Brassica genome sequencing meeting: Saskatoon, Canada   Thursday 27 August 2009

Present:
Chair: Ian Bancroft (IB) ian.bancroft@bbsrc.ac.uk
Secretary: J. Chris Pires (JCP) piresjc@missouri.edu
Jacqui Batley (JB) j.batley@uq.edu.au
Dave Edwards (DE)   Dave.Edwards@uq.edu.au
Xiaowu Wang (XW) wangxw@mail.caas.net.cn
Shengyi Liu (SL) liusy@oilcrops.cn
Jinling Meng (JM) jmeng@mail.hzau.edu.cn
Isobel Parkin (IP) parkini@agr.gc.ca
Andy Sharpe (AS) Andrew.Sharpe@nrc-cnrc.gc.ca
Others present from Saskatoon group and Shengyi Liu's graduate student

Minutes from January 2009 PAG to confirm – We approved distributed minutes as final.
Action: Isobel Parkin will alert Graham King to post on the web

Xiaowu Wang (XW): Two updates on activities since sent out *B. rapa* assembly to our
community. First, the version sent out prior to this meeting did not include 3 GB of 180 bp
insert library reads. After adding this data, the contig size increased from 25 kb to 46 kb, and the
scaffold number was reduced to 39 super-contigs. Second, there have been continuing activities
on genetic anchoring – used Lim/Park maps from Korea initially.
However, Graham King has indicated there are some discrepancies between the Lim and Park
maps, which may impact alignment of the *B. rapa* assembly.
IB commented that the discrepancies in BAC positioning on the genetic maps is due to the fact
that Yong Pyo tracked the Chiifu allele to the sequenced BAC, which removed errors due to
duplication in the genome.
Yong Pyo Lim has sent some SSR markers to XW, but less than 200, which is far less than
needed for giving good anchoring of most of the scaffolds.

Jinling Meng (JM): mapping the Korean markers onto Tapidor (TN) population as another
check. Some markers from Lim and some from Park groups in Korea (~180 BACs mapped by
YPL).

Dave Edwards' (DE) evaluation report:
Examined PE reads from A1, A3, A8 and A9 provided by XW against Illumina paired read data
that DE has sequenced from *B. rapa*, *B. oleracea* and *B. nigra*, and compared the four
chromosomes with Arabidopsis. Across the four scaffolds, there are 18 positions that need re-
examination due to inconsistency between Arabidopsis and Brassica and low coverage (not
necessarily incorrect, but need validation). JB will target these for additional work (genetic
mapping): 6 on A9, 5 on A3, 4 on A8, and 3 on A1.
Second, scanning across paired end (PE) data, there are some indications of a few discrepancies
between scaffold and reference assembly. We can find discrepancies down to 10s of bases but
not sure how many (estimate 10s to 100s) from among these 4 chromosomes – but this seems
easy to correct and look like some small-scale indels or lack of coverage (and can be small
groups of Ns). We should not see these types of discrepancies for Illumina data. It looks like

some repeats were masked during the assembly process and these show up as strings of N's in the contigs. The length of N's seems to correspond to the amount of missing sequence and it should be relatively easy to go back and to fill in the N's with real sequence using available sequence data.

XW: This is true and when get 454 data, we will confirm this with respect to SSR.

IB: It looks like most SSRs have been accurately assembled with Illumina sequencing.

DE: We have some markers to add to improve assembly but otherwise we will focus on cleaning up small indels. We have our own Australian sequencing data and public BACs to validate scaffolds (BACs were not used as part of assembly –this is also what Bancroft, Parkin, and Koreans doing). Also there will be more tests with assembly to Arabidopsis.

IB: Boulos Chaloub will be providing 454 PE data in September that can be used to improve the assembly. Genescope will use large insert libraries but don't know details. They will sequence to 5x but don't know how many plates this will be but guess 5-10 runs (100,000 euro budget). The Genescope pipeline is well established for 454 and all PE reads. Not sure if reads will be 125 bp each end or longer.

XW/DE: Include 454 reads in assembly or map onto Illumina reads.
IB: Genescope could also reassemble all raw data.
XW: China BGI can also do this (as doing for potato).
DE: Error types very different for 454 and Illumina data so weird branching in de Brujin graphs could result in merging the data prior to assembly.Ideally, do 454 and Illumina *de novo* assemblies separately and then map reads onto each other.

