



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2009

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

Wicker, T ; Taudien, S ; Houben, A ; Keller, B ; Graner, A ; Platzer, M ; Stein, N

DOI: <https://doi.org/10.1111/j.1365-313X.2009.03911.x>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-29488>

Journal Article

Accepted Version

Originally published at:

Wicker, T; Taudien, S; Houben, A; Keller, B; Graner, A; Platzer, M; Stein, N (2009). 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. *The Plant Journal*, 59(5):712-722.

DOI: <https://doi.org/10.1111/j.1365-313X.2009.03911.x>

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

Thomas Wicker¹, Stefan Taudien², Andreas Houben³, Beat Keller¹, Andreas Graner³,
Matthias Platzer², Nils Stein^{3,4}

¹ Institute of Plant Biology, University Zurich, Zollikerstrasse 107, CH-8008 Zurich

² Leibniz Institute for Age Research -Fritz Lipmann Institute (FLI), Beutenbergstr. 11, D-07745 Jena

³ Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, D-06466 Gatersleben

⁴ Corresponding author

Corresponding author's address

Dr. Nils Stein

Group Genome Diversity, Dep.Genebank

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Corrensstr. 3

D-06466 Gatersleben

phone:+49-(0)39482-5522

fax: +49-(0)39482-5595

email: stein@ipk-gatersleben.de

Author's email addresses :

Thomas Wicker wicker@botinst.uzh.ch
Stefan Taudien stau@fli-leibniz.de
Andreas Houben houben@ipk-gatersleben.de
Beat Keller bkeller@botinst.uzh.ch
Andreas Graner graner@ipk-gatersleben.de
Matthias Platzer mplatzer@fli-leibniz.de

Suggested running title: 454 whole-genome snapshot in barley

Keywords: 454 whole-genome snapshot, gene content, repetitive DNA, genome size evolution, conserved non-coding sequence, fluorescence in situ hybridisation.

Word count

Title page	162
Summary	194
Introduction	992
Results	2475
Discussion	2051
Methods	647
Figure Legends	203
Tables	110
Acknowledgements	63
Legends Supplementary material	48
<u>Total Manuscript</u>	6945
References	1287

Summary

The genomes of barley and wheat, two of the world's most important crops, are very large and complex due to their high content of repetitive DNA. In order to obtain a whole-genome sequence sample, we performed two runs of 454 (GS20) sequencing on genomic DNA of barley cv. Morex which yielded an approximate 1% of a haploid genome equivalent. Almost 60% of the sequences were comprised of known transposable element (TE) families while another 9% represented novel repetitive sequences. We also discovered high amounts of low-complexity DNA and non-genic low-copy DNA. We identified almost 2,300 protein coding gene sequences and more than 660 putative conserved non-coding sequences. Comparison of the 454 reads with previously published genomic sequences suggested that TE families are distributed unequally along chromosomes. This was confirmed by *in situ* hybridisations of selected TEs. A comparison of these data on the barley genome with a large sample of publicly available wheat sequences showed that several TE families which are highly abundant in wheat are absent from the barley genome. This finding implies that the TE composition of their genomes differs dramatically, despite their very similar genome size and their close phylogenetic relationship.

Introduction

The tribe of the Triticeae contains some of the world's most important crops such as barley and wheat. Their haploid genomes have a size of approximately 5,700 Mbp and contain at least 80% repetitive DNA (Bennett and Smith, 1976). Despite their large genomes, the gene content of diploid Triticeae genomes is expected to be similar to that of the rice genome which contains 31,429 genes according to the manually curated gene set of the rice annotation project (RAP-DB, rapdb.dna.affrc.go.jp). The repetitive fraction of the Triticeae genomes consists primarily of transposable elements (TEs). Over 200 TE families were discovered in genomic sequences (mostly BACs) during the past few years and deposited in the database for Triticeae repeats (TREP, wheat.pw.usda.gov/ITMI/Repeats). Because of the large size and repetitiveness of their genomes, long genomic Triticeae sequences have been rather rare. At the time this study was done, only 23 genomic barley sequences larger than 20 kb which total slightly more than 4.2 Mbp were deposited at GenBank. The available genomic barley sequences are possibly biased due to sampling because most published sequences originated from map-based cloning projects, specifically targeting regions in the distal parts of chromosomes where most genes are located (Qi *et al.*, 2004).

In contrast to barley, genomic resources for wheat have been growing recently due to an international effort to sequence the entire wheat genome (International Wheat Genome Sequencing Consortium, IWGSC). Currently, there are 359 large genomic sequences from wheat with a cumulative size of more than 44 Mbp publicly available (about 0.6% of a haploid genome equivalent). The sample of wheat sequences is less biased because almost 250 BAC clones had been selected randomly (Devos *et al.*, 2005, Bennetzen, Devos and SanMiguel, personal communication). However, most of these sequences are not annotated and knowledge on TE content and composition comes mostly from hand-annotated sequences that were generated in map-based cloning efforts. Thus, the actual overall composition of the Triticeae genomes was extrapolated from limited datasets (Sabot *et al.*, 2005) and no whole-genome survey has been performed so far. Nevertheless, a recent

study, based on 36 bp Solexa reads and mathematically defined repeats, concluded that at least in the gene-rich regions, knowledge of TE sequences and genome composition in Triticeae is fairly comprehensive (Wicker *et al.*, 2008).

Wheat and barley have diverged approximately 11.6 million years ago (Chalupska *et al.*, 2008) and their close phylogenetic relationship is reflected in strong conservation of gene order and content as well as in their composition of the TE fraction. Several high-copy TE families were discovered in both barley and wheat, suggesting that similar dynamics shaped the genomes of both species. For example, the elements of the *BARE1* clade (*BARE1* in barley and *WIS* and *Angela* in wheat) have been shown to contribute at least 10% to the barley genome (Vicent *et al.*, 1999; Kalendar *et al.*, 2000; Soleimani *et al.*, 2006) and they are expected to be present in similar amounts in the wheat genome.

The only other large plant genome where repetitive DNA has been extensively studied is maize. There, it was shown that the genome was expanded mainly by the amplification of LTR retrotransposons (SanMiguel and Bennetzen, 1998). These high-copy TE families likely were present already in the common ancestor of several maize species because the most abundant TE families were found in all species examined (Takahashi *et al.*, 1999; Meyers *et al.*, 2001). Most high-copy families were found more or less evenly distributed across the genome although some (especially DNA transposons) were found somewhat enriched in distal chromosomal regions (Messing *et al.*, 2004; Bruggmann *et al.*, 2006). A similar pattern was found in the recently published sorghum genome (Paterson *et al.*, 2009).

Previous studies have also shown that the repetitive fractions of the Triticeae and other grass genomes are highly dynamic with genomic DNA permanently being created through TE amplification and removed through deletions. This "genomic turnover" results in a rapid reshuffling of intergenic sequences within a few million years and virtually no TE older than 4 or 5 million have been identified (Wicker *et al.*, 2003; SanMiguel *et al.*, 2002; Swigonová *et al.*, 2005; Wicker and Keller, 2007). Additionally, comparison of genomic sequences from different wheat species indicated that differences in TE activities lead to differential

compositions of TE content of wheat genomes (Sabot *et al.*, 2005) but no genome-wide study has been undertaken so far to compare the precise composition of their repetitive fraction.

