doi:10.1038/nature11543

SI Guide

Supplemental Note 1. Access to datasets

Table S1: Accession IDs and internet URLs for access to original datasets

Type of data	Database	Accession IDs / Study IDs/ URLs
WGS short reads raw data		
Barke	EMBL/ENA	ERP001450
Bowman	EMBL/ENA	ERP001449
Igri	EMBL/ENA	ERP001433
Haruna Nijo	EMBL/ENA	ERP001451
H. spontaneum B1K-04-12	EMBL/ENA	ERP001434
Morex	EMBL/ENA	ERP001435
WGS assembly		
Barke	EMBL/ENA	CAJV010000001-CAJV012742077
Bowman	EMBL/ENA	CAJX010000001-CAJX012077901
Morex	EMBL/ENA	CAJW010000001-CAJW012670738
BES		
HVVMRXALLeA	EMBL/ENA	HF140858-HF362636, HE975059-HE977519
HVVMRXALLhA	EMBL/ENA	HF000001-HF140857
HVVMRXALLhB	EMBL/ENA	HE867107-HE939654
HVVMRXALLrA	EMBL/ENA	HE939655-HE956691
HVVMRXALLmA	EMBL/ENA	HF362637-HF479769
BAC short reads raw data		
Illumina sequenced BACs	NCBI GenBank	SRA047913
Roche/454 sequenced BACs	EMBL/ENA	http://www.ebi.ac.uk/ena/data/view/ERP000238
BAC assembled sequences		
Illumina sequenced BACs	NCBI Genbank	AC250421.1-AC252610.1
	HarvEST	http://www.harvest-web.org/utilmenu.wc?job=RTRVFORM&db=MOREX_HV3_9
Roche/454 sequenced BACs	NCBI Genbank	AC247243.1-AC247289.1, AC247294.1-AC250420.1, AC252611.1-AC253531.1
	GABI-PD	http://www.gabipd.org/projects/Barley_BAC_Contigs
RNAseq		
RNAseq data of diverse cultivars	EMBL/ENA	ERP001573
RNAseq data of Morex developmental stages	EMBL/ENA	ERP001600
Genetic map data		
GBS data short reads raw data Morex x Barke DH population	NCBI Genbank	SRP010876.1
GBS data genotyping info Morex x Barke DH population	GrainGenes	http:// wheat.pw.usda.gov/cgi-bin/graingenes/report.cgi?class=mapdata;name=Barley,+ Morex+x+Barke+DH,+GBS
iSelect Morex x Barke RIL population	James Hutton Institute	http://bioinf.hutton.ac.uk/waugh/iselect
Sequence search		
BLAST analysis to WGS assemblies, BES, BAC	IPK Gatersleben	http://webblast.ipk-gatersleben.de/barley/
Physical map visualization		
CrowsNest	Helmholtz Center Munich, MIPS	http://seacow.helmholtz-muenchen.de/cgi-bin/gb2/gbrowse/Barley_PhysMap/

EMBL-ENA = EMBL/ENA Sequence Read Archive (SRA) (use: http://www.ebi.ac.uk/ena/data/view/accessionID), BES = BAC end sequence

Besides access to the original raw data as specified in Table S1, integrated meta-data sets (i.e. compilations of FPC contigs associated to genetic markers and sequence resources etc.) can be obtained via FTP download from:

ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/

Supplemental Note 2. Physical map construction

S2.1 Physical map

Table S2: Assembly statistics of barley HICF map

# of input BACs[1]	571,007
# clones in contigs	517,202
# of singleton clones	53,805
# of contigs	9,265
Assembly length (Gbp)	4.99
∅ contig size (Kbp)	538
N50[*]	1,587
L50[**] (Kbp)	904

[*]= number of contigs representing 50% of the total assembly length, [**]= sort all contigs from longest to shortest. L50 gives the length of the shortest of all long contigs that make up 50% of the assembly length.

S2.2 Selection of "gene-bearing" BAC clones: A total of 83,831 "gene-bearing" BACs in a barley cv. Morex library[2] were identified using genic probes. This is 26.8% of the complete 6.3x library and encompasses an estimated 75% of all gene-bearing BACs in this library. Most probes were applied as pools of nearly 200 simultaneous 36 bp overgos[3], resulting in lists of BACs associated with lists of genes rather than specific gene-BAC relationships. Information on additional gene-bearing BACs from prior work using hybridization or PCR amplification methods was also included in the compilation. These gene-bearing BACs were fingerprinted using a four-dye method[4] and assembled computationally using a compartmentalized approach[5]. Contigs in the resulting physical map (available at http://phymap.ucdavis.edu/barley) account for appr. 1,700 Mbp of the barley genome, roughly four times the size of the rice genome. A minimal tiling path that comprises almost 15,000 clones was computed. The minimal tiling path was rearrayed and used for BAC-by-BAC sequencing using a combinatorial. method[6]. About 70,000 of the gene-bearing BACs (among the clones from library HVVMRXALLhA, Table S3) were re-fingerprinted in the present study and thus included in the genome-wide HICF map of barley (Table S2).

Supplemental Note 3. Genomic sequencing

S3.1 BAC end sequencing (BES): Paired-end sequencing of bacterial artificial chromosome (BAC) clones was either performed according to published procedures[7] or by the following approach: Barley BAC clones were grown overnight in 384 –plates containing 180 µl of 2xLB (12.5 ug/ml chloramphenicol) medium. DNA was isolated by standard alkaline lysis plasmid-miniprep techniques adjusted for BACs and 384-plates (Millipore) and subsequently used as a template for cycle sequencing performed for each BAC end in a 10 µl reaction containing 0.92 X, 0.08 X BigDye Terminator Mix and 0.32 µM sequencing vector-specific primer using BigDye Terminator chemistry (Applied Biosystem, Foster City, USA). Sequencing reactions were conducted using a thermal cycler (Applied Biosystems 9800 Fast Thermal Cycler, Foster City, USA) with the following protocol: 35 cycles of a denaturation at 96°C for 10 sec, followed by annealing at 50°C for 5 sec and an extension at 60° C for 4 min. The products were purified from salts and non incorporated ddNTPs by ethanol precipitation (add 3µl of 125mM EDTA pH 8.0 and 30µl 100% EtOH, shaking of the plates, incubation of 15 min) and centrifugation (2900 x g, 55 min). After removal of supernatant and residual traces of EtOH, DNA pellets were resuspended (8µl of Hi-Di Formamide, Applied Biosystems, Foster City, USA) and sequencing reactions were resolved on a 96-capillary sequencing device (3730xl DNA Analyzer, Applied Biosystems, Foster City, USA). Sequences were quality trimmed by using software LUCY[8] (www.tigr.org/softlab) at standard parameter settings providing both vector sequence and cloning sites. Short reads (i.e., < 100bp) were automatically removed. In addition, BLASTN searches were performed to detect and discard sequences with high homology to organelle DNA and barley pathogen sequences available at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/).

Table S3: statistics of BES

Library[1]	# seqs	# BACs	∑ length (bp)
HVVMRXALLhA	140,857	73,298	98,648,831
HVVMRXALLhB	72,548	36,904	50,818,577
HVVMRXALLeA	224,239	119,071	145,873,939
HVVMRXALLmA	117,133	64308	68,455,174
HVVMRXALLrA	17,037	10,942	9,751,362
∑	**571,814**	**304,523**	**373,547,883**

S3.2 BAC shotgun sequencing and assembly

BAC clones were sequenced either using Roche/454 or Illumina GAIIx or Hiseq platforms. Sequence assemblies can be retrieved from: GABI primary database (http://www.gabipd.org/projects/Barley_BAC_Contigs); HarvEST database (http://harvest-web.org/utilmenu.wc?job=RTRVFORM&db=MOREX_HV5).

Table S4: BAC sequencing and assembly statistics

DataSet	# BACs	# Contigs	Σ Contigs (bp)	Avg BAC length (Kbp)	# Contigs per BAC**	L50 (bp)**
454 gene*	3158	67,353**	378,733,823**	120	21	30,987
454 random	937	18,898**	133,910,243**	142	20	43,074
Illumina*	2183	276,320	249,750,056	114	32	6,852

*BACs were pre-selected for the presence of genes[3],**=considering contigs of >500 bp only

S3.3 Genomic shotgun sequencing and assembly

Illumina whole genome shotgun sequencing: Illumina paired-end (PE) and mate-pair (MP) libraries with fragment lengths of ~500 bp (PE) and 2.5 kb (MP) were prepared from genomic DNA of different barley *(Hordeum vulgare* ssp. *vulgare)* cultivars ('Morex', 'Barke', 'Bowman', 'Igri'), and wild barley (*Hordeum vulgare* ssp. s*pontaneum*) (selection from B1K-04-12[9] purified by 3 cycles of single seed descent). DNA was fragmented mechanically (nebulization (PE) or shearing, Hydroshear, Digilab Inc., Holliston, MA, USA). Size selected DNA fragments were purified from excised agarose gels and checked for fragment length distribution by Agilent DNA 7500 chips (Agilent Technologies, Germany). PE library preparations were performed according to the manufacturer's instructions (Illumina "Preparing Samples for Paired-End Sequencing", Part # 1005063 Rev. A June 2008). For MP library preparation, a hybrid Roche/Illumina protocol was used as follows. First, circularization was performed according to the Roche Paired End Library Protocol, using Roche circularization adaptors (Paired End Library Preparation Method Manual 20 kb and 8 kb Span; Roche Diagnostics, October 2009, steps 3.1 to 3.7.3). Subsequently, the circularized fragments were fragmented again by nebulization and processed for library preparation following the Illumina protocol ("Preparing 2–5kb Samples for Mate Pair Library Sequencing", Part # 1005363 Rev. B February 2009, page 25 ff). Samples were sequenced on an Illumina GAIIx using Illumina's paired-end cluster generation and cycle sequencing kits, following the recipes for 2x100 and 2x150 cycles, respectively. Sequences were extracted by the Genome Analysis Pipeline CASAVA. Duplicon fractions were estimated by an in-house perl program. Sequencing statistics are provided in Table S5.

Table S5: Summary of barley WGS sequencing data

DNA	library type	size	# of libraries	read length	pairs	singlets	Gb	approx. genome coverage
Morex [A]	PE	500 bp	3	2x100 bp	7,31E+08		147,6	29,5
	MP	2500 bp	2	2x100 bp	2,15E+07	2,42E+07	6,4	1,3
			6	2x150 bp	2,78E+08	3,81E+08	122,8	24,6
Σ			11		1,03E+009	4,05E+008	276,8	55,4
Barke [A]	PE	500 bp	1	2x100 bp	3,34E+07		6,8	1,4
				2x150 bp	1,28E+08		38,8	7,8
			2	2x150 bp	3,47E+08		104,8	21,0
Σ			3		5,09E+008		150,3	30,1
Bowman (FLI) [A]	PE	500 bp	1	2x100	2,12E+08		42,8	8,6
Bowman (JHI) [A]	PE	500 bp	1	2x100	6,76E+08		136,5	27,3
Σ			2		8,88E+08		179,3	35,9
Spontaneum [B]	PE	500 bp	1	2x100	1,85E+08		37,3	7,5
Igri [B]	PE	500 bp	1	2x100	1,81E+08		36,5	7,3
HarunaNijo [C]	Single End		1	1x500		6,57E+07	32,9	6,6

PE = paired end, MP = mate pair, [A] = Illumina GAIIx, [B] = Illumina Hiseq 2000, [C] = Roche FLX Titanium

Roche 454 whole genome shotgun sequencing: Genomic DNA of cv. Haruna Nijo was isolated from leaf tissue using Qiagen DNeasy Plant mini kit (QIAGEN). 3-5 µg of genomic DNA was used for shotgun library development with GS Titanium General Library Preparation Kit (Roche Applied Science) according to the manufacturer's protocol (Roche Applied Science). Three shotgun libraries with average fragment sizes range of 600-960 bp were sequenced using a Roche 454 genome sequencer FLX on Titanium picotiter-plate.

De novo assembly of GAIIx whole genome shotgun sequencing data: Read pair distance for libraries was determined by mapping pairs against the chloroplast DNA (cpDNA) sequence of barley (NC_008590) using BWA[10] (Figure S1). Read distance was extracted in case both reads mapped to the cpDNA. Minimum and maximum distance was first quartile minus 1.5-times inter-quartile-range (IQR) and third quartile plus 1.5 times IQR, respectively. Since BWA expects read pairs to be in a "forward-backward" position for mate-pair libraries the reverse complement had to be built as additional step. Sequences were quality trimmed and *de novo* assembled using CLC Assembly Cell 3.2.2 (http://www.clcbio.com/). Independent assemblies were generated for WGS datasets of cultivars Morex, Bowman and Barke, respectively (Table S6).

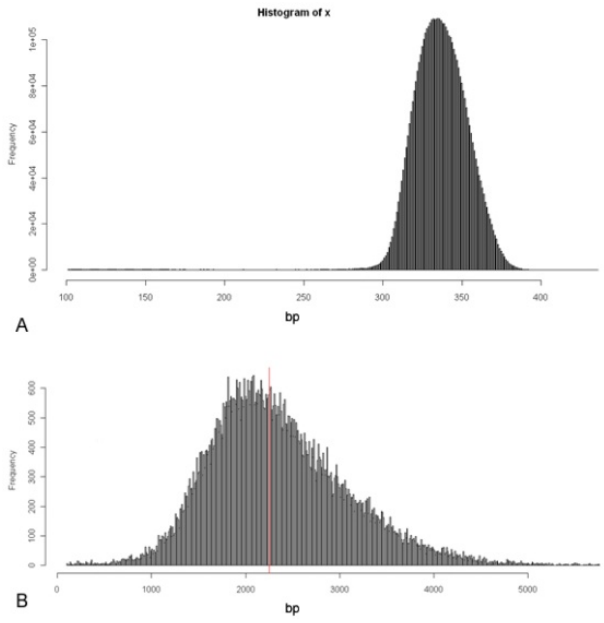


Figure S1: Fragment size distribution of a PE and MP shotgun sequencing libraries of barley.

Paired-end (A) and mate-pair reads (B) were mapped to the published sequence of barley chloroplast DNA[11]. Average sizes of just below 350 and 2,500 bp were obtained. The red line shows the median of observed distances

Table S6: statistics of whole genome shotgun sequence assembly

Assembly #	Morex	Bowman	Barke
# sequences	2,670,738	2,077,901	2,742,077
# bp	1,868,648,155	1,779,486,241	2,019,369,188
largest contig (bp)	36,084	37,442	38,386
# contigs > 10k	8,319	9,265	4,290
# contigs > 1k	376,261	398,468	471,808
N50 size (bp)	1,425	1,986	1,419
N50 number	264,958	204,101	323,608

Assessment of WGS assembly quality: To evaluate the completeness of genes in the assemblies, all contigs were compared to publicly available barley full length cDNAs[12] (fl-cDNA) by Blast resulting in local alignments for each fl-cDNA. For Morex, 82% of all fl-cDNAs (23449 out of 28,266) were covered by more than 95% (98% identity). For Bowman and Barke, the fractions were lower (71% and 26%, respectively) due to lower sequencing depth (36x and <20x, respectively) (Figure S2). The level of representation of the barley gene space within the whole genome shotgun (WGS) sequence assemblies was furthermore determined on the basis of a comparison to all 454 sequenced and assembled gene-bearing BAC clones as provided in this study. BAC assemblies were masked against repetitive DNA (see Online Methods). The remaining un-masked sequences larger than 200 bp were aligned to the whole genome shotgun assembly of Morex. Of 69,702 contig-fragments, 55,478 (79%) were 100% covered by local alignments, 65,012 (93%) are covered by 95% or more, respectively. 98.92% of the combined length of repeat-masked contig-fragments (92,253,103 bp) was covered by local alignments (Figure S3).

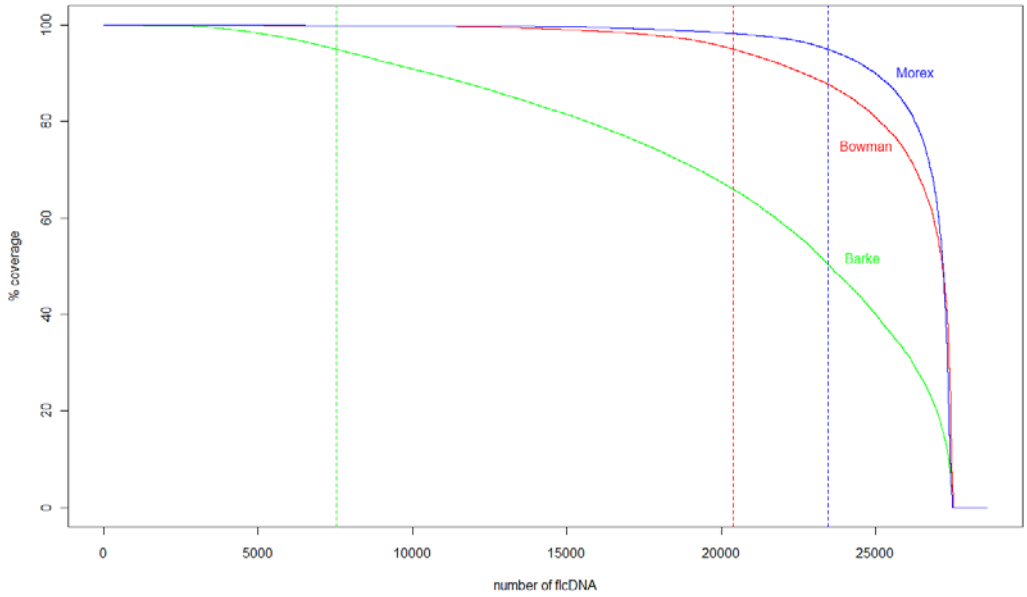


Figure S2: **Coverage of barley full length cDNAs (fl-cDNA) as determined by local alignments to WGS contigs.** Barley fl-cDNAs[12] were aligned to the assembled WGS data of cultivars Morex, Barke and Bowman. The percentage of bases that are included in any local alignment (identity >= 95%) are displayed. X-Axis shows fl-cDNAs ordered by coverage. Y-Axis represents the percentage coverage by local alignments. Dashed vertical lines mark the number of fl-cDNAs that are represented in the WGS-assemblies by more than 95%.

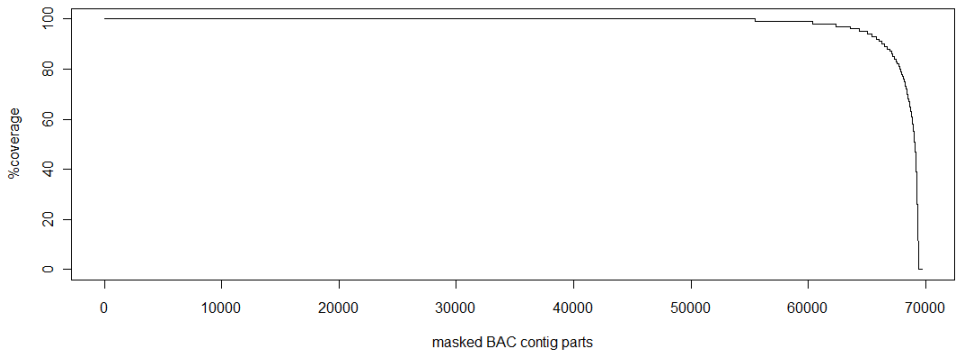


Figure S3: Coverage of repeat-masked BAC-fragments with WGS contig sequences. Repeat-masked contig fragments larger than 200 bp were aligned to the assembled WGS sequences of cultivar 'Morex'. X-axis shows the BAC-contig-fragments ordered by % coverage.

Supplemental Note 4. Integration of genetic / physical map and sequence resources

S4.1 The strategy of genomic integration

The physical map of barley obtained after HICF comprised 9,265 FPC contigs. To unlock the full potential of this genome-wide map we intended to anchor as many as possible of these contigs to a genetic map framework. We achieved this in a multi-stepped and multi-tiered stratification approach (Figure S4) that took full advantage of the genomic sequence resources generated within this study but also by the wealth of barley genetic maps that have been published before. The strategy involves six major steps:

Step1: Association of genomic sequence information (supplemental note 3) to the FPC map: MDR/kmer-masked[13] BAC end sequences as well as shotgun sequenced BAC clones provided anchor points for integrating WGS *de novo* assemblies from the cultivars 'Morex', 'Bowman' and 'Barke'.

Step2: Integration of genetic marker information: Genetic markers were either assigned directly to BAC clones experimentally by screening of multidimensional pools of BAC libraries (online methods) or by stringent sequence similarity searches against the integrated sequence datasets.

Step3: Integration of genetic marker info from published maps which could not be assigned to FPC contigs: Genetic markers were compared to the WGS assemblies delivering additional sequence tags that could be associated to the genetic/physical framework established in steps 1 and 2.

Step 4: Interpolation of genetic maps to build a genetic backbone: Genetic maps in barley were developed in a wide variety of experimental populations. In order to exploit the full potential the genetic maps had to be interpolated on the basis of commonly shared marker/gene information and by considering physical / genetic linkage as provided by the presented FPC map. This genetic marker backbone was used to anchor the FPC map and sequence data.

Step 5: Exploiting gene information not represented in genetic maps of barley: Genes represented in barley fl-cDNA resources[12] or annotated from WGS assemblies (supplemental note 7) may not be represented in genetic maps of barley. We used chromosome-specific sequence information[14] (Chromosome arm assignment, CarmA) to assign such genes / WGS contigs into chromosome-arm bins.

Step 6: Syntenic stratification: We used the result of step 4 and compared the physical/genetic order of the integrated genes to sequenced grass genomes. This provided us with high confidence conserved syntenic genome blocks. These were used to position CarmA processed genes/markers from step 5.

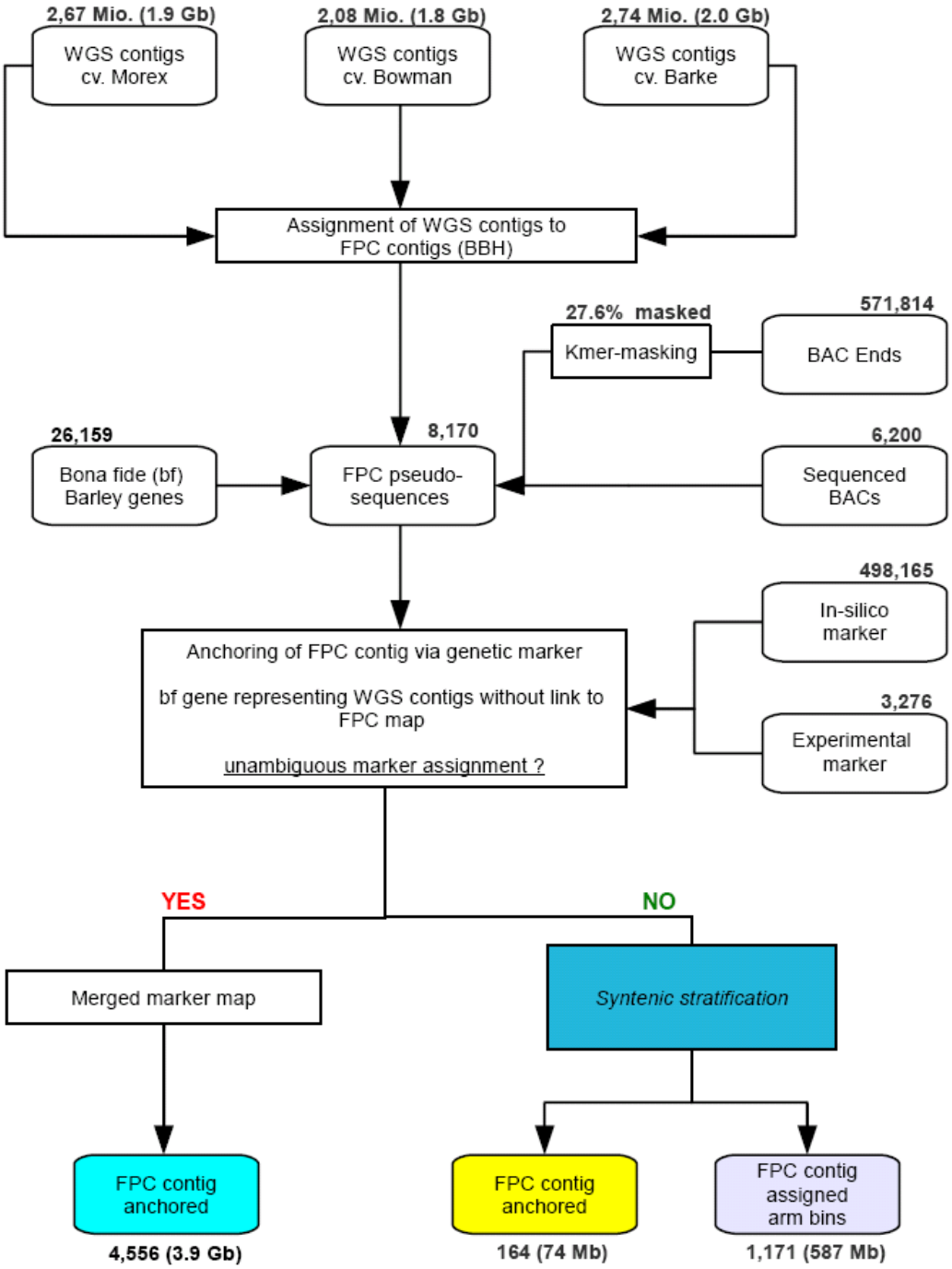


Figure S4: Schematic workflow for layering and integration of barley genomic data

FPC contigs were associated with sequence information from whole genome shotgun sequence assemblies via information and sequence anchors provided by BAC end sequences and shotgun sequenced BAC clones. FPC contigs that could not be anchored to the genetic framework directly by this approach were subjected to a gene anchoring pipeline utilizing sorted chromosome arm sequence information (CarmA) and syntenic stratification.

S4.2 Integration of genomic sequence resources

9,265 FPC contigs with an average length of 538 kb each, that include 517,202 BACs (6,295 BACs sequence = 766.4 Mb) and 571,814 sequenced BAC ends (BES=373.5 Mb) were combined to form the barley genome scaffold. Whole genome shotgun (WGS) assemblies (WGS contigs) from three different barley cultivars ('Morex', 'Bowman', 'Barke') were associated via sequence homology (VMatch; www.vmatch.de) to BES or sequenced BACs [minimum hit length 200, 3 mismatches allowed, best bidirectional hit (BBH)]. This provided an additional 307.5, 313.4 and 268.8 megabases of FPC associated genomic sequence information for the three barley cultivars Morex, Bowman and Barke, respectively (Figure S5, Table S7). To illustrate the extent of sequence resources integrated and associated to the FPC contigs, examples from three known gene loci and three randomly chosen unknown loci, represented by FPC contigs of similar size, are given in Table S8.

