THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

Finding function in complex crop genomes

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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

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

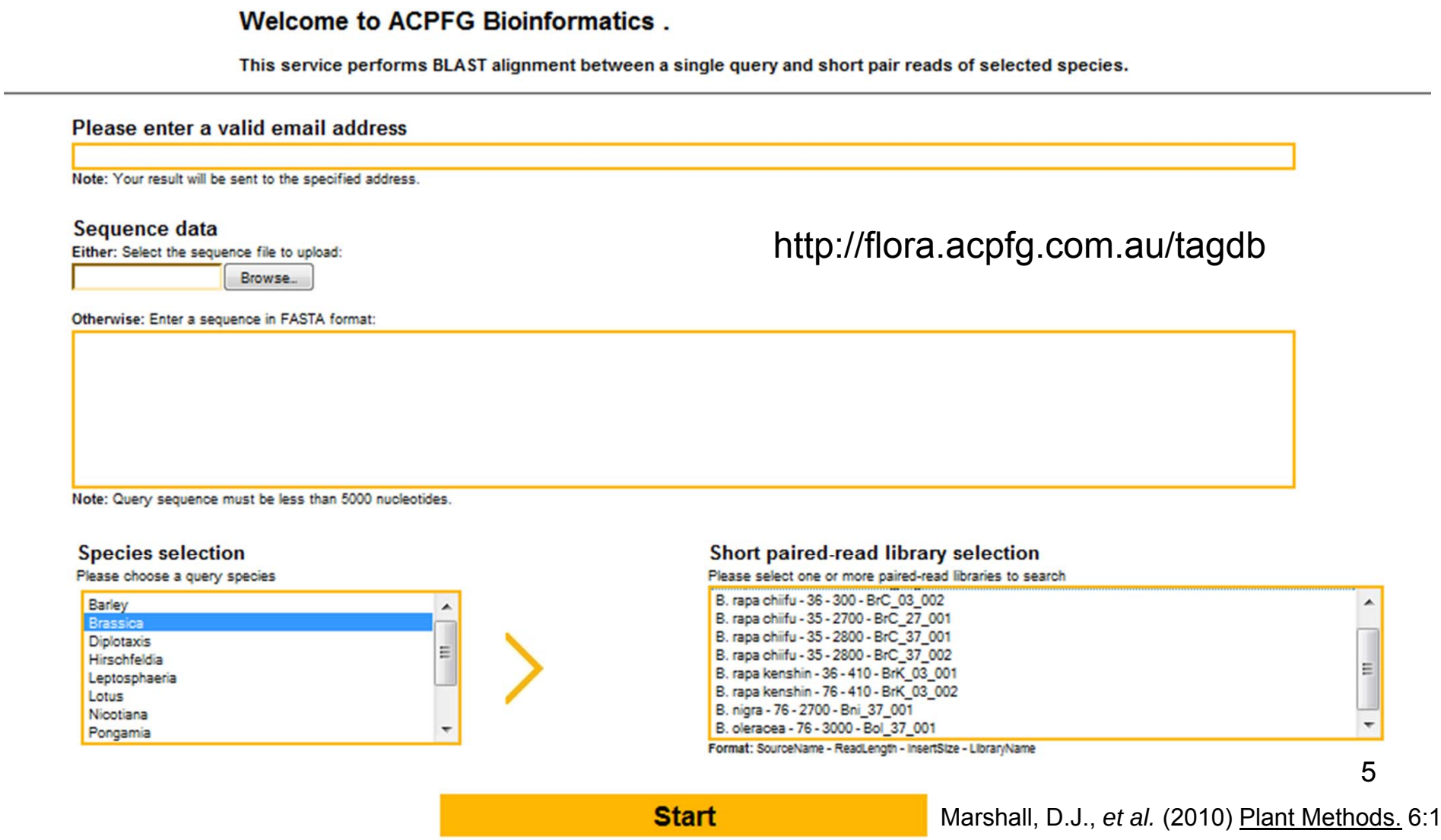
Illumina paired reads

- Illumina GAIIx/HiSeq 2000
- Read length (100 bp)
- Insert size 300 - 500 bp

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

TAGdb



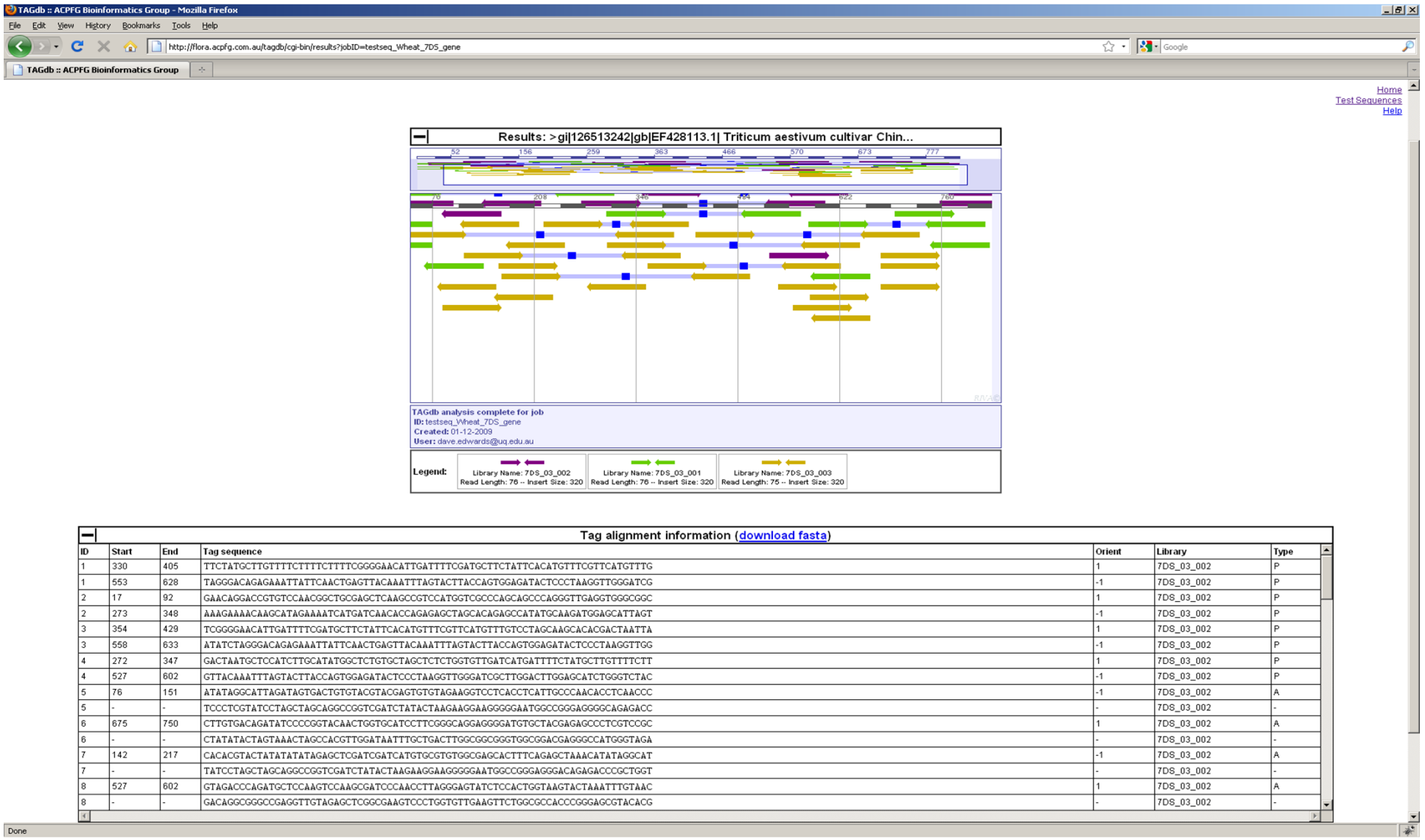
http://flora.acpfg.com.au/tagdb

Marshall, D.J., *et al.* (2010) Plant Methods. 6:19

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

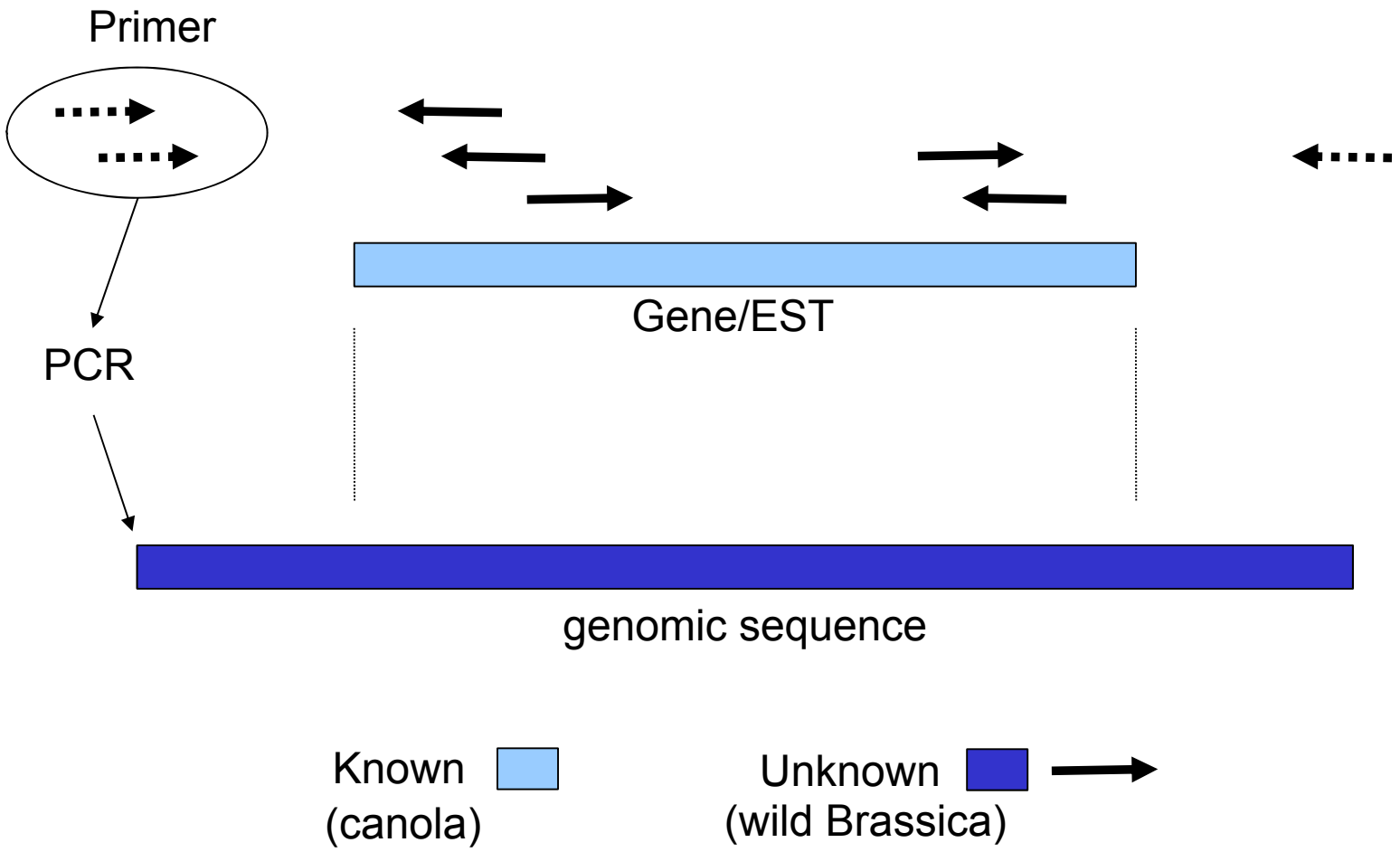
TAGdb output



Tag alignment information (download fasta)

ID	Start	End	Tag sequence	Orient	Library	Type
1	330	405	TTCTATGCTTGTTTTCTTTTCTTTTCGGGGAACATTGATTTTCGATGCTTCTATTCACATGTTTCGTTCATGTTTG	1	7DS_03_002	P
1	553	628	TAGGGACAGAGAAATTATTCAACTGAGTTACAAATTTAGTACTTACCAGTGGAGATACTCCCTAAGGTTGGGATCG	-1	7DS_03_002	P
2	17	92	GAACAGGACCGTGTCCAACGGCTGCGAGCTCAAGCCGTCCATGGTCGCCCAGCAGCCCAGGGTTGAGGTGGGCGGC	1	7DS_03_002	P
2	273	348	AAAGAAAACAAGCATAGAAAAATCATGATCAACACCAGAGAGCTAGCACAGAGCCATATGCAAGATGGAGCATTAGT	-1	7DS_03_002	P
