

The Br GSP:
Progress on first-pass annotation
and seed BAC selection

UK-BRC 23rd May 2007

Martin Trick
Computational & Systems Biology Dept
John Innes Centre

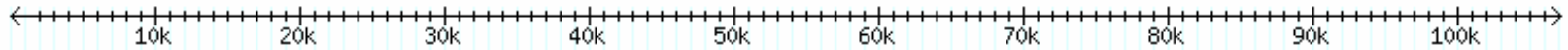
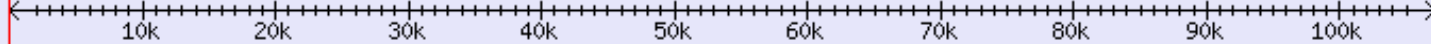
Brassica Genome Gateway 2007

<ul style="list-style-type: none">● Genome sequencing 522 seed BACs now annotated [Read me - soon!] Do It Yourself annotation - [Read me] [database] BLAST vs BrGenDB BAC tracking database - from PGG Australia	BAC name search e.g. <input type="text" value="KBrH001H24"/> <input type="button" value="Go"/> BAC feature search e.g. <input type="text" value="Flowering"/> <input type="button" value="Go"/>	January 2007 download MBGP Steering and BrGSP Committee meetings at PAG XV (PDF 150K)
<ul style="list-style-type: none">● BBSRC Brassica IGF Project Information Database A genome contigs C genome contigs		14th November 2006 download MBGP Steering Committee meeting at 15th Crucifer Genetics Workshop (PDF 50K)
<ul style="list-style-type: none">● BBSRC BrassicaDB <input type="button" value="Contents"/> Browse BrassicaDB ATIDB with Brassica features - update 13/10/06 [Read me] BrassicaDB sequence updates - more than a million seqs FTP site - download Brassica sequences, GFF files etc. BrassicaDB BLAST server		11th May 2006 more 15th Crucifer Genetics Workshop, 30 September - 4 October 2006
<ul style="list-style-type: none">● Multinational Brassica Genome Project BBSRC Brassica IGF Project (JIC/HRI/Bath/Birmingham, UK) IMSORB: oilseed rape programme (EU/China) B. oleracea GSS database (TIGR, US) AAFC Comparative Genome Viewer (SRC, Canada) www.brassica-rapa.org (NIAB/CNU, Korea) BAC library screening and distribution (JIC, UK) PGG Bioinformatics (PGG, Australia) Brassica.info (R-Res, UK)		11th May 2006 download UK-BRC meeting, 24th May 2006 Programme with links to presentations (PDF 25K)
<ul style="list-style-type: none">● Other links UK Brassica Research Community - [Post a message or contact the UK BBSRC Brassica IGF Project]		10th March 2006 download MBGP Draft White Paper (PDF 500K)
		11th June 2003 more Concept note: Brassica Genome Sequencing

<http://brassica.bbsrc.ac.uk>

Example first pass annotation

Overview of KBrB049E19



IGF probe matches (BLAST)

At1g16790

Genet. <http://brassica.bbsrc.ac.uk/tmp/align.7345>

File Edit Select View Format Colour Calculate Help

SSRs

KBrB049E19_79142_82985/1-640 MELDPEDVFRDEDEDPE SQFFQKEKEASKEFVVYLIDASP KMF S STCPSEEE - KQES
 At1g16970.1/1-621 MELDPD DVFRDEDEDPE NDF FQKEKEASKEFVVYLIDASP KMF C STCPSEEE - KQES



Consensus

MELDP - DVFRDEDEDPE - - FQKEKEASKEFVVYLIDASP KMF - STCPSE - E - KQES

Sequence position 1 11.0

Java Applet Window

View protein alignment

Strand	Feature	Start	End	Score	Accession	Description
-	Initial	829				
-	Internal	828				
-	Internal	826				
-	Internal	824				
-	Internal	82063	81958	2	At1g16970.1	ESFIKDIGSQNGIVSDSRENSLYSALWVAQ ALLRKG
-	Internal	81831	81714	1	At1g16970.1	GSSKTADKRIFLPTNEDDPFGSMRISVKED MIRTTLQRAK
-	Internal	81575	81492	0	At1g16970.1	DAQDLGISIELLPLSHPDQFDISLPYK
-	Internal	81413	81355	0	At1g16970.1	DLIGLNGDELTEFMPFVQ
-	Internal	81178	81001	1	At1g16970.1	KLEDMDQLKGRVLA KRIAKRITPMICDGV SIELNGYALLRPATPGTIITWLDSTINLPIK

