Multinational Brassica Genome Project (MBGP) Steering Committee Meeting

held in Melbourne, Australia on Thursday October 7, 2016
{as a side-meeting of the *Brassica 2016* conference)

Present: Rod Snowdon, Dave Edwards, Shengyi Liu, Joerg Schondelmaier, Martin Ganal, Jacqui Batley ,Bhavna Hurgobin, Yuxuan Yuan, Armin Scheben, Regine Delourme, Anita Severn-Ellis, Philipp Bayer, Thierry Rouxel, Harsh Raman, Shunxue Tang, Jun Zou, Graham King, Abdul Baten, Amine Abbadi, Guusje Bonnema, Venki Pegadaraju, Luis cel Rio, Katy Navabi, Theo Borm, Annaliese Mason

Apologies: Boulos Chalhoub, who is present MBGP chair but passes on his apologies that he was unable to attend due to visa problems)

Chair: Rod Snowdon (on behalf of Boulos Chalhoub)

Minutes: Annaliese Mason / Corrections: Rod Snowdon, Graham King, Dave Edwards, Jacqui Batley, Martin Ganal, Shengyi Liu

1) Introduction (Rod)

General introduction for new participants, background of MBGP history, aims and mission:

- MBGP conceived as an organ for interested researchers to come together and share information and see what is happening in different parts of the world, plan for future work etc.
- *B. rapa* genome sequence was the first "project".
- Coordinated completion/analysis of other genomes, as well as adoption of chromosome and functional gene nomenclature, UniGene EST set leading to establishment of Affymetrix GeneChip and subsequent Illumina 60k and 90k arrays etc have been successful outcomes.
- Now all genomes are sequenced, it is possible to sequence easily and do more genomes at low cost in a short time, and many new Brassica reference genomes, resequenced genomes and other "omics" data are being generated worldwide.
- Now the main question is what should the MBGP do? In recent years the meeting has become more of an information session (i.e. "who has done what during the past year") rather than a "project", and we should think about what the international Brassica research community really needs most and come with ideas as a community for how we can best achieve these needs.
- Main question to answer in meeting: What (multinational) initiatives should we go for in order to generate new resources, pool information and efforts to achieve added value from our ongoing research activities.

2) Approval of minutes from previous meeting. Approved

3) New international projects: None of the participants reported new major international collaboration projects initiated since the January (PAG) meeting.

Rod reiterated that the mandate of the group should be to **develop new international cooperation** in areas where the community sees a deficit. Many ongoing activities were listed in the last minutes from Jan 2016, including information on ongoing sequencing projects.

The group who just published *B. juncea* and *B. nigra* genome sequences (Nat Genet August 2016) apologised by email for not being able to attend the meeting but requested that this publication be publicised in the meeting.

Shengyi Liu – *B. rapa* new genome version is in progress (but not discussed with Xiaowu Wang for this information release), integrated with PacBio data, Xiaowu also planned update of *B. oleracea* genome sequence. Will finish the Chinese *B. napus* variety genome (Zhongshuang 11) soon.

Guusje – would like reference-like genomes for different morphotypes. Two projects underway but breeding companies involved. Aim to genotype 1000 *B. oleracea* accessions, wild and cultivars, select subset for resequencing (160 or so). Project proposal under review; if successful data will be embargoed until 6-9 months after end of project. Using GBS for SNP genotyping

Rod – Recent paper in Front Plant Sci from Mike Gore's group at Cornell, GBS with 782 *B. napus* accessions from USDA collection– Sequence data have been deposited at NCBI under BioProject PRJNA298631 (see http://journal.frontiersin.org/article/10.3389/fpls.2016.00525/full)

Guusje – radish resequencing (17 genotypes) apparently completed (see https://www.ncbi.nlm.nih.gov/pubmed/27377547)

Shengyi Liu – already resequenced 700 *B. napus* cultivars at 7 – 20x coverage, selected 150 for in-depth sequencing for pangenomic analysis (20 – 30 initial selection for sequencing at >17x). Data will be made public in 3 stages, beginning with a GWAS paper to be published soon. Recombination data available end of 2016, in-depth sequencing of the 150 lines aim to release data by middle of 2017

Rod – recent release of the ABC 90K Illumina SNP chip described at Brassica 2016 conference in Illumina presentation, release of new "optimised" 15k SNP Illumina chip by Trait Genetics (Germany) described in presentation from Martin Ganal.

