Analysis of Homeolog Specific Transcription in Hexaploid Wheat

Lars Grønvold

Norwegian University of Life Sciences
Department of Plant Science (IPM)
Master Thesis 60 credits 2013

Acknowledgement

This master thesis completes my two year Master‘s Degree in Bioinformatics at the Norwegian University of Life Sciences. The work was carried out at CIGENE as part of the wheat project which represents the Norwegian effort in the International Wheat Genome Sequencing Consortium and aims to generate the reference sequence for the wheat chromosome 7B.

I would like to thank my supervisor Prof. Odd-Arne Olsen and co-supervisor Simen Rød Sandve for their guidance and for giving me the opportunity to join the wheat team and analyze your expensive dataset. Also thanks to Prof. Torgeir R. Hvidsten, Dr. Bujie Zhan and everybody else who helped me on the way.

Abstract

Bread wheat (*Triticum aestivum*, BBAADD) is one of the most important food-crops world-wide. The genome of bread wheat is allohexaploid, meaning that it contains three related diploid genomes (termed A, B and D). This implies that most wheat genes exist in three similar copies (i.e. homeologs), which together with the fact that the genome mainly consists of repeats from transposable elements, makes the wheat genome very hard to sequence and assemble. Currently there is an effort by the International Wheat Genome Sequencing Consortium (IWGSC) to generate a wheat genome reference sequence. An enabling factor is the use of flow cytometry to isolate the individual chromosome arms which then can be sequenced and assembled separately. The final reference sequence is still a long way from being finished, however a short-read shot-gun sequence assembly of each chromosome arm, referred to as chromosome survey sequencing (CSS) assembly, has been made available from the IWGSC. Using the CSS assembly as reference for analyses of RNA-sequencing (RNA-seq) data enables us, for the first time, to distinguish and quantify the transcription level of the homeologous genes. In this study I use RNA-seq data sampled from the starchy endosperm, aleurone layer, and transfer cells of the developing wheat endosperm to analyze the homeolog-specific aspects of the hexaploid transcriptome. I found that three quarters of the genes have an expression bias towards either one of the sub-genomes, but that no sub-genome is favored on a general level. The pattern of homeolog bias shows no correlation with functional gene-groups, but 28% of the genes show developmentally controlled homeolog-specific regulation when comparing the sampled tissues. There is also evidence supporting that the D-genome has a different distribution of expression levels than A and B. This could be explained by the polyploidization history, or indicate the presence of hybridization in a D-genome ancestor. Greater understanding of the mechanisms that govern homeolog-specific gene regulation can have an impact on the way breeding of allopolyploids is performed. This study illustrates the importance of homeolog specific reference sequences, and the potential hexaploid wheat has as a model to study mechanism of gene-regulation in allopolyploids.

Table of content

1 Introduction 1

1.1 The hexaploid bread wheat genome 1

1.2 The wheat endosperm 2

1.3 Genetic effects of allopolyploidy 3

1.4 Homeolog specific expression 4

1.5 Study aims 5

2 Materials and Methods 7

2.1 Plant material, sample preperation and sequencing 7

2.2 Chromosome Survey Sequences 8

2.3 Genome zipper 8

2.4 RNA-seq analysis – an experimental algorithm to calculate homeolog specific expression 9

2.4.1 Step 1: Align reads 10

2.4.2 Step 2: Extract reads mapping to zipper contigs 10

2.4.3 Step 3: Find homeoreads 11

2.4.4 Step 4: Filter biased reads 11

2.4.5 Step 5: Count perfect matches 12

2.4.6 Step 6: Estimate homeolog read-counts 12

2.5 Differential expression analysis 14

2.6 Principal component analysis (PCA) 15

2.7 Detecting reciprocal silencing 16

2.8 Gene set enrichment using Gene Ontology categories 17

2.9 Filtering contigs based on hit-counts in orthologous regions 18

3 Results: 19

3.1 RNA sequencing and read mapping 19

3.2 PCA analysis confirms good data quality 20

3.3 Sub-genomes contribute equally to expression per chromosome 21

3.4 Most genes show significant homeolog specific expression 22

3.5 Hierarchical gene clustering and expression heatmap...23

3.6 Reciprocal silencing of homeologs between tissues...25

3.7 Differences in homeolog bias levels between tissues...26

3.8 Homeolog expression distribution show that D genome is less dominant 27

3.9 Gene set enrichment fail to show genome asymmetry of GO terms 29

4 Discussion 32

4.1 Differential expression of homeologs are common in the wheat genome...........................32

4.2 Large scale sub-genomic expression bias..33

4.3 Reciprocal silencing 33

4.4 The special expression distribution of the D genome...34

4.5 Gene set enrichment does not suggest genomic asymmetry in wheat................................34

4.6 Limitations of the homeoreads method 35

4.6.1 Using contigs to represent genes 35

4.6.2 Arbitrary choice of reference zipper 36

4.6.3 Inflated variance from ambiguous hits...36

4.6.4 Effect of discarding multireads within sub-genome 37

4.6.5 Is 2bp mismatch a good choice?...37

5 Conclusion 39

6 References...40

Appendix A Scripts for the homeoreads pipeline 44

1 Introduction

1.1 The hexaploid bread wheat genome

Bread wheat (AABBDD, *Triticum aestivum*) is a ~17GB allohexaploid that was formed through two relatively recent polyploidization events (i.e. genome duplication). Allopolyploids are formed through species hybridization and contain two different diploid genomes, as opposed to autopolyploids which arise through duplication of a single genome. The three genomes that make up the bread wheat genome are referred to as the A, B and D genomes. The diploid progenitors of the three genomes diverged from a common ancestor about 2.5-4.5 million years ago (Huang et al., 2002). The current model of hexaploid wheat evolution is that the tetraploid *Triticum dicoccoides* (AABB) formed about 0.5 million years ago when *Triticum urartu* (AA) crossed with an unknown species (BB) related to *Aegilops speltoides* (Salse et al., 2008). Subsequently, around 10k years ago (Nesbitt and Samuel, 1996), tetraploid wheat (AABB) hybridized with diploid goat grass (DD, *Aegilops tauschii*) to form the hexaploid bread wheat we know today (Figure 1A).

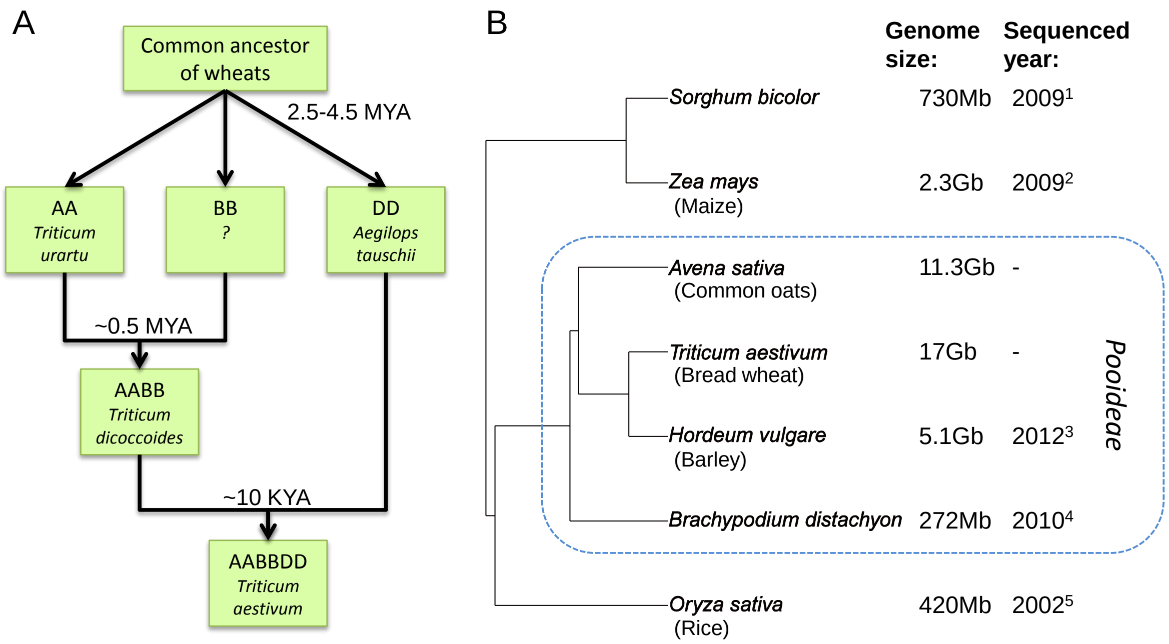
A
Common ancestor of wheats
2.5-4.5 MYA
AA Triticum urartu
BB ?
DD Aegilops tauschii
~0.5 MYA
AABB Triticum dicoccoides
~10 KYA
AABBDD Triticum aestivum
B
Genome size:
Sequenced year:
Sorghum bicolor 730Mb 2009¹
Zea mays (Maize) 2.3Gb 2009²
Avena sativa (Common oats) 11.3Gb -
Triticum aestivum (Bread wheat) 17Gb -
Hordeum vulgare (Barley) 5.1Gb 2012³
Brachypodium distachyon 272Mb 2010⁴
Oryza sativa (Rice) 420Mb 2002⁵
Pooideae

Figure 1: (A) Evolution of hexaploid wheat genome from its diploid ancestors (B) Relationship with other important grass species (Poaceae). (Grass Phylogeny Working Group et al., 2001; [1]Paterson et al., 2009; [2]Schnable et al., 2009; [3]Mayer et al., 2012; [4]The International Brachypodium Initiative, [5]2010; Yu et al., 2002)

Wheat belongs to the Poaceae family, also called Gramineae or true grasses. Of the Poaceae species, several important agricultural species like rice, maize and sorghum have been sequenced

(Figure 1B). However in the sub-family Pooideae, which includes wheat, sequencing efforts are lagging behind because of their typically large genomes and high levels of repeat content (Kellogg and Bennetzen, 2004). Brachypodium, with its uncharacteristically small genome, has been selected as a model system for the Pooideae and was sequenced and annotated in 2010.

Currently the International Wheat Genome Sequencing Consortium (IWGSC) is working to establish a reference genome for bread wheat. To overcome the problem of polyploidy they are using flow cytometry to isolate the individual chromosome arms (Dolezel et al., 2007), which then can be sequenced separately. The strategy of the IWGSC is to do BAC-by-BAC sequencing of the minimal tiling path of each of the isolated chromosome arms to create a high quality wheat genome assembly; however this work is labor intensive, expensive, and not expected to be finalized before 2015. Hence, to be able to provide the wheat breeders and research community with a catalogue of wheat genes the IWGSC launched a wheat chromosome survey sequencing (CSS) project in 2010. This project aims at generating chromosome arm shot-gun assemblies of the hexaploid wheat genome. Although these assemblies are highly fragmented they provide the possibility to, for the first time, study the molecular biology of hexaploid wheat at the nucleotide resolution within and between the A, B and D sub-genomes.

1.2 The wheat endosperm

The endosperm makes up the nutritious fraction of the wheat grain; hence understanding endosperm biology, especially with respect to baking quality, is of major importance for future wheat breeding (Olsen et al., unpublished). The wheat endosperm is also interesting as a model to study general aspects of developmental biology in plants and polyploids in particular. In angiosperms the megagametophyte goes through a double fertilization process where both the egg that gives rise to the embryo and the central cell that gives rise to the endosperm is fertilized at the same time (Raghavan, 2003). The endosperm can therefore be considered an organism on its own, which develops next to the embryo. Its biological role is to provide nutrition to the developing embryo and seedling. Endosperm consists of three different types of tissues (Olsen, 2004) of which the major part is the starchy endosperm that contains starch and storage proteins. Along the surface on the basal side is a layer of transfer cells that connect the endosperm with the vascular tissue of the maternal plant and have the role of transferring nutrients from the plant into the endosperm (Olsen, 2004). Along the rest of the surface the starchy endosperm is surrounded by a single layer of aleurone cells, which produce enzymes that break down the starch and protein in the starchy endosperm when the seed germinates (Olsen, 2004).

As part of the IWGSC effort to annotate the wheat genome and to improve our understanding of endosperm biology in the hexaploid bread wheat, RNA-sequencing of the endosperm tissues at different timepoints during grain development has been performed in the research group of Odd-Arne Olsen at IPM/UMB. This RNA-seq dataset is the basis for my master thesis project.

1.3 Genetic effects of allopolyploidy

The conditions in the cells of a newly formed allopolyploid are very different from its diploid progenitors. There are the physical constraints from having twice as much DNA and having to replicate it, effects of having an extra dose of every gene, and unintended interactions between the regulatory networks of the different genomes. To investigate the effects of polyploidy studies often use neopolyploids, i.e. synthetic polyploids that have been produced in the lab. This makes it possible to study the early effects in the first few generations and also facilitates direct comparisons with the diploid progenitors, such as in differential expression experiments. For natural allopolyploids (such as wheat) the original progenitor may have diverged or died out since the polyploidization event, making it difficult or often impossible to do direct comparisons between a polyploid genome and its diploid ancestors.

One of the difficulties a new polyploid has to overcome is incorrect chromosome pairing during meiosis (Zhang et al., 2013). The homeologous chromosomes derived from each progenitor are typically so similar that they are mistakenly paired, leading to aneuploidy in the next generation. In wheat, the gene *Ph1* (Pairing homeologous) is associated with an increased ability to correctly pair homologous chromosomes (Wall et al., 1971). Despite the presence of the *Ph1* gene, neopolyploid wheat still suffer frequent aneuploidy as is shown in a comprehensive study of chromosomal variation in 11 consecutive selfed-generations of neopolyploid wheat (Zhang et al., 2013). Feldman et al. (1997) identified sequences specific to each chromosome that occur in all three diploid progenitors but occur only in one copy in hexaploid wheat. They suggested that elimination of these sequences occur in the developing allopolyploids and provide the physical basis for diploid-like stable homologous pairing. It has been shown that wheat allopolyploids, both natural and first generation neopolyploids, contain less DNA than the sum of their progenitors, suggesting that loss of DNA occur during/after polyploidization (Eilam et al., 2008).

Another difficulty allopolyploids have to overcome is the effect of altered gene expression resulting from having an extra gene copy and interactions between the regulatory networks of the diverged genomes. The altered expression may be advantageous, deleterious, or have no effect. Reproducible physical gene deletions, activation and suppression of gene transcription as well as changes in cytosine methylation patterns (with putative gene regulatory consequences) have been observed in neopolyploid wheat (Shaked et al., 2001; Kashkush et al., 2002; He et al., 2003). Later studies have used microarrays (Affymetrix GeneChip Wheat Genome Array) to measure genome-wide transcription levels of neopolyploids and their progenitors (Akhunova et al., 2010; Chagué et al., 2010; Qi et al., 2012). These studies conclude that even though most genes show additive expression in the neopolyploid compared to progenitors, a significant fraction (7% or 19%) displays non-additive expression patterns. Furthermore, studies of tetraploid cotton have given rise to the idea of parental genomic expression dominance where the expression profile of the polyploid typically is closer to one parent than the other (Rapp et al., 2009; Yoo et al., 2013). Akhunova et al. (2010) have reported evidence for this kind of dominance by the A+B genome in re-synthesized wheat.

1.4 Homeolog specific expression

Very few studies have looked at homeolog specific expression in wheat, probably because of the technical difficulty of distinguishing and quantifying homeolog specific transcripts prior to the availability of the CSS assembly. Attempts have been made to use microarrays to measure genome specific expression in re-synthesized polyploid wheat by identifying probes that specifically bind to either of the parental tetraploid AB or diploid D genome. However, apart from being unable to distinguish between the A and B homeologs, verification of a subset of the probes revealed that the method was error-prone (15-20% probes were incorrectly classified) (Akhunova et al., 2010).

Mochida et al. (2004) managed to quantify homeolog specific expression of 90 genes in hexaploid wheat by sequencing. This study first identified SNPs from EST databases and then used nullisomic/tetrasomic wheat strains that lack a specific chromosome to determine the chromosome of origin. By sequencing the transcripts, they could identify the pattern of SNPs that are unique for each homeolog. They found that homeolog silencing varied between the 10 tissues they had sampled and that the level of silencing was independent of chromosome or sub-genome, concluding that silencing occurs on the basis of individual genes.

The use of homeolog specific SNPs (homeoSNPs) as markers to identify and quantify homeolog specific expression have also been used in allotetraploid cotton (Chaudhary et al., 2009). Note that a homeoSNP can typically only distinguish between two homeologs, which works great for tetraploids, but for hexaploid wheat several homeoSNPs are required to identify the originating sub-genome of a transcript.

Another method that has been used to evaluate the extent of homeolog specific expression is single-strand conformation polymorphism (SSCP). Using SSCP in combination with nullisomic/tetrasomic wheat strains, Bottley et al. (2006) investigated the homeolog specific expression of 236 genes in both leaf and root tissues of hexaploid wheat. In leaf tissue, 27% of the genes had one silenced homeolog, while 26% of the genes in the root had one or two homeologs silenced. In several cases the homeologs were reciprocally silenced, i.e. one was silenced in the leaf and the other in the root. This phenomenon of tissue specific reciprocal silencing of homeologs has also been observed in allotetraploid cotton (Chaudhary et al., 2009; Adams et al., 2003). Reciprocal silencing represents a type of subfunctionalization as each homeolog is specialized to function in different tissues. Subfunctionalization is thought to preserve duplicate genes and subsequently allow neofunctionalization of each copy as they adapt to the specific needs in the tissues where they are expressed (Rastogi and Liberles, 2005; Lynch and Force, 2000).

