



# Canadian Canola Genome Sequencing Initiative (CanSeq)

**Andy Sharpe**

**5<sup>th</sup> Applying Genomics to Canola Improvement  
Workshop  
11<sup>th</sup> December 2008**



# *Why is Brassica genome sequencing project so important?*

A **foundational resource** for Brassica crop species:

- 'Road map' of the genome and genes
- Identify novel genes
- DNA markers (SNPs and SSRs) for marker-assisted selection (MAS) and enhanced trait development
- Regulatory sequences
- Basis for global mutation discovery (TILLING)
- Re-sequencing different genotypes





*B. oleracea*  
n = 9

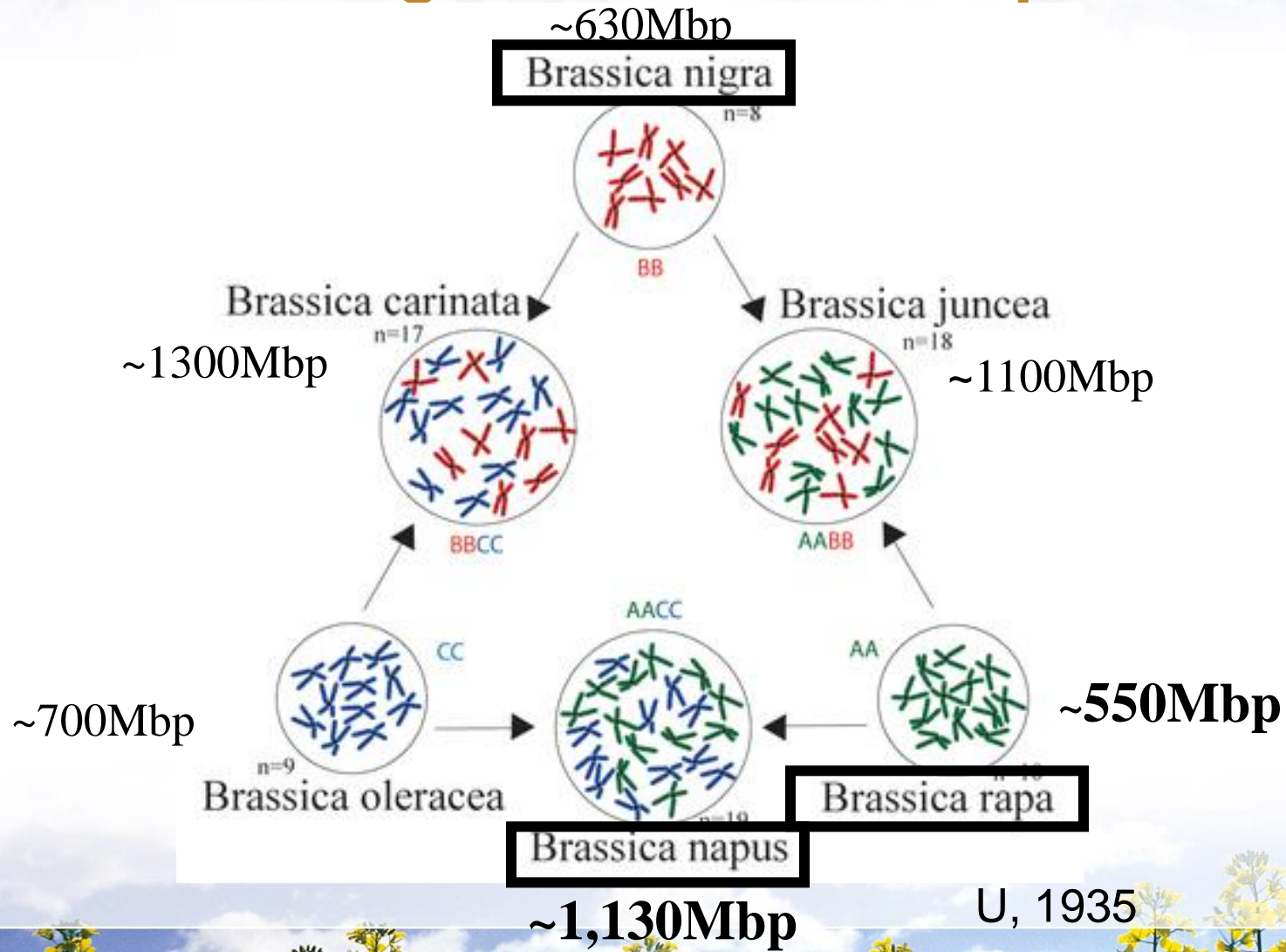


*B. rapa*  
n = 10



*B. napus*  
n = 19

# U's Triangle for Brassica species



# Why *Brassica rapa* (A genome)?

- Parental genome of *B. napus* & *B. juncea*
