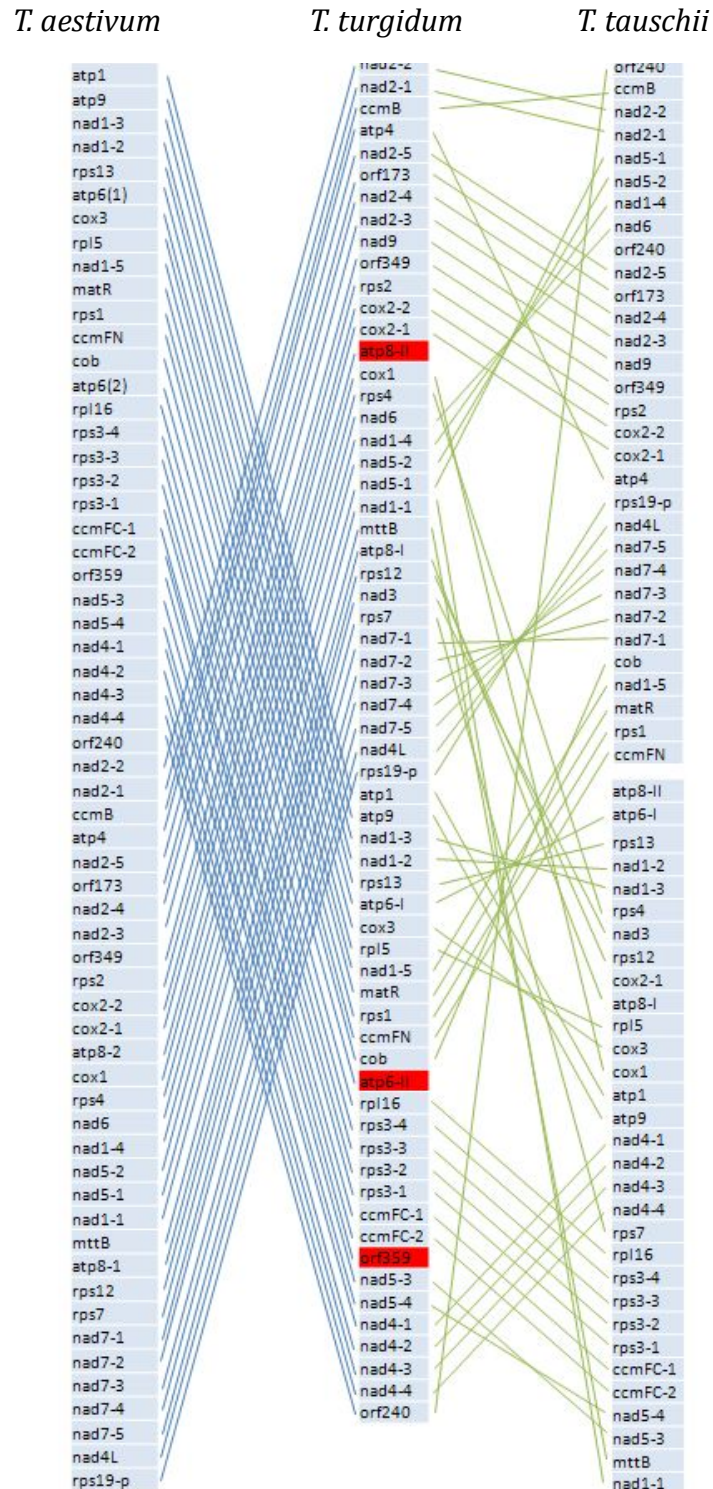
Figure 17. Graphical representation of the synteny observed in *T. aestivum, T. turgidum* and *T. tauschii* genomes. Genes labeled in red are not present in our final assembly of *T. tauschii* genome. Full gene order conservation between *T. turgidum* and *T. aestivum*, contrasted with multiple gene synteny breakages with *T. tauschii*.

Discussion

The study of nuclear-cytoplasmic interactions among various *Triticum* species started with the work of Kihara (1951). A main result of these studies was the proper identification of ancestral relationship between *Triticum* species. Unfortunately, detailed mitochondrial genomic sequence data of both proposed parental species of hexaploid wheat (*T. aestivum*) have been lacking. This crucial information was only available for *T. aestivum* (Ogihara *et al*., 2005). To create a full picture of cytoplasmic genome evolution and to explain observed diversity (Kihara, 1951), it is critical to obtain mitochondrial genomic sequences from other members of the tribe including the parental species of the most important; hexaploid wheat. To achieve part of this goal, mitochondria of *T. turgidum* and *T. tauschii*, considered to be the parents of *T. aestivum*, were sequenced and described in this study. Comparison of these genomes with other sequenced mitochondrial *Triticum* genomes allowed us to determine changes in the mitochondria as a result of polyploidization and evolution.

Triticum aestivum NC compatibility and its ancestral relationships

Triticum aestivum was created by multiple hybridization events, beginning with hybridization of diploid species to form a tetraploid species and ending with a second hybridization to form the hexaploid genome of *T. aestivum* (Figure 16). As a result, *T. aestivum* is a hybrid of all genomes of its ancestors, *Ae. speltoides* (AA genome), *T. urartu* (BB) and *T. tauschii* (DD), creating hexapolid genome AABBDD. However, during polyploidization only the female cytoplasm was maintained. It was recognized that *T. turgidum* donated the cytoplasm to *T. aestivum* (Kihara, 1951). As our results indicate, we were able to show that *T. aestivum* and *T. turgidum* share the same

68

mitochondrial genome (Figure 17), where the full synteny of gene order is observed. However, when we compared both species to *T. tauschii* we observed multiple differences in structure. *Triticum turgidum* in turn inherited its cytoplasm from *Ae. speltoides* (Feldman *et al*., 1995). It was also determined that the cytoplasm of A genome donor is not compatible with the *T. turgidum* nuclear genome (Tsunewaki *et al*., 1999). Artificial transfer of cytoplasm was attempted by Maan, 1991, where transfer of a *T. longissimum* cytoplasm to *T. turgidum* was performed; however, incompatibility between both species was recognized. Our results indicate that *T. tauschii* and the *T. longissimum* share highly similar cytoplasmic genomes in structure (Figure A2), size (Table C2) and gene content (Table C4, C5) with only overall genome SNP polymorphism differences (Table C3). Later it was determined that compatibility between *T. turgidum* and *T. longissimum* cytoplasm can be restored using chromosome 1D, which originated from *T. tauschii* (DD) (Maan, 1992a). Maan proposed that all *Triticum* species carry genes responsible for compatibility between the nucleus and cytoplasm (Maan, 1991 and 1992a). Sequence similarity between the mitochondrial genomes of *T. tauschii* and *T. longissimum* and the presence of nuclear encoded compatibility gene(s) is confirmed by the compatibility of the *T. aestivum* nucleus and *T. turgidum* nucleus with chromosome 1D when the *T. longissimum* cytoplasm is present.

    We used *de novo* genome assembly to analyze gene space identified in *T. aestivum* sequencing project. All genes recognized previously were found in both species, confirming content conservation of mitochondrial genomes. The majority of genes did not show nucleotide polymorphism. However, multiple differences were identified, such as major gene differences in *atp6* (Figure D4), *nad9* (Figure D6), *nad6* (Figure D7),

69

rps19-p (Figure D8) and *cob* genes in *T. tauschii*. The same polymorphism was reported in the alloplasmic line of *T. aestivum* (Liu *et al*., 2011) and (lo) durum line (Paper I). Overall, we observed a high conservation of gene content and a lack of polymorphism in the majority of mitochondrial genes across sequenced *Triticum* species, which agrees with the high conservation of mitochondrial function across all species (Douce, 1985). However, this is not the only kind of conservation observed. Our data suggests the major gene differences observed are widely present in the cytoplasm pool of *Triticum* when compared to other research (Liu *et al.,* 2011). These facts create a picture of a stable but diversified population of mitochondrial genomes in *Triticum* species. In our assembly of the *T. tauschii* genome, we did not find a second copy of *atp6* or the other subunits of *rrn26*, *18* and *5*. Final assembly of the *T. tauschii* genome carries a single copy of *atp6,* reflecting a lack of a possible duplication event in relation to *atp6* evolution. It was established that existence of two copies of *atp6* in mitochondrial genome in *Arabidopsis* was a result of duplication (Marienfeld *et al*., 1996). Another reason for the lack of a copy of that gene is the presence of separate plasmons caring that particular allele of a gene in separation to main the mitochondrial genome.

Triticum tauschii **chromosome 1D restorer of NC compatibility**

Sequencing of *T. turgidum* mitochondria provided us with an opportunity to describe the changes that arose in its new hexaploid background during 10,000 years of most recent wheat's evolution. We did not observe any dramatic changes in structure, gene content or polymorphism. Only 5 SNPs were identified in gene space and 40 SNPs in total between the two species. When homologous regions of *T. tauschii* and *T. aestivum* were analyzed, we identified 27 SNPs in gene space and 679 SNPs overall

difference. It is important to underline that 3 out of 5 SNPs were in *rps1* and *rps2,* and 10

out of 27 were present in ribosomal genes (*rps13, rpl5, rps1, rps2 and rps4*). In general,

the majority of polymorphism in gene space was in ribosomal coding genes. A very

similar observation was made by Liu *et al.,* 2011, where most of the polymorphism was

observed in the ribosomal gene of the alloplasmic line of *T. aestivum* with *Ae. crassa*

cytoplasm. That brings a possibility that ribosomal genes can be "hot spots" for

mutations, and somehow facilitate the adaptation of cytoplasm in a new nuclear

background. However, the role and possible outcome of these mutations would need to

be investigated. As mentioned before, *T. aestivum* inherited the cytoplasm from *T.*

turgidum. This information, with connection to a relatively low amount of polymorphism

between both species, indicates that the hybridization event and addition of a new

genome (DD) from *T. tauchii* did not create a great evolutionary pressure. Multiple

genomes in hexaploid wheat allowed *T. aestivum* to adapt relatively easily to a new

environment (Feldman *et al*., 1995) and a better response to the alloplasmic line creation

(Maan, 1991 and 1992a).

Heteroplasmy of mitochondrial genomes in relation to paternal leakage

During the polymorphism search, we were able to find differences in SNPs

between raw and final assembly (Table 4), which indicate that additional polymorphism

exists within mitochondrial gene space but was not detected using only *de novo*

assembly. We identified 4 alleles of *atp8,* but only two copies of that gene are present in

the final assembly of the *T. turgidum* (and *T. aestivum*) mitochondrial genomes. The SNP

pattern found in *T. tauschii* matches the two alleles in *T. turgidum*. Additionally, we

identified *atp6* (Figure D4), *nad9* (Figure D6), *nad6* (Figure D7) and *rps19*-p (Figure D8)

to have allele variations as identified in the raw assembly data of *T. tauschii* and *T. turgidum*. A gene space polymorphism search also showed that genes *nad3* and *rps1* have two alleles in *T. turgidum* and one of them is the same as *T. tauschii*. It is a common phenomenon that multiple mitotypes are recognized in plant species, as demonstrated in *Brassica napus,* where substochiometric shift was recognized as a reason for observed CMS in that species (Chen *et al*., 2011). Those results would be on par with results obtained by Hattori *et al.,* 2002 where both maternal and paternal mitotypes prevailed, depending on the line used as a cytoplasm donor. Bentley *et al*., 2010 also recognized parental mitotypes in *Silene vulgaris* representing 2.5% of mitochondria population. The low amount of mitochondrial DNA carrying paternal genotype would explain that: by only using raw *de novo* sequencing data it was possible to recognize paternal leakage. It is a recognized fact that in hybrids of wheat with *Aegilops* species, one particular plasmon is preferentially maintained in higher levels than other plasmons under different nuclear backgrounds (Tsukamoto *et al.,* 2000). Additionally, we were able to detect indications of both alleles of *atp6*-T and *atp6*-L in durum cytoplasm, together with other genes with major allelic differences between *T. tauschii* and *T. turgidum*. That information strongly suggests the presence of at least two major (possibly parental) mitotypes in *T. durum*. It would be a valuable resource to recheck the raw sequencing data from previous projects to find possible remaining mitotypes and establish paternal relationships between *Triticum* species, especially in case of *T. aestivum*; however, this can probably be achieved in the case of projects where high sequencing depth was obtained, since the amount of paternal DNA can be very low and possibly undetectable. Based on this data it is possible that the mitochondrial genome of *T. tauschii*, was

transferred and is currently present in the *T. aestivum* cytoplasm. Considering the fact that hexaploid wheat is a result of two hybridization events it should be possible to detect three major "parental" mitotypes, reflecting hybridization events, leading to hexaploid wheat.

Gene synteny between *T. aestivum* and its parents

Comparison of the gene order showed complete conservation between tetraploid and hexaploid wheat. During assembly analysis, it was found that there is a short region (about 1,200 bp) where no sequencing reads aligned against the reference sequence. An extension process of two contigs from *T. turgidum*, created by removal of the reference sequence of *T. aestivum*, allowed us to join those two contigs at opposite ends. Joining of contigs at the previous position was not possible. One possible reason is the highly polymorphic region between both genomes prohibiting the assembler software from creating the proper alignment. Comparison of the *T. tauschii* genome to other species showed multiple rearrangements. When gene order was compared to *T. tauschii* gene blocks, we were able to identify the particular genes within each block. Existence of gene blocks in mitochondria is a common phenomenon observed in many species and is related to the cotranscription of genes (Hoffmann *et al*., 1999). It was found that recombination points between some blocks are conserved in plant species, such as in potatoes (Susely *et al*., 2006). Existence of gene blocks, which are conserved across sequenced *Triticum* species, could indicate that the process of recombination is not random (gene order maintained) but rather controlled by some nuclear component.

Conclusions

The next generation sequencing method allowed us to have a close look at the mitochondrial population and analyze the evolutionary changes. Strong indication of paternal leakage and the heteroplasmy related to it was observed. Using similar methods, it would be possible to observe and track stoichiometric changes in mitochondrial composition, which would bring an answer on the role of heteroplasmy and its relation to male sterility and other phenotypic abnormalities observed in alloplasmic *Triticum* lines. Careful analysis and identification of mitotypes will allow proper identification of phylogenetic relationships and explain the evolution of the *Triticae* tribe. Detailed descriptions of mitochondrial diversity will assist in further explaining of biology of nuclear–cytoplasmic interactions.

CONCLUSIONS

To explain the phenotypic abnormalities observed in the alloplasmic durum line, we sequenced and assembled its mitochondrial genome along with its parents *T. longissimum* and *T. turgidum*. Additionally, we obtained the mitochondrial sequence from *T. tauschii*, one of the ancestral parents for *T. aestivum,* to confirm its evolutionary relationship.

