PLANT METHODS

REVIEW Open Access

Analysing complex *Triticeae* genomes – concepts and strategies

Manuel Spannagl*, Mihaela M Martis, Matthias Pfeifer, Thomas Nussbaumer and Klaus FX Mayer*

Abstract

The genomic sequences of many important *Triticeae* crop species are hard to assemble and analyse due to their large genome sizes, (in part) polyploid genomes and high repeat content. Recently, the draft genomes of barley and bread wheat were reported thanks to cost-efficient and fast NGS technologies. The genome of barley is estimated to be 5 Gb in size whereas the genome of bread wheat accounts for 17 Gb and harbours an allo-hexaploid genome. Direct assembly of the sequence reads and access to the gene content is hampered by the repeat content. As a consequence, novel strategies and data analysis concepts had to be developed to provide much-needed whole genome sequence surveys and access to the gene repertoires. Here we describe some analytical strategies that now enable structuring of massive NGS data generated and pave the way towards structured and ordered sequence data and gene order. Specifically we report on the GenomeZipper, a synteny driven approach to order and structure NGS survey sequences of grass genomes that lack a physical map. In addition, to access and analyse the gene repertoire of allo-hexaploid bread wheat from the raw sequence reads, a reference-guided approach was developed utilizing representative genes from rice, *Brachypodium distachyon*, sorghum and barley. Stringent sub-assembly on the reference genes prevented collapsing of homeologous wheat genes and allowed to estimate gene retention rate and determine gene family sizes. Genomic sequences from the wheat sub-genome progenitors enabled to discriminate a large number of sub-assemblies between the wheat A, B or D sub-genome using machine learning algorithms. Many of the concepts outlined here can readily be applied to other complex plant and non-plant genomes.

Keywords: *Triticeae* genomes, Grass genomes, Wheat genome, Barley genome, GenomeZipper, Genome analysis

Review

Introduction

The *Triticeae* tribe comprises some of the most economically important crops including bread wheat, barley and rye. Bread wheat ranked third in world crop production with 681 million tons in 2011 [1], making it an indispensable source for our everyday diet. Domestication history of *Triticeae* dates back several thousand years. They consequently have a complex genetic history [2].

The genomes of many *Triticeae* species including wheat and barley appear to be extremely challenging to assemble and analyse due to their genome size, high repeat content, complex transposable element structure and, in part, polyploid genome [3,4].

With an estimated genome size of ~5 Gb the barley genome is significantly larger than the human genome, however exceeded by the bread wheat genome with ~17 Gb. Bread wheat contains an allo-hexaploid genome with three sub-genomes, namely the A, B and D sub-genome. It has been speculated that the bread wheat genome originated from hybridization between cultivated tetraploid emmer wheat (AABB) and diploid goat grass (DD) about 8000 years ago [5].

Complementing the genome size, many *Triticeae* genomes show a very high degree of repetitive elements (~80% in bread wheat [6]). These repetitive stretches can span several 100kbs and have a complex architecture and composition. Consequently assembly of long scaffolds or even whole chromosome sequences from NGS survey sequences using current technology is still an open problem [Review [7]]. Although synteny is pronounced in grasses and in general the gene order appears to be well conserved [8], repeat and TE activities

* Correspondence: manuel.spannagl@helmholtz-muenchen.de; k.mayer@helmholtz-muenchen.de
MIPS/IBIS, Helmholtz Center Munich, National Research Center for Environment and Health, Ingolstaedter Landstr. 1, Neuherberg, Germany

as well as structural rearrangements contribute to the formation of pseudogenes, gene fragments and changes in local gene order [4].

With the availability of economic and rapid NGS technologies whole-genome sequence surveys of many grass genomes including a number of *Triticeae* species were generated recently [9,10]. While the direct assembly of reads into pseudo-chromosomes or scaffolds is hampered by the genomes' repetitiveness and size, the gene inventory, gene order and chromosomal positioning of genetic elements such as genes and markers is of high interest not only for breeders but also helps to shed light on the evolutionary history of the respective plants and the *Triticeae* in general.

Consequently, numerous novel strategies and concepts were developed over the last few years to order [11,12], analyse [9] and compare complex *Triticeae* genomes even in the light of the challenges and limitations described. Here, we highlight a few of these concepts that were applied to analyse the recently published genome sequences of barley [10] and bread wheat [9] and describe the methodology used in more detail.

Many of the concepts and strategies described and discussed here are not restricted to the *Triticeae* but can be applied to other complex grass and plant genomes that have not been sequenced and analysed so far due to their genome size and/or polyploid nature.

A strategy for the comprehensive analysis of polyploid genomes. an ortholome approach for the analysis of hexaploid bread wheat

The orthologous group assembly (OA) is a strategy which aims to identify the gene repertoire of polyploid genomes based on low- and medium coverage, long-read (454) whole genome shotgun data. This approach was applied for the comprehensive sequence analysis of hexaploid bread wheat [9] and facilitated the identification of 94,000-96,000 wheat genes. Due to a stringent assembly protocol, rare sequence polymorphisms are sufficient in order to maintain and distinguish distinct copies of homeologous genes, which might be collapsed by a brute-force *de novo* assembly. In contrast to traditional assembly approaches, the orthologous group assembly focuses on the gene space and uses conserved sequence homology to genes of closely related plant species of smaller size and repeat content. Briefly, orthologous genes of multiple species are grouped and, for each gene family, one representative protein is selected thereby defining an orthologous group representative (OGR). Subsequently, based on conserved sequence homology, raw sequencing reads are associated to the OGRs. These associations define sequence read collections for each individual OGR. Each of these sequence bins is independently assembled using stringent parameters that can be estimated from *in silico* simulations of whole genome sequencing experiments. The assembled gene fragments are re-aligned and ordered along the OGR to facilitate estimation of the gene copy number of the target genome and further simplifies downstream analysis (Figure 1).