Applet jalview.bin.JalviewLite started



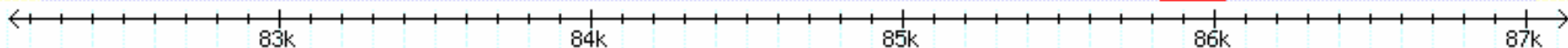
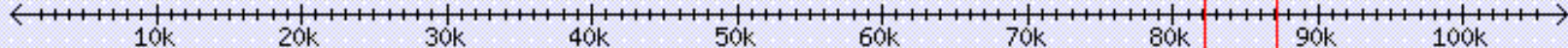
Close

Close

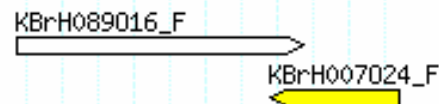
AT1G16970



Overview of KBrB049E19



BAC end matches (BLAST)



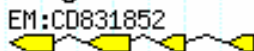
ClinterHMM predictions



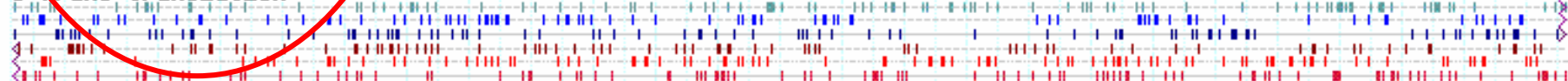
Arabidopsis gene model alignments (BLAT)



EST alignments (BLAT)



6-frame translation



Brassica Genome Gateway 2007

● **Genome sequencing**

BAC name search e.g.

[522 seed BACs](#) now annotated [Read me - soon!]

[Do It Yourself annotation](#) - [Read me] [database]

[BLAST vs BACends](#)

[BAC tracking database](#) - from PGG Australia

BAC feature search e.g.

● **BBSRC Brassica IGF Project**

[Information](#)

[Database](#)

[A genome contigs](#)

[C genome contigs](#)

● **BBSRC BrassicaDB**

[Browse BrassicaDB](#)

[ATIDB with Brassica features](#) - update 13/10/06 [Read me]

[BrassicaDB sequence updates](#) - more than a million seqs

[FTP site](#) - download Brassica sequences, GFF files etc.

[BrassicaDB BLAST server](#)

● **Multinational Brassica Genome Project**

[BBSRC Brassica IGF Project](#) (JIC/HRI/Bath/Birmingham, UK)

[IMSORB: oilseed rape programme](#) (EU/China)

[B. oleracea GSS database](#) (TIGR, US)

[AAFC Comparative Genome Viewer](#) (SRC, Canada)

[www.brassica-rapa.org](#) (NIAB/CNU, Korea)

[BAC library screening and distribution](#) (JIC, UK)

[PGG Bioinformatics](#) (PGG, Australia)

[Brassica.info](#) (R-Res, UK)

● **Other links**

[UK Brassica Research Community](#) - [Post a message or ...]

January 2007
MBGP Steering and BrGSP
Committee meetings at PAG XV
(PDF 150K)

14th November 2006
MBGP Steering Committee meeting at
15th Crucifer Genetics Workshop
(PDF 50K)

11th May 2006
15th Crucifer Genetics Workshop, 30
September - 4 October 2006

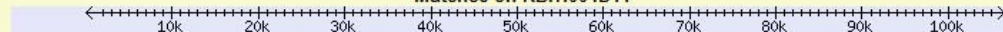
11th May 2006
UK-BRC meeting, 24th May 2006
Programme with links to
presentations (PDF 25K)

10th March 2006
MBGP Draft White Paper (PDF 500K)

11th June 2003
Concept note: Brassica Genome
Sequencing

<http://brassica.bbsrc.ac.uk>

Matches on KBrH004D11



Matches

KBrH004...

