Finding function in complex crop genomes

David Edwards
University of Queensland, Australia
Dave.Edwards@uq.edu.au
Outline

- Second generation DNA sequencing technology
- TAGdb
- Candidate gene discovery
- Brassica repeat identification
- Brassica SNP discovery
Second-generation sequencing (2GS)

- Illumina GAIIx and HiSeq2000
  - ↑↑↑ sequence
  - ↓ money
  - ↓ time
  - ↓ read-length
  - ↑ computation
Illumina paired reads

Insert size

- Illumina GAIIx/HiSeq 2000
- Read length (100 bp)
- Insert size 300 - 500 bp
Welcome to ACPFG Bioinformatics.
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:
Browse...

Otherwise: Enter a sequence in FASTA format:

Note: Query sequence must be less than 5000 nucleotides.

Species selection
Please choose a query species:
Barley
Cereals
Dipotaxis
Hirschfeldia
Leptosperma
Lotus
Nicotiana
Pongonia

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

B. rapa chifu - 36 - 300 - BrC_03_002
B. rapa chifu - 35 - 2700 - BrC_22_001
B. rapa chifu - 35 - 2800 - BrC_37_001
B. rapa chifu - 35 - 2800 - BrC_37_002
B. rapa chifu - 35 - 2800 - BrC_37_003
B. rapa chifu - 35 - 300 - BrC_37_002
B. rapa chifu - 35 - 300 - BrC_37_003
B. rapa chifu - 35 - 410 - BrC_03_001
B. rapa chifu - 36 - 410 - BrC_03_001
B. nigra - 76 - 2700 - Bn_37_001
B. oleracea - 76 - 3000 - Bo_37_001

Format: SourceName - ReadLength - insertsize - LibraryName

Start


TAGdb output
TAGdb – Gene discovery

Sequencing SYM genes in Brassicas: NSP2

Alice Hayward
Jacqueline Batley
Brassica genomes

- Illumina GAIIx and Hi-Seq data for:
  - *B. rapa* BA, XA
  - *B. oleracea* BC
  - 8 *B. napus* cultivars
  - Wild Brassica species
Repeat detection and annotation

High-covered regions of short reads and their corresponding annotation.
The following 382 regions match your request.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
<th>Position</th>
<th>Match Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniRef90_Q6GUQ5</td>
<td>protein:blastp</td>
<td>Brassinosteroid LRR receptor kinase n:4 Tax:Solanum RepID:BRI1_SOLLCC</td>
<td>Chr1_BA_V4.0:227116..230562</td>
<td>10</td>
</tr>
<tr>
<td>UniRef90_Q6GUQ5_1</td>
<td>cds:blastp</td>
<td>Brassinosteroid LRR receptor kinase n:4 Tax:Solanum RepID:BRI1_SOLLCC</td>
<td>Chr1_BA_V4.0:227116..230562</td>
<td>10</td>
</tr>
<tr>
<td>Chr1_BA_V4.0.snap.416</td>
<td>gene:SNAP</td>
<td>Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH</td>
<td>Chr1_BA_V4.0:1284157..1288031</td>
<td>10</td>
</tr>
<tr>
<td>UniRef90_COLGS2_1</td>
<td>cds:blastp</td>
<td>Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH</td>
<td>Chr1_BA_V4.0:1284211..12884750</td>
<td>10</td>
</tr>
<tr>
<td>UniRef90_COLGS2</td>
<td>protein:blastp</td>
<td>Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH</td>
<td>Chr1_BA_V4.0:1284211..1288028</td>
<td>10</td>
</tr>
<tr>
<td>UniRef90_COLGS2_2</td>
<td>cds:blastp</td>
<td>Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH</td>
<td>Chr1_BA_V4.0:1284826..1285635</td>
<td>10</td>
</tr>
<tr>
<td>UniRef90_COLGS2_3</td>
<td>cds:blastp</td>
<td>Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH</td>
<td>Chr1_BA_V4.0:1285759..1287093</td>
<td>10</td>
</tr>
<tr>
<td>UniRef90_COLGS2_4</td>
<td>cds:blastp</td>
<td>Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH</td>
<td>Chr1_BA_V4.0:1287516..1288028</td>
<td>10</td>
</tr>
</tbody>
</table>
Candidate gene discovery

