Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools

Wicker, T ; Krattinger, S G ; Lagudah, E ; Komatsuda, T ; Pourkheirandish, M ; Matsumoto, T ; Cloutier, S ; Reiser, L ; Kanamori, H ; Sato, K ; Perovic, D ; Stein, N ; Keller, B

DOI: https://doi.org/10.1104/pp.108.129734

Running head: Intraspecies diversity in Triticeae

Corresponding author

Beat Keller

Institute of Plant Biology

University of Zurich

Zollikerstrasse 107

8008 Zurich Switzerland

Phone: +41 44 634 82 30

Fax: +41 44 634 82 04

email: bkeller@botinst.uzh.ch

Journal research category: Genetics, Genomics and Molecular Evolution

Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight in the structure of diploid and hexaploid Triticeae gene pools

Thomas Wicker[1,2], Simon Krattinger[1,2], Evans S. Lagudah[3], Takao Komatsuda[4] , Mohammad Pourkheirandish[4] , Takashi Matsumoto[4] , Sylvie Cloutier[5], Laurenz Reiser[2], Hiroyuki Kanamori[6], Kazuhiro Sato[7], Dragan Perovic[8,9], Nils Stein[8], Beat Keller[2,10]

[1]Both authors contributed equally to this study

[2]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[3]CSIRO Plant Industry, GPO Box 1600, Canberra, ACT, 2601, Australia

[4]National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan

[5]Cereal Research Centre, Agriculture and Agri-Food Canada, 195 Dafoe Road, Winnipeg, MB, Canada R3T 2M9

[6]Institute of Society for Techno-Innovation of Agriculture, Forestry and Fisheries (STAFF), Tsukuba 305-0854, Japan

[7]Research Institute for Bioresources, Okayama University, Japan

[8]Leibniz Institute of Plant Genetics and Crop Plant Research, D-06466 Gatersleben, Germany

[9]Present address: Julius-Kuehn-Institute (JKI), Institute for Resistance Research and Stress Tolerance, Erwin-Baur-Str. 27, D-06484 Quedlinburg, Germany

[10]Corresponding author

Financial sources:

This work was supported by the Swiss National Science Foundation grant 3100-105620 to BK, by grants from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation TRC1004 and Green Technology Project GD3006), by GRDC grant CSP00063 and 00099, by BMBF 0312280A (GABI-MAP) and by PROBRAIN to KS.

Present address of Dragan Perovic:

Julius-Kuehn-Institute (JKI), Institute for Resistance Research and Stress Tolerance, Erwin-Baur-Str. 27, D-06484 Quedlinburg, Germany

Corresponding author: Beat Keller (bkeller@botinst.uzh.ch)

Abstract

A large number of wheat and barley varieties have evolved in agricultural ecosystems since domestication. Because of the large, repetitive genomes of these Triticeae crops, sequence information is limited and molecular differences between modern varieties are poorly understood. To study intraspecies genomic diversity, we compared large genomic sequences at the *Lr34* locus of the wheat varieties *Chinese Spring*, *Renan*, *Glenlea* and diploid wheat *Aegilops tauschii*. Additionally, we compared the barley loci *Vrs1* and *Rym4* of the varieties *Morex*, *Cebada Capa* and *Haruna Nijo*. Molecular dating showed that the wheat D genome haplotypes diverged only a few thousand years ago while some barley and *Ae. tauschii* haplotypes diverged more than 500,000 years ago. This suggests gene flow from wild barley relatives after domestication, whereas this was rare or absent in the D genome of hexaploid wheat. In some segments, the compared haplotypes were very similar to each other but for two varieties each at the *Rym4* and *Lr34* loci, sequence conservation showed a breakpoint which separates a highly conserved from a less conserved segment. We interpret this as recombination breakpoints of two ancient haplotypes, indicating that the Triticeae genomes are a heterogeneous and variable mosaic of haplotype fragments. Analysis of insertions and deletions (InDels) showed that large events caused by transposable element insertions, illegitimate recombination or unequal crossing-over were relatively rare. Most InDels were small and caused by template slippage in short homopolymers of only a few bp in size. Such frequent polymorphisms could be exploited for future molecular marker development.

Introduction

The Triticeae tribe contains some of the world's most important crops, among them wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*). Wheat and barley diverged approximately 11.6 million years ago (MYA) and have further diversified since then into several subspecies (Chalupska et al., 2008). Modern bread wheat has a hexaploid genome with A, B and D genomes resulting from hybridisation. The first hybridisation event combined the genomes of the wild wheat species *T. urartu* (A genome) and a probably extinct close relative of *Aegilops speltoides* (B genome) into the tetraploid *T. turgidum ssp. dicoccoides* (Feldmann, 2001). Much more recently, approximately 10,000 years ago, a second polyploidisation event with the diploid *Ae. tauschii* gave rise to hexaploid wheat *T. aestivum* (Feldmann, 2001). This second hybridisation occurred in the early stages of human agriculture. A recent study revealed lower levels of polymorphism in the D genome than in the A and B genomes, indicating that later gene flow from tetraploid to hexaploid species was frequent while it was very limited between the diploid *Ae. tauschii* and hexaploid wheat (Chao et al., 2008).

In contrast to wheat, barley is a diploid species and there is evidence for extensive gene flow from wild to cultivated forms (Pickering and Johnston, 2005). Since domestication, human selection has produced thousands of wheat and barley varieties adapted to many different environments. Modern breeding lines have complex pedigrees but there is little or no molecular information about intraspecific diversity at the genome level.

Intraspecific variation at the haplotype level has been studied to varying degrees in different plant species. The most comprehensive data are available for Arabidopsis where oligonucleotide re-sequencing arrays allowed comparison of multiple ecotypes at a whole-genome level (Clark et al., 2007; Zeller et al., 2008). These studies showed that nucleotide substitutions are irregularly distributed across the genome and that about 4% of genomic sequences were absent in some ecotypes relative to the reference genome. Additionally, regions with almost no sequence diversity were interpreted as results of recent selective sweeps (Clark et al., 2007). In rice, genome-wide comparisons of the subspecies

indica and *japonica* revealed an overall colinearity of genes but reported also numerous polymorphic transposable element (TE) insertions and even differences in gene order and content (Han and Xue, 2003). Variability has been studied in maize on a smaller scale, where dramatic differences between lines have been observed (Fu and Dooner, 2002; Brunner et al., 2005). Both studies showed that not only did the lines differ in the majority of TE insertions, but also in their gene content. Some of the differences in genic sequences were shown to result from transposable element activity (Lai et al., 2005; Morgante et al., 2005).

The high repeat content of more than 80% (Bennett and Smith, 1976) and the large diploid genome size of approximately 5,700 Mbp have so far effectively prevented large-scale genomic sequencing of wheat and barley so that the amount of publicly available genomic sequences is quite limited. At the time this study was done, 416 genomic Triticeae sequences larger than 50 kb were publicly available. Most of them are unfinished BAC sequences which are not annotated and where sequencing is still in progress. One study comparing the *Rph7* locus of the two barley cultivars *Morex* and *Cebada Capa* found numerous TE insertion polymorphisms and concluded that the two loci have diverged roughly within the past 1 million years (Scherrer et al., 2005). Similarly, for the leaf rust resistance locus *Lr10*, two haplotypes exist which differ strongly in the presence/absence of genes and which have diverged at least 1 MYA (Isidore et al., 2005a). Additionally, a PCR screen of 26 wheat lines revealed a relatively low average SNP density of 1 SNP every 335 bp. However, they were very irregularly distributed among the tested loci (Ravel et al., 2006). In summary, these data indicate that the Triticeae gene pool is genetically diverse and contains haplotypes that may be older than the actual species.

The studies cited above were focused on genes in the A genome which, as part of tetraploid wheat, has been part of the polyploid wheat gene pool for a long time. The *Lr34* locus, which is described in this study, is located on the D genome. This genome is the most recent addition to hexaploid wheat and shows therefore the lowest degree of polymorphism. Two recent studies showed a surprisingly high level of polymorphism between the D genome of hexaploid wheat and the diploid wheat *Ae. tauschii*, suggesting that the *Ae. tauschii* line that was used to

produce the respective BAC library has diverged at least 1 MYA from the donor of the D genome (Chantret et al., 2005; Gu et al., 2006). However, so far, no sequence data was available that would allow a comparison of D genome sequences from within the hexaploid wheat gene pool.

