

## **Supplementary Note 1: Access to datasets**

**page 3**

Supplementary Table 1.1: Accession IDs and URLs for access to datasets

## **Supplementary Note 2: Chromosome conformation capture sequencing (Hi-C/TCC)**

**page 5**

2.1 Hi-C library construction and sequencing

2.2 TCC library construction and sequencing

Supplementary Figure 2.1: Quality controls during Hi-C/TCC library construction

Supplementary Figure 2.2: Distance-dependent decay and Rabl configuration

## **Supplementary Note 3: Transcriptome sequencing and annotation of transcribed regions**

**page 11**

3.1 PacBio Isoseq data collection

3.2 PacBio Isoseq error correction

Supplementary Table 2.1: Information on Isoseq libraries

Supplementary Table 2.2: PacBio Isoseq error correction.

Supplementary Table 2.3: PacBio Isoseq length statistics

Supplementary Figure 3.1: Illumina NextSeq length distribution

Supplementary Figure 3.2: PacBio Isoseq length distribution after error correction

3.3 Generation of Illumina RNA-seq data from 16 barley tissues

3.4 Automated gene annotation

Supplementary Table 3.4: RNA-seq read mapping statistics

3.5 Characteristics and number of predicted gene models

Supplementary Figure 3.3: Distribution of gene locus size and transcripts per locus

Supplementary Figure 3.4: Distribution of transcript size and exons per transcript

Supplementary Figure 3.5: Gene completeness and predicted protein length of HC and LC genes

Supplementary Figure 3.6: Comparison to previous annotation

3.6 Validation of gene models

3.7 Prediction of long non-coding RNAs

Supplementary Figure 3.7: Annotation of long non-coding RNAs

Supplementary Figure 3.8: Classification of TE-derived and nonTE-derived lncRNAs in barley

### 3.8 Discovery and validation of microRNA loci

#### Supplementary Note 4: Analysis of gene families

page 32

##### 4.1 OrthoMCL gene family clustering

Supplementary Table 4.1: Over-represented GO terms in the GO category "biological process" in selected barley gene family subsets relative to other species

Supplementary Table 4.2: Over-represented GO terms in the GO category "cellular component" in selected barley gene family subsets

Supplementary Table 4.3: Over-represented GO terms in the GO category "molecular function" in selected barley gene family subsets

Supplementary Table 4.4: Boundaries of genomic compartments

Supplementary Table 4.5: GO term enrichment analysis between genomic compartments

##### 4.2 Analysis of $\alpha$ -amylases

Supplementary Table 4.6: Coding sequences used to search for  $\alpha$ -amylase genes in the barley genome

Supplementary Table 4.7 Sequences used for the phylogenetic analysis of the  $\alpha$ -amylase multi-gene family in grass genomes

Supplementary Figure 4.1: Sequence alignment of promoter regions of *amy1* genes

Supplementary Figure 4.2: Sequence alignment of promoter regions of *amy2* genes

##### 4.3 Analysis of SWEET and VPE genes

Supplementary Figure 4.3: Expression of *SWEET11a* and *SWEET11b*

#### Supplementary Note 5: Diversity analysis

page 44

Supplementary Table 5.1: List of barley cultivars subject to exome sequencing

#### Supplementary References

page 45

## Supplementary Note 1: Access to datasets

Supplementary Table 1.1: Accession IDs and URLs for access to datasets.

Data set	Repository <sup>1</sup>	Accession number/Digital Object Identifier/URL
<b>BAC short read raw data</b>		
0H MTP Illumina PE	ENA	PRJEB11991, PRJEB9427
0H MTP Roche 454	ENA	PRJEB9104
0H MTP Illumina MP	ENA	PRJEB11992
1H MTP Illumina PE	ENA	PRJEB8576
1H MTP Roche 454	ENA	PRJEB9097
1H MTP Illumina MP	ENA	PRJEB8579, PRJEB8580
2H MTP Illumina PE	ENA	PRJEB11758, PRJEB9428
2H MTP Roche 454	ENA	PRJEB9098
2H MTP Illumina MP	ENA	PRJEB11798
3H MTP Illumina PE	ENA	PRJEB8577
3H MTP Roche 454	ENA	PRJEB9099
3H MTP Illumina MP	ENA	PRJEB8579, PRJEB8580
4H MTP Illumina PE	ENA	PRJEB8578
4H MTP Roche 454	ENA	PRJEB9100
4H MTP Illumina MP	ENA	PRJEB8579, PRJEB8580
5H MTP Illumina PE	ENA	PRJEB9429
5H MTP Roche 454	ENA	PRJEB9101
5H MTP Illumina MP*	ENA	PRJEB10963
6H MTP Illumina PE	ENA	PRJEB9430
6H MTP Roche 454	ENA	PRJEB9102
6H MTP Illumina MP	ENA	PRJEB12096
7H MTP Illumina PE	ENA	PRJEB9431
7H MTP Roche 454	ENA	PRJEB9103
7H MTP Illumina MP	ENA	PRJEB11489
Non-MTP Illumina PE	ENA	PRJEB9619
Non-MTP Roche 454	ENA	PRJEB9062
Non-MTP Illumina MP	ENA	PRJEB8579, PRJEB8580
<b>Assembled sequences</b>		
MTP BAC assemblies	ENA	PRJEB13020
Assemblies of gene-bearing BACs	NCBI	PRJNA198204
All BAC assemblies	IPK	doi:10.5447/IPK/2016/21 (before contamination removal) doi:10.5447/IPK/2016/12 (after contamination removal)
Non-redundant sequence	IPK	doi:10.5447/IPK/2016/30
Pseudomolecule sequence	IPK	doi:10.5447/IPK/2016/34, doi:10.5447/IPK/2016/35
Split pseudomolecule sequence	IPK	doi:10.5447/IPK/2016/36, doi:10.5447/IPK/2016/37
<b>Chromosome conformation capture data</b>		
Hi-C and TCC reads	ENA	PRJEB14169
Hi-C map	IPK	doi:10.5447/IPK/2016/20
<b>Transcriptome sequencing</b>		
Illumina RNAseq reads	ENA	PRJEB3149, PRJEB14349
PacBio Isoseq reads	ENA	PRJEB14446
<b>Exome capture data</b>		
Exome capture sequencing of 96 barley	ENA	PRJEB14445
<b>Annotation of transcribed regions</b>		
High-confidence (HC) gene set GTF file	IPK	doi:10.5447/IPK/2016/38
Low-confidence (LC) gene set GTF file	IPK	doi:10.5447/IPK/2016/46
HC gene set CDS (all and representative only)	IPK	doi:10.5447/IPK/2016/40, doi:10.5447/IPK/2016/43
LC gene set CDS (all and representative only)	IPK	doi:10.5447/IPK/2016/48, doi:10.5447/IPK/2016/51
HC gene set transcripts (all and repr. only)	IPK	doi:10.5447/IPK/2016/39, doi:10.5447/IPK/2016/42
LC gene set transcripts (all and repr. only)	IPK	doi:10.5447/IPK/2016/47, doi:10.5447/IPK/2016/50
HC gene set proteins (all and repr. only)	IPK	doi:10.5447/IPK/2016/41, doi:10.5447/IPK/2016/44
LC gene set proteins (all and repr. only)	IPK	doi:10.5447/IPK/2016/49, doi:10.5447/IPK/2016/52
HC gene set functional descriptions	IPK	doi:10.5447/IPK/2016/45
LC gene set functional descriptions	IPK	doi:10.5447/IPK/2016/53
Long non-coding RNAs	IPK	doi:10.5447/IPK/2016/18
microRNA loci	IPK	doi:10.5447/IPK/2016/15

<b>Annotation of repetitive regions</b>		
GTF file with the positions of mobile elements		doi:10.5447/IPK/2016/16
<b>BioNano map data</b>		
BioNano raw data	IPK	doi:10.5447/IPK/2016/32
BioNano genome map	IPK	doi:10.5447/IPK/2016/31
<b>Genetic map data</b>		
GBS raw data of Morex x Barke RILs	ENA	PRJEB14130
POPSEQ raw data	ENA	PRJEB13028, ERP002184
GBS map of Morex x Barke RILs	IPK	doi:10.5447/IPK/2016/29
Updated POPSEQ map (WGS assembly)	IPK	doi:10.5447/IPK/2016/33
Updated POPSEQ map (pseudomolecules)	IPK	doi:10.5447/IPK/2016/17
<b>Sequence search and map visualization</b>		
BLAST server <sup>2</sup>	IPK	<a href="http://webblast.ipk-gatersleben.de/barley_ibsc/">http://webblast.ipk-gatersleben.de/barley_ibsc/</a>
BARLEX – the barley genome explorer	IPK	<a href="http://barlex.barleysequence.org/">http://barlex.barleysequence.org/</a>

<sup>1</sup>ENA: European Nucleotide Archive of the European Molecular Biology Laboratory (EMBL), <http://www.ebi.ac.uk/ena/>; NCBI: NIH genetic sequence database (Genbank) of the US National Center for Biotechnology Information; IPK: Plant Genomics and Phenomics Research Data Repository (ref. 1) hosted by IPK Gatersleben, <http://edal.ipk-gatersleben.de/repos/pgp/>.

## Supplementary Note 2: Chromosome conformation capture sequencing (Hi-C/TCC)

### 2.1 Hi-C library construction and sequencing

#### *Plant growth and crosslinking of cells*

Barley plants (*H. vulgare* cv. “Morex”) were grown for 7 days in a greenhouse cabinet as described previously<sup>2</sup>. Freshly harvested leaves (24 g) were crosslinked, and a nuclei suspension was isolated essentially as described before<sup>3</sup>. Briefly, 2 g leaves were cut into 2 cm pieces, transferred into a 50 ml tube and vacuum infiltrated (1 h, 400 mbar) in 15 ml nuclei isolation buffer (NIB) supplemented with 2% formaldehyde. Crosslinking was stopped by adding glycine to 0.125 M and vacuum infiltration (5 min, 400 mbar).

#### *Cell disruption, nuclei isolation and chromatin digestion*

The material was washed, ground in liquid nitrogen, resuspended in NIB containing protease inhibitors and filtered as described<sup>3</sup>. The nuclei suspension was spun (3000 g, 15 min, 4°C), and the pellet was resuspended in 300 µl ice-cold NIB supplemented with protease inhibitors and 1.5 M sucrose (NIBS). The suspension was layered on 1.5 ml sucrose cushion (NIB plus protease inhibitors and 1.7 M sucrose), and spun in a 2 ml Eppendorf tube (16000 g, 1h, 4°C). The sedimented nuclei were resuspended in 100 µl NIBS and checked for quality and quantity as described earlier<sup>3</sup>. For Hi-C library construction, about 10<sup>7</sup> nuclei were collected by centrifugation (1900 g, 5 min, 4°C) and washed twice with ice-cold 500 µl 1 x NEB2 buffer (New England Biolabs). Nuclei were purified (1900 g, 5 min, 4°C), resuspended in 260 µl 1 x NEB2 buffer and split into five 50 µl aliquots. To each tube 312 µl 1x NEB 2 buffer and 38 µl 1% SDS was added. The mixture was incubated for 10 min at 65 °C, chilled on ice, supplemented with 44 µl 10% Triton X-100 to quench the SDS and digested with 400 units *HindIII* as described<sup>5</sup>.

#### *Marking of DNA ends with biotin-14-dCTP, blunt-end ligation and DNA purification*

Four tubes containing digested chromatin were marked with biotin-14-dCTP (Thermo Fisher Scientific) using a fill-in reaction and Klenow enzyme<sup>4</sup> to produce blunt-end DNA molecules. One tube was kept unused on ice serving as a 3C control. The blunt-end ligation for the Hi-C library, the normal 3C ligation and the subsequent DNA purification were as described<sup>4</sup>. The four Hi-C tubes were pooled and the 3C control was kept separate.

#### *Quality controls*

The concentration of the ligation products was determined using a Qubit Fluorometer (Life Technologies). Ligation of filled-in *HindIII* sites (AAGCTT) results in the creation of sites for the restriction enzyme *NheI* (GCTAGC). To control for effectiveness of fill-in and blunt-end ligation, a PCR fragment (534 bp) from two adjacent genomic barley *HindIII* restriction fragments was generated using Q5 Hot Start High-Fidelity DNA Polymerase (NEB). The primers employed were: 5'-ATCTTCATGCGAGGCAGAGT-3' and 5'-ACCGTTGAACCATCTTCAGG-3'. Products were digested with *HindIII* and *NheI* and analysed on a 2% agarose gel (Supplementary Fig. 2.1a).

#### *Removal of biotin from un-ligated ends and DNA purification*

In order to remove the biotin-14-dCTP from non-ligated DNA ends, five µg DNA were subjected to the exonuclease activity of the T4 DNA polymerase<sup>4</sup>. The DNA was purified by phenol:CHCl<sub>3</sub> (1:1)

extraction, precipitated and washed as described<sup>4</sup>. The DNA was dissolved in 133 µl TE and stored on ice.

#### *DNA fragmentation*

For the fragmentation a Covaris S220 device (Covaris Ltd.) was used (130 µl DNA in Covaris microTUBES, 10% Duty Factor, 175 W Peak Incident Power, 200 cycles per burst and 1 to 3 cycles of 1 min time). After each cycle the size was verified using the Agilent Bioanalyzer High Sensitivity DNA Kit or standard 2 % agarose gel electrophoresis. If necessary, an additional cycle was added in order to generate fragments in the range of 200-300 bp.

#### *Size fractionation using AMPure XP beads*

The size of the DNA was fractionated using different ratios of sample volumes and AMPure XP beads as described<sup>5</sup>. The 1.1 x fraction containing fragments with a size between 150 and 300 bp was used for Illumina adapter ligation.

#### *End repair, A-tailing and biotin pull-down*

Blunt-end repair, A-tailing and biotin pull-down were performed as described<sup>5</sup>. For the pull-down, 5 µl My-One C1 streptavidin bead solution (Dynabeads) were used per µg of Hi-C ligation products. The final bead suspension containing the Hi-C DNA was resuspended in a total volume of 38.8 µl in 1x T4 DNA ligation buffer (Fermentas) containing 5% PEG 4000.

#### *Illumina adapter ligation, titration of PCR-cycles and large-scale PCR*

Adapters were ligated by adding 38.8 µl Hi-C DNA bead suspension, 6 µl Illumina PE adapter, 1.1 µl 10x T4 DNA ligation buffer (Fermentas), 1.1 µl 50% PEG 4000 (Fermentas), 1 µl water and 2 µl T4 DNA ligase (5U/µl; Fermentas). The reaction was incubated 1 h at 22°C. Products were purified and resuspended in 20 µl 1x NEB2 buffer as described<sup>5</sup>. Test PCR reactions (9, 12, 15 and 18 cycles) were used to find the optimal PCR cycles to produce sufficient library for sequencing, without generating unwanted byproducts. The reactions (25 µl) contained 1.5 µl streptavidin bead-bound Hi-C DNA, 1.25 µl forward and reverse primer (10 µM), 5 µl 5x Q5 reaction buffer, 2.0 µl 2.5 mM dNTP and 0.25 µl Hot Start High-Fidelity DNA Polymerase (2U/µl; NEB). After an initial incubation at 98°C (30 sec), amplification of the DNA was performed (98°C for 10 s, 66°C for 30 s, and 72°C for 30 s), followed by a final 3 min extension at 72°C. The forward and reverse primers employed for the amplification were 5'- AAT GAT ACG GCG ACC ACC GAG AT-3' and 5'- CAA GCA GAA GAC GGC ATA CGA -3'. In order to select the PCR cycles, size distribution and quantity of the PCR products were analysed using the Agilent Bioanalyzer High Sensitivity DNA Kit. For the large-scale PCR the remainder of the bead-bound Hi-C DNA was used in eight separate 25 µl PCR-reactions as described for the trial PCR. The products were pooled and purified with AMPure XP beads<sup>4</sup>. The final Hi-C library was eluted in 35 µl 10 mM Tris-Cl, 0.1 mM EDTA (pH 8.0). To estimate the fraction with biotinylated junctions, 100 ng of the library were digested with *NheI* (Supplementary Fig. 2.1b.).

#### *Illumina sequencing*

The Hi-C library was quantified using Real-Time PCR and sequenced (paired-end, 2 x 100 cycles) using the HiSeq Illumina system as described previously<sup>6</sup>.