Isobel Parkin/Andy Sharpe's Evaluation Report
A2 and A10 BACs – 150 of these BACs that Canada would like to check against the A2 and A10 scaffolds. Canada will send BAC sequence to China and in return for access to the A2 and A10 scaffolds.
Have carried out additional targeted BAC sequencing for some triplicated regions: A2/A10/A3 (FLC, etc), A1/A8/A3 (Ian regions) and duplicated segment on A7.

Martin Trick and Ian Bancroft Evaluation Report (PowerPoint presentation, later sent by IB)
1. Overall alignment
1a. Overall alignment of sequences to A1
Dot plot of chromosome A1 looks good – A1 has 150 or so BACs assembled into 12 scaffolds that were compared with the 9 A1 IVG-BGI scaffolds. Clearly could see that most regions aligned as predicted (most of sequences fall on diagonal).
Caveats: All BACs sequenced to Phase II so there will be gaps in JIC BAC assemblies. However, the scaffolds order has been confirmed by genetic anchoring.
1b. Overall alignment of sequences to A8

Dot plot of A8 identified clear difference between data types. The difference identified a false fusion introduced in the IVF-BGI build. So there was obviously value in comparing the two data sets in this fashion.

1c. Alignment of scaffold and BAC sequences based on comparative gene annotation showed numerous anomalies BUT they were in broad agreement. Example – 2 copies from Phase 2 Sanger sequencing seen only in one copy at IVF-BGI build (4 kb + size piece).

2. Sequence level

Some base mismatches, and problems with homopolymers (BGI has one less bp in string of As for example). Assume Sanger data is correct because F and R reads available.

DE: Sanger and 454 sequencing more prone to this type of error than GAII, but should confirm which is correct.

However, most repetitive regions are successfully resolved.

Since previously only aiming for Phase II, the more interesting part will be looking at annotation.

3. Detailed Annotation

Martin Trick finds that annotation pipeline can scale from BAC size sequences to 1.5 MB. Identification of SSR regions is also good and got nearly all of them.

Most assembly gaps do not affect gene models corresponding to orthologues of genes annotated in Arabidopsis. They often contain gene models corresponding to transposons.

But some assembly gaps do affect gene models corresponding to orthologues of genes annotated in Arabidopsis, although all those identified correspond to non-collinear genes (likely repetitive sequences/transposons, really) .

Significant amounts of InDel variation observed between the IVF-BGI build and BrGSP data.

DE: indels also observed and probably due to Illumina collapsed repeats, can resolve by local assembly and comparison with 454 data

JCP: how resolve these differences?

Ian: Rather than going in and edit case by case, we should improve automated assembly.

Jacqui Batley: When performing *in silico* and wet lab validation for the SNPs, we identified that there is plastid sequence (mt and cp). Some of these 300 bp fragments were all plastid, for others there was just part of the sequence hitting plastid. Even if is plastid sequence in the genome, would not be suitable for mapping due to the repetitive nature. Will be validating and mapping the SNPs using the GoldenGate assay.

Should not be much cp since DNA extracted from callus. Also, some parts of chloroplast not covered and other parts of cp.

DE has assembled whole cp genome and most of Br mt genome (have scaffolds, just not full assembly). Could use this information to survey scaffolds for cp insertions. We will do this as part of the annotation process.

IB: So, what next? More evaluation work or is IVF assembly acceptable?

DE: We are not going to start sequencing BACs. To ensure that we have a good final reference genome we need more finishing and polishing. Advocate reviewing any possible misassembled scaffolds and including Jacqui's mapping data should assist with validating the genome sequence. Need to fix end regions as requirement for annotation to call a good draft genome. Not a lot of indels but easy to find and fix. SNP variation that Ian observed can be looked at in more detail – could be error on Sanger side or Illumina side.

In summary, we don't want to return to BAC-by-BAC sequencing but just need to fix IVF assembly.
AS: Agree – just need to fix superscaffolds.

IB: Dense genetic anchoring will fix false fusions.

JB: She can focus SNP mapping (or at least ignore very good regions).
Will complete small GoldenGate test first (384 SNP validation check) and then move to 1536 later after more pilots since have 10% attrition rate.

AS: Wellesbourne has SSR data? Graham King could add information from what mapped in *Brassica napus*.

IB: Ideally you want alleles to match in CKDH and this is what Jacqui will be tracking.

XW: Has supplied 600-700 SNP markers to Jacqui but many will not work so XW will supply more sequence data to Jacqui and if possible can include additional information from others for example Canada *B. napus* A genome as well as Yong Pyo Lim markers.
Aim for three markers per scaffold if possible

AS: We will have 1536 public SNP markers available for *B. napus* in December, but have markers in hand already that can be checked for current scaffolds. Run through SSRs and SNPs that are already developed and see if provided value for the genetic anchoring.