Divergence times of sequences from different species (or varieties) can be estimated based on the number of nucleotide substitutions in intergenic regions. Triticeae are especially suited for such analyses because of their high transposable element (TE) content. TEs and other intergenic regions are believed to be largely free from selection pressure (Petrov, 2001) and therefore accumulate mutations at a basic rate which was estimated to be 1.3 E-8 per nucleotide site per year (Ma and Bennetzen, 2004). If a particular TE has inserted in the ancestor of two species (or varieties), their divergence time can be estimated by the number of nucleotide substitutions that have accumulated in that TE. The same principle can also be applied to estimate the age (i.e. insertion time) of Long Terminal Repeat (LTR) retrotransposons (SanMiguel et al., 1998) because their LTRs are identical at the time of insertion and accumulate mutations to a degree that is proportional to their age.

When sequences diverge over time, they do not only accumulate nucleotide substitutions but also insertions and deletions. In Triticeae, the most prominent source of insertions are TEs, which can change size and sequence organisation of a genomic region dramatically within a relatively short evolutionary time (SanMiguel et al., 2002; Wicker et al., 2003). Major DNA losses are caused by unequal homologous crossing-over (Shirasu et al., 2000; Vitte and Panaud, 2003) and illegitimate recombination (Devos et al., 2002; Wicker et al., 2003). While illegitimate recombination can occur between dispersed homologous sequences of only a few (usually 1-10) base pairs, unequal crossing-over events require two highly similar stretches of at least dozens, usually hundreds, of base pairs in size. Unequal crossing-over can occur between LTRs of retrotransposons (Shirasu et al., 2000; Vitte and Panaud, 2003) but also between any other kind of tandem repeats (Wicker et al., 2007). It can also increase the size of a genomic region through expansion of existing tandem repeat arrays or the creation of tandem LTR retrotransposons (Wicker et al., 2007). Small InDels in low complexity sequences

such as homopolymers or simple sequence repeats (SSRs) are usually caused by template slippage during DNA replication (Ramakrishna, et al., 2002, Lovett, 2004). Although template slippage can occur also between DNA fragments longer than a few bp (Lovett, 2004), it is generally assumed that larger events are caused by unequal crossing-over (Shirasu et al., 2000; Vitte and Panaud, 2003; Ma et al., 2007; Wicker et al., 2007).

Here, we present a comparative analysis of the levels of intraspecific variation in two barley loci *Vrs1* and *Rym4* and the wheat locus *Lr34*. The *Vrs1* locus regulates spike morphology in barley, whereas the *Rym4* and *Lr34* are involved in disease resistance. For the *Rym4* and *Lr34* loci, we had sequences from three varieties and for *Vrs1* from two. Additionally, we analysed the *Lr34* locus in the diploid *Ae. tauschii*. At two of the studied loci, we found ancient haplotype fragments in the otherwise highly conserved sequences. We identified several major sequence differences which were caused by illegitimate recombination and unequal crossing-over. Additionally, we found that template slippage in short homopolymers is the major source of small InDels in closely related sequences.

Results

For our comparative analyses of intraspecific genome variability, we used large genomic sequences from the two barley loci *Vrs1* (Komatsuda et al., 2007) and *Rym4* (Stein et al., 2005; Wicker et al., 2005). For the former, sequences from the varieties *Morex* and *Haruna Nijo* were used while for the latter sequences from three varieties (*Morex*, *Haruna Nijo* and *Cebada Capa*) were studied. These three barley varieties have very different breeding histories and pedigrees with probably very little overlap in modern breeding history. *Morex* is an elite breeding line mostly derived from the North American gene pool (http://genbank.vurv.cz/barley/pedigree/krizeni3.asp?id='1488'). Cultivar *Cebada Capa* is an old breeding line from Argentina from around the year 1920 (Grando and Macpherson, 2005), whereas *Haruna Nijo* was released from Sapporo Brewery Ltd., Japan in 1981 (Saisho et al., 2007). In wheat, we studied sequences from the three varieties *Chinese Spring*, *Renan* and *Glenlea* as well as from the diploid donor of the D genome, *Ae. tauschii* (accession *AL8/78*).

The three hexaploid wheat lines analysed originate from very different gene pools: *Chinese Spring* is an old landrace from China whereas *Glenlea* is a Canadian line with a very complex pedigree based on mostly CIMMYT and North American material (http://genbank.vurv.cz/wheat/pedigree/krizeni3.asp?id='19332'). Finally, the cultivar *Renan* is a French winter wheat representing elite wheat material with a pedigree mostly based on European lines (http://genbank.vurv.cz/wheat/pedigree/krizeni1_n.asp?oper=Like&name=Renan&acc=name). The sequences from wheat and from the barley varieties *Haruna Nijo* and *Cebada Capa* are made public here for the first time. The two barley loci contain two putative genes each while the wheat *Lr34* locus contains eight (Table I).

The barley varieties *Morex* and *Haruna Nijo* are very similar in genomic organisation at the *Vrs1* locus but show hundreds of sequence differences

The recently published sequence of the *Vrs1* locus from the barley variety *Morex* (Komatsuda et al., 2007) was compared with a completely overlapping 147 kb BAC sequence of the same locus from the barley variety *Haruna Nijo*. The two sequences are overall very similar to each other since only 459 nucleotide substitutions were detected (i.e. approximately 3.1 nucleotide substitutions per 1 kb). The SNPs are distributed somewhat irregularly across the sequence, however, at such low numbers even a purely random model distribution shows great fluctuations (Figure 1a). The fact that more than 78% of the compared region is derived from TEs allowed to precisely estimate the divergence time of the two loci by using the method of SanMiguel et al. (1998). For this estimate, we excluded the *Hox-1* gene and a conserved non-coding sequence (*CNS-1,* identified through comparison with rice, see methods) plus 1 kb of upstream and downstream region of both. The remaining 139,760 bp contained 446 substitutions (264 transitions and 182 transversions) which translates into an estimated divergence time of 123,020 years (+-5829 years), thus much earlier than domestication.

The two sequences contain a total of 61 insertions and deletions (InDels), 34 of them are deletions in the *Morex* sequence and 27 are deletions in the *Haruna Nijo* sequence. The vast

majority (54) of them are InDels of 10 bp or less, while large InDels are rare. 35 of them are single bp InDels. To get clues as to the molecular mechanism that caused the InDels, we analysed the surrounding sequences of all 61 InDels (Table II). We found that template slippage is with 49 cases the most frequent cause for InDels in the sequence analysed. We considered template slippage as the molecular mechanism whenever the inserted/deleted sequence was found duplicated or multiplied in its entirety in the immediately neighbouring sequences. For example, an InDel motif GGA is embedded in a GGA microsatellite sequence (Figure 1b). The largest InDel we attributed to template slippage is 14 bp in size. The 14 bp motif is one unit of a tandem array of 5 and 4 units in *Haruna Nijo* and *Morex*, respectively (Figure 1b). For five InDels ranging in size from 6 to 390 bp, we could identify short direct repeats flanking the breakpoints of the InDels and indicating illegitimate recombination as the molecular mechanism (Figure 1c). The most striking differences in size between the two sequences were apparently caused by unequal crossing-over. In one large event, the internal domain of an *RLC_BARE1* element was eliminated in the *Morex* sequence (see below). The second event resulted in the presence/absence of two 70 bp units in a large tandem repeat array (Figure 1d). For four InDels, it is not clear by which mechanism they were caused.

Comparison of three barley varieties at the *Rym4* locus reveals a breakpoint between two ancient haplotypes

The three genomic sequences that cover the barley *Rym4* locus from the three varieties *Morex*, *Haruna Nijo* and *Cebada Capa* have a common overlap of approximately 87 kb. This region is highly repetitive and contains only two genes (*EIF4E-1* and *MCT-1*, Table I) which are located tail-to-tail immediately adjacent to each other and are surrounded by a large number of transposable elements (Figure 1e). *Haruna Nijo* and *Cebada Capa* are very similar to each other in the *Rym4* region as they show only 70 nucleotide substitutions and 28 InDels in a total of 92,305 aligned bases. Twenty-one InDels can be attributed to template slippage in simple sequence repeats. For six, the mechanism is unclear and in one case, an unequal crossing over event led to the expansion of a direct repeat array by two units in *Haruna Nijo* (Table II).

Excluding the region containing the two genes plus 1 kb of up-and downstream region, a total of 81,760 bp could be aligned. This fraction contains 69 nucleotide substitutions, corresponding to an estimated divergence time of 32,478 years (+-3,525).

