Finding function in complex genomes

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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

Outline

- Assembled Brassica genomes
- Identify and annotate genes
- Identify genome variation
- Associate genome variation with traits

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

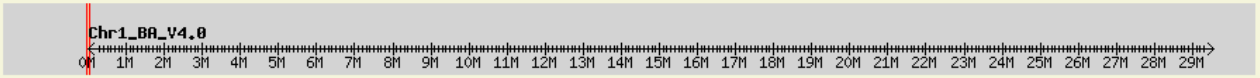
Examples: Chr1_BA_V4.0:1..10000, Chr1_XA_V3.0:1..10000.

Data Source: hiruko brassica_private

Download Track Data Configure... Go

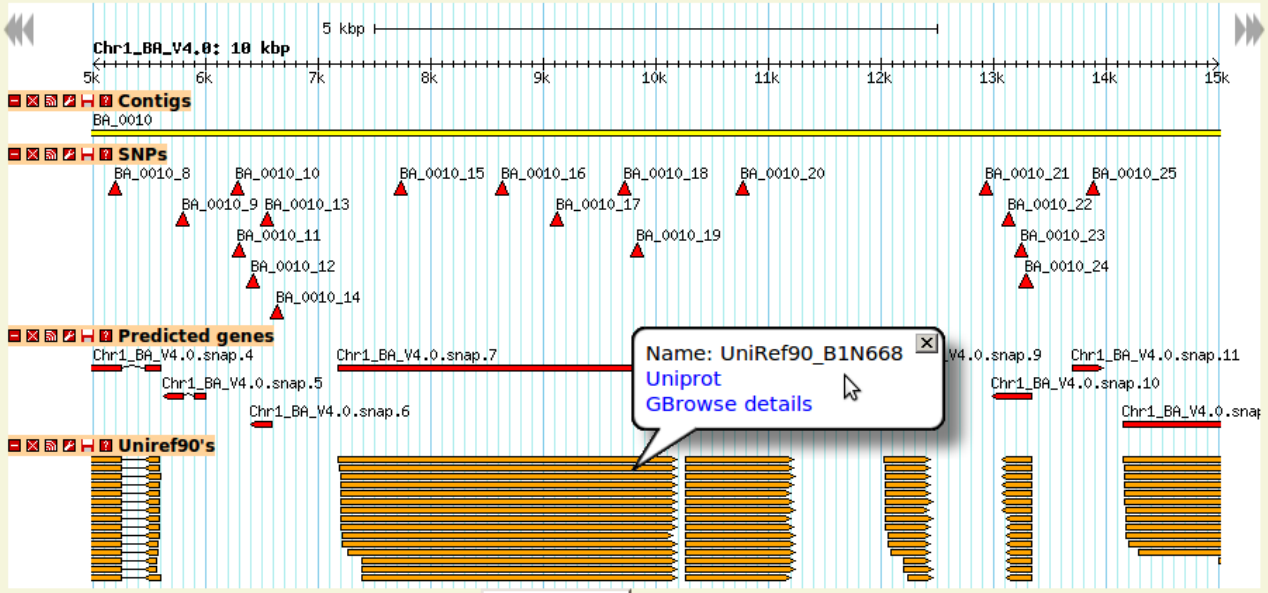
Scroll/Zoom: Show 10 kbp **Flip**

Overview



Region

Details



Select Tracks Clear highlighting

Select Tracks

Applications Places System 3.19 22 °C Fri Oct 8, 1:27 PM uqlsmits

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

hiruko b... | uqlsmit... | uqlsmit... | uqlsmit... | Inbox (7... | uqlsmit... | brassica... | Jupiter | uqlsmit... | Save Sc... | Compos...

Finding function

- Which alleles have been selected for in different germplasm?
- Can we associate genes/alleles with traits
- 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?

Finding function

- Genotyping by sequencing
- Screen broad germplasm
    - Diversity sets
    - Segregating populations
- Infinium array

