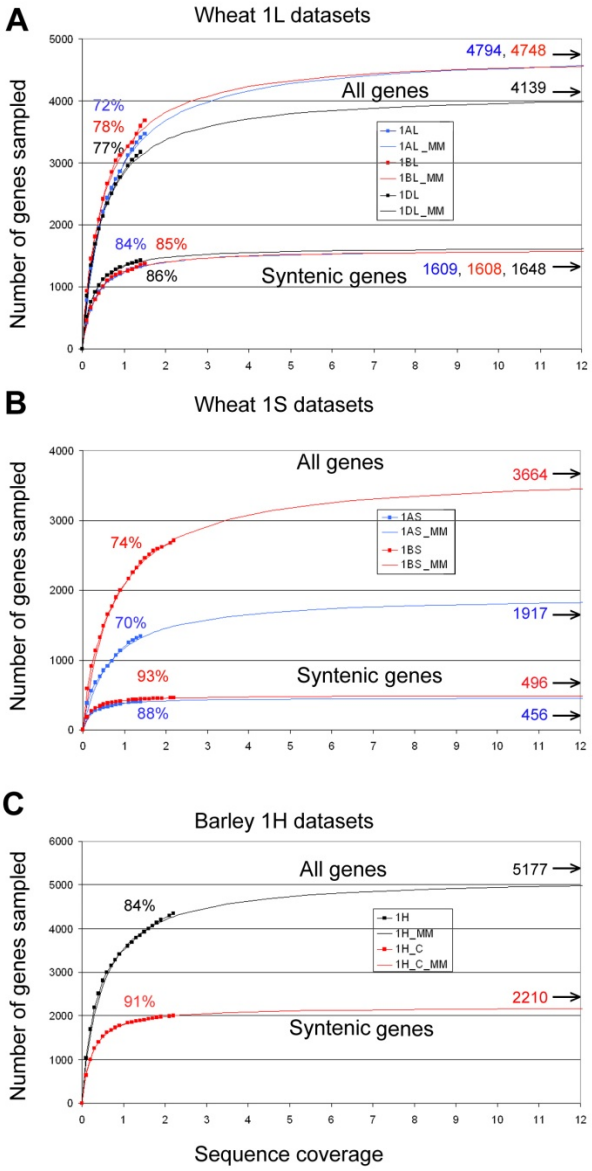
Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives

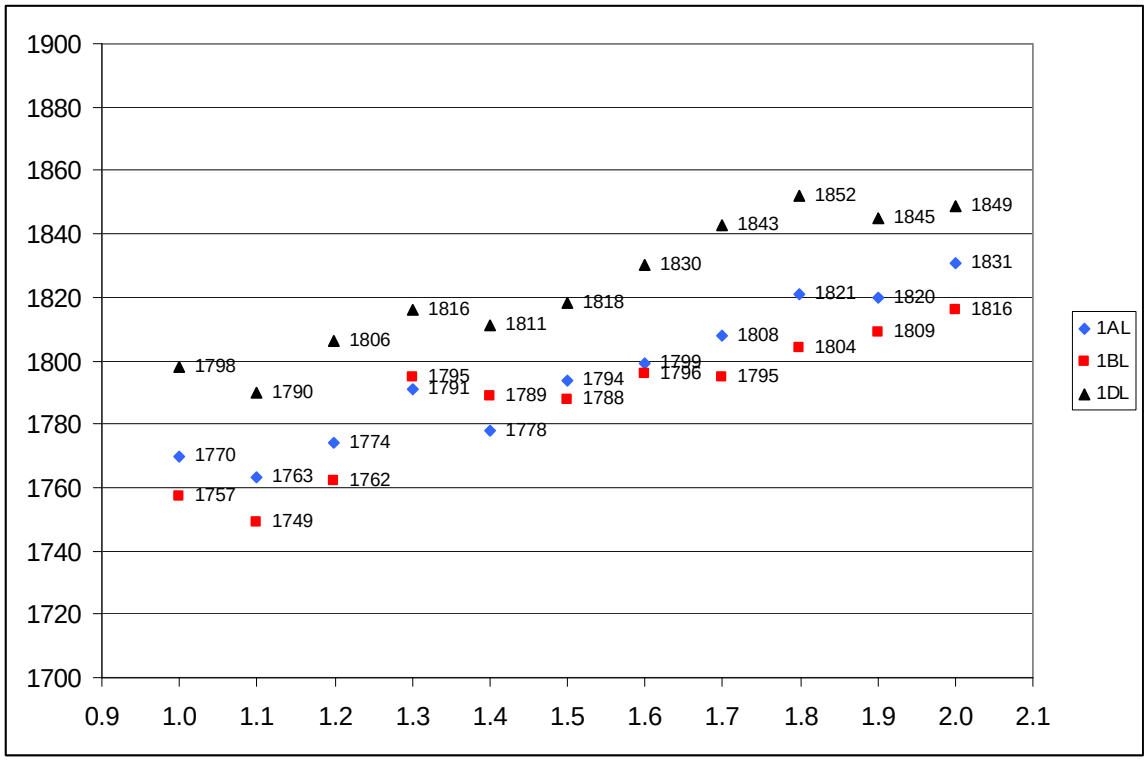
Wicker, Thomas ; Mayer, K F X ; Gundlach, H ; Martis, M ; Steuernagel, B ; Scholz, Uwe ; Simková, H ; Kubaláková, M ; Choulet, F ; Taudien, S ; Platzer, M ; Feuillet, C ; Fahima, T ; Budak, H ; Dolezel, J ; Keller, B ; Stein, N