  - Similar traits, controlled by homologous genes
- Simpler genome organization
  - Effectively halves level of duplication for simpler data analysis
- Extensive resources already developed
  - Korea initiated the project in 2003; major crop for this country

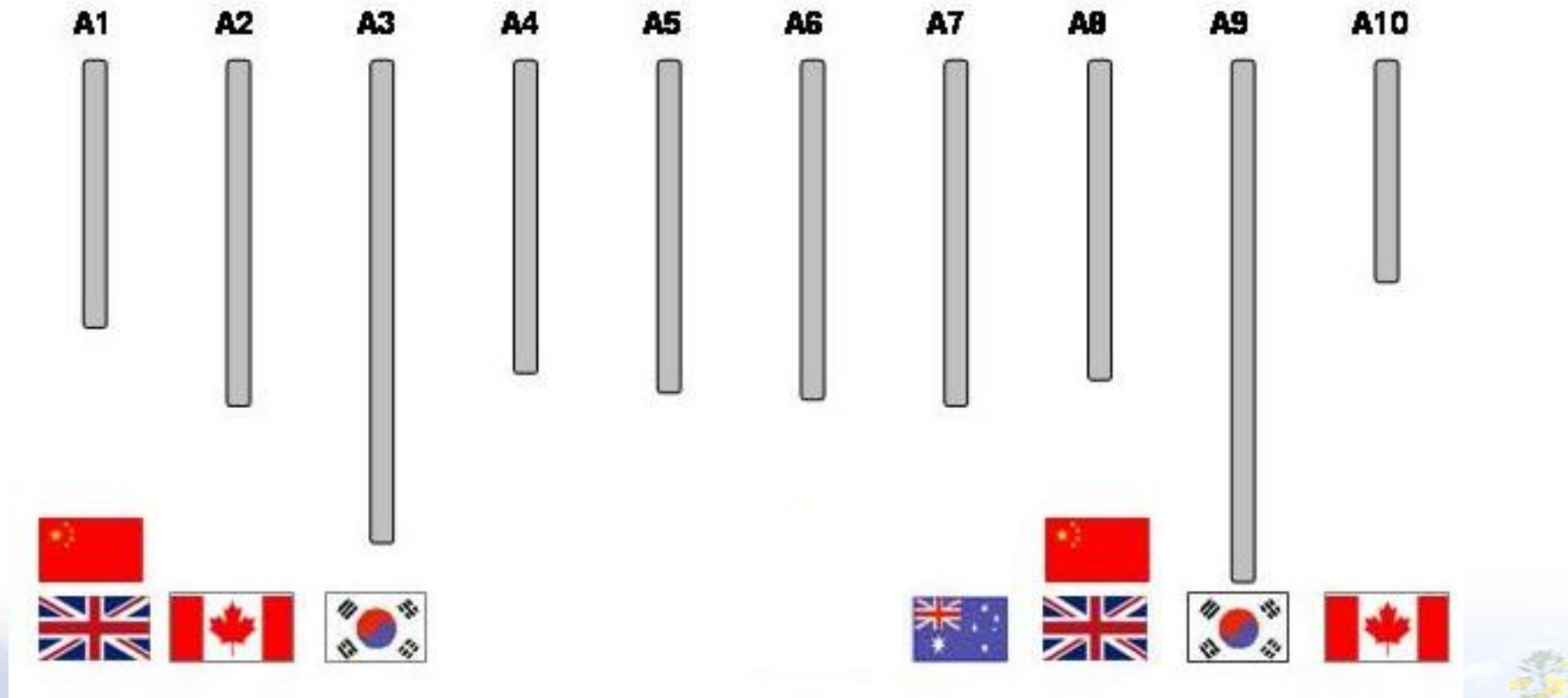


# CanSeq Project Overview

- NRC-PBI / AAFC joint project
  - funding from NRC, AAFC, Genome Alberta and industry partners
  - Cargill, Dow AgroScience, KWS and Rapool Ring (others?)
  - \$2.5M project
- Sequencing of A2 & A10 chromosomes of *B. rapa*
- Draft Sequence of *B. napus*
- Re-sequencing of 10 *B. napus* lines
- *de novo* whole genome shot-gun *B. oleracea* sequencing
- Resource development for *B. nigra* (B genome)
  - BAC end sequencing and fingerprinting