The rise of new sequencing technologies such as 454 pyrosequencing (Margulies *et al.*, 2005) has decreased costs of sequencing and allows for a rapid and cost-effective sampling of genomes (for review see Mardis 2008, Shendure and Ji 2008). Besides the reduced costs, 454 sequencing has the advantage that no genomic libraries have to be constructed, which could potentially discriminate against certain sequences. Indeed, sequence coverage with 454 reads was shown to be much more even than with Sanger reads derived from the classic shotgun approach (Wicker *et al.*, 2006). As 454 is a young technology, only few studies have been performed on entire plant genomes. 454 whole-genome shotgun sampling was recently used to characterise the repetitive fraction of the pea and soybean genomes (Macas *et al.*, 2007; Swaminathan *et al.*, 2007). These studies provided a large amount of data on repetitive sequences in two genomes which had been poorly characterised up to that point.

In this study, we took that approach further by comparing the outcome of a whole genome snapshot from barley with a set of well-characterised genomic sequences. This led to conclusions on the relative abundance of TEs in the barley genome and provided a broad insight into the gene space of barley. Based on 454 reads and publicly available sequences we found uneven distribution of certain TE families along chromosomes, a finding which was confirmed by *in situ* hybridisations. Additionally, a comparison with genomic sequences from wheat revealed strong differences in the TE compositions of the two genomes despite their very similar size.

Results

Two runs of GS20 sequencing provide approximately 1% of a haploid barley genome equivalent

The two GS20 runs produced a total of 571,509 reads with an average size of 103 bp. They amount to 58.91 Mb which corresponds to slightly more than 1% of the haploid barley genome (assuming a size of 5,700 Mbp, Bennett and Smith, 1976). We refer to the produced 454 sequence library as a "whole-genome snapshot". To determine the sequence composition of the genome snapshot, the individual short reads were subjected to various analyses, including a series of BLAST searches against different databases.

The first step was the identification of sequence reads that are composed primarily of simple sequence repeats (SSRs) and other low-complexity sequences by identifying sequences with under-represented bases (see Experimental procedures). We identified a total of 11,876 reads (2.08% of the snapshot) that largely consisted of low-complexity motifs (e.g. large microsatellites). Chloroplast and mitochondrial DNA contributed 4.80% and 0.15% to the 454 snapshot, respectively. In addition to the genomic DNA that is contained in the organelles, nuclear genomes frequently contain insertions of fragments of organellar DNA (Leister, 2005). The rice genome contains approximately 0.18 – 0.24% of each chloroplast and mitochondrial DNA insertions (International Rice Genome Sequencing Project, 2005). If barley contains similar amounts of organellar insertions in its nuclear genome, the mitochondrial sequences could indeed represent genomic insertions while the larger numbers of chloroplast sequences are more likely to originate from chloroplasts that were carried over during DNA preparation.

14 transposable element families contribute 50% to the barley genome

To identify known repetitive elements, all individual 454 reads were used in BLASTN searches against the TREP database. The 342,476 reads (59.92% of the total snapshot) which produced significant hits with E-values below $10E-6$ were sorted according to the TE family they belong to. In total, 248 previously known TE families are represented in the genome snapshot. The contribution of each TE family to the total snapshot was calculated to estimate their overall abundance in the barley genome (Table 1). 50.31% of the sequences

were contributed by only 14 TE families (Table 1) while the residual 234 families contributed another 9.61% (not shown). Twelve of the top 14 TE families are LTR retrotransposons, 10 of which belong to the *Gypsy* and two to the *Copia* superfamily. As expected, the *Copia* element *BARE1* tops the list, contributing more than 12% to the total genome and thereby confirming previous estimates (Vicient *et al.*, 1999; Kalendar *et al.*, 2000; Soleimani *et al.*, 2006). Second and third most abundant are the two *Gypsy* families *Sabrina* and *WHAM*, which contribute almost 8.5% and 5.5% to the total sequences, respectively. Only two DNA (Class 2) transposon families were found in the top 14 TE families (*Balduin* and *Caspar*). Both belong to the *CACTA* superfamily and they rank 7th and 8th, contributing 2.32% and 2.1% to the total snapshot, respectively.

As TREP and other repeat databases do not contain all of the probably thousands of TE families in the barley genome, new families (i.e. families that do not share enough homology at the DNA level), were identified by BLASTX against PTREP, the protein division of TREP. This approach has the limitation that only coding regions of TEs can be identified and non-coding parts or regions encoding highly divergent proteins can not be detected. The BLASTX against PTREP yielded only 0.77% of the sequences. An additional 0.15% of TE protein encoding sequences were identified during the search for genes (see below), bringing the total TE content of the 454 snapshot to 60.77%. The low number of newly identified TE families indicates that the TREP database had contained most of the abundant Triticeae TE families (Table 1).

The 454 snapshot contains coding sequences of almost 2,000 genes and many putative conserved non-coding sequences

Potential gene coding sequences were identified by BLASTX of the 454 reads against a database containing all 66,710 predicted rice proteins from the TIGR rice genome (version 5) and the dataset of the rice annotation project database (RAP-DB) which contains 30,192 proteins. The TIGR dataset includes at least 20,000 TE-related sequences while the TE

content in the RAP-DB set should be minimal. We used both datasets because it would allow to identify putative genes as well as additional TE related sequences. The BLAST searches against TIGR identified 2,445 while the searches against RAP-DB identified 2,399 potential coding sequences. Interestingly, only 1,777 reads were identified by both BLAST searches. All 668 reads that were identified only by the TIGR dataset were considered TE related and added to the repetitive fraction.

A cross-blast of the RAP-DB reads against the TIGR database identified an additional 252 reads that were designated as retrotransposon or transposon in the TIGR dataset. These were also added to the repetitive fraction, leaving the final gene count at 2,147

To discover additional putative genic sequences, those 454 reads which had not been characterised as repeats or genes by now were used in BLASTN searches against the entire rice genome. The search yielded 845 sequences which produced significant alignments (>50 bp, E-value <10E-6). Because all 454 reads with homology to rice CDS or proteins had been filtered out, these sequences represented potential conserved non-coding sequences (CNS) or non-annotated genes. Interestingly, the vast majority (770) of these reads were low-copy with only 1-5 copies in the rice genome and another 50 reads had 6-20 copies. Only 25 had more than 20 copies and were added to the repetitive dataset.

To test if the 820 candidate CNS with 20 or less copies could still encode proteins, we screened the alignments of barley and rice sequences for base mismatches that are separated from each other by multiples of 3. In protein coding sequences the third position of the codon is more likely to differ, due to the degeneration of the genetic code. We found that 158 sequence alignments were significantly enriched at $p < 0.05$ for such mismatch spacings. These sequences were considered protein encoding and added to the total gene count, bringing it to 2,305 (Figure 1). The other 687 reads were classified as putative CNS (Figure 1). Additionally, all 687 putative CNS were used in BLASTN searches against a collection of 560 small nucleolar, small nuclear and noncoding RNA sequences from plants.

Only four of them produced significant BLASTN hits. A total of 41 of the CNS are conserved in Arabidopsis.

