Multinational Brassica Genome Project (MBGP)

Steering committee meeting at PAG 2019 San Diego

Date: Sunday January 13th 2019
Time: 1.30 – 3pm
Location: Garden Salon 2
Minutes (prepared by Jacqui Batley and Graham King)

Present: Graham King, Dave Edwards, Jacqui Batley, Rod Snowdon, Lenka Havlickova, Ian Bancroft, Yong Pyo Lim, Isobel Parkin, Jenny Lee, Liang Song, Anne La Perche, Hong An, Armin Scheben, Shannon Greer, Seongmin Hong, Lixi Jiang, Philipp Bayer, Judith Irwin, Chris Pires, Mike Barker, Fangning Zhang, Andy Yuan, Katie Greenham, Guy Naamati, Ronan O'Malley, Zachary Stansell, Yuan Yuxiang, Jean-Marc Aury, Lawrence Bramham, Bruno Contreras-Moreira (*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 July 2018

Participants unanimously approved the Minutes of the last MBGP meeting (Brassica 2018, St Malo, July 2018)

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

Graham King suggested only presenting extra activities/resources beyond what was presented at the meeting in July. These are resources available to share, not private data.

Genoscope: Have new *Brassica* assemblies. They are also sequencing a *Brassica napus* genotype– a resynthesised hybrid. This genome sequence will probably be available before mid-2019.

Katie Greenham: Has recently been awarded a JGI Science program grant. Within this they will generate a *B. rapa* pan genome, using 6 different representatives from different crop types. This will be using PacBio, but she will also do re-sequencing for other lines. They anticipate having assemblies by the end of the year. She is interested in abiotic stress responses, they have already perfomed a cold stress response time-course, every 4 hours for RNASeq analysis, also glucosinolate profiling. It will take a bit longer for the RNASeq data to become available. She has also completed the *B. rapa* R500 PacBio genome assembly, currently this is at the stage of being written up, but access is available.

- Graham King asked whether the above will go into NCBI as a genome assembly, and offered to assist with this if required. It is important as this is a requirement prior to the genome being included in Ensembl Plants. Other genomes will go into Phytozome.

Chris Pires: along with Mike Barker he has generated pan transcriptomes for *B. oleracea, B. rapa* and *B. napus*. There are also wild C genome relatives that they want to sequence. In the *B. napus* pan genome they are interested in rutabaga (introgressing R-genes from rutabaga). He also has a really nice genome from *Crambe abyssinica*.

Ian Bancroft: has a further 30 *B. napus* DFFS transcriptome sequences, these will be in SRA.

Philipp Bayer: Described Daisychain (which will be presented during conference), this links gene ids in different annotations. http://daisychain.appliedbioinformatics.com.au/
Cropsnpdb has also been developed, this is a repository for SNP array data for all Brassica species. Can upload the SNP data yourself, or send to Philipp and he can put in.
http://appliedbioinformatics.com.au/index.php/SNParrayDB

Rod Snowdon: Jenny and Hamid are finishing off *B. napus* Express assembly, using nanopore and PacBio data. They are also sequencing a synthetic *B. napus* with Nanopore, which will be made public. Express is currently in scaffolds, and they are making into pseudomolecules. They don't have sufficient RNASeq from Express – which is needed for annotation.

- GK asked if the Express assembly will go into NCBI as a genome – RS said it will not go in until it is in pseudochromosome form, which is probably a year away. Scaffolds can be provided in the meantime.

Dave Edwards: The NRGene Darmor assembly is almost done (collaboration with Boulos, Rod, Jacqui), we think most errors are fixed, although there are still some regions with minor question marks. If anyone wants to help fix, we are happy to share the assembly. This was used as the foundation for the *B. napus* pan genome. Currently he is comparing with *B. oleracea* and *B.rapa* pan genomes, but want this to be community effort, happy to share assemblies, draft manuscript etc. The aim is to submit in around 6 months. Also described the Brassica information system, where the follow up is similar to wheat. There was a WIS meeting which showed that WIS is being used as a model for other species. WIS wants to have feedback from communities to share what works, what doesn't etc. Anyone is welcome to attend these WIS meetings

Jacqui Batley: Continuing to work on R gene identification from different species/annotations. Found a problem with repeat masking of R genes within genomes (Bayer et al. 2018. Nature Plants 4:762-765)

Graham King: We have demonstrated the value of high quality proteomic data (LC-MS/MS) for allocating proteins expressed in eg seed aligned to specific genomic loci/alleles. AU$30 a sample.