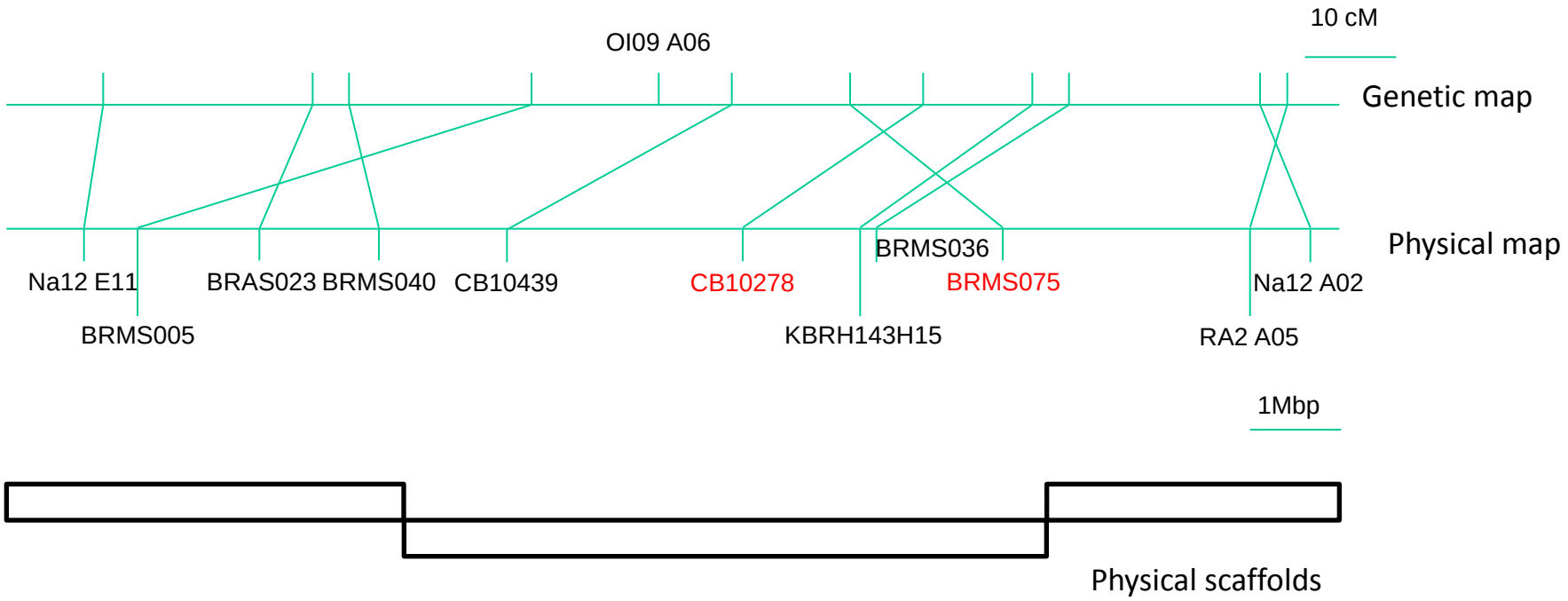
Brassica Sequencing

- Illumina GAIIx and Hi-Seq data for:

- *B. rapa* BA, XA
- *B. oleracea* BC
- 14 *B. napus* cultivars
- Diploid A, B and C
- Wild Brassica species

The *Brassica rapa* Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa. Nature Genetics*