The barley genome contains large amounts of low-or moderately repetitive DNA

At this point, 391,092 of the 571,509 reads had been classified as either SSR, repeat, gene or CNS. The leftover 180,417 sequences (31.6%) which have not been classified at this point were used for an assembly (see methods). This resulted in 51,972 reads being assembled into 16,040 contigs. We considered all these 51,972 reads as repetitive, because the 1% genome coverage which the 454 snapshot provides renders it very unlikely that many low-copy sequences would be covered multiple times. The leftover 128,445 reads which remained singletons were considered low- or single copy. A total of 15,727 contigs were shorter than 200 bp, 246 were 200-300 bp in size, while only 67 were longer than 300 bp. Five of these large contigs were derived from ribosomal DNA (rDNA).

Most of the large contigs had a relatively high coverage (i.e. they were assembled from a large number of reads). The average coverage of the largest 100 contigs ranged from 10 to 180-fold. A 180-fold coverage translates into an approximate copy number of 18,000 (because of the 1% coverage of the genome with 454 reads). The 16,040 contigs mostly represent sequences that we could not characterise any further. Only 52 of them showed homology to TE sequences. However, the mere fact that the 51,970 reads could be assembled into contigs indicated that they are repetitive. Thus, they were added to the set of repetitive sequences, bringing the total repeat content of the snapshot to 69.15% (Figure 1).

As the assembly of the 454 reads was relatively stringent, we wanted to investigate whether some of the resulting contigs may actually belong to the same repeat family (i.e. that they share an 80% sequence identity). We found 607 contigs with significant homology to others (see methods). These 607 contigs could be clustered into 117 supercontigs with an average size of 295 bp. The longest contig was 1083 bp. Only one of them showed homology to a *Gypsy* retrotransposon.

Some repeats are unequally distributed across the barley genome

For an earlier study, we had selected a set of nine large genomic sequences from barley, totalling more than 2 Mbp, and manually annotated them carefully to identify as many TE sequences as possible (Wicker *et al.*, 2008). These nine sequences (hereafter be referred to as "reference set") originate from distal chromosomal (i.e. gene-rich) regions. Because 454 provides a very even sequencing coverage (Wicker *et al.*, 2006), we wanted to test whether we could identify TE sequences that are unevenly distributed along chromosome arms.

Thus, we compared the relative abundance of the top 14 TE families in the 454 reads with their occurrence in the reference set. As expected, most dominant in both the 454 snapshot and the reference set were well-known high-copy TE families such as *BARE1*, *Sabrina* and *WHAM* (Figure 2), although *Sabrina* and *WHAM* were somewhat under-represented in the reference set. A surprising exception is *BAGY2* which is the 4th most abundant TE family in the 454 snapshot, but was found only once in the reference set, corresponding to an almost 10-fold under-representation (Figure 2). The family *WHAM* is also about two-fold under-represented in the reference set. In contrast, *Caspar* and *Jeli* were more than three and two-fold over-represented, respectively.

To verify these results, we derived probes for fluorescence *in situ* hybridisations (FISH) from *BAGY2*, *Caspar* and *Jeli*, as these elements were either under- or over-represented in the reference set. *Jeli*-specific signals were found scattered throughout the chromosomes, indicating a sporadic but more or less even distribution (Figure 3a). In contrast, *Caspar* specific signals were highly enriched in the sub-telomeric regions while *BAGY2* signals were found throughout all chromosomes, except in sub-telomeric regions (Figure 3b and c). In addition to the previously described TE sequences, five probes derived from large contigs of the assembly (see above) were also used for *in situ* hybridisations. These five were chosen because of their high copy number and because they could not be classified as any known repeat type. Four of the five probes (10010, 100156, 10027 and 06796) resulted in relatively

uniform labelling of the chromosomes (Supplementary Figure 1). The signals of probe 100156 were less abundant in subtelomeric regions. Probe 00112 produced the most intense signals in the centromeric regions of one chromosome pair (Figure 3d). Additional hybridisation with an SSR marker identified that centromeric signal to be specific for chromosome 3H (not shown).

Diploid wheat has a very different genome composition than barley

Analogous to the barley reference set of genomic sequences, we prepared a set of manually-annotated genomic sequences from diploid wheat *Triticum monococcum* (Wicker *et al.*, 2008). Its repetitive fraction was compared to the data from the barley 454 genome snapshot. Because barley and wheat diverged only approximately 11.6 MYA (Chalupska *et al.*, 2008) they contain some very similar TE families. For example, the TE families *WHAM*, *Sabrina* and *BARE1* from barley are 75-80% identical at the DNA level with their wheat counterparts. It should be noted here that mainly for historical reasons the wheat homologs of *BARE1* are called *WIS* and *Angela*, although their strong sequence homology would justify classification of all three into just one family. Thus, *BARE1*, *WIS* and *Angela* were treated as the same family for this study (*BARE1*). As in barley, *BARE1* is the most abundant TE family, contributing 16.71% to the *T. monococcum* reference sequences (Figure 2). Second and third most abundant are *Sabrina* and *WHAM* which occur at similar frequencies in the *T. monococcum* reference set as in the barley sequences. In total, 11 TE families contribute 50% to the *T. monococcum* sequences (not shown). Interestingly, that list contains five TE families (*Fatima*, *Jorge*, *Daniela*, *Erika* and *Claudia*) which were practically absent from the 454 barley snapshot as they only produced 51 (*Jorge*) to 354 (*Daniela*) significant BLASTN hits in the barley 454 reads. *Fatima*, *Daniela*, and *Erika* are LTR retrotransposons of the *Gypsy* Superfamily while *Claudia* is a *Copia* element. *Jorge* is a class 2 transposon of the *CACTA* superfamily.

It is possible that the absence of BLAST hits for these five TE families is simply caused by sequence divergence. However, if one assumes that these five TE families diverged at a similar pace as *WHAM*, *Sabrina* and *BARE1*, one can expect that elements from barley should be 75-80% identical to those from wheat and therefore they should be unambiguously detected. To measure the influence of sequence divergence on the number of sequences hit, we used *BARE1*, *Sabrina* and *WHAM* sequences from both barley and wheat in BLASTN searches against the 454 barley snapshot. These three TE families are abundant in both species. For all three families, the elements from barley produced approximately twice as many hits to the 454 snapshot than those from wheat. For example, a *Sabrina* element from barley showed significant homology to 28,048 of the 454 sequences while the one from wheat hit only about half as many (13,798). Taking into account this reduction of identification sensitivity, the number of *Fatima*, *Jorge*, *Daniela*, *Erika* and *Claudia* type TE's in the barley snapshot should be about the double of the BLASTN hit numbers (~ 100 and ~700, respectively). This, however, is still a negligible amount.

There was also the possibility that the reference sequence set from *T. monococcum* is biased and contains a high number of *Fatima*, *Jorge*, *Daniela*, *Erika* and *Claudia* elements by pure chance. Thus, we broadened the sample by determining their contribution to all publicly available large genomic sequences (i.e. larger than 20 kb) from wheat. The majority of the 359 sequences from wheat (*Triticum*) species which were available at NCBI are randomly picked BAC clones and should therefore sample the genome relatively evenly (Bennetzen, Devos and SanMiguel, personal communication). The overall contribution to the complete sequence set and to the reference set was estimated by measuring the cumulative size of all BLASTN alignments of the five TE families in both sequence sets. This is a very rough approach and not comparable in quality to hand-annotation. For example, divergent parts of TEs are almost certainly not recognised while in hand-annotation they can be easily identified. Nevertheless, as shown in Supplementary Table 1, the five TE families contributed similar amounts to both the reference set and the entire public dataset of wheat sequences.

