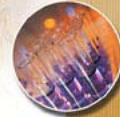




Agriculture and  
Agri-Food Canada

Agriculture et  
Agroalimentaire Canada

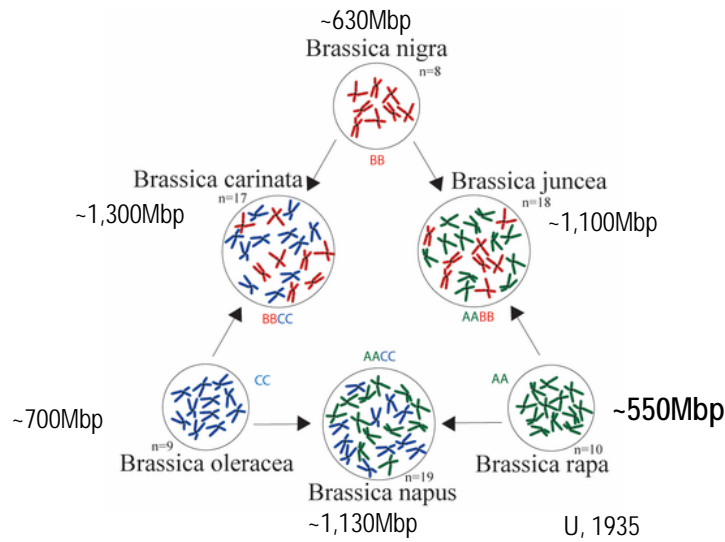


## Harnessing Molecular Diversity in Brassica crops

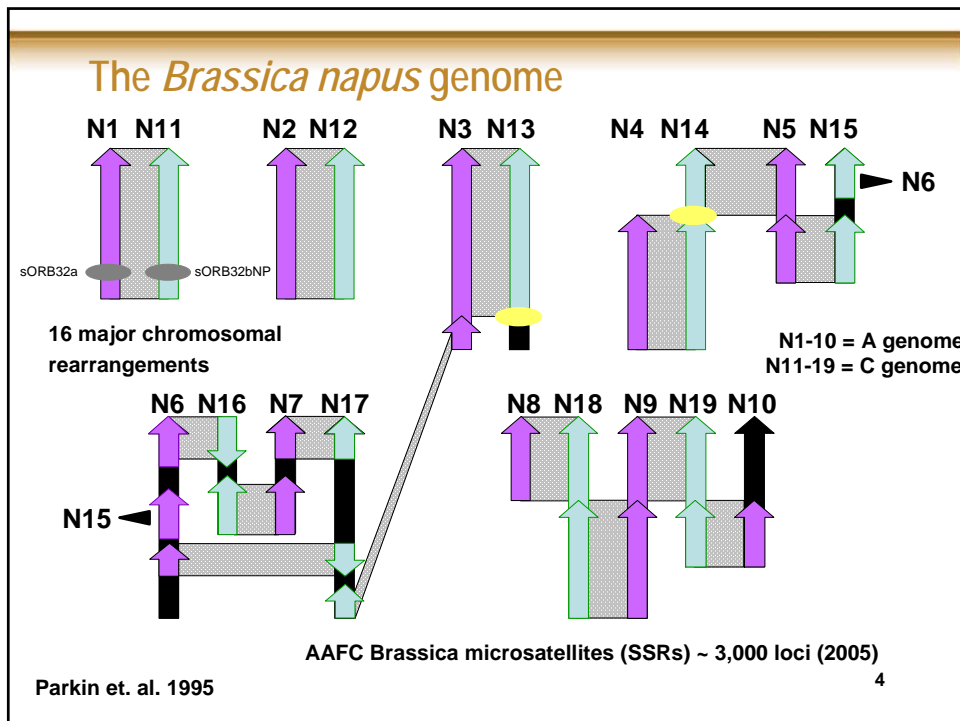
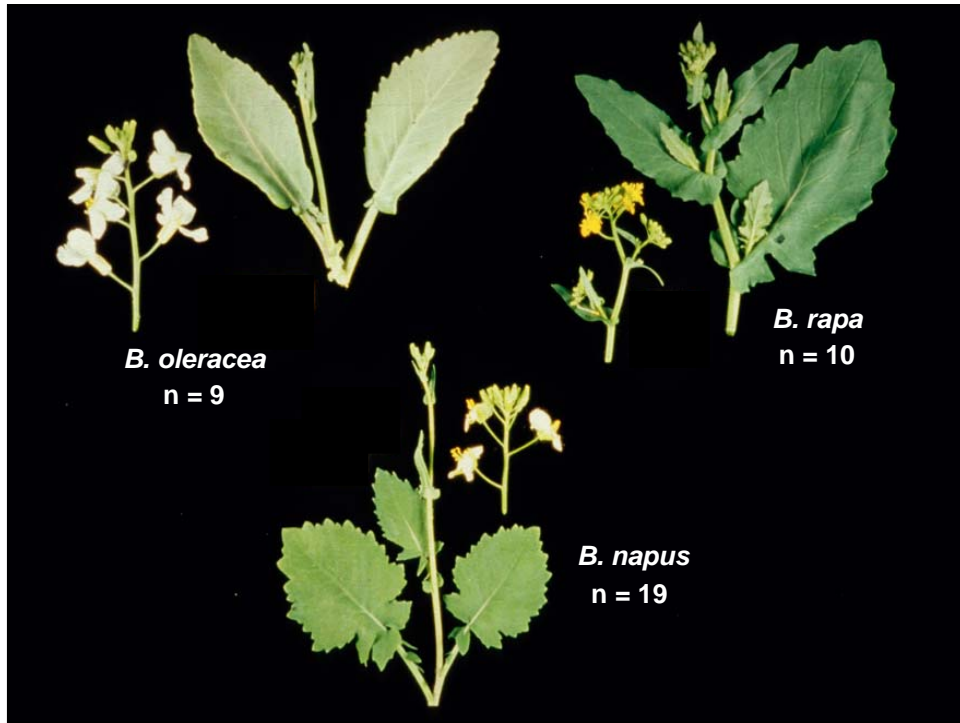
Andrew Sharpe  
AAFC Saskatoon Research Centre  
December 13<sup>th</sup> 2008



