SNP discovery from amphidiploid species and transferability across the Brassicaceae

Jacqueline Batley
University of Queensland, Australia
j.batley@uq.edu.au
• Objectives
• *Brassicas*
• Genome Sequencing
• SNP discovery
• SNP validation
• Cross species transferability
• Future work
Objectives

- Development of bioinformatics tool for SNP discovery and annotation
- Establish cost effective discovery and validation of SNPs within the amphidiploid *B. napus*
- Assess association of SNPs with genes for agronomic traits
- Assess the extent of LD within *B. napus*
- Assess genetic diversity of important agronomic genes within cultivated *Brassica* spp. and wild relatives
- Establish a strategy for SNP discovery from other large and complex genomes.
Methodology

- Paired end sequence from parents of mapping populations
- SNP discovery
- Genotyping using golden gate
- SNPs genetically and physically mapped
- Cross species amplification to other Brassicaceae members
Brassicas
Diversity genomics

• Characterising genomic and phenotypic diversity in cultivated and wild plant species and their pathogens
  ➢ Brassicaceae, *Leptosphaeria maculans*
• Investigating genetic variation in crops and wild relatives
• Investigating the evolution of plant pathogen interactions
• Identifying novel genes and genetic markers for traits of interest, such as disease resistance
• Germplasm collections are valuable gene pools

• Assessing genetic and genomic diversity within these collections:
  ➢ assign lines and populations to diverse groups
  ➢ study the evolutionary history of wild relatives
  ➢ verify pedigrees and fill in the gaps in incomplete pedigree or selection history
  ➢ monitor changes in allele frequencies in cultivars or populations
  ➢ help narrow the search for new alleles at loci of interest.
Domestication bottlenecks

- *B. napus* canola, *B. juncea* mustard and *B. carinata* are allopolyploids.
- Rare natural polyploids only incorporate a limited genetic diversity from progenitor diploids.
- Wide genetic diversity in *B. rapa*, *B. nigra*, *B. oleracea* progenitors and wild relatives, options to enhance canola and mustard.
- A range of strategies is available to realise the genetic potential of the *Brassicaceae*.
Sequence data

• Illumina GAIIx and Hi-Seq data for:
  • 8 *B. napus* cultivars
  • 2 *B. rapa* cultivars
  • *B. oleracea*
  • 3 Brassicaceae

• Funding for 100+ Brassicas
Brassica genome sequencing

- *B. rapa* ssp. Pekinensis var. Chiifu
- 10 chromosomes, ~550 Mbp
- Multinational *Brassica* genome sequencing committee originally agreed BAC by BAC sequencing approach
- >100,000 BAC end sequences
- >600 BACs sequenced
- Genome sequenced using Illumina GAIIx
B. rapa SNP discovery and genotyping

- Illumina paired end sequence from parents of mapping populations
- SNP discovery
- Genotyping using golden gate
- Physical mapping
- Cross species amplification to other Brassicaceae members
SNP validation
B. rapa SNPs

- 768 SNPs
- 95% SNPs correctly called
- Lower coverage more likely to be true SNPs
Cross species transferability

- SNPs from A genome amplified and sequenced from:
  - C genome
  - Parents of *B. napus* mapping populations
  - Brassicaceae
Genotyping

- Illumina Golden gate system
- 384 SNPs
  - 2 *B. rapa* mapping populations
  - Parents of *B. napus* mapping populations
  - Selection of wild Brassicaceae
SNP Genotyping
SNP Genotyping
• Assess relationships within the Brassicaceae
• Correlate this with morphological and interspecific hybridisation data
Brassicaceae diversity - SNPs

Genome key
A : Green
B : Blue
C : Red
BC : Purple
AC : Aqua
AB : Pink
Tertiary: Orange
AC predicted : Black
B. napus SNP discovery

• Custom algorithm developed for SNP discovery from Illumina data for amphidiploid species
• Distinguish between inter and intra genomic SNPs
**B. napus SNP discovery**

- Custom algorithm developed for SNP discovery from Illumina data for amphidiploid species
- Distinguish between inter and intra genomic SNPs

| XA_0011r | 1252  | 1252 | 3 | S=G=2;M1=G=3;Sr=X=0;A=G=3;J=T=3;M2=G=1;Bn=X=0;E=X=0; | T;G; |
| XA_0011r | 1379  | 1379 | 5 | S=T=2;M1=T=3;Sr=X=0;A=T=1;J=C=3;M2=X=0;Bn=X=0;E=C=2; | C;T; |
| XA_0011r | 2036  | 2036 | 4 | S=G=1;M1=G=2;Sr=X=0;A=G=1;J=T=8;M2=T=3;Bn=X=0;E=T=6; | T;G; |
| XA_0011r | 4921  | 4921 | 2 | S=X=0;M1=X=0;Sr=X=0;A=T=8;J=X=0;M2=X=0;Bn=X=0;E=C=2; | C;T; |
| XA_0011r | 5070  | 5070 | 4 | S=X=0;M1=G=2;Sr=X=0;A=G=2;J=A=6;M2=X=0;Bn=X=0;E=X=0; | A;G; |
| XA_0011r | 5273  | 5273 | 3 | S=C=4;M1=C=5;Sr=X=0;A=C=6;J=G=2;M2=X=0;Bn=X=0;E=G=1; | C;G; |
| XA_0011r | 5442  | 5442 | 8 | S=T=1;M1=X=0;Sr=X=0;A=T=7;J=C=5;M2=X=0;Bn=C=1;E=C=3; | C;T; |
| XA_0011r | 5512  | 5512 | 7 | S=G=3;M1=G=3;Sr=X=0;A=G=5;J=A=4;M2=X=0;Bn=A=2;E=A=1; | A;G; |
| XA_0011r | 5976  | 5976 | 11 | S=T=8;M1=T=1;Sr=X=0;A=T=2;J=C=6;M2=X=0;Bn=C=2;E=C=3; | C;T; |
| XA_0011r | 5992  | 5992 | 10 | S=A=9;M1=A=1;Sr=X=0;A=A=3;J=G=5;M2=X=0;Bn=G=2;E=G=3; | A;G; |
**B. napus SNP discovery**
SNPs in GBrowse
# B. napus SNP discovery

