bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Anatomy and transcription dynamics of wheat ribosomal RNA loci revealed by optical mapping and RNA sequencing

Zuzana Tulpová[1], Aleš Kovařík[2], Helena Toegelová[1], Veronika Kapustová[1], Pavla Navrátilová[1], Eva Hřibová[1], Jan Vrána[1], Jiří Macas[3], Jaroslav Doležel[1], Hana Šimková[1]

[1]Institute of Experimental Botany of the Czech Academy of Sciences, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ-779 00 Olomouc, Czech Republic

[2]Czech Academy of Sciences, Institute of Biophysics, Královopolska 135, CZ-61265 Brno, Czech Republic

[3]Biology Centre, Czech Academy of Sciences, Institute of Plant Molecular Biology, Branišovská 31, CZ-37005 České Budějovice, Czech Republic

Abstract

Three out of four RNA components of ribosomes are encoded by 45S rDNA loci, whose transcripts are processed into 18S, 5.8S and 26S ribosomal RNAs. The loci are organized as long head-to-tail tandem arrays of nearly identical units spanning over several megabases of sequence. Due to this peculiar structure, the exact copy number, sequence composition and expression status of the 45S rDNA remain elusive even in the genomics era, especially in complex polyploid genomes harbouring multiple loci. We employed a novel approach to study rDNA loci in polyploid wheat, comprising chromosome flow sorting, optical mapping and high-throughput rRNA sequencing. This enabled unbiased quantification of rDNA units in particular arrays. Total number of tandemly organised units in wheat genome was 4390, with 64.1, 31.4, 3.9 and 0.7% located in short arms of wheat chromosomes 6B, 1B, 5D and 1A, respectively. Optical maps revealed high regularity of tandem repeats in 1B and 5D, while the 6B array showed signatures of non-rDNA sequences invasions; 1A locus harbored highly rearranged repeats with many irregularities. At the expression level, only 1B and 6B loci contributed to transcription at roughly 2:1 ratio. The 1B:6B ratio varied among five analysed tissues (embryo, coleoptile, root tip, primary leaf, adult leaf), being the highest (2.64:1) in adult leaf and lowest (1.72:1) in coleoptile. In conclusion, a fine genomic organisation and tissue-specific expression of wheat rDNA loci was deciphered, for the first time, in a complex polyploid species. The results are discussed in the context of rDNA evolution and expression.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Keywords

45S rDNA, *Triticum aestivum*, NOR, chloroplast, mitochondria, flow cytometry, RNA-seq, Iso-Seq

Introduction

Ribosomal DNA (rDNA) is a general term for genes encoding several types of ribosomal RNA (rRNA), the crucial components of ribosomes. 45S rDNA locus, specific for eukaryotic genomes, is associated with the nucleolus and is termed nucleolus organising region (NOR). 5S rRNA gene loci are located apart from NOR locus in a majority of organisms. Both nuclear 5S and 45S rDNA units are mostly organized as long head-to-tail tandem arrays. The 45S rDNA unit is composed of conserved genes for 18S, 5.8S and 26S rRNA, separated by internal transcribed spacers (ITS1, ITS2). A smaller part of the unit belongs to a non-transcribed intergenic spacer (IGS) composed of several types of repeats of variable size and sequence. The 45S rDNA is transcribed by polymerase I into one large precursor, which is then processed into the three types of rRNA. The 5S rRNA genes are approximately120 bp long and, interspersed by non-transcribed spacers of a few hundred base pairs, form arrays spanning over several kilobases, which makes them accessible to long-read sequencing technologies such as PacBio (Symonova et al. 2017). On the contrary, unit size of the 45S rDNA is 8-13 kb across yeast, plant and animal genomes (Gerlach and Bedbrook, 1979, Pruitt and Meyerowitz, 1986, Nelson et al. 2019) and its arrays are assumed to span several hundred kilobases to hundred megabases of sequence (Todd et al. 2017, Handa et al. 2018), which precludes their complete assembling from next-generation-sequencing data. Absence of the 45S rDNA arrays in genome assemblies hampers their analysis, which is especially challenging in complex and polyploid genomes comprising multiple rRNA gene loci, as is the case of bread wheat. Bread wheat (*Triticum aestivum* L.) is an allohexaploid species with a genome of 16.58 Mb (Doležel et al. 2018), composed of three subgenomes (2n=6x=42, AABBDD). The huge genome size and polyploidy and, consequently, a lack of a genome sequence complicated genome analysis for a long time. On the other hand, the large chromosomes and complex genome composition attracted attention of cytogeneticists and stimulated research on wheat NORs by microscopic and other techniques. These studies made use of a collection of aneuploid lines generated for cv. Chinese Spring (Sears et al. 1954) that enabled assigning NORs to particular chromosomes and studying their behaviour.

As early as in late 1950's, Crosby (1957) reported that there were at least four different chromosomes in hexaploid wheat (1A, 1B, 6B and 5D) able to form nucleoli. In 'Chinese Spring' (CS), large nucleoli were assigned to chromosomes 1B and 6B, while smaller ones

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(micronucleoli) were associated with other chromosomes, mostly the 5D (Darvey and Driscoll, 1972). Nevertheless, the activity of the minor loci has not been generally confirmed. Subsequently, 45S rDNA clusters were positioned on metaphase chromosomes by *in situ* hybridisation (ISH), which placed all previously described loci to short arms of respective chromosomes and identified additional minor rDNA loci on wheat chromosomes 7D and 3D (summarised in Dubcovsky and Dvořák, 1995).

First estimates for numbers of rRNA genes harboured by the bread wheat genome were done by Flavell and co-workers (summarised in Flavell and O'Dell, 1976). By rRNA/DNA hybridisation assays, the authors estimated the 18S-25S rRNA genes to make up 0.087% wheat 'Chinese Spring' genome, which they interpreted as 9150 rRNA genes for a diploid nucleus. Out of them, 5500 (60%), 2700 (30%) and 950 (10%) were assigned to 6B, 1B and other chromosomes (predominantly 5D and 1A), respectively. Surprisingly, rRNA gene number in 1B and 6B correlated negatively with nucleolar volume in root tip cells of CS wheat where the volume of nucleolus on chromosome 1B was twice that on chromosome 6B (Martini and Flavell, 1985), indicating doubled transcription activity of the 1B NOR compared to that of 6B. This study also demonstrated that the minor NORs in 5D and 1A were increased their activity when the major NORs were deleted.

In the most recent study on wheat NORs, Handa et al. (2018) used quantitative PCR (qPCR) and fluorescence *in situ* hybridisation (FISH) to quantify rDNA units in bread wheat genome. A total copy number of the rDNA units in the genome was estimated to be 11,160, of which 30.5%, 60.9% and 8.6% were assigned to *Nor-B1* (1B), *Nor-B2* (6B) and other NORs, respectively. The estimate for total length of the NORs was 100 Mb. Besides, the authors compared rDNA units extracted from the reference genome of bread wheat, IWGSC RefSeq v1.0 (IWGSC 2018), and identified four rDNA subtypes, differing in the composition of the 3' external transcribed spacer (3' ETS), and analysed their expression, which was not found proportional to their representation in the genome. Moreover, the authors identified positions of the major 45S rDNA loci on chromosomes 1B and 6B, taking presence of rDNA clusters in the sequence as a clue.

BLAST search for the 18S, 5.8S and 26S rRNA genes in the IWGSC RefSeq v1.0 genome, carried out in Ensembl Plants (https://plants.ensembl.org/Triticum_aestivum/Tools/Blast), reveals several additional clusters of rDNA, located outside the proposed positions of *Nor-B1* and *Nor-B2*, mainly on the short arms of chromosomes 1A (1AS), 1B (1BS), 5D (5DS), 7D (7DS) and long arm of 1B (1BL), some of which might correspond to the minor loci identified previously by ISH (Mukai et al. 1991). None of the clusters, including those assigned to *Nor-B1* (1B) and *Nor-B2* (6B), consisted of a regular array of rDNA units. These observations call

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for a detailed study to confirm the presence of rDNA arrays in the predicted positions and characterize sequence organization of the particular loci. We showed previously (Kapustová et al. 2019) that Bionano optical mapping, a technology that visualises short sequence motives, typically CTTAGGG, along DNA molecules of 150 kb to 1 Mb (Lam et al. 2012), recognizes rDNA arrays in a genomic context thanks to presence of the sequence motives in rRNA genes.

Here we used this technology, combined with separation of individual rDNA-bearing wheat chromosome arms by flow cytometry (Staňková et al. 2016), to search for the presence of rDNA arrays in the anticipated positions in 1BS and 6BS chromosome arms as well as in 1AS and 5DS that are expected to bear minor rDNA loci. Importantly, the analysis of optical map raw data for tandemly organised repeats provided a novel approach to quantify rDNA units, which discriminates the functional units organised in arrays from the dispersed and fragmented ones, which are probably non-functional pseudogenes. Participation of the major and minor loci on rRNA synthesis was assessed in five tissues by analysing RNA sequencing data for specific chromosomal variants in the 26S rRNA. Besides, this data provided insights into the dynamic contribution of particular organelles to the overall rRNA production.