Thus, the high frequency of the five TE families in the reference set is not merely coincidental since they are present at similar frequencies in the mostly randomly selected wheat BACs.

Discussion

Our analysis of 454 reads that cover about 1% of a haploid barley genome equivalent provided a new and unbiased view of the composition of Triticeae genomes. The repetitive fraction of Triticeae genomes is probably better characterised than for any other plant genome. We have used these resources for comparison and integration of biological information on the nature of the Triticeae genomes and to evaluate the current knowledge on their genome composition. Our analysis therefore provided additional elements to similar 454 based whole-genome surveys as they were performed for pea and soybean (Macas *et al.*, 2007; Swaminathan *et al.*, 2007). It was also the first study that took this approach to a grass genome and demonstrated that by massively parallelized DNA sequencing, repeat composition of large and highly repetitive genomes can be investigated at a high-resolution. The data provided a quantitative assessment of all major TE families in the barley genome and we gained information on differential distribution of TE families along chromosomes. Additionally, we unravelled strong differences in the composition of TE families between barley and wheat.

The 454 snapshot sampled many genes and conserved non-coding sequences

The 454 sampling of the barley genome provided 2,305 reads that contained protein coding sequences of genes. The high number of reads produced by 454 sequencing provided an excellent sample depth, as we were able to sample 7.4% of all barley genes (assuming a total gene number of 31,000 genes) while the actual number of bases sequenced amount to only about 1% of a haploid genome equivalent. Nevertheless, the actual number of protein coding sequences in our 454 snapshot might be even higher than 2,305. This assumption is based on the calculation of the sampling depth: the average size of coding sequences in the

rice RAP-DB dataset is 1,723 bp. Therefore, the 31,000 rice genes correspond to about 54 Mbp of actual coding sequence. If barley has the same number and average size of genes, only about 0.95% of its 5,700 Mbp genome would actually be protein coding sequences. Thus, we would expect 5,430 protein coding gene sequences (0.95%) instead of 2,305. This lower than expected gene content can be due to several reasons: First, although rice is the closest relative to barley for which the genome is sequenced, the two plants have diverged approximately 50 million years ago (Paterson *et al.*, 2004). Fast-evolving genes may have diverged to a degree that hampers their homology recognition by BLAST with relatively short 454 reads. Second, many coding sequences might also have been missed because they were only partially covered by the reads and thus did not produce significant BLAST hits.

It is intriguing that we identified more than 660 sequences which do not seem to have coding capacity but still remained conserved during the roughly 50 million years since barley and rice diverged (Paterson *et al.*, 2004). Interestingly, most of these CNS have very low copy numbers, indicating that they do not simply represent conserved parts of transposable elements. Forty-one of them were also conserved in Arabidopsis. This indicates that most CNS diverge probably more quickly than gene sequences. Nevertheless, their conservation in different plant species implies a biological function. Additionally, a few studies have shown that CNS can be found in colinear positions in different species (Bossolini *et al.*, 2007; Pourkheirandish *et al.*, 2007), further supporting the possibility of a conserved function.

The repetitive landscape of the barley genome is extremely diverse and not homogeneous

The 454 snapshot provided very precise estimates for the abundance of different TE families in the barley genome. The results clearly showed that very few TE families have reached such enormous copy numbers that they completely dominate the repetitive fraction while dozens of other, previously described TE families contribute only a few percent to the total genome. As expected, the "usual suspects" *BARE1*, *Sabrina* and *WHAM* clearly topped the

list. All three have long been known to contribute large portions to the barley genome but actual copy number estimates were only available for *BARE1* prior to this study (Vicient *et al.*, 1999; Kalendar *et al.*, 2000; Soleimani *et al.*, 2006), while the abundance of all other TE families so far had to be inferred from very limited datasets (Sabot *et al.*, 2005). Interestingly, two Class 2 transposons (*Caspar* and *Balduin*) were found in the top 10 most abundant elements, contributing more than 4.5% to the total genomic sequences studied. This in contrast to the situation in maize where all known *CACTA* families combined make only about 1% of the genome (Messing *et al.*, 2004).

The known repetitive elements accounted for about 60% of the 454 reads, raising the question about the nature of the other roughly 30% which are not genes, CNS, rDNA or organellar DNA. The fact that we could assemble over 16,000 contigs from these uncharacterised reads indicated that there may be thousands of TE families populating the barley genome at low or moderate copy numbers. Different TE families often share no detectable sequence homology at the DNA level, and can only be detected based on their coding sequences (e.g. by BLASTX). But TEs usually contain considerable stretches of non-coding DNA (e.g. LTRs or TIRs) which can not be detected by homology search. We assume that many of these 16,000 sequence contigs actually represented non-coding portions of TEs.

More than 22% of the 454 reads remained completely uncharacterised. One must expect that many of them are still repetitive, at a 1% genome coverage, sequences with less than 100 copies are very likely to remain singletons. Thus, the situation in barley might not be fundamentally different from that in smaller genomes such as rice. There, dozens of LTR retrotransposon families have been described but only relatively few reached very high copy numbers. Most are present in low or moderate copy numbers (Vitte *et al.*, 2007). The difference between large and small plant genomes might therefore be merely quantitative as only a few high-copy TE families actually determine genome size.

However, a recent study indicated that a relatively large percentage of the barley genome might also be simply non-coding low-copy DNA (Wicker *et al.*, 2008). With the current data, one can also not exclude that some of the uncharacterised reads are simply sequencing artefacts. More extensive studies with larger datasets will be needed to answer this question.

Some TE families are distributed unevenly across the genome

A surprising finding was that two TE families (*BAGY2* and *WHAM*) were clearly under-represented while two others (*Caspar* and *Jeli*) were over-represented in the reference sequence set. This indicated that the previously published sequences are not representative for the composition of the entire barley genome and that TE composition varies along the chromosomes. Previous studies in maize found that most high-copy TE families are more or less randomly distributed across the maize genome (Meyers *et al.*, 2001, Bruggmann *et al.*, 2006). Some TE families are distributed unevenly, especially DNA transposons were found to be enriched in gene-rich regions (Bureau and Wessler, 1994, Bruggmann *et al.*, 2006). Also in the recently published genome of Sorghum, retrotransposons have the tendency to cluster in proximal regions while DNA transposons are enriched in distal chromosomal regions (Paterson *et al.*, 2009).

In this respect, our results showed that the situation in barley is similar to that in other grass genomes. However, our targeted approach of comparing whole-genome 454 reads with previously published sequences allowed an efficient identification of TEs with an unusual distribution, despite the lack of a complete genome sequence. We could therefore specifically design probes for *in situ* hybridisations which allowed us to demonstrate that some repetitive sequences indeed are very unevenly distributed along chromosomes. Our data showed that major high-copy TE families like *Caspar* tend to prefer distal (i.e. gene-rich) regions. The inverse situation was found with *BAGY2* and probe 00112 which were distributed throughout all chromosomes except the sub-telomeric and pericentromeric regions. Interestingly, probe 00112 appeared to be even chromosome specific. Although we

tested only a relatively small number of repetitive sequences *in situ*, three of them showed very distinct chromosomal distribution. These results suggest that there may be a large number of repetitive sequences with insertion site preferences. Thus, complex models for the interaction of TEs with their host genome may be necessary to explain our observations.