KBrH004...

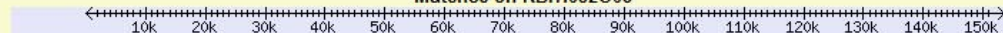
KBrH004D11.14 At5g10140.1 (E=1.7e-25) FLC (FLOWERING LOCUS C)

KBrH004D11:43,541..46,799 (3259 bp) score=7.46

KBrH004D11.14 At5g10140.1 (E=4.0e-06) FLC (FLOWERING LOCUS C)

KBrH004D11:43,541..46,799 (3259 bp) score=7.46

Matches on KBrH052O08



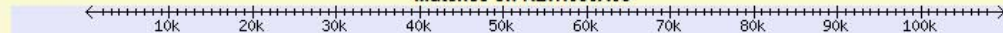
Matches

KBrH052...

KBrH052O08.27 At5g10140.1 (E=5.7e-25) FLC (FLOWERING LOCUS C)

KBrH052O08:97,426..100,370 (2945 bp) score=7.46

Matches on KBrH080A08



Matches

KBrH080...

KBrH080...

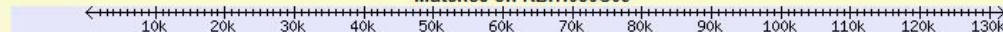
KBrH080A08.10 At5g10140.1 (E=3.8e-26) FLC (FLOWERING LOCUS C)

KBrH080A08:40,166..48,456 (8291 bp) score=7.46

KBrH080A08.23 At5g10140.1 (E=6.5e-17) T31P16.130, MADS-box protein **flowering** locus F (FLF), GO:0016566 [specific transcriptional repressor activity] GO:0008372 [cellular_component unknown] GO:0009910 [negative regulation of flower development] GO:0003700 [transcription factor activity]

KBrH080A08:48,241..48,654 (414 bp) score=5.97

Matches on KBrH080C09



Matches

KBrH080...

KBrH080...

KBrH080...

KBrH080C09.26 At1g77080.5 (E=9.6e-23) MAF1 (MADS AFFECTING FLOWERING 1), AGL27 MAF1 AT1G77080.4 AT1G77080.2 AT1G77080.3 AT1G77080 AT1G77080 FLM FLOWERING LOCUS M MADS AFFECTING FLOWERING 1 AT1G77080 AT1G77080 AT1G77080 [biological_process, regulation of transcription, DNA-dependent (GO:0006355)] [molecular_function, transcription factor activity (GO:0003700)] [biological_process, regulation of flower development (GO:0009909)] [biological_process, negative regulation of flower development (GO:0009910)]

KBrH080C09:91,587..95,296 (3710 bp) score=8.94

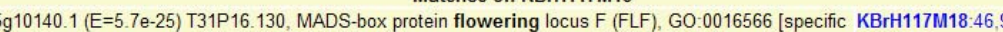
KBrH080C09.27 At5g65070.1 (E=1.9e-06) MAF4 (MADS AFFECTING FLOWERING 4)

KBrH080C09:103,328..107,787 (4460 bp) score=7.30

KBrH080C09.27 At5g65080.1 (E=6.8e-22) MAF5 (MADS AFFECTING FLOWERING 5 VARIANT I)

KBrH080C09:103,328..107,787 (4460 bp) score=7.22

Matches on KBrH117M18



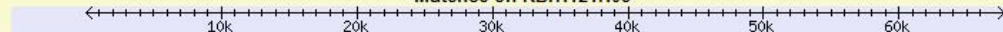
Matches

KBrH121...

KBrH117M18.16 At5g10140.1 (E=5.7e-25) T31P16.130, MADS-box protein **flowering** locus F (FLF), GO:0016566 [specific transcriptional repressor activity] GO:0008372 [cellular_component unknown] GO:0009910 [negative regulation of flower development] GO:0003700 [transcription factor activity]

KBrH117M18:46,939..49,858 (2920 bp) score=5.97

Matches on KBrH121H08



Matches

KBrH121...