3	354	429	TCGGGGAACATTGATTTTCGATGCTTCTATTCACATGTTTCGTTCATGTTTGTCCTAGCAAGCACACGACTAATTA	1	7DS_03_002	P
3	558	633	ATATCTAGGGACAGAGAAATTATTCAACTGAGTTACAAATTTAGTACTTACCAGTGGAGATACTCCCTAAGGTTGG	-1	7DS_03_002	P
4	272	347	GACTAATGCTCCATCTTGCATATGGCTCTGTGCTAGCTCTCTGGTGTTGATCATGATTTTCTATGCTTGTTTTCTT	1	7DS_03_002	P
4	527	602	GTTACAAATTTAGTACTTACCAGTGGAGATACTCCCTAAGGTTGGGATCGCTTGGACTTGGAGCATCTGGGTCTAC	-1	7DS_03_002	P
5	76	151	ATATAGGCATTAGATAGTGACTGTGTACGTACGAGTGTGTAGAAGGTCCTCACCTCATTGCCCAACACCTCAACCC	-1	7DS_03_002	A
5	-	-	TCCCTCGTATCCTAGCTAGCAGGCCGGTCGATCTATACTAAGAAGGAAGGGGGAATGGCCGGGAGGGGCAGAGACC	-	7DS_03_002	-
6	675	750	CTTGTGACAGATATCCCCGGTACAACTGGTGCATCCTTCGGGCAGGAGGGGATGTGCTACGAGAGCCCTCGTCCGC	1	7DS_03_002	A
6	-	-	CTATATACTAGTAAACTAGCCACGTTGGATAATTTGCTGACTTGGCGGCGGGTGGCGGACGAGGGCCATGGGTAGA	-	7DS_03_002	-
7	142	217	CACACGTACTATATATATAGAGCTCGATCGATCATGTGCGTGTGGCGAGCACTTTCAGAGCTAAACATATAGGCAT	-1	7DS_03_002	A
7	-	-	TATCCTAGCTAGCAGGCCGGTCGATCTATACTAAGAAGGAAGGGGGAATGGCCGGGAGGGACAGAGACCCGCTGGT	-	7DS_03_002	-
8	527	602	GTAGACCCAGATGCTCCAAGTCCAAGCGATCCCAACCTTAGGGAGTATCTCCACTGGTAAGTACTAAATTTGTAAC	1	7DS_03_002	A
8	-	-	GACAGGCGGGCCGAGGTTGTAGAGCTCGGCGAAGTCCCTGGTGTTGAAGTTCTGGCGCCACCCGGGAGCGTACACG	-	7DS_03_002	-