### The Brassica crop species



2



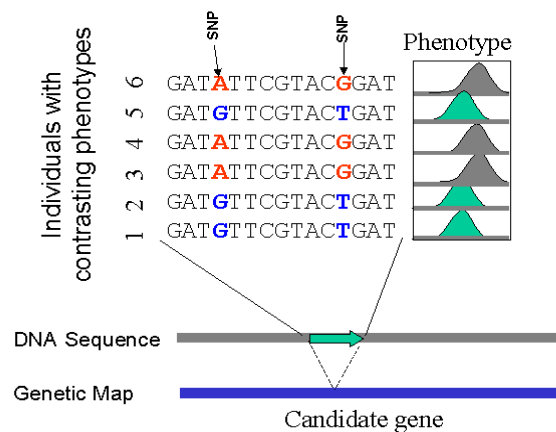
## Single Nucleotide Polymorphisms (SNPs)

- Advantage of SNP markers
  - assay a single locus (bi-allelic)
  - amenable to multiplexing and automation
  - marker-assisted selection
  - discover in sequence databases
- SNP characteristics
  - found in gene coding and non-coding DNA
  - nucleotide substitutions (e.g. C to T)
  - nucleotide insertions and deletions (“indels”)
  - SNPs in genes can lead to amino acid changes
  - potential to associate SNP alleles with phenotype

5

## SNP Haplotypes (Haploid Genotype)

- Multiple co-inherited SNPs at single alleles
- Haplotypes can extend kilobases – HapMaps rather than single SNPs



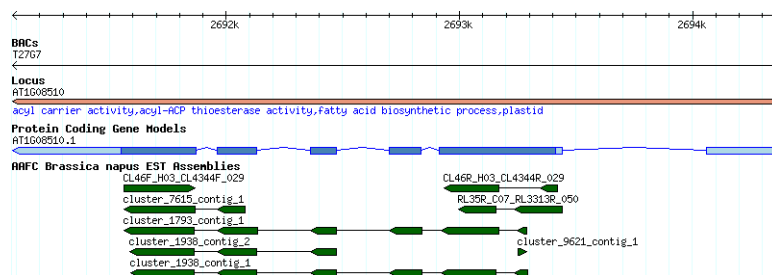
Rafalski 2002

## Brassica SNP Discovery

- eSNP discovery in databases (DNA LandMarks)
  - random genomic sequence data
  - active genes only – expressed sequence tags (ESTs)
  - requires access to sequence data from different genotypes
- Re-sequencing SNP discovery (AAFC)
  - Re-sequencing PCR amplified segments of DNA
  - amplification from diverse genotypes at single loci
  - sequence data from a single genotype is adequate
- AAFC / DNA LandMarks Brassica SNP Discovery Project
  - Initiated in 2006, industry and AAFC MII funding
  - Goal of 2,500 mapped SNPs