Genetic map

Physical map

Physical scaffolds
Candidate gene discovery

- Sequencing confirmed the presence of a large (402 nt) insertion in the 3' region of Ag Spectrum

**Skipton**

**Ag Spectrum**

Insertion in 3’ UTR
Stop codon at position 44 – truncated protein
B. napus SNP discovery

- Illumina paired end sequence from 8 B. napus cultivars
- Map reads to reference using SOAP
- Identify varietal genomic SNPs using custom algorithm
  - input: BAM files
  - output: text, goldengate, GFF3, and VCF
- Identified > 1 million SNPs
- Validated accuracy > 96%
B. napus 6K infinium

- 5306 genome wide SNPs
- Genotype >2000 lines
SNP density across A01

2 Mb

100 Kbp
Wheat genomic SNP discovery
SNP density

• Causes of SNP density variation

  • Constraints on variation tolerance in expressed genes

  • Breeding and selection, reducing SNP density in regions and fixing alleles

  • Diverse crossing/introgression, increasing variation in regions
SNP density around genes

Expressed gene
SNP density around genes

10Kbp 5'

10Kbp 3'
Which genes have low SNP density

**Example genes:**

- Q6NL05 Ethylene-responsive transcription factor ERF015
- P49592 Protein Dr1 homolog
- O49550 Dof zinc finger protein DOF4.5
- O95780 zinc finger protein 682
- Q7FJS2 Homeobox-leucine zipper protein HDG5
- Q8GXT3 Transcription factor bHLH123
- Q9SAH7 Probable WRKY transcription factor 40
- Q9SZ69 Zinc finger A20 and AN1 domain-containing stress-associated protein 7
- Q39081 Transcription factor CAULIFLOWER
- Q7XJK6 Agamous-like MADS-box protein AGL36
- Q9S7L5 Ethylene-responsive transcription factor ERF018
- Q1PDN3 Heat stress transcription factor A-6a
- Q9SJ41 Zinc finger CCCH domain-containing protein 18
- Q9FX25 Auxin response factor 13
- Q5RJC5 Zinc finger CCCH domain-containing protein 67
- Q8L500 Two-component response regulator-like PRRR9
- Q8GZ13 Transcription factor BEE 1
- Q8L7A4 Probable ADP-ribosylation factor GTPase-activating protein AGD11
- Q8SB4 WUSCHEL-related homeobox 3
- Q9SGJ6 Dehydration-responsive element-binding protein 1E
- Q38828 Auxin-responsive protein IAA10
- Q9LV52 Heat stress transcription factor C-1
- Q38827 Auxin-responsive protein IAA9
- Q6J9Q2 Ethylene-responsive transcription factor ERF086

**Gene classes:**

1. transcription (7.33)
2. ion transport (3.42)
3. signal transduction / membrane receptors (2.87)
4. F-box protein (2.73)
5. transposable element (2.27)
6. response to auxin / stimuli (2.03)
7. chromatin assembly, DNA damage/repair (1.93)
8. auxin transporter (1.91)
9. developmental processes (1.81)
10. immune response (1.71)
SNP density

- Causes of SNP density variation

  - Constraints on variation tolerance in expressed genes
  
  - Breeding and selection, reducing SNP density in regions and fixing alleles
  
  - Diverse crossing/introgression, increasing variation in regions
• Which alleles have been selected for in different germplasm?

• Have all favourable alleles been fixed?

• Have unfavourable alleles been dragged along for the ride?
  • What is the impact of linkage drag?

• Can this information be used for selecting parents/progeny?

• Can this be applied for breeding better canola?
Acknowledgements

Paul Berkman
Kenneth Chan
Chris Duran
Michael Imelfort
Kaitao Lai
Hong Lee
Edmund Ling
Michal Lorenc
Sahana Manoli
Pradeep Ruperio
Jiri Stiller

Dominic Eales
Lars Smits
Jacqueline Batley
Alice Hayward
Emma Campbell
Jessica Dalton-Morgan
Reece Tollenaere
Harsh Raman
Bart Lambert
Benjamin Laga

Contact:
Dave.Edwards@uq.edu.au