Results showed multiple differences in genome structure, open reading frames and nucleotide polymorphism in the alloplasmic line when compared to parental species. Despite multiple genome rearrangements, gene synteny between species was identified as conserved gene blocks. Evidence of paternal leakage was also recognized. Multiple mitotypes based on single nucleotide polymorphism (in *T. longissimum*, *T. tauschii*, (lo) durum) and major gene differences are present in the *T. turgidum* cytoplasm. We confirmed that *T. turgidum* was the maternal donor of mitochondria to *T. aestivum*. Overall, the majority of the mitochondrial genes are highly conserved in all species analyzed; however, we observed number of variations in the alloplasmic line that likely formed as a result of its development.

Mitochondria in *Triticum* species create a picture of conserved gene content with few major gene differences. Our results allow functional analysis of gene space and formulate the basis for explaining some of the phenotypic variations observed in the alloplasmic lines. Identification of various cytoplasms (also particular mitotypes within organism or tissue type) could also allow development of strategies to select for favorable types during wheat hybrid creation. Analysis of the (lo) durum mitochondrial genome

indicated a complex series of rearrangements and variations where a single genetic variation cannot be established as the likely cause of abnormal phenotypes in this line. Looking at the overall conservation of mitochondrial genomes and the presence of multiple mitotypes, the most probable cause for an aberrant phenotype would be substoichiometric shift. Quantitative analysis of a mitochondrial population is necessary to establish this reason as underlying the phenotypic variation in the alloplasmic line. Here, we also propose a method to establish ancestral relationships among the *Triticeae* tribe based on the mitochondrial genome, using reference assembly.

REFERENCES

Abdelnoor RV, Yule R, Elo A, Christensen A, Meyer-Gauen G, Mackenzie S: **Substoichiometric Shifting in the Plant Mitochondrial Genome is influenced by a Gene Homologous to MutS.** *Proc. Natl Acad. Sci.* 2003 **100**(10): 5968-5973.

Adams KL, Palmer JD: **Evolution of mitochondrial gene content: gene loss and transfer to the nucleus.** *Molecular Phylogenetics and Evolution* 2003, **29**(3): 380-395.

Adams KL, Qiu YL, Stoutemyer M, Palmer JD: **Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution.** *Proc. Natl. Acad. Science* 2002, **99**(15): 9905-9912.

Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang C, Du F, Xu D, Gibson M, Cifrese J, Clifton SW, Newton KJ: **Comparison among two fertile and three male-sterile mitochondrial genomes of maize.** *Genetics* 2007, **177**: 1173-1192.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD: **Insights into the evolution of mitochondrial genome size from complete sequences of *citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae).** *Mol. Biol. Evol*. 2010, **27**(6): 1436-1448.

Andersson GE, Kurland CG: **Reductive evolution of resident genomes.** *Trends in microbiology* 1998, **6**(7): 263-268.

Arrieta-Montiel MP, Lyznik A, Woloszanska M, Janska H, Tohme H, Mackenzie SA: **Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean.** *Genetics* 2001, **158**(2): 851-864.

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA: **Diversity of the *Arabidopsis* mitochondrial genome occurs via nuclear-controlled recombination activity.** *Genetics* 2009, **183**(4): 1261-1268.

Avital G, Buchshtav M, Zhidkov I, Tuval Feder J, Dadon S, Rubin E, Glass D, Spector TD, Mishmar D: **Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins.** *Hum. Mol. Genet* 2012*, **21**(19): 4214-24

Baum M, Laugudah ES, Appels R: **Wide crosses in cereals**. *Annu. Rev. Plant Physiol. and Plant Mol.* 1992, **43:** 117-43.

Bentley KE, Mandel JR, McCauley DE: **Paternal leakage and heteroplasmy of mitochondrial genomes in** *Silene vulgaris***: evidence from experimental crosses.** *Genetics* 2010, **185**(3): 961-968.

Bentolila S, Stefanov S: **A Reevaluation of rice mitochondrial evolution based on the complete sequence of male-fertile and male-sterile mitochondrial genomes.** *Plant Physiology* 2012, **158**(2): 996-1017.

Bonen L, Bird S: **Sequence analysis of the wheat mitochondrial** *atp6* **gene reveals a fused upstream reading frame and markedly divergent N termini among plant ATP6 proteins.** *Gene* 1988, **71**(1): 47-56

Bonfield JK, Smith KF, Staden RA: **A new DNA sequence assembly program.** *Nucleic Acids Research* 1995, **23:** 4992-4999.

Boore JL: **Animal mitochondrial genomes.** *Nucleic Acids Research* 1999, **27**: 1767-1780.

Carver T, Berriman M, Tivey A, Patel A, Böhme A, Barrell BG, Parkhill J, Rajandream MA: **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database.** *Bioinformatics* 2008, **24**(23): 2672-2676.

Chan DC: **Mitochondria: dynamic organelles in disease, againg, and development.** *Cell* 2006, **125:** 1241-1252.

Chase CD: **Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions.** *Trends Genet* 2007, **23**(2): 81-90.

Chen J, Guan R, Chang S, Du T, Zhang H, Xing H: **Substoichiometrically different mitotypes coexist in mitochondrial genomes of** *Brassica napus* **L.** *PLoS One* 2011, **6**(3): e17662. doi:10.1371/journal.pone.0017662.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Research* 2004, **14**: 1147-1159.

Cho Y, Mower JP, Qiu Y-L, Palmer JD: **Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants.** *PNAS* 2004, **101**(51): 17741-17746.

Cui P, Liu H, Lin Q, Ding F, Zhuo G, Hu S, Liu D, Yang W, Zhan K, Zhang A, Yu J: **A complete mitochondrial genome of wheat (***Triticum aestivum* **cv. Chinese Yumai), and fast evolving mitochondrial genes in higher plants.** *Journal of Genetics* 2009, **88**(3): 299-307.

Darracq A, Varré JS, Maréchal-Drouard L, Courseaux A, Castric V, Saumitou-Laprade P, Oztas S, Lenoble P, Vacherie B, Barbe V, Touzet P: **Structural and content diversity of mitochondrial genome in beet: a comparative genomic analysis.** *Genome Biol. Evol.* 2011, **3**: 723-36.

Dieterich JH, Braun HP, Schmitz UK: **Alloplasmic male sterility in *Brassica napus* (CMS 'Tournefortii-Stiewe') is associated with a special gene arrangement around a novel *atp9* gene.** *Mol. Gen. Genomics* 2003, **269**: 723-732.

Douce R: **Mitochondria in higher plants: structure, function and biogenesis Orlando.** *Academic Press* 1985.

Dubcovsky J, Dvorak J: **Genome plasticity a key factor in the success of polyploid wheat under domestication.** *Science* 2007, **316:** 1862-1865.

Dubkina NV, Heinemeyer J, Sunderhaus S, Boekema EJ Braun HP: **Respiratory chain supercomplexes in the plant mitochondrial membrane.** *Thrends in Plant Science* 2006, **11**(5): 232-240.

Ducos E, Touzet P, Boutry M: **The male sterile G cytoplasm of wild beet displays modified mitochondrial respiratory complexes.** *The Plant Journal*. 2001, **26**(2): 171-180.

Elias EM. 1995. Durum Wheat Products. p. 23-31. **In Seminar on Durum wheat Quality in the Mediterranean Region**. CIHEAM/ICARDA/CIMMYT, Zaragoza, Spain. 17-19 Nov., 1993. Mediterranean Agronomic Institute.

Elson JL, Lightowlers RN: **Mitochondrial Heteroplasmy and Disease.** 2006. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net [doi: 10.1038/npg.els.0006079]

Fehrenbacher KL, Yang HC, Huckaba TM, Pon LA: **Live cell imaging of mitochondrial movement along actin cables in budding yeast.** *Curr Biol* 2004, **14**(22): 1996-2004.

Feldman M. **Wheats**. In: Smartt J, Simmonds NW, editors. **Evolution of crop plants**. Harlow, UK: Longman Scientific and Technical; 1995. p. 185-192.

Feldman M. 2001. **'The origin of cultivated wheat'**, in Bonjean, A. and Angus , W. (ed.) *The origin of cultivated wheat*. Paris: Lavoisier Publishing.

Gissi C, Iannelli F, Pesole G: **Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species.** *Heredity* 2008, **101**: 301-320.

Goloenko MI, Teljatnicova AA, Davydenko OG: **Some nuclei-cytoplasmic combinations of barley substituted lines collection.** *Barley Genetics Newletter* 2000, **30:** 28-31.

Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova KD, Nekrutenko A: **Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study.** *Genome Biology* 2011, **12:** R59-75

Gray MW: **Evolution of organellar genomes.** *Curr. Opin. Genet. Dev.* 1999, **9**(6): 678-687.

Hanson MR. **Plant mitochondrial mutations and male sterility.** *Annu. Rev. Genet.* 1991, **25**: 461–486.

Hanson MR, Bentolila S: **Interactions of mitochondrial and nuclear genes that affect male gametophyte development**. *The Plant Cell* 2004, **16**: 154-169.

Haouazine N, Takvorian A, Jubier MF, Michel F, Lejeune B. 1993. **The *nad6* gene and the exon d of *nad1* are co-transcribed in wheat mitochondria.** *Curr. Genet.* **24**(6): 533-538.

Hattori N, Kitagawa K, Takumi S, Nakamura C: **Mitochondrial DNA heteroplasmy in wheat, aegilops and their nucleus-cytoplasm hybrids.** *Genetics* 2002, **160:** 1619-1630.

Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J Millar AH: **Experimental analysis of the arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and Indicates plant-specific mitochondrial proteins.** *The Plant Cell* 2004, **16:** 241-256.

Heijne G: **Mitochondrial targeting sequences may form amphiphilic helices.** *The EMBO Journal* 1986, **5**(6): 1335-1342.

Herrmann JM: **Converting bacteria to organelles: evolution of mitochondrial protein sorting.** *Trends Microbiol.* 2003, **11**(2): 74-79.

Hey J: **What's So Hot about Recombination Hotspots?** *PloS Biol.* 2004, **2**(6) 0730-0733.

Hodgins KA, Rieseberg L, Otto SP: **Genetic control of invasive plants species using selfish genetic elements.** *Evolutionary Applications* 2009, **2**(4): 555-569.

Hoffmann M, Dombrowski D, Guha C, Binder S**: Cotranscription of the *rpl5-rps14*-cob *gene* cluster in pea mitochondria.** *Molecular and General Genetics* 1999, **261**(3): 537-545

Hosseini R: **The possible role of *atp6* gene in cytoplasmic male sterility in WA (Wild Abortive) type of rice (*Oryza sativa* L.).** *Iranian Journal of Biotechnology* 2010, **8**(2): 117-120.

Hu F: **The mediterranean diet and mortality - olive oil and beyond.** *New England Journal of Medicine* 2003, **348**(26): 2595-2596.

Iwabuchi M, Kyozuka J, Shimamoto K: **Processing followed by complete editing of an altered mitochondrial *atp6* RNA restores fertility of cytoplasmic male sterile rice.** *EMBO Journal* 1993, **12**(4): 1437-1446.

Kalavachara V, Hossain K, Maan SS, Kianian SF: **Nuclear cytoplasmic interactions in wheat.** *Recent. Res. Devel. Genet. Breeding* 2005, **2**: 71-84.

Kanazowa A, Tsutsumi A, Hirai A: **Reversible changes in the composition of the population of mtDNAs during.** *Genetics* 1994, **138**: 865-870.

Kawaura S, Saeki A, Masumara T, Morita S, Ogihara Y: **Heteroplasmy and expression of mitochondrial genes in alloplasmic and euplasmic wheat.** *Genes Genet Syst.* 2011, **86**(4): 249-55.

Kazama T, Toriyama K: **A pentatricopeptide repeat-containing gene that promotes the processing of aberrant *atp6* RNA of cytoplasmic male-sterile rice.** *FEBS Lett.* 2003, **544**(1-3): 99-102.

Kihara H: **Substitution of nucleus and its effects on genome manifestations.** *Cytologia* 1951, **16:** 177–193.

Köhler RH, Horn R, Lossl A, Zetsche K: **Cytoplasmic male sterility in sunflower is correlated with the co-transcription of a new open reading in sunflower is correlated with the co-transcription of a new open reading.** *Mol. Gen. Genet.* 1991, **227:** 369-376.

Leeuwen TV, Vanholme B, Pottelberge S, Nieuwenhuyse P, Nauen R, Tirry L, Denholm I: **Mitochondrial heteroplasmy and the evolution of insecticide resistance: Non-Mendelian inheritance in action.** *PNAS* 2008, **105**(16): 5980-5985.

Leister D: **Origin, evolution and genetic effects of nuclear insertions of organelle DNA**. *Trends Genet* 2005, **21**: 655–663.

Leon P, Arroyo A, Mackenzie SA: **Nuclear control of plasmid and mitochondrial development in higher plants.** *Annu. Rev. Plant. Physiol. Plant. Mol. Biol.* 1998, **49:** 453-480.

Lough AN, Roark LM, Kato A, Ream TS, Lamb JC, Birchler JA, Newton KJ: **Mitochondrial DNA transfer to the nucleus generates extensive insertion.** *Genetics* 2008, **178:** 47-55.

Liu CG, Wu YW, Hou H, Zhang C, Zhang Y, McIntosh RA: **Value and utilization of alloplasmic common wheats with Aegilops crassa cytoplasm.** *Plant Breeding.* 2002, **121**(5): 407-410.

Liu H, Cui P, Zhan K, Lin Q, Zhuo G, Guo X, Ding F, Yang W, Liu D, Hu S, Yu J, Zhang A: **Comparative analysis of mitochondrial genomes between a wheat K-type cytoplasmic male sterility (CMS) line and its maintainer line.** BMC Genomics 2011, **12:** 163-177

Logan DC: **Mitochondrial fusion, division and positioning in plants.** *Biochem. Soc. Trans.* 2010, **38**(3): 789-795.

Maan SS. 1975. **Cytoplasmic variability and speciation in *Triticinae*.** pp. 255-281. In Prairie: A Multiple View, edited by M.K. Wali. University of North Dakota Press, Grand Forks, ND.

Maan SS: **Genetic analyses of male fertility restorer in wheat I. Chromosome location of *Rf* genes.** *Crop Science* 1984, **24:** 17-20.

Maan SS: 1991. **Nucleo-cytoplasmic genetics of wheat**, pp. 175–194 in International Symposium on Nuclear and Organellar Genetics of Wheat, edited by T. Sasakuma and T. Kinoshita. Hokkaido University, Sapporo, Japan.

Maan SS: **Transfer of the species specific cytoplasm (scs) from *Triticum*.** *Genome* 1992a, **35:** 238-243.

Maan SS: **The *scs* and *Vi* genes correct a syndrome of cytoplasmic effects in alloplasmic durum wheat**. *Genome* 1992c, **35:** 772-779.