# Multinational *B. rapa* Genome Sequencing Project (MBrGSP)



# *B. rapa* sequencing – 1<sup>st</sup> Phase

End sequencing of *B. rapa* large insert BAC libraries: Canada made 2<sup>nd</sup> largest contribution using AAFC and NRC funds

Plates	No. clones	Group	Country	Status
KBrH001 - KBrH015	5,760	NIAB	S. Korea	completed
<b>KBrH016 - KBrH050</b>	<b>13,440</b>	<b>AAFC + NRC</b>	<b>Canada</b>	<b>completed</b>
KBrH051 - KBrH062	4,608	JIC	UK	completed
KBrH063 - KBrH087	9,600	DPI	Australia	completed
KBrH088 - KBrH117	11,520	JIC, Bath	UK	completed
KBrH118 - KBrH136	7,296	U. Bielefeld	Germany	Completed
KBrH137 - KBrH144	3,072	CNU	S. Korea	Completed
KBrB001 - KBrB096	36,864	NIAB, CNU	S. Korea	Completed
<b>KBrB097 - KBrB132</b>	<b>13,824</b>	<b>AAFC + NRC</b>	<b>Canada</b>	<b>completed</b>



## *B. rapa* Sequencing – 2<sup>nd</sup> Phase

- 2 *B. rapa* chromosomes (R2 and R10) - Canada
- Existing genes of interest and intra-genomic duplication
- ~800 BACs (~120MB of DNA) sequenced in pools of 12 BACs
- Combination of existing Sanger and Next Generation Sequencing technology (NGS)
- Roche 454 FLX Genome Sequencer