Interestingly, when comparing the pattern of homeolog specific silencing between different wheat cultivars there seems to be a great deal of variation. In a study of 15 genes in 16 wheat cultivars, Bottley and Koebner (2008) found that 8 genes showed homeolog specific silencing and that out of the 16 cultivars, only 2 cultivars had the same pattern of homeolog specific silencing. They suggest that this variation in homeolog gene expression is linked to epigenetic silencing through methylation. Methylation-based homeolog expression differences have been shown in a study of the three homeologs of two MADS-box genes in wheat, WSEP and WLHS1. The B homeolog of WLHS1 is silenced by cytosine-methylation, while the A homeolog has lost its function but is still expressed (Shitsukawa et al., 2007).

1.5 Study aims

The recently generated survey sequences from the IWGSC CSS project in combination with RNA-seq data makes it possible to analyze genome-wide homeolog specific expression with unprecedented precision and accuracy.

The first goal of this thesis is to develop a pipeline/algorithm to determine homeolog specific transcript abundance from the endosperm RNA-seq data. Secondly, the homeolog expression levels generated by the pipeline will be used to explore and analyze aspects of homeolog expression bias, with focus on three topics:

1. **Sub-genome dominance on whole genome or chromosome level** – Does any of the sub-genomes show any over-all expressed bias?

2. **Reciprocal silencing between tissues** – This phenomenon has been observed in smaller studies before (Mochida et al., 2004; Bottley et al., 2006), but how prevalent is it?
3. **Genomic asymmetry of certain traits/functions –** It has been suggested that many traits are preferentially controlled by a single sub-genome (Feldman et al., 2012). I will use gene ontology (GO) to test if any gene function or process is over-represented in any sub-genome.

Note that, although the RNA-seq data is from the endosperm, the goal of this thesis is not to investigate any aspects of endosperm biology.

2 Materials and Methods

2.1 Plant material, sample preperation and sequencing

(Note that sampling and sequencing was carried out before my involvement in the project)

Bread wheat cv. Chinese Spring was grown in pots in two separate rooms. Each room had 75 pots with three seedlings per pot. Plants were tagged at anthesis and ears were harvested at 10, 20 and 30 days post anthesis (DPA). Only the 20 DPA samples are included in this analysis. The middle part of each ear was harvested and stored at -80°C until use. Seeds were dissected using dry ice and covered by RNA*later*©-ICE under the microscope. The embryo was removed and the seeds cut in slices for isolation of the aleurone layer (AL), transfer cells (TC) and starchy endosperm (SE). Due to difficulties in dissection of the different tissues, there is some contamination of SE in the TC samples and possibly some in the AL samples, whereas SE samples should be pure (Figure 2). Dissected tissues were put in liquid nitrogen and stored at -80°C for RNA isolation. Tissues from 15 pots were pooled before RNA isolation, with two replicates per room, giving a total of four replicates per tissue.

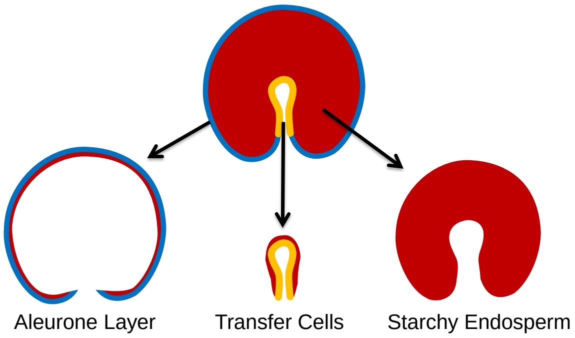

Figure 2: Dissection of the wheat endosperm tissues. Note that there can be some contamination of SE in the AL, and TC samples whereas SE samples should be pure.

Total RNA was extracted from frozen plant material using the RNeasy Lipid Tissue Mini kit (QIAGEN). For starchy endosperm and transfer cells we did a pre-extraction step in order to remove starch. The concentration of RNA was measured using a Nanodrop 8000 spectrophotometer (ND8000, Thermo Scientific, Wilmington, USA). RNA integrity was assessed on an Agilent 2100

Bioanalyzer (DE54704553, Agilent Technologies, Inc., CA, USA) using an RNA 6000 LabChip kit. RNA samples were stored at -80°C until sent for sequencing. RNA samples with good quality and quantity were sent for sequencing on Illumina HiSeq2000 at Norwegian Sequencing Centre (www.sequencing.uio.no).

Table 1 lists the samples that have been included in this study.

Table 1: List of samples included in the analysis.

SampleID	Room#	DPA	Tissue	Rep#
Room1_20DPA_AL_1	1	20	AL	1
Room1_20DPA_AL_2	1	20	AL	2
Room1_20DPA_AL_31	1	20	AL	3.1*
Room1_20DPA_AL_32	1	20	AL	3.2*
Room2_20DPA_AL_1	2	20	AL	1
Room2_20DPA_AL_3	2	20	AL	3
Room1_20DPA_Ref_1	1	20	Ref**	1
Room1_20DPA_Ref_2	1	20	Ref**	2
Room1_20DPA_SE_1	1	20	SE	1
Room1_20DPA_SE_2	1	20	SE	2
Room2_20DPA_SE_1	2	20	SE	1
Room2_20DPA_SE_2	2	20	SE	2
Room1_20DPA_TC_1	1	20	TC	1
Room1_20DPA_TC_2	1	20	TC	2
Room2_20DPA_TC_1	2	20	TC	1
Room2_20DPA_TC_2	2	20	TC	2

* Technical replicates

** Mix of AL, SE and TC

2.2 Chromosome Survey Sequences

As a reference sequence for the transcriptional analyses I used the bread wheat CSS assemblies generated by the IWGSC (www.wheatgenome.org/). The CSS assemblies are derived from shotgun sequencing of whole chromosomes or chromosome arms that have been sorted by flow cytometry (Dolezel et al., 2007). Due to the high level of repeats in the wheat genome, the CSS assembly consists of mostly short contigs (N50 = 2292 base pairs).

2.3 Genome zipper

The genome zipper is a virtual map of wheat genes based on combining genetic maps with information of syntenic relationships between the related sequenced grass species *Oryza sativa*, *Sorghum bicolor* and *Brachypodium distachyon*. The zipper was constructed by the IWGSC using the method described in Mayer et al. (2009). Syntenic regions are discovered by mapping the wheat survey sequences against the related grass genomes which are then anchored to a scaffold of mapped genetic markers. The resulting genome zipper is represented as an ordered table for each chromosome that contains the names of the orthologous genes and the contig(s) they relate to in the survey sequence. Some zipper-loci can refer to several contigs which reflect genes that has been split into several contigs or genes existing in multiple copies. However, to simplify the analysis I am only using a single contig to represent each gene.

2.4 RNA-seq analysis – an experimental algorithm to calculate homeolog specific expression

RNA-seq allows expression levels to be estimated by counting the number of RNA-seq reads that map to each of the genes in a reference genome. In my case, I use the CSS assembly as reference. In the CSS assembly, the identity of the chromosome arms and sub-genomes is known, but gene models are not available. It is therefore necessary to identity regions that contain the genes. Furthermore, to be able to compare the expression between homeologs of each gene it is necessary to also locate the regions in each sub-genome that contain the homeologous genes.

The method I developed solves the problem of locating the homeologs by using the ID-numbers of the mapped reads to identify those reads that map to homeologous regions. Homeologous regions are defined by their high sequence similarity, and in gene coding regions they are expected to be about 98-99% identical. The reads are mapped to each of the sub-genomes with two base-pair mismatch allowed so that it is likely that the 101 base-pair long reads map to all three homeologous gene-copies. By comparing read IDs, the reads that map exactly once to all three sub-genomes are picked out. These reads essentially represents a multiple alignment of the regions in the three homeologs that are at least 98% identical and are hereby referred to as **homeoreads**. Since the samples are taken from the same wheat cultivar (Chinese spring) as the reference, the homeoreads will map perfectly to the originating sub-genome. The presence of nucleotide mismatches between the three homeologs (i.e. homeoSNPs) is used to determine the sub-genome of origin.

However, the homeoreads only define the homeologous regions, it is also necessary to locate the genes. To do this, I use the contigs referred to in the genome zipper. Using the genome zipper has the advantage of avoiding non-coding transcribed elements and ribosomal RNA as well as having an estimated position along the chromosome and the identity of orthologous genes in sequenced grass genomes. Expression levels per homeolog-triplet are calculated by counting the number homeoreads mapping to each contig in the reference zipper, which is converted into an estimate of homeolog specific expression based on the nucleotide mismatches between the three homeologs. The method can be divided into 6 steps (Figure 3) which are described in the following sections. See Appendix A for the source code of steps 2-6, which are implemented as a set of scripts that can be run from a main R script.

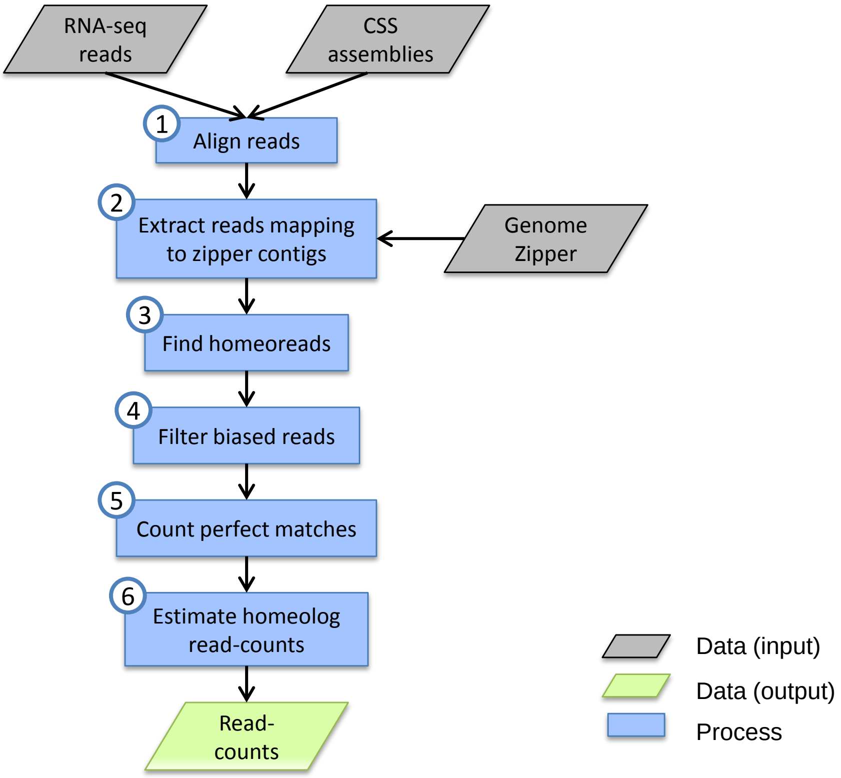

Figure 3: Homeoread pipeline flow-chart

2.4.1 Step 1: Align reads

The CSS sequences from each chromosome arm are pooled together to get three reference genome assemblies – one for each of the three sub-genomes A, B and D. The reads are mapped against each of the three sub-genomes in three separate runs using TopHat (Trapnell et al., 2009) with two base-pairs mismatch allowed. Reads that map several places within sub-genomes (multireads) are removed.

2.4.2 Step 2: Extract reads mapping to zipper contigs

The zipper for one of the sub-genomes is used as reference to select contigs that contain genes with known orthologs. The choice of zipper reference genome is arbitrary (I used the B-genome). For the selected sub-genome, only the reads that map to the contigs from the reference zipper are kept. These reads are assumed to have been transcribed from the genes we want to study, and in step 3 they will be used to find the homeologous regions in the other sub-genomes.

2.4.3 Step 3: Find homeoreads

By the using the IDs of the reads that mapped to the reference genome zipper contigs, I identify and keep only reads that also map to both of the two other sub-genomes. Note that paired reads are treated as two independent single reads. The remaining reads (homeoreads) map to all three homeologs of the zipper genes.

2.4.4 Step 4: Filter biased reads

Homeologous regions with certain combinations of homeoSNPs can create a read count that is biased towards one of the sub-genomes. A homeoSNP is a nucleotide in a homeologous aligned region that is different in one sub-genome compared to the two other. For example, if sub-genome A has a 'T' nucleotide and both sub-genomes B and D has a 'C' nucleotide, then there is a homeoSNP in sub-genome A. Bias occurs when a 101bp region contains 3-4 homeoSNPs where one homeolog has no homeoSNPs and the two other homeologs have 2 and 1-2 homeoSNPs (Figure 4). In this case, if a read is transcribed from the first genome with no homeoSNPs, it will successfully align since it has no more than 2bp mismatch with the two other genomes. However, if a read is transcribed from one of the two sub-genomes with homeoSNPs, they will fail to align because they will have 3 mismatches. Since this kind of bias only occurs when certain patterns of homeoSNPs are present, it is possible to detect biased homeoreads and discard them. Mismatch information, including position and type of each mismatch, is contained in the SAM file (output from TopHat) and is used to exclude gene regions that create homeoSNP-bias.

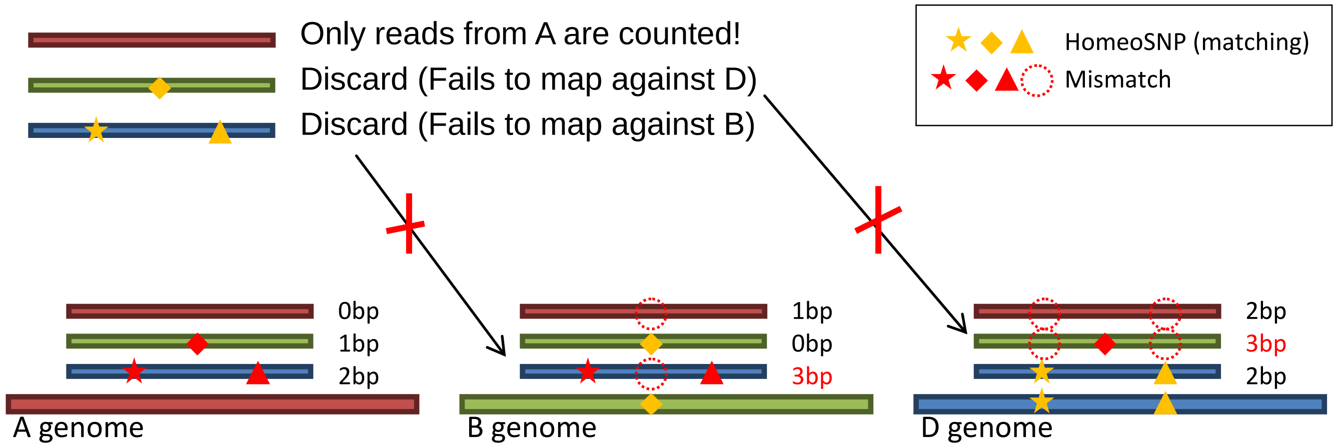
Only reads from A are counted!
Discard (Fails to map against D)
Discard (Fails to map against B)
HomeoSNP (matching)
Mismatch
0bp
1bp
2bp
1bp
0bp
3bp
2bp
3bp
2bp
A genome
B genome
D genome

Figure 4: Example of homeoSNP pattern that cause read-count bias. Three reads from the same homeologous region of the three sub-genomes mapped against each sub-genome. The reads originating from B and D fails to map because they have more than 2bp mismatch to at least one sub-genome.

2.4.5 Step 5: Count perfect matches

The reads are divided into 7 groups depending on which sub-genome(s) they map perfectly to: the A, B or D specific (unambiguous) reads; the AB, AD or BD ambiguous (maps to two of the sub-genomes); or the completely ambiguous ABD reads that map to all three sub-genomes. The number of reads that fall into each of these groups is counted for each zipper gene.

Figure 5 provides an example of read count assignment. In this example there are 6 reads that are mapped without any mismatches to sub-genome A. Four of these reads marked in red have mismatches when aligned against either B or D and is therefore counted as A specific. The purple colored read has mismatches only against B and is therefore counted as AD ambiguous. The white read maps perfectly to all three. Note that the black reads have more than two mismatches in at least one of the sub-genomes and therefore will not be included in the analysis (discarded in step 3).

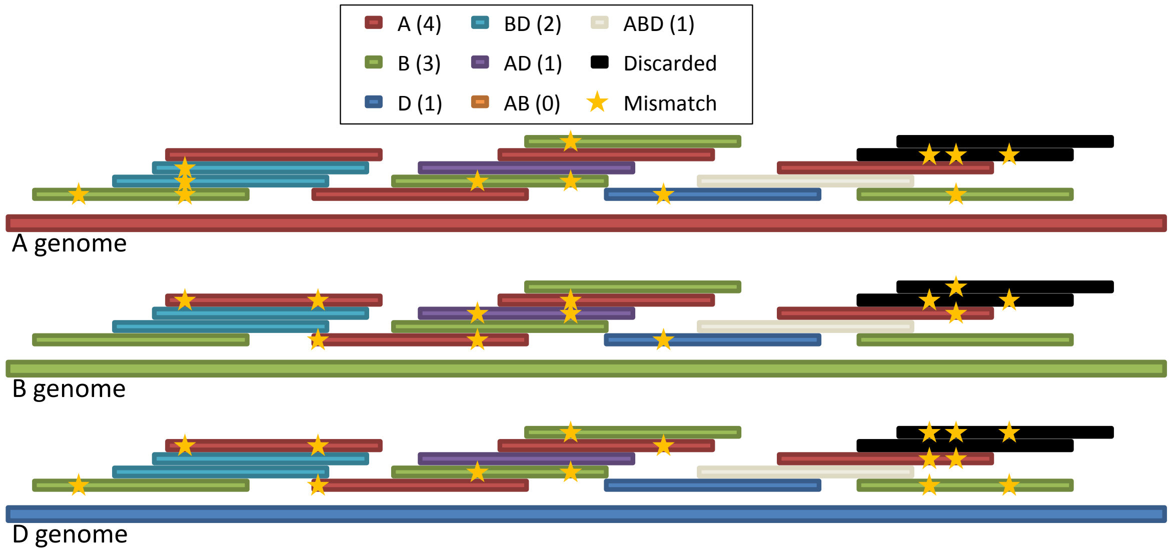

Figure 5: Example of homeoreads mapped to the three sub-genomes. The colors of the reads indicate the inferred origin of the reads. The number of reads per group is given in the legend. The stars indicate where there is a mismatch between a base-pair in the read and the genome. The black colored reads fails to map to at least one of the sub-genomes since it has more than 2bp mismatch and is therefore excluded from the analysis.

