Linkage illuminates a complex genome

John K McKay & Jan E Leach

High-resolution linkage analysis enabled by transcriptome sequencing brings order to the genome of a polyploid crop.

High-throughput sequencing technologies hold great promise for improving agricultural productivity through targeted genetic modifications to optimize desirable traits. But for many important crops, including bread wheat, cotton, oilseed rape, banana and sugarcane, the sheer size and complexity of the genome still present a formidable barrier to analysis. The combination of polyploidy, large gene families and repetitive sequences makes it difficult to identify the single-nucleotide polymorphisms (SNPs) that serve as markers for constructing genetic linkage maps—an essential first step in genomics-based crop improvement. In this issue, Bancroft *et al.*[1] show that high-throughput transcriptome sequencing can be used to generate high-resolution linkage maps and gene models for a tetraploid crop, oilseed rape (*Brassica napus*). Features such as allopolyploidy and large-scale gene duplication make *B. napus* an excellent case study for evaluating approaches to dissect complex genomes, suggesting that the authors' strategy will be useful for characterizing other large crop genomes.

Genetic linkage maps are used to identify the chromosomal positions of genes that control important agronomic traits and thereby allow crop breeders to track and select for alleles in segregating populations, a process called marker-assisted selection. Large-scale SNP discovery projects, driven by high-throughput sequencing, have greatly improved the linkage maps of diploid crops, such as rice, soybean and maize[2,3]. For these diploid crops, high-quality assembled genomes provided the reference for deducing SNP variation discovered by short-read high-throughput sequencing. In contrast, there are no complete genome sequences for most of the much larger polyploid crop genomes. A notable exception is the 844-MB autotetraploid potato genome, completion of which was enabled by the availability of a homozygous doubled-monoploid clone[4]. Besides the challenges posed by the sizes of polyploid genomes, the closely related chromosomes (homoeologs) of recently formed polyploid genomes share many sequences that are almost identical, making assignment of short-read sequences to the appropriate chromosomes difficult. To complicate matters further, most SNP variation in *B. napus* and other allopolyploids represents interhomoeolog polymorphism and is therefore not informative for creating linkage maps.

Bancroft *et al.*[1] tackled these issues by working with a doubled haploid mapping population of *B. napus*. *B. napus*, a major source of cooking oil, is an allotetraploid species derived from two diploid progenitors: *Brassica rapa* ($n = 10$) and *Brassica oleracea* ($n = 9$)[5]. For each chromosome in members of the doubled haploid mapping population, there are two homologs, which are identical, and two homoeologs, which are quite different, with the difference reflecting the time since divergence of *B. rapa* and *B. oleracea*[6].

Bancroft *et al.*[1] sequenced cDNA libraries prepared from leaf tissue collected from 37 doubled haploid lines, and identified transcript SNPs by aligning the transcript-derived sequences with unigenes from the reference genomes of the diploid progenitors *B. rapa* and *B. oleracea*. The success of their efforts depended largely on the fact that, unlike sequence assembly, linkage analysis is essentially unaffected by allopolyploidy and repeated sequences as long as homoeologous recombination is rare and genome-specific alleles can be identified. Bancroft *et al.*[1] found sufficient variation in many of the expressed genes to increase the number of SNP markers from *B. napus* to 23,037—an order of magnitude increase over the previous map for *B. napus*.