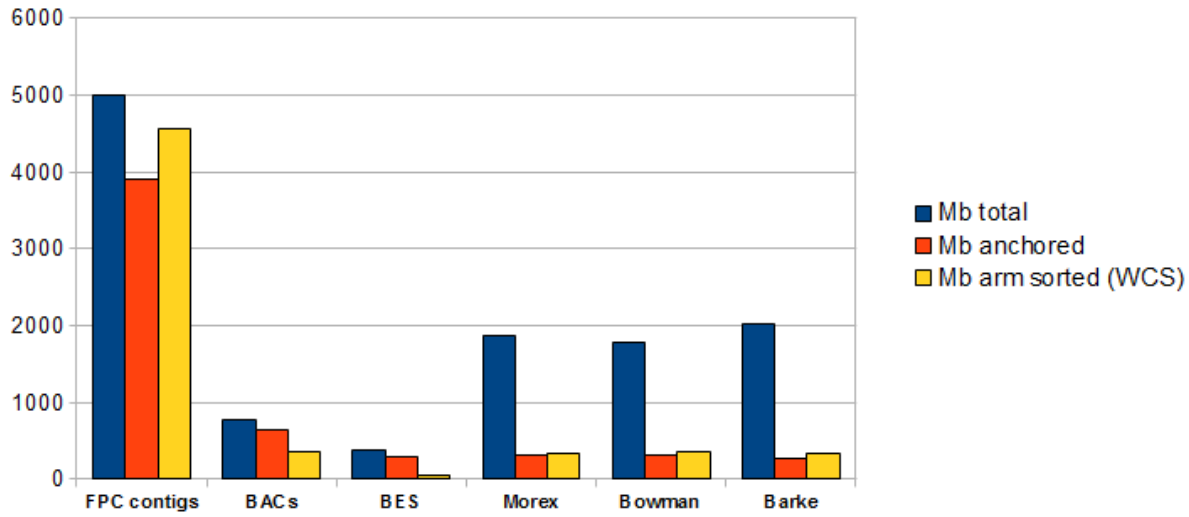


Figure S5: **Genomic sequence resources of barley associated to the FPC map**. Different genomic sequence data resources were used for integration into the genomic framework. The figure illustrates the cumulative length of the individual resources that was available initially (blue bars), that could be assigned to chromosome arm / chromosome bins (yellow bars) by the CarmA approach (S4.4), or that could be integrated into the physical/genetic map scaffold (red bars). The y-axis provides the Mb scale for the different sequence resources, however, in case of the FPC contigs the scale indicates the length spanned by physical map contigs. (BES: BAC end sequences).

Table S7: Summary of WGS contigs integrated to the barley physical/genetic framework on the basis of FPC contigs (homology to BES and sequenced BACs)

Chromosome	anchored WGS contigs per assembly (number)			anchored WGS contigs per assembly (Mbp)		
	Morex	Bowman	Barke	Morex	Bowman	Barke
1H	13,065	12,096	12,135	36	35	31
2H	18,330	17,155	16,911	51	51	44
3H	16,197	15,301	15,163	45	46	40
4H	16,223	15,610	15,685	43	46	39
5H	16,188	14,810	15,043	45	45	39
6H	15,087	14,191	13,741	39	40	34
7H	17,896	16,959	16,274	48	49	41
Σ	112,986	106,122	104,952	307	312	268

Table S8: Example for the status of integrating sequence resources into the physical/genetic barley genome framework

	known gene locus			unknown random locus		
	Rym4[15]	Ppd-H1[16]	Int-c[17]			
FPC contig	contig_44421	contig_2992	contig_621	contig_48090	contig_749	contig_38578
size of FPC contig (kb)	794,8	1324,3	3176,9	758,9	1335,5	3209,1
chromosome arm	3HL	2HS	4HS	5HL	6HL	1H
genetic position (cM)	148,4	22,2	22,2	53,4	93,1	106,2
physical position (Mb)	556,8	24,8	16,4	395,9	509,7	439,8
# anchored WGS contigs (cv. Morex)	24	62	139	16	40	116
cumulative length anchored Morex WGS contigs (kb)	24.4 (3.1 %)	206.3 (15.6%)	365.3 (11.5 %)	21.4 (2.8%)	131.2 (9.8%)	473.8 (14.8%)
# anchored WGS contigs (cv. Bowman)	22	60	106	13	40	124
cumulative length anchored Bowman WGS contigs (kb)	24.8 (3.1 %)	195.6 (14.8%)	348.2 (11.0 %)	24.8 (3.3 %)	108.7 (8.1%)	421.9 (13.1%)
# anchored WGS contigs (cv. Barke)	18	57	128	16	28	113
cumulative length anchored Barke WGS contigs (kb)	17.0 (2.1%)	159.3 (12.0%)	320.8 (10.1%)	37.4 (4.9%)	76.7 (5.7%)	292.7 (9.1%)
# sequenced BACs	2	6	10	1	3	8
cumulative length sequenced BACs (kb)	147.4 (18.5%)	817.6 (61.7%)	1162 (36.6%)	99.6 (13.1%)	357.3 (26.8%)	935.8 (29.2%)
# BES	58	164	291	54	125	322
cumulative length BES (kb)	38.9 (4.9%)	106.7 (8.1 %)	179.5 (5.7%)	33.0 (4.3%)	75.9 (5.7%)	205.6 (6.4%)
anchored HC genes	1	15	20	1	7	34

FPC = fingerprinted contig, WGS = whole genome shotgun, BAC = bacterial artificial chromosome, BES = BAC end sequence, HC = high confidence

S4.3 Genetic map resources for anchoring of the physical map

Experimental (direct) anchoring Several experimental populations and genetic maps have been published for barley that provide large numbers of markers for experimental anchoring of physical map contigs to a genetic scaffold. 3,276 markers from six maps were experimentally anchored to BAC clones of the physical map by PCR[18], microarray[19]or Illumina Golden Gate Oligo Pool Assay-based[6,20] screening of multidimensional BAC DNA pools (Table S9).

Table S9: Summary of experimentally anchored genetic markers

Genetic map code	Marker type/screening procedure	No. of marker assigned to FPC contigs
MM1[21]	TDM/microarray	230
MM2[14]	STS / PCR, MD BAC pools	297
MM3[22]	STS / PCR, MD BAC pools	1736
MM4[23]	STS / PCR, MD BAC pools	145
MM5[20]	BOPA / BAC pools	596
MM6[24]	TDM / microarray	342
Σ		3,276

The PCR-based screening of multi-dimensional BAC DNA pools can lead to false positive anchoring information due to cross contamination of DNA pools. We assessed the rate of false associations between experimentally screened markers and respective BAC addresses. From all 4,556 FPC contigs that were finally integrated into the physical/genetic barley genome framework a total of 1,521 non-redundant contigs carried information of at least a single experimentally anchored marker. A false positive rate of 2 – 8 % was determined depending on the screening approach and the marker resource (Table S10). Such false positive marker/BAC relationships were not considered for any further downstream analysis.

Table S10: Determination of consistency of experimental marker anchoring
Consistency check has been undertaken in context to overall anchoring information provided for FPC contigs integrated to the physical/genetic barley genome framework

Genetic map code	total # FPC contigs with experimental marker info[1]	total # integrated FPC contigs with experimental marker info	ratio of integrated / total # of FPC contigs with experimental marker info	# FPC contigs with consistent experimental anchoring info	# FPC contigs with contradicting experimental anchoring info	Ratio of FPC contigs with consistent exp. anchoring / total # integrated FPC contigs with experimental marker info
MM1	207	199	0.96	184	15	0.92
MM2	165	151	0.92	145	6	0.96
MM3	1119	1036	0.93	985	51	0.95
MM4	154	146	0.95	137	9	0.94
MM5	473	445	0.94	439	6	0.99
MM6	309	296	0.96	276	20	0.93

[1]= contigs carrying the number of markers as specified in Table S9

In silico (indirect) anchoring

A larger dataset of genetic markers was used for sequence-based anchoring. This included all markers from the same six maps used for experimental anchoring plus additional maps (Table S11). The largest number of markers was provided by maps developed by genotyping-by-sequencing (GBS)[25]. This included previously published 34,396 SNPs and 241,159 dominant tag markers on the Barley Oregon Wolfe DH population[26]. We used the same approach here to genotype the Morex x Barke DH population[20] and added 21,384 SNPs and 184,796 dominant tags to this genetic map. Raw data for the Morex x Barke DH GBS is submitted under NBCI SRA study # SRP010876.1. The SNP mapping data can be obtained through GrainGenes (*wheat.pw.usda.gov*). Altogether, 498,165 in-silico marker sequences and 3,276 experimental markers were combined. These were related to 8,170 (~4.9 Gb) FPC contigs using sequence homology criteria (hit length >=55, at maximum 1 error, best scoring hit).

Table S11: Summary of in silico genetic markers
Experimental marker maps and in silico marker maps referencing the same genetic map where indicated via a common digit [e.g. SM4 and MM4(Table S9)]. Experimental maps are indicated with the letter 'M', in silico maps with the letter 'S'.

Genetic map code	Marker type	Average marker length (bp)	# of marker sequences in map	length of genetic map (cM)	# marker assigned to anchored FPC contig
SM2[14]	WCS read	489	622	NA	237
SM3[22]	EST	519	5,780	2,092	944
SM4[23]	EST	534	1,052	1,111	144
SM5[27]	EST/Golden Gate	203	2,994	1,335	1,036
SM6[28]	EST/iSelect	141	5,481	991	2,643
SM7[26]	GBS/SNP	63	241,159	1,001	48,708
SM8[26]	GBS/PA	63	34,396	1,007	8,477
SM9	GBS/SNP	60	184,796	986	44,166
SM10	GBS/PA	61	21,384	1,001	6,524
SM11[29]	EST assembly	1,401	501	1,248	238
Σ			498,165		113,117

WCS=454 shotgun read obtained from sorted chromosomes , EST= expressed sequence tag, PA= presence/absence polymorphism

Interpolation of genetic maps to build a genetic marker backbone

To create a consensus map, the SM6 map[28] was taken as the backbone map. The SM6 map was considered as most suitable as it provides the highest marker density and resolution (*N*=360, RIL/F8) for a single bi-parental mapping population. Overlaps between the different

marker maps have been identified based on matching sequence associations on FPCs. Joints have been used to align marker maps and intercalate non-FPC anchored markers. In summary, 113,117 contig-associated *in* silico markers and all 3,276 experimentally anchored markers were integrated into this genetic map backbone. Experimental markers were assigned to physical map contigs using Vmatch (www.vmatch.de, hit length >= 55 bp; max. one error; best single hit). For anchoring, only physical map contigs were considered with an unambigious CarmA assignment (supplemental note S4.4) which was implemented to correct for putatively false marker assignments. For every respective genetic marker map the median overall cM position was taken and resulting values were interpolated to a cM position of map SM6 using the R method 'approxfun' (http://www.R-project.org). Comparisons of structure and consistency between the different marker maps and the stratified and combined marker map demonstrate the robustness of the stratification of different marker maps (Figure S6). In general very pronounced colinearity between the marker maps was observed with only few marker datapoints offset from the diagonal. Using the dense genetic marker scaffold we were able to anchor and position 4,556 physical map contigs with a cumulative length of 3.9 Gb (Table S12). Sequence homology to the utilized markers allowed direct association of 25,457 (103 Mb), 27,740 (115 Mb) and 27,963 contigs (89 Mb) of the 'Morex', 'Bowman' and 'Barke' WGS assemblies, respectively. Thus approximately one-third of the entire length of all anchored WGS contigs could be directly integrated into the physical/genetic genome framework on the basis of the integrated marker sequence information (Table S13).

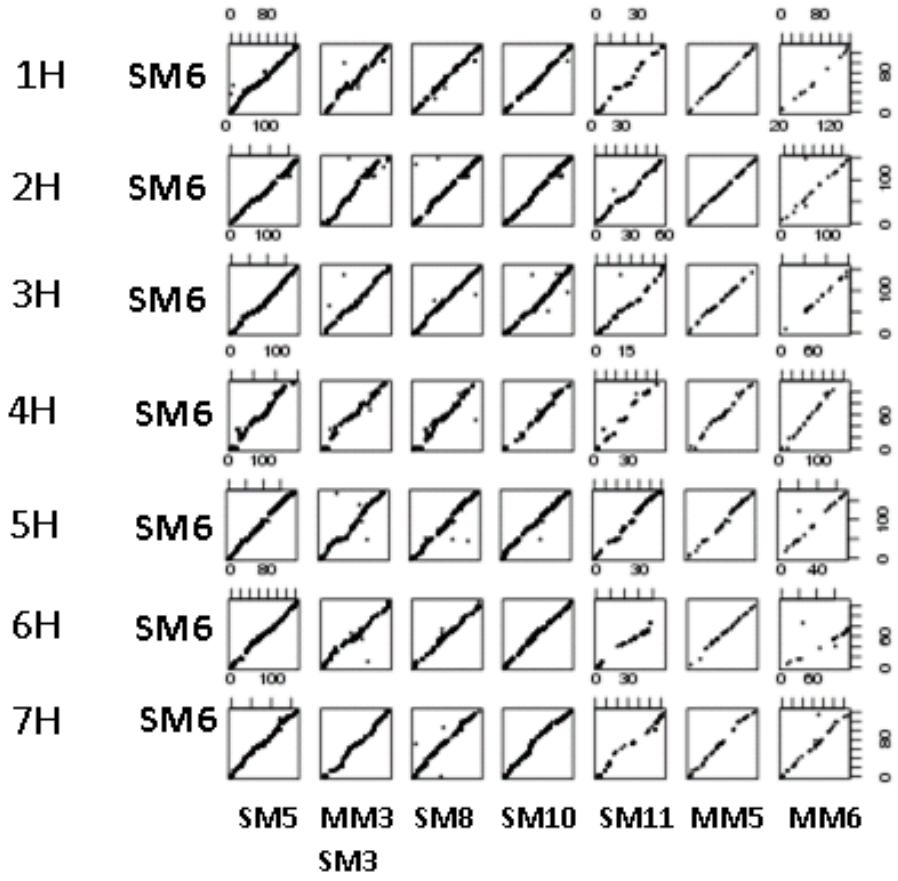


Figure S6: Comparison of selected marker maps against the reference map.

Comparison and visualization of physical map contigs that contain both markers from the backbone map (SM6) and from the respective experimental/*in silico* maps (SM5, MM3/SM3, SM8, SM10, SM11, MM5, MM6). This enables to estimate the performance of contig anchoring as both markers should share comparable relative positions within the individual maps. As shown, only few of the anchored physical map contigs are offset from the expected diagonal which indicates the colinearity of the maps.

Table S12: Summary statistics of genetically anchored FPC contigs

Chromosome	No. of FPC contigs	Cumulative FPC contig length (Mb)	Estimated length of chromosome (Mb)	FPC map length / estimated length of chromosome (%)
1H	560	464	622	75%
2H	715	628	790	79%
3H	668	565	755	75%
4H	596	543	729	74%
5H	688	560	760	74%
6H	644	539	689	78%
7H	685	602	755	80%
Σ	4,556	3,900	5,100	76%

Table S13: Summary of WGS contigs integrated to the physical/genetic genome framework on the basis of molecular markers

Chromosome	anchored WGS contigs per assembly (number)			anchored WGS contigs per assembly (Mbp)		
	Morex	Bowman	Barke	Morex	Bowman	Barke
1H	3,137	3,123	2,988	12	13	9
2H	3,984	4,464	4,487	17	20	15
3H	4,139	4,612	4,754	17	19	15
4H	2,474	2,801	2,758	10	12	9
5H	4,008	4,327	4,471	17	18	14
6H	3,678	4,074	3,807	14	16	12
7H	4,037	4,339	4,698	16	17	15
Σ	25,457	27,740	27,963	103	115	89

The diverse genetic maps utilized for anchoring contributed to different extent to the anchoring of FPC contigs. By far the largest number of FPC contigs (4,378 of 4,556, 96%) were anchored on the basis of GBS-based maps (SM7 – 10) (Table S14). These markers were represented by 64bp long sequence tags bearing the risk of false positive anchoring based on multi-copy sequences in the genome of barley. Such limitations were compensated for by highly redundant anchoring information from multiple genetic maps and marker resources. Apart from the GBS maps, the MM3/SM3, SM5 and SM6 maps are the major contributors for the anchoring of FPC contigs. They account for 26 and 16 percent of all anchored contigs, respectively (Table S14). For 4,387 FPC contigs at least one GBS marker is available demonstrating the impact of the GBS maps. By comparison, SM6 allowed 1201 FPC contigs to be anchored.

Table S14: Number of FPC contigs associated to different marker maps

Map	# of FPC contigs tagged	Tagged FPC contigs / Σ of tagged FPC contigs (%)
MM1	199	4
MM2+SM2	231	5
MM3+SM3	1,167	26
MM4+SM4	243	5
MM5	445	10
MM6	296	6
SM5	718	16
SM6	1,201	26
SM7	3,983	87
SM8	2,241	49
SM9	3,038	67
SM10	1,540	34
SM7-10	4,378	96
SM11	200	4
Σ	4,556	100

S4.4 Chromosome arm assignment (CarmA) of genomic sequences

Whole chromosome arm shotgun sequences (WCS) for individual chromosomes and chromosome arms were generated previously[14]. We used this as reference index to assign WGS contigs and FPC contigs (on the basis of associated sequence resources) to a specific chromosome arm. This allowed us to assign FPC contigs to chromosome arms even in the absence of associated genetic marker information. Furthermore, this provided an independent option for cross validation of rare cases of contradicting map assignments obtained from genetic markers originating from different regions in the barley genome. Sequence comparisons of WGS sequence contigs and other FPC contig-associated genomic sequences were performed against WCS data (minimal sequence hit length of >= 100 bp, at least 99% identity to kmer-masked WCS data). The sum of matching basepairs had to exceed potential crossmatching chromosome arm sequence information by a factor of 1.2 to serve as robust chromosome arm binning information. This approach enabled us to assign 6,437 FPC contigs (4561 Mb) to chromosome/arm bins. Thus 1,881 additional, previously non-anchored FPC contigs (~0.66 Gbp) were assigned to individual chromosome arms. Including this strategy increased the percentage of FPC contigs that could at least be assigned to any chromosome arm to more than 91%. The CarmA approach was also applied to assign annotated high confidence (HC) barley genes (supplemental note 7) and EST assemblies into chromosomes of barley (Table S15).

Table S15: High confidence (HC) barley genes assigned to barley chromosome arms

Chromosome	# of HC genes assigned to chromosomes based on CarmA	HarvEST35[30] unigenes assigned to chromosomes
1H	3,111	4,386
2H	3,562	4,626
3H	3,489	4,498
4H	2,589	3,471
5H	3,480	4,436
6H	2,538	3,268
7H	3,365	4,263
Σ	22,134	28,948

S4.5 Syntenic stratification of genes into the barley genome framework

Not all of the high confidence transcribed genes annotated from the whole genome shotgun sequence assembly of cultivar Morex could be associated to the sequence enriched FPC map directly. If such genes were represented by genetic markers from any of the maps used for anchoring it was still possible to assign them into the genetic/physical genome framework. For a substantial portion of the non-integrated genes (4,692) the CarmA assignment provided chromosome arm bin information. Grass genomes are highly related and share regions of extended and highly conserved synteny. It was previously shown that synteny between barley and sequenced model grass genomes can be exploited to determine a structured and ordered gene index for a genome[14]. We compared the genetically anchored and physically ordered marker (genes) backbone (supplemental note S4.3) to the reference genomes of *Oryza sativa, Sorghum bicolor* and *Brachypodium distachyon* (hit length 100, similarity >=75 Percent, best bidirectional hit (BBH)) to test integrity of the marker scaffold and to identify conserved syntenic segments between barley and the respective reference genomes. 15,719 HC genes associated to a FPC contig or WGS contigs were used to define syntenic segments between Barley and *Brachypodium*, rice and *Sorghum* by plotting orthologus gene pairs (BBH, similarity >=75 %, >=hit length 30aa) between the integrated barley contig map and the reference genomes of *Brachypodium*, rice and *Sorghum* (Figures S7-S9). The observed extensive syntenic blocks confirm previous marker map-based results[14,20,31] and illustrate the robustness of the genetic stratification strategy. We exploited this information to complement our genetic stratification strategy for placing further barley genes into the genome framework by a synteny driven approach ('syntenic stratification') (Figure S10). Barley HC genes that had chromosome arm assignments and represented orthologs of model grass genes located in the respective corresponding synteny blocks thus could be integrated into the physical / genetic context. An additional 3,743 genes were positioned based on these criteria. In summary, sixty percent of the HC genes were anchored to a genetic position, an additional 14% were assigned via 'syntenic stratification'

and a further 18% were -assigned to chromosomes/chromosome arms using CarmA. In summary 92% of all HC genes were assigned a chromosomal position (Table S16).

Table S16: Anchoring of barley HC genes using a marker centric approach, syntenic stratification and CarmA

Anchoring strategy	positioned HC genes	HC barley genes total	[%]
Genetic stratification	15,719	26,159	60.09%
Syntenic stratification	3,743	26,159	14.31%
CarmA	4,692	26,159	17.94%
Σ	24,154	26,159	92.34%

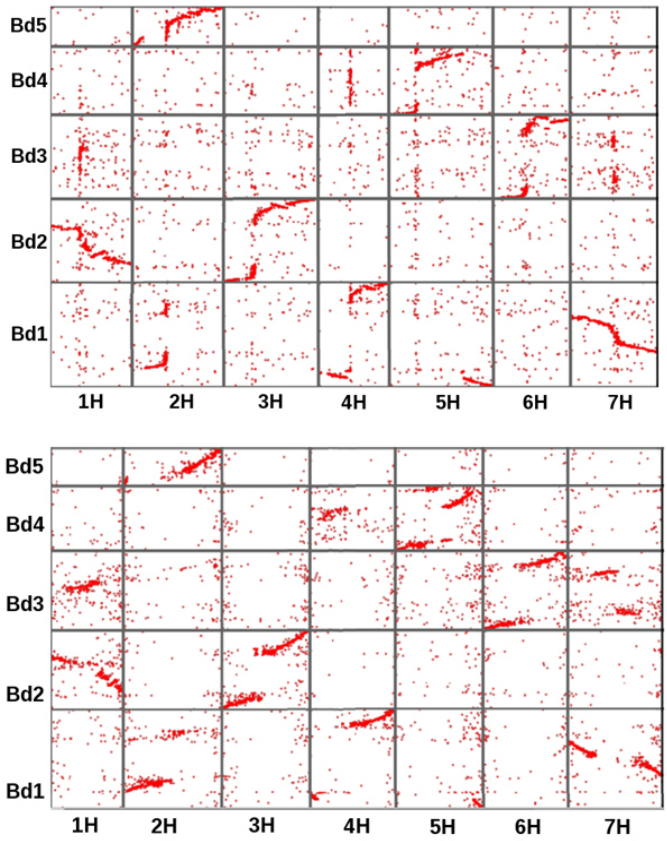


Figure S7: Comparison of the barley physical/genetic genome framework to the genome of *Brachypodium distachyon*

Upper panel compares the genetic scale (cM) and the lower panel the physical scale (Mbp) of the barley genome. Only HC genes of barley with best bidirectional blast hits (tblastx) against *Brachypodium* genes at a cutoff of at least 75% identity over at least 100bp genic sequence were used.

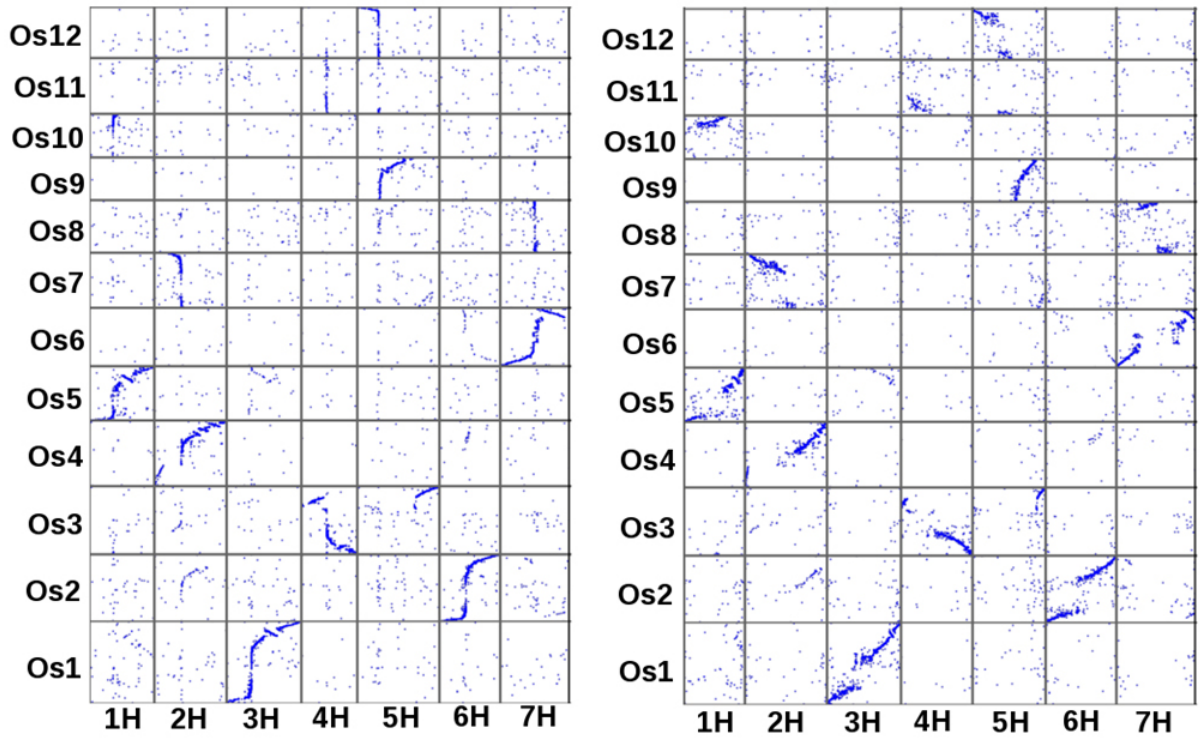


Figure S8: Comparison of the barley physical/genetic genome framework to the genome of *Oryza sativa*

Left panel compares the genetic scale (cM) and the right panel the physical scale (Mbp) of the barley genome. HC genes of barley with best bidirectional blast hits (tblastx) against rice genes at a cutoff of at least 75% identity over at least 100bp genic sequence were are shown.