7

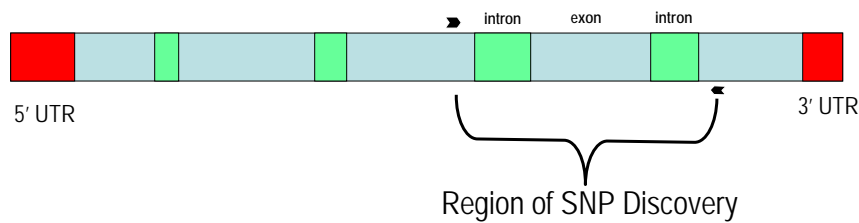
## Brassica / Arabidopsis Genome Browser



- GBrowse enabled genome viewer
  - Displays Brassica ESTs against Arabidopsis gene models
  - Annotation can be inferred

8

## AAFC SNP Discovery



- Identify conserved exons using AAFC / DLM Brassica ESTs and Arabidopsis annotation
  - infer exon/intron boundaries
- Amplify target regions from multiple members of gene families
- Clone and sequence different members and design locus specific primer pairs in introns (more variation)

9

## Consequences of the polyploid *B. napus* genome

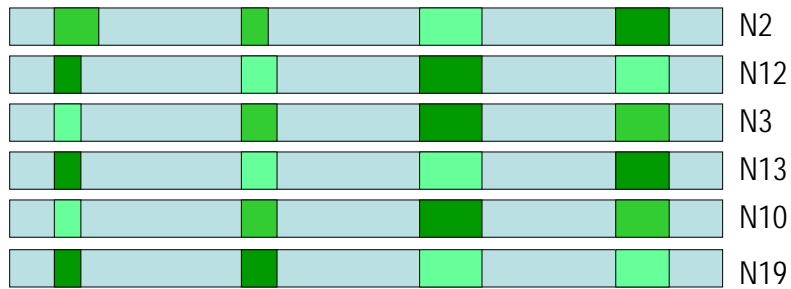
- A and C genomes evolved from a hexaploid ancestor (triplicated)
- 'Families' of related genes across the genome (4-6 members)
- Related genes distributed in conserved blocks
- Multiple members of families expressed
- Traits controlled by related genes

Likely 80,000-120,000 expressed genes in *B. napus*

(26,000 genes predicted in *Arabidopsis*)

10

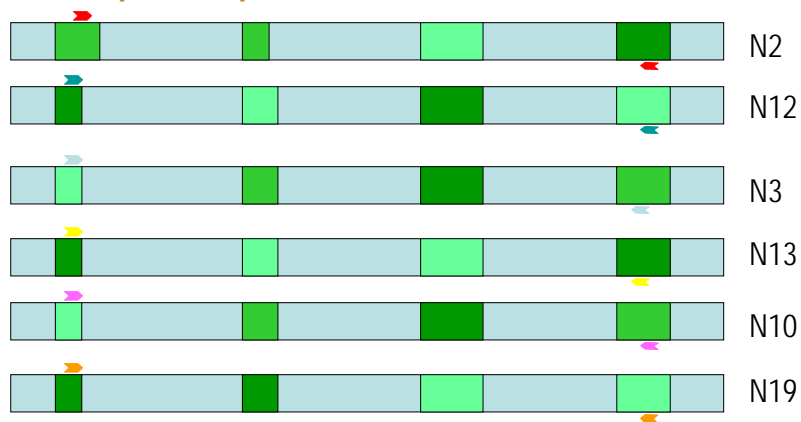
## Clone and sequence members of gene families



- 2-6 different amplicons may be produced from a single genotype using conserved primers
- 24 clones for each gene family sequenced (bi-directional)
- Currently 450 gene families selected for primer design using an automated pipeline

11

## Locus specific primers



- Use intronic variability and design locus specific PCR and sequence amplicons in 32 inbred Brassica lines directly (M13 and T7 tails for routine sequencing)
- Currently 219 locus specific primer pairs using automated software

