Barley remains dated to the dawn of agriculture, having been found at several archaeological sites. In addition to indications that barley was an important food crop, recent excavations have fuelled speculation that beverages from fermented grains may have motivated early Neolithic hunter–gatherers to erect some of humankind’s oldest monuments. Moreover, brewing beer may also have played a role in the eastward spread of the crop after its initial domestication in the Fertile Crescent.

Since 2012, both genetic research and crop improvement in barley have benefited from a partly ordered draft sequence assembly. This community resource has underpinned gene isolation and population genomic studies. However, these and other efforts have also revealed limitations of the current draft assembly. The limitations are often direct consequences of two characteristic genomic features: the extreme abundance of repetitive elements, and the severely reduced frequency of meiotic recombination in pericentromeric regions.

These factors have limited the contiguity of whole-genome assemblies to kilobase-sized sequences originating from low-copy regions of the genome. Thus, a detailed investigation of the composition of the repetitive fraction of the genome—including expanded gene families—and of the distribution of targets of selection and crop improvement in (genetically defined) pericentromeric regions has been beyond reach.

Here we present a map-based reference sequence of the barley genome including the first comprehensively ordered assembly of the pericentromeric regions of a Triticaceae genome. The resource highlights a conspicuous distinction between distal and proximal regions of chromosomes that is reflected by the intranuclear chromatin organization. Moreover, chromosomal compartments are differentiated by an exponential gradient of gene density and recombination rate, striking contrasts in the distribution of retrotransposon families, and distinct patterns of genetic diversity.
A chromosome-scale assembly of the barley genome

We adopted a hierarchical approach to generate a high-quality reference genome sequence of the barley cultivar Morex, a US spring six-row malted barley. First, a total of 87,075 bacterial artificial chromosomes (BACs) were sequenced, mainly using Illumina paired-end and mate-pair technology and assembled individually from 4.5 terabases of raw sequence data12–14 (Supplementary Note 1). In a second step, overlaps between adjacent clones15 were detected and validated by physical map information16, a genetic linkage17 and a highly contiguous optical map18 to construct super-scaffolds composed of merged assemblies of individual BACs (Table 1 and Extended Data Table 1). This increased the contiguity as measured by the N50 value (the scaffold size above which 50% of the total length of the sequence was included in the assembly) from 79 kb to 1.9 Mb. Scaffolds were assigned to chromosomes using a population sequencing (POPSEQ) genetic map19. Finally, we used three-dimensional proximity information obtained by chromosome conformation capture sequencing19–21 (Hi-C) to order and orient BAC-based super-scaffolds (Supplementary Note 2 and ref. 22). The final chromosome-scale assembly of the barley genome consists of 6,347 ordered super-scaffolds composed of merged assemblies of individual BACs, representing 4.79 Gb (~95%) of the genomic sequence content, of which 4.54 Gb have been assigned to precise chromosomal location in the Hi-C map (Table 1).

Mapping of transcriptome data and reference protein sequences from other plant species to the assembly identified 83,105 putative gene loci including protein-coding genes, non-coding RNAs, pseudogenes and transcribed transcripts (Fig. 1, Extended Data Fig. 1, Extended Data Table 2 and Supplementary Note 3). These loci were filtered further and divided into 39,734 high-confidence genes (with four different sub-categories) and 41,949 low-confidence genes on the basis of sequence homology to related species (Methods and Supplementary Note 3.4). Moreover, we predicted 19,908 long non-coding RNAs (Supplementary Note 3.7) and 792 microRNA precursor loci (Supplementary Note 3.8). The high co-linearity between the Hi-C-based pseudomolecules and linkage and cytogenetic maps22 as well as the conserved order of syntenic genes in pericentromeric regions compared with model grass Brachypodium distachyon (Extended Data Fig. 2a) corroborated the quality of the assembly. Extrapolating from a set of conserved euchromatic core genes23, we estimate that the predicted gene models represent 98% of the cultivar Morex barley gene complement (Extended Data Fig. 2b).

Organization of chromatin

Barley has served as a model for traditional cytogenetics11; but relating chromosomal features to unique sequences has been challenging, requiring the cloning of repeat-free probes24. The reference sequence allowed us to employ the Hi-C data to interrogate the three-dimensional organization of chromatin in the nucleus. As in other eukaryotes25,26, the spatial proximity of genomic loci as measured by Hi-C link frequency is highly dependent on their distance in the linear genome (Fig. 2a). However, we observed an elevated link frequency at distances above 200 Mb and a pronounced anti-diagonal pattern in the intrachromosomal Hi-C contact matrices (Fig. 2b and Extended Data Fig. 3a), indicating an increased adjacency of regions on different chromosome arms. We interpret this pattern as reflective of the so-called Rabl configuration27 of interphase nuclei, where individual chromosomes fold back to juxtapose the long and short arms, with centromeres and telomeres of all chromosomes clustering at opposite poles of the nucleus (Fig. 2c and Supplementary Fig. 2.2). Fluorescence
Centromeres and telomeres are presented by red and green circles, links between 1 Mb windows on chromosome 1H on a logarithmic scale.

Figure 2 marks associated with heterochromatin in large, repeat-rich genomes. Resolution, but is consistent with cytogenetic mapping of histone insertion site preferences. On a genome (Fig. 1b).

Contacts between loci on different chromosomes followed a similar pattern (Fig. 2d) supported this hypothesis. Principal component analysis of the intrachromosomal matrix showed that the first three principal components cumulatively explained ~70% of the variation and differentiated (1) distal from proximal regions, (2) interstitial from both distal and proximal regions and (3) the long arms from the short arms (Fig. 2f and Extended Data Fig. 4a). A linear model taking into account the genomic distance between two loci, as well as their relative distance from the centromere, accounted for 79% of the variation (Extended Data Fig. 4b) in the intrachromosomal proximity matrix at 1 Mb resolution.

Contacts between loci on different chromosomes followed a similar pattern (Fig. 2e and Extended Data Fig. 3b): a prominent cross pattern supporting a juxtaposition of long and short arms. In contrast to intrachromosomal matrices, contact probabilities between loci on, for instance, the short arm of one chromosome are equal for loci on both the short and the long arm on another chromosome having the same relative distance to the centromere: that is, facing each other in the interphase nucleus. We also observed a higher contact frequency between telomere–near regions, as has been observed in Arabidopsis.

To test whether pairs of homologous chromosomes are positioned closer to each other than to non-homologues, we performed diploid Hi-C on leaf tissue from F1 hybrids between the cultivars Morex and Barke, and assigned the resultant Hi-C links between 1 Mb windows on chromosomes 1H (x axis) and 2H (y axis) on a logarithmic scale. A principal component analysis of the normalized contact matrix at 1 Mb resolution of chromosome 1H was conducted.

The genomic context of repetitive elements

Large plant genomes consist mainly of highly similar copies of repetitive elements such as long terminal repeat (LTR) retrotransposons and DNA transposons. Our hierarchical sequencing strategy reduced the algorithmic complexity of assembling a highly repetitive genome from short reads. Instead of resolving complex repeat structures on the whole-genome level, we reconstructed the sequences of 100–150 kb BACs. This allowed us to disentangle nearly identical copies of highly abundant repetitive elements, as evidenced by the good representation of both mathematically defined repeats and retrotransposon families (Extended Data Fig. 2c, d). Homology-guided repeat annotation with a Triticeae-specific repeat library identified 3.7 Gb (80.8%) of the assembled sequence as derived from transposable elements (Table 1, Fig. 1a and Extended Data Table 3), most of which were present as truncated and degenerated copies, with only 10% of mobile elements intact and potentially active.