The comparison of *Morex* and *Haruna Nijo* yielded much more complex results, indicating that the *Rym4* locus of *Morex* is the most divergent of the three. At the level of sequence organisation, *Morex* and *Haruna Nijo* differ in numerous insertions and deletions (Figure 1e). The Morex sequence contains 8 TEs which are absent in *Haruna Nijo*, while the latter contains two specific TEs (Figure 1e). Three of the additional TEs in *Morex* are nested in other TEs (Figure 1e). There are seven full-length retrotransposons for which their insertion times could be estimated. Only one of them is common to both sequences and has inserted approximately 1.87 MYA. The other six, which are all only present in the *Morex* sequence, have inserted between 0.87 and 0.05 MYA (Figure 1e). This data would imply that the two loci diverged sometime between 1.87 and 0.87 MYA, the time span between the youngest retrotransposon common to both and the oldest one that appears in only one variety.

We studied the distribution of nucleotide substitutions between the loci by excising *in silico* from both sequences those regions (TEs and other InDels) which are present in only one of the two. This resulted in a hypothetical ancestor sequence that contains only those regions which are found in both varieties (Figure 1e). Interestingly, SNP frequency in the left part of the sequence is more than 5 times higher than in the right part. The first approximately 25 kb of the sequence have an average of 23.7 SNPs per 1000 bp. This value drops abruptly to 4.1 SNPs per 1000 bp between position 25,000 and 26,000 and remains at this low level for the rest of the sequence (Figure 1e).

We interpret this drop in SNP frequency as the breakpoint of two ancient haplotypes that were recombined in *Morex* but not in *Haruna Nijo*. When the TE fractions of the strongly conserved and the less conserved parts were used separately for molecular dating, their estimated divergence times were 159,621 (+-10,335) and 927,306 (+-39,229) years ago, respectively. Based on these numbers we developed a model of the evolution of the two varieties (Figure 1f). We propose that two different barley lineages have diverged approximately 930,000 years

ago, giving raise to two different haplotypes 1 and 2. This divergence time estimate is based on the number of nucleotide substitutions found in the 25 kb at the left end of the hypothetical ancestor sequence. The lineages giving rise to *Haruna Nijo* and part of the *Morex* sequence diverged approximately 160,000 years ago (based on nucleotide substitutions on the right part of the sequence). The lineage leading to *Morex* later recombined with haplotype 2, resulting in a chimerical sequence that includes regions from both haplotypes. The breakpoint of the recombination event lies 3 to 5 kb upstream of the *HvEIF4E-1* gene (Figure 1e). The results of the comparison of *Morex* with *Cebada Capa* were basically identical because *Cebada Capa* and *Haruna Nijo* are much more similar to each other at the *Rym4* locus than to *Morex* (not shown).

Wheat varieties *Chinese Spring*, *Renan* and *Glenlea* differ in their degree of conservation at the *Lr34* locus

The largest wheat sequence we analysed has a size of 207 kb, derived from the *Lr34* locus of the hexaploid wheat variety *Chinese Spring*. The sequence contains eight genes, including one conserved non-coding sequence (CNS) and one pseudogene (*CYP-1*, Table I) which was disrupted by the insertion of a DNA transposon of the *Harbinger* superfamily (Figure 2). This sequence was compared to two sequences from the varieties *Glenlea* and *Renan*. The *Glenlea* sequence has a size of 147 kb and covers all of the right part of the *Chinese Spring* sequence but does not cover the first approximately 60 kb at the left end (Figure 2). The *Chinese Spring* and *Glenlea* sequences are extremely similar to each other. Neither contains a major insertion or deletion that would clearly distinguish it from the other and we detected one single nucleotide substitution which was confirmed by re-sequencing in 146,245 aligned bases. When all gene sequences (plus 1 kb of up and downstream regions) were removed from the alignment, 107,182 bp were left for a divergence time estimate. Using this figure, the two sequences were estimated to have diverged roughly within the past 700 years (359 years +- 359). In contrast to the low number of SNPs, a total of 30 InDels were found. Twenty-eight of them are due to template slippage, one to unequal crossing-over and one has no clear cause.

Renan is with 91 kb the shortest sequence that was available. It is completely covered by the *Chinese Spring* sequence but overlaps in only 42 kb with the sequence of *Glenlea* (Figure 2a). In total, we could align 81,862 bp between *Chinese Spring* and *Renan* and found 13 nucleotide substitutions. The *Renan* sequence contains four putative genes (total gene space of 27,2 kb, including 1 kb of up-and downstream sequences). Nine nucleotide substitutions were found in the 54,6 kb of non-genic sequence, translating into a divergence time estimate of 6,339 years (+-2,113 years).

We detected a total of 26 InDels between *Renan* and *Chinese Spring* and 6 InDels between *Renan* and *Glenlea*. All except two can be attributed to template slippage. The two other InDels are the product of illegitimate recombination. The main difference is a large 9.8 kb deletion affecting parts of an *RLC_Angela* retrotransposon plus part of its flanking region. The deletion is present in *Glenlea* and *Chinese Spring* while *Renan* still contains the 9.8 kb fragment. In this respect, *Renan* is more similar to *Ae. tauschii* which also does not contain that deletion (InDel γ, Figure 2, see below). A second, less dramatic difference between *Renan* and the two other sequences is a 20 bp deletion which is found only in *Renan* but not in *Glenlea* and *Chinese Spring*. These two diagnostic deletions put the *Chinese Spring* and *Glenlea* sequences phylogenetically closer together, independently from the divergence time estimates. In summary, the three wheat varieties *Chinese Spring*, *Renan* and *Glenlea* are overall very similar to one another, although *Renan* is clearly the most divergent of the three.

Comparison of hexaploid wheat and *Ae. tauschii* reveals a complex history of recombination

The *Ae. tauschii* sequence has a size of 180 kb and is completely covered by the *Chinese Spring* sequence. It contains one sequence gap of unknown size in a *CACTA* transposon of the *TAT1* family (Figure 2), but the order of the two sequence contigs could be inferred based on the alignment with the *Chinese Spring* sequence. *Ae. tauschii*, although overall similar, shows several major differences to the hexaploid wheat sequences. These differences include the insertions of two *CACTA* transposons and a non-LTR retrotransposon in hexaploid wheat

(InDels α and β, Figure 2a) as well as the absence of a major unequal crossing-over event in an LTR retrotransposon (InDel δ, see below). Additionally, as mentioned above, the absence of a large deletion makes it more similar to *Renan* than to the two other sequences (InDel γ).

Interestingly, the *Ae. tauschii* sequence also shows an uneven distribution of nucleotide substitutions, similar to the barley haplotypes described above: a large region of approximately 80 kb in the left half of the sequence is highly conserved between hexaploid wheat (i.e. *Chinese Spring*) and *Ae. tauschii* (Figure 2b) as the two species differ in only 87 nucleotide positions. This region includes three of the genes and one non-LTR retrotransposon. However, the situation is more complex than at the barley *Rym4* locus and more difficult to interpret. From the SNP density plot (Figure 2b), it appears that there are at least three different levels of sequence conservation: a highly conserved region in the left half (Region A), a highly variable segment in the middle (Region D) and intermediate conservation in the centre and the right half of the sequence (Regions B and E). We interpret these data as combination of at least two, maybe three, haplotypes of different ages. The exact breakpoints could not be determined, but one appears to be close to the *LRK-1* gene (border between Regions A and B, Figure 2b). Comparisons of non-genic sequences to the left of that island produced divergence time estimates of roughly 57,000 years while those to the right of it indicate a divergence time of approximately 400,000 -500,000 years. In the right part, more TE insertions and other large InDels can be found, which fits to the higher divergence time estimate.

In addition, there is one region (Region C) that is so divergent in the two species that sequences could not be aligned reliably. This region was therefore excluded from the overall sequence alignment and compared separately (Figure 2c). Region C lies between the two *LRK* genes and has a size of 1800 bp in *Chinese Spring* and 2800 bp in *Ae. tauschii*. Some parts of Region C are conserved between *Chinese Spring* and *Ae. tauschii*, however, at a lower level of sequence identity. Additionally, it contains several unique fragments that indicate multiple insertion and deletion events (Figure 2c).

Unequal crossing-over events occurred frequently in both wheat and barley and can be reconstructed precisely through comparative analysis

In total, we identified nine unequal crossing-over events which caused differences in the compared sequence. These events must have occurred in the relatively recent evolutionary history since the divergence of the sequences studied. All detected unequal crossing-over events occurred in non-genic regions. Seven of them affected relatively small arrays of tandem repeats while two had a major impact on the size of the regions because they occurred between the LTRs of retrotransposons. One of them resulted in a solo-LTR while the other produced a tandem element: The barley *Haruna Nijo* sequence at the *Vrs1* locus contains a full-length *RLC_BARE1* retrotransposon close to its left end, while the *Morex* sequence contains a solo-LTR at this position. Based on the LTR divergence of the full-length element, we estimated *RLC_BARE1* to have inserted at its location approximately 190,000 years ago (+-63,000 years). The standard deviation is relatively high as the two LTRs differ in only nine positions. Because the solo-LTR in the *Morex* sequence is a hybrid of the two LTRs of the full-length element, comparison of specific SNPs allowed to narrow down the putative breakpoint of the unequal crossing-over to a stretch of 252 bp (Figure 3a).