## 2.2 TCC library construction and sequencing

### *Plant growth, crosslinking of cells, cell disruption and nuclei isolation*

Barley plants were grown (*H. vulgare* cv. “Morex” and F<sub>1</sub> hybrids between cultivars Morex and Barke) in a greenhouse cabinet as described for HiC library construction. In the same greenhouse compartment, etiolated plants (cv. “Morex” only) were grown for 7 days in a light-tight container. The initial steps from harvest of leaves to the checks for quality and quantity of nuclei were as described for Hi-C library construction. For a TCC library about 10<sup>7</sup> nuclei were collected by centrifugation (1900 g, 5 min, 4°C) and washed twice with 900 µl ice-cold wash buffer<sup>7</sup>. Following purification (1900 g, 5 min, 4°C), the nuclei were resuspended in 250 µl ice-cold wash buffer.

### *Chromatin biotinylation, HindIII digestion and dialysis*

To the nuclei suspension, 95 µl 2% SDS was added, and the nuclei were incubated (65°C, 10 min) to solubilize the crosslinked chromatin. Chromatin was biotinylated using EZlink Iodoacetyl-PEG2-Biotin (IPB, Pierce Protein Research Products), followed by neutralization of SDS<sup>7</sup>. The DNA was digested with HindIII and dialysed<sup>7</sup>.

### *Tethering*

The dialysed DNA was divided into 5 equal aliquots, and the volume was adjusted to 500 µl using phosphate buffered saline supplemented with 0.01% Tween 20<sup>7</sup>. Four tubes were used for TCC library construction. One tube served as a 3C control. The DNA of each aliquot was immobilized to low surface coverage, and free streptavidin was saturated with biotin as described<sup>7</sup>.

### *Filling of DNA ends, blunt-end ligation of immobilized DNA and DNA purification*

The 3C control was diluted with 100 µl 1x NEB2 buffer and kept unused on ice. The four tubes containing digested chromatin were marked with biotin-14-dCTP (Thermo Fisher Scientific) using Klenow enzyme<sup>7</sup>. By including dGTPαS (1:1 mixture of R<sub>p</sub> and S<sub>p</sub> isomers; Jena Bioscience GmbH) in the fill-in reaction, the blunt-ends were modified with a phosphorothioate bond located 5' to the biotinylated cytosine residue. The reaction was stopped by adding 5 µl 0.5 M EDTA. Beads were washed, purified, resuspended and transferred to a conical 15 ml tube as described<sup>7</sup>. For ligation of TCC samples, 3.9 ml water, 250 µl 10x T4 DNA ligase reaction buffer (NEB), 180 µl 10% Triton X-100, 100 µl 1 M Tris-Cl (pH 7.4), 50 µl 10 mg/ml BSA (NEB) and 12 µl T4 DNA ligase (5U/µl, Fermentas) were added. For the 3C control, 3 µl T4 DNA ligase (5 U/µl, Fermentas) was added only. The tubes containing the ligations were incubated horizontally and rocked (100 rpm) overnight at 16 °C. The reaction was stopped by adding 200 µl 0.5 M EDTA. The beads with the ligation products were isolated using a magnet, and the liquid was discarded.

### *Reversion of crosslink, DNA extraction and quality control*

The formaldehyde crosslink was reversed using proteinase K, followed by DNA extraction and precipitation as described<sup>7</sup>. Pellets were resuspended in 15 µl 10 mM Tris-Cl (pH 8.0), and the four TCC samples were pooled. The quality and quantity of the ligation products was controlled as described for Hi-C libraries (Supplementary Fig. 2.1a).

### *Removal of biotin from non-ligated DNA ends*

The 3'-hydroxyl termini from duplex DNA (5 µg) were digested with Exonuclease III as described<sup>7</sup>, thereby removing the biotin from non-ligated DNA ends. The reaction was stopped by adding 2 µl 0.5 M EDTA, 2.5 µl 4 M NaCl and incubation at 70°C for 20 min. The final volume was adjusted to 100 µl using water.

#### *DNA fragmentation*

DNA was sheared using a Covaris S220 device (Covaris Ltd.) as described for Hi-C library construction. The DNA was precipitated by adding 1.8 volumes AMPure XP beads as described by the manufacturer (Beckman Coulter Inc.). DNA was eluted in 52 µl EBT [10 mM Tris-Cl (pH 8.0), 0.05% (v/v) Tween 20].

#### *End repair, A-tailing and biotin pull-down*

Blunt-end repair was as described for Hi-C. Products were purified by adding 1.8 volumes AMPure XP beads (Beckman Coulter Inc.). DNA was eluted in 39 µl 10 mM Tris-Cl, 0.1 mM EDTA (pH 8.0). A-tailing was performed as described for Hi-C. The reaction was terminated by adding 1 µl 0.5 M EDTA and 20 min incubation at 65°C. Pull-down of biotinylated DNA and washing of the beads was as described<sup>7</sup>. The bead suspension containing the TCC DNA was resuspended for ligation in a total volume of 38.8 µl in 1x T4 DNA ligation buffer (Fermentas) containing 5% PEG 4000.

#### *Illumina adapter ligation, titration of PCR-cycles, large-scale PCR and size-selection*

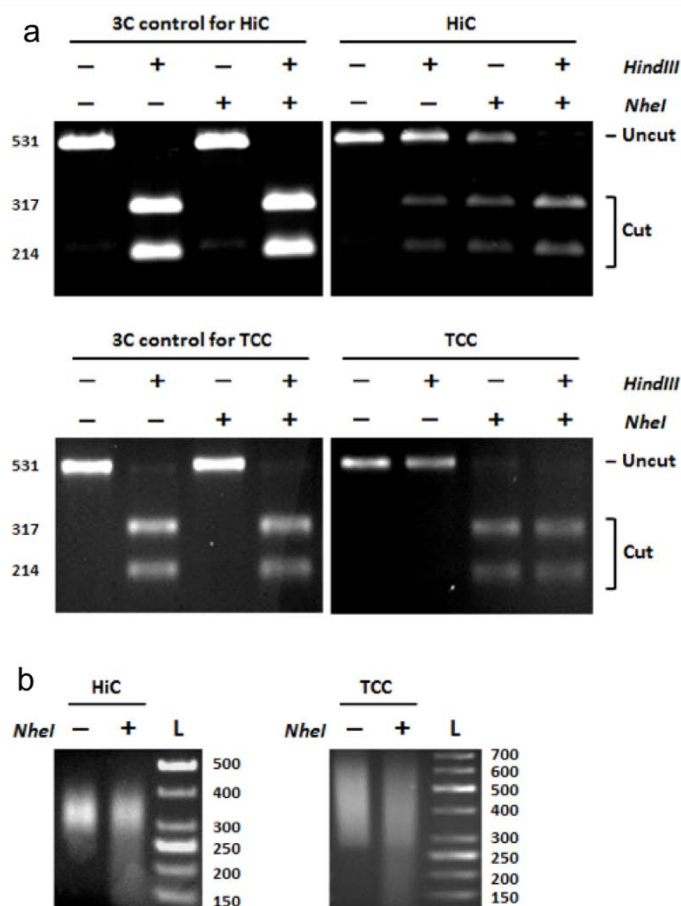
Illumina adapter ligation, test-PCR, large-scale PCR and reaction clean-up were as described for Hi-C. PCR products were eluted in 27 µl EBT and quantified using the Qubit dsDNA BR Assay Kit (Life Technologies). The libraries were size-separated using agarose gel electrophoresis, and the range between 350 and 550 bp was isolated<sup>8</sup>. The final TCC library was eluted in 50 µl 10 mM Tris-Cl (pH 8.5). The fraction with biotinylated junctions was estimated as described for Hi-C (Supplementary Fig. 2.1b).

#### *Illumina sequencing*

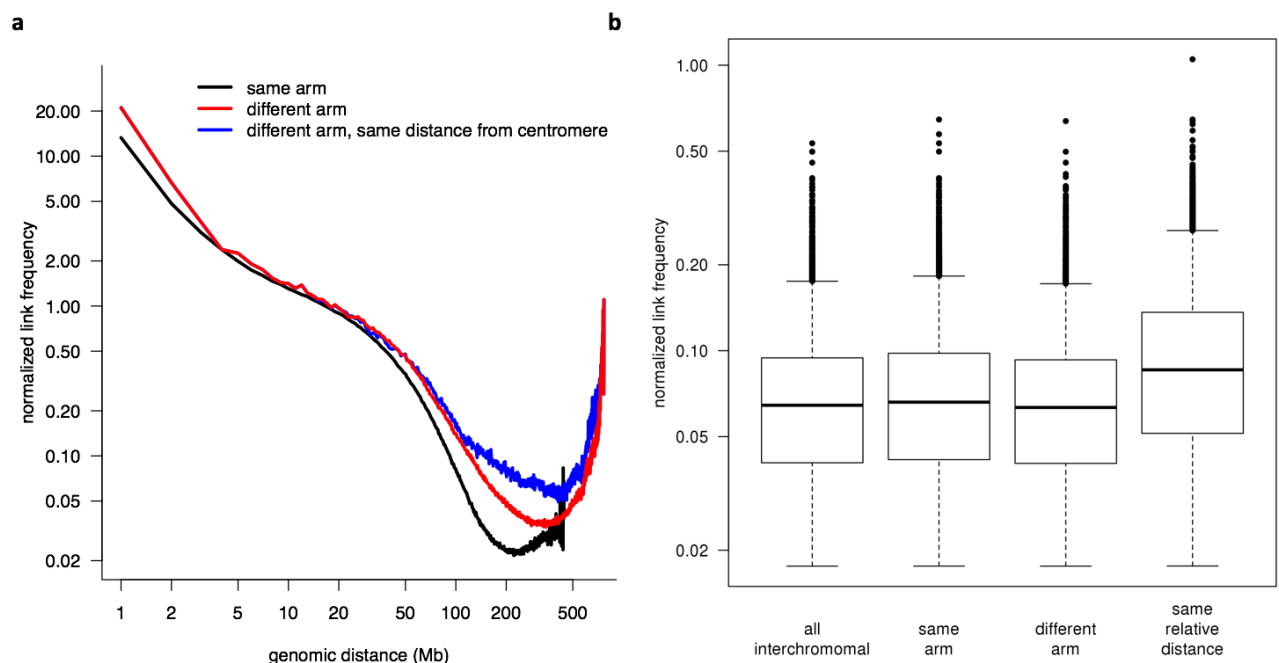
The TCC libraries were quantified and sequenced as described for Hi-C libraries.



**Supplementary Figure 2.1:** Quality controls during Hi-C/TCC library construction. **(a)** Control for marking and ligation of ends in Hi-C and TCC. The ligation junction of two close genomic barley *HindIII* fragments was PCR amplified and digested. In the 3C controls the *HindIII* overhangs were ligated, thus creating a *HindIII* restriction sequence in the amplicon. In contrast, the Hi-C/TCC junctions were derived from blunt-end ligation of filled-in *HindIII* sites, and were therefore recleaved by *NheI* only. In typical experiments 50 - 80 % of the amplicons from the Hi-C/TCC samples contained a *NheI* site. **(b)** Quality control of the final Hi-C/TCC library. The library (100 ng) was digested with *NheI* and compared to the uncut control. The smaller size distribution provides an estimate for the fraction of the library containing true Hi-C/TCC ligation products. DNA was separated using a standard 2% agarose gel. The size of the products and the DNA ladder (L) are indicated (bp).



**Supplementary Figure 2.2:** Distance dependent decay and Rabl configuration. **(a)** The median normalized count of HiC links between 1 Mb bins at a certain genomic distance (x-axis) is plotted on the y-axis as in Fig. 2a, but considering only pairs of bins from the same chromosome arm (black), different chromosome on the same chromosome (red), or bins at the same relative distance from the centromere on the same chromosome (blue). Contact frequencies at large distances in the linear genome ( $> 200$  Mb) are highest for pairs of bins with the same relative distance from centromere, i.e. those pairs corresponding to the anti-diagonal in Extended Data Fig. 4a. **(b)** Distribution of interchromosomal normalized link frequencies are shown for [from left to right] (i) all pairs of 1 Mb bins on different chromosomes, (ii) pairs of bins on the same chromosome arm (e.g. bin 1 on 1HS and bin 2 on 3HS), (iii) pairs of bins on the different chromosome arms (e.g. bin1 on 1HS and bin on 2HL), and (iv) pairs of bins with the same relative distance from the centromere of the respective chromosomes. The higher contact probability of case (iv) reflect the pronounced diagonals and anti-diagonals in Extended Data Fig. 4b and support the notion of greater spatial proximity between loci on the different chromosomes that are juxtaposed in the Rabl configuration. Boxplots were generated the R function boxplot() with default parameters.



## Supplementary Note 3: Transcriptome sequencing and annotation of transcribed regions

### 3.1 PacBio Isoseq data collection

Long-range sequencing of RNA for the purpose of gene annotation was performed using the SMRT (Single Molecule, Real-Time) technology of Pacific Biosciences (<http://www.pacb.com/smrt-science/smrt-sequencing/>). Library preparation followed the protocol provided by Pacific Biosciences: Isoform Sequencing (Iso-Seq™) using the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin™ Size Selection System. For each sample, 1µg of total RNA was used for reverse transcription (Clontech laboratories Inc.). The optimal number of cycles for large-scale PCR was determined to be 16. After large-scale PCR, resulting amplicons were divided into four aliquots in order to process different fractions of DNA fragments: I. without size selection; II. 1-2kb; III. 2-4kb; and IV. 3-8kb.

DNA fraction I derived from large-scale PCR was used directly for library preparation. Isolation of DNA fractions II-IV was done using BluePippin (SageScience Inc., USA). DNA fraction II extracted by the BluePippin was introduced subsequently into library preparation. DNA fractions III and IV were amplified in a second large-scale PCR after extraction by BluePippin (IIIa and IVa). Furthermore, for depletion of short fragments amplicons derived from the second large-scale PCR were processed, in parallel to IIIa and IVa, using a BluePippin (IIIb and IVb). The library preparation approach resulted in six libraries listed in Supplementary Table 3.1.

In total, 16 SMRT cells were sequenced using a PacBio RSII. The number of SMRTcells per library is given in Supplementary Table 3.1.

The data was further processed using the RS\_IsoSeq protocol of the SMRT Analysis System to obtain full-length non-chimeric reads. The number and the average lengths of the full-length reads per library are also shown in Supplementary Table 3.1.

### 3.2 PacBio Isoseq error correction

#### *PacBio Isoseq data preparation*

The polished high and low confidence read-of-insert sequences from runs A03\_1-4 (3-8Kb), B03\_1-4 (2-4Kb) and C03\_1-3 (1-2Kb) (Supplementary Table 3.1) were merged into a single FASTA file. These sequences represent the output from Quiver, the consensus calling tool in PacBio's proprietary SMRTPipe software (<https://github.com/PacificBiosciences/SMRT-Analysis/wiki/SMRT-Pipe-Reference-Guide-v2.1>), which generates consensus sequences from multiple reads-of-insert representing the same transcript. Assuming some overlap between different size libraries, merging libraries will result in a dataset that is redundant in so far as the same transcript may be represented more than once. The merged dataset contained 122,190 sequences.

#### *Error correction*

Even though reads of this type have been subjected to two separate error correction procedures (the multiple alignment of the read-of-insert subreads, and the subsequent consensus calling by Quiver) they may still contain errors at this stage. The PacBio sequencing technology's prevalent type of error are small indels, and these are particularly unfavourable in transcript sequences as they produce

frame shifts, which may affect predicted translations. The merged sequences were therefore subjected to a further round of read correction with Illumina NextSeq data (TruSeq Total RNA v2 library, single flow cell of 2x 150 bp reads) generated from the same original RNA sample. We used proovread<sup>9</sup> (version 2.13.8) to carry out this task.

To preprocess the data in line with the recommendations for this software, we first quality trimmed the Illumina NextSeq reads to a base quality of 20 from both ends with trimmomatic<sup>10</sup> (version 0.30), whilst simultaneously removing Illumina adapter sequences. We then merged overlapping paired end reads with the FLASH read overlapper<sup>11</sup> using default settings. The resulting overlapped reads were combined with the remaining non-overlapping forward and reverse reads, resulting in a single file with 25,066,985 sequences with a size range between 35 and 257 bp (mode 151 bp). Supplementary Fig. 3.1 shows a size frequency distribution for this read set.

The proovread software supports custom parallelisation through partitioning of the input sequences, and we used its *SeqChunker* helper application to split the PacBio Isoseq sequences into ten input files containing exactly 12,219 sequences each. The proovread subjobs were then executed in parallel on ten nodes of our compute cluster, using default parameters. Subjobs took approx. four hours each to compute, using four threads each.

The proovread output consists of both trimmed and untrimmed sequences (see documentation at <https://github.com/BioInf-Wuerzburg/proovread>), where the untrimmed sequences represent complete corrected PacBio reads but inclusive of any uncorrected or poorly corrected regions. We decided to use the trimmed output, which consists of only the high accuracy PacBio reads, where uncorrected/low quality regions with no Illumina support have been removed.