# 454 FLX System at NRC-PBI

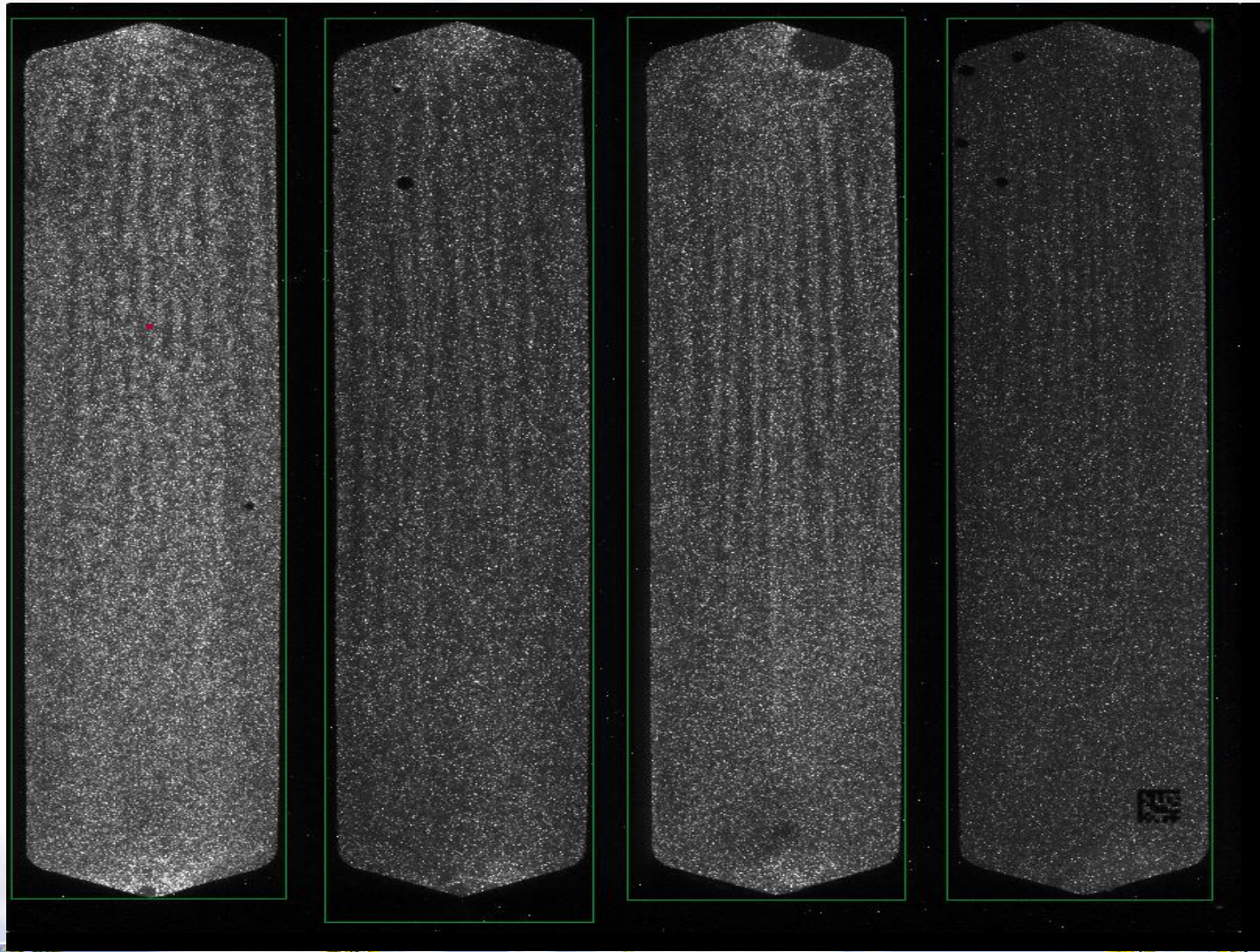


# 454 FLX System Benefits

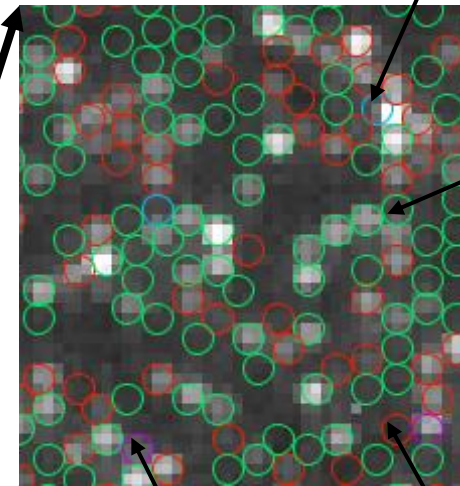
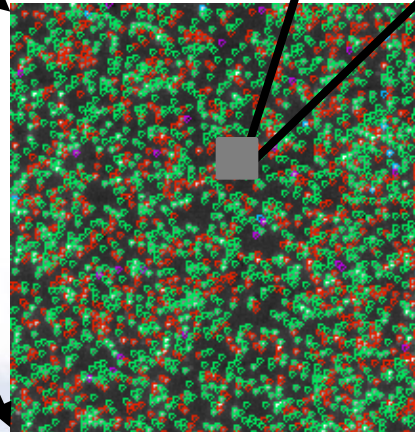
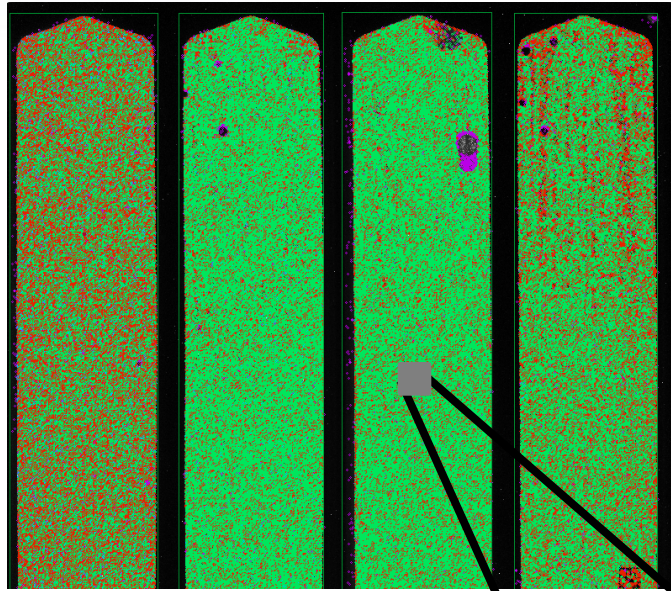
- >400,000 reads per run, 250bp read length, >99.5%
- >100 million bases (100MBp) per 8 hour run
- Eliminates cloning and colony-picking
- Reduced cost/raw base
- Variety of template DNA (amplicons, gDNA, cDNA, BACs, etc.)
- Variety of applications (*de novo* sequencing, re-sequencing, SNP discovery, etc.)
- Sequence multiple samples in one run with 12 available Multiplex Identifiers (MID) adaptors