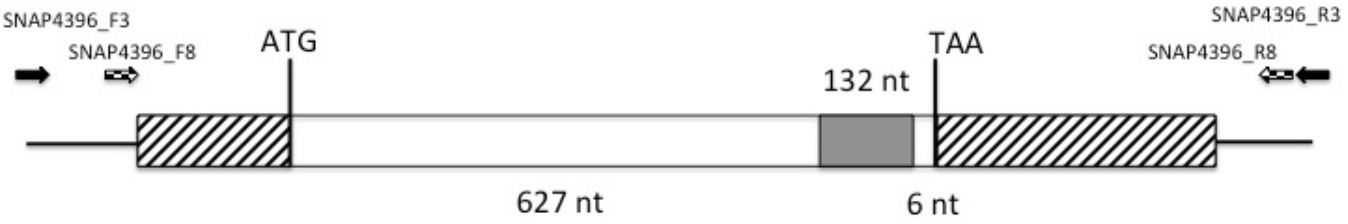
Candidate gene discovery



Candidate gene discovery

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

Skipton



Ag Spectrum

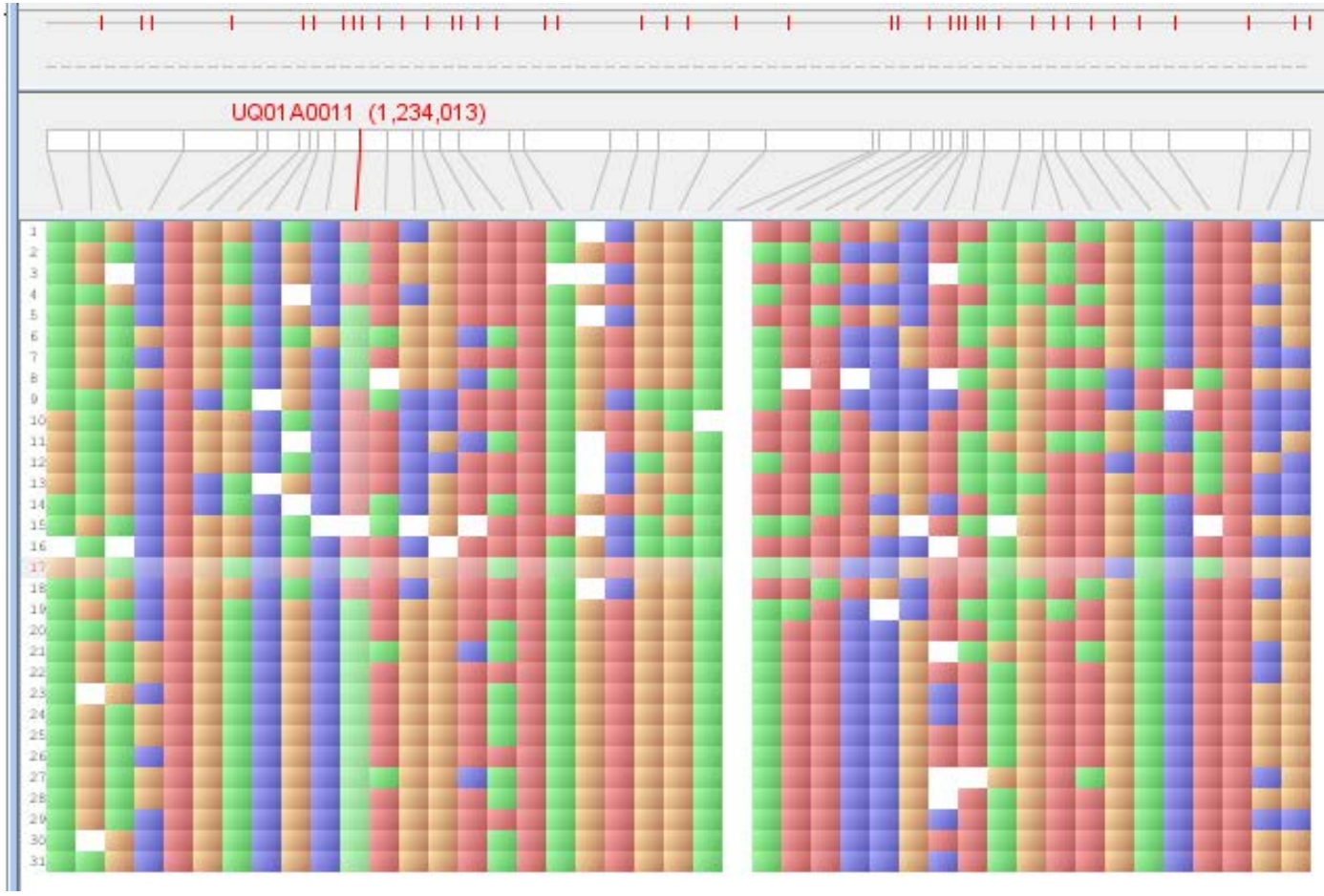
Insertion in 3' UTR
Stop codon at position 44 – truncated protein

