A physical, genetic and functional sequence assembly of the barley genome.

This is the author's manuscript

Original Citation:
A physical, genetic and functional sequence assembly of the barley genome. / Mayer KF; Waugh R; Langridge P; Close TJ; Wise RP; Graner A; Matsumoto T; Sato K; Schulman A; Muehlbauer GJ; Stein N; Ariyadasa R; Schulte D; Poursarebani N; Zhou R; Steuernagel B; Mascher M; Scholz U; Shi B; Langridge P; Madishetty K; Svensson JT; Bhat P; Moscou M; Resnik J; Close TJ; Muehlbauer GJ; Hedley P; Liu H; Morris J; Waugh R; Frenkel Z; Korol A; Bergès H; Graner A; Stein N; Steuernagel B; Scholz U; Taudien S; Felder M; Groth M; Platzer M; Stein N; Steuernagel B; Scholz U;

Availability:
This version is available http://hdl.handle.net/2318/125624 since 2016-06-29T14:49:43Z

Published version:
DOI:10.1038/nature11543

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
Barley (Hordeum vulgare L.) is among the world’s earliest domesticated and most important crop plants. It is diploid with a large haploid genome of 5.1 gigabases (Gb). Here we present an integrated and ordered physical, genetic and functional sequence resource that describes the barley gene-space in a structured whole-genome context. We developed a physical map of 4.98 Gb, with more than 3.90 Gb anchored to a high-resolution genetic map. Projecting a deep whole-genome shotgun assembly, complementary DNA and deep RNA sequence data onto this framework supports 79,379 transcript clusters, including 26,159 ‘high-confidence’ genes with homology support from other plant genomes. Abundant alternative splicing, premature termination codons and novel transcriptionally active regions suggest that post-transcriptional processing forms an important regulatory layer. Survey sequences from diverse accessions reveal a landscape of extensive single-nucleotide variation. Our data provide a platform for both genome-assisted research and enabling contemporary crop improvement.

A sequence–enriched barley physical map
We constructed a genome-wide physical map of the barley cultivar (cv.) Morex by high-information-content fingerprinting and contiguous assembly of 571,000 bacterial artificial chromosome (BAC) clones (~14-fold haploid genome coverage) originating from six independent BAC libraries. After automated assembly and manual curation, the physical map comprised 9,265 BAC contigs with an estimated N50 contig size of 904 kilobases and a cumulative length of 4.98 Gb (Methods, Supplementary Note 2). It is represented by a minimum tiling path (MTP) of 67,000 BAC clones. Given a genome size of 5.1 Gb, more than 95% of the barley genome is represented in the physical map, comprising favourably to the 1,036 contigs that represent 80% of the 1 Gb wheat chromosome 3B.

We enhanced the physical map by integrating shotgun sequence information from 5,341 gene-containing and 937 randomly selected BAC clones (Methods, Supplementary Notes 2 and 3, and Supplementary Table 4). These provided 1,136 megabases (Mb) of genomic sequence integrated directly into the physical map (Supplementary Table 3). This framework facilitated the incorporation of whole-genome shotgun sequence data and integration of the physical and genetic maps. We generated whole-genome shotgun sequence data from genomic DNA of cv. ‘Morex’ by short-read Illumina GAIIx technology, using a combination of 300 base pairs (bp) paired-end and 2.5 kb mate-pair libraries, to >50-fold haploid genome coverage (Supplementary Note 3.3). De novo assembly resulted in sequence contigs totalling 1.9 Gb. Due to the high proportion of repetitive DNA, a substantial part of the whole-genome shotgun data collapsed into relatively small contigs characterized by exceptionally high read depths. Overall, 376,261 contigs were larger than 1 kb (N50 = 264,958 contigs, N50 length = 1,425 bp). Of these, 112,989
(308 Mb) could be anchored directly to the sequence-enriched physical map by sequence homology.

We implemented a hierarchical approach to further anchor the physical and genetic maps (Methods, Supplementary Note 4). A total of 3,241 genetically mapped gene-based single-nucleotide variants (SNV) and 498,165 sequence-tag genetic markers allowed us to use sequence homology to assign 4,556 sequence-enriched physical map contigs spanning 3.9 Gb to genetic positions along each barley chromosome. An additional 1,881 contigs were assigned to chromosomal bins by sequence homology to chromosome-arm-specific sequence data sets (Supplementary Note 4.4). Thus, 6,437 physical map contigs totalling 4.56 Gb (90% of the genome), were assigned to chromosome arm bins, the majority in linear order. Non-anchored contigs were typically short and lacked genetically informative sequences required for positional assignment.