Materials and Methods

Plant material

Seeds of double ditelosomic stocks of bread wheat (*Triticum aestivum*, L.) cv. Chinese Spring, dDt1AS, dDt1BS, dDt6BS and dDt5DS (Sears and Sears, 1978), were kindly provided by Prof. Bikram Gill (Kansas State University, Manhattan, USA). Seeds of bread wheat cv. Chinese Spring with a standard karyotype were provided by Dr. Pierre Sourdille (INRAE, Clermont-Ferrand, France).

Bionano optical maps

Short arms of chromosomes 1A (1AS), 1B (1BS), 6B (6BS) and 5D (5DS) were purified by flow sorting from corresponding telosomic lines according to Kubaláková et al. (2002) and used for HMW DNA preparation as described in Staňková et al. (2016). Bionano optical maps (OMs) of chromosome arms 1AS, 1BS and 6BS were constructed using NLRS labelling chemistry (Nt.*BspQI* enzyme) and data were generated on the Irys platform (Bionano Genomics, San Diego, USA). The 5DS chromosome arm was labelled by DLS chemistry (DLE-1 enzyme) and molecules were analysed on the Saphyr platform (Bionano Genomics). DNA labelling and data collection followed protocols of Bionano Genomics https://bionanogenomics.com/products/bionano-prep-kits/, with minor modifications. Optical

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

map of the 1AS was assembled by Bionano IrysSolve 2.1.1 software, using optArguments_human.xml file. Assemblies of 1BS, 6BS and 5DS OMs were done by Bionano Solve 3.4 applying cmaps generated from fasta files of 1BS, 6BS and 5DS pseudomolecules of the IWGSC RefSeq v1.0 (IWGSC, 2018) as a reference. Assembly parameters specified by optArguments_nonhaplotype_noEs_noCut_irys.xml file were used for 1BS and 6BS map assembly whereas optArguments_nonhaplotype_noEs_noCut_DLE_saphyr.xml file was used for 5DS OM. Information about input DNA amount and labelling, quantity and quality of data used for the OM assembly, and metrics of particular chromosomal OM assemblies are stated in Suppl. Table S1. Map-to-sequence alignments were visualised in Bionano Access v1.5.

Quantification of rDNA units in arrays

An algorithm included in the IrysView 2.0 software package (Bionano Genomics) was used to identify labelled tandemly organized repeats in size-filtered (>150 kb) raw data. Regular arrays of six and more repeat units were considered in the analysis. Repeat stretch tolerance of 0.1 was applied for 5DS (one DLE1 site per unit) while more relaxed 0.19 tolerance was used for 1BS and 6BS data generated with Nt.BspQI enzyme (two merging recognition sites in the unit). Detected repeats were quantified and their unit size and frequency in the dataset were plotted in a histogram for visual analysis. During chromosome flow sorting, the telosomics targeted in our study were contaminated by mixture of other chromosomes/arms, totalling 9.6-19.7% of the sorted fraction (Table 1), which was considered in recalculating the proportion of rDNA units per pure chromosomal fraction. Since the 1BS, 6BS and 5DS arms differ in size significantly (Table 1), their mutual contamination was unlikely and was not considered in the recalculation.

Reconstruction of wheat rDNA units

A wheat consensus and three chromosome-specific 45S rDNA units were reconstructed from published short-read data using *RepeatExplorer* pipeline (Novák et al., 2013), as previously described in Kapustová et al. (2019). The consensus sequence was generated from whole-genome Illumina data of wheat Synthetic W7984 (SRP037990, Chapman et al. 2015) while the chromosome-specific units were reconstructed from Illumina reads for 1BS (ERX250504, IWGSC 2014) and Roche/454 reads for 6BS (DRX007672, Tanaka et al. 2013) and 5DS arms (ERA296180, Akpinar et al. 2015). For the unit reconstructions, a random data set of 4,471,956 filtered reads, corresponding to 0.13x genome coverage, was used for the wheat rDNA consensus, and reads corresponding to 0.15-0.77x coverages of particular chromosome arms were applied for the chromosomal consensuses. If a chromosomal rDNA unit was not reconstructed to a contiguous sequence after one round of a graph-based

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

clustering, the available contigs were used to extract all putative reads homologous to 45S rDNA using BLASTN. A second round of clustering was done on a data set enriched for putative rDNA reads and lacking other abundant repeat types. Finally, particular components of the 45S rDNA unit (genes and spacers) were ordered and annotated according to a rye 45S rDNA (JF489233.1, Fluch et al., 2012) as shown in Suppl. Table S2.

Proposal and validation of diagnostic SNPs

Chromosome-arm-specific SNPs were identified after aligning consensual rDNA sequences of 1BS, 6BS and 5DS in Geneious v7.1.2 (http://www.geneious.com). Diagnostic value of all proposed SNPs was tested by cross-mapping of available telosomic-derived Illumina reads (IWGSC, 2014) from 1AS (ERX250501), 1BS (ERX250504), 6BS (DRR008486) and 5DS (ERX250533) to individual chromosomal rDNA consensuses. Representation of particular SNPs in three wheat 'Chinese Spring' assemblies (SRX2994097, Zimin et al., 2017; ERX1700146, Clavijo et al., 2017; SRX3059308, IWGSC, 2018) was assayed by mapping whole-genome Illumina reads generated in particular projects to the wheat consensual rDNA unit. Reads were filtered for quality and length and aligned to wheat rDNA consensus sequence using Bowtie2 aligner with default parameters (Langmead et Salzberg, 2012). The resulting sam files were converted to bam files, sorted by SAMtools package version 1.6 and 0.1.18, respectively and viewed in Integrative Genomics Viewer (IGV) v2.7.2, which was also used to quantify particular SNP variants. To verify proposed locus-specific haplotypes, we exploited merged 2x250bp (PE450) reads from whole-genome libraries, generated for the IWGSC RefSeq v1.0 genome (IWGSC, 2018), and aligned them to the wheat rDNA consensus. Reads overlapping all three diagnostic SNPs were extracted by SAMtools and checked for haplotype constitution using JVarkit Biostar214299 utility (Lindenbaum 2015).

RNA-seq

Five plant tissues - embryo, coleoptile (including plumule), root tips, primary leaf and mature leaf - in two biological replicates for each were used for the transcriptomic analysis. Seeds were incubated for 5 days at 4°C in Petri dishes on wet cellulose covered by filter paper, followed by incubation at 26°C in the dark for 4-6 hours for embryo and 24 hours for coleoptile and root tips. After 24-hour incubation at 26°C, five germinating seeds were planted in garden soil and maintained in a growth chamber under long-day (16 hours' day light) conditions and 20°/16°C for five days or until the flag leaf was visible (Zadoks' stage 37) to collect the primary and the mature leaf, respectively. Three embryos, 2-5 coleoptiles, 10 root tips about 2 mm in length and up to 100 mg of the primary leaf tissue and the mature

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

leaf blade, respectively, were collected per sample and flash frozen in liquid nitrogen. Total RNA was isolated using RNEasy Plant Mini Kit (Qiagen, Germany) and RNA quality was checked on Bioanalyzer using Agilent RNA 6000 Pico Kit (Agilent Technologies, Santa Clara, USA). Sequencing libraries were prepared by NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, USA) and pair-end sequenced on the Illumina NextSeq550 System. Summary of the sequencing output is provided in Suppl. Table S4.

Calling of transcript variants

Raw RNA-seq data were quality controlled using fastqc, trimmed with trimgalore/0.6.2 using default settings and rid of chloroplast and mitochondrial DNA by mapping to the respective references (Middleton et al., 2014, accession number KC912694.1; IWGSC, 2018, accession number MH051716.1) using HiSat2 v2.1.0 (Suppl. Table S4). Two million (one million of pairs) of clean reads were used as an input for SNV calling. Out of these 90-96 % (Suppl. Table S4) were mapped to the wheat rDNA consensus unit, which included the ETS-18S-ITS1-5.8S-ITS2-26S subregions (Suppl. Figure S2a). Mapping was carried out using commands in the CLC genomics workbench (Qiagen, Germany) with the following parameters: Match score – 1, mismatch cost – 2, insertion cost – 3, deletion cost – 3, length fraction – 0.5, similarity fraction – 0.8. Read tracks were visually checked in the program window and coverage graphs were constructed.

Variants were called via the 'Probabilistic Variant Detection' function tool in CLC using default settings. SNPs were filtered as follows: minimum read coverage – 400, count (the number of countable reads supporting the allele) - 40, frequency (the ratio of "the number of 'countable' reads supporting the allele" to "the number of 'countable' reads covering the position of the variant"): ≥10% (high frequency SNPs).