12

## Panel of 32 Brassica accessions

No.	Species	Variety	Crop Type	No.	Species	Variety	Crop Type
1	<i>B. napus</i>	DH12075	Spring	17	<i>B. rapa</i>	ACdsc	Spring
2	<i>B. napus</i>	PSA12	Resyn	18	<i>B. rapa</i>	Candle	Spring
3	<i>B. napus</i>	Quantum	Spring	19	<i>B. rapa</i>	R-o-18	Spring
4	<i>B. napus</i>	Paroll	Spring	20	<i>B. rapa</i>	Maleksberger	Spring
5	<i>B. napus</i>	Dunkeld	Spring	21	<i>B. rapa</i>	SVAnte	Spring
6	<i>B. napus</i>	Global	Spring	22	<i>B. rapa</i>	Reward	Spring
7	<i>B. napus</i>	Westar_10	Spring	23	<i>B. rapa</i>	Pluto	Winter Turnip
8	<i>B. napus</i>	Zhongyou_821	Chinese Winter	24	<i>B. rapa</i>	Per	Winter Turnip
9	<i>B. napus</i>	Ningyou_7	Chinese Winter	25	<i>B. oleracea</i>	A12DHd	Chinese Kale
10	<i>B. napus</i>	Tapidor	Winter	26	<i>B. oleracea</i>	Bejo1	Cauliflower
11	<i>B. napus</i>	Capitol	Winter	27	<i>B. oleracea</i>	Bejo2	Kohl Rabi
12	<i>B. napus</i>	Lirajet	Winter	28	<i>B. oleracea</i>	HRIGRU4816	Intermediate
13	<i>B. napus</i>	Mohican	Winter	29	<i>B. oleracea</i>	HxB_1280	Cabbage
14	<i>B. napus</i>	Samourai	Winter	30	<i>B. oleracea</i>	Coral_Queen	Kale
15	<i>B. napus</i>	Hamburger	Winter	31	<i>B. oleracea</i>	Purple_Vienna	Kohl Rabi
16	<i>B. napus</i>	Express	Winter	32	<i>B. oleracea</i>	Senna	White Kale

Highlighted *B. napus* lines indicate parents of DH mapping populations

13

## Haplotypes from *B. napus* accessions (Is102B)

Atg36220 (Is102B) ferulic acid hydroxylase (FAH)

## SNP haplotype analysis in *B. napus*

- A or C genome origin of loci in *B. napus* deduced from comparison with *B. rapa* and *B. oleracea* sequences
- Polymorphism identified between DH12075 x PSA12, Quantum x Paroll, Ningyou 7 x Tapidor and Hamburger x Samurai
  - 49% polymorphic in DH12075 x PSA12 (SG population)
  - 70% polymorphic in all crosses
- Average of 2.2 haplotypes per locus in *B. napus*
- Polymorphism Information Content (PIC) values range from 0.12 - 0.73 with an average of 0.39 for loci with more than one haplotype
- Higher number of transitions than transversions, indels quite common
- Higher diversity in *B. oleracea* and *B. rapa*

15

## Mapping SNPs on *B. napus* reference maps

- Length Polymorphism (MegaBACE)
  - need significant length difference (indels)
  - M13 and T7 tails on locus specific primer pairs for labelling
- Melting Curve Analysis (StepOne Real-time PCR)
  - locus specific primers and an internal probe covering the SNP(s)
  - LC Green melting dye (saturation dye) for higher resolution melting curve
  - issues with reproducibility
- Haplotyping (sequencing with 3730xl DNA Analyzer at NRC-PBI)
  - optimised protocol using high throughput analyzer
  - very robust (multiple SNPs)
  - current method for mapping loci
  - multiplexing not possible

16

## SNP haplotyping in SG mapping population (Is155C)

Sequencher [Contig[155C\_SG]]

File Edit Select Contig Sequence View Window Help

Overview Summary Cut Map Find Show Chromatograms ReAligner

SNP\_SEQUENCING\_PLATE3\_R\_817\_07SEP2007\_079  
 SNP\_SEQUENCING\_PLATE3\_R\_818\_07SEP2007\_079  
 SNP\_SEQUENCING\_PLATE3\_R\_819\_07SEP2007\_080  
 SNP\_SEQUENCING\_PLATE3\_R\_820\_07SEP2007\_080  
 SNP\_SEQUENCING\_PLATE3\_R\_017\_07SEP2007\_077  
 SNP\_SEQUENCING\_PLATE3\_R\_018\_07SEP2007\_077  
 SNP\_SEQUENCING\_PLATE3\_R\_D20\_07SEP2007\_079  
 SNP\_SEQUENCING\_PLATE3\_R\_F17\_07SEP2007\_075  
 SNP\_SEQUENCING\_PLATE3\_R\_F19\_07SEP2007\_076  
 SNP\_SEQUENCING\_PLATE3\_R\_F20\_07SEP2007\_076  
 SNP\_SEQUENCING\_PLATE3\_R\_H17\_07SEP2007\_075  
 SNP\_SEQUENCING\_PLATE3\_R\_H18\_07SEP2007\_075  
 SNP\_SEQUENCING\_PLATE3\_R\_H19\_07SEP2007\_074  
 SNP\_SEQUENCING\_PLATE3\_R\_H20\_07SEP2007\_074  
 SNP\_SEQUENCING\_PLATE3\_R\_I17\_07SEP2007\_071  
 SNP\_SEQUENCING\_PLATE3\_R\_I18\_07SEP2007\_071  
 SNP\_SEQUENCING\_PLATE3\_R\_I19\_07SEP2007\_072  
 SNP\_SEQUENCING\_PLATE3\_R\_I20\_07SEP2007\_072  
 SNP\_SEQUENCING\_PLATE3\_R\_L17\_07SEP2007\_069  
 SNP\_SEQUENCING\_PLATE3\_R\_L18\_07SEP2007\_069  
 SNP\_SEQUENCING\_PLATE3\_R\_L20\_07SEP2007\_070  
 SNP\_SEQUENCING\_PLATE3\_R\_N17\_07SEP2007\_067  
 SNP\_SEQUENCING\_PLATE3\_R\_N18\_07SEP2007\_067  
 SNP\_SEQUENCING\_PLATE3\_R\_N19\_07SEP2007\_069  
 SNP\_SEQUENCING\_PLATE3\_R\_N20\_07SEP2007\_066  
 SNP\_SEQUENCING\_PLATE3\_R\_P17\_07SEP2007\_065  
 SNP\_SEQUENCING\_PLATE3\_R\_P18\_07SEP2007\_065  
 SNP\_SEQUENCING\_PLATE3\_R\_P19\_07SEP2007\_066  
 SNP\_SEQUENCING\_PLATE3\_R\_P20\_07SEP2007\_066