Median 20-mer frequencies were used to partition the seven barley chromosomes into three zones (Fig. 1 and Extended Data Fig. 5a), reminiscent of the three compartments of wheat chromosome 3B. The distal zone 1 was characterized by an enrichment of low-copy regions, a high gene content and frequent meiotic recombination. Zone 2, occupying the interstitial regions of chromosomes, had the highest 20-mer frequencies and intermediate gene density. Surprisingly, the abundance of repetitive 20-mers decreased in the proximal zone 3, where older mobile elements with diverged, and thus unique, sequences predominated (Fig. 1). The three zones also differed in the composition of the gene space (Extended Data Table 2b and Supplementary Note 3). For example, genes involved in defence response and reproductive processes were preferentially found in distal regions, while proximal regions contained more genes related to housekeeping processes, such as photosynthesis and respiration, compared with other parts of the genome (Fig. 1b).

Transposable element groups exhibited pronounced variation in their insertion site preferences (Fig. 3a and Extended Data Fig. 5b). On a global scale, most miniature inverted-repeat transposable elements

**Figure 2** Chromosome conformation capture analysis. a. Distance-dependent decay of contact probability. b. Intrachromosomal contact matrix. The intensity of pixels represents the normalized count of Hi-C links between 1 Mb windows on chromosome 1H on a logarithmic scale. c. Schematic model of the Rabl configuration of interphase chromosomes. Centromeres and telomeres are presented by red and green circles, respectively. d. Leaf interphase nucleus of barley. Chromatin was stained blue with 4′,6-diamidino-2-phenylindole (DAPI). Fluorescence in situ hybridization was performed with probes specific for centromeres (red) and telomeres (green). Scale bar, 5 μm. e. Interchromosomal contact matrix. The intensity of pixels represents the normalized count of Hi-C links between 1 Mb windows on chromosomes 1H (x axis) and 2H (y axis) on a logarithmic scale. A principal component analysis of the normalized contact matrix at 1 Mb resolution of chromosome 1H was conducted. f. The first and second eigenvectors are plotted against each other. Each point represents a 1 Mb window. Closer proximity to the centromere is indicated by a darker colour. Windows from the short and long arms are coloured blue and red, respectively.
Gene families expanded in barley were tested for overrepresentation of Gene Ontology terms compared with sorghum, rice, *Brachypodium* and *Arabidopsis*. Among the most significant results were terms related to defence response and disease resistance (NBS-LRR and thionin genes), as well as thioredoxin genes (Supplementary Note 4.1).

In the following, we focused on a detailed analysis of gene families having particular importance for malting quality. Germinating barley grains possess high diastatic power: that is, the combined ability of a complex of enzymes to mobilize fermentable sugars from starch. Key diastatic enzymes include α-amylases. The genome of barley cultivar Morex contains 12 α-amylase (amy) family sequences (Supplementary Note 4.2 and Extended Data Table 4a), which can be classified into four subfamilies. Gene duplication events have occurred in the subfamilies amy1 and amy2 (Fig. 4b), located on chromosomes 6H and 7H, respectively. The existence of these duplications had been speculated earlier, but could not be analysed further because of high sequence similarity between the copies. The reference assembly contained five full-length amy1 subfamily genes, four of which, here designated as amy1_1-a–d, shared >99.8% identity at the nucleotide level including introns. Locus-specific PCR confirmed earlier suggestions of multiple, highly similar amy1_1 genes (Extended Data Fig. 6 and Supplementary Note 4.2). Given the relevance of α-amylase activity to the brewing process, the high variability of the amy1_1 multiple gene locus (Extended Data Fig. 6) observed in landraces and elite lines, including modern malting cultivars, is remarkable.

The accumulation of fermentable carbohydrates in the grain depends on the transfer of sugars from maternal tissue into the developing seeds. In contrast to the two routes of nutrient transfer in rice seeds—the nucellar projection and nucellar epidermis—delivery of assimilates into barley grains occurs predominantly via the nucellar projection and requires active transporters. The family of SUGARS WILL EVENTUALLY BE EXPORTED TRANSPORTER (SWEET) transmembrane proteins mediating sugar efflux consists of 23 members in barley (Extended Data Table 4b and Supplementary Note 4.3). There is a small extension of the sugar-transporting SWEET11, SWEET13, SWEET14 and SWEET15 subfamilies, with two or more genes for each subgroup compared with only a single orthologue in rice and *Arabidopsis* (Extended Data Table 4b). Duplication of SWEET11 was most likely followed by neofunctionalization as evidenced by divergent expression patterns. Both SWEET11a and SWEET11b were highly expressed in maternal seed tissue, but differed in the distribution of expression domains (Fig. 4c and Extended Data Fig. 7). Genes encoding a family of vacuolar processing enzymes, which are essential for programmed cell death in maternal tissue and starch accumulation in the grain (Supplementary Note 4.3 and V.R., unpublished observations) showed a similar expansion in barley (Extended Data Table 4c), pointing to the central role of the nucellar projection for grain filling in the Triticaceae.

These examples of genes involved in sugar transport and metabolism illustrate that the high-quality reference genome sequence can serve as a springboard for the in-depth analysis of the evolutionary history of gene duplications, their relation to morphological and physiological innovations, and their impact on crop performance.

**Molecular diversity and haplotype analysis**

To explore how the new barley genome assembly could be exploited for genetics and breeding, we generated exome sequence data from 96 European elite barley lines, half with a spring growth habit, half with a winter one (Supplementary Table 5.1). We investigated the extent and partitioning of molecular variation within and between these groups using 71,285 single-nucleotide polymorphisms (SNPs). Plotting diversity values in 100 SNP windows both in linear order (Fig. 5a) and according to physical distance (Fig. 5b) revealed marked contrasts in the levels and distribution of diversity both within and between gene pools. In spring types, extensive regions on...
a potential tandem gene duplication resulted in two amy3 gene copies. Moderate copy number extension was observed in sorghum and rice where most species further contain only a single amy3 gene copy per genome, genomes. These genes are located on distinct chromosomes and hence of gene copy numbers and duplication events across species. Note that for barley, hexaploid wheat, clusters for a species.

Total number of proteins that were included into the OrthoMCL analysis. Contrasting local patterns of diversity. For example, intense selection during breeding for different end-use sectors that the lack of observed variation in elite germplasm is a signature maintain higher diversity across these regions owing to the presence extensive pericentromeric regions. Chromosomes 3H, 4H and 6H was a large region on 5H in the winter gene pool. For these chromosomes 1H, 2H and 7H were virtually devoid of diversity, as was a large region on 5H in the winter gene pool. For these chromosomes, this results in a single gene-pool-specific haplotype across the extensive pericentromeric regions. Chromosomes 3H, 4H and 6H maintain higher diversity across these regions owing to the presence of multiple similarly extensive haplotypes. This is even more evident when diversity is plotted on a physical scale (Fig. 5b). We presume that the lack of observed variation in elite germplasm is a signature of intense selection during breeding for different end-use sectors (principally malting versus feed barley), and the virtual absence of allelic re-assortment during meiosis owing to restricted recombination in the pericentromeric regions.