The tandem element resulting from unequal crossing-over was found close to the left end of the region compared between *Chinese Spring* and *Ae. tauschii*. A large 14 kb LTR retrotransposon of the *Gypsy* superfamily was found in its original full-length form in *Ae. tauschii*, while in *Chinese Spring*, it has undergone an unequal crossing-over event. The unequal homologous recombination occurred between the two LTRs, resulting in a tandem element with three LTRs and two internal domains. Additionally, one unit of the tandem element was subsequently partially deleted (Figure 3b).

Transition rates in CG and CNG sites are evenly distributed in the sequences analysed

Methylation of cytosine in CG and CNG sites increases the likelihood of spontaneous transitions from C to T (reviewed by Walsh and Xu, 2006). Because intergenic (i.e. TE) sequences in Triticeae are often methylated (SanMiguel et la., 2002), it was possible that some

of the strong variation in SNP density is due to mutations in CG and CNG sites. Therefore, we compared the number of transitional base substitutions in CG and CNG sites (Cmet) with the number of transitions in other positions. The ratio TrCmet (Cmet/total numer of transitions) was calculated in sliding windows across the compared sequences. As shown in Figure 4, we found no evidence for enrichment of Cmet sites in specific TEs in any of the comparisons. This is in contrast to what was described previously (SanMiguel et al., 2002). The average TrCmet was very similar in all three comparisons, ranging from 29.6% in the *Chinese Spring*/*Ae. tauschii* comparison to 30% in the *Morex*/*Haruna nijo* comparison at the *Vrs1* locus. These values are relatively low and within the range of the 33%-43% that were previously reported for introns of genes (SanMiguel et al., 2002). Two intergenic regions (1 and 3, Figure 4) showed a low absolute number of SNPs but a high TrCmet ratio. Additionally, two genic sequences (2 and 4, Figure 4) showed very low TrCmet values.

Discussion

Compared to other plant species such as Arabidopsis or rice, sequence information from Triticeae crops is still very limited. In addition, only relatively few BAC libraries from different varieties are available. The germplasm analyzed in this study represents a broad range of Triticeae species and varieties, allowing for comparisons within diploid and hexaploid species as well as for cross comparison. By including sequences from BAC libraries each from three wheat and barley varieties, we exploited all resources available to date for barley and hexaploid wheat. The D genome of wheat is especially interesting as it shows a low level of diversity and provides a background with low levels of gene flow. In addition, loci on the D genome can be easily compared with sequences from the diploid donor species *Ae. tauschii*.

The analysis of intergenic sequences proved to be very efficient for understanding the evolutionary history of the loci studied. Since genes are under selection pressure, they do not diverge at a constant rate and can not be used directly for divergence time estimates. Additionally, genes evolve more slowly and allow for much fewer genomic rearrangements than intergenic sequences. In contrast, TE and other non-genic sequences keep a trace record

of all mutations and rearrangements. Therefore, the high repeat content of Triticeae genomes, although an obstacle for sequencing, makes them ideally suited to study the molecular mechanisms of genome evolution.

The extensive stretches of TE sequences allowed very precise estimates of divergence times. However, these have to be considered with caution for two reasons. First, for sequences which differ in only a few bp, all SNPs have to be carefully checked either by re-sequencing or at least by examining the quality of sequences and assembly in the respective positions. For this study, we re-sequenced the regions of all SNPs between the wheat varieties *Chinese Spring*, *Glenlea* and *Renan* to exclude the possibility of sequencing errors. Second, because TE sequences are often heavily methylated in Triticeae, an increased mutation rate in CG and CNG sites has to be expected (Walsh and Xu, 2006). In the regions studied, we did not find any indication for an increased frequency of SNPs in CG and CNG sites in TE sequences. This is in contrast to a previous study which found dramatic differences between TE and genic sequences (SanMiguel et al., 2002). Thus, precise divergence time estimates should always include rigorous evaluation of sequence quality and analysis of SNP frequencies in potential DNA methylation sites.

The Triticeae gene pool as a hodgepodge of ancient haplotypes

It is amazing that the comparison of only three loci from a few varieties resulted in the identification of multiple haplotypes. Our comparative analysis therefore suggests that haplotypes of distant evolutionary origin are common in the Triticeae gene pool. By analysing SNP frequencies, we found that different parts of the sequences have diverged at different time points. The identified haplotype divergences cover a relatively broad time span. Haplotypes 1 and 2 at the barley *Vrs1* locus diverged approximately 160,000 and 930,000 years ago while the two haplotype segments identified in *Ae. tauschii* diverged approximately 50,000 and 500,000 years ago. The ages of the two older haplotypes are comparable to the previously described ancient haplotypes at the wheat *Lr10* locus (Isidore et al., 2005a).

The evolutionary interpretation becomes more complex if the data are evaluated in more detail. For example, we estimated that the *Vrs1* sequences of *Haruna Nijo* and *Morex* diverged approximately 120,000 years ago which is significantly different from the 160,000 years divergence for the "younger" haplotype at the *Rym4* locus. Thus, the sequences of the *Vrs1* and *Rym4* loci must represent yet again haplotypes that diverged at different points in time. The more divergent part of the *Ae. tauschii* sequence might also represent several distinct segments as the SNP distribution along the sequence is very uneven when compared with *Chinese Spring*. Additionally, the *Ae. tauschii* sequence contains an extremely divergent DNA segment between the two *LRK* genes. Because of its short length, this is more likely to be the result of a gene conversion event rather than a double crossing-over. It seems that this region was introgressed from yet another, more divergent haplotype or from a different locus in the genome (e.g. a homoeologous locus of the A or B genome). An example of the latter was recently published (Zhang and Dubcovsky, 2008).

Interestingly, previous studies placed the divergence of the D genome of hexaploid wheat with *Ae. tauschii* approximately between 550,000 and 900,000 years (Chantret et al., 2005; Gu et al., 2006), indicating the existence of even more divergent haplotypes than those described in this study. Our finding of the "younger" haplotype which diverged only about 50,000 years ago indicates that the D genome pool has a large genetic diversity.

The level of polymorphism observed between the three wheat varieties is much lower than between the barley varieties. This is in perfect agreement with the fact that the hybridization event adding the D genome to tetraploid wheat occurred only about 8,000 years ago (Feldman, 2001), with very little or no gene flow from the wild to the cultivated D genome afterwards (Chao et al., 2008). Indeed, our data indicate that the *Lr34* loci of *Glenlea* and *Chinese Spring* diverged probably within the past 700 years, clearly after the formation of hexaploid wheat. The *Renan Lr34* locus is more diverse and might have diverged in the early stages of agriculture about 6,300 years ago. Interestingly, *Renan* shares characteristics with *Ae. tauschii* at the *Lr34* locus, such as the absence of a large deletion. This deletion might have occurred less than 6,300 years ago in the lineage leading to *Chinese Spring* and *Glenlea,* but before their

divergence about 700 years ago. Thus, the age of the haplotypes reflects at the molecular level the evolutionary history of the young hexaploid wheat D genome which is only a few thousand years old. Nevertheless, the haplotypes have clearly distinct sequences, reflecting their origin in lines with highly divergent breeding histories and different gene pools.

In contrast, the barley haplotypes are much older. There are two possible explanations which are not mutually exclusive. It has been hypothesized (Zohary, 1996) that barley domestication was a multiple event. This would result in the presence of distinct, old haplotypes within the gene pool of domesticated barley. Alternatively, the extensive gene flow from the wild barley species *H. spontaneum* can explain the presence of old haplotypes in cultivated barley as well as in recombined sequences of very different ages. A similar situation with recombined haplotypes is also found at the *Lr34* locus of *Ae. tauschii.*

We conclude that the molecular analysis of intraspecies diversity in wheat and barley at the haplotype level confirms and extends our knowledge on the evolutionary history of these two crop species. Our data demonstrate that divergence time can only be determined for haplotypes or segments of them, but not for varieties. Therefore, we are confronted with an emerging picture of a gene pool that consists of a large number of haplotypes which have frequently recombined in the past. However, these events were not frequent enough as to completely homogenise the entire gene pool, instead still allowing for distinct haplotype fragments to be identified.

