



NGS Platforms - Applications in Plant Genomics

Andrew Sharpe

Plant Biotechnology Institute

NGS / Bioinfo Workshop January 20th 2011



National Research
Council Canada

Conseil national
de recherches Canada

Canada

Roche 454 FLX System

- Titanium upgrade in 2009
- >1M reads per run, 450bp read length, >99.5%
- >500 million bases (500Mbps) per run
- Eliminates cloning and colony-picking
- Variety of template DNA (amplicons, gDNA, cDNA, BACs, etc.)
- Shotgun (450bp) and long paired end / mate (3-40kb) libraries
- Variety of applications (*De novo* sequencing, transcriptome, amplicon, metagenomic, etc.)
- Sequence multiple samples in one run with 12 available Multiplex Identifiers (MID) adaptors



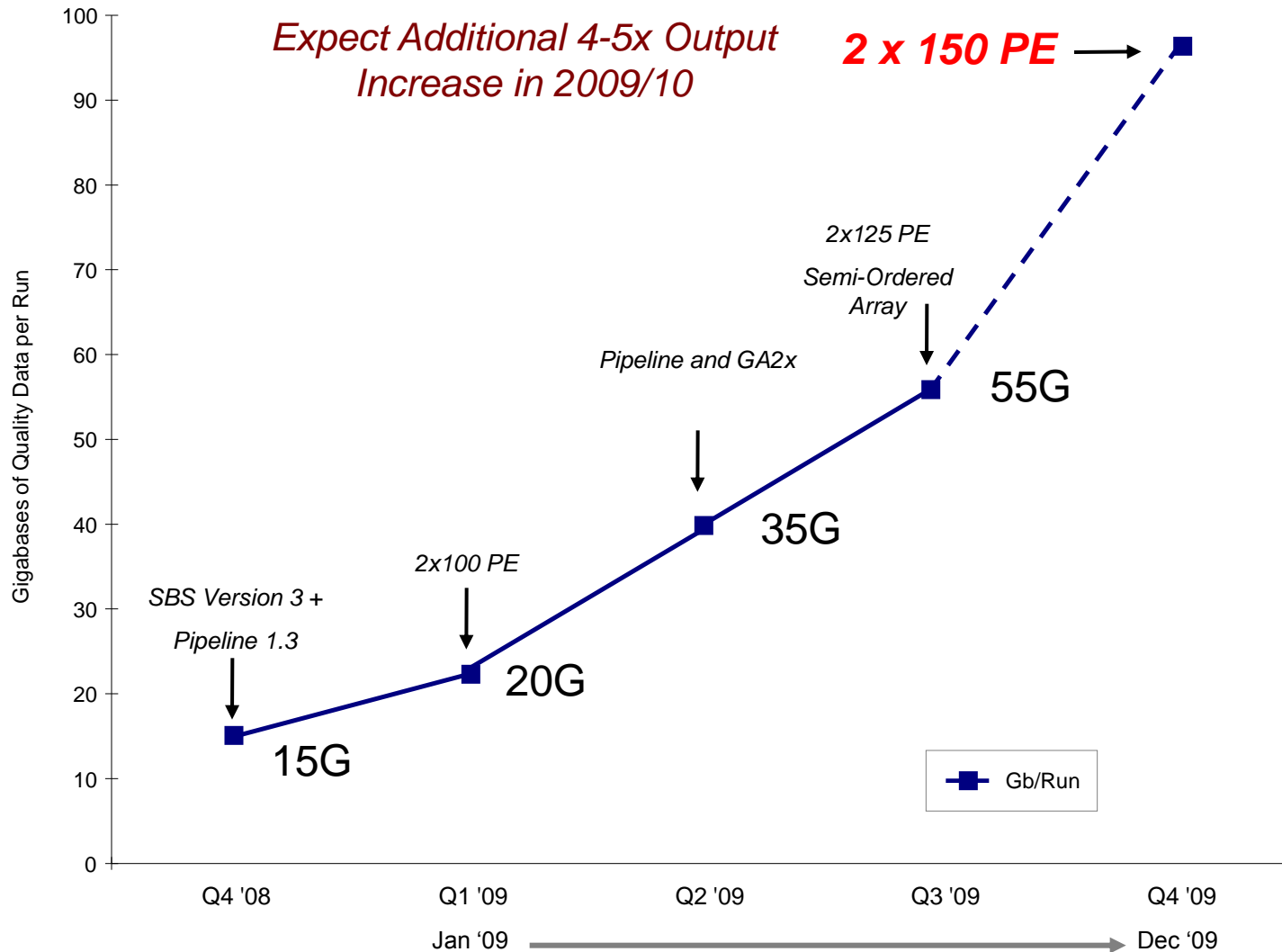
Illumina Genome Analyzer Ix (2009)

- >50 million reads per flowcell (single read)
- >1.5GB per single read flowcell (36bp read)
- >3.0GB per PE flowcell (36 bp read)
- >750MB/day
- 2 day single read run, 4 day PE run
- Supported read length: 50bp
- System enabled for 75bp+ reads
- Short Insert Paired End (200-600bp)
- Long Insert Mate Pairs (3-5kb)
- 12 index for multiplexing
- *De novo*, re-sequencing, RNA-Seq, smallRNA, CHIP Seq, Meth-Seq



Illumina GAIIx improvement (2010)

95Gb – Available
(Aug 2010)