Mapping of Iso-Seq data

In order to identify possible rare transcripts originating from 5DS or other minor rDNA loci, available wheat CS Iso-Seq data from PRJEB15048 (Clavijo et al., 2017) were mapped to the wheat rDNA consensus and 5DS rDNA unit, respectively. Total of 817,892 high-quality (CCS) Iso-Seq reads, divided into six sets corresponding to six plant developmental stages - leaf, root, seed, seedling, stem and spike - were mapped using minimap2 software with default parameters. Resulting sam files were converted into bam files and sorted using SAMtools. The data were visualised and variant proportion was estimated using IGV.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Relative rRNA quantification

A hybrid reference wheat genome was generated by concatenating repeat- and rDNA-masked IWGSC RefSeq v1.0 genome sequence, the wheat consensus 45S rDNA unit, and the wheat chloroplast (KC912694.1) and mitochondrial (MH051716.1) genomes. Trimmed RNA-seq data were mapped to this reference using HISAT2 v2.1.0 (-k 1), followed by running the featureCounts function from the package Subread v1.5.2 in a paired-end mode. Per-genome counts were obtained using counts and lengths of corresponding wheat nuclear (18S, 26S), chloroplast (16S, 23S) and mitochondrial (18S, 26S) rRNA genes. For mitochondria, comprising three and two non-clustered copies of 18S and 26S rRNA genes, respectively, counts for each of the genes were summed. Other rRNA types were not considered. DESeq2 (Love et al. 2014) was used for relative quantification of the nuclear/chloroplast/mitochondrial 18/16S and 26/23S rRNA gene transcripts using embryo as a reference tissue as well as for calculating the FPKM (fragments per kilobase of exon per million mapped reads) values. The values from the two replicas of each sample were averaged for the final plots and table.

Results

Positioning and characterisation of rDNA arrays

Using bread wheat whole-genome Illumina reads (Chapman et al. 2015) and *RepeatExplorer* pipeline, we reconstructed a wheat consensus 45S rDNA unit with the length of 8193 bp (Suppl. Data 1). We searched the unit for GCTCTTC and CTTAAG motifs, which are the recognition sites of Nt.BspQI and DLE-1 enzymes, respectively, used to label DNA molecules on optical mapping platforms of Bionano Genomics. The rDNA unit comprised one CTTAAG and two GCTCTTC sites, all of them located in the 26S rRNA gene. Since the two GCTCTTC motifs are positioned just 1126 bp apart, they could not be spatially discriminated on the Irys platform that we used to generate the BspQI maps, and were expected to be recorded as a single label. Thus, each of the enzymes was predicted to introduce one label per unit and an array of tandemly organized rDNA units was expected to generate a regular labelling pattern with a spacing of 8-9 kb.

To verify this prediction and to visualise particular rDNA arrays in their genomic context, we constructed chromosome-arm-specific Bionano optical maps for wheat 1BS and 6BS chromosome arms, which bear the major rRNA multigene loci, and for 5DS and 1AS arms, carrying minor rRNA loci. Parameters of the chromosomal maps are given in Suppl. Table S1. The optical map contigs were aligned to the IWGSC RefSeq v1.0. Among the 369 1BS OM contigs, we found several comprising a regular label pattern with label spacing of 8.8-10

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

kb. Only 1BS contig 447, having a longer segment with an irregular pattern, could be hereby aligned to the 1BS pseudomolecule at 80.7 – 81.5 Mb (Figure 1). A non-aligned 200-kb part of this contig had a regular pattern with 9.4-kb label spacing that overhung towards the telomere, indicating a presence of an rDNA array at this position. This presumption was supported by the presence of a scaffold boundary at 80.65 Mb and of a cluster of incomplete rDNA units between 80.65-80.88 Mb. Another cluster of rDNA fragments was found at 111.1-111.2 Mb of the 1BS pseudomolecule but this region was continuously covered by an OM contig without a regular pattern, excluding presence of a regular rDNA array at this position. We analogously assigned rDNA arrays to positions 111.58 Mb and 19.01 Mb of 6BS and 5DS pseudomolecules, respectively. In both the 6BS and the 5DS, we found rDNA-comprising OM contigs locating the rDNA arrays from both the distal and the proximal side (Figure 1). The 5DS optical map indicated that 5DS scaffold 34514 (5D:19005535..19736721 bp), flanking the rDNA array at the proximal side, was misoriented in the RefSeqv1.0 pseudomolecule, artificially breaking the rDNA array into two parts. Clusters of incomplete rDNA units flanking the rRNA multigene loci indicated that the 1BS and 6BS rRNA genes were oriented from the centromere to the telomere while the 5DS rRNA locus had the opposite orientation. In all these loci, clusters of rDNA fragments were found only upstream of the regular arrays while repeat-rich gene deserts were characteristic of the downstream regions. The rDNA-bearing OM contigs showed a highly regular pattern in 1BS and 5DS but the regularity has been disrupted in several positions by label-free gaps of tens of kilobases in the 6BS, suggesting that the rDNA array has been invaded by larger blocks of transposable elements or interspersed by shorter arrays of unlabelled tandem repeats. We also searched for the characteristic pattern in OMs of the 1AS arm but failed in finding a single contig with a clear ~9-kb pattern. Only 1AS OM contig 42 comprised segments resembling the rDNA pattern with more condensed spacing (around 8.6 kb) and some irregularities (Figure 1). This contig aligned with a high confidence to the start of the 1A pseudomolecule, thus confirming correctness of the sequence assembly in this region. Closer inspection of the DNA sequence revealed a presence of 29 rDNA units comprising all of 18S, 5.8S and 26S rRNA genes, featured by high sequence and length diversity and variable orientation (Suppl. Figure S1a, c). This degenerated rDNA array spanned over 430 kb. Ribosomal DNA units were interspersed by rDNA fragments and telomeric repeats of diverse orientation (Suppl. Figure S1b), indicating a dynamic character of the region and a loss of functionnMajority of molecules involved in the assembly of the distal part of the OM contig 42 terminated with label-free overhangs, suggesting possible presence of a regular telomeric sequence in immediate vicinity of the rDNA array.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Quantification of rDNA units in arrays

The optical map contigs comprised segments with rDNA-specific patterns spanning over 200-370 kb, but it was obvious that these values did not correspond to the full lengths of the rDNA arrays, but rather reflected the sizes of DNA molecules used for optical map assembly, which did not exceed 500 kb. Nevertheless, an approximate quantification could be done by the IrysView software, which includes algorithms that calculate proportion of nucleotides comprised in tandemly organised repeats of a particular size. The results of repeat analysis done by this software feature revealed that repeats of 8-10 kb, corresponding to the size of wheat 45S rDNA units, were the major labelled tandem repeats in chromosome arms 1BS, 6BS and 5DS, while representation of other labelled tandem repeats in arrays larger than five units was negligible (Figure 2). Consequently, we performed quantification of tandem-organized rDNA units in each rRNA locus, employing information on percentage of nucleotides in repeats, sizes of the telosomics and purities of particular sorted fractions (Table 1). Our analysis indicated that the largest array of rDNA units was located in the 6BS chromosome arm (2813 units, 26.87 Mb), followed by arrays in 1BS (1378 units, 12.96 Mb), 5DS (170 units, 1.63 Mb) and 1AS (29 complete units, 0.43 Mb), being the smallest locus. The total of 4390 units occupied approximately 42 Mb of the wheat CS genome space. Relative quantification of the units assigned 31.4 and 64.1% of them to the major NORs in 1BS and 6BS, respectively. The remaining 3.9 and 0.7% belong to the minor loci in 5DS and 1AS, respectively.

Table 1. Quantification of 45S rDNA units

Quantification of arrayed 45S rDNA units from optical map raw data

Chromosome arm	1BS	6BS	5DS	1AS**	Total
Arm size (Mb)*	314	415	258	275	
% contaminating chromosomes	15	19.7	15	9.6	
% nucleotides in repeats	3.51	5.2	0.54		
Mb in repeats - contaminated sample	11.2	21.58	1.39		
Mb in repeats - pure sample	12.96	26.87	1.63	0.43	41.89
45S rDNA unit size (kb)***	9.4	9.55	9.56	8.6	
No. 45S rDNA units in arrays	1378	2813	170	29	4390
% total number units in arrays	**31.4**	**64.1**	**3.9**	**0.7**	

* Arm sizes according to Šafář et al., 2010

** 1AS rDNA quantified from the sequence

*** Unit size estimated from the optical map

Quantification of 26S rDNA haplotypes across 5DS-1, 6BS-2 and 1BS-1 SNPs

Chromosome arm	1BS	6BS	5DS	1AS + others	Non-specified
Haplotype	CGT	CTC	TGC	CGC	Other combinations
No. haplotypes	2177	3667	240	135	21
% haplotypes	**34.9**	**58.8**	**3.8**	**2.2**	**0.3**

Haplotypes are discriminated based on SNPs in positions 5300 bp, 5348 bp and 5533 bp of the wheat rDNA consensus sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Chromosome-specific rDNA units and diagnostic SNPs