Insertions and deletions are abundant and caused by a variety of molecular mechanisms

The intraspecies comparative analyses provided an opportunity to study seven unequal crossing-over events in great detail. For all of them, obvious template sequences such as LTRs of retrotransposons or arrays of tandem repeats could be identified. The two unequal crossing-over events in LTR retrotransposons nicely illustrate the potential of unequal homologous recombination to either expand or contract a region. Solo-LTRs are reported frequently in plants (Shirasu et al., 2000; Vitte and Panaud, 2003; Ma et al., 2007) but tandem

elements appear to be less abundant (Sentry and Smyth, 1989). It is rare that one has orthologous loci available to study such events in detail.

Five incidents of unequal crossing-over affected arrays of tandem repeats. All these arrays were either found in TEs or in non-coding intergenic sequences. From the available data, we conclude that such tandem arrays frequently expand and contract through unequal crossing-over. As it was recently described for leucine-rich repeats of resistance gene analogs (Wicker et al., 2007), initial pairs of tandem repeats can be caused by illegitimate recombination. Unequal homologous recombination can then lead to a virtual runaway amplification of such repeats, resulting in the observed large arrays of direct repeats. We do not know if these arrays have a biological function. Judging from their occurrence mainly in non-coding DNA, it seems that such arrays are part of the "noise" of genome evolution.

As in previous studies (Shirasu et al., 2000; Wicker et al., 2001; SanMiguel et al., 2002), TE insertions were found to be the major cause of size increase in the regions studied. A particularly striking example is the expansion of the *Rym4* locus in the barley variety *Morex* by more than 70 kb in less than 1 million years. The wheat *Lr34* locus has expanded considerably by the insertion of two TEs within the past 500,000 years. We also identified several deletions which were apparently caused by illegitimate recombination, ranging in size from a few bp to almost 10 kb. This supports previous observations which indicated that illegitimate recombination is an important source of major changes in the plant genomes and that it can partially compensate the expansion of the genome by TE insertions (Devos et al., 2002; Wicker et al., 2003; Ma and Bennetzen, 2006; Wicker et al., 2007).

SSRs or microsatellites have long been used as molecular markers and are formed by relatively long stretches of di, tri-or tetranucleotide repeats which are polymorphic because of template slippage. Indeed, we also found that most InDels were apparently caused by template slippage, resulting in a short DNA motif being repeated in different numbers between two genomic sequences. Template slippage was by far the most abundant cause of InDels between sequences that diverged only recently (e.g. *Glenlea* and *Chinese Spring*). However, two observations were unexpected: First, we found that most of the motifs that were affected

by template slippage were relatively short homopolymers (i.e. stretches of only 3 or 4 identical nucleotides). This is surprising because it is commonly found that SSRs are only polymorphic if they have a certain length. Second, InDels greatly outnumbered nucleotide substitutions in the *Glenlea* vs. *Chinese Spring* comparison. This suggests that template slippage is a more frequent source of polymorphism than nucleotide substitution. This finding has implications for the design of highly polymorphic molecular markers for breeding germplasm, particularly for the wheat D genome with its low degree of polymorphism: a strategy allowing the efficient isolation of short homopolymers, e.g. by re-sequencing arrays, might result in a very large number of potential molecular markers, as such short motifs are more frequent than the typically used SSRs and SNPs.

Conclusion

Although the regions studied in this work might not be fully representative for the whole genomes, e.g. because of a putatively strong selection pressure on these regions involved in disease resistance or spike morphology, there are several possible implications for the future genetic analysis of agronomical traits in the Triticeae crops. The finding of recombined haplotypes has to be considered in association mapping, as single, “haplotype-specific” markers might actually detect different, recombined haplotypes. As we have shown here, comparative analysis allows to detect such recombined haplotypes. A future, large-scale analysis of haplotypes (including both genic and intergenic regions) in Triticeae gene pools will be greatly supported by next generation sequencing technologies such as 454, Solexa and oligonucleotide re-sequencing arrays similar to those used in Arabidopsis (Clark et al., 2007; Zeller et al., 2008). Such data will provide a basis for genome-wide mapping of intraspecies variability and haplotypes in the Triticeae gene pool with the ultimate goal of obtaining haplotype maps including both genic and intergenic regions. Such maps could help identify large chromosomal segments containing multiple traits or alleles in specific varieties. An essential requirement for such broad and systematic analyses will be the production of advanced drafts of complete reference genomes for wheat and barley. Two international

efforts are currently on the way to produce physical maps for the wheat variety *Chinese Spring* (International Wheat Genome Sequencing Consortium, wheatgenome.org/) and for the barley variety *Morex* (International Barley Sequencing Consortium, public.iastate.edu/~imagefpc/IBSC Webpage/).

We have found that despite a great diversity and great age of the ancient haplotypes, sequence diversity within genes or differences in gene content were minimal. This contrasts with findings in maize where a large number of gene fragments distinguished varieties (Fu and Dooner, 2002; Brunner et al. 2005). This suggests that the gene space is relatively stable in the Triticeae gene pool and that the movement of gene fragments is probably maize-specific and not simply a consequence of a large genome size. Therefore, the high phenotypic diversity observed in the wheat and barley gene pools is likely to derive strongly from SNPs in coding and regulatory regions. However, especially in hexaploid wheat, loss of genes might be more frequent because of genomic redundancy between the three subgenomes (Dubcovsky and Dvorak, 2007) and sample size in this study may be too small to allow a quantitive statement on that phenomenon. In addition, the observed insertion polymorphisms caused by transposons even between closely related haplotypes might have substantial effects on the expression of neighbouring genes and contribute significantly to phenotypic variability.

Material and Methods

Shotgun sequencing

BAC clones were obtained by screening of publicly available BAC libraries of the respective varieties (Table III). Identified BACs were confirmed either by Southern hybridization to fingerprint blots obtained after single restriction of BAC DNA or by PCR amplification of specific genic or genomic markers from the corresponding BACs. BAC shotgun sequencing was done on an ABI3730 automated sequencer (Applied Biosystems, Foster City, CA). Base calling and quality trimming of the sequences were done using PHRED version 0.020425.c (Ewing et al. 1998) and the initial assembly of BAC sequences was done with the PHRAP assembly engine version 1.080721 (provided by P. Green and available at http://www.phrap.org). Gaps in the

BAC sequences were closed by by primer walking on shotgun clones or by PCR amplication of fragments from BAC DNA.

Sequence analysis

For sequence analysis, programs from the EMBOSS package (http://emboss.sourceforge.net/), CLUSTAL W (Thompson *et al.*, 1994), and DOTTER (Sonnhammer and Durbin, 1995) were used. In a first step, all known repetitive elements were identified through BLAST (Altschul et al., 1997) against the database for Triticeae repetitive elements (TREP) (wheat.pw.usda.gov/ITMI/Repeats) and annotated. The remaining sequence, not annotated, was screened for the presence of putative genes by BLASTX against all rice and Arabidopsis proteins and BLASTN against all Triticeae ESTs. Conserved non-coding sequences (CNS) were identified by BLASTN search of the sequences that were still not annotated at that point against the entire TIGR rice genome (version 5, http://rice.plantbiology.msu.edu/). Identified repetitive elements were submitted to the TREP database. Pairwise alignment of large genomic regions was done by aligning a series of 10 kb fragments with the EMBOSS program WATER. The sequences pairs were the concatenated into one contiguous pairwise alignment and numbers of nucleotide substitutions were determined by an original Perl program. Molecular dating was done according to SanMiguel et al. (1998), but applying a synonymous substitution rate of 1.3E-8 (Ma and Bennetzen, 2004). Here, it is important to note that most studies, except those published within the past two or three years, have used a lower substitution rate of 6.5E-9 (Gaut et al., 1996), resulting in more ancient divergence times. The results can be easily converted by dividing the divergence times by two. We did so when we cite the following studies: (Isidore et al., 2005a; Gu et al., 2006).

For sequences which differed in only very few base pairs, primers were designed and all regions containing SNPs were re-sequenced independently.

The sequences described in this study were deposited at GenBank under the accession numbers FJ436983 and FJ436984 - FJ436986.