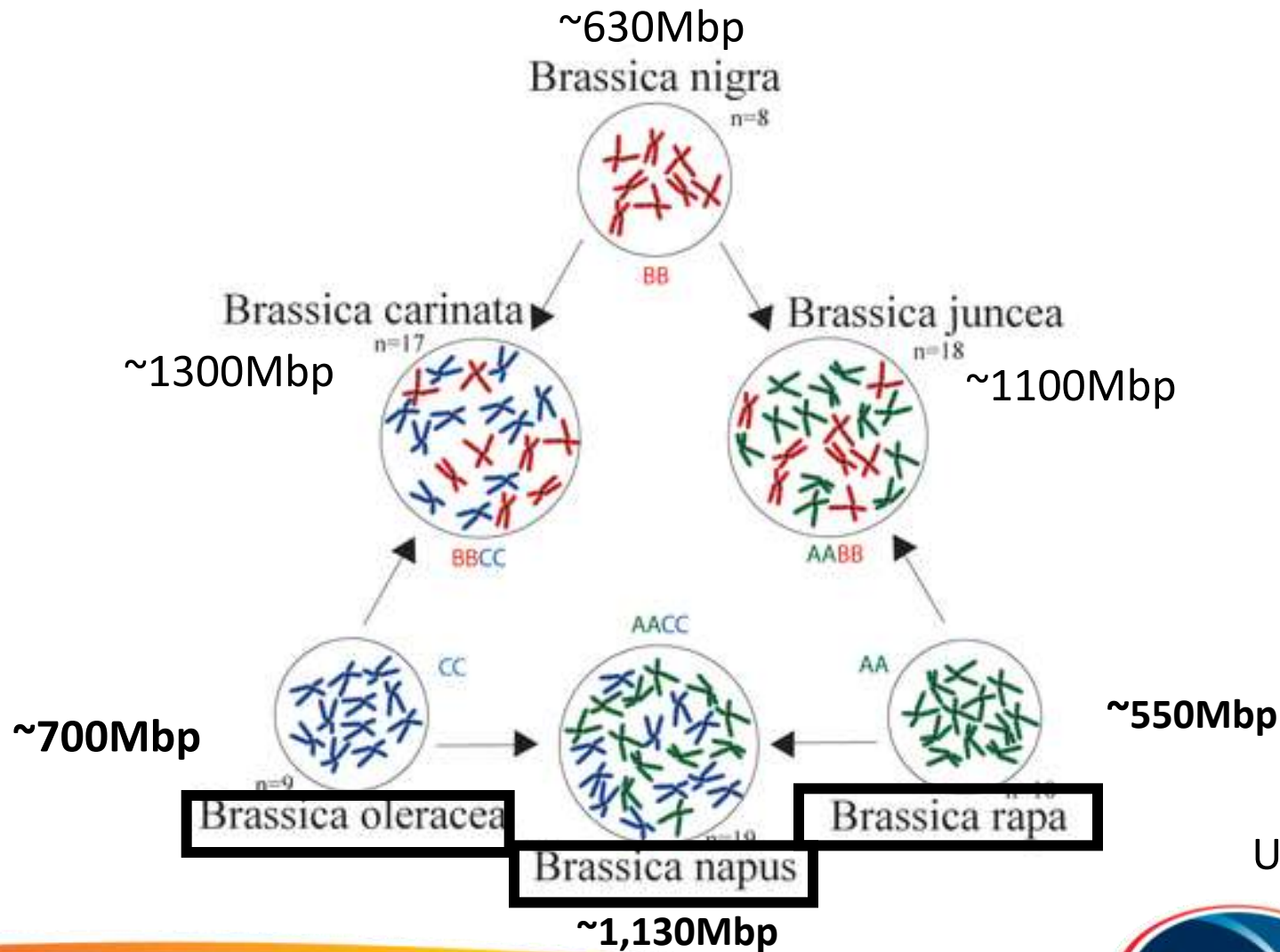


HiSeq 2000 Platform (2011)



- Upgraded platform to existing Illumina GAllx platform
- Produces 200Gb data per run on two flow cells – independent; going up to **400Gb by end of 2010; 600Gb 2011**
- Utilizes existing infrastructure for GAllx – cluster station, bioinfo, etc
- Support more flexible sequencing of Illumina samples – eg. one flow cell small RNA , other genomic paired
- Schedule flexibility – 4 flow cells instead of 2; HiSeq faster than GAllx - 8 days for 100bp paired runs
- New library kits – faster & higher multiplexing

U's Triangle for Brassica species



U, 1935

Why is sequencing Brassica genomes 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 – promoter / enhancers
- Basis for global mutation discovery
- Re-sequencing different genotypes

CanSeq Project (2009-12)

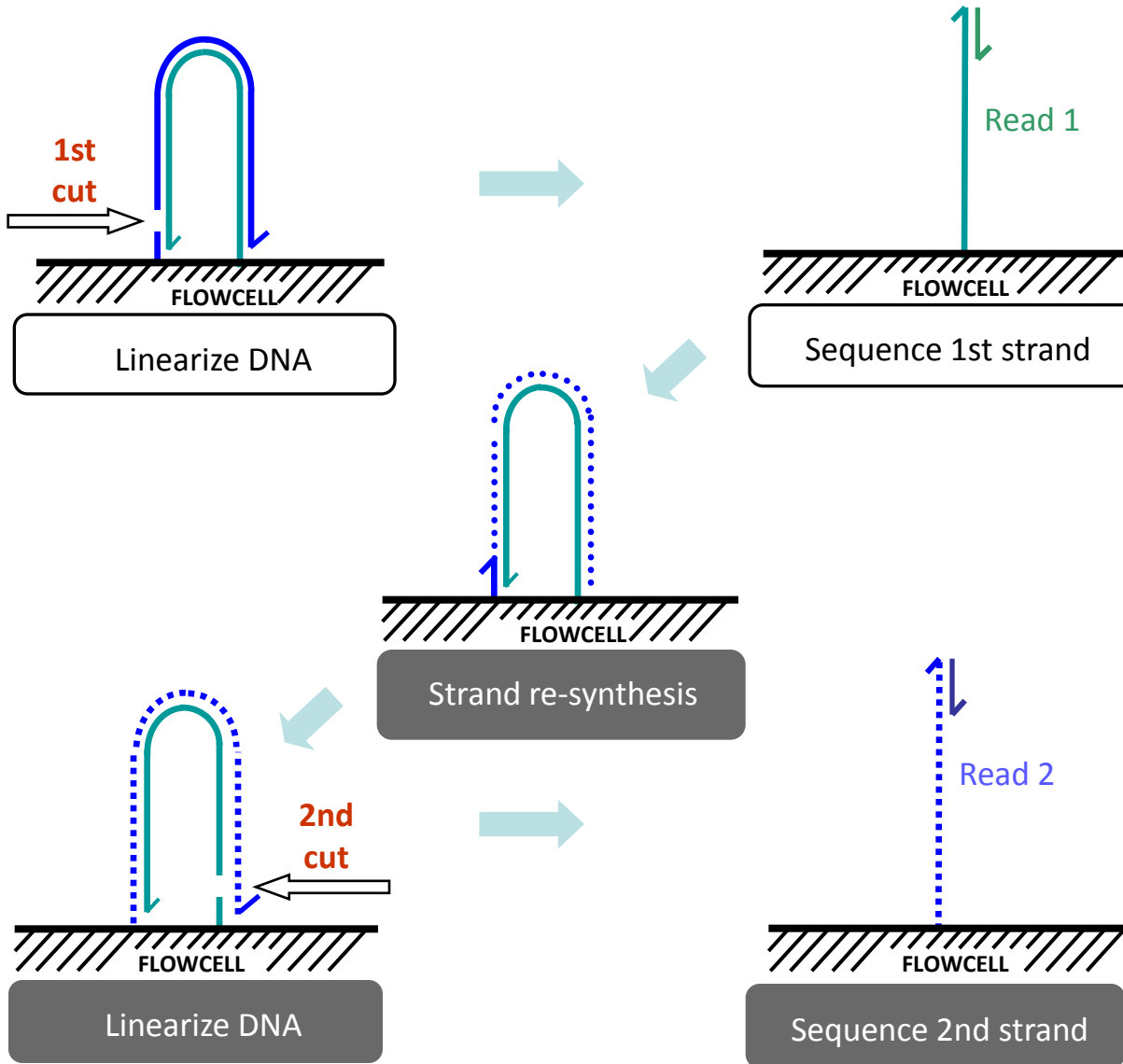
- Canadian Canola Genome Sequencing Project (CanSeq)
- National Research Council / Agriculture Canada joint project
 - funding from NRC, AAFC, Genome Alberta and 9 industry partners
 - \$2.5M project
- Validate China *B. rapa* Chiifu draft (96 BACs in triplicated regions)
- Re-sequence oilseed *B. rapa* genotype using new draft (Illumina @ PBI)
- *de novo* whole genome sequence of *B. oleracea* genome (TO1000)
 - US (Pires / Town), UK (Barker) & France (Chalboub) (454/Illumina)
- Draft sequence of *B. napus* – use *B. rapa* and *B. oleracea*
- Re-sequencing of 16 *B. napus* lines
- Resource development for *B. nigra* (B genome)
 - BAC end sequencing and fingerprinting
 - Draft sequence of *B. nigra* genome
 - Draft sequence of *B. juncea* and *B. carinata*

Why *Brassica oleracea* (C genome)?

