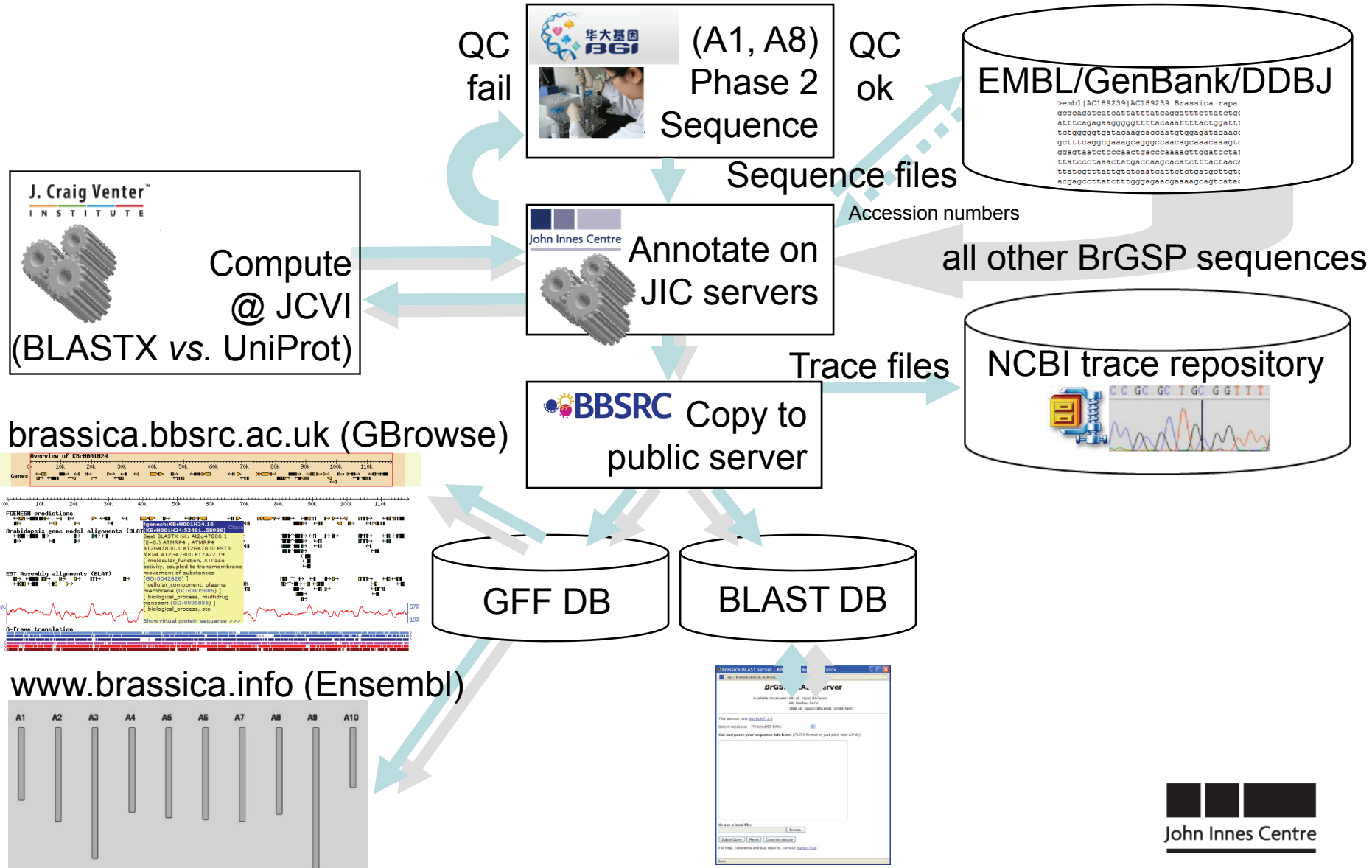


Update on *B. rapa* genome annotation

Nizar Drou & Martin Trick
Computational & Systems Biology Dept.
John Innes Centre, UK

Annotation pipeline



New *ab initio* genefinder

- SNAP – **S**emi-HMM-based **N**ucleic **A**cid **P**arser (Korf, UCSD)

post-processed by

- PASA – **P**rogram to **A**ssemble **S**pliced **A**lignments (Haas, TIGR & Broad Institute MIT)

Following tracks remain available for the enthusiasts...

- Augustus (Stanke, Göttingen)
- FGENESH (Softberry, NJ)
- GENSCAN (Burge, MIT)
- GlimmerHMM & GeneZilla (Majoros, TIGR)