The trimmed regions from the ten subjobs were merged into a single file, resulting in a total of 123,875 corrected sequences. This number is greater than that of the input dataset, which presumably is the result of proovread splitting up chimaeric reads into subregions. Supplementary Table 3.2 shows some basic statistics for the Isoseq reads before and after the read correction step. Supplementary Fig. 3.2 shows a frequency distribution for the length of the trimmed, error corrected Isoseq reads. The latter is essentially trimodal, reflecting merging of the three different size libraries. The total residue count was reduced by ~ 10% in the trimmed reads. This loss was caused by the removal of unsupported regions. Also, both the minimum and maximum sequence length in the trimmed reads were lower than those of the original reads, which appeared to be the result of chimeric reads being split. The average transcript length was slightly lower in the trimmed data when compared to the original (1,326 vs. 1,480, respectively).

#### *Quantifying the effects of error correction*

The trimmed, error corrected PacBio reads were compared to the original input data using a number of different metrics. First, we mapped the Illumina NextSeq data back to both the uncorrected and the corrected data using Bowtie2<sup>12</sup>. As the Isoseq reference sequences were potentially redundant, we chose the *-a* mapping mode in Bowtie2, which allows reads to map to all of their potential mapping locations. This enables reads to map to multiple copies of the same transcript, thereby avoiding a scenario where a lowly expressed transcript is duplicated and the few available reads are all mapped to one of the copies, creating an uncovered region which is then flagged up as erroneous.

To ensure that reads were consistently aligned around indels, we ran the GATK's IndelRealigner<sup>13</sup> over the resulting Bowtie2 mappings. The resulting mappings were then subjected to coverage analysis with the GATK's DepthOfCoverage tool, and the percentage of reference bases that had no Illumina support (zero coverage) was calculated from the output. These percentages were 18.3% and 2.9% for the uncorrected and corrected reference, respectively, which suggests a clear improvement resulting from the error correction.

We also ran the VarScan variant caller<sup>14</sup> (version 2.3.7) over both mappings to establish the number of insertion and deletions (indels) before and after error correction. The results from this are shown in Supplementary Table 3.2. The error correction resulted in a dramatic reduction of the number of indels from 59,765 (for the whole dataset of 122,190 uncorrected sequences) to only 127 in the corrected sequences.

In addition, we also measured how error correction affects the potential of the reads to be translated correctly. We used custom written Java code that produces six-frame translations using the Biojava code libraries<sup>15</sup> and selects the longest open reading frame (ORF) available among the translations. Protein-coding regions were defined as ORFs downstream of the 5'-proximal ATG codon in the longest ORF. The results of this are shown in Supplementary Table 3.3. As expected, both the mean and maximum peptide length was increased in the corrected reads when compared to the uncorrected reads (by 10.5 % and 7.1 % respectively), which suggests that repair of broken reading frames caused by indels in the PacBio data produces a direct improvement in the length of translated peptides.

**Supplementary Table 3.1:** Information on Isoseq libraries.

library	large-scale PCR	BluePippin	2 <sup>nd</sup> large-scale PCR	BluePippin	SMRT cells	P1 reads*	Number of full-length non-chimeric reads	Avg. length of full-length non-chimeric reads [bp]
I	16 cycles	-	-	-	3	280,397 (93,466)	118,090	1,188
II	16 cycles	1-2kb	-	-	3	259,629 (86,543)	99,915	1,433
IIIa	16 cycles	2-4kb	12 cycles	-	1	100,872	34,908	1,545
IIIb	16 cycles	2-4kb	12 cycles	2-4kb	4	376,962 (94,241)	113,306	2,000
IVa	16 cycles	3-8kb	12 cycles	-	1	51,550	16,676	1,249
IVb	16 cycles	3-8kb	15 cycles	3-8kb	4	322,212 (80,553)	93,944	1,610

\*Summarized P1 read number across all SMRTcells; averages are given in parentheses.

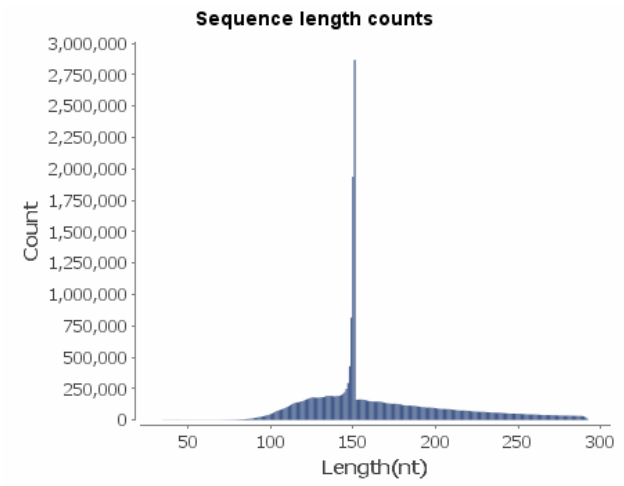
**Supplementary Table 3.2:** PacBio Isoseq error correction. Basic statistics for the PacBio Isoseq read before and after read correction with Illumina Nextseq data. Indels were filtered to only include those where at least 90% of the reads contained the variant.

	uncorrected Isoseq reads	corrected Isoseq reads
Number of sequences	122,190	123,875
Residue count	180,910,855	164,276,222
Minimum sequence length (bp)	305	110
Maximum sequence length (bp)	31,775	5,689
Average sequence length (bp)	1,480.6	1,326.2
Number of insertions and deletion	59,765	127

**Supplementary Table 3.3:** PacBio Isoseq length statistics. Sequence length statistics for peptides translated from the longest open reading frames of the uncorrected and error corrected PacBio Isoseq reads.

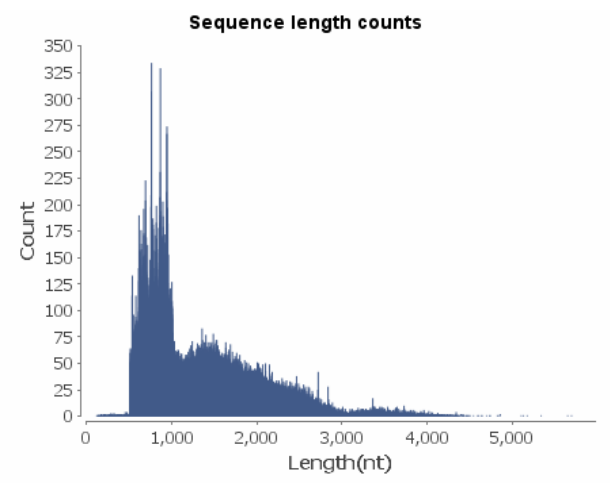
	uncorrected	corrected
Number of peptide sequences	122,189	123,848
Mean length	281.2	310.7
Mode of length	210	210
Median length	214	242
Maximum length (bp)	1,602	1,715
Minimum length (bp)	2	1

**Supplementary Figure 3.1:** Illumina NextSeq length distribution. Frequency distribution of sequence lengths for the quality trimmed Illumina Nextseq reads after applying the FLASH read overlacer.





**Supplementary Figure 3.2:** PacBio Isoseq length distribution after error correction. Frequency distribution of sequence lengths for the trimmed, error corrected PacBio Isoseq reads.



### 3.3 Generation of Illumina RNA-seq data from 16 barley tissues

#### *Barley material*

Plants of barley cv. 'Morex' were grown under controlled conditions at the James Hutton Institute, or the Leibniz Institute of Plant Genetics and Crop Plant Research. All tissues were sampled from each of three biological replicates, which were either individual plants or pooled dissected tissues where material was limiting. Tissue was flash frozen in liquid nitrogen and stored at -80 °C. In total, 16 barley tissues were isolated (48 samples in total) for generation of RNA-seq data:

Vegetative tissues: Leaf tissue (LEA) was sampled from seedlings at 17 days after planting (dap), and root tissue at both 17 dap (ROO1) and 28 dap (ROO2). The third stem internode (NOD) was dissected at 42 dap, and senescing leaf (SEN) isolated at 56 dap. Seedlings were grown to 10 dap in the dark to isolate etiolated leaf (ETI) material. Epidermal strips (EPI) were made from 28 dap plant leaf tissue. All material was grown in pots under controlled glasshouse conditions.

Inflorescence tissues: Whole developing inflorescence tissue was sampled at 30 dap (INF1) and 50 dap (INF2). From plants 42 dap, inflorescences were dissected to isolate lodicule (LOD), lemma (LEM) and palea (PAL). Rachis (RAC) was isolated from plants 35 dap. All material was grown in pots under controlled glasshouse conditions.

Developing grain tissues: Whole developing grain (caryopsis) was sampled at 5 days post-anthesis (dpa; CAR5) and 15 dpa (CAR15). Only the central caryopses were sampled from each spike. All material was grown in pots under controlled glasshouse conditions.

Germinating grain tissues: Mature grain was germinated in petri plates in the lab and in the dark. After 4 days, embryonic tissue (including mesocotyl and seminal roots; EMB) was dissected.

#### *RNA extraction*

Total RNA was extracted from all frozen samples (200 mg each) using TriReagent (Sigma-Aldrich) as recommended, with an additional phenol/chloroform extraction and ethanol precipitation step. Quality was checked by spectrophotometry, and integrity using the RNA 6000 NanoChip on a 2100 Bioanalyzer (Agilent Technologies).

#### *Library preparation and Illumina RNA sequencing*

RNA-seq libraries were constructed and sequenced at Earlham Institute, 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 was purified from 1 µg total RNA using two rounds of purification with oligo-dT magnetic beads. Purified mRNA was fragmented by addition of 5x fragmentation buffer and heating at 94°C for 8 minutes, yielding fragments of ~250 nt. First strand cDNA was synthesised using random hexamers, and second strand cDNA synthesised by adding GEX second strand buffer, dNTPs, RNase H and DNA polymerase I, followed by incubation for 2.5 h at 16°C. Second strand cDNA was then subjected to end repair, A-tailing, and adapter ligation with barcoded adapters. Purified cDNA was enriched by 15 cycles of PCR (10 s, 98°C; 30 s, 60°C; 30 s, 72°C) using PCR Primer Mix Cocktail and PCR Master Mix. Samples were purified using AMPure XP Beads and eluted in 30 µl Resuspension Buffer. Purified cDNA libraries were QC'd using DNA 100 Chip

(Agilent Technologies 2100 Bioanalyzer). Libraries were normalised to 10 nM and pooled with equal molarity (in pools of 8 samples each, and as replicate blocks).

Each pool was diluted to 2 nM with NaOH and 5 µl of diluted library transferred into 995 µl HT1 (final concentration of 10 pM). Normalised libraries were then loaded onto an Illumina cBot cluster generation system. Each pool of libraries was clustered onto two lanes of an Illumina flow cell using a TruSeq Paired-End Cluster Kit v3, following the PE Amplification Linearization Blocking Hybridisation v8 recipe. Flow cells were loaded onto Illumina HiSeq 2000 instruments according to the manufacturer's instructions. Paired-end sequencing was performed generating 2 x 100 bp reads using TruSeq SBS kit v3 sequencing chemistry, Illumina software HCS 1.4 and RTA 1.12.2

### 3.4 Automated gene annotation

#### *Transcript prediction*

The gene annotation pipeline (Extended Data Fig. 1) combined information of splice site-aware alignments with reference proteins, RNA-seq based gene structure predictions, alignment of IsoSeq reads and alignments of full-length cDNAs (flcDNAs). We aligned predicted protein sequences from barley<sup>16</sup> and closely related grass species *B. distachyon*<sup>17</sup>, rice<sup>18</sup> (*O. sativa*), and sorghum<sup>19</sup> (*S. bicolor*) as well as predicted open reading frames from full length cDNA sequences from barley<sup>20</sup> against chromosome sequences using the splice-aware alignment software GenomeThreader<sup>21</sup> (version 1.6.2; parameters used: -species rice -gcmincoverage 30 -prseedlength 7 -prhdist 4 -force) and resulting transcript structure predictions were then merged using Cuffcompare from the Cufflinks<sup>22</sup> package. The RNA-seq data contained reads from 48 samples originating from two different sets. The first RNA-seq data set was published previously<sup>16</sup> and the second RNA-seq set was newly generated. We aligned all RNA-seq reads against the chromosomes using Hisat2<sup>23</sup> with default parameters. More than 1.9 billion Illumina RNA-seq read pairs from 48 samples were mapped to the barley genome. The average overall alignment ratio per sample was 85.36 %. Using Cufflinks, we defined structural information for 276,048 transcript sequences which were clustered into 63,075 potential gene positions (loci). An overview about the RNA-seq data and alignment rates per sample is given in Supplementary Table 3.4. We used GMAP<sup>24</sup> to align the IsoSeq reads against the genome. Results were stored in a SAM file. The Python script *collapse\_isoforms\_by\_sam.py* from the PacBio repository ([https://github.com/PacificBiosciences/cDNA\\_primer](https://github.com/PacificBiosciences/cDNA_primer)) was used to predict transcript structures from the SAM file and to remove redundant transcripts. We used GMAP to align flcDNA sequences to chromosomes and received a single gff3 file that contained transcript structures. We then applied a custom script to convert the gff3 formatted file into a gff2 formatted file. We clustered the RNA-seq-based transcript structures with the reference-based gene model predictions, the structural information from the IsoSeq alignments as well as with structural information from flcDNA sequence alignments and defined a consensus transcript set by using Cuffcompare from the Cufflinks package.

#### *Prediction and selection of open-reading frames (ORFs)*

Merging of transcript sequences from RNA-seq, reference proteins, flcDNAs and Isoseq sequences resulted in 344,248 transcript sequences, which were clustered into 83,105 potential gene positions. A custom script was used to extract transcript sequences based on their coordinates and to store them in a single fasta file. We applied Transdecoder (version rel\_16Jan; parameters: -m 30 –retain\_long\_orfs 90 –search\_pfam pfam.AB.hmm.bin) to determine putative open reading frames as

well as corresponding peptide translations including prediction of Pfam domains. TransDecoder (<https://transdecoder.github.io>) often reported several alternative predicted peptides for each transcript. To select a single best translation per transcript, we used BLASTP to compare all predicted peptides with a comprehensive protein database which contained high confidence protein sequences from *Arabidopsis thaliana*<sup>25</sup>, maize<sup>26</sup>, *Brachypodium distachyon*<sup>17</sup>, rice<sup>18</sup> and sorghum<sup>19</sup>. BLASTP hits with an e-value below  $10^{-5}$  were considered as significant hits. We then applied a sequential filtering approach to find a single best translation for each transcript. In each filtering step all remaining peptide sequences were sorted according to a specific attribute/category and those translations within the highest category (or with highest value) were retained for the next filtering step. The six filtering steps were:

1. Three categories: 1.) Neither homology support nor Pfam domains, 2.) without homology support and with Pfam domains, 3.) with homology support
2. Total length of translation
3. Four categories: 1.) CDS without start and without stop codon, 2.) CDS without start and with stop codon, 3.) CDS with start codon and no stop codon, 4.) CDS with start and with stop codon
4. Number of Pfam domains
5. Number of significant BLAST hits
6. Start position on the chromosome

#### *Preliminary confidence assignment*

Stringent confidence classification was applied to all predicted genes to discriminate between loci representing high-confidence (HC) protein-coding genes and less reliable low-confidence (LC) genes, which potentially consisted of gene fragments, putative pseudogenes and non-(protein)-coding transcripts. We assigned confidence values to a gene model in a two-step procedure by using the same criteria and methods described previously<sup>27</sup>. First of all, we considered genes with transcripts that showed significant sequence homology (BLASTN with e-value below  $10^{-10}$ ) to a library of repeats and transposable elements as low confidence genes. Secondly, we compared the predicted peptide sequences against the protein data sets of barley, *B. distachyon*, rice, sorghum, *Arabidopsis thaliana* and predicted protein sequences from the barley flcDNAs using BLASTP and considered hits with an E-value below  $10^{-10}$  as significant. For each gene, we selected the best-matching reference protein as template sequence and defined the transcript sequence with maximum coverage of the template sequence as a gene representative. Genes were defined as high confidence (HC) genes if they had a significant BLAST hit to reference proteins and if their representative protein had a similarity to the respective template sequence above a threshold which we determined on the basis of the origin of template sequences (> 60% for *A. thaliana*, sorghum and rice, > 65 % for *B. distachyon*, and > 90 % for barley).