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

B. napus 6K infinium

- 5306 genome wide SNPs
- Genotype >2000 lines



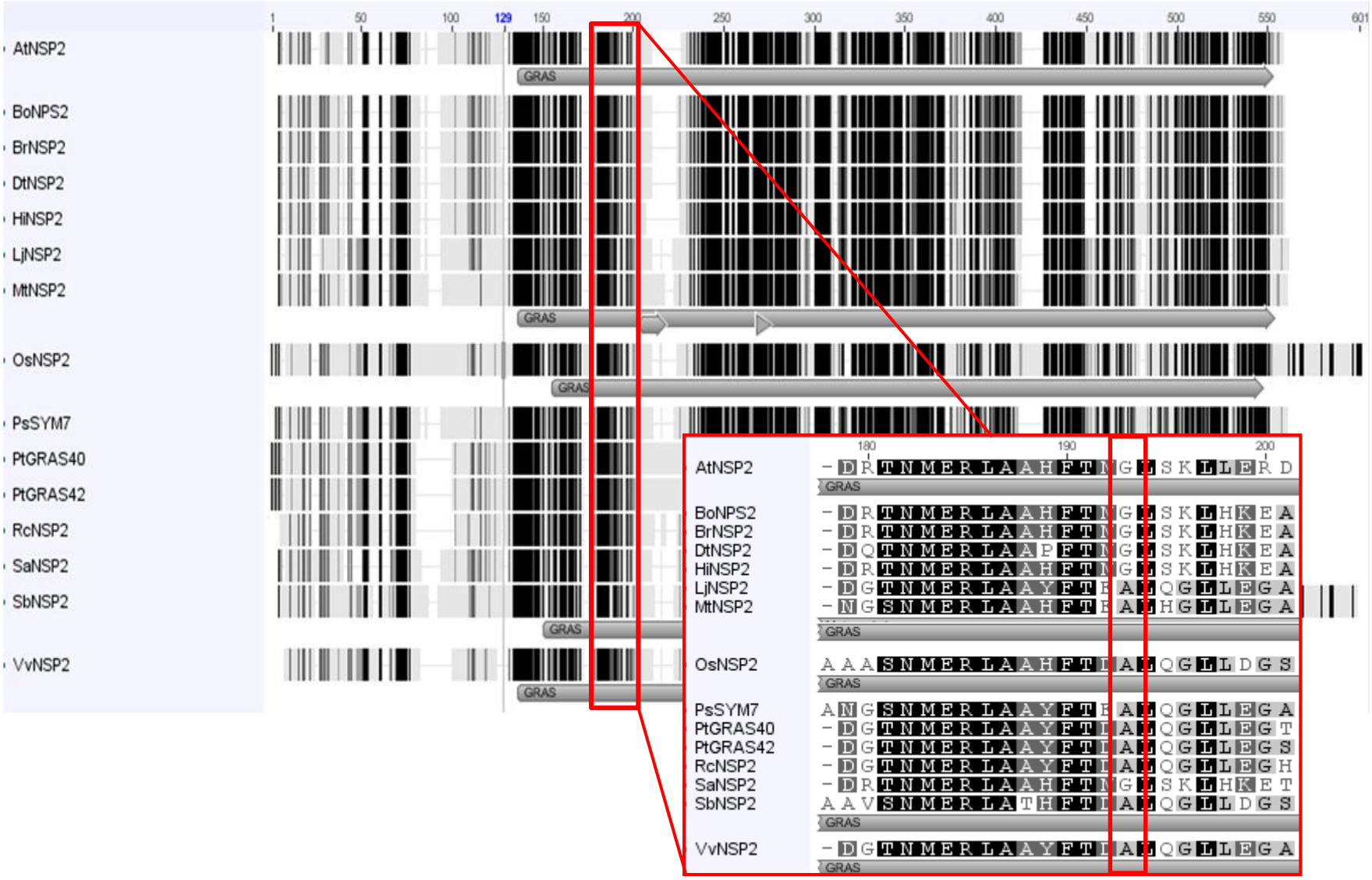
MBGP SNP consortium

- A, C and AC genomes
- 60,000 SNPs
- Genome wide coverage
- SNP content by 1 December 2011
- Delivery April 2012

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

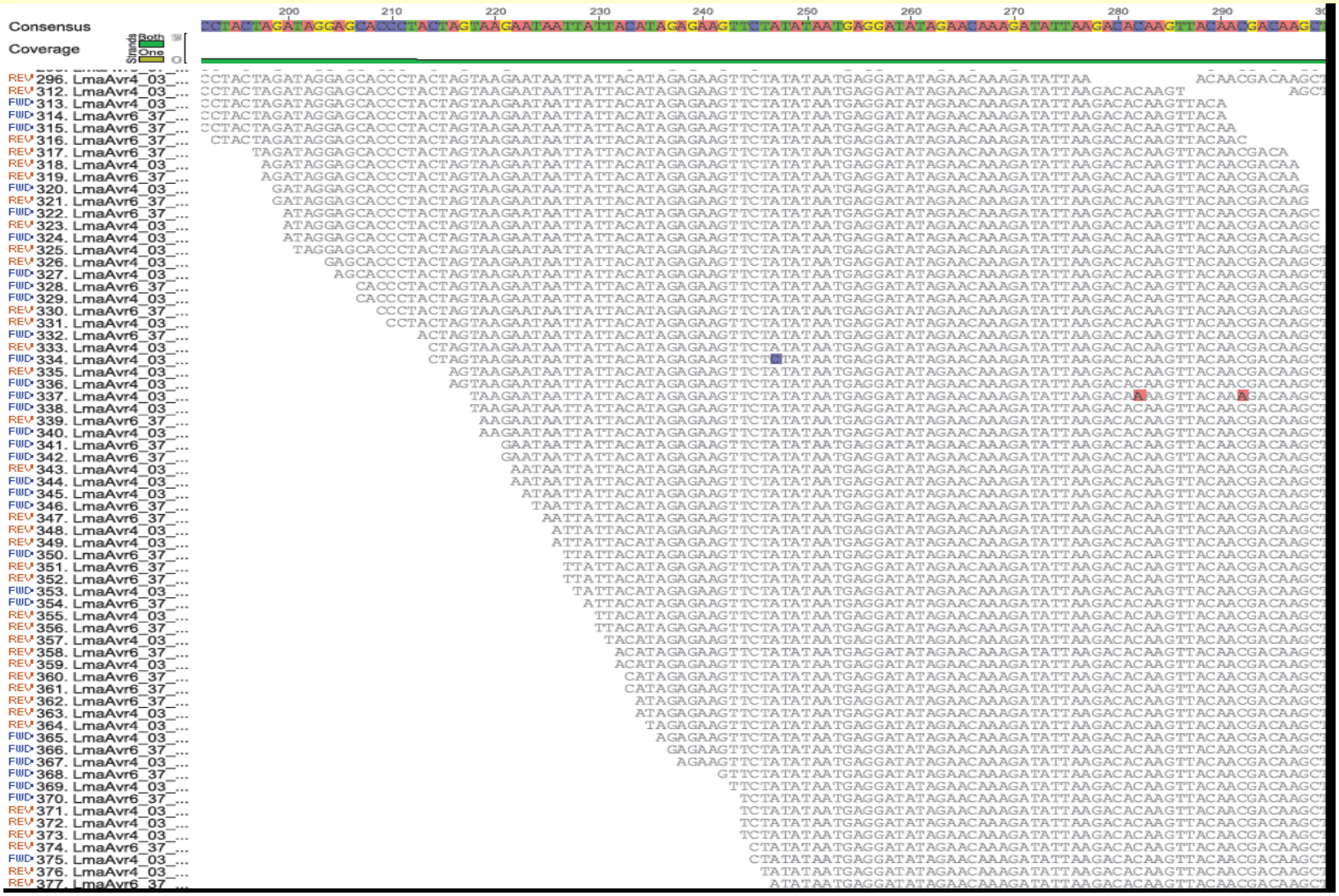
Sequencing *SYM* genes in *Brassicas*: *NSP2*