2.4.6 Step 6: Estimate homeolog read-counts

An estimate of the number of reads originating from each homeolog is made based on the counts of both ambiguously and unambiguously mapped reads. Reads that map ambiguously to two or three homeologs are divided proportionally between them based on following model:

For each set of homeologous genes, depending on the locations and number of homeoSNPs, there will be a certain chance that a homeoread originates from a region that is either: unique for all homeologs (p_0); unique for A but B and D are identical (p_{BD}); unique for B but A and D are identical (p_{AD}); unique for D but A and B are identical (p_{AB}); or identical for all homeologs (p_{ABD}).

The problem can be presented as a geometrical problem (Figure 6). x_A, x_B and x_D are the estimated numbers of reads originating from homeolog A, B and D that we want to find. The known variables are the areas marked with different colors that represent the read counts for the 7 different groups named: $n_A, n_B, n_D, n_{AB}, n_{AD}, n_{BD}$ and n_{ABD}.

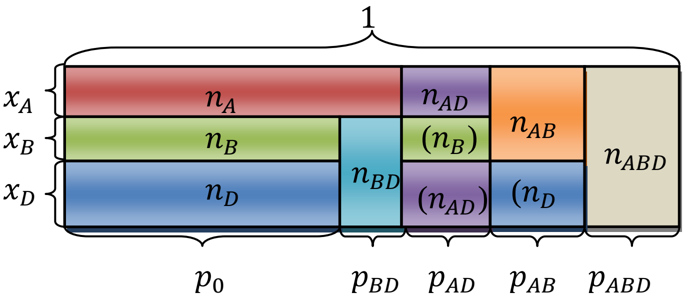

Figure 6: Geometrical representation of the homeolog read count estimation problem. The areas shown with different colors are known and represent the counts of reads that map to a specific homeolog or ambiguously to several of the homeologs, ($\boldsymbol{n_A}, \boldsymbol{n_B}, \boldsymbol{n_D}, \boldsymbol{n_{AB}}, \boldsymbol{n_{AD}}, \boldsymbol{n_{BD}}$ and $\boldsymbol{n_{ABD}}$). $\boldsymbol{p_0}, \boldsymbol{p_{BD}}, \boldsymbol{p_{AD}}, \boldsymbol{p_{AB}}$ or $\boldsymbol{p_{ABD}}$ represents the proportion of the aligned homeologous regions that will yield reads that are unique for all homeologs, indistinguishable between homeologs B and D, A and D, A and D or any of the homeologs, respectively. $\boldsymbol{x_A}, \boldsymbol{x_B}$ and $\boldsymbol{x_D}$ represents the estimated number of reads originating from each of the homeologs.

The proportions of the ambiguous reads that originates from each sub-genome can be calculated if the total number of reads originating from each sub-genome were known. For example: If x_A and x_B were known, then the part of n_{AB} that contributes to x_A would be $n_{AB}\frac{x_A}{x_A+x_B}$

Following that logic we can define these equations:

$$x_A = n_A + n_{AB}\frac{x_A}{x_A + x_B} + n_{AD}\frac{x_A}{x_A + x_D} + n_{ABD}\frac{x_A}{x_A + x_B + x_D}$$

$$x_B = n_B + n_{AB}\frac{x_B}{x_A + x_B} + n_{BD}\frac{x_B}{x_B + x_D} + n_{ABD}\frac{x_B}{x_A + x_B + x_D}$$

$$x_D = n_D + n_{AD}\frac{x_D}{x_A + x_D} + n_{BD}\frac{x_D}{x_B + x_D} + n_{ABD}\frac{x_D}{x_A + x_B + x_D}$$

These equations are solved numerically by making an initial estimate that is gradually improved by applying the above formulas repeatedly until they converge. The initial estimates used are based on an assumed equal amount of A, B and D counts:

$$x_A = n_A + \frac{n_{AB}}{2} + \frac{n_{AD}}{2} + \frac{n_{ABD}}{3}$$

$$x_B = n_B + \frac{n_{AB}}{2} + \frac{n_{BD}}{2} + \frac{n_{ABD}}{3}$$

$$x_D = n_D + \frac{n_{AD}}{2} + \frac{n_{BD}}{2} + \frac{n_{ABD}}{3}$$

2.5 Differential expression analysis

The R package “DESeq” (Anders and Huber, 2010) is used to test for differential expression. A typical feature for transcriptome data is the relatively few replicates per treatment (result of high price), and this makes it difficult to identify statistically significant expression differences. To improve the statistical inference, DESeq assumes constant variance for genes with similar read-counts. The variance between all replicates for all factors and genes are fitted to a model that only depends on the read-count. This model takes into account both the shot-noise and the biological variance between the samples. The shot-noise, which comes from randomly picking a subset of RNA molecules from a larger pool, follows a poisson distribution and it accounts for most of the variance in low abundance transcripts. For high-abundance transcripts, the variance is dominated by the biological variance. Because the technical replicates (Room1_20DPA_AL_31 and Room1_20DPA_AL_32) do not contain any biological variance, they are pooled and treated as a single sample (Room1_20DPA_AL_3), as was suggested in Anders and Huber (2010).

The DESeq package is designed to compare expression levels between the same genes in several samples. In this study however, the comparison is made between the homeologs within the same sample. This is possible because only the homeoreads are counted, which means that even if the length of the homeologs vary, only the parts that are common are compared. Each sample is divided into three sub-genomic pseudo-samples separating the read-count contribution of the A, B and D homeologs. DESeq treats these pseudo-samples as independent samples with different levels of the “subgenome” factor.

Normalization is performed using DESeq’s *estimateSizeFactors( )*function which assigns a scaling coefficient to each sample to correct for any technical bias that affects the sequencing depth. Since the three sub-genomic pseudo-samples are actually from a single sample, the scaling coefficient acquired from the DESeq normalization procedure is averaged for each sample so that the three pseudo-samples are equally scaled.

To test for differential expression DESeq have implemented a generalized linear model (GLM) regression procedure. GLM regression makes it possible to use variables that don't follow a normal distribution, such as expression data. Each sample is assigned three factors; "subgenome", "tissue" and "room". Since there is more than one factor, tests have to be done by comparing different regression models. One model that accounts for the effect that is tested, and a reduced model where the effect is not accounted for. These tests do not specifically test for differential expression but rather tests if a specific factor has a significant effect on the expression levels. The models tested in this study are listed in Table 4 in the results section.

P-values from GLM-tests are adjusted using the Benjamini-Hochberg procedure to control the false discovery rate. This adjusted P-values are equivalent to the Q-value and can be interpreted as the expected proportion of false positives, e.g. if 1000 genes are called significant with adjusted $P<0.01$ it is expected that 10 of these are false positives.

2.6 Principal component analysis (PCA)

An indication of good quality for experimental data is that samples from different conditions are more different than when comparing replicate samples. Each sample can be viewed as a point in a coordinate system with the same number of dimensions as there are genes, i.e. a vector containing the expression values of the genes. The distance between these points reflects the level of similarity/difference between the samples. If data quality is good, replicates for each condition should form clear clusters which can be visualized by projecting the multidimensional points onto a two-dimensional plane. To maintain the maximum amount of information, the projection plane is defined by the first two principal components (PCs). The first PC is the line that goes through the cloud of points in the multidimensional space that minimizes the sum of square of distances between the line and all points. Therefore the maximal variance is along the PC. The following PCs are found the same way but must be perpendicular to all previous PCs.

Principal components were calculated in R using the *prcomp( )* function. Expression values are transformed first using the variance stabilizing transformation function from the DESeq package, which attempts to make variance constant regardless of the total expression level. Centering is perform by the *prcomp( )* function, but not scaling, as the variance stabilizing transformation ensures equal variance.

2.7 Detecting reciprocal silencing

Given two tissues X and Y, reciprocal silencing of homeologs (A, B, and D) between tissues is present when homeologs switch between being expressed in different tissues:

X-tissue: A=0, B=10, D=12

Y-tissue: A=11, B=0, D=0

In the actual data complete silencing of homeologs is not common. Hence, reciprocal silencing will mostly be present in situations with partial silencing (Chaudhary et al., 2009). It is therefore necessary to define a test for partial "reciprocality".

The GLM-model implemented in DESeq allows for testing of interaction effect between sub-genome and tissue. If reciprocal silencing is present in the data, this would be expressed as a strong interaction effect in the GLM-model tests. However, a significant interaction effect is not sufficient to indicate reciprocal silencing. For example, consider the hypothetical gene that has the expression level ratio 1:1:2 of homeologs A:B:D in SE and 2:2:3 in AL. Since the relative expression between the homeologs differs in the two tissues it could test positive for interaction effect but it clearly isn't reciprocal. To be able to quantify "reciprocality" between three values I found it useful to represent the expression levels of A, B and D homeologs using an alternative set of parameters. The idea is borrowed from computer graphics, where colors can be represented by their red, green, and blue components but also by the parameters *hue, saturation*, and *value* (HSV). Since there are three homeologs expression values – A, B, and D – they can represent the colors red, green, and blue respectively, which then are converted to HSV. In this alternative representation, saturation (S) reflects the level of homeolog expression bias: 100% means that one or two homeologs are completely silenced, 0% means that all homeologs are expressed equally, and 50% means that there is a 2x fold difference between the least and most expressed homeologs. The hue (H) represents a measure of which homeologs that dominates the expression (Figure 7). For example, if only one of the homeologs A, B, or D are expressed, H will be 0°,120° or 240°, respectively (Figure 7). The *value* parameter (V) represents the expression level of the dominant homeolog.

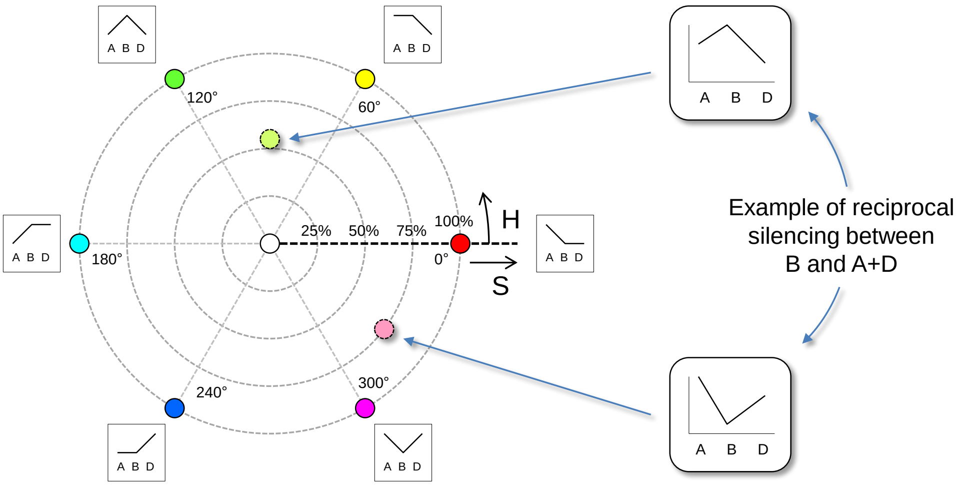

Figure 7: HSV parameterized homeolog-triplet expression profiles. Homeolog expression bias is defined by S (saturation). The dominant homeolog(s) is defined by the angle H (hue). Reciprocal expression profiles are characterized by difference in H.

Reciprocal homeolog silencing between tissues is characterized by differences in H. Say that a gene is expressed only in the B sub-genome in one tissue while in another tissue only A and D homeologs are expressed. The difference in H would then be 180°. I define any difference in H greater than 120° as reciprocal silencing. Given that reciprocal silencing only makes sense if there is homeolog specific expression bias in the first place, I require that S is at least 50% for both homeologs. An example of two homeolog expression profiles that pass these criteria are shown in Figure 7.

The mean homeolog expression count across all replicates (including room replicates) for each tissue is calculated before converting to HSV. Reciprocal silencing is found by pair-wise comparisons of the three tissues, where a gene is classified as reciprocal if their HSV values are within the chosen thresholds. Because low expression levels are subject to high variance from shot-noise I try to limit the number of false positives by only including the genes where both tissues have V>20.

2.8 Gene set enrichment using Gene Ontology categories

Gene set enrichment is a method that is used to look for associations of experimental conditions with groups of genes that perform a specific function or are involved in a specific process or pathway. In this study I used GO terms to define the gene sets. GO terms are standardized terms used to annotate gene function, and they are organized in a hierarchical manner so that each term is related to a lower level more broadly describing term. All GO terms can be traced down to the three main categories “molecular function”, “biological process” or “cellular compartment”. The R package “TopGO” (Alexa and Rahnenfuhrer, 2010) was used to run the analysis. This package implements the elimination algorithm described in (Alexa et al., 2006) which improves the test by using the relationships between the GO terms.

The Kolmogorov–Smirnov test (K-S test) implemented in the TopGO package was used to test for the association between the GO terms and sub-genome or tissue specific expression patterns. Each gene is scored using the Pearson-correlation between the sub-genome/tissue and the expression values. For example, to test for genes that are regulated specifically in sub-genome A, the expression values for A:B:D is correlated with 1:0:0. In this way genes that are only expressed in A will be positively correlated and genes that are only expressed in B+D will be negatively correlated. For each GO term the distribution of correlation-scores of genes with that GO term is compared with the distribution of scores of the rest of the genes using the K-S test. The test will pick out GO terms that have a significantly high proportion of genes with high or low correlation-scores.

For comparison, the K-S tests are run both with and without the elimination algorithm. The Fisher exact test is also used, where the selection criteria is that correlation-score is at least 0.5. Note that the K-S test detects both over-expressed and under-expressed genes, while the Fisher test only detects over-expressed.

GO annotations were inferred from the annotations of the orthologous *B. distachyon* genes downloaded from: ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/ontology/go/.

2.9 Filtering contigs based on hit-counts in orthologous regions

Because the expression values are counted per contig and not per gene, miss-annotations can occur if the contig contains more than one gene. The *B. distachyon* gene sequences was therefore aligned with the contigs using “tblastx” in order to find the regions within the contigs that are orthologous with the *B. distachyon* genes specified in the genome zipper. The RNA-seq read-alignments were then parsed to count the number of hits that mapped inside the orthologous regions. For the GO analysis, contigs were filtered based on the proportion of reads that mapped to the orthologous regions. Only contigs with > 90% hits inside orthologous regions was used.

3 Results:

3.1 RNA sequencing and read mapping

The RNA sequencing generated 476 million read-pairs from 16 samples (Table 2). After discarding reads mapping to several locations within sub-genome (~25%) 34%, 40% and 42% of the reads mapped to sub-genomes A, B, and D, respectively.

Table 2: RNA-sequencing and read mapping statistics

		Mapped reads*			
Sample	Read-pairs	A	B	D	Homeoreads**
Room1_20DPA_AL_1	32919785	35 %	40 %	44 %	6.7 %
Room1_20DPA_AL_2	30833988	36 %	42 %	44 %	7.1 %
Room1_20DPA_AL_31	32374902	36 %	42 %	44 %	7.1 %
Room1_20DPA_AL_32	32685090	36 %	42 %	45 %	7.0 %
Room1_20DPA_Ref_1	34617242	33 %	39 %	40 %	3.7 %
Room1_20DPA_Ref_2	30517594	33 %	39 %	40 %	3.9 %
Room1_20DPA_SE_1	30009734	31 %	37 %	39 %	2.6 %
Room1_20DPA_SE_2	29714230	30 %	36 %	40 %	2.5 %
Room1_20DPA_TC_1	18586985	36 %	41 %	43 %	5.2 %
Room1_20DPA_TC_2	31121623	35 %	41 %	44 %	5.0 %
Room2_20DPA_AL_1	27753881	38 %	43 %	45 %	7.3 %
Room2_20DPA_AL_3	31365012	37 %	42 %	44 %	6.7 %
Room2_20DPA_SE_1	26664432	30 %	36 %	37 %	2.6 %
Room2_20DPA_SE_2	27602634	32 %	37 %	39 %	2.8 %
Room2_20DPA_TC_1	29885904	33 %	39 %	40 %	4.0 %
Room2_20DPA_TC_2	29668161	33 %	39 %	40 %	3.5 %

* 2bp mismatch allowed. Not including multireads

** Reads that map to reference zipper contigs and all three sub-genomes

The pipeline used to analyze differential expression (see methods section) only counts homeoreads that map to the genome zipper, which means that only 2.5-7.3% of the total reads were used (Table 2). Of the homeoreads used, 49.3% mapped specifically to only one of the sub-genomes (most informative), 33.4% mapped perfectly to two of the sub-genomes (less informative), and 17.3% mapped perfectly to all sub-genomes and hold no information regarding homeologous differential expression (Table 3).

Table 3: Proportion of homeoreads that map perfectly to each sub-genome or combination of sub-genomes

Specificly mapped			Ambiguously mapped			
A	**B**	**D**	**AB**	**AD**	**BD**	**ABD**
17.2 %	16.8 %	15.3 %	8.5 %	12.3 %	12.6 %	17.3 %

In total, the wheat genome zipper (B sub-genome) contains 15194 gene loci (contigs). Out of these, 6182 contigs (41%) have no homeoreads mapped to them at all, either because they are not transcribed or because the gene has been deleted from one or two of the homeologous sub-genomes. Only the 6858 contigs with a certain minimum of mapped homeoreads (>10 mapped reads in at least one of the samples) are included in the following analyses.