#### *Final confidence assignment and manual refinement*

We predicted function for all genes (high confidence and low confidence) using the AHRD pipeline (<https://github.com/groupschoof/AHRD>) on the basis of one representative protein sequence for each gene. Low confidence genes with predicted function were transferred into the high confidence gene set and high-confidence genes that were annotated as transposable elements were transferred into the set of low-confidence genes. During gene annotation, we noticed putatively duplicated regions in the genome assembly that were masked as described<sup>28</sup>. High- or low-confidence genes

that were completely contained within such duplicated regions were removed. High-confidence genes partially overlapping duplicated regions were transferred into to low-confidence gene set. Finally, high confidence genes with representative proteins that were identical fragments of other high confidence proteins with less than 50 % length were transferred into low confidence set, while they were considered as likely candidates for pseudogenes. Based on functional annotation and final refinement, we defined ten confidence subclasses:

- HC\_G: high confidence genes with predicted function
- HC\_TE?: high confidence genes which might be transposable element due to conflicting information
- HC\_u: high confidence gene without function assignment
- HC\_U: high confidence gene with annotation of unknown function (based on homology to other proteins with unknown function)
- LC\_M: genes that were manually removed from the set of high confidence genes
- LC\_nof: genes without open reading frame
- LC\_TE: genes that are annotated as transposable elements
- LC\_u: genes without functional annotation
- LC\_U: genes with annotation of unknown function (based on homology to other proteins with unknown function)

An overview about the computational pipeline for confidence assignment is provided in Extended Data Fig. 1.

**Supplementary Table 3.4:** RNA-seq read mapping statistics.

Library Name	Sample Description	Number of read pairs	Alignment rate
LIB1742	Etiolated seedling	41,384,846	94.56%
LIB1743	Etiolated seedling	43,303,706	93.94%
LIB1744	Etiolated seedling	49,531,901	93.59%
LIB1745	Lemma	47,998,226	94.08%
LIB1746	Lemma	53,749,132	93.67%
LIB1747	Lemma	60,119,172	93.29%
LIB1748	Lodicule	43,755,966	94.33%
LIB1749	Lodicule	51,604,095	93.96%
LIB1750	Lodicule	46,999,900	94.02%
LIB1751	Palea	45,245,738	93.78%
LIB1752	Palea	42,748,577	93.30%
LIB1753	Palea	49,742,794	93.56%
LIB1754	Peeled epidermis	40,499,667	93.86%
LIB1755	Peeled epidermis	47,346,160	93.17%
LIB1756	Peeled epidermis	45,557,424	92.13%
LIB1757	Rachis	44,035,817	94.00%
LIB1758	Rachis	50,796,023	93.77%
LIB1759	Rachis	48,845,702	92.55%
LIB1760	Root	42,890,166	93.86%
LIB1761	Root	51,380,220	93.66%
LIB1762	Root	49,903,681	93.45%
LIB1763	Senescing leaves	36,620,432	93.25%
LIB1764	Senescing leaves	40,033,204	93.32%
LIB1765	Senescing leaves	40,559,378	93.09%
CAR05_biorep-1	Developing caryopses	24,772,003	85.96%
CAR05_biorep-2	Developing caryopses	26,425,521	86.25%
CAR05_biorep-3	Developing caryopses	23,165,002	86.58%
CAR15_biorep-1	Developing caryopses	31,862,147	75.00%
CAR15_biorep-2	Developing caryopses	29,663,308	74.09%
CAR15_biorep-3	Developing caryopses	21,312,783	75.43%
EMB_biorep-1	Embryonic tissue	24,400,586	84.10%
EMB_biorep-2	Embryonic tissue	27,788,521	83.03%
EMB_biorep-3	Embryonic tissue	18,565,903	86.67%
INF1_biorep-1	Developing inflorescences	28,105,744	85.68%
INF1_biorep-2	Developing inflorescences	26,065,488	86.52%
INF1_biorep-3	Developing inflorescences	22,584,784	87.72%
INF2_biorep-1	Developing inflorescences	32,411,952	85.05%
INF2_biorep-2	Developing inflorescences	32,497,495	87.16%
INF2_biorep-3	Developing inflorescences	22,898,169	88.36%
LEA_biorep-1	Shoots from seedlings	24,190,028	58.40%
LEA_biorep-2	Shoots from seedlings	22,140,874	58.18%
LEA_biorep-3	Shoots from seedlings	21,093,621	68.06%
NOD_biorep-1	Third internode	31,645,395	84.81%
NOD_biorep-2	Third internode	166,184,951	90.20%
NOD_biorep-3	Third internode	80,672,694	88.48%
ROO_biorep-1	Root tissue	23,385,380	55.30%
ROO_biorep-2	Root tissue	38,739,551	55.74%
ROO_biorep-3	Root tissue	33,639,235	34.23%
<b>Total/average</b>		<b>1,948,863,062</b>	<b>85.36%</b>

### 3.5 Characteristics and number of predicted gene models

We identified 39,734 high- and 41,949 low-confidence gene models in the barley genome assembly. A potential function was assigned to genes based on their representative proteins. Using this procedure, we were able to assign a potential function to the majority (n=31,899) of high confidence genes (subclass HC\_G). Some high confidence genes were likely candidates for transposable elements (subclass HC\_TE?) but remained in the set of high confidence genes due to significant homology with high confidence genes from other species or from previous gene annotation. Most low confidence genes (n=27,922) had no predicted function (subclass LC\_u). A great amount of low confidence genes (n=8,975) were annotated as transposable elements (subclass LC\_TE) and 948 low confidence genes were likely candidates for being transposable elements (subclass LC\_TE?). A small amount of low confidence genes (n=403) had no open reading frame of at least 150 bp length (subclass LC\_nof). The number of HC genes per chromosome varied between 4,380 for chromosome 4H and 6,518 for chromosome 2H and we predicted 2,157 high confidence genes on unanchored scaffolds (U) (Extended Data Table 2).

In the following, we describe the structural properties of the gene predictions.

#### *Loci*

The high confidence genes had a mean length of 6,010 nt and 50.0 % of HC genes had a length of more than 2,258 nt. The mean length of low confidence genes was significantly shorter at 2,328 nt. Number of transcripts per gene varied between high and low confidence genes as well; 13,977 high confidence genes had a single predicted transcript, while the majority of high confidence genes (51.1 %) had more than two predicted transcripts. In contrast, 78.4 % of the low confidence genes had only a single predicted transcript (Supplementary Fig. 3.3).

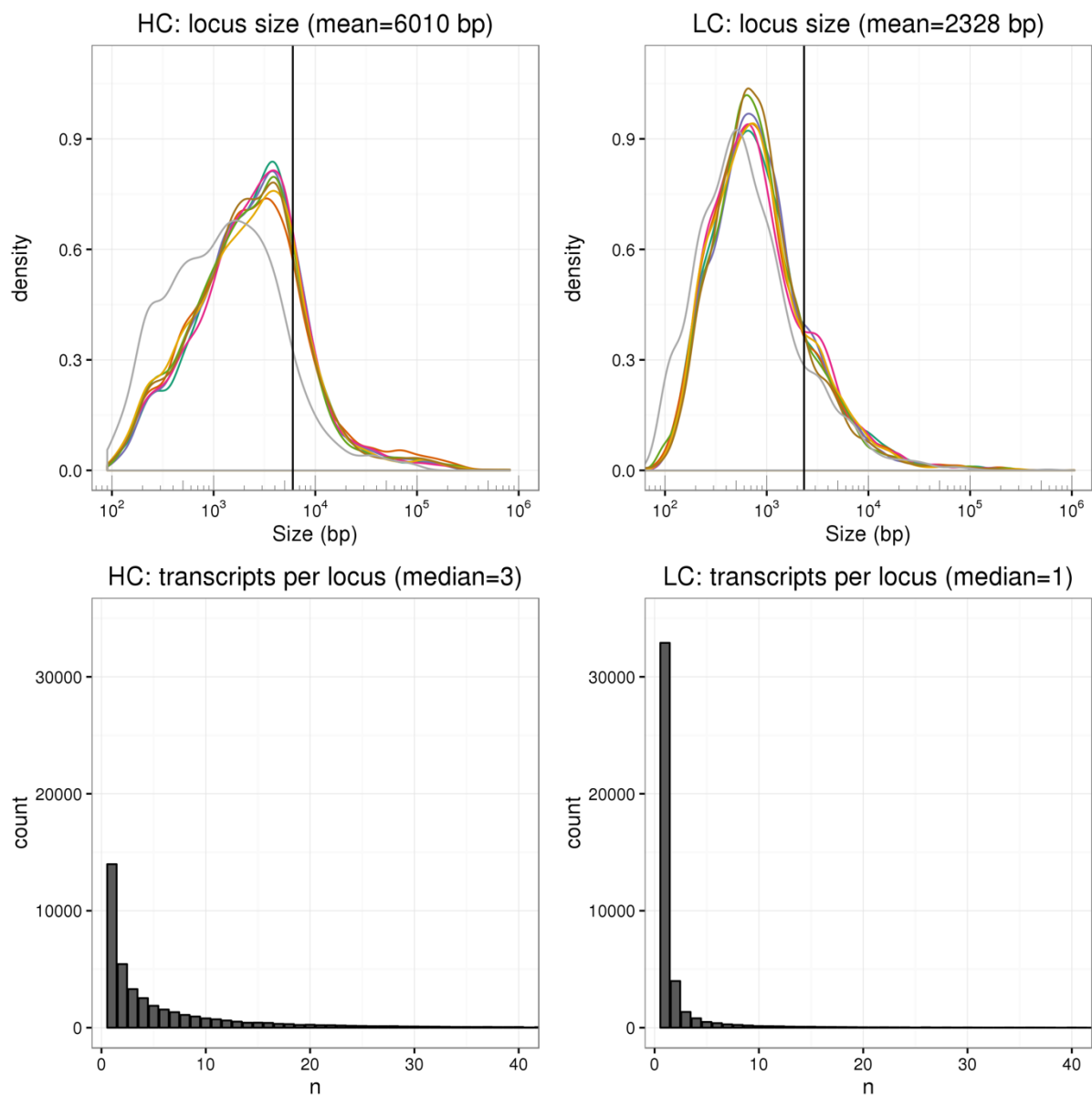
#### *Transcripts*

The mean length of spliced transcripts from high confidence genes was 1,927 nt and 48.6 % of high confidence transcripts were composed of more than five exons. In contrast, transcripts from low confidence genes were significantly shorter at 1,478 nt and had a lower number of exons (Supplementary Fig. 3.4).

#### *Proteins*

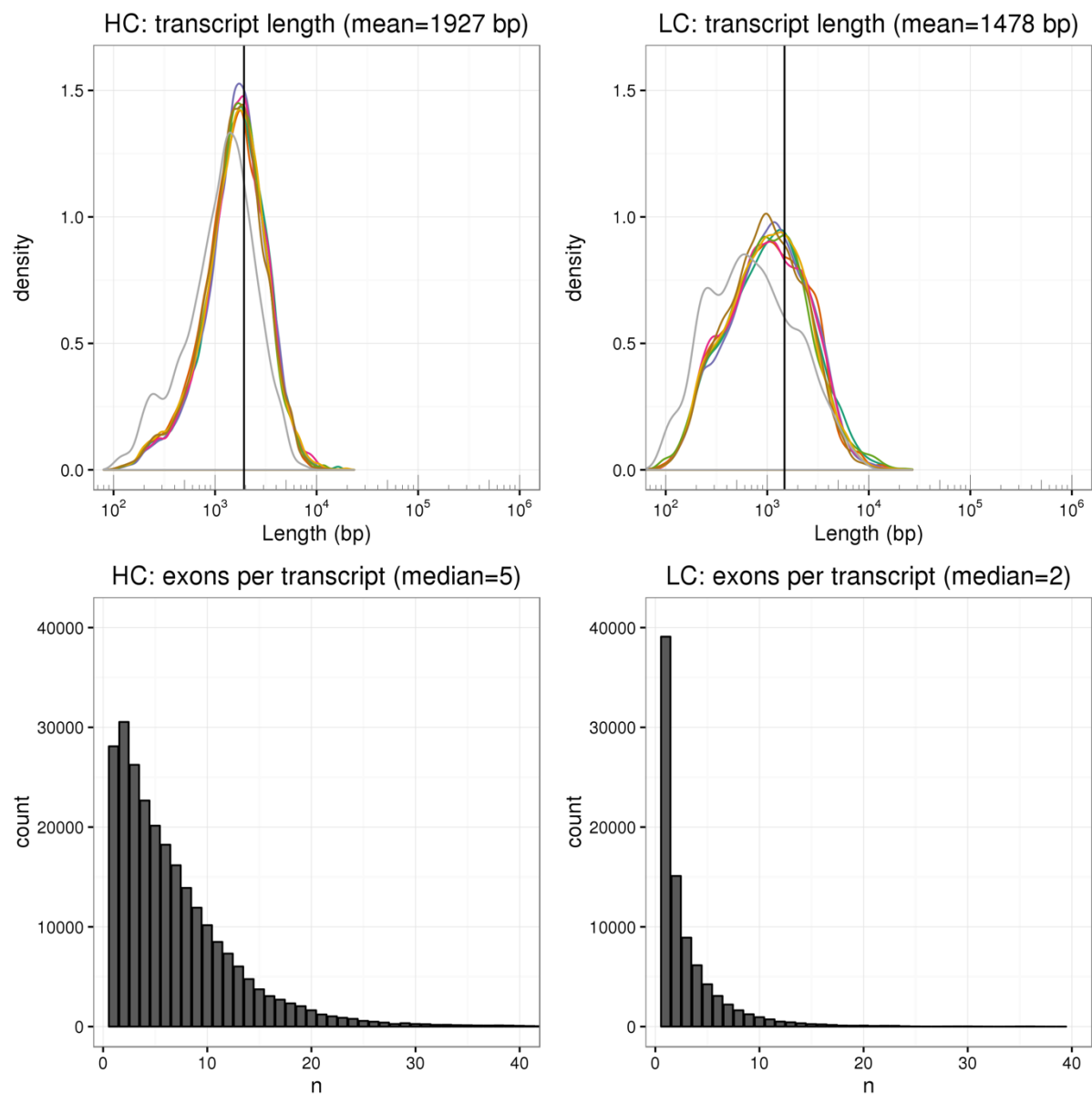
Most of the high confidence genes (69.0 %) had at least one completely annotated protein with start and stop codon, and 45.4 % of low confidence genes had a least one protein with start and stop codon. Overall, 53.6 % of high confidence proteins and 48.6 % of low confidence proteins were annotated with start and stop codon. The mean length of high confidence proteins was 360 AA and the mean length of low confidence proteins was significantly lower with 174 AA (Supplementary Fig. 3.5).

Supplementary Figure 3.3: Distribution of gene locus size and transcripts per locus.