610 620 630 640 650 660 670 680

Multiple diagnostic SNPs provides very robust segregation data

## SNP haplotyping converted to segregation data (Is153B)

Segregation Data for top of N3

*pW153-eNP	+-----+-----+-----+-----+
*sN1830a	+-----+-----+-----+-----+
*pW153-b	+-----+-----+-----+-----+
*p07-c	+-----+-----+-----+-----+0+--
<b>*Is153B</b>	<b>0-----+0+-----+-----+-----+0+--</b>
*sC079a	+-----+-----+-----+-----+0+--
*sN13036	+-----+-----+-----+-----+-----
*sn2031a	+-----+-----+-----+-----+-----

- SNP haplotype segregation data is integrated into existing maps
- existing genetic maps comprised of many RFLP and SSR markers
- dense maps allow placement of new loci with few individuals

## Brassica SNP Discovery Project Goals

- SNP Discovery
  - 500 gene families, 2,000 loci screened and 1,500 mapped loci (AAFC)
  - 1,500 bi-allelic SNP markers screened and 1,000 mapped loci (DLM)
  - high throughput SNP genotyping on Sequenom MassARRAY (DLM)
  - 1/3<sup>rd</sup> of all data will be made publicly available
- Development of large mutagenized *B. napus* population (AAFC)
- Induced and natural variation in 100 selected genes (AAFC and DLM)
  - induced variation in the mutagenized *B. napus* population
  - natural variation in diverse Brassica germplasm
- Utilization of Next Generation Sequencing Platforms (AAFC)
  - pilot project to establish SNP discovery on a selected platform
  - possibility of simultaneous SNP discovery and genetic mapping
  - possibility of rapid mutation discovery