- Parental genome of *B. napus* & *B. carinata*
Similar traits, controlled by homologous genes
- Simpler genome organization
Effectively halves level of duplication for simpler data analysis
- Extensive resources already developed
TO1000 – ½ genome coverage of shotgun Sanger, BAC end sequences
- Strategy to utilize a hybrid assembly – Illumina, 454 and Sanger data with the CABOG assembler

Illumina Paired-End Sequencing

Cluster amplification



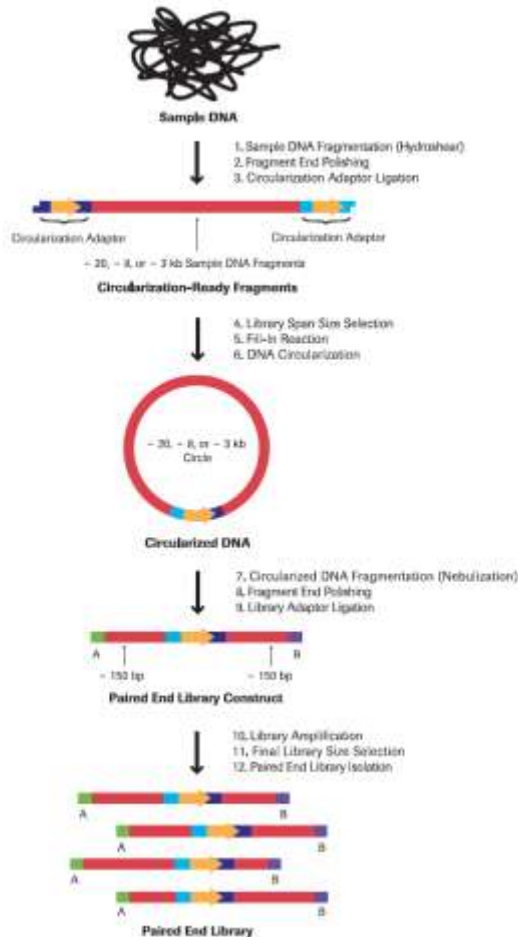
Roche 20kb and 40kb paired-end libraries

- Illumina mate libraries (2 x 35bp) not functional for Celera Assembler + high levels of redundancy and chimerism (lack of “join” index in circularized molecules)

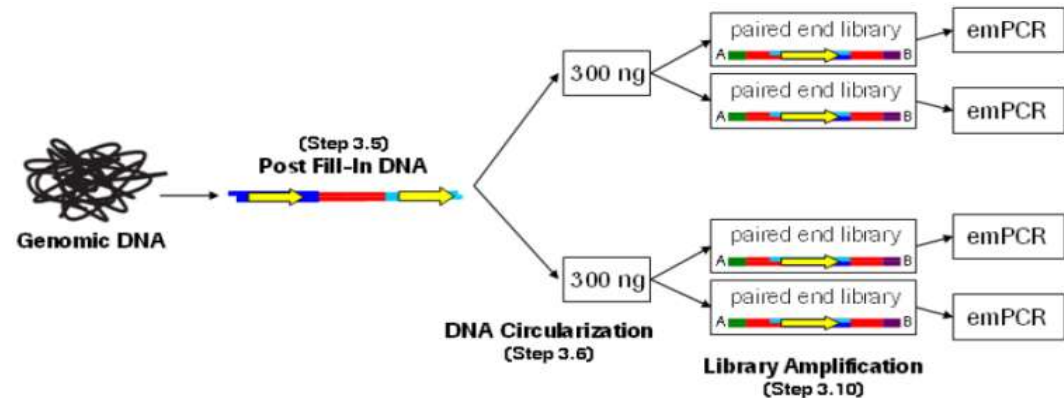
- The starting material was 45 µg of high quality nuclei prep genomic *Brassica oleracea* TO-1000 DH3 DNA.

- Two slices were extracted from the one lane (20 kb and 40 kb). From this point, the two libraries were processed side by side.

- Eight libraries were prepared, four libraries each of 20 and 40 kb spans, as libraries generated were doubled at the DNA circularization and library amplification steps.



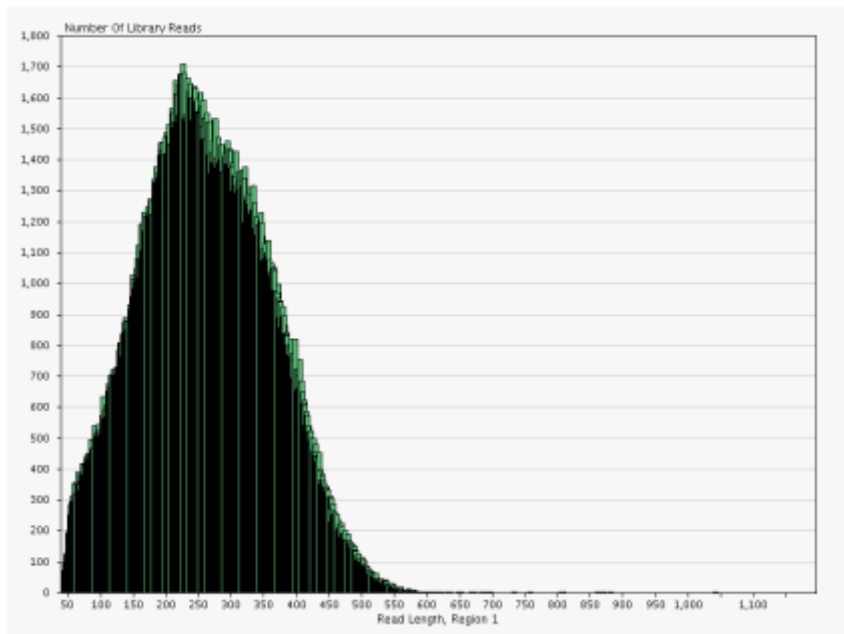
150bp paired “mates”



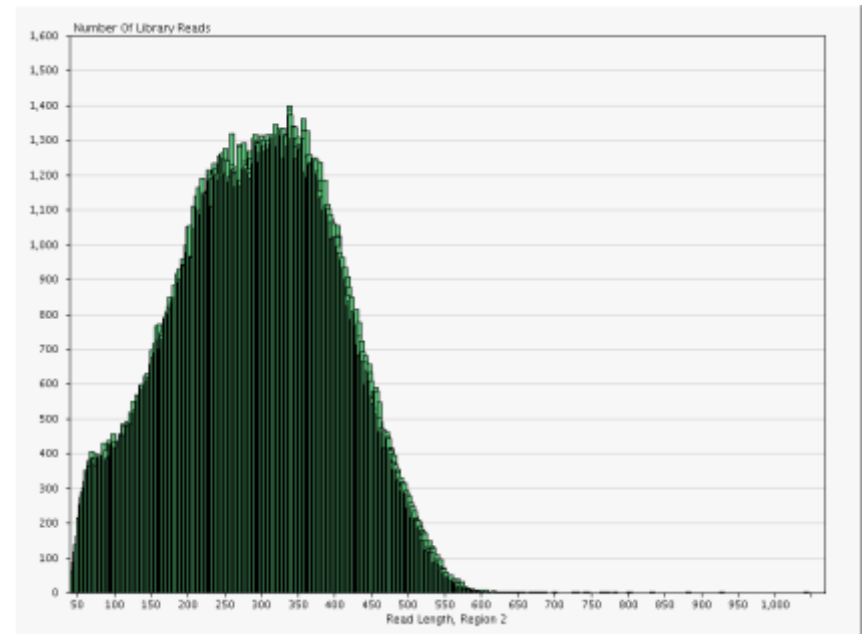
Roche 20kb and 40kb paired-end libraries



- Difficult to achieve high resolution separation for DNA molecules much larger than 10 kb – trapping of smaller fragments.
- Field Inversion gel electrophoresis (FIGE) to resolve larger fragment more accurately.