Abstract: All six arms of the group 1 chromosomes of hexaploid wheat (Triticum aestivum) were sequenced with Roche/454 to 1.3- to 2.2-fold coverage and compared with similar data sets from the homoeologous chromosome 1H of barley (Hordeum vulgare). Six to ten thousand gene sequences were sampled per chromosome. These were classified into genes that have their closest homologs in the Triticeae group 1 syntenic region in Brachypodium, rice (Oryza sativa), and/or sorghum (Sorghum bicolor) and genes that have their homologs elsewhere in these model grass genomes. Although the number of syntenic genes was similar between the homologous groups, the amount of nonsyntenic genes was found to be extremely diverse between wheat and barley and even between wheat subgenomes. Besides a small core group of genes that are nonsyntenic in other grasses but conserved among Triticeae, we found thousands of genic sequences that are specific to chromosomes of one single species or subgenome. By examining in detail 50 genes from chromosome 1H for which BAC sequences were available, we found that many represent pseudogenes that resulted from transposable element activity and double-strand break repair. Thus, Triticeae seem to accumulate nonsyntenic genes frequently. Since many of them are likely to be pseudogenes, total gene numbers in Triticeae are prone to pronounced overestimates.

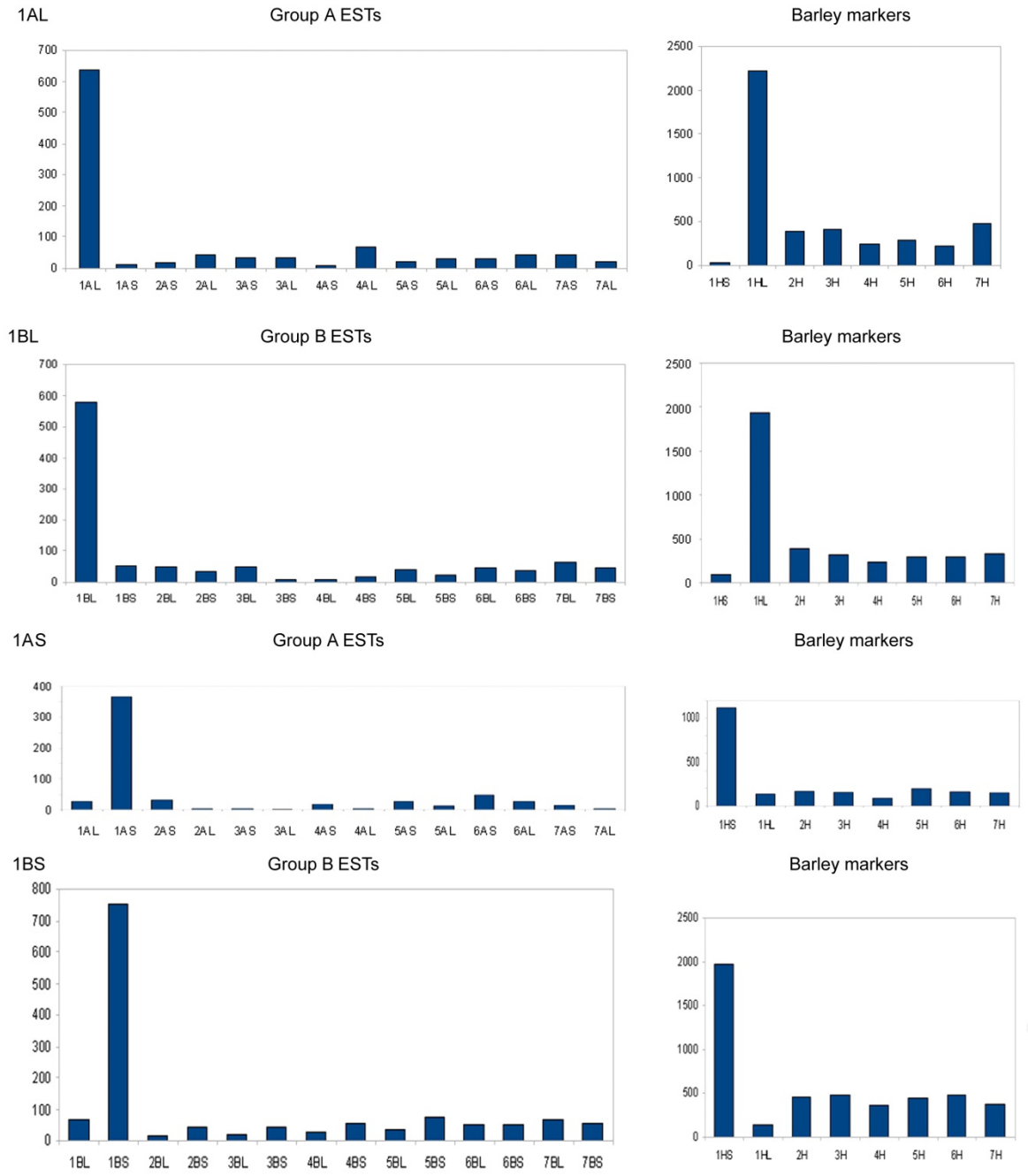
DOI: https://doi.org/10.1105/tpc.111.086629



