Open access • Journal Article • DOI:10.1007/S11032-015-0253-1

BARLEYMAP: physical and genetic mapping of nucleotide sequences and annotation of surrounding loci in barley — Source link

Carlos Pérez Cantalapiedra, Carlos Pérez Cantalapiedra, Ridha Boudiar, Ana M. Casas ...+2 more authors

Institutions: Autonomous University of Barcelona, Spanish National Research Council

Published on: 20 Jan 2015 - Molecular Breeding (Springer Netherlands)

Topics: Genomics and Population

BARLEYMAP: physical and genetic mapping of nucleotide sequences and annotation of surrounding loci in barley

Carlos P. Cantalapiedra[1,2*], Ridha Boudiar[1], Ana M. Casas[1], Ernesto Igartua[1] and Bruno Contreras-Moreira[1,3*]

[1]Estación Experimental de Aula Dei (EEAD-CSIC), Avda. Montañana, 1005, 50059 Zaragoza, Spain.

[2]Plant Biology and Biotechnology PhD Program, Universitat Autònoma de Barcelona, Spain.

[3]Fundación ARAID, calle María de Luna 11, 50018 Zaragoza, Spain.

* Corresponding authors:

Email: cpcantalapiedra@eead.csic.es, bcontreras@eead.csic.es

Phone: +34 976716089

Fax: +34 976716145

Abstract

The BARLEYMAP pipeline was designed to map both genomic sequences and transcripts against sequence-enriched genetic/physical frameworks, with plant breeders as the main target users. It reports the most probable genomic locations of queries after merging results from different resources, so that diversity obtained from re-sequencing experiments can be exploited. In addition, the application lists surrounding annotated genes and markers, facilitating downstream analyses. Pre-computed marker datasets can also be created and browsed to facilitate searches and cross-referencing. Performance is evaluated by mapping two sets of long transcripts and by locating the physical and genetic positions of four marker collections widely used for high-throughput genotyping of barley cultivars. In addition, genome positions retrieved by BARLEYMAP are compared to positions within a conventional genetic map for a population of recombinant inbred lines (RIL), yielding a gene order accuracy of 96%. These results reveal advantages and drawbacks of current *in-silico* approaches for barley genomics. A web application to make use of barley data is available at http://floresta.eead.csic.es/barleymap. The pipeline can be set up for any species with similar sequence resources, for which a fully-functional standalone version is available for download.

Keywords

Barley, marker, genetic and physical maps, genotyping-by-sequencing, gene annotation, sequence mapping

Introduction

The main challenge for users of genomic data for applied purposes is the efficient use of the enormous amount of data generated by sequencing (Boller 2013). To aid geneticists and breeders of the *Triticeae* crops, some of the most important species for food security, several tools and data repositories have been developed recently, including HarvEST (Close et al. 2007), the T3 toolbox (http://triticeaetoolbox.org) or the Genome Zippers (Mayer et al. 2011).

The public release of the sequence-enriched genetic and physical map of barley (*Hordeum vulgare* L.) is being exploited for different purposes and already benefits breeding programs and companies worldwide, which previously had to rely solely on genetic maps and synteny-driven predictions. However, the current genomic assemblies are highly fragmented, as barley contains a major fraction of repeated sequences which hinder the assembly process (International Barley Genome Sequence Consortium 2012) (IBSC). Moreover, the anchored sequences come from different cultivars and sequencing methods, increasing the richness as well as the complexity of the reference map. In addition, another sequence-enriched map, based on one of the previous assemblies, has been published recently (POPSEQ, Mascher et al. 2013).

Due to that complexity, it can be a daunting task for plant breeders to place arbitrary nucleotide sequences within the barley genome and to identify nearby genes and genetic markers, useful for tasks such as genetic map assessment or map-based cloning. Furthermore, it is expected that some sequences will have multiple matches due to the presence of putative duplicated chromosome segments, paralogs and pseudogenes, as

65 well as possible inconsistencies in the assembly (Muñoz-Amatriain et al. 2013;
66 Poursarebani et al. 2013).

67 The described genomic patchwork is not exclusive to barley, as genomes from other
68 species have been and are currently being assembled with the aid of sequence-enriched
69 maps, especially since the advent of Next Generation Sequencing methods and when
70 dealing with highly repetitive genomes. Examples of the last are some species related to
71 barley: *Brachypodium distachyon* (International Brachypodium Initiative 2010),
72 *Aegilops tauschii* (Jia et al. 2013) and hexaploid wheat (*Triticum aestivum* L., Paux et
73 al. 2008; Paux et al. 2012). Among dicots, examples include grapevine (*Vitis vinifera*
74 L., Jaillon et al. 2007), potato (*Solanum tuberosum* L., Sharma et al. 2013) or
75 allotetraploid cotton (*Gossypium hirsutum* L., Yu et al. 2014).