Acknowledgements

We thank Jelena Perovic for her excellent technical assistance.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and psi-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci **274:** 227–274

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17: 343–360

Chalhoub B, Belcram H, Caboche M (2004) Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. Plant Biotechnol J **2:** 181–188

Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R, Gornicki P (2008) *Acc* homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci USA **105:** 9691–9696

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, et al. (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). Plant Cell **17:** 1033–1045

Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo M-C, Dubcovsky J (2008) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. Mol Breeding DOI 10.1007/s11032-008-9210-6

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, et al. (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana.* Science **317:** 338–342

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res **12:** 1075–1079

Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. Science **316:** 1862-1866. Erratum in: Science. **318:** 393

Ewing B, Hillier L, Wendl M, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8:** 175-185

Feldman, M (2001) Origin of cultivated wheat. In AP Bonjean, WJ Angus, eds, The World Wheat Book: A History of Wheat Breeding, Ed 1. Lavoisier Publishing, pp. 3–56.

Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci USA **99:** 9573–9578

Gaut, BS, Morton, BR, McCaig, BC, Clegg, MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc Natl Acad Sci USA **93:** 10274-10279

Grando S, Macpherson H (2005) Food barley: Importance, uses and local knowledge. Proceedings of the International Workshop on Barley Improvement **1:** 156–156

Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, et al. (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics **174:** 1493–1504

Han B, Xue Y (2003) Genome-wide intraspecific DNA-sequence variations in rice. Curr Opin Plant Biol **6:** 134–138

Isidore E, Scherrer B, Chalhoub B, Feuillet C, Keller B (2005a) Ancient haplotypes resulting from extensive molecular rearrangements in the wheat a genome have been maintained in species of three different ploidy levels. Genome Res **15:** 526–536

Isidore E, Scherrer B, Bellec A, Budin K, Faivre-Rampant P, Waugh R, Keller B, Caboche M, Feuillet C, Chalhoub B (2005b). Direct targeting and rapid isolation of BAC clones spanning a defined chromosome region. Funct Integr Genomics **5:** 97– 103

Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, et al. (2007) Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. Proc Natl Acad Sci USA **104:** 1424–1429

Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA **102:** 9068–9073

Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol Microbiol **52:** 1243–1253

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA **101:** 12404–12410

Ma J, Bennetzen JL (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci USA **103:** 383–388

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet **23:** 134–139

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by Helitron-like transposons generate intraspecies diversity in maize. Nat Genet **37:** 997–1002

Nilmalgoda SD, Cloutier S, Walichnowski AZ (2003) Construction and characterization of a bacterial artificial chromosome (BAC) library of hexaploid wheat (*Triticum aestivum* L.) and validation of genome coverage using locus-specific primers. Genome **46:** 870–878

Petrov DA (2001) Evolution of genome size: new approaches to an old problem. Trends Genet **17:** 23–28

Pickering R, Johnston PA (2005) Recent progress in barley improvement using wild species of *Hordeum.* Cytogenet Genome Res **109:** 344–349

Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics. **162:**1389-1400

Ravel C, Praud S, Murigneux A, Canaguier A, Sapet F, Samson D, Balfourier F, Dufour P, Chalhoub B, Brunel D, et al. (2006) Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). Genome **49:** 1131–1139

Saisho D, Myoraku E, Kawasaki S, Sato K, Takeda K (2007) Construction and characterization of a bacterial artificial chromosome (BAC) library from the japanese malting barley variety '*Haruna Nijo*'. Breeding Science **57:** 29–38

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet **20:** 43–45

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct Integr Genomics **2:** 70–80

Scherrer B, Isidore E, Klein P, soon Kim J, Bellec A, Chalhoub B, Keller B, Feuillet C (2005) Large intraspecific haplotype variability at the *rph7* locus results from rapid and recent divergence in the barley genome. Plant Cell **17:** 361–374

Sentry JW, Smyth DR (1989) An element with long terminal repeats and its variant arrangements in the genome of *Lilium henryi.* Mol Gen Genet **215:** 349–354

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res **10:** 908–915

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167:** GC1–G10

Stein N, Perovic D, Kumlehn J, Pellio B, Stracke S, Streng S, Ordon F, Graner A (2005) The eukaryotic translation initiation factor 4e confers multiallelic recessive bymovirus resistance in *Hordeum vulgare* (L.). Plant J **42:** 912–922

Thompson, JD, Higgins, DG, Gibson, TJ (1994) CLUSTAL W, improving the sensitivity of progressive multiple sequence alignment through sequence weighting. position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

Vitte C, Panaud O (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol Biol Evol **20:** 528–540

Walsh CP, Xu GL (2006) Cytosine methylation and DNA repair. Curr Top Microbiol Immunol **301:** 283-315

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J **26:** 307–316

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu ZD, Dubcovsky J, 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, Yahiaoui N, Keller B (2007) Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. Plant J **51:** 631–641

Wicker T, Zimmermann W, Perovic D, Paterson AH, Ganal M, Graner A, Stein N (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-EIF4e* locus: recombination, rearrangements and repeats. Plant J **41:** 184–194

Zhang W, Dubcovsky J. (2008) Association between allelic variation at the Phytoene synthase 1 gene and yellow pigment content in the wheat grain. Theor Appl Genet **116:** 635-645

Zeller G, Clark RM, Schneeberger K, Bohlen A, Weigel D, Rätsch G (2008) Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. Genome Res **18:** 918–929

Zohary D. (1996) The mode of domestication of the founder crops of Southwest Asian agriculture. In DR Harris, ed, The origins and spread of agriculture and pastoralism in Eurasia, Ed 1. University College London Press, London, pp 142–158.

Figure Legends

Figure 1. Comparison of two loci *Vrs1* (**a.** through **d.**) and *Rym4* (**e.** and **f.**) from the barley varieties *Morex* and *Haruna Nijo*. **a.** Density of nucleotide substitutions between *Morex* and *Haruna Nijo* in a 1000 bp sliding window with 100 bp sliding steps. The red line is the observed SNP density while the grey line represents the simulation of purely randomly distributed SNPs. UECO: positions in which unequal homologous crossing-over between direct repeats occurred. **b.** Examples of InDels between *Morex* (top) and *Haruna Nijo* (bottom) which were presumably caused by template slippage. **c.** Deletion that shows a 3 bp illegitimate recombination signature. **d.** Dotplot alignment of a tandem repeat array. The arrow indicates a part of the array that has presumably undergone an unequal crossing-over event, resulting in the presence of five repeat units in *Morex* and three in *Haruna Nijo*. **e.** Comparsion of the *Rym4* locus from the two barley varieties *Morex* and *Haruna Nijo*. The map depicts the sequence organisation of the hypothetical ancestor sequence. Transposable elements which have subsequently inserted in *Morex* (top) and *Haruna Nijo* (bottom) are indicated as coloured boxes with arrows pointing to their insertion sites. Estimated times of insertions in million years are indicated inside the elements. SNP density is depicted as in **a.** The higher SNP frequency in the left part of the sequence indicates haplotype recombination. **f.** Model for the evolution and recombination of two ancient haplotypes.

Figure 2. Comparison of orthologous *Lr34* loci from three hexaploid wheat varieties and *Ae. tauschii*. a. Aligned maps of the four sequences (colours correspond to those in Figure 1). Gaps in one sequence indicate deletions in that sequence or the insertions of TEs in another. Positions of nucleotide substitutions between *Glenlea* and *Chinese Spring* are indicated as red vertical bars underneath the *Glenlea* map. Those between *Renan* and *Chinese Spring* are indicated as blue bars. Genes are numbered underneath the *Ae. tauschii* map and their transcriptional orientations are indicated with arrows. α, β, γ and δ indicate large InDels. The question mark indicates a gap of unknown size in the *Ae. tauschii* sequence b. Density of nucleotide substitutions between *Chinese Spring* and *Ae. tauschii* in a 1000 bp sliding window

with 100 bp sliding steps. The regions A through E are discussed in the text. c. Detailed map of a region highly variable between *Chinese Spring* and *Ae. tauschii*. Stretches that could be aligned are connected with turquoise and pink areas.

Figure 3. Unequal crossing-over events in LTR retrotransposons in barley and wheat. **a.** unequal crossing-over in a *RLC_BARE1* element resulting in a solo-LTR. SNPs by which the two LTRs in the full-length element differ are indicated with vertical bars. The grey areas indicate in which regions the solo-LTR contains SNPs which correspond to either the 5' or the 3'LTR of the full-length element. The 252 bp region in the solo-LTR between the two grey areas is where the unequal crossing-over must have occurred. The positions of non-diagnostic SNPs is indicated by asterisks in the solo-LTR sequence. These SNPs probably originated after the divergence of the two varieties. Thus, it is not possible to determine whether they were introduced in the *Haruna Nijo* or the *Morex* sequence. **b.** The *RLC_Geneva* is present in *Ae. tauschii* as a regular full-length element. In *Chinese Spring*, an unequal crossing-over event created a tandem element which was later affected by an internal deletion.