# *PicoTiterPlate Image – 4 region run*



# Signal Processed Data

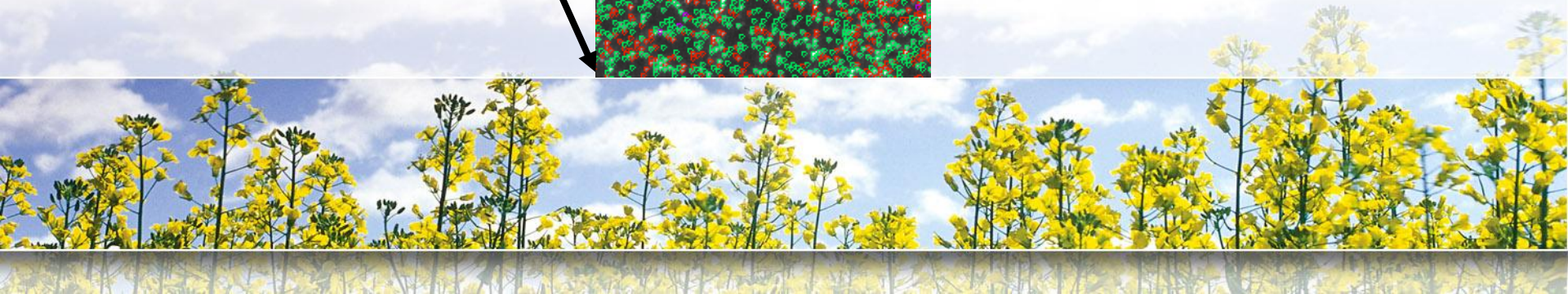


Control well

Passed Well

Failed well

Key pass well



# Sequencing Multiple BACs on the 454

- MIDs (**M**ultiplex **I**Dentifiers):

- MID1 ACGAGTGCGT
- MID2 ACGCTCGACA
- MID3 AGACGCACTC
- MID4 AGCACTGTAG
- MID5 ATCAGACACG
- MID6 ATATCGCGAG
- MID7 CGTGTCTCTA
- MID8 CTCGCGTGTC
- MID9 TAGTATCAGC
- MID10 TCTCTATGCG
- MID11 TGATACGTCT
- MID12 TACTGAGCTA

- Each 454 read can be resolved to a single MID key which corresponds to a particular BAC in a pool of 12 BACs
- 454 Newbler software used for assembly after data partition



# *BAC Pooling on a 2-region 454 run*

Fragment Library Quality Assessment and Quantitation  
Agilent 2100 Bioanalyzer



Dilute 24 Fragment Libraries



Add equal volumes of 12 diluted libraries / pool



## Pool I

MID 1	BAC 1
MID 2	BAC 2
MID 3	BAC 3
MID 4	BAC 4
MID 5	BAC 5
MID 6	BAC 6
MID 7	BAC 7
MID 8	BAC 8
MID 9	BAC 9
MID 10	BAC 10
MID 11	BAC 11
MID 12	BAC 12