76 Here we present a generic software platform designed to exploit genetic and physical
77 information from sequence-enriched maps. As such, it can be configured to work with
78 different sequence databases and maps, and thus it may take advantage of re-sequencing
79 data. The application can be used with two types of input:

80 1) DNA sequences, which are aligned to genome assemblies to estimate their likely
81 genomic positions. Two strategies are supported, allowing users to map either: i)
82 arbitrary genomic sequences and/or ii) transcripts or Expressed Sequence Tags
83 (ESTs), allowing for possible introns in the alignment.

84 2) Standard marker identifiers, so that users can have immediate access to pre-
85 computed positions of markers. For example, those widely used in high-
86 throughput genotyping experiments for a given species.

87 The BARLEYMAP pipeline, available at http://floresta.eead.csic.es/barleymap,
88 provides researchers a simple mapping report with details on genetic and physical

position of markers, as well as additional results with surrounding genes and known markers from other datasets. Here it is benchmarked and implemented as a web tool with barley data, although its use can be extended, with the standalone version, to any other species with similar genomic resources available.

Materials and Methods

Pipeline outline

The BARLEYMAP pipeline (Figure 1a) was mainly implemented in Python 2.6 and includes SplitBlast, a Perl script for distributing BLAST jobs (Contreras-Moreira and Vinuesa 2013). It has two main commands: [Align sequences] and [Find markers]. The first one uses a batch of FASTA-formatted DNA sequences as input, which are aligned by means of Blastn:Megablast from the BLAST package (Altschul et al. 1997), GMAP (Wu and Watanabe 2005) or both. The "auto" mode calls both programs sequentially: input sequences are first aligned by Blastn, and those which do not yield alignments over customizable sequence identity and query coverage thresholds (default: 98% and 95%, respectively) are then passed to GMAP. Results from both programs are filtered. In the case of Blastn, only the alignments with the best bit score are kept. Lacking bit scores, GMAP results are filtered by defining bad hits as those with both identity and coverage worse than those of other hits, as well as those marked as chimera. The alignment step is performed against one or more sequence databases (DBs in Figure 1a). These can be queried independently, merging the results afterwards, or by using a hierarchical strategy, in which only those queries not found in one DB are searched in

111 the next ones (Figure 1b). The [Find markers] command instead takes a list of query
112 identifiers as input and retrieves their alignment targets from pre-computed datasets.

113 For the mapping step, the positions of targets in one or more genetic/physical maps are
114 looked up and transferred to the initial queries. Results that provide the same location
115 for a given query are merged into a single record. Once map positions have been
116 compiled, the output report is augmented with genes or genetic markers anchored to
117 those genome regions. Finally, the user has toggle controls to append to the results the
118 functional annotation of those genes, as well as the genes to which the additional
119 markers hit.

120 **Barley data configuration and application distribution**

121 BARLEYMAP was originally configured to work with barley data. Whole Genome
122 Shotgun (WGS) assemblies of cultivars Morex, Barke and Bowman, as well as Morex
123 Bacterial Artificial Chromosome (BAC) contigs and BAC-End sequences (BES) from
124 the IBSC (2012), are employed as DBs. Genetic positions are retrieved separately from
125 two recently published maps: the genetic/physical framework from the IBSC and the
126 POPSEQ map of Morex contigs (Mascher et al. 2013). For the first one, mapping
127 positions were obtained from the AC datasets and assigned to the DBs depending on the
128 original source of the anchored sequence. As pre-calculated datasets, several collections
129 of genetic markers were compiled: i) Infinium® iSelect 9K (Comadran et al. 2012), ii)
130 DArTs™ (Wenzl et al. 2006), iii) DArTseq™ (Diversity Arrays Technology, Australia;
131 Kilian et al. 2012) and iv) a set of SNPs generated via genotyping-by-sequencing (GBS)
132 for the Oregon Wolfe Barley (OWB) population (Poland et al. 2012). All of them were
133 aligned to the DBs by means of BARLEYMAP [Align sequences]. Cultivar Haruna
134 Nijo full-length cDNAs (flcDNAs, Matsumoto et al.2011) and HarvEST assembly #36

cDNA sequences (Close et al. 2007), including 32,331 unigenes and 37,817 singletons, were aligned to the DBs as well. The default values of identity and coverage described above were used as thresholds for the alignments in all cases, performing both Blastn and GMAP steps for aligning against every DB independently. For comparison purposes, the previous datasets were also located using the hierarchical search with BARLEYMAP [Find markers] over the WGS assemblies (Morex, Barke and Bowman), BACs and BES references, in that order.