Parallel evolution of dynamic genome size equilibrium in barley and wheat

The common perception is that barley and wheat have a very similar composition of their repetitive fraction because they have very similar diploid genome sizes and diverged only about 11 MYA (Chalupska *et al.*, 2008). Indeed, we found that some high-copy TE families such as *Sabrina*, *WHAM* and the *BARE1/Angela/WIS* clade contribute similar fractions to the genomes of barley and wheat. This is reminiscent of the finding that the same major high-copy retrotransposon families are found in all maize species studied (Takahashi *et al.*, 1999, Meyers *et al.*, 2002).

A highly interesting and novel aspect of our study is that several TE families that are highly abundant in the available genomic sequences of wheat but virtually absent in barley. Apparently, several TE families (e.g. *Fatima*, *Jorge*, *Daniela*, *Erika* and *Claudia*) have become very successful in wheat during the past roughly 11 million years while they have gone practically extinct in barley. We believe that we were able to observe this phenomenon because of the high resolution of the 454 genome snapshot that allowed us to look beyond the main high-copy TE families, and we speculate that a similar situation may also be found when different maize species were compared using 454 sequencing.

Curiously, however, the differential rise and decline of TE families had basically no effect on the overall diploid genome sizes of barley and wheat. Thus, the overall rates of TE amplification and removal have stayed in balance, even if the composition of active TE families differs strongly. It was shown that the balance of TE amplification and DNA deletion determines the size of a genome (Devos *et al.*, 2002; Wicker *et al.*, 2003; Pereira, 2004; Wicker and Keller, 2007), giving rise to the “increase-decrease” model of genome size

evolution (Vitte and Panaud, 2005). Bursts of TE activity can expand genomes tremendously (SanMiguel and Bennetzen, 1998; Swigonová, et al., 2005; Piegu et al., 2006) but genome expansion can be also counteracted by removal of DNA through unequal crossing-over (Shirasu *et al.*, 2000) or illegitimate recombination (Devos *et al.*, 2002; Wicker *et al.*, 2003).

Our findings add the novel aspect to the increase-decrease model that genome size can remain in a dynamic equilibrium for several million years (~11 in the case of barley and wheat) while the TE composition of the genome might undergo drastic changes. This suggests the existence of regulatory forces that keep genome size constant over long periods of time. For example, one can speculate that the physical size of chromosomes has to stay within a relatively narrow range as to allow “every day” pairing and recombination between individuals of a population. Thus, intergenic regions could be turned over rapidly as described in the increase-decrease model as long as overall genome size does not change too much. Drastic changes in genome size (as is, for example, observed between different grass) species might therefore be restricted to actual species formation events. Future studies targeting the genomes of Triticeae and other grass species for whole-genome snapshots will be necessary to further elucidate this phenomenon.

Conclusion

The presented data showed the potential of new high-throughput sequencing technology to survey entire complex genomes and to allow broad insight into many different aspects of genome evolution and organisation. Using larger sample sizes and upgraded sequence technologies (e.g. Roche/454 GSFLX Titanium which produces reads of about 400 bp), more detailed quantitative and qualitative analyses will be possible in the near future. The data also demonstrated that the gene space even in very large and complex genomes can be effectively sampled with a modest financial effort. Improved sequencing technology and decreasing costs will allow us to expand this kind of study to several different Triticeae species. The larger amount of data will provide great level of detail in assessing of gene

content, number of CNS and overall repeat makeup of the different genomes. Such studies will make it possible to develop detailed models of the mechanisms and dynamics of genome evolution on Triticeae and other plants.

Experimental procedures

454 sequencing

Preparation and sequencing of the 454 sequencing library was essentially performed according to manufacturer's instructions (GS20 Shotgun DNA Library Preparation Kit / Sequencing Kit, Roche Diagnostics, March 2007). About 4 µg of barley cv. Morex whole genome DNA were fragmented for 1 min at 3 bar N₂ by nebulizers to an average fragment length of 700 bp (determined on DNA-1000 LabChip, Agilent). The library was quantified by RNA-1000 LabChip (Agilent) and the Quant-IT RiboGreen RNA assay (Invitrogen). The reads analysed in this study were generated on a GS20 by two runs (two segments each) with 70x75 picotiterplates and 4 out of 8 segments ("titration") of a 40x75 picotiterplate.

Sequence analysis

All analyses were performed on LINUX systems. Stand alone BLAST was obtained from NCBI (www.ncbi.nlm.nih.gov/). For classification of 454 reads, BLAST databases were created, including the TIGR rice genome version 5 (rice.plantbiology.msu.edu), RAP-DB (rapdb.dna.affrc.go.jp), TREP (wheat.pw.usda.gov/ITMI/Repeats/) and organellar DNA from plants and animals. BLAST searches of individual 454 reads were performed with the help of a custom Perl script which immediately evaluated the BLAST report and deleted it again, to save disc space. Sequences with hits were collected in one file, those without hits in another. The "no hit" file is then used for BLAST searches against other databases. Throughout this study, we considered BLAST hits with E-values smaller than 10E-6 as significant, except for the CNS search where we used the more stringent criterion of hits >50 bp.

Because low-complexity sequences (e.g. microsatellites) are inherently non-random patterns, they were identified with a Perl program that finds sequences with under-represented bases. Under-represented was defined by the base frequency of less than 0.07 in a stretch of 100 bp. In a random sequence, the chance for a particular base (A, C, T, G) occurring 6 or less times is less than 1 in a million.

BLASTN searches for known TE families were performed against the complete TREP database (totalTREP) which was chosen because it contains several copies of known high-copy elements instead of consensus sequences or just single representatives like in the non-redundant TREP database (nrTREP). The reason for this was that different members of TE families show a certain degree of sequence variation. The TE family definition from Wicker et al. (2007a) was used.

Assembly of 454 sequences

The assembly of the subset of 454 reads which could not be classified as either genic or repetitive sequences was done with MIRA V2.9.25 (Chevreux et al., 1999, http://chevreux.org/projects_mira.html) using default parameters for 454 data. Assemblies were done with sequence text files only without using quality data ("notraceinfo" option).

Contigs which belong to the same repeat family (i.e. that they share an 80% sequence identity) but were not similar enough to be assembled with MIRA were identified in an all vs. all BLASTN search. Contigs with more than five significant hits to other contigs were selected for clustering with CLUSTALW at default settings. The aligned multiple sequence file was screened with an original Perl script for groups of sequences that share regions of more than 50 bp and 80% sequence identity. These groups were then used to construct consensus sequences. One additional round of all vs. all comparison with DOTTER was done to identify further overlapping supercontigs which were then combined by hand into single supercontigs.

Probe preparation and fluorescence in situ hybridization (FISH)

Primer pairs (Supplementary Table 2) were designed based on assembled 454-contigs and used to amplify the corresponding genomic fragments. PCR products were cloned and sequenced. Clones with the highest similarity to the original 454-derived sequences were selected for FISH and labeled with digoxigenin-11-dUTP by nick-translation. Preparation of mitotic chromosomes and subsequent in situ hybridization was performed as previously described (Houben *et al.*, 2006). Hybridization was detected after incubation with conjugated fluorescence anti-DIG antibodies. The fluorescent signals were recorded by cooled sensitive charged-couple device (CCD) camera and pseudo-colored using Adobe Photoshop software.

Sequence deposition

The complete set of 454 sequences has been deposited at GenBank (accession numbers XXXXX – XXXXXX). The dataset can also be obtained from the authors via FTP upon request.