Consistent with genome sequences of other grass species the peri-centromeric and centromeric regions of barley chromosomes exhibit significantly reduced recombination frequency, a feature that compromises exploitation of genetic diversity and negatively impacts genetic studies and plant breeding. Approximately 1.9 Gb or 48% of the genetically anchored physical map (3.9 Gb) was assigned to these regions (Fig. 1 and Supplementary Fig. 11).

Repetitive nature of the barley genome

A characteristic of the barley genome is the abundance of repetitive DNA. We observed that approximately 84% of the genome is comprised of mobile elements or other repeat structures (Supplementary Note 5). The majority (76%) in random BACs of these consists of retrotransposons, 99.6% of which are long terminal repeat (LTR) retrotransposons. The non-LTR retrotransposons contribute only 0.31% and the DNA transposons 6.3% of the random BAC sequence. In the fraction of the genome with a high proportion of repetitive elements, the LTR Gypsy retrotransposon superfamily was 1.5-fold more abundant than the Copia superfamily, in contrast to observations in both Brachypodium and rice. However, gene-bearing BACs were slightly depleted of retrotransposons, consistent with Brachypodium where young Copia retroelements are preferentially found in gene-rich, recombinogenic regions from which inactive Gypsy retroelements have been lost by LTR–LTR recombination. Overall, we see reduced repetitive DNA content within the terminal 10% of the physical map of each barley chromosome arm (Fig. 1). Class I and II elements show non-quantitative reverse-image distribution along barley chromosomes (Fig. 1), a feature shared with other grass genomes and shown by fluorescence in situ hybridization (FISH) mapping. Not surprisingly, the whole-genome shotgun assembly shows a lower abundance of LTR retrotransposons (average 53%) than gene-bearing BACs. That LTR retrotransposons are long highly repetitive and often nested supports our assumption that short reads either collapsed or did not assemble. Short interspersed elements (SINEs), short (80–600 bp) non-autonomous retrotransposons that are highly repeated in barley, showed no differential exclusion from the assemblies. However, miniature inverted-repeat transposable elements (MITEs), small non-autonomous DNA transposons, were twofold enriched in the whole-genome shotgun assemblies compared with BES reads or random BACs, consistent with the gene richness of the assemblies and their association with genes. Both MITEs and SINEs are 1.5 to 2-fold enriched in gene-bearing BACs which could indicate that SINEs are also preferentially integrated into gene-rich regions, or because they are older than LTR retroelements, may simply remain visible in and around genes where retro insertions have been selected against.

Transcribed portion of the barley genome

The transcribed complement of the barley gene space was annotated by mapping 1.67 billion RNA-seq reads (167 Gb) obtained from eight stages of barley development as well as 28,592 barley full-length cDNAs to the whole-genome shotgun assembly (Methods, Supplementary Notes 6, 7 and Supplementary Tables 20–22). Exon detection and consensus gene modelling revealed 79,379 transcript clusters, of which 75,258 (95%) were anchored to the whole-genome shotgun assembly (Supplementary Notes 7.1.1 and 7.1.2). Based on a gene-family-directed comparison with the genomes of Sorghum, rice, Brachypodium and Arabidopsis, 26,159 of these transcribed loci fall into clusters and have homology support to at least one reference genome (Supplementary Fig. 16); they were defined as high-confidence genes. Comparison against a data set of metabolic genes in Arabidopsis thaliana indicated a detection rate of 86%, allowing the barley gene-set to be estimated as approximately 30,400 genes. Due to lack of homology and missing support from gene family clustering, 53,220 transcript loci were considered low-confidence (Table 1). High-confidence and low-confidence barley genes exhibited distinct characteristics: 75% of the high-confidence genes had a multi-exon structure, compared with only 27% of low-confidence genes (Table 1). The mean size of high-confidence genes was 3,013 bp compared with 972 bp for low-confidence genes. A total of 14,481 low-confidence genes showed distant homology to plant proteins in public databases (Supplementary Notes 7.1.2, 7.1.4 and Supplementary Fig. 18), identifying them as potential gene fragments known to populate Triticeae genomes at high copy number and that often result from transposable element activity. A total of 15,719 high-confidence genes could be directly associated with the genetically anchored physical map (Supplementary Note 4). An additional 3,743 were integrated by invoking a conservation of synteny model (Supplementary Note 4.5) and a further 4,692 by association.
with chromosome arm whole-genome shotgun data (Supplementary Note 4.4 and Supplementary Table 15). Importantly, the N50 length of whole-genome shotgun sequence contigs containing high-confidence genes was 8,172 bp, which is generally sufficient to include the entire coding sequence, and 5’ and 3’ untranslated regions (UTRs). Overall, 24,154 high-confidence genes (92.3%) were associated and positioned in the physical/genetic scaffold, representing a gene density of five genes per Mb. Proximal and distal ends of chromosomes are more gene-rich, on average containing 13 genes per Mb (Fig. 1).