Finally, barley genes, including introns and up to 5,000 bp upstream of each transcript, were extracted from the Morex assembly, by means of custom scripts using the GTF data for High Confidence (HC) and Low Confidence (LC) genes from the MIPS FTP site (ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data). Those two gene sets were used as targets for matching of all the markers from the pre-computed datasets. The same thresholds described above to align markers to the reference DBs were applied, using the hierarchical search to prioritize hits on the HC dataset. Functional annotations were also downloaded from the MIPS FTP site.

The standalone version of BARLEYMAP is distributed with the pre-computed barley datasets to support the [Find markers] mode without further requirements (the total package is ~15 MB). The attached documentation explains the configuration required to run the [Align sequences] mode and to add custom DBs, maps or datasets, including those from any other organism for which similar sequence-based mapping resources are available. The BARLEYMAP web application relies on a CherryPy web server to handle client requests, and enables the user to query all the barley resources described above. When several DBs are chosen by the user, the web application runs the hierarchical search by querying the WGS assemblies of cultivars Morex, Bowman and Barke; Morex BAC contigs and BES, in that order.

Genetic map construction

The performance of BARLEYMAP was benchmarked against a newly developed genetic map for the barley population SBCC073 x Orria. SBCC073 is a Spanish landrace-derived inbred line (from Archidona, Málaga, Spain), with high yield under drought (Yahiaoui et al. 2014). Orria [(((Api x Kristina) x M66.85) x Sigfrido's) x 79W40762] is a semi-dwarf cultivar selected in Spain from a CIMMYT nursery, which is highly productive across most Spanish regions. This cross was carried out within the Spanish National Breeding Program. This is a population of 101 BC1F5 lines, originally developed to carry out quantitative trait locus (QTL) studies, which was genotyped with a DArTseq^{TM} GBS assay. One BC1F5 line was discarded on the basis of high percentages of heterozygous data. Therefore, the final mapping population comprised 100 lines. A genetic map was constructed in a two-step process, using first Joinmap 4 (Van Ooijen 2006) and then MSTMap (Wu et al. 2008).Resulting linkage groups were assigned to barley chromosomes based on the genomic positions assigned by BARLEYMAP.

The same polymorphic SNP markers were also queried by means of BARLEYMAP [Find markers] to both IBSC and POPSEQ maps, in hierarchical mode, to obtain *in-silico* maps. Spearman rank correlations were calculated between positions in the resulting genetic map and positions in the genetic/physical maps of IBSC and POPSEQ, using GenStat 16 (Payne 2009).

Results

Alignment of barley transcripts

To test the alignment step of BARLEYMAP (Figure 1a), the "auto" mode was selected to match long transcripts against the WGS assemblies of cultivars Morex, Barke and Bowman, as well as against the BAC contigs and BES from the IBSC, in that order by means of the hierarchical search. Of 28,620 flcDNAs from cultivar Haruna Nijo (Matsumoto et al. 2011), 60% were successfully aligned, with 68.5% of the alignments obtained by GMAP (Figure 2). Applying the same method, at least one hit was found for 59% out of 70,148 HarvEST cDNA sequences, with almost 60% of them aligned by Blastn. 79% and 86% of the previous hits were matched against the first queried database, the WGS assembly of cultivar Morex. The rest, 3,578 and 5,725 queries respectively, could only be matched in the remaining references.

Alignment of barley markers

A second benchmark consisted of mapping diverse collections of genetic markers, described in Materials and Methods, which are widely used by geneticists and breeders:

1) 7,864 Infinium® iSelect SNPs.
2) 2,000 Diversity Array Technology presence-absence (PAV) markers (DArTs™).
3) 24,061 GBS markers, including both SNP and PAV markers (DArTseq™)
4) 34,396 GBS SNP markers from the OWB population.