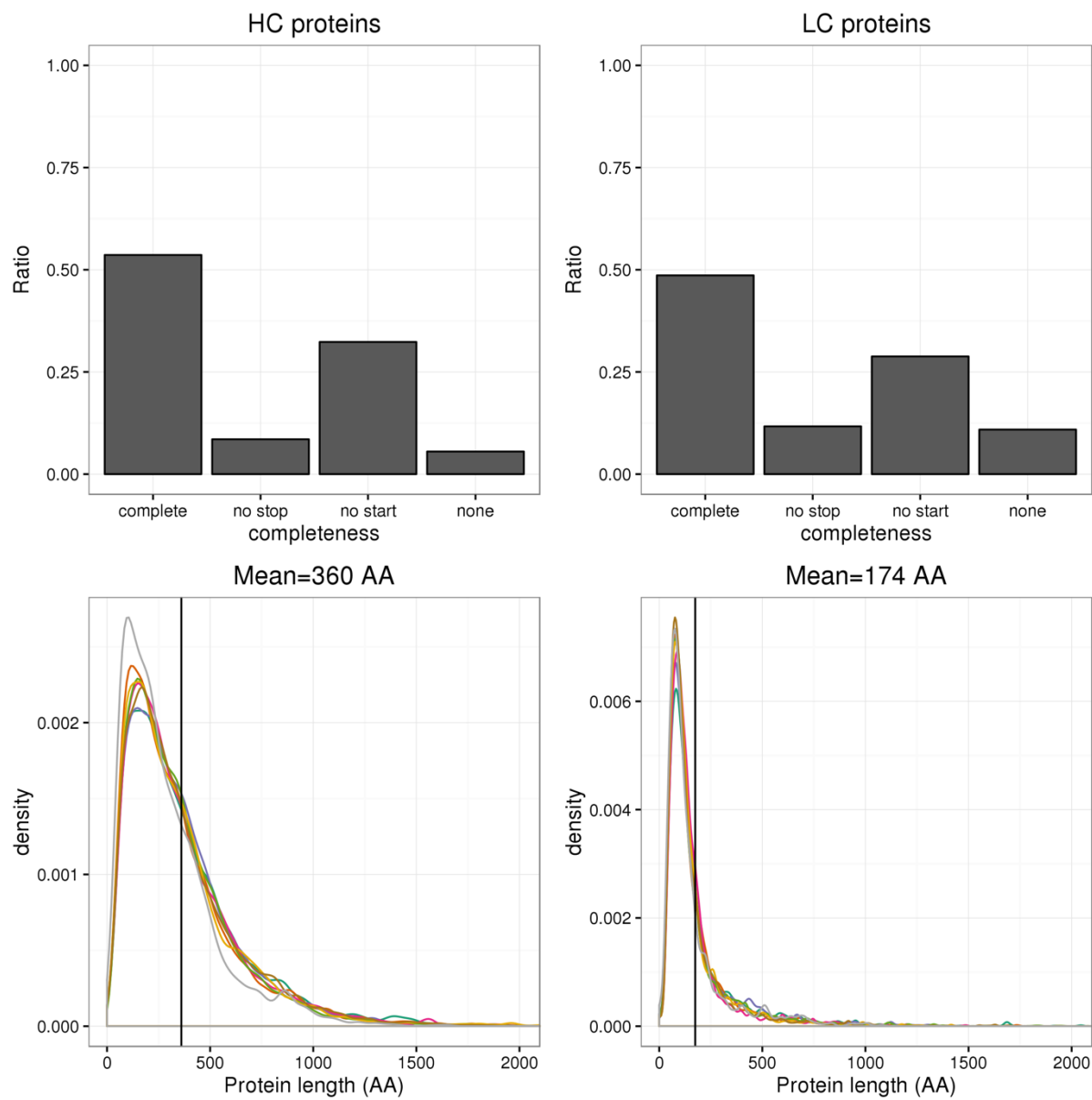




**Supplementary Figure 3.4:** Distribution of transcript size and exons per transcript.



**Supplementary Figure 3.5:** Gene completeness and predicted protein length of HC and LC genes.



### 3.6 Validation of gene models

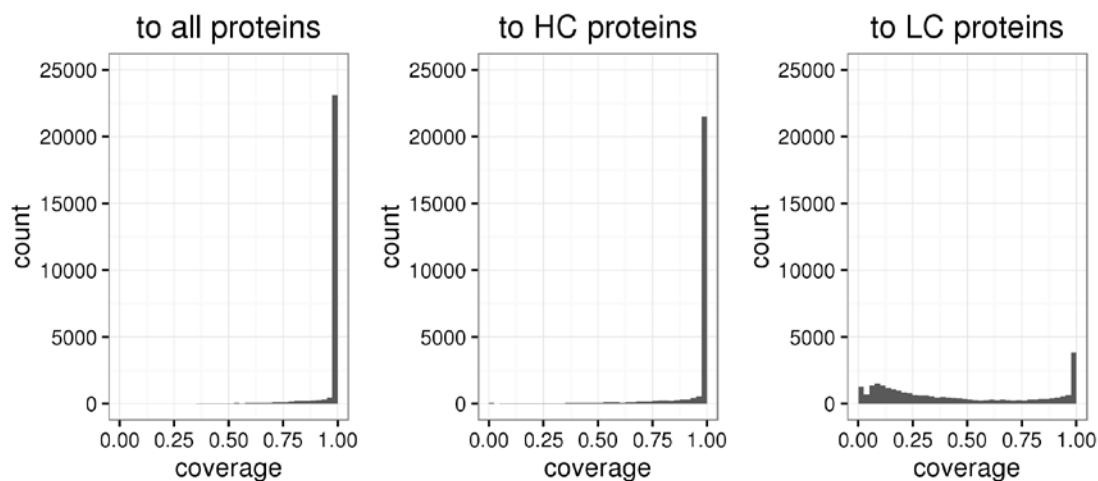
#### *Comparison to genes models in the WGS assembly of barley cv. Morex*

We used BLAST to compare the representation of previously annotated barley proteins<sup>16</sup> (“the 2012 genes/proteins”) with our new gene annotation (Supplementary Fig. 3.6). A large proportion of the 2012 proteins (96.4 %) were represented by our new protein predictions with at least 75 % coverage. Thereby, 92.2 % were represented by high confidence proteins and 28.3 % were represented by low confidence genes. The overlap between high and low confidence genes can be explained by sets of pseudogenes in the low-confidence gene set and duplicated regions in the assembly. A total of 1,058 2012 proteins were represented by new LC proteins and not by new HC proteins. Of these, 572 proteins had a best hit to a low confidence protein from subclasses LC\_TE or LC\_TE?, which indicates that some of these proteins might be transposable elements that were annotated as high confidence proteins previously.

#### *Benchmarking Universal Single-Copy Orthologs (BUSCO)*

We validated the completeness of the genome assembly as the basis for gene annotation using BUSCO gene set and software<sup>29</sup> (early release, plantdb). We thereby tested BUSCO genes for abundance in the set of predicted representative proteins and reported results for all predicted genes, all HC genes and all LC genes (Extended Data Fig. 2b).

**Supplementary Figure 3.6:** Comparison gene annotations. Structural overlap between previously annotated barley proteins on the whole-genome shotgun assembly of cv. Morex and the new gene annotation for HC and LC genes on the map-based reference assembly.



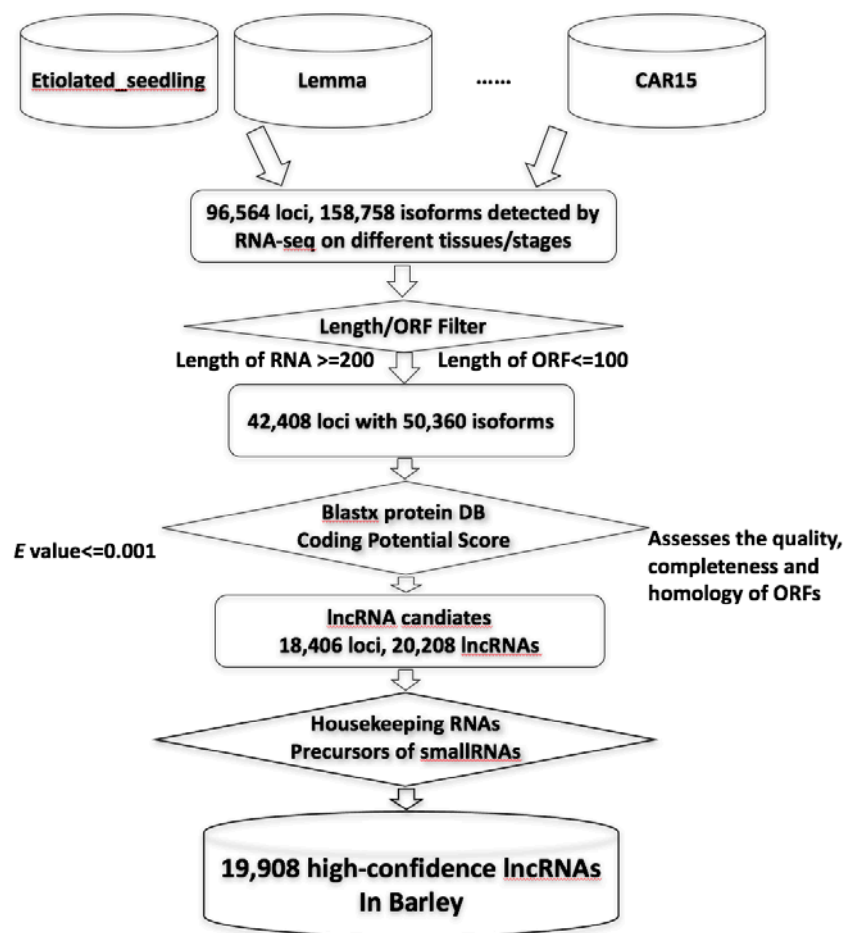
### 3.7 Prediction of long non-coding RNAs

RNA-seq data from 16 different tissues/stages of Morex was used to annotate barley long non-coding RNAs (lncRNAs) in the Morex reference genome. All RNA-seq reads were mapped to the reference genome using Tophat2<sup>30</sup>, allowing only 1 bp mismatch, and only uniquely mapped reads were employed for further analyses. Then, Cufflinks<sup>22</sup> was employed to detect *de novo* transcribed loci and to assemble transcript isoforms. The bioinformatics pipeline LncRNA-Finder<sup>31</sup> was used to filter all transcript isoforms for the identification of lncRNAs (Supplementary Fig. 3.7). Briefly, all assembled transcript isoforms were subject to the filters as follows: (1) size selection, which excluded transcripts smaller than 200bp; (2) open reading frame filter to remove transcripts with long open reading frames (ORFs), which are likely to be protein coding. Transcripts that encode ORFs of 100 or less amino acids or incomplete ORFs were considered as lncRNA candidates; (3) known protein domain filter to eliminate transcripts with potential protein-coding ability (cutoff E-value $\leq$ 0.001) using the alignment against the Swiss-Protein database; (4) Protein-coding-score test for the detection of quality, completeness, and sequence similarity of the ORF to proteins, which filtered the protein-coding prone transcripts; and (5) Elimination of housekeeping lncRNAs and precursors of small RNAs. The housekeeping lncRNA databases including tRNA, rRNA, snRNAs, snoRNAs, and signal recognition particle (7SL/SRP) were collected from the NCBI Nucleotide database using the query command “*Hordeum vulgare* [Organism] AND (rrna[filter] or trna[filter] or snrna[filter] or snorna[filter])”, while the barley small RNAs were obtained from the publicly available small RNA-seq dataset<sup>32</sup>. lncRNA candidates with significant alignment (E $\leq$ 0.001) with barley housekeeping lncRNAs and small RNAs were eliminated. Finally, a total of 19,908 high confidence barley lncRNAs were identified.

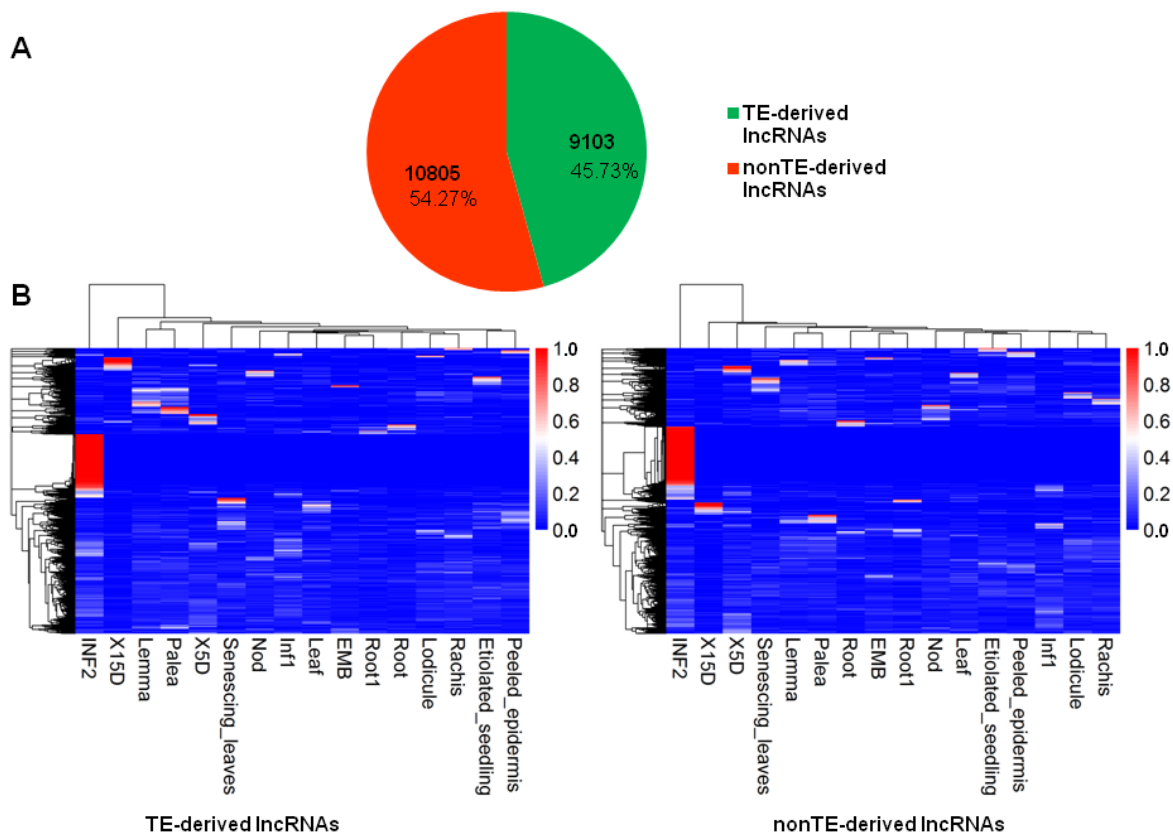
The majority of lncRNAs reported were found to be derived from transposable elements (TE)<sup>33</sup>. To discriminate the TE-derived barley lncRNAs from the nonTE-derived barley lncRNAs, we intersected the coordinates of lncRNAs with the barley TE annotation file using BEDTools (<http://bedtools.readthedocs.io/en/latest/>). If the proportion of lncRNAs intersected with TEs was more than 0.5, the lncRNA was considered a TE-derived lncRNAs. Otherwise, it was classified as a nonTE-derived lncRNA.

High quality RNA-seq reads of each biological replicate for each tissue/stage were mapped onto the annotated lncRNA region space of the barley genome using RSEM (ref. 34) with parameters “-p 8 --bowtie2 --estimate-rspd --append-names --output-genome-bam”. The statistic “Transcripts Per Million” (TPM) was adopted as the proxy of normalized expression-level for each barley lncRNA across different tissues/stages. The relative expression-levels in different tissues for TE-derived lncRNAs and nonTE-derived lncRNAs were visualized and compared in a heatmap (Supplementary Fig. 3.8).

**Supplementary Figure 3.7:** Annotation of long non-coding RNAs. Schematic diagram of the computational pipeline used for the identification of long non-coding RNAs in the barley genome assembly.



**Supplementary Figure 3.8:** Classification of TE-derived and nonTE-derived lncRNAs in barley. (A) The proportion of TE-derived and the other barley lncRNAs. (B) Expression-levels of barley lncRNAs across 16 different tissues/stages. Hierarchical clustering (Ward's method) of expression for the TE-derived lncRNAs and nonTE-derived lncRNAs that were expressed in at least one tissue suggests that tissue-specific expression for lncRNAs is common. Per-transcript normalization was applied to allow for visualization of relative expression in different tissues for all barley lncRNAs. The color scale ranges blue (low expression) to red (high expression).



### 3.8 Discovery and validation of microRNA loci

Nearly 50 million raw barley Illumina small RNA-seq reads were obtained from previous studies<sup>35-37</sup>. Adapter sequences were clipped using FASTX-Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)). Unclipped reads and/or clipped reads shorter than 18 nucleotides were excluded from downstream analyses. This yielded 33.1 million adapter-clipped reads. Redundant reads were then collapsed using the mapper.pl script<sup>38</sup>, producing 10.3 million non-redundant sequences. These were then aligned onto the barley pseudomolecules with zero mismatches using Bowtie<sup>39</sup>. Candidate barley microRNA (miRNA) loci were predicted using miRDeep-Plant (miRDeep-P) as previously described<sup>40</sup>. Predicted barley microRNA loci were excluded if: a) they have more than 20 copies in the barley genome<sup>41</sup>; b) contain Ns (ambiguous base calls) overlapping the loop region of the miRNA hairpin; c) the predicted precursors is shorter than 58 nt in length; d) candidate precursors have sequence similarity to the chloroplast genome, protein-coding genes, and repeats; e) candidate miRNA loci with a mature sequence longer than 24 nt; f) expressed in a single barley small RNA library; and g) have less than

five reads mapped onto the pre-miRNA. Precursor sequences passing all filtering criteria were retained as barley microRNAs.

To validate barley microRNAs three strategies were employed: i) screen sequence similarity against known pre-miRNA and mature miRNA sequences deposited in miRBase (release 21); ii) map candidate pre-miRNAs onto selected plant genomes; and iii) predict downstream microRNA regulatory targets and use degradome-seq libraries<sup>35,42</sup> to validate predicted targets.

Annotated barley pre-miRNAs and mature sequences were BLASTN<sup>41</sup> against miRBase (release 21)<sup>44</sup>. Mature sequences with sequence similarity to barley or other species microRNAs deposited in miRBase were classified as 'known' microRNAs, while those with no sequence similarity were classified as 'novel' microRNAs.