Croses between spring and winter barleys are rarely performed as they are considered to disrupt the gene-pool-specific gene complexes required for general performance (such as phenological adaptations) and end-use quality. Contrasting local patterns of diversity outside the pericentromeric regions therefore also most likely reflect the outcome of selection within alternative gene pools. We explored this further by comparing diversity in eight characterized genes whose variant alleles are important for conditioning barley’s seasonal growth habit (Supplementary Note 5). Of the eight genes, HvCEN is uniquely ‘locked’ in the pericentromeric region of chromosome 2H where alternative alleles at a single SNP confer both differences in days-to-heading45 and strong latitudinal differentiation46. The extensive pericentromeric haplotype in spring barleys (Fig. 5) may stem from selection for this single HvCEN SNP. While strong selection for other favourable alleles locked in the same region in spring barley cannot be ruled out, the virtual absence of recombination severely restricts exploitation of diversity across the entire region. Despite our focus here on life-history traits, strong selection for other traits mapping to pericentromeric regions45,46, including good malting quality in the spring gene pool on chromosomes 1H and 7H, would probably also reduce diversity in these regions. Interestingly, we are unaware of any phenotypic trait in the winter gene pool that would

Figure 4 | Expansion of agronomically important gene families. a, OrthoMCL clustering of the barley high-confidence gene complement with B. distachyon, rice, sorghum and Arabidopsis thaliana genes. Numbers in the sections of the Venn diagram correspond to numbers of clusters (gene groups). The first number below the species name denotes the total number of proteins that were included into the OrthoMCL analysis for each species. The second number indicates the number of genes in clusters for a species. b, Phylogenetic tree of 68 full-length α-amylase protein sequences derived from amy genes identified in the genomes of barley, hexaploid wheat, B. distachyon, rice, sorghum and maize. Each wheat subgenum was considered separately to facilitate the comparison of gene copy numbers and duplication events across species. Note that for the amy4 subfamily, two to three genes per genome were identified in all genomes. These genes are located on distinct chromosomes and hence most probably did not originate from tandem gene duplications. While most species further contain only a single amy3 gene copy per genome, moderate copy number extension was observed in sorghum and rice where a potential tandem gene duplication resulted in two amy3 gene copies.
result in strong selection for a single pericentromeric haplotype on chromosome 5H.

We next explored patterns of linkage disequilibrium across the entire genome. As expected for two highly inbred and elite crop gene pools, we observed extensive linkage disequilibrium on all chromosomes in both spring and winter barleys (Extended Data Fig. 8). The number of discrete haplotype blocks in this germplasm set varied from 86 to 161 per chromosome (Extended Data Fig. 8). Surprisingly, the two-row spring gene pool, generally considered to be narrowest owing to intense positional cloning, provided a better contextualization of population community resource for cereal genetics and genomics. It will facilitate approaches.

will be one of the last such efforts to follow the laborious BAC-by-BAC map-based assemblies48,49, we believe that the barley genome project that the quality of whole-genome shotgun assemblies is on a par with technologies such as optical mapping18 and chromosome-scale scaffolding with Hi-C21. The latter technology was key to resolving the linear order of sequence scaffolds in pericentromeric regions. We anticipate the adoption of Hi-C-based genome mapping in other Triticeae species, such as bread and durum wheat and their wild relatives. Now that the quality of whole-genome shotgun assemblies is on a par with map-based assemblies48,49, we believe that the barley genome project will be one of the last such efforts to follow the laborious BAC-by-BAC approach.

The barley reference genome sequence constitutes an important community resource for cereal genetics and genomics. It will facilitate positional cloning, provide a better contextualization of population genomic datasets and enable comparative genomic analysis with other Triticeae in non-recombining regions that have been inaccessible to analysis of gene collinearity until now. The exciting methodological advances in sequence assembly and genome mapping have enabled even large and repeat-rich genomes to be unlocked48,50 and hold the promise of constructing reference-quality genome sequences, not only for a single cultivar, but also for representatives of major germplasm groups.

Discussion

To assemble a highly contiguous reference genome sequence for barley, we combined hierarchical shotgun sequencing, a strategy previously used for assembling large and complex plant genomes51,52, with novel technologies such as optical mapping48 and chromosome-scale scaffolding with Hi-C21. The latter technology was key to resolving the linear order of sequence scaffolds in pericentromeric regions. We anticipate the adoption of Hi-C-based genome mapping in other Triticeae species, such as bread and durum wheat and their wild relatives. Now that the quality of whole-genome shotgun assemblies is on a par with map-based assemblies48,49, we believe that the barley genome project will be one of the last such efforts to follow the laborious BAC-by-BAC approach.

The barley reference genome sequence constitutes an important community resource for cereal genetics and genomics. It will facilitate positional cloning, provide a better contextualization of population genomic datasets and enable comparative genomic analysis with other Triticeae in non-recombining regions that have been inaccessible to analysis of gene collinearity until now. The exciting methodological advances in sequence assembly and genome mapping have enabled even large and repeat-rich genomes to be unlocked48,50 and hold the promise of constructing reference-quality genome sequences, not only for a single cultivar, but also for representatives of major germplasm groups.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 August 2016; accepted 3 March 2017.

1. van Zeist, W. & Bakker-Heeres, J. A. H. Archaeological studies in the Levant 1. Neolithic sites in the Damascus basin: Aswad, Gharafel, Ramad. Palaeohistoria 24, 165–256 (1985).
2. Riehl, S., Zeidi, M. & Conard, N. J. Emergence of agriculture in the foothills of the Zagros Mountains of Iran. Science 341, 65–67 (2013).
3. Dietrich, O., Heun, M., Notroff, J., Schmidt, K. & Zarnkow, M. The role of cult and feasting in the emergence of Neolithic communities. New evidence from Göbekli Tepe, south-eastern Turkey. Antiquity 86, 674–695 (2012).
4. Hayden, B., Canuel, N. & Shanse, J. What was brewing in the Natufian? An archaeological assessment of brewing technology in the Epipaleolithic. J. Archaeol. Method Theory 20, 102–150 (2013).
5. Wang, J. et al. Revealing a 5,000-year-old beer recipe in China. Proc. Natl Acad. Sci. USA 113, 6444–6448 (2016).
6. Zohary, D., Hopf, M. & Weiss, E. Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin (Oxford Univ. Press, 2012).
7. International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. Nature 491, 711–716 (2012).
8. Yang, P. et al. PROTEIN DISULFIDE ISOMERASE LIKE 5-1 is a susceptibility factor to plant viruses. Proc. Natl Acad. Sci. USA 111, 2104–2109 (2014).
9. Pourkheirandish, M. et al. Evolution of the grain dispersal system in barley. Cell 162, 527–539 (2015).
10. Russell, J. et al. Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. Nat. Genet. 48, 1024–1030 (2016).
11. Künzel, G., Korzun, L. & Meister, A. Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. Genetics 154, 397–412 (2000).
12. Beian, S. et al. Multiplex sequencing of bacterial artificial chromosomes for assembling complex plant genomes. Plant Biotechnol. J. 14, 1511–1522 (2016).
13. Muñoz-Amatrana, M. et al. Sequencing of 15,622 gene-bearing BACs clarifies the gene-dense regions of the barley genome. Plant J. 84, 216–227 (2015).
14. Ounis, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16, 236 (2015).
15. Colmsee, C. et al. BARLEX - the barley Draft Genome Explorer. Mol. Plant 8, 964–966 (2015).
METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Sequencing and assembly of individual BAC clones. Barley genome sequencing relied exclusively on shotgun sequencing of 88,731 BAC clones using high-throughput next-generation sequencing by-synthesis25. This comprised 15,661 so-called gene-bearing BAC clones, preselected mainly by overpro- baid hybridization of the presence of transcribed genes and fingerprinted for definition of a minimum tiling path of the barley gene space. These gene-space minimum tiling path BAC clones were sequenced as combinatorial pools by Illumina short-read technology and, after quality trimming of de-convoluted reads, were assem- bled using Velvet version 1.2.09 as previously described31. The remaining 73,070 BACs were selected from a minimum tiling path representing the physical map of the barley genome32. Minimum tiling path BAC clones assigned to different barley chromosomes were sequenced at one of four sequencing centres, relying on highly multiplexed paired-end and mate-pair sequencing libraries using either the Roche 454 Titanium or the Illumina MiSeq, HiSeq2000 and HiSeq2500 platforms (Supplementary Note 1 and ref. 51). In brief, sequencing reads were de-convoluted on the basis of the used BAC-specific barcode sequence tags and assembled with sequencing centre-specific assembly pipelines. BAC clones sequenced on the Roche 454 Titanium platform were assembled with MIRA26 according to previ- ously described procedures32,53. Illumina HiSeq2000 paired-end sequencing data (2 × 100 nucleotides) of BAC clones were assembled either with CLC Assembly Cell version 4.0.6 beta (http://www.clcbio.com/products/clc-assembly-cell/) set to default parameters32, SOAPdenovo version 2.01 (ref. 35) or the ABySS assembler (version 1.5.1f). Sequence contigs of the de novo BAC assemblies larger than 500 base pairs (bp) were scaffolded using mate-pair sequencing information either generated from BAC DNA–derived 8 kb insert mate-pair sequencing libraries or from 2 kb, 5 kb or 10 kb genomes-derived mate-pair libraries. This was achieved by either using BWA mem version 0.7.4 (ref. 56) with default parameters for read mapping, followed by scaffolding individual BACs using SPASE version 3.0 Standard57, or with SOAPAligner soap2 version 2.21 and using SOAPdenovoa scaffold version 2.01.