Definition of an orthologous gene set

In a first step, orthologous gene clusters were computed for the reference genomes of three grass genomes (*Brachypodium distachyon* [13], *Sorghum bicolor* [14] and *Oryza sativa* [15]) originating from different grass sub-families plus publicly- available barley full-length cDNAs. The OrthoMCL software version 1.4 [16] was used to calculate pairwise sequence similarities between all input protein sequences using BLASTP [17]. Markov clustering of the resulting similarity matrix defines the orthologous cluster structure.

A total of 86,944 coding sequences from these four grasses were clustered into 20,496 gene families. 9,843 clusters contained sequences from all four genomes.

In a second step, one representative gene model was selected from each of the 20,496 orthologous gene clusters, using the following strategy: 1. BLASTX [17] of all contigs from the bread wheat LCG assembly against all grass genes used in the OrthoMCL analysis; 2. Select the gene in each cluster that concentrates the most distinct wheat contigs; 3. If genes pool the same number of wheat contigs, select the one with the longest protein sequence as the representative gene model.

Genes with associated wheat contigs were identified for all but 445 gene clusters resulting in a total of 20,051 representative gene models.

Allocating the sequencing reads to orthologous gene representatives

The 454 sequence reads are clustered to the orthologous gene representatives using conserved sequence similarity. Thereby, pre-processing of the raw 454 sequencing data is a helpful approach to reduce computational complexity. Especially repetitive sequences, which represent up to 80% of the grass genomes [18,19], considerably extend search space, and thus memory as well as time requirements. Moreover, increased gene copy numbers caused by repetitive mechanisms and transposable element activity complicate and hampers downstream gene family analysis. For a fivefold whole genome shotgun dataset from hexaploid wheat using this strategy, 77% of raw sequence data was removed and 24 Gb out of 83 Gb of sequence kept for the orthologous group assemblies.

Afterwards reads were aligned to the orthologous gene representatives using BLASTX [17] and reported alignments filtered for minimum alignment length (minimum of 30 amino acids (AA)) and minimum alignment identity. Thereby, different alignment identity thresholds were

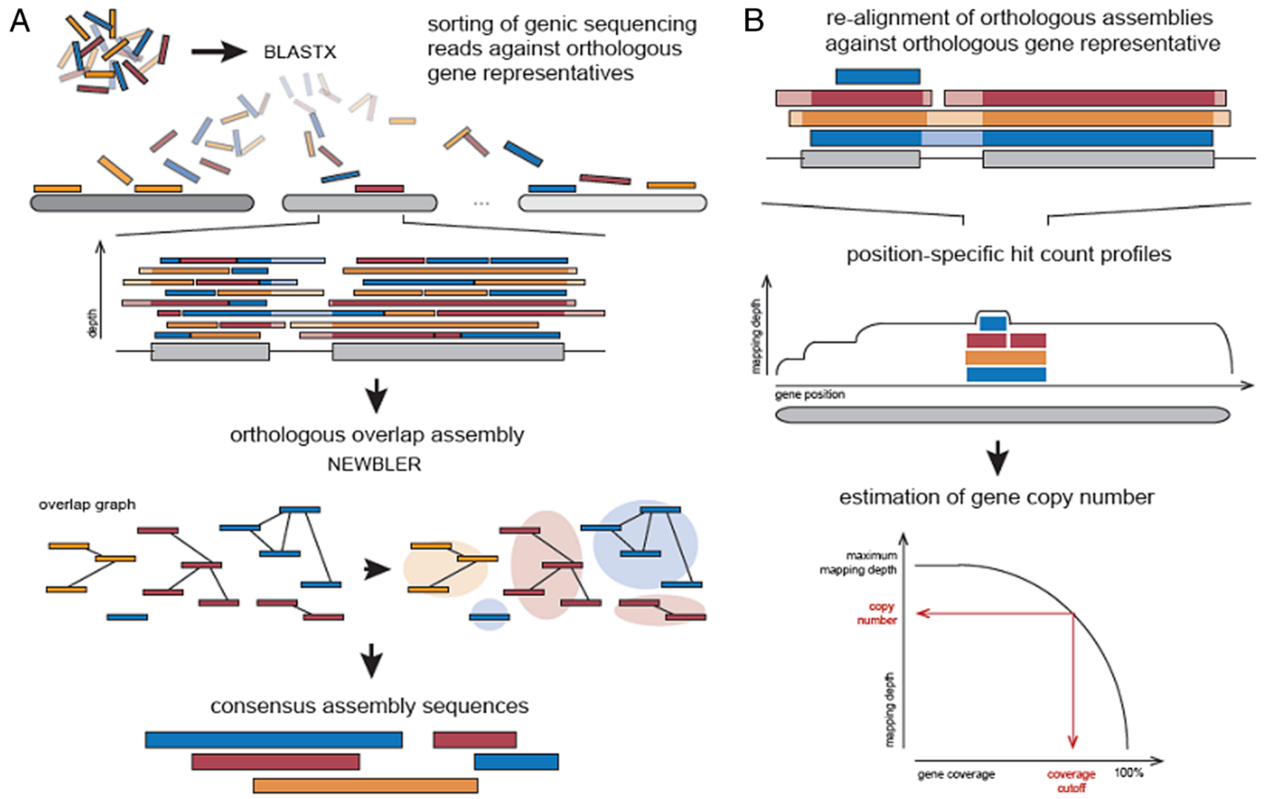
A
BLASTX
sorting of genic sequencing reads against orthologous gene representatives
...
depth
orthologous overlap assembly
NEWBLER
overlap graph
consensus assembly sequences
B
re-alignment of orthologous assemblies against orthologous gene representative
position-specific hit count profiles
mapping depth
gene position
estimation of gene copy number
maximum mapping depth
copy number
mapping depth
gene coverage
coverage cutoff
100%