To evaluate the conservation of novel barley microRNAs in other plants species the available genome assemblies for 15 selected plant species were downloaded from [ftp.ensemblgenomes.org](http://ftp.ensemblgenomes.org): *Aegilops tauschii* (GCA\_000347335.1.30), *Arabidopsis lyrata* (v.1.0.30), *Arabidopsis thaliana* (TAIR10.30), *Brachypodium distachyon* (v.1.0.30), *Glycine max* (V1.0.30), *Medicago truncatula* (GCA\_000219495.2.30), *Musa acuminata* (MA1.30), *Oryza indica* (AMS465v1.30), *Oryza sativa* (IRGSP-1.0.30), *Setaria italica* (JGIv2.0.30), *Sorghum bicolor* (Sorbi1.30), *Triticum aestivum* (IWGSC1.0+popseq.30), *Triticum urartu* (GCA\_000347455.1.30), *Vitis vinifera* (IGGP\_12x.30) and *Zea mays* (AGPv3.30). Identified barley pre-miRNAs and mature sequences were mapped onto other plant genomes using BLAST+<sup>45</sup> and its 'blastn-short' task option.

## Supplementary Note 4: Analysis of gene families

### 4.1 OrthoMCL gene family clustering

Gene family clusters were defined from the barley high-confidence class genes and the annotated gene sets of three grasses from diverse grass sub-families and *Arabidopsis thaliana* using OrthoMCL software version 2.0. In the 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:

- Barley high-confidence genes (see Supplementary Note 3): **39,734** genes
- *Sorghum bicolor* v2.1: **33,032** genes
- *Brachypodium distachyon* v2.1: **31,694** genes
- Rice MSU7.0: **39,049** genes
- *Arabidopsis thaliana* TAIR10: **27,416** genes

Splice variants were removed from the data sets, keeping the representative/longest protein sequence prediction, and data sets were filtered for internal stop codons and incompatible reading frames. A total of 170,925 coding sequences from these five species were clustered into 24,337 gene families. In total, 8,608 clusters contained sequences from all five genomes. An overview about the gene family structure is given in Fig. 4a.

#### *GO over-/under-representation for specific groups/singletons*

Over-/under-representation of gene ontology (GO) terms in barley gene families compared to the other gene sets was analysed via hypergeometric testing using the functions GOstats and GSEABase from the Bioconductor R package<sup>46</sup> against a universe of all genes with GO annotations. REVIGO<sup>47</sup>, which removes redundant and similar terms from long GO lists by semantic clustering was applied to visualise the enrichment results.

#### *Ensembl Compara gene family clustering*

Gene families with barley-specific gene duplications as compared to other plant species were extracted from the ENSEMBL Compara pipeline<sup>48</sup> and analysed for over- or under-represented GO terms. Results are listed in Supplementary Tables 4.1-4.3.

#### *Expanded gene families in OrthoMCL and GO over-/under-representation*

We extracted gene models from three different distinct OrthoMCL subsets:

- a.) Barley genes in groups/clusters which are barley-specific (cluster/group contains only genes from barley) and cluster size > 1: "Barley-specific" set
- b.) Barley genes in groups/clusters where the barley gene copy number is significantly (p-value<0.05) expanded relative to any of the other species contained or the cluster (size>1)



only consists of barley genes (this is subset referred to in a.)): “Barley-expanded” set

- c.) Barley genes in groups/clusters where the barley gene copy number is significantly ( $p$ -value $<0.05$ ) expanded relative to any of the other species contained within the clusters (size $>1$ ) consisting only of barley genes (this is subset referred to in a.)): “Barley-expandedNOTexclusive” set

The individual gene sets were analysed for over- and under-represented GO terms from all GO categories “biological process”, “molecular function” and “cellular component”. Results are summarized and visualized in Supplementary Tables 4.1-4.3.

**Supplementary Table 4.1:** Over-represented GO terms in the GO category “biological process” in selected barley gene family subsets.

This table is provided as a separate Microsoft Excel spreadsheet.

**Supplementary Table 4.2:** Over-represented GO terms in the GO category “cellular component” in selected barley gene family subsets.

This table is provided as a separate Microsoft Excel spreadsheet.

**Supplementary Table 4.3:** Over-represented GO terms in the GO category “molecular function” in selected barley gene family subsets.

This table is provided as a separate Microsoft Excel spreadsheet.

**Supplementary Table 4.4:** Boundaries of genomic compartments

This table is provided as a separate Microsoft Excel spreadsheet.

**Supplementary Table 4.5:** GO term enrichment analysis between genomic compartments.

This table is provided as a separate Microsoft Excel spreadsheet.

## 4.2 Analysis of $\alpha$ -amylases

### *Identification of $\alpha$ -amylase encoding genes in the genome of barley cv. Morex*

The individual genes of the  $\alpha$ -amylase (*amy*) multi gene family were identified as follows: firstly, nucleotide sequences of known amy genes were downloaded from GenBank for *Arabidopsis thaliana*<sup>49-51</sup>, *Oryza sativa*<sup>52-54</sup> and *Hordeum vulgare*<sup>55-56</sup>. They were used for BLAST searches<sup>43</sup> against the genome databases of *Zea mays*, *Brachypodium distachyon* and *Sorghum bicolor*. The amy orthologues from these plants were identified from the hits with significant E-value (*E-values less than e-20*), high sequence identities (more than 90%) and sufficient coverage (more than 60% of query sequence) matching any one of the query sequences. Secondly, the nucleotide sequences of identified amy genes from *Sorghum bicolor*, *Brachypodium distachyon* and *Zea mays* were downloaded from genomic resources as listed in Supplementary Table 4.6. Together with the sequences from *A. thaliana*, *O. sativa* and *H. vulgare*, they were used to search the barley BAC assembly database using BLAST software with default parameters in order to obtain the BAC contigs bearing amy genes. The BAC contig IDs with corresponding amy genes are listed in Extended Data Table 4a. Thirdly, the amy gene sequences in the BAC contigs were identified after translating to peptides using GENSCAN program<sup>57</sup> and blasting to NCBI non-redundant protein sequence database. The peptide sequences identified from BACs were then used to blast the barley protein sequence database (Hv\_IBSC\_PGSB\_r1\_proteins\_HighConf.fa and Hv\_IBSC\_PGSB\_r1\_proteins\_LowConf.fa). The amy genes corresponding to each BAC contig were identified as top hits. Finally, the derived protein sequences from the searches outlined above were evaluated after alignments with two 3D structures of  $\alpha$ -amylase (AMY) proteins (1AMY and 1HT6<sup>56,57</sup>). Both of these proteins contain several characteristic AMY features including an  $(\alpha/\beta)_8$  barrel domain A, two  $\beta$  sheet domain B, five  $\beta$ -sheet domain C (carbohydrate binding domain) and three catalytic amino acids (Asp, Glu and Asp). When the proteins contained sequences of the domain A, the respective genes were considered as barley amy genes. Extended Data Table 4a provides a full list of all amy genes identified during this search. Two of the identified gene sequences in BAC contigs HVVMRXALLhB0076E06\_C1 and HVVMRXALLhA0174I01\_C3 are not considered when referring to the total count of amy genes, as these sequences are either located in a masked region of the assembly or are redundant to another amy gene sequence.

### *Identification of the barley AMY orthologs in other grass genomes*

GenomeThreader<sup>23</sup> (version 1.6.5) with parameters *seedlength 9* and *max gap width* of 50000 was used to identify putative orthologs and their genomic locations for all 12 barley AMY proteins in other grasses with completed genome sequences. This search included: *O. sativa* (MSU7 genome assembly), *B. distachyon* (assembly version 2.0), *S. bicolor* (assembly version 2.0), *Z. mays* (assembly AGPv3.21), and *Triticum aestivum* (bread wheat – hexaploid, TGAC assembly version 1). Both GenomeThreader and BLAST searches obtained the same numbers of AMY orthologs in other grass species. In bread wheat, AMY orthologs were identified with GenomeThreader only with the respective coding sequences on chromosomes 2AL, 2BL, 2DL, 3B, 3DL, 5AL, 5BL, 5DL, 6AL (3 copies), 6BL (5 copies), 6DL (3 copies), 7AL, 7BL (3 copies), 7DL (2 copies), as well as on the unanchored scaffolds U\_3AL\_7AS, U\_7DL\_7DS and U\_7AL\_6AL, U\_2AS\_7AL.

To evaluate the GenomeThreader mapping we also mapped the barley AMY predictions back to the barley genome assembly (described in this manuscript) and obtained 12 predicted genome locations for the protein set.

### Phylogenetic analysis of $\alpha$ -amylase multi-gene family in grass genomes

Phylogenetic analysis of the  $\alpha$ -amylase (amy) multi-gene family in grasses was done using the Phylogeny.fr platform<sup>60</sup>. A multiple sequence alignment was created with the protein sequences deduced from all full length or near full length amy gene sequences identified in the genomes of *H. vulgare*, *T. aestivum*, *O. sativa*, *S. bicolor*, *Z. mays*, and *B. distachyon* (see Supplementary Note 4.2, section “Identification of the barley AMY orthologs in other grass genomes” and Supplementary Table 4.7) using Muscle software in full mode. Proteins derived from truncated sequences (sequences which were missing either start or stop codon and which were more than 50% shorter than any other sequences identified in the respective subfamily) were not included in the alignment. This led to the exclusion of three truncated amy1-like gene copies on chromosome 6D in wheat and to the exclusion of an amy3-like sequence from *S. bicolor* (Sb02g026625). Additionally, the sequence derived from *Hvamy1\_1e* was omitted from the alignment, as the cv. Morex reference assembly is missing sequence information in the coding region for this gene.

Prior to phylogenetic tree construction the multiple sequence alignment was curated with Gblocks, settings were as follows: minimum sequence for flanking positions = 85%; maximum contiguous non-conserved position = 8; minimum block length = 10; gaps in final block = half. After the step of alignment curation, a phylogenetic tree was created using PhyML. Statistical testing of branch support was done by bootstrapping (n=100), the applied amino acid substitution model was WAG and the number of substitution rate categories was set to 4. The gamma distribution parameter and the proportion of invariable sites were estimated. The resulting tree is shown in Fig. 4b.

### *amy1\_1* copy number evaluation by PCR

Barley (*Hordeum vulgare* L.) cultivars Morex, Barke, Bowman, Masan Naked 1, Akashinriki and Etincel were grown in a greenhouse to the second leaf stage (growth conditions: 18°C under 16 h light/8 h dark cycles, light intensity was set to a photon flux of 300  $\mu\text{mol m}^{-2} \text{sec}^{-1}$ ). Genomic DNA was extracted from green leaf material, and *amy1\_1* promoter fragments were PCR-amplified using the REDExtract-N-Amp™ Plant PCR Kit (Sigma-Aldrich, St. Louis, MO, USA) according to the manufacturer's instructions, respectively.

PCR with primers CD52\_amy1fw (5'-CGTAGCAGTGCAGCGTGAAGTCATAG-3') and CD53\_amy1rc (5'-CATTCGTTTCGATGAGCATTCAATTCGTGAGGG-3') was performed for 40 cycles (initial denaturation at 94°C/3 min followed by 40 cycles of 94°C/30 s, 59°C/30 s, and 72°C/60 s for extension, with a final extension step of 72°C/10 min). The PCR for the *amy1\_1a* copy-specific promoter fragment was performed for 40 cycles (initial denaturation at 94°C/3 min followed by 40 cycles of 94°C/30 s, 61°C/30 s, and 72°C/90 s for extension, with a final extension step of 72°C/10 min) using primer combination CD54\_fw1a (5'-CAGCGTGAAGTCATAGATAGACTGCTTATCACG-3') and CD58\_amy1rc (5'-GGACAAGATCTTACCTGAAAGAGGACTTGCC-3'). The *amy1\_1b* copy-specific promoter fragment was amplified using primers CD55\_fw1b (5'-GTCAGTTGGATCTGCTCCGGCCATTG-3') and CD58\_amy1rc, while the *amy1\_1c* copy specific promoter fragment was amplified with primers CD56\_fw1c (5'-GGCGATAACGTCTCCGGCTAGC-3') and CD58\_amy1rc. PCR products obtained with primers CD52\_amy1fw and CD53\_amy1rc were resolved using agarose gel electrophoresis (UltraPure™ Agarose; Invitrogen/ThermoFisher Scientific, Waltham, MA, USA), 100V for 4 h (4% agarose) and copy-specific PCR products for 3 h (2% agarose) and visualized with ethidium bromide staining. PCR products were purified using the NucleoSpin® Gel and PCR Clean-Up Kit (Macherey-Nagel GmbH &

Co. AG, Germany) according to the manufacturer's instructions for Sanger sequencing (StarSEQ, Mainz, Germany).

#### *Analysis for conserved sequence motifs in promoter regions*

Sequence identities in the amy1 promoter regions 500 bp upstream of the translational start codon ATG are high within the amy1 subfamily with 99–100% identity for *amy1\_1a*, *amy1\_1b* and *amy1\_1d*, but slightly lower for *amy1\_1c* and *amy1\_2* (64–84%). All amy1 promoters (except *amy1\_1e*, which is truncated) contain conserved GA-responsive element (GARE) TAACAACTCCGG and a TATCCA(C/T) box, which are necessary for GA responses<sup>61</sup>. A pyridine box CCTTTT may be involved in the enhancement of transcription. A cAMP-like responsive element (TGAGCTC) is conserved (Supplementary Fig. 4.1).

Promoters (regions 500 bp upstream of the start codon) within the amy2 subfamily generally share 60–68% sequence identity. They contain a conserved GA-responsive element (GARE) TAACAGAG(T/G)C(C/T)GG, which is necessary for GA responses<sup>61–63</sup>. Both *amy2\_1* and *amy2\_3* have a conserved pyrimidine box (CCTTTT) and a TATCCAT box, which control transcript levels, while a cAMP-like responsive element (TGAGCTC) and an O2S (CTTGXXTCATC) motif are present upstream of *amy2\_2*. A TTCCCATGGA(A/G)...TGCC box responsible for transcript levels is found in *amy2\_1* (Supplementary Fig. 4.2). This analysis shows that gene expression regulation by GA might be very different among the three amy2 genes and also different from the amy1 subfamily.

#### *Role of $\alpha$ -amylases in barley and significance of the reference sequence based de novo gene annotation*

Barley is the primary cereal used for the production of beer, which begins with the malting of barley grains. Malting is a seed germination process during which the aleurone layer produces and secretes a series of starch degrading enzymes into the starchy endosperm of the grains. This is followed by mashing, the enzymatic conversion of starch into fermentable sugars through the action of the aforementioned enzymes. The mash is then boiled with hops, and yeast is added to ferment the sugars into alcohol to produce beer.

The barley endosperm starch is composed of the two  $\alpha$ -D-glucose homo-polymers amylose and amylopectin. While amylose is a linear molecule of  $\alpha$ -1,4-linked glucose molecules, amylopectin is a larger molecule with additional  $\alpha$ -1,6 branching points<sup>64</sup>. Starch degradation during malting and mashing is catalyzed by a series of enzymes, namely  $\alpha$ - and  $\beta$ -amylases, possessing endo- and exo-1,4-hydrolytic activity respectively, and the 1,6-hydrolytic enzyme limit dextrinase<sup>65</sup>. The ability of a batch of germinated barley, the malt, to enzymatically break down its own starch into fermentable sugars is referred to as diastatic power. Malt derived from modern elite malting barley cultivars usually possesses sufficient diastatic power to reduce its own starch to sugar several times.

The large amylopectin molecules are broken down to smaller units, linear and branched maltodextrins, by the endo-hydrolytic activity of  $\alpha$ -amylase.  $\beta$ -amylase removes maltose from the non-reducing end of maltodextrins and starch. Limit dextrinase cleaves  $\alpha$ -1,6 branching points<sup>64</sup>. While  $\beta$ -amylase is produced during seed development and stored in the endosperm in an inactive form, both  $\alpha$ -amylase and limit dextrinase are secreted from the aleurone layer of the barley grain upon germination<sup>66,67</sup> and  $\alpha$ -amylase is present in large, non-limiting quantities in barley malt<sup>65</sup>.

Due to their central role in the breakdown of starch during malting, barley  $\alpha$ -amylases and the genes encoding these enzymes are of specific interest in the context of malting and brewing. These genes were initially distinguished by the isoelectric point of the respective proteins as high and low pI isoenzymes<sup>68</sup> and two gene families encoding the high pI and low pI enzymes were found to be located on chromosome 6H and 7H, respectively. It should be noted that in the past no consistent nomenclature of barley amy genes and the deduced proteins was applied in literature. Herein the gene family on chromosome 6H is designated as *amy1* and the gene family on chromosome 7H as *amy2* following Brown and Jacobsen<sup>72</sup>. The *amy3* gene was labelled as such, as it was identified as an orthologue of the rice *amy3* gene. The remaining two barley amy genes were designated as *amy4*.