3.2 PCA analysis confirms good data quality

Principal component analysis (PCA) was performed to evaluate the quality of the RNA-seq data. Normally, only the first two PCs are used to show the clustering of samples, but by using the first four PCs (accounting for 66% of the variance) I was able to improve visualization of the similarity between tissues and sub-genome (Figure 8). PC1+PC4 and PC2+PC3 separate replicate samples for each tissue and sub-genome, respectively, into clear clusters. There is no indication of bad data quality (e.g. sample swapping or technical issues). However, clusters of room-replicates are slightly separated, especially for TC, indicating that there is a subtle effect of room-conditions on the transcription.

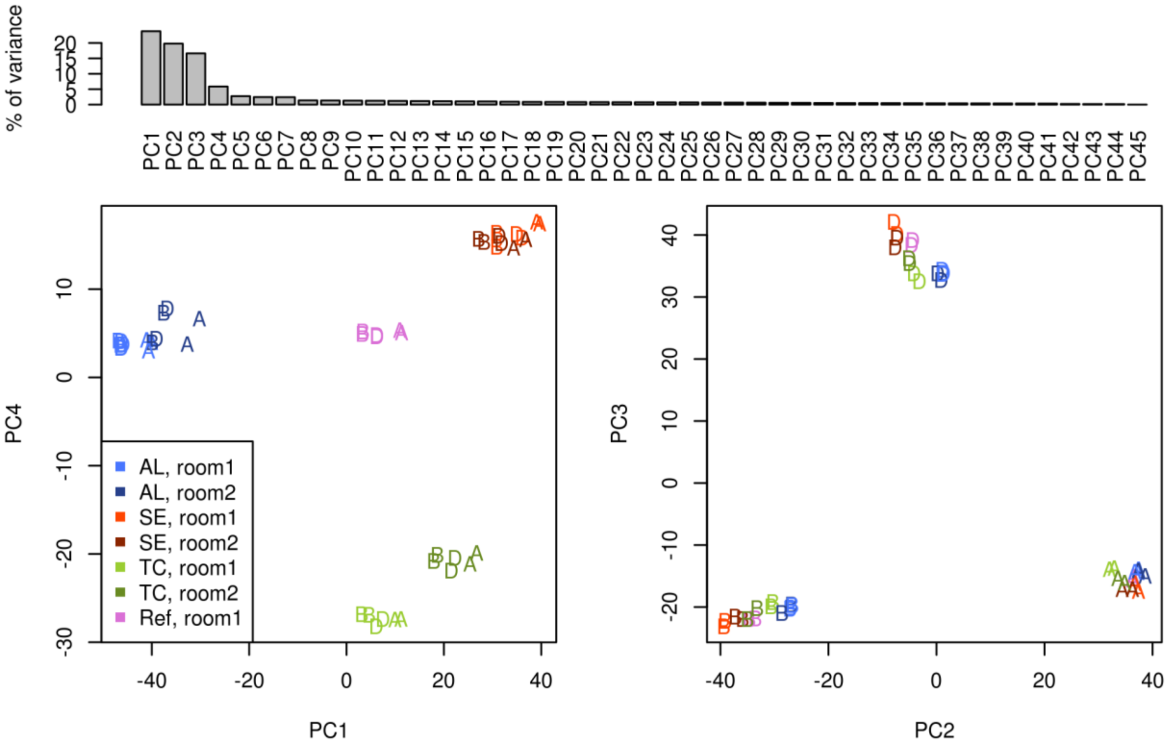

Figure 8: Principal component analysis. The letters in the plots refer to the contribution of the corresponding sub-genomes A, B and D.

3.3 Sub-genomes contribute equally to expression per chromosome

To investigate if any of the sub-genomes dominate the gene expression on the chromosome level, the total expression per chromosome for all samples was evaluated (Figure 9A-B). Figure 9B shows a normalized chromosome expression measure calculated by dividing each homeolog read count with the total count for the homeolog triplet. This prevents expression of high abundance transcript (50% of the reads map to the 7% most abundant transcripts) to dominate and override expression patterns from low abundance transcripts. From the data presented in Figure 9 each sub-genome is transcribed at about the same level for each chromosome arm but that some chromosomes show a generally higher expression level per gene than others. For example, chromosome 1 has a higher total read count than chromosome 2 although it has fewer genes. Figure 9C shows the distribution of how much each homeolog contributes to the expression of each gene. Each homeolog typically contributes around one third of the transcripts for each gene, with very few contributing more than half. In conclusion, there seems to be no overall expression level dominance of any sub-genome at the chromosome level.

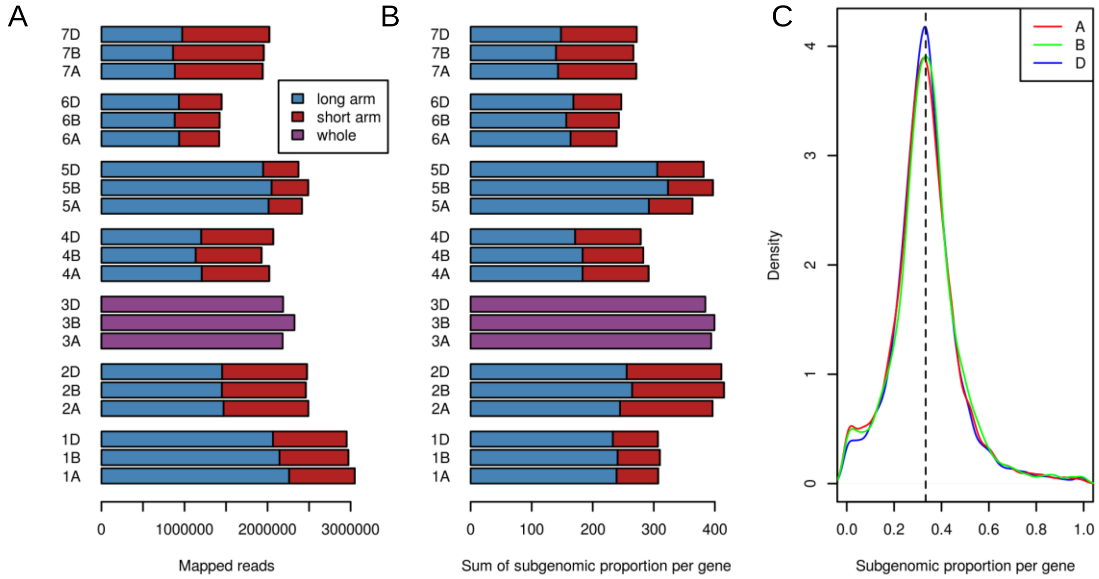

Figure 9: Sub-genomic expression levels (A) Total read count for all samples per chromosome. (B) Normalized per homeolog triplet. (C) Density plot of the normalized homeolog triplet expression, I.e. proportion of reads per sub-genome

3.4 Most genes show significant homeolog specific expression

To test whether the difference in expression between A, B and D homeologs is significant, the regression model that accounts for room, tissue and sub-genome is compared with a reduced model that only accounts for room and tissue effects. 5072 out of the 6858 genes tested show significant effect (adjusted P<0.01) of sub-genome on expression level. Similar test were done for tissue, room and interaction effect between sub-genome and tissue (Table 4). Although many genes show significant sub-genome effect, the fold change between the highest and lowest expressed homeologs is not necessarily very high. As Figure 10 shows, genes with a mean read count over 100 can show statistically significant difference even though the most expressed homeolog is less than 25% higher than the least expressed homeolog.

Table 4: Proportion of genes showing significant differential expression between sub-genomes, tissues, room or interaction effect between tissue and sub-genome and the corresponding regression models used.

Model:	Effect tested	Significant genes*
room + tissue + subgenome + tissue:subgenome	interaction	28 %
room + tissue + subgenome	subgenome	74 %
room + tissue	tissue	67 %
room + subgenome	room	15 %
tissue + subgenome		

* Adjusted p-value < 0.01. Out of 6858 expressed genes

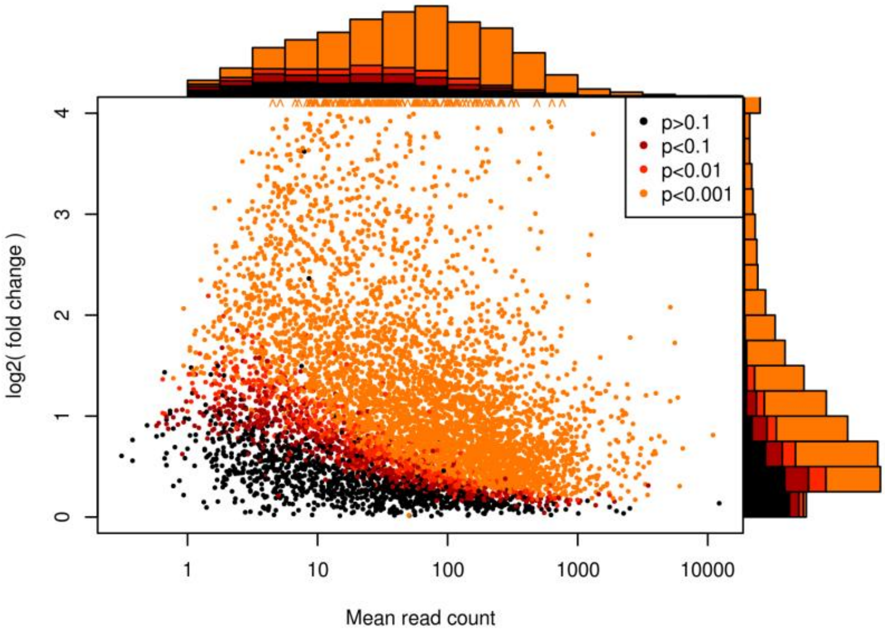

Figure 10: Read count fold change between highest and lowest expressed homeolog (mean over all samples). Colors indicate adjusted P-values for the sub-genome effect.

3.5 Hierarchical gene clustering and expression heatmap

Figure 11 shows a heatmap of the gene expression for each sub-genome and sample. Genes are ordered according to hierarchical clustering based on expression correlation. The heatmap gives an overview of the expression patterns and clusters of genes with sub-genome or tissue specific expression. The main discernible features of the gene expression clusters are either homeolog specific expression or tissue specific expression. A large portion of genes are differently expressed between SE and AL tissues (Figure 11, regions annotated on the left), while these same genes seem to be expressed at an intermediary level in TC. Another pattern that can be noted is that relatively few genes are specifically expressed in TC. Of the genes that show sub-genomic dominance there seems to be about equally many that are expressed by a single sub-genome as are expressed in two.

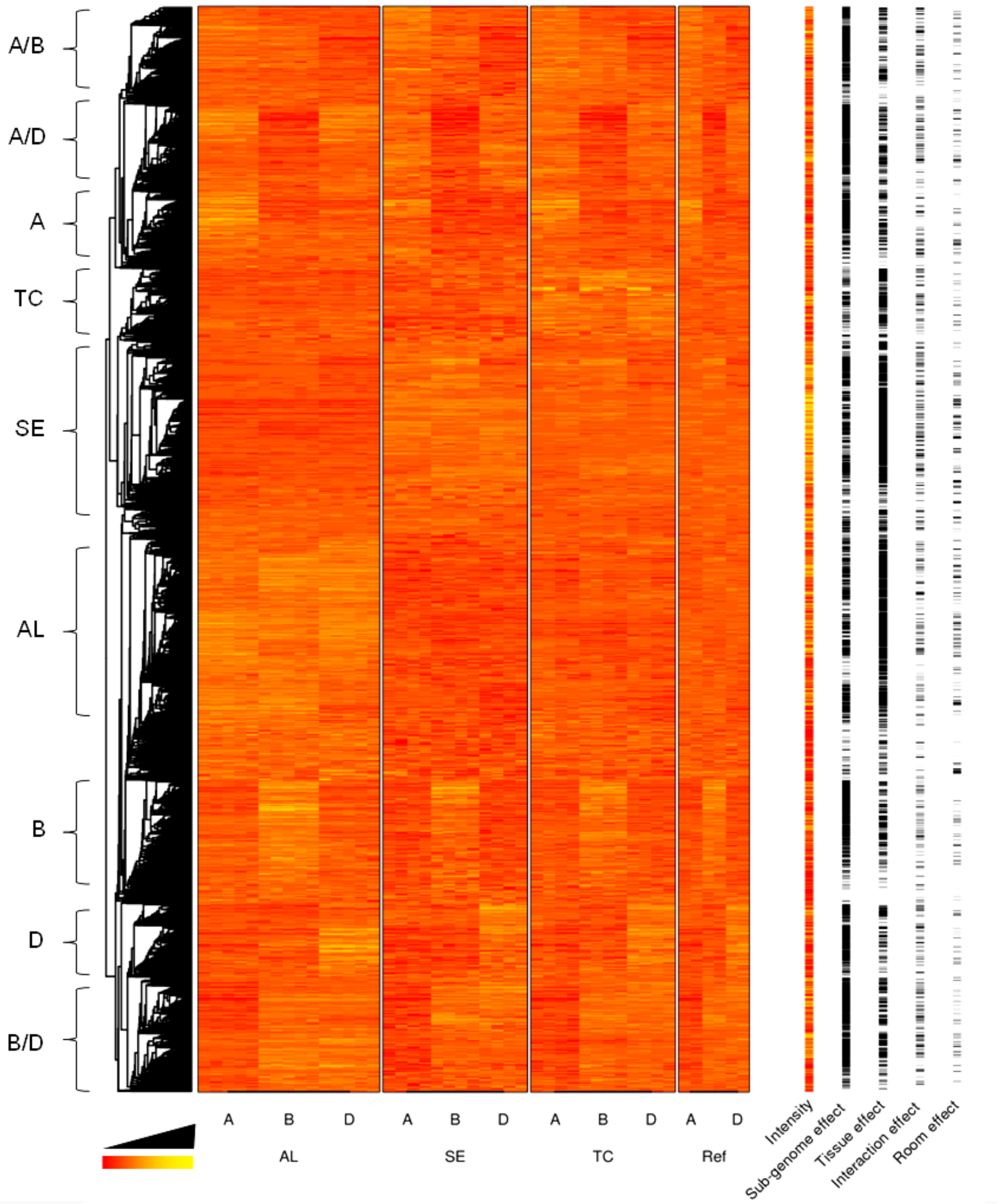

Figure 11: Clustering and heat-map of expression data. Expression levels are scaled so that the total for each gene remains constant. The "intensity" column shows the scaling factor which reflects total expression level (yellow = high expression). The effect columns show the p-values of the different tests for each gene (black = significant). The dominant features are annotated on the left.

3.6 Reciprocal silencing of homeologs between tissues

Earlier studies have reported that in allopolyploids (cotton and wheat) some genes utilize different homeologs in different tissues (e.g. the A homeolog is expressed in the leaf while the B homeolog is expressed in the root). In my dataset there is no clear-cut on/off regulation of expression hence it was necessary to choose threshold values to classify a gene as “reciprocal” or not (see methods section). The number of genes that were classified as reciprocal ranged from 0.9-0.3% in the three tissues (Table 5). Figure 12 show plots of expression profiles for a selection of these reciprocal silenced genes. By closer inspection of the single cases of reciprocal silencing, it seems that some of the genes are not truly reciprocal (Figure 12). An example of an expression profile that looks like reciprocal silencing is the one plotted in row 3, column 4. This gene has strong A genome bias in TC while in AL the A homeolog is suppressed and both B and D homeologs are up-regulated. Another example is found in row 1, column 3, which is dominated by B in AL and A+D in SE. Several of the other profiles are more doubtful examples reciprocal silencing, either because they have high variation between replicates, or because all homeologs are expressed at much higher levels in one of the tissues (e.g. row 4, column 2).