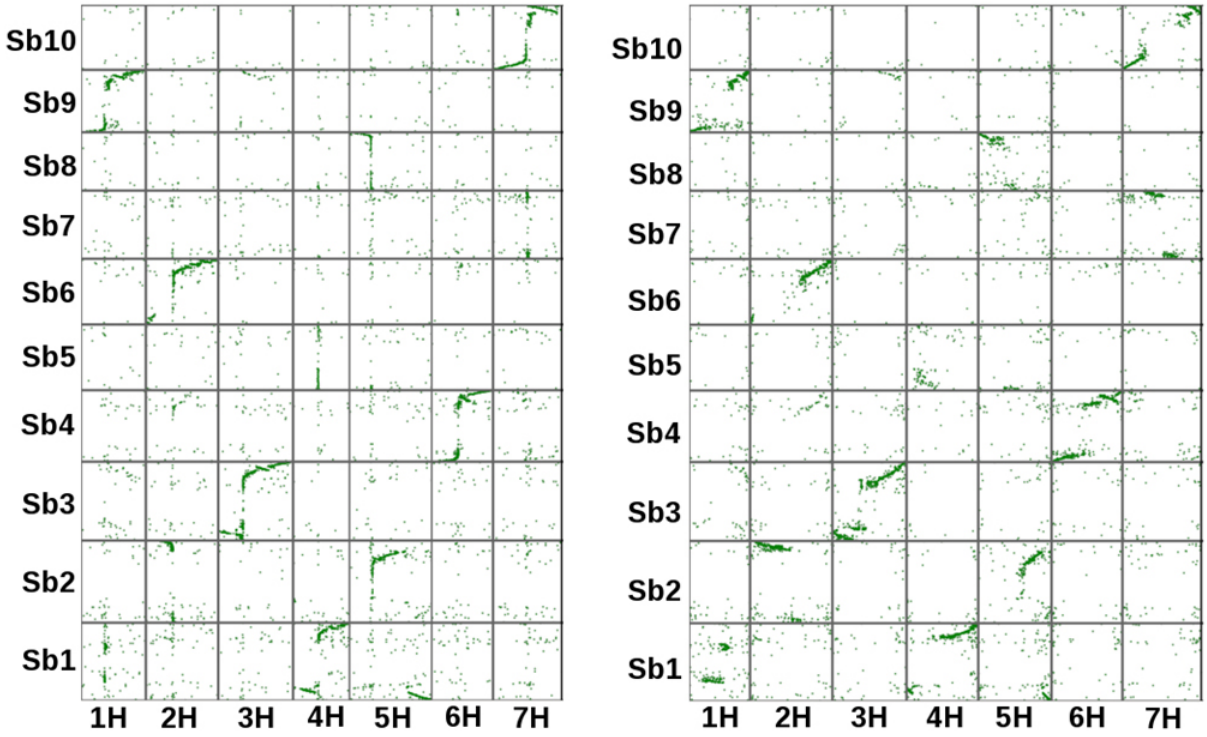


Figure S9: Comparison of the barley physical/genetic genome framework to the genome of *Sorghum bicolor*.

Left panel compares the genetic scale (cM) and the right panel the physical scale (Mbp) of the barley genome. HC genes of barley with best bidirectional blast hits (tblastx) against *Sorghum* genes at a cutoff of at least 75% identity over at least 100bp genic sequence are shown.

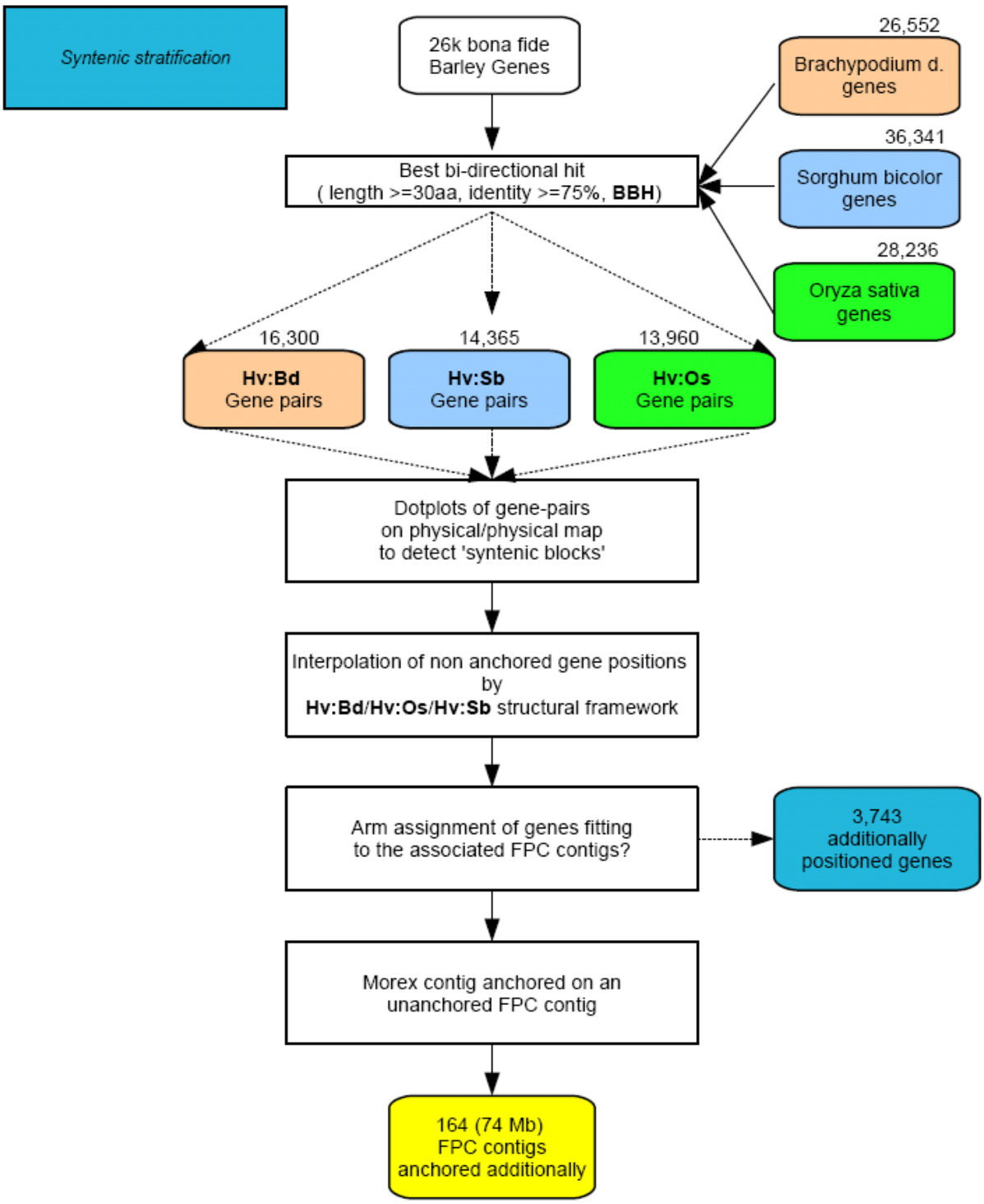


Figure S10: Strategy of syntenic stratification for anchoring genes and FPC contigs into the barley physical / genetic genome framework

S4.6 Relationship of genetic and physical distance; gene distribution along barley chromosomes

We used all HC genes (Supplemental Note S7) that were integrated into the stratified genome framework to analyse gene distribution along the seven barley chromosomes (Table S17, Figure S11). A pronounced correlation of genetic resolution and physical distance that is significantly higher in barley telomeres than in pericentric and centromeric regions was observed. Approximately ~1.9 Gb (49 %) of all anchored FPC contigs are located in chromosomal regions with highly reduced recombination frequencies. These regions contain as many as 3,474 (22%) HC genes.

Table S17: Barley genes located in centromeric and peri-centromeric regions of the barley physical/genetic genome framework The number of HC genes that have been anchored to the genetically defined centromeric and pericentromeric regions (+/- 5cM from genetically defined centromere) is given.

Barley chromosome	No. of HC genes within +/- 5cM of genetic centromere	No. HC genes positioned by genetic stratification[Table S14] only
1H	529	2,082
2H	591	2,664
3H	459	2,478
4H	463	1,749
5H	506	2,553
6H	431	1,868
7H	495	2,325
Σ	**3,474**	**15,719**

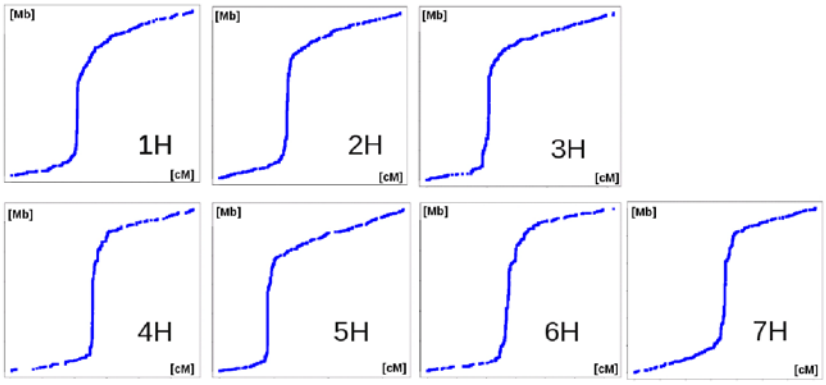


Figure S11: Correlation of genetic (cM) and physical distance (Mb) along barley chromosomes.

Genetic distances were plotted against physical distances as derived from anchored and positioned FPC contig sizes. Regions with low recombination expand to approximately 1.9 Gb (49%) of the barley genome in the centromeric and pericentric regions of all seven barley chromosomes.

Supplemental Note 5. Analysis of repetitive DNA of the barley genome

S5.1 Repeat detection and analysis: Content and composition of barley repetitive DNA was analyzed in the various sequence datasets generated and utilized in this study (BES, WGS reads and assemblies, BAC sequences) by comparing against the MIPS Repeat library mipsREdat_v9.0_Poaceae which has been enriched for barley specific repetitive DNA sequences (Table S18). The relative frequency of different repeat classes was compared between different datasets indicating an underrepresentation of highly repetitive sequences in the WGS assemblies and a slight overrepresentation of class II mobile elements in 'gene-bearing' BAC clones (Table S19). For measurements and visualization of genome-wide distribution of repetitive DNA elements along the genetically anchored physical map only BES and FPC contig associated sequence resources were considered. Presence and classification of repetitive DNA (class I or II) was associated with the respective FPC contigs and plotted for the respective chromosomes.

Table S18: Annotation summary of repetitive of barley genomic DNA resources

Different types of repetitive elements are listed and the percentual contribution of the respective element types is given for the individual categories and sequence collections

% of all bp	WGS reads subset of 850 Mb	BES 374 Mb	random BACs 63Mb	gene bearing BACs 379 Mb	Bowman WGS assembly 1.8 Gb	Morex WGS assembly 1.9Gb	Barke WGS assembly 2.0 Gb
MDR: unique/low copy (20mer <10x)	30,63	29,34	30,17	34,00	57,78	56,91	50,19
MDR: medium/high copy (>=10 to <1000x)	33,88	37,15	36,14	33,61	33,27	31,36	35,40
MDR: very high copy (20mer >= 1000x)	35,49	33,51	33,69	32,38	8,95	11,73	14,41
Mobile Element	**81,58**	**81,80**	**82,10**	**74,05**	**61,04**	**58,89**	**58,45**
Class I: Retroelement (RXX)	**75,33**	**76,91**	**75,47**	**67,06**	**54,99**	**52,83**	**52,72**
LTR Retrotransposon (RLX)	75,10	76,62	75,15	66,44	54,41	52,25	52,19
Copia (RLC)	15,32	12,50	13,66	14,50	8,73	8,48	8,46
Gypsy (RLG)	22,29	21,78	20,84	16,99	19,09	18,16	17,96
Gypsy/Copia ratio	1,45	1,74	1,53	1,17	2,19	2,14	2,12
unclassified LTR	37,49	42,35	40,66	34,95	26,58	25,61	25,77
non-LTR Retrotransposon	0,22	0,29	0,31	0,63	0,58	0,58	0,53
Class II: DNA Transposon (DXX)	**5,60**	**4,59**	**6,21**	**6,36**	**5,22**	**5,25**	**5,00**
Retro-TE/DNA-TE ratio	13,44	16,77	12,16	10,54	10,54	10,07	10,55
DNA Transposon Superfamily (DTX)	5,39	4,34	5,91	5,78	4,62	4,63	4,45
CACTA superfamily (DTC)	5,16	4,06	5,53	5,17	4,00	4,01	3,89
hAT superfamily (DTA)	0,02	0,01	0,03	0,03	0,02	0,02	0,02
Mutator superfamily (DTM)	0,12	0,13	0,21	0,29	0,27	0,28	0,26
Tc1/Mariner superfamily (DTT)	0,02	0,03	0,03	0,06	0,06	0,06	0,06
PIF/Harbinger (DTH)	0,06	0,09	0,09	0,18	0,21	0,20	0,17
unclassified	0,02	0,01	0,02	0,05	0,05	0,05	0,04
MITE (DXX)	0,18	0,20	0,23	0,47	0,53	0,52	0,48
Helitron (DHH)	0,01	0,03	0,03	0,06	0,03	0,04	0,04
unclassified DNA transposon	0,02	0,03	0,04	0,04	0,04	0,07	0,04
Class I/Class II- ratio	**13,44**	**16,77**	**12,16**	**10,54**	**10,54**	**10,07**	**10,55**
Unclassified Element (XXX)	0,65	0,30	0,43	0,63	0,83	0,81	0,74
Simple Sequence Repeat	0,32	0,21	0,35	0,10	0,22	0,21	0,22
rRNA gene	0,29	0,53	0,11	0,12	0,01	0,06	0,12

Table S19: Representation of different repetitive DNA classes in barley genomic DNA resources

An enrichment factor for the respective repeat element types and their presence in the various sequence resources has been calculated. Random, non-assembled WGS data have been used as the reference. The numerical values give the degree of enrichment for the respective elements in the various sequence collections. A value of 1 represents equal representation, negative values indicate depletion and values >1 indicate enrichment of the respective element types.

% of all bp	BES 374 Mb	random BACs 63Mb	gene bearing BACs 379 Mb	Bowman WGS assembly 1.8 Gb	Morex WGS assembly 1.9Gb	Barke WGS assembly 2.0 Gb
MDR: unique/low copy (20mer <10x)	-1,04	-1,02	1,11	1,89	1,86	1,64
MDR: medium/high copy (>=10 to <1000x)	1,10	1,07	-1,01	-1,02	-1,08	1,04
MDR: very high copy (20mer >= 1000x)	-1,06	-1,05	-1,10	-3,97	-3,03	-2,46
Mobile Element	1,00	1,01	-1,10	-1,34	-1,39	-1,40
Class I: Retroelement (RXX)	1,02	1,00	-1,12	-1,37	-1,43	-1,43
LTR Retrotransposon (RLX)	1,02	1,00	-1,13	-1,38	-1,44	-1,44
Copia (RLC)	-1,23	-1,12	-1,06	-1,75	-1,81	-1,81
Gypsy (RLG)	-1,02	-1,07	-1,31	-1,17	-1,23	-1,24
unclassified LTR	1,13	1,08	-1,07	-1,41	-1,46	-1,45
non-LTR Retrotransposon	1,28	1,40	2,79	2,59	2,58	2,36
Class II: DNA Transposon (DXX)	-1,22	1,11	1,14	-1,07	-1,07	-1,12
DNA Transposon Superfamily (DTX)	-1,24	1,10	1,07	-1,17	-1,16	-1,21
CACTA superfamily (DTC)	-1,27	1,07	1,00	-1,29	-1,29	-1,33
hAT superfamily (DTA)	-1,32	1,56	1,75	1,24	1,31	1,25
Mutator superfamily (DTM)	1,11	1,80	2,45	2,32	2,34	2,23
Tc1/Mariner superfamily (DTT)	1,18	1,37	2,59	2,86	2,83	2,64
PIF/Harbinger (DTH)	1,58	1,54	3,20	3,59	3,51	2,95
unclassified	-1,08	1,51	3,05	3,24	3,19	2,80
MITE (DXX)	1,08	1,26	2,59	2,89	2,82	2,62
Helitron (DHH)	2,20	2,19	5,33	2,95	3,39	3,12
unclassified DNA transposon	1,23	1,98	2,12	1,78	3,18	1,73
Unclassified Element (XXX)	-2,14	-1,51	-1,03	1,29	1,26	1,14
Simple Sequence Repeat	-1,50	1,12	-3,30	-1,43	-1,54	-1,43
rRNA gene	1,79	-2,70	-2,38	-20,90	-4,95	-2,40

Supplemental Note 6. RNA sequencing

S6.1 Plant Material and RNA extraction

All plant material was prepared from August to October 2011 in a glasshouse at JHI, Dundee. No artificial lights were used and automatic watering was utilised. For growing plants, a roller-bench was subdivided in the three areas representing three replicates. In the centre of each area, 10 cm from the bench surface, an RFID temperature logger miniNOMAD® (model OM-84-TMP from Omega Engineering Ltd) was placed to record temperatures every hour from 10:00 am August 3 to 10:00 am October 16. In total 1,800 temperature measurements for each logger were recorded as shown below (Figure S12).

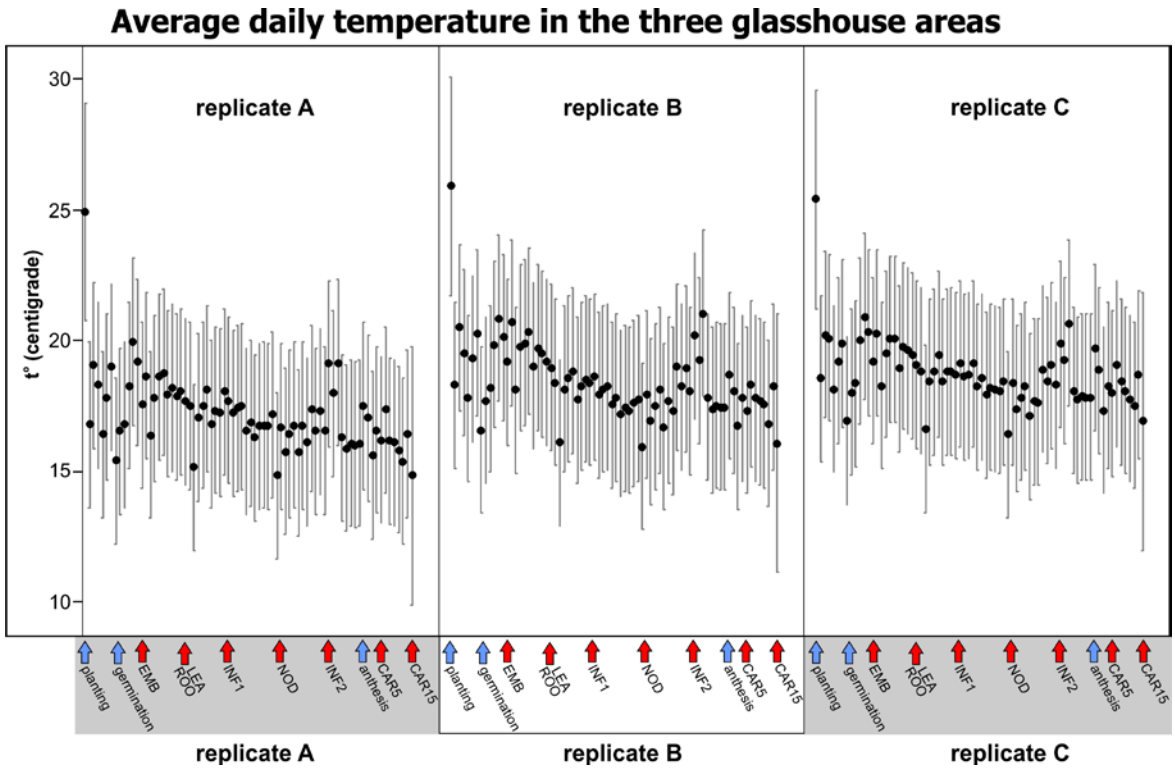


Figure S12: Growth conditions for three replicates of tissue sampling for RNA-seq

Black dots show daily mean temperatures for 75 days of the experiment. Error bars are 95% Bonferroni corrected confidence intervals for the mean (n=24). Arrows below show experiment start, anthesis (blue) and sampling points (red).

Tissue sampling

Approximately 400 barley cv. Morex seeds were surface sterilized in 1% (v/v) bleach solution for 20 minutes, followed by extensive rinsing with tap water. Seeds were subdivided in three parts -one was used for germination on the plates to obtain EMB samples, the next was used to plant in the pots with perlite (for LEA and ROO samples), and the remaining in the pots with compost to grow plants for INF, NOD and CAR tissues. Tissue sampling was always performed between 10 and 11 am. All sampled tissues were flash frozen in liquid nitrogen and put in -80°C freezer for RNA extraction.

In total, 8 barley tissues were prepared for RNA-seq (Figure S13, Table S20) as follows:

EMB: In total 20 sterilized seeds were spread across three stacked sheets of wet filter paper and covered with three sheets of wet filter paper in the large 23x23 cm square petri dishes. Petri dishes were wrapped in foil and placed in the same glasshouse area where seeds were planted. After four days of germination, embryonic tissue (coleoptile, mesocotyl and seminal roots) from nine similarly looking germinating seeds were dissected.

LEA, ROO: Sampling of the LEA and ROO tissue was made when upper parts of the plants grown in pots with perlite reached c. 10 cm length (17 days after planting). Plants were carefully removed from the pots, remains of the perlite were washed away, and roots and shoots were separated by cutting just below and above the root/shoot junction. Remains of the embryo were not sampled. Roots and shoots from four plants were used per sampling.

INF1, INF2: The main tillers of c. 30 day old plants were used to dissect developing inflorescences (INF1). Only those inflorescences that were about 5 mm long were collected for RNA extraction. In total, 15-20 inflorescences were dissected per sample. For obtaining 10-15 mm long inflorescences, main tillers from c. 50 day old plants were used.

NOD: The third internode (4-6 cm long) was dissected from 42 days old plants. Leaf blades and sheaths were removed, and the internode was cut out from just below the 4th and just above the 3rd nodes.

CAR5, CAR15: To obtain developing caryopses (CAR), anthesis time was determined using c. 60 day old plants by removing anthers from the central florets at the centre of the spike and examining their colour and determining whether they have shed. Spikes with yellow anthers, but not shedding, were marked and examined subsequently for shedding to set the anthesis time. Five days after anthesis, 3-5 mm long caryopses (CAR5) were dissected from the central florets of the central spike. Fifteen caryopses were collected per sample. Ten days later, c. 10 mm caryopses were dissected from spikes to obtain CAR15 samples.

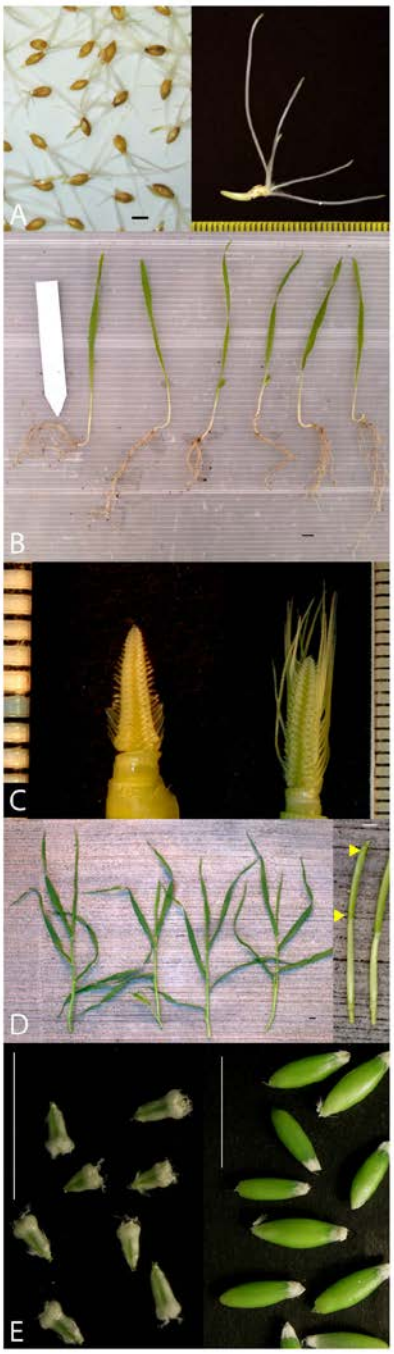


Figure S13: overview plant tissues for RNA-seq

Barley tissues sampled for RNA-seq: (A) germinating grain (4 day) embryos (EMB); (B) roots (ROO) & shoots from seedlings(LEA) (10 cm stage); (C) early developing inflorescences [5 (INF1) & 15 mm (INF2)]; (D) developing tiller internodes (NOD) (six-leaf stage; sectioned between arrows); (E) early developing grain [5 (CAR5) & 15 dpa (CAR15)]. Bar represents 10 mm; rulers show mm.

Table S20: Plant material and sequence reads from RNA-seq

Tissue abbreviation	Tissue description	Paired-end reads
EMB	4-day embryos dissected from germinating grains	141,510,020
ROO	Roots from the seedlings (10 cm shoot stage)	191,528,332
LEA	Shoots from the seedlings (10 cm shoot stage)	134,849,046
INF1	Young developing inflorescences (5mm)	153,512,032
INF2	Developing inflorescences (1-1.5 cm)	175,615,232
NOD	Developing tillers at six – leaf stage, 3rd internode	557,006,080
CAR5	Developing grain, bracts removed (5 DPA)	148,725,125
CAR15	Developing grain, bracts removed (15 DPA)	165,676,476

S6.2 RNA-seq of developmental stages

RNA extraction

Total RNA was extracted from c. 200 mg dissected frozen tissue from each sample using 2 ml TriReagent (Sigma-Aldrich) as recommended, with two additional phenol-chloroform purification steps followed by ethanol precipitation. RNAs were quality checked using the RNA 6000 Nano kit on a 2100 Bioanalyzer (Agilent). All RNA samples were judged high quality with an RNA Integrity Number (RIN) >8.

Library construction and Illumina sequencing

All RNA-seq libraries were constructed and run at 'The Genome Analysis Centre' (Norwich, UK). The Illumina TruSeq RNA Sample preparation kit (Illumina Inc.) was used according to the manufacturer's protocol. In brief, poly-A containing mRNA molecules were purified from 1 ug total RNA using poly-T oligo attached magnetic beads using two rounds of purification. The purified mRNA was fragmented by addition of 5x fragmentation buffer (Illumina, Hayward, CA) and was heated at 94°C in a thermocycler for 8 minutes. The fragmentation yields fragments of ~250 bp. First strand cDNA was synthesised using random hexamers to eliminate the general bias towards 3' end of the transcript. Second strand cDNA synthesis was done by adding GEX second strand buffer (Illumina, Hayward, CA), dNTPs, RNaseH and DNA polymerase I followed by incubation for 2.5 h at 16°C. Second strand cDNA was further subjected to end repair, A-tailing, and adapter ligation with barcoded adapters in accordance with the manufacturer supplied protocols. Purified cDNA templates were enriched by 15 cycles of PCR for 10 s at 98°C, 30 s at 60°C, and 30 s at 72°C using PCR Primer Mix Cocktail and PCR Master Mix (Illumina, Hayward, CA). The samples were cleaned using AMPure XP Beads and eluted in 30 µl Resuspension Buffer as per manufacturer's instructions (QIAGEN, CA). Purified cDNA libraries were quantified using

Bioanalyzer DNA 100 Chip (Agilent Technology 2100 Bioanalyzer). The libraries were normalised to 10 nM and pooled equimolarly in pools of 8 samples per pool.
Each pool of libraries was diluted to 2nM with NaOH solution and 5 µL of diluted library transferred into 995 µL HT1 (high salt buffer supplied by Illumina) to give a final concentration of 10 pM. 120 µL of normalised library was then transferred into a 200 µL strip tube and placed on ice before loading onto the Illumina cBot cluster generation system. Each pool of libraries was clustered onto two lanes of an Illumina flowcell. Flow cells were clustered using TruSeq Paired-End Cluster Kit v3, following the Illumina PE_Amplification_Linearization_Blocking_Hybridisation_v8 recipe. On completion of the clustering procedure, the flow cell was loaded onto the Illumina HiSeq2000 instrument according to the manufacturer's instructions. A paired end sequencing protocol was performed generating 2 x 100bp reads using TruSeq SBS kit v3 sequencing chemistry, Illumina software HCS 1.4 and RTA 1.12.