XW: We would like to have the feedback of marker-anchoring of the scaffolds which were released for validation.

DE: We have also checked our *B. napus* markers across the scaffolds and will continue with this.

AS: Additional evaluation of triplicated regions needed?

IB: Not really needed.

IP: Check on A7 duplication that is more recent – notoriously hard place to put together. Disease resistance genes there, a tandem duplicate like FLC5. Could be specific to *B. napus* and not in *B. rapa* A7?

AS: Also, additional 454 data into project from France needed. If get large inserts that would be great. Is more data needed for the project? Probably not need even Kenshin data.

XW: China has 14 kb inserts but improving like improved 10 kb data.

IB: Boulos, tell him he needs to get data to community by October?

AS: Is there need for more gene space? Validate amount of gene space characterized?

IP: You were missing 4,000 Arabidopsis genes but not genes?
Are these peri-centromeric genes or ?

IB: What is missing is not transcribed at high level.

XW: We have covered 98% of genes. We will do 3 Gb of transcriptome data (RNA-Seq) from a range of tissues (pooled, not indexed). Will do this over next two weeks.

JM: We just need improvement, do not need to go to BAC-by-BAC, should make reference for *Brassica rapa* if not other Brassica species.

IB: For physical resources, it sounds like it will cover it.
Do you need more input on computational/informatics aspect?

DE: We all agree that GFF3 is standard file format, can share annotation using CVS.

IB: Our funding is out and just can do what we have done here. Compare two annotation schemes after scaffolds done (China/Australia annotation compare to UK annotation).

Matt: the annotation should always be additive and if using Gbrowse the more the merrier.

JCP: What about public curation and community annotation?

IB: Do this later and just get really good automated annotation out now, not enough Brassica researchers interested in manual annotation and that will come later.

DE: Doing a lot of work on transposons across Brassica and Brassicaceae (current and novel).

IB: WHAT CAN WE GET INTO PUBLIC DOMAIN BEFORE PAPER PUBLISHED?
So not just people in this room benefit but into public domain?
We do not want to be perceived as benefiting ourselves unfairly.

DE: We don't want to release until we publish, but publish quickly.

IP: Will you include BAC data in the paper?

XW: Yes, will include all BAC data – but no idea if Korea will release. Just use their public Korea 600 BACs. BAC ends data are valuable for assembling; the fully sequenced BAC data will be only valuable for validation.
Australia 454 data is public but not deconvoluted so worthless.
Timeline proposed: Get finalized sequence: China transcriptome and French data by October 31.
Mapping data by October 31? (will aim for this!)
Finish first version of manuscript by November, and submit by December.

IB/JCP: MIAMI convention and BrGSP says to make data public; do it after accepted or before?

XW: Cucumber paper in revision at Nature Genetics – first next-gen paper
(Sanger data and 454 data also, 4x Sanger), so *B. rapa* paper cannot have "technology" angle.
JCP: B. rapa paper – how make high impact paper?

XW: Triplication and gene loss story – what genes gone and what keep and why?
Classification of which gene families are kept or lost, and still functional (Ian Bancroft agrees).

JCP: "Freeling" analysis of which genes kept or retained, and if lost are they transposed or not.
Take Bancroft BAC comparison-style papers to bigger picture since can now say if genes lost
not just in BACs but if lost from genome or moved.

IB: Second story is chromosomal number/block evolution (similar to Brachypodium
presentation). (avoid genome evolution versus exploiting Arabidopsis)
Popular to link to wheat or food security – or health benefits of Brassica?
What is plan for paper to get to high profile journal?
Sexy science and pitching correct is hard. Take on board advice that we can give you – don't go
too far over word limit. We can recommend whom to approach at Nature and Science.

ACTION: UNANIMOUS VOTE-AGREE THAT IVF-BGI SEQUENCING IS ACCEPTABLE
ENOUGH TO ABANDON BAC-BY-BAC APPROACH.

OTHER ITEMS:

C genome:

DE: we are not setting up a separate C genome project but can offer sequence data and
bioinformatics support as a resource, willing to contribute to and collaborate on projects.
Xiaowu Wang doing his genome – and back-to-back publications with USA/Canada project?

JCP: We are sequencing *Brassica oleracea* TO 1000 DH3 genotype (USA/Canada/UK project).
These are our current collaborations and open to possibilities (back-to-back or combined
publications, several options still open).

IB: Conclude, no need for formal agreement for C genome sequencing now, just major players
stay in touch.