 Genome-wide three-dimensional chromatin conformation capture sequencing. To generate physical scaffolding information for the BAC sequence based genome assembly, as proposed in ref. 21, Hi-C and tethered conformation capture (TCC) sequencing data were generated from 7-day-old leaf tissue of greenhouse-grown barley plantlets by adapting previously published procedures (Supplementary Note 2). In brief, for Hi-C, freshly harvested leaves were cut into 2 cm pieces and vacuum infiltrated in nuclei isolation buffer supplemented with 2% for- maldehyde. Crosslinking was stopped by adding glycine and additional vacuum infiltration. Fixed tissue was frozen in liquid nitrogen and ground to powder before re-suspending in nuclei isolation buffer to obtain a suspension of nuclei. About 10⁵ purified nuclei were digested with 400 units of HindIII as described previously59. Digestion chromatin was marked by incubating with biotin-14-dCTP and Klenow enzyme using a fill-in reaction59 resulting in blunt-ended repaired DNA fragments. Biotin-14-dCTP from non-ligated DNA ends was removedby reversing the formaldehyde crosslink using protease K70. Biotin-14-dCTP from non-ligated DNA ends was removed by using Exonuclease III19. Hi-C and TCC products were mechanically sheared from the magnetic beads by reversing the formaldehyde crosslink using protease K70. Biotin-14-dCTP from non-ligated DNA ends was removed by using Exonuclease III19. Hi-C and TCC products were mechanically sheared to fragment sizes of 200–300 bp by applying ultrasound using a Covaris S220 device followed by size-fractionation using AMPure XP beads. DNA fragments in the range between 150 and 300 bp were blunt-end repaired and A-tailed before purification through biotin–streptavidin-mediated pull-down19. Illumina paired-end adapters were ligated to the Hi-C and TCC protocols, respectively, followed by PCR amplification, pooling of PCR products and purification with AMPure XP beads before quantification of Hi-C/TCC libraries by qPCR for Illumina HiSeq2500 PE100 and HiSeq2000 PE100 libraries.

 Nanochannel-based genome mapping. Long-range scaffolding of genome sequence assemblies was facilitated by BioNano genome maps generated by nanochannel electrophoresis of fluorescently labelled high-molecular mass DNA obtained from flow-sorted chromosomes59. High-molecular mass DNA was prepared from 3.5 × 10⁷ purified chromosomes (whole genome) of barley cultivar Morex essentially following published procedures60,61. The purified chromosomes were embedded in agarose miniplugs to achieve approximate concentrations of 1 million chromosomes per 40 μl volume before being treated with proteinase K as described previously61. DNA was labelled at Nt.BspQI nicking sites (GTCTTCC) by incorporation of fluorescent-dUTP nucleotide analogues using Taq polymerase as described previously59. The labelled DNA was analysed on the Irys platform (BioNano Genomics) in 191 cycles in total, generating 243 Gb of data exceeding 150 kb. On the basis of the label positions on single DNA molecules, de novo assembly was performed by a pairwise comparison of all single molecules and graph building52. The parameter set for large genomes was used for assembly with the IrysView software. A P value threshold of 10⁻¹⁴ was used during the pairwise assembly, 10⁻¹⁰ for extension and refinement steps and 10⁻¹⁴ for merging contigs. A whole-genome map of 4.3 Gb was obtained (Extended Data Table 1).

 Data integration for constructing pseudomolecules. The construction of pseudomolecules representing the seven barley chromosomes followed an iterative, mainly automated procedure which involved the integration of the following major datasets: (1) sequence assemblies of 87,075 unique, successfully sequenced and assembled BAC clones; (2) BAC assembly information from a genome-wide physical map of barley56; (3) 571,814 end-sequences of BAC clones; (4) a dense linkage map assigning genetic positions to 791,177 contigs of a whole-genome shotgun assembly of barley cultivar Morex19; (5) Hi-C/TCC sequence information; and (6) the optical map of the genome of barley cultivar Morex. A schematic outline of the procedure is presented elsewhere22. In the first step, overlaps between individual BAC assemblies were searched with Megablast65 by either applying ‘stringent’ or ‘permissive’ alignment criteria22 and by combin- ing with the high density genetic map information. On the basis of this initial analysis, a BAC overlap graph was constructed by use of the R package igraph63 considering the above-listed additional datasets in subsequent iterative steps. Building the overlap graph focused first on overlaps obtained under ‘stringent’ search criteria for BACs within individual physical map contigs (FP contigs) and then subsequently also between independent FP contigs. Subsequently, overlaps obtained under ‘permissive’ criteria were evaluated while checking for cumulative evidences provided by the additional datasets supporting the overlap information62. Ordering and orienting of the resultant sequence scaffolds were achieved by integrating the overlap graph with Hi-C/TCC data61. Before the construction of pseudomolecules, we (1) identified genes incomplete or missing in the non-redundant sequence, but represented by (a) BAC sequence that had been excluded from the construction of the non-redundant sequence, or by (b) Morex WGS contigs, and (2) performed a final scan for contaminant sequences. Then a single FASTA file containing a single entry for each barley chromosome (a ‘pseudomolecule’) and an additional entry combining all sequences not anchored to chromosomes was constructed22.

 Three-dimensional chromatin conformation analysis. Mapping of Hi-C/TCC reads and assignment to restriction fragments were performed as described elsewhere22. Briefly, raw reads were trimmed with cutadapt66. Trimmed Hi-C reads were mapped to the barley pseudomolecule sequence with BWA mem (version 0.7.12)66. Duplicate removal and sorting were performed with NovoSort (http:// www.novocraft.com/products/novosort/). Mapped reads were assigned to restric- tions fragments with BEDTools67, tabulated with custom AWK scripts and imported into R (https://www.r-project.org/). Raw counts of Hi-C links were aggregated in 1 Mb bins and normalized separately for intra- and interchromosomal contacts using HiCNorm68. Contact probability matrices were plotted using standard R functions49. Principal component analysis was performed with the R function prcomp() on the matrix of log-transformed normalized Hi-C link counts between 1 Mb fragments.

 We fitted the linear model log₁₀(nl) ~ log₁₀(dist) + abs(cen_dist1 – cen_dist2) + abs(cen_dist1 – fam_dist1) + abs(cen_dist2 – fam_dist2) + abs(cen_dist1 – fam_dist2) (using the R function lm()). Here, nl is the normalized link count between two 1 Mb bins, dist is their distance in the linear genome, cen_dist1 and cen_dist2 are the relative distances from the centromere for both loci, arm1 and arm2 are the chromosome arm assignment of both loci, and apo1 and apo2 are the relative distances of both loci from the ends of the chromosome arm (that is, apo1 is close to zero if locus 1 is either near the centromere or the telomere, and close to one if locus 1 resides in interstitial regions). TCC reads of Morex × Barke F₂ hybrids were mapped to a synthetic reference representing the parental genomes. An in silico Barke assembly was created by inserting SNPs discovered by aligning Barke WGS reads to the Morex reference assembly with BWA MEM66 using variant calling with SAMtools69. SNPs were then inserted into the Morex reference using the FastaAlternateReferenceMaker of GATK21. TCC reads of the hybrid were then mapped to the synthetic reference as described above. Only uniquely alignable read pairs were considered. Hi-C link counts were tabulated at the level of chromosomes.
Fluorescence in situ hybridization was performed with *H. vulgare* nuclei as described earlier32 using Arabidopsis-type telomere and barley centromere-specific [AGGGAGG] repeats (probe)73.