Lixi Jiang: Has undertaken resequencing of a collection of *B. napus* rapeseed germplasm - 991 accessions. These data have been mapped to the *Darmor-bzh* reference. This has just been published in Molecular Plant. Rod has written a spotlight article for this. Drawback is the low depth. They have selected a small core collection of germplasm that can represent the whole collection, and are doing deep re-seq (>30x) of these lines, so they can compare with different reference genomes of different morphotypes. They would like a spring type reference genome. He also has Ningyou7 (semi winter), and ZS11 (semi winter). They would like to collaborate to do the analysis of SV – PAV and CNV.

Yong Pyo Lim: Has collected several hundred cruciferae from different species from Korea. He has sequenced a core collection of lines (117) from *B. rapa* >20x using Novoseq. Has >24 morphological traits and >17 chemical component traits and is now doing GWAS. In radish he has sequenced 6 inbred lines to 30x and is currently doing the analysis. They are also doing RNASeq – they have tissue, including specific targets of disease resistance – fusarium wilt and clubroot. For nutrition genomics they aim to have better glucosinolate content. He has also developed a red Chinese cabbage, this has different anthocyanins, they are looking to find the underlying genes. They have been undertaking genome editing of *B. rapa*, for the FLC gene, this data is almost ready to submit. Through a collaboration with European countries – Horizon 2020 – he is studying blackrot resistance in *oleracea*, this involves 1000 accessions, looking for good resistance.

Seongmin Hong: Bioinformatics analysis, studying 200 different *B. rapa* lines, 150 with phenotype data and doing GWAS with these, as well as Structure analysis and phylogenetics. Phenotypes are quantitative and qualitative. Trying to compare with previous results with aim of developing functional markers.

Armin Scheben: For Cropsnpdb (mentioned by Philipp Bayer) they are accepting genotype data – this database has been accepted for publication in Plant Journal.

4. Reference Genome Assemblies

GK: There is an opportunity to revise which *Brassica* genomes appear in Ensembl Plants as references, along with processing through the Compara pipeline. At the moment the *B. rapa* assembly (Chiifu-401 IVFCAASv1) is from 2011, and later versions are not present. Original versions are also only available for *B. oleracea* (TO1000 v2.1 BOL, INSDC Assembly GCA_000695525.1, May 2014) and *B. napus* (Darmor-*bzh* sequence_level 2, AST_PRJEB5043_v1, INSDC Assembly GCA_000751015.1, Sep 2014).

There has been a change of staff at EBI, and Bruno Contreras-Moreira is happy to work with us a community, so that the MBGP can prioritise what goes in and when. Graham will coordinate providing a schedule to EBI in consultation with genome consortia. There is a pre-condition that genome assemblies are registered within NCBI/ENA and gff annotation files are available. They are processed through a pipeline and compared with other plant genomes (Compara), and this then becomes fixed. The system generates explicit information on introns etc, which you can't always get from genbank, which has some advantages. Two genomes from France have been registered in NCBI/ENA, and hopefully the NRGene assembly will be soon. The *B. rapa* R-o-18 (now with anchored pseudochromosomes) effort led by Graham will also be registered as an assembly, and some additional functional annotation added.

There is also a real benefit in ensuring all submitted assembles are processed through Ian Bancroft's quality control (QC) process (generating graphical genotypes, ordering through genome sequence assembly – He & Bancroft, 2018, Nature Genetics 50:1496–7), to identify any systematic or local errors, which can then be corrected prior to registering as a later version. This can be used to check that the assembly fits with the genetic map, and can be used to ratify the quality and integrity of an assembly to a consistent criterion. Ian is happy to work with anyone on this. He also now has a *B. carinata* map for doing B genome assembly validation.