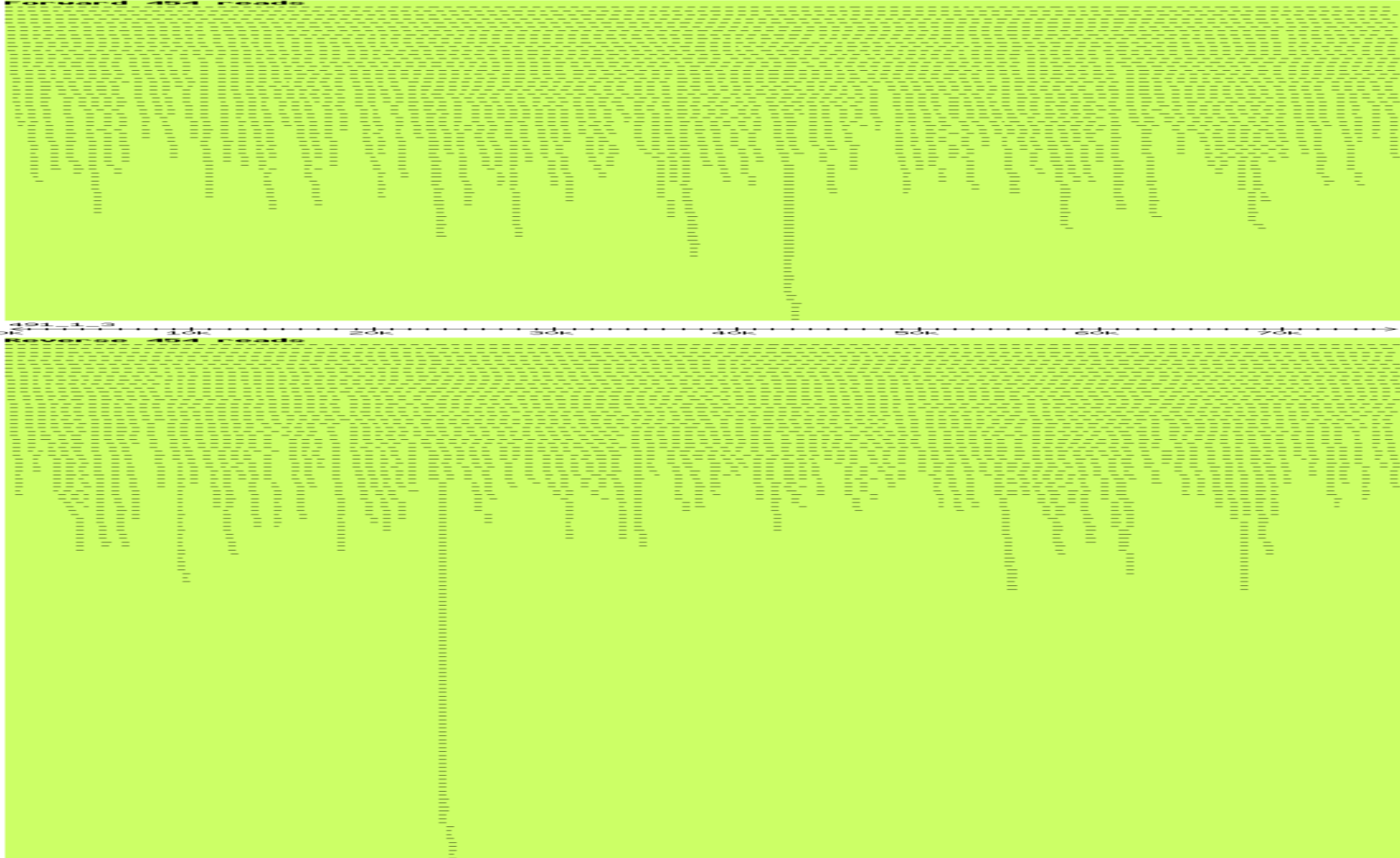


## Pool II

MID 1	BAC 13
MID 2	BAC 14
MID 3	BAC 15
MID 4	BAC 16
MID 5	BAC 17
MID 6	BAC 18
MID 7	BAC 19
MID 8	BAC 20
MID 9	BAC 21
MID 10	BAC 22
MID 11	BAC 23
MID 12	BAC 24

# *B. rapa* BAC KBrH088K06

Assembled data: Genome Viewer detail with 454 reads from one contig



# B. rapa BAC KBrH088K06

Assembled data: Genome Viewer detail with collinearity with Arabidopsis

Microsoft PowerPoint - [4Dec08.ppt]

http://napus.agr.gc.ca/cgi-bin/gbrowse-1.69/gbrowse/KBrH088K06-collinear/#search

KBrH088K06\_phasecolinear: scaf... contig-img.cgi (PNG Image, 570x2196 ...)