As observed for transcripts, a significant number of Infinium (30%) and DArT (16%) markers could only be confidently aligned with GMAP (Figure 2). However, this proportion was tiny for GBS markers, especially for DArTseq SNPs, which were mostly aligned by Blastn. Nonetheless, around 1,400 OWB GBS markers were aligned by GMAP.

205 Although these markers are short DNA sequences, their alignments produced mostly
206 single hits (over 98%) when searched independently in the WGS assemblies of cultivars
207 Morex, Barke and Bowman. However, such percentage was smaller for BAC contigs
208 and BES references (64% and 88%, respectively). Using the hierarchical method, this
209 percentage was near 99% for every marker dataset (Table 1).

210 The databases yielding the highest number of aligned markers were the WGS
211 assemblies (OnlineResource1, Figure S1), with those from cultivars Morex and
212 Bowman being slightly more informative than the one from cultivar Barke. The number
213 of markers aligned to BAC contigs and BES references was smaller in comparison. In
214 all cases, the use of the hierarchical search method resulted in a larger number of
215 markers available for position retrieval.

216 ***Mapping of aligned markers to barley genetic/physical maps***

217 Markers aligned to sequence DBs (Table 1) were then assigned genetic positions
218 retrieved from the IBSC and POPSEQ sequence-enriched maps (Online Resource
219 2).While POPSEQ comprises only contigs from the Morex assembly, IBSC map
220 positions can be retrieved for contigs from up to five different DBs. Thus, in the latter
221 case, marker positions were obtained either i) by merging the results from their
222 alignment to each DB independently or ii) from the hits obtained with the hierarchical
223 method (see Materials and Methods). As summarized in Table2, the highest number of
224 markers was mapped to the IBSC map, with 59% of them having a single map position.
225 In contrast, the POPSEQ results had the least number of mapped markers, but 99% of
226 them had a single map position. Regarding the hierarchical search, it misses ~4,300
227 marker positions with respect to IBSC, but a large majority of the sequences mapped
228 (99%) had a single map position, just as observed for POPSEQ.

A significant fraction of all the mapped markers lie on identical genetic positions and do not contribute to effectively resolve genomic intervals. Thus, considering only unique genetic locations, the hierarchical search method yields the maximum number of landmarks, with 6,908. This advantage of the hierarchical method when compared to the IBSC results comes at the cost of masking markers with multiple positions in different DBs. However, the information lost is mostly redundant, as revealed by the analysis of the positions of markers: for markers with multiple locations in the same DB reported by both search methods, 102 out of 140 (73%) lay in different chromosomes; for those removed by the hierarchical method (15,493) only 8% are in different chromosomes and most of the remaining are less than 5 cM apart, as shown in Online Resource 1, Figure S2.

Matching of genetic markers to barley genes

By taking the IBSC gene annotations, the sequences of genes, including introns and up to 5,000 bp upstream of each transcript, were obtained from the WGS assembly of cultivar Morex, yielding 62,426 HC and 69,299 LC sequences. A total of 68,321 markers from the datasets in Table 1 were matched to these gene sequences with the [Align sequences] command, hierarchical search and default parameters, as explained in Materials and Methods. Of these, 39.23% matched currently annotated genes, with 68% being HC genes.

Validating genetic maps of barley populations

The population SBCC073 x Orria yielded 2,483 polymorphic SNPs. These were filtered according to presence of missing data (<10%), heterozygotes (<10%), or allelic frequency of the donor parent (SBCC073) over 75%. After filtering, 1,227 SNPs were used to construct a genetic map. In a first step, linkage groups were created with

253 software Joinmap using the maximum likelihood algorithm. Then, in a second step, the
254 distances between markers were recalculated based on the Kosambi's mapping function
255 using MSTMap, which works more efficiently when the number of markers is large. A
256 total of 11 linkage groups were thus identified, representing 4 whole chromosomes (1H,
257 3H, 4H and 5H) and 3 fragmented ones (chromosome 2H in 3 groups, chromosomes 6H
258 and 7H in 2 groups each). Linkage groups were assigned to chromosomes, and the
259 resulting genetic positions of the 1,227 SNP markers compared to the positions assigned
260 to them by BARLEYMAP by hierarchically searching against either POPSEQ or IBSC
261 references. Correlation analyses, summarized in Figure 3 and Online Resource 1, Table
262 S1, reveal that locus order in the genetic map derived from the population is largely
263 similar to the implicit ordering of positions automatically assigned by the [Find
264 markers] command. The weighted averages obtained across linkage groups for
265 POPSEQ and IBSC were 0.92 and 0.96, respectively.There were nonetheless three
266 exceptions: i) a small linkage group made of 10 markers for which the genetic map is
267 necessarily less consistent than for larger groups; ii) linkage group 4H and; iii) linkage
268 group 6H.2. For these last two groups there was good agreement with only one of the
269 two physical maps used, pointing to local discrepancies between the data from IBSC
270 and POPSEQ (see Figure 3).