Figure 4. Distribution of SNP that were potentially caused by cysteine methylation (in CG or CNG sites) in barley and wheat varieties. Graphs and axes' units are labelled for the barley *Rym4* locus and are analogous for the *Vrs1* (middle) and *Lr34* loci (bottom). Genes are indicated as white boxes while TEs are shaded. Transcriptional orientation of genes is indicated with arrows. A CNS in the *Vrs1* locus is indicated with and asterisk. Regions that show especially high or low densities of SNPs in CG or CNG sites are labelled with numbers 1 through 4 and referred to in the text. SNP frequencies were calculated in sliding windows of 3,000 bp.

Table I. Overview of genes in the sequences studied.

Gene name	Locus	Description
Hox-1	*Vrs1*	Homeobox gene
CNS-1	*Vrs1*	Conserved non-coding sequence
EIF4E-1	*Rym4*	Eukaryotic translation initiation factor
MCT-1	*Rym4*	Pseudouridine synthase
Hex-1	*Lr34*	Hexose carrier
ABC-1	*Lr34*	ABC transporter
CYT-1	*Lr34*	Cytochrome P450
CNS-1	*Lr34*	Conserved non-coding sequence
LRK-1	*Lr34*	Lectin receptor-type kinase
LRK-2	*Lr34*	Lectin receptor-type kinase
CYT-2	*Lr34*	Cytochrome P450
CYP-1	*Lr34*	Cysteine protease[a]

[a]Pseudogene, contains TE insertion

Table II. Mechanisms that cause insertions and deletions in barley varieties.

Mechanism	*Vrs1*	*Rym4*[a]	*Rym4*[b]
Template slippage	49	51	21
Unequal crossing-over	2	1[c]	1[c]
Illegitimate recombination	5	8	0
Unclear	4	14	6
Total	**61**	**74**	**28**

[a]comparison of *Haruna Nijo* and *Morex*

[b]comparison of *Haruna Nijo* and *Cebada Capa*

[c]same event

Table III. BAC libraries from which clones were used for this study.

Species	Variety/accession	BAC library
H. vulgare	*Haruna Nijo*	Saisho et al., 2007
H. vulgare	*Cebada Capa*	Isidore et al., 2005b
H. vulgare	*Morex*	unpublished
T. aestivum	*Chinese Spring*	Chalhoub et al., 2004
T. aestivum	*Glenlea*	Nilmalgoda et al., 2003
T. aestivum	*Renan*	unpublished
Ae. tauschii	*AL8/78*	unpublished

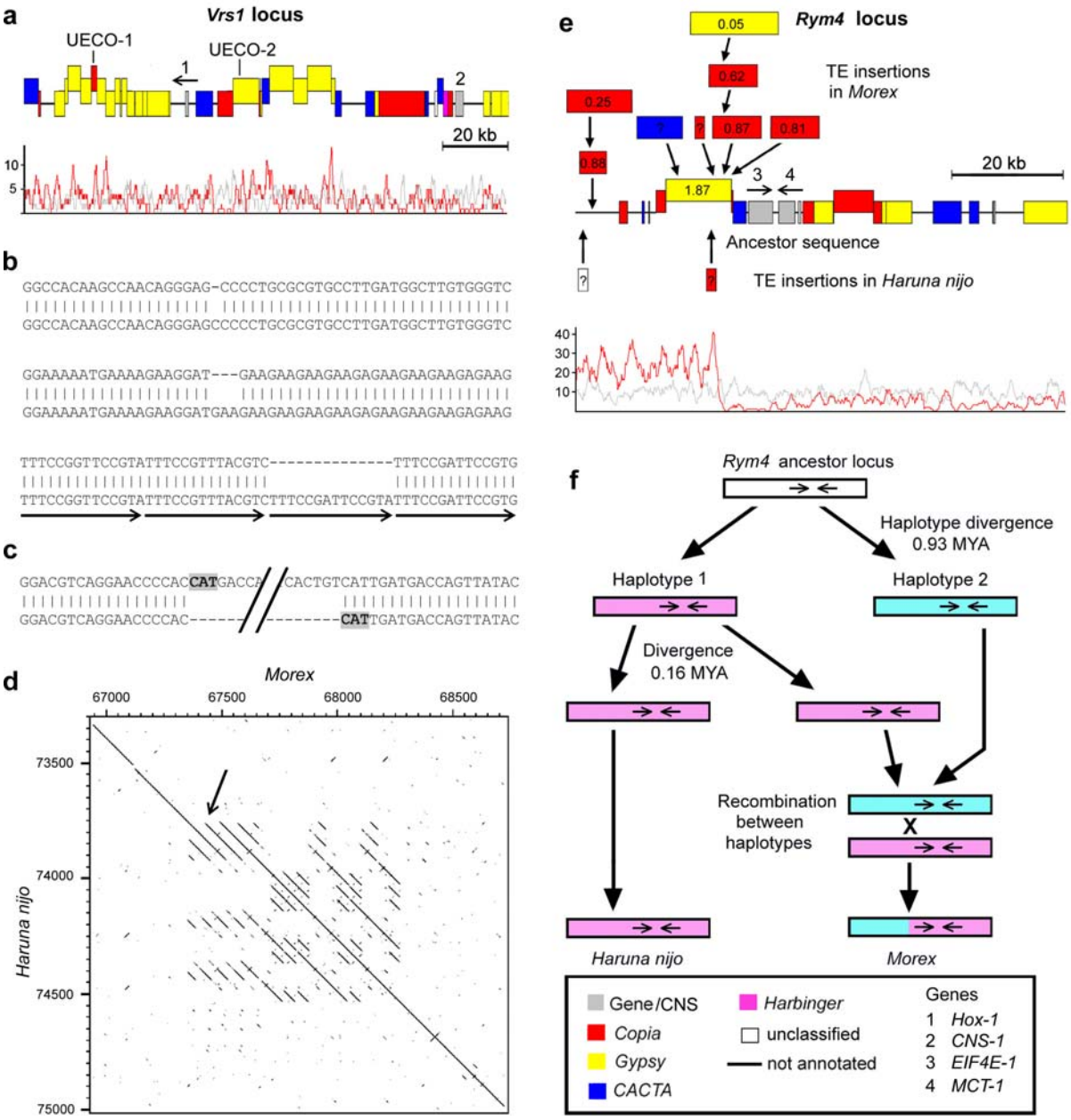

Figure 1. Comparison of two loci *Vrs1* (**a.** through **d.**) and *Rym4* (**e.** and **f.**) from the barley varieties *Morex* and *Haruna Nijo*. **a.** Density of nucleotide substitutions between *Morex* and *Haruna Nijo* in a 1000 bp sliding window with 100 bp sliding steps. The red line is the observed SNP density while the grey line represents the simulation of purely randomly distributed SNPs. UECO: positions in which unequal homologous crossing-over between direct repeats occurred. **b.** Examples of InDels between *Morex* (top) and *Haruna Nijo* (bottom) which were presumably caused by template slippage. **c.** Deletion that shows a 3 bp illegitimate recombination signature. **d.** Dotplot alignment of a tandem repeat array. The arrow indicates a part of the array that has presumably undergone an unequal crossing-over event, resulting in the presence of five repeat units in *Morex* and three in *Haruna Nijo*. **e.** Comparsion of the *Rym4* locus from the two barley varieties *Morex* and *Haruna Nijo*. The map depicts the sequence organisation of the hypothetical ancestor sequence. Transposable elements which have subsequently inserted in *Morex* (top) and *Haruna Nijo* (bottom) are indicated as coloured boxes with arrows pointing to their insertion sites. Estimated times of insertions in million years are indicated inside the elements. SNP density is depicted as in **a.** The higher SNP frequency in the left part of the sequence indicates haplotype recombination. **f.** Model for the evolution and recombination of two ancient haplotypes.