## This is an automated Gbrowse deployment for a Brassica rapa BAC (KBrH088K06 phase colinear)

Showing 17 kbp from scaffold00001, positions 97,000 to 114,000

**Instructions**  
**Searching:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed.  
**Navigation:** Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.  
**Examples:** KBrH088K06, scaffold00001, 491\_1\_4, 491\_1\_3, 491\_1\_2.

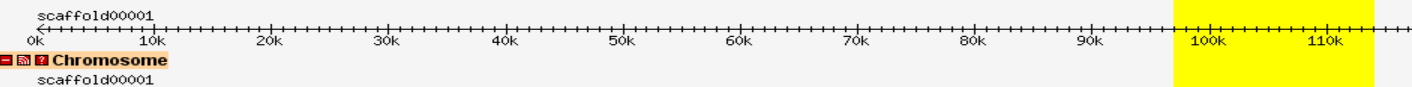
[Bookmark this] [Upload your own data] [Hide banner] [Share these tracks] [Link to Image] [High-res Image] [Help] [Reset]

**Search**  
**Landmark or Region:** scaffold00001:97000..114000 Search

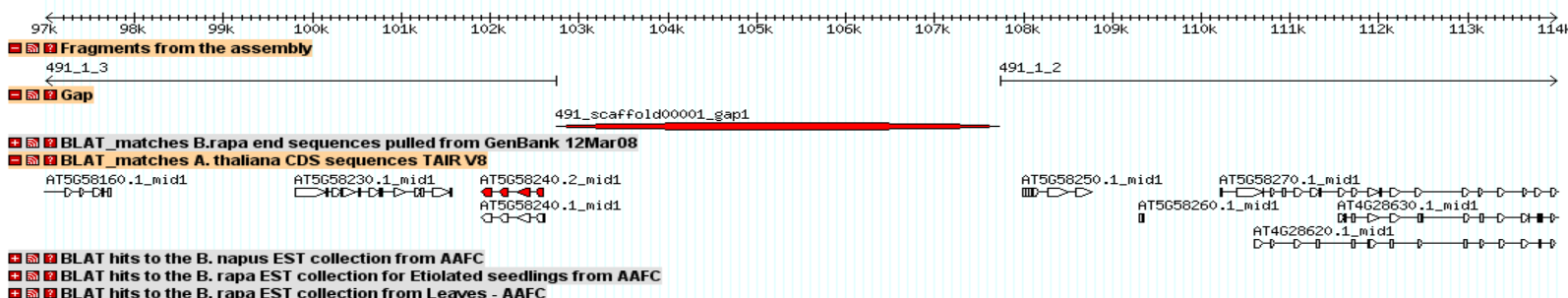
**Data Source**  
KBrH088K06\_phasecolinear

Scroll/Zoom: <<< << < Show 17 kbp > >> >>>  Flip

**Overview**



**Details**



**Tracks**

**Overview**  All on  All off

Chromosome

**Assembly**  All on  All off

Fragments from the assembly  Gap

**DNA**  All on  All off

6-Frame Translation

**Gene**  All on  All off

Find:

Update Image

# CanSeq Project Progress

- *B. rapa* sequencing for R2 and R10
  - “Seed” BACs and then “walking” BAC by BAC
  - 454 FLX and sub-clone sequencing (Sanger – 3730xl)
  - Sequencing 24 BACs per FLX run
  - 60 BACs now sequenced – submitted to GenBank
- 454 Titanium upgrade – December 10<sup>th</sup> 2008
  - 450bp and 1M reads = 500Mbp per run,
  - updated software + new computer cluster; better/faster assembly
- *B. napus* sequencing – 2009 / 2010
  - *de novo* whole genome sequencing / re-sequencing
  - use *B. rapa* sequence as a reference
  - shotgun approach using 454 Titanium (15x fold genome coverage)
  - **Illumina Genome Analyzer II at NRC-PBI in 2009**



# Management of Project Data

- Project Website:
  - immediate data access for project partners, password protected
  - public accessibility for *B. rapa* data
- *B. rapa* data:
  - released to public domain according to MBrGSP release policy
- *B. napus* data (draft and 10 genotypes):
  - release to public domain after proprietary period
- *B. nigra* and *B. oleracea* data
  - release to public domain

