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; Himmelbach A; Taudien S; Felder M; Platzer M; Lonardi S; Duma D; Alpert M; Cordero F; Beccuti M; Ciardo G; Ma Y; Wanamaker S; Close TJ; Stein N; Cattonaro F; Vendramin V; Scalabrin S; Radovic S; Wing R; Schulte D; Steuernagel B; Schmutzer T; Mascher M; Scholz U; Taudien S; Platzer M; Sato K; Marshall D; Bhat P; Stein N; Mayer KF; Waugh R; Brown JW; Schulman A; Langridge P; Platzer M; Fincher GB; Muehlbauer GJ; Sato K; Close TJ; Wise RP; Stein N. - In: NATURE. - ISSN 0028-0836. - 491(2012), pp. 711-716.

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

The International Barley Genome Sequencing Consortium*

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.

Cultivated barley, derived from its wild progenitor Hordeum vulgare ssp. spontaneum, is among the world’s earliest domesticated crop species and today represents the fourth most abundant cereal in both poorer countries and in areas where food is harvested (http://faostat.fao.org). Approximately three-quarters of global production is used for animal feed, 20% is malted for use in alcoholic and non-alcoholic beverages, and 5% as an ingredient in a range of food products. Barley is widely adapted to diverse environmental conditions and is more stress tolerant than its close relative wheat. As a result, barley remains a major food source in poorer countries, maintaining harvestable yields in harsh and marginal environments. In more developed societies it has recently been classified as a true functional food. Barley grain is particularly high in soluble dietary fibre, which significantly reduces the risk of serious human diseases including type II diabetes, cardiovascular disease and colorectal cancers that afflict hundreds of millions of people worldwide. The USA Food and Drug Administration permit a human health claim for cell-wall polysaccharides from barley grain.

As a diploid, inbreeding, temperate crop, barley has traditionally been considered a model for plant genetic research. Large collections of germplasm containing geographically diverse elite varieties, landraces and wild accessions are readily available and undoubtedly contain alleles that could ameliorate the effect of climate change and further enhance dietary fibre in the grain. Enriching its broad natural diversity, extensive characterized mutant collections containing all of the morphological and developmental variation observed in the species have been generated, characterized and meticulously maintained. The major impediment to the exploitation of these resources in fundamental and breeding science has been the absence of a reference genome sequence, or an appropriate enabling alternative. Providing either of these has been the primary research challenge to the global barley community.

In response to this challenge, we present a novel model for delivering the genome resources needed to reinforce the position of barley as a model for the Triticeae, the tribe that includes bread and durum wheats, barley and rye. We introduce the barley genome gene space, which we define as an integrated, multi-layered informational resource that provides access to the majority of barley genes in a highly structured physical and genetic framework. In association with comparative sequence and transcriptome data, the gene space provides a new molecular and cellular insight into the biology of the species, providing a platform to advance gene discovery and genome-assisted 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 contig 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, comparing 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), and 304,523 BAC-end sequence (BES) pairs (Supplementary Table 3). These provided 1,136 megabases (Mb) of genomic sequence integrated directly into the physical map (Supplementary Tables 3 and 4). 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

*A list of authors and their affiliations appears at the end of the paper.

©2012 Macmillan Publishers Limited. All rights reserved
(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). 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 (~10 kb), 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 *Arabidopsis*, *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
First, we observed evidence for extensive alternative splicing. Of confidence barley genes seemed to be differentially regulated between developmental or tissue sample. More importantly, 36–55% of high-dynamic pathogen populations29,30. It is noteworthy that the highly Biased allocation to recombination-rich regions provides the genomic over-represented (1,3)-

Table 1 | Characteristics of high-confidence and low-confidence gene sets in barley