19

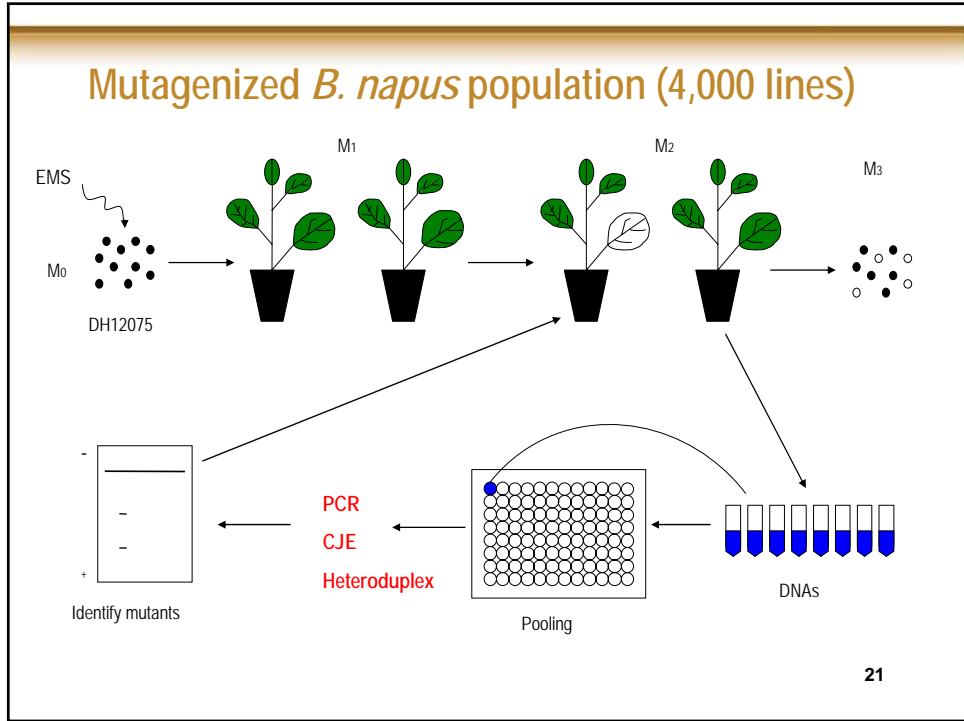
## Induced variation in *B. napus*

- Heavily mutagenized population (4,000 lines) generated using EMS
- Irreversible point mutations in DNA
- High throughput analysis possible (TILLING methodology)
  - requires locus specific primer pairs for detection
  - identifies induced alleles in any target gene
- Crossing and phenotype determination after detection

## Natural variation in Brassica germplasm

- Identification of natural alleles in diverse germplasm (ecoTILLING)
- Same technique but higher level of allelic diversity detected

20



### Induced mutations detected in one FAD2 gene

M2 plants carrying mutations	M1 parent	Size forward fragment (bp)	Size reverse fragment (bp)	Mutation	Transition	Genotype of M2	Codon change	Amino acid change
N21-1	21	158	156	fad2-1	G/C to A/T	het	gcg to gca	A to A
N21-3	21	158	156	fad2-1	G/C to A/T	het	gcg to gca	A to A
N21-2	21	168	144	fad2-2	C/G to T/A	het	acg to atg	T to M
N21-6	21	168	144	fad2-2	C/G to T/A	hom	acg to atg	T to M
N24-2	24	106	208	fad2-3	C/G to T/A	het	cag to cac	Q to H
N24-7	24	106	208	fad2-3	C/G to T/A	het	cag to cac	Q to H

Estimate of mutation frequency  
 2 genes: 7 mutations in 73 M1 plants  
 1 mutation in 108bp or  
 1 mutation per 7.9kb per M1 plant

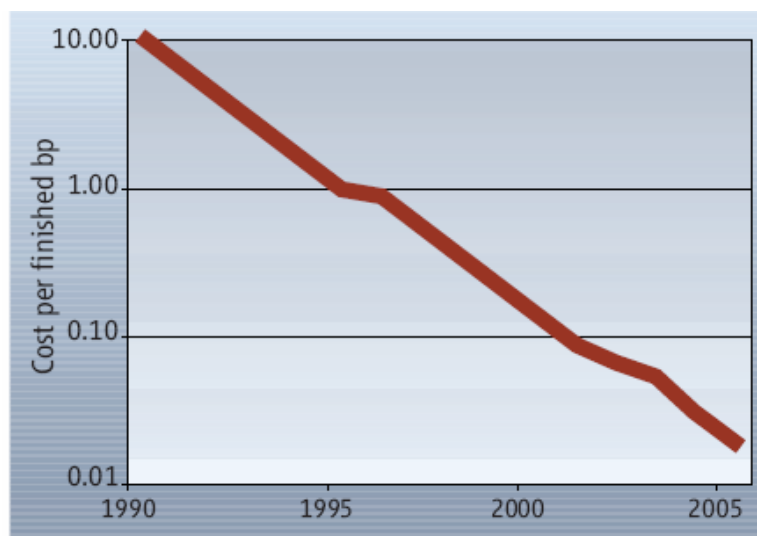
22

## Quantum leaps in DNA sequencing technology

- The first generation of automated sequencers read about 5000 base pairs per day. Today's machines, which use an improved version of the same technique, sequence about a million bases a day (1Mbp)
- The next (or second) generation will sequence 100 to 300 million bases a day (100-300Mbp)
- The third generation machines will sequence 3 billion bases a day (3Gbp) and more in the near future

