Multinational Brassica Genome Project (MBGP)

Steering committee meeting at PAG 2020 San Diego

Date: Sunday January 12th 2020
Time: 1.30 – 3pm
Location: Club Room, Handlery Hotel
Minutes (prepared by Jacqui Batley and Graham King)

Present: Graham King, Dave Edwards, Jacqui Batley, Rod Snowdon, Sebastien Faure, Ian Bancroft, Yong Pyo Lim, Isobel Parkin, Andy Sharpe, Philipp Bayer, Matthew Nelson, Chris Pires, Soodeh Tirnaz, Andy Yuan, Katie Greenham, Sarah Turner, Guy Kol (*Apology to those having participated and not appearing on the list*)

Chair: Graham King

1. **Introduction and welcome**

General introduction for participants.

2. **Approval of minutes from June 2019**

Participants unanimously approved the Minutes of the last MBGP meeting (International Rapeseed Congress, Berlin, June 2019)

3. **Reports on progress/notable milestones/announcements from members**

These are resources available to share, not private data.

Chris Pires: has *B. oleracea* and wild C genome sequences, along with rutabaga, Siberian kale and wild B. rapa.

Katie Greenham: the R500 pacbio genome assembly is now available on CoGe. There are 6 *B. rapa* morphotypes that have been sequenced using Pacbio. Transcriptome data is starting to become available from a cold stress response time-course.

Sebastien Faure: Is undertaking Oxford nanopore and HiC for *B. napus* Aviso – this is ongoing.

Rod Snowdon: The *B. napus* Express assembly, using nanopore and Pacbio data, is finished. The paper is submitted and the genome assembly can be made available. This will be submitted to NCBI.

Dave Edwards: The A, C and AC pangenomes and comparisons are complete. The manuscript is in a draft stage and he is happy to share it and collaborate if anyone would like to contribute. Has a new project on deep learning/ machine learning and needs datasets (image or genome/phenome) and would like to collaborate if anyone has datasets available.

Graham King: SCU has funding for *B. juncea* re-sequencing, including lines from Canada. Tracking flowering time and stature genes – interested if anybody has specific leads in *B. juncea*? *B. rapa* R-o-18 genome is finished, currently working through Ian's pipeline. This will be published and will be in databases/browsers including Ensembl. SCU team also working on seed protein proteomics and interested in collaborating if anyone else working in same area.

Jacqui Batley: has funding to look at novel sources of blackleg resistance, including from wild Brassicaceae. Has just re-sequenced the commercial Australian cultivars (~100), data will be available. This is just Illumina sequencing to around 15x. In future the Brassicaceae and introgression lines will be sequenced, all data will be available. Has a second project aimed at clarifying the blackleg resistance gene nomenclature. This is a collaborative project, initially with Canada, but the aim is to expand. As part of this a set of standard isolates and plant lines are being assembled for distribution and the international blackleg collection is being updated. If anyone has isolates they would like to contribute please get in contact, these will be sequenced and must be available for distribution.

Isobel Parkin: The *B. napus* pan genome with NRGene is underway. There are 10 genomes to be sequenced, 7 have been completed to date. If anyone is interested in accessing or contributing, please contact Isobel or NRGene. Has also finished a nanopore assembly of a NAM founder line, along with nanopore re-sequencing of the all NAM founder lines (~50), will be looking at SV. Has sequenced another 4 species of *Camelina*.

Sarah Turner: has re-sequenced 172 accessions of *B. oleracea*/wild C genome lines. This data will be available following publication (are at 10-20X, coverage higher in the wild species). There is also phenotypic data available (leaf shape)

Yong Pyo Lim: Undertaking GWAS in *B. rapa*. Has sequenced 132 homozygous lines, each to 70X coverage. Has 72X data for Chiifu. Developed 64K SNPs, which are being used for glucosinolate research. Has a functional study to identify candidate genes for *Fusarium* rot, club root, glucosinolate content and anthocyanins. Once published the data will be public. Has collected hundreds of *B. napus* inbred lines, screening for clubroot resistance, has a QTL, is now developing markers. Also working on blackrot in *B. oleracea*. This is a joint program with the EU2020 "Organic breeding" program, a 4 year project led by collaborators in Italy and involving 17 countries, aim is to develop diagnostics using different races.