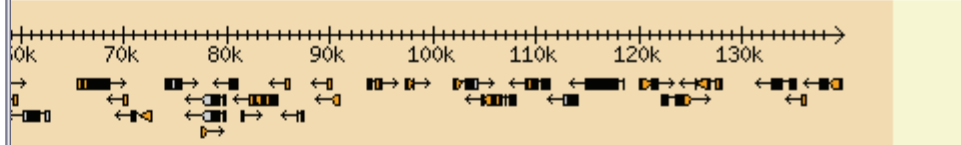
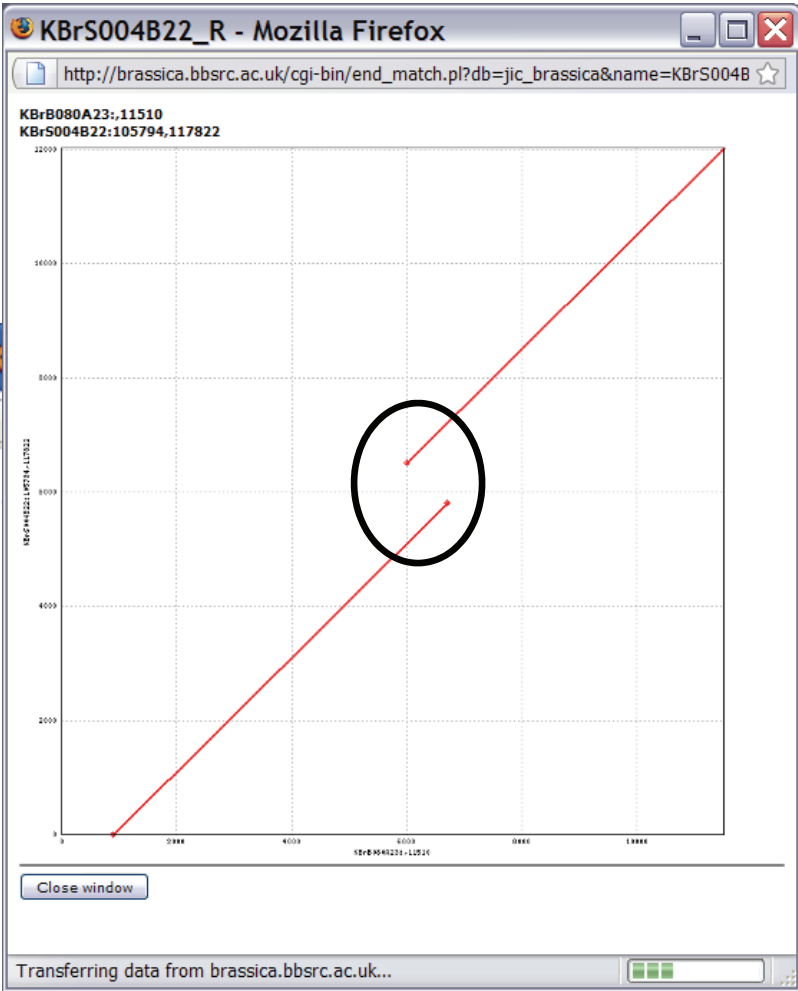
SNAP/PASA gene models

- PASA-corrected SNAP gene models, using alignments to ~800k Brassica ESTs
 - PASA requires 95% identity over 90% of transcript length & canonical donor/acceptor sites
 - PASA calls 5' and 3' UTRs
 - PASA can call alternative splicing variants (“gene instantiations”)
 - SNAP models pass through if no EST evidence
 - All other genefinder predictions will remain available to the enthusiasts

Genome coverage @ 7/5/09

- Total number of annotated BACs: 976
- Coverage (redundant): 116,947,567 bp
- Number of *ab initio* predicted genes: 18,890
- Number supported by EST evidence: 9,521

- Number of PASA/SNAP-annotated CDS: 88,928
- Covering: 17,409,673 bp
- => Coding sequence density = 14.9%



06... [Close]

name=KBrB080A23:11510

KBrS004B... [Close]

http://brassica.bbsrc.ac.uk

Clicked KBrS004B22_R
View BAC end alignment...

Start Jalview

msatfinder:KBrB080A23.11
(KBrB080A23:11517..11532)

align.17625

n.17161

430 440 450 460

AATCTGATTTAAAAAAAAAAAAACATATAAATATGAACTCAAGACTTAT

AATTTGATTTAAAAAAAAAAAAACATATAAATATGAACTCAAGACTTAT

AGAAGCAGCTCCGTTGCT-TAAT-TGATTTAAAAAAAAAAAAACATATAAATATGAACTCAAGACTTAT

Cognate c

Sequence 1 ID: KBrB080A23_10823_11510 Nucleotide: Adenine (442)

Consensus

MGNQKLKWT-EEEEALLAGVVRKHGPGKWKNI LRDP E-A-QLS-RSNIDLKDKWRNLSV

Sequence 1 ID: KBrB080A23_3619_6068 Residue: LEU (6)

52%

7%

B. napus BACend sequences

- 90,746 *B. napus* BES available (average read 441 bp) – submitted to EMBL/GenBank/DDBJ
- BLASTN analysis vs completed KBr BACs:
 - 45,912 JBnB hits (@ \geq 92% identity)
 - Added to public annotation server 5/5/09

Brassica Genome Gateway

• Genome sequencing

[963 BACs](#) have now been annotated - [\[Read me\]](#)
List the [BACs](#) annotated in the last 31 days
Quick [BLAST](#) vs. BACends or completed BACs
[BAC registry](#) - from RRes, UK
['Do It Yourself' annotation](#) - [\[Read me\]](#) [\[database\]](#)
[Latest news: 90,579 B. napus BES BLASTable](#)

BAC name search e.g.

BAC feature search e.g.

• Brassica 95k unigene set

[Read me](#)
Quick [BLAST](#) vs. unigenes

Search unigenes e.g.