Update on brassica.info website

Graham King: Still exists: it is very dependent on people sending Graham information. Aware some things are out of date, happy to link to other information. Used to host data from the UK via CropStoreDB – replaced by pointing to the improved Brassica Information Portal. Maintaining information about standards and nomenclature where these exist. New issues of Cruciferae Newsletter published annually. Any ideas for improvement please contact Graham

4) Progress on a "*Brassica* informatics platform"

Rod – something discussed at last meeting, had a proposal underway to try and get funding in Germany to link high-density genetic maps/QTL to diverse Brassica genomes and SNP polymorphisms. Idea was an alignment/browser tool to navigate from genetic maps to *Brassica* genomes, consensus QTL, associated genes and their polymorphisms/expression data, etc. This

proposal had good support from academics and breeders in various countries worldwide, but was unfortunately rejected by the funding agency so Rod is currently exploring other possibilities to realise this idea. At the last MBGP meeting the URGI crop informatics platform (e.g. "wheat information system" was presented by French participants, as this was already being implemented for *Brassica* napus (however with proprietary data from the RAPSODYN project that may not be released until 2018/2019). However the database system, metadata standards, expertise and tools were already there, so it was agreed to be a good idea to link everything to that platform so that when the French data is available the data can be integrated.

Graham King – wheat database is very difficult to navigate, hard to find information. It appears there have been personnel changes at URGI in recent months and the schema is in flux (see comment from Dave below).

Dave Edwards – Wheat Information System is is still in early stages so not much information in yet, main philosophy is to look at databases and datasets out there and index through a central portal; just aim to capture data first before imposing standards and encouraging community to follow standards. Still early days but appears successful. Rice people following same model, seems to be being adopted. Very supportive of Brassica community also adopting this model. Wheat community is still in the process of trying to get funding, but should be able to index databases that are available. Aim is not to make a massive database but to provide a portal which allows navigation/linking out to different resources, at least providing a sustainable long-term system to capture information and then have people develop systems to use it.

The pragmatic way forward would seem to be:
a) create inventory
b) create data warehouse of datasets available
c) then see to what extent can organise/formalise, but (b) is the key step.

Jacqui – is the idea to extract material from e.g. published supplementary information, to provide information or to link to information elsewhere? A lot of genotyping data is hidden in supplementaries and not easily available nor usable.

Dave – aiming to capture all this information and provide a link for that data. Currently no way of comparing that data to other data though, but at least knowing what information is available will be a good start to adopting/suggesting metadata standards for collation of compatible datasets.

Rod – The real aim should be to eventually develop standards to allow direct data comparison, because otherwise there is no value adding to own data. One idea is to find databases that are publically available and try and integrate. **This should be on the agenda as one of the most critical projects for the future**!

Theo – Need to try and encourage as many people in the community as possible people to use standard formats for their open-source data – to achieve this it is critical to adopt robust metadata standards.

Dave – there are multiple standards, which to go for? Choice was either to have no data submission or just accept multiple standards to start with – need to see it as being useful first. SDA now loosening view because no one has time to change data, need to build up awareness first.

Joerg – genomics or trait association? Sequence or phenotypes?

Shengyi Liu – metabolomics project starting, if anyone is interested then we can get together, we will probably do limited compounds so collaboration would be good to expand range. Primary and secondary metabolites, glucosinolates

Guusje – needs a lot of money to produce data, and there are specialist research groups, so joint metabolomics is probably not so feasible

[Rod: Metabolites and traits should probably be discussed and treated separately, the first priority should be to standardise all the available data which has been collected using standardised methods, e.g. Illumina SNP arrays or sequencing data]

Rod – many people here have large data sets, either based on the 60K array or GBS data etc., QTL and trait data that can be annotated to genomes. We've never had any way of comparing between open-source population data – we could use this to compare between QTLs in different environments, genome polymorphisms etc.