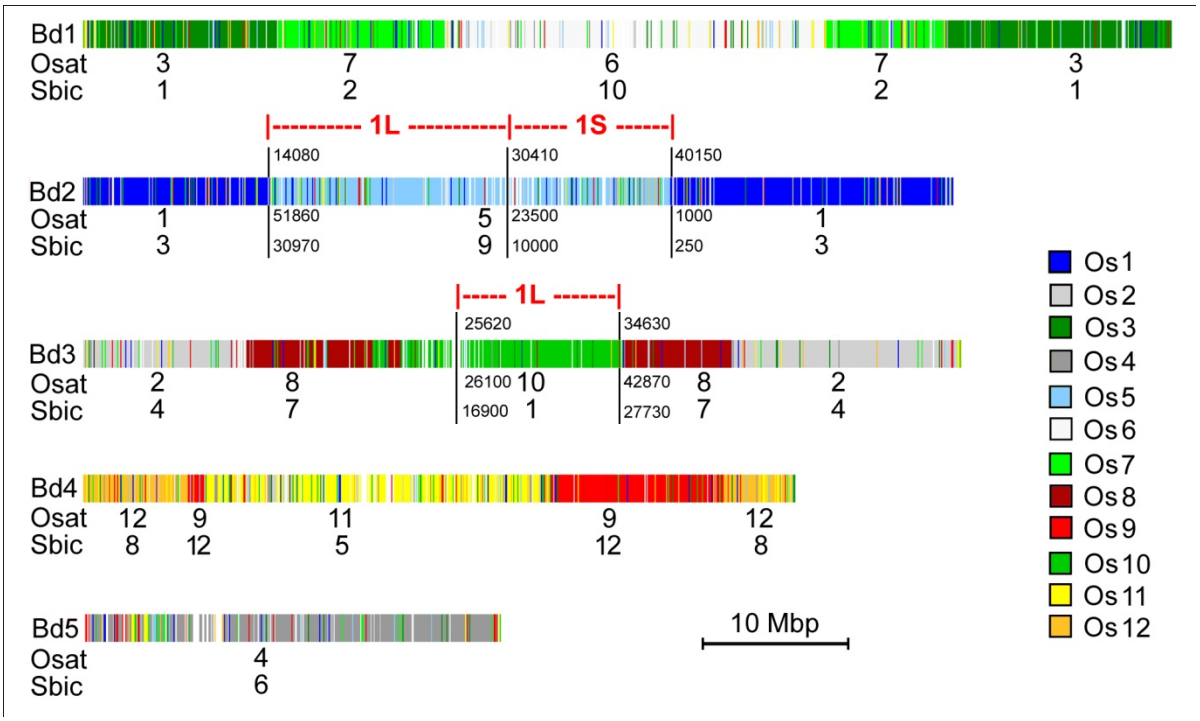
Supplemental Figure 1. Relationship between sequence coverage and number of genes sampled by Roche/454. For the series, we used different sized portions of the datasets. **A.** Datasets from the long arm of wheat chromosome 1. The datasets for the syntenic genes are very similar while the complete gene set varies more. The curves can, in principle, be described by a Michaelis-Menten (MM) saturation function with the saturation value being the actual number of genes on a chromosome. The values for the observed data are indicated with black, blue or red rectangles while the fitted MM-curve is a solid line. The saturation values (corresponding to the actual number of genes) for the extrapolated curves are shown at the very right. The percentages indicate what fraction of the saturation gene number has been sampled at the highest coverage value. **B.** Datasets for the short arms of wheat chromosomes 1A and 1B. The dataset for chromosome arm 1DS was not used (see text for explanation). 1AS and 1BS have a similar number of syntenic genes but differ vastly in the number of non-syntenic ones. **C.** Dataset for the entire chromosome 1 of barley.