Figure 1 Schematic representation of the orthologous group assembly workflow and the protocol for the estimation of gene copy number. Grey boxes represent the protein sequence of orthologous group representative, whereas lines connecting boxes depict exon boundaries. Coloured boxes visualize sequencing reads and assembled sequences, respectively. The colour code groups sequences that originate from the same genome and light colouring visualize non-coding regions. **A** The orthologous assembly algorithm sorts the raw sequencing reads to corresponding orthologous gene representatives based on sequence similarity (BLASTX). Then, each sequence bin is separately assembled using NEWBLER, an overlap graph assembler which identifies overlapping sequence reads first and then creates a consensus assembly sequence based on the overlap graph. **B** For estimating the gene copy number the orthologous assembly sequences were re-aligned to the corresponding template (BLASTX) and, thus, ordered along its protein sequence. The alignments are transferred into a position-specific hit count profile that counts the number of distinct sub-assemblies mapped to each amino acid of the template protein. Based on the cumulative coverage distribution of the hit-count profile, the final gene copy number is determined as the maximum number of distinct sub-assemblies covering a defined proportion of the template gene.

applied accounting for the different evolutionary distances between bread wheat and the reference species used for the orthologous group analysis. Overall, 4 million (6%) of the wheat repeat-filtered sequencing reads passed the applied alignment criteria, and approximately two-third (68%) of these matched a single OGR. For the remaining 454 reads, that map multiple representatives, only the first-best match (FBH), were considered. Almost all (19,483) of the selected orthologous representatives were detected by at least one 454 sequence read indicating the good representation of the majority of wheat genes in the whole genome shotgun sequence dataset.

Generating gene-centric "sub-assemblies" using the Newbler overlap assembler

Sequence information and quality scores were extracted for the 4 million aligned 454 reads from the original sequencing library files. For each orthologous group, individual assemblies were computed using the Newbler de novo assembly software, which generates larger contigs based on overlaps between reads [20]. Thereby, the assembly parameter specifying the minimum alignment identity (*mi*) to accept overlaps between reads has to be selected with caution. On the one hand, a too relaxed *mi* parameter may cause a collapse of homeologous sequence copies, but, on the other hand, too stringent *mi* parameter would overestimate gene copies due to sequencing errors. Two methods applied for selection of the best *mi* parameter are discussed in the following section. The final set of "sub-assembly" sequences was created by combining singletons (454 reads that do not overlap with any other reads) and the assembled contigs (> = 100 bp).

The assembly results are strongly dependent on the chosen minimum alignment identity parameter. Whereas, 76% of reads were assembled into contigs applying 97% minimum alignment identity, the number of assembled reads dropped to 51% applying a minimum alignment identity of 100%, respectively (Figure 2). On the contrary, the number of 454 reads remaining singletons almost doubles between 97% and 100% *mi*. This fact already indicates that small-scale adjustments of this parameters influence the assembly results and the subsequent downstream analysis. Since this parameter is thus of critical importance in the following we discuss two method for the selection of the minimum overlap identity parameter in the following sections.

Estimation of gene copy number based on gene coverage

In order to estimate the gene copy number, the orthologous sub-assemblies were re-aligned to the corresponding template gene. The same alignment parameters used for mapping the raw sequencing reads were applied. Thereby, spliced alignments of sub-assemblies were additionally considered and all consecutive high-scoring segment pairs (HSPs) along the template protein sequence accepted. Using this strategy, we account stretches of non-coding sequence regions of the sub-assemblies which span and/or reach into introns or 5′ and 3′ UTRs. Then, the alignments of the sub-assemblies were transferred into a position-specific hit count profile by counting the number of aligned sub-assemblies located at a specific amino acid position of the template sequence. The algorithm converts the hit count profiles into a cumulative coverage distribution. The distribution curve ranges from 1 to the maximum hit-count in the profile by only considering template positions that are tagged by one or more sub-assemblies. Based on that profile, the gene copy number is determined as the hit count assigned to C% of the OG representatives, whereby C is defined as the minimum fraction of the covered template. As already indicated above, the estimated gene copy number is dependent on the assembly parameters. As shown in Figure 3, the more stringent the *mi* parameter was chosen, the higher the estimated gene copy number.