Brassica BAC annotation - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://brassica.bbsrc.ac.uk/cgi-bin/gbrowse/jic_brassica?name=GO%3A0006355

Gateway IGF IMSORB ATIDB Annotation DB Brassica-ATIDB jilin4 WAN Agrenet jicbio status Weather F1 FIA timing Intranet Charge

DNA binding (GO:0003677) [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]

Matches on KBrB002K01

KBrB002K01.20	At5g18000.1 (E=1.6e-38) DNA binding / transcription factor , AT5G18000.1 AT5G18000 MCM23.7 MCM23_7 [cellular_component, cellular component unknown (GO:0008372)] [molecular_function, DNA binding (GO:0003677)] [molecular_function, transcription factor activity (GO:0003700)] [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]	KBrB002K01:72,959..74,423 (1465 bp)	score=3.00
KBrB002K01.20	At5g18000.1 (E=1.3e-18) DNA binding / transcription factor , AT5G18000.1 AT5G18000 MCM23.7 MCM23_7 [cellular_component, cellular component unknown (GO:0008372)] [molecular_function, DNA binding (GO:0003677)] [molecular_function, transcription factor activity (GO:0003700)] [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]	KBrB002K01:72,959..74,423 (1465 bp)	score=3.00
KBrB002K01.20	At5g18000.1 (E=4.9e-26) DNA binding / transcription factor , AT5G18000.1 AT5G18000 MCM23.7 MCM23_7 [cellular_component, cellular component unknown (GO:0008372)] [molecular_function, DNA binding (GO:0003677)] [molecular_function, transcription factor activity (GO:0003700)] [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]	KBrB002K01:72,959..74,423 (1465 bp)	score=3.00

Matches on KBrB003A10

KBrB003A10.5	At3g05670.1 (E=1.6e-19) DNA binding / protein binding / ubiquitin-protein ligase/ zinc ion binding , AT3G05670.1 AT3G05670 F18C1.6 F18C1_6 [molecular_function, DNA binding (GO:0003677)] [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]	KBrB003A10:10,186..10,491 (306 bp)	score=3.15
KBrB003A10.6	At3g05670.1 (E=4.8e-39) DNA binding / protein binding / ubiquitin-protein ligase/ zinc ion binding , AT3G05670.1 AT3G05670 F18C1.6 F18C1_6 [molecular_function, DNA binding (GO:0003677)] [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]	KBrB003A10:10,586..13,343 (2758 bp)	score=3.15
KBrB003A10.6	At3g05670.1 (E=3.7e-126) DNA binding / protein binding / ubiquitin-protein ligase/ zinc ion binding , AT3G05670.1 AT3G05670 F18C1.6 F18C1_6 [molecular_function, DNA binding (GO:0003677)] [biological_process, regulation of transcription, DNA-dependent (GO:0006355)]	KBrB003A10:10,586..13,343 (2758 bp)	score=3.15

Done

Extending from existing seed BACs

Overview of KBrH001H24

http://brassica.bbsrc.ac.uk/tmp/align.13320

File Edit Select View Format Colour Calculate Help

50 60 70 80 90

KBrH001H24_101430_101950V1-521 AAGCTTATTATTTCTACCATTTTAAATCGGTTTGGTTCTTTTGTAGTAGTAAGTTA

KBrH131G16_F/1-564 AAGCTTATTATTTCTACCATTTTAAATCGGTTTGGTTCTTTTGTAGTAGTAAGTTA

Consensus

AAGCTTATTATTTCTACCATTTTAAATCGGTTTGGTTCTTTTGTAGTAGTAAGTTA

Sequence 1 ID: KBrH001H24_101430_101950 Nucleotide: Adenine (19)

Java Applet Window

about: - Mozilla F...

Clicked **KBrB016B23_F**
View BAC end alignment...

Start Jalview

KBrB016B23 may map w
(other end **KBrB016B23**
113874..114298)

Close window

Applet jalview.bin...

about: - Mozilla F...