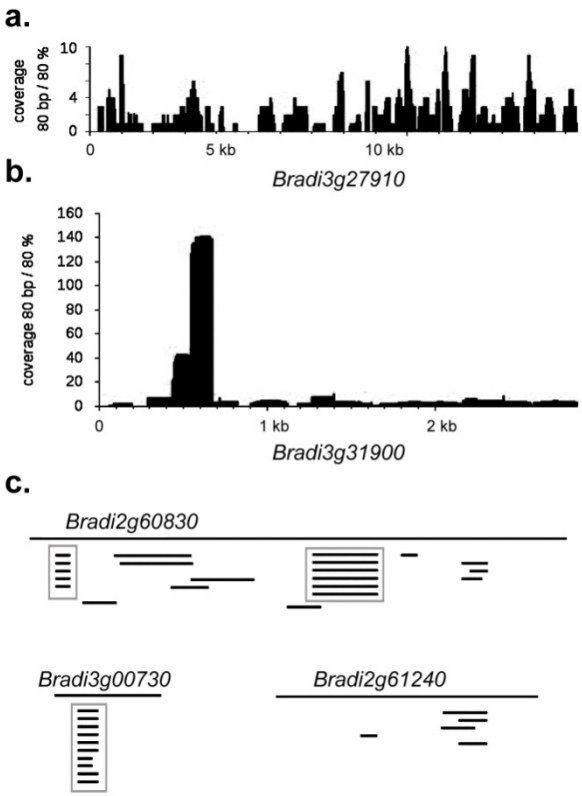
Supplemental Figure 2. Relationship of estimated physical chromosome size to extrapolated gene numbers on chromosome arms. The goal was to test how much the extrapolated number of genes would vary if size estimates of chromosomes had been inaccurate (and therefore the sequence coverage differed from the expected value). The x-axis displays the range of coverage assuming different size estimates of the chromosome arms and the y-axis shows the number of genes extrapolated from the sequence datasets. The differences in extrapolated gene numbers are surprisingly small for a coverage ranging from 1.0x to 2.0x.



Supplemental Figure 3. Evaluatuon of purity of the sorted wheat chromosome arms. To save space, only the results for chromosome 1A and 1B are shown. Results for 1DS are comparable, while 1DL turned out to be a re-arramged chromosome (see text). The graphs on the left side show the number of 454 reads that were hit when ESTs maped to groups of homoeologous chromosomes were used in BLASTN searches against the 454 datasets. The graphs on the right side show the number of 454 reads that were hit when pools of chromosome specific barley ESTs were used in BLASTN searches against the 454 datasets



Supplemental Figure 4. Regions syntenic to Triticeae chromosome arms 1S and 1L in *Brachypodium*, rice and sorghum. The five bars represent the five *Brachypodium* chromosomes (Bd1 through Bd5). Areas of different colours refer to different syntenic chromosomal regions in rice and sorghum. The large numbers underneath the maps indicate to which chromosome in rice (Osat) and sorghum (Sbic) an region corresponds. The vertical bars indicate the boundaries of the Triticeae group 1 syntenic regions of 1L and 1S. The gene identifier numbers at the boundaries are given for *Brachypodium* (above the maps), rice and sorghum (underneath the maps). Different colours indicate the syntenic relationships to rice chromosomes.



Supplemental Figure 5. Examples for 454 coverage of *Brachypodium* genes. **A.** Coverage of gene *Bradi3g27910* with 454 reads. The BLASTN search identified a total of 148 reads that produce hits longer than 80 bp. The coverage is relatively even. **B.** Gene *Bradi3g31900* is covered unevenly with 454 reads, indicating a strong distortion because of the amplification of the DNA. **C.** Removal of redundancy caused by amplification artifacts. The full size of the CDS is indicated with a black bar. 454 reads are indicated as bars underneath the gene. Multiple reads covering the same region (indicated by grey boxes) of the gene are interpreted as artifacts and all except one are removed from the records. In the case of *Bradi3g00730*, the gene is, therefore covered by only one read after removal of redundant reads.