<table>
<thead>
<tr>
<th></th>
<th>High confidence</th>
<th>Low confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes</td>
<td>26,159</td>
<td>53,220</td>
</tr>
<tr>
<td>Gene loci positioned on barley cultivar</td>
<td>24,243 (93%)</td>
<td>51,015 (96%)</td>
</tr>
<tr>
<td>Morex assembly*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single exon</td>
<td>5,954 (25%)</td>
<td>37,395 (73%)</td>
</tr>
<tr>
<td>Multi exon</td>
<td>18,289 (75%)</td>
<td>13,620 (27%)</td>
</tr>
<tr>
<td>Number of distinct exons†</td>
<td>184,710</td>
<td>107,768</td>
</tr>
<tr>
<td>Mean number of distinct exons per gene</td>
<td>7.62</td>
<td>2.11</td>
</tr>
<tr>
<td>Number of genes with alternative transcript variants</td>
<td>13,299 (55%)</td>
<td>8,214 (16%)</td>
</tr>
<tr>
<td>Total number of predicted transcripts</td>
<td>62,426</td>
<td>69,266</td>
</tr>
<tr>
<td>Mean number of transcripts per gene</td>
<td>2.58</td>
<td>1.36</td>
</tr>
<tr>
<td>Mean gene locus size (first to last exon)</td>
<td>3,013 bp</td>
<td>972 bp</td>
</tr>
<tr>
<td>Mean transcript size (UTR, CDS)</td>
<td>1,878bp</td>
<td>931 bp</td>
</tr>
<tr>
<td>Mean exon size</td>
<td>454 bp</td>
<td>536 bp</td>
</tr>
<tr>
<td>Gene loci not positioned on barley cv.</td>
<td>1,916 (7%)</td>
<td>2,205 (4%)</td>
</tr>
<tr>
<td>Morex assembly*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tagged by unmapped RNA-seq reads</td>
<td>1,657 (86%)</td>
<td>1,127 (51%)</td>
</tr>
<tr>
<td>Not tagged by unmapped RNA-seq reads</td>
<td>259 (14%)</td>
<td>1,078 (49%)</td>
</tr>
</tbody>
</table>

* Gene locus representatives are (1) RNA-seq based transcript or (2) barley fl-cDNA that were mapped to the barley cultivar Morex assembly or tagged by RNA-seq based transcript during clustering.
† Exons of two or more transcripts were counted once if they have identical start and stop positions.
§ Gene locus representatives are barley fl-cDNAs that were not mapped to the barley cultivar Morex assembly and not matched by any RNA-seq based transcript.

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) 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, colocalizing with the MLA powdery mildew resistance gene cluster87. Biased allocation to recombination-rich regions provides the genomic environment for generating sequence diversity required to cope with dynamic pathogen populations43,44. It is noteworthy that the highly over-represented (1,3)-β-glucan synthase genes have also been implicated in plant-pathogen interactions45.

**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 spatial-temporal RNA-seq samples (Fig. 2a), slightly lower than reported for rice43 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 species33,34. Premature termination codons activate the NMD pathway3, which leads to rapid degradation of PTC+ transcripts, and have been associated with transcriptional regulation during disease and stress response in human and Arabidopsis, respectively34,36–39. 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 7.4). 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 divergence41. 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 processes41,42.

**Natural diversity**

Barley was domesticated approximately 10,000 years ago1. 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 diversity44. 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...
non-redundant high-confidence genes upregulated in comparison to all other genes incorporating PTC. Transcripts containing PTC were considered.

Predicted coding sequence could not be completely projected to genomic transcript model (partial mapping of fl-cDNA).

Morex contigs. Entire predicted coding sequence (100%) was transferred to transcript model on barley cultivar *M. spontaneum* may serve here as a reservoir of genetic diversity, using this diversity may itself be compromised by restricted recombination and the consequent inability to disrupt tight linkages between desirable and deleterious alleles. Surprisingly, the short arm of chromosome 4H had a significantly lower SNV frequency than all other barley chromosomes (Supplementary Fig. 33). This may be a consequence of a further reduction in recombination frequency on this chromosome, which is genetically (but not physically) shortest. Reduced SNV diversity was also observed in regions we interpret to be either the consequences of recent breeding history or could indicate landmarks of domestication (Fig. 3).