TAGdb – Gene discovery

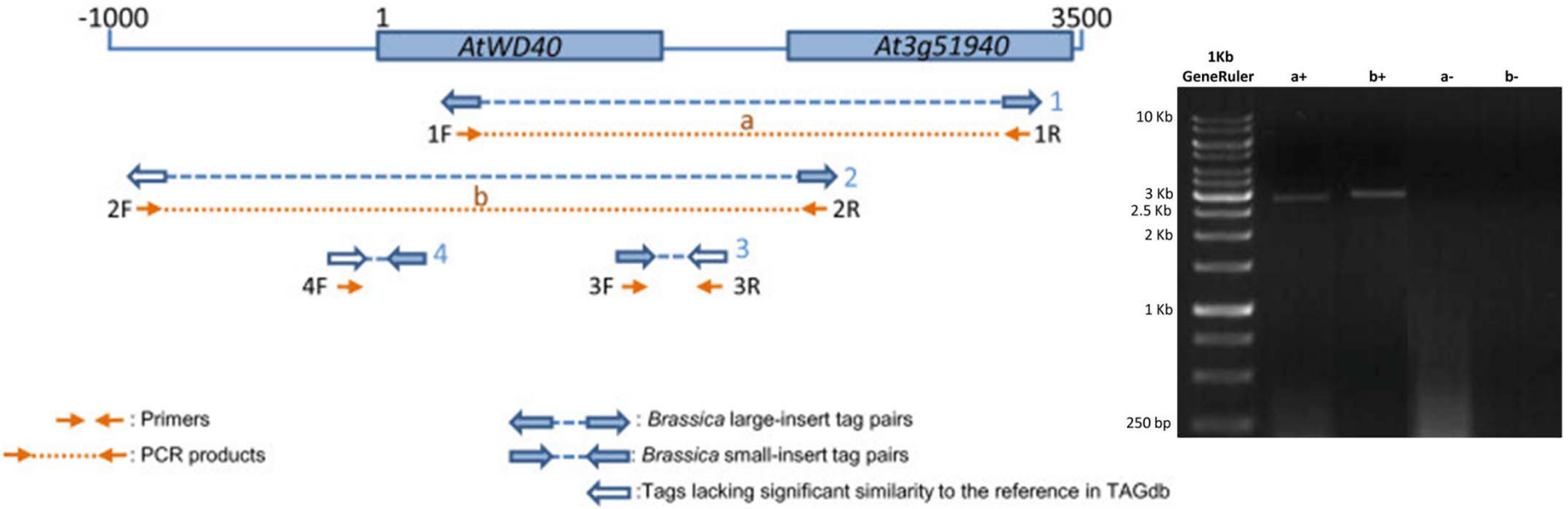


Marshall, D.J., *et al.* (2010) Plant Methods. 6:19

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

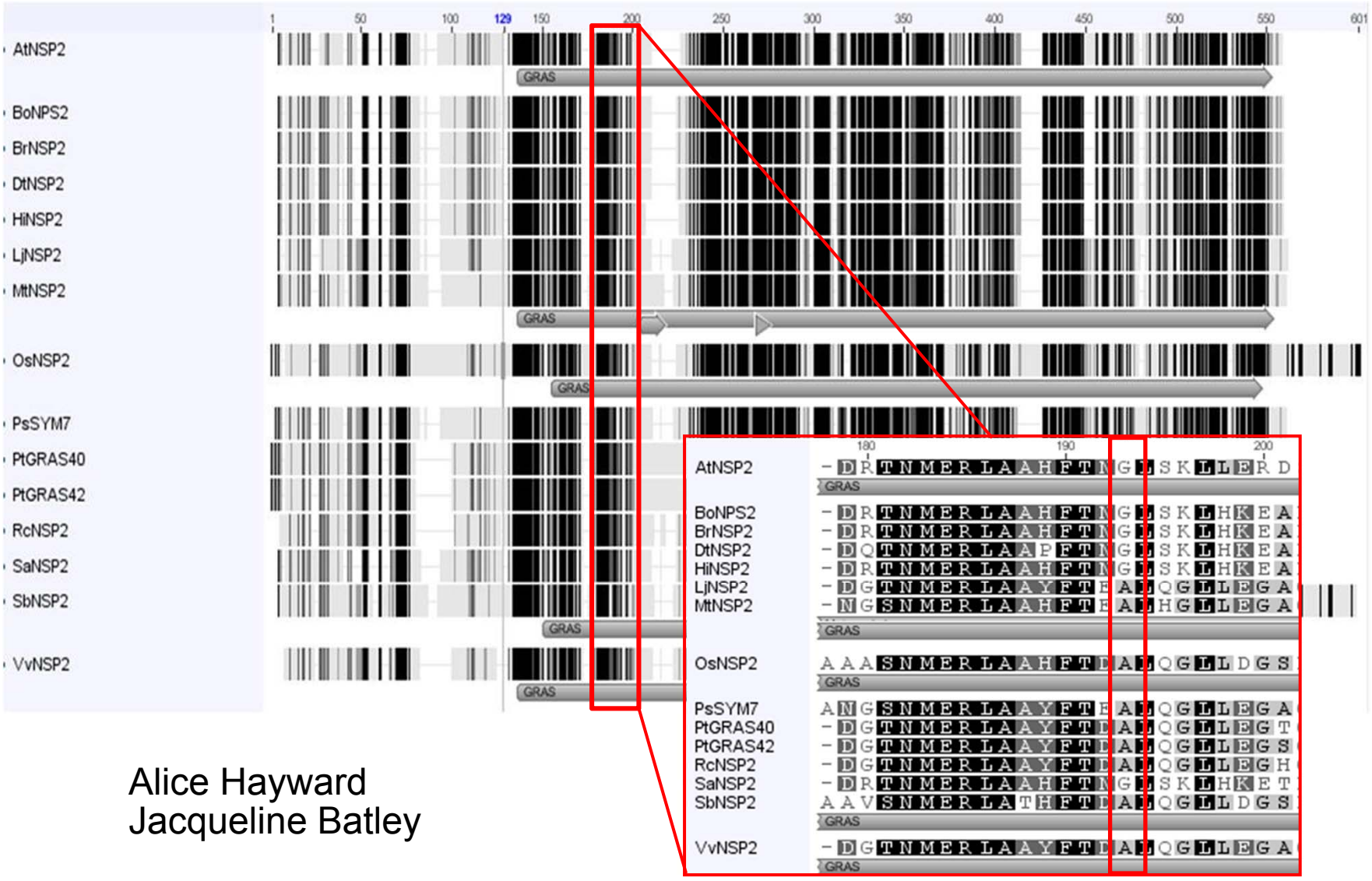
TAGdb



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