Table 5: Number of genes classified as reciprocal homeolog silencing between tissues

	AL/SE	**AL/TC**	**SE/TC**
Genes tested	4189	4321	4279
Reciprocal	38(0.9%)	17(0.4%)	13(0.3%)

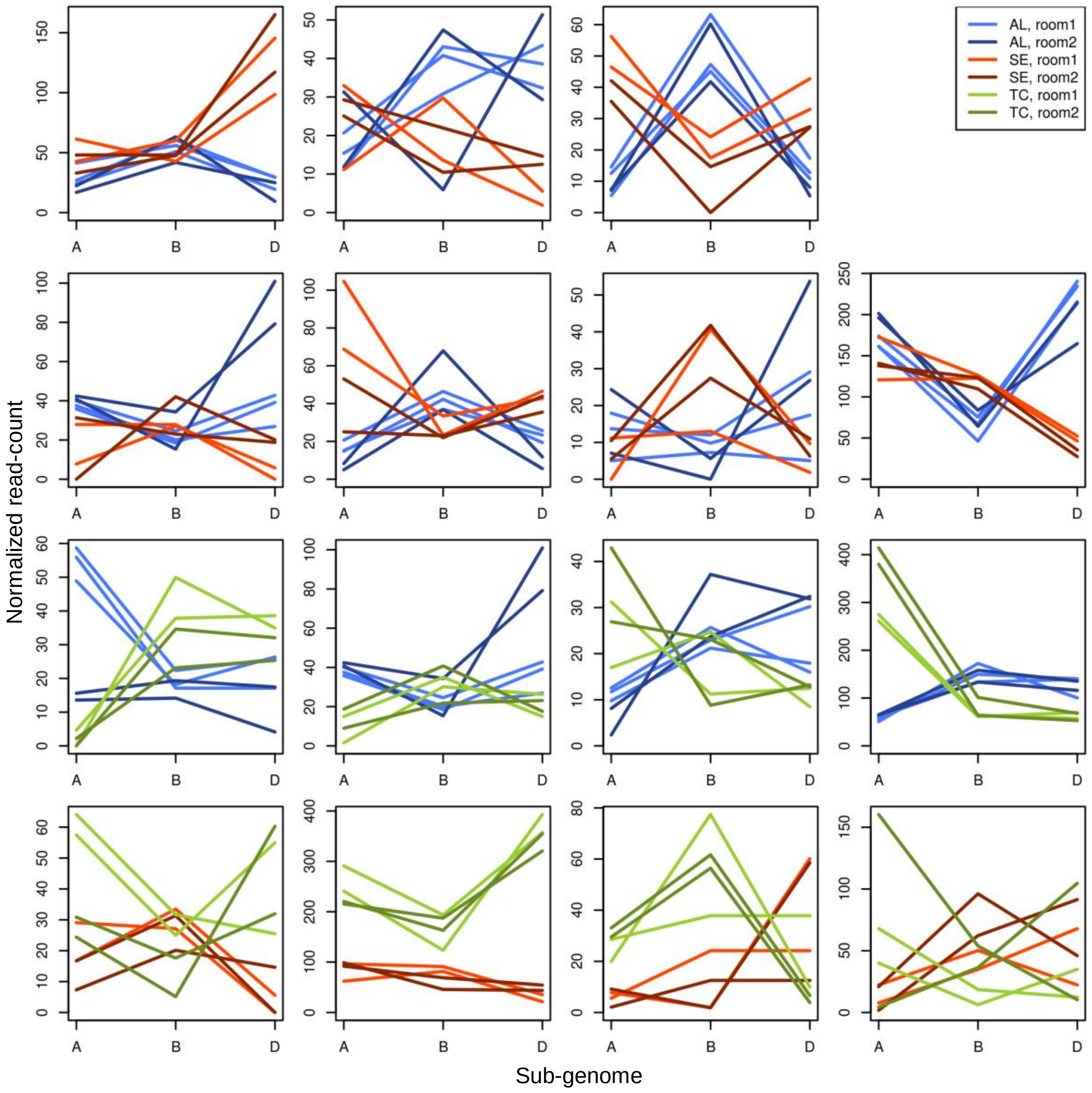

Figure 12: Expression levels (read-counts) for a selection of genes that are classified as showing signs of reciprocal homeolog silencing between tissues.

3.7 Differences in homeolog bias levels between tissues

The HSV-transformed expression values (see methods) can be useful to visualize general patterns of the homeolog expression bias. Saturation, i.e. the level of homeolog bias, is equivalent to the inverse of the fold change between the highest and lowest expressed homeolog. Distribution of homeolog bias values for each tissue indicates that homeolog bias is most prevalent in SE (Figure 13A). However, the rightwards shift in the distribution for SE and TC may be an artifact resulting from the lower total read-counts from those tissues (Figure 13C). Low read-counts give higher spread from shot-noise, which in turn increases the chance of getting higher saturation. Nevertheless, that the TC and SE tissues have very similar low read count (Figure 13C) but very different distributions in homeolog bias (Figure 13A) support the presence of true biological differences in homeolog bias between tissues.

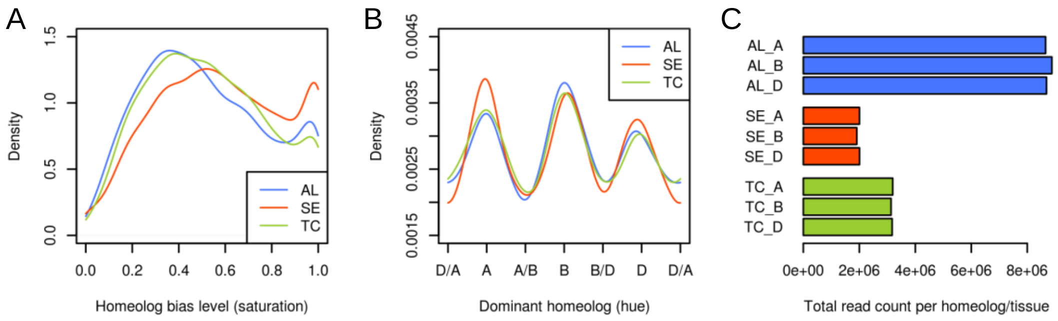

Figure 13: Tissue specific expression distribution. (A) Distribution of homeolog bias level. (B) Distribution of "hue", which indicates the dominant homeolog. Only genes with significant sub-genome effect included. Hue is given in degrees (0°-360°) and 15° was used as bandwidth when calculating kernel density. (C) Total raw read count per homeolog/tissue (not normalized).

3.8 Homeolog expression distribution show that D genome is less dominant

The hue tells us which homeolog that is dominating expression for each gene. Its distribution (Figure 13B) should not be as affected by total read counts as the homeolog bias (Figure 13A). The most striking feature of the hue distribution is the clear peaks and valleys which represents single homeolog over- and under-expression. This indicates that expression of each gene tends to be dominated by a single homeolog. A possible cause is that homeolog expression has diverged in an exponential manner, i.e. if up-regulation causes expression to double, the equivalent down-regulation would make the expression halved. If homeolog expression changes occur in this exponential manner then random changes is more likely to result in a single dominant homeolog than not.

Another feature of the hue distribution is that the peak around D is less distinct, i.e. the peak is lower and the surrounding valleys are not so deep. This indicates that the D genome is less dominant and tends to be co-expressed with A or B more frequently than A and B are co-expressed.

Log-transforming the read-counts before converting to HSV removes the effect caused by exponential expression. The transformed hue plot (Figure 14A) shows a clear dip around D while from A to B there are some smaller inconsistent fluctuations. To test if the there is any statistical significance to the observed pattern, the same graph is calculated with the read-counts for the A, B and D homeologs randomly shuffled. This is repeated to generate a null distribution under the null hypothesis that the identity of the homeologs has no consequence. This null distribution is shown as the grey area in Figure 14B which marks the interval for which at any given value of hue, 99% of the permutations falls within. For all three tissues, the dip at D goes below the 99% interval of the random distribution, indicating that it is very unlikely to have occurred by chance. This test is complicated by the fact that all points along the kernel density graph is somehow dependent on each other, e.g. if the graph goes up at one point it must go down at another so that the sum remains 1. Assuming that the D genome has an effect, it is possible to test if there is any significant differences between A and B only. Figure 14C shows the distribution when only the A and B homeologs are shuffled. This time the graph remains (mostly) inside the 99% interval, indicating that A and B are more similar to each other than D.

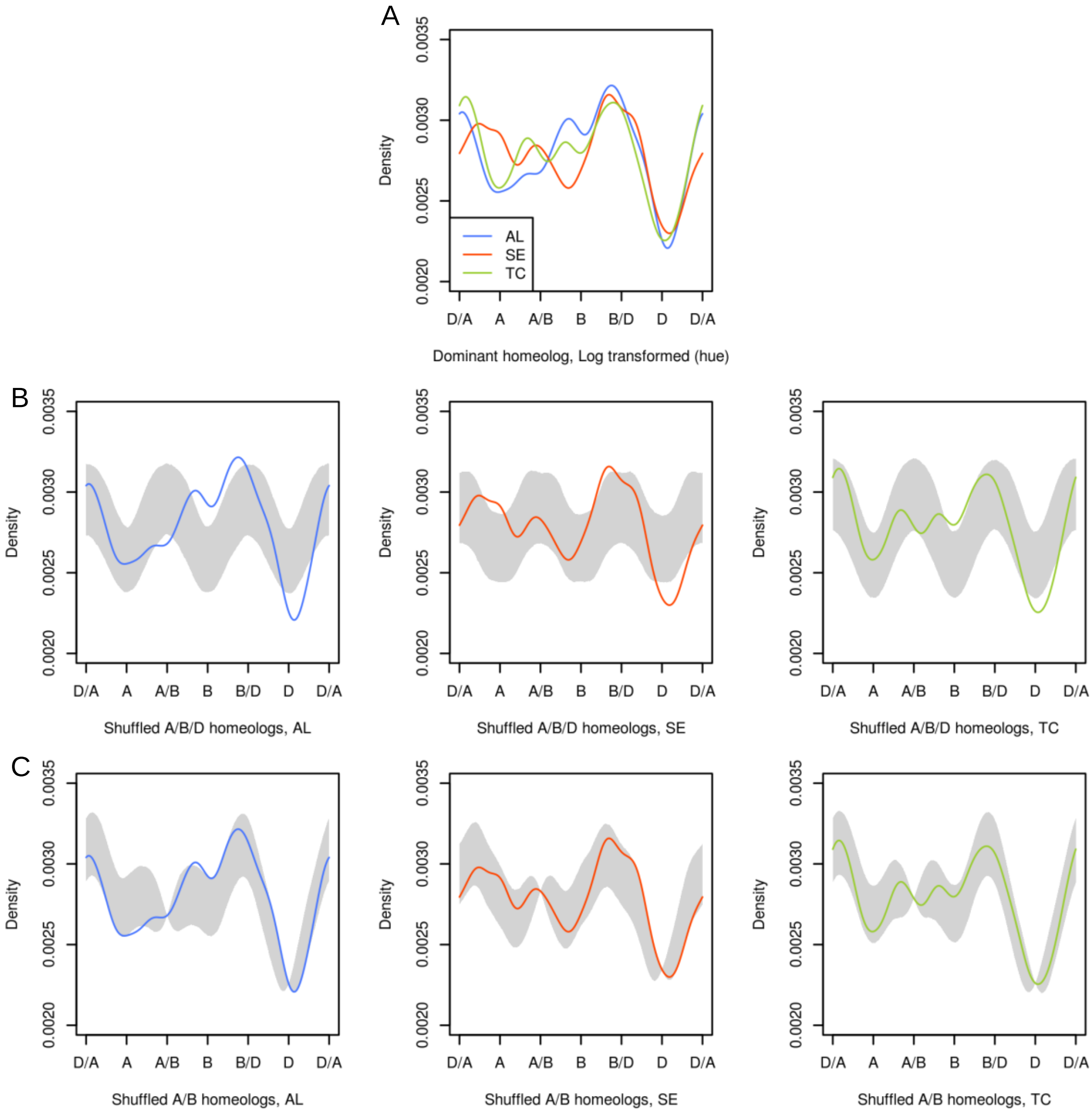

Figure 14: Log transformed hue. (A) Distribution of hue for each tissue. (B) Comparison with random distribution resulting from shuffling the A, B and D homeolog. Gray area marks the interval for each hue that contains 99% of the permutations at that hue. (C) Same, but this time only the A and B homeolog is shuffled.

3.9 Gene set enrichment fail to show genome asymmetry of GO terms

Gene set enrichment was done in R using the package topGO. GO annotation were assigned via the *B. distachyon* GO annotation as the orthologous *B. distachyon* genes are defined in the genome zipper. After excluding 43% of the genes from filtering based on orthologous hits (see methods), there was 2311 (33%) genes left with GO annotations. Analysis was run to look for specific gene regulation in each of the three sub-genomes and each of the three tissues. In addition, the analysis was run using both the "biological process" and "molecular function" ontologies. The top 10 results for each of the 12 runs are shown in Table 6-9. No adjusted P-values for sub-genome specific GO tests were significant at the 0.1 level. In the tissue specific tests, there was only a few significant GO's, all in AL (Table 7). These GO's were transcription regulation, nuclear acid binding, protein binding and zinc ion binding. Note that, for those GO's, fewer genes than expected had a correlation coefficient higher than 0.5, which indicates that the significant K-S test is based on negative correlation, i.e. genes with those GO terms tend to be down-regulated in AL compared to SE and TC.

Table 6: Gene set enrichment analysis of tissues using biological process GO terms

	GO ID	GO Term (Biological Process)	Annotated	cor > 0.5	Expected	Fisher	KS	elimKS	elimKS adjusted
AL	GO:0006355	regulation of transcription, DNA-depende...	172	24	38.63	1	1.3E-06	1.3E-06	0.0031
	GO:0044260	cellular macromolecule metabolic process	857	149	192.46	1	3E-09	0.00052	0.62
	GO:0006396	RNA processing	68	7	15.27	1	0.002	0.00203	1.00
	GO:0045037	protein import into chloroplast stroma	4	0	0.9	1	0.0029	0.00292	1.00
	GO:0090304	nucleic acid metabolic process	375	51	84.22	1	1.6E-07	0.00359	1.00
	GO:0010021	amylopectin biosynthetic process	3	0	0.67	1	0.0056	0.00562	1.00
	GO:0006461	protein complex assembly	31	2	6.96	1	0.0062	0.00623	1.00
	GO:0010051	xylem and phloem pattern formation	12	0	2.69	1	0.0073	0.00726	1.00
	GO:0006468	protein phosphorylation	195	40	43.79	0.78	0.0095	0.00946	1.00
	GO:0035670	ovule-producing ovary development	11	0	2.47	1	0.0103	0.01034	1.00
SE	GO:0055085	transmembrane transport	136	6	10.71	0.97	0.00065	0.00065	0.95
	GO:0006633	fatty acid biosynthetic process	23	2	1.81	0.55	0.00296	0.00296	0.95
	GO:0055114	oxidation-reduction process	199	19	15.67	0.21	0.00353	0.00353	0.95
	GO:0009239	enterobactin biosynthetic process	9	0	0.71	1	0.00496	0.00496	0.95
	GO:0006790	sulfur compound metabolic process	23	0	1.81	1	0.00552	0.00552	0.95
	GO:0019748	secondary metabolic process	15	0	1.18	1	0.00769	0.00769	0.95
	GO:0010255	glucose mediated signaling pathway	3	0	0.24	1	0.0077	0.0077	0.95
	GO:0006631	fatty acid metabolic process	31	2	2.44	0.71	7.4E-05	0.0098	0.95
	GO:0072522	purine-containing compound biosynthetic ...	41	1	3.23	0.97	0.0098	0.0098	0.95
	GO:0042542	response to hydrogen peroxide	10	0	0.79	1	0.0112	0.0112	0.95
TC	GO:0042364	water-soluble vitamin biosynthetic proce...	12	0	0.71	1	0.00012	0.00012	0.26
	GO:0046394	carboxylic acid biosynthetic process	78	1	4.59	0.99	1.2E-06	0.00029	0.26
	GO:0055114	oxidation-reduction process	199	6	11.71	0.98	0.00038	0.00038	0.26
	GO:0009853	photorespiration	6	0	0.35	1	0.00044	0.00044	0.26
	GO:0044242	cellular lipid catabolic process	10	0	0.59	1	0.0006	0.0006	0.29
	GO:0042445	hormone metabolic process	14	0	0.82	1	0.00081	0.00081	0.31
	GO:0008610	lipid biosynthetic process	65	0	3.83	1	9.4E-06	0.00092	0.31
	GO:0006725	cellular aromatic compound metabolic pro...	50	2	2.94	0.8	0.00159	0.00159	0.48
	GO:0009081	branched-chain amino acid metabolic proc...	12	1	0.71	0.52	0.00195	0.00195	0.52
	GO:0006979	response to oxidative stress	34	1	2	0.87	0.00224	0.00224	0.54

Table 7: Gene set enrichment analysis of tissues using molecular function GO terms

	GO ID	GO Term (Molecular Function)	Annotated	cor > 0.5	Expected	Fisher	KS	elimKS	elimKS adjusted
AL	GO:0003676	nucleic acid binding	497	52	106.43	1	4.8E-13	2E-07	0.00025
	GO:0005515	protein binding	423	64	90.58	1	4.6E-06	4.6E-06	0.0028
	GO:0008270	zinc ion binding	303	58	64.89	0.87	1.2E-05	1.2E-05	0.0049
	GO:0003677	DNA binding	245	28	52.47	1	9.6E-06	5.6E-05	0.017
	GO:0008026	ATP-dependent helicase activity	33	3	7.07	0.98	0.0027	0.0027	0.66
	GO:0043565	sequence-specific DNA binding	49	5	10.49	0.99	0.004	0.004	0.74
	GO:0003723	RNA binding	132	17	28.27	1	0.0023	0.0043	0.74
	GO:0046527	glucosyltransferase activity	17	4	3.64	0.51	0.0057	0.0057	0.83
	GO:0005524	ATP binding	500	98	107.07	0.88	0.0062	0.0062	0.83
	GO:0000049	tRNA binding	3	0	0.64	1	0.0082	0.0082	0.99
SE	GO:0016747	transferase activity, transferring acyl ...	37	1	2.96	0.96	0.00032	0.00032	0.39
	GO:0051539	4 iron, 4 sulfur cluster binding	4	0	0.32	1	0.00274	0.00274	1.00
	GO:0004620	phospholipase activity	7	0	0.56	1	0.0038	0.0038	1.00
	GO:0022891	substrate-specific transmembrane transpo...	150	8	12.01	0.93	0.0041	0.0041	1.00
	GO:0008667	2,3-dihydro-2,3-dihydroxybenzoate dehydr...	9	0	0.72	1	0.00451	0.00451	1.00
	GO:0016491	oxidoreductase activity	212	19	16.97	0.33	0.0025	0.00964	1.00
	GO:0019200	carbohydrate kinase activity	12	1	0.96	0.63	0.01006	0.01006	1.00
	GO:0042562	hormone binding	5	0	0.4	1	0.01013	0.01013	1.00
	GO:0004312	fatty acid synthase activity	6	0	0.48	1	0.01018	0.01018	1.00
	GO:0015075	ion transmembrane transporter activity	122	7	9.77	0.87	0.01214	0.01214	1.00
TC	GO:0030170	pyridoxal phosphate binding	22	1	1.39	0.76	0.00014	0.00014	0.17
	GO:0016491	oxidoreductase activity	212	9	13.35	0.93	0.00012	0.0003	0.18
	GO:0016651	oxidoreductase activity, acting on NADH ...	9	0	0.57	1	0.00161	0.00161	0.65
	GO:0003824	catalytic activity	1522	78	95.83	1	3.6E-06	0.00226	0.68
	GO:0015291	secondary active transmembrane transport...	42	2	2.64	0.75	0.00339	0.00339	0.77
	GO:0005198	structural molecule activity	107	5	6.74	0.82	0.00455	0.00455	0.77
	GO:0016614	oxidoreductase activity, acting on CH-OH...	52	2	3.27	0.85	0.00507	0.00507	0.77
	GO:0022890	inorganic cation transmembrane transport...	80	6	5.04	0.39	0.00524	0.00524	0.77
	GO:0015662	ATPase activity, coupled to transmembran...	16	0	1.01	1	0.00609	0.00609	0.77
	GO:0016810	hydrolase activity, acting on carbon-nit...	20	0	1.26	1	0.00641	0.00641	0.77