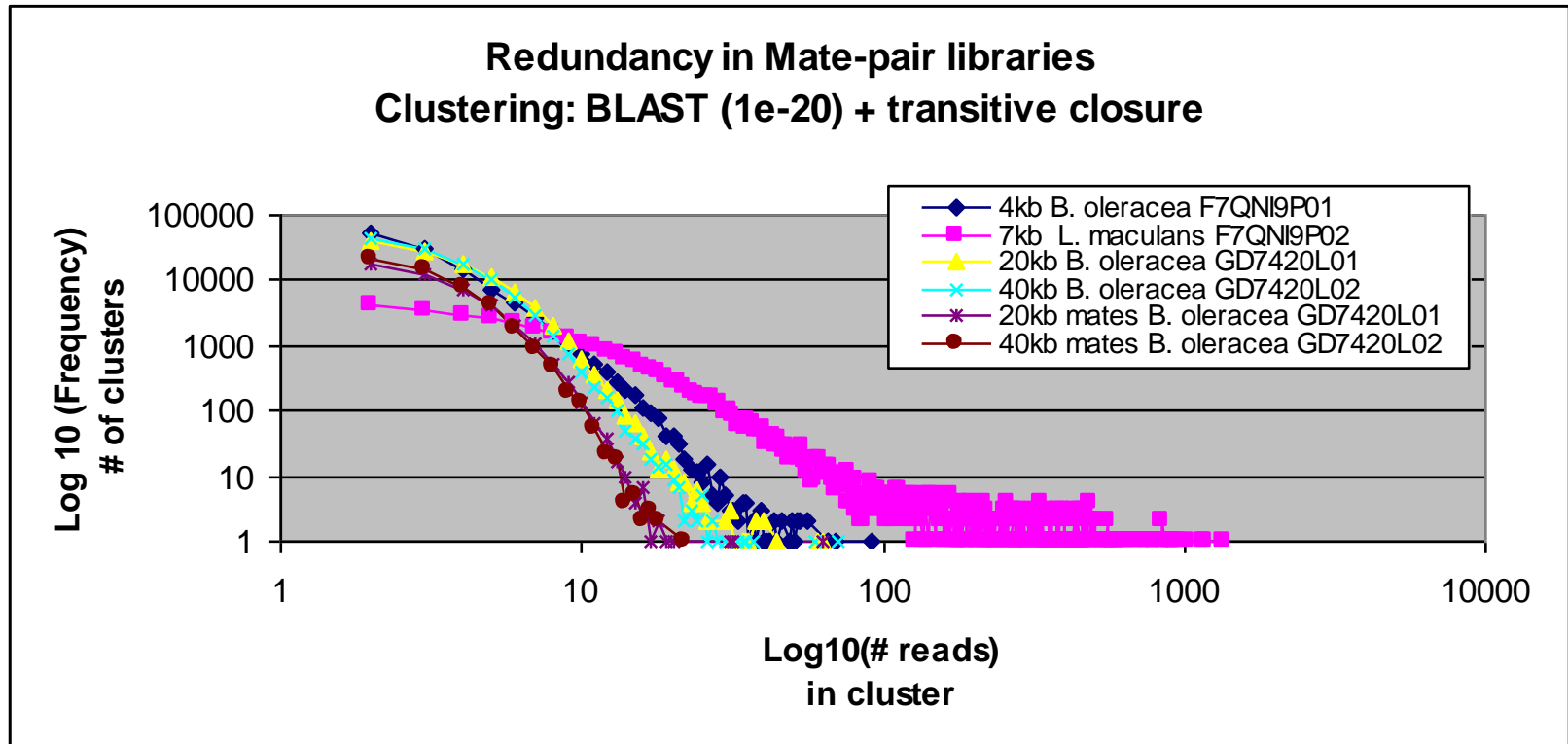
20A (Region 1, 20kb)



40A (Region 2, 40kb)

- Libraries sequenced on 454; 225 Mbp generated (smaller fragments = lower yield)

Checking for redundancy in Mate-pair libraries

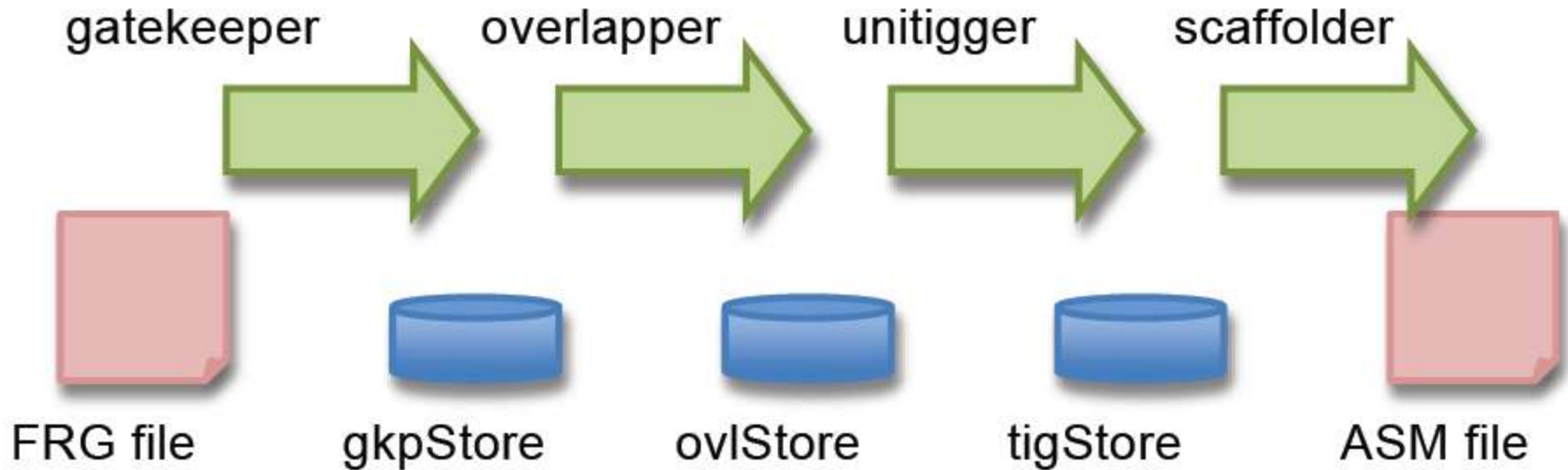


Data used in TO1000 assembly – 39Gb

Date	Source	Platform	RunName	Type	NumReads	TotalBases	AveLength	Expected Mate Distance (bp)	Note
06/05/09	AAFC/PBI	454 Titan	oleracea	wgs	1,334,905	536,840,143	402	--	FVF0U4Z
29/05/09	AAFC/PBI	454 Titan	oleracea2	wgs	1,384,005	495,952,143	358	--	FWMLSQ
05/06/09	AAFC/PBI	454 Titan	oleracea3	wgs	1,413,282	482,609,060	341	--	FW1DCZT
10/06/09	AAFC/PBI	454 Titan	oleracea4	wgs	1,393,411	500,221,257	359	--	FXAQPTL
06/07/09	AAFC/PBI	454 Titan	oleracea5	wgs	1,426,328	514,851,915	361	--	FYQTF2Y
12/08/09	AAFC/PBI	454 Titan	oleracea6	wgs	1,369,089	486,935,298	356	--	F0OWSTU
21/12/09	AAFC/PBI	454 Titan	BoleraceaLMPE	matepair	455,573	140,326,412	308	8K	F7QNI9P01
06/07/10	Boulos	454 Titan	APZ_AOTA	matepair	601,013	190,189,186	316	8K	GG3RUFO02
06/07/10	Boulos	454 Titan	APZ_AOTA	matepair	544,556	184,940,616	340	8K	GILAAW301
06/07/10	Boulos	454 Titan	APZ_AOTA	matepair	535,989	178,587,329	333	8K	GILAAW302
06/07/10	Boulos	454 Titan	APZ_EOTA	matepair	514,521	182,060,415	354	20K	GINB5JB01
06/07/10	Boulos	454 Titan	APZ_EOTA	matepair	520,245	200,549,290	385	20K	GIQ3JIC01
06/07/10	Boulos	454 Titan	APZ_EOTA	matepair	530,701	202,070,844	381	20K	GHKJ99Y01
01/06/10	JCVI	454 Titan	JCVI	matepair	554,538	181,306,728	327	20K	GHJ4PM201
01/06/10	JCVI	454 Titan	JCVI	matepair	649,109	258,728,068	399	20K	GHU82RK01
19/03/10	AAFC/PBI	454 Titan	oleracea20-40pe	matepair	417,345	108,640,128	260	20K	GD7420L01
19/03/10	AAFC/PBI	454 Titan	oleracea20-40pe	matepair	398,801	116,326,485	292	40K	GD7420L02
06/07/10	Boulos	454 Titan	APZ_FOTA	matepair	430,432	144,547,237	336	40K	GINB5JB02
06/07/10	Boulos	454 Titan	APZ_FOTA	matepair	493,330	157,192,314	319	40K	GHJ87TI01
06/07/10	Boulos	454 Titan	APZ_FOTA	matepair	436,073	161,447,685	370	40K	GIQ3JIC02
06/07/10	Boulos	454 Titan	APZ_GOTA	matepair	494,103	174,286,511	353	40K	GIOZC5F01
06/07/10	Boulos	454 Titan	APZ_GOTA	matepair	458,223	141,882,556	310	40K	GHJ87TI02
06/07/10	Boulos	454 Titan	APZ_GOTA	matepair	476,097	154,083,221	324	40K	GIOZC5F02
01/06/10	JCVI	454 Titan	JCVI	matepair	536,202	164,093,252	306	40K	GHJ4PM202
10/10/09	NCBI	Sanger	TO1434 BES	BAC-end	85,318	66,965,071	785	100K?	BOT01
10/10/09	JCVI	Sanger	BOG	paired	415,522	374,851,381	902	--	** pair-end libraries of various sizes
10/10/09	WUGSC	Sanger	WUGSC	paired	168,390	119,921,637	712	--	** pair-end libraries of various sizes
17/03/10	Guy Barker	Illumina	Illumina March2010	paired	45,803,322	3,481,052,472	76	500	Lane4
01/06/10	Guy Barker	Illumina	Illumina June2010	paired	45,803,322	5,038,365,420	110	600	Lane1 - possible problem with Nextera library so use as WGS
01/06/10	Guy Barker	Illumina	Illumina June2010	paired	41,976,840	4,617,452,400	110	600	Lane - 2possible problem with Nextera library so use as WGS
01/06/10	Guy Barker	Illumina	Illumina June2010	paired	46,146,534	5,076,118,740	110	455	Lane5
01/06/10	Guy Barker	Illumina	Illumina June2010	paired	46,353,346	5,098,868,060	110	455	Lane6
01/06/10	Guy Barker	Illumina	Illumina June2010	paired	41,593,810	4,575,319,100	110	652	Lane7
01/06/10	Guy Barker	Illumina	Illumina June2010	paired	40,589,874	4,464,886,140	110	652	Lane8