Hybridization experiments suggested the presence of six or more high pI and three highly similar low pI genes<sup>56,69</sup>. More recent work identified four distinct  $\alpha$ -amylase genes in barley and showed expression of two genes during grain filling<sup>70</sup>. Still until today the exact copy number of  $\alpha$ -amylase genes in barley remained unknown<sup>71</sup> and the identification of the individual members of the two gene families on 6H and 7H as well as additional  $\alpha$ -amylase genes on other chromosomes was – to the best of our knowledge – still missing.

The current analysis hence confirms the presence of multi-gene families on chromosome 6H and 7H, as already suggested in earlier studies<sup>69,72</sup>, and it also explains why resolving the actual copy number of the amy genes present in the barley genome remained challenging until today. Five of the six *amy1* genes identified in this study share 99.5 to 100% sequence identity with each other at the protein level and more than 99.8% to 100% sequence identity at the nucleotide level, in between start and stop codons including the intron sequences, rendering separation of different copies by molecular biology methods like PCR or RT-PCR practically impossible. These almost identical gene copies were designated *amy1\_1a*, *\_1b*, *\_1c*, *\_1d*, and *\_1e*. In previous work the *amy1\_1* gene was referred to as *amy6-4* (ref. 56). The sixth *amy1*-like gene copy, *amy1\_2*, is located on chromosome 6H, too, and shares only 83% sequence identity with the *amy1\_1* sequence. The *amy1\_2* gene has been designated as *amy46* in earlier studies<sup>56</sup>. While proteins encoded by the *amy1* and the *amy2* gene families share about 72% sequence identity with each other, proteins encoded by the *amy3* and the *amy4* genes are notably different and share only 65% and 20 – 40% sequence identity, respectively, with the *amy1\_1* encoded protein.

Both *amy1* and *amy2* genes are expressed in aleurone cells upon treatment with GA (ref. 56), which coincides with the role of the encoded  $\alpha$ -amylases in starch breakdown during germination. The presence of at least three full length *amy2* and five full length *amy1* gene sequences in the barley genome might hence have positively influenced the applicability of barley in industrial beer production. An analysis of the *amy1* subfamily promoter sequences revealed that the regulatory motifs in *amy1* genes are identical within the subfamily, implicating that the transcription of all genes of the subfamily may be regulated by similar mechanisms. The gene duplication events within the *amy1* family may therefore have led to an increase in gene expression and possibly increased  $\alpha$ -amylase enzyme activity during germination in barley.

**Supplementary Table 4.6:** Coding sequences used to search for  $\alpha$ -amylase genes in the barley genome.

Species	Gene ID	URL
<i>Hordeum vulgare</i>	K02637.1 ( <i>amy6-4</i> )	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
	J04202.1 ( <i>amy46</i> )	
	X05166.1 ( <i>amy32b</i> )	
<i>Arabidopsis thaliana</i>	AT1G76130.1	<a href="https://www.arabidopsis.org/">https://www.arabidopsis.org/</a>
	AT1G69830.1	
	AT4G25000.1	
<i>Oryza sativa</i>	LOC_Os08g36900.1	<a href="http://rice.plantbiology.msu.edu/analyses_search_blast.shtml">http://rice.plantbiology.msu.edu/analyses_search_blast.shtml</a>
	LOC_Os08g36910.1	
	LOC_Os04g33040.1	
	LOC_Os01g51754.1	
	LOC_Os02g52700.1	
	LOC_Os02g52710.1	
	LOC_Os09g28420.1	
	LOC_Os09g28400.1	
	LOC_Os06g49970.2	
	LOC_Os01g25510.1	
	LOC_Os01g25510.1	
<i>Zea mays</i>	GRMZM5G863596	<a href="ftp://ftp.ensemblgenomes.org/pub/plants/release-30">ftp://ftp.ensemblgenomes.org/pub/plants/release-30</a>
	GRMZM2G138468	
	GRMZM2G422938	
	GRMZM2G103055	
	GRMZM2G070172	
	GRMZM2G074781	
	GRMZM2G422938	
	GRMZM2G081502	
<i>Brachypodium distachyon</i>	Bradi3g58010.1	<a href="http://www.plantgdb.org/BdGDB/cgi-bin/blastGDB.pl">http://www.plantgdb.org/BdGDB/cgi-bin/blastGDB.pl</a>
	Bradi1g35050.1	
	Bradi4g32140.1	
	Bradi5g08800.1	
	Bradi2g48150.1	
<i>Sorghum bicolor</i>	Sb03g032830.1	<a href="http://www.plantgdb.org/SbGDB/cgi-bin/blastGDB.pl">http://www.plantgdb.org/SbGDB/cgi-bin/blastGDB.pl</a>
	Sb07g023020.1	
	Sb06g015110.1	
	Sb04g034140.1	
	Sb02g026620.1	
	Sb07g023020.2	
	Sb02g023790.1	
	Sb02g026610.1	
	Sb04g034150.1	
	Sb02g023250.1	
	Sb02g026625.1	
	Sb02g026625.1	

**Supplementary Table 4.7:** Sequences used for the phylogenetic analysis of  $\alpha$ -amylase multi-gene family in grass genomes.

This table is provided as a separate Microsoft Excel spreadsheet.



**Supplementary Figure 4.1:** Sequence alignment of promoter regions of amy1 genes. Gene IDs are described in Extended Data Table 4a.

```

amy1_1a: TAAACCCCTTTTGGGGTTGATCATGTACAAAACATATAACCACTCCCAGTTGAGTAGTTTCCGTGTTCTTGCAAATTCCTTGGCTTGC 87
amy1_1b: TAAACCCCTTTTGGGGTTGATCATGTACAAAACATATAACCACTCCCAGTTGAGTAGTTTCCGTGTTCTTGCAAATTCCTTGGCTTGC 87
amy1_1c: AATTAGTGAAGCAATCTATATTTCTTGCAACACATACTCCTACCTCAGCAATTGAATGCTCTGCAACGAATCAATATTGGA 85
amy1_1d: TAAACCCCTTTTGGGGTTGATCATGTACAAAACATATAACCACTCCCAGTTGAGTAGTTTCCGTGTTCTTGCAAATTCCTTGGCTTGC 87
amy1_1e: ----- 87
amy1_2 : GAATTCGATCTCTGGCAGCACTTATGTCCGGTTTATCCCTCTCGAGAAAGGCCACTCATCCAGGTTATTCCAGGAAATTTGCGCAGGAATTT 91

amy1_1a: CTACAGACATACAGTTGCGGTAGATGAAGGTTTGTAAATGTAACCACAGCACACTATTTCGATGAAAAATGCTCGA 162
amy1_1b: CTACAGACATACAGTTGCGGTAGATGAAGGTTTGTAAATGTAACCACAGCACACTATTTCGATGAAAAATGCTCGA 162
amy1_1c: TATGTAGATCTCTTCGGACTGAAAAAGTTTGAAACTGCTAGCCACAGCACACTATTTCATGAAAAATGCTCGA 162
amy1_1d: CTACAGACATACAGTTGCGGTAGATGAAGGTTTGTAAATGTAACCACAGCACACTATTTCATGAAAAATGCTCGA 162
amy1_1e: ----- 162
amy1_2 : CTGACCCGGATTCTCGCTTTGTTAACTGAAATGCGCAAGTAACCGTCAGTTGGCGTCAGATCTTACGTTGCAACAGGATAACTGACAGGA 182

amy1_1a: ATGTTCTGTCTCAGAAAAACAGAGGTTTCAGGATAACTGACGGTCTGATTGACCGGTGCCTTCTTATGGAAGGCGAAGGCTGCCTC 248
amy1_1b: ATGTTCTGTCTCAGAAAAACAGAGGTTTCAGGATAACTGACGGTCTGATTGACCGGTGCCTTCTTATGGAAGGCGAAGGCTGCCTC 248
amy1_1c: ATGTTCTGTCTCAGAAAAACAGAGGTTTCAGGATAACTGACGGTCTGATTGACCGGTGCCTTCTTATGGAAGGCGAAGGCTGCCTC 248
amy1_1d: ATGTTCTGTCTCAGAAAAACAGAGGTTTCAGGATAACTGACGGTCTGATTGACCGGTGCCTTCTTATGGAAGGCGAAGGCTGCCTC 248
amy1_1e: ----- 248
amy1_2 : ATTATCTGATTCTGAGGAATTCAGAGTTTCAGGAGGATAATGACGTGGTATTGCGCGGTGCCTTCTCATGGAAGCCGGTG 265

          CCTTTT      TGAGCTC      TAACAACTCCGG      TATCCA      C
amy1_1a: CATCTACATCACTTGGGCATTGAATCGCCTTTTGAGCTCACCGTACCGGCCGATAACAAACTCCGGCCGACATATCCA-----CTGGC 331
amy1_1b: CATCTACATCACTTGGGCATTGAATCGCCTTTTGAGCTCACCGTACCGGCCGATAACAAACTCCGGCCGACATATCCA-----CTGGC 331
amy1_1c: CATCTACATCACTTGGGCATTGAATCGCCTTTTGAGCTCACCGTACCGGCCGATAACAAACTCCGGCCGACATATCCA-----CTGGC 331
amy1_1d: CATCTACATCACTTGGGCATTGAATCGCCTTTTGAGCTCACCGTACCGGCCGATAACAAACTCCGGCCGACATATCCA-----CTGGC 331
amy1_1e: ----- 331
amy1_2 : -----CTCATCTCATTCGCTTTTGAGCTCACCGCACCGGCCGATAACAAACTCCGGCCGACATATCCATCGATGCACGGC 340

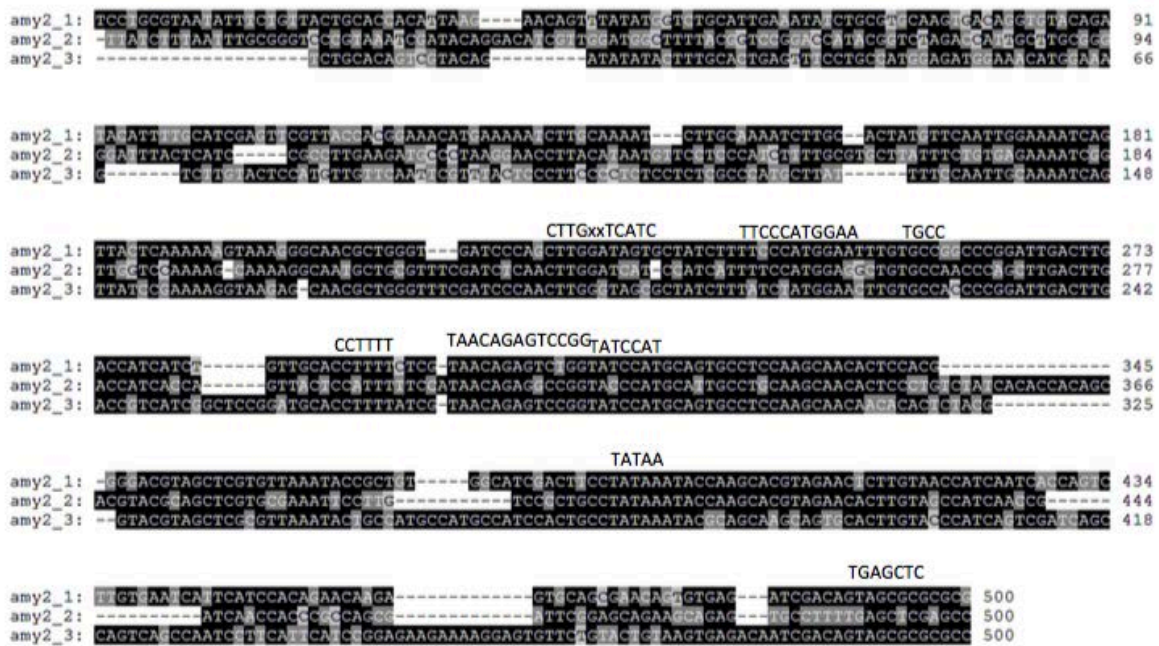
          TATAA
amy1_1a: CCAAAGGAGCATTCAAGCCGAGCACACGAGAAAGTGATTGCAAGTTGCACACCGGCAGCAATTCGGGCATGCTGCAGCACACTATAAATA 422
amy1_1b: CCAAAGGAGCATTCAAGCCGAGCACACGAGAAAGTGATTGCAAGTTGCACACCGGCAGCAATTCGGGCATGCTGCAGCACACTATAAATA 422
amy1_1c: CCAAAGGAGCATTCAAGCCGAGCACACGAGAAAGTGATTGCAAGTTGCACACCGGCAGCAATTCGGGCATGCTGCAGCACACTATAAATA 422
amy1_1d: CCAAAGGAGCATTCAAGCCGAGCACACGAGAAAGTGATTGCAAGTTGCACACCGGCAGCAATTCGGGCATGCTGCAGCACACTATAAATA 422
amy1_1e: -----CGAGCACACGAGAAAGTGATTGCAAGTTGCACACCGGCAGCAATTCGGGCATGCTGCAGCACACTATAAATA 73
amy1_2 : CCAAGGAGCATTGAAGCCGAGCACACCGGAATATGTTCTGCAAGTTGCCACCG-----GCATGCTGCAGCACACTATAAATA 419

amy1_1a: CCTGGCCAGACACACAAGCTGAATGCATCAGTTCTCCATCGTACTCTTCGAGAGCACAGCAAGAGAGAGCTGAAGAAC 500
amy1_1b: CCTGGCCAGACACACAAGCTGAATGCATCAGTTCTCCATCGTACTCTTCGAGAGCACAGCAAGAGAGAGCTGAAGAAC 500
amy1_1c: CCTGGCCAGACACACAAGCTGAATGCATCAGTTCTCCATCGTACTCTTCGAGAGCACAGCAAGAGAGAGCTGAAGAAC 500
amy1_1d: CCTGGCCAGACACACAAGCTGAATGCATCAGTTCTCCATCGTACTCTTCGAGAGCACAGCAAGAGAGAGCTGAAGAAC 500
amy1_1e: CCTGGCCAGACACACAAGCTGAATGCATCAGTTCTCCATCGTACTCTTCGAGAGCACAGCAAGAGAGAGCTGAAGAAC 151
amy1_2 : CCTGGCCAGACACACAAGCTGAATGCATCAGTTCTCCATCGTCTCTCTCCAGAGCACAGCTAGCTAGAGCTCAAGATC 500

```



**Supplementary Figure 4.2:** Sequence alignment of promoter regions of amy2 genes. Gene IDs are described in Extended Data Table 4a.



### 4.3 Analysis of SWEET and VPE genes

#### *Identification of SWEET encoding genes in the genome of barley cv. Morex*

In angiosperms, the embryo and endosperm, established after double fertilization, are developed fully covered by maternal tissues, which not only protect filial organs against detrimental environment but also deliver nutrients from mother plant<sup>73</sup>. As has been experimentally shown, the main conduit for maternal nutritional supplies for the developing barley grains is through the nucellar projection and endosperm transfer cells<sup>74</sup>. This differs in rice grains with two pathways involved in the transport of nutrients within the developing caryopsis: one via a pathway analogous to the nucellar projection pathway of barley and the other via the nucellar epidermis<sup>75</sup>. Because endosperm and embryo are symplastically isolated from maternal tissues, assimilate transport must occur apoplastically with activation of numerous transporters. The recently described family of membrane proteins, called SWEETs, transports specific sugar molecules across the membrane down a concentration gradient<sup>76</sup>. Contrary to vertebrate genomes (human, rat, *Danio rerio*, *Xenopus laevis*) with only a single *SWEET* gene, plant genomes carry multiple *SWEET* genes suggesting that they are essential for plant survival and indicate for diverse functionalization<sup>77</sup>. It has been experimentally proven that SWEETs from clade III are able to transport sucrose<sup>78</sup>.