Sequencing

- Isolates sequenced
  - 06 MGPP641
  - 04 MGPS021
- Illumina GAIIx paired end reads
  - 06-41 -> 140x coverage
  - 04-21 -> 86x coverage
- Aligned to *L. maculans* genome
- SNP prediction performed

Alignment of tags to reference genome:



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

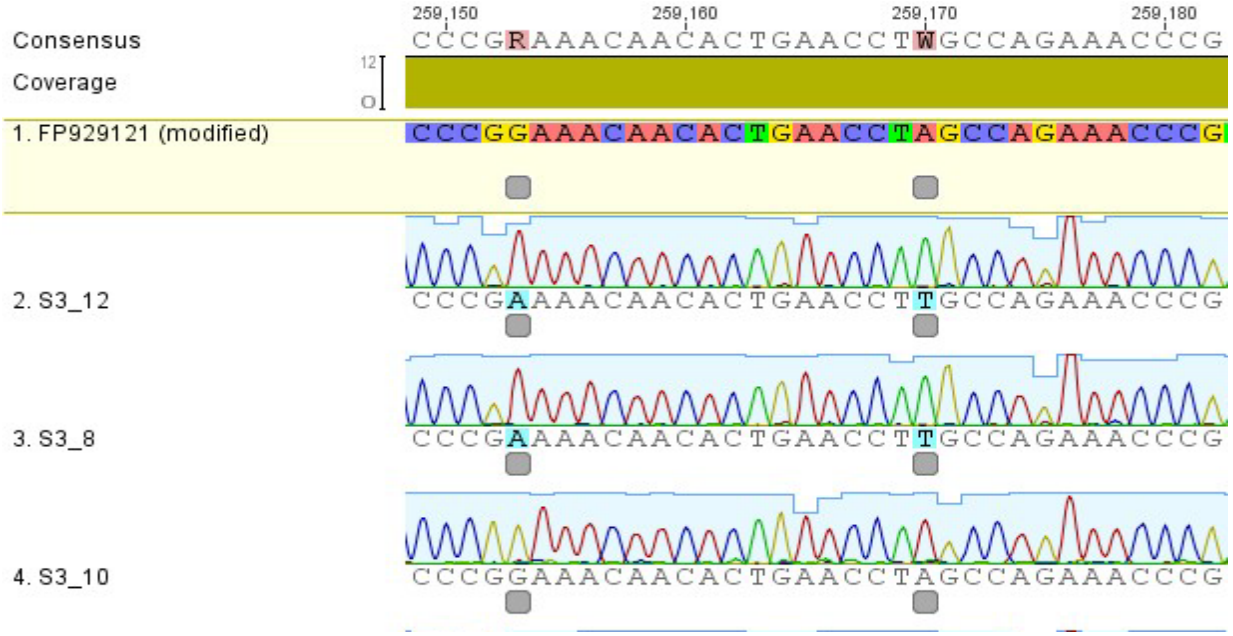
SNP identification

- SNPs predicted using SGSautoSNP
- 21,814 SNPs predicted
- 2.07 SNPs per kb

SNP type	Base change	Count	%
Transition	A > G	8905	40.82
Transition	C > T	8758	40.15
Transversion	A > C	1088	4.99
Transversion	A > T	1021	4.68
Transversion	C > G	1038	4.76
Transversion	G > T	1004	4.6

SNP Validation

- SNPs predicted using SGSautoSNP
- 20 SNPs validated using Sanger sequencing across 25 different isolates
- 384 SNPs chosen for Goldengate assay



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

SNP Validation

SNP number	SNP score	Ref	21	41	14
S1	14	G	A	G	G
S2	10	C	C	T	T
S3	18	A	A	T	T
S5	5	G	A	G	G
S6	11	A	T	A	T
S7	4	C	T	C	C
S8	2	A	G	A	A
S9	8	A	A	C	A
S11	11	G	G	T	n/a
S14	13	A	A	G	A
S15	23	G	G	A	G
S16	15	G	C	G	G
S17	8	C	C	G	C
S18	5	G	A	G	G