Clicked **KBrH058E07_F**
View BAC end alignment...

Start Jalview

Other end **KBrH058E07_R** ma
sequenced clone(s): **KBrB085**

Close window

Applet jalview.bin...

about: - Mozilla F...

Clicked **KBrH131G16_F**
View BAC end alignment...

Start Jalview

Close window

Applet jalview.bin...

Candidate BAC for extension

with another
seed BAC
(adding BAC
seeds)

about: - Mozilla F...

Start Jalview

Overlap with finished BAC **KBrH015H17**

Close window

Applet jalview.bin...

Selection of new seeds

- Problem: the ~500 seed BACs selected show clustering and, already, many overlaps are suspected or demonstrated
- Premise: clones inferred to span breakpoints with respect to the Arabidopsis sequence will be a useful new source of seed BACs for extension
- The Parkin *et al.* map contains about 80 such breakpoints - there are likely to be many more revealed at the microscale
- Method: programmatically interrogate database recording all BAC end-mappings, to discover significant associations between non-contiguous chromosome bins

The breakpoint sniffer

Brassica Genome Gateway Arabidopsis thaliana genome (includes TAI... **Breakpoint analysis**

Breakpoint analysis

Chromosome bin size: 500kb ▾

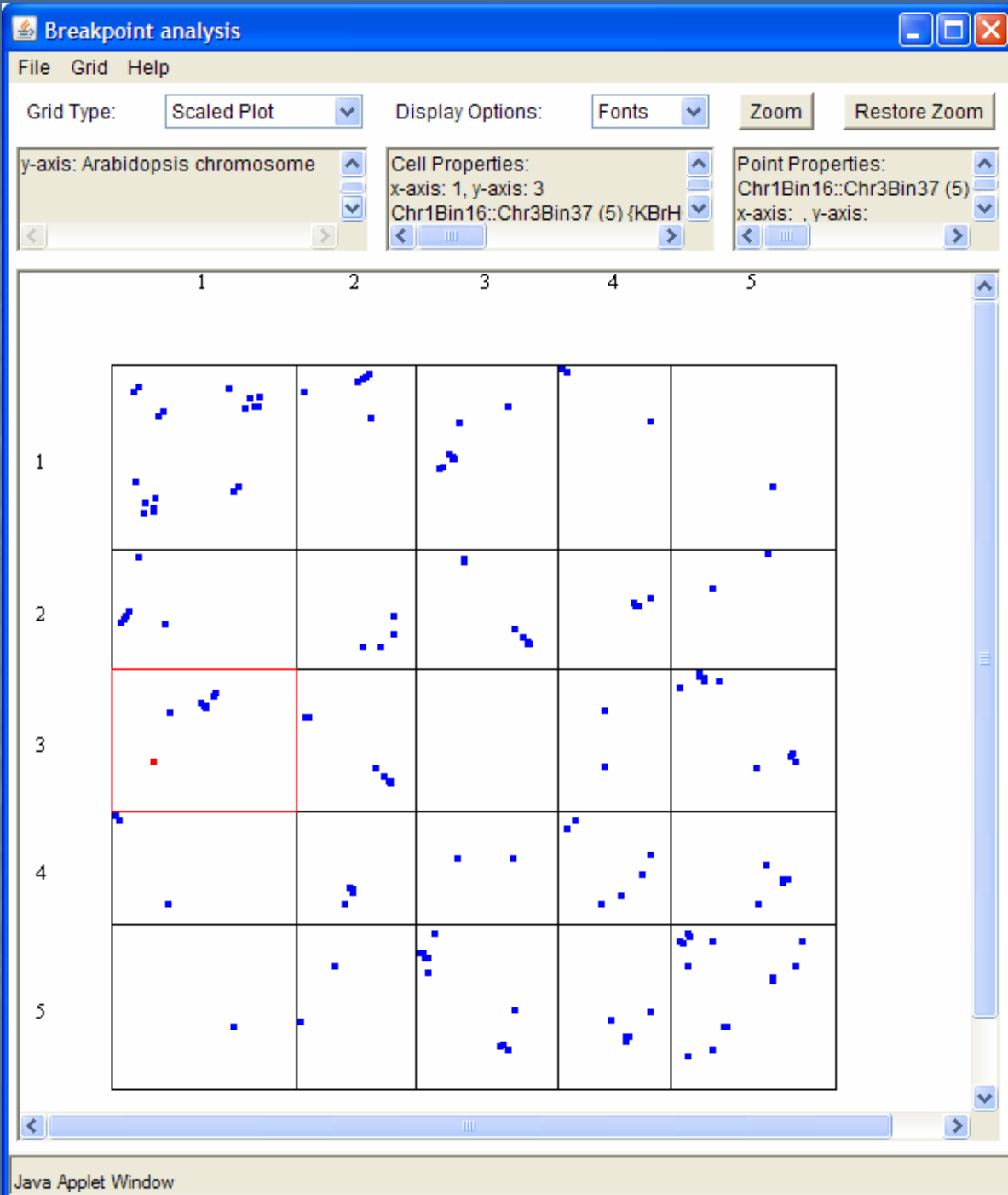
Microsynteny range: only consider bins separated by greater than 500kb ▾