**Automated annotation of transcribed regions.** Automated gene annotation of the barley reference sequence assembly was based on four datasets providing independent gene evidence information (Supplementary Note 3). This included (1) RNA sequencing (RNA-seq) data; (2) reference protein predictions from barley3, rice3, *B. distachyon*37 and *S. bicolor*76; (3) published barley full-length complementary DNA (cDNA) sequences77; and (4) newly generated barley PacBio Iso-Seq data. Previously published2 and newly generated RNA-seq data were derived from a total of 16 different tissues, each with three biological replicates, including seven vegetative, six inflorescence, two developing grain and one germinating grain tissues. RNA-seq libraries were sequenced on Illumina HiSeq2000 in paired-end 2 × 100 nucleotides (PE100) mode (Supplementary Note 3). To support gene calling in general, and the identification of alternative splice forms in particular, enriched full-length transcript information was generated by the Iso-Seq method using the PacBio RS II system and DNA Sequencing Chemistry 4.0 version 2 (Supplementary Note 3). RNA-seq-based transcript structures, reference-based gene model predictions, structure information from Iso-Seq alignments as well as structure information from flcDNA sequence alignments were clustered into a consensus transcript set using Cuffcompare74 (Supplementary Note 3). Predicted transcript sequences were automatically extracted into a single FASTA file on the basis of respective coordinates in the genome assembly. Putative open reading frames and corresponding peptide sequences, including prediction of Pfam domains, were obtained by applying TransDecoder (https://transdecoder.github.io), which also resulted in reports about predicted alternative peptides per transcript (Supplementary Note 3). A single best translation per transcript was selected on the basis of BLASTT similarity comparison of all predicted peptides to a comprehensive protein database containing high-confidence protein sequences from *A. thaliana*80, maize47, *B. distachyon*37, rice3 and *S. bicolor*76, followed by additional filtering procedures (Supplementary Note 3). Functional descriptions (‘human readable descriptions’) were generated for all potential genes using the AHRD pipeline (https://github.com/groupschool/AHRD/) on the basis of one representative protein sequence for each gene locus. Gene candidates were then classified into high- and low-confidence genes and further subdivided into nine classes, each supported by different levels of gene evidence (Supplementary Note 3). High-confidence protein-coding genes either showed significant sequence homology to a reference protein or were associated with a predicted function. Low-confidence genes were characterized by (1) having no or only weak sequence homology to reference proteins and no predicted function, (2) they were candidates for transposons or (3) they lacked an open reading frame of a minimal length (Supplementary Note 3). Completeness of gene-space representation was evaluated with the BUSCO pipeline23 (Extended Data Fig. 2b).

**Feature distributions along the chromosomes.** A sliding window approach with a window size of 4 Mb and a shift of 0.8 Mb was used to display the distribution of different genome components and other features such as GC content or recombination rate along the chromosomes. The sliding window data were smoothed with the python function scipysignal.gaussian (p1 = 40, p2 = 10 for Fig. 1a; p1 = 15, p2 = 3 for Fig. 2a). The boundaries of genomic compartments (Fig. 1) are given in Supplementary Table 4.4.

**Annotation of the non-genic part of the genome.** Transposable elements were detected and classified by homology search with Vmatch (http://www.vmatch.de) using the Iso-Seq method against the BLAST database52. The filtering applied was as follows: (1) only alignments with at least one novel key character were retrieved; (2) multi-nucleotide polymorphisms or short insertions/deletions, while the remaining marker positions were ignored; (3) only markers with a putative function were retained (e.g. transposon, retrotransposon, DNA transposon, and SINE). The resulting data were filtered for high-confidence key characters and then either shortening (<90% coverage and ≥50 bp rest length) or removing low-scoring overlaps.

The identification of full-length LTR retrotransposons with LTRharvest82 resulted in 143,957 non-overlapping candidate sequences using the following parameter settings: ‘overlaps best = seed 30 -minlen100 = maxlen100 -mindist1000 = maxdist10000 -similar 85 -mintsd 4 -maxtsd 20 -motif tga -motifism 1 -vic 60 -xdrop 5 -mat 2 -mis 2 -ins 3 -del 3’. All candidates were annotated for Pfam A domains with hammer software83 and stringently filtered for false positives by several criteria, the main ones being the presence of at least one typical retrotransposon domain (for example, RT, RH, INT, GAG) and a tandem repeat context below 25%. This resulted in an initial set of 24,952 LTR retrotransposons. Insertion ages of the LTR retrotransposons were calculated according to the method of ref. 84 by the diversity of 5′ and 3′ LTRs that had been identical at the time of transposition. We used a grass-specific mutation rate of 1 × 10⁻⁸. The average age of all full-length LTR elements was calculated in 4 Mb windows and plotted in Fig. 1a. The frequency of 20% were determined using Tallymer85.

Phylogenetic analysis of Gypsy elements was performed on predicted protein sequences deposited at the TREP database52. Protein domains in predicted open reading frames were identified with Pfam86, SignalP87 and COILS88. For the analysis of transposable element content in up- and downstream regions of genes, 10 kb immediately flanking the predicted coding sequences of all high-confidence genes were extracted from the genome assembly. The genomic segments were then used in BLASTN searches against the TREP database52. After an initial annotation, previously unclassified or poorly characterized transposable element families were re-analysed and new consensus sequences were constructed. Analysis of up- and downstream regions was then repeated with the updated TREP database. The transposable element family producing the longest BLASTN hit was determined for every 20th base position of each 10 kb segment, resulting in 500 data points for each up- and downstream region of the high-confidence genes.

**Gene family analysis.** Gene family clusters were defined from 39,734 barley high-confidence class genes and the annotated gene sets of Rice MSU7.0 (39,049 genes, http://rice.plantbiology.msu.edu/), *B. distachyon* version 3.1 (31,694 genes, https://phytozome.jgi.doe.gov/pz/portal.html#info?alias=Org_Bdistachyon), *S. bicolor* version 3.1 (33,032 genes, https://phytozome.jgi.doe.gov/pz/portal.html#info?alias=Org_Sbicolor) and *A. thaliana* TAIR10 (27,416 genes, https://www.arabidopsis.org/) using OrthoMCL89 software version 2.0. Splice variants were removed from the datasets, keeping only the representative/longest protein sequence prediction, and datasets were filtered for internal stop codons and incompatible reading frames. In the first step, pairwise sequence similarities between all input protein sequences were calculated using BLAST29 with an e-value cut-off of 10⁻⁵. Markov clustering of the resulting similarity matrix was used to define the orthologue cluster structure, using an inflation value (=1) of 1.5 (OrthoMCL default). Gene families with barley-specific gene duplications, compared with other plant species, were extracted from the ENSEMBL Compara pipeline90. Over- and under-representation of Gene Ontology terms between barley and other plant species (Supplementary Tables 4.1–4.3) and between genomic compartments (Supplementary Table 4.5) were analysed with a hypergeometric test using the functions GOstats and GSEA from the Bioconductor R package91 against a universe of all genes with Gene Ontology annotations. REVIGO92, which removes redundant and similar terms from long Gene Ontology lists by semantic clustering, was applied to visualize the enrichment results. Expansion of three barley gene families encoding α-amylases, the vacuolar processing enzyme VPE2 protein subfamily and the sugar transporters SWEET1 subfamily, with specific importance in barley grain filling/seed development or barley germination/malting, were analysed in greater detail using BLAST searches (versus genome and gene prediction) as well as GenomeTeamread mappings to the barley genome assembly. Further details are provided in Supplementary Note 4. In situ hybrids for SWEET genes were performed as described previously93.