Makaroff CS, Apel JI, Palmer JD: **The *atp6* coding region has been disrupted and a novel Reading frame generated in the mitochondrial genome of cytoplasmic male-sterile radish.** *The J. of Biol. Chem.* 1989, **264**(20): 11706-11713

Manchekar M, Scissum-Gunn K, Song D, Khazi F, McLean SL, Nielsen BL: **DNA recombination activity in soybean mitochondria.** *J. Mol. Biol.* 2006, **356**(2): 288-299.

Marienfeld J, Useld M, Brandt P, Brennicke A: **Genomic recombination of the mitochondrial *atp6* gene in *Arabidopsis thaliana* at the protein processing site creates two different presequences.** *DNA Research.* 1996, **3**(5): 287-290

Martin W, Herrmann RG: **Gene transfer from organelles to the nucleus: how much, what happens, and why?** *Plant Physiol.* 1998, **118:** 9-17.

Mbantenkhu M, Wang X, Nardozzi JD, Wilkens S, Hoffman E, Patel A, Cosgrove MS Chen XJ: **Mgm101 is a rad52-related protein required for mitochondrial DNA.** *J. Biol. Chem.* 2011, **286**(49): 42360-70.

McClintock B: **The significance of responses of the genome to challenge.** *Physiology of Medicine.* 1983, **226**(4676): 792-801.

McFadden ES, Sears ER: **The origin of *Triticum* spelta and its free-threshing hexaploid relatives.** *J Hered* 1946, **37:** 81–89

Millar HA, Heazlewood JL, Kristensen BK, Braun HP, Møller IM: **The plant mitochondrial proteome.** *Trends in Plant Science* 2005, **10**(1): 36-43.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet - next generation sequence assembly visualization.** *Bioinformatics* 2010, **26**(3): 401-402.

Mohr S, Schulte-Kappert E, Odenbach W, Oettler G, Kück U: **Mitochondrial DNA of cytoplasmic male-sterile *Triticum timopheevi*: rearrangement of upstream sequences of the *atp6* and *orf25* genes.** *Theor. Appl. Genet.* 1993, **86:** 259-268.

Nevo E, Chen G: **Drought and salt tolerances in wild relatives for wheat and barley improvement.** *Plant, Cell and Environment* 2010, **33**(3): 670-685

Ogihara Y, Yamazaki Y, Murai K, Kanno A, Terachi T, Shiina T, Miyashita N, Nasuda S, Nakamura C, Mori N, Takumi S, Murata M, Futo S, Tsunewaki K: **Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome.** *Nucleic Acids Research* 2005, **33**(19): 6235-6250.

Okamoto K, Shaw JM: **Mitochondrial morphology and dynamics in yeast and multicellular eukaryotes.** *Annu. Rev. Genet.* 2005, **39:** 503-536.

Palmer JD, Herbon LA: **Plant mitochondrial DNA evolves rapidly in structure but slowly in sequence.** *J. Mol. Evol.* 1988, **28:** 87-97.

Parkinson CL, Mower JP, Qiu YL, Shirk AJ, Song K, Young ND, dePamphilis CW Palmer JD: **Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae.** *BMC Evol. Biol.* 2005, **5**(1): 73-85.

Paumard P, Vaillier J, Coulary B, Schaeffer J, Soubannier V, Mueller DM, Brèthes D, di Rago JP, Velours J: **The ATP synthase is involved in generating mitochondrial cristae morphology.** *EMBO Journal* 2002, **21**(3): 221-230.

Reader SM, Miller TE: **The introduction into bread wheat of a major gene for resistance to powdery mildew from wild emmer wheat.** *Euphytica* 1991, **53:** 57–60

Rozen S. Skaletsky HJ: 2000. 'Primer3 on the WWW for General Users and for Biologist Programmers', in Krewetz, S. and Misener, S. (ed.) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, Totowa, NJ: Humana Press.

Rychlik W: **OLIGO 7 primer analysis software.** *Methods in Mol Biol* 2007, **402:** 35-60.

Sambrook J, Fritsch EF Maniatis T. 1989. **Molecular Cloning: A Laboratory Manual**, NY: Cold Spring Harbor Laboratory Press.

Sandoval P, León G, Gómez I, Carmona R, Figueroa P, Holuigue L, Araya A, Jordana X: **Transfer of RPS14 and RPL5 from the mitochondrion to the nucleus in grasses.** *Gene* 2004, **324:** 139-147.

Schneider G, Fecher U: **Advances in the prediction of protein targeting signals.** *Proteomics* 2004, **4**(6): 1571-1580.

Scotti N, Marechal-Drouard L, Cardi T: **The *rpl5-rps14* mitochondrial region: a hot spot for DNA rearrangements in *Solanum* spp. somatic hybrids.** *Current Genetics* 2004, **45:** 378-382.

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA: **Plant mitochondrial recombination surveillance requires unusual *RecA* and *MutS* homologs.** *Plant Cell* 2007, **19:** 1251–1264

Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat. Biotechnol* 2008, **26:** 1135-1145.

Shewry PR: **Wheat**. *J. Exp. Bot.* 2009, **60**(6): 1537-1553.

Shinada T, Kikuchi Y, Fujimoto R, Kishitani S: **An alloplasmic male-sterile line of Brassica oleracea harboring the mitochondria from *Diplotaxis muralis* expresses a novel chimeric open reading frame, *orf72*.** *Plant Cell Physiol.* **47**(4): 549-53

Siedow JN, Umbach AL: **Plant mitchondrial electron transfer and molecular biology.** *The Plant Cell* 1995, **7:** 821-831.

Sirrenberg C, Endres M, Folsch H, Stuart RA, Neupert W, Brunner M: **Carrier protein import into mitochondria mediated by the intermembrane proteins Tim10/Mrs11 and Tim12/Mrs5.** *Nature* 1998, **391**(6670): 912–915.

Small JD, Isaac PG, Leaver CJ: **Stoichiometric differences in DNA molecules containing the *atpA* gene suggest mechanisms for the generation of mitochondrial genome diversity in maize.** *The EMBO Journal* 1987, **6**(4): 865-869.

Smith BD: **Documenting plant domestication: The consilience of biological and archaeological approaches.** *PNAS*. 2001, **98**(4): 1324-1326.

Soltani A, Ghavami F, Noyszewski AK, Anantharama R, Meingardt S, Kianian SF: **Overexpression of two mitochondrial orfs can lead to cytoplasmic male sterility in alloplasmic durum wheat.** Poster presentation 2011, ASBP.

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M: **The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants.** *Mol. Genet. Genomics* 2005, **272**(6): 603-615.

Susely F, Tada S, Souza AP: **A recombination point is conserved in the mitochondrial genome of higher plant species and located downstream from the *cox2* pseudogene in *Solanum tuberosum* L**. *Genet. and Molecular Biol.* 2006, **29**(1): 83-89

Tao D, Xu P, Zhou J, Deng X, Li J, Deng W: **Cytoplasm affects grain weight and filled-grain ratio in indica rice.** *BMC Genetics* 2011, **12:** 53-60.

Triboush SO, Danilenko NG, Davydenko OG: **A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower.** *Plant Mol. Biol. Rptr.* 1998, **16:** 183–189.

Tsukamoto N, Askura N, Hattori N, Takumi K, Mori N: **Identification of paternal mitochondrial DNA sequences in the nucleus-cytoplasm hybrids of tetraploid and hexaploid wheat with D and D2 plasmons from *Aegilops* species**. *Curr. Genet.* **38**: 208-217

Tsunewaki K, 1988: **Genetic diversity of the cytoplasm in *Triticum* and Aegilops**. In: T. E. Miller, and R. M. D. Koebner (eds), Proc. 7th Int. Wheat Genet Symp., p. 35-62. Bath Press, Bath.

Tsunewaki K: **Genome-plasmon interactions in wheat.** *Jpn. J. Genet.* 1993, **68:** 1-34.

Tsunewaki K, Wang GZ, Matsuoka Y: **Plasmon analysis of *Triticum* (wheat) and *Aegilops*. 1. Production of alloplasmic wheats and their fertilities.** *Genes Genet. Syst.* 1996, **71:** 293-311.

Tsunewaki K, Shimada T, Matsuoka Y: **Transfer of *Triticum urartu* cytoplasm to emmer wheat is difficult, if not impossible.** *Wheat. Information* 1999, **88:** 27-31.

Tsunewaki K, Wang GZ, Matsuoka Y: **1. Plasmon analysis of *Triticum* (wheat) and *Aegilops*. 2. Characterisation and classification of 47 plasmons based on their effects on common wheat phenotype.** *Genes Genet. Syst.* 2002, **77:** 409-427.

Turmel M, Otis C, Lemieux C: **The complete mitochondrial DNA sequence of Mesostigma viride identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants.** *Mol. Biol. Evol.* 2002, **19**(1): 24-38.

Ullrich H, Lättig K, Brennicke A, Knoop V: **Mitochondrial DNA variations and nuclear RFLPs reflect different genetic similarities among 23 Arabidopsis thaliana ecotypes.** *Plant Mol. Biol.* 1997, **33**(1): 37-45.

Vitart V, de Paepe R, Mathieu C, Chetrit P, Vedel F: **Sequences in a nuclear, male sterile mutant regeneration from protoplast culture in *Nicotiana sylvestris*.** *Molec. and General Genetics* 1992, **233:** 193-200.

Wang DY, Zhang Q, Liu Y, Lin ZF, Zhang SX, Sun MX: **The levels of male gametic mitochondrial DNA are highly regulated in angiosperms with regard to mitochondrial inheritance.** *The Plant Cell* 2010, **22**(7): 2402-2416.

Weider C, Stamp P, Christov N, Husken A, Foueillassar X, Camp KH, Munsch M: **Stability of cytoplasmic male sterility in maize under different environmental conditions.** *Crop Sci.* 2009, **49:** 77-84.

Weihe A, Aptiz J, Pohlheim F, Salinas-Hartwig A, Börner T: **Biparental inheritance of plastidial and mitochondrial DNA and hybrid variegation in Pelargonium**. *Mol Genet Genomics* 2009, **282**(6): 587-593.

Welch ME, Danell MZ, McCauleyL DE: **Variable populations within variable populations: quantifying mitochondrial heteroplasmy in natural populations of the gynodioecious plant *Silene vulgaris*.** *Genetics* 2006, **174**(2): 829-837.

Wolfe KH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs**. *Proc. Natl. Acad. Sci.* 1987, **84**(24): 9054-9058.

Woloszynska M: **Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes - though this be madness, yet there's method in't.** *J. Exp. Bot.* 2010, **61**(3): 657-671.

Wu Y, Zhand C, Liu C, Shuxin R, Yan Z: **Breeding technology of alloplasmic wheat.** *Science in China Series C: Life Sciences* 1998, **41**(5): 449-458.

Zaegel V, Guermann B, LeRet M, Andres C, Meyer D, Erhardt M, Canaday J, Gualberto JM, Imbault P: **The plant-specific ssDNA binding protein OSB1 is involved in the stoichiometric transmission of mitochondrial DNA in Arabidopsis.** *The Plant Cell* 2006, **18:** 3548–3563.

Zick M, Rabi R, Reichert AS. **Cristae formation-linking ultrastructure and function of mitochondria.** *Biochim. Biophys. Acta.* 2009, **1793**(1): 5-19.

APPENDIX A. MITOCHONDRIAL GENOME REFERENCE ASSEMBLY AS A TOOL TO STUDY PHYLOGENETIC RELATIONSHIPS – PROPOSAL

Here I would like to propose the use of mitochondrial genome reference assembly as an easy and relatively fast method to establish phylogenetic relationships between closely related species. This proposal is based on data and experience gained from sequencing four mitochondrial genomes from the *Triticeae* tribe, that is *T. turgidum*, *T. tauschii*, *T. longissimum*, (lo) durum and also the previously sequenced *T. aestivum* mitochondrial genome (Ogihara *et al*. 2004). For details about sequence data and assembly process please refer to the Material and Methods chapter, Paper I and II of this dissertation.

There are two main ways to perform genome assembly, which can be applied to various kinds of sequencing techniques: *de novo* and reference assembly. A *de novo* assembly uses sequencing reads (read length depends on the sequencing technology used) to create full-length sequences, based only on sequence information provided by the reads itself (including one-end or pair-end information). A reference assembly is based on the assembly of sequencing reads against a previously established "backbone sequence". The backbone sequence guides an assembly process. These two approaches can differ in obtained results: *de novo* assembly allows the creation of novel genomic sequences for which a sequence was not previously known. Reference assembly creates the possibility to build genomes, which are similar, but carry some level of polymorphism (most often polymorphism recognized as single nucleotide polymorphism). It is relatively easy and fast to obtain results from mapping assembly,

since the alignment of sequencing reads is based on sequence information already given by the references sequence used for assembly.

Multiple proposals of *de novo* sequencing of thousands genomes, such as projects like the i5k initiative (Arthropod Sequencing Initiative) or the project announced in Nature (The 1000 Genome Project Consortium), where the goal is to sequence 1092 of human genomes (BGI-Shenzhen, China), will rely on only *de novo* assembly as a method.

Why is *de novo* assembly valuable even in the case of species whose genomes are assembled? The main reason for that is to properly describe nucleotide and sequence variation, which cannot be detected using simple reference assembly. This variation can certainly be lost during sequencing projects which try to describe population of genomes which are highly diversified, such as mitochondrial population. Here, simple reference assembly can only provide information about one of many mitotypes present in a cell. Structural and nucleotide polymorphism present between mitotypes will not be detected until proper *de novo* assembly is performed.

However, *de novo* assembly is very challenging to the biologist from multiple perspectives. These problems include: size of genomes, complexity (e.g. number and length of repeated regions), sequencing errors, length of sequencing reads, quality assessment of finished assemblies, computational resources of particular projects and access to a highly skilled software designer who could address project-specific needs and integrate that information when particular problems arise. In general, the amount of data and ease with which we can obtain sequencing reads is not connected with ease we can assemble that data and actually obtain biologically relevant information. Proposed here is

a new use of reference assembly, with the use of a backbone sequence to guide the search for new genomic information, which would allow much easier recognition of any polymorphism found in new analyzed sequences. The differences which would be found during the process would give us a possibility to describe relative distance between analyzed species.

Phylogeny is the study of relationships between genes, either proteins or whole organisms, which are derived from common ancestor. In other words, phylogenetic studies allow researchers to establish evolutionary relationships. The most common way to establish this kind of relationship is to build multiple alignments of nucleotide/protein gene sequences between various species. These are most often based on a single gene sequence of conserved, non-protein coding genes such as parts of ribosomal proteins. Based on the multi-alignment distance between all sequences established, phylogenetic tree depicting relationships between species is drawn.