Ensembl Plants do not discard old assemblies, so they will always be available in an archive.

IP will submit her first *B. nigra* genome (Ian already done the QC)

Dave Edwards: The PAV from the pan genome is within gff3 as a pie symbol, which shows how frequent gene is among range of lines. This is useful to have – can it be hosted?

- It was confirmed that Ensembl will look at it, and if not can link out or in from specific genes.
- External references -can be managed in Ensembl – just add into database as xref, so perhaps this could point to existing JBrowse instances.

JB: highlighted that repeat masking is a problem. GK asked whether we could insert a warning in pre-ambles on Ensembl Plant pages for some of these genomes, so that future PhD students and others are aware.

5. Brassica Information System (BIS)

GK had generated and circulated a questionnaire (based on WIS). There were 39 respondents, of which one was from an industry representative. This was used to identify areas of interest (could vote more than once). All information is presented in the attached document. For curation, there were fairly consistent scores for the first questions. Quite a few people would like an integrated system, which is a reasonable objective, but won't happen overnight. It was encouraging that there were significant number of responses indicating research groups had a range data to contribute to BIS. The information provided help identify who has data in particular areas, and so they will be approached for contributions. As expected there were a greater number of potential users that providers.

There was considerable interest in access to services (eg mutation, transformation). It was also encouraging to see buy-in in terms of standardisation of naming, and wish for look up of legacy names. Not many respondents were keen on restricted data analysis, with an embargo of 6 months seeming reasonable to most.

Overall, the questionnaire provides a useful guide for where to go next; specifically in the next stage of the BIS Roadmap in identifying and building an inventory of data sources and work towards ensuring their long-term persistence.

Dave Edwards – for wheat Gramene/Gaingenes have been available, so there are interfaces to search everything via the WIS portal: this includes search pages and Solr indexes for each (DE has experience of setting this up). Now there are also flat-file resources. In these systems you click out to the databases and can search data in more complex ways, and the BIS for brassica would be able to benefit from this so don't have to reinvent the wheel. DE wants to know who has databases for which he can establish indexes. Dave will contribute his, it is easy to set up the front page, can then identify other data to go in. It won't all be done at once.

GK – asked if responses from the questionnaire were the similar as for wheat, DE confirmed they were.

AL -people at URGI can help. *DE* – stressed the need to coordinate.

ACTIONS:

- GK will contact those respondents who indicated they had data resources to share to compile and initial inventory for sharing.
- DE will work with GK and others to identify current online resources (eg BIP, genomes in JBrowse, et al) that could initiate a BIS 'global search' via Solr indexing.
- To resolve how this is hosted directly or via alias under the brassica.info domain
- GK to provide an update on progress via brassica.info mailing list and in website by May.

6. Standards, ontologies and Brassica trait dictionary

BraTO: The team involved will try and continue to define new traits. GK had been unable to find the trait dictionary in Crop Ontology site. He will let them know by email. Unclear if someone at Earlham there to do it. JI – might have got access.

Gene model nomenclature

At St Malo (item 6, Action: it was agreed that "*Proposals/suggestions to be provided prior to, and then discussed, .. then to be agreed at next meeting (Jan 2019), and then disseminated*"). GK had called for ideas in late 2018 and received a suggestion that has since been refined by a sub-group of the steering committee. This was circulated on the brassica.info mailing list prior to this meeting, and GK asked for views of suitability.

IB – There is importance in recognising the value of names that tell you the context of genes in relative order in chromosomes (genome pseudomolecules). Although there are pan genomes, not all genomes are the same. The nomenclature for a gene-model should be established in relation to originating genome, rather than species, and hence based on the canonical diploids. It should then indicate the species it relates to.