Table 8: Gene set enrichment analysis of sub-genomes using biological function GO terms

	GO ID	GO Term (Biological Process)	Annotated	cor > 0.5	Expected	Fisher	KS	elimKS	elimKS adjusted
A genome	GO:0006944	cellular membrane fusion	4	0	0.46	1	0.0018	0.0018	1.00
	GO:0044087	regulation of cellular component biogene...	4	0	0.46	1	0.0066	0.0066	1.00
	GO:0044403	symbiosis, encompassing mutualism throug...	6	1	0.69	0.52	0.0067	0.0067	1.00
	GO:0048437	floral organ development	23	3	2.64	0.5	0.0075	0.0075	1.00
	GO:0048467	gynoecium development	13	1	1.49	0.8	0.0096	0.0096	1.00
	GO:0016485	protein processing	7	0	0.8	1	0.0098	0.0098	1.00
	GO:0009646	response to absence of light	6	0	0.69	1	0.0103	0.0103	1.00
	GO:0006465	signal peptide processing	4	0	0.46	1	0.0114	0.0114	1.00
	GO:0048440	carpel development	12	1	1.38	0.77	0.0156	0.0156	1.00
	GO:0032271	regulation of protein polymerization	3	0	0.34	1	0.0233	0.0233	1.00
B genome	GO:0009733	response to auxin stimulus	36	2	4.07	0.93	0.0024	0.0024	1.00
	GO:0008283	cell proliferation	10	0	1.13	1	0.0034	0.0034	1.00
	GO:0009909	regulation of flower development	22	2	2.48	0.73	0.008	0.008	1.00
	GO:0006869	lipid transport	8	0	0.9	1	0.0084	0.0084	1.00
	GO:0010229	inflorescence development	5	0	0.56	1	0.0086	0.0086	1.00
	GO:0009817	defense response to fungus, incompatible...	5	0	0.56	1	0.0117	0.0117	1.00
	GO:0006310	DNA recombination	13	1	1.47	0.79	0.0121	0.0121	1.00
	GO:0009648	photoperiodism	10	1	1.13	0.7	0.0134	0.0134	1.00
	GO:0006323	DNA packaging	12	0	1.36	1	0.0143	0.0143	1.00
	GO:0006333	chromatin assembly or disassembly	14	0	1.58	1	0.0158	0.0158	1.00
D genome	GO:0009873	ethylene mediated signaling pathway	9	0	0.83	1	0.0012	0.0012	1.00
	GO:0007062	sister chromatid cohesion	4	0	0.37	1	0.0032	0.0032	1.00
	GO:0009112	nucleobase metabolic process	4	0	0.37	1	0.0052	0.0052	1.00
	GO:0006075	(1->3)-beta-D-glucan biosynthetic proces...	4	0	0.37	1	0.0058	0.0058	1.00
	GO:0009850	auxin metabolic process	9	0	0.83	1	0.0162	0.0162	1.00
	GO:0009629	response to gravity	10	0	0.93	1	0.0192	0.0192	1.00
	GO:0006206	pyrimidine nucleobase metabolic process	3	0	0.28	1	0.0194	0.0194	1.00
	GO:0009870	defense response signaling pathway, resi...	3	0	0.28	1	0.0211	0.0211	1.00
	GO:0006613	cotranslational protein targeting to mem...	6	0	0.56	1	0.024	0.024	1.00
	GO:0006614	SRP-dependent cotranslational protein ta...	6	0	0.56	1	0.024	0.024	1.00

Table 9: Gene set enrichment analysis of sub-genomes using molecular function GO terms

	GO ID	GO Term (Molecular Function)	Annotated	cor > 0.5	Expected	Fisher	KS	elimKS	elimKS adjusted
A genome	GO:0003677	DNA binding	245	22	26.95	0.88	0.0032	0.0052	1.00
	GO:0003690	double-stranded DNA binding	5	0	0.55	1	0.0073	0.0073	1.00
	GO:0008026	ATP-dependent helicase activity	33	4	3.63	0.5	0.0116	0.0116	1.00
	GO:0070035	purine NTP-dependent helicase activity	33	4	3.63	0.5	0.0116	0.0116	1.00
	GO:0030145	manganese ion binding	5	0	0.55	1	0.0202	0.0202	1.00
	GO:0004702	receptor signaling protein serine/threon...	9	0	0.99	1	0.0207	0.0207	1.00
	GO:0005057	receptor signaling protein activity	9	0	0.99	1	0.0207	0.0207	1.00
	GO:0016772	transferase activity, transferring phosp...	325	27	35.75	0.96	0.0217	0.0217	1.00
	GO:0003899	DNA-directed RNA polymerase activity	11	1	1.21	0.72	0.0263	0.0263	1.00
	GO:0034062	RNA polymerase activity	11	1	1.21	0.72	0.0263	0.0263	1.00
B genome	GO:0051537	2 iron, 2 sulfur cluster binding	5	0	0.57	1	0.0046	0.0046	1.00
	GO:0016705	oxidoreductase activity, acting on paire...	25	0	2.87	1	0.0137	0.0137	1.00
	GO:0004519	endonuclease activity	16	0	1.83	1	0.0141	0.0141	1.00
	GO:0004550	nucleoside diphosphate kinase activity	2	0	0.23	1	0.0207	0.0207	1.00
	GO:0004518	nuclease activity	29	4	3.32	0.43	0.0214	0.0214	1.00
	GO:0008171	O-methyltransferase activity	3	0	0.34	1	0.0218	0.0218	1.00
	GO:0016209	antioxidant activity	16	3	1.83	0.28	0.0245	0.0245	1.00
	GO:0005509	calcium ion binding	49	3	5.62	0.93	0.0248	0.0248	1.00
	GO:0003909	DNA ligase activity	2	0	0.23	1	0.0253	0.0253	1.00
	GO:0003910	DNA ligase (ATP) activity	2	0	0.23	1	0.0253	0.0253	1.00
D genome	GO:0003924	GTPase activity	42	0	3.82	1	0.0031	0.0031	1.00
	GO:0003843	1,3-beta-D-glucan synthase activity	4	0	0.36	1	0.0055	0.0055	1.00
	GO:0070011	peptidase activity, acting on L-amino ac...	113	5	10.28	0.98	0.0148	0.0148	1.00
	GO:0003746	translation elongation factor activity	6	0	0.55	1	0.0173	0.0173	1.00
	GO:0008233	peptidase activity	116	6	10.55	0.96	0.0319	0.0319	1.00
	GO:0016838	carbon-oxygen lyase activity, acting on ...	2	0	0.18	1	0.0355	0.0355	1.00
	GO:0016853	isomerase activity	47	3	4.27	0.82	0.0362	0.0362	1.00
	GO:0016866	intramolecular transferase activity	14	2	1.27	0.37	0.0395	0.0395	1.00
	GO:0004091	carboxylesterase activity	5	0	0.45	1	0.0397	0.0397	1.00
	GO:0008234	cysteine-type peptidase activity	19	1	1.73	0.84	0.0412	0.0412	1.00

4 Discussion

4.1 Differential expression of homeologs are common in the wheat genome

With the CSS assembly it is now for the first time possible to measure homeolog specific expression on a whole-genome level. About three quarters of the expressed genes showed noticeably different expression of its homeologs (Table 4), although the fold change is not very high for most genes (Figure 10). The high number of significant tests without large expression differences (about half had less than 2x fold change) reflects the high sensitivity of the DESeq test applied, and the proportion of genes that show sub-genome effect could possibly be even higher than 75% if more replicates or tissues had been included. Even though it is evident that most homeologs are expressed differently it is less clear what this means regarding evolution and partitioning of gene expression in the polyploidy wheat genome. Since the A, B, and D genomes ancestors diverged about 2-4 million years ago, these genomes evolved as diploids for several million years before they came together in polyploidy wheat. It is therefore likely that much of the expression level variation observed reflect fixed differences between the diploid ancestors (ancestral) rather than expression differences that have evolved after polyploidization (i.e. true homeolog divergence). To differentiate between these types of homeolog expression differences it is necessary to include diploid ancestors (or close relatives) in future experiment.

Although we cannot reveal the nature of homeolog expression divergence from this study it is worth commenting on possible mechanisms that might be involved. Since all homeologs are subject to the exact same cellular (trans-acting) conditions, sub-genomic expression bias is only possible through local (cis-acting) differences between the homeologs. These differences could be mutations in promoter/regulatory regions, methylation, chromatin remodeling or any sequence changes that affects the stability of the mRNA. Since polyploid wheat is fairly young (~500.000 years) (Salse et al., 2008), we would expect that true homeolog divergence should more often have evolved through DNA-methylation or chromatin modifications divergence rather than mutations in the DNA (mutation rate is very low). Several studies have linked methylation with homeolog suppression (Kashkush et al., 2002; Shitsukawa et al., 2007). Bottley and Koebner (2008) found different sub-genomic expression patterns in different wheat cultivars. They suggested that the variation is caused by epigenetic factors, which are inheritable and can affect phenotype and therefore be important for breeding. The data in my study cannot be used to look for epigenetic effects directly, but for a further study it could be interesting to identify the promoter and regulatory regions of the genes to see if any sequence variation there can explain the variation in homeolog expression. To investigate the claim that there is variations between cultivars, one could re-examine the RNA-seq dataset from a recent study that sampled the same tissues but from a different cultivar (Gillies et al., 2012).

4.2 Large scale sub-genomic expression bias

Figure 9 shows that among the homeoreads, there is no clear sign of homeolog specific expression bias on whole genome level or in each chromosome-arm. It is therefore more likely that homeolog regulation occurs on individual genes. However, the homeoreads only accounts for genes that exist in one copy in each sub-genome. When looking at the total number of mapped reads to each of the sub genomes in Table 2, there is a statistically significant difference; there are consistently more reads mapped to D genome than B, and there are more mapped to B than A. There are many possible explanations for this observed bias. Some genomes can have experienced more gene-loss than others or there could be differences in quality of genome assemblies. Another explanation could be differential expression of rRNA between homeologous chromosomes as a result of nucleolar dominance (Pikaard, 1999).

4.3 Reciprocal silencing

Tissue-specific reciprocal silencing is interesting as it could represent a form of rapid subfunctionalization (Adams et al., 2003). Bottley et al. (2006) reported that between root and leaf tissues in hexaploid wheat, 5 out of the 236 genes they tested were reciprocally silenced, and at least one of them showed complete reciprocal silencing. In my study, by comparing the homeolog specific expression between the three tissues in the endosperm, 0.3-0.9% of the genes was classified to be (partially) reciprocally silenced. This figure is comparable with the proportion of reciprocally silenced genes found in Bottley et al. (2006).

The earlier studies identifying complete reciprocal silencing in cotton (Adams et al. 2003) and wheat (Bottley et al. 2006) used SSCP to score expression or silencing. A later study on cotton (Chaudhary et al., 2009) using Sequenome MassARRAY technology tested far more genes and tissues than the Adams et al. (2003) study, but could only find partial reciprocal silencing. Similarly I did not find any cases of complete reciprocal silencing in wheat, suggesting that SSCP tends to overestimate the level of silencing and that reciprocal silencing is typically only partial. In conclusion I can only say that reciprocal silencing does occur between the tissues in the endosperm for some genes, but only at a level of partial silencing.

4.4 The special expression distribution of the D genome

On the overall level each sub-genome is equally expressed (see Figure 9C), but from Figure 13B and Figure 14 it seem like the D genome has a different expression pattern as there are fewer genes dominated specifically by the D genome compared to the A and B genome while there are more genes that are equally dominated by D+A or D+B. In other words, the expression pattern of the D homeologs seems to be slightly more similar to either A or B than the A and B are to each other. Another case where the D genome is different is seen in Table 3, where there are more reads that map to both sub-genomes D+A or D+B, than there are reads that map to both A+B. This indicates that the D genome sequence is more similar to A and B than the A and B are to each other.

One explanation for this pattern can be that during the time the A and B genome has evolved as a tetraploid, the extra gene-copies could have created a genetic redundancy that allows for mutations to accumulate at a higher rate and thus also lead to diverged homeolog expression levels. However, it is unlikely that enough mutations could accumulate during that relatively short evolutionary period. Another explanation is that these differences where present before polyploidization. An unpublished phylogenetic study looking at the origins of the wheat sub-genomes have shown that the D genome may have evolved from a hybridization between the ancestors of the A and B genomes (S. R. Sandve, personal communication). The phylogenies show that similar numbers of D homeologs are closest related to A and B genomes, which fits well with my observations of sequence and expression level similarity.

4.5 Gene set enrichment does not suggest genomic asymmetry in wheat

One goal of the gene set enrichment test was to test for genomic asymmetry (Feldman et al., 2012), in which some functions or traits are controlled by a single sub-genome. The assumption is that the asymmetrically controlled functions are annotated with specific GO terms. Test results indicated that no GO functional group of genes was specifically up or down-regulated in any of the sub-genomes (Table 8-9). In the topGO manual (Alexa and Rahnenfuhrer, 2010) they do comment that P-values from enrichment analysis often are not very extreme and that multiple testing correction of P-values could mask out interesting results. However, this does not change the fact that randomizing the data would result in equally (un)extreme P-values.

Gene ontology analysis of RNA-seq expression data from the SE and AL tissues of wheat has been performed before (Gillies et al., 2012), where they showed that the two tissues display different molecular functions. Such divergence in GO terms between tissues is expected based on the highly different biological functions of the SE and AL tissues. Hence, I used GO enrichment tests for tissue specific regulation as a control for the power/sensitivity of the GO term test applied in my thesis. The low proportion of GO terms significant for tissue specific expression in my study (Table 6-7) therefore seem to indicate low method sensitivity or that some other factor influence the GO analysis. However, the comparison with Gillies et al., (2012) might not be valid because of differences in time of tissue sampling (they sampled at 6, 9 and 14 DPA) compared to the data analyzed in this thesis.

There are many different ways to perform gene set enrichment, all of which can give different results and conclusions (Hung et al., 2012). In this study, both Fisher exact test and K-S test was applied, which resulted in highly different P-values. The Fisher test was performed for genes that had a certain level of positive correlation the condition, while the K-S test is based on the correlation-score which can test significant on negative correlations as well as positive correlations. How the GO annotations were acquired may also have affected the results. The GO annotations where inferred from the annotations of orthologous *B. distachyon* genes. In the Gillies et al. (2012) study they used blast2GO, which uses blast to find annotated orthologous genes. They also used a GO slimmer which reduces the number of GO terms by using a select set of broader terms.

Finally, it is possible that genomic asymmetry is caused by different genome content due to loss of genes or divergent gene family sizes in certain sub-genomes and not by homeolog specific suppression of expression. This study only includes the genes that exist in one copy each of the three sub-genomes and therefore cannot account for asymmetry in genome content. It is also possible that asymmetry in phenotypic traits are caused by functional mutations in certain sub-genomes while sub-genomic expression levels are unaffected and therefore not detectable in this study.

4.6 Limitations of the homeoreads method

The pipeline developed to calculate expression levels is designed to give unbiased relative levels of expression between homeologous genes when no gene-models are available. The use of homeoreads to find homeologous regions and the use of genome zipper contigs to select genes has some inherent limitation that could be improved upon:

4.6.1 Using contigs to represent genes

The homeoreads method relies on the assumption that each gene resides within a single contig and that there are no other transcribed elements within that contig. These assumptions are simplifications and do not hold for all genes. Where the survey sequence assembly is highly fragmented it is likely that the exons of a gene will end up in different contigs (*GENE1_B* in Figure 15A). In that case, since only a single contig represents the gene, only a part of the gene is considered and the absolute expression level will be underestimated. As the aim of this thesis is to analyze expression between sub-genomes, an underestimate of the total expression level *per se* does not affect our conclusions. However, lower read counts do increase the uncertainty of the expression measurements. Another situation occur where the assembly is less fragmented and several genes or transcribed elements co-reside in a single large contig (Figure 15B). Assuming that sub-genome B is used as reference and we want to measure the expression of GENE2 which resides in Contig_B_4, since GENE3_B is inside the same contig, we end up measuring the sum of the expression of both genes. As it is illustrated, only GENE3 is expressed and is therefore measured instead. To circumvent this problem, an extra filtering step was performed before the GO analysis that ensures that there is no mistaken gene identity.