**Analysis of sequence and haplotype diversity.** Ninety-six two-row spring (n = 48) and winter (n = 48) homoyzogous inbred elite barley lines (Supplementary Table 5.1) were subjected to exome capture using the barley Roche NimbleGen exome capture liquid array44 against Illumina HiSeq 2500 platform. An average of 2 × 21,876,780 paired-end Illumina reads per sample was generated. This corresponds to approximately 72% coverage of the 61 Mb exome capture space.

The raw Illumina reads were mapped to the reference sequence with BWAMEM version 0.7.10 (ref. 66), using a stringent mismatch setting of ≤2% mismatches per read. Variant calling was performed with the Genome Analysis Tool Kit (GATK)53 version 3.4.0, following the GATK Best Practices pipeline (https://www.broadinstitute.org/gatk/guide/best-practices.php). This included read-duplication, indel realignment, base quality score recalibration and variant calling with the latest version of the HaplotypeCaller. The workflow was implemented in a C-shell script. The Table assembly view76 was used for visual spot checks of mappings and SNPs calls.

Variant discovery resulted in 15,982,580 variants in total, of which 943,959 were multi-nucleotide polymorphisms or short insertions/deletions (indels), while the remainder represented SNPs. For subsequent genetic analysis, we first reduced the total variant dataset by applying rigorous filtering criteria to produce a highly robust subset of 72,563 SNPs distributed across all seven barley chromosomes. The filtering applied was as follows: (1) ≥8× coverage for ≥50% of the samples; (2) ≥95% of samples represented at each SNP locus; (3) ≥5% minor allele frequency at the level of the sample: that is, counting sample genotypes rather than individual reads. The resulting dataset was filtered using plink command. These filters reduced false-positive variant calls by removing spurious variant calls resulting from systematic read mis-mapping. Of this filtered dataset, a subset of 3,500 randomly sampled markers from each chromosome was analysed with the Haploview software96. This subsampling was required as Haploview was
unable to generate the required plots when larger data volumes were used as input. 

Haploview was run on duplicates, using the GABRIEL blocks method. The genotype calls were also imported into the genotype visualization software Flapjack\(^5\) to produce chromosome-scale images of haplotype diversity within the spring and winter pools. Diversity statistics were calculated in GenALEx version 6.502 (ref. 98) and rolling averages based on 100 adjacent SNPs were plotted in Microsoft Excel 2010.

Data availability. The genotype assembly for barley has been deposited in the Plant Genomics and Phenomics Research Data Repository under digital object identifier http://dx.doi.org/10.5447/IPK/2016/34. Accession numbers for all deposited datasets are listed in Supplementary Note 1. The barley genome assembly has been deposited on the IPK barley Blast Server (http://webblast.ipk-gatersleben.de/barley/_ibsc/). All other data are available from the corresponding authors upon reasonable request.

51. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. In Computer Science and Biology: Proc. 99th German Conference on Bioinformatics (eds Hofestät, R. et al. 45–56 (GCBB, 1999).

52. Steyaert, B. et al. De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome assembly in the complex genome of barley. BMC Genomics 10, 547 (2009).

53. Taudien, S. et al. Sequencing of BAC pools by different next generation sequencing platforms and strategies. BMC Res. Notes 4, 411 (2011).

54. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18 (2012).

55. Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123 (2009).

56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

57. Boetzer, M., Henkel, C. V., Jansen, H. L., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579 (2011).

58. Belton, J. M. et al. Hi-C—a comprehensive technique to capture the conformation of genomes. Methods 58, 268–276 (2012).

59. Stariková, H. et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol. J. 14, 1523–1531 (2016).

60. Lysák, M. A. et al. Flow karyotyping and sorting of mitotic chromosomes of barley (Hordeum vulgare L.). Chromosome Res. 7, 431–444 (1999).

61. Simkó, H., Čhalíková, J., Vrána, J., Lysák, M. & Doležel, J. Preparation of HMW artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. Plant J. 84, 385–394 (2015).

62. Aliyeva-Schnorr, L. et al. Cyto genetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. Plant J. 84, 385–394 (2015).

63. Budakova, S. et al. Sequence organization of barley centromeres. Nucleic Acids Res. 29, 5029–5035 (2001).

64. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature 436, 793–800 (2005).

65. International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463, 763–768 (2010).

66. Paterson, A. H. et al. The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551–566 (2009).

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

68. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578 (2012).

69. Attschel, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

70. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000).

71. Spannagl, M. et al. PGSB PlantsDB: updates to the database framework for comparative plant genome research. Nucleic Acids Res. 44 (D1), D1141–D1157 (2016).

72. Ellinghaus, D., Kurtz, S. & Willihoit, L. LRHast: an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18 (2008).

73. Eddy, S. R. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195 (2011).

74. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. Nat. Genet. 20, 43–45 (1998).

75. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans. Comput. Biol. Bioinformatics 10, 645–656 (2013).

76. Bateman, A. et al. The Pfam protein families database. Nucleic Acids Res. 32, D138–D141 (2004).

77. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786 (2011).

78. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. Science 252, 1162–1164 (1991).

79. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).

80. Bolster, D., Staines, D. M., Pritchard, E. & Kersey, P. Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomics data. Methods Mol. Biol. 1374, 115–140 (2016).

81. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004).

82. Supek, F., Bošnjak, M., Škunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. PLoS ONE 6, e21800 (2011).

83. Rudnich, V., Weier, D., Rudnich, R., Weschke, W. & Weber, H. Development of maternal seed tissue in barley is mediated by regulated cell expansion and cell disintegration and coordinated with endosperm growth. J. Exp. Bot. 62, 1217–1227 (2011).

84. 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).

85. Milne, I. et al. Tablet—next generation sequence assembly visualization. Bioinformatics 26, 401–402 (2010).

86. Barrett, J. C., Fry, B., Miller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265 (2005).

87. Milne, I. et al. Flapjack—graphical genotype visualization. Bioinformatics 26, 2529–2531 (2010).

88. Peakall, R. & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28, 2537–2539 (2012).

89. The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788 (2014).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Extended Data Figure 1 | Gene annotation pipeline. a, Gene annotation pipeline combined gene evidence information from four data sources. Open reading frames were then predicted for 83,105 gene candidates. b, Gene candidates were classified into high-confidence (HC) and low-confidence (LC) genes on the basis of homology to reference proteins and alignment to library of repeat elements. Additional filtering procedures were applied before defining the final gene sets. Arrows between boxes with counts of high-confidence and low-confidence genes in each step indicate re-classifications (high-confidence to low-confidence, or low-confidence to high-confidence).
Extended Data Figure 2 | Assembly validation. a, Conserved gene order between barley (y axis) and B. distachyon (x axis). b, Completeness of the gene annotation as assessed by BUSCO. c, Representation of repetitive k-mers in reads and assemblies. d, Representation of full-length LTR retrotransposons in sequence assemblies of plant genomes with different sizes (represented by black points). The map-based reference sequence of barley reported in the present paper is shown in blue. Red dots correspond to shotgun assemblies of the barley genome and wheat chromosome 3B.

| Gene set | Complete genes | Fragmented genes | Missing genes |
|----------|----------------|------------------|---------------|
| All gene models | 97.6% | 1.3% | 1.2% |
| High-confidence genes only | 94.7% | 1.9% | 3.5% |
| Low-confidence genes only | 10.5% | 4.5% | 85.0% |
| Gene models annotated on WGS assembly | 95.3% | 2.9% | 1.8% |
Extended Data Figure 3 | Hi-C contact matrices. a, Intrachromosomal contacts. b, Interchromosomal contacts. Darker red indicates a higher contact probability.
Extended Data Figure 4 | Global patterns in Hi-C contact matrices. 

a, Principal component analysis of intrachromosomal Hi-C contact matrices. The eigenvectors of the first three principal components are plotted. Centromere positions are marked with a red line. 

b, Proportion of variance explained by linear models incorporating position informational in the linear genome fitted to the Hi-C contact matrices. 

c, Hi-C link counts in Morex × Barke F1 hybrids within the same chromosome, between homologous chromosomes and between non-homologous chromosomes.