Using *RepeatExplorer* pipeline, we reconstructed chromosome-specific consensual units from DNA sequences of telosomes 1BS, 6BS and 5DS, carrying the regular rDNA arrays. This resulted in units of 8326, 7910 and 7887 bp for 1BS, 6BS and 5DS arms, respectively (Suppl. Data 1). The length differences were due to the diversity in intergenic spacer (IGS), which is composed of a variable number of tandemly organised repeats, hampering an accurate assembly of the IGS from short-read data. The lengths of 18S, 5.8S and 26S rRNA genes and of internal transcribed spacers ITS1 and ITS2 were identical. Mutual comparison of the chromosomal consensual sequences revealed a high sequence diversity in IGS and several SNPs in both ITS1 and ITS2. While the 18S and 5.8S rRNA gene sequences were identical among the three telosomes, total of eight putative arm-specific SNPs were identified in the 26S rRNA gene: two 6BS-specific, two 1BS-specific and four 5DS-specific. Diagnostic value of all these SNPs was tested by cross-mapping Illumina reads from 1AS, 1BS, 6BS and 5DS to the three chromosomal consensuses. This assay led to the selection of two diagnostic SNPs for each rRNA multigene locus. The 5DS-2 SNP variant was also found in 1A and other minor rDNA clusters outside the B genome (Suppl. Table S2), thus we designated it as “non-B”. Further, we analysed the frequency of allelic variants of the proposed diagnostic SNPs in the genome of wheat cv. Chinese Spring by investigating raw data that had been used for RefSeq v1.0 assembly (IWGSC, 2018), and also Illumina reads used for another two CS genome assemblies (Triticum 3.1, Zimin et al., 2017; TGACv1, Clavijo et al., 2017) (Suppl. Table S3). Pairs of SNPs featuring 1BS and 6BS, respectively, had highly consistent representations in IWGSC RefSeq v1.0 and Triticum 3.1 data, while showing slight inconsistencies in TGACv1, perhaps due to a smaller number of available reads. The RefSeq v1.0 analysis showed that the 1BS, 6BS and 5DS repeat unit variants represented 36.2, 56.6 (averaged values for 1BS-1+2 and 6BS1+2) and 4.2% (5DS-1) of the whole-genome data, respectively. Similar representation of particular chromosomal loci has been found in raw data of the other CS projects.

To verify the proposed locus-specific haplotypes, we made use of merged 450-bp Illumina pair-end reads, generated for the IWGSC RefSeq v1.0 assembly, and aligned them to the wheat consensus rDNA unit. Visualisation of the alignment in Geneious v7.1.2 enabled inspection of haplotypes in the merged reads (Figure 3). We found 6420 merged 450-bp reads covering triplet of chromosome-specific SNPs 5DS-1, 6BS-2 and 1BS-1 that enable discrimination of haplotypes proposed for 1BS, 6BS and 5DS, respectively. Out of the 6420 haplotypes, 2177 (34.9%) corresponded to the predicted allelic constitution of 1BS, 3667 (58.8%) to 6BS and 240 (3.8%) to 5DS. Another haplotype, observed in 135 (2.2%) reads was found in most units of the 1AS array and in small clusters of degenerated rDNA units in

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

7DS and other arms, which besides that comprised units with the 1BS haplotype. Only 21 (0.3%) 450-bp reads had a different combination of SNPs (Table 1). The results of the haplotype analysis thus supported validity of the proposed diagnostic SNPs and the chromosome-specific haplotypes.

Locus-specific transcription of 26S rDNA

To assess locus-specific transcription, we used short Illumina reads from deep RNA sequencing and long PacBio reads (Iso-Seq data) of low sequence coverage. A typical profile of rRNA-seq read coverage along the rDNA unit is shown in Suppl. Figure S2b. As expected, the read coverage was high in genic regions and low in the non-coding regions. Less than 0.1% reads were mapped to the ETS region.

The analysis of high (>10%) frequency variants in short Illumina RNA-seq data obtained from five tissues – embryo, coleoptile, root tip, primary leaf and mature leaf – confirmed the presence of significant transcript variants at the diagnostic SNPs in both major rDNA loci in the B genome, namely 6BS-1 (5090 bp in wheat rDNA consensus), 6BS-2 (5348), 1BS-1 (5533) and 1BS-2 (8010), totally accounting for >90% of variation in wheat rRNA. In all tissues, the 1BS rRNA variants predominated those of the 6BS in ratio roughly 2:1, which slightly varied among tissues (Suppl. Table S5, Figure 4a). The highest frequency of 6BS-specific variants was in the coleoptile (35 and 37.6% in positions 5090 and 5348, respectively) while the highest frequency of the 1BS-specific variants was in the mature leaf (71.9 and 75.1% in positions 5533 and 8010, respectively).

Our failure to identify allelic variants at 5DS-1 and 5DS-2 SNP positions suggested the absence of transcription outside the major loci located in the wheat B genome. To validate this finding, we made use of published Iso-Seq data of CS wheat (Clavijo et al. 2017) with reads of several kilobases in length, having potential to reliably identify possible 5DS transcripts, and analysed them in the IGV. A majority of transcripts in all datasets started at position 2351 bp of the rDNA consensus, which was previously identified as a transcription start site (Vincentz and Flavell, 1989), and the rRNA transcripts did not extend beyond 8154 bp (Suppl. Figure S3a), preceding a polypyridine CCCTCCCCC tract. The datasets for particular developmental stages (leaf, root, seed, seedling, stem and spike) differed in read length and amount of data mapped to the 45S rDNA unit. The most beneficial showed the dataset of “Seed” comprising 1951 reads mapping to the 5DS rDNA unit, which was used to search for transcripts of 5DS and possibly other minor rRNA loci. In five out of 1079 „Leaf“ Iso-Seq reads mapping to position of the 5DS-1 SNP in the 5DS rDNA unit, we could unambiguously recognize the 5DS-specific haplotype based on the 5DS-1, several 5DS-specific SNPs in ITS2 and in two cases also 5DS-2 and 1BS-2 SNPs (Suppl. Figure S3b,c).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Additional two reads out of 425 mapping to the terminal part of the 26S rRNA gene, covering 5DS-2 and 1BS-2 but not 5DS-1, suggested a non-B origin of the transcripts, and another three reads, covering the entire 26S rRNA gene, had a non-B non-5DS combination of SNP alleles indicating possible 1A origin. To summarize, we observed among the "Seed" transcripts of CS wheat possible non-B variants of 26S rRNA with frequency around one percent.

Relative rRNA quantification

Relative quantification of nuclear, chloroplast and mitochondrial ribosomal RNA across five tissues was based on FPKM estimates for the large rRNAs (16/18S and 23/26S) and was calculated from the mapped total-RNA-seq datasets. The overall rRNA synthesis culminated in the primary leaf, boosted by increased transcription from both the nucleus and the chloroplasts (Figure 4b, Suppl. Table S6). High level of chloroplast rRNA synthesis persisted in the leaf until maturity (analysed at flag leaf appearance) while the production of nuclear rRNA has been decreased at that stage. Contribution of mitochondria to the total rRNA production was negligible in all tissues. A differential analysis using embryo as a reference tissue (Suppl. Figure S4, Suppl. Table S7) confirmed the dramatic increase of chloroplast rRNA types in the leaf samples, in line with the proliferation of chloroplasts and a high photosynthetic activity. Also production of the mitoribosomal RNA and the nuclear rRNA were most increased in the primary leaf (Suppl. Figure S4). Root tip and mature leaf samples had slightly increased expression of nuclear rRNA genes relative to the embryo.

Discussion

Instability in rDNA copy number has been well documented in both plants and animals at the inter-population and even inter-individual levels (Rogers and Bendich, 1987; Rabanal et al. 2017, Nelson et al. 2019). In our study, we used a new approach to quantify 45S rDNA in the genome, which was based on calculating tandemly organised repeats of corresponding size in optical map raw data. Our total estimate - 4390 rDNA units in arrays - was remarkably similar to that of Flavell and O'Dell (1976) who, based on classical filter hybridisation techniques, estimated 4575 copies per haploid genome in wheat cv. Chinese Spring. The accordance between both studies suggests robustness of both quantification approaches and points to a relative stability of rDNA copy number, even between standard karyotype of CS wheat and derived telosomic lines used in our study, which have undergone multiple meiotic cycles since their generation in 1970's (Sears and Sears, 1978). Nevertheless, a recent study of Handa et al. (2018) reported 11,160 rDNA units in the CS genome, which is roughly 2.5 higher estimate than that of Flavell and O´Dell (1976) and ours. The incongruence between the studies is explained by different approaches used for rDNA

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

quantification rather than inter-individual variability. The hybridisation techniques used by Flavell and O'Dell (1976) are likely to pick up only longer DNA sequences while q-PCR used by Handa and co-workers targeted short fragments of a few hundred base pairs. The optical mapping applied in the current study considers only units organised in arrays, which are likely to be functional, while q-PCR cannot discriminate the arrayed genes from the pseudogenised ones that are dispersed, truncated or recombined. Of note, BLAST search for the consensual rDNA unit in the IWGSC RefSeq v1.0 genome, carried out in Ensembl Plants (https://plants.ensembl.org/Triticum_aestivum/Tools/Blast), provided 6089 hits distributed evenly on all wheat chromosomes (Suppl. Figure S5). Vast majority of them were located outside arrays and represent rDNA fragments. Taking this into account, we consider our approach more relevant for assessing sequences with potential to contribute to rRNA production. Still we admit that our predicted number of units might be a slight underestimate due to presence of false negative/positive labels occurring in the raw data and possibly disrupting the regularity of the label pattern.