GK – Had received feedback from Elke Diederichsen that the proposed name string were too long, and hence there was concern that would not be widely adopted. This was not generally seen as an obstacle, given that the 'root' of the gene model may be used for comparative reference of loci, but the longer genotype-specific form to describe an allele.

DE – There needs to be translation of legacy names. The Wheat community is putting effort into this and we may be able to learn from this. Should we start from a standard reference?

IB said this is why important to go from a genome perspective (diploid), so there is no HE. He also asked which would be regarded as the core pan-genomes and when can these be adopted to fix the core nomenclature?

IP questioned whether the same 5 digit number would be in the A genome and the A from the AC genome?

IB responded 'yes'. IP thinks this might have issues in *B. napus*. GK asked that if there was a translocation, would we keep the original diploid order?

IB stated that we will use the diploid pan-genome to try and get over genes not being present.

GK suggested that if this happens then we add in a name from the bottom of the list, and this would indicate that the relevant gene was not derived from the diploid pan genome. The name would be based on the *B. napus* position

IP reminded that there were not yet good stable pan genomes, so when will it be done? What do we do in the meantime?

GK suggested that we need an intermediary step, as currently there are names that are the same but they are not the same genes.

DE mentioned that his group had developed Daisychain (http://daisychain.appliedbioinformatics.com.au/) to find the gene in other references, irrespective of what the gene is called.

IB stated that there are not many new genes coming into assemblies now and he is concerned about order

DE/GK asked what how are genes currently named? It is important to establish a stable id.

Action: IB will generate a flow diagram of what is proposed

GK stated that we need to try and avoid homonyms (two distinct objects having the same name).

DE- suggested whether we could name on sequence identity instead? GK said why not associate a DOI instead?

JI stated it would be more useful to know where the gene is on genome, more than knowing potential function. (*functional naming is covered in drill-down annotation and separate nomenclature system established in 2008* - http://brassica.info/tools/data_standards.html)

IB- stated that alien introgression could be a new genome letter

RS suggested that naming on function may raise false expectation amongst users

GK reminded the meeting that in St Malo we had an *Action* (see above) that the nomenclature would be fixed at this MBGP meeting, and he would like to see a timetable developed. GK summarised, stating that proposal discussed is workable, but not achievable right now as we don't have a set of diploid pan genomes that can be used.
IB thinks that we are close enough.
GK asked how do we measure to say it is good enough?
DE thinks that very few genes are missing, the issue is positioning. Others may have references that would help place them. Currently in pan genomes only 40% new genes are placed back in the genome.
IB is doing work to see if there is conserved synteny that can place them. As long as there are less than 10 genes in a block, then it is not a problem (due to numbering redundancy).
IP noted that annotation varies considerably, and currently separate groups do it their own way, with variations around a theme. How do we make sure things are not missed, genes are not split etc? The question is how do we check this quality within the pan genome?
DE suggested that perhaps this could be part of the process.
IB noted that this is usually resolved with the pan genome. Sequences are there, just not called as models.

Action: DE to distribute his pan genomes as a first step

GK suggested that to have process, 3 independent labs, contribute to a proposed naming list that is acceptable, and to accept it will not be 100% perfect - even 90% is good. This is a pragmatic way forward whilst we work to improve the quality over the next decade.
IP asked who will maintain the record of annotation?.
DE can put into Daisychain and link to other references .
(GK note: the acceptable naming of gene models will be incorporated into the Ensembl Plants reference genomes)

KG noted that genomes are being deposited, and asked if there is seed availability? GK stated that in an ideal world yes, and the brassica community now have an open invitation to deposit at ABRC (https://abrc.osu.edu/). KG says this needs to be encouraged. Ideally there will also be a voucher herbarium specimen lodged. ABRC also mentioned in the Brassica workshop that they want donations and really want to push this.

Action: GK to collate the relevant information on agreed gene model nomenclature as discussed and send the proposed protocol to the MBGP mailing list.