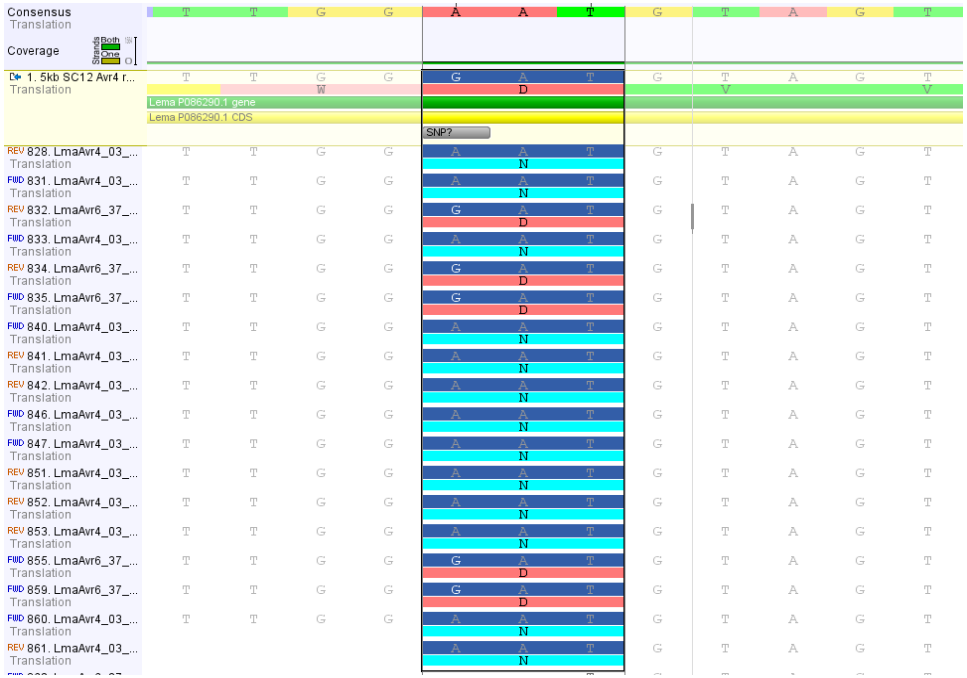
L. maculans Goldengate assay

- 96 different isolates (including replicates)
- 384 SNPs
  - Genome-wide spread
    - 76 Supercontigs

SNP number	SuperContig	Position	Genic location	SNP Type	Base changes	PIC
SNP1	0	369650	CDS	Transition	G >A	0.4352
SNP2	20	433691	CDS	Transition	C > T	0.1472
SNP3	25	259169	Intergenic	Transversion	A > T	0.4862
SNP5	5	996624	CDS	Transition	G > A	0.1472
SNP6	60	506	Intergenic	Transversion	A > T	0.4966
SNP7	9	1524477	Genic	Transition	C > T	0.3085
SNP8	9	789721	Intergenic	Transversion	A > G	0.0671
SNP9	2	838930	Intergenic	Transversion	A > C	0.3299
SNP11	4	1362010	Intergenic	Transversion	G > T	0.4898
SNP14	8	1296096	Intergenic	Transition	A > G	0.4959
SNP15	12	352844	CDS	Transition	G > A	0.4445
SNP16	12	838085	Intergenic	Transversion	G > C	0.2188
SNP17	12	1107690	Intergenic	Transversion	C > G	0.4133
SNP18	14	1268037	CDS	Transition	G > A	0.0643

SNPs in coding region

- Differences in tag sequences between 06-41 and 04-21
  - base change D (Aspartic acid) to N (Asparagine)