Supplemental Note 7. Gene annotation, gene family and comparative analysis and expression analysis

S7.1 Construction of a high confidence (HC) barley gene set

S7.1.1 RNA-seq mapping and transcript reconstruction

To annotate transcriptional active regions of the barley genome, RNAs from eight different samples were sequenced using Illumina transcriptome sequencing technology. Three biological replicates were generated for each sample (Table S21).

Table S21: RNA-seq libraries used for gene prediction and transcriptional characterization of the barley genome.

The table gives the amount of RNA-seq data analysed for 8 different barley cv. Morex samples that represent different tissues or developmental stages. For abbreviations and description of the tissues used please refer to table S20.

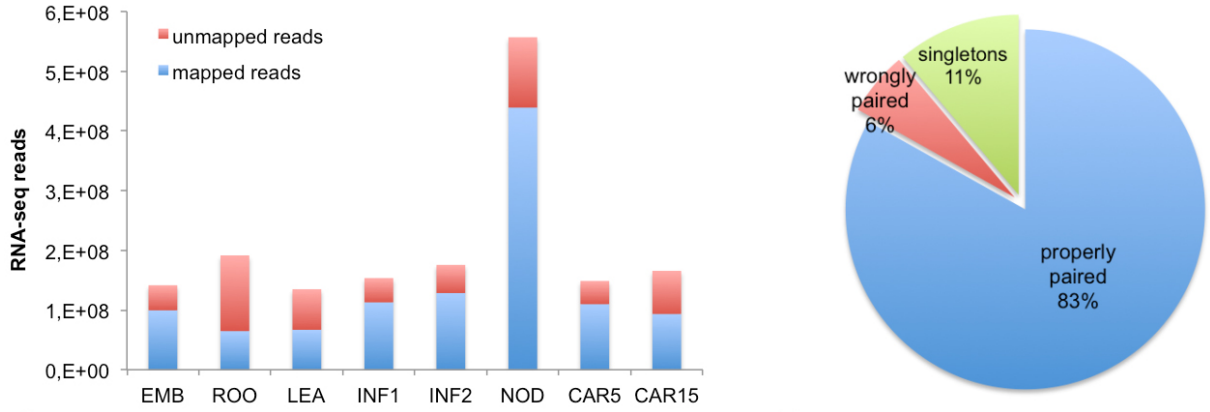
Sample	Batch	insert size (bp)	Read-pairs	Sequence (bp)
EMB	Batch 2	137	15,951,510	3,190,302,000
		134	20,685,672	4,137,134,400
		140	12,175,190	2,435,038,000
	Batch 1	137	8,449,076	1,689,815,200
		134	7,102,849	1,420,569,800
		140	6,390,713	1,278,142,600
	Σ	**137**	**70,755,010**	**14,151,002,000**
ROO	Batch 2	132	15,249,190	3,049,838,000
		155	27,933,445	5,586,689,000
		142	21,497,606	4,299,521,200
	Batch 1	132	8,136,190	1,627,238,000
		155	10,806,106	2,161,221,200
		142	12,141,629	2,428,325,800
	Σ	**143**	**95,764,166**	**19,152,833,200**
LEA	Batch 2	136	15,668,787	3,133,757,400
		135	15,890,875	3,178,175,000
		135	13,559,058	2,711,811,600
	Batch 1	136	8,521,241	1,704,248,200
		135	6,249,999	1,249,999,800
		135	7,534,563	1,506,912,600
	Σ	**135**	**67,424,523**	**13,484,904,600**
INF1	Batch 2	155	18,358,276	3,671,655,200
		146	19,509,221	3,901,844,200
		139	14,871,585	2,974,317,000
	Batch 1	155	9,747,468	1,949,493,600
		146	6,556,267	1,311,253,400
		139	7,713,199	1,542,639,800
	Σ	**147**	**76,756,016**	**15,351,203,200**
INF2	Batch 2	162	21,176,491	4,235,298,200
		158	24,417,426	4,883,485,200
		174	14,940,145	2,988,029,000
	Batch 1	162	11,235,461	2,247,092,200
		158	8,080,069	1,616,013,800
		174	7,958,024	1,591,604,800
	Σ	**165**	**87,807,616**	**17,561,523,200**
NOD	Batch 2	151	20,722,455	4,144,491,000
		265	53,311,727	10,662,345,400
		188	166,184,951	33,236,990,200
	Batch 1	151	10,922,940	2,184,588,000
		265	27,360,967	5,472,193,400
	Σ	**208**	**278,503,040**	**55,700,608,000**
CAR5	Batch 2	254	16,158,577	3,231,715,400
		299	19,631,567	3,926,313,400
		273	15,184,667	3,036,933,400
	Batch 1	254	8,613,426	1,722,685,200
		299	6,793,954	1,358,790,800
		273	7,980,335	1,596,067,000
	Σ	**275**	**74,362,526**	**14,872,505,200**
CAR15	Batch 2	171	20,446,172	4,089,234,400
		202	21,079,514	4,215,902,800

	175	13,157,378	2,631,475,600
	171	11,415,975	2,283,195,000
Batch 1	202	8,583,794	1,716,758,800
	175	8,155,405	1,631,081,000
Σ	**183**	**82,838,238**	**16,567,647,600**
	Σ	**834,211,135**	**166,842,227,000**

First, the paired-end 100bp-reads of the three replicates of a sample were pooled and each sample data set was aligned against the library based repeat-masked assembly of barley cv. Morex 50x WGS data using Bowtie (v 0.12.7) and TopHat (v1.4.0) with default settings and the previously determined mean inner distance between mate pairs (Figure S14, Table S22, S23). Based on the spliced alignments of the RNA-seq reads, Cufflinks (v1.3.0) was used to assemble the mapped reads. The different RNA Seq samples were treated separately. The individual predicted transcript models were combined with Cuffcompare (v1.3.0) that takes the Cufflinks output structures and creates a non-redundant set of transcript structures. Thereby, structures available in two or more samples sharing identical intron structures were united and counted only once.

CuffDiff implements methods that address potential problems caused by overdispersion among biological replicates. The methods used refuse significance testing for genes/transcript that have replicate expression levels outside of the confidence interval of 0.01 generated when pooling the replicates together. Thus for replicates that show slightly reduced correlation with the two other replicates we corrected for aberrant confidence intervals during the analytical procedure.

Unmapped RNA-seq reads (both mates of pairs were not mapped to Morex WGS assembly) were mapped against 28,592 public barley fl-cDNAs by using TopHat and the number of mapped reads was counted for each fl-cDNA. In all samples most RNA-seq reads were mapped to barley fl-cDNA accession AK24813.1 (Table S24). Especially for samples ROO and LEA that show low mapping rate on the barley Morex WGS assembly between 18% (ROO) and 14% (LEA) respectively of the total RNA-seq data sets were mapped to this fl-cDNA. BLAST comparison of AK24813.1 against the NCBI non. red. database was performed. The first best matching entry, JF489233.1, is described as ribosomal RNA gene of *Secale cereale* (including external transcribed space, 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8 S ribosomal RNA gene and internal transcribed spacer 2 (complete sequence) as well as partial sequence of 26S RNA gene).



A B

Figure S14: Mapping statistics of RNA-seq reads against barlex cv. Morex assemblies

A) Number of mapped and unmapped RNA-seq reads for each sample. B) Mapping statistics of paired-end reads: Properly-paired: mapping of both reads according to given mate pair distance; wrongly paired: both mate pairs mapped but with different inner distance than given; singletons: only one read of the mate pair was mapped.

Table S22: Mapping of RNA-seq reads to barley cv. Morex WGS contigs.

General sequencing and mapping statistics of RNA Seq data generated for different barley tissues and developmental stages. For the sample description please refer to Table S20.

Sample	Input sequencing data			Non-redundant mapped reads								TopHat mappings (inclusive multiple read mappings)						
	Pairs	Reads	Sequence (Gb)	Unique		Multiple		Mapped		Unmapped		Total read mappings	Properly paired		Wrong paired		Singletons	
				#	%	#	%	#	%	#	%	#	#	%	#	%	#	%
EMB	70,755,010	141,510,020	14.151	98,994,193	99.45	545,758	0.55	99,539,951	70.34	41,970,069	29.66	100,246,128	86,754,426	86.54	3,440,678	3.43	10,051,024	10.03
ROO	95,764,166	191,528,332	19.153	64,361,295	99.49	328,822	0.51	64,690,117	33.78	126,838,215	66.22	65,101,610	51,616,750	79.29	6,155,956	9.46	7,328,904	11.26
LEA	67,424,523	134,849,046	13.485	66,394,476	99.29	477,297	0.71	66,871,773	49.59	67,977,273	50.41	67,995,613	58,170,678	85.55	2,742,300	4.03	7,082,635	10.42
INF1	76,756,016	153,512,032	15.351	112,191,489	99.39	690,893	0.61	112,882,382	73.53	40,629,650	26.47	113,715,602	92,576,248	81.41	8,255,322	7.26	12,884,032	11.33
INF2	87,807,616	175,615,232	17.562	127,650,216	99.41	759,243	0.59	128,409,459	73.12	47,205,773	26.88	129,322,406	101,425,752	78.43	12,693,552	9.82	15,203,102	11.76
NOD	278,503,040	557,006,080	55.701	436,890,594	99.48	2,288,698	0.52	439,179,292	78.85	117,826,788	21.15	442,101,368	397,175,796	89.84	9,054,756	2.05	35,870,816	8.11
CAR5	74,362,526	148,725,052	14.873	109,268,653	99.49	565,099	0.51	109,833,752	73.85	38,891,300	26.15	110,545,709	82,078,056	74.25	12,819,664	11.6	15,647,989	14.16
CAR15	82,838,238	165,676,476	16.568	92,300,444	98.83	1,091,865	1.17	93,392,309	56.37	72,284,167	43.63	94,722,797	64,245,252	67.82	9,367,206	9.89	21,110,339	22.29
Σ	834,211,135	1,668,422,270	166.842	1,108,051,360	99.39	6,747,675	0.61	1,114,799,035	66.82	553,623,235	33.18	1,123,751,233	934,042,958	83.12	64,529,434	5.74	125,178,841	11.14

Table S23: Gene structure prediction: Cufflinks and Cuffcompare results.

The table gives an overview on the transcribed loci and genomic characteristics for the RNA Seq based gene and exon detection in the different spatiotemporal samples. For the sample description please refer to the description given in Table S20.

	EMB	ROO	LEA	INF1	INF2	NOD	CAR5	CAR15	Merged
predicted genes	50,396	46,541	48,479	50,545	60,795	60,418	52,220	46,257	86,549
single exon genes	25,400 (50%)	21,944 (47%)	24,360 (50%)	26,927 (53%)	36,104 (59%)	34,026 (56%)	26,643 (51%)	22,918 (50%)	51,442 (59%)
multi exon genes	24,996 (50%)	24,547 (53%)	24,119 (50%)	23,618 (47%)	24,691 (41%)	26,392 (44%)	25,577 (49%)	23,339 (50%)	35,107 (41%)
predicted distinct exons*	157,626	151,676	153,835	156,189	170,948	171,992	164,500	149,697	307,266
exons per gene	3.13	3.26	3.17	3.09	2.81	2.85	3.15	3.24	3.55
max exons per gene	51	51	53	51	51	51	52	50	72
predicted transcripts	57,666	53,315	56,203	59,149	71,031	69,361	61,331	54,386	142,763
genes with alternative transcripts	5,753 (11%)	5,394 (12%)	5,963 (12%)	6,547 (13%)	7,576 (12%)	6,828 (11%)	6,962 (13%)	6,229 (13%)	21,084 (24%)
transcripts per gene	1.14	1.15	1.16	1.17	1.17	1.15	1.17	1.18	1.65
max transcripts per gene	7	6	6	7	6	6	6	7	32
mean gene size (first to last exon) [bp]	1,683	1,762	1,698	1,666	1,574	1,566	1,722	1,754	1,550
mean transcript (UTR + CDS)[bp]	1,072	1,096	1,082	1,094	1,109	1,057	1,135	1,127	1,315
mean exon size [bp]	348	342	347	360	403	377	370	356	474

Table S24: Analysis of RNA-seq reads not mapped to the barley cv. Morex WGS assembly

Tissue	Most tagged FL-cDNA	# of RNA-seq reads	% of RNA-seq reads
EMB	AK248318.1	2,499,207	1.77
ROO	AK248318.1	34,756,265	18.15
LEA	AK248318.1	19,119,677	14.18
INF1	AK248318.1	2,043,690	1.33
INF2	AK248318.1	1,395,732	0.79
NOD	AK248318.1	3,152,464	0.57
CAR5	AK248318.1	817,829	0.55
CAR15	AK248318.1	4,007,457	2.42

S7.1.2 Clustering of barley fl-cDNAs and RNA-seq gene models for gene family analysis

To create a comprehensive barley gene set, 28,592 publicly available fl-cDNAs were merged with 142,763 predicted RNA-seq transcripts. First, a non-redundant set of fl-cDNAs was created and thereby 5,243 fl-cDNAs were removed using CD-Hit clustering at 98% identity level (-c98 –n 8). For both, RNA-seq transcripts and fl-cDNAs open reading frames (ORFs) were predicted using OrfPredictor[32]. 227 RNA-seq transcripts and 6 fl-cDNAs were excluded from further analysis because of failed ORF prediction.

Second, genomic clustering was performed. Fl-cDNAs were mapped to the reference WGS assembly of barley cv. Morex using Est2Genome (>=95% identity and minimum coverage >=90%). Strand information was assigned to RNA-seq transcript structures and to mapped fl-cDNAs as defined by the ORF prediction. Structure information of mapped fl-cDNAs and RNA-seq gene models were combined using cuffcompare and a final structure set of 82,121 loci incorporating 140,720 transcripts identified. For further analysis representative transcripts of the loci were extracted as defined by the maximum length ORF. In total, 7,834 gene loci comprised of clusters of mapped fl-cDNAs and RNA-seq transcripts and 375 fl-cDNA specific clusters were identified and the transcript or fl-cDNA with maximum length coding sequence extracted for OrthoMCL analysis (steps 1 and 2 in Figure S15).

Third, sequence homology based clustering of the remaining 73,912 gene loci and 14,068 unmapped fl-cDNAs was performed. Therefore, the transcript structure with maximum length of coding sequence was selected as representative sequence for each gene locus (in case of conflicts, sequences starting with methionine were given higher priority) and were aligned to the fl-cDNAs (BLASTN) considering only the first-best (alignment length >=100bp and alignment identity >=95%). Fl-cDNAs (3,534) and RNA-seq transcripts (53,105) without significant alignments were added to the OrthoMCL data set (steps 3 and 4). Partial matches (coverage of query and subject <50%) between fl-cDNAs and RNA-seq transcripts caused by matching paralogous or high conserved sequence motifs were not considered. Affected sequences (587 fl-cDNAs and 3,982 RNA-seq transcripts) were excluded from the clustering procedure and added to the OrthoMCL analysis (steps 5 and 6).

In step 7, 722 out of 16,825 RNA-seq transcripts replaced a total of 707 fl-cDNAs for OrthoMCL-clustering because the RNA-seq transcripts include ≥95% of the fl-cDNA and showed extended nucleotide sequences. Due to gene fragmentation caused by unassembled reference contigs, 16,103 RNA-seq transcripts were completely included in 9,827 fl-cDNAs. In these cases, the fl-cDNA was extracted as representative sequence for OrthoMCL analysis (steps 8 and 9).

Based on that clustering pipeline, a working gene set of 79,379 sequences was created including 17,718 fl-cDNAs and 61,669 RNA-seq transcripts of which 75,258 (95%) have an assigned locus position on the reference barley cv. Morex contigs.

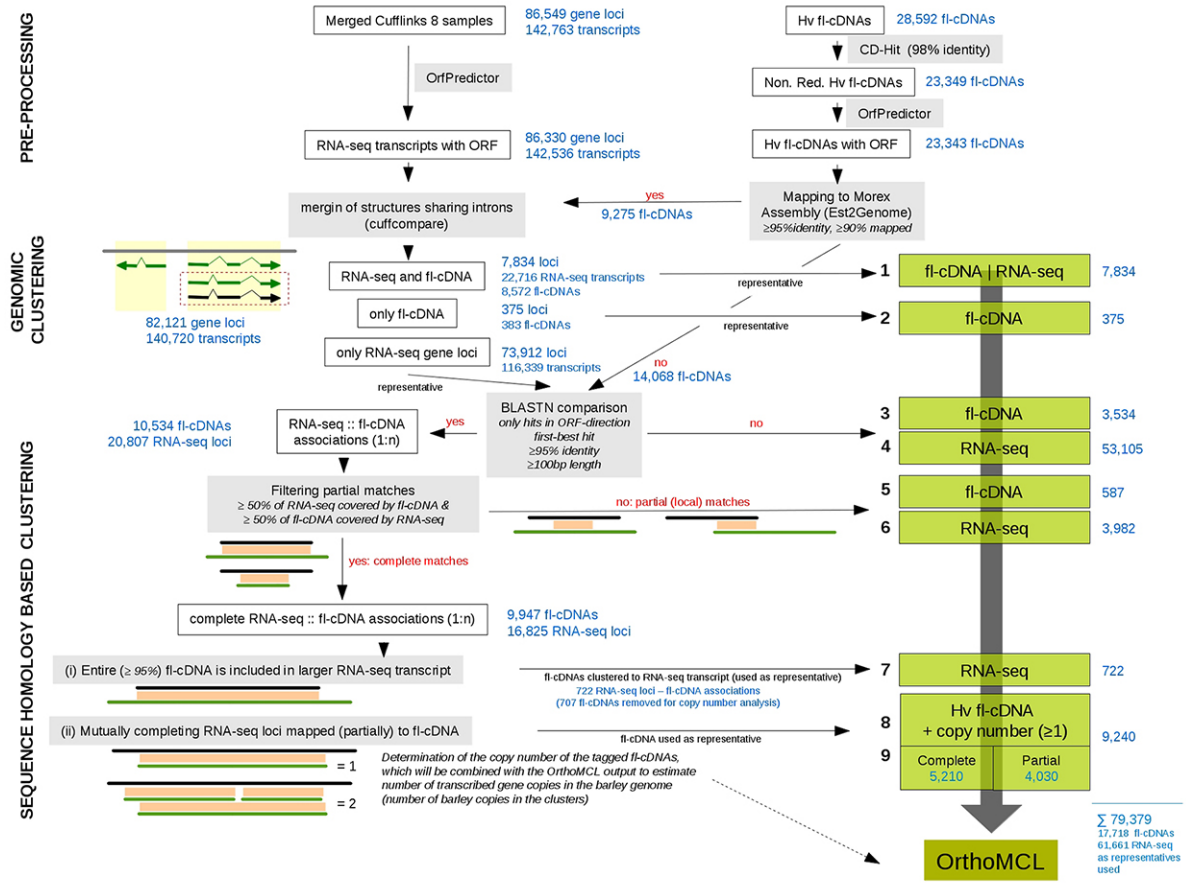


Figure S15: Schematic outline of the pipeline for clustering barley fl-cDNAs and RNA-seq transcripts. Flowchart of the different filtering and combination steps for the combination of fl-cDNA and RNA-seq based structural gene annotation. For details please refer to section S7.1.2

S7.1.3 Analysis of barley gene families and comparison against the gene complements of *Brachypodium distachyon*, *Sorghum bicolor, Oryza sativa* and *Arabidopsis thaliana*

S7.1.3.1 Construction of an orthologous grass gene set for barley

To define gene family clusters from the barley working transcript set (see section S7.1.2) and the sequenced genomes of three grasses from diverse grass sub-families and *Arabidopsis thaliana*, the OrthoMCL software version 2.0[33] was used. In a first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an e-value cut-off of 1e-05. Markov clustering of the resulting similarity matrix was used to define the ortholog cluster structure, using an inflation value (-I) of 1.5 (OrthoMCL default).

The input datasets were:

Brachypodium distachyon: v1.2 MIPS (http://mips.helmholtz-muenchen.de/plant/*Brachypodium*/index.jsp)

Sorghum bicolor: v1.4 MIPS (http://mips.helmholtz-muenchen.de/plant/*Sorghum*/index.jsp)

Oryza sativa: RAP2 (http://mips.helmholtz-muenchen.de/plant/rice/index.jsp)

Hordeum vulgare: working transcript set (combined and processed CUFFLINK and barley fl-cDNA; see section S7.1.2)

Arabidopsis thaliana: TAIR10 (www.*Arabidopsis*.org)

Splice variants were removed from the data set, keeping the longest protein sequence prediction, and data sets were filtered for internal stop codons and incompatible reading frames. A total of 195,329 coding sequences from these five species were clustered into 24,343 gene families. 8,717 clusters contained sequences from all four genomes. An overview of the clusters is shown in Figure S16.

To define a high-confidence set of coding sequences for barley, we extracted all barley sequences in clusters together with at least one representative from at least one of the other reference organisms. A total of 23,162 barley transcripts was identified with another 8,071 transcripts in barley-specific clusters (clusters containing only barley transcripts with at least two members; these still can have good homology to sequences from other species but cluster together because of much higher between-similarity, e.g. young duplications). To detect barley representatives in this 'barley-specific' set with significant sequence homology to sequences from the reference organisms we used BLASTP with an e-value cut-off of 10^{-5}. Another 2,997 sequences passed this criteria resulting in a total of 26,159 'high-confidence' (=homology-supported) barley transcripts.

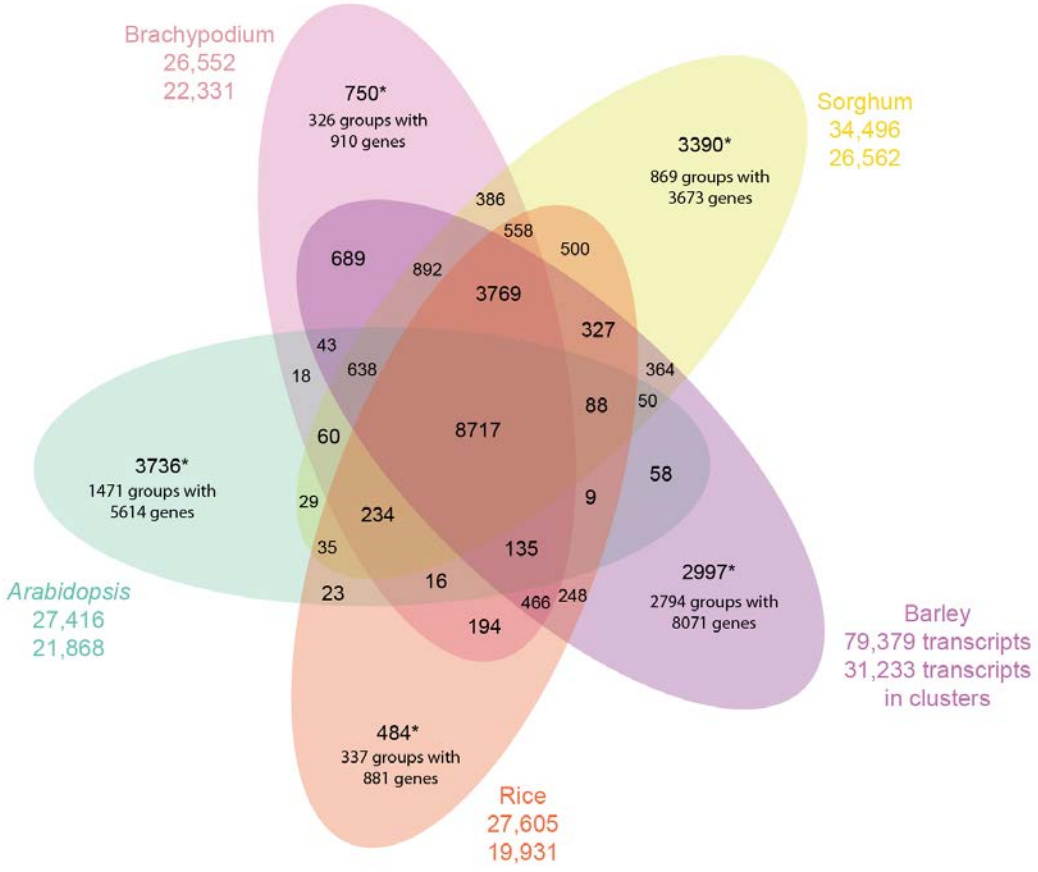


Figure S16: Distribution of orthologous gene families in barley, rice, *Sorghum*, *Arabidopsis* and *Brachypodium*.

Gene family comparisons were performed with OrthoMCL[34]. The following gene annotations were used: barley transcripts, RAP2[35,36], TAIR10[37], MIPSv1.4[38] and MIPSv1.2[39]. A total of 101,994 sequences from the five different organisms was clustered into 24,343 families (plus singletons). In each intersection of the Venn diagram the number of gene-groups ("families") are represented. Of the 79,379 transcripts predicted for barley, 31,233 were clustered in a total of 19,287 families; 8,717 families are common to all five genomes. Sections labeled by asterisks (*) contain genes clustered in groups only within one species (2 members or more). The numbers given in these sections represent the number of genes in these clusters with significant BLASTP hits ($\leq 10^{-5}$) to the other compared species.

S7.1.3.2 Expanded and contracted gene families in barley

We computed PFAM domain signatures and GO terms for the barley working transcript set with InterproScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/[40]. Only GO terms from the molecular function category were considered. To identify GO terms over- and under-represented in expanded barley orthologous groups (OG) (as determined by OrthoMCL) we used the GOstats[41] R package from Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/GOstats.html). Significant terms are reported for p-values =< 0.05. To identify over- and under-represented PFAM domains in expanded barley OG groups Bonferroni correction for multiple testing was applied. Expanded barley gene families were identified by their copy number distribution in the respective cluster using a binomial probability distribution with a significance level < 0.05.
GO terms and PFAM domains over- and underrepresented in barley-expanded gene clusters are given in Table S25, which is provided as a separate excel file.

The number of NBS-LRR genes in barley: From the most current *Brachypodium* gene set (version 1.2)[39] we identified 116 genes annotated as NBS-LRR genes. We used BLASTP with a minimum of 40% identity and 70% alignment query coverage and found 131 distinct copies in the barley HC gene set satisfying these criteria. To search for putative additional family members in barley and *Sorghum* / rice / *Brachypodium* we queried the OrthoMCL gene family clusters (S7.1.3.) with the previously identified 131 barley NBS-LRR genes. A total of 191 barley HC genes were identified in OrthoMCL clusters containing at least one barley gene. This compares to 153 genes in *Sorghum*, 176 genes in *Brachypodium* and 161 genes in rice. Based on these observations barley exhibits a slight expansion of NBS-LRR related genes as was also observed in the PFAM domain enrichment analysis (Table S25).