### 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 fulfilling criteria for PTC+ as detected in different spatial and temporal RNA-seq samples (Supplementary Note 7.4).

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Alternative splicing and transcripts containing PTCs in high-confidence genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General statistics of alternative splicing in high-confidence genes</strong></td>
<td></td>
</tr>
<tr>
<td>High-confidence genes with RNA-seq data to monitor alternative splicing</td>
<td>24,243</td>
</tr>
<tr>
<td>Predicted transcripts at high-confidence genes</td>
<td>62,426</td>
</tr>
<tr>
<td>Transcripts with complete CDS structures*</td>
<td>62,256</td>
</tr>
<tr>
<td>Transcripts with partial CDS structures†</td>
<td>170</td>
</tr>
<tr>
<td>Genes with alternative transcripts</td>
<td>13,299</td>
</tr>
<tr>
<td>Predicted transcripts derived from genes with alternative splicing</td>
<td>51,482</td>
</tr>
<tr>
<td><strong>Premature stop codon analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Predicted transcripts used for PTC analysis‡</td>
<td>51,338</td>
</tr>
<tr>
<td>Transcripts without PTC</td>
<td>41,461 (81%)</td>
</tr>
<tr>
<td>Transcripts containing PTC</td>
<td>9,877</td>
</tr>
<tr>
<td>PTC caused by intron retention</td>
<td>5,286 (10%)</td>
</tr>
<tr>
<td>PTC + transcripts predicted to be NMD++-sensitive</td>
<td>4,591 (9%)</td>
</tr>
<tr>
<td>Gene loci incorporating PTC + NMD transcripts</td>
<td>2,466</td>
</tr>
</tbody>
</table>

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

CDS, coding sequence.
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 SNV’s in non-overlapping 50-kb intervals of concatenated ‘Morex’ genomic scaffold; range, zero to maximum number of SNV’s 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[45-46]. 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[45]. 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[48], 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
The International Barley Genome Sequencing Consortium (IBSC)

Principal investigators Klaus F. X. Mayer1, Robbie Waugh2, Peter Langridge3, Timothy J. Close4, Roger P. Wise5, Andreas Graner6, Takashi Matsumoto7, Kazuhiro Sato8, Alan Schulman9, Gary J. Muehlbauer10, Nils Stein6

Physical map construction and direct anchoring Ruvinu Arjyanadass1, Daniela Schulte6, Naser Poursarebani11, Runouz Zhon12, Burkhard Steuer4, Martin Mascher6, Uwe Scholz13, Bujun Shi13, Peter Langridge3, Kavitha Madishetty13, Jan T. Svensson3, Prasanna Bhat4, Matthew Moscou5, Josh Resnik4, Timothy J. Close4, Gary J. Muehlbauer10, Peter Hedley13, Hui Liu13, Jenny Morris3, Robbie Waugh2, Zeew Frenkel13, Avraham Korol11, Hélène Bergès12, Andreas Graner6, Nils Stein (leader)6

Genomic sequencing and assembly Burkhard Steuer4, Uwe Scholz13, Stefan Tautien14, Marcus Felder14, Marco Groth14, Matthias Platzer13, Nils Stein (leader)6

BAC assembly and sequencing Burkhard Steuer4, Uwe Scholz13, Axel Himmelbach2, Stefan Tautien14, Marius Felder14, Matthias Platzer13, Stefano Lonardi18, Denis Dumia14, Matthew Alpert12, Francesca Cordero12, Marco Beccuti14, Gianfranco Ciardio18, Yaqin Ma14, Steve Wanamaker4, Timothy J. Close (co-leader)4, Nils Stein (leader)6

Genome annotation Manuel Spannagl19, Matthias Pfeifer1, Heiner Gundlach14, Nils Stein (leader)6