In comparison with sequenced model plant genomes, gene family analysis (Supplementary Note 7.1.3) revealed some gene families that exhibited barley-specific expansion. We defined the functions of members of these families using gene ontology (GO) and PFAM protein motifs (Supplementary Table 25). Gene families with highly overrepresented GO/PFAM terms included genes encoding (1,3)-β-glucan synthases, protease inhibitors, sugar-binding proteins and sugar transporters. NB-ARC (a nucleotide-binding adaptor shared by APAF-1, certain R gene products and CED-4[2]) domain proteins, known to be involved in defence responses, were also overrepresented, including 191 NBS-LRR type genes. These tended to cluster towards the distal regions of barley chromosomes (Supplementary Fig. 17), including a major group on barley chromosome 1HS, co-localizing with the MLA powdery mildew resistance gene cluster[3]. Biased allocation to recombination-rich regions provides the genomic environment for generating sequence diversity required to cope with dynamic pathogen populations[29,30]. It is noteworthy that the highly over-represented (1,3)-β-glucan synthase genes have also been implicated in plant–pathogen interactions[41].

### Regulation of gene expression

Deep RNA sequence data (RNA-seq) provided insights into the spatial and temporal regulation of gene expression (Supplementary Note 7.2). We found 72–84% of high-confidence genes to be expressed in all spatiotemporal RNA-seq samples (Fig. 2a), slightly lower than reported for rice[3] where ~95% of transcripts were found in more than one developmental or tissue sample. More importantly, 36–55% of high-confidence barley genes seemed to be differentially regulated between samples (Fig. 2b), highlighting the inherent dynamics of barley gene expression.

Two notable features support the importance of post-transcriptional processing as a central regulatory layer (Supplementary Notes 7.3 and 7.4). First, we observed evidence for extensive alternative splicing. Of the intron-containing high-confidence barley genes, 73% had evidence of alternative splicing (55% of the entire high-confidence set). The spatial and temporal distribution of alternative splicing transcripts deviated significantly from the general occurrence of transcripts in the different tissues analysed (Fig. 2c). Only 17% of alternative splicing transcripts were shared among all samples, and 17–27% of the alternative splicing transcripts were detected only in individual samples, indicating pronounced alternative splicing regulation. We found 2,466 premature termination codon-containing (PTC+) alternative splicing transcripts (9.4% of high-confidence genes) (Fig. 2d and Table 2), similar to the percentage of nonsense-mediated decay (NMD)-controlled genes in a wide range of species[33,34]. Premature termination codons activate the NMD pathway[35], which leads to rapid degradation of PTC+ transcripts, and have been associated with transcriptional regulation during disease and stress response in human and Arabidopsis, respectively[34,36–39]. The distribution of PTC+ transcripts was strikingly dissimilar, both spatially and temporally, with only 7.4% shared and between 31% and 40% exclusively observed in only a single sample (Fig. 2d). Genes encoding PTC+ containing transcripts show a broad spectrum of GO terms and PFAM domains and are more prevalent in expanded gene families. These observations support a central role for alternative splicing/NMD-dependent decay of PTC+ transcripts as a mechanism that controls the expression of many different barley genes.