A relative proportion of loci from particular chromosomes was determined both from the optical map data and from representation of locus-specific haplotypes in whole genome data (Table 1). Our results obtained by both approaches were concordant, except of a slight shift (34.9 vs. 31.4%) in favour of 1B in the haplotype data. This may reflect the fact that the haplotype analysis does not exclude the non-arrayed units and larger (> 450 bp) unit fragments dispersed across the genome where the 1BS haplotype predominates, likely because of a higher transcription activity of the locus. The 1BS : 6BS ratio of 31.4 : 64.1, is in a good agreement with 30/30.5 : 60/60.9 obtained in previous studies (Flavell and O'Dell, 1976; Handa et al. 2018) but our OM data ascribe a smaller percentage to other chromosomes (4.6 vs. 10/8.6%), for the reasons explained above.

Aiming to analyses locus/chromosome-specific transcription, we identified several SNPs in the 26S rDNA that can reliably discriminate the chromosome-specific haplotypes. Their chromosomal specificity was validated by analysing sequence reads obtained from flow-sorted chromosome arms, which showed that 94-99% reads obtained from both 1BS and 6BS carried the expected SNP variant. This proportion was slightly lower (86-87%) for the 5DS, probably due to presence of dispersed units/fragments that bear mostly the B-genome SNP variants and influence the result more here than in the chromosome arms harbouring the larger arrays. Availability of locus-specific SNPs along the 26S rRNA gene can be beneficial for epigenetic analyses of particular loci. Occurrence of haplotypes with a high locus specificity.is not supportive of frequent interchromosomal conversion events between particular arrays, which is in agreement with findings of Lassner et al. (1987) who did not find evidence of homogenisation between B- and D-genome rDNA based on comparing IGS

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sequences from the 6B and 5D loci. On the other hand, Handa et al. (2018), analysing part of the IGS (3' ETS region), identified two rDNA subtypes that are shared between 1B and 6B chromosomes, suggesting interchange between the two chromosomes.

We carried out transcriptomic analysis across five wheat tissues with the aim to identifiy differential transcription, observed in some allopolyploid species of other genera (*Brassica*, *Tragopogon*, *Solanum*). Our effort to distinguish locus-specific transcription was based on validated chromosome-specific SNPs located in the 26S rRNA gene, which was well covered by reads, in contrast to the 3' ETS region, used in the study of Handa et al. (2018), to which mapped a negligible part of both RNA-seq and Iso-Seq reads. Our results for 1B- and 6B-specific transcription supported findings of Martini and Flavell (1976) who measured volumes of both types of nucleoli and predicted 2:1 transcription ratio for 1B and 6B loci, respectively. Since Handa and co-workers analysed rRNA subtypes rather than transcripts of particular arrays, comparison of both studies is unfeasible except for their subtype 3, which was mostly present on chromosome 5D and showed negligible transcription in a standard CS wheat, similarly to our findings for 5DS-specific transcripts. Both Handa and co-workers and Martini and Flavell observed activation of the minor loci, mainly Nor-D3 in 5D, after removal of one of the major NORs. This is in agreement with our observation that the 5DS locus is represented by a regular array of compactly organised rDNA units, which suggests their full functionality. We did not find clear evidence for 26S rDNA expression from the 1AS locus that is featured by large scale rearrangement and truncated copies, indicating extensive pseudogenisation. This suggests that silencing of this locus is most likely permanent. We do not know whether pseudogenisation was a cause or a consequence of long term inactivity. However, presence of multiple telomeric repeat insertions (Suppl. Figure S1) evidences that the locus has been subjected to double strand breaks and healing with telomerase in its evolutionary history. In any case, several non-truncated units seem to be present, harbouring shorter IGS than in other loci. Given that lengths of IGS were correlated with transcription activity in wheat (Sardana et al. 1993), it is likely that the 1AS units became underdominant in an early hybrid and that the rearrangements arrived later in evolution, which is supported by observation of Guo et al. (2014) that the A genome loci were silenced in synthetic lines of wheat and eventually lost in subsequent generations. That study clearly demonstrated nucleolar dominance of the wheat B genome in newly formed allopolyploids. However, reasons for predominance of the 1B over the 6B transcription has not been clarified yet. Handa et al. (2018) suggested a regulation mechanism based on rDNA sequence difference because their predominantly expressed subtype was found on both 1B and 6B chromosomes. In our study, the higher expression was clearly associated with the 1B chromosome, which allows

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

speculating about alternative mechanisms of the regulation, such as involvement of trans factors or position of the loci in the 3D nucleus.

In our transcriptomic study across five tissues, we observed partial developmental silencing of the 6B locus. Our hypothesis that contribution of the secondary (6B) locus raises in tissues with increased demand on rRNA production has not been confirmed; mature leaf, showing the lowest proportion of 6BS transcripts, had a comparable relative expression of 26S rRNA genes as root tip and coleoptile samples. An alternative explanation for the transcription dynamics lays in a faster turnover of the 6BS rDNA transcripts in the mature leaf. This option is supported by the fact that the tissue-dependent dynamics of 6BS transcripts correlates with proportion of 45S rDNA-mapped reads in the total RNA-seq dataset rid of mitochondrial and chloroplast RNA (Suppl. Table S4); mature leaves had the lowest (90.2%) proportion of 45S rDNA-mapping reads while coleoptile had the highest one (96.4), in line with proportion of the 6BS-specific 26S rRNA variant. This suggests that the proportional loss of rRNA to total nuclear RNA in leaf tissues might be due to loss of the 6BS variant, which may undergo a faster transcript turnover / posttranscription modifications that prevent mapping of the 6BS rRNA reads to the consensus rDNA unit.

Data Accessibility

Raw RNA-seq data for five tissues are available in the SRA under BioProject PRJNA657991. Optical maps of 1AS, 1BS, 6BS and 5DS wheat chromosome arms (bnx and cmap files) are available on request.

Acknowledgements

We are grateful to M. Mascher for sharing merged 450-bp Illumina reads of wheat. We thank to Zdeňka Dubská, Jitka Weiserová and Eva Jahnová for excellent technical assistance. The work was supported by the Czech Science Foundation (grant No. 17-17564S) and by ERDF project "Plants as a tool for sustainable global development" (No. CZ.02.1.01/0.0/0.0/16_019/0000827). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures, and by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

Competing Interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure Legends

Figure 1. Positioning of 45S rDNA arrays by optical maps.

Optical maps (blue bars) generated from rDNA-bearing chromosomes 1AS, 1BS, 6BS and 5DS were aligned to corresponding pseudomolecules of IWGSC RefSeq v1.0 wheat genome (green bars), digested *in silico* with enzymes used to construct the optical map. The vertical lines represent BspQI (red) and DLE-1 (blue) recognition sites labelled in the optical map. Regular pattern with 8.4-9.6-kb spacing (highlighted by a yellow line) reveals presence of a tandemly organized repeat of 45S rDNA. Only the irregular rDNA array in 1AS is present in the sequence while the regular arrays in the other chromosomes are missing in the pseudomolecules. The relevance of the arrays for the rDNA has been supported by presence of rDNA fragments found in the flanking genome regions (highlighted by violet lines). Scaffold 34514, flanking the 5DS rDNA array on the distal side, is misoriented in the RefSeq v1.0 assembly. The red arrows indicate scaffold boundaries in the reference genome.

Figure 2. Quantification of 45S rDNA units in arrays from optical map data.

(a) Visualisation of a single molecule with array of 45S rDNA units on the Irys platform. (b)-(d) Plots for number of repeat units vs. repeat unit size in optical-map single-molecule data for chromosome arm 1BS (b), 6BS (c) and 5DS (d). Peaks corresponding to the size of a 45S rDNA unit in particular chromosome arm are indicated by red arrows.

Figure 3. Analysis of chromosome-specific haplotypes in 45S rDNA.

Merged 450-bp Illumina reads of wheat cv. Chinese Spring were mapped to interval 5080-5570 bp in the 45S rDNA consensus (top), comprising diagnostic SNPs 6BS-1 (5090 bp), 5DS-1 (5300 bp), 6BS-2 (5348 bp) and 1BS-1 (5533 bp). The SNPs are highlighted blue. Note the consistency between the two 6BS-specific SNPs. Region 5300-5532 bp, comprising 5DS-1, 6BS-2 and 1BS-1, is sufficient for haplotype discrimination (bottom).