Yong Pyo Lim also presented an update on behalf of other Korean Researchers who could not attend. The group from the National Institute of Agriculture Sciences, RDA has produced a new *B. rapa* Chiifu assembly from ~81.7-fold coverage of Pacbio Sequel sequencing reads (37.03Gb). The resulting assembly consists of 510 contigs, with a contig N50 of 3.9 Mb and a total length of 380.7 Mb. They have also produced a *B. juncea* assembly from Dolsangot16-1_DH line. This was from ~106.8-fold coverage of Pacbio Sequel reads (99.3 Gb). The resulting assembly consisted of 945 contigs, with a contig N50 of 3.63 Mb and a total length of 882.4Mb (98.3% of BUSCOv3.0). There was also 10X Genomics and Dovetail Hi-C added, covering 90.8% (801.3 Mb) of the sum length of 18 scaffolds.

Ian Bancroft: undertaking QC on various genome assemblies. Has also developed re-arrangements from Chiifu and *B. oleracea* TO1000 – this has formed the basis for the pan genomes (covered in more detail in next item on agenda)

Guy Kol: NRGene is part of the pan-genome project with Isobel. Graham King asked if there would be other Brassica pangenomes (eg *B. juncea*). Guy noted that there may be additional *B. napus* added, but there were no other species planned.

Andy Sharpe: had Genome Canada proposal on *B. napus* seed protein which unfortunately didn't get funded, however some of the matching funds are still available, so the project will move forward in some form.

4. Diploid Pan genome harmonisation and nomenclature

Previously in MBGP steering committee meetings there was agreement to establish a consensus *Brassica* gene-model naming system based on the diploid-pan-genomes (http://brassica.info/tools/data_standards.html). Ian Bancroft, Dave Edwards and Graham King had been moving this forward during 2019. Graham circulated a global pan genome progress report from Ian Bancroft in December to relevant members that had contributed genomes.

Ian has developed the optimal organisation for the diploid A, B and C genomes. Through careful mapping against the polyploids and GOGG (genome ordered graphical genotype) analysis they were able to identify chimeric scaffolds. The pipeline has been tested and shown to work well, including with newer assemblies. The HDEM and Z1 genomes matched perfectly and additional scaffolds were added. When a mis-assembly was corrected for 1 linkage map, other maps were also corrected. They are confident that there is the same basic genome organisation for the A, B and C genomes, irrespective of the species; there are just deviations in individual lines. There are currently ~190K gene models across A, B, C genomes, of which ~104K are high confidence (some are non-coding).

As previously agreed, the aim is to adopt a standard nomenclature for gene names, following the system in Arabidopsis, with the order of gene names as go through the genome and appended at the end with the name of the species and cultivar, eg C01p010030.1_BnaDAR (http://brassica.info/tools/data_standards.html. There also needs to be sufficient flexibility within the naming system to allow for insertion of additional gene models as they are identified. Graham King noted that at present there are 9 intervening spaces between the genes where new ones can be inserted and questioned whether 10 is a big enough space? Dave Edwards noted that it will depend whether the missing genes are random or a chunk.

Graham King asked how close we were in having telomere to telomere coverage. Ian noted that there are scaffolds near the telomeres but not for all chromosomes.

Dave Edwards noted that he has provided pan genome content to Ian. Ian said they need to discuss how to incorporate those genes not currently present. At present for many of the extra ones we do not know the positions at present. Graham King suggested that we check the underlying evidence. Ian Bancroft suggested that we look at co-linearity and flanking sequences. If they are co-linear they can be put in.