Acknowledgements

This work was supported by grant 0314000A (Barlex) of the Federal Ministry of Education and Research of Germany (BMBF) to NS and by grants 3100A0-122242/1 to TW and 3100-105620 to BK of the Swiss National Science Foundation. TW, BK and NS are cooperation partners in the EU COST action FA0604 ‘Tritigen’. We thank K. Kumke and O. Weiß for excellent technical support.

References

Bennett, M.D. and Smith, J.B. (1976) Nuclear DNA amounts in angiosperms. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **274**, 227–274.

- Bossolini, E., Wicker, T., Knobel, P.A. and Keller, B.** (2007) Comparison of orthologous loci from small grass genomes *Brachypodium* and rice, implications for wheat genomics and grass genome annotation. *Plant J.*, **49**, 704–717.
- Bruggmann, R., Bharti, A.K., Gundlach, H., Lai, J., Young, S., Pontaroli, A.C., Wei, F., Haberer, G., Fuks, G., Du, C., Raymond, C., Estep, M.C., Liu, R., Bennetzen, J.L., Chan, A.P., Rabinowicz, P.D., Quackenbush, J., Barbazuk, W.B., Wing, R.A., Birren, B., Nusbaum, C., Rounsley, S., Mayer, K.F. and Messing, J.** (2006) Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* **16**, 1241-1251.
- Bureau, T.E. and Wessler, S.R.** (1994) Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. *Proc. Natl. Acad. Sci. USA*, **91**, 1411–1415.
- Chalupska, D., Lee, H.Y., Faris, J.D., Evrard, A., Chalhoub, B., Haselkorn, R. and Gornicki, P.** (2008) *Acc* homoeoloci and the evolution of wheat genomes. *Proc Natl Acad Sci USA*, **105**, 9691–9696.
- Chevreux, B., Wetter, T. and Suhai, S.** (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: *Proceedings of the German Conference on Bioinformatics (GCB)*, **99**, 45-56.
- Devos, K.M., Brown, J.K.M. and Bennetzen, J.L.** (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.*, **12**, 1075–1079.
- Devos, K.M., Ma, J., Pontaroli, A.C., Pratt, L.H. and Bennetzen, J.L.** (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc. Natl. Acad. Sci. USA*, **102**, 19243–19248.
- Houben, A., Orford, S.J. and Timmis, J.N.** (2006) In situ hybridization to plant tissues and chromosomes. *Methods. Mol. Biol.*, **326**, 203–218.
- International Rice Genome Sequencing Project.** (2005) The map-based sequence of the rice genome. *Nature*, **436**:793-800.
- Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H.** (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci. USA*, **97**, 6603–6607.
- Leister, D.** (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet.*, **21**, 655–663.
- Macas, J., Neumann, P. and Navratilova, A.** (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome, comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics.*, **8**, 427.
- Mardis, E.R.** (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133-141.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bembien, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al.** (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

- Messing, J., Bharti, A.K., Karlowski, W.M., Gundlach, H., Kim, H.R., Yu, Y., Wei, F., Fuks, G., Soderlund, C.A., Mayer, K.F. and Wing, R.A.** (2004) Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA.*, **101**, 14349-14354.
- Meyers, B.C., Tingey, S.V., and Morgante, M.** (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.*, **11**, 1660-1676
- Paterson, A.H., Bowers, J.E., Chapman, B.A., Peterson, D.G., Rong, J. and Wicker, T.** (2004) Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. *Curr. Opin. Biotechnol.*, **15**, 120–125.
- Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberler, 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, IV., Lyons, E., Maher, CA., Martis, M., Narechania, A., Ollilar, R.P., Penning, BW., Salamov, AA., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob-ur-Rahman., Ware, D., Westhoff, P., Mayer, K.F., Messing, J. and Rokhsar, D.S.** (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**:551-556.
- Pereira, V.** (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biol.*, **5**, R79.
- Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A. and Panaud, O.** (2006) Doubling genome size without polyploidization, dynamics of retrotransposition-driven genomic expansions. *Genome Res.*, **16**, 1262-1269.
- Pourkheirandish, M., Wicker, T., Stein, N., Fujimura, T., Komatsuda, T.** (2007) Analysis of the barley chromosome 2 region containing the six-rowed spike gene *Vrs1* reveals a breakdown of rice-barley micro collinearity by a transposition. *Theor. Appl. Genet.*, **114**, 1357–1365.
- Qi, L.L., Echalié, B., Chao, S., Lazo, G.R., Butler, G.E., Anderson, O.D., Akhunov, E.D., Dvorak, J., Linkiewicz, A.M., Ratnasiri, A., et al.** (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics*, **168**, 701–712.
- Sabot, F., Guyot, R., Wicker, T., Chantret, N., Laubin, B., Chalhou, B., Leroy, P., Sourdille, P. and Bernard, M.** (2005) Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Mol. Genet. Genomics*, **274**, 119–130.
- SanMiguel, P., and Bennetzen, J.L.** (1998) Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons *Ann. Bot.*, **82 (Supplement 1)** , 37-44.
- SanMiguel, P.J., Ramakrishna, W., Bennetzen, J.L., Busso, C.S. and Dubcovsky, J.** (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). *Funct. Integr. Genomics*, **2**, 70–80.
- Shendure, J. and Ji, H.** (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135-1145.

- Shirasu, K., Schulman, A.H., Lahaye, T. and Schulze-Lefert, P.** (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.*, **10**, 908–915.
- Soleimani, V.D., Baum, B.R. and Johnson, D.A.** (2006) Quantification of the retrotransposon *BARE-1* reveals the dynamic nature of the barley genome. *Genome*, **49**, 389–396.
- Swaminathan, K., Varala, K. and Hudson, M.E.** (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*, **8**, 132.
- Swigonová, Z., Bennetzen, J.L., Messing, J.** (2005) Structure and evolution of the *r/b* chromosomal regions in rice, maize and sorghum. *Genetics*, **169**, 891–906.
- Takahashi, C., Marshall, J.A., Bennett, M.D., Leitch, I.J.** (1999) Genomic relationships between maize and its wild relatives. *Genome*, **42**:1201–1207.
- Vicient, C.M., Kalendar, R., Anamthawat-Jonsson, K. and Schulman, A.H.** (1999) Structure, functionality, and evolution of the *BARE-1* retrotransposon of barley. *Genetica*, **107**, 53–63.
- Vitte, C. and Panaud, O.** (2005) LTR retrotransposons and flowering plant genome size, emergence of the increase/decrease model. *Cytogenet Genome Res.*, **110**, 91–107.
- Vitte, C., Panaud, O., Quesneville, H.** (2007) LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, **8**, 218.
- Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.D., Dubcovsky, J. and Keller, B.** (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell*, **15**, 1186–1197.
- Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B. and Stein, N.** (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, **7**, 275.
- Wicker, T. and Keller, B.** (2007) Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.*, **17**, 1072–1081.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al.** (2007a) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
- Wicker, T., Narechania, A., Sabot, F., Stein, J., Thi Ha Vu, G., Graner, A., Ware, D., Stein, N.** (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518.

Short legends for supporting information

Supplementary Figure 1. Metaphase cells of barley cv. Morex after fluorescence *in situ* hybridisations with selected repetitive sequences.

Supplementary Table 1. TEs families in public and reference sequence set from wheat.