Extended Data Figure 5 | Distributions of genomic features and the context of repetitive elements. a, b, Panels a and b are analogous to Figs 1a and 2a. Grey vertical connector bars and dashed lines inside sub-panels between sub-panels for each chromosome indicate centromere positions.
Extended Data Figure 6 | Experimental strategy to distinguish individual amyl1_1 copies by PCR from genomic DNA through polymorphisms in the extended promoter regions of amyl1_1 full-length copies. a, Experimental strategy, primers CD52_amy1fw and CD53_amy1rc bind in the extended promotor region of all full-length amyl1_1 copies (expected amplicon sizes are 225 bp for amyl1_1a, 299 bp for amyl1_1b and amyl1_1d and 336 bp for amyl1_1c). Forward primers CD54_fw1a, CD55_fw1b and CD56_fw1c are designed to specifically amplify copies amyl1_1a, amyl1_1b and amyl1_1c, respectively when used with reverse primer CD58_amy1rc, which binds in the coding region of all amyl1_1 copies. Expected amplicon sizes are 1,024 bp (amyl1_1a), 1,026 bp (amyl1_1b) and 757 bp (amyl1_1c). Primer pair (CD55_fw1b–CD58_amy1rc) further binds to copy amyl1_1d: here, sequences of the expected amplicons contain sufficient polymorphisms to distinguish these copies from each other. Positions of selected sequence polymorphisms and deleted regions suitable to distinguish single copies are indicated as black vertical bars and gaps, respectively. Numbering was done in respect of copy amyl1_1b. b, PCR amplification of amyl1_1 promoter regions in six barley cultivars and landraces. As expected, a PCR for cultivar Morex, using universal primers CD52_amy1fw and CD53_amy1rc, resulted in three amplicons of the expected sizes 225, 299 and 336 bp (compare a), which was confirmed by Sanger sequencing. Further primers CD52_amy1fw and CD53_amy1rc were used to amplify the amyl1_1 extended promoter region in various barley cultivars. These experiments indicate polymorphic variation in, or even absence of, single promoters of amyl1_1 in the different cultivars. The cultivars analysed differ in row type (six-rowed: cultivars Morex, Masan Naked 1, Akashinriki, Etincel; two-rowed: cultivars Barke, Bowman, Masan Naked 1, Akashinriki; winter barley: cultivar Etincel) and geographic origin (North America: cultivars Morex, Bowman; Europe: cultivars Barke, Etincel; Asia: cultivars Masan Naked 1, Akashinriki). The cultivars Masan Naked 1 and Akashinriki depict landraces used for food, Bowman was classified as non-malting barley, while Morex, Barke and Etincel represent modern malting barley. c, Copy-specific PCR amplification of amyl1_1 extended promoter regions. PCR amplification and Sanger sequencing identified three amyl1_1 copies in barley cultivar Morex: amyl1_1a (CD54_fw1a–CD58_amy1rc), amyl1_1b (CD55_fw1b–CD58_amy1rc) and amyl1_1c (CD56_fw1c–CD58_amy1rc). Additionally, sequencing revealed two polymorphic sites in PCR amplicon amyl1_1b (CD55_fw1b–CD58_amy1rc) at positions 721 bp (T/C) and 1175 bp (C/T) (see a), indicating the presence of one or two additional amyl1_1b-like copies in the genome of the analysed individual. The presence of copy amyl1_1d could not be confirmed. The reason for that might have been sequence deviations in the cultivar Morex accession used for BAC library construction versus that used for the presented experiments, or differences in PCR efficiency for amplification of copies amyl1_1b and amyl1_1d.
Extended Data Figure 7 | SWEET gene expression. a, Control experiment for mRNA in situ hybridizations shown in Fig. 3c. In situ hybridization with sense probes for SWEET11a (top) and SWEET11b (bottom). Scale bars, 100 μm. b, Expression of SWEET11a and SWEET11b. Results of qPCR in different plant organs and in the developing grains at 7 days after flowering (DAF).
Extended Data Figure 8 | Haplotype blocks in sets of 48 samples each of elite two-row spring barley lines (top half of each chromosome’s figure) and winter barley lines (bottom half), separately for each chromosome. We restricted the number of SNPs per chromosome by randomly choosing 3,500 to fit with the maximum permitted by the software. The red and green plots in the centre of each chromosome figure represent whole-canvas dumps produced with the Flapjack software. Markers are arranged in columns in linear order along the chromosome; red pixels represent reference alleles, while green pixels represent alternative alleles. Each row represents a barley cultivar; these have been sorted top to bottom by year of introduction (ascending). The Flapjack plots are framed by cropped linkage disequilibrium plots generated with the HaploView software. Colour intensity conveys the extent of linkage between pairs of markers (red, highest). Approximate centromere positions are indicated by semi-opaque grey squares. The triangles with the thin black outline represent haplotype blocks as computed by HaploView. In some regions, extensive stretches exist where no blocks were detected (for example, chr2H, spring lines in top half, near centromere). These generally present highly monomorphic regions where there is no evidence for multiple haplotypes, and consequently blocks were not called.
### Extended Data Table 1 | Hi-C and optical map datasets for chromosome-scale assembly

#### a Summary of Hi-C libraries

| Library | Number of all reads | Number of mapped reads | Links between restriction fragments |
|---------|---------------------|------------------------|-----------------------------------|
| HiC1    | 229,672,122         | 63,133,030             | 7,449,949                         |
| HiC2    | 334,742,791         | 79,745,191             | 7,663,777                         |
| HiC4    | 183,044,989         | 53,818,372             | 4,983,859                         |
| HiC5    | 178,785,306         | 58,212,813             | 2,439,898                         |
| HiC6    | 219,294,615         | 63,853,743             | 5,594,744                         |
| TCC2    | 260,968,878         | 55,242,411             | 7,431,165                         |
| TCC4    | 182,033,300         | 35,964,622             | 6,336,274                         |
| TCC5    | 204,856,338         | 42,544,941             | 7,913,758                         |
| TCC7    | 236,976,831         | 65,188,433             | 7,197,767                         |
| TCC8    | 226,042,216         | 71,397,037             | 4,380,187                         |
| TCC9    | 237,059,303         | 49,879,999             | 8,877,701                         |
| TOTAL   | 2,493,476,689       | 638,980,592            | 70,269,079                        |

#### b Raw data and assembly statistics of the optical map.

| Statistic                  | Value     |
|----------------------------|-----------|
| Number of molecules > 150 kb | 774,557   |
| Molecule N50               | 340 kb    |
| Number of contigs          | 2,875     |
| Assembly length            | 4,289 Mb  |
| Average contig coverage    | 57-fold   |
| Fraction of molecules aligned to assembly | 85 %     |
Extended Data Table 2  |  Statistics on gene annotation and genomic compartments