General overview of the Celera Assembler



- Originally designed for Sanger shotgun data
- Adapted for 454 @ JCVI (CABOG; Miller et al 2008) and recently for Illumina (CA 6.1 Release April 2010)
- TO1000 is one of the largest assemblies to exploit CA (Turkey genome – Dalloul et al. *PLoS Biology*, 2010; 8(9): e1000475 DOI)

Assembly using CA

- Hardware co-located to NRC-PBI
 - From AAFC
 - 1 x Dell R900
 - 16 CPU cores
 - 256 Gb RAM
 - Solid State Drive 640 Gb
 - From PBI
 - Initially 30 Penguin Computing 1U servers
 - These were dropped out of the pool due to performance issues
 - 5 x Dell R900
 - 24 CPU cores
 - 128 Gb RAM each
 - Isilon Disk array for space
 - Network interconnect (1GbE)

Contigs

All contigs \geq 400bps

sequences 80,789

Sum (bps) 351,751,498

Mean 4,354

Median 2,923

Mode 1,461

Minimum 400

Maximum 403,037

N50 6,394

Scaffolds

All scaffolds \geq 400bps

sequences 17,655

Sum (bps) 488,253,615

Mean 27,655

Median 2,303

Mode 1,178

Minimum 408

Maximum 9,908,692

N50 3,823,124

Degenerates

All degenerates ≥ 400 bps

sequences 175,356

Sum of data 111,833,642

Mean 638

Median 528

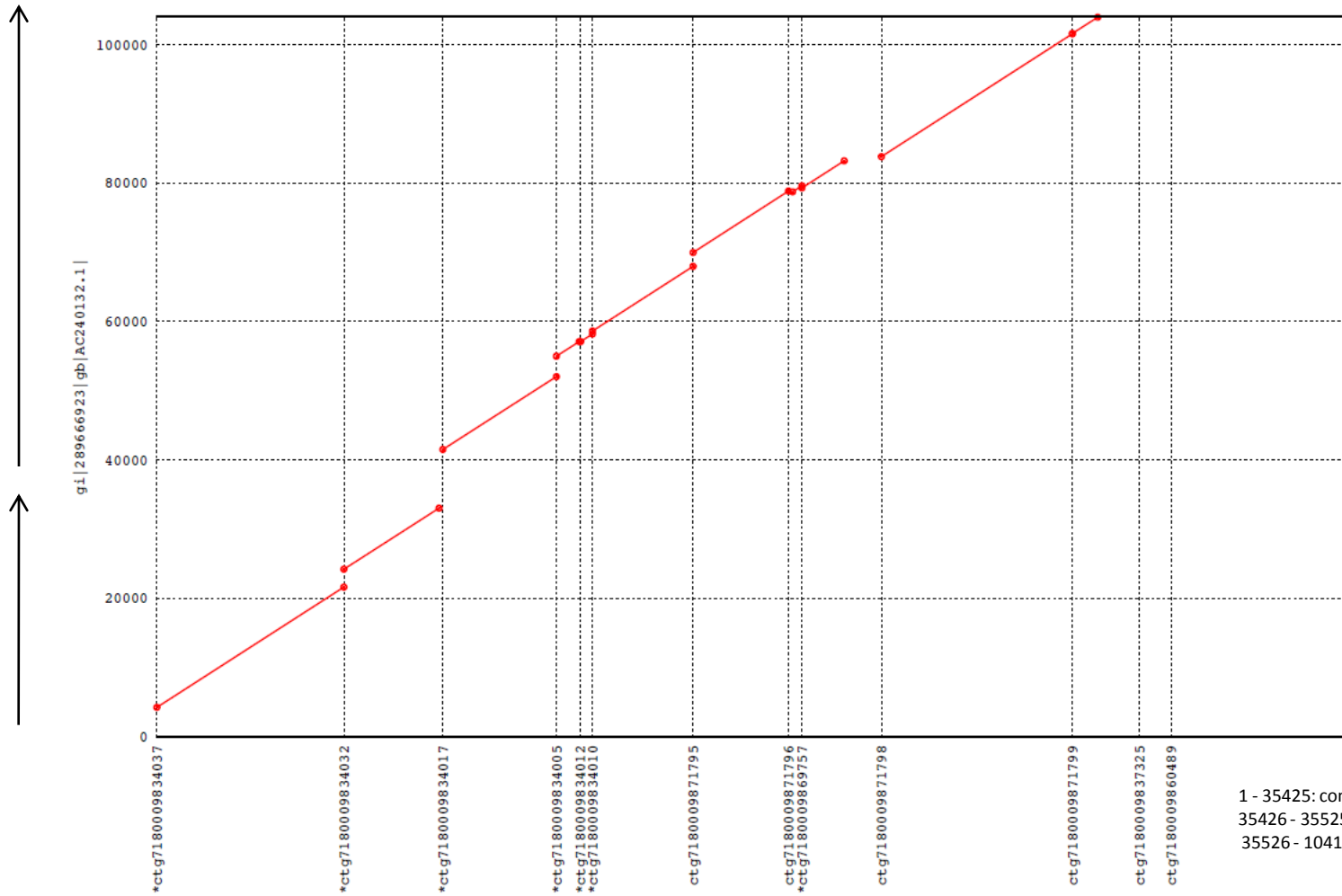
Mode 411

Minimum 400

Maximum 24,602

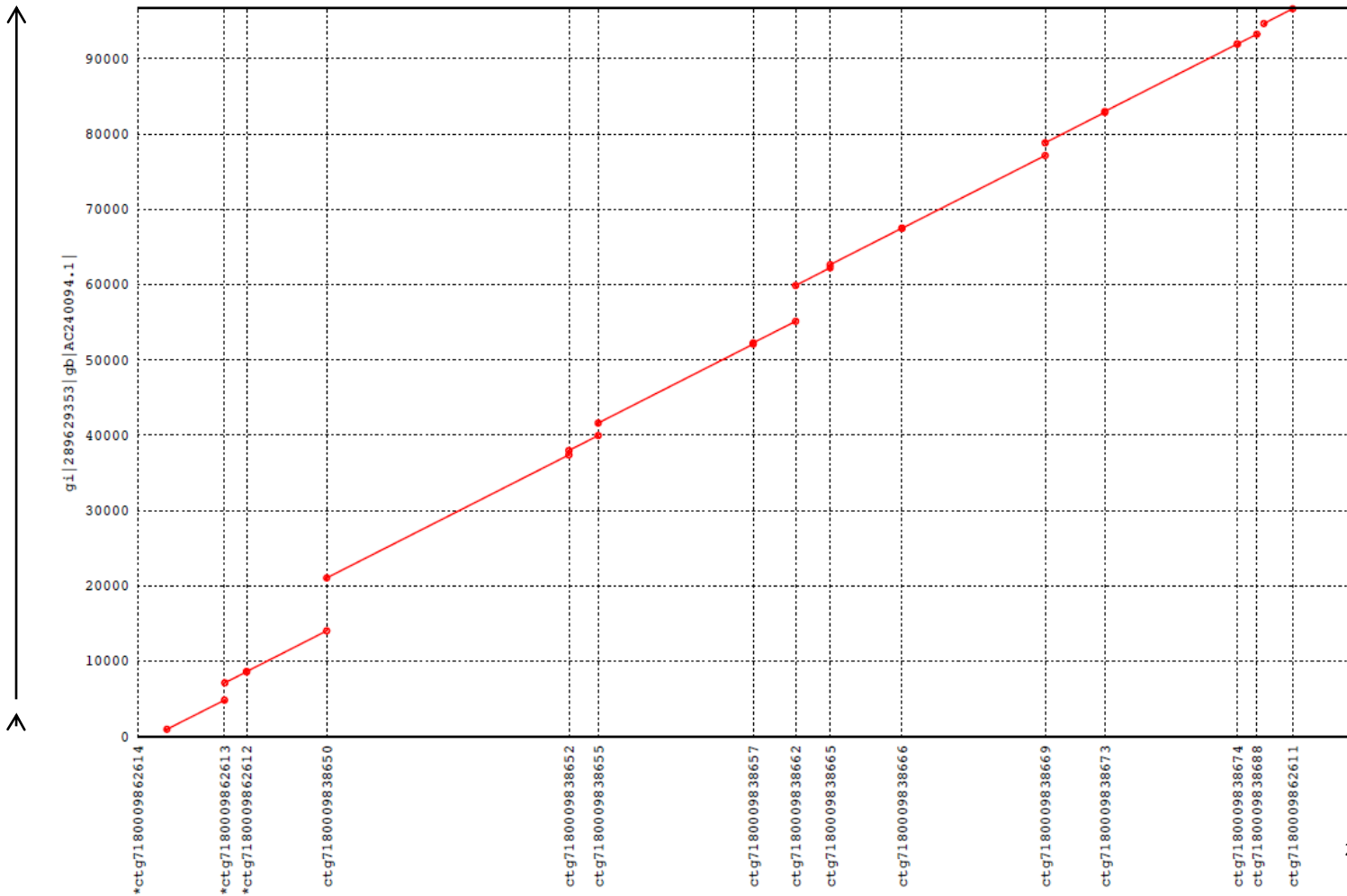
N50 612

CA contigs vs. a *B. oleracea* BAC – 2 pieces



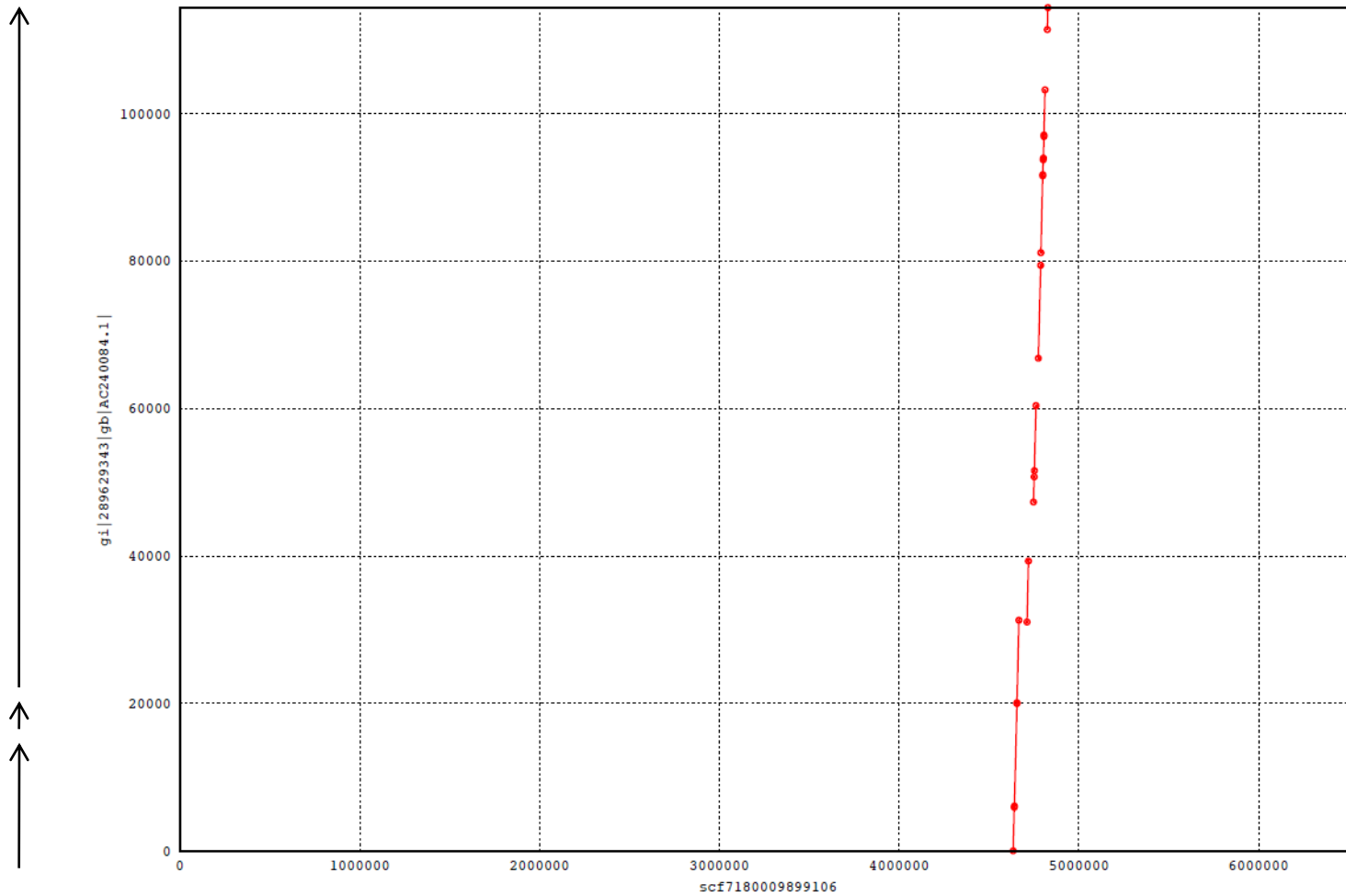
1 - 35425: contig of 35425 bp in length
 35426 - 35525: gap of unknown length
 35526 - 104145: contig of 68620 bp in length.

CA contigs vs. a *B. oleracea* BAC – 2 pieces



1 - 2894: contig of 2894 bp in length
 2895 - 2994: gap of unknown length *
 2995 - 96771: contig of 68620 bp in length.