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

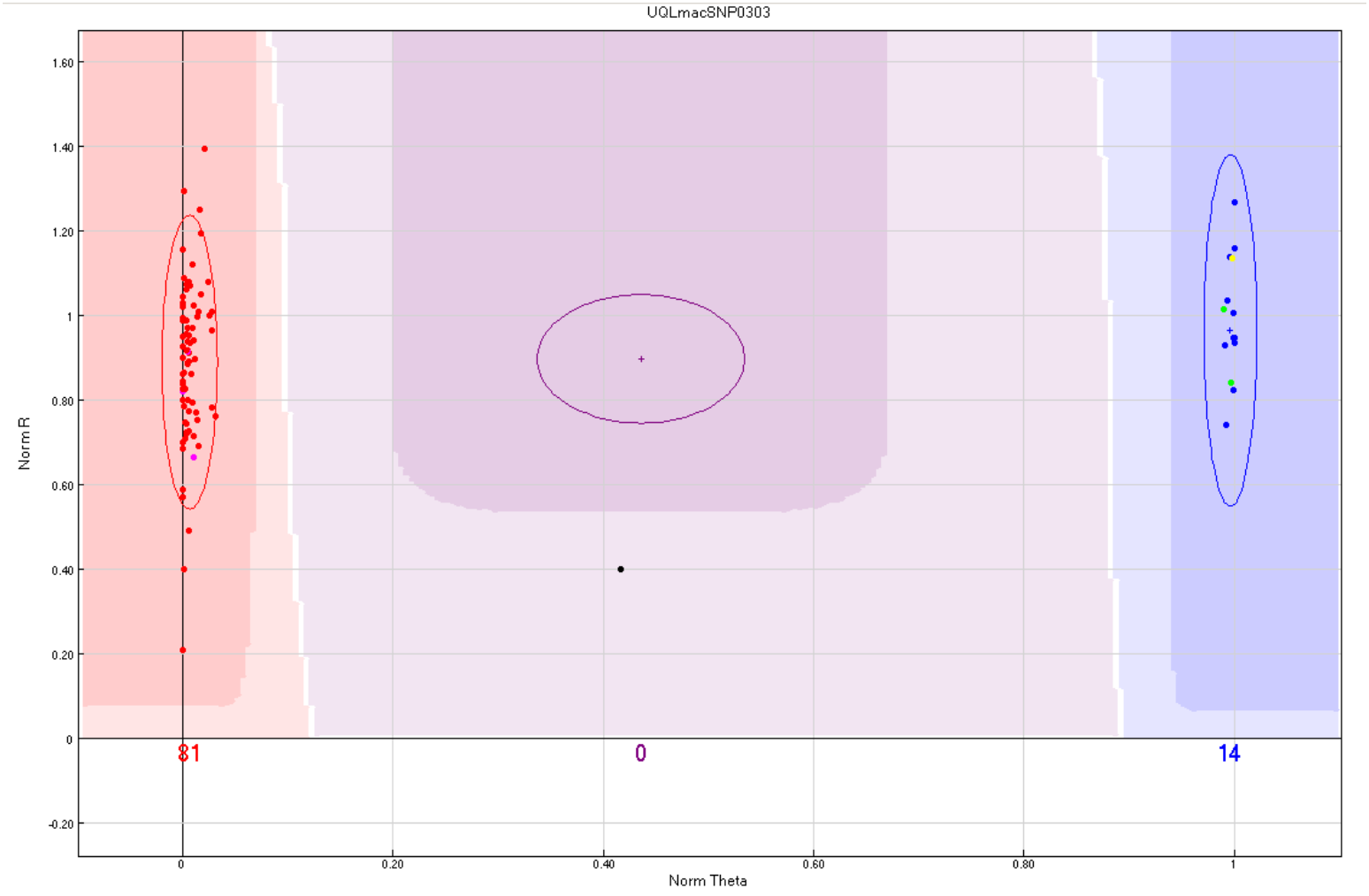
ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

Isolate	Reference number	Isolate name	Year Cultured	Year plant was grown	Stubble resistance source	Stubble cultivar	Stubble collection site	State
	04MGP-P029	*L. maculans*	2004	2003	Polygenic	Pinnacle	Geelong	VIC
	07VTJH020	*L. maculans*	2007	2006	Juncea	JC05007	Horsham	VIC
3	09SMJ087	Not yet determined	2009	2008	Juncea	OasisCL	Kaniva	VIC
4	05MGPP033	*L. maculans*	2005	2004	Poly	Skipton	Yeelanna	SA
5	89/C16	*L. maculans*	1988	1987	unknown	B. napus	Millicent	SA
6	04MGP-S006	*L. maculans*	2004	2003	Sylvestris	Surpass 400	Eyre Peninsula	SA
7	10SMJ041	Not yet determined	2010	2009	Juncea	Oasis	Tamworth	NSW
8	06MGPP019	*L. maculans*	2006	2005	Polygenic	Beacon	Wagga Wagga	NSW
9	LM300	*L. maculans*	2002	2001	Polygenic	TI1 Pinnacle	Mt Barker	WA
10	LM592	*L. maculans*	2003	2002	Polygenic	TI1 Pinnacle	Mt Barker	WA
11	06S012	*L. maculans*	2006	-	Napus	ATR-Beacon	Bordertown	SA
12	06J085	*L. maculans*	2006	-	Juncea	(cultivar unknown)	Horsham	VIC
13	05P032	*L. maculans*	2005	-	Napus	Skipton	Yeelana	SA
14	IBCN13	*L. maculans*	1991	-	Napus	(cultivar unknown)	Australia	-
15	IBCN15	*L. maculans*	1988	-	Napus	(cultivar unknown)	Streatham	VIC
16	IBCN16	*L. maculans*	1988	-	Napus	(cultivar unknown)	Australia	-
17	IBCN17	*L. maculans*	1988	-	Napus	(cultivar unknown)	Millicent	SA
18	IBCN18, M1	*L. maculans*	1988	-	Napus	(cultivar unknown)	Penshusrt	VIC
19	IBNC75	*L. maculans*	1987	-	Napus	(cultivar unknown)	Australia	-
20	IBCN76	*L. maculans*	1987	-	Napus	(cultivar unknown)	Mt Barker	WA
21	N/A	*L. maculans*	2005	-	Napus	Surpass501TT	Mt Barker	WA
22	N/A	*L. maculans*	2005	-	Napus	Beacon	Mt Barker	WA
23	PHW1223	*L. maculans*	1987	-	Napus	(cultivar unknown)	Australia	-
Ref	v23.1.3	*L. maculans*	Mid-1990	-	-	-	-	-
21'	04MGPP021	*L. maculans*	-	-	-	-	-	-
41'	06MGPP041	*L. maculans*	-	-	-	-	-	-