7. Scope for multinational collaborative projects

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

IB is expanding his *B. juncea* diversity panel transcriptome

Through a UK-Australia collaboration grant the Brassica Information Portal and interface database that has been developed at Earlham Institute will move to Graham's lab, and depending on available resources will be reconfigured so that additional data from a wider range of sources can be added. GK will also see if he can additional capability can be developed.

IP asked if there was interest in the community for a full reference sequence pan genome, based on NRgene, of between 10-15 lines. IP will do 2 benchmark lines, plus there is darmor available through RS/DE/JB/BC, INRA can commit to 2 genomes. There are also companies involved. Does the community want to add additional genomes? There is also a need to identify genotypes to cover the diversity present. If anyone is interested in contributing, including academic contributions, then they can be considered. For example Ian's contribution of genome validation. IP will be in touch with people about lines (60K data). Pick lines that will maximise diversity

8. Any other items/meeting announcements

- Brassica 2020: This meeting will now be in Saskatoon, not USA. IP to organise. RS said that he is happy to host in 2020 in Giessen
- YPL is the working leader for Brassica ISHS. If anyone is interested in holding a symposium then please let him know,
- IRC in Berlin June 16-19 2019. Abstracts due mid-March.

GK announced that there is a postdoc pre-breeder/breeder (brassica focus) position open at SCU, which will be advertised in Feb 2019. If interested please follow up with Graham

9. New Chair:

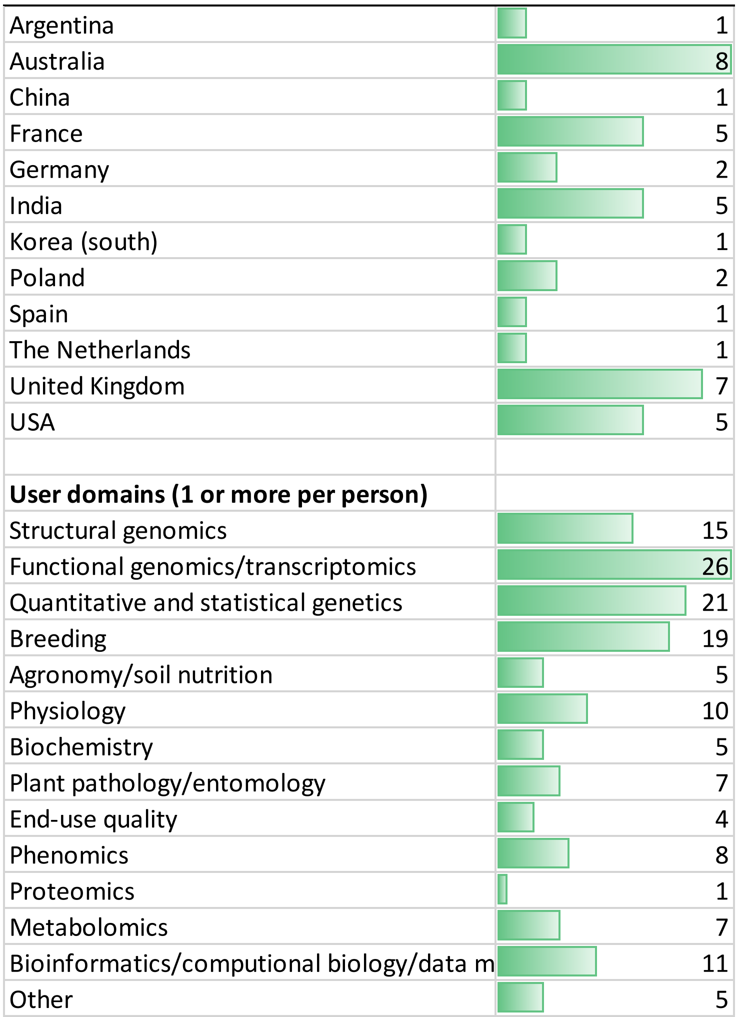
It was agreed that GK continue for a further year, as nomenclature issues are just gaining traction and he is best placed to follow them through to completion. GK suggested that he have a shadow for a year before handing over, RS suggested JB, which was agreed.