Second, recent reports have highlighted the abundance of novel transcriptionally active regions in rice that lack homology to protein-coding genes or open reading frames (ORFs)[40]. In barley as many as 27,009 preferentially single-exon low-confidence genes can be classified as putative novel transcriptionally active regions (Supplementary Note 7.1.4). We investigated their potential significance by comparing the homology of barley novel transcriptionally active regions with the rice and Brachypodium genomes that respectively represent 50 and 30 million years of evolutionary divergence[41]. A total of 4,830 and 2,450 novel transcriptionally active regions yielded a homology match to the Brachypodium and rice genomes, respectively (intersection of 2,046; BLAST P value ≤ 10−5), indicating a putative functional role in pre-mRNA processing or other RNA regulatory processes[41,42].

### Natural diversity

Barley was domesticated approximately 10,000 years ago[1]. Extensive genotypic analysis of diverse germplasm has revealed that restricted outcrossing (0–1.8%)[43], combined with low recombination in pericentromeric regions, has resulted in modern germplasm that shows limited regional haplotype diversity[44]. We investigated the frequency and distribution of genome diversity by survey sequencing four diverse barley cultivars (’Bowman’, ’Barke’, ’Igri’ and ’Haruna Nijo’) and an H. spontaneum accession (Methods and Supplementary Note 8) to a depth of 5–25-fold coverage, and mapping sequence reads against the barley cultivar ‘Morex’ gene space. We identified more than 15 million non-redundant single-nucleotide variants (SNVs). H. spontaneum contributed almost twofold more SNV than each of the cultivars (Supplementary Table 28). Up to 6 million SNV per accession could be assigned to chromosome arms, including up to 350,000 associated with exons (Supplementary Table 29). Approximately 50% of the exon-located SNV were integrated into the genetic/physical framework (Fig 3, Supplementary Table 30 and Supplementary Fig. 31), providing a platform to establish true genome-wide marker technology for high-resolution genetics and genome-assisted breeding.

We observed a decrease in SNV frequency towards the centromeric and peri-centromeric regions of all barley chromosomes, a pattern that seemed more pronounced in the barley cultivars. This trend was supported by SNV identified in RNA-seq data from six additional cultivars mapped onto the Morex genomic assembly (Supplementary Note 8.2). We attribute this pattern of eroded genetic diversity to low recombination in the pericentromeric regions, which reduces effective population size and consequently haplotype diversity. Whereas
**Table 2** | Alternative splicing and transcripts containing PTCs in high-confidence genes

<table>
<thead>
<tr>
<th>General statistics of alternative splicing in high-confidence genes</th>
<th>24,243</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-confidence genes with RNA-seq data to monitor alternative splicing</td>
<td>62,426</td>
</tr>
<tr>
<td>Transcripts at high-confidence genes</td>
<td>170</td>
</tr>
<tr>
<td>Transcripts with complete CDS structures*</td>
<td>13,299</td>
</tr>
<tr>
<td>Transcripts with partial CDS structures†</td>
<td>51,482</td>
</tr>
<tr>
<td>Genes with alternative transcripts</td>
<td>3,068</td>
</tr>
<tr>
<td>Predicted transcripts deriving from genes with alternative splicing</td>
<td>1,147</td>
</tr>
<tr>
<td>Premature stop codon analysis</td>
<td>41,461 (81%)</td>
</tr>
<tr>
<td>Predicted transcripts used for PTC analysis‡</td>
<td>9,877</td>
</tr>
<tr>
<td>Transcripts containing PTC</td>
<td>5,286 (10%)</td>
</tr>
<tr>
<td>PTC caused by intron retention</td>
<td>4,591 (9%)</td>
</tr>
<tr>
<td>Gene loci incorporating PTC + NMD transcripts</td>
<td>2,466</td>
</tr>
</tbody>
</table>

* Entire predicted coding sequence (100%) was transferred to transcript model on barley cultivar Morex contigs.
† Predicted coding sequence could not be completely projected to genomic transcript model (partial mapping of fli-cDNA).
‡ Only transcripts with structures for entire coding sequence on barley cultivar Morex WGS assembly were considered.