A combination of reference assembly and phylogenetic studies can be made. That information would be biologically relevant and important and would allow researcher to establish phylogenetic relationships between closely related species based on a few fully sequenced mitochondrial genomes, without the tedious and expensive process of building hundreds of mitochondrial genomes *de novo*. It is known that mitochondria are highly conserved, but their structure is variable between even closely related species or even within a particular cell, which makes *de novo* assembly very difficult. In our studies we obtained mitochondrial genomic DNA from four *Triticum* species. The first step in our genome analysis was to perform the mapping assembly, against the *T. aestivum* mitochondrial genome; we also used negative control from *Tripsacum dactyloides* to

demonstrate how assembly will perform based on the backbone from a species distant to

Triticum. The Figure A2 represents the results of that assembly. Analysis of the results

from genome assembly allowed us to immediately recognize that *T. durum* and *T.*

aestivum share highly similar mitochondrial genomes, with the exception of the region

designated as E (Figure A2). Additionally, we can see that *T. tauschii* and *T. dactyloides*

do not share similarity almost in any region across the genome. Results from *T.*

longissimum and *T. tauschii* shows that these genomes share similar regions to *T.*

aestivum and also that they are very similar to each other. For the purpose of that

discussion, on Figure A2 regions A-D were designated, and show places without read

coverage. Approximate size of these regions is very similar between both species.

Additionally, we also described in similar fashion the (lo) durum line. In this case, a lack

of regions C and D is observed and size difference between the region marked as A.

Region B seems to be present in all three species. Summary of those results are in Table

A1.

Table A1. Summary of the mapping assembly results of all four sequenced species. Mapping assembly performed based on *T. aestivum*. Letters designate regions of no sequence read coverage. Minus and plus sign mark the absence or presence of given region. Dollar sign designates the region with size change in relation to other species.

Specie	A	B	C	D
T. aestivum	-	-	-	-
T. turgidum	-	-	-	-
T. longissimum	+	+	+	+
(lo) durum	+$	+	-	-
T. tauschii	+	+	+	+

Figure A1 shows an example of data that is gathered after performing reference assembly. Detection of SNPs within the regions similar to both species would be possible, allowing an even more precise description of recognized polymorphism.

```
1 ATGAC
2 ATAACTAGT
3 ATAACTAGTAATTTGCATGGTAGGATACAT                                              CATCAGTT
4 ATGACTAGTAAATTGCATGGTAGGATACATACCCA                                         CATCAGTT
5 ATAACTAGTAATTTGCATG                                               GATAGGCCATCAGTT
6 ATGACTAGTAAATTGCA                                      TCCAGTACATCTGATAGTCCATCAGTT
7 ATGACTAGTAATTTGCATGGTAGGATGCATACCCATTA              GTACATCTGATAGTCCATCAGTT
8 ATAACTAGTAAATTGCATGGTAGG                         CAGTACATCTGATAGGCCATCAGTT
9 ATGACTAGTAATTTGCATGGTAGGATACATACC                         CTGATAGTCCATCAGTT
B ATAACTAGTAAATTGCATGGTAGGATACATACCCATTACCTTACCATTGGCATATACATTTCCAGTACATCTGATAGGCCATCAGTT
S ++-++++++++-+++++++++++++++++###################################################++++++
```

Figure A1. An ideogram representing the search algorithm used for detection of "gaps" in mapping assembly. Picture shows fragment of mapping assembly against backbone sequence. Gap (labeled as green pound sign) and three SNPs (labeled with minus sign red) were found; coverage for each nucleotide position also will be established. Numbers are assigned to each sequencing read; B stands for backbone sequence; S is a score in relation to SNP and gap detection.

It's important to note that observed similarities between described species, were also confirmed by a more in-depth analysis of gene comparison, ORF search and SNP analysis, which showed the same relationships between all five species as described in the mapping assembly analysis. Accuracy confirmation of the described method was done with a mapping assembly of the *T. longissimum* genome, with use of backbone from *T. tauschii*, where both genomes share identical genome structure and content. Based on those results it would be possible to develop software, with use of the described algorithm, which would allow to precisely identify regions of coverage/no coverage, describe size differences between species, possible coverage variation for each region, and single polymorphism differences within similar regions. Collection of that information would allow the creation of phylogenetic trees, which in the end give the possibility to establish evolutionary distances within that species. This algorithm would

create a great tool to assess the diversity of mitochondrial genomes, based on a single reference genome, without the need of difficult and unpredictable *de novo* assembly, additionally lowering cost of project. At the same time, this technique would provide a robust method to describe relationships between cytoplasms within e.g. *Triticum* tribe, greatly improving studies of cytoplasm diversity.
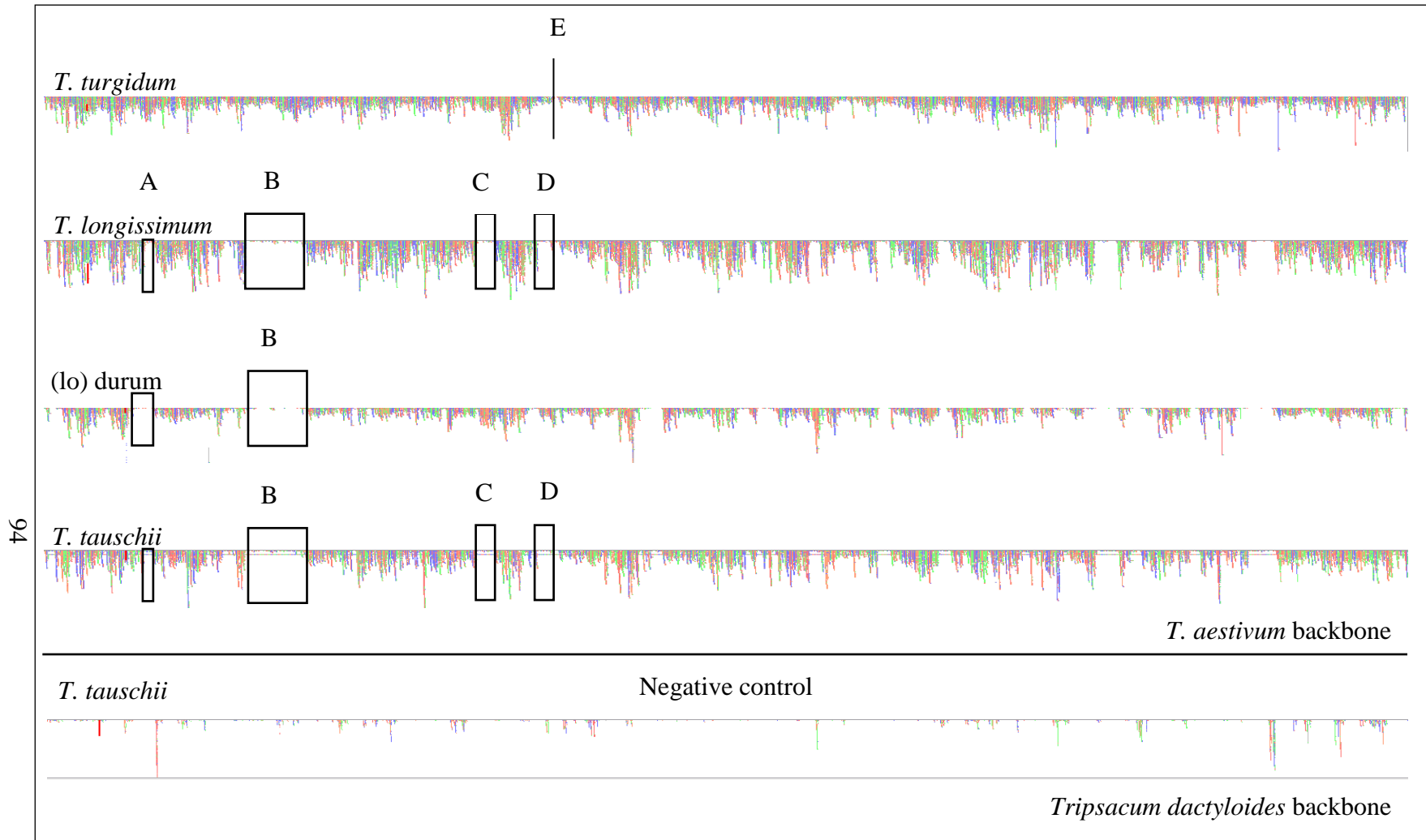
Figure A2. Graphical representation of mapping assembly between four *Triticum* species. Species compared are *T. turgidum*, *T. longissimum*, (lo) durum, *T. tauschii* using the *T. aestivum* mitochondrial genome as a backbone sequence. For the control of the assembly process we used *Tripsacum dactyloides* as a backbone sequence and assemble the *T. tauschii* sequencing reads against it. Colorful picks across genomes represent sequencing reads assembled in that particular region, or regions of no reads coverage (examples depicted by black boxes). Additionally, differences between genomes are visible in coverage pattern, showing highly conserved regions of mitochondrial genomes. Picture not normalized, and can't be used for exact quantitative comparison. Letter (A-E) designate no coverage regions

APPENDIX B. DEVELOPMENT OF THE ASSEMBLY WORKSTATION

Constantly improving the next generation DNA sequencing (NGS) methods, also known as deep sequencing, brings more genomic data than ever before, providing great opportunity to accelerate biological research (Shendure and Ji, 2008). NGS technologies are highly dependent on the computer processing unit (CPU), available random access memory (RAM) and hard disk drive (HDD) storage capacity to properly process and describe large amounts of raw sequencing data. Additionally, extensive post processing is required, which also is highly demanding. Processes such as contig editing, joining, alignment and other related data manipulation creates a need for appropriate hardware.

At the time we obtained mitochondrial sequencing data, there was no possibility to use the computers available in our laboratory or any of the multicore computer clusters available at the North Dakota State University (NDSU) Campus, because of hardware requirements of current assembly software. Reasons of that was insufficient CPU processing power, capacity of RAM per CPU, and assembly software that was not optimized to utilize multicore CPUs. In this case, a 96-node parallel cluster available at NDSU, provided a speed of assembly as fast as one of the clusters nodes (CPU). Additionally, each of nodes didn't have sufficient RAM capabilities where 56 nodes had 1 gigabyte, 8 nodes had 6 gigabytes and 32 nodes had 2 gigabytes of RAM.

We decided to build a computer based on Intel i7 950 CPU (3.04 GHz) four core processor with Hyper-Threading technology (Figure B1). That particular platform allowed us to have up to 24 GB of RAM per CPU, however for this project we used only 12 GB of RAM, which was sufficient to assemble acquired sequencing data, and also perform tasks connected with data editing and annotation. We were able to run *de novo*

assembly in about 8 hours, mapping assembly in about 4 hours (depending on assembly

parameters and sample) using MIRA assembly software. The multicore structure of the

processor and a large amount of RAM memory allowed us to run assembly for three

species at the same time, greatly improving time for checking various assembly

parameters, and to choose optimal assembly conditions.

The assembled computer will be able to facilitate future assembly projects.

Operating systems and main software used during assembly and post processing:

Windows 7 64- bit, Ubuntu 64-bit, MIRA, Staden Package, Tablet software.

Figure B1. Computer workstation built for mitochondrial genome assembly. Computer was capable of running three mitochondrial assemblies simultaneously. General specification: Intel i7 950 (3.06GHz), 12 GB (6 x 2GB) DDR3 RAM Triple Chanel, 640GB HDD.

Table C1. PCR primers for DNA quantification and *atp6* gene sequence confirmation. These primers were used to establish the level of chloroplast and nuclear DNA contamination in the mitochondrial DNA samples and to differentiate between alleles of *atp6* gene.

Genome (gene)	Primer name	Sequence 5' - 3'	Amplicon size (bp)	Primer source
Chloroplast (*psb60*)	psbB-60F	ATGGGTTTGCCTTGGTATCGTGTTCATAC	355	Heinze B. 2007
	psbB-61R	TCCCAATAYACCCAATGCCAGATAG		
Mitochondria (*nad3*)	nad3-F	GCACCCCTTTTCCATTCATA	357	This dissertation
	nad3-R	TCGGAATTTGCACCTATTTG		
Nucleus (retro-junction)	cfp1589	GTTTCGTTCATGCGATATGTGCC	321	Paux E. e*t al.* 2008
	cfp1589	ATCTCTACCCCCCGCCCC		
atp6	atp6LO-F	TCCGCAGCTTGTGAACCGGC	843	This dissertation
	atp6LO-R	GCAGTGGGACTCCCGCAGGTA		
	atp6TU-F	AACGCCAACGCCTGAGCGAA	642	This dissertation
	atp6TU-R	GCAGTGGGACTCCCGCAGGTA		

Table C2. Final assembly results for mitochondrial genomes of various *Triticum* species

Species	Coverage (x)	Genome size (bp)	Read length (bp)	# of contigs
T. tauschii	116	399,374	468	2
T. turgidum	114	451,925	450	1
T. longissimum	61	399,074	442	2
(lo) durum	133	432,606	436	1

Table C3. Number of single nucleotide variations (SNVs) and SNV blocks found within a particular species

Species	SNP*	SNV density (1SNV/bp)	SNV block (#SNV)$	1-10x#
T. turgidum	93	1/4,859	22 (151)	0.3%(1,283 bp)
(lo) durum	61	1/7,091	15 (155)	0.9%(3,903 bp)
T. longissimum	98	1/6,542	27 (246)	1%(3,962 bp)
T. tauschii	52	1/4,294	20(171)	0.3%(1,186 bp)

*SNV is single nucleotide variation found in a particular position, only nucleotide changes were considered (insertion/deletions were excluded from analysis).

$Block describes genomic regions with an increased density of SVNs (with distance between single SNVs lower than 15 bp).

#Percent of regions with coverage of 1-10x in the assembled genomes

Table C4. Nucleotide variation(s) in ribosomal protein coding genes. Listed are the nucleotide changes found in (lo) durum, *T. longissimum* and *T. tauschii* with *T. turgidum* as the backbone sequence. The SNP's characteristic for alloplasmic line is underlined.

		T. turgidum	*T. longissimum*	*T. tauschii*	(lo) durum	Amino acid change
Ribosomal proteins	*rps1-1**	+	-	-	-	-
	rps1-2	-	C/A^{33}, C/T^{397}	C/A^{33}, C/T^{397}	C/A^{33}, C/T^{397}	no change, T/I
	rps2-1	+	+	+	-	Y/F
	rps2-2	-	-	-	$\underline{A/T^{233}}$	
	rps4-1	+	-	-	-	-
	rps4-2	-	G/T^{146}, T/G^{236}	G/T^{146}, T/G^{236}	-	R/L, L/R
	rps4-3	-	-	-	G/T^{146}, T/G^{236}, $\underline{C/A^{495}}$	R/L, L/R + no change
	rps13-1	+	-	-	-	-
	rps13-2	-	A/C^{45}	A/C^{45}	-	no change
	rps13-3	-	-	-	A/C^{45}, $\underline{A/C^{170}}$	no change, E/A
	rpl5-1^	+	-	-	-	-
	rpl5-2	-	G/T^{28}	G/T^{28}	G/T^{28}	
rRNA	*His tRNA-1*	+	-	-	-	-
	His tRNA-2	-	C/T^{10}, T/C^{22}	C/T^{10}, T/C^{22}	C/T^{10}, T/C^{22}	not applicable
Variation^&		0	6	6	9	

Arabic numbers indicate the SNP position relative to the start codon
+/- Indicates the presence/absence of gene allele in a particular genome
* Gene with multiple alleles. Reported here are the most abundant alleles
^& Summary of nucleotide variation observed only in genes with one known allele of that gene within a species

Table C5. Nucleotide variation(s) within the coding genes of the electron transport complexes. Listed are the nucleotide changes found in (lo) durum, *T. longissimum* and *T. tauschii* with *T. turgidum* as the backbone sequence. The SNP's characteristic for each line are underlined.