Sequencing *SYM* genes in *Brassicas*: *NSP2*

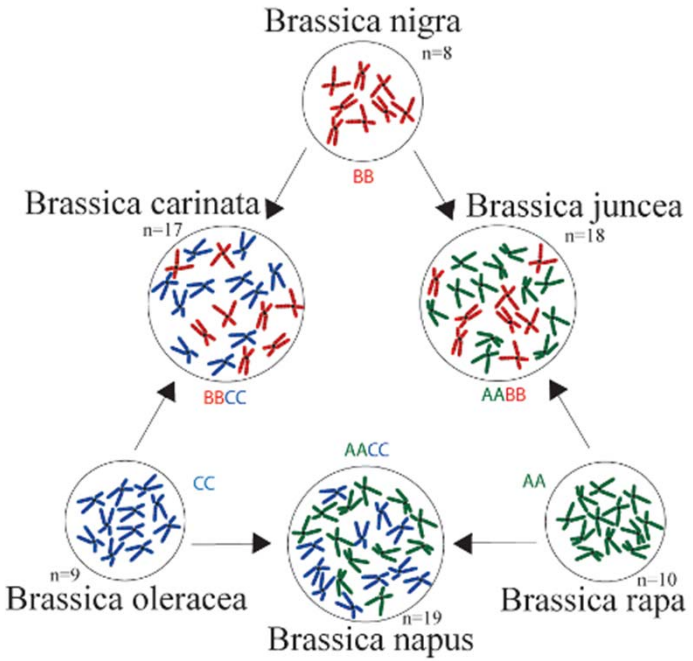


Alice Hayward
Jacqueline Batley

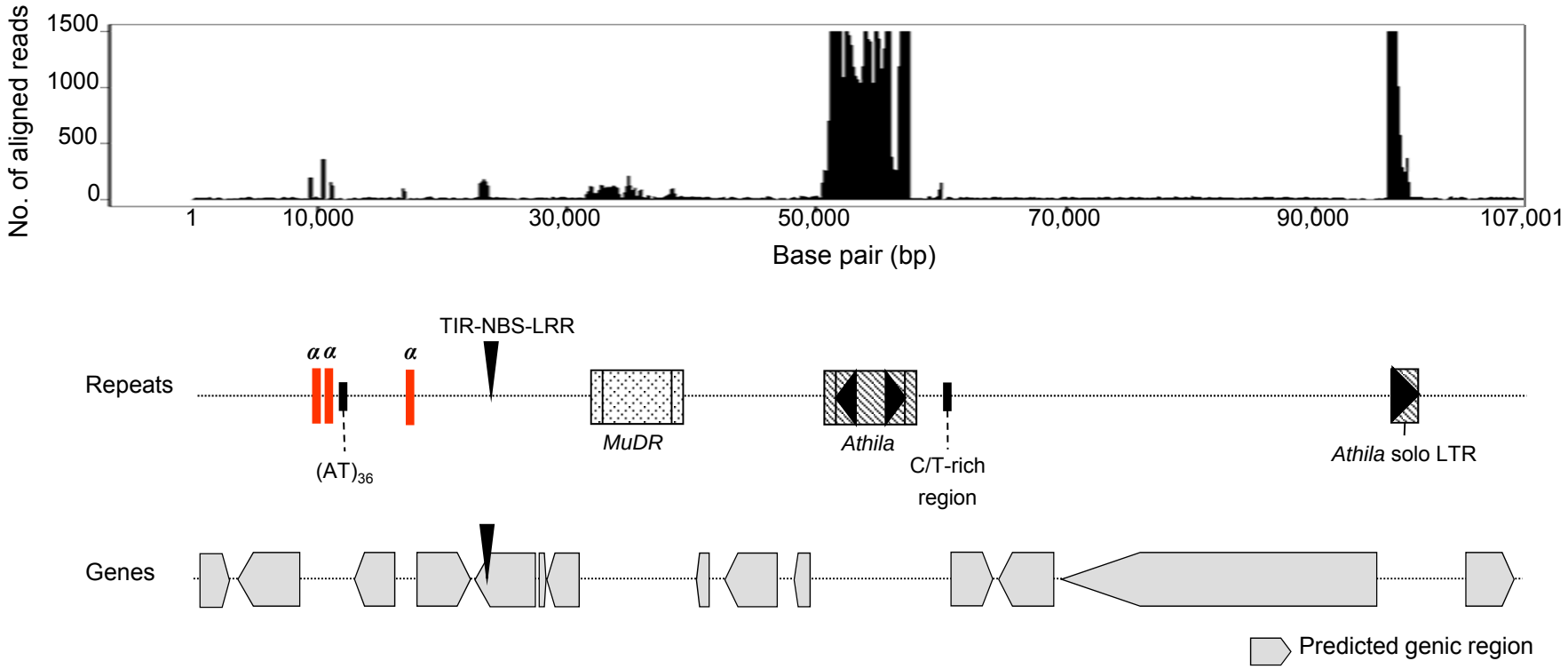
Brassica genomes

- 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

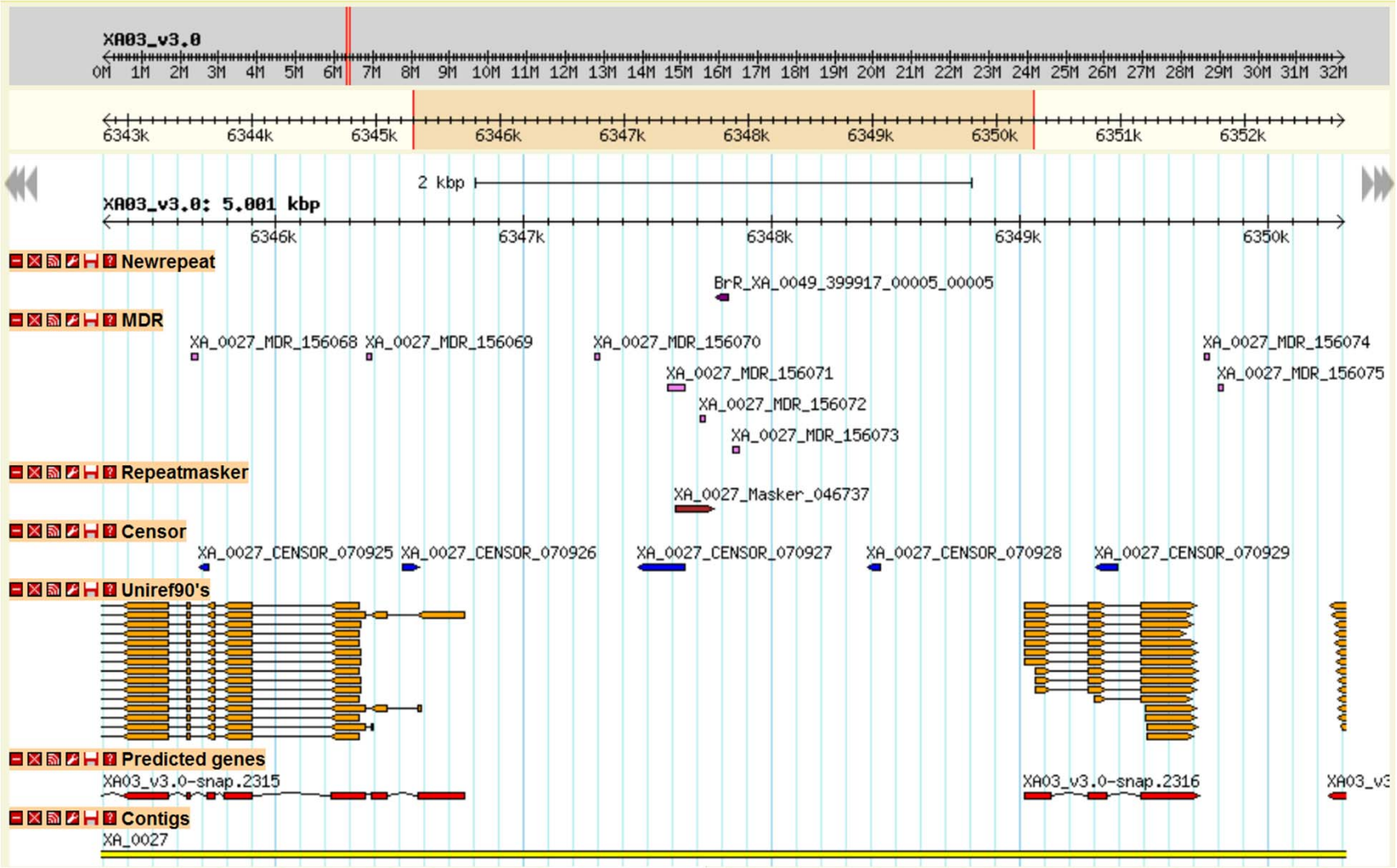