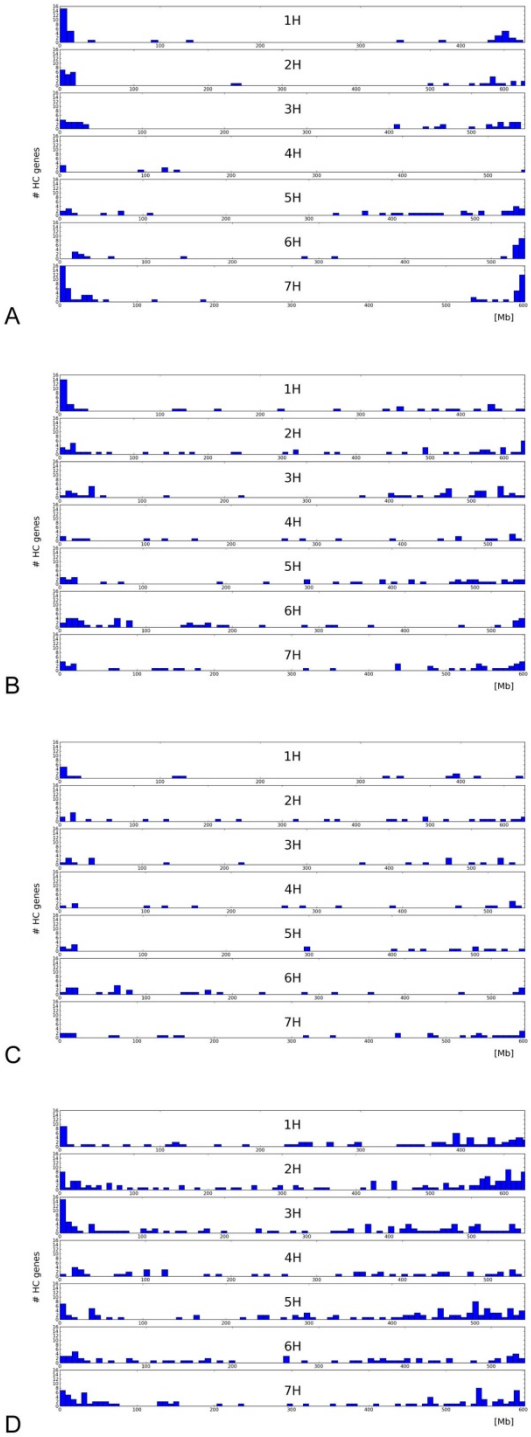


Figure S17: Physical distribution of selected expanded gene families.

Expanded gene families containing the PFAM domains (A) NB-ARC domain (PF00931), (B) Leucine Rich Repeat (PF00560), (C) Leucine rich repeat N-terminal domain (PF08263) and (D) Protein kinase domain (PF00069) were visualized in the context of the barley physical / genetic scaffold. Genes with PF00931 are strongly clustered towards the terminal (highly recombinogenic) parts of the chromosomes whereas genes containing PF00069 are more evenly distributed throughout the chromosomes. A cluster for PF00931 and PF00560 on 1HS coincides with the presence of the *Mla* powdery mildew resistance gene cluster.

S7.1.4 Identification of barley specific transcripts, nTARs and pseudogenes/remote homologs.

To identify barley specific transcripts as well as nTARs (novel Transcriptional Active Regions[42]) and pseudogenes/remote homologs, barley transcripts that were not assigned to the high-confidence gene set were filtered in a multi-step analysis pipeline (Figure S18). The input data set of 53,220 barley transcripts (49,420 RNA-seq transcripts predictions and 3,800 fl-cDNAs) was processed to identify barley transcripts that share homology with publicly available fl-cDNA sequences from wheat only ("triticeae-specific" transcriptome set), and to identify transcripts specific to barley. For the triticeae-specific transcriptome set a total of 7,999 barley transcripts was identified with a significant BLASTN hit (e-value < 10e-05) against the wheat fl-cDNA library but with no significant BLASTX hit (e-value < 10e-05) against a comprehensive set of angiosperm reference protein sequences, mainly from finished genome projects including all publicly available grass sequences.

Barley transcripts with no significant homology to plant protein sequences nor sequences included in the NCBI nonRED database (BLAST; e-value cut-off of 10e-05) were also compared against the genome sequences of rice (MSU_IRGSP_v7 release 31 Oct 2011) and *Brachypodium* (BLASTN; e-value cut-off of 10e-05). We found 4,830 barley transcripts that matched 13,118 locations on the *Brachypodium* genome and 2,450 barley transcripts at 5,844 locations on the rice genome. Among these, 2,046 barley transcripts were identified both on the rice and on the *Brachypodium* genome. Using more stringent parameters (alignment length >=50% of the originating barley transcript) 282 barley transcripts were identified on the *Brachypodium* genome sequence and 124 barley transcripts on the rice genome. The intersection comprised 90 barley transcript that were found both on the rice and the *Brachypodium* genome sequences.

For the remaining 16,560 barley transcripts with no homology support by any of the tested databases and sequences we computed PFAM domains using InterproScan. Only for 12 distinct transcripts PFAM domains were predicted.

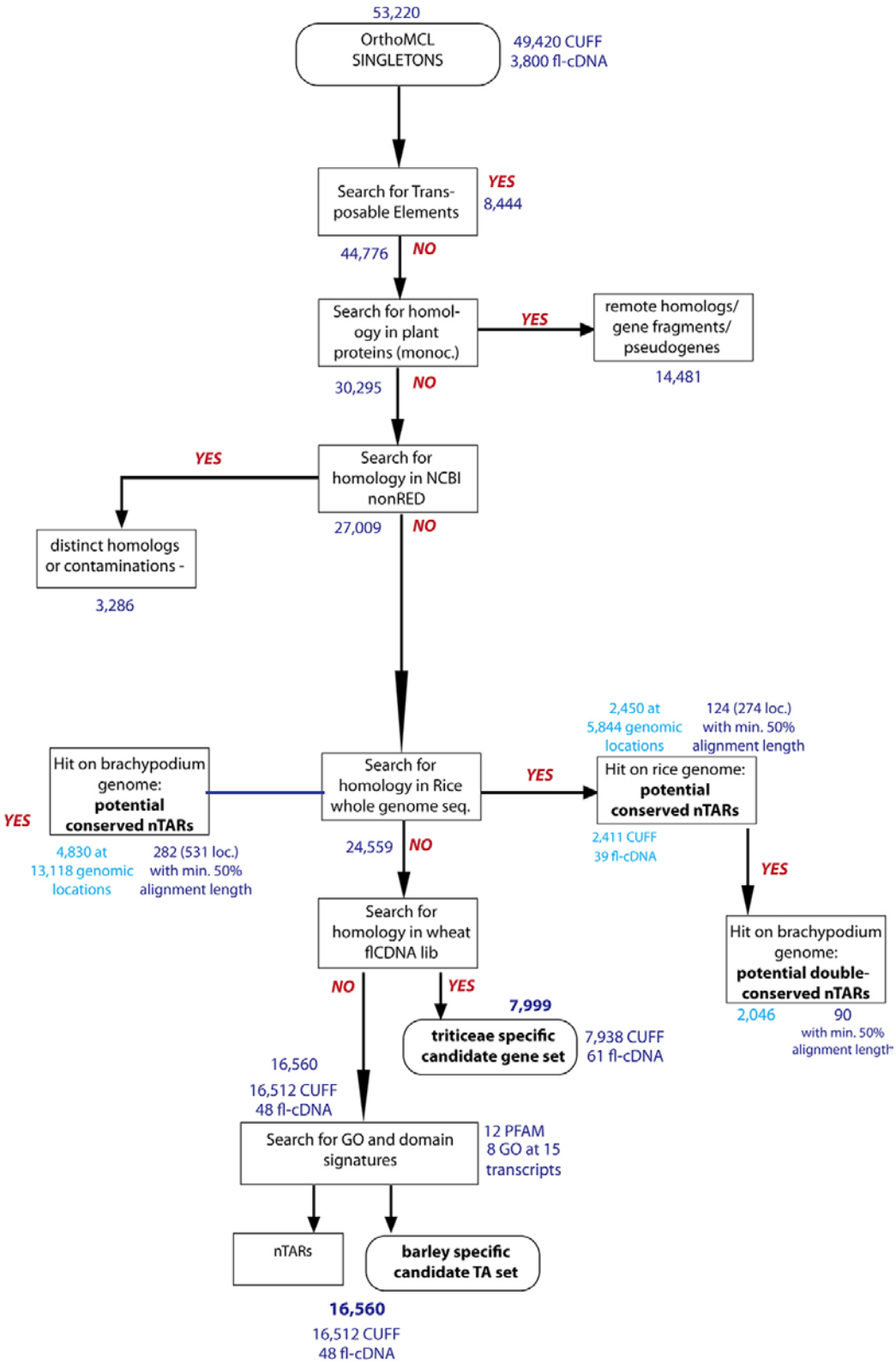


Figure S18: Schematic outline of analytical steps used to filter and analyse barley transcripts that did not classify as protein coding genes.

For details on the individual analytical steps please refer to section S7.1.4.

S7.2 Expression analysis of the barley transcriptome

S7.2.1. High confidence vs. less confidence genes (HC vs LC genes)

Based on the OrthoMCL-analysis for each element of the working gene set one representative transcribed gene locus was assigned. In case of mutually supported RNA-seq transcripts and fl-cDNA, the RNA-seq transcript with maximum fl-cDNA coverage was chosen.

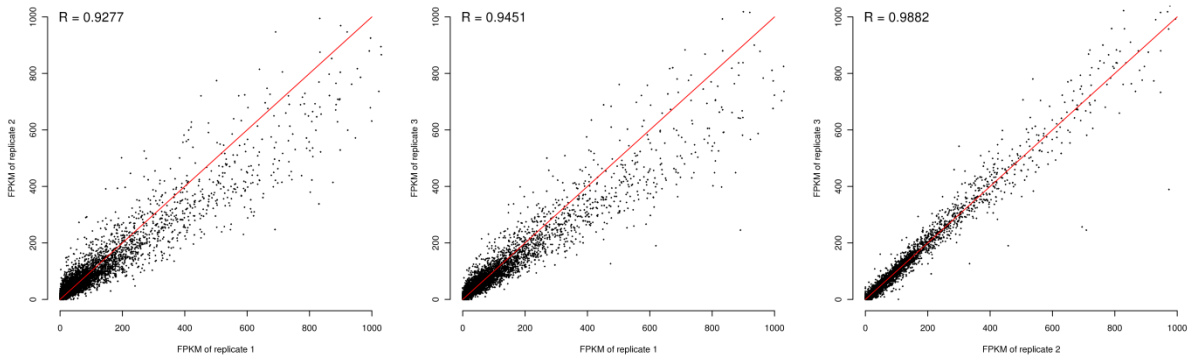
S7.2.2 Definition of representative gene structures for each gene locus

For expression analysis, one representative transcript structure for each gene was chosen. For gene loci with alternatively spliced transcripts, the isoform with maximum ORF extension and 5' end was selected. For each of the eight samples, RNA-seq reads of replicates were separately mapped against the reference contigs (barley cv. Morex assembly).

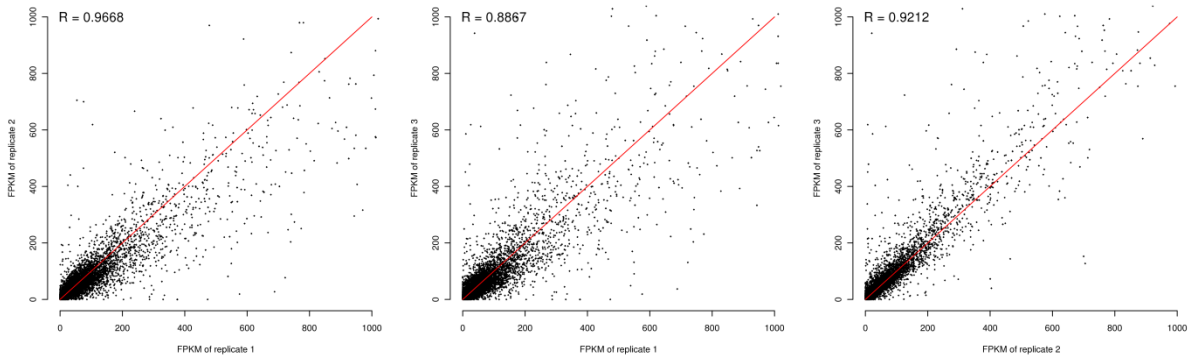
S7.2.3 Comparison of FPKM values between replicates

Correlation of expression level between replicates was tested. For each of the eight samples, RNA-seq reads of replicates were mapped against the reference WGS contig sequences (Morex assembly) using TopHat (8 x 3 mappings). Representative gene structures were taken and Cufflinks used to predict FPKM (fragments per kilobase of exon per million fragments mapped) expression values considering each replicate mapping of the eight different samples (3 x 8 FPKM calculations) (parameters: -G -b –compatible-hits-norm –F 0.00001). Subsequentially FPKM expression levels were pairwise compared for each experimental condition (Figure S19). Replicates show a high correlation within an experimental condition with a coefficient of correlation above 0.80 except for the comparison of replicate 1 and 3 in the LEA sample (corr. 0.74 – Table S20, Figure S19).

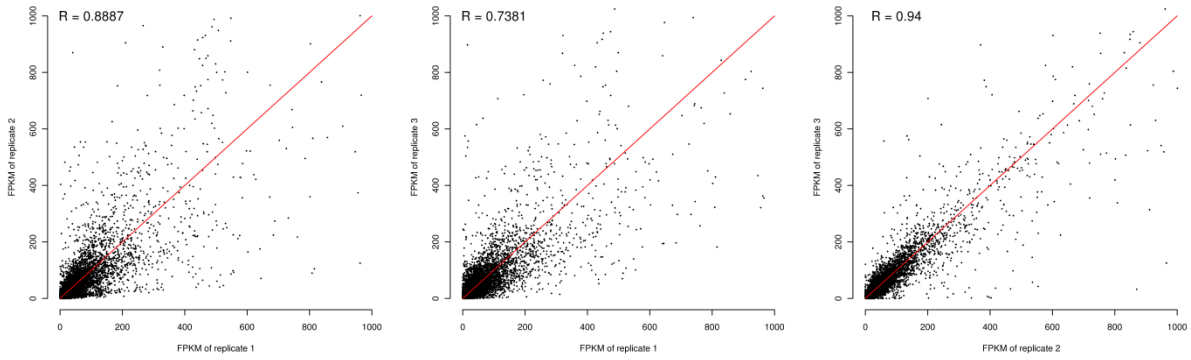
A) EMB



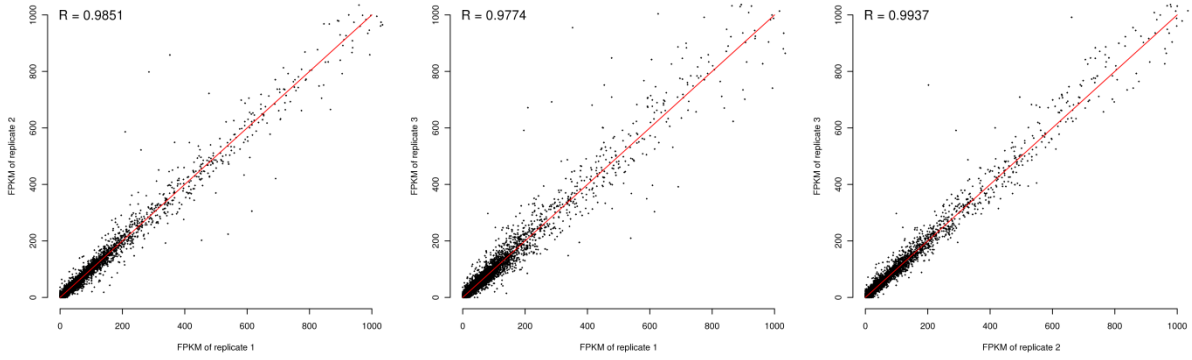
B) ROO



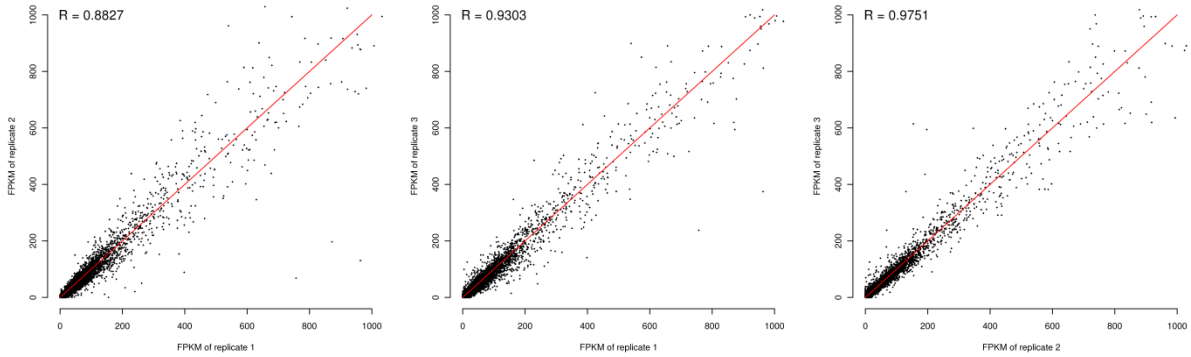
C) LEA



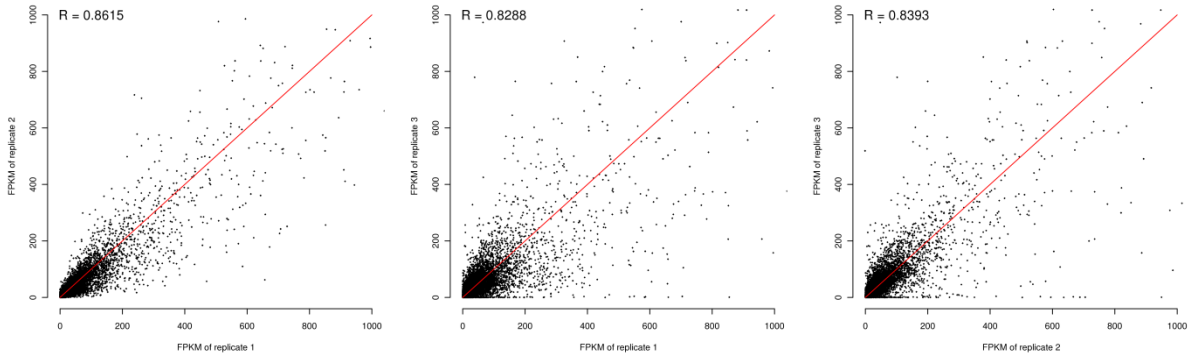
D) INF1



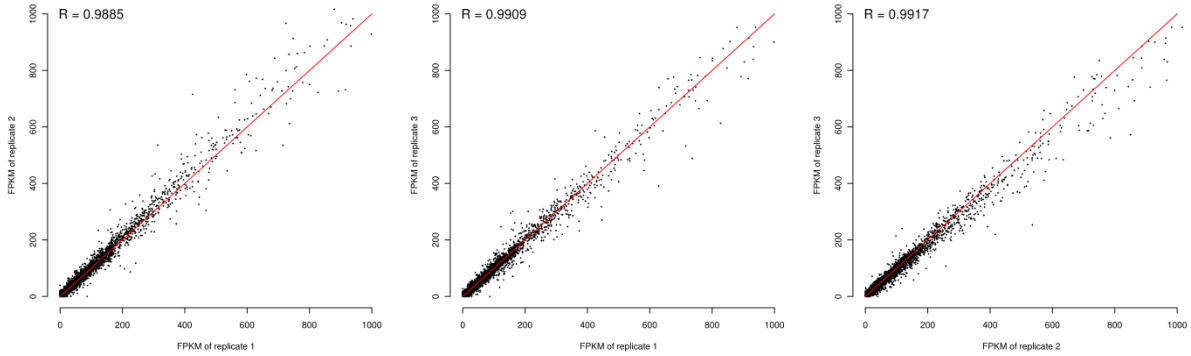
E) INF2



F) NOD



G) CAR5



H) CAR15

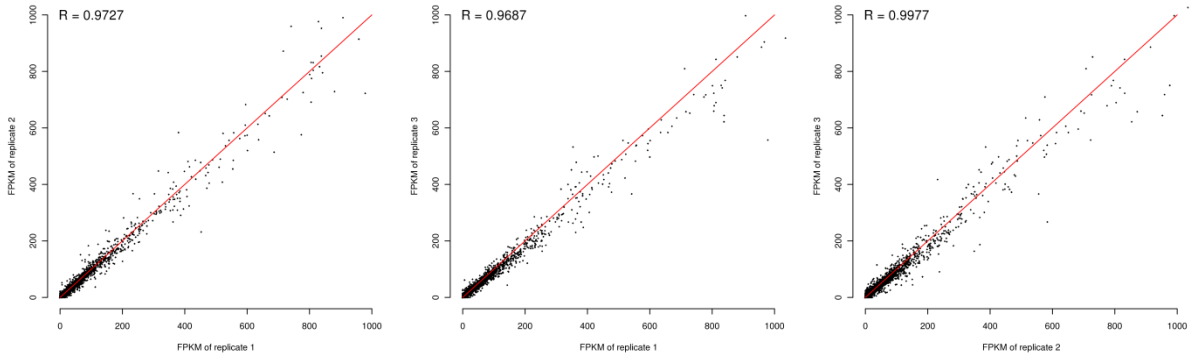


Figure S19: Correlation of FPKM expression levels between replicates of each sample.
For details please refer to section S7.2.3

A) CAR5 (replicate 1 = CAR5-2, replicate 2 = CAR5-3, replicate 3 = CAR5-4)
B) EMB (replicate 1 = EMB-A, replicate 2 = EMB-B, replicate 3 = EMB-C)
C) ROO (replicate 1 = ROO-A, replicate 2 = ROO-B, replicate 3 = ROO-C)
D) CAR15 (replicate 1 = CAR15-1, replicate 2 = CAR15-2, replicate 3 = CAR15-4)

E) LEA (replicate 1 = LEA-A, replicate 2 = LEA-B, replicate 3 = LEA-C)
F) INF2 (replicate 1 = INF2-F, replicate 2 = INF2-G, replicate 3 = INF2-H)
G) NOD (replicate 1 = NOD-A, replicate 2 = NOD-B, replicate 3 = NOD-C)
H) INF1 (replicate 1 = INF1-A, replicate 2 = INF1-B, replicate 3 = INF1-C)
For definition of the tissues/developmental stages please refer to Table S20

S7.2.4 FPKM calculation and determination of differentially expressed genes

Cuffdiff was applied to calculate transcript expression levels in FPKM and to find significant changes in transcript expression (parameters: --upper-quartile-norm, --frag-bias-correct). Thereby cuffdiff incorporates mapping information of each RNA-seq read and determines a count dispersion model for each experimental condition that describes variances of fragment counts across replicates. In 99.8% (24,196) of the gene loci, cuffdiff successfully calculated FPKM values for all eight samples. Only these genes were used for further analysis.

S7.2.5 Tissue-specific analysis of expression of the barley transcriptome

S7.2.5.1 Hierarchical clustering analysis

Hierarchical clustering analysis of the transcriptional profile between different tissues, developmental stages and inflorescence states was undertaken. FPKM values were log2-transformed and clusters were calculated using the hclust command in R and the default complete linkage method.

S7.2.5.2 Identification of a FPKM threshold

Analysis of FPKM expression level of all transcripts will include many transcripts with a FPKM value close to zero. Cufflinks gene structure prediction and determined FPKM values was compared to identify cases in which no gene structure was created but FPKM >0 predicted. For around 3,000 representative gene loci, no cufflinks structure was predicted but a FPKM value > 0 identified. These cases were caused by a small number of RNA-seq reads mapping at a gene locus but not at a sequencing depth sufficient to derive a genomic gene structure. Comparison of the gene expression values reveals significant higher expression levels for gene loci with structural information (median FPKM of 8.08, bottom fifth percentile of 0.339). Thus for further analysis a minimum expression level of 0.4 was used to consider a gene as expressed in one of the samples.

S7.2.5.3 Gene expression patterns in tissues

A Z-score analysis was performed to identify gene expression patterns in the different tissues, developmental stages and inflorescence states profiled by RNA-seq. The Z-score is a numerical measure that reflects the number of standard deviations of the expression level

of a gene in a specific sample is from the mean expression level in all samples. Therefore, Z-score of a specific gene i in tissue t (Zi,t) was calculated by the formula Zi,t = (Xi,t – Xmean,t)/SDt, where Xi,t is the log2-transformed transcript level, Xmean,t the average transcript level for the given gene in all tissues and SD the standard deviation of transcript level across all tissues (Figure S20).

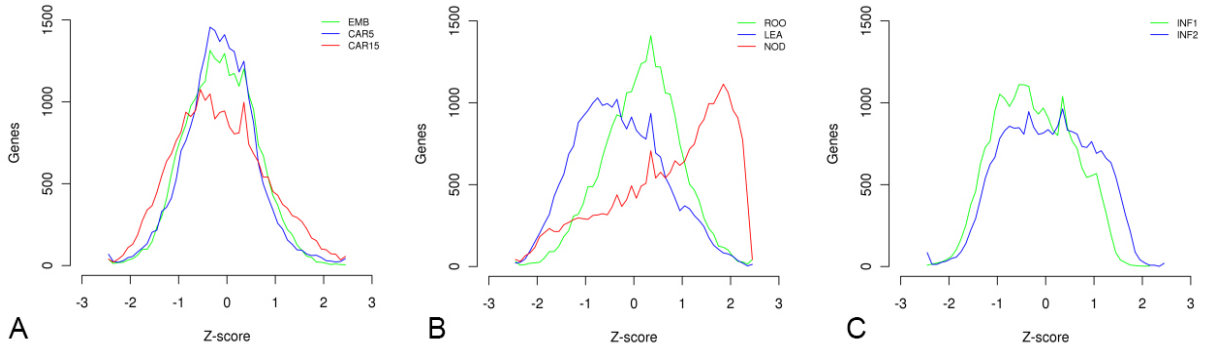


Figure S20: Frequency distribution of the determined Z-Scores for each tissue/sample. Please refer to section S7.2.5.3 for details. A) Developmental stages (EMB, CAR5, CAR15), B) tissues (ROO, LEA, NOD), C) inflorescence developmental stages (INF1, INF2)

S7.2.5.4 Tissue-specific gene expression

To identify genes expressed only in one spatiotemporal sample measured, expression values were compared between the eight different developmental stages, tissues and inflorescence treatments. A minimum FPKM level of 0.4 was required to consider genes as expressed. For each gene locus (HC transcript cluster and non-HC transcript cluster), the number of samples in which genes are expressed were counted (Figure S21). For 436 HC (HC genes) and 2,985 non-HC gene loci (LC genes), no significant expression was detected. These gene loci represent mapped fl-cDNA loci that lack RNA-seq support or have very low expression levels. Expression in different RNA-seq samples were different for LC and HC genes. LC gene expression tend to be more restricted to one particular sample with 22% of LC genes that have support from only one sample. 58% of HC genes were found to be represented in all samples while for LC genes this is observed for only 32% of the genes.