Appendix: results of BIS Questionnaire

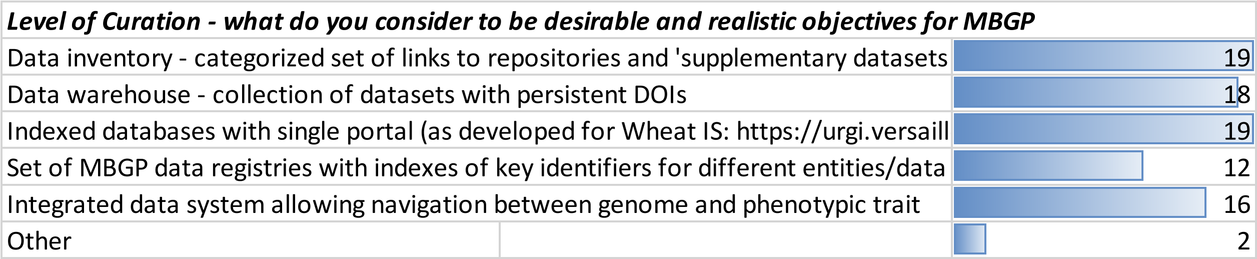
MBGP Brassica Information System Questionnaire 2018/19

Preliminary report 7/1/2019 (compiled by Graham King)

Total respondents = 39 (all identified as University/academic apart from one industry)