Repetitive DNA analysis Heiner Gundlach14, Cédric Mojsy18, Jaakko Tanskanen18, Simone Scalabrin15, Andrea Zuccolo18, Vera Vendramin18, Michele Morgante15,16, Klaus F. X. Mayer (co-leader)1, Alan Schulman (leader)9

Transcriptome sequencing and assembly Matthias Pfeifer1, Manuel Spannagl19, Pete Hedley13, Jenny Morris3, Joanne Russell20, Arnis Drukova21, David Marshall3, Michaa Bayer21, David Swarbrick21,13, Dhariana Sampath12,13, Sarah Aylng13, Melanie Faber13, Mario Caccamo13, Takashi Matsumoto7, Tsuyoshi Tanaka18, Kazuhiro Sato8, Roger P. Wise5, Timothy J. Close4, Steve Wanamaker4, Gary J. Muehlbauer10, Nils Stein6, Klaus F. X. Mayer (co-leader)4, Robbie Waugh (leader)2

Re-sequencing and diversity analysis Burkhard Steuer4, Thomas Schmutz15, Martin Mascher6, Uwe Scholz13, Stefan Tautien14, Matthias Platzer13, Kazuhiro Sato8, David Marshall3, Michaa Bayer21, Robbie Waugh (co-leader)2, Nils Stein (leader)6

Writing and editing of the manuscript Klaus F. X. Mayer (co-leader)4, Robbie Waugh (co-leader)4, John W. S. Brown21, Alan Schulman (leader)9, Peter Langridge3, Matthias Pfeifer1, Geoffrey B. Fischer17,18, Martin Neumann19, Heiner Gundlach14, Nils Stein (leader)6

NIPS/Berlin/IBIS, Helmholtz Zentrum Munchen, D-85764 Neuhberg, Germany. 2The James Hutton Institute, Invergowrie, Dundee DD2 5DE, UK. 3Australian Centre for Plant Functional Genomics, University of Adelaide, Glen Osmond 5064, Australia. 4The Department of Botany & Plant Sciences, University of California, Riverside, California 92521, USA. 5USDA-ARS, Department of Plant Pathology & Microbiology, Iowa State University, Ames, Iowa 50011-1020, USA. 6Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany. 7National Institute of Agrobiological Sciences, 21, 2-1, Kannondai, Tsukuba Ibaraki 305-8602, Japan. 8Okayama University, Kurashiki 710-0046, Japan. 9Mitt Agrofood Research and Institute of Biotechnology, University of Helsinki, FI-00014 Helsinki, Finland. 10University of Minnesota, Department of Agronomy and Plant Genetics, Department of Plant Biology, St Paul, Minnesota 55101, USA. 11University of Haifa, Haifa 31905, Israel. 12NRA-CNRG, Avezzoe CS 52627, France. 13Leibniz Institute of Age Research- Fritz Lippmann Institute (FLI), D-07745 Jena, Germany. 14Department of Computer Science & Engineering, University of California, Riverside, California 92521, USA. 15Istituto di Genomica Applicata, Via J. Linussio 51, 33100 Udine, Italy. 16Dipartimento di Scienze della Terra, Universita di Trento, Povo, 38122 Trieste, Italy. 17Department of Science and Technology, 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 Svezia 185, 10149 Torino, Italy.
METHODS

Building the physical map. BAC clones of six libraries of cultivar 'Morex' were analysed by high information content fingerprinting (HiCF)\(^{44}\). A total of 571,000 edited profiles was assembled using FPC v9.2 (Supplementary Table 2) (Sulston score threshold of 10\(^{-9}\), tolerance = 5, tolerated Q clones = 10%). Nine iterative automated re-arrangements were performed at successively reduced stringency (Sulston score of 10\(^{-8}\) to 10\(^{-10}\)). A final step of manual merging of FPC contigs was performed at lower stringency (Sulston score threshold 10\(^{-10}\))\(^{-22}\) considering genetic anchoring for markers with a genetic distance ≤ 5 cM. This produced 9,265 FPC contigs (approximately 14-fold haplode genome coverage) (Supplementary Table 2).