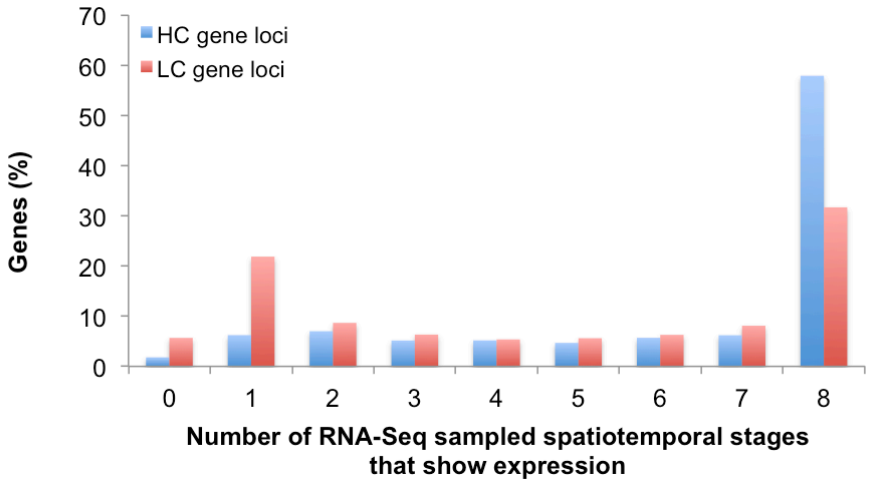


Figure S21: HC and LC gene expression supported in different RNA-seq samples
The majority of HC genes is supported by RNAseq data of all eight analysed tissue samples. In contrast the proportion of LC genes supported by a single tissue sample is significantly higher.

S7.2.5.5 Pairwise comparison for significant increase of gene expression in one tissue

The eight samples were tested for a significant increase in gene expression in a pairwise manner. The transcript counts of every tissue were compared to every other tissue using the cuffdiff expression test with FDR correction. Only genes with FPKM ≥0.4 were considered as expressed and a FDR-corrected test-value of lower 0.01 was required.

S7.2.6 Expression analysis of nTARs

Expression profiles of gene loci classified as nTARs were analysed. For 35% of the nTARs RNA-seq evidence was found in all eight samples and approximately one fourth was expressed in only one sample (Figure S22A). With the exception of tissue INF2 comparable numbers of nTARs were expressed in all samples (Figure S22B). Compared to HC gene loci, nTARs show a decrease in gene expression level (Table S26).

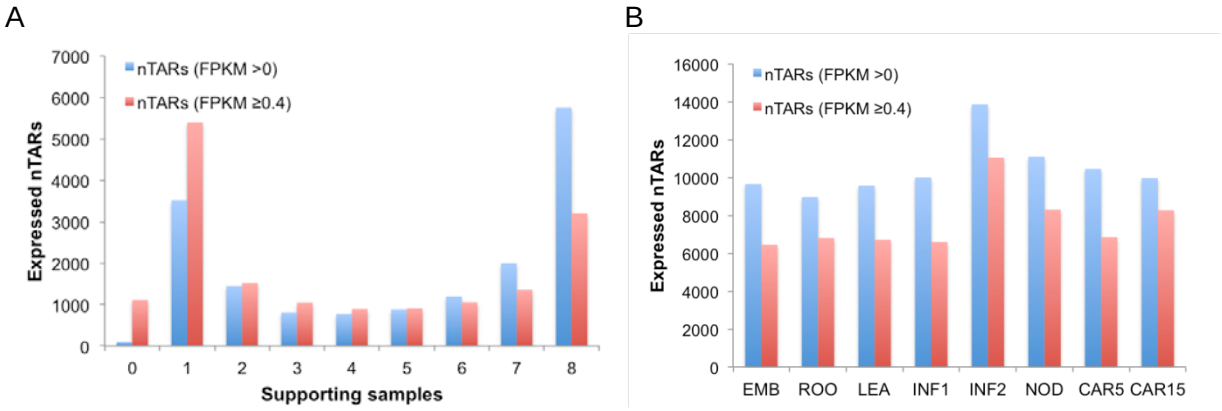


Figure S22: RNA-seq support of nTARs among samples. A) Number of samples supporting an expressed nTAR. B) Number of nTARs expressed in a specific sample

Table S26: Median expression levels (FPKM) for nTARs compared to HC genes

Sample	nTARs	HC genes
EMB	0.18	4.32
ROO	0.18	5.69
LEA	0.21	3.88
INF1	0.19	3.06
INF2	0.81	3.93
NOD	0.43	9.45
CAR5	0.23	3.98
CAR15	0.41	3.21

S7.3 Alternative splicing of HC genes

Alternative splicing for HC gene loci was evaluated. All predicted transcript structures were considered and identical exon-intron structures detected from different samples were collapsed to generate non-redundant gene variant counts.

S7.4 Analysis of premature termination codons (PTCs) in alternative spliced transcripts

Premature termination codons (PTCs) are stop codons that occur before the authentic stop codon and can generate truncated proteins or trigger nonsense mediated decay (NMD). PTCs can activate NMD by virtue of their distance from the 3' end of the transcript (Long 3'UTR) or if they are >50 nucleotides upstream of an exon/exon splice junction[43-45]. OrfPredictor[32] derived CDS information was transferred to structures of 51,388 alternative spliced transcripts. In total, 38,204 transcripts were found to contain a stop-codon located in the last 3' exon and 13,134 transcripts were found to contain at least one non-coding 3' exon,

respectively. For 9,877 transcripts, a PTC resided >50nt upstream of the exon/exon junction indicating that the corresponding transcripts could be potential candidates for NMD. These PTC+ candidate transcripts were found for 4,469 HC gene loci. Recently, it has been suggested that many transcripts in *Arabidopsis* with retained introns represent partially or incompletely spliced transcripts and are not subject to NMD. Retention of an intron is likely to generate PTCs due to the relative UA-richness of plant introns. Assuming that transcripts with retained introns are also not sensitive to NMD in barley, we identified these transcripts. For 5,286 transcripts, the PTC-containing exon can be also explained by two or more spliced exons of another transcript of the same gene and thus might be related to retained introns. Excluding these intron retained transcripts a final PTC set of 4,591 transcripts could be identified which are located at 2,466 HC gene loci.

The rules for classifying transcripts as PTC+ were also applied to transcripts without evidence for alternative splicing. We found only 1% (130) HC structures to be classified as PTC+ candidates, thus the percentage and number of genes that are classified as alternatively spliced and PTC+ [4591 genes (9%); see Table 2] significantly deviate. We therefore concluded that PTC+ is significantly correlated to alternatively spliced transcripts.

S7.4.1 Analysis of premature termination codons (PTCs) in different barley cultivars

RNA-seq data sets of barley varieties Morex, Barke, Betzes, Optic, Quench, Sergeant and Tocada (Supplemental Note 8.2) were used to analyze effects of variation in different barley cultivars. RNA samples are from same barley tissues (4 day germinating embryo, radicle and coleoptile), passed same experimental procedure (Sections S7.2.1 and S7.2.2) and showed comparable sequencing depth among the different cultivars used (Table S27). These samples represent one single experimental condition that is comparable to the EMB data set used for gene calling (Table S20).

First, RNA-seq reads of each genotype were mapped against the library based repeat-masked assembly of barley cv. Morex 50x WGS using Bowtie (v0.12.7) and TopHat (v1.4.0) with default settings (Figure S23). Cufflinks (v1.3.0) was used to assemble the mapped reads and open reading frames (ORFs) were predicted for all transcript assemblies by using OrfPredictor. Strand information was assigned to RNA-seq transcript structures and transcripts without predicted open reading frame were not used for further analysis. Cuffcompare (v1.3.0; Parameters: -V -C -r -s) was used to cluster and compare the predicted gene models of the barley cultivar RNA-seq data sets with the reference barley annotation (Section S7.1).

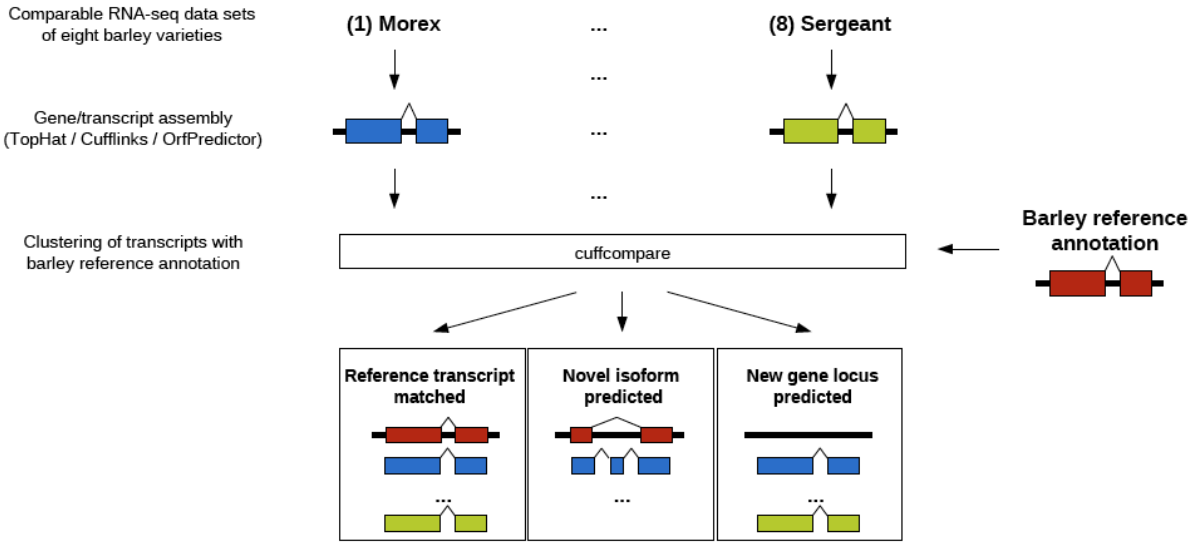


Figure S23: Workflow for comparative analysis of barley transcript structure in different barley cultivars.

No pronounced differences of the predicted gene structures were observed between the different barley cultivars (Table S27). Between 77% and 88% of the available RNA-seq reads were aligned to the reference assembly of barley cv. Morex and comparable numbers of gene and transcript structures were predicted. On average 81% of predicted gene models based on RNA-seq of different barley varieties are in very high accordance with the reference barley annotation. Minor fractions of the predicted cultivar gene models represented novel transcript isoforms (5% in average) and newly identified intergenic gene models (0.66% in average). Thereby, the Morex sample showed reduced detection rate of novel intergenic structures (0.38%) as expected because the reference gene structures were also based on RNA-seq data of that barley variety. However, some of the cultivar gene models (13 % in average) showed inconsistencies to the reference annotation (e.g. possible pre-mRNA fragments, polymerase run-on fragments, exonic overlap with reference on the opposite strand, multiple ambiguous classification) that were probably caused by the relatively low sequencing depth of the barley variety data sets compared to the multi-tissue data set used for creating of the reference annotation.

Table S27: Mapping of RNA-seq reads of seven different cultivars to barley cv. Morex WGS contigs, gene assembly and comparison to barley reference annotation.

	Morex	Quench	Optic	Barke	Tocada	Betzes	Sergeant
Data sets							
RNA-seq reads	23,250,889	26,946,706	23,252,182	25,663,186	23, 868,881	22,204,022	24,408,462
Sequence (Gb)	1.74	2.02	1.74	1.92	1.79	1.67	1.84
Mapping statistics							
Mapped RNA-seq reads	18,310,515	20,682,627	18,614,359	20,249,898	19,631,696	18,335,232	19,818,159
Mapped RNA-seq reads (%)	79	77	80	79	82	83	81
Mapped Sequence (Gb)	1.37	1.55	1.40	1.52	1.47	1.38	1.49
Gene assembly and structure prediction statistics							
Predicted gene loci	39,463	40,206	40,255	40,584	39,498	37,397	40,193
Alternative spliced gene loci	2,719 (7%)	2,802 (7%)	2,928 (7%)	2,897 (7%)	2,864 (7%)	1,749 (5%)	2,929 (7%)
Predicted transcripts	42,659	43,529	43,698	44,030	42,896	39,351	43,655
Comparison of predicted transcripts to the barley working gene set (including HC and LC gene loci)							
Matching reference annotation*	34,705 (81%)	34,951 (80%)	34,944 (80%)	35,095 (80%)	34,540 (81%)	32,599 (83%)	34,947 (80%)
Potentially novel isoforms**	2,003 (5%)	1,998 (5%)	2,191 (5%)	2,201 (5%)	2,031 (5%)	1,370 (3%)	2,160 (5%)
New intergenic transcripts***	164 (0.4%)	362 (0.8%)	292 (0.7%)	343 (0.8%)	267 (0.6%)	214 (0.5%)	359 (0.8%)
Other classifications****	5,787 (14%)	6,218 (14%)	6,271 (14%)	6,391 (15%)	6,058 (14%)	5,168 (13%)	6,189 (14%)
Comparison of predicted transcripts to the HC gene set							
Matched HC gene loci	16,543	16,651	16,704	16,793	16,457	16,447	16,749
Matched HC transcripts	18,868	19,089	19,177	19,232	18,853	18,317	19,188
Normally spliced transcripts	5,779	5,784	5,846	5,911	5,711	5,899	5,890
Alternative spliced transcripts	12,448	12,671	12,666	12,665	12,497	11,836	12,667
PTC+ transcripts	641	634	665	656	645	582	631
Novel alternative transcripts	1,473	1,433	1,607	1,607	1,491	1,047	1,564
Alternative spliced transcripts	1,282 (87%)	1,207 (84%)	1,343 (84%)	1,353 (84%)	1,237 (83%)	943 (90%)	1,307 (84%)
PTC+ transcripts	190 (13%)	226 (16%)	264 (16%)	254 (16%)	254 (17%)	104 (10%)	257 (16%)
Normalized comparison of predicted transcripts to the HC gene set (normalized to max. number of mapped reads)							
Matched HC gene loci	18,686	16,651	18,560	17,152	17,338	18,553	17,480
Matched HC transcripts	21,313	19,089	21,308	19,643	19,862	20,662	20,025
Normally spliced transcripts	6,528	5,784	6,496	6,037	6,017	6,654	6,147
Alternative spliced transcripts	14,061	12,671	14,073	12,936	13,166	13,351	13,220
PTC+ transcripts	724	634	739	670	680	657	659
Novel alternative transcripts	1,664	1,433	1,786	1,641	1,571	1,181	1,632
Alternative spliced transcripts	1,448 (87%)	1,207 (84%)	1,492 (84%)	1,382 (84%)	1,303 (83%)	1,064 (90%)	1,364 (84%)
PTC+ transcripts	215 (13%)	226 (16%)	293 (16%)	259	268 (17%)	117 (10%)	268 (16%)

* the transcripts a complete match of the intron chain of a reference transcripts or completely contained in a reference structure (cuffcompare classification codes '=' and 'c')
** the transcript is a potentially novel isoform (at least one splice junction is shared with a reference) (cuffcompare classification code 'j')
*** the transcript is new (intergenic) (cuffcompare classification code 'u')
**** the transcript can't be described by 1-3 (eg. classified as possibly pre-mRNA fragment, anti-sense transcript, possibly polymerase run-on fragment, multiple ambiguous classifications) (cuffconpare classification codes 'e', 'i', 'o', 'p', 'r', 'x', 's' and '.')

All used barley cultivar samples showed comparable overlap to the barley reference annotation with some small variances which might be caused by slightly different sample sizes indicating that no pronounced differences in expression and splicing patterns exist between different barley cultivars (Figure S24).

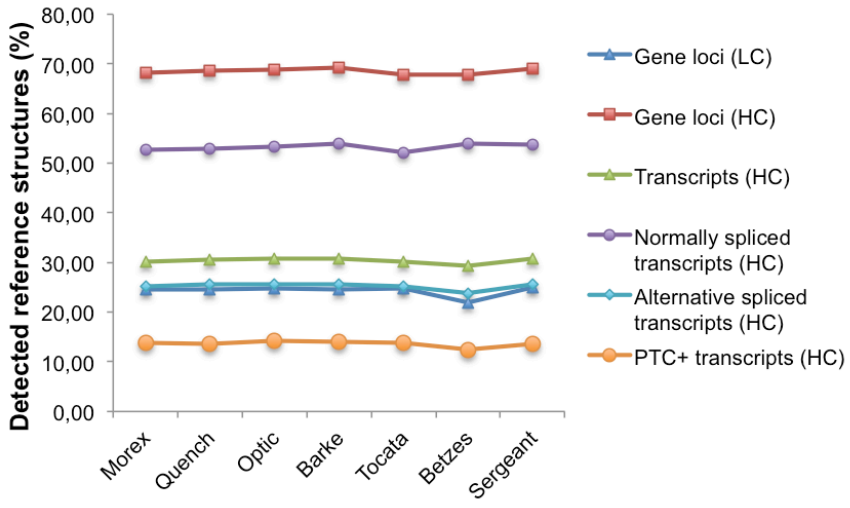


Figure S24: Distribution of barley reference genes and transcripts that were detected by RNA samples of seven barley varieties.

S7.4.2 Analysis of different barley gene splice variants on sequenced BAC clones

Nucleotide sequences of the 62,427 predicted HC transcripts (24,243 HC gene loci) that are located in the assembly of barley cv. Morex 50x WGS were mapped to BAC clones (766Mb) by using BLAT v34 (Parameter: -t dna -q rna -minIdentity 95 -maxGap=0 -fine). Thereby, a minimum alignment identity of 95% and complete (100%) mapping of the transcript sequence was required. Only the mapping with highest alignment identity was used in case of multiple mapping locations of a barley transcript.

In total, mapping positions for 13,098 (20%) transcripts on BAC sequences were found . Similar percentages were found for alternative spliced [10,225 out of 51,338 (20%)] and PTC+ transcripts [905 out of 4,591 (20%)]. Furthermore, the positioning of the stop codon that terminates the predicted open reading frame was evaluated for these alignments. In >99% of the alignments, the position of the predicted coding sequence was confirmed by the BAC clone. This confirmed that the predicted stop codons (especially the stop codons classified as premature termination codons) are not artificially introduced and caused by assembly errors in the WGS assembly (Figure S25).

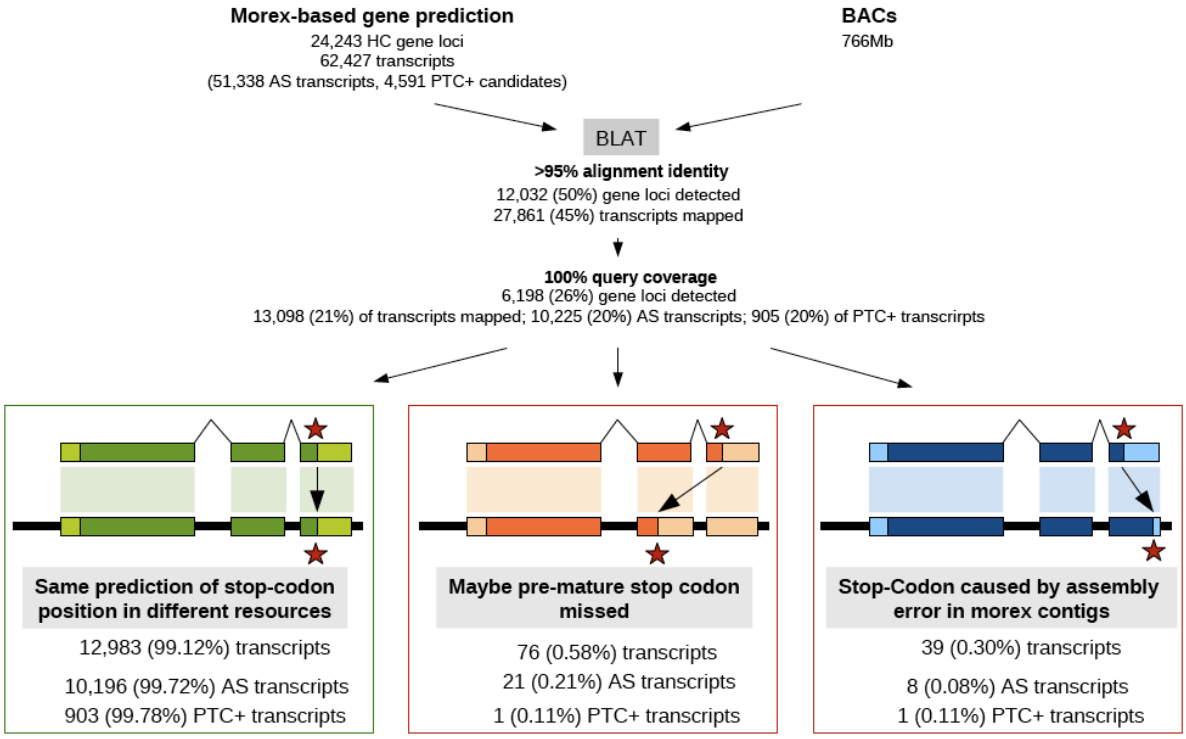


Figure S25: Comparison of barley cv. Morex-based gene prediction to BAC clones

S7.4.3 Barley PTC+ gene splice variants in relation to gene family size

To test whether barley PTC+ barley transcripts occur more or less frequently in expanded barley gene families (as determined in S 7.1.3.2) we determined the gene ratios in OrthoMCL clusters that contain and lack PTC+ barley HC genes. The ratios were computed for all clusters containing at least one barley HC gene and at least one gene from *Sorghum*/rice/*Brachypodium* (Sb/Os/Bd). In case of multiple species represented in the cluster, analysis has been carried out separately for each barley- Sb/Os/Bd combination. We plotted the frequency of gene copy number ratios in bins for all OrthoMCL clusters satisfying the criteria in Figure S26, separately for barley PTC-containing and PTC-free clusters and species combination**.** Our data indicate that barley PTC+ transcripts occur more frequently in expanded barley gene groups, in comparison to OrthoMCL clusters without barley PTC-related transcripts (Figure S26).

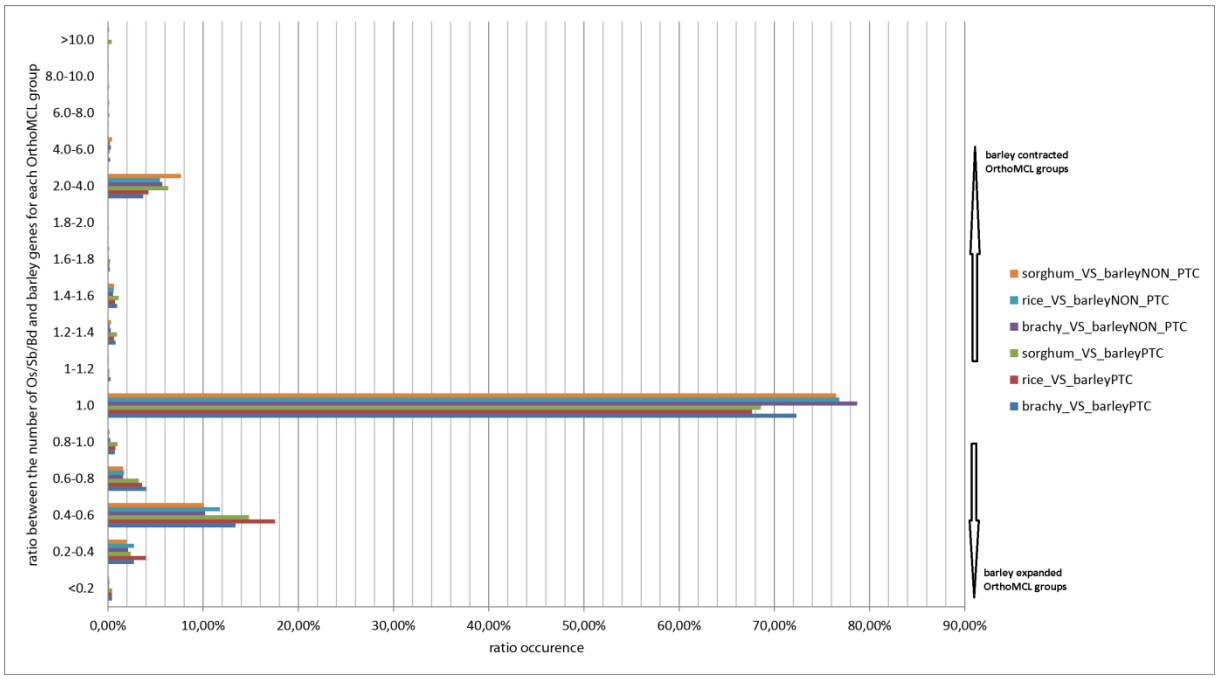


Figure S26: Expanded and reduced gene families in barley and their relation to the presence of PTC+ transcripts.

The ratio is defined by the number of Sb/Os/Bd genes in the particular OrthoMCL cluster versus the number of barley HC genes present in the same cluster. Values greater than 1.0 account for a reduced copy number in barley whereas values smaller than 1.0 indicate expanded copy numbers in barley. A ratio of 1.0 indicates that the number of e.g. Brachypodium and barley genes in that cluster is equal (balanced cluster). NON_PTC: gene(s) with AS but without a premature termination codon, PTC: gene(s) with AS and containing a premature termination codon.

Supplemental Note 8. Analysis of sequence variation

S8.1 Single nucleotide variations in whole genome shotgun data. The whole genome shotgun sequence assemblies and the genetically anchored sequence resources offered the unique possibility of performing a genome-wide survey for single nucleotide variation (SNV) between different barley accessions. Mining for variations was always done between two cultivars. The *de novo* assembly of Morex served as reference for SNV mining in the cultivars Barke, Bowman, Igri, Haruna Nijo and a single accession of wild barley (*H. spontaneum*). Illumina paired end reads of each accession (Table S5) were mapped to the reference assembly using BWA (version: 0.5.9-r16)[46]. Amplification artifacts (Duplicons) have been removed and putative variations have been called using samtools (version 0.1.18 r982)[47]. The raw set of potential variations has been generated using vcftools (version 0.1.17)[48]. Artificial variations caused by inserted Ns have been removed and remaining variations were filtered using the vcfutils module of bcftools. We are aware of the draft nature of the *de novo* assembly of 'Morex', thus limited sequence accuracy in regions of low read coverage or where leading into repetitive DNA could easily result in false positive SNV calls. In order to reduce this risk and to identify critical positions we mapped all available paired-end reads of the reference cultivar 'Morex' to its whole genome assembly. All nucleotide positions containing variations or any kind of ambiguous read-support have been discarded from further comparison to the other cultivars. 12,065,380 positions in Morex were thus discarded from further analysis. The positions discarded during this step reflect the characteristics of the draft nature of the assembly caused by the complexity of the barley genome. The majority of masked Morex/Morex-polymorphic sites are located in short contigs or/and towards the end of contigs (Figure S27). 67% of those positions are in contigs shorter than the N50-length and 66% of the positions are at a distance of less than 200 bases from the ends of a contig. In contrast, if comparing high-confidence variations between Bowman and Morex, only 17% of such variations are located in contigs shorter than the N50-length and only 27% are at a distance of less than 200 bases from the contig end (Figure S27). Importantly, if looking at the physical/genetic distribution of masked Morex/Morex polymorphic sites a more or less even distribution along the chromosomes could be observed (Figure S28). All other reported variations between any accession and Morex or Bowman, respectively (Tables S28-S30), have been identified in the remaining un-masked positions. In order to further minimize the ratio of false positive variations, SNV positions were filtered for a VCF-tools generated phred-like quality score of 50 or more. Additionally, all reads of the mapping cultivar must support the alternative allele unambiguously.