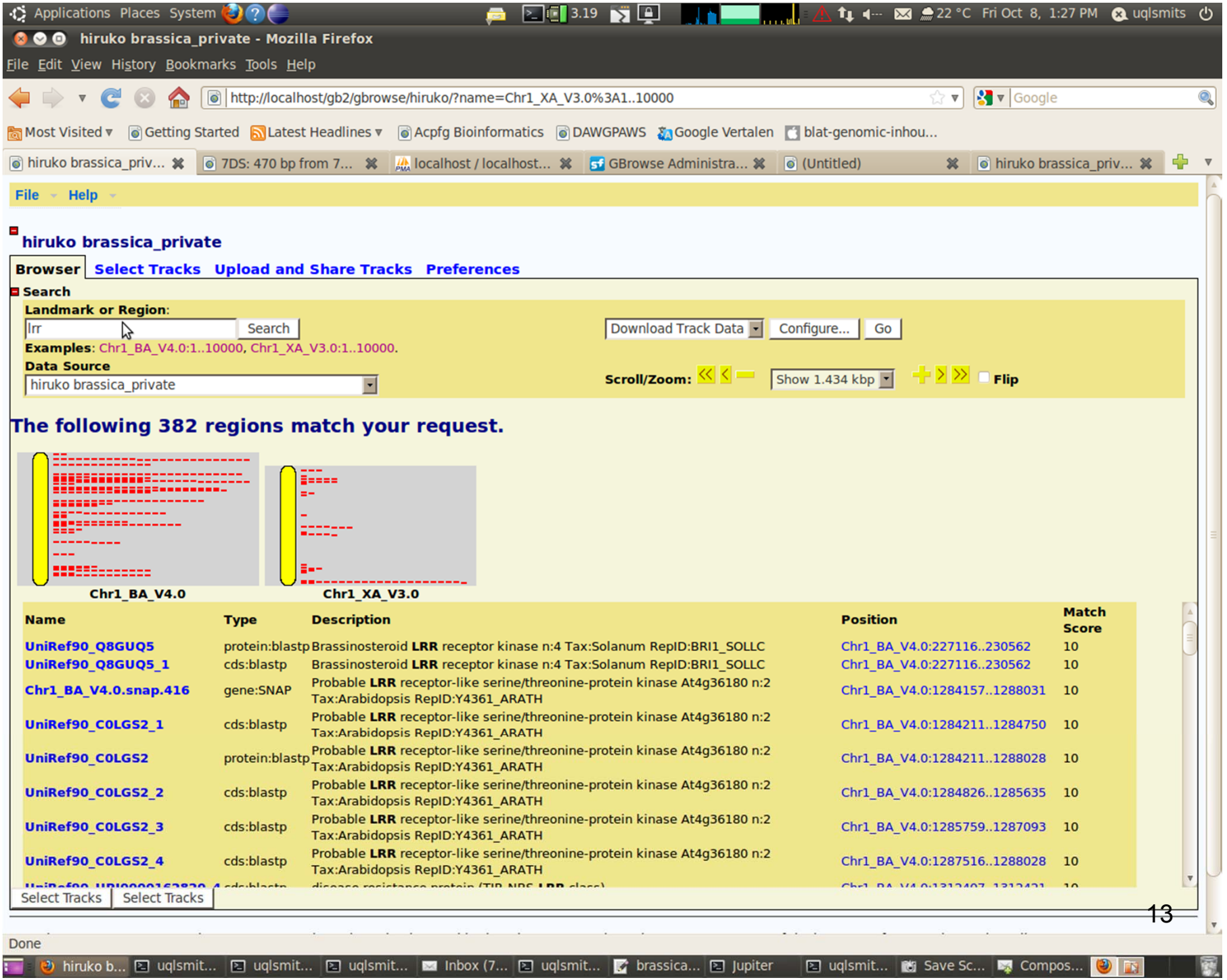
High-covered regions of short reads and their corresponding annotation.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

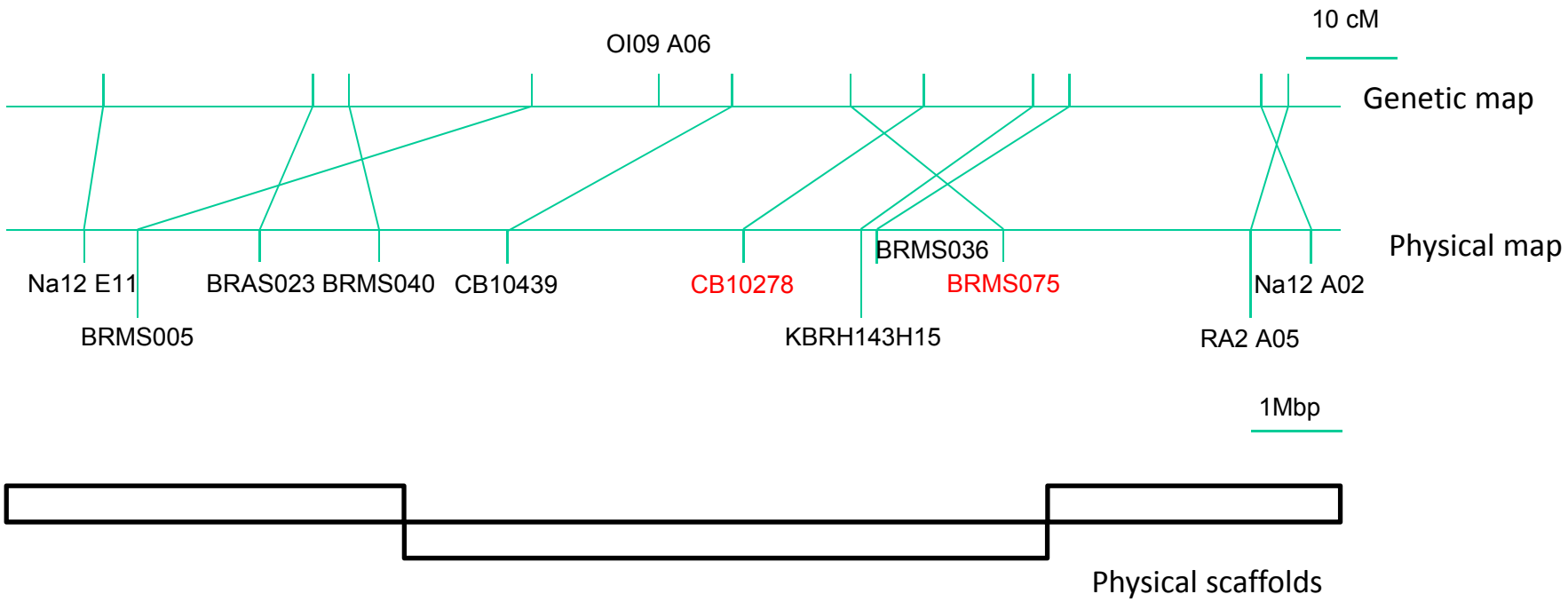
ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