Supplementary Table 2. Primers and accession numbers for *in situ* hybridisation probes.

Table 1. Families of TEs and their contribution to 454 reads TE family Superfamily 454 reads

<u>TE family</u>	<u>Superfamily</u>	<u>454 reads</u>	<u>Fraction[%]</u>
<i>BARE1</i>	<i>Copia</i>	72549	12.69
<i>Sabrina</i>	<i>Gypsy</i>	48303	8.45
<i>WHAM</i>	<i>Gypsy</i>	31476	5.50
<i>BAGY2</i>	<i>Gypsy</i>	29439	5.15
<i>Sukkula</i>	<i>Gypsy</i>	15327	3.59
<i>Maximus</i>	<i>Copia</i>	14457	2.52
<i>Balduin</i>	<i>CACTA</i>	13284	2.32
<i>Caspar</i>	<i>CACTA</i>	12054	2.10
<i>Jeli</i>	<i>Gypsy</i>	11685	2.04
<i>Laura</i>	<i>Gypsy</i>	11319	1.98
<i>Haight</i>	<i>Gypsy</i>	6163	1.07
<i>Inga</i>	<i>Copia</i>	5880	1.02
<i>Vagabond</i>	<i>Gypsy</i>	5687	0.99
<u><i>Sabine</i></u>	<u><i>Gypsy</i></u>	<u>5119</u>	<u>0.89</u>
Subtotal	-	287982	50.31
Other known families ^a	-	54494	9.54
<u>Novel TE families^b</u>	-	<u>4420</u>	<u>0.82</u>
Total	-	342476	60.77

^aMatch at the DNA level in TREP

^bNo match at DNA level

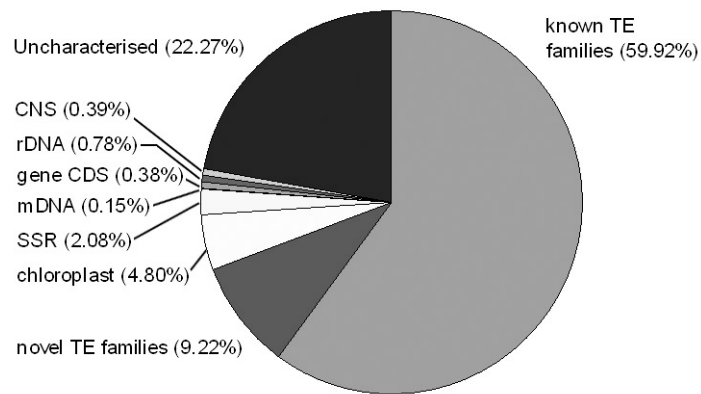


Figure 1. Composition of 571,509 sequences of a 454 barley whole-genome snapshot. The largest fraction (known TE) includes only TEs which have significant sequence identity at the DNA level with known TEs. The novel repeats include sequences which have been classified as repetitive by various analyses (see text). CNS: conserved non-coding sequence, CDS: coding sequence, SSR: simple sequence repeat, rDNA: ribosomal DNA, mDNA: mitochondrial DNA.

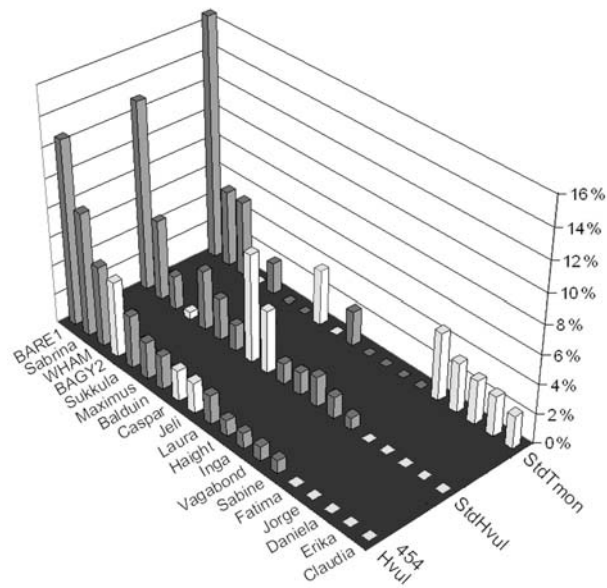


Figure 2. Abundance of TE families in barley 454 snapshot sequences and sets of manually-annotated large genomic sequences (Reference sets) from barley (ReHvul) and diploid wheat *T. monococcum* (ReTmon). The different TE families are shown on the x-axis. The y-axis shows their contribution to the individual sequence sets in percent. TE families which are over or under-represented in some datasets are highlighted in white.