<table>
<thead>
<tr>
<th>Base Change</th>
<th>Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&gt;G</td>
<td>transition</td>
<td>105045</td>
</tr>
<tr>
<td>C&gt;T</td>
<td>transition</td>
<td>105513</td>
</tr>
<tr>
<td>A&gt;C</td>
<td>transversion</td>
<td>42480</td>
</tr>
<tr>
<td>A&gt;T</td>
<td>transversion</td>
<td>49287</td>
</tr>
<tr>
<td>C&gt;G</td>
<td>transversion</td>
<td>29828</td>
</tr>
<tr>
<td>G&gt;T</td>
<td>transversion</td>
<td>42217</td>
</tr>
</tbody>
</table>
### B. napus SNP discovery

<table>
<thead>
<tr>
<th>Base Change</th>
<th>Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&gt;G</td>
<td>transition</td>
<td>105045</td>
</tr>
<tr>
<td>C&gt;T</td>
<td>transition</td>
<td>105513</td>
</tr>
<tr>
<td>A&gt;C</td>
<td>transversion</td>
<td>42480</td>
</tr>
<tr>
<td>A&gt;T</td>
<td>transversion</td>
<td>49287</td>
</tr>
<tr>
<td>C&gt;G</td>
<td>transversion</td>
<td>29828</td>
</tr>
<tr>
<td>G&gt;T</td>
<td>transversion</td>
<td>42217</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Base Change</th>
<th>Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&gt;G</td>
<td>transition</td>
<td>24207</td>
</tr>
<tr>
<td>C&gt;T</td>
<td>transition</td>
<td>24375</td>
</tr>
<tr>
<td>A&gt;C</td>
<td>transversion</td>
<td>10158</td>
</tr>
<tr>
<td>A&gt;T</td>
<td>transversion</td>
<td>12254</td>
</tr>
<tr>
<td>A&gt;T</td>
<td>transversion</td>
<td>12254</td>
</tr>
<tr>
<td>C&gt;G</td>
<td>transversion</td>
<td>6621</td>
</tr>
<tr>
<td>G&gt;T</td>
<td>transversion</td>
<td>9918</td>
</tr>
</tbody>
</table>
B. napus SNP density

![Graph showing SNP density distribution with Series1 dataset.](image-url)
B. napus SNP validation

- Sanger sequence validation of predicted SNPs
- 9/10 correctly predicted
- Using results to refine algorithm
Brassicaceae
Gene discovery

Primer

PCR

Gene/EST

Genomic sequence

Known
(Arabidopsis)

Unknown
(Brassica)
Welcome to ACPF Bioinformatics TAGdb.

This service performs BLAST alignment between a single query and short pair reads of selected species.

Please enter a valid email address

Note: Your result will be sent to the specified address.

Sequence data

Either: Select the sequence file to upload:

Otherwise: Enter a sequence in FASTA format:

Note: Query sequence must be less than 5000 nucleotides.

Species selection

Please choose a query species:

Barley
Bansilo
Nootana
Pongame
Rye
Wheat

Short paired-read library selection

Please select one or more paired-read libraries to search:

A. rape chufu - 36 - 140 - RbC_03_001
B. rape chufu - 36 - 300 - RbC_03_002
C. rape chufu - 35 - 2700 - RbC_27_001
D. rape chufu - 35 - 3500 - RbC_37_001
E. rape chufu - 35 - 2800 - RbC_07_002
F. rape chufu - 36 - 410 - RbC_03_001
G. rape chufu - 36 - 410 - RbC_03_002
H. nigrum - 70 - 2700 - RbC_27_001

Format: SpeciesName - ReadLength - InsertSize - LibraryName

Start

TAGdb output
Gene discovery

• Finding the genes for the traits

• Integration of genetic data with genomic data
  • Mapping of QTL regions to genomic data

... 

Annotation
Gene discovery

• Finding the genes for the traits

• Integration of genetic data with genomic data
  • Mapping of QTL regions to genomic data
Gene discovery - application
Future work

• SNP identification and genotyping of cultivated and wild Brassicaceae
• Large scale SNP discovery and genotyping for fine mapping and LD studies
• Identify which Brassicaceae to sequence
• Use next generation sequencing data, molecular markers and morphological variation to study diversity across *Brassica* species and wild relatives
Next generation sequencing data is suitable for gene, promoter and SNP discovery in non-sequenced and orphan species.

SNPs can be applied for gene discovery and evolution in crop species and wild relatives.

High throughput genotyping can be used for fine mapping and LD studies.
Acknowledgements

Emma Campbell
Christina Delay
Megan McKenzie
Reece Tolleneare
Joanne McLanders
Manuel Zander
Alice Hayward

Paul Berkman
Chris Duran
Kaitao Lai
Michal Lorenc
Sahana Manoli
Adam Skarszewski
Lars Smits
Jiri Stiller
David Edwards

Bob Redden
Harsh Raman
Xiaowu Wang