SupplementalTable 1. Pairs of BAC clones of which one was mapped to chromosome 1H and the other one elsewhere in the genome.

1H copy	Comment	non-1H copy	Rem	*Brachypodium*[a]	Acc numbers
407O08*	Pseudo	432B18	Intact	Bradi1g00850	ERR013911,ERR013563
393K06	Pseudo	767M04	Intact	Bradi1g01110	ERR014029,ERR014259
48N22	Pseudo	401J16	Pseudo	Bradi1g01170	ERR014486,ERR013316
416C24	Pseudo	498E18	Intact	Bradi1g02670	ERR015286,ERR013976
272P03	Pseudo[b]	178I23	Pseudo	Bradi1g04870	ERR014838,ERR015277
225A16	Pseudo	456F24	Pseudo	Bradi1g06350	ERR014696,ERR014673
18A09	Pseudo[b]	22J09	Pseudo[b]	Bradi1g07750	ERR014249,ERR014784
522J06	Pseudo	56M07	Intact	Bradi1g09320	ERR014801,ERR020498
69G11	Pseudo	311L15	Pseudo	Bradi1g11180	ERR014500,ERR013785
816C23	Intact	486P08	Pseudo	Bradi1g12160	ERR013667,ERR014433
549F09	Intact	295L22	Pseudo	Bradi1g15030	ERR013668,ERR014090
23C11	Pseudo	151K05	Pseudo	Bradi1g28380	ERR013883,ERR015284
222O19	Pseudo	271P06	Intact	Bradi1g28540	ERR015021,ERR014831
311L15	Intact	686N02	Intact	Bradi1g29670	ERR013785,ERR015208
271P06	Pseudo	710M11	Pseudo	Bradi1g30150	ERR014831,ERR014155
311L15	Pseudo[b]	207K08	Pseudo[b]	Bradi1g31580	ERR013785,ERR015106
90K18	Pseudo	188N02	Pseudo	Bradi1g34150	ERR014551,ERR015232
773L18	Pseudo	329M04	Pseudo	Bradi1g38760	ERR013762,ERR013570
751P17	Intact	395O12	Pseudo	Bradi1g49140	ERR014957,ERR014724
363N16	Pseudo	500E15	Intact	Bradi1g49300	ERR013727,ERR014791
490K16	Pseudo	582D13	Intact	Bradi1g51460	ERR013801,ERR014611
18A09	Pseudo[b]	237G01	Intact	Bradi1g51650	ERR014249,ERR013945
174D12	Pseudo	87M05	Pseudo	Bradi1g52150	ERR015235,ERR014103
373K06	Pseudo	718N07	Pseudo	Bradi1g56780	ERR015087,ERR013381
635J11	Pseudo	475K23	Pseudo	Bradi1g67430	ERR014173,ERR015292
311L15	Pseudo[b]	324K14	Pseudo[b]	Bradi1g68120	ERR013785,ERR015046
316D17	Pseudo	78K03	Pseudo	Bradi1g69590	ERR014873,ERR014488
632I20	Pseudo	437L18	Intact	Bradi1g70040	ERR013764,ERR014034
312K12	Pseudo[c]	694I09	Intact	Bradi1g71060	ERR015063,ERR013912
696J09	Pseudo	466L17	Intact	Bradi2g02490	ERR014040,ERR014269
83KhA0073M11	Pseudo	49K16	Pseudo	Bradi2g42610	ERR013471,ERR014547
12J01	Pseudo	10L22	Intact	Bradi2g52770	ERR014778,ERR014786
814K22	Intact	385G18	Intact	Bradi2g53080	ERR013676,ERR014653

* To save space, the prefix “HVVMRXALLhA” as well as leading zeroes were omitted from the BAC identifiers

[a]Closest *Brachypodium* homolog

[b]Gene fragment duplicated by TE capture

[b]Gene fragment duplicated by reverse transcription

Supplemental Table 2. Removal of redundancy in the 454 data with homology to *Brachypodium* genes.

Dataset	Total reads[a]	nr reads[b]	Redundant
1AS	13692	8423	38.5%
1AL	30463	21178	30.5%
1BS	29295	18722	36.1%
1BL	32722	21968	32.9%
1DL	25652	19139	25.4%
1H	62601	39892	36.3%

[a]Total number of 454 reads that hit *Brachypodium* genes.

[b]Number of reads after removal of amplification artefacts.

[c]Fraction of reads that represent amplification artefacts