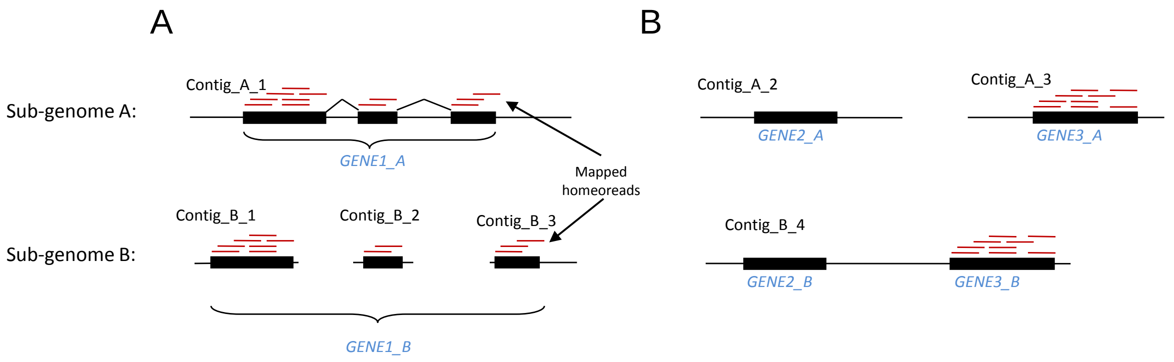

Figure 15: Illustration of pitfalls of assuming that one contig = one gene. (A) Fragmentation of exons. (B) Multiple genes in one contig.

4.6.2 Arbitrary choice of reference zipper

Figure 15 illustrate examples where the choice of reference zipper can have an impact on the calculated expression levels. A comparison of the expression levels obtained when using each of the sub-genomes as reference showed that most expressed genes give varied counts and that only about a quarter of the expressed genes are completely unaffected. This tells us that the situations like those illustrated in Figure 15 are common. However, the relative homeolog expression levels should not be affected and most major conclusions from this study are expected to be valid if the reference genome is changed. Unfortunately, I did not have time to re-run the major analyses using a different reference genome due to time constraints.

4.6.3 Inflated variance from ambiguous hits

When a read maps perfectly to several of the sub-genomes it is not known from where it actually originated. The read-count of these ambiguously mapped reads is divided proportionally to each sub-genome based on the reads that are mapped specifically to only one sub-genome. When there are only a few reads that map specifically, the variance from shot-noise of the specifically mapped reads will propagate to the calculation of expression. Since DESeq is designed to estimate the shot-noise based on the raw read-counts, the software will underestimate the expression variance when given the mixed read-counts. The result is that genes with low proportion of specifically mapped reads will have a slightly higher level of false positives, but it is unclear to what extent.

4.6.4 Effect of discarding multireads within sub-genome

Multireads are reads that map to several positions in the genome. This occurs when a gene or parts of a gene exists in several copies. The homeoreads method uses the fact that reads from homeologous genes map to all three homeologs. If a gene exists in more than one copy within the sub-genome, it would be difficult to determine which of them to count as the homeologous copy. To avoid that problem, all multireads within the sub-genome are discarded. One consequence of this is that genes that have paralogs will not be counted or will have reduced read-counts, but there is also a possibility that homeolog specific bias is introduced for some genes. As an example, take a gene that has three homeologs A, B and D. At some point in evolution, the B homeolog is copied, creating the copy B*. It is possible that in a 100bp region there is one homeoSNP in each of A,B and D and that B* carries the same SNP as B plus one extra SNP. In that case, reads originating from B or B* will map to each other and be discarded as multireads. Reads originating from A and D, on the other hand, will not map to B* because there are 3 mismatches. The result is that reads from A or D will be counted while reads from B will not, in other words: the count is biased. It is not clear to what extent this kind of bias occurs in this study, but it is reasonable to expect a few false-positives as a consequence of this.

4.6.5 Is 2bp mismatch a good choice?

When the reads are mapped they are allowed to have 2bp mismatch. This is allowed so that the reads can map to all three homeologs, not only the one they originate from. This number was chosen arbitrarily (it is the default setting). Since reads will fail to map to all three homeologs if the region contains too many homeoSNPs, allowing more mismatches should be consider.

By assuming that homeoSNPs are distributed uniformly we can calculate the expected proportion of reads that will fail to map to all three homeologs. Based on the observed number of specifically and ambiguously mapped reads we find that the chance of a base pair having a homeoSNP is about 2.5%. When only 2bp mismatch is allowed this means that about 34% of the reads will fail to map to one or more of the three homeologs. By using the same model with 3bp mismatch allowed there will be only 20% of reads that fails to map. The additional homeoreads acquired by increasing the allowed number of mismatches will mostly map specifically and therefore contribute to a more accurate estimation of the homeolog specific expression. From these calculations it seems to be better to allow 3bp mismatch or even more. However, there might be a higher chance of multireads when allowing more mismatches and the mapping procedure might get more time consuming. Unfortunately, there was no time to re-run the mapping with altered parameters.

5 Conclusion

It is evident that the use of RNA-seq in combination with the newly available CSS assembly as reference makes it possible to measure the genome-wide homeolog specific expression levels at unprecedented accuracy and precision. This provides new and exciting opportunities for researchers to understand polyploid genome function in general, and hexaploid wheat biology in specific.

Analysis of the homeolog specific expression showed that most genes have a reproducible homeolog specific expression bias, although the level of expression difference may not be very strong. For many genes the homeolog specific expression is developmentally controlled and there are signatures of reciprocal silencing suggesting partitioned use of homeologs in different cellular context. The homeolog bias seems to affect genes individually as there is no clear bias on chromosome level and there was no evidence for genomic asymmetry in relation to different molecular or biological functions. The most striking result was that the D genome exhibits a different expression distribution, possibly reflecting an evolutionary history involving hybridization in the D-genome lineage.

However, in this study it was not possible to know if the observed homeologous expression patterns are a result of polyploidization or if the pattern was established already in the progenitor species of each sub-genome. Previous studies using microarrays have solved this by using synthetic hexaploid wheat and include samples from the parental plants for comparison. To do such a study again, only using RNA-seq instead, would be of great help to further improve knowledge of expression regulation in hexaploid wheat. Another gap in the knowledge is how much homeolog specific expression varies between different cultivars of wheat. An RNA-seq study that compares different cultivars could also investigate the importance of epi-genetic effects as a mediator of expression variance. If homeolog specific expression varies without differences in DNA-sequence it would be a good idea to investigate methylation levels in those genes.

6 References

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proceedings of the National Academy of Sciences of the United States of America **100**: 4649–54.

Akhunova, A.R., Matniyazov, R.T., Liang, H., and Akhunov, E.D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC genomics **11**: 505.

Alexa, A. and Rahnenfuhrer, J. (2010). topGO: Enrichment analysis for Gene Ontology.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics (Oxford, England) **22**: 1600–7.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology **11**: R106.

Bottley, A. and Koebner, R.M.D. (2008). Variation for homoeologous gene silencing in hexaploid wheat. The Plant journal□: for cell and molecular biology **56**: 297–302.

Bottley, A., Xia, G.M., and Koebner, R.M.D. (2006). Homoeologous gene silencing in hexaploid wheat. The Plant journal□: for cell and molecular biology **47**: 897–906.

Chagué, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.-M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., and Chalhoub, B. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. The New phytologist **187**: 1181–94.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J. a, Verma, N., Springer, N.M., and Wendel, J.F. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). Genetics **182**: 503–17.

Dolezel, J., Kubaláková, M., Paux, E., Bartos, J., and Feuillet, C. (2007). Chromosome-based genomics in the cereals. Chromosome research□: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology **15**: 51–66.

Eilam, T., Anikster, Y., Millet, E., Manisterski, J., and Feldman, M. (2008). Nuclear DNA amount and genome downsizing in natural and synthetic allopolyploids of the genera Aegilops and Triticum. Genome / National Research Council Canada = Génome / Conseil national de recherches Canada **51**: 616–27.

Feldman, M., Levy, A.A., Fahima, T., and Korol, A. (2012). Genomic asymmetry in allopolyploid plants: wheat as a model. Journal of experimental botany **63**: 5045–59.

Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, a a, and Vega, J.M. (1997). Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. Genetics **147**: 1381–7.

Gillies, S. a, Futardo, A., and Henry, R.J. (2012). Gene expression in the developing aleurone and starchy endosperm of wheat. Plant biotechnology journal **10**: 668–79.

Grass Phylogeny Working Group, Barker, N.P., Clark, L.G., Davis, J.I., Duvall, M.R., Guala, G.F., Hsiao, C., Kellogg, E.A., and Linder, H.P. (2001). Phylogeny and Subfamilial Classification of the Grasses (Poaceae). Annals of the Missouri Botanical Garden **88**: pp. 373–457.

He, P., Friebe, B.R., Gill, B.S., and Zhou, J.-M. (2003). Allopolyploidy alters gene expression in the highly stable hexaploid wheat. Plant molecular biology **52**: 401–14.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proceedings of the National Academy of Sciences of the United States of America **99**: 8133–8.

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in bioinformatics **13**: 281–91.

Kashkush, K., Feldman, M., and Levy, A. a (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics **160**: 1651–9.

Kellogg, E.A. and Bennetzen, J.L. (2004). The evolution of nuclear genome structure in seed plants. American journal of botany **91**: 1709–25.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–73.

Mayer, K.F.X., Taudien, S., Martis, M., Simková, H., Suchánková, P., Gundlach, H., Wicker, T., Petzold, A., Felder, M., Steuernagel, B., Scholz, U., Graner, A., Platzer, M., Dolezel, J., and Stein, N. (2009). Gene content and virtual gene order of barley chromosome 1H. Plant physiology **151**: 496–505.

Mayer, K.F.X., Waugh, R., Brown, J.W.S., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., Wise, R.P., and Stein, N. (2012). A physical, genetic and functional sequence assembly of the barley genome. Nature **491**: 711–6.

Mochida, K., Yamazaki, Y., and Ogihara, Y. (2004). Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. Molecular genetics and genomics□: MGG **270**: 371–7.

Nesbitt, M. and Samuel, D. (1996). Hulled Wheats. In Proc. 1st Internat. Workshop Hulled Wheats, International Plant Genetic Resources Institute, S. Padulosi, K. Hammer, and J. Heller, eds, pp. 41–100.

Olsen, O. (2004). Nuclear endosperm development in cereals and Arabidopsis thaliana. The Plant cell **16 Suppl**: S214–227.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H.,

Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I. V., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457**: 551–556.

Pikaard, C.S. (1999). of transcription. **1385**: 478–483.

Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., Zhang, H., Pang, J., Han, F., and Liu, B. (2012). Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. BMC biology **10**: 3.

Raghavan, V. (2003). Some reflections on double fertilization, from its discovery to the present. New Phytologist **159**: 565–583.

Rapp, R.A., Udall, J.A., and Wendel, J.F. (2009). Genomic expression dominance in allopolyploids. BMC biology **7**: 18.

Rastogi, S. and Liberles, D. a (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC evolutionary biology **5**: 28.

Salse, J., Chagué, V., Bolot, S., Magdelenat, G., Huneau, C., Pont, C., Belcram, H., Couloux, A., Gardais, S., Evrard, A., Segurens, B., Charles, M., Ravel, C., Samain, S., Charmet, G., Boudet, N., and Chalhoub, B. (2008). New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides. BMC genomics **9**: 555.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. a, Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science (New York, N.Y.) **326**: 1112–5.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, a a (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. The Plant cell **13**: 1749–59.

Shitsukawa, N., Tahira, C., Kassai, K.-I., Hirabayashi, C., Shimizu, T., Takumi, S., Mochida, K., Kawaura, K., Ogihara, Y., and Murai, K. (2007). Genetic and epigenetic alteration among three homoeologous genes of a class E MADS box gene in hexaploid wheat. The Plant cell **19**: 1723–37.

The International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature **463**: 763–8.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England) **25**: 1105–11.

Wall, A.M., Riley, R., and Gale, M.D. (1971). The position of a locus on chromosome 5B of Triticum aestivum affecting homoeologous meiotic pairing. Genetical Research **18**: 329–339.

Yoo, M.-J., Szadkowski, E., and Wendel, J.F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity **110**: 171–80.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science (New York, N.Y.) **296**: 79–92.

Zhang, H., Bian, Y., Gou, X., Zhu, B., Xu, C., Qi, B., Li, N., Rustgi, S., Zhou, H., Han, F., Jiang, J., Von Wettstein, D., and Liu, B. (2013). Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. Proceedings of the National Academy of Sciences of the United States of America **110**: 3447–52.

Appendix A Scripts for the homeoreads pipeline

The following scripts are used to calculate the estimated homeolog specific expression as described in the method section. It is assumed that TopHat read alignment has been performed (step 1 in the pipeline, see methods). As input it requires an ordered list of contig names (reference zipper contigs) and a table that contains the description of each sample (room, DPA, tissue, rep#) as well as paths to the BAM files generated by TopHat when mapping against sub-genomes A, B and D.

Script hierarchy:

- **main.R** – Runs step 2-6 of the pipeline for all samples
    - **extractZipperAndFindHomeoreads.R** – Step 2-3. Uses the following scripts:
        - **scripts/reformatSam**
        - **scripts/extractZipperContigs**
        - **scripts/extractHomeoReads**
        - **scripts/getSubset.py**
        - **scripts/addPairedEndSuffix.py**
    - **scripts/removeBiasedReads.R** – Step 4
    - **scripts/countReadGroups.R** – Step 5
    - **scripts/calcExprRecursive.R** – Step 6

```r
#
# Main script for running homeoread method
#

source("scripts/extractZipperAndFindHomeoreads.R")
source("scripts/removeBiasedReads.R")
source("scripts/countReadGroups.R")
source("scripts/calcExprRecursive.R")

#
# folder structure:
# /                # current (project) dir, contains main script(s)
# /scripts/        # sourced R-scripts, shell/python scripts
# /input/          # All input data: bamTbl, raw genezipper, BdAnnot, Bd.fasta ...
# /output/         # All files generated by scripts
# /output/Room1_20DPA_AL_3.1/  # (example generated path for a sample)


# calcHomeoreadExpression function
#
# Parameters:
#   bams: Vector of characters specifying the path of the BAM files mapped against
#         sub-genome A, B and D respectively
#   outdir: Directory to save output files
#   contigsFile: Path to file containing list of reference contigs
#   refGenome: Specify which genome which is used as reference ("A","B" or "D")
calcHomeoreadExpression <- function(bams,outdir,contigsFile,refGenome){

  # create directory for output files
  dir.create(outdir)

  # Extract zipper contigs and find homeoreads
  homeoreads <- extractZipperAndFindHomeoreads(bams,outdir,contigsFile,refGenome)

  # Filter biased reads
  homeoreads <- removeBiasedReads( homeoreads )

  # Count perfect matches
  countTbl <- countReadGroups( homeoreads, contigsFile )

  # Estimate expression from each homeolog
  expTbl <- calcExprRecursive(countTbl)

  # Store counts and estimated expression
  write.table( cbind(countTbl,expTbl[,2:5]), file=file.path(outdir, "counts.txt"),
               sep="\t", col.names=T, row.names=F, quote=F)

  return(expTbl[,2:4])
}



# load list of BAM files with description
bamTbl <- read.table("input/bamFileDesc.txt",header=T,stringsAsFactors=F)


refGenome <- "B"
contigsFile <- "output/simpleZipper/GZcontigsB.txt"

# Run homeoread method for each sample
for( i in 1:nrow(bamTbl) ){
  outdir <- file.path("output",bamTbl$SampleID[i])
  bams <- as.character( bamTbl[i,c("bamA","bamB","bamD")] )
  calcHomeoreadExpression(bams,outdir,contigsFile,refGenome)
}

# Put all expression values in one table
fullExpTbl <- NULL
for( i in 1:nrow(bamTbl) ){
  outdir <- file.path("output",bamTbl$SampleID[i])
  tbl <- read.table( file=file.path(outdir, "counts.txt"), sep="\t", header=T,stringsAsFactors=F)
  exp <- tbl[,10:12]
  exp <- round(exp)
  names(exp) <- paste(bamTbl$SampleID[i],c("A","B","D"),sep="_")
  if( is.null(fullExpTbl) )
    fullExpTbl <- exp
  else
    fullExpTbl <- cbind(fullExpTbl,exp)
}

write.table(fullExpTbl, "output/fullExpTbl.txt", sep="\t", col.names=T, row.names=F, quote=F)
```

main.R

```r
# Parameters:
```

```
#   bams: Vector of characters specifying the path of the BAM files mapped against
#         sub-genome A, B and D respectively
#   outdir: Directory to save output files
#   contigsFile: Path to file containing list of reference contigs
#   refGenome: Specify which genome which is used as reference ("A","B" or "D")
extractZipperAndFindHomeoreads <- function(bams,outdir,contigsFile,refGenome){
  subGenome=c("A","B","D")

  # Extract zipper contigs
  refBAM <- bams[subGenome==refGenome]
  refSAM <- file.path(outdir,"ref.sam")
  system(paste("scripts/extractZipperContigs",refBAM,contigsFile,refSAM))


  # Find homeoreads
  homeoBAMs <- bams[subGenome!=refGenome]

  # extract reads from BAM files that have same read ID as those mapped to reference zipper
  homeoSAM1 <- file.path(outdir,"homeo1.sam")
  homeoSAM2 <- file.path(outdir,"homeo2.sam")
  system(paste("scripts/extractHomeoReads", homeoBAMs[1], refSAM, homeoSAM1))
  system(paste("scripts/extractHomeoReads", homeoBAMs[2], refSAM, homeoSAM2))

  # Order sam-files according to sub-genome
  samFiles <- character(3)
  names(samFiles) <- subGenome
  samFiles[refGenome] <- refSAM
  samFiles[subGenome!=refGenome] <- c(homeoSAM1, homeoSAM2)

  # load sam-file data
  hitsA <- read.table(file=pipe( paste("scripts/reformatSam", samFiles["A"]) ),
                      col.names = c("read","contig","pos","CIGAR","MD"),
                      stringsAsFactors=F)
  hitsB <- read.table(file=pipe( paste("scripts/reformatSam", samFiles["B"]) ),
                      col.names = c("read","contig","pos","CIGAR","MD"),
                      stringsAsFactors=F)
  hitsD <- read.table(file=pipe( paste("scripts/reformatSam", samFiles["D"]) ),
                      col.names = c("read","contig","pos","CIGAR","MD"),
                      stringsAsFactors=F)

  # make table of reads using hits to A as reference
  idxAinD <- match(hitsA$read,hitsD$read)
  idxAinB <- match(hitsA$read,hitsB$read)

  homeoreads <- data.frame( read=hitsA$read,
                            contig.A=hitsA$contig,
                            contig.B=hitsB$contig[ idxAinB ],
                            contig.D=hitsD$contig[ idxAinD ],
                            MD.A = hitsA$MD,
                            MD.B = hitsB$MD[ idxAinB ],
                            MD.D = hitsD$MD[ idxAinD ],
                            CIGAR.A = hitsA$CIGAR,
                            CIGAR.B = hitsB$CIGAR[ idxAinB ],
                            CIGAR.D = hitsD$CIGAR[ idxAinD ],
                            stringsAsFactors=F)
  homeoreads <- homeoreads[!(is.na(idxAinB) | is.na(idxAinD)),] # remove reads that are not also in B
      and D

  return(homeoreads)
}
```