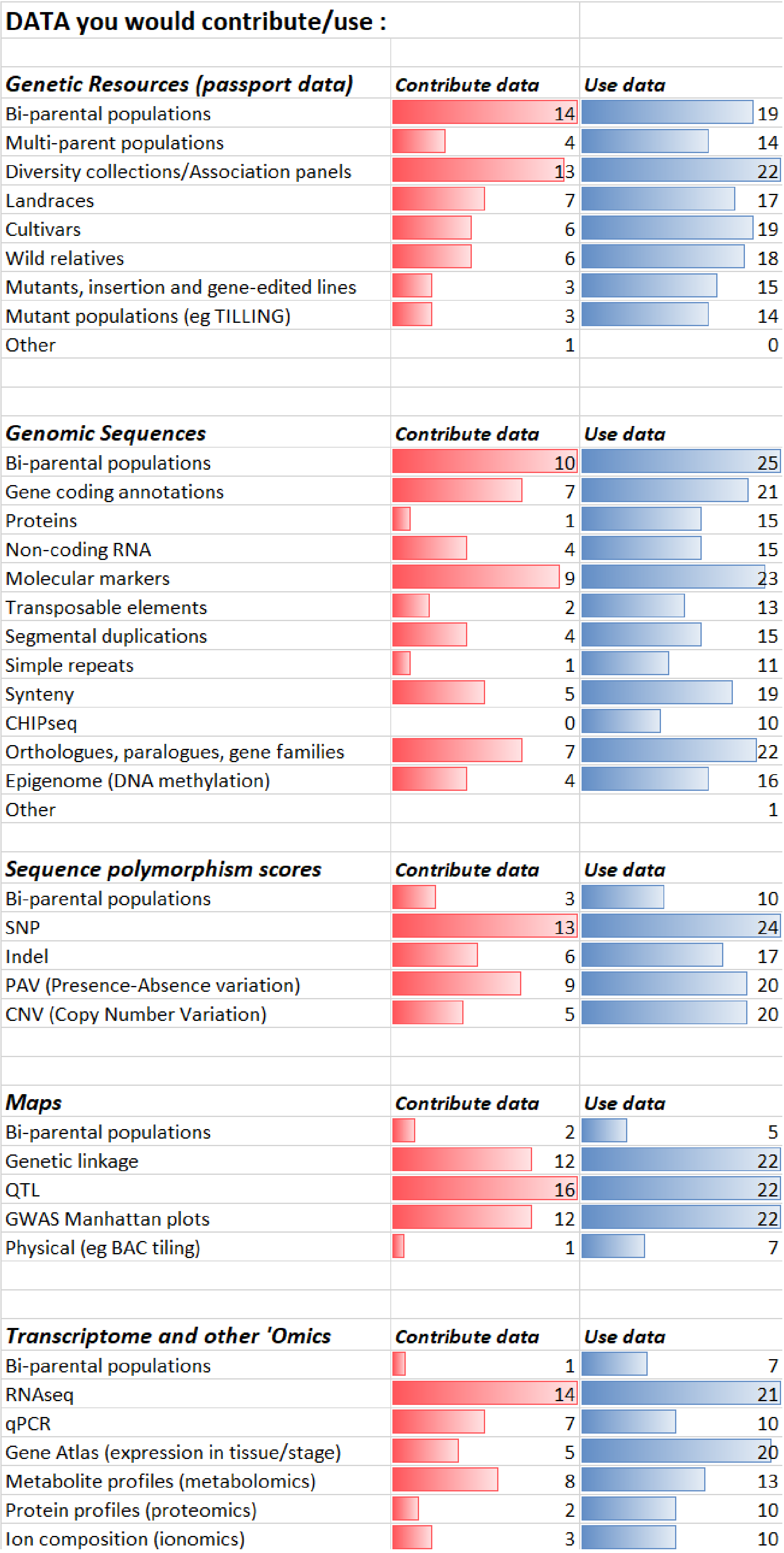


Argentina	1
Australia	8
China	1
France	5
Germany	2
India	5
Korea (south)	1
Poland	2
Spain	1
The Netherlands	1
United Kingdom	7
USA	5
User domains (1 or more per person)	
Structural genomics	15
Functional genomics/transcriptomics	26
Quantitative and statistical genetics	21
Breeding	19
Agronomy/soil nutrition	5
Physiology	10
Biochemistry	5
Plant pathology/entomology	7
End-use quality	4
Phenomics	8
Proteomics	1
Metabolomics	7
Bioinformatics/computional biology/data m	11
Other	5



Level of Curation - what do you consider to be desirable and realistic objectives for MBGP	
Data inventory - categorized set of links to repositories and 'supplementary datasets	19
Data warehouse - collection of datasets with persistent DOIs	18
Indexed databases with single portal (as developed for Wheat IS: https://urgi.versaill	19
Set of MBGP data registries with indexes of key identifiers for different entities/data	12
Integrated data system allowing navigation between genome and phenotypic trait	16
Other	2

Comment (other): I regard 3 & 5 as desirable, the other options are more realistic in a 5-year timeframe but could be regarded as intermediate steps?



DATA you would contribute/use :		
Genetic Resources (passport data)	***Contribute data***	***Use data***
Bi-parental populations	14	19
Multi-parent populations	4	14
Diversity collections/Association panels	13	22
Landraces	7	17
Cultivars	6	19
Wild relatives	6	18
Mutants, insertion and gene-edited lines	3	15
Mutant populations (eg TILLING)	3	14
Other	1	0
Genomic Sequences	***Contribute data***	***Use data***
Bi-parental populations	10	25
Gene coding annotations	7	21
Proteins	1	15
Non-coding RNA	4	15
Molecular markers	9	23
Transposable elements	2	13
Segmental duplications	4	15
Simple repeats	1	11
Synteny	5	19
CHIPseq	0	10
Orthologues, paralogues, gene families	7	22
Epigenome (DNA methylation)	4	16
Other		1
Sequence polymorphism scores	***Contribute data***	***Use data***
Bi-parental populations	3	10
SNP	13	24
Indel	6	17
PAV (Presence-Absence variation)	9	20
CNV (Copy Number Variation)	5	20
Maps	***Contribute data***	***Use data***
Bi-parental populations	2	5
Genetic linkage	12	22
QTL	16	22
GWAS Manhattan plots	12	22
Physical (eg BAC tiling)	1	7
Transcriptome and other 'Omics	***Contribute data***	***Use data***
Bi-parental populations	1	7
RNAseq	14	21
qPCR	7	10
Gene Atlas (expression in tissue/stage)	5	20
Metabolite profiles (metabolomics)	8	13
Protein profiles (proteomics)	2	10
Ion composition (ionomics)	3	10