23

## Decreased cost of finished DNA sequence



Service 2006, Science <sup>24</sup>

## Next Generation DNA Sequencing

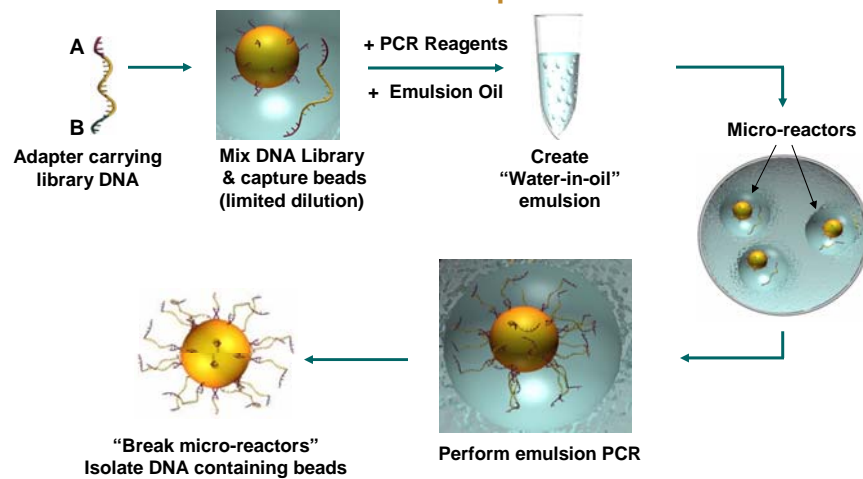
=> Three main technologies:

- Roche 454 FLX (100Mbp, 250bp => 1Gb, 450bp in 2008)
- Illumina 1G Genetic Analyzer (1Gbp, 35bp reads)
- Applied Biosystems SOLiD Sequencer (3Gbp, 35bp reads)



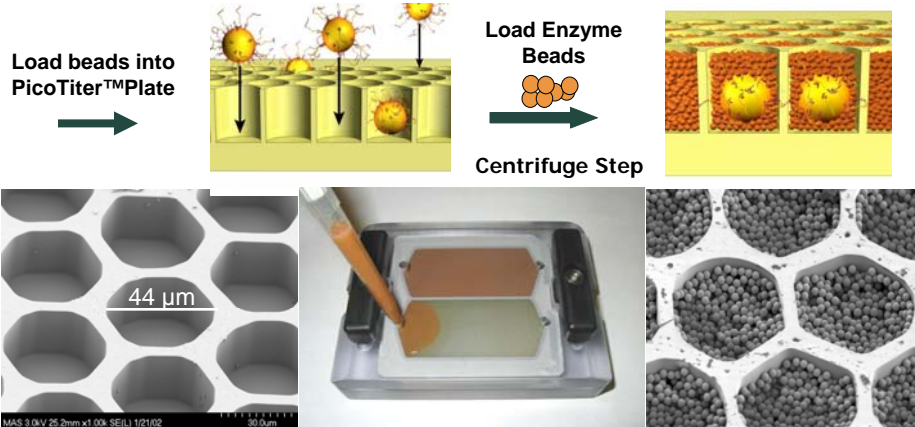
25

## Emulsion Based Clonal Amplification



- Generation of millions of clonally amplified sequencing templates on each bead
- No cloning and colony picking

## Depositing DNA beads into the PicoTiter™ Plate



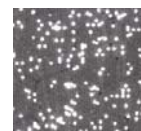
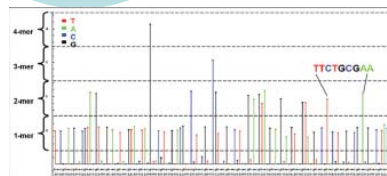
www.roche-applied-science.com

454 LIFE SCIENCES

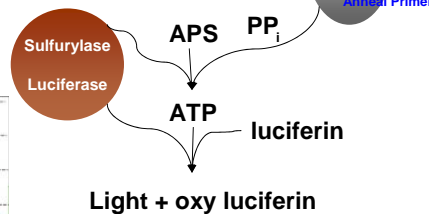
## Sequencing-By-Synthesis

- Simultaneous sequencing in hundreds of thousands of picoliter-size wells
- Pyrophosphate signal generation

DNA Capture Bead Containing Millions of Copies of a Single Clonal Fragment



A A T C G G C A T G C T A A A G T C A T



www.roche-applied-science.com

454 LIFE SCIENCES

## Applications for Next Generation DNA Sequencing

- *De novo* DNA sequencing
  - BACs, whole genomes (*B. rapa* genome)
- DNA re-sequencing
  - different genotypes of already sequenced genome
- Gene expression (transcriptome) studies
  - cDNA, SAGE
- **SNP discovery / mutation discovery**
- Small RNA discovery
- Metagenomics

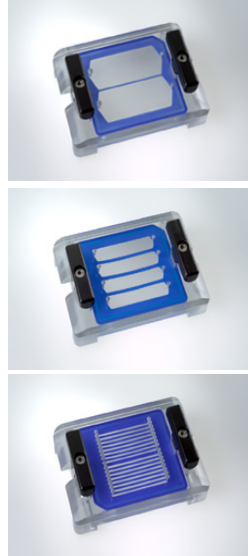
29

## Platforms are optimal for different applications

- *De novo* sequencing of complex genomes
  - 454 FLX has better assembly with 250bp reads
- Deep transcriptome profiling
  - SOLiD has higher number of reads vs 454 FLX