• BBSRC Brassica IGF Project

[Project information](#)
[Search the database](#)
View [A genome contigs](#) and [C genome contigs](#)

• BBSRC BrassicaDB

[Browse BrassicaDB](#)
[ATIDB with Brassica features](#) - [\[Read me\]](#)
[BrassicaDB sequence updates](#) - refreshed daily
[Brassica BLAST server](#) - all DB options
[FTP site](#) - download sequences, GFF files etc.

Search BrassicaDB

• Multinational Brassica Genome Project

[BBSRC Brassica IGF Project](#) (JIC/HRI/Bath/Birmingham, UK)
[IMSORB: oilseed rape programme](#) (EU/China)
[BAC library screening and distribution](#) (JIC, UK)
[B. oleracea GSS database](#) (TIGR, US)
[AAFC Comparative Genome Viewer](#) (SRC, Canada)
[www.brassica-rapa.org](#) (NIAB/CNU, Korea)
[PGG Bioinformatics](#) (PGG, Australia)
[Brassica.info](#) (R-Res, UK)

• Other links

[UK Brassica Research Community](#) - [\[Post a message or subscribe to the UK-BRC mailing list \]](#)
[GARNet](#) - UK-based Arabidopsis genomics
[AtEnsembl](#) - Arabidopsis genome browser

April 2008 [download](#)
[BrGSP committee meeting at PAG XVI](#)
[- draft minutes \(PDF 95KB\)](#)

February 2008 [download](#)
[MBGP committee meeting at PAG XVI](#)
[- draft minutes \(PDF 80KB\)](#)

May 2007 [more](#)
[UK-BRC meeting, 23 May 2007](#)
[presentations \(PDFs\)](#)

January 2007 [download](#)
[MBGP and BrGSP committee](#)
[meetings at PAG XV \(PDF 150KB\)](#)

11th June 2003 [more](#)
[Concept note: Brassica Genome](#)
[Sequencing](#)

Archive [more](#)
[Older news items](#)

Last modified: Tue Apr 21 15:45:52 BST 2009

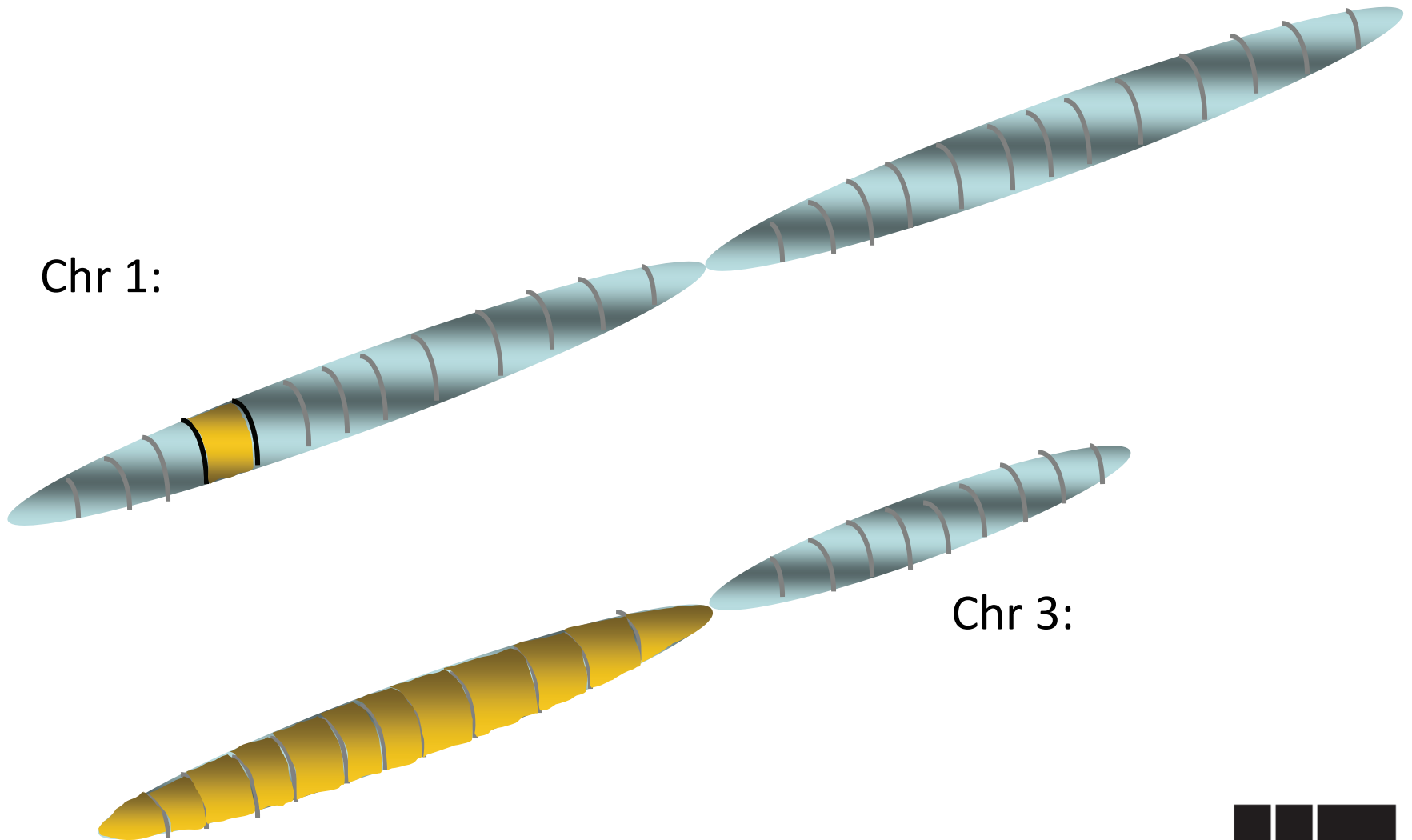
[contact us](#)