Data exchanges with other databases/portals	
Ensembl Plants	20
Brassica Information Portal	23
Cyverse (iPlant)	11
Elixir	2
NCBI / ENA	18
TAIR / IAIC (Arabidopsis portal)	21
URGI	6
www.brassica.info (MBGP portal)	24
http://brassicadb/org/brad/ (BRAD)	20
Other	4

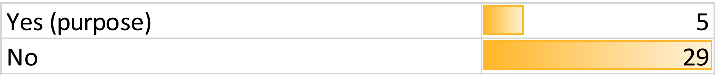
Other: https://brassibase.cos.uni-heidelberg.de/ and http://www.brassicagenome.net/ (2)

Comment: Particularly for existing Brassica platforms I would prefer to discuss how/what would be exchanged and whether/how a BIS might incorporate existing structures. Portals which are less well maintained may become redundant, others with good ideas but which are not yet overflowing with data (BIP?) might be enriched by more exchange?

In which services would you be interested ?		
Data browsing	27	
Database integration (link data together)	21	
Ready access to experimental meta-data	20	
Navigate trait to genome (GWAS/QTL)	25	
Download data files	23	
Download community analysis tools	17	
BLAST or other alignment servers	23	
Genome viewers	23	
Synteny / collinearity viewer	23	
Paralogue catalogue / finder	21	
Complex query capability	12	
Analysis workflows / pipelines	17	
Computing capacity	10	
Other	1	*Trait to plant accession*
Which tools or data standards would you use ?		
MBGP look-up data REGISTRIES for key iden	22	
MIAPPE (for phenotyping)	16	
BraTO (Brassica Trait Ontology)	20	
MBGP standardised gene-model per genome	23	
Look up table of legacy gene-models	16	
Others	0	

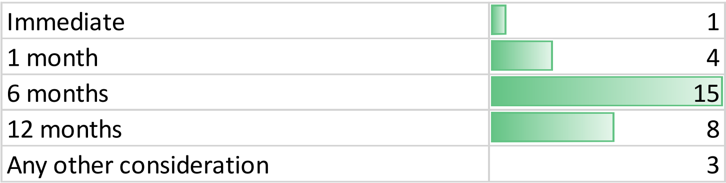
Restricted data access. Would you be interested in such functionality?

(no if wished to ensure all data submitted to be public)



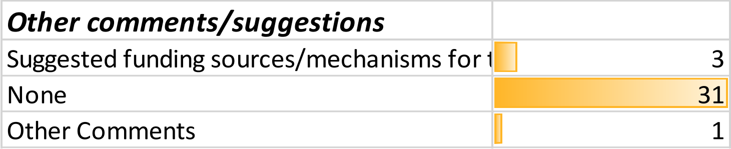
Yes (purpose)	5
No	29

Reasonable embargo period?



Immediate	1
1 month	4
6 months	15
12 months	8
Any other consideration	3

Other = upon release by the authors; Immediately upon acceptance of publication (if relevant)



Other comments/suggestions	
Suggested funding sources/mechanisms for t	3
None	31
Other Comments	1

Suggestions for BIS:

- *H2020 CORNET: Must include at least two EU countries (e.g. DE/UK/NL and/or FR?).*
  - *PRO: Application process relatively simple for EU and funding quota is unusually high at ~50%!*
  - *CON: Must include interested 5 small/medium industry partners in each participating country so very hard to raise the numbers, would need lobby work to bring in breeders, biotech companies, IT startups, etc...*
- *Sponsorship from industry players with potential interest in better access to large-scale public datasets, pangenomes etc., e.g. for use in deep learning applications (Bayer, BASF, Syngenta, KWS, Limagrain, NPZ, Rijk Zwaan)*
- *Feel free to contact urgi-contact@inra.fr to get the source code of the WheatIS portal that have been developed to be generic for all plant species.*

Report compiled from Qualtrics summary data by Graham King, SCU, Jan 7th 2019