scripts/extractZipperAndFindHomeoreads.R

```
#!/bin/bash

# Extract only the interesting fields (i.e. QNAME, RNAME, POS, CIGAR, MD) from a sam file.

# validate parameters
if [[ $# != 1 ]]
then
  echo "Invalid number of parameters!"
  echo ""
  echo "Usage: reformatSam <sam-file>"
  exit 1
fi

cat $1 | perl -lne 'm|([^\t]+)\t[^\t]+\t([^\t]+)\t([^\t]+)\t[^\t]+\t([^\t]+)\t.+\tMD\:Z\:([^\t]+)|g&&
    print "$1\t$2\t$3\t$4\t$5"'
```

scripts/reformatSam

```
#!/bin/bash

# Read bamfile, remove all multi-reads, keep only reads that map to given contigs
# and add pair-end suffix ("_1"/"_2") if not present.
# Write result to outfile in sam format

# validate parameters
if [[ $# != 3 ]]
then
  echo "Invalid number of parameters!"
  echo ""
  echo "Usage: extractZipperContigs <bam-file> <contig-file> <output-file>"
  exit 1
fi

BAMfile=$1
contigs=$2
outfile=$3

pbin=$(dirname $0)
# samtools is not in PATH when executing from R, so add it now.
export PATH=/local/genome/bin:$PATH

echo "samtools view $BAMfile | grep \"NH:i:1$\" | python $pbin/getSubset.py -c 3 $contigs | python $pbin/addPairedEndSuffix.py > $outfile"
samtools view $BAMfile | grep "NH:i:1$" | python $pbin/getSubset.py -c 3 $contigs | python $pbin/addPairedEndSuffix.py > $outfile
```

scripts/extractZipperContigs

```
#!/bin/bash

# Read bamfile, remove all multi-reads, add pair-end suffix ("_1"/"_2") if not present
# and keep only reads have same read IDs as in given SAM file.
# Write result to outfile in sam format

# validate parameters
if [[ $# != 3 ]]
then
  echo "Invalid number of parameters!"
  echo ""
  echo "Usage: extractHomeoReads <bam-file> <sam-file> <output-file>"
  exit 1
fi

BAMfile=$1
SAMfile=$2
outfile=$3

pbin=$(dirname $0)
# samtools is not in PATH when executing from R, so add it now.
export PATH=/local/genome/bin:$PATH

echo "samtools view $BAMfile | grep \"NH:i:1$\" | python $pbin/addPairedEndSuffix.py | python $pbin/getSubset.py -c 1 $SAMfile > $outfile"
samtools view $BAMfile | grep "NH:i:1$" | python $pbin/addPairedEndSuffix.py | python $pbin/getSubset.py -c 1 $SAMfile > $outfile
```

scripts/extractHomeoReads

```
import sys
from optparse import OptionParser

usage = "Usage: python %prog <list of identifiers to extract>"
parser = OptionParser(usage)
parser.add_option("-c", dest="c", type="int", default=1, metavar="N",
                  help="Specify column to search for identifers [default: %default]")
parser.add_option("-i", "--invert",
                  action="store_true", dest="invertSubset", default=False,
                  help="keep lines that are NOT in list")
(opt,args) = parser.parse_args()
if len(args) != 1:
    parser.error("Incorrect number of arguments")

# load the identifiers
items_file = args[0] # file containing list of identifiers in first column
items = set()
with open(items_file, ’rb’) as f:
    for line in f.readlines():
        items.add(line.split("\t")[0].rstrip())

# Write only lines (from stdin) that contains one of the specified identifiers in specified column to stdout
c = opt.c - 1
line = sys.stdin.readline()
if opt.invertSubset:
    while line:
        if not line.split("\t")[c] in items:
            sys.stdout.write(line)
        line = sys.stdin.readline()
else:
    while line:
        if line.split("\t")[c] in items:
            sys.stdout.write(line)
        line = sys.stdin.readline()
```

scripts/getSubset.py

```
import sys

# append paired-end suffix ("_1"/"_2") to sam file
while 1:
    try:
        line = sys.stdin.readline()
    except KeyboardInterrupt:
        break
    if not line:
        break
    c = line.split("\t")
    if not c[0][-2] == '_':
        c[0] = c[0]+"_"+str((int(c[1])&0xC0)>>6)
    sys.stdout.write("\t".join(c))
```

scripts/addPairedEndSuffix.py

```
library("inline")

#  checkReadBias function
#  ----------------------.    .-.     .-.     .-.
#                         \   /   \   /   \   /   '._.'
#                          '-'     '-'     '-'
# Check for bias that occurs when potential homeoreads from the other two sub-genomes
# would have failed to map because of too many mismatches.
#
# Example: A homeoread matches perfectly to sub-genome A, when mapped against
# the other sub-genomes it has the following pattern of mismatches:
#   Sub-genome B: --T----C--         MD="2T4C2"   snpPos[] = {2,7}   snp[] = {'T','C'}
#   Sub-genome D: -G--------         MD="1G8"     snpPos[] = {1}     snp[] = {'G'}
# If a read was transcribed from the corresponding locus in B and mapped against D
# or vice versa, it would have 3 mismatches and would have been discarded, thus any
# reads transcribed from this locus would be result in biased read counts.
#
# The pattern of mismatches is given in the MD field from the tophat result, so the
# first thing to do is to decode the MD string to a position (snpPos[]) and base (snp[])
# for each mismatch aka SNP (see examples).
#
# Note that deletions and insertions are ignored except that knowledge of insertions
# are used to correct the SNP positions. Insertions are described in the CIGAR field.
#
# The number of mismatches between B and D is the sum of number of SNPs for both
# minus the SNPs that are the same for both. E.g:
#   Sub-genome B: -C---T----         MD="1C3T4"   snpPos[] = {1,5}   snp[] = {'C','T'}
#   Sub-genome D: -----T----         MD="5T4"     snpPos[] = {5}     snp[] = {'T'}
# Here the total number of SNPs is 3 but 2 are the same SNP. Mapping B against D would
# yield 3-2=1 mismatch.
#
# Often the reference sequence for one sub-genome is reverse complementary with the
# other sub-genome. In that case the SNP pattern needs to be reversed and the bases
# switched before comparing. E.g
#   Sub-genome B: ----A---G-         MD="4A3G1"   snpPos[] = {4,8}   snp[] = {'A','G'}
#   Sub-genome D: -----T----         MD="5T4"     snpPos[] = {5}     snp[] = {'T'}
# Here it looks like mapping B agianst D would yield 3 mismatches, but actually it is
# the same as the previous example except that sub-genome B has been reversed.
#
# The number of mismatches is calculated for both forward and reverse, if both are
# higher than 2 then it flagged as biased.
#
# parameters:
#  MD1: vector of characters containing the MD field of one of the other sub-genomes
#  MD2: vector of characters containing the MD field of the second other sub-genome
#  CIGAR1: vector of characters containing the CIGAR field of one of the other sub-genomes
#  CIGAR2: vector of characters containing the CIGAR field of the second other sub-genome
# Return value:
#  vector of logical that is TRUE for biased reads.
checkReadBias <- cfunction(signature(MD1="character", MD2="character", CIGAR1="character", CIGAR2="character"),
"
SEXP res;
  int nprotect = 0, c, num, pos, nSnp, refGen;
  int n = length(MD1);
  PROTECT(res = allocVector(LGLSXP, n)); nprotect++;
  int *pRes = INTEGER(res);
  int snpPos[2][10] = {0,0,0,0}; // SNP positions (2 reference genomes)
  char snp[2][10] = {0,0,0,0}; // SNP base type (2 reference genomes)
  int totSnp[2] = {0,0}; // number of SNPs found per ref. genome
  const char *pStr;

  for( int i = 0; i < n; i++ ){
    for( refGen = 0; refGen < 2; refGen++ ){
      if( refGen == 0 )
        pStr = CHAR(STRING_ELT(MD1, i));
      else
        pStr = CHAR(STRING_ELT(MD2, i));

      pos = 0; nSnp = 0;
      for( int j = 0; pStr[j] != 0; j++){
        switch(pStr[j]){
          case ’^’: // deletion
            do {
              j++;
            } while(pStr[j] != 0 && pStr[j] >= ’A’ && pStr[j] <= ’Z’ );
            j--;
            break;

          case ’A’:
          case ’C’:
          case ’G’:
          case ’T’:
          case ’N’: // SNPs
            snpPos[refGen][nSnp] = pos; // store SNP position
            snp[refGen][nSnp] = pStr[j]; // store SNP base
            pos++;  // increment base position count
            nSnp++; // increment snp count
            break;

          default:
            if( pStr[j] >= ’0’ && pStr[j] <= ’9’ ){
              num = (int)(pStr[j] - ’0’);
              j++;
              while(pStr[j] >= ’0’ && pStr[j] <= ’9’) {
                num = 10*num + (int)(pStr[j] - ’0’);
                j++;
              };
              j--;
            }
            pos += num; // increment base position count
            break;
        }
      } // next j (character in MD)
      totSnp[refGen] = nSnp; // store number of SNPs found

      if( refGen == 0 )
        pStr = CHAR(STRING_ELT(CIGAR1, i));
      else
        pStr = CHAR(STRING_ELT(CIGAR2, i));

      pos = 0;
      for( int j = 0; pStr[j] != 0; j++){
        // get number
        num = 0;
        while(pStr[j] >= ’0’ && pStr[j] <= ’9’) {
          num = 10*num + (int)(pStr[j] - ’0’);
          j++;
        };

        switch(pStr[j]){
          case ’M’: // Match
            pos += num; // update position
            break;

          case ’I’: // Insertion. Adjust snpPos accordingly
            for ( nSnp = 0; nSnp < totSnp[refGen]; nSnp++ ){
              if(snpPos[refGen][nSnp] >= pos){ // if SNP comes after insertion
                snpPos[refGen][nSnp] += num; // increment SNP pos with number of insertions
              }
            }
            pos += num; // update position
            break;

          case ’D’: // Deletion
          case ’N’: // Junction (intron)
          default:
            break;
        }
      } // next j (character in CIGAR)
    } // next refGen

    // Compare the SNP patterns and calculate number of mismatches
    // for both forward and reverse

    // assume first that all SNPs yields mismatches
    int cMismatchForward = totSnp[0] + totSnp[1];
    int cMismatchReverse = totSnp[0] + totSnp[1];

    // reduce mismatch count for SNPs that are same in both
    for ( int snp0 = 0; snp0 < totSnp[0]; snp0++ ){
      for ( int snp1 = 0; snp1 < totSnp[1]; snp1++ ){
        if( snpPos[0][snp0] == snpPos[1][snp1] ){  // SNP at same pos
          if( snp[0][snp0] == snp[1][snp1] ){      // same SNP
            cMismatchForward -= 2; // reduce mismatch count
            break;
          }
        }
        // reverse check
        if( snpPos[0][snp0] == 100-snpPos[1][snp1] ){  // SNP at reverse pos
          if( ( snp[0][snp0] == ’C’ && snp[1][snp1] == ’G’ ) ||
              ( snp[0][snp0] == ’T’ && snp[1][snp1] == ’A’ ) ||
              ( snp[0][snp0] == ’A’ && snp[1][snp1] == ’T’ ) ||
              ( snp[0][snp0] == ’G’ && snp[1][snp1] == ’C’ ) ||
              ( snp[0][snp0] == ’N’ && snp[1][snp1] == ’N’ ) ){  // complementary SNP
            cMismatchReverse -= 2; // reduce mismatch count
            break;
          }
        }
      }
    }

    // if both counts are higher than 2 then read is flagged as biased.
    if( cMismatchForward > 2 && cMismatchReverse > 2)
      pRes[i] = 1; // more than 2 mismatches. Biased
    else
      pRes[i] = 0; // 2 or less mismatches. Unbiased


  } // next i (homeoread)


  UNPROTECT(nprotect);
  return res;
")

removeBiasedReads <- function( homeoreads ){
  # use MD field to see which reads that have perfect hits in A, B and D
  inA <- homeoreads$MD.A == "101"
  inB <- homeoreads$MD.B == "101"
  inD <- homeoreads$MD.D == "101"
  # Which reads have only perfect hits in A, B or D
  onlyA <- inA & !inB & !inD
  onlyB <- !inA & inB & !inD
  onlyD <- !inA & !inB & inD

  isBiasedA <- checkReadBias(homeoreads$MD.B, homeoreads$MD.D, homeoreads$CIGAR.B, homeoreads$CIGAR.D)
  isBiasedB <- checkReadBias(homeoreads$MD.A, homeoreads$MD.D, homeoreads$CIGAR.A, homeoreads$CIGAR.D)
  isBiasedD <- checkReadBias(homeoreads$MD.A, homeoreads$MD.B, homeoreads$CIGAR.A, homeoreads$CIGAR.B)

  isBiased <- (isBiasedA & onlyA) | (isBiasedB & onlyB) | (isBiasedD & onlyD)

  return(homeoreads[!isBiased,])
}
```

scripts/removeBiasedReads.R

```
countReadGroups <- function(homeoreads, contigsFile){

  # Make table of contigs with number of hits per group
  tbl <- as.data.frame(table( inA = homeoreads$MD.A == "101",
                              inB = homeoreads$MD.B == "101",
                              inD = homeoreads$MD.D == "101",
                              contig = homeoreads$contig.B))

  # Reshape table
  tbl$abd <- as.logical(tbl$inA) + 2*as.logical(tbl$inB) + 4*as.logical(tbl$inD)
  tbl2 <- reshape(tbl[,c("contig","Freq","abd")],
                 idvar=c("contig"),timevar="abd",direction="wide")
  colnames(tbl2) <- c("contig","abd","Abd","aBd","ABd","abD","AbD","aBD","ABD")

  # load genome zipper contigs
  contigGZ <- read.table( file=contigsFile,sep="\t",stringsAsFactors=F)$V1
  # complete the table by adding zeros for all contigs that don’t have expression
  missingContigs <- contigGZ[!(contigGZ %in% tbl2$contig)]
  zeros <- numeric(length(missingContigs))
  zeroTbl <- data.frame(contig=missingContigs,abd=zeros,Abd=zeros ,aBd=zeros, ABd=zeros,
                       abD=zeros, AbD=zeros, aBD=zeros, ABD=zeros )
  tbl3 <- rbind(tbl2,zeroTbl)
```

```r

  # sort according to genome zipper
  row.names(tbl3) <- tbl3$contig
  return( tbl3[ contigGZ , ] )
}
```

scripts/countReadGroups.R

```r
calcExprRecursive <- function(countTbl, MAX_RECURSIONS = 100, CONVERGE_DIFF = 0.01){
  # make empty table for results
  expTbl <- data.frame( contig = countTbl$contig, expA = numeric(nrow(countTbl)),
                        expB = numeric(nrow(countTbl)), expD = numeric(nrow(countTbl)),
                        recursions = numeric(nrow(countTbl)))

  for( irow in 1:nrow(countTbl)){
    A = countTbl$Abd[irow]; B = countTbl$aBd[irow]; D = countTbl$abD[irow]
    AB = countTbl$ABd[irow]; BD = countTbl$aBD[irow]; AD = countTbl$AbD[irow]
    ABD = countTbl$ABD[irow]

    # Handle cases where all homeologs has 0 reads to avoid divide by 0
    if( AB==0 & BD==0 & AD==0 & ABD==0 ){
      expTbl$expA[irow] <- A
      expTbl$expB[irow] <- B
      expTbl$expD[irow] <- D
    }else{

      # initial estimates
      Xa <- A + AB/2 + AD/2 + ABD/3
      Xb <- B + AB/2 + BD/2 + ABD/3
      Xd <- D + AD/2 + BD/2 + ABD/3

      # recursively estimate new expression values until they converge
      for( i in 1:MAX_RECURSIONS){
        Xa.next <- A + AB*Xa/(Xa+Xb) + AD*Xa/(Xa+Xd) + ABD*Xa/(Xa+Xb+Xd)
        Xb.next <- B + AB*Xb/(Xa+Xb) + BD*Xb/(Xb+Xd) + ABD*Xb/(Xa+Xb+Xd)
        Xd.next <- D + AD*Xd/(Xa+Xd) + BD*Xd/(Xb+Xd) + ABD*Xd/(Xa+Xb+Xd)
        if( abs(Xa - Xa.next) < CONVERGE_DIFF  &
            abs(Xb - Xb.next) < CONVERGE_DIFF  &
            abs(Xd - Xd.next) < CONVERGE_DIFF ) {
          break
        }

        Xa <- Xa.next
        Xb <- Xb.next
        Xd <- Xd.next
      }
      expTbl$recursions[irow] <- i

      expTbl$expA[irow] <- Xa.next
      expTbl$expB[irow] <- Xb.next
      expTbl$expD[irow] <- Xd.next
    }
  }
  return(expTbl)
}
```

scripts/calcExprRecursive.R