Sequence identity threshold: only consider mappings with better than 60% ▾ % identity

Apply heuristic filter Log filtering process

Coincidence threshold: min number of clones shared across bins 3 ▾

[Martin Trick](#)
Last modified: Wed May 16 17:32:20 BST 2007



A1 Clone

	A	B	C	D	E	F	G	H	I	J
1	Clone	Chr Coordinate	Gene neighbour	Strand			Chr Coordinate	Gene neighbour	Strand	
91	Chr1Bin21::Chr2Bin30									
92	KBr004O13	1	10,273,628	AT1G29350	+		2	14,541,242	AT2G34480	-
93	KBr013P23	2	14,802,238	AT2G35100	-		1	10,473,578	AT1G29910	+
94	KBr060K20	2	14,722,655	AT2G34880	-		1	10,465,959	AT1G29890	+
95										
96	Chr1Bin16::Chr3Bin37									
97	KBrH001J23	3	18,446,951	AT3G49700	+	3	1	7,603,452	AT1G21650	+
98	KBrH041I11	3	18,480,274	AT3G49790	-	3	1	7,756,275	AT1G22020	-
99	KBrH087G17	3	18,458,433	AT3G49730	+	3	1	7,617,084	AT1G21690	+
100	KBrH101A22	1	7,728,059	AT1G21970	-	3	3	18,483,322	AT3G49800	-
101	KBrH103L20	3	18,413,836	AT3G49640	+	3	1	7,648,028	AT1G21750	+
102										
103	Chr1Bin23::Chr3Bin17									
104	KBrB050L10	1	11,214,382	AT1G31320	+	3	3	8,290,364	AT3G23230	-
105	KBrB052M01	1	11,379,643	AT1G31772	-	3	3	8,161,794	AT3G22980	+
106	KBrB117D12	1	11,447,568	AT1G31880	-	3	3	8,208,809	AT3G23080	+
107	KBrH034H11	3	8,274,765	AT3G23175	-	3	1	11,195,772	AT1G31300	+
108										
109	Chr1Bin36::Chr3Bin13									
110	KBrB082G02	3	6,171,409	AT3G18035	-	3	1	17,930,931	AT1G48490	+
111	KBrH022J23	3	6,414,100	AT3G18640	+	3	1	17,958,785	AT1G48560	+
112	KBrH087E17	3	6,428,234	AT3G18680	-	3	1	17,988,480	AT1G48635	+
113	KBrH126C10	3	6,209,667	AT3G18110	-	3	1	17,972,796	AT1G48600	+
114										
115	Chr1Bin37::Chr3Bin14									
116	KBrB043K10	1	18,379,594	AT1G49640	-	3	3	6,880,826	AT3G19820	-
117	KBrH017B07	1	18,233,609	AT1G49270	+	3	3	6,688,373	AT3G19290	+
118	KBrH098G11	1	18,420,793	AT1G49760	-	3	3	6,798,282	AT3G19570	-
119										
120	Chr1Bin38::Chr3Bin14									