**a** Gene annotation statistics for high-confidence (HC) and low-confidence (LC) genes.

|                      | 1H  | 2H  | 3H  | 4H  | 5H  | 6H  | 7H  | Un  | TOTAL |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| No. of HC genes      | 4,834 | 6,518 | 5,760 | 4,380 | 6,165 | 4,544 | 5,576 | 2,157 | 39,734 |
| No. of LC genes      | 4,911 | 6,259 | 6,035 | 4,720 | 6,420 | 4,994 | 6,712 | 1,898 | 41,949 |
| No. of HC transcripts| 30,711 | 40,432 | 38,322 | 29,388 | 37,877 | 28,293 | 36,709 | 7,538 | 248,270 |
| No. of LC transcript | 10,754 | 13,287 | 12,589 | 10,331 | 12,471 | 10,354 | 12,795 | 3,275 | 85,856 |
| Mean length of HC genes | 5,450 | 7,533 | 5,836 | 5,472 | 6,013 | 6,091 | 6,319 | 3,195 | 6,010 |
| Mean length of LC genes | 2,460 | 2,561 | 2,145 | 2,253 | 2,361 | 2,322 | 2,286 | 1,982 | 2,328 |
| Median no. of transcript per HC gene | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 |
| Median no. of transcript per LC gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Mean length of HC transcripts | 1,990 | 1,876 | 1,992 | 1,963 | 1,926 | 1,961 | 1,888 | 1,475 | 1,927 |
| Mean length of LC transcripts | 1,595 | 1,484 | 1,532 | 1,487 | 1,534 | 1,453 | 1,360 | 1,156 | 1,478 |
| Median no. of exon per HC transcript | 6 | 5 | 6 | 6 | 5 | 5 | 5 | 4 | 5 |
| Median no. of exon per LC transcript | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| Mean length of HC proteins | 380 | 351 | 364 | 366 | 357 | 361 | 362 | 298 | 360 |
| Mean length of LC proteins | 191 | 173 | 184 | 166 | 179 | 164 | 165 | 164 | 174 |

**b** Genomic compartments across all chromosomes

|                      | ZONE 1 | ZONE 2 | ZONE 3 |
|----------------------|--------|--------|--------|
|                      | distal | interstitial | proximal |
| Size                 | 433 Mb | 3,075 Mb | 1,076 (Mb) |
|                      | (9 %)  | (63.6 %) | (22.3 %) |
| Number of genes      | 9,725  | 24,516 | 3,336 |
|                      | (24.5 %) | (91.7 %) | (8.4 %) |
| Gene density per Mb  | 22.5   | 8.0    | 3.1    |
| Transposon content   | 64.2 % | 82.1 % | 83.7 % |
| LTR/DNA-TE ratio     | 6.1    | 18.7   | 16.8   |
| Gypsy/Copia ratio    | 0.6    | 1.3    | 1.8    |
### Extended Data Table 3 | Repeat annotation statistics

| Mobile Element (TXX) | % of genome | % of TE bp | number | number % | size (Mb) | average length (bp) |
|---------------------|-------------|------------|--------|----------|-----------|---------------------|
| Class I: Retroelement (RXX) | 80.8 | 100.0 | 3,408,238 | 100 | 3,695 | 1,084 |
| LTR Retrotransposon (RLX) | 75.2 | 93.1 | 2,881,139 | 84.5 | 3,439 | 1,194 |
| Gypsy (RLG) | 16.0 | 19.8 | 588,579 | 17.3 | 732 | 1,243 |
| unclassified LTR (RLX) | 21.3 | 26.3 | 765,584 | 22.5 | 972 | 1,270 |
| non-LTR Retrotransposon (RXX) | 37.7 | 46.6 | 1,505,759 | 44.2 | 1,723 | 1,144 |
| Class II: DNA Transposon (DXX) | 0.3 | 0.3 | 21,217 | 0.6 | 12 | 581 |
| DNA Transposon Superfamily | 0.3 | 0.3 | 19,173 | 0.6 | 12 | 605 |
| SINE (RSX) | 0.0 | 0.0 | 2,044 | 0.1 | 1 | 355 |
| Class II: DNA Transposon Superfamily | 5.3 | 6.5 | 473,797 | 13.9 | 241 | 509 |
| CACTA superfamily (DTC) | 5.0 | 6.2 | 418,583 | 12.3 | 230 | 550 |
| hAT superfamily (DTA) | 4.7 | 5.9 | 375,421 | 11.0 | 217 | 578 |
| Muator superfamily (DTM) | 0.01 | 0.01 | 607 | 0.0 | 0 | 402 |
| Tcl/Mariner superfamily (DTT) | 0.01 | 0.01 | 8,199 | 0.2 | 1 | 134 |
| PIIF/Harbinger (DTH) | 0.08 | 0.10 | 9,007 | 0.3 | 4 | 402 |
| unclassified (DTX) | 0.03 | 0.03 | 6,413 | 0.2 | 1 | 191 |
| Class II: DNA Transposon Superfamily | 0.20 | 0.25 | 52,112 | 1.5 | 9 | 178 |
| Class II: DNA Transposon Superfamily | 0.03 | 0.04 | 1,843 | 0.0 | 1 | 818 |
| unclassified DNA transposon | 0.01 | 0.01 | 1,459 | 0.0 | 1 | 350 |
| Unclassified Element (TXX) | 0.32 | 0.40 | 53,302 | 1.6 | 15 | 274 |
| Retro-TE/DNA-TE ratio | 14.2 | 6.1 |
| Gypsy/Copia ratio | 1.3 | 1.3 |
Extended Data Table 4 | Information on gene families associated with malting quality

| Gene family | ID     | Chr | Strand | Coordinates on pseudomolecule (start to stop codon) | BAC sequence contig | Historical nomenclature | Copy-specific PCR primer for promoter region amyF_1 |
|-------------|--------|-----|--------|--------------------------------------------------|---------------------|------------------------|------------------------|
| α-amylases  | amy4_1 | 2H  | plus   | 511,664,000 – 511,567,683 | ma02311c11_c8 | N/A                    | N/A                    |
|             | amy4_2 | 3H  | minus  | 513,498,473 – 513,485,531 | ea00111c11_c1 | N/A                    | N/A                    |
|             | amy3   | 5H  | plus   | 517,452,674 – 517,544,307 | a01714b14_c3 | N/A                    | N/A                    |
|             | amy1_1 | 6H  | minus  | 533,880,485 – 533,879,015 | h0060c26_c2 | amy6_4^2              | CD54_fw1a              |
|             | amy1_2 | 6H  | minus  | 534,112,867 – 534,114,337 | e0332p17_c7 | amy6_4^2              | CD55_fw1b              |
|             | N/A    | 6H  | minus  | 534,258,381 – 534,259,057 | h0076e08_c8 | amy6_4^2              | N/A                    |
|             | amy1_1 | 6H  | minus  | 534,499,529 – 534,498,059 | e178f18_c4 | amy6_4^2              | CD56_fw1c              |
|             | amy2_1 | 6H  | minus  | 542,857,506 – 542,858,990 | e203j18_c8 | amy6_4^2              | N/A                    |
|             | amy2_2 | 7H  | minus  | 556,169,883 – 556,167,920 | h0261m10_c2 | amy2b^d              | N/A                    |
|             | amy2_3 | 7H  | minus  | 557,386,785 – 557,397,068 | a0332a16_c1 | N/A                    | N/A                    |
|             | N/A    | 7H  | minus  | 557,428,810 – 557,427,021 | a0332a16_c1 | N/A                    | N/A                    |
|             | amy1_1 | 7H  | minus  | 184,040,968 – 184,042,438 | a0174k01_c6 | amy6_4^2              | CD55_fw1b              |
|             | amy1_2 | 7H  | minus  | 195,047,130 – 195,048,600 | b0054j14_c4 | amy6_4^2              | N/A                    |
|             | amy1_3 | 7H  | minus  | 196,262,594 – 196,261,798 | b006b02_c14 | amy6_4^2              | N/A                    |