www.brassicagenome.net




Applications Places System
hiruko brassica_private - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/gb2/gbrowse/hiruko/?name=Chr1_XA_V3.0%3A1..10000
Most Visited Getting Started Latest Headlines Acpfg Bioinformatics DAWGPAWS Google Vertalen blat-genomic-inhou...
hiruko brassica_priv... 7DS: 470 bp from 7... localhost / localhost... GBrowse Administra... (Untitled) hiruko brassica_priv...
File Help
hiruko brassica_private
Browser Select Tracks Upload and Share Tracks Preferences
Search
Landmark or Region:
lrr
Search
Examples: Chr1_BA_V4.0:1..10000, Chr1_XA_V3.0:1..10000.
Data Source
hiruko brassica_private
Download Track Data
Configure...
Go
Scroll/Zoom:
Show 1.434 kbp
Flip
The following 382 regions match your request.
Chr1_BA_V4.0
Chr1_XA_V3.0
Name
Type
Description
Position
Match Score
UniRef90_Q8GUQ5 protein:blastp Brassinosteroid LRR receptor kinase n:4 Tax:Solanum RepID:BRI1_SOLLC Chr1_BA_V4.0:227116..230562 10
UniRef90_Q8GUQ5_1 cds:blastp Brassinosteroid LRR receptor kinase n:4 Tax:Solanum RepID:BRI1_SOLLC Chr1_BA_V4.0:227116..230562 10
Chr1_BA_V4.0.snap.416 gene:SNAP Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH Chr1_BA_V4.0:1284157..1288031 10
UniRef90_C0LGS2_1 cds:blastp Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH Chr1_BA_V4.0:1284211..1284750 10
UniRef90_C0LGS2 protein:blastp Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH Chr1_BA_V4.0:1284211..1288028 10
UniRef90_C0LGS2_2 cds:blastp Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH Chr1_BA_V4.0:1284826..1285635 10
UniRef90_C0LGS2_3 cds:blastp Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH Chr1_BA_V4.0:1285759..1287093 10
UniRef90_C0LGS2_4 cds:blastp Probable LRR receptor-like serine/threonine-protein kinase At4g36180 n:2 Tax:Arabidopsis RepID:Y4361_ARATH Chr1_BA_V4.0:1287516..1288028 10
Select Tracks
Select Tracks
Done

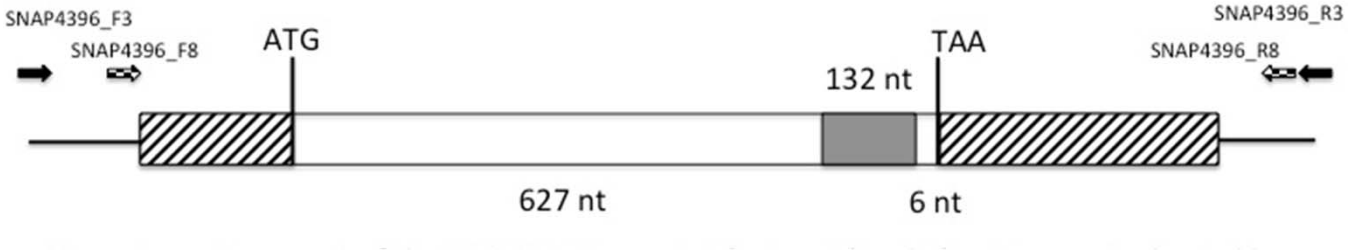
Candidate gene discovery



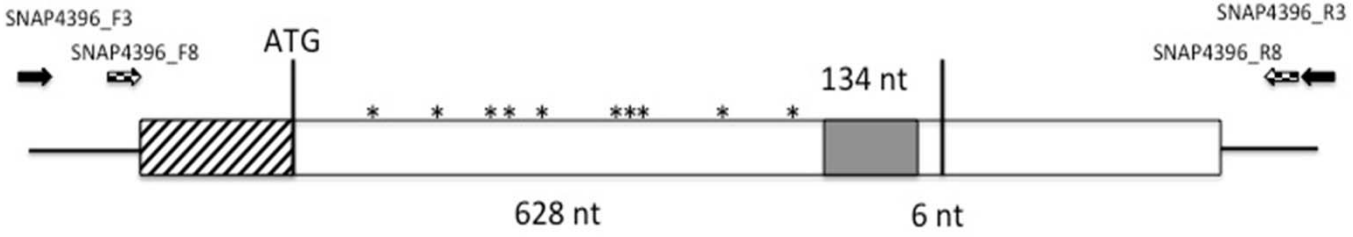
Candidate gene discovery

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

Skipton



Ag Spectrum



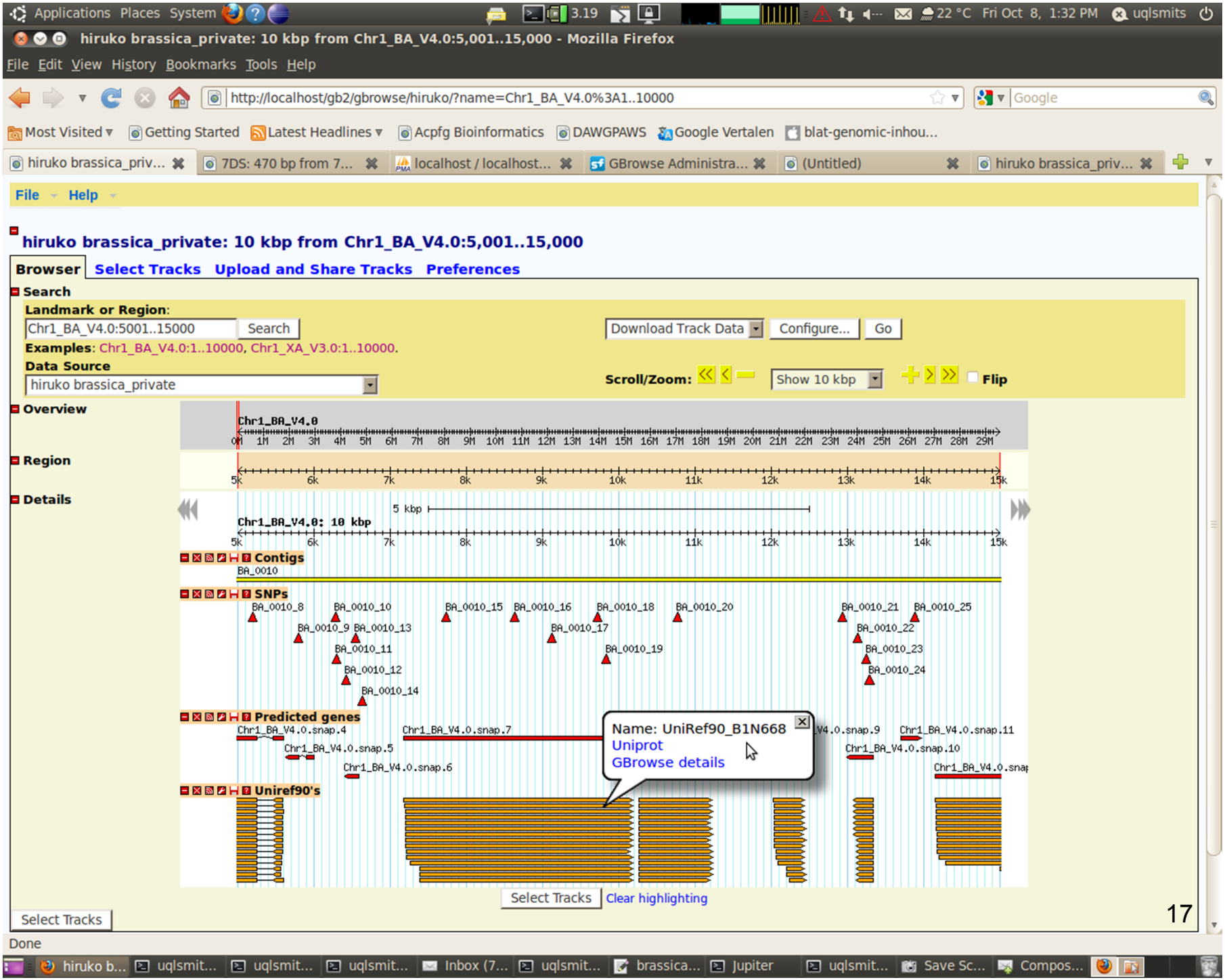
Insertion in 3' UTR
Stop codon at position 44 – truncated protein