For the visualization of genome-wide SNV frequency between survey-sequenced culivars/accessions and the reference 'Morex', high-confidence SNV were counted for 50 Kb intervals of concatenated and anchored reference sequence. The maximum values of SNV /

50 Kb were used to set the scale in Figure 3 of the main manuscript and differed for the 5 tested samples: Barke = 57, Bowman = 74, Haruna Nijo = 48, Igri = 58, *H. spontaneum* = 57. The global visualization of SNV frequency in cultivated and wild barley accessions allowed to visually depict obvious patterns of variation and showed for all samples including *H.spontaneum* an erosion of diversity in the peri-centromeric regions. For cultivar Bowman (Figure 3, red SNV histogram) at least two contigous regions (red arrowheads) lacked SNV diversity with cv. 'Morex'. This highlighted regions that most likely are identical by descent due to a partially shared pedigree (http://genbank.vurv.cz/barley/pedigree/). Other regions showed clear genotype-specific erosion of SNV frequency (green arrowheads) which could be a function of breeding history. Small regions were identified that exhibited erosion of diversity in all compared cultivars but not in wild barley (black arrowhead) which could hint at regions with domestication dependent erosion of SNV frequency, however, a larger number of samples will need to be analysed in the future to provide statistical power for the detection of domestication related changes of SNV frequency.

In order to make sure that different variation-densities in specific regions are not caused by lack of mapping coverage, we correlated the number of observed variations per 50k-interval of anchored WGS-contigs with the average coverage of the same interval. Using Spearman' rank correlation we observed a week negative correlation between observed variations and coverage (Figure S29). The explanation is most probably our strict filtering of positions with ambiguous read support, which is more likely to happen at higher coverage. However, this observed trend is of minor relevance to the prominent regions of lack of erosion of SNV frequency (Figure 3). A visual inspection of the mapping coverage for each line in those regions does not reveal any particular variation in coverage over successive 50k-intervals. (Figure S30)

Sensitivity/False positives

To assess the sensitivity of the SNV survey pipeline, we compared our results to a set of 3,972 SNV-markers that were experimentally mapped between genotypes Morex and Barke on the Illumina iSelect[28]. Sequences from the iSelect manifest file carrying the experimentally assayed SNV information were aligned to the Morex assembly for determination of SNV positions in regard of the Morex WGS contigs. The sensitivity was calculated by the fraction of iSelect SNVs that could be identified by our pipeline. Of all 3,972 initial sequences 3,934 could be anchored to a Morex WGS contig. 3,457 SNVs (87%) were consistently detected using our SNV detection pipeline (see above). 308 SNVs (8%) had to be discarded during our filtering steps for unusable positions of the Morex WGS assembly, ambiguous read-support or score and 207 (5%) had not been detected at all. Thus sensitivity can be estimated by 3,457/3,972 = 0.8703 (87%)

For estimating the percentage of falsely predicted SNVs, 300 variations between Morex and Barke were randomly chosen from specific subsets of positions: 100 positions within a predicted exon with phred-like quality score of 100 or more; 100 positions within exons with phred-like quality score between 50 and 100; 100 random positions. PCR Primers were automatically designed in batch mode using BatchPrimer3[49] for each of the 300 variations and amplification products have been cycle-sequenced using Big Dye Terminator v3.0 (Applied Biosystems, Foster City, USA) chemistry and a capillary sequencing device (ABI 3730xl, Applied Biosystems, Foster City, USA). For 249 re-sequenced variations high quality sequence reads were available for Morex and Barke. Nineteen of the assayed SNV positions showed Barke reads supporting both alleles. Fifteen of these 19 were reassessed in the WGS re-sequenced alleles showing ambiguity for Barke as well. As a consequence the filtering step for ambiguous read-support was introduced to our SNV pipeline (see above). For the remaining 230 re-sequenced variations, 220 (96%) could be validated.

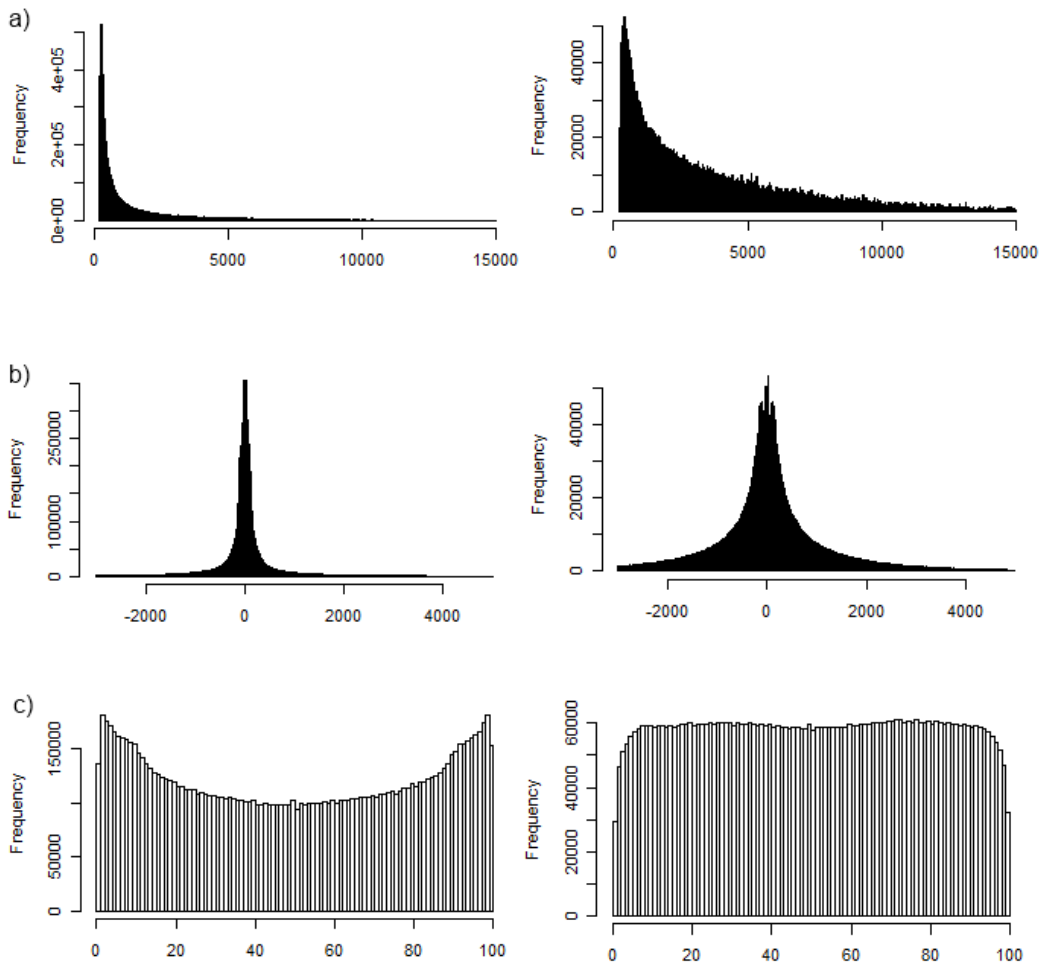


Figure S27: Charateristics of low confidence (reference genotype/reference genotype) and high confidence (test genotype/reference genotype) SNV data-mined from whole genome survey-sequencing data.

Left panels illustrate the characteristics of low confidence SNV (Morex/Morex) whereas right panels illustrate characteristics of high confidence (Bowman/Morex) SNV positions. a) SNV frequency in context of contig length. b) Minimal distance of SNV position in respect to a contig end (negative values refer to 3' end, positive values refer to the 5' end, respectively). c) Relative positions of SNV on contigs illustrated as percent of contig-length. Low confidence SNV are predominantly located towards ends of WGS contigs and are preferentially located on small WGS contigs, both indicating an association to repetitive DNA of barley.

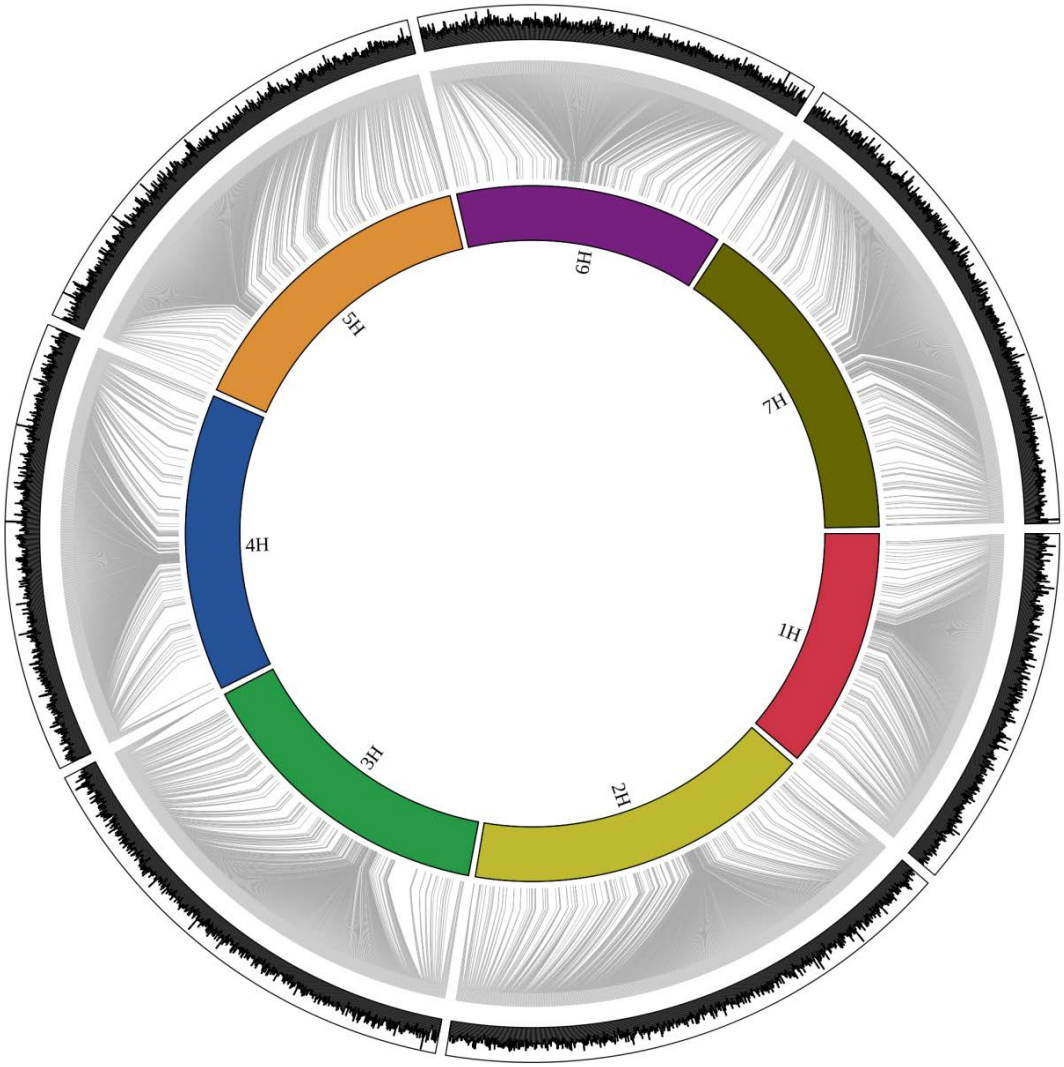


Figure S28: Histogram of genome wide distribution of low confidence (Morex/Morex) SNV positions.

The inner circle indicates the seven chromosomes of barley . Black lines connect genetic and physical maps of barley chromosomes. The outer ring histogram displays the positions per 50,000 base-pairs (integrated to the genetically anchored physical map) showing variations between the Morex WGS contig sequences and mapped sequence reads of paired-end Morex WGS reads (28-fold genome coverage). These low-confidence SNV positions were excluded from further SNV data mining in genomic survey-sequencing and RNA-seq data.

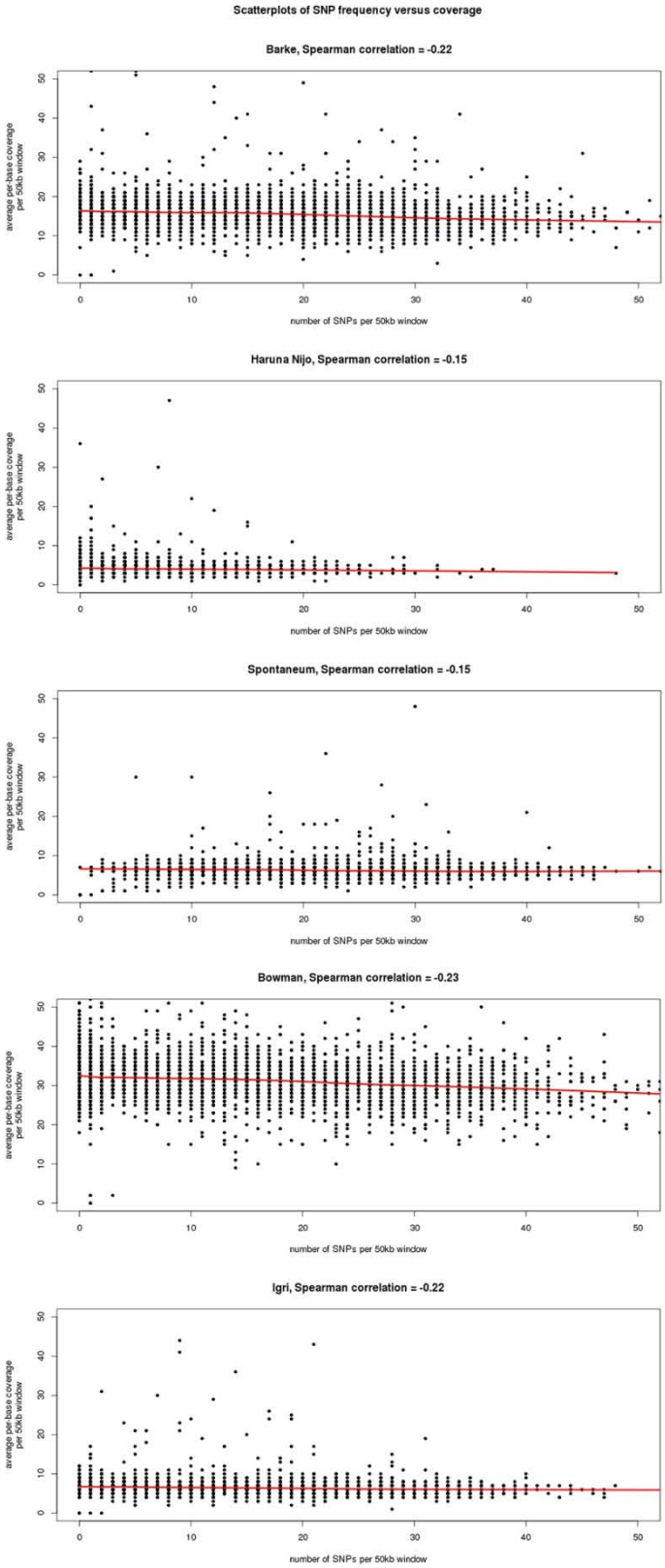


Figure S29: Scatterplots of SNV frequency versus survey sequencing coverage.
For each individual analysed accession / cultivar a weak negative correlation of observed variations per 50 Kb interval of anchored WGS contigs with the average coverage of the same interval can be observed.

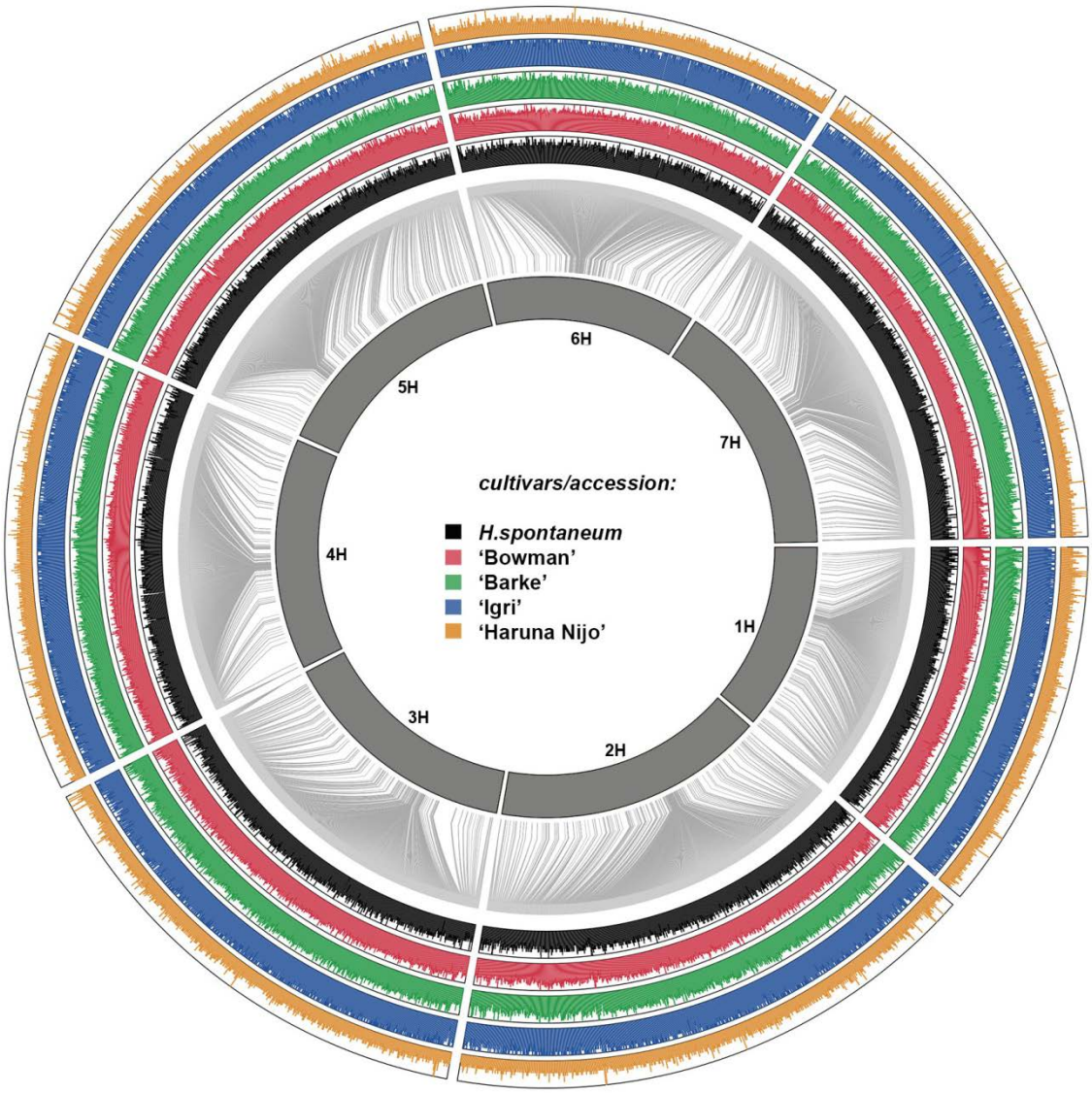


Figure S30: Survey sequence coverage of anchored WGS contigs of reference 'Morex' per mapped cultivar/accession.

The five outer rings of histograms give the average coverage of mapped reads per 50 Kb interval of physically mapped contigs of the WGS assembly of Morex. The scale for each line is defined individually from 0 to third quartil plus 1.5 interquartil-range. (Wild barley=9, Bowman=44, Barke=22, Igri=7, Haruna Nijo=9).

Table S28: Number of SNV between barley accessions and the reference cultivar 'Morex' in WGS contigs assigned to chromosome-arm bins

	Anchored sequence	Number of high confidence SNV				
Arm	bp	Bowman	Barke	Haruna Nijo	Igri	*H. spontaneum*
1H	125,155,387	446,844	436,529	100,510	405,603	842,153
2HS	59,040,496	162,046	231,124	65,281	210,407	387,863
2HL	81,391,919	274,721	373,123	103,742	400,848	556,560
3HS	56,813,423	120,639	151,303	45,823	133,880	390,418
3HL	81,619,445	377,344	367,185	81,083	363,949	590,575
4HS	59,260,128	85,867	87,175	19,935	66,815	297,747
4HL	74,665,245	214,259	207,088	56,146	197,346	535,652
5HS	43,847,007	304,071	169,848	67,043	167,767	205,917
5HL	87,107,404	345,349	399,261	130,617	362,650	597,399
6HS	53,389,480	167,708	212,331	38,341	113,624	307,070
6HL	64,794,614	228,948	332,376	52,262	222,310	457,594
7HS	68,990,988	190,199	357,454	109,645	337,971	501,147
7HL	67,526,997	194,525	326,533	78,294	274,776	521,035
Σ	923,602,533	3,112,520	3,651,330	948,722	3,257,946	6,191,130
N/A	945,045,622	1,521,891	1,744,730	98,019	1,307,253	2,516,031

*= total amount of WGS data anchored to chromosomes, N/A= not assigned to any chromosome

Table S29: Number of SNV in exons of the HC gene set and assigned to chromosome-arm bins

	Anchored exon sequence	Number of high confidence SNV				
Arm	bp	Bowman	Barke	Haruna Nijo	Igri	*H. spontaneum*
1H	10,562,810	30,164	26,470	10,410	24,433	43,246
2HS	4,874,658	12,182	13,922	6,311	13,267	20,368
2HL	7,510,525	21,070	21,408	9,524	22,631	30,573
3HS	4,688,890	8,805	10,970	5,147	9,721	19,362
3HL	7,902,136	24,914	23,122	10,414	24,999	35,480
4HS	4,011,013	6,906	5,858	2,161	4,884	13,440
4HL	5,347,280	11,461	9,183	4,473	9,604	20,486
5HS	3,132,610	9,162	9,298	2,873	9,094	11,204
5HL	9,721,752	17,253	29,257	12,043	29,532	45,655
6HS	4,363,171	14,770	12,408	5,429	11,249	18,188
6HL	5,209,963	11,903	14,068	4,138	11,128	21,657
7HS	6,764,918	12,823	22,828	10,068	22,938	34,482
7HL	6,117,637	11,148	20,108	5,224	13,757	30,126
Σ	80,207,363	192,561	218,900	88,215	207,237	344,267
N/A	9,686,391	17,392	17,368	4,462	15,341	26,611

*= total amount of WGS data anchored to chromosomes, N/A= not assigned to any chromosome

Table S30: Number of FPC anchored SNV in cultivars and wild barley

SNV	Bowman	Barke	Haruna Nijo	Igri	Spontaneum
all	4,634,411	5,396,060	1,046,741	4,565,199	8,707,161
exon-based	209,953	236,268	92,677	222,578	370,878
all FPC anchored*	1,108,380	1,312,832	410,444	1,166,770	2,139,293
exon FPC anchored**	86,410	99,513	45,062	94,318	157,381
Read Mapping Coverage	24	12	5	5	5

*=data visualized in Figure 3, **= data visualized in Figure S31

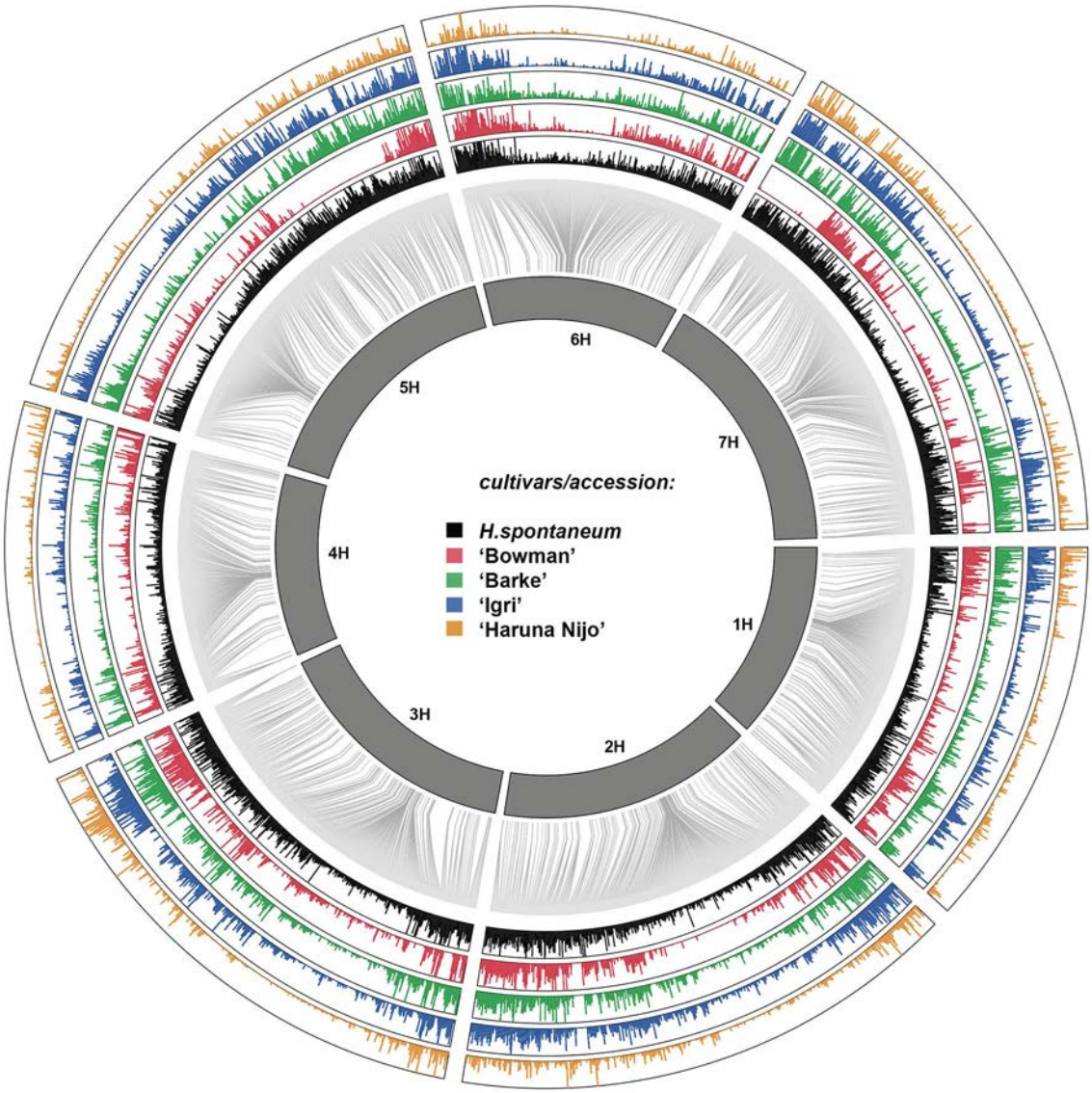


Figure S31. Exon-based SNV frequency in WGS data from barley cultivars and wild barley.