271

272 Discussion

273 Plant breeders have relied upon large numbers of de novo genetic maps and consensus
274 maps to deduce information about the relative position of their markers in relation to
275 others. The lack of common markers between maps has hindered the progress towards
276 the identification of genes or QTL underlying relevant traits for breeding. The era of

277 abundant sequence data is providing the opportunity to identify numerous new markers,
278 which are implemented in relatively cheap and high-throughput platforms, widely used
279 by the community. This is the case of GBS protocols or array genotyping systems based
280 on data from SNP calling pipelines.

281 In addition, such diversity of markers makes it possible to construct high-resolution
282 genetic maps, which, within genome sequencing projects, are used in conjunction with
283 physical maps to anchor sequences from shotgun or BAC sequencing. These resources
284 may not constitute a complete genome, but often contain a high proportion of the genes
285 of an organism, correctly placed in linear order. Many of the absent assembled contigs
286 come from highly repetitive, less gene abundant regions (International Barley Genome
287 Sequence Consortium 2012). Thus, exploiting such sequence-enriched maps can be of
288 help when locating genetic markers, when relating and comparing different maps to
289 each other, or in map-based cloning. This must be done with caution, since the actual
290 genotype or population under analysis could be more or less closely related to the
291 sequence references or could even bear local rearrangements (Farré et al. 2012).
292 Moreover, these sequence-enriched maps tend to have specific features for different
293 species, since each genome project may opt to use one or several genotypes as
294 references, or could use different sequencing technologies and sources. For these
295 reasons, it would be helpful to have tools flexible enough to help fill the gap between
296 specific genomic databases and the data used by plant breeders.

297 General resources, such as Ensembl Plants (Kersey et al. 2014), or more specific ones,
298 as the IPK Barley server (http://webblast.ipk-gatersleben.de/barley/viroblast.php), can
299 certainly be of help for these tasks. However, they are purely sequence-based and do not
300 make explicit use of the genetic maps underlying the physical assembly. Therefore, they
301 do not filter alignment matches in order to summarize mapping results, thus not

302 considering possible redundant positions as well as those with non-consistent locations
303 along the genome, originated from subtle differences among data sources. In addition,
304 the choice of BLAST as the only search engine complicates mapping transcripts. While
305 BLAST is able to generate local alignments that may be used to reconstruct a complete
306 spliced alignment, there is extensive literature reporting the importance of using
307 specialized algorithms for performing spliced alignments. The reason is not only for the
308 convenience of obtaining directly a full-length alignment, including its overall statistics,
309 but furthermore to consider micro-exons, large introns, donor/acceptor splice sites and
310 other features related to spliced sequences that could facilitate its correct identification.
311 This is especially important in the presence of paralogs, pseudogenes and segmental
312 duplications in the entire genome, which can hinder joining together local alignments,
313 and can be addressed better with programs which perform both the mapping and
314 alignment steps in a single job (see Gotoh 2008 and references therein). Finally, these
315 resources fail to include collections of genetic markers routinely used by breeders for
316 genotyping their plant materials. On the other hand, HarvEST(Close et al. 2007),
317 another important barley resource, does include SNP markers and IBSC positions of
318 Morex genes and homologs in other grasses, but cannot be used to interactively map
319 selected DNA sequences within the genome.

320 A unique feature of BARLEYMAP is the integration of alignment to sequence
321 references and mapping to genetic and physical frameworks. Being designed to
322 facilitate the access to positional information, BARLEYMAP concentrates in hiding the
323 underlying redundancy and complexity by means of a series of filters. First, it allows
324 the user to directly filter alignment results by percent identity and query coverage. Then,
325 it considers that the user should be typically interested in the best alignment result,
326 which is automatically selected by the BARLEYMAP web server (behaviour which