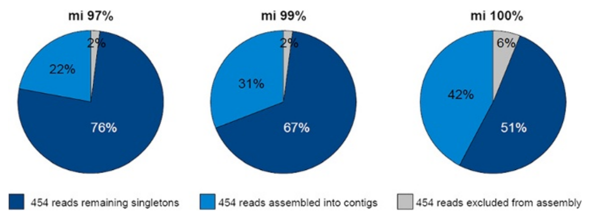

Figure 2 Newbler assembly statistics of 454 reads for different minimum overlap identity parameters. Three orthologous assemblies were performed using 97% minimum alignment identity (*mi*), 99% and 100%. For each assembly, the number of 454 remaining singletons (no significant overlap to any other 454 read), the number of assembled 454 reads and the number of excluded 454 reads (ultra-short reads, outlier reads, repeat reads) were counted.

Estimation of assembly parameters and evaluation of the copy number predictions

As already shown in Figure 2 and Figure 3, the correct selection of the Newbler parameter –*mi*, which specifies the minimum identity in order to combine two reads during the assembly, is a major factor affecting both the assembly outcome and the downstream gene family analysis based on predicted gene copy numbers. To estimate the correct Newbler parameterisation two simulation experiments were applied: (i) simulation of an whole-genome sequencing experiment for a diploid reference genome of similar genome size and repeat and gene composition and (ii) an *in silico* generation of an hexaploid gene set with similar sequence differences as in the homeologous genes of the target species.

Simulation of a whole genome sequencing experiment

Established reference assemblies or complete genome sequences of a reference species with similar genome size and structure can be used for the *in silico* simulation reflecting the experimental setting as achieved for the target genome sequence. Thereby, genomes with different degree of polyploidy could be used for this evaluation by adjusting the expectation of the reference gene family size. For the bread wheat genome analysis, the repeat masked genome sequence of maize (ZMb73, version 5b.60; http://www.maizegdb.org) was used and first, the “real” gene family size (gene count) identified by OrthoMCL clustering of the 39,656 annotated maize proteins and the reference protein data sets. However, because highly similar gene sequences, which could not be discriminated with the applied identity cut-offs during the sequence assembly, would artificially lead to significant underestimation of predicted gene counts, CD-Hit [21] nucleotide similarity

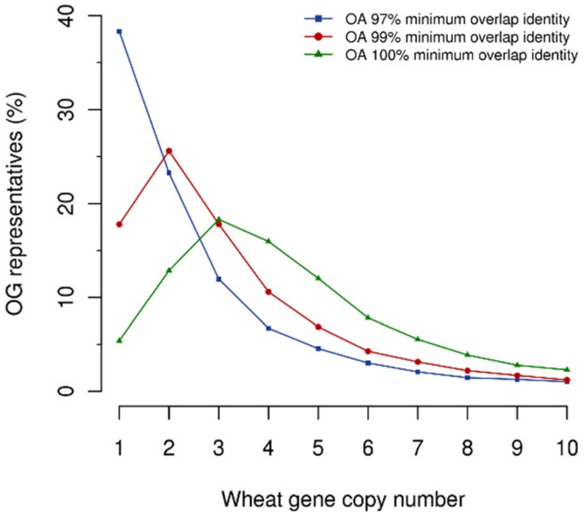

Figure 3 Wheat gene copy number in dependency of the applied assembly minimum alignment parameter.

clustering was performed at different similarity levels (97%, 99% and 100%, respectively) and all but one of the redundant sequences was retained. We selected more than 15,000 OrthoMCL cluster including one or more maize genes as well as exactly one OGR. Thus, counting the number of grouped maize proteins provided reference values for subsequent analysis and comparisons.

Next, the input data set for the orthologous assembly was designed and a collection of raw 454 sequencing reads created from the maize reference genome assembly applying with 5× genome coverage along with a uniform error rate of 0.5% [22]. Then, the above described protocol for the orthologous assembly was carried out with varying minimum overlap identity between 97% mi and 100% *mi*, respectively, and the gene copy number was predicted for the orthologous assembly.

Simulation of a polyploidy gene catalogue

Complementary to the above described evaluation method, a second evaluation was implemented, which in addition to the gene count simulates the polyploid effect size in the whole genome sequencing experiment. For this purpose, the transcript sequences of the 28,236 rice gene models (RAP2 annotation), including coding sequences as well as non-coding 3′ and 5′ UTRs and introns, were aligned against the orthologous gene representatives. Thereby the gene copy number in the rice data set was determined for each OG representative. Then, the aligned rice transcripts were triplicated and random single nucleotide variants (SNVs) were introduced to simulate the sequence similarity of the homeologous genes in the polyploid target genome. For example, for bread wheat nucleotides were randomly exchanged with p = 0.01 (1 nucleotide change per 100 bp) in coding transcript sequences and p = 0.04 (4 nucleotide exchanges per 100 bp) in non-coding transcript sequences, respectively. Then, 454-like shotgun reads were simulated (5× genome coverage), re-mapped against their corresponding OG representatives, sub-assembled with varying minimum overlap identity (97% mi, 99% mi and 100% mi) and, finally, the gene copy number predicted.

As expected, depending on the required stringency for read overlaps, the copy number prediction largely differs. In both evaluation methods 454 reads of different gene copy were collapsed by using 97% *mi*, whereas, requiring perfect alignment overlaps (100% *mi*) clearly results in an overestimation of the gene family sizes. As expected, we observed nearly perfect agreement of the expected 1:1 relationship between the expected and observed gene family size in the maize simulation, as well as the 1:3 relationship in the polyploidy simulation by using 99% alignment identity. This parameterisation allows compensating for sequencing errors, by simultaneously maintaining distinct gene copies that share high sequence similarity in coding regions. However, highly similar gene copies from large gene families may still collapse into single assemblies resulting in slight overall gene number underestimates.