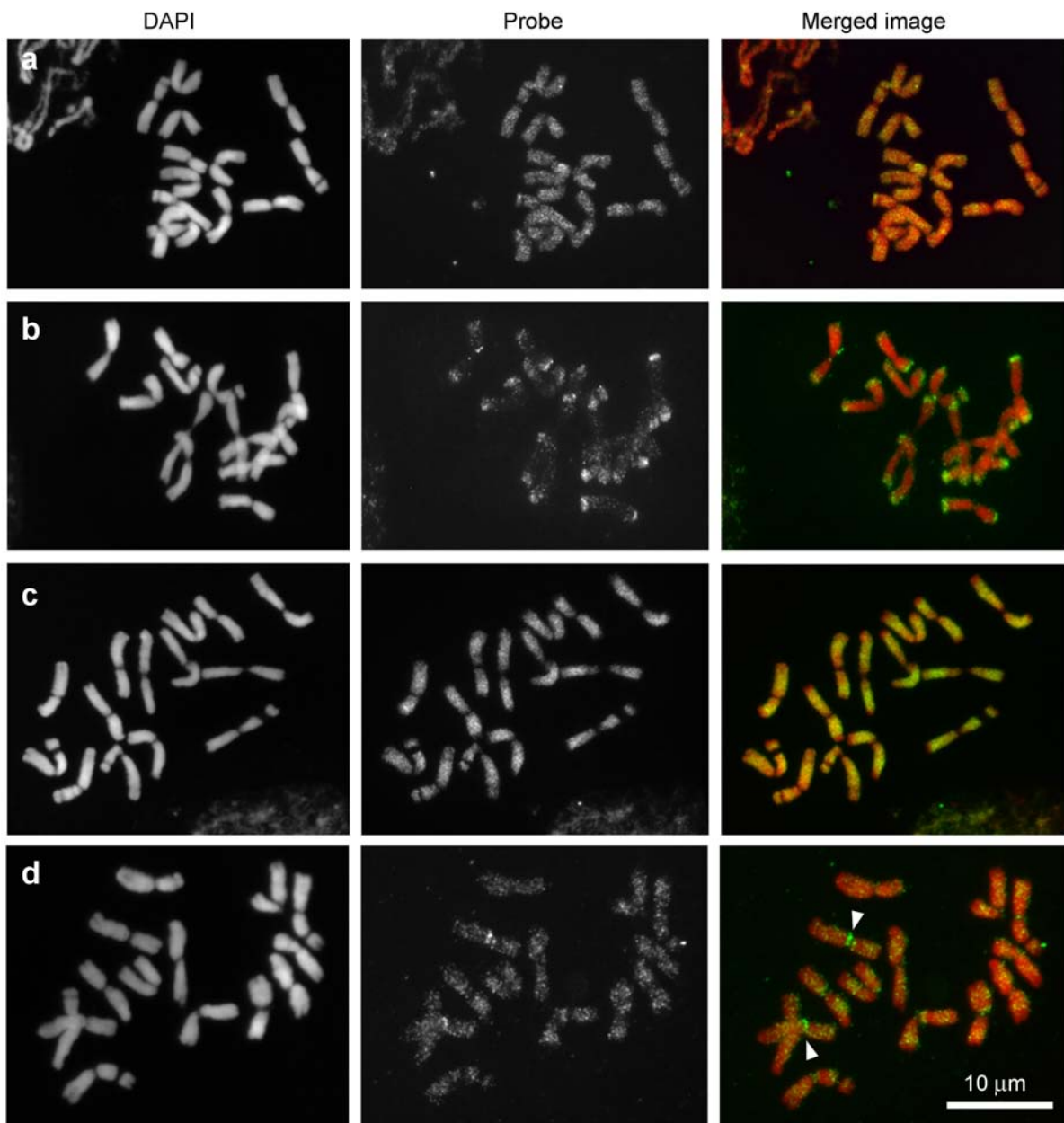
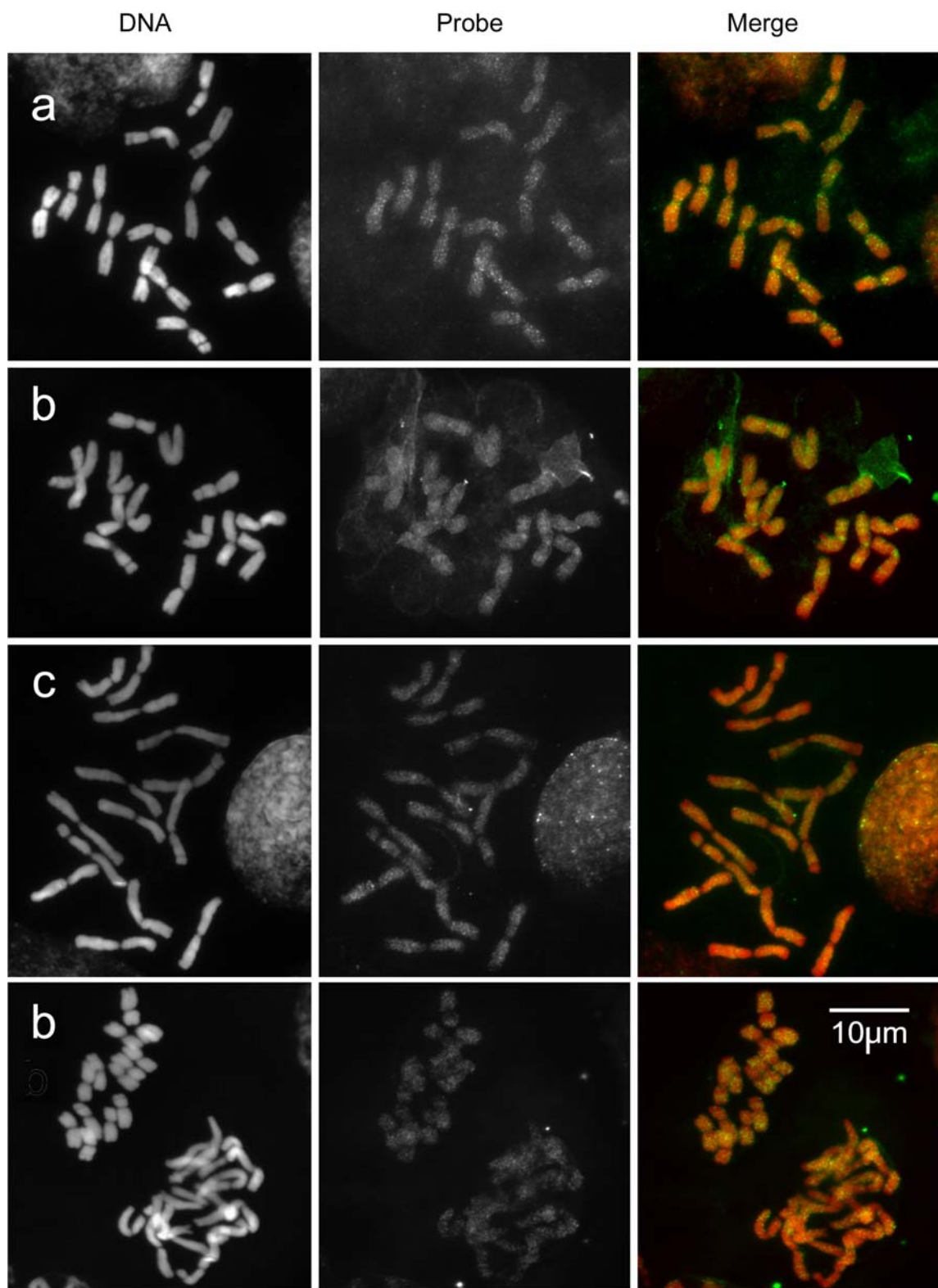


Figure 3. Metaphase cells of barley cv. Morex after fluorescence *in situ* hybridisations with selected repetitive sequences. **a.** Hybridisation with the *Jeli* LTR sequence produces a scattered signal across all chromosomes. **b.** *Caspar* transposons are highly enriched in the subtelomeric regions while **(c.)** *BAGY2* specific signals are distributed throughout all chromosomes except the sub-telomeric and pericentromeric regions. **d.** The strongest hybridisation signal of probe 00112 is located specifically on the chromosome 3H centromere (arrowheads).



Supplementary Figure 1. Metaphase cells of barley cv. Morex after fluorescence *in situ* hybridisations with selected repetitive sequences **a.** 10010, **b.** 10156, **c.** 10027 and **d.** 06796. All repeats probed located throughout all chromosomes.

Supplementary Table 1. Families of TEs and their contribution to the entire public dataset of large *Triticum* sequences and to a set of nine hand-annotated sequences from *T. monococcum*.

TE family	all public sequence [%]	reference set [%]
<i>Claudia</i>	1.05	1.42
<i>Daniela</i>	1.18	1.48
<i>Erika</i>	3.41	5.54
<i>Fatima</i>	3.30	2.18
<i>Jorge</i>	1.69	1.01

Supplementary Table 2. Primers and accession numbers for in situ hybridisation probes.

Probe	5'primer	3'primer	acc. No.
00112	TGCATTTCCACCAGATAG	CACCATGCATTTGAAGAA	FJ175361
10156	ACCGATACAAACACAAAGCACACC	GCGCCACGTCTACCCCCTATT	FJ175362
10027	GTTTGGGGTTTGGTTATCCTT	CTGTCATGCTTAGCTCTTAGTT	FJ175363
10010	TTCAAATAAATTCACACAAA	ACCCCTCCAAAACCATTC	FJ175364
Caspar	CCCCCTCTCCCCTCTTCCTT	CCCCTCTTCTCCGTGCTTGGTC	FJ175365
Jeli	ACCGGCCGTTGTGAAGG	GGGGCCGACTGCTGGAGAAGAA	FJ175366
BAGY2-2	GAGGGGAAATCGTCGCCATCGTC	CCCTCTCGCAAGCATCCGCAACTA	FJ175367
06796	TGTGGTGTGCAAGTTCAAG	AGTGCGGGATCCCTAAAGAC	FJ175368