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

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%

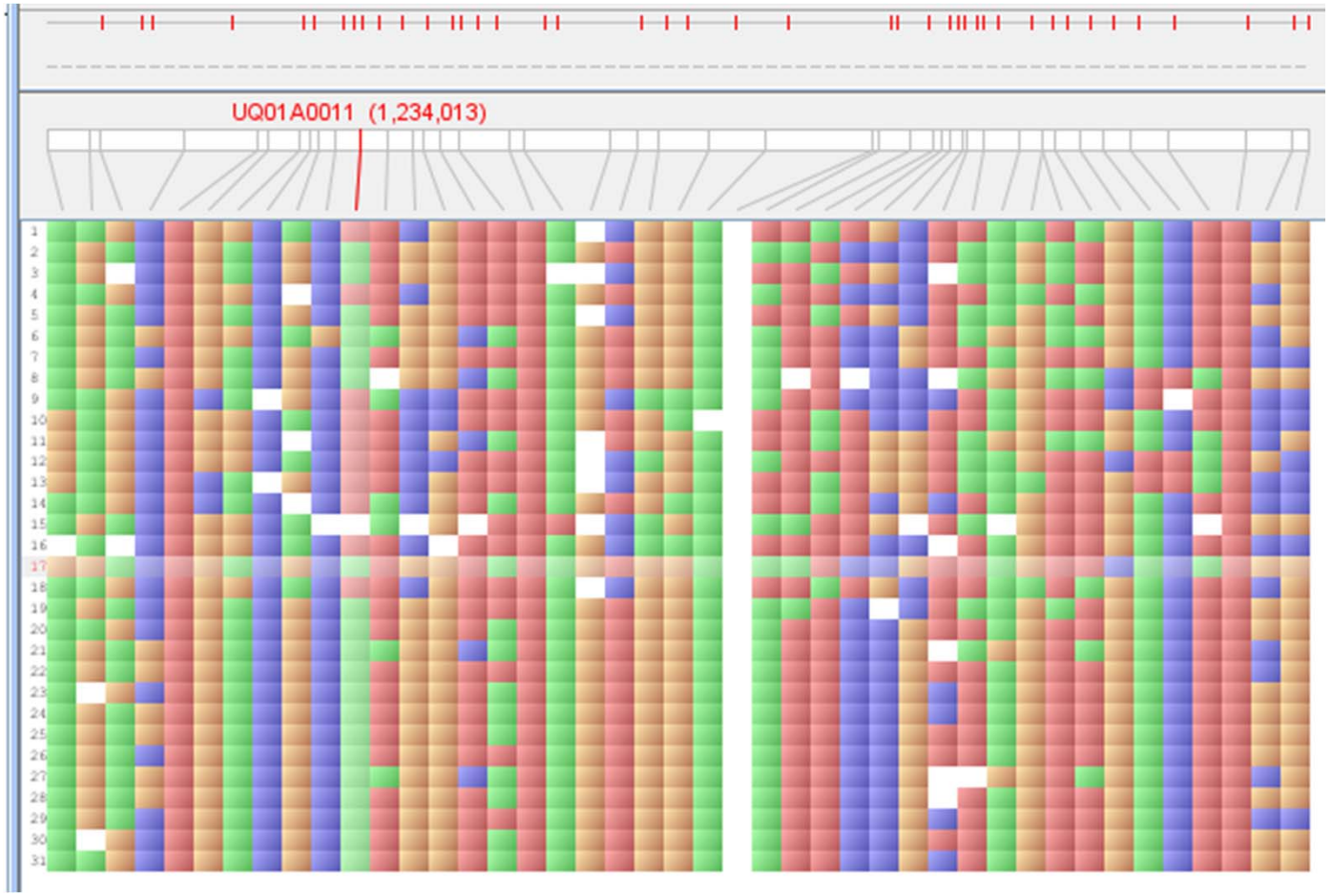

Applications Places System
hiruko brassica_private: 10 kbp from Chr1_BA_V4.0:5,001..15,000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/gb2/gbrowse/hiruko/?name=Chr1_BA_V4.0%3A1..10000
File Help
hiruko brassica_private: 10 kbp from Chr1_BA_V4.0:5,001..15,000
Browser Select Tracks Upload and Share Tracks Preferences
Search
Landmark or Region:
Chr1_BA_V4.0:5001..15000
Search
Download Track Data
Configure...
Go
Examples: Chr1_BA_V4.0:1..10000, Chr1_XA_V3.0:1..10000.
Data Source
hiruko brassica_private
Scroll/Zoom:
Show 10 kbp
Flip
Overview
Region
Details
Chr1_BA_V4.0
Chr1_BA_V4.0: 10 kbp
5 kbp
Contigs
BA_0010
SNPs
BA_0010_8
BA_0010_10
BA_0010_15
BA_0010_16
BA_0010_18
BA_0010_20
BA_0010_21
BA_0010_25
BA_0010_9
BA_0010_13
BA_0010_17
BA_0010_22
BA_0010_11
BA_0010_19
BA_0010_23
BA_0010_12
BA_0010_24
BA_0010_14
Predicted genes
Chr1_BA_V4.0.snap.4
Chr1_BA_V4.0.snap.7
Chr1_BA_V4.0.snap.9
Chr1_BA_V4.0.snap.11
Chr1_BA_V4.0.snap.5
Chr1_BA_V4.0.snap.10
Chr1_BA_V4.0.snap.6
Name: UniRef90_B1N668
Uniprot
GBrowse details
Uniref90's
Select Tracks
Clear highlighting
Select Tracks
Done

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

B. napus 6K infinium

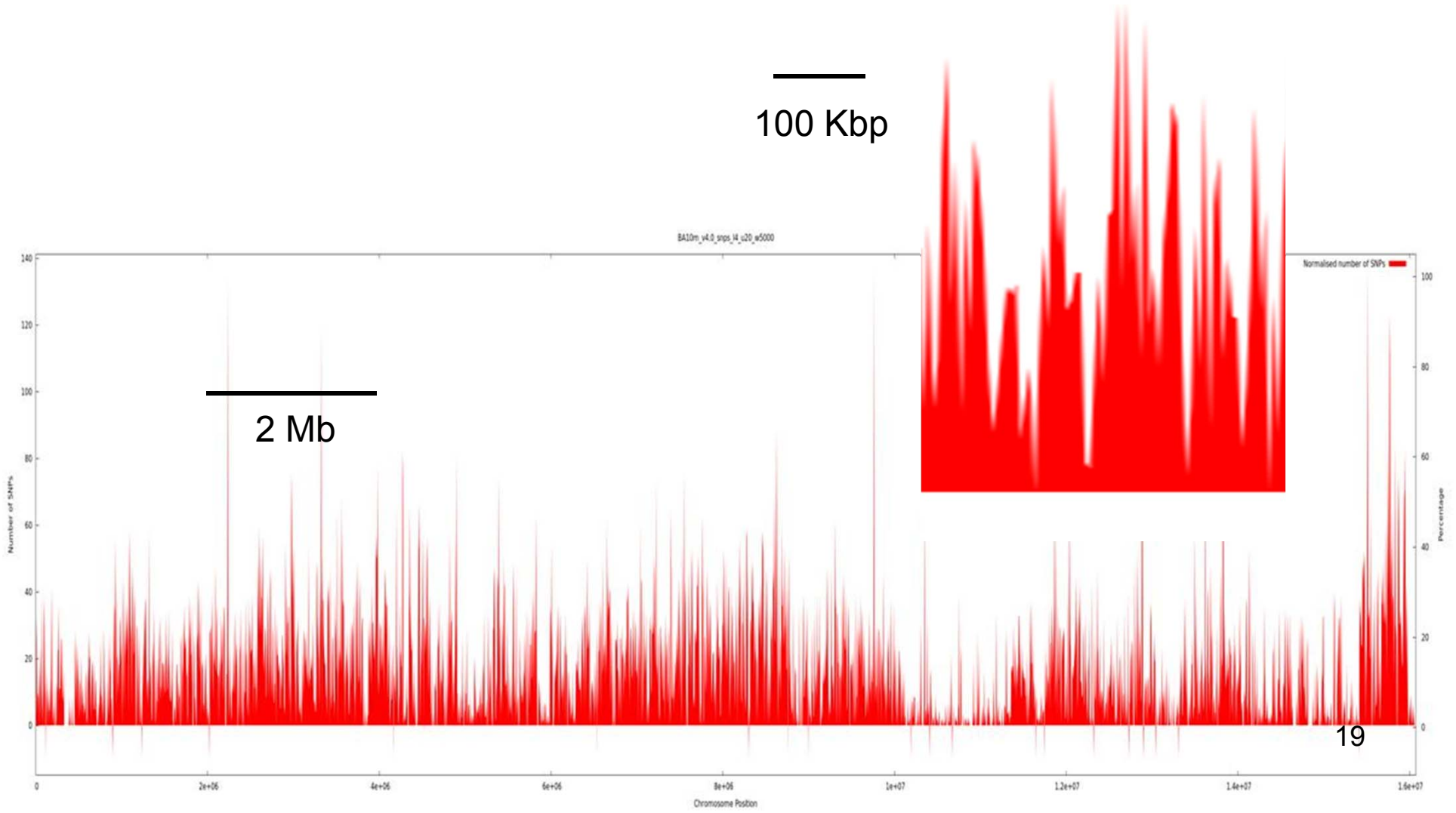
- 5306 genome wide SNPs
- Genotype >2000 lines