Sub-genome classification of bread wheat transcripts using machine learning

Besides sheer size and high repeat content hexaploidy makes the genome of bread wheat extremely challenging to analyse. Being able to differentiate between homologous genes of the three wheat sub-genomes (*A*, *B* and *D*) is of high importance not only for marker design and breeding but also to address open questions in the evolution and domestication of bread wheat.

Recent NGS sequencing approaches of the bread wheat genome generated a 5-fold 454 survey sequence, without being able to separate A, B and D sub-genome directly [9]. In principle, this can be facilitated applying a chromosome sorting technology [23], however, sorted sequences were only available for wheat linkage group 1 at this time [4].

As a consequence, an alternative approach was established to classify wheat sub-assembly sequences for the A, B and D sub-genomes. Wheat sub-assemblies were generated by a stringent assembly of reads mapped to representative (for orthologous groups defined by OrthoMCL [16]) genes from the reference organisms *Brachypodium distachyon* [13], *Hordeum vulgare, Oryza sativa* [15] *and Sorghum bicolor* [14] as well as the genome sequences of the D genome donor species *Ae. tauschii* [24], and the A genome relative *Triticum monococcum* (NCBI archive SRP004490.3), and cDNA sequence assemblies from *Ae. speltoides* (Trick&Bancroft, unpublished data) a member of the Sitopsis section to which the putative B genome donor belongs. Expecting that A- related sub-assemblies are more related to *T. monococcum* sequences, D- related sub-assemblies to *Ae. tauschii*, and B-related sub-assemblies to *Ae. speltoides,* sequence similarities of the sub-assemblies to each of these datasets would define and discriminate their origin.

In a first step, sequence similarities of each sub-assembly sequence to the wheat progenitor sequences were computed using BLAST [17]. Only sub-assemblies with BLAST hits to all three wheat progenitor sequences were considered for classification. Although a classification into A, B or D sub-genome derived transcripts seemed possible for many sub-assemblies by applying simple similarity cut-offs or rules, fixed similarity cut-offs appeared not suitable to separate the majority of sub-assemblies with confidence [9].

Consequently, several machine learning approaches were applied to the similarity matrix and evaluated for their performance. A major prerequisite represents the identification of a suitable training and test data set. We made use of wheat group 1 chromosome sequences which were separated into their sub-genomes (A, B and D) using flow-sorted chromosomes [4]. We extracted all wheat sub-assemblies associated with wheat group 1 chromosome sequences and classified them into A, B or D depending on their best hit. Every sub-genome classification was then complemented with the sub-assemblies' similarities to *T. monococcum*, *Ae. speltoides* and *Ae. tauschii* sequences to create a training set compatible to all non-chr1 related sequences.

We applied a number of machine learning algorithms (e.g. Logistic Regression, Naive Bayes, Decision Trees and Support Vector Machine algorithms) from the WEKA package [25] (http://www.cs.waikato.ac.nz/ml/weka/) to this training set and evaluated the results by stratified k-fold cross-validation. The best compromise between precision and recall was observed for the Support Vector Machine algorithm (libSVM). This trained libSVM classifier was used to classify the set of genic wheat sub-assemblies into A-, B- or D-related sequences. The results of this classification are summarized in Table 1. Below defined libSVM probability estimate thresholds the classification was considered unreliable (e.g. in cases where a sub-assembly sequence matches equally well to all three progenitor sequence sets).

With the recently published draft genome sequences of *Aegilops tauschii* [26] and *Triticum urartu* [27] two additional WGS datasets from wheat progenitors have since become available. As *T.urartu* is thought to represent the wheat A sub-genome progenitor even better than *T. monococcum* [28], using these datasets in an analysis update could potentially further improve the separation of sub-assemblies generated from the A sub-genome and therefore re-fine the overall result.

An alternative approach was taken in the separation of homeologs in the tetraploid wheat (AABB; "pasta wheat") transcriptome, published by Krasileva et al. [29]. Here, a post-assembly pipeline including polymorphism identification, phasing of SNPs, read sorting and re-assembly of phased reads was used to separate homeologs.

Physical, genetic and functional assembly of the barley genome

Barley is diploid and diverged from wheat approximately 12 million years ago (Mya) [30]. Barley is the first cereal genome where an anchored physical map has been reported on the basis of the entire genome. The physical map was constructed by combining different clone libraries using FPC [31] to assemble clones into contigs.

Table 1 Support Vector Machine classification results on wheat genic sub-assemblies

Classification category	Wheat sub-assembly sequences with intact open reading frame
Used for classification	**462,803 (100%)**
Classified as A	**94,949 (20.5%)**
Classified as D	**113,065 (24.4%)**
Classified as B	**97,923 (21.2%)**
Not classified	**157,166 (33.9%)**

The genome size of 5 Gb made barley the largest genome for which a physical map has been constructed. 570,000 BAC clones were assembled into 9,265 fingerprinted contigs (contigs). Several clone libraries were combined to prevent that genomic regions would remain unrepresented. End-merging of contigs allowed to bridge overlapping contigs together, possible when marker evidences on both ends would indicate a merge. A wealth of different marker resources mostly linked to transcripts [32] but also by using *genotyping by sequencing* (GBS) technology helped to assign a majority of contigs to a genetic position. To increase the robustness of the resulting map, experimental markers were included. Apart from the construction of the physical map three different barley cultivars were sequenced under high coverage resulting in *whole genome shotgun* (WGS) assembly with approximately 2 gigabases (Gb) each. These WGS contigs, apart from forming the basis to derive gene models also helped to extend sequences on the physical map by extending clone end sequences. The WGS contig decorated physical map was then taken to infer a genetic position for physical contigs. Altogether 4 Gb out of the 5 Gb cumulative contig map length was anchored.