The following people confirmed that they have substantial, open-source data (especially from genotyping with the 60k SNP array, GBS or other sequence-annotated data) that they would like to share (no strings attached) in order to gain added value from putting all public data together: Rod Snowdon, Harsh Raman, Jun Zou, Graham King, Guusje Bonnema, Jacqui Batley, Regine Delourme, Dave Edwards.

ACTION: Rod and Dave will communicate with these people to find out what the most common forms/formats of open-source data are, these should be the starting point for developing metadata standards to which other existing datasets could be adapted without too much effort. Armin (UWA) could be involved in doing this. Ideally we should aim to achieve this by the end of 2016 so that we can reach out to others at the MBGP meeting at PAG in January and encourage other researchers to also add their data to a Brassica Information System.

Dave suggested to use the name "Brassica IS" as analog to Wheat IS, Rice IS, etc.

Theo - Are genotypes going to be available to underlie sequence data? Guusje – can encode genotypes, mainly just want the information without necessarily the genotypes (if proprietary)

- 60K data should at least be able to be integrated based on flanking sequences mapped to reference genome(s)
- Dave Edwards – standard markers, should be very easy.

Martin Ganal – has a lot of SNP data from own activities and many groups which could be made available with permission from these groups in a standardized way., but there are different ways of reading out data from the 60K array and other service providers do might do this differently. This could create problems with data integration. All SNP array data originating from TraitGenetics are standardised, and it may be possible to ask others genotyping groups with the same array to adopt

the same formats and use the same allele calling (cluster file), however this isn't so easy for existing published datasets. It certainly helps to at least always have flanking sequences associated with their SNP calls.

Dave – first stage should be to communicate with the groups who have data and use the most common format. It may also be valuable to develop a central repository and then encourage journals to adopt this as the "accepted" repository for uploading of published data. With regard to a standardised format for 60K SNP data, the pragmatic approach is the "the majority will win", i.e. just collect data first then decide on standards based on what is most commonly used.

Martin Ganal – species and type would be essential information for such a database.

Dave – minimum would be contact details to get more information from contributors.

Venki - Could we use one of the existing resources like array express? Use this as a primary resource and link to brassica.info etc.?

Dave – first step to find out what's out there, data structure should be relatively straightforward since it's a standard array.

Rod – Besides just SNP data, it would be very useful to get maps from all public mapping populations along with their sequence-annotated markers in form of linkage maps, with SNP calling matrices associated if available.

ACTION: Rod and Regine will talk with the people developing the French platform about potential for collaboration and implementation of international open-source data.

5) Gene model nomenclature

Graham briefly presented the history of the discussion about standardising gene model nomenclature across different Brassica species. Before the release of the *B. napus* genome the MBGP agreed that it would be extremely helpful to assign standardised gene names for future releases. However so far *B. napus* Darmor-bzh has been the only one that publically adopted the nomenclature. B. oleracea TO1000 was released just a week or two later and it was too late to modify the nomenclature in Ensembl to adopt this standard system. Now it is extremely difficult or impossible (especially for people not familiar with Brassicas) to find corresponding gene homologs between Bn, Br and Br (and others).

Graham – one idea is an index table, which we are happy to host on www.brassica.info. This could be done internally, but would need validation (i.e. lead investigators responsible for each genome) to make it the main reference point. A registry of gene models could then be available?

Dave –need to compare between species, really challenging to identify orthologs or paralogs. A perhaps better idea (which enables present annotations to retain present systems without so much confusion) would be a graphical rather than a tabular reference, using an automatic method based on physical position and sequence information.

Graham – gene model system in Arabidopsis (e.g. At1g69120) has been very robust over past 16 years, but we may also need a DOI per gene?