	Gene	*T. turgidum*	*T. longissimum*	*T. tauschii*	(lo) durum	Amino acid change
Complex I	*nad3-1**	+	-	-	-	-
	nad3-2	-	T/C ^{185}	T/C ^{185}	T/C ^{185}	L/P
Complex IV	*cox3-1*	+	-	-	-	-
	cox3-2	-	A/C ^{157}, <u>GA/TC ^{687}</u>	A/C ^{157}, <u>GA/TC ^{687}</u>	-	I/L, QM/HL
	cox3-3	-	-	-	A/C ^{157}	I/L
Complex V	*atp1-1*	+	-	-	-	-
	atp1-2	-	A/T ^{1431}	A/T ^{1431}	A/T ^{1431}	no change
	atp8-I-1*	+	-	-	-	-
	atp8-I-2	-	C/A ^{239}	C/A ^{239}	-	
	atp8-I-3	-	-	-	C/T ^{39}, A/G ^{149}, G/A ^{166}	
	atp8-II-1*	+	-	-	-	-
	atp8-II-2	-	C/A ^{239}, G/A ^{463}, T/C ^{467}	C/A ^{239}, G/A ^{463}, T/C ^{467}	-	
	atp8-II-3	-	-	-	G/A ^{462}, T/C ^{467}	
Other proteins	*mttB-1*	+	+	+	-	-
	mttB-2	-	-	-	<u>T/G ^{41}</u>	L/W
Variations$^{\&}$		0	4	4	3	

Arabic numbers indicate the SNP position relative to the start codon
+/- Indicates the presence/absence of gene allele in a particular genome
* Gene with multiple alleles. Reported here are only the most abundant alleles
$^{\&}$ Summary of nucleotide variation observed only in genes with one known allele of that gene within species

Table C6. Characteristics of the *de novo* assemblies used for gene space description

Name of assembly run	Species	Minimum coverage	# of best contigs selected	Largest contig (bp)	N50 (bp)	Total consensus (bp)	Total average coverage (x)
Run10Bastien3							
	T. turgidum	18	221	32,888	7,063	978,532	59
	(lo) durum	39	70	36,597	13,476	482,735	136
	T. longissimum	3	2,409	26,024	4,752	6,226,807	9
	T. tauschii	33	105	36,496	9,732	519,397	100
Run4							
	T. turgidum	29	179	31,874	8,303	767,867	90
	(lo) durum	37	49	44,554	24,461	431,313	142
	T. longissimum	3	2,419	27,701	4,949	6,212,505	9
	T. tauschii	36	63	31,694	17,841	435,711	112
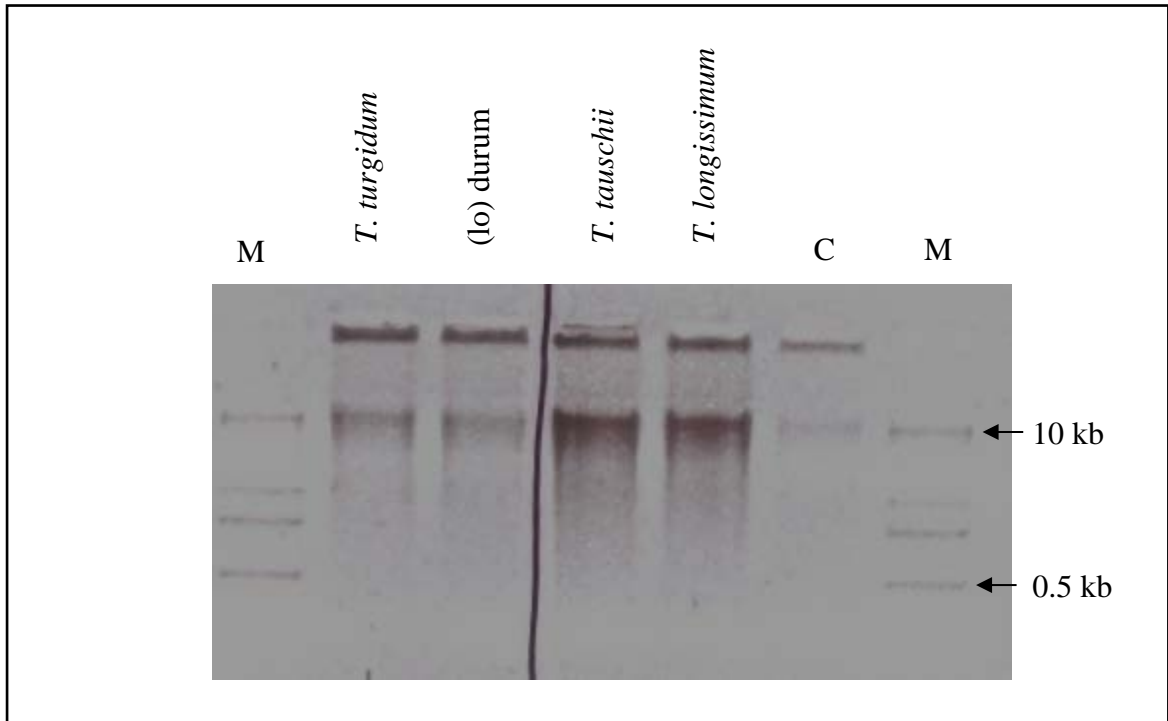
Figure D1. Mitochondrial DNA amplification, gel separation. MtDNA was amplified using Rolling Circle Amplification (RCA) with random hexamer primers and separated in a 1% agarose gel. M-size markers, C-Lambda DNA was used as amplification control.
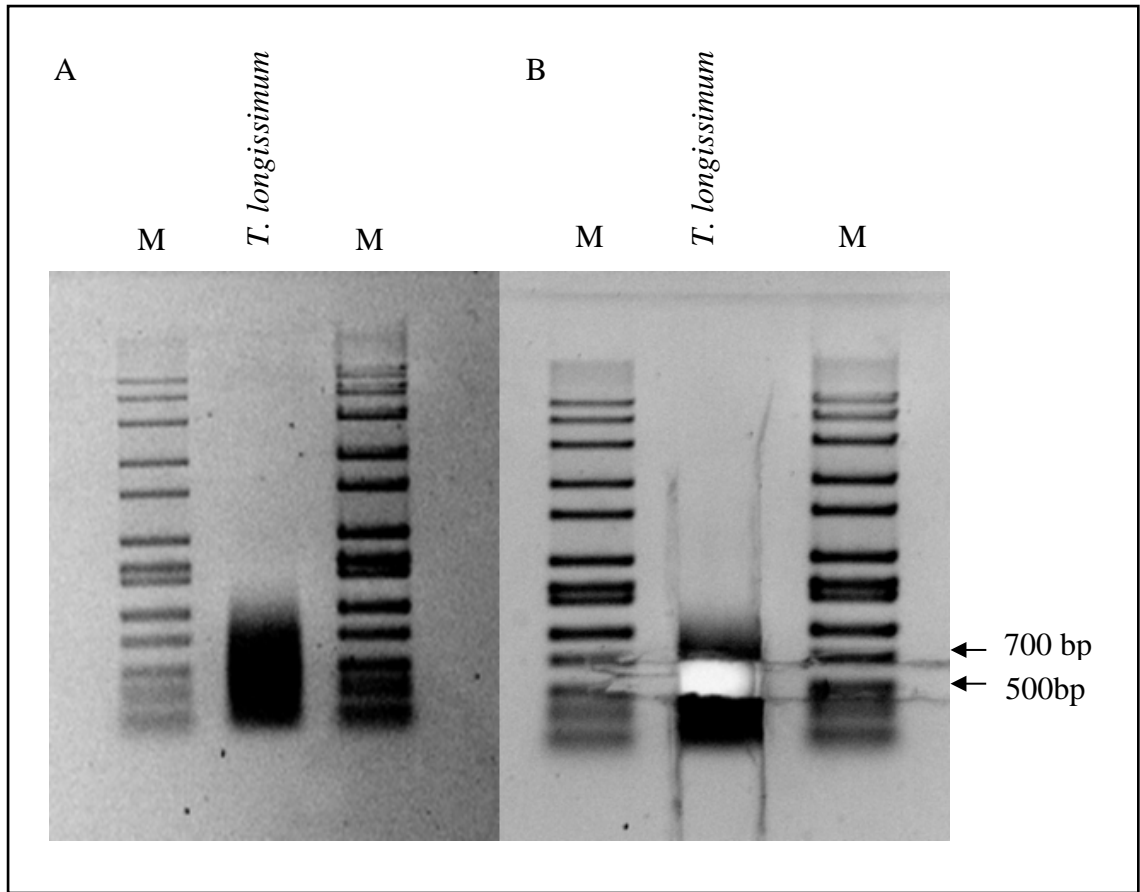
Figure D2. Nebulization results of mitochondrial genomic DNA (A). The sample picture shown here comes from *T. longissimum*. Fragments from 500 bp to 700 bp were used for sequencing (B). M-size markers.

Figure D3. Sequencing quality check and read length distribution. A. *T. turgidum*, B. (lo) durum, C. *T. tauschii* and D. *T. longissimum*. Graphs made with use of Geneious software (www.geneious.com).

A

```
         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
            10        20        30        40        50        60        70        80        90
T.turgidum      TCCTTTCAAA AGTGAGCGAG CAGAGCTGAA AAAGATGGAG TTACCTGGAG ATGAG--ATT TCTTTCTACG GATATGAAGG ATAGAAA---
(lo) durum      TCCTTTCAAA AGTGAGCGAG CAGAGCTGAA AAAGATGGAG TTACCTGGAG ATGAACAAAT TCGATTTAAT AACACAAACA GTACAAAGCG
T.longissumum   TCCTTTCAAA AGTGAGCGAG CAGAGCTGAA AAAGATGGAG TTACCTGGAG ATGAACAAAT TCGATTTAAT AACACAAACA GTACAAAGCG
T.tuashii       TCCTTTCAAA AGTGAGCGAG CAGAGCTGAA AAAGATGGAG TTACCTGGAG ATGAACAAAT TCGATTTAAT AACACAAACA GTACAAAGCG

         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
            100       110       120       130       140       150       160       170       180
T.turgidum      TATGCTATTT GC-TGCTATT ACAACGAAT- -----CAACC -------AAT TCGTAGTAAG TGTTCCCGTC TTCCCGAT-- CTACATGATT
(lo) durum      TCTCTCACGT CCATTCTTTT ATTTTAGATT TTTCCCAACC TGCGTTGGAT CCGCAGCTTG TGAACCGGCT TTTGGAAGGG GTGGAGGCAC
T.longissumum   TCTCTCACGT CCATTCTTTT ATTTTTGATT TTTCCCAACC TGCGTTGGAT CCGCAGCTTG TGAACCGGCT TTTGGAAGGG GTGGAGGCAC
T.tuashii       TCTCTCACGT CCATTCTTTT ATTTTTGATT TTTCCCAACC TGCGTTGGAT CCGCAGCTTG TGAACCGGCT TTTGGAAGGG GTGGAGGCAC

         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
            190       200       210       220       230       240       250       260       270
T.turgidum      TTTTCCCAAC CAACA----- --TCTCTCAG AACTTTGCT- -ATAACGCC- --AAAC---- ---------- -TTGGATATA ACGCCAACGC
(lo) durum      CTACCTCGAA AGACATGGAC CTCCTTTTAA AAATGTATGG AATTGCGCCC GAAAATACAC AGGCTTTGCT TTTTGAAATC ACAAAAATGG
T.longissumum   CTACCTCGAA AGACAGGGAC CTCCTTTTAA AAATGTATGG AATTGCGCCC GGAAATACAC AGGCTTTGCT TTTTGAAATC ACAAAAATGG
T.tuashii       CTACCTCGAA AGACAGGGAC CTCCTTTTAA AAATGTATGG AATTGCGCCC GGAAATACAC AGGCTTTGCT TTTTGAAATC ACAAAAATGG

         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
            280       290       300       310       320       330       340       350       360
T.turgidum      CTGAG-CGAA TTGCCGGCGT CACAAT--AG TTTTACAA-- ---ATAGAAG AGTATTTGGG CC-AAAATGA G--TCCGAAC AGGGAGCAGT
(lo) durum      TAGAATCATA CATGCGGCAG GAAAATTTGG TTTTGCCACC TGCATACGAT GTTACCTGGT TTTGTAACCA GTTTGTGGGC GTGGATTCTT
T.longissumum   TAGAATCATA CATGCGGCAG GCAAATTTGG TTTTGCCACC TGCATACGAT GTTACCTGGT TTTGTAACCA GTTTCTGGGC GTGGATTCTT
T.tuashii       TAGAATCATA CATGCGGCAG GCAAATTTGG TTTTGCCACC TGCATACGAT GTTACCTGGT TTTGTAACCA GTTTCTGGGC GTGGATTCTT

         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
            370       380       390       400       410       420       430       440       450
T.turgidum      CAATTTAGCT --AGAACAGT ATTGGGAGCC CGCCACCGAA ATGGCGA--- -----AACTT GG----CAGG GCATATTAG- -------AGG
(lo) durum      TTAAGGATCT AAAGGATACT TTCGTAGATT TACATTTCAA GGGGCGCGAA AGCTTAACTT GGAACACGGG CTTTGTTAAT TTTTATAACG
T.longissumum   TTAAGGATCT AAAGGATACT TTCGTAGATT TACATTTCAA GGGGCGCGAA AGCTTAACTT GGAACACGGG CTTTGTTAAT TTTTATAACG
T.tuashii       TTAAGGATCT AAAGGATACT TTCGTAGATT TACATTTCAA GGGGCGCGAA AGCTTAACTT GGAACACGGG CTTTGTTAAT TTTTATAACG
```

Figure D4. The full *atp6* gene sequence including conserved regions of mitochondrial genome surrounding both alleles of *atp6* gene. Region colored blue represents a pre-sequence part of *atp6* gene, starting from ATG codon. Regions colored green represent conserved region and gray bars shows polymorphism found within core region of gene