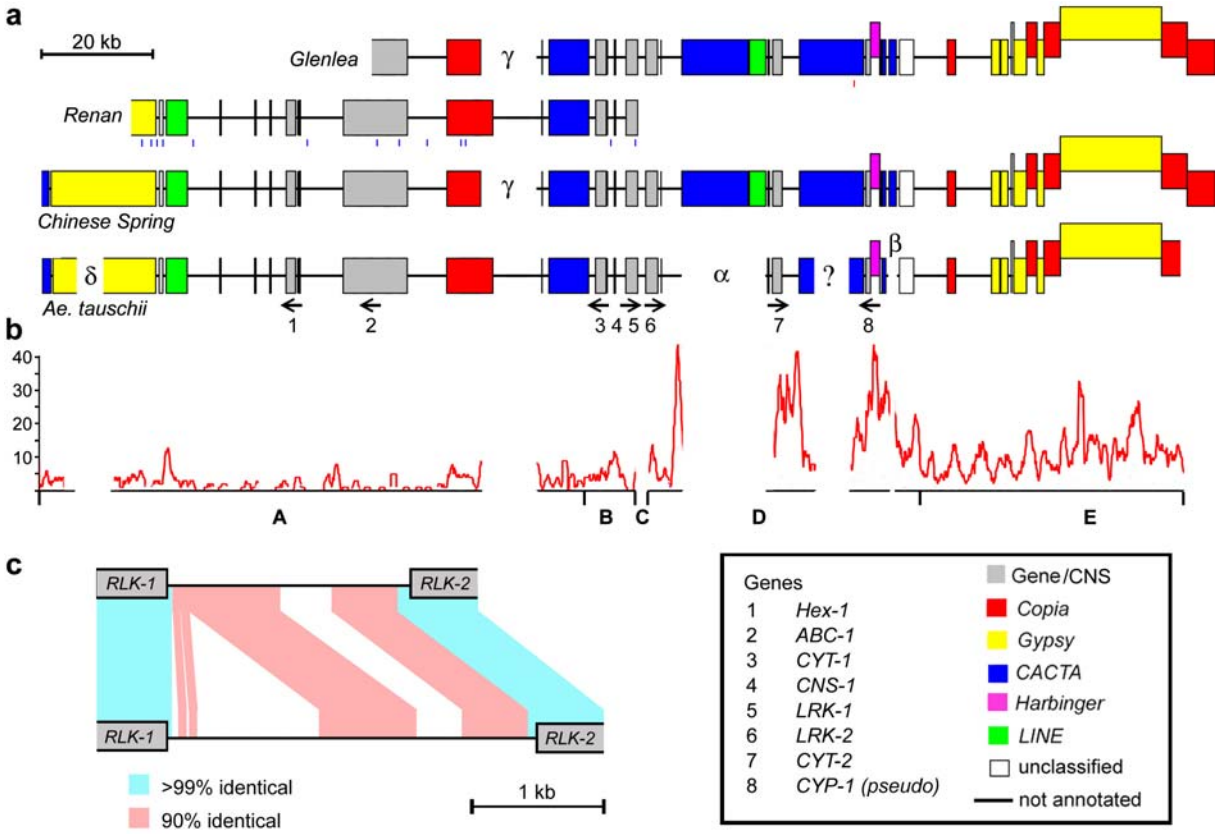

Figure 2. Comparison of orthologous *Lr34* loci from three hexaploid wheat varieties and *Ae. tauschii*. a. Aligned maps of the four sequences (colours correspond to those in Figure 1). Gaps in one sequence indicate deletions in that sequence or the insertions of TEs in another. Positions of nucleotide substitutions between *Glenlea* and *Chinese Spring* are indicated as red vertical bars underneath the *Glenlea* map. Those between *Renan* and *Chinese Spring* are indicated as blue bars. Genes are numbered underneath the *Ae. tauschii* map and their transcriptional orientations are indicated with arrows. α, β, γ and δ indicate large InDels. The question mark indicates a gap of unknown size in the *Ae. tauschii* sequence b. Density of nucleotide substitutions between *Chinese Spring* and *Ae. tauschii* in a 1000 bp sliding window with 100 bp sliding steps. The regions A through E are discussed in the text. c. Detailed map of a region highly variable between *Chinese Spring* and *Ae. tauschii*. Stretches that could be aligned are connected with turquoise and pink areas.

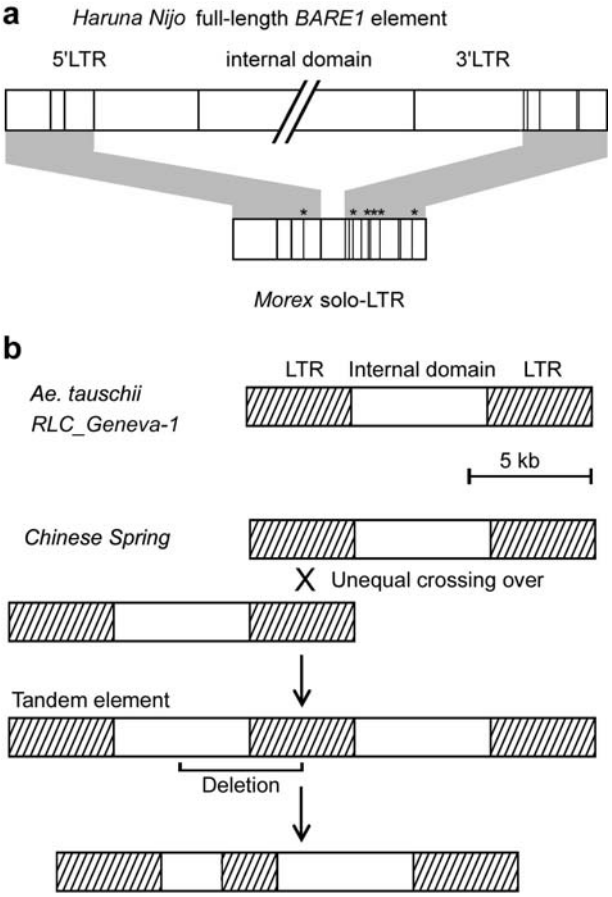

Figure 3. Unequal crossing-over events in LTR retrotransposons in barley and wheat. **a.** unequal crossing-over in a *RLC_BARE1* element resulting in a solo-LTR. SNPs by which the two LTRs in the full-length element differ are indicated with vertical bars. The grey areas indicate in which regions the solo-LTR contains SNPs which correspond to either the 5' or the 3'LTR of the full-length element. The 252 bp region in the solo-LTR between the two grey areas is where the unequal crossing-over must have occurred. The positions of non-diagnostic SNPs is indicated by asterisks in the solo-LTR sequence. These SNPs probably originated after the divergence of the two varieties. Thus, it is not possible to determine whether they were introduced in the *Haruna Nijo* or the *Morex* sequence. **b.** The *RLC_Geneva* is present in *Ae. tauschii* as a regular full-length element. In *Chinese Spring*, an unequal crossing-over event created a tandem element which was later affected by an internal deletion.

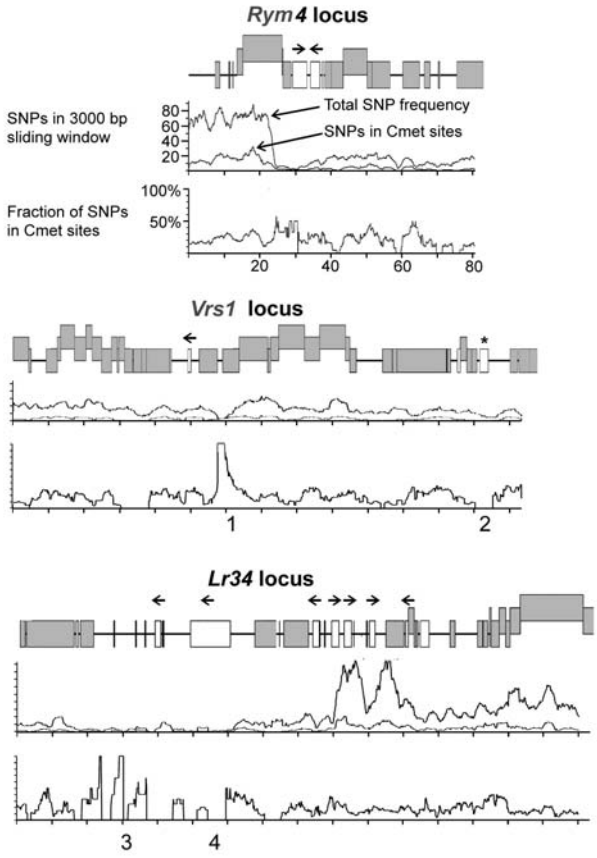

Figure 4. Distribution of SNP that were potentially caused by cysteine methylation (in CG or CNG sites) in barley and wheat varieties. Graphs and axes' units are labelled for the barley *Rym4* locus and are analogous for the *Vrs1* (middle) and *Lr34* loci (bottom). Genes are indicated as white boxes while TEs are shaded. Transcriptional orientation of genes is indicated with arrows. A CNS in the *Vrs1* locus is indicated with and asterisk. Regions that show especially high or low densities of SNPs in CG or CNG sites are labelled with numbers 1 through 4 and referred to in the text. SNP frequencies were calculated in sliding windows of 3,000 bp.