may be disabled in the standalone application). Moreover, it provides an explicit control on the presence of results from multiple mapping queries in the final report, avoiding redundant results both from the alignment and the mapping steps. In the first case, different hits to the same contig will share the same genetic and physical anchored position. In the second one, different contigs may be anchored to the same position, therefore yielding redundant results. Additionally, it facilitates the interpretation of unmapped queries, by separating those with alignment hit from those without it. The combined use of Blastn and GMAP allows BARLEYMAP to align transcripts, and markers derived from them, as demonstrated here by aligning flcDNAs, ESTs, and several genetic marker collections. Moreover, the use of a hierarchical method for alignment provides a reasonable compromise between the use of a single DB and the direct merging of results from the independent alignment to several DBs. In the first case, a number of queries may be absent, depending on the completeness of the assembly or presence-absence polymorphisms. For instance, cultivar Morex, as a spring cultivar, lacks the *VrnH2* gene (von Zitzewitz et al. 2005). Being an incomplete reference, other genes might only be found in alternative datasets, as the subset of flcDNAs (21%) that cannot be confidently aligned to Morex but are found in other references. The second approach, the alignment of every sequence to every reference, in addition to being a time-consuming process, produces queries with multiple targets and redundancy, both difficult to identify and fix, and can significantly reduce the number of useful markers associated to a single, unambiguous map location. The hierarchical method reduces computing time by aligning only the remaining unaligned sequences. In addition, queries with multiple mappings will arise only when the different locations are found in the same DB. As a drawback, the hierarchical method could be masking true multiple alignments (for example copy-number variation polymorphisms) in the case of

markers for which different targets are found in different DBs. However, most of those multiple positions seem to be very close to each other and are almost completely removed when using the hierarchical method. This suggests that such multiple positions are mostly artificial, generated by the independent mapping to different assemblies and sources. For efficiency and to ease downstream analysis, the web application uses only the hierarchical method when querying several DBs. The standalone application gives the user full control on using or not the hierarchical method.

BARLEYMAP allows barley geneticists and breeders to exploit their new and existing genotyping data in an accessible and time-saving manner, by integrating different marker types and flexible annotation retrieval in a single framework. It does so efficiently, as demonstrated by the good agreement between the orders of a purpose-built genetic map and the positions derived from BARLEYMAP (Online Resource1, Table S1). According to these observations it would be tempting to skip the mapping step altogether for any new population under study, and to proceed for further analyses using directly the positions derived from sequence-enriched genetic/physical maps. This benchmark suggests that analyses based on positions such as those produced by BARLEYMAP from currently available barley resources would produce reasonable results. However, the different outcome obtained by aligning the GBS markers to the two main genomic resources (IBSC and POPSEQ) advise against using such information as the gold standard for position, at least until the accuracy of barley references improves, and even then maybe only for genotypes close enough to the existing references.

A similar statement can be made for fine mapping purposes. Despite the fact that it can be of great help to use knowledge about surrounding genes and markers provided by BARLEYMAP, when working with a marker defined interval, the positions and relative

order of such features should be assessed carefully due to the technical and biological variability that might exist in the reference data (Hofmann et al. 2013; Liu et al. 2014).

Finally, BARLEYMAP allows research groups to use custom databases, maps and pre-computed datasets of markers, so that they may work with their own data and share it in a light-weight manner. Therefore, it provides a framework that ranges from a ready-to-work application for the retrieval of positional data from barley resources, up to a customizable pipeline that allows working with sequence-based positional data, if available, from any organism.

Acknowledgements

We thank Andrzej Kilian for his help with DArT markers.

Funding: This work was funded by DGA - Obra Social La Caixa [grant number GA-LC-059-2011] and by the Spanish Ministry of Science and Innovation [projects AGL2010-21929 and RTA2009-00006-C04-02]. Carlos P. Cantalapiedra is funded by [grant BES-2011-045905 linked to project AGL2010-21929]. Ridha Boudiar was supported by a Master's fellowship from IAMZ-CIHEAM.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Boller B (2013) Interview with Beat Boller, President of EUCARPIA, the European Association for Research on Plant Breeding. International Innovation (Environment):42-43.

Close TJ, Wanamaker S, Roose ML, Lyon M (2007) HarvEST. Methods Mol Biol 406:161-177.

Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, Shaw P, Bayer M, Thomas W, Marshall D, Hedley P, Tondelli A, Pecchioni N, Francia E, Korzun V, Walther A, Waugh R (2012) Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. Nat Genet 44:1388-1392.

Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79:7696-7701.

Farré A, Cuadrado A, Lacasa-Benito I, Cistué L, Schubert I, Comadran J, Jansen J, Romagosa I (2012) Genetic characterization of a reciprocal translocation present in a widely grown barley variety. Mol Breed 30:1109-1119.