B

```
         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             460       470       480       490       500       510       520       530       540
T.turgidum   ATATTCGGGC GGGTGGTGGT AT------GG ATAATTTTAT CCAGAATCTG CCTGGTGCCT ACCCGGAAAC CCCATTGGAT CAATTTGCCA
(lo) durum   AAATATGGAG GGCCGAAGAC TCCACGCCGG GTAGCGCTGT AAGTGCTCCG AATCAGCTTT TTTTGGAAAA CCCATTGGAT CAATTTGCCA
T.longissumum AAATATGGAG GGCCGAAGAC TCCCCGCCGG GTAGCGCTGT AAGTGCTCCG AATCAGCTTT TTTTGGAAAA CCCATTGGAT CAATTTGCCA
T.tuashii    AAATATGGAG GGCCGAAGAC TCCCCGCCGG GTAGCGCTGT AAGTGCTCCG AATCAGCTTT TTTTGGAAAA CCCATTGGAT CAATTTGCCA


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             550       560       570       580       590       600       610       620       630
T.turgidum   TTATCCCAAT AATTGATCTT CATGTGGGCA ACTTTTATTT ATCATTTACA AATGAAGTCT TGTATATGCT GCTCACTGTC GTTTTGGTCG
(lo) durum   TTTACCCAAT AATTGATCTG CATGTGGGCA ACTTTTATTT TACATTTACA AATGCAGTCT TGTATATGCT GCTCACTGTC GTTTTGGTCG
T.longissumum TTTACCCAAT AATTGATCTT CATGTGGGCA ACTTTTATTT TACATTTACA AATGCAGTCT TGTATATGCT GCTCACTGTC GTTTTGGTCG
T.tuashii    TTTACCCAAT AATTGATCTT CATGTGGGCA ACTTTTATTT TACATTTACA AATGCAGTCT TGTATATGCT GCTCACTGTC GTTTTGGTCG


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             640       650       660       670       680       690       700       710       720
T.turgidum   TTTTTCTTTT TTTTGTTGTT ACGAAAAAGG GAGGTGGAAA GTCAGTGCCA AATGCATGGC AATCCTTGGT CGAGCTTATT TATGATTTCG
(lo) durum   TTTTTCTGTT TTTTGTTGTT ACGAAAAAGG GAGGTGGAAA GTCAGTGCCA AATGCATGGC AATCCTTGGT CGAGCTTATT TATGATTTCG
T.longissumum TTTTTCTGTT TTTTGTTGTT ACGAAAAAGG GAGGTGGAAA GTCAGTGCCA AATGCATGGC AATCCTTGGT CGAGCTTATT TATGATTTCG
T.tuashii    TTTTTCTGTT TTTTGTTGTT ACGAAAAAGG GAGGTGGAAA GTCAGTGCCA AATGCATGGC AATCCTTGGT CGAGCTTATT TATGATTTCG


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             730       740       750       760       770       780       790       800       810
T.turgidum   TGCTGAACCT GGTAAACGAA CAAATAGGTG GTCTTTCCGG AAATGTGAAA CAAAAGTTTT TCCCTCGCAT CTCGGTCACT TTTACTTTTT
(lo) durum   TGCTGAACCT GGTAAACGAA CAAATAGGTG GTCTTTCCGG AAATGTGAAA CAAAAGTTTT TCCCTCGCAT CTCGGTCACT TTTACTTTTT
T.longissumum TGCTGAACCT GGTAAACGAA CAAATAGGTG GTCTTTCCGG AAATGTGAAA CAAAAGTTTT TCCCTCGCAT CTCGGTCACT TTTACTTTTT
T.tuashii    TGCTGAACCT GGTAAACGAA CAAATAGGTG GTCTTTCCGG AAATGTGAAA CAAAAGTTTT TCCCTCGCAT CTCGGTCACT TTTACTTTTT


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             820       830       840       850       860       870       880       890       900
T.turgidum   CGTTATTTCG TAATCCCCAG GGTATGATAC CCTTTAGCTT CACAGTGACA AGTCATTTTC TCATTACTTT GGCTCTTTCA TTTTCCATTT
(lo) durum   CGTTATTTCG TAATCCCCAG GGTATGATAC CCTTTAGCTT CACAGTGACA AGTCATTTTC TCATTACTTT GGCTCTTTCA TTTTCCATTT
T.longissumum CGTTATTTCG TAATCCCCAG GGTATGATAC CCTTTAGCTT CACAGTGACA AGTCATTTTC TCATTACTTT GGCTCTTTCA TTTTCCATTT
T.tuashii    CGTTATTTCG TAATCCCCAG GGTATGATAC CCTTTAGCTT CACAGTGACA AGTCATTTTC TCATTACTTT GGCTCTTTCA TTTTCCATTT
```

Figure D4. The full *atp6* gene sequence including conserved regions of mitochondrial genome surrounding both alleles of *atp6* gene (continued).

C

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   910        920        930        940        950        960        970        980        990
T.turgidum      TTATAGGCAT TACGATCGTT GGATTTCAAA GACATGGGCT TCATTTTTTT AGCTTCTTAT TACCTGCGGG AGTCCCACTG CCGTTAGCAC
(lo) durum      TTATAGGCAT TACGATCGTT GGATTTCAAA GACATGGGCT TCATTTTTTT AGCTTCTTAT TACCTGCGGG AGTCCCACTG CCGTTAGCAC
T.longissumum   TTATAGGCAT TACGATCGTT GGATTTCAAA GACATGGGCT TCATTTTTTT AGCTTCTTAT TACCTGCGGG AGTCCCACTG CCGTTAGCAC
T.tuashii       TTATAGGCAT TACGATCGTT GGATTTCAAA GACATGGGCT TCATTTTTTT AGCTTCTTAT TACCTGCGGG AGTCCCACTG CCGTTAGCAC


                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  1000       1010       1020       1030       1040       1050       1060       1070       1080
T.turgidum      CTTTCTTAGT ACTCCTTGAG CTAATCTCTT ATTGTTTTCG TGCATTAAGC TTAGGAATAC GTTTATTTGC TAATATGATG GCCGGTCATA
(lo) durum      CTTTCTTAGT ACTCCTTGAG CTAATCTCTT ATTGTTTTCG TGCATTAAGC TTAGGAATAC GTTTATTTGC TAATATGATG GCCGGTCATA
T.longissumum   CTTTCTTAGT ACTCCTTGAG CTAATCTCTT ATTGTTTTCG TGCATTAAGC TTAGGAATAC GTTTATTTGC TAATATGATG GCCGGTCATA
T.tuashii       CTTTCTTAGT ACTCCTTGAG CTAATCTCTT ATTGTTTTCG TGCATTAAGC TTAGGAATAC GTTTATTTGC TAATATGATG GCCGGTCATA


                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  1090       1100       1110       1120       1130       1140       1150       1160       1170
T.turgidum      GTTTAGTAAA GATTTTAAGT GGGTTTGCTT GGACTATGCT ATTTCTGAAT AATATTTTCT ATTTCATAGG AGATCTTGGT CCCTTATTTA
(lo) durum      GTTTAGTAAA GATTTTAAGT GGGTTTGCTT GGACTATGCT ATTTCTGAAT AATATTTTCT ATTTCATAGG AGATCTTGGT CCCTTATTTA
T.longissumum   GTTTAGTAAA GATTTTAAGT GGGTTTGCTT GGACTATGCT ATTTCTGAAT AATATTTTCT ATTTCATAGG AGATCTTGGT CCCTTATTTA
T.tuashii       GTTTAGTAAA GATTTTAAGT GGGTTTGCTT GGACTATGCT ATTTCTGAAT AATATTTTCT ATTTCATAGG AGATCTTGGT CCCTTATTTA

                  .         .         .         .         .         .         .         .         .         .
                                                                                                                60
T.turgidum      TAGTTCTAGC ATTAACCGGT CTGGAATTAG GTGTAGCTAT ATCACAAGCT CATGTTTCTA CGATCTCAAT TTGTATTTAC TTGAATGATG
(lo) durum      TAGTTCTAGC ATTAACCGGT CTGGAATTAG GTGTAGCTAT ATCACAAGCT CATGTTTCTA CGATCTCAAT TTGTATTTAC TTGAATGATG
T.longissumum   TAGTTCTAGC ATTAACCGGT CTGGAATTAG GTGTAGCTAT ATCACAAGCT CATGTTTCTA CGATCTCAAT TTGTATTTAC TTGAATGATG
T.tuashii       TAGTTCTAGC ATTAACCGGT CTGGAATTAG GTGTAGCTAT ATCACAAGCT CATGTTTCTA CGATCTCAAT TTGTATTTAC TTGAATGATG


                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  1270       1280       1290       1300       1310       1320       1330       1340       1350
T.turgidum      CTACAAATCT CCATCAAAAT GAGTCATTTC ATAATTGAAT AAAAACGAGG AGCCGAAGAT TCTAGGGGGC TACAGCTGCG CTTTTGCAGC
(lo) durum      CTACAAATCT CCATCAAAAT GAGTCATTTC ATAATTGAAT AAAAACGAGG AGCCGAAGAT TCTAGGGGGC TACAGCTGCG CTTTTGCAGC
T.longissumum   CTACAAATCT CCATCAAAAT GAGTCATTTC ATAATTGAAT AAAAACGAGG AGCCGAAGAT TCTAGGGGGC TACAGCTGCG CTTTTGCAGC
T.tuashii       CTACAAATCT CCATCAAAAT GAGTCATTTC ATAATTGAAT AAAAACGAGG AGCCGAAGAT TCTAGGGGGC TACAGCTGCG CTTTTGCAGC
```

Figure D4. The full *atp6* gene sequence including conserved regions of mitochondrial genome surrounding both alleles of *atp6* gene (continued).

Figure D5. The BlastN results using the *T. longissimum* mitochondrial *atp6* gene region. Results show a lack of similarity between the pre-sequence region of *atp6* gene (from START codon), and the conserved flanking sequences compared to *T. aestivum* mitochondrial genome (first four hits) and *T. timophevii*. The core region reveals a high conservation among other species (black box). (NCBI, 2011; http://blast.ncbi.nlm.nih.gov/Blast.cgi)

A

```
         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             10        20        30        40        50        60        70        80        90
T. turgidum      ATGCTCTGTA TAATACTTTT CCCCGAGCGA TGGTTTAGCG GATTCGGAAT TGTAACCAAG CATCCTGGGT TCTATACCCG ATTCAACACT
(lo) durum       ATGCTCTGTA TAATACTTTT CCCCGAGCGA TGGTTTAGCG GATTCGGAAT TGTAACCAAG CATCCTGGGT TCTATACCCG ATTCAACACT
T.longissimum    ATGCTCTGTA TAATACTTTT CCCCGAGCGA TGGTTTAGCG GATTCGGAAT TGTAACCAAG CATCCTGGGT TCTATACCCG ATTCAACACT
T. tauschii      ATGCTCTGTA TAATACTTTT CCCCGAGCGA TGGTTTAGCG GATTCGGAAT TGTAACCAAG CATCCTGGGT TCTATACCCG ATTCAACACT


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             100       110       120       130       140       150       160       170       180
T. turgidum      AGAGCATGCA GCCGATCCTG GATACATAAC TCTAAAAAGT GTGTGTGCAG TTTTGGATCT TTATTGGTAG CCAGTCTTTC ACTTCTGCCT
(lo) durum       AGAGCATGCA GCCGATCCTG GATACATCAC TATATAAAGT GTGCATGCAG TTTTGGATCT TTATTGGTAG CCAGTCTTTC ACTTCTGCCT
T.longissimum    AGAGCATGCA GCCGATCCTG GATACATAAC TATATAAAGT GTG----CAG TTTTGGATCT TTATTGGTAG CCAGTCTTTC ACTTCTGCCT
T. tauschii      AGAGCATGCA GCCGATCCTG GATACATAAC TATATAAAGT GTG----CAG TTTTGGATCT TTATTGGTAG CCAGTCTTTC ACTTCTGCCT
                                                                                        STOP


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             190       200       210       220       230       240       250       260       270
T. turgidum      CTCCACTCCC ATGCCTTTCT TGGTCGGACC AACCCAACCG GCGATTTCCG ACAAGTCTTT CTGCTTAGAG CAAGAAGCGG AACCAAAATA
(lo) durum       CTCCACTCCC ATGCCTTTCT TGGTCGGACC AACCCAACCG GCGATTTCCG ACAAGTCTTT CTGCTTAGAG CAAGAAGCGG AACCAAAATA
T.longissimum    CTCCACTCCC ATGCCTTTCT TGGTCGGACC AACCCAACCG GCGATTTCCG ACAAGTCTTT CTGCTTAGAG CAAGAAGCGG AACCAAAATA
T. tauschii      CTCCACTCCC ATGCCTTTCT TGGTCGGACC AACCCAACCG GCGATTTCCG ACAAGTCTTT CTGCTTAGAG CAAGAAGCGG AACCAAAATA


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             280       290       300       310       320       330       340       350       360
T. turgidum      AAGCTTTCTT TATTTTCATT TATGGATAAC CAATCCATTT TCCAATATAG TTGGGAGATT TTACCCAAGA AATGGGTACA TAAAATGAAA
(lo) durum       AAGCTTTCTT TATTTTCATT TATGGATAAC CAATCCATTT TCCAATATAG TTGGGAGATT TTACCCAAGA AATGGGTACA TAAAATGAAA
T.longissimum    AAGCTTTCTT TATTTTCATT TATGGATAAC CAATCCATTT TCCAATATAG TTGGGAGATT TTACCCAAGA AATGGGTACA TAAAATGAAA
T. tauschii      AAGCTTTCTT TATTTTCATT TATGGATAAC CAATCCATTT TCCAATATAG TTGGGAGATT TTACCCAAGA AATGGGTACA TAAAATGAAA


         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             370       380       390       400       410       420       430       440       450
T. turgidum      AGATCGGAAC ATGGGAATAG ATCTTATACC AATACTGACT ACCCATTTCC ATTGTTGTGC TTTCTAAAAT GGCATACCTA TACAAGGGTT
(lo) durum       AGATCGGAAC ATGGGAATAG ATCTTATACC AATACTGACT ACCCATTTCC ATTGTTGTGC TTTCTAAAAT GGCATACCTA TACAAGGGTT
T.longissimum    AGATCGGAAC ATGGGAATAG ATCTTATACC AATACTGACT ACCCATTTCC ATTGTTGTGC TTTCTAAAAT GGCATACCTA TACAAGGGTT
T. tauschii      AGATCGGAAC ATGGGAATAG ATCTTATACC AATACTGACT ACCCATTTCC ATTGTTGTGC TTTCTAAAAT GGCATACCTA TACAAGGGTT
```

Figure D6. The full *nad9* gene nucleotide sequence multi-alignment. Three SNP's were recognized (light gray boxes) in comparison to *T. turgidum*, one of them (dark gray box) was found only in (lo) durum line. A four-nucleotide deletion in *T. longissimum* creates a STOP codon at the base 157, an additional dinucleotide change (CA/TG[134-135]) change in (lo) durum is indicated by the (orange box).

109

B