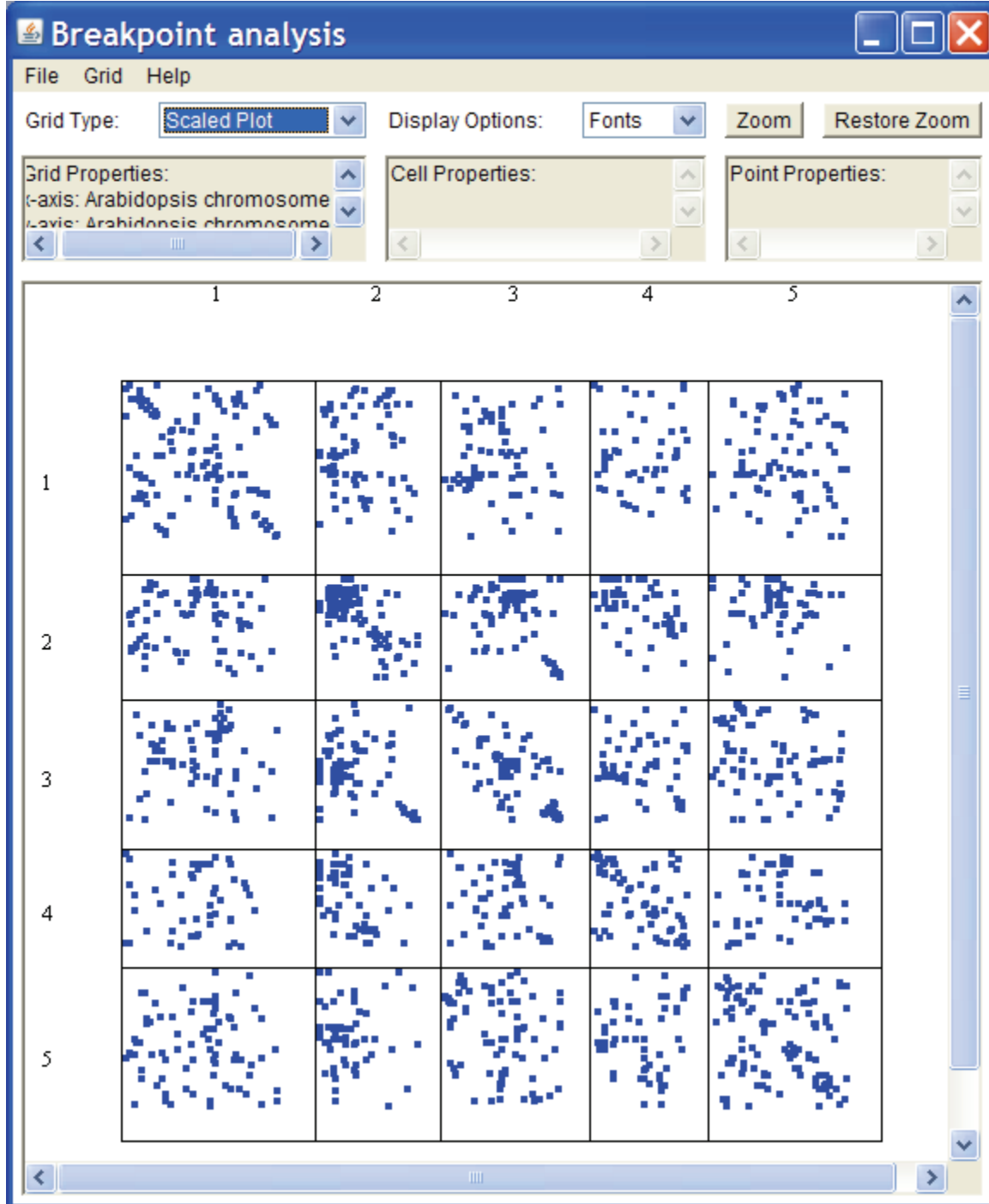


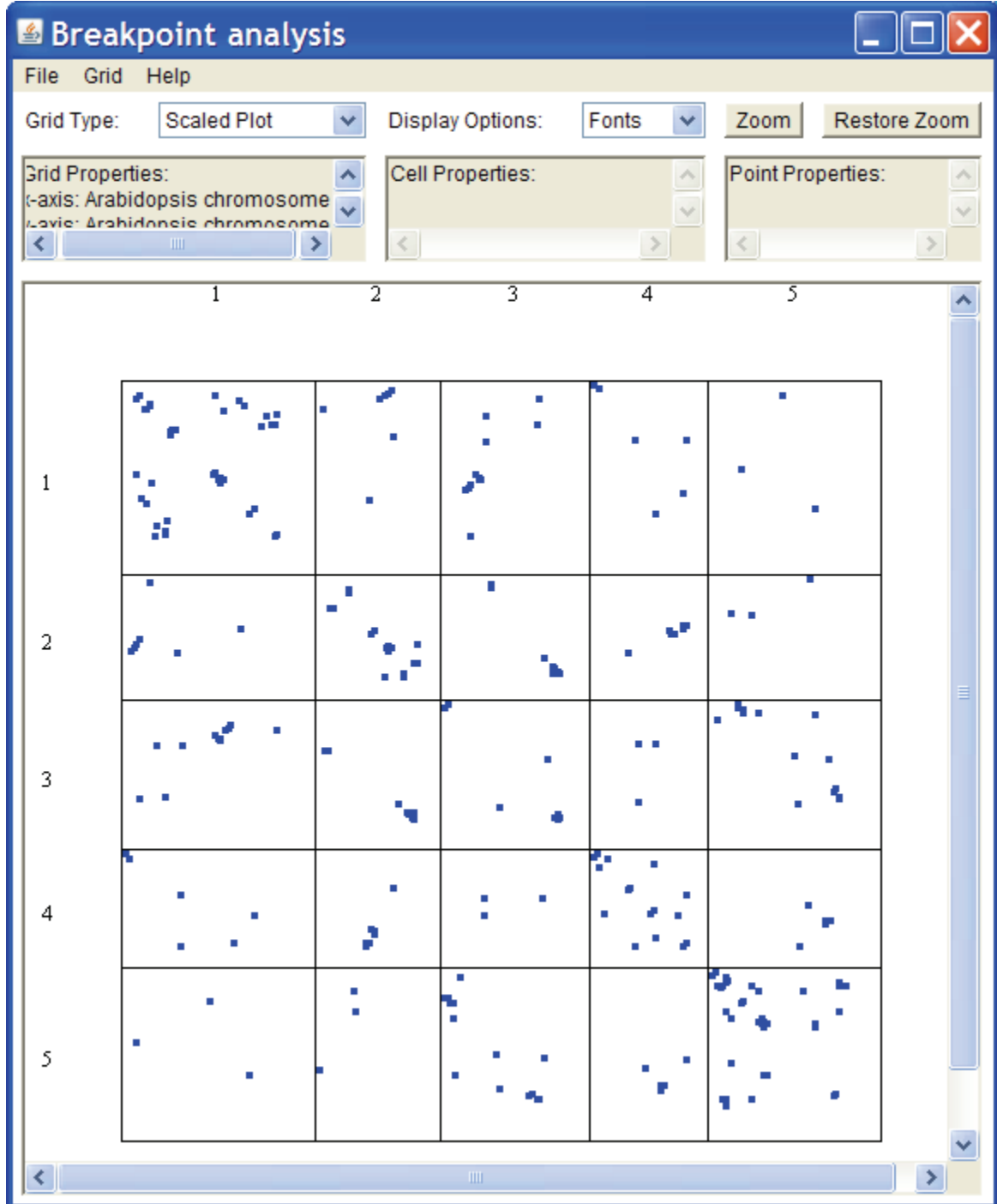
Synteny breakpoints

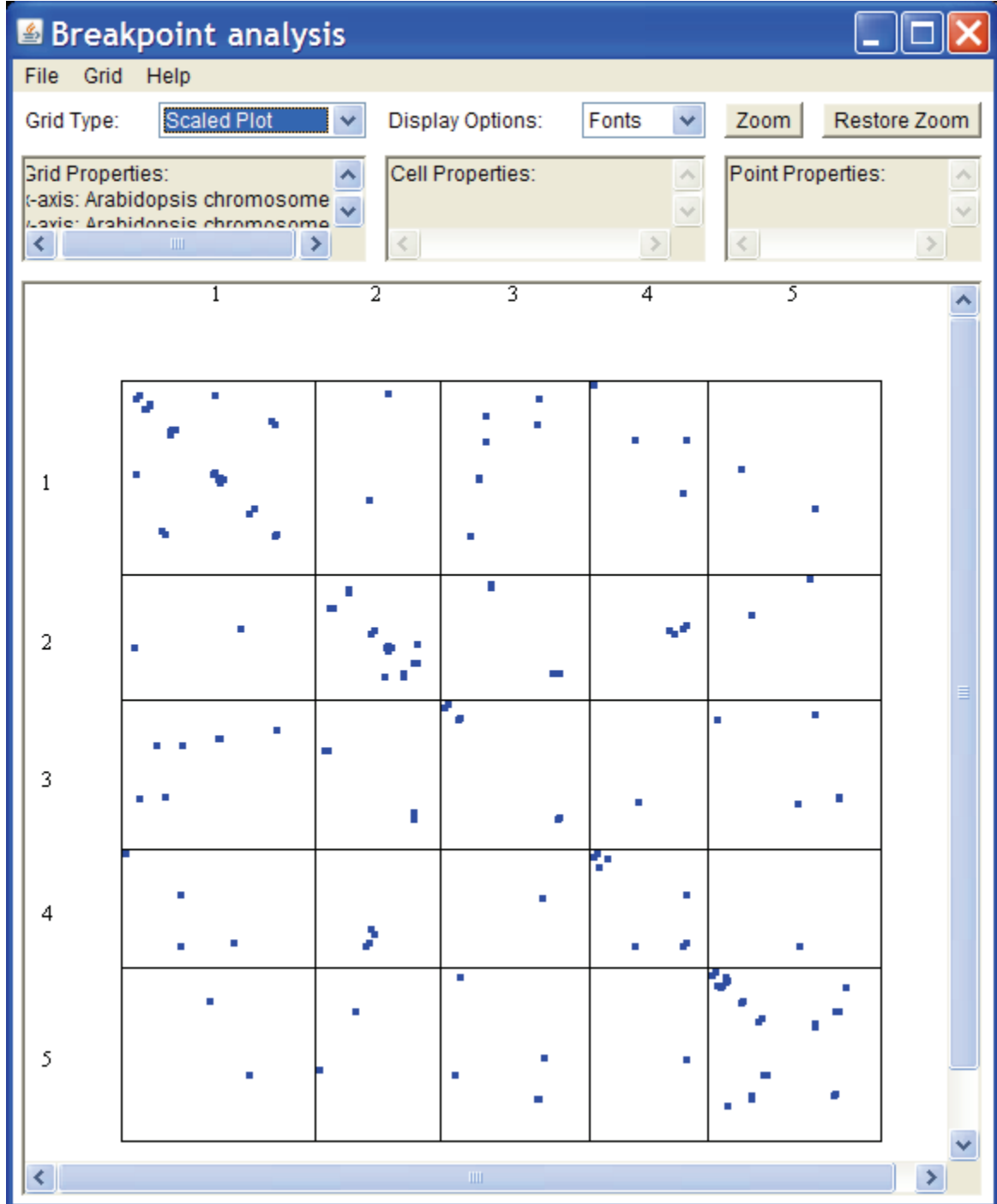
- Starting seed BACs selected by *in silico* mapping onto Arabidopsis pseudomolecules
 - Clustered in Brassica genespace
 - Not enough free ends for extension
- Select clones spanning synteny breakpoints
 - Biologically interesting
 - New source of seed points
- Heuristic algorithm developed to infer breakpoints and identify candidate clones