Shengyi Liu/XW/China – we have 500 Mb of C genome assembled with 454 and Illumina.
C genome reference: Jinzao Sheng (02-12) (Golden Early, in English)– Chinese cabbage – 50 x
coverage!
Also sequenced Beijiing Zaoshu (397) (Beijing Early, in English) – less coverage, more for SSR
and SNP (other parent of mapping population of Jinzao Sheng x Beijing Zaoshu 397).We would
very likely re-sequence Chinese kale as well.
SL: We are open to collaborations.

B genome (and AB and BC genomes)

Dave Edwards/Jacqui B/Australia doing some B genome sequencing for gene discovery in wild Brassicaceae. Jacqui interested in B genome diversity and related species (not just B genome), and Victoria/DPI – Bob Redden interested in evolution of this and has temperate collection. We have lots of hypotheses and questions.

Parkin/Sharpe/Canada leading a DH line *B. nigra*, and interest in *B. carinata*, and *B. juncea* (mostly grow *B. juncea*, and Dijon mustard, *B. carinata* interest for drought tolerance and inflorescence; and *B. carinata* is future crop for industrial oils).
(also *B. juncea* grown in Australia).
Andy - Draft assembly of B genome by 454 and Illumina platforms – PBI DH genotype with BAC library, and BES in progress. Sequence over next 6 months,
And then *B. carinata* and *B. juncea* after that (but industrial partners so may not all be public)

AC genome: *B. napus*

Shengyi Liu/OCRI/China:
B. napus sequence **Zhangshuang 11** INBRED – 20 x coverage data ready. Based on assembly of A and C genomes and tentative assembly of napus AC (if possible with Illumina 20x + 454 2x + BAC ends), and comparison of the three species, to decide whether do *de novo* sequencing/assembling for *B. napus* AC genome and if yes, do more sequencing.
Mapping population (F2 or more than 200) is from plant breeding group, another parent is 73290 inbred line. Looking for more SSRs or decoding codes used in different groups or different genetic maps.
Looking to do low coverage of other 11-12 *B. napus* genomes (send some to Ian Bancroft for transcriptomes). Open to collaborations or just publish together with China *B. oleracea*?
(We think this may be first *B. napus* assembled)

Parkin/Sharpe/ Canada projects: Resequenced *B. napus* after A and C genomes.
Draft assembly of public reference ***B. napus* – DH 12075.**
Also 16 "private" *B. napus* genotypes – company interest in that data so some data will be proprietary; confidential for 5 years.
Canada-Snowdon-Bancroft project: 1536 Brassica SNP as public Golden Gate system – release early 2010. Looking at SNPs across 400 lines; these are public.

Bancroft/UK: Agreed to work with Boulous Chalhoub on *B. napus* transcriptomes and Jinleng Meng project (genomic and transcriptomes).
His interest is in diversity analysis – 100 lines of 3 week pre-vernalization leaf transcriptome.
Should all be DFFS lines: choose from physiological and morphological variation.
Generally interested in A and C diversity.
Sequence **Tapidor** and **Ningyou7** with Jinling Meng (*de novo* assembly).
Interested in *de novo* data, structural differences and some half dozen more transcriptomes.

Australia: SNP discovery and sequencing in *B. napus* – 3 year release policy after May 2009/data generation, but can collaborate on specific projects. Can't say what are proprietary lines, but can name public lines. Project with Regine Delomoure as partner (JB's fellowship) – **Darmor** and **Yudal** to be sequenced over 3 years by JB.

JM: He suggested that the community should work together like on *B. rapa* project.
Suggested C genome and AC from Shengyi Liu should not be in one paper?
Or China-Australia effort join with C genome paper of USA/Canada/(UK) project?
And China-Australia with AC genome papers with Bancroft/Jinleng Meng/Canada?

JCP: We should submit *B. rapa* paper first, and then discuss next steps for C genome, AC genome, and various B genome papers.
BRASSICACEAE wide projects to circulate as wide as possible:
Need to post to Brassica.info list (whole genome or transcriptome)

Shengyi Liu/OCRI/China: We will sequence other members of wild *Brassicaceae*

JB: will provide her list of wild *Brassicaceae* sequencing.

JCP: Preview of USA DOE-JGI *Brassicaceae* taxa to sequence (will provide list later).
Pires also starting to sequence chloroplasts (and transcriptomes) across family.

DE/JB: Blackleg – public Australian effort that he invites all to join.

IB: conclude meeting, congratulations to IVF/OCRI/BGI/China team on very good *B. rapa* assembly.

END