CA scaffold vs *B. oleracea* BAC



1 - 17342: contig of 17342 bp in length
17343 - 17442: gap of unknown length
17443 - 22662: contig of 5220 bp in length
22663 - 22762: gap of unknown length
22763 - 114384: contig of 91622 bp in length.

(6.5 Mb scaffold)

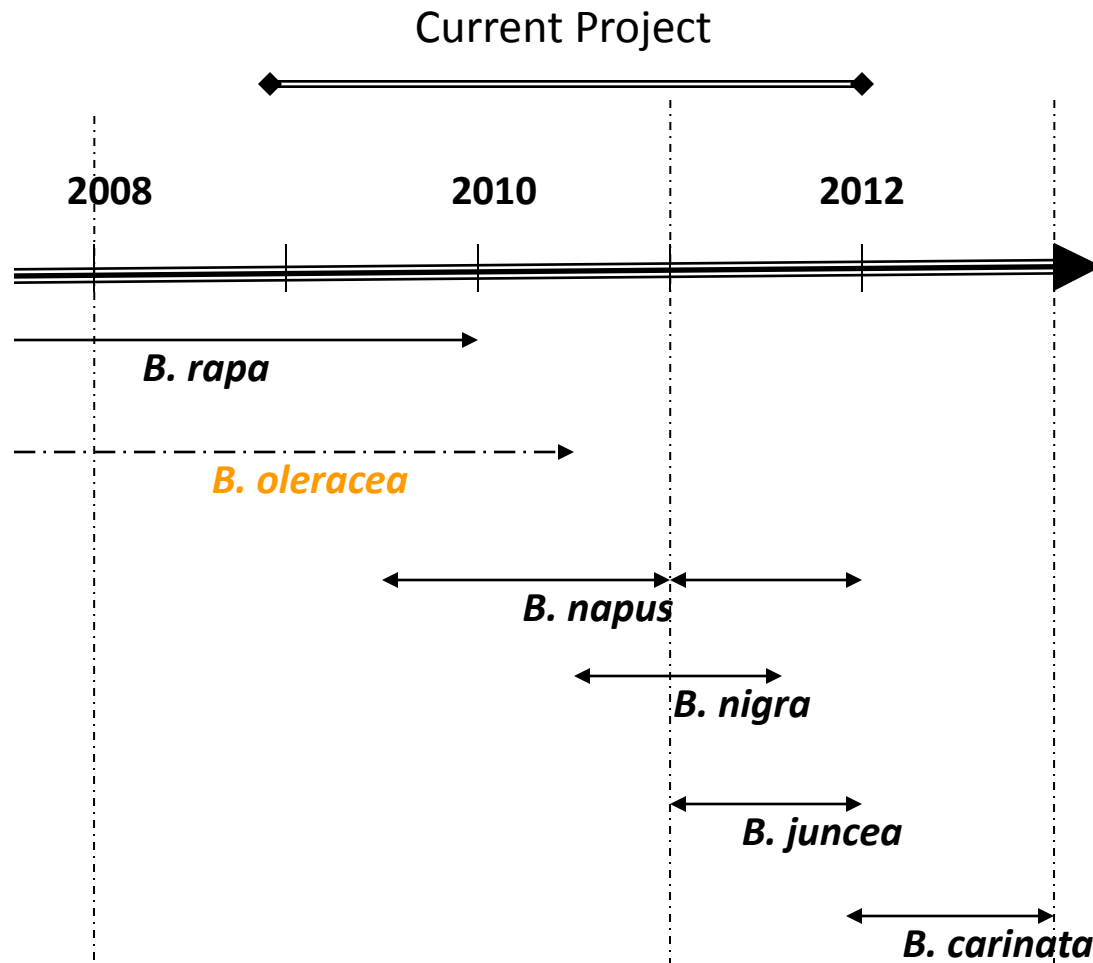
Gene space coverage

- Using all coding sequences from TAIR what is the coverage in *B. oleracea* TO1000?
 - BLAST 1e-10
- 89% of TAIR CDS sequences have at least 1 match to the assembled data
- This is the same as the *B. rapa* Chiifu BGI assembly (89%) – approx. 280Mbp assembly

Next Steps

- Further fine tuning of CA assembly
 - Insert size ranges, run-time
 - Genetic anchoring – high density map via “genotyping by sequencing”
- Annotation & Analysis
 - Evidence Modeler (EVM) Brian J. Haas, et al. 2009
 - Genome architecture
- Comparison to other *de novo* C genome assembly and *de novo* A genome assembly
- *B. oleracea* re-sequencing – Early Big cabbage data from CSHL (Martiennson / McCombie; 50Gb) and other genotypes

Vision for elucidating the genomes of all Brassica U triangle species

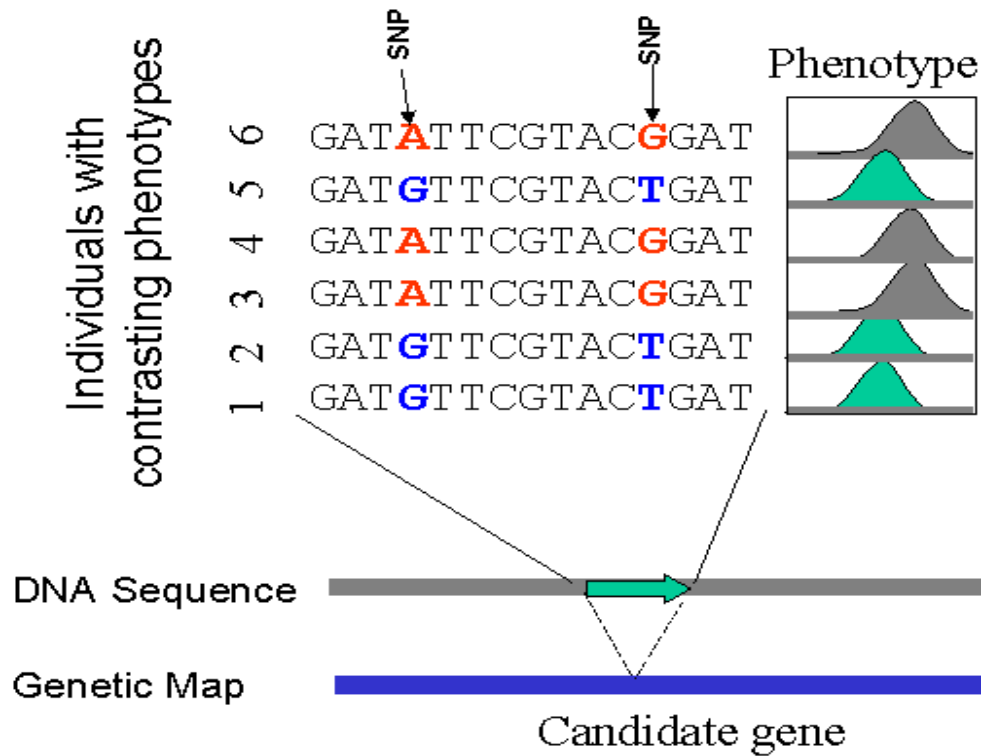


Single Nucleotide Polymorphisms (SNPs)

- Advantage of SNP markers
 - assay a single locus (bi-allelic)
 - amenable to mutiplexing and automation
 - marker-assisted selection
 - discovery 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

SNP Haplotypes (Haploid Genotype)