Figure 4. Tissue-specific contribution of particular loci to rRNA production.

(a) Representation of 6BS-specific (yellow) and 1BS-specific (blue) SNP variants in 26S rDNA estimated in raw data of IWGSC RefSeq v1.0 genome and in 26S rRNA from five tissues. (b) Relative contribution of nuclear, mitochondrial and chloroplast genomes to overall rRNA production, expressed as FPKM (Fragments Per Kilobase of exon per Million reads mapped) values for each of the 16S/18S and 23/26S rRNA types. Values from two replicas of each sample have been averaged.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Files

Suppl. Table S1. Optical map statistics

Suppl. Table S2. Frequency of allelic variants in diagnostic SNPs in Illumina reads of flow-sorted chromosome arms

Suppl. Table S3. Frequency of allelic variants in diagnostic SNPs in Illumina reads from three wheat 'Chinese Spring' genome projects

Suppl. Table S4. RNA-seq data

Suppl. Table S5. Variant analysis of 26S rRNA in different tissues

Suppl. Table S6. FPKM values for nuclear, chloroplast and mitochondrial 16S/18S and 23/26S rRNAs across five tissues

Suppl. Table S7. DeSeq2 results table with log2 fold changes and adjusted p-values behind the heatmap and its color-code

Suppl. Figure S1. Degenerated ribosomal DNA array in the terminal part of chromosome arm 1AS. (a) The 1AS rDNA array is composed of both complete (comprising all of 18S, 5.8S and 26S rRNA genes; blue boxes), and incomplete units. Separated18S and 26S rRNA genes and gene fragments are marked purple and yellow, respectively. (b) Ribosomal DNA units (blue boxes) are interspersed by telomeric repeats (red triangles) of various orientation. (c) Comparison of 29 complete 45S rDNA units building the 1AS array shows a high degree of variability. Grey bars represent individual rDNA units, disagreements are highlighted by black lines. Gene annotation is shown above the alignment.

Suppl Figure S2. Wheat consensus 45S rDNA unit and its coverage by RNA-seq reads (a) Scheme of the consensus rDNA unit. Coding regions are depicted as brown arrows. Putative transcription start site is indicated in the position 1141 bp upstream from the 18S rRNA gene. (b) RNA-seq read coverage plot along the rDNA unit. Note near absence of the ETS/IGS and ITS1/ITS2 transcripts. Both 18S and 26S rRNA genes were relatively equally covered except of a GC-rich subregion close to the 26S rRNA 5'-terminus. (c) Annotation of the rDNA unit.

Suppl. Figure S3. Mapping of IsoSeq reads to wheat consensus and 5DS rDNA units. (a) Mapping of 801 CCS IsoSeq reads from PRJEB15048 (Clavijo et al. 2017), sample „Leaf", to the wheat 45S rDNA consensus. Coverage of the unit by reads is shown in the top, example of reads in the middle and rDNA unit annotation in the bottom of the figure. (b) Total

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of 1951 CCS IsoSeq reads, sample „Seed“, mapped to the 5DS consensual 45S rDNA unit. (c) In five out of 1079 „Seed“ reads mapping to 5DS-1 position, we could unambiguously recognize the 5DS-specific haplotype. A read showing complete match with the 5DS rDNA consensus, both in ITS2 and 26S rRNA, is marked by red asterisk. The colored lines indicate allelic variants with respect to the applied reference.

Suppl. Figure S4. Nuclear, chloroplast and mitochondrial rRNA - Log2 fold change heatmap across five tissues. The DESeq2 dataset clustered heatmap shows log2 fold transformed differences in amounts of 18/16S and 26/23S rRNA in each of the tissues related to the embryo. Color code of the values indicates padj values where red values correspond to $p < 0.05$ that is considered as statistically significant. Blue shades mark increased expression whereas pink marks decreased expression related to the embryonic rRNA as indicated by the scale.

Suppl. Figure S5. Graphical outcome of a BLAST search for sequences homologous to wheat consensus rDNA unit in the IWGSC RefSeq v1.0 genome. The search was carried out in Ensembl Plants (https://plants.ensembl.org/Triticum_aestivum/Tools/Blast). Position of the irregular rDNA array in 1A is marked by a rectangle.

Suppl. Data 1. Sequences of wheat 45S rDNA consensus and chromosome-specific consensual rDNA units

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

Akpinar, B. A. *et al.* (2015) 'Molecular organization and comparative analysis of chromosome 5B of the wild wheat ancestor Triticum dicoccoides', *Scientific Reports*. doi: 10.1038/srep10763.

Chapman, J. A. *et al.* (2015) 'A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome', *Genome Biology*, 16, p. 26. doi: 10.1186/s13059-015-0582-8

Clavijo, B. J. *et al.* (2017) 'An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations', *Genome Research*, 27(5), pp. 885–896.

Crosby, A. R. (1957) 'Nucleolar activity of lagging chromosomes in wheat', *American Journal of Botany*, pp. 813–822. doi: 10.1002/j.1537-2197.1957.tb08268.x.

Darvey, N. L. and Driscoll, C. J. (1972) 'Nucleolar behaviour in *Triticum*', *Chromosoma*, 36, pp. 131-139. doi: 10.1007/bf00285208.

Doležel, J. *et al.* (2018) 'One major challenge of sequencing large plant genomes is to know how big they really are', *International Journal of Molecular Sciences*, p. 3554. doi: 10.3390/ijms19113554.

Dubcovsky, J. and Dvořák, J. (1995) 'Ribosomal RNA multigene loci: nomads of the Triticeae genomes', *Genetics* 140, pp. 1367-1377.

Flavell, R. B. and O'Dell, M. (1976) 'Ribosomal RNA genes on homoeologous chromosomes of groups 5 and 6 in hexaploid wheat', *Heredity*, pp. 377–385. doi: 10.1038/hdy.1976.102.

Flavell, R. B. and O'Dell, M. (1979) 'The genetic control of nucleolus formation in wheat', *Chromosoma*, 71, pp. 135–152. doi: 10.1007/bf00292819.

Fluch, S. *et al.* (2012) 'Sequence composition and gene content of the short arm of rye (*Secale cereale*) chromosome 1', *PloS ONE*, 7(2), p. E30784. doi: 10.1371/journal.pone.0030784.

Gerlach, W. L. and Bedbrook, J. R. (1979) 'Cloning and characterization of ribosomal RNA genes from wheat and barley', *Nucleic Acids Research*, 7(7), pp. 1869–1885. doi: 10.1093/nar/7.7.1869.

Guo, X., Han, F. (2014) 'Asymmetric epigenetic modification and elimination of rDNA sequences by polyploidization in wheat', *The Plant Cell*, 26, pp. 4311-4327. doi: 10.1105/tpc.114.129841

Handa, H. *et al.* (2018) 'Structural features of two major nucleolar organizer regions (NORs), Nor-B1 and Nor-B2, and chromosome-specific rRNA gene expression in wheat', *The Plant Journal*, 96(6), pp. 1148–1159. doi: 10.1111/tpj.14094.

International Wheat Genome Sequencing Consortium (IWGSC) (2014) 'A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome', *Science*, 345(6194), p. 1251788. doi: doi: 10.1126/science.1251788.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Kapustová, V. *et al.* (2019) 'The Dark Matter of Large Cereal Genomes: Long Tandem Repeats', *International Journal of Molecular Sciences*, p. 2483. doi: 10.3390/ijms20102483.

Kubaláková, M. *et al.* (2002) 'Flow karyotyping and chromosome sorting in bread wheat (Triticum aestivum L.)', *Theoretical and Applied Genetics*, pp. 1362–1372. doi: 10.1007/s00122-002-0888-2.

Lam, E. T. *et al.* (2012) 'Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly', *Nature Biotechnology*, 30(8), pp. 771–776. doi: 10.1038/nbt.2303.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Lassner, M. *et al.* (1987) 'Hypervariation associated with a 12-nucleotide direct repeat and inferences on intergenomic homogenization of ribosomal-RNA gene spacers based on the DNA-sequence of a clone from the wheat NOR-D3 locus', *Genome*, 29, pp. 770-781. doi 10.1139/g87-130.

Love, M. I. *et al.* (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Martini, G. and Flavell, R. (1985) 'The control of nucleolus volume in wheat, a genetic study at three developmental stages', *Heredity*, pp. 111–120. doi: 10.1038/hdy.1985.15.

Middleton, C. P. *et al.* (2014) 'Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe', *PLoS ONE*, p. e85761. doi: 10.1371/journal.pone.0085761.

Mukai, Y. *et al.* (1991) 'Physical mapping of the 18S.26S rRNA multigene family in common wheat: Identification of a new locus', *Chromosoma*, pp. 71–78. doi: 10.1007/bf00418239.