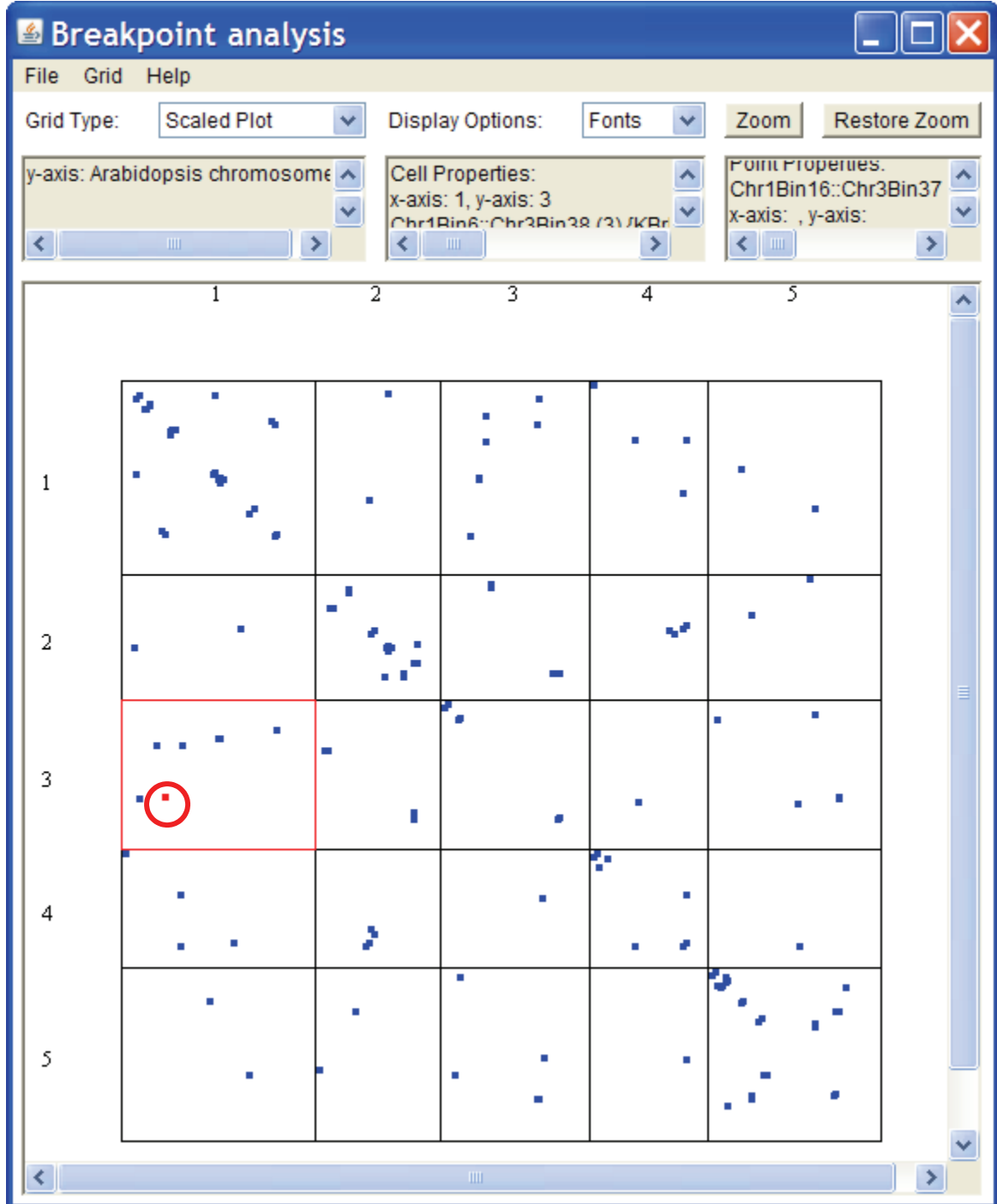
Breakpoint inference



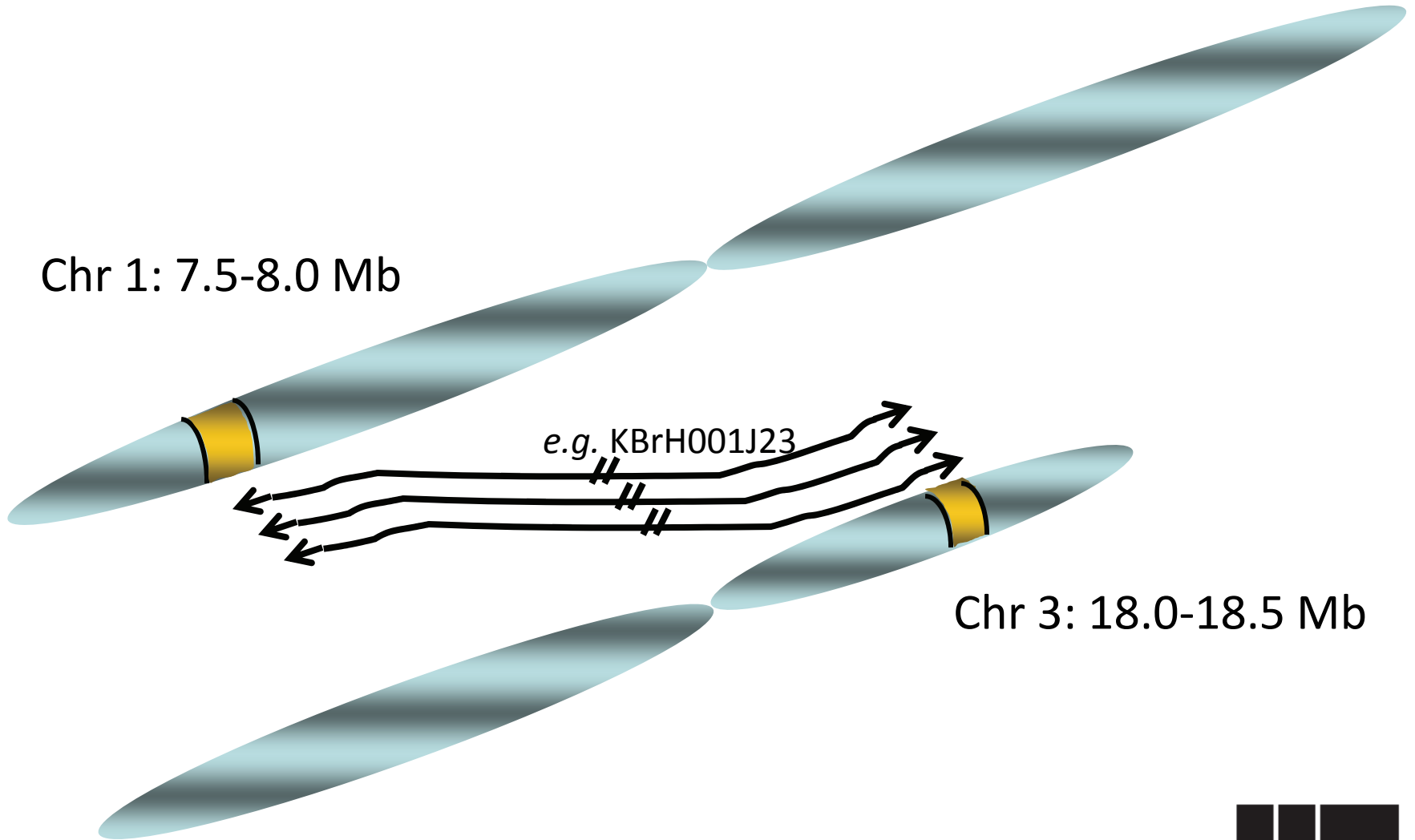








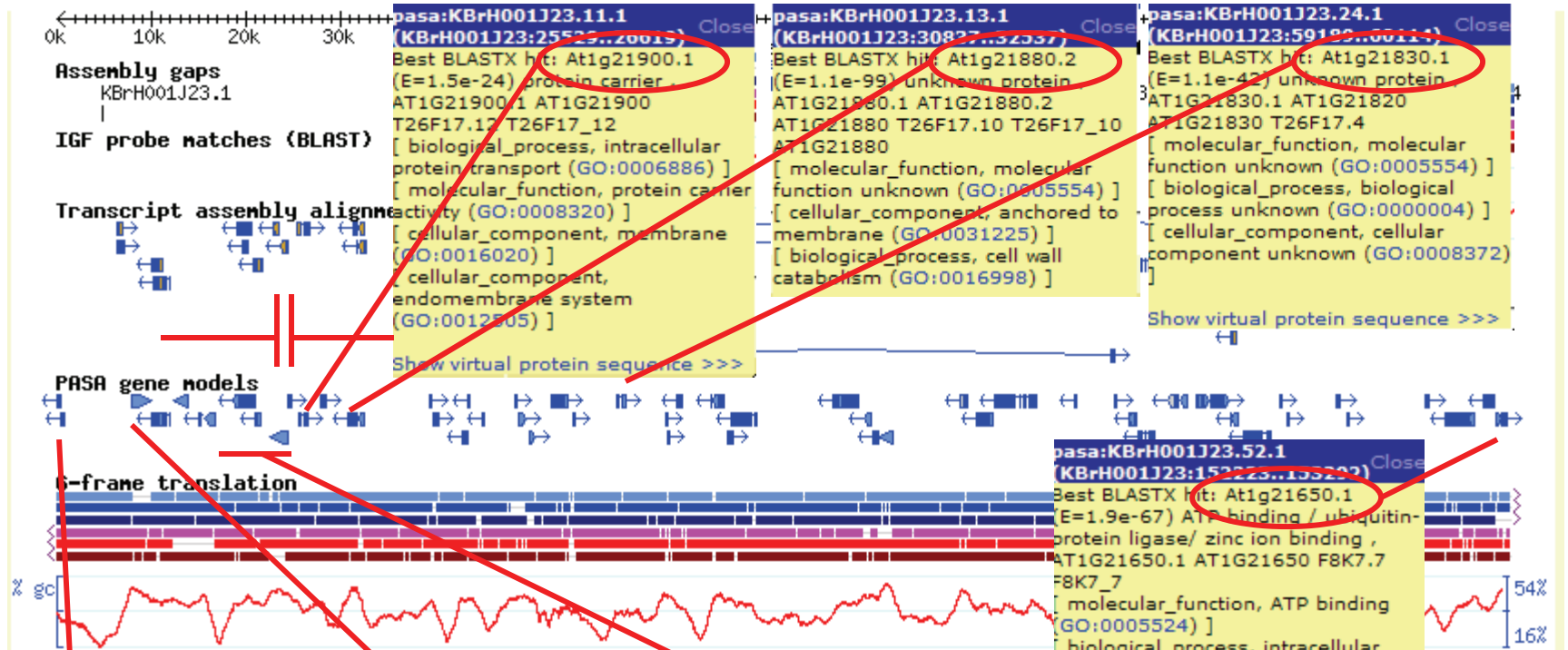
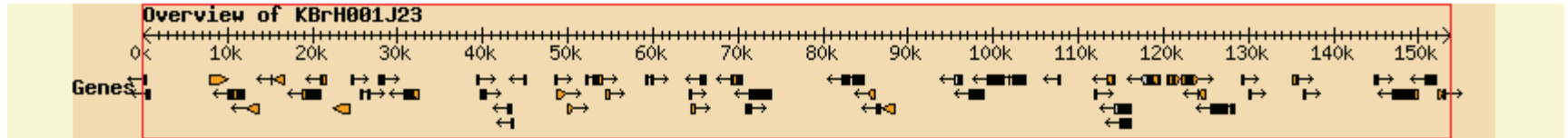
Chr 1: 7.5-8.0 Mb



e.g. KBrH001J23

Chr 3: 18.0-18.5 Mb

Breakpoint validation



pasa:KBrH001J23.2.1 (KBrH001J23:407..780) Close
 Best BLASTX hit: At3g49700.1 (E=2.6e-21) ETOX (ETHYLENE OVERPRODUCING 3)
 Show virtual protein sequence >>>

pasa:KBrH001J23.3.1 (KBrH001J23:7880..10053) Close
 Best BLASTX hit: At3g49710.1 (E=0.) unknown protein
 AT3G49710.1 AT3G49710 T16K5.60
 [molecular_function, molecular function unknown (GO:0005554)]
 [cellular_component, cellular

pasa:KBrH001J23.7.1 (KBrH001J23:18790..20850) Close
 Best BLASTX hit: At3g49730.1 (E=2.7e-45) GTP binding
 AT3G49730.1 AT3G49730 T16K5.80
 [cellular_component, mitochondrion (GO:0005739)]

pasa:KBrH001J23.52.1 (KBrH001J23:15225..15320) Close
 Best BLASTX hit: At1g21650.1 (E=1.9e-67) ATP binding / ubiquitin-protein ligase / zinc ion binding
 AT1G21650.1 AT1G21650 F8K7.7 F8K7_7
 [molecular_function, ATP binding (GO:0005524)]
 [biological_process, intracellular protein transport (GO:0006886)]
 Show virtual protein sequence >>>