**Discussion**

The size of Triticeae cereal genomes, due to their highly repetitive DNA composition, has severely compromised the assembly of whole-genome shotgun sequences and formed a barrier to the generation of high-quality reference genomes. We circumvented these problems by integrating complementary and heterogeneous sequence-based genomic and genetic data sets. This involved coupling a deep physical map with high density genetic maps, superimposing deep short-read whole-genome shotgun assemblies, and annotating the resulting linear, albeit punctuated, genomic sequence with deep-coverage RNA-derived data (full-length cDNA and RNA-seq). This allowed us to systematically delineate approximately 4 Gb (80%) of the genome, including more than 90% of the expressed genes. The resulting genomic framework

Figure 2 | Atlas of barley gene expression. a, Barley gene expression in different spatial and temporal RNA-seq samples (Supplementary Notes 6, 7). Numbers refer to high-confidence genes. b, Dendrogram depicting relatedness of samples and colour-coded matrix showing number of significantly upregulated high-confidence genes in pairwise comparisons. Σ, total number of non-redundant high-confidence genes upregulated in comparison to all other samples. Height, complete linkage cluster distance (log\(_2\) (fragments per kilobase of exon per million fragments mapped)); see Supplementary Note 7.2.3.1. c, Distribution and overlap of alternatively spliced barley transcripts between RNA-seq samples. d, Distribution and overlap of alternative splicing transcripts fulfilling criteria for PTC+ as detected in different spatial and temporal RNA-seq samples (Supplementary Note 7.4).
Figure 3 | Single nucleotide variation (SNV) frequency in barley. Barley chromosomes indicated as inner circle of grey bars. Connector lines give the genetic/physical relationship in the barley genome. SNV frequency distribution displayed as five coloured circular histograms (scale, relative abundance of SNVs within accession; abundance, total number of SNVs in non-overlapping 50-kb intervals of concatenated ‘Morex’ genomic scaffold; random, zero to maximum number of SNVs per 50-kb interval). Selected patterns of SNV frequency indicated by coloured arrowheads (for further details see Supplementary Note 8). Colouring of arrowheads refers to cultivar with deviating SNV frequency for the respective region.

provides a detailed insight into the physical distribution of genes and repetitive DNA and how these features relate to genetic characteristics such as recombination frequency, gene expression and patterns of genetic variation.

The centromeric and peri-centromeric regions of barley chromosomes contain a large number of functional genes that are locked into recombinationally ‘inert’ genomic regions. The gene-space distribution highlights that these regions expand to almost 50% of the physical length of individual chromosomes. Given well-established levels of conserved synteny, this will probably be a general feature of related grass genomes that will have important practical implications. For example, infrequent recombination could function to maintain evolutionarily selected and co-adapted gene complexes. It will certainly restrict the release of the genetic diversity required to decouple advantageous from deleterious alleles, a potential key to improving genetic gain. Understanding these effects will have important consequences for crop improvement. Moreover, for gene discovery, forward genetic strategies based on recombination will not be effective in these regions.

Whereas alternative approaches exist for some targets (for example, by coupling resequencing technologies with collections of natural or induced mutant alleles), for most traits it remains a serious impediment. Some promise may lie in manipulating patterns of recombination by either genetic or environmental intervention. Quite strikingly, our data also reveal that a complex layer of post-transcriptional regulation will need to be considered when attempting to link barley genes to functions. Connections between post-transcriptional regulation such as alternative splicing and functional biological consequences remain limited to a few specific examples, but the scale of our observations suggests this list will expand considerably.

In conclusion, the barley gene space reported here provides an essential reference for genetic research and breeding. It represents a hub for trait isolation, understanding and exploiting natural genetic diversity and investigating the unique biology and evolution of one of the world’s first domesticated crops.

**METHODS SUMMARY**

Methods are available in the online version of the paper.

Full Methods and any associated references are available in the online version of the paper.

Received 2 May: accepted 30 August 2012.

Published online 17 October 2012.


©2012 Macmillan Publishers Limited. All rights reserved
Principal investigators Klaus F. X. Mayer 1, Robbie Waugh 2, Peter Langridge 3, Timothy J. Close 4, Roger P. Wise 5, Andreas Graner 6, Takashi Matsumoto 6, Kazuhiro Sato 6, Alan Schulman 7, Gary J. Muehlbauer 8, Nils Stein 9

Physical map construction and direct anchoring Ruvini Ariyadasa 10, Daniela Schulte 10, Naser Pourasrari 10, Runouz Zn 11, Burkhard Steuernagel 11, Martin Mascher 12, Uwe Scholz 12, Bujin Shi 13, Peter Langridge 3, Kawitha Madishetty 14, Jan T. Svensson 14, Prasanna Bhat 15, Mathew Moscou 15, Josh Resnik 16, Gary J. Muehlbauer 16, Peter Hedley 17, Hui Liu 17, Jenny Morris 17, Robbie Waugh 17, Zeew Frenkel 17, Avraham Korol 18, Hélène Bergès 18, Andreas Graner 6, Nils Stein (leader) 9