The anchoring and ordering of FPC and sequence contigs along the barley genome demonstrates that GBS technology combined with a transcript derived map is powerful to develop a rich and deep physical map even for the complex and large barley genome. Along with the annotation of genes it describes the functional and physical assembly of a cereal genome.

The GenomeZipper approach

The GenomeZipper (GZ) [11,12] is a synteny driven approach to order and structure NGS survey sequences of grass genomes that lack a physical map. The approach can be applied to a variety of different data sets, i.e. 454 reads, contigs, or scaffolds grouped in individual chromosomes or chromosome arms. The approach exploits the widely conserved synteny among grasses [33] and uses corresponding syntenic intervals, as defined by heterologous, corresponding marker intervals among the species under investigation and suitable reference genomes to deduce a tentative ordering of genes in the corresponding regions. This approach can be undertaken for smaller regions of interest, whole chromosome arms and chromosomes on a genome wide level. The outcome is a virtual gene order map which integrates gene-based marker maps as well as conserved syntenic information from at least one sequenced model grass genome and NGS data. Initially applied on the barley genome [11,12], the genome zipper approach has now also been used with other grass species, such as wheat chromosome 4A [34], rye chromosome B [35] and Lolium [36].

The GenomeZipper approach consists of three discrete steps: repeat filtering, detection of syntenic conserved regions and an integration step which provides an anchored, information-rich scaffold (Figure 4).

In the first step, filtering repeat regions out of the query sequence reduces computational effort in gene

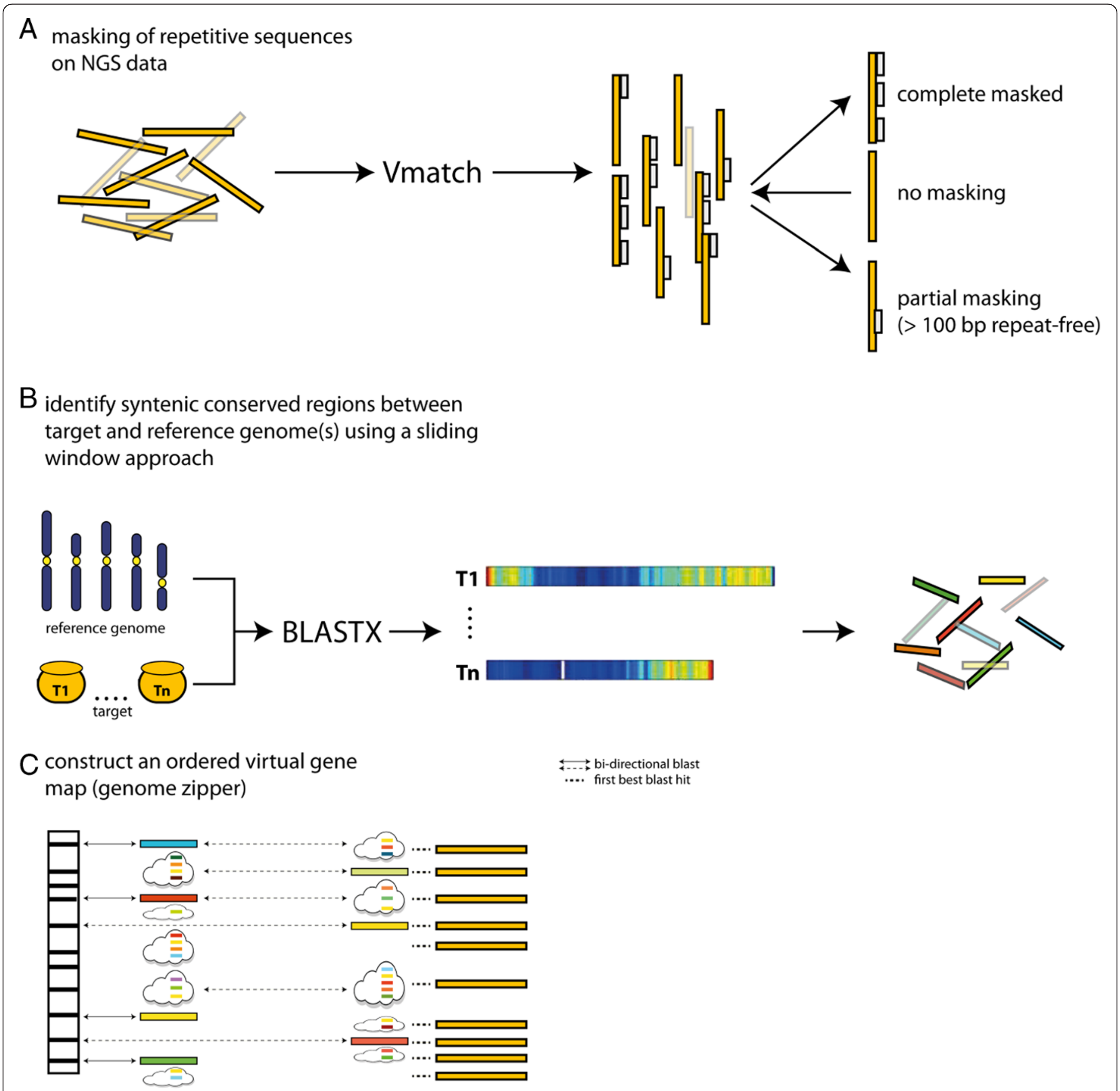
A masking of repetitive sequences on NGS data
Vmatch
complete masked
no masking
partial masking (> 100 bp repeat-free)
B identify syntenic conserved regions between target and reference genome(s) using a sliding window approach
reference genome
T1
Tn
target
BLASTX
T1
Tn
C construct an ordered virtual gene map (genome zipper)
bi-directional blast
first best blast hit