```
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                460        470        480        490        500        510        520        530        540
T. turgidum     CAAGTTTCGA TCGATATTTG CGGAGTGGAT CATCCCTCTC GAAAACGAAG ATTTGAAGTT GTCCATAATT TACTGAGTAC TCGGTATAAC
(lo) durum      CAAGTTTCGA TCGATATTTG CGGAGTGGAT CATCCCTCTC GAAAACGAAG ATTTGAAGTT GTCCATAATT TACTGAGTAC TCGGTATAAC
T.longissimum   CAAGTTTCGA TCGATATTTG CGGAGTGGAT CATCCCTCTC GAAAACGCAG ATTTGAAGTT GTCCATAATT TACTGAGTAC TCGGTATAAC
T. tauschii     CAAGTTTCGA TCGATATTTG CGGAGTGGAT CATCCCTCTC GAAAACGCAG ATTTGAAGTT GTCCATAATT TACTGAGTAC TCGGTATAAC


            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                550        560        570        580        590        600        610        620        630
T. turgidum     TCACGCATTC GTGTACAAAC AAGTGCAGAC GAAGTAACAC GAATATCTCC GGTAGTCAGT CTATTTCCAT CAGCCGGCCG GTGGGAGCGA
(lo) durum      TCACGCATTC GTGTACAAAC AAGTGCAGAC GAAGTAACAC GAATATCTCC GGTAGTCAGT CTATTTCCAT CAGCCGGCCG GTGGGAGCGA
T.longissimum   TCACGCATTC GTGTACAAAC AAGTGCAGAC GAAGTAACAC GAATATCTCC GGTAGTCAGT CTATTTCCAT CAGCCGGCCG GTGGGAGCGA
T. tauschii     TCACGCATTC GTGTACAAAC AAGTGCAGAC GAAGTAACAC GAATATCTCC GGTAGTCAGT CTATTTCCAT CAGCCGGCCG GTGGGAGCGA


            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                640        650        660        670        680        690        700        710        720
T. turgidum     GAAGTATGGG ATATGTCTGG TGTTTCTTCC ATCAATCATC CGGATTTACG CCGTATATCA ACAGATTATG GTTTCGAGGG TCATCCATTA
(lo) durum      GAAGTATGGG ATATGTCTGG TGTTTCTTCC ATCAATCATC CGGATTTACG CCGTATATCA ACAGATTATG GTTTCGAGGG TCATCCATTA
T.longissimum   GAAGTATGGG ATATGTCTGG TGTTTCTTCC ATCAATCATC CGGATTTACG CCGTATATCA ACAGATTATG GTTTCGAGGG TCATCCATTA
T. tauschii     GAAGTATGGG ATATGTCTGG TGTTTCTTCC ATCAATCATC CGGATTTACG CCGTATATCA ACAGATTATG GTTTCGAGGG TCATCCATTA


            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                730        740        750        760        770        780        790        800        810
T. turgidum     CGAAAAGACT TTCCTCTGAG TGGATATGTG GAAGTACGCT ATGATGATCC AGAGAAACGT GTGGTTTCTG AACCCATTGA GATGACCCAA
(lo) durum      CGAAAAGACT TTCCTCTGAG TGGATATGTG GAAGTACGCT ATGATGATCC AGAGAAACGT GTGGTTTCTG AACCCATTGA GATGACCCAA
T.longissimum   CGAAAAGACT TTCCTCTGAG TGGATATGTG GAAGTACGCT ATGATGATCC AGAGAAACGT GTGGTTTCTG AACCCATTGA GATGACCCAA
T. tauschii     CGAAAAGACT TTCCTCTGAG TGGATATGTG GAAGTACGCT ATGATGATCC AGAGAAACGT GTGGTTTCTG AACCCATTGA GATGACCCAA


            ....|....| ....|....| ....|....| ....|....| ....|....| ....
                820        830        840        850        860
T. turgidum     GAATTTCGCT ATTTCGATTT TGCTAGTCCT TGGGAACAGC GTAGCGACGG ATAA
(lo) durum      GAATTTCGCT ATTTCGATTT TGCTAGTCCT TGGGAACAGC GTAGCGACGG ATAA
T.longissimum   GAATTTCGCT ATTTCGATTT TGCTAGTCCT TGGGAACAGC GTAGCGACGG ATAA
T. tauschii     GAATTTCGCT ATTTCGATTT TGCTAGTCCT TGGGAACAGC GTAGCGACGG ATAA
```

Figure D6. The full *nad9* gene nucleotide sequence multi-alignment (continued).

110

A

```
                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        10        20        30        40        50        60        70        80        90
T. turgidum         ATGCGTCTTC TTGCTCCAGC ATTCAAGTTC CATTTCAAGG GAGGACGACG TACCATGATA CTTTCTGTTT TGTCGAGCCC TGCTTTGGTC
(lo) durum          ATGCGTCTTC TTGCTCCAGC ATTCAAGTTC CATTTCAAGG GAGGACGACG TACCATGATA CTTTCTGTTT TGTCGAGCCC TGCTTTGGTC
T. longissimum      ATGCGTCTTC TTGCTCCAGC ATTCAAGTTC CATTTCAAGG GAGGACGACG TACCATGATA CTTTCTGTTT TGTCGAGCCC TGCTTTGGTC
T. tauschii         ATGCGTCTTC TTGCTCCAGC ATTCAAGTTC CATTTCAAGG GAGGACGACG TACCATGATA CTTTCTGTTT TGTCGAGCCC TGCTTTGGTC


                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        100       110       120       130       140       150       160       170       180
T. turgidum         TCTGGTTTGA TGGTTGTACG TGCTAAAAAT CCGGTACATT CCGTTTTGTT TCCCATCCTA GTCTTTTGCG ACACTTCTGG TTTACTTATT
(lo) durum          TCTGGTTTGA TGGTTGTACG TGCTAAAAAT CCGGTACATT CCGTTTTGTT TCCCATCCTA GTCTTTTGCG ACACTTCTGG TTTACTTATT
T. longissimum      TCTGGTTTGA TGGTTGTACG TGCTAAAAAT CCGGTACATT CCGTTTTGTT TCCCATCCTA GTCTTTTGCG ACACTTCTGG TTTACTTATT
T. tauschii         TCTGGTTTGA TGGTTGTACG TGCTAAAAAT CCGGTACATT CCGTTTTGTT TCCCATCCTA GTCTTTTGCG ACACTTCTGG TTTACTTATT


                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        190       200       210       220       230       240       250       260       270
T. turgidum         TTGTTAGGTC TCGACTTCTC CGCTATGATC TCCCCAGTAG TTCATATAGG AGCTATTGCC GTTTCATTCC TATTCGTGGT TATGATGTTC
(lo) durum          TTGTTAGGTC TCGACTTCTC CGCTATGATC TCCCCAGTAG TTCATATAGG AGCTATTGCC GTTTCATTCC TATTCGTGGT TATGATGTTC
T. longissimum      TTGTTAGGTC TCGACTTCTC CGCTATGATC TCCCCAGTAG TTCATATAGG AGCTATTGCC GTTTCATTCC TATTCGTGGT TATGATGTTC
T. tauschii         TTGTTAGGTC TCGACTTCTC CGCTATGATC TCCCCAGTAG TTCATATAGG AGCTATTGCC GTTTCATTCC TATTCGTGGT TATGATGTTC


                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        280       290       300       310       320       330       340       350       360
T. turgidum         AATATTCAAA TAGCGGAGAT TCACGAAGAA GTATTGCGCT ATTTACCAGT GAGTGGTATT ATTGGACTGA TCTTTTGGTG GGAAATGTTC
(lo) durum          AATATTCAAA TAGCGGAGAT TCACGAAGAA GTATTGCGCT ATTTACCAGT GAGTGGTATT ATTGGACTGA TCTTTTGGTG GGAAATGTTC
T. longissimum      AATATTCAAA TAGCGGAGAT TCACGAAGAA GTATTGCGCT ATTTACCAGT GAGTGGTATT ATTGGACTGA TCTTTTGGTG GGAAATGTTC
T. tauschii         AATATTCAAA TAGCGGAGAT TCACGAAGAA GTATTGCGCT ATTTACCAGT GAGTGGTATT ATTGGACTGA TCTTTTGGTG GGAAATGTTC


                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        370       380       390       400       410       420       430       440       450
T. turgidum         TTCATTTTAG ATAATGAAAC CATTCCATTA CTACCAACCC ACAGAAATAC GACCTCTCTG AGATATACGG TTTATGCCGG AAAGGTACGA
(lo) durum          TTCATTTTAG ATAATGAAAC CATTCCATTA CTACCAACCC ACAGAAATAC GACCTCTCTG AGATATACGG TTTATGCCGG AAAGGTACGA
T. longissimum      TTCATTTTAG ATAATGAAAC CATTCCATTA CTACCAACCC ACAGAAATAC GACCTCTCTG AGATATACGG TTTATGCCGG AAAGGTACGA
T. tauschii         TTCATTTTAG ATAATGAAAC CATTCCATTA CTACCAACCC ACAGAAATAC GACCTCTCTG AGATATACGG TTTATGCCGG AAAGGTACGA
```

Figure D7. The full *nad6* gene nucleotide sequence multi-alignment. Three SNPs were recognized and two dinucleotide changes (light gray boxes) in comparison to *T. turgidum*. Beginning from the 703 bp position there is the start of highly polymorphic gene region.

B

```
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 460       470       480       490       500       510       520       530       540
T. turgidum      AGTTGGACTA ATTTGGAAAC ATTGGGCAAT TTACTTTATA CCTACTATTC CGTCTGGTTT TTGGTTTCTA GTCTGATTTT ATTAGTTGCT
(lo) durum       AGTTGGACTA ATTTGGAAAC ATTGGGCAAT TTGCTTTATA CCTACTATTC CGTCTGGTTT TTGGTTTCTA GTCTGATTTT ATTAGTAGCT
T. longissimum   AGTTGGACTA ATTTGGAAAC ATTGGGCAAT TTGCTTTATA CCTACTATTC CGTCTGGTTT TTGGTTTCTA GTCTGATTTT ATTAGTAGCT
T. tauschii      AGTTGGACTA ATTTGGAAAC ATTGGGCAAT TTGCTTTATA CCTACTATTC CGTCTGGTTT TTGGTTTCTA GTCTGATTTT ATTAGTAGCT

            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 550       560       570       580       590       600       610       620       630
T. turgidum      ATGATTGGGG CTATAGTACT TACTATGCAT AGGACTACAA AGGTGAAAAG ACAGGATGTA TTCCGACGAA ATGCCTTGGA TTCTAGGAGC
(lo) durum       ATGATTGGGG CTATAGTACT TACTATGCAT AGGACTACAA AGGTGAAAAG ACAGGATGTA TTCCGACGAA ATGCCTTGGA TTCTAGGAGG
T. longissimum   ATGATTGGGG CTATAGTACT TACTATGCAT AGGACTACAA AGGTGAAAAG ACAGGATGTA TTCCGACGAA ATGCCTTGGA TTCTAGGAGG
T. tauschii      ATGATTGGGG CTATAGTACT TACTATGCAT AGGACTACAA AGGTGAAAAG ACAGGATGTA TTCCGACGAA ATGCCTTGGA TTCTAGGAGG

            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 640       650       660       670       680       690       700       710       720
T. turgidum      CATATAATGA ACAGGACTAT TTCTCCTTTT GGCCATAGCC ATAGAAGAAG CTTCTCCTCC GGAGCGGGGG GACCG---CC TGACAATTAC
(lo) durum       AATATAATGA ACAGGACTAT TTCTCCTTTT GGCCATAGCC ATAGAAGAAG CTTCTCCTCC AAAGCGGAGG GAGGGGAATC TCAGGATTCC
T. longissimum   AATATAATGA ACAGGACTAT TTCTCCTTTT GGCCATAGCC ATAGAAGAAG CTTCTCCTCC AAAGCGGAGG GAGGGGAATC TCAGGATTCC
T. tauschii      AATATAATGA ACAGGACTAT TTCTCCTTTT GGCCATAGCC ATAGAAGAAG CTTCTCCTCC AAAGCGGAGG GAGGGGAATC TCAGGATTCC

            ....|....| ....|....| ....|....|
                 730       740       750
T. turgidum      AAAGAAACCT TTAAAATGTG GATTTAGTGC
(lo) durum       TATGACCCTG CTTATATCGA TTTTTTAAGA
T. longissimum   TATGACCCTG CTTATATCGA TTTTTTAAGA
T. tauschii      TATGACCCTG CTTATATCGA TTTTTTAAGA
```

Highly polymorphic region

Figure D7. The full *nad6* gene nucleotide sequence multi-alignment (continued).