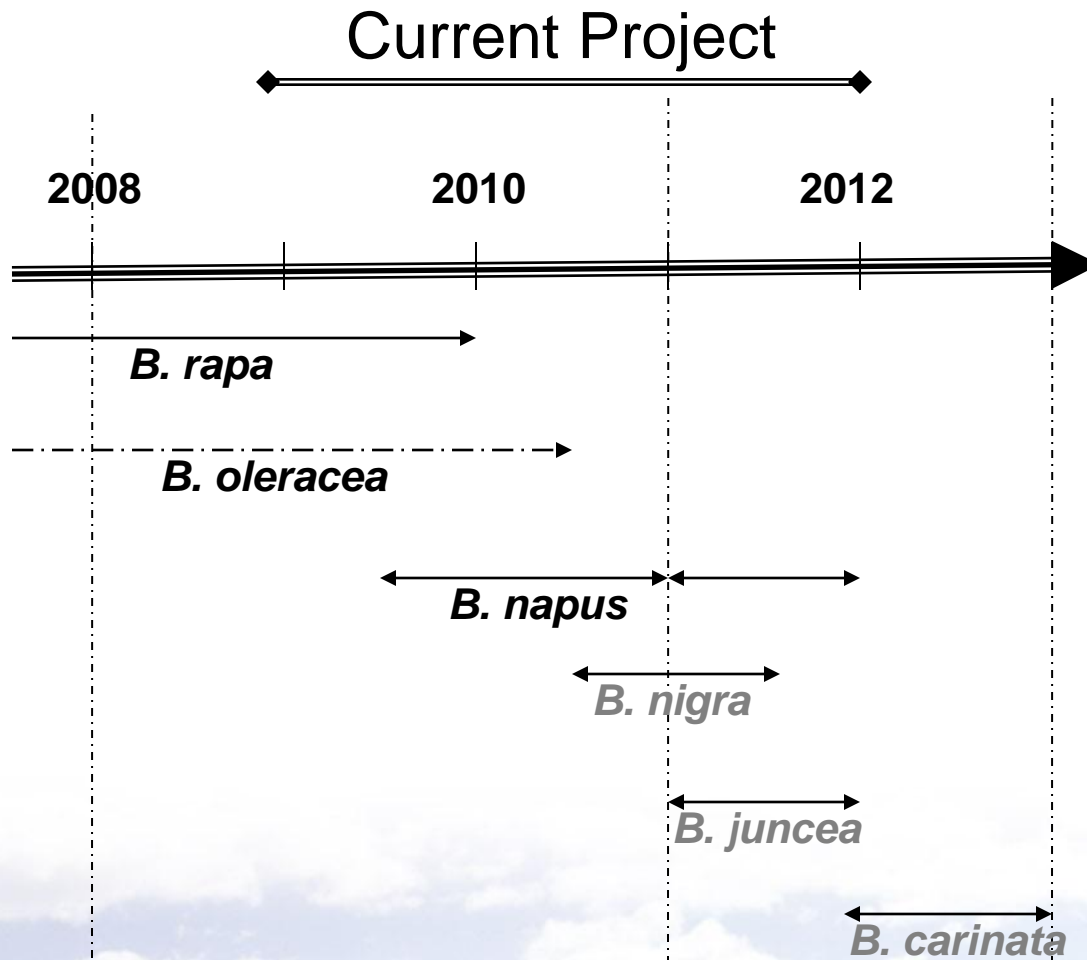


# *Benefits to Industry*

- Early access to *B. napus* sequence data
- Discovery of novel genes for GM application
- Non-GM developments:
  - DNA markers
  - Novel mutation
  - Interspecies transfer of traits
- Role in establishing future priorities and strategies



# Time frame for DNA sequencing of all Brassica U triangle species



# *Acknowledgements*

- NRC-PBI
  - Dr. Faouzi Bekkaoui
  - Jacek Nowak and Kevin Koh
  - Carrie Haimanot and Inge Roewer
- AAFC-SRC
  - Dr. Isobel Parkin
  - Matt Links
  - Rob Wood
- Genome Alberta and Industry Partners
- Wilf Keller and Genome Prairie