Gotoh O (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res 36:2630-2638.

Hofmann K, Silvar C, Casas AM, Herz M, Buttner B, Gracia MP, Contreras-Moreira B, Wallwork H, Igartua E, Schweizer G (2013) Fine mapping of the Rrs1 resistance locus against scald in two large populations derived from Spanish barley landraces. Theor Appl Genet 126:3091-3102.

International Barley Genome Sequence Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711-716.

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

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P, French-Italian Public

Consortium for Grapevine Genome C (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Yang H, Liu X, He Z, Mao L (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91-95.

Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DS, Humphrey J, Kerhornou A, Khobova J, Langridge N, McDowall MD, Maheswari U, Maslen G, Nuhn M, Ong CK, Paulini M, Pedro H, Toneva I, Tuli MA, Walts B, Williams G, Wilson D, Youens-Clark K, Monaco MK, Stein J, Wei X, Ware D, Bolser DM, Howe KL, Kulesha E, Lawson D, Staines DM (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res 42:D546-552.

Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C, Aschenbrenner-Kilian M, Evers M, Peng K, Cayla C, Hok P, Uszynski G (2012) Diversity arrays technology: a generic genome profiling technology on open platforms. Methods Mol Biol 888:67-89.

Liu H, Bayer M, Druka A, Russell JR, Hackett CA, Poland J, Ramsay L, Hedley PE, Waugh R (2014) An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. BMC Genomics 15:104.

Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Munoz-Amatriain M, Close TJ, Wise RP, Schulman AH, Himmelbach A, Mayer KF, Scholz U, Poland JA, Stein N, Waugh R (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J 76:718-727.

Matsumoto T, Tanaka T, Sakai H, Amano N, Kanamori H, Kurita K, Kikuta A, Kamiya K, Yamamoto M, Ikawa H, Fujii N, Hori K, Itoh T, Sato K (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. Plant Physiol 156:20-28.

Mayer KF, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubalakova M, Suchankova P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo T, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Dolezel J, Waugh R, Stein N (2011) Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249-1263.

Muñoz-Amatriain M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KF, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N (2013) Distribution, functional

466 impact, and origin mechanisms of copy number variation in the barley genome.
467 Genome Biol 14:R58.
468 Paux E, Sourdille P, Mackay I, Feuillet C (2012) Sequence-based marker development in
469 wheat: advances and applications to breeding. Biotechnol Adv 30:1071-1088.
470 Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S,
471 Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H,
472 Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A
473 physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101-104.
474 Payne RW, Murray, D.A., Harding, S.A., Baird, D.B. & Soutar, D.M. (2009) GenStat for
475 Windows (12th Edition) Introduction. VSN International, Hemel Hempstead
476 Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic
477 maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing
478 approach. PLoS One 7:e32253.
479 Poursarebani N, Ariyadasa R, Zhou R, Schulte D, Steuernagel B, Martis MM, Graner A,
480 Schweizer P, Scholz U, Mayer K, Stein N (2013) Conserved synteny-based anchoring
481 of the barley genome physical map. Funct Integr Genomics 13:339-350.
482 Sharma SK, Bolser D, de Boer J, Sonderkaer M, Amoros W, Carboni MF, D'Ambrosio JM, de
483 la Cruz G, Di Genova A, Douches DS, Eguiluz M, Guo X, Guzman F, Hackett CA,
484 Hamilton JP, Li G, Li Y, Lozano R, Maass A, Marshall D, Martinez D, McLean K,
485 Mejia N, Milne L, Munive S, Nagy I, Ponce O, Ramirez M, Simon R, Thomson SJ,
486 Torres Y, Waugh R, Zhang Z, Huang S, Visser RG, Bachem CW, Sagredo B, Feingold
487 SE, Orjeda G, Veilleux RE, Bonierbale M, Jacobs JM, Milbourne D, Martin DM, Bryan
488 GJ (2013) Construction of reference chromosome-scale pseudomolecules for potato:
489 integrating the potato genome with genetic and physical maps. G3 3:2031-2047.
490 Van Ooijen JW (2006) JoinMap 4, software for the calculation of genetics linkage maps in
491 experimental populations. Kyazma B.V., Wageningen, Netherlands.
492 von Zitzewitz J, Szucs P, Dubcovsky J, Yan L, Francia E, Pecchioni N, Casas A, Chen TH,
493 Hayes PM, Skinner JS (2005) Molecular and structural characterization of barley
494 vernalization genes. Plant Mol Biol 59:449-467.
495 Wenzl P, Li H, Carling J, Zhou M, Raman H, Paul E, Hearnden P, Maier C, Xia L, Caig V,
496 Ovesna J, Cakir M, Poulsen D, Wang J, Raman R, Smith KP, Muehlbauer GJ, Chalmers
497 KJ, Kleinhofs A, Huttner E, Kilian A (2006) A high-density consensus map of barley
498 linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC
499 Genomics 7:206.
500 Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA
501 and EST sequences. Bioinformatics 21:1859-1875.