Genomic sequencing and assembly Burkhard Steuernagel 11, Uwe Scholz 11, Stefan Taudien 11, Marcus Felder 11, Marco Groth 11, Matthias Platzer 11, Nils Stein (leader) 9

BAC accession and assembly Burkhard Steuernagel 11, Uwe Scholz 11, Axel Himmelbach 19, Stefan Taudien 11, Marcus Felder 11, Matthias Platzer 11, Stefano Lonardi 12, Denis Duma 14, Matthew Alpert 14, Francesca Cordoro 14,22, Marco Beccuti 14, Gianfranco Ciardi 14, Yaqin Ma 14, Steve Wanamaker 17, Timothy J. Close (co-leader) 8, Nils Stein (leader) 9

BAC-end sequencing Federica Çatonnar 14, Vera Vendramini 17, Simone Scalabrin 15, Slobodanka Radovick 14, Rod Wing 11, Daniela Schulte 11, Burkhard Steuernagel 11, Michele Morganet 15,16, Nils Stein, Robbie Waugh (leader) 2

Integration of physical/genetic map and sequence resources Thomas Nussbaumer 16, Heidrun Gundlach 8, Miaelia Martin 16, Ruvini Ariyadasa 10, Naser Pourasrari 10, Burkhard Steuernagel 11, Uwe Scholz 11, Roger P. Wise 5, Jesse Poland 16, Nils Stein 9, Klaus F. X. Mayer (leader) 1

Gene annotation Manuel Spannagl 16, Matthias Pfeifer 17, Heidrun Gundlach 8, Klaus F. X. Mayer (leader) 1

Repetitive DNA analysis Heidrun Gundlach 8, Cédric Moisy 23, Jaakko Tanskanen 13, Simone Scalabrin 15, Andrea Zuccolo 15, Vera Vendramini 17, Michele Morganet 15,16, Klaus F. X. Mayer (co-leader) 8, Alan Schulman (leader) 9

Transcriptome sequencing and analysis Matthias Pfeifer 16, Manuel Spannagl 16, Peter Hedley 17, Jenny Morris 17, Joanne Russell 17, Amis Druka 16, David Marshall 17, Micha Bayer 15, Geoffrey B. Fincher 16, Mario Caccamo 19, Takashi Matsumoto 6, Tsuyoshi Tanaka 6, Kazuhiro Sato 6, Roger P. Wise 5, Timothy J. Close 8, Steve Wanamaker 17, Gary J. Muehlbauer 6, Nils Stein 9, Klaus F. X. Mayer (co-leader) 8, Robbie Waugh (leader) 2

Re-sequencing and diversity analysis Burkhard Steuernagel 11, Thomas Schmutzer 11, Martin Mascher 11, Uwe Scholz 11, Stefan Taudien 11, Matthias Platzer 11, Kazuhiro Sato 6, David Marshall 17, Micha Bayer 17, Robbie Waugh (co-leader) 8, Nils Stein (leader) 9

Writing and editing of the manuscript Klaus F. X. Mayer (co-leader) 8, Robbie Waugh (co-leader) 8, John W. S. Brown 20, Alan Schulman 9, Peter Langridge 3, Matthias Platzer 11, Geoffrey B. Fincher 16, Mario Caccamo 19, Takashi Matsumoto 6, Tsuyoshi Tanaka 6, Kazuhiro Sato 6, Roger P. Wise 5, Timothy J. Close 8, Roger P. Wise 5, Nils Stein (leader) 9

The International Barley Genome Sequencing Consortium (IBSC)