30

## SNP and mutation discovery using 454 FLX

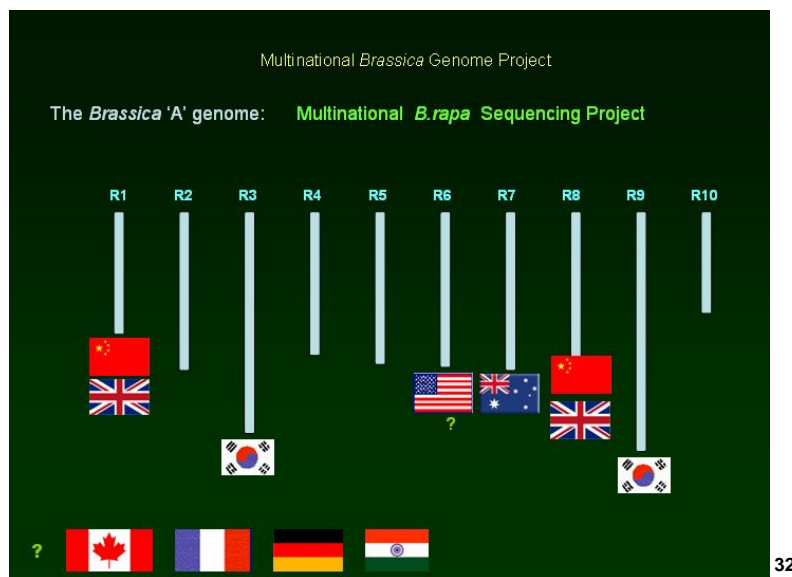


- 250-450bp reads will provide SNP haplotypes
- Gasket and "index" tags will allow multiplexing
  - pooling of different amplified products
  - pooling from different genotypes
- SNP discovery
  - 32 diverse lines
- SNP mapping
  - 32 parental and DH progeny lines
- Mutation discovery
  - induced variation in mutagenized populations
  - natural variation in large germplasm panels
  - 3-D pooling strategies

www.roche-applied-science.com

454 LIFE SCIENCES

## *B. rapa* Sequencing Project – NRC / AAFC proposal



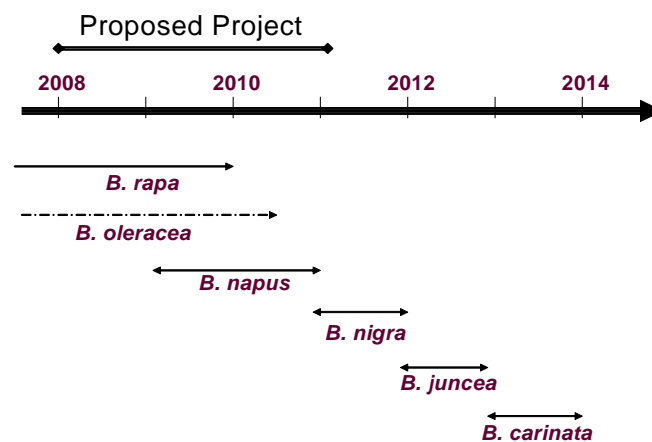
32

## *B. rapa* Sequencing Project – NRC / AAFC proposal

- **Multinational component**
  - Phase II of the Multinational Brassica Genome project
  - 2 *B. rapa* chromosomes (R2 and R10)
  - ~800 BACs (~120MB of DNA) sequenced in pools of 5 BACs
  - existing Sanger (3730xl) and new 454 FLX sequencer @ NRC-PBI
- ***B. napus* genome sequencing**
  - use *B. rapa* sequence as “reference”
  - Shotgun approach using 454 FLX only
  - 15 runs of 1 GB each with 450bp reads
  - Approximate 14 x genome coverage
- **Total budget of \$1.6M**

33

## Estimation for completion of the DNA sequencing of all the Brassica U triangle species



34

## Acknowledgements

- Brassica SNP Discovery:
  - Derek Lydiate (mutagenized *B. napus* population)
  - Christine Sidebottom, Wayne Clarke, Daijun Yang and Brent Mooney
  - Wing Chueng and Charles Pick @ DNA LandMarks
  - Larry Pelcher @ NRC-PBI
  - Industry Partners and AAFC MII Funds
- Canadian Canola Genome Sequencing Initiative (CanSEQ)
  - Larry Pelcher, Faouzi Bekkaoui, Isobel Parkin and Wilf Keller
  - Jerome Konecni and Reno Pontarollo @ Genome Prairie

35