502 Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic
503 linkage maps from the minimum spanning tree of a graph. PLoS Genet 4:e1000212.
504 Yahiaoui S, Cuesta-Marcos A, Gracia MP, Medina B, Lasa JM, Casas AM, Ciudad FJ, Montoya
505 JL, Moralejo M, Molina-Cano JL, Igartua E (2014) Spanish barley landraces
506 outperform modern cultivars at low-productivity sites. Plant Breeding 133:218–226.
507 Yu JZ, Young CJL, Pepper AE, Li F, Yu S, Buyyarapu R, Sharma GC, Hinze LL, Percy RG
508 Toward Cotton Molecular Breeding: Challenges and Opportunities. In: International
509 Plant & Animal Genome XXII San Diego, CA, USA, 2014. p W604
510
511

Figure legends

Figure 1. The BARLEYMAP pipeline. **a)** Two types of input can be queried: identifiers (query IDs) or FASTA sequences. The alignment modes allow to query for genomic and/or transcript sequences. The "auto" mode uses both Blastn:Megablast and GMAP (dotted arrows inside "modes" box). This will be repeated for each sequence reference (DB), independently, unless the hierarchical search is specified, in which case only unaligned queries will be searched in the remaining DBs. If those do not align against any DB, they will be discarded, along with secondary alignments, alignments without position (unmapped) and GMAP chimeras (dotted arrows). Alternatively, alignment targets can be recovered from pre-computed data. Map positions of the targets will be associated to the queries, and after several filtering steps, enrichment with surrounding genes and markers will be performed. Finally, annotation of genes maybe appended to the results. **b)** An example with marker i_11_10679, from the Infinium dataset. First, it is searched by means of sequence alignments against the barley shotgun assemblies. With the hierarchical search (right track), the marker is found in the Morex assembly, so no other DBs are queried. The position (chr: chromosome; cM: genetic position in centimorgan; bp: physical position in base pairs) of the Morex contig, which is the target of the alignment, is retrieved from the IBSC map and finally reported. If DBs are queried independently (left track), all the results are kept, and the position of such contigs retrieved. Finally, as the redundancy filter cannot distinguish between actual different positions and erroneous differences, it reports a marker with multiple positions. Circled numbers are used to relate the different steps from a) and b) flowcharts.

Figure 2. Percentage of sequences found by either Blastn or GMAP, using the hierarchical method to align every dataset to barley sequence references.

Figure 3. 2D scatter plots comparing the RIL population map (X axis) against the IBSC and POPSEQ *in-silico* maps (Y axis). Positions of marker loci in cM. The positions of the IBSC genetic/physical map (grey crosses) and the POPSEQ map (black circles) were obtained using the hierarchical method of BARLEYMAP [Find markers].

Table 1. Genetic markers aligned by BARLEYMAP to barley sequence references, using the hierarchical search method. The proportion of matched queries with a single alignment hit is shown as well.

Marker sets	Markers	Aligned (%)	Single target (%)
DArTs	2,000	1,340 (67.0)	1,334 (99.6)
DArTseq PAVs	15,526	7,498 (48.3)	7,456 (99.4)
DArTseq SNPs	8,535	6,876 (80.6)	6,832 (99.4)
OWB SNPs	34,396	22,992 (66.8)	22,731 (98.9)
Infinium	7,864	7,304 (92.9)	7,291 (99.8)
Total	68,321	46,010 (67.3)	45,644 (99.2)

550

551 **Table 2.** Result of mapping all the 68,321 markers from Table 1 to the IBSC and
552 POPSEQ maps. For IBSC, results obtained by the independent and hierarchical search
553 strategies are shown.

Map / Search type	markers with map position	markers with single position	unique genetic positions
IBSC / Independent	38,528	22,891	5,675
POPSEQ / Morex assembly	30,330	30,232	2,721
IBSC / Hierarchical	34,203	34,063	6,908

554

555