There are 17 *SWEETs* encoded in the Arabidopsis genome and 21 in rice. The barley genome encodes a total of 23 *SWEET* genes (Extended Data Table 4b). We identified a small expansion in the *SWEET11* sub-group to two genes (*HvSWEET11a* and *HvSWEET11b*) in barley compared to one gene in rice and Arabidopsis. We analyzed the expression and transcript localization of the *SWEET11a* and *SWEET11b* genes in barley grains by qRT-PCR and in situ hybridization. Both genes were predominantly transcribed in vascular bundle and nucellar projection (Fig. 4c, Extended Data Fig. 7a), both tissues devoted to assimilate transfer in maternal seed part, albeit their levels of expression and topographical localization were slightly different. Besides developing grains, the *SWEET11a* and *SWEET11b* transcripts were found in anthers but were barely detectable in vegetative tissues and organs (Extended Data Fig. 7b). To conclude, the expansion of *SWEET11* genes in barley might be linked with sexual reproduction and is required for more effective assimilate transfer within maternal seed parts (and probably from sporophytic generation to gametophyte in anthers and gynoecia).

#### *Identification of VPE encoding genes in the genome of barley cv. Morex*

Effective nutrient transfer requires programmed cell death (PCD) at margins of nucellar projection of cereal grains<sup>79,80</sup>. In both animal and plant kingdoms, the molecular basis for PCD relies on caspase-like activities. However, because plant genomes do not harbor genes encoding true caspases, other proteases were found to exhibit caspase-like activities. Vacuolar processing enzyme (VPE, also referred to as legumain) was shown to possess caspase-1-like activity<sup>81</sup> and is required for diverse types of PCD in plants<sup>82</sup>.

The barley genome encodes eight *VPE* genes while the *VPE* gene family in rice consists of 5 members. The barley subfamily *HvVPE2a-HvVPE2d* consists of four very similar genes (sharing 84.7-95.2% identity in ORFs at nucleotide level) indicating a recent genic expansion. *HvVPE2a*, *HvVPE2b* and *HvVPE2c* genes are located closely on the chromosome 2 while the position of *HvVPE2d* is more distant at the same chromosome. During grain development, *HvVPE2a*, *HvVPE2b* and *HvVPE2d* showed similar expression patterns being exclusively detected in nucellus and nucellar projection with an expression maximum between 7 and 10 DAF<sup>83</sup>. Expression of *HvVPE2c* was deregulated and

barely detectable. This gene shows also the lowest similarity to the other three genes (84.7-85.4% identity at nucleotide level) indicating early duplication and further modification of gene activity in barley seeds. Simultaneous repression of *HvVPE2a*, *HvVPE2b* and *HvVPE2d* results in delay of PCD in the nucellus and nucellar projection and lower starch accumulation in endosperm probably due to compromised assimilate transfer (V. Radchuk, unpublished results).

#### *Gene copy number determination for SWEET and VPE gene families*

GenomeThreader<sup>23</sup> version 1.6.5 with parameters seedlength 9 and max gap width of 50000 was used to identify putative orthologs and their genomic locations for all previously known *SWEET* (Extended Data Table 4b) and *VPE* genes<sup>83</sup> (Extended Data Table 4c) in the latest barley genome and the rice genome (MSU7 genome assembly). Following that, the predicted genomic locations of *VPE2a-d* were checked for overlap with predicted gene loci in both the high-confidence and low-confidence gene classes. Gene loci with maximum positional overlap with the predicted GenomeThreader locations were selected, regardless of their predicted open reading frame (and hence predicted protein sequence). As recommended<sup>6</sup>, barley *SWEET* genes are named according to their closest rice sequence homologs followed by a letter designation to distinguish closely related genes.

To evaluate the GenomeThreader mapping we also used BLAST to search for the previously known *SWEET* and *VPE* genes in the latest barley genome and the rice genome (MSU7 genome assembly), confirming the results of the GenomeThreader analysis.

## Supplementary Note 5: Diversity analysis

We investigated patterns of diversity in eight genes known to influence the seasonal growth habit of barley. Four of these genes were represented in the barley exome capture design<sup>8</sup>. For these, unbiased heterozygosity was low in winter barleys (WB) and high in spring barleys (SB) at *HvELF3*<sup>84</sup> (WB: 0.089, SB: 0.323) and *HvVRN-H1*<sup>85</sup> (WB: 0.028, SB: 0.361), and *vice versa* at *HvCEN*<sup>86</sup> (WB: 0.32, SB: 0.002) and *HvPPD-H1 (HvPRR7)*<sup>87</sup> (WB: 0.28, SB: 0.013), as expected based on their dominant/recessive nature, the function of their ancestral alleles and the strong geographical structuring of causal mutations observed in landraces sampled from across the geographic range of the species<sup>88</sup>. For the remaining four, we assessed diversity in adjacent genes on the pseudomolecule. Average unbiased heterozygosity at adjacent loci was low in WB and high in SB at the *HvCBF* gene cluster<sup>89</sup> (WB: 0.05, SB: 0.315), where functional alleles are required for tolerance to frost, at *HvVRN-H2*<sup>90</sup> (WB: 0.0, SB: 0.314) where dominant functional alleles repress flowering in WB, and at *HvPPD-H2*<sup>91</sup> (WB: 0.027, SB: 0.19), which promotes flowering in weakly vernalised WB and SB under increasing photoperiods. *HvVRN-H3 (HvFT)*<sup>92</sup> (WB: 0.113, SB: 0.18s) exhibited moderate diversity in both genepools, a feature that has been proposed to provide natural variation in vernalisation requirement and increase adaptive capacity. If we assume a strong requirement for specific alleles at all loci that show a contrasting pattern of diversity between SB and WB genepools, the chances of returning an optimal genotype from a winter by spring cross is vanishingly small.

**Supplementary Table 5.1:** List of barley cultivars subjected to exome sequencing.

This table is provided in a separate Microsoft Excel spreadsheet.

## Supplementary References

1. Arend, D. *et al.* PGP repository: a plant phenomics and genomics data publication infrastructure. *Database* **2016**, baw033 (2016).
2. Zimmermann, G., Bäuml, H., Mock, H.-P., Himmelbach, A. & Schweizer, P. The multigene family encoding germin-like proteins of barley. Regulation and function in basal host resistance. *Plant Physiology* **142**, 181-192 (2006).
3. Hövel, I., Louwers, M. & Stam, M. 3C technologies in plants. *Methods* **58**, 204-211 (2012).
4. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
5. Belton, J.M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-76 (2012).
6. Mascher, M. *et al.* Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J* **76**, 494-505 (2013).
7. Kalhor, R., Tjong, H., Jayatilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* **30**, 90-8 (2012).
8. Himmelbach, A., Knauft, M. & Stein, N. Plant sequence capture optimised for Illumina sequencing. *Bio Protoc* **4**, e1166 (2014).
9. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-3011 (2014).
10. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170 (2014).
11. Magoc, T. & Salzberg, S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-63 (2011).
12. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
13. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
14. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576 (2012).
15. Prlić, A. *et al.* BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **28**, 2693-2695 (2012).
16. International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-6 (2012).
17. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-8 (2010).
18. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793-800 (2005).
19. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-6 (2009).
20. Matsumoto, T. *et al.* Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol* **156**, 20-8 (2011).
21. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965-978 (2005).
22. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
23. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357-360 (2015).
24. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).

25. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).
26. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-5 (2009).
27. International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
28. Beier, S. *et al.* Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Scientific Data* doi:10.1038/sdata.2017.44 (2017).
29. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-2 (2015).
30. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
31. Li, L. *et al.* Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biology* **15**, 1 (2014).
32. Schreiber, A.W., Shi, B.-J., Huang, C.-Y., Langridge, P. & Baumann, U. Discovery of barley miRNAs through deep sequencing of short reads. *BMC Genomics* **12**, 1 (2011).
33. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics* **9**, p.e1003470 (2013).
34. Li B., Dewey C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
35. Curaba, J., Spriggs, A., Taylor, J., Li, Z. & Helliwell, C. miRNA regulation in the early development of barley seed. *BMC Plant Biol* **12**, 120 (2012).
36. Lv, S. *et al.* Identification and characterization of microRNAs from barley (*Hordeum vulgare* L.) by high-throughput sequencing. *Int J Mol Sci* **13**, 2973-84 (2012).
37. Hackenberg, M. *et al.* A comprehensive expression profile of microRNAs and other classes of non-coding small RNAs in barley under phosphorous-deficient and -sufficient conditions. *DNA Res* **20**, 109-25 (2013).
38. Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**, 37-52 (2012).
39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
40. Yang, X. & Li, L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* **27**, 2614-5 (2011).
41. Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P. & Anderson, T.A. Conservation and divergence of plant microRNA genes. *Plant J* **46**, 243-59 (2006).
42. German, M.A., Luo, S., Schroth, G., Meyers, B.C. & Green, P.J. Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc* **4**, 356-62 (2009).
43. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).
44. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154-8 (2008).
45. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
46. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
47. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
48. Bolser, D., Staines, D.M., Pritchard, E. & Kersey, P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Methods Mol Biol* **1374**, 115-40 (2016).



49. Doyle, E.A., Lane, A.M., Sides, J.M., Mudgett, M.B. & Monroe, J.D. An alpha-amylase (At4g25000) in Arabidopsis leaves is secreted and induced by biotic and abiotic stress. *Plant Cell Environ* **30**, 388-98 (2007).
50. Seung, D. *et al.* Arabidopsis thaliana AMY3 is a unique redox-regulated chloroplastic alpha-amylase. *J Biol Chem* **288**, 33620-33 (2013).
51. Stanley, D., Fitzgerald, A.M., Farnden, K.J. & MacRae, E.A. Characterisation of putative alpha-amylases from apple (*Malus domestica*) and Arabidopsis thaliana. *BIOLOGIA-BRATISLAVA* **57**, 137-148 (2002).
52. Huang, N., Sutliff, T.D., Litts, J.C. & Rodriguez, R.L. Classification and characterization of the rice alpha-amylase multigene family. *Plant Mol Biol* **14**, 655-68 (1990).
53. Mitsui, T., Ueki, Y. & Igaue, I. Biosynthesis and secretion of  $\alpha$ -amylase by rice suspension-cultured cells: purification and characterization of  $\alpha$ -amylase isozyme H. *Plant physiology and biochemistry* **31**, 863-874 (1993).
54. Ranjhan, S., Litts, J.C., Foolad, M.R. & Rodriguez, R.L. Chromosomal localization and genomic organization of alpha-amylase genes in rice (*Oryza sativa* L.). *Theor Appl Genet* **82**, 481-8 (1991).
55. Karrer, E.E., Litts, J.C. & Rodriguez, R.L. Differential expression of alpha-amylase genes in germinating rice and barley seeds. *Plant Mol Biol* **16**, 797-805 (1991).
56. Khursheed, B. & Rogers, J.C. Barley alpha-amylase genes. Quantitative comparison of steady-state mRNA levels from individual members of the two different families expressed in aleurone cells. *J Biol Chem* **263**, 18953-60 (1988).
57. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).
58. Kadziola, A., Abe, J., Svensson, B. & Haser, R. Crystal and molecular structure of barley alpha-amylase. *J Mol Biol* **239**, 104-21 (1994).
59. Robert, X. *et al.* The structure of barley alpha-amylase isozyme 1 reveals a novel role of domain C in substrate recognition and binding: a pair of sugar tongs. *Structure* **11**, 973-84 (2003).
60. Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**, W465-9 (2008).
61. Gubler, F. & Jacobsen, J.V. Gibberellin-responsive elements in the promoter of a barley high-pl alpha-amylase gene. *Plant Cell* **4**, 1435-41 (1992).
62. Lanahan, M.B., Ho, T.H., Rogers, S.W. & Rogers, J.C. A gibberellin response complex in cereal alpha-amylase gene promoters. *Plant Cell* **4**, 203-11 (1992).
63. Rogers, J.C., Lanahan, M.B. & Rogers, S.W. The cis-acting gibberellin response complex in high pl alpha-amylase gene promoters. Requirement of a coupling element for high-level transcription. *Plant Physiol* **105**, 151-8 (1994).
64. Bahaji, A. *et al.* Starch biosynthesis, its regulation and biotechnological approaches to improve crop yields. *Biotechnol Adv* **32**, 87-106 (2014).
65. Bamforth, C.W. Barley and malt starch in brewing: A general review. *Technical Quarterly-Master Brewers Association of the Americas* **40**, 89-97 (2003).
66. Zeeman, S.C., Kossmann, J. & Smith, A.M. Starch: its metabolism, evolution, and biotechnological modification in plants. *Annu Rev Plant Biol* **61**, 209-34 (2010).
67. Shahpiri, A., Talaei, N. & Finnie, C. Spatio-temporal appearance of alpha-amylase and limit dextrinase in barley aleurone layer in response to gibberellic acid, abscisic acid and salicylic acid. *J Sci Food Agric* **95**, 141-7 (2015).
68. Bak-Jensen, K.S. *et al.* Spatio-temporal profiling and degradation of alpha-amylase isozymes during barley seed germination. *FEBS J* **274**, 2552-65 (2007).
69. Muthukrishnan, S., Gill, B.S., Swegle, M. & Chandra, G.R. Structural genes for alpha-amylases are located on barley chromosomes 1 and 6. *J Biol Chem* **259**, 13637-9 (1984).
70. Radchuk, V.V. *et al.* Spatiotemporal profiling of starch biosynthesis and degradation in the developing barley grain. *Plant Physiol* **150**, 190-204 (2009).

71. Polakova, K.M., Kucera, L., Laurie, D.A., Vaculova, K. & Ovesna, J. Coding region single nucleotide polymorphism in the barley low-pl, alpha-amylase gene Amy32b. *Theor Appl Genet* **110**, 1499-504 (2005).
72. Brown, A. & Jacobsen, J. Genetic basis and natural variation of  $\alpha$ -amylase isozymes in barley. *Genetical Research* **40**, 315-324 (1982).
73. Radchuk, V. & Borisjuk, L. Physical, metabolic and developmental functions of the seed coat. *Front Plant Sci* **5**, 510 (2014).
74. Melkus, G. *et al.* Dynamic  $(1)(3)C/(1)H$  NMR imaging uncovers sugar allocation in the living seed. *Plant Biotechnol J* **9**, 1022-37 (2011).
75. Oparka, K.J. & Gates, P. Transport of assimilates in the developing caryopsis of rice (*Oryza sativa* L.) : The pathways of water and assimilated carbon. *Planta* **152**, 388-96 (1981).
76. Eom, J.S. *et al.* SWEETs, transporters for intracellular and intercellular sugar translocation. *Curr Opin Plant Biol* **25**, 53-62 (2015).
77. Yuan, M. & Wang, S. Rice MtN3/saliva/SWEET family genes and their homologs in cellular organisms. *Molecular plant* **6**, 665-674 (2013).
78. Chen, L.Q. *et al.* Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. *Science* **335**, 207-11 (2012).
79. Radchuk, V. *et al.* Jekyll encodes a novel protein involved in the sexual reproduction of barley. *Plant Cell* **18**, 1652-66 (2006).
80. Yin, L.L. & Xue, H.W. The MADS29 transcription factor regulates the degradation of the nucellus and the nucellar projection during rice seed development. *Plant Cell* **24**, 1049-65 (2012).
81. Kuroyanagi, M. *et al.* Vacuolar processing enzyme is essential for mycotoxin-induced cell death in *Arabidopsis thaliana*. *Journal of Biological Chemistry* **280**, 32914-32920 (2005).
82. Hatsugai, N., Yamada, K., Goto-Yamada, S. & Hara-Nishimura, I. Vacuolar processing enzyme in plant programmed cell death. *Frontiers in Plant Science* **6**, 234 (2015).
83. Tran, V., Weier, D., Radchuk, R., Thiel, J. & Radchuk, V. Caspase-like activities accompany programmed cell death events in developing barley grains. *PLoS One* **9**, e109426 (2014).
84. Zakhrebekova, S. *et al.* Induced mutations in circadian clock regulator Mat-a facilitated short-season adaptation and range extension in cultivated barley. *Proc Natl Acad Sci U S A* **109**, 4326-31 (2012).
85. von Zitzewitz, J. *et al.* Molecular and structural characterization of barley vernalization genes. *Plant Mol Biol* **59**, 449-67 (2005).
86. Comadran, J. *et al.* Natural variation in a homolog of *Antirrhinum CENTRORADIALIS* contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet* **44**, 1388-92 (2012).
87. 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**, 1031-4 (2005).
88. Russell, J. *et al.* Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nature Genetics* **48**, 1024-30 (2016).
89. Skinner, J.S. *et al.* Structural, functional, and phylogenetic characterization of a large CBF gene family in barley. *Plant molecular biology* **59**, 533-551 (2005).
90. Yan, L. *et al.* The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science* **303**, 1640-4 (2004).
91. Faure, S., Higgins, J., Turner, A. & Laurie, D.A. The FLOWERING LOCUS T-like gene family in barley (*Hordeum vulgare*). *Genetics* **176**, 599-609 (2007).
92. Yan, L. *et al.* The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc Natl Acad Sci U S A* **103**, 19581-6 (2006).