The National Academy of Sciences, 2-1-2, Kannondai, Tsukuba Ibaraki 305-8602, Japan. 8Okayama University, Graduate School of Agriculture, Science and Technology, 010-0013 Okayama, Japan. 9Leibniz Institute of Plant Genetics and Crop Plant Research, 07745 Jena, Germany. 10Argonne National Laboratory, 9800 S. Cass Ave., Argonne, IL 60439, USA. 11Department of Computer Science & Engineering, University of California, Riverside, 12University of California, Riverside, CA 92521, USA. 12USDA-ARS, Department of Plant Pathology & Microbiology, Iowa State University, Ames, Iowa 50011-1020, USA. 13Leibniz Institute of Plant Genetics and Crop Research (IPK), D-06466 Seeland OT Gatersleben, Germany. 14National Institute of Agrobiological Sciences, 2-1-2, Kannondai, Tsukuba Ibaraki 305-8602, Japan. 15Laboratory of Bioinformatics, 4200 University Town, Okayama 700-8530, Japan. 16Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. 17Department of Plant Sciences, University of Dundee at The James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK. 18ARC Centre of Excellence in Plant Cell Walls, University of Adelaide, Waite Campus, Glen Osmond, South Australia 5064, Australia. 19Department of Computer Science, Corso Svizzera 185, 10149 Torino, Italy.
Whole-genome shotgun sequence assembly. Of 600–1,000 bp) was sequenced using Roche 454 GSFLX Titanium chemistry. Illumina GAIIx and Hiseq 2000. Genomic DNA of cultivar 'Haruna Nijo' (size range 454 GS FLX or FLX Titanium51,52. Sequences were assembled using MIRA v3.2.0 to reveal gene sequence information for physical map anchoring. An additional 2,183 gene-bearing BACs (Supplementary Table 4.1). 3,072 SNP markers on Illumina GoldenGate assays45 leading to 2,032 PCR-based markers from published genetic maps59,60 were PCR-screened (presence/absence) tags for the two doubled haploid populations Oregon Wolfe (in the press).

Gene annotation, functional categorization and differential expression. Gene family clustering was undertaken using OrthoMCL (Supplementary Note 7.2) and the latter providing also information of alternatively spliced transcripts. Structural gene calls were combined and the longest ORF for each locus was used as representative for gene family analysis (Supplementary Note 7.1.2).

Gene family clustering was undertaken using OrthoMCL (Supplementary Note 7.1.3) by comparing against the genomes of Oryza sativa (RAP2), Sorghum bicolor, Brachypodium distachyon (v 1.4) and Arabidopsis thaliana (TAIR10 release). Analysis of differential gene expression (Supplementary Note 7.2) was performed on RNA-seq data using CuffDiff49. Analysis of sequence diversity. Genome-wide SNV was assessed by mapping (BWA v0.5.9-r164) the original sequence reads of sequenced genotypes to de novo assembly of cultivar 'Morex'. Sequence reads from RNA-seq were mapped against the 'Morex' assembly. Details are provided in Supplementary Note 8.

Genetic framework for anchoring. The genetic framework for anchoring was based on a single-nucleotide variation (SNV) map53 (Supplementary Note 4.3) which provided the highest marker density (3,973) and resolution (N = 360, RIL/F8) for a single bi-parental mapping population in barley. Additional high-density genetic marker maps (Supplementary Note 4.3) were compared and aligned on the basis of shared markers. Furthermore, we used genotyping-by-sequencing (GBS)54 to generate high-density genetic maps comprising 34,396 SNVs and 21,384 SNVs as well as 241,159 and 184,796 dominant (presence/absence) tags for the two doubled haploid populations Oregon Wolfe barley49 and Morex × Barke45, respectively. Altogether 498,165 marker sequence tags were used (Supplementary Table 1).

Transcriptome sequencing. Eight tissues of cultivar 'Morex' (three biological replicates each) eukaryotic stages of the barley life cycle from germinating grain to maturing caryopsis were selected for deep RNA sequencing (RNA-seq). Plant growth, sampling and sequencing is detailed in Supplementary Information (Supplementary Note 6). Further mRNA sequencing data was generated from eight additional spring barley cultivars within a separate study and was used here for sequence diversity analysis (Supplementary Note 8.2).

GENETIC FRAMEWORK OF THE PHYSICAL MAP. The genetic framework for anchoring the physical map of barley was built on a single-nucleotide variation (SNV) map53 (Supplementary Note 4.3) which provided the highest marker density (3,973) and resolution (N = 360, RIL/F8) for a single bi-parental mapping population in barley. Additional high-density genetic marker maps (Supplementary Note 4.3) were compared and aligned on the basis of shared markers. Furthermore, we used genotyping-by-sequencing (GBS)54 to generate high-density genetic maps comprising 34,396 SNVs and 21,384 SNVs as well as 241,159 and 184,796 dominant (presence/absence) tags for the two doubled haploid populations Oregon Wolfe barley49 and Morex × Barke45, respectively. Altogether 498,165 marker sequence tags were used (Supplementary Table 1).