Figure 4 GenomeZipper workflow. The GenomeZipper approach can be divided into three individual steps which can be run independently: repeat masking **(A)**, detection of syntenic conserved blocks **(B)** and the 'zipping'/integration **(C)** of all data sets into a virtual linear gene order model. **A)** The repetitive sequences (grey boxes) were filtered out from the NGS data (orange boxes) using Vmatch and the MIPS REdat Poaceae v9.0 library. Only sequences without repetitive elements or with at least 100 base pairs repeat-free regions were considered for the next analysis steps. **B)** The syntenic conserved regions between target and reference genome(s) were determined using BLASTX and a sliding window approach. The sequences were aligned against the reference genome(s) and the highly conserved genes (coloured boxes) were extracted and used for the construction of the gene map. **C)** The virtual gene map is constructed by integrating the syntenic conserved genes of one or multiple reference genomes and the NGS survey sequences along a backbone build by a genetic marker map of the same organisms as the target or a very closely related one. In a first step all conserved genes with a bi-directional blast hit to the gene-based marker are integrated into the zipper backbone. The remaining conserved genes (depicted in the clouds) are indirectly incorporated using the order deduced from the corresponding reference genome. The NGS data (orange boxes) are anchored to the ordered map by first best blast hits.

space estimation. In barley and wheat the repetitive amount was identified by aligning the 454 reads against the MIPS REdat Poaceae repeat library using Vmatch (http://vmatch.de).

In the next step, the conserved homologs between query sequence and one or multiple reference genomes are determined. For barley, three model grass genomes - *Brachypodium distachyon* [13], rice [15] and *Sorghum bicolor* [14] - were used to identify homologous regions. The sequence comparisons were done using BLASTX and only first best hits with at least 75%/70% sequence identity and a minimal alignment length of 30 amino acids were considered. The syntenic conserved regions are defined by the density of homologous matches between query and reference genome using a sliding window approach.

During the last step, the NGS data is structured and ordered using a high resolution genetic map and orthologous genes obtained in the previous step. Thereby the intervals as defined by genetically ordered markers are used as a scaffold to project the likely order of corresponding genes in these intervals. Gene order in these intervals is deduced from the order found in the respective reference genomes, whereby evolutionary closest reference genomes get highest rank. Once ordered, additional evidences, such as full length cDNAs and/or ESTs can be attached to the ordered gene scaffold.

The linear ordered gene maps provide a valuable resource for a variety of applications: (i) for marker development and to assist positional cloning [37], (ii) for comparative analyses of the conserved gene space [4], and (iii) to resolve the structure of a genome/chromosome and to establish the colinearity between grass genomes[34,35].

Conclusions

Next generation sequencing technologies now start to enable to decipher large plant genomes such as those from many *Triticeae* (wheat, barley, rye) which until recently were difficult to access due to severe technological and economic restrictions. The assembly and analysis of these complex genomes remains a challenge and requires novel concepts and strategies. Here, we outlined and described a number of these concepts, that were developed and used to analyse and order genes from the recently published genome sequences of barley and hexaploid wheat. For barley, a physical and genetic map integration approach allows to positionally anchor ~21,000 genes. A complimentary approach, the GenomeZipper concept, makes use of the conserved gene order between grass reference genomes and many monocot crop genomes to anchor and order genes by an *in silico* approach in the complex wheat and barley genomes.

The hexaploid nature with three highly homologous sub-genomes makes the genome of bread wheat extremely challenging to assemble and analyse. To access the gene inventory, a set of orthologous representative genes was constructed from related and finished reference grass genomes. Wheat NGS reads from a 454 5× whole genome sequence survey were mapped onto these orthologous representatives and separately assembled in a stringent way to avoid collapsing of homologous (sub-genome derived) genes. Genic sub-assembly sequences were subsequently classified into A, B or D- sub-genome derived with a machine-learning assisted approach making use of differing sequence similarities to the A, B and D subgenomes progenitor species.

The approaches and concepts outlined here may be readily applied to other complex genomes, even beyond plants, where direct sequence assembly and analysis is hampered by size and/or polyploidy but related, less complex reference genomes are available.

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

MS wrote the introduction, conclusions, and contributed to the wheat sections. MM contributed to the GenomeZipper section. MP contributed to the wheat sections and TN contributed to the barley section. KM supervised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors want to thank all collaboration partners involved in the individual studies reviewed here. In particular, all members from the International Barley Sequencing Consortium (IBSC), the UK wheat consortium (Neil Hall, Anthony Hall and Mike Bevan et al.) as well as IPK Gatersleben (Nils Stein et al.).