Genomic sequencing. BAC-end sequencing (BES). BAC insert ends were sequenced using Sanger sequencing (Supplementary Note 2.1). Vector and quality trimming of sequence trace files was conducted using LUCY\(^{50}\) (http://www.jcvi.org/cms/research/software/). Short reads (that is, <100 bp) were removed. Organelar DNA and barley pathogen sequences were filtered by BLASTN comparisons to public sequence databases (http://www.ncbi.nlm.nih.gov/).

BAC shotgun sequencing (BAGseq). Seed BACs of the FPC map were sequenced to reveal gene sequence information for physical map anchoring. 4,095 BAC clones were shotgun sequenced in pools of 2\(^{\pm}\)8 individually barcoded BACs on Roche/454 GS FLX or FLX Titanium\(^{54,55}\). Sequences were assembled using MIRA v3.2.0 (http://www.jcvi.org/projects_mira.html) at default parameters with features 'accurate', '454', 'genome', 'denovo'. An additional 2,183 gene-bearing BACs (Supplementary Note 4.3) which provided the highest marker density (3,973) and 59.6% of these MD pools was used for multiplexed screening of 42,302 barley EST-seq pools and 99.5% of the remaining pools

 shotgun sequencing (BACseq). Seed BACs of the FPC map were sequenced to reveal gene sequence information for physical map anchoring. 4,095 BAC clones were shotgun sequenced in pools of 2\(^{\pm}\)8 individually barcoded BACs on Roche/454 GS FLX or FLX Titanium\(^{54,55}\). Sequences were assembled using MIRA v3.2.0 (http://www.jcvi.org/projects_mira.html) at default parameters with features 'accurate', '454', 'genome', 'denovo'. An additional 2,183 gene-bearing BACs (Supplementary Note 3.2) were sequenced using Illumina HiSeq 2000 in 91 combinatorial pools\(^{13}\). Deconvoluted reads were assembled using VELVET\(^{53}\). Assembly statistics are given in Supplementary Table 4.

Whole-genome shotgun sequencing. Illumina paired-end (PE; fragment size ~350 bp) and mate-pair (MP; fragment size ~2.5 kb) libraries were generated from fragmented genomic DNA\(^{48}\) of different barley cultivars ('Morex', 'Barke', 'Bowman', 'Igri') and an S3 single-seed selection of a wild barley accession B1K-40-1255 (Supplementary Note 4.3) which provided the highest marker density (3,973) and 307 Mb of 'Morex' whole-genome shotgun contigs (Supplementary Note 11). Repeat and repeats were detected using RepeatMasker\(^{50}\) (http://www.repeatmasker.org/). Independent de novo assemblies were performed from data of cultivars 'Morex', 'Bowman' and 'Barke'.

Transcriptome sequencing. Eight tissues of cultivar 'Morex' (three biological replicates each) earmarking 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).

Genetic framework of the physical map. The genetic framework for anchoring the physical map of barley was built on a single-nucleotide variation (SNV) map\(^{48}\) (Supplementary Note 4.3) which provided the highest marker density (3,973) and resolution (N = 360, RIL/8) 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)\(^{49}\) to generate high-density genetic maps by comparing the population structure of wild barley. Genetic diversity was assessed using the GenomeMapper\(^{49,50}\) platform and strategies. BMC Res. Notes 4, 411 (2011).

Analysis of repetitive DNA and repeat masking. A total of 571,000 marker sequence reads from RNA-seq were mapped to the 'Morex' assembly. Details are provided in Supplementary Note 8.

Gene annotation, functional categorization and differential expression. Analysis of sequence diversity. Genome-wide SNV was assessed by mapping (BWA v0.5.9-r1656) the original sequence reads of sequenced genotypes to de novo assemblies. Sequence reads from RNA-seq were mapped to the 'Morex' assembly. Details are provided in Supplementary Note 8.

©2012 Macmillan Publishers Limited. All rights reserved.