Novak, P. *et al.* (2013) 'RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads', *Bioinformatics*, pp. 792–793. doi: 10.1093/bioinformatics/btt054.

Pruitt, R. E. and Meyerowitz, E. M. (1986) 'Characterization of the genome of Arabidopsis thaliana', *Journal of Molecular Biology*, 187(2), pp. 169–183. doi: 10.1016/0022-2836(86)90226-3.

Rabanal F.A. et al. (2017) 'Unstable Inheritance of 45S rRNA Genes in *Arabidopsis thaliana*'. *G3 (Bethesda)*. 7(4), pp.1201-1209. doi: 10.1534/g3.117.040204

Rogers S.O. and Bendich A.J. (1987) 'Heritability and variability in ribosomal RNA genes of *Vicia faba*'. *Genetics*, 117(2), pp. 285-295.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Šafář, J. *et al.* (2010) 'Development of chromosome-specific BAC resources for genomics of bread wheat', *Cytogenetic and Genome Research*, 129(1-3), pp. 211–223. doi: 10.1159/000313072.

Sardana, R. et al. (1993) 'Correlation between the size of the intergenic regulatory region, the status of cytosine methylation of rRNA genes and nucleolar expression in wheat', *Molec. Gen. Genet.* 236, pp. 155–162. doi: 10.1007/BF00277107

Sears, E. R. (1954) 'The aneuploids of common wheat', *Missouri Agricultural Experimental Station Research Bulletin*, 572, pp. 1-59.

Sears, E., Sears, L. (1978) 'The telocentric chromosomes of common wheat', *Proceedings 5th International Wheat Genetics Symposium*, S. Ramanujam, Ed. Indian Agricultural Research Institute, New Delhi, 1, pp. 389–407.

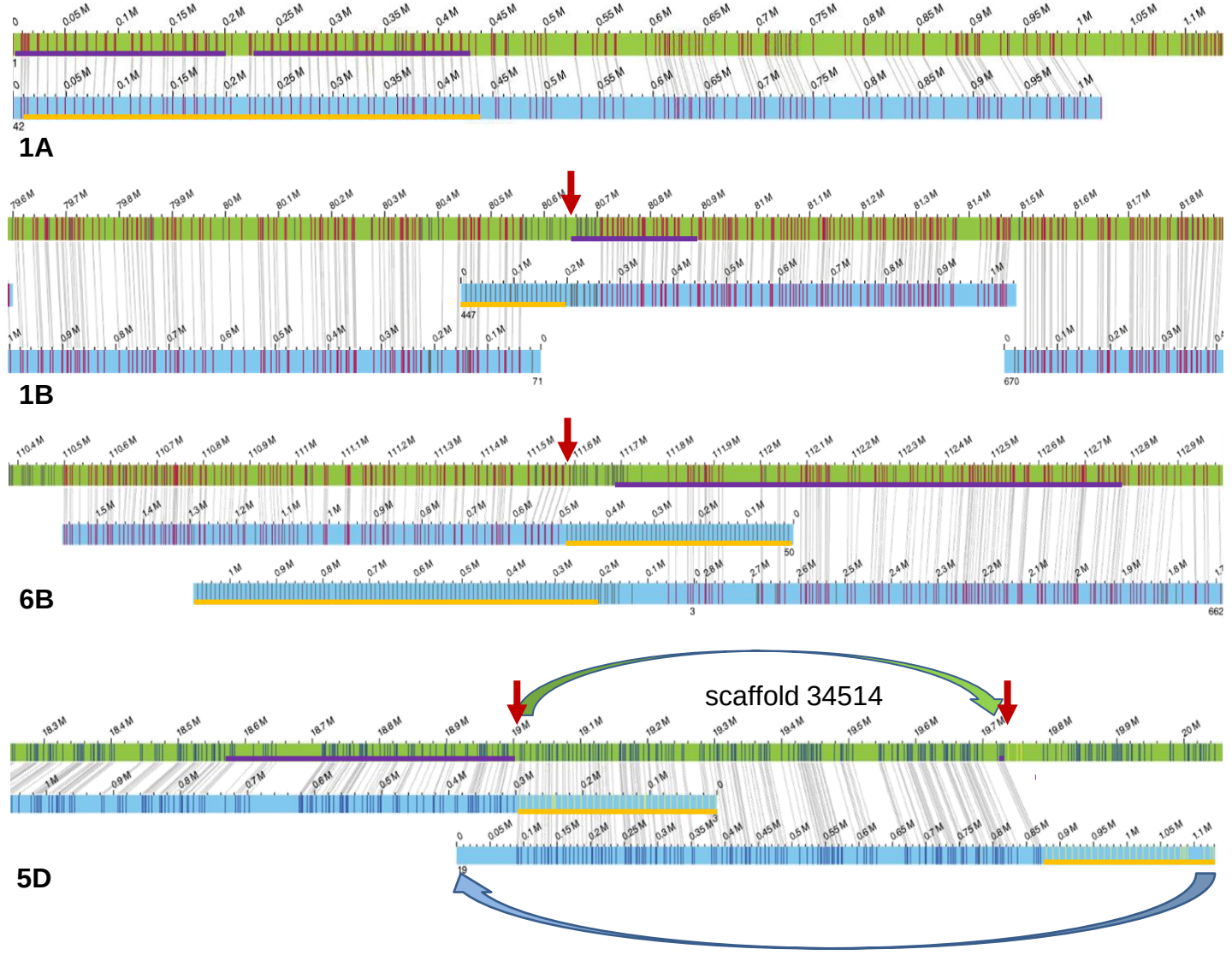
Staňková, H. *et al.* (2016) 'BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes', *Plant biotechnology journal*, 14(7), pp. 1523–1531. doi: 10.1111/pbi.12513.

Symonová, R. *et al.* (2017) 'Higher-order organisation of extremely amplified, potentially functional and massively methylated 5S rDNA in European pikes (Esox sp.)', *BMC genomics*, 18(1), p. 391. doi: 10.1186/s12864-017-3774-7.

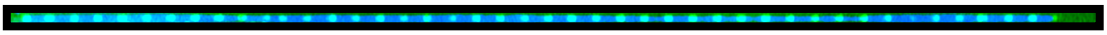
Tanaka, T. *et al.* (2014) 'Next-generation survey sequencing and the molecular organization of wheat chromosome 6B', *DNA Research*, 21(2), pp. 103–114. doi: 10.1093/dnares/dst041.

Todd, M. P. *et al.* (2017) 'Comprehensive definition of genome features in Spirodela polyrhiza by high-depth physical mapping and short-read DNA sequencing strategies', *The Plant Journal* 89(3), pp. 617–635. doi: 10.1111/tpj.13400.

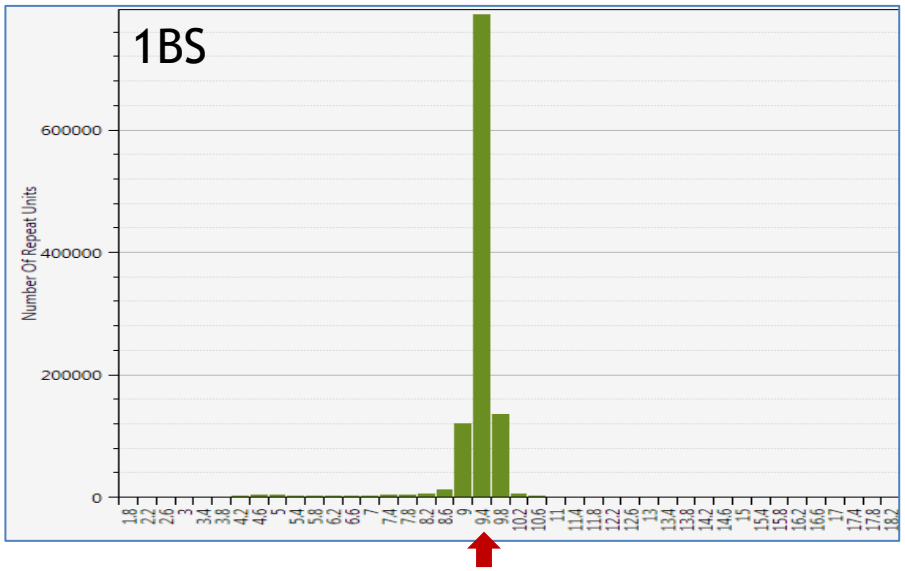
Zimin, A. V. et al. (2017) 'The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*', *GigaScience*, 6(11), pp. 1–7. doi: 10.1093/gigascience/gix097.