```
        ....|....| ....|....| ....|....| ....|....| ....|....|
               10         20         30         40         50
T. turgidum      AGAGAAAGTC TGAGGAGCAG GAAAATTTGG TCACGTAGAT CTTCTATTTC
T. longissimum   AGAGAAAGTC TGAGGAGCAG GAAAATTTGG TCACGTAGAT CTTCTATTTC
(lo) durum       AGAGAAAGTC TGAGGAGCAG GAAAATTTGG TCACGTAGAT CTTCTATTTC


        ....|....| ....|....| ....|....| ....|....| ....|....|
               60         70         80         90        100
T. turgidum      GCCGGAATTC GTTGATTGCT CCGTACTCAT TTACAATGGA AAAACTCCTG
(lo) durum       GCCGGAATTC GTTGATTGCT CCGTACTCAT TTACAATGGA AAAACTCCTG
T. longissimum   GCCGGAATTC GTTGATTGCT CCGTACTCAT TTACAATGGA AAAACTCCTG


        ....|....| ....|....| ....|....| ....|....| ....|....|
              110        120        130        140        150
T. turgidum      TTCGTTGTAA GATCACTGAA GG-------- -TCATAAATT TGGAGAGTTT
(lo) durum       TTCGTTGTAA GATCACTGAA GGAAAGGTTG GTCATAAATT TGGAGAGTTT
T. longissimum   TTCGTTGTAA GATCACTGAA GGAAAGGTTG GTCATAAATT TGGAGAGTTT


        ....|....| ....|....| ....|....| ....|....| ....|....|
              160        170        180        190        200
T. turgidum      GCTTTTACAC GGAGACGAAG ACCCTATCGA ACAAATAGAG GAAAGGGGAA
(lo) durum       GCTTTTACAC GGAGACGAAG ACCCTATCGA ACAAATAGAG GAAAGGGGAA
T. longissimum   GCTTTTACAC GGAGACGAAG ACCCTATCGA ACAAATAGAG GAAAGGGGAA


        ....|..
T. turgidum      AAAGTAA
(lo) durum       AAAGTAA
T. longissimum   AAAGTAA
```

Figure D8. The result of *rps19*-p multi-alignment. It was found that *T. turgidum* has 9 bp deletions in relation to the (lo) durum and *T. longissimum* alleles, where both of these species share the same allele.
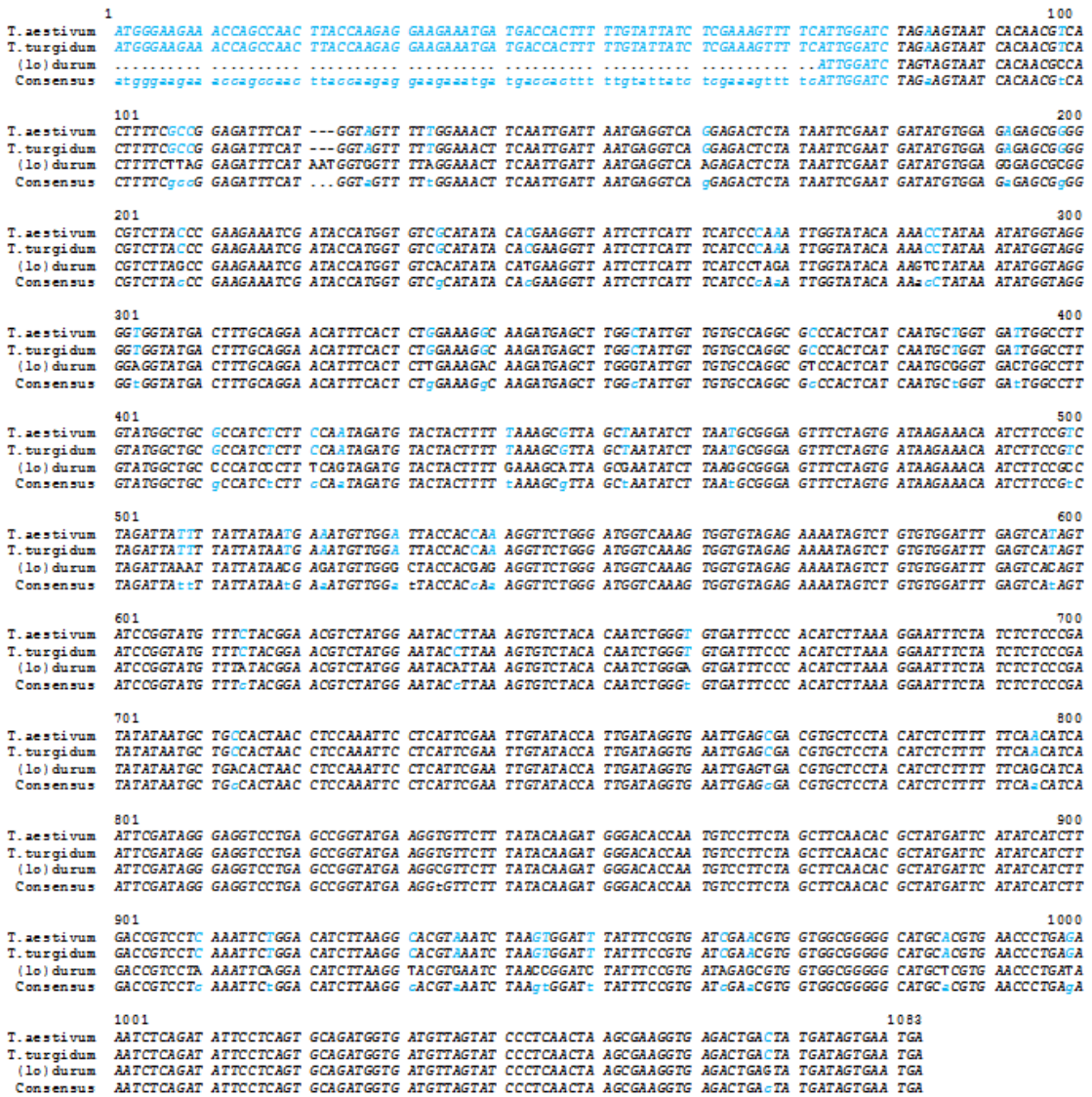
```
            1                                                                                                    100
T.aestivum  ATGGGAAGAA ACCAGCCAAC TTACCAAGAG GAAGAAATGA TGACCACTTT TTGTATTATC TCGAAAGTTT TCATTGGATC TAGAAGTAAT CACAACGTCA
T.turgidum  ATGGGAAGAA ACCAGCCAAC TTACCAAGAG GAAGAAATGA TGACCACTTT TTGTATTATC TCGAAAGTTT TCATTGGATC TAGAAGTAAT CACAACGTCA
(lo)durum   .......... .......... .......... .......... .......... .......... .......... ..ATTGGATC TAGTAGTAAT CACAACGCCA
Consensus   atgggaagaa accagccaac ttaccaagag gaagaaatga tgaccacttt ttgtattatc tcgaaagttt tcATTGGATC TAG.AGTAAT CACAACGtCA

            101                                                                                                  200
T.aestivum  CTTTTCGCCG GAGATTTCAT ---GGTAGTT TTTGGAAACT TCAATTGATT AATGAGGTCA GGAGACTCTA TAATTCGAAT GATATGTGGA GAGAGCGGGG
T.turgidum  CTTTTCGCCG GAGATTTCAT ---GGTAGTT TTTGGAAACT TCAATTGATT AATGAGGTCA GGAGACTCTA TAATTCGAAT GATATGTGGA GAGAGCGGGG
(lo)durum   CTTTTCTTAG GAGATTTCAT AATGGTGGTT TTAGGAAACT TCAATTGATT AATGAGGTCA AGAGACTCTA TAATTCGAAT GATATGTGGA GGGAGCGCGG
Consensus   CTTTTCgccG GAGATTTCAT ...GGT.GTT TTtGGAAACT TCAATTGATT AATGAGGTCA gGAGACTCTA TAATTCGAAT GATATGTGGA G.GAGCGgGG

            201                                                                                                  300
T.aestivum  CGTCTTACCC GAAGAAATCG ATACCATGGT GTCGCATATA CACGAAGGTT ATTCTTCATT TCATCCCAAA TTGGTATACA AAACCTATAA ATATGGTAGG
T.turgidum  CGTCTTACCC GAAGAAATCG ATACCATGGT GTCGCATATA CACGAAGGTT ATTCTTCATT TCATCCCAAA TTGGTATACA AAACCTATAA ATATGGTAGG
(lo)durum   CGTCTTAGCC GAAGAAATCG ATACCATGGT GTCACATATA CATGAAGGTT ATTCTTCATT TCATCCTAGA TTGGTATACA AAGTCTATAA ATATGGTAGG
Consensus   CGTCTTAcCC GAAGAAATCG ATACCATGGT GTCgCATATA CAcGAAGGTT ATTCTTCATT TCATCCc.A.A TTGGTATACA AA.cCTATAA ATATGGTAGG

            301                                                                                                  400
T.aestivum  GGTGGTATGA CTTTGCAGGA ACATTTCACT CTGGAAAGGC AAGATGAGCT TGGCTATTGT TGTGCCAGGC GCCCACTCAT CAATGCTGGT GATTGGCCTT
T.turgidum  GGTGGTATGA CTTTGCAGGA ACATTTCACT CTGGAAAGGC AAGATGAGCT TGGCTATTGT TGTGCCAGGC GCCCACTCAT CAATGCTGGT GATTGGCCTT
(lo)durum   GGAGGTATGA CTTTGCAGGA ACATTTCACT CTTGAAAGAC AAGATGAGCT TGGGTATTGT TGTGCCAGGC GTCCACTCAT CAATGCGGGT GACTGGCCTT
Consensus   GGtGGTATGA CTTTGCAGGA ACATTTCACT CTgGAAAGgC AAGATGAGCT TGGcTATTGT TGTGCCAGGC GcCCACTCAT CAATGCtGGT GAtTGGCCTT

            401                                                                                                  500
T.aestivum  GTATGGCTGC GCCATCTCTT CCAATAGATG TACTACTTTT TAAAGCGTTA GCTAATATCT TAATGCGGGA GTTTCTAGTG ATAAGAAACA ATCTTCCGTC
T.turgidum  GTATGGCTGC GCCATCTCTT CCAATAGATG TACTACTTTT TAAAGCGTTA GCTAATATCT TAATGCGGGA GTTTCTAGTG ATAAGAAACA ATCTTCCGTC
(lo)durum   GTATGGCTGC CCCATCCCTT TCAGTAGATG TACTACTTTT GAAAGCATTA GCGAATATCT TAAGGCGGGA GTTTCTAGTG ATAAGAAACA ATCTTCCGCC
Consensus   GTATGGCTGC gCCATcTCTT cCAa.TAGATG TACTACTTTT tAAAGCgTTA GCt.AATATCT TAAtGCGGGA GTTTCTAGTG ATAAGAAACA ATCTTCCGtC

            501                                                                                                  600
T.aestivum  TAGATTATTT TATTATAATG AAATGTTGGA TTACCACCAA AGGTTCTGGG ATGGTCAAAG TGGTGTAGAG AAAATAGTCT GTGTGGATTT GAGTCATAGT
T.turgidum  TAGATTATTT TATTATAATG AAATGTTGGA TTACCACCAA AGGTTCTGGG ATGGTCAAAG TGGTGTAGAG AAAATAGTCT GTGTGGATTT GAGTCATAGT
(lo)durum   TAGATTAAAT TATTATAACG AGATGTTGGG CTACCACGAG AGGTTCTGGG ATGGTCAAAG TGGTGTAGAG AAAATAGTCT GTGTGGATTT GAGTCACAGT
Consensus   TAGATTAttT TATTATAAtG A.ATGTTGG. tTACCACc.A. AGGTTCTGGG ATGGTCAAAG TGGTGTAGAG AAAATAGTCT GTGTGGATTT GAGTCAtAGT

            601                                                                                                  700
T.aestivum  ATCCGGTATG TTTCTACGGA ACGTCTATGG AATACCTTAA AGTGTCTACA CAATCTGGGT GTGATTTCCC ACATCTTAAA GGAATTTCTA TCTCTCCCGA
T.turgidum  ATCCGGTATG TTTCTACGGA ACGTCTATGG AATACCTTAA AGTGTCTACA CAATCTGGGT GTGATTTCCC ACATCTTAAA GGAATTTCTA TCTCTCCCGA
(lo)durum   ATCCGGTATG TTTATACGGA ACGTCTATGG AATACCATTAA AGTGTCTACA CAATCTGGGA GTGATTTCCC ACATCTTAAA GGAATTTCTA TCTCTCCCGA
Consensus   ATCCGGTATG TTTcTACGGA ACGTCTATGG AATACcTTAA AGTGTCTACA CAATCTGGGt GTGATTTCCC ACATCTTAAA GGAATTTCTA TCTCTCCCGA

            701                                                                                                  800
T.aestivum  TATATAATGC TGCCACTAAC CTCCAAATTC CTCATTCGAA TTGTATACCA TTGATAGGTG AATTGAGCGA CGTGCTCCTA CATCTCTTTT TTCAACATCA
T.turgidum  TATATAATGC TGCCACTAAC CTCCAAATTC CTCATTCGAA TTGTATACCA TTGATAGGTG AATTGAGCGA CGTGCTCCTA CATCTCTTTT TTCAACATCA
(lo)durum   TATATAATGC TGACACTAAC CTCCAAATTC CTCATTCGAA TTGTATACCA TTGATAGGTG AATTGAGTGA CGTGCTCCTA CATCTCTTTT TTCAGCATCA
Consensus   TATATAATGC TGcCACTAAC CTCCAAATTC CTCATTCGAA TTGTATACCA TTGATAGGTG AATTGAGcGA CGTGCTCCTA CATCTCTTTT TTCA.CATCA

            801                                                                                                  900
T.aestivum  ATTCGATAGG GAGGTCCTGA GCCGGTATGA AGGTGTTCTT TATACAAGAT GGGACACCAA TGTCCTTCTA GCTTCAACAC GCTATGATTC ATATCATCTT
T.turgidum  ATTCGATAGG GAGGTCCTGA GCCGGTATGA AGGTGTTCTT TATACAAGAT GGGACACCAA TGTCCTTCTA GCTTCAACAC GCTATGATTC ATATCATCTT
(lo)durum   ATTCGATAGG GAGGTCCTGA GCCGGTATGA AGGCGTTCTT TATACAAGAT GGGACACCAA TGTCCTTCTA GCTTCAACAC GCTATGATTC ATATCATCTT
Consensus   ATTCGATAGG GAGGTCCTGA GCCGGTATGA AGGtGTTCTT TATACAAGAT GGGACACCAA TGTCCTTCTA GCTTCAACAC GCTATGATTC ATATCATCTT

            901                                                                                                  1000
T.aestivum  GACCGTCCTC AAATTCTGGA CATCTTAAGG CACGTAAATC TAAGTGGATT TATTTCCGTG ATCGAACGTG GTGGCGGGGG CATGCACGTG AACCCTGAGA
T.turgidum  GACCGTCCTC AAATTCTGGA CATCTTAAGG CACGTAAATC TAAGTGGATT TATTTCCGTG ATCGAACGTG GTGGCGGGGG CATGCACGTG AACCCTGAGA
(lo)durum   GACCGTCCTA AAATTCAGGA CATCTTAAGG TACGTGAATC TAACCGGATC TATTTCCGTG ATAGAGCGTG GTGGCGGGGG CATGCTCGTG AACCCTGATA
Consensus   GACCGTCCTg AAATTCtGGA CATCTTAAGG cACGT.AATC TAAgtGGATt TATTTCCGTG ATcGA.CGTG GTGGCGGGGG CATGC.CGTG AACCCTGAgA

            1001                                                                   1083
T.aestivum  AATCTCAGAT ATTCCTCAGT GCAGATGGTG ATGTTAGTAT CCCTCAACTA AGCGAAGGTG AGACTGACTA TGATAGTGAA TGA
T.turgidum  AATCTCAGAT ATTCCTCAGT GCAGATGGTG ATGTTAGTAT CCCTCAACTA AGCGAAGGTG AGACTGACTA TGATAGTGAA TGA
(lo)durum   AATCTCAGAT ATTCCTCAGT GCAGATGGTG ATGTTAGTAT CCCTCAACTA AGCGAAGGTG AGACTGAGTA TGATAGTGAA TGA
Consensus   AATCTCAGAT ATTCCTCAGT GCAGATGGTG ATGTTAGTAT CCCTCAACTA AGCGAAGGTG AGACTGAcTA TGATAGTGAA TGA
```

Figure D9. Multi-alignment of nucleotide sequences of *orf359* from *T. turgidum, T. aestivum* and (lo) durum. O*rf359* is not present in *T. longissimum* and *T. tauschii*. Blue color indicates observed polymorphism between species.