The 5 genotypes are *H. spontaneum* (black), Bowman (red), Barke(green), Igri(blue) and Haruna Nijo (yellow). The visualized reference consists of all parts of contigs of the WGS Morex assembly that can be anchored to the physical map and are predicted to be transcribed to an exon (35,818,725 bp). For each chromosome all sequences are concatenated and histograms show the number of single nucleotide variations between each line and Morex for intervals of 10 Kb. The histograms are scaled between 0 and 37.5 variations per 10 Kb. 37.5 is the third quartile plus 1.5 interquartile-range of all observed variations per 10 Kb of the most diverse accession, *H. spontaneum*. Connectors (gray) show the relation between the genetic map (centi-morgan) and anchored bases.

S8.2 SNV frequencies in RNA-seq data

S8.2.1 Barley material and RNA extraction

Seeds were germinated from each of 9 spring barley varieties (Barke, Betzes, Bowman, Derkado, Intro, Optic, Quench, Sergeant and Tocada) and Morex on filter paper moistened with sterile water in petri dishes. Following 4 days incubation in the dark at 20^{o}C, developing radicle (c. 10-20 mm) and embryo tissues were dissected and flash frozen in liquid nitrogen. Total RNA was extracted from c. 200 mg mixed tissue from each genotype using 2 ml TriReagent (Sigma-Aldrich) as recommended, with two additional phenol-chloroform purification steps. RNAs were quality checked using the RNA 6000 Nano kit as instructed on a 2100 Bioanalyzer (Agilent), with all RNA samples having an RNA Integrity Number (RIN) >8.

S8.2.2 Transcriptome sequencing

Total RNA samples were submitted to The Sir Henry Wellcome Functional Genomics Facility, University of Glasgow, for RNA-seq processing using standard recommended Illumina GAII procedures. For each sample, one lane of single-end 54 nt or 76 nt RNA-seq was performed, generating between c. 6M and 27M reads respectively (Table S31).

Table S31: Read numbers and read lengths for all samples used

sample	read numbers untrimmed	read numbers trimmed	mean trimmed read length
Barke	25,663,186	25,538,142	72
Betzes	22,204,022	22,118,902	72.5
Bowman	7,257,869	7,251,430	48.8
Derkado	5,932,230	5,930,692	39.8
Intro	6,066,180	6,064,803	40.4
Morex	26,664,480	26,292,853	68.9
Optic	23,252,182	23,164,817	71.5
Quench	26,946,706	26,795,748	69.8
Sergeant	24,480,462	24,288,466	72.1
Tocada	23,868,881	23,764,572	72.2
Total	165,671,718	164,917,572	62.8

S8.2.3 Read preparation

The raw Illumina reads were obtained in two separate sequencing runs carried out more than a year apart, and as a result reads obtained in the earlier run were significantly shorter (54 nt) and less numerous than those from the latter run (76 nt) as Illumina read lengths and read volumes continue to increase significantly. All raw reads were quality trimmed to a quality score (*phred* equivalent) of 20 or greater. This removes stretches of sequence from either read end where the base qualities are consistently below the threshold. This step

significantly reduces the risk of obtaining false positives during SNV discovery. Read numbers and length statistics are detailed in Table S31.

S8.2.4 Read mapping

The trimmed Illumina reads were mapped against 142,763 exons predicted from the Morex v.3 genomic assembly (supplemental note 7) using the Bowtie read mapper version 0.12.7 64-bit[50] at following parameter settings:

--phred64-quals (Input qualities are ASCII chars equal to the Phred quality plus 64)
-a (Report all valid alignments per read)
--best (reported singleton alignments are "best" in terms of stratum)
--strata (report only those alignments that fall into the best stratum)
-v 1 (Report alignments with at most 1 mismatch)
--sam (Print alignments in SAM format)
-p 4 (Launch 4 parallel search threads)
-a (Report all valid alignments per read)

Samples were mapped separately and then merged with the samtools v. 0.1.18[47] 'merge' command into a combined mapping file which contained the mapping data from all 9 samples other than Morex. Duplicate reads were removed with the samtools 'rmdup' command, which leads to a substantial reduction of the false positive rate.

S8.2.5 SNV discovery and filtering

Raw SNVs were called using samtools mpileup and bcftools v. 0.1.17 using the following command line settings:
samtools mpileup -D -u -f <reference data> <bam file> | bcftools view -vcg - > SNVs.vcf

Raw SNV calls were filtered with samtools' vcfutils.pl script to exclude SNVs that had less than two instances of the minor allele. Custom written code was used to also filter out any low confidence SNVs based on the phred-like quality score in the VCF file, eliminating all SNVs with a quality score of less than 50.

S8.2.6 Validation of Illumina mappings

Accurate read mappings are of paramount importance for the accuracy of any downstream analysis, in particular SNV discovery. To ascertain the validity of the Illumina read mappings the SNV discovery was based on, custom written Java code was used to compare genotype

calls extracted from the Illumina read mappings with existing genotype data for the same samples obtained with the Illumina Golden Gate genotyping assay (BOPA - barley oligo pooled assay[20]). Results were in the range of approx 96-99% agreement (Table S32), depending on the sample. Mappings were also inspected visually using the Tablet assembly viewer[51] to check for obvious indications of misassembly.

Table S32. Validation rates for the Illumina read mappings by sample, compared to existing benchmark genotype data from the Illumina Golden Gate genotyping assay

sample	FALSE	TRUE	no NGS data	no benchmark data	% correct
Betzes	18	945	75	13	98.13
Bowman	3	420	559	69	99.29
Intro	2	168	711	170	98.82
Optic	6	478	80	487	98.76
Sergeant	27	956	59	9	97.25
Tocada	34	940	69	8	96.51
Derkado	1	327	720	3	99.70
Barke	22	962	62	5	97.76
Quench	7	474	76	494	98.54

Table S33: SNV frequency in RNA-seq data of diverse barley cultivars in exons assigned to chromosome-arm bins

	Anchored exon sequence	Number of high confidence SNP									
Arm	bp	Bowman*	Barke	Betzes	Derkado*	Intro*	Optic	Quench	Sergeant	Tocada	rnaSeqAllLines
1H	10,562,810	969	4,401	4,149	559	363	4,283	4,442	4,462	4,209	5,013
2HS	4,874,658	294	2,162	1,797	230	159	1,933	1,969	2,008	1,834	2,073
2HL	7,510,525	633	3,904	3,565	473	324	4,119	4,238	4,083	3,594	3,895
3HS	4,688,890	232	1,501	1,403	163	139	1,464	1,529	1,474	1,353	1,498
3HL	7,902,136	514	3,376	3,019	399	312	3,119	3,160	3,006	3,302	3,356
4HS	4,011,013	243	982	871	135	85	1,043	861	1,180	903	994
4HL	5,347,280	448	1,739	1,869	167	128	1,735	1,747	1,914	1,534	1,763
5HS	3,132,610	215	1,260	710	112	147	1,254	1,334	1,309	1,221	1,223
5HL	9,721,752	403	4,267	3,584	500	338	4,043	4,293	4,014	3,723	3,608
6HS	4,363,171	320	1,696	1,136	199	125	1,621	1,733	1,525	1,619	1,702
6HL	5,209,963	461	2,511	1,794	266	234	2,303	2,456	2,435	2,295	2,661
7HS	6,764,918	315	2,788	2,707	420	249	2,922	2,967	3,112	2,924	2,710
7HL	6,117,637	256	2,188	1,631	241	165	2,158	2,283	1,634	1,852	2,135
Σ	80,207,363	5,303	32,775	28,235	3,864	2,768	31,997	33,012	32,156	30,363	32,631
N/A	9,686,391	711	3,072	2,604	549	463	3,211	3,223	3,036	3,076	4,065

*= low coverage, N/A= not assigned to any chromosome, data used for visualization in Figure S32

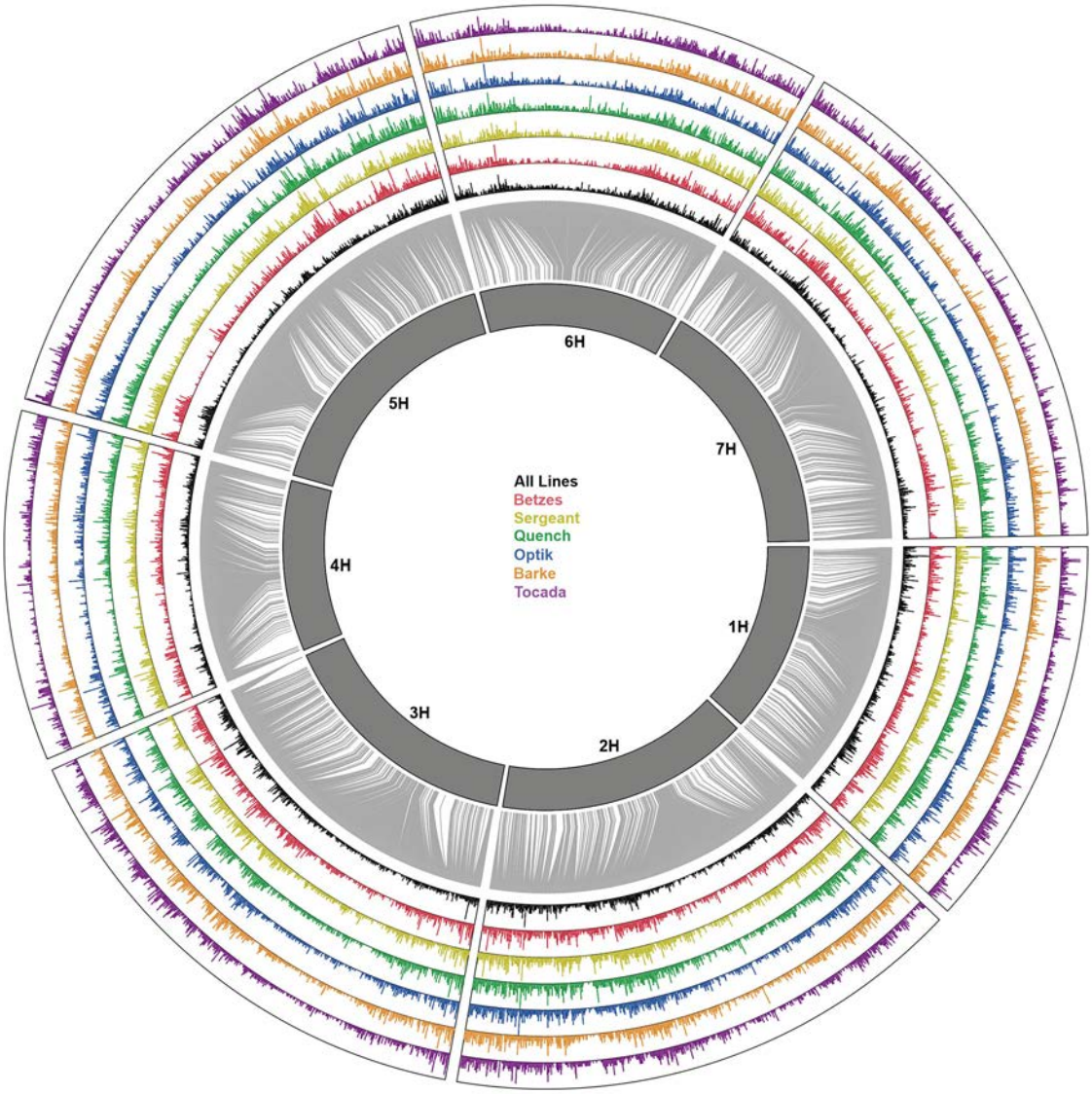


Figure S32: RNA-seq based SNV frequency in barley cultivars.

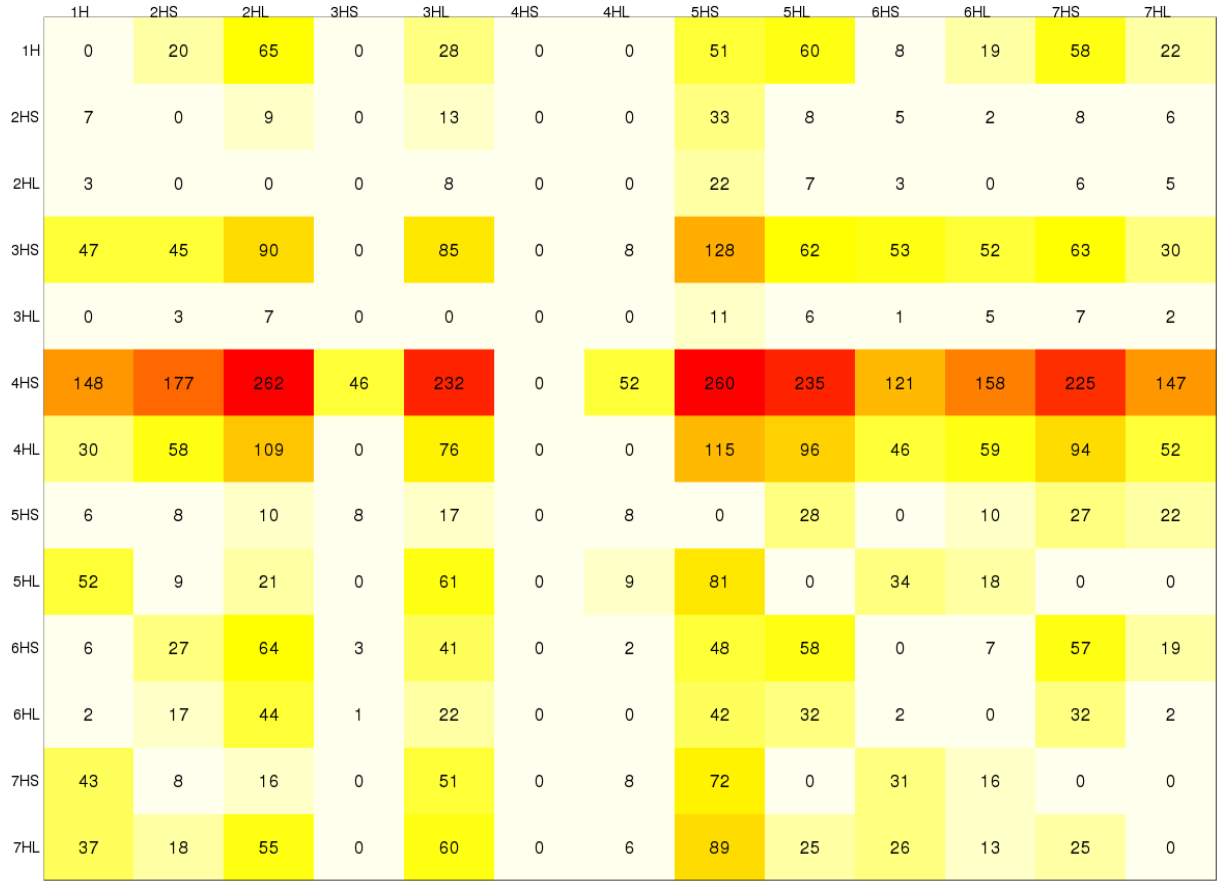
The inner histogram (black) shows the non-redundant SNV detected over all genotypes. Individual cultivar RNA-seq based SNV frequency is displayed in further outer concentric cirles of colored histograms: genotypes are Betzes (red), Sergeant (yellow), Quench (green), Optic (blue), Barke (orange) and Tocada (purple). SNV were determined by mapping RNA-seq reads to predicted exons of the Morex reference WGS assembly contigs associated to the genetically anchored physical map of barley (35,818,725 bp). For each chromosome all anchored sequences were concatenated and histograms show the number of single nucleotide variations between each cultivar and Morex for intervals of 10 Kb. Connectors (gray) show the relationship between genetic map (cM) and anchored bases. Data for cultivars Bowman, Derkado and Intro are not shown, due to a lack of coverage.

S8.3 SNV frequency analysis per chromosome

After obtaining genome-wide SNV information on the basis of survey-sequencing and RNA-seq data we observed a reproducible trend for chromosome 4H overall exhibiting lower SNV frequency if compared to all other barley chromosomes. This could not be expected on the basis of the size of the chromosome (4H is the third smallest chromosome). A similar but not as pronounced trend was also observed in the single *H. spontaneum* accession utilized for re-sequencing.

SNV frequencies in exons were obtained from re-sequencing data of cultivars 'Bowman', 'Barke', Igri', Haruna Nijo' and *H. spontaneum*. For each genotype, pairwise comparisons of SNV frequencies between chromosome arms were performed. First, SNVs were averaged over 50kb sized regions. For each pair of chromosome arms (A,B) the mean SNV frequency on all regions of A (M_A) was compared to the mean SNV frequency on all regions of chromosomes B (M_B). The non-parametric Wilcoxon rank-sum test was applied to test the null-hypothesis that the M_A is greater than M_B. Let $p_g(A,B)$ denote the p-values of this test for a pair of chromosomes (A,B) in a genotype g. Then for each pair of chromosomes we can define the sum of logarithmized p-values over genotypes $P(A,B)= \sum_g -\log10(p_g(A,B))$ as a measure of SNV scarcity on chromosome arm A as compared to chromosome arm B over all four genotypes.

Pairwise comparisons were visualized (Figure S33) in the R statistical environment http://www.R-project.org. using the function "labeledHeatmap" from the package WGCNA[52].



	1H	2HS	2HL	3HS	3HL	4HS	4HL	5HS	5HL	6HS	6HL	7HS	7HL
1H	0	20	65	0	28	0	0	51	60	8	19	58	22
2HS	7	0	9	0	13	0	0	33	8	5	2	8	6
2HL	3	0	0	0	8	0	0	22	7	3	0	6	5
3HS	47	45	90	0	85	0	8	128	62	53	52	63	30
3HL	0	3	7	0	0	0	0	11	6	1	5	7	2
4HS	148	177	262	46	232	0	52	260	235	121	158	225	147
4HL	30	58	109	0	76	0	0	115	96	46	59	94	52
5HS	6	8	10	8	17	0	8	0	28	0	10	27	22
5HL	52	9	21	0	61	0	9	81	0	34	18	0	0
6HS	6	27	64	3	41	0	2	48	58	0	7	57	19
6HL	2	17	44	1	22	0	0	42	32	2	0	32	2
7HS	43	8	16	0	51	0	8	72	0	31	16	0	0
7HL	37	18	55	0	60	0	6	89	25	26	13	25	0

Figure S33: Pairwise comparisons of exonic SNV frequencies between different chromosomes of barley.

Matrix entries are the SNV scarcity values (for details see section S8.3) of a row as compared to a column. For example, the entry 262 in row “4HS” and column “2HL” indicates that the SNV frequency on chromsosome arm 4HS is significantly smaller than the SNP frequency on 2HL. Heatmap colors from white (low scarcity) to red (high scarcity) were used to highlight different scarcity values

S8.4 Data visualization

Visualization of genetic variations

Displays were generated using Circos (http://**circos**.ca/). Unless otherwise stated in figure legends, the length of each (pseudo)chromosome is displayed as cumulative length of all whole genome shotgun contigs of the reference cultivar that could be anchored to the physical map. The histograms display the number of SNV per 50,000 base-pairs of concatenated whole genome shotgun contigs. The centi-morgan position for drawing the connectors have been calculated by dividing the length of each pseudo-chromosome by the maximum centi-morgan value for each chromosome.

Supplemental References

1 Schulte, D. *et al.* BAC library resources for map-based cloning and physical map construction in barley (*Hordeum vulgare* L.). *BMC Genomics* **12**, 247 (2011).

2 Yu, Y. *et al.* A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theor Appl Genet* **101**, 1093-1099 (2000).

3 Madishetty, K., Condamine, P., Svensson, J. T., Rodriguez, E. & Close, T. J. An improved method to identify BAC clones using pooled overgos. *Nucleic Acids Res* **35**, e5-, doi:10.1093/nar/gkl920 (2007).

4 Luo, M.-C. *et al.* High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378 (2003).

5 Bozdag, S., Close, T. & Lonardi, S. A compartmentalized approach to the assembly of physical maps. *BMC Bioinformatics* **10**, 217 (2009).

6 Lonardi, S. *et al.* Barcoding-free BAC pooling enables combinatorial selective sequencing of the barley gene space. *arXiv:1112.4438v1 [q-bio.GN]* (2012).

7 Kim, H. *et al.* Comparative physical mapping between *Oryza sativa* (AA genome type) and *O. punctata* (BB genome type). *Genetics* **176**, 379-390 (2007).

8 Chou, H.-H. & Holmes, M. H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093-1104 (2001).

9 Hübner, S. *et al.* Strong correlation of the population structure of wild barley (*Hordeum spontaneum*) across Israel with temperature and precipitation variation. *Mol Ecol* **18**, 1523-1536 (2009).

10 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851 - 1858 (2008).

11 Saski, C. *et al.* Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theor Appl Genet* **115**, 571-590 (2007).

12 Matsumoto, T. *et al.* Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Pysiol* **156**, 20-28 (2011).

13 Kurtz, S., Narechania, A., Stein, J. & Ware, D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008).

14 Mayer, K. F. X. *et al.* Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**, 1249-1263 (2011).

15 Stein, N. *et al.* The eukaryotic translation initiation factor 4E confers multiallelic recessive bymovirus resistance in *Hordeum vulgare* (L.). *Plant J* **42**, 912-922 (2005).

16 Turner, A., Beales, J., Faure, S., Dunford, R. P. & Laurie, D. A. The Pseudo-Response Regulator Ppd-H1 Provides Adaptation to Photoperiod in Barley. *Science* **310**, 1031-1034 (2005).

17 Ramsay, L. *et al. INTERMEDIUM-C*, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1*. *Nat Genet* **43**, 169-172 (2011).

18 Ariyadasa, R. & Stein, N. Advances in BAC based physical mapping and map integration strategies in plants. *J Biomed Biotechnol* **2012**, Article ID 184854, doi:10.1155/2012/184854 (2012).

19 Liu, H. *et al.* Highly parallel gene-to-BAC addressing using microarrays. *BioTechniques* **50**, 165-174 (2011).

20 Close, T. J. *et al.* Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10**, 582 (2009).

21 Chen, X. *et al.* An eQTL analysis of partial resistance to *Puccinia hordei* in barley. *PLoS ONE* **5**, e8598 (2010).

22 Sato, K., Nankaku, N. & Takeda, K. A high-density transcript linkage map of barley derived from a single population. *Heredity* **103**, 110-117 (2009).

23 Stein, N. *et al.* A 1000 loci transcript map of the barley genome – new anchoring points for integrative grass genomics. *Theor Appl Genet* **114**, 823-839 (2007).

24 Potokina, E. *et al.* Gene expression quantitative trait locus analysis of 16,000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J* **53**, 90-101 (2008).

25 Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379 (2011).
26 Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J.-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* **7**, e32253 (2012).
27 Muñoz-Amatriaín, M. *et al.* An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Gen* **4**, 238–249 (2011).
28 Comadran, J. *et al.* A homologue of *Antirrhinum CENTRORADIALIS* is a component of the quantitative photoperiod and vernalization independent *EARLINESS PER SE 2* locus in cultivated barley *Nat Genet* **under review** (2012).
29 Moscou, M. J., Lauter, N., Steffenson, B. & Wise, R. P. Quantitative and qualitative stem rust resistance factors in barley are associated with transcriptional suppression of defense regulons. *PLoS Genet* **7**, e1002208 (2011).
30 Close, T. J., Wanamaker, S., Roose, M. L. & Lyon, M. in *Plant Bioinformatics* Vol. 406 *Methods in Molecular Biology* (ed David Edwards) 161-177 (Humana Press, 2008).
31 Thiel, T. *et al.* Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evol Biol* **9**, 209 (2009).
32 Min, X. J., Butler, G., Storms, R. & Tsang, A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res* **33**, W677-680 (2005).
33 Fang, G., Bhardwaj, N., Robilotto, R. & Gerstein, M. B. Getting started in gene orthology and functional analysis. *PLoS Comp Biol* **6**, e1000703 (2010).
34 Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res* **13**, 2178-2189 (2003).
35 Project, R. A. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36**, D1028-D1033 (2008).
36 Itoh, T. *et al.* Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. *Genome Res* **17**, 175-183 (2007).
37 Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**, D1202-D1210 (2012).
38 Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
39 The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768 (2010).
40 Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* **396**, 59-70 (2007).
41 Beißbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464-1465 (2004).
42 Lu, T. *et al.* Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res* **20**, 1238-1249 (2010).
43 Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**, 189-192 (2003).
44 Kertész, S. *et al.* Both introns and long 3'-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic Acids Res* **34**, 6147-6157 (2006).
45 Kalyna, M. *et al.* Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res* **40**, 2454-2469 (2012).
46 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2010).
47 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
48 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
49 You, F. *et al.* BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**, 253 (2008).
50 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Gen Biol* **10**, R25 (2009).
51 Milne, I. *et al.* Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401-402 (2010).
52 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

Supplemental Acknowledgements

We gratefully acknowledge: 1) excellent technical assistance by Susanne König, Manuela Knauft, Uli Beier, Anne Kusserow, Katrin Trnka, Ines Walde, Sandra Driesslein, Ingelore Dommes, Tatjana Sretenovic, 2) sharing of gene-bearing BAC addresses by Andris Kleinhofs, Edmundo Rodriguez, Pascal Condamine, Harkamal Walia, Hung Le, Tao Jiang, Jie Zheng, Serdar Bozdag, Yonghui Wu, Saghai Maroof, J Perry Gustafson, Lothar Altschmied, Shane Heinen, Muharrem Dilbirligi, Kulvinder Gill, Lol Cooper, Patrick Hayes, Peggy Lemaux, Phil Bregitzer, Catherine Feuillet, Anders Falk, Timothy Sutton, Nick Collins, David Laurie, Leila Feiz, Thomas Blake, Heather Witt, Frank You, Mingcheng Luo, 3) helpful discussion on the sequencing and analysis of RNA-seq data by Jonathan Wright, Jane Rogers, 4) the GABI-PD team (http://www.gabipd.org), especially Doreen Pahlke, for data submission of BAC sequencing raw data to EMBL/ENA and hosting the BAC contigs, 5) Doreen Stengel, Steffen Flemming, Stephan Weise, Sebastian Beier for sequence raw data management and data submission of WGS and BAC raw data to EMBL/ENA, 6) Arnaud Bellec, Sonia Vautrin and the team of http://cnrgv.toulouse.inra.fr/ for BAC library management and rearraying of MTP clones, 7) Dave Kudrna and Yeisoo Yu for their support in BES sequencing.