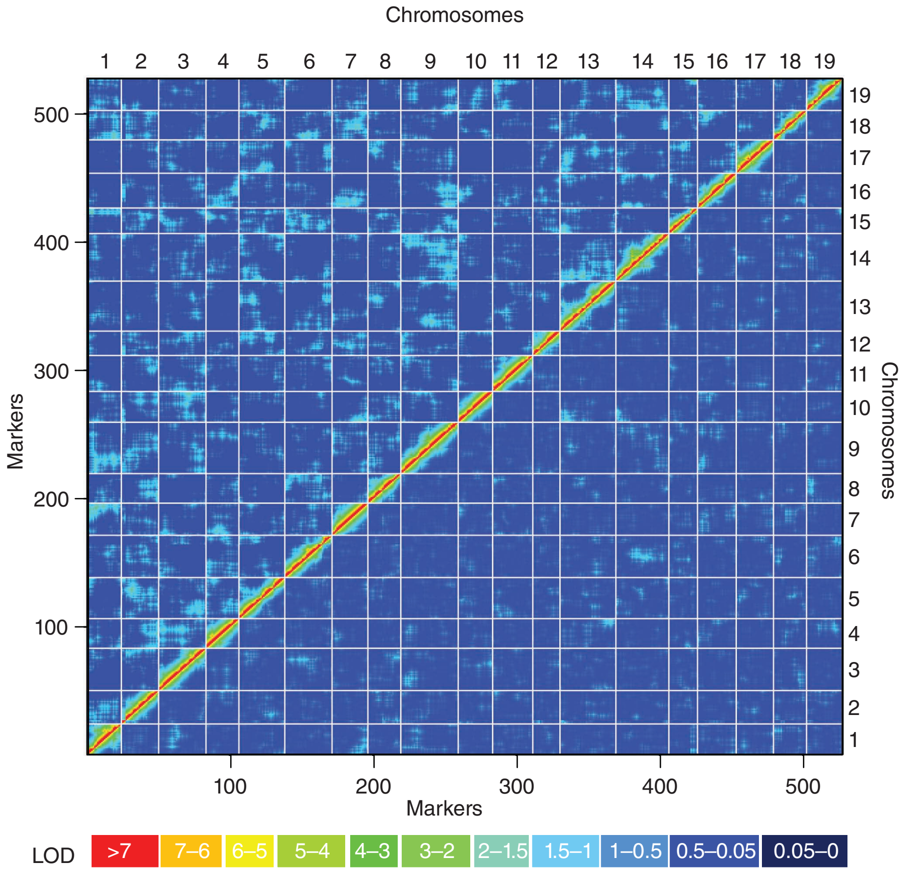

Figure 1 Graphical representation of the high-quality linkage map of *B. napus*. Bancroft *et al.*[1] aligned transcript sequences of *B. napus* with those of unigenes from its two diploid progenitors, *B. rapa* and *B. oleracea*, for which genome-wide sequence assemblies are available. Alignment of 527 marker bins along the 19 chromosomes of *B. napus* reveals pairwise recombination fractions (left triangle) and LOD (logarithm of odds) scores (lower right triangle) among all pairs of markers, plotted here using R/qtl[14]. Red, large LOD (small recombination fraction); dark blue, small LOD (high recombination fraction). The red along the diagonal and the lack of red off the diagonal indicate the high quality and strong statistical support of the map produced by Bancroft *et al.*[1].

John K. McKay and Jan E. Leach are at Colorado State University, Fort Collins, Colorado, USA.
e-mail: jkmckay@colostate.edu

These polymorphisms were aggregated into recombination bins, and linkage information from the mapping population was used to order these bins on each chromosome. **Figure 1** shows the excellent statistical support for the linkage map, which is especially notable insofar as it was obtained using so few lines.

The power of the authors' linkage mapping approach can be glimpsed in several additional results. First, they used the linkage map and allelic variation observed in the doubled haploid progenies, parents and progenitors to correct genome scaffold assemblies of the progenitor species *B. rapa* and *B. oleracea*. They were able to correct 32 errors in the previous assemblies of *B. rapa* and *B. oleracea* genomes, all in repetitive regions. Second, they identified a set of unigenes in *B. napus* with inconsistent or no sequence similarity with the progenitor scaffolds; these regions appear unique to *B. napus*. Third, mapping the SNPs onto seven ancestors of the *B. napus* doubled haploid population allowed them to predict the genomic regions that underlie the breeding history. Finally, they documented homoeologous recombination events, a first step toward evaluating the importance of these events to marker-assisted and genomic selection of germplasm during breeding programs.

Bancroft *et al.*[1] also corroborate the value of genetically tractable model plants for crop breeding. *Brassica* crops are logical candidates for applying the vast amount of genetic and genomic information on *Arabidopsis thaliana* ($n = 5$) because they share a common ancestor as recent as ~20 million years ago[7]. Strong conservation of genic space between *A. thaliana* and several *Brassica* species, revealed by DNA sequence comparisons[8], has long suggested that the Brassicaceae are ideally suited to showcase the potential of comparative genomics for crop improvement[9]. Taking advantage of the existing infrastructure for knowledge transfer from *A. thaliana* to *Brassica* crops, Bancroft *et al.*[1] aligned their dense *B. napus* linkage map to the genome of *A. thaliana*. This analysis confirmed tracts of synteny as well as chromosomal rearrangements between *A. thaliana* and each of *B. napus*, *B. rapa* and *B. oleracea*, which had been reported previously[9]. These similarities should facilitate the transfer of functional annotations of loci in *A. thaliana* to their homologous counterparts in *B. napa*. In other crops, similar efforts are already underway to sample the genomes of wild relatives to take advantage of the power of comparative genomics[2,8]. It seems likely that coupling transcriptome sequencing with linkage analysis may facilitate progress in these and subsequent efforts.