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

SNP density across A01



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

Wheat genomic SNP discovery

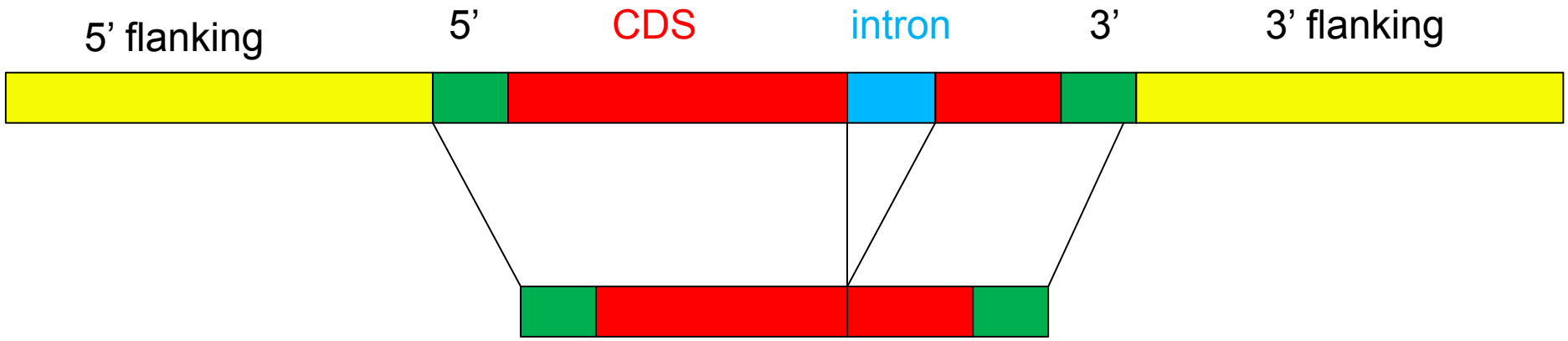
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

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

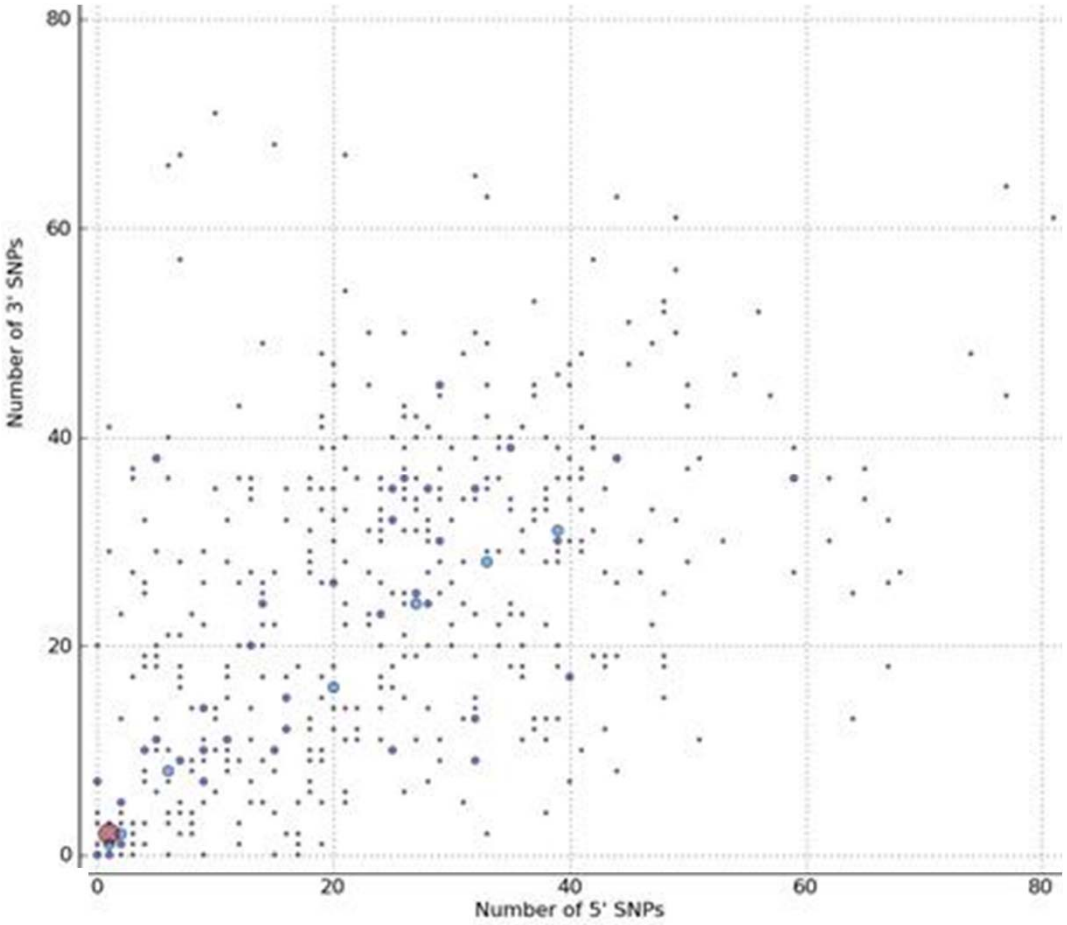
SNP density around genes



Expressed gene

SNP density around genes

10Kbp 3’



10Kbp 5’

Which genes have low SNP density

Example genes:

Q6NLD5	Ethylene-responsive transcription factor ERF015
P49592	Protein Dr1 homolog
O49550	Dof zinc finger protein DOF4.5
O95780	zinc finger protein 682
Q9FJS2	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 APRR9
Q8GZ13	Transcription factor BEE 1
Q8L7A4	Probable ADP-ribosylation factor GTPase-activating protein AGD11
Q9SIB4	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

Finding function

- 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