Dave – we need to use just physical location and sequence information, because we are never going to be absolutely sure if we have the exact ortholog

Shengyi: It is a good idea to assign DOI, but it is better to assign these after the main brassica genomes have been updated, because our current assemblies miss too many genes and there are too many positions in the genomes for which we do not yet know the gap sizes. Further, small RNAs (including microRNA) are also genes, should we consider to name them consistently.

Graham – Agreed, but need to standardise the labels of gene models in reference genomes to link to this information, to avoid duplication of gene model names (homonym) within the assemblies etc.

Rod/Dave – suggest to start with the *Brassica* pangenome and use that as the reference, or vice versa (the pangenome would come out of a species-wide Brassica annotation)

Graham – we can retrospectively apply gene model names to the reference genomes?

Graham – Agreed. Arabidopsis has been very stable with its gene model nomenclature – adopted nomenclature a long time ago, we should be able to do it retrospectively even though we didn't do this early enough. The protocol for this is (implemented by Boulos Chalhoub *for B.napus* Darmor-*bzh* is at http://www.brassica.info/info/genome_annotation.php)

Annaliese – can we use this in future for the new references coming out?

Graham – can we vote on this?

The chair requested that we ask to have the standardised gene nomenclature system (as previously agreed by the MBGP) communicated to the people updating existing Brassica genome references, urging them to adopt the standardised nomenclature to enable better cross-species comparability –

The meeting agreed to this without dissent (i.e. vote passed)

– ACTION: *Dave Edwards and Graham King will communicate the above with people responsible for updating existing Brassica reference genomes*

Shengyi Liu noted that previous versions might be updated without too much trouble as new, improved references are released.

Dave – second action should be the development of a graphical database to integrate genomes and identify orthologs/homeologs ***– ACTION:*** *Dave Edwards will investigate potential possibilities for graphical database for cross-species gene annotation/ indexing*

6) Upcoming meetings

- During the Brassica 2016 conference the following upcoming Brassica meetings were announced:

1) International Rapeseed Congress, Berlin, September 2019 (chair: Wolfgang Friedt)
2) 20th Crucifer Genetics Workshop – "Brassica 2018" – in St. Malo, France: Organisation Regine Delourme & colleagues
3) 4th Glucosinolates Summit, organisation Guusje Bonnema and colleagues
4) A "Brassica 2017" meeting (ISHS?) was also announced at Brassica 2016, to be held in north-western Spain (organisation Fernando Branco?). Seems to be somewhat unclear how this meeting fits with the 4-yearly "Brassica" conferences or the biennial Crucifer Genetics Workshops.

Also two upcoming "non-Brassica" meetings that might be of interest for MBGP participants:

- September 3-7, 2017: 4th International Symposium on Genomics of Plant Genetic Resources, Giessen, Germany (conference chair: Rod Snowdon)
- Agriculture and Climate Change meeting, March 26-28, 2017: Sitges, Barcelona. Speaking slots still available if anyone is interested in presenting (contact Dave Edwards, conference chair)

7) Epigenomes?

- Rod/Amine – just got funding for whole-genome bisulfite resequencing (along with corresponding genome, siRNA, mRNA) in 100 Bn genotypes. Still to be decided when/how data will be published, but would at least want to publish (pangenome) reference positions of epigenetic marks even If genotype information remains proprietary.
- Venki - Differentially regulated genes as a means for selection now under emphasis in the animal community
- 60K SNP format – majority will win, will just collect data first then decide on standards

8) Closing remarks

Rod – Thanks all for very constructive input and discussions. **Conclusion from this meeting** is that the MBGP ***is*** a "project" after all, and has very important initiatives which need to be realised for the greater good of the entire international community. This can only be achieved by re-intensification of the level of collaboration and communication within the international research community, development and adoption of common standards and compatibility/sharing of open-source data.

Chair especially thanks Jacqui for organising Brassica 2016; the next MBGP-SC meeting will be as usual in San Diego in January, most likely same time and place as last year: PAG Sunday from 1:30 pm until 3 pm in Dover Room.

Meeting closed 7:50pm