Although the approach described by Bancroft *et al.*[1] will certainly be invaluable for dissecting many other polyploid genomes, more complex genomes will likely offer fresh challenges. Particularly difficult will be genomes with a high content of repeats and transposons, such as wheat, or a high degree of heterozygosity, such as switchgrass and many trees. Some of this complexity can be avoided by targeting only transcribed sequences, but an obvious limitation of the approach is that it would miss the many important traits controlled by variation in noncoding regions[10]. Future efforts will benefit from newer sequencing-related technologies, such as barcoding, complexity reduction schemes[11], normalized and large-insert libraries and better assembly methods[12]. Current debate centers on the relative merits of *de novo* and reference-based assembly. Even the relatively simple genome of *A. thaliana* eludes *de novo* assembly, and current efforts to resequence and assemble 1,001 *A. thaliana* genomes are relying on reference-guided assembly[13]. Experimental designs like that of Bancroft *et al.*[1], which take advantage of linkage, are a promising solution to the assembly problem. Given the current rate of innovation in all areas relevant to high-throughput sequencing, we are confident that linkage-based methods will become increasingly affordable for all plant species. Genomic and molecular analyses of model and crop species seem poised to not only improve agricultural productivity but also provide fundamental insights into genome organization and adaptive evolution.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Bancroft, I. *et al. Nat. Biotechnol.* **29**, 762–766 (2011).
2. Jackson, S.A. *et al. New Phytol.* published online, doi: 10.1111/j.1469-8137.2011.03804.x (27 June 2011).
3. Feuillet, C. *et al. Trends Plant Sci.* **16**, 77–88 (2011).
4. The Potato Genome Sequencing Consortium. *Nature* **475**, 189–195 (2011).
5. U, N. *Jap. J. Bot.* **7**, 389–452 (1935).
6. Parkin, I.A., Sharpe, A.G., Keith, D.J. & Lydiate, D.J. *Genome* **38**, 1122–1131 (1995).
7. Lagercrantz, U. *Genetics* **150**, 1217–1228 (1998).
8. Cheung, F. *et al. Plant Cell* **21**, 1912–1928 (2009).
9. Schranz, M.E., Lysak, M.A. & Mitchell-Olds, T. *Trends Plant Sci.* **11**, 535–542 (2006).
10. Elshire, R.J. *et al. PLoS ONE* **6**, e19379 (2011).
11. Andolfatto, P. *et al. Genome Res.* **21**, 610–617 (2011).
12. Gnerre, S. *et al. Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
13. Schneeberger, K. *et al. Proc. Natl. Acad. Sci. USA* **108**, 10249–10254 (2011).
14. Broman, K.W. & Sen, S. *A Guide to QTL Mapping with R/qtl* (Springer, New York, 2009.)

First CHO genome

Florian M Wurm & David Hacker

An ancestor of the Chinese hamster ovary cell lines used for production of recombinant therapeutics has been sequenced.

Chinese hamster ovary (CHO) cells were originally chosen for commercial protein production because they were considered safe (as they do not propagate most human pathogenic viruses), allowed easy transfer of foreign DNA into their genome and grew rather quickly and robustly. They have since become the workhorse for industrial manufacture of recombinant therapeutic proteins and have even surpassed some microbial systems in productivity, delivering 3–10 g/l of high-value product from highly optimized processes. In this issue, Xu *et al.*[1] report the draft genome sequence of an ancestral cell line from which many CHO cell lines in industrial use today were derived. This work opens the door to a better understanding of these important immortalized cells, but the long history of study of the genetic instability of CHO cells suggests that many more lines will have to be sequenced before protein manufacturing can be improved through genomics.

Florian M. Wurm and David Hacker are at the Swiss Federal Institute of Technology Lausanne (EPFL), Faculty of Life Sciences and Faculty of Basic Sciences, Laboratory of Cellular Biotechnology, Lausanne, Switzerland. e-mail: florian.wurm@epfl.ch.

The 'CHO' designation encompasses a large number of very different cell lines. Substantial genetic heterogeneity has accumulated in these lines since the isolation of the first CHO cells in 1956 (**Fig. 1**), when Theodore Puck recovered a spontaneously immortalized population of fibroblast cells from the cultured ovarian cells of a partially inbred Chinese hamster[2]. There is reason to believe that CHO cells have a clonal origin (although this was not mentioned in the original paper), as the first CHO cells and all subsequently derived cell lines are deficient in proline synthesis. The cell line sequenced by Xu *et al.*[1], CHO-K1, was generated from the original CHO cell line by single-cell cloning in 1957.

The use of CHO cells in the biotech industry began after the isolation of cell lines harboring mutations in the dihydrofolate reductase (DHFR) gene, a genetic defect that facilitates