Received: 17 May 2013 Accepted: 10 July 2013
Published: 6 September 2013

References

1. United States: *Department of agriculture. World agricultural supply and demand estimates.* Report No. WASDE-511; 2012.
2. Purugganan MD, Fuller DQ: **The nature of selection during plant domestication.** *Nature* 2009, **457:**843–848.
3. Eilam T, Anikster Y, Millet E, Manisterski J, Sagi-Assif O, Feldman M: **Genome size and genome evolution in diploid Triticeae species.** *Genome* 2007, **50:**1029–1037.
4. Wicker T, Mayer KF, Gundlach H, Martis M, Steuernagel B, Scholz U, Simkova H, Kubalakova M, Choulet F, Taudien S, *et al*: **Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives.** *Plant Cell* 2011, **23:**1706–1718.
5. Nesbitt M, Samuel D: **Hulled wheats.** In *Proc. 1st Internat. Workshop hulled wheats, international plant genetic resources institute.* Edited by Padulosi S, Hammer K, Heller J.; 1996:41–100.
6. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, *et al*: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces.** *Plant Cell* 2010, **22:**1686–1701.
7. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly.** *Nat Methods* 2009, **6:**S6–S12.
8. Moore G, Devos KM, Wang Z, Gale MD: **Cereal genome evolution. Grasses, line up and form a circle.** *Curr Biol* 1995, **5:**737–739.
9. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, *et al*: **Analysis of the bread wheat genome using whole-genome shotgun sequencing.** *Nature* 2012, **491:**705–710.
10. Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, *et al*: **A physical, genetic and functional sequence assembly of the barley genome.** *Nature* 2012, **491:**711–716.

11. Mayer KF, Taudien S, Martis M, Simkova H, Suchankova P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, *et al*: Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* 2009, **151:**496–505.
12. Mayer KF, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, *et al*: Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 2011, **23:**1249–1263.
13. Initiative IB: Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* 2010, **463:**763–768.
14. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al*: The Sorghum bicolor genome and the diversification of grasses. *Nature* 2009, **457:**551–556.
15. Project IRGS: The map-based sequence of the rice genome. *Nature* 2005, **436:**793–800.
16. Li L, Stoeckert CJ Jr, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003, **13:**2178–2189.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, **215:**403–410.
18. Flavell RB, Bennett MD, Smith JB, Smith DB: Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet* 1974, **12:**257–269.
19. Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N: A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J* 2009, **59:**712–722.
20. Miller JR, Koren S, Sutton G: Assembly algorithms for next-generation sequencing data. *Genomics* 2010, **95:**315–327.
21. Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006, **22:**1658–1659.
22. Richter DC, Ott F, Auch AF, Schmid R, Huson DH: MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One* 2008, **3:**e3373.
23. Dolezel J, Vrana J, Safar J, Bartos J, Kubalakova M, Simkova H: Chromosomes in the flow to simplify genome analysis. *Funct Integr Genomics* 2012, **12:**397–416.
24. Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, *et al*: A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. *Proc Natl Acad Sci U S A* 2013, **110:**7940–7945.
25. Frank E, Hall M, Trigg L, Holmes G, Witten IH: Data mining in bioinformatics using Weka. *Bioinformatics* 2004, **20:**2479–2481.
26. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, *et al*: Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 2013, **496:**91–95.
27. Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, *et al*: Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* 2013, **496:**87–90.
28. Dvorak J, Terlizzi P, Zhang HB, Resta P: The evolution of polyploid wheats: identification of the A genome donor species. *Genome* 1993, **36:**21–31.
29. Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, Soria M, Wang S, Consortium I, Akhunov E, *et al*: Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol* 2013, **14:**R66.
30. Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R, Gornicki P: Acc homoeoloci and the evolution of wheat genomes. *Proc Natl Acad Sci U S A* 2008, **105:**9691–9696.
31. Soderlund C, Longden I, Mott R: FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* 1997, **13:**523–535.
32. Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, Shaw P, Bayer M, Thomas W, Marshall D, *et al*: Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet* 2012, **44:**1388–1392.
33. Gale MD, Devos KM: Comparative genetics in the grasses. *Proc Natl Acad Sci U S A* 1998, **95:**1971–1974.
34. Hernandez P, Martis M, Dorado G, Pfeifer M, Galvez S, Schaaf S, Jouve N, Simkova H, Valarik M, Dolezel J, Mayer KF: Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J* 2012, **69:**377–386.
35. Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Simkova H, *et al*: Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci USA* 2012, **109:**13343–13346.
36. Pfeifer M, Martis M, Asp T, Mayer KF, Lubberstedt T, Byrne S, Frei U, Studer B: The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiol* 2013, **161:**571–582.
37. Lupken T, Stein N, Perovic D, Habekuss A, Kramer I, Hahnel U, Steuernagel B, Scholz U, Zhou R, Ariyadasa R, *et al*: Genomics-based high-resolution mapping of the BaMMV/BaYMV resistance gene rym11 in barley (Hordeum vulgare L.). *Theor Appl Genet* 2013, **126:**1201–1212.

doi:10.1186/1746-4811-9-35
Cite this article as: Spannagl *et al.*: Analysing complex *Triticeae* genomes – concepts and strategies. *Plant Methods* 2013 **9**:35.