Ian also asked that if anyone notices that the current genome annotation is missing genes known to exist in those genotypes, please let him know and he will re-instate them (see Action below). Specifically blocks of genes missing are important. Graham asked if there was an automated system to put the genes in, Ian stated that there needs to be a three letter genome extension and the position where it needs to be placed (for pan genomes this should be PAA, PAB etc so know if comes from the pan genome and from which species). Dave Edwards and Philipp Bayer noted that this was compatible with their DaisyChain software.

Ian Bancroft has also applied for funding with EBI to develop a comparative browser, if funded it should be available quickly in public domain. Graham King stated that even if the grant is not funded, individual genomes should still go into a browser.

Ian Bancroft will be writing a publication describing this and anyone who has contributed to the effort will be a co-author (aim for resource in Plant Cell).

Graham King asked if the meeting accepted this approach and are we happy as a community to move forward and fix this as common nomenclature, which will be retrospectively assigned to existing genomes and adopted for all new genomes. All agreed.

ACTION: Graham King proposed a deadline of mid March for the review/checking process, working firstly with people who contributed the available genomes.

ACTION: Dave Edwards to send pan A genome to Ian Bancroft.

Following the initial checks in March, the system will then be adopted, small changes can still be performed at this stage if necessary. Graham noted that this was a substantial move forward and a landmark for the Brassica community.

5. Brassica Information System

Graham King had asked during 2019 which datasets are available and in what form. Many of those available fell into the "other" category, for example supplementary data, but also as available spreadsheets and other files. At present there were few online databases that could be indexed, but there are resources eg at BRAD (China) and BIP (UK, ported to Australia). Graham has an inventory of this data and could host datasets on www.brassica.info server once have sign-off for Creative Commons (eg CC BY 4.0) license from contributors. His team will work with Dave Edwards' team for serving the setting up indexing (see below).

Dave Edwards attended the Wheat Information System meeting and there is now a revised model for indexing. Previously there was a limitation of requiring an Adobe SOLR index of databases, which had been implemented for wheat. The new version from INRAE/URGI can now take Excel, mySQL etc and they are happy to extend the wheat model to other species (e.g Brassicas). They are also expanding the search capabilities to link back to the database or repository. This is an easy way forward. Graham will work with Dave and aim to get initial datasets through the system.

Graham King asked where the files will be hosted. Will it be somewhere central (e.g. brassica.info). Dave Edwards noted that DOIs can be created for documents so that they can persist and will have version control. Graham also asked if there was a template for metadata? The issue with this is the more requirements there are, the less data will become available if it doesn't fit the template. Dave Edwards noted that if the file is supplementary data, it can just link back to the original source file from the journal.

Graham King stated that the Brassica Information portal (BIP) has now been ported to SCU and is able to be run independently in a 'container' which simplifies complex software library dependencies etc. His group are still in contact with the original developers. If anyone would like to host it on another machine, Graham can arrange to provide a containerised version. This will also be available on GitHub during 2020.

6. Scope for multinational collaborative projects

Opportunities for co-funding and international funding were discussed (updates from June).

Matthew Nelson: has a bilateral collaboration with RRes, this involves reciprocal visits with Pete Eastmond.

Jacqui Batley**:** as mentioned in earlier updates Australia/Canada funding for blackleg research: developing a set of standardised blackleg isolates and hosts (*B. napus* plants); understand diversity

and make a set of standards, link back to sequences. Hoping to link out to other countries in the future. All isolates, seeds and sequence will be available

7. **Any other items/meeting announcements**

Brassica 2020: 27-30 September in Saskatoon. Website is now up and running. (http://cruciferseq.ca/Brassica2020)

Action: Isobel to send an announcement to the Brassica.info mailing list

Yong Pyo Lim: there was an ISHS brassica meeting planned in India this year, but he will suggest that this moves to 2021 so there is no overlap with Brassica2020.

8. **New chair:**

The committee thanked Graham for his excellent work as chair, and noted that he had forged substantial progress for the community. As agreed in 2019, Jacqui Batley will become the new chair.