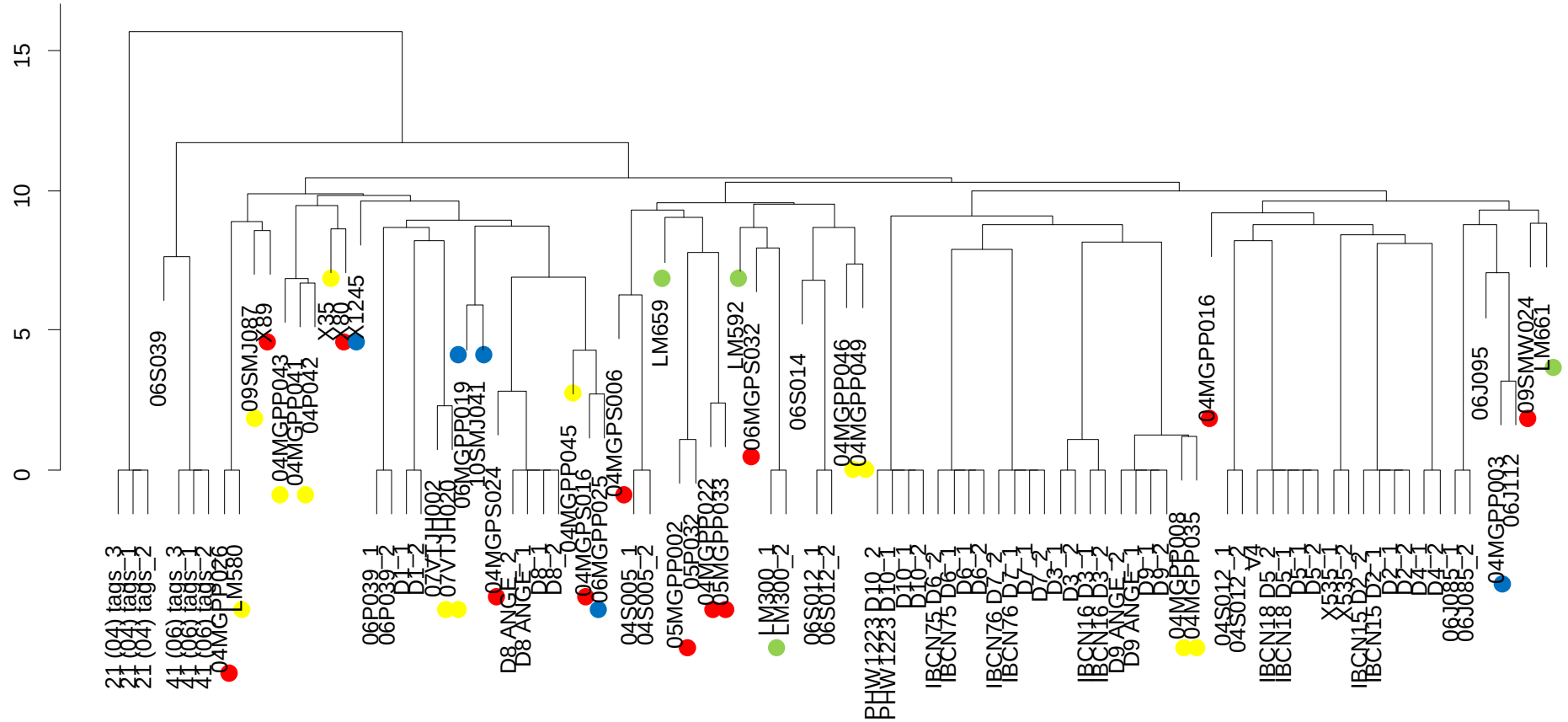
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ACPFG
AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

Clusters



Goldengate Preliminary Analysis



Presence/Absence variation

- Sequencing further isolates
- Identification of sequence reads present in some genomes and not others
- Annotation of these regions
- Association with known information about the isolates