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



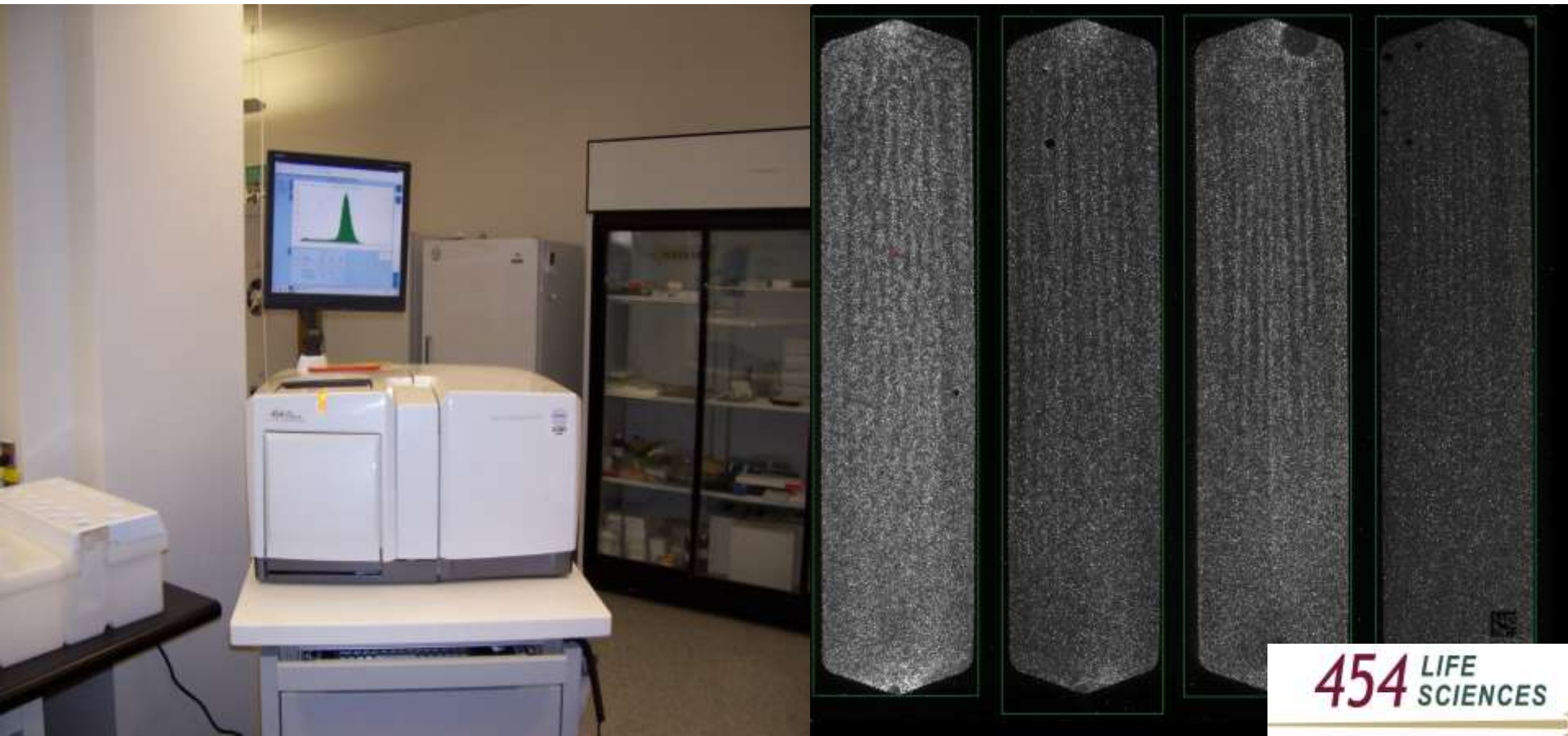
Rafalski 2002

U. of Saskatchewan CDC / NRC-PBI

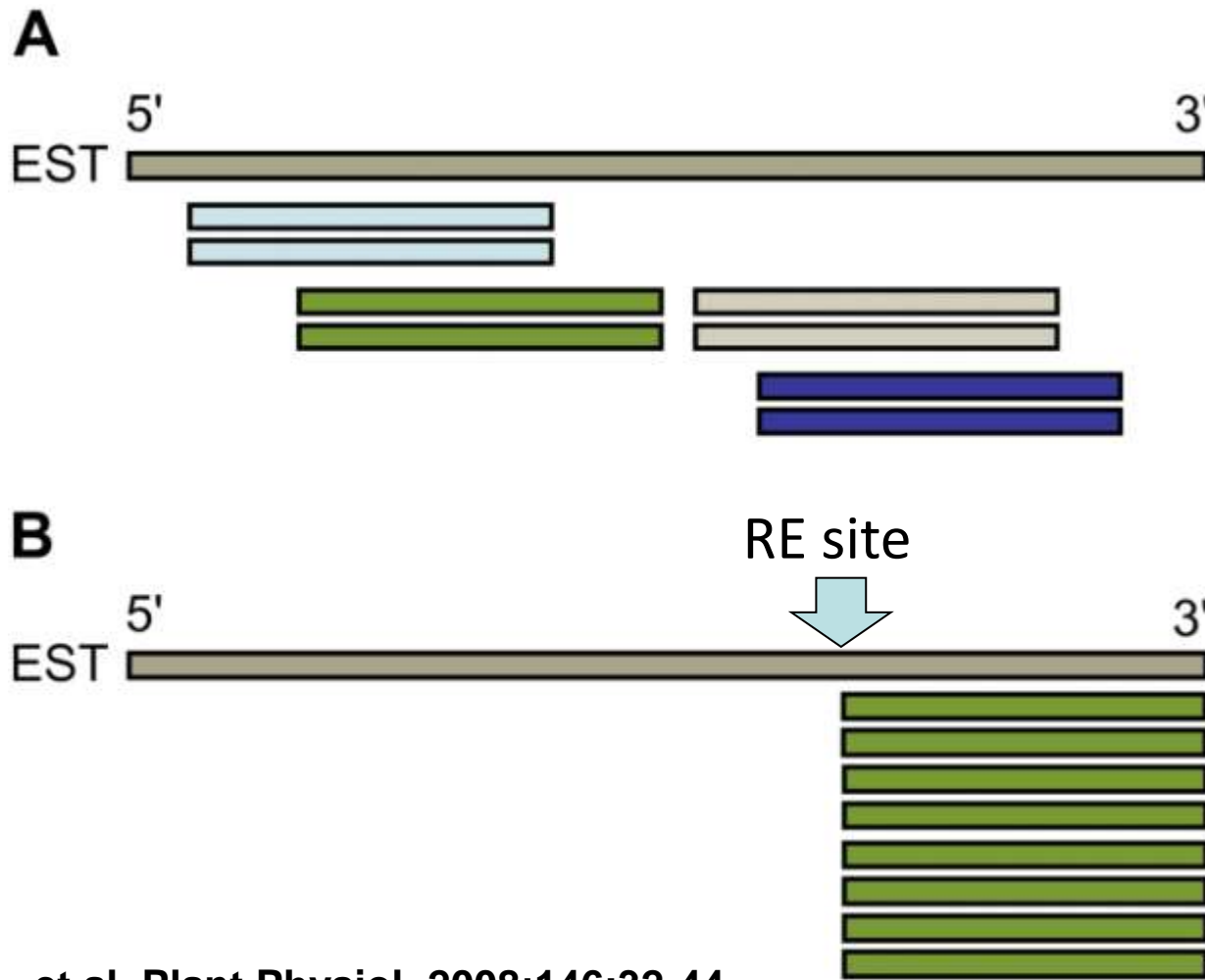
- **EST and SNP discovery (2008-2010)**
 - high volume sequencing
 - seed related **ESTs** in lentil, chickpea, field pea and bean
 - single genotype and range of vegetative and seed related tissues
 - **SNP discovery** using 454 cDNA profiling in all four species – 8 genotypes in each, 5 tissues each
 - ~400-500K reads in each genotype in each species
 - *in-silico* SNP discovery from 4M reads per species
- **Pulse SNP genotyping platforms**
 - development and utilization of Illumina arrays
 - single-plex assays (KASPar)

454 Titanium SNP discovery

- Titanium upgrade 2009 at NRC-PBI
- 450bp reads, ~500Mbp total / run
- 3' anchored transcript profiling – SNP discovery