1A
1B
6B
5D
scaffold 34514

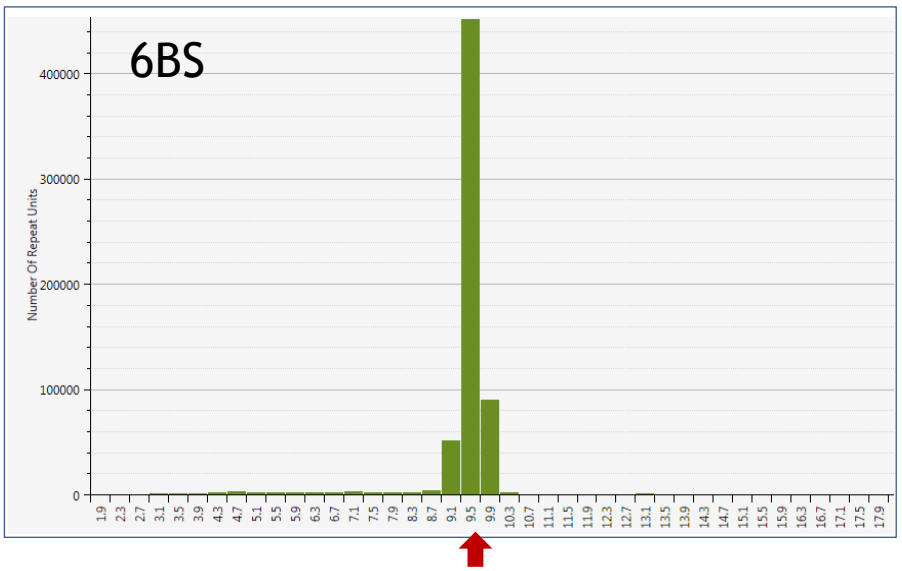
a



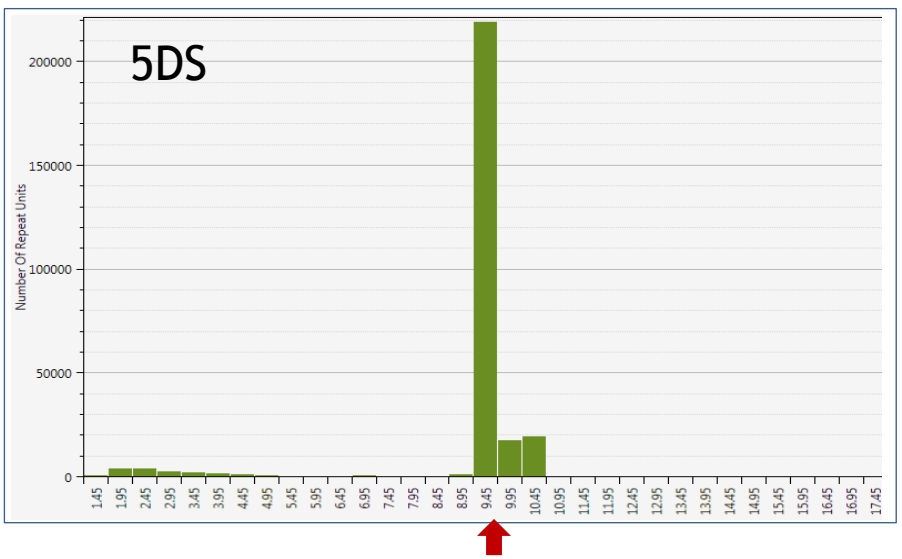
b



c



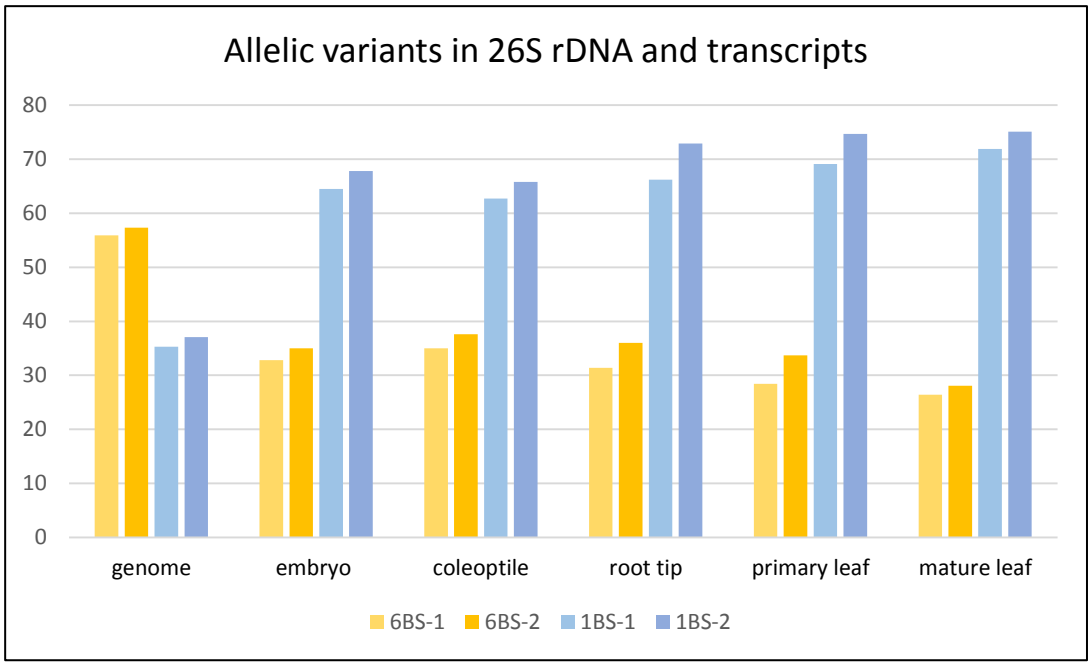
d



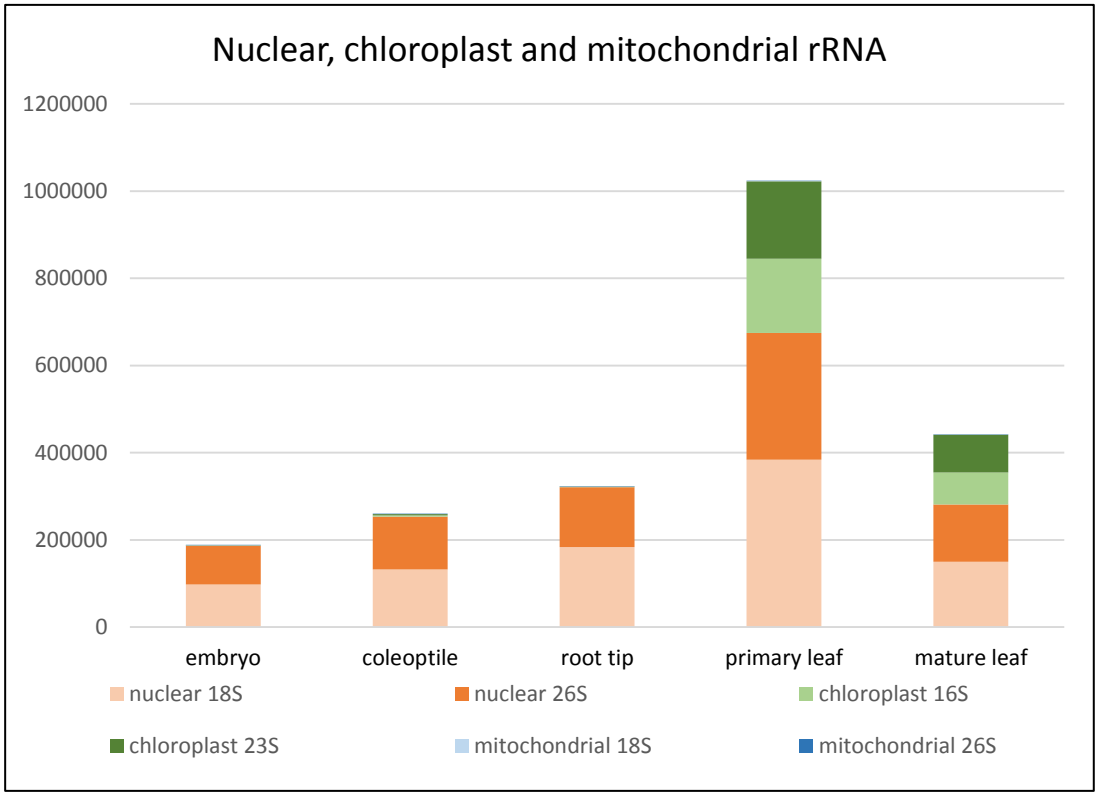
Repeat unit size (kb)

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.29.273623; this version posted August 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

a


Allelic variants in 26S rDNA and transcripts
80
70
60
50
40
30
20
10
0
genome
embryo
coleoptile
root tip
primary leaf
mature leaf
6BS-1
6BS-2
1BS-1
1BS-2

b


Nuclear, chloroplast and mitochondrial rRNA
1200000
1000000
800000
600000
400000
200000
0
embryo
coleoptile
root tip
primary leaf
mature leaf
nuclear 18S
nuclear 26S
chloroplast 16S
chloroplast 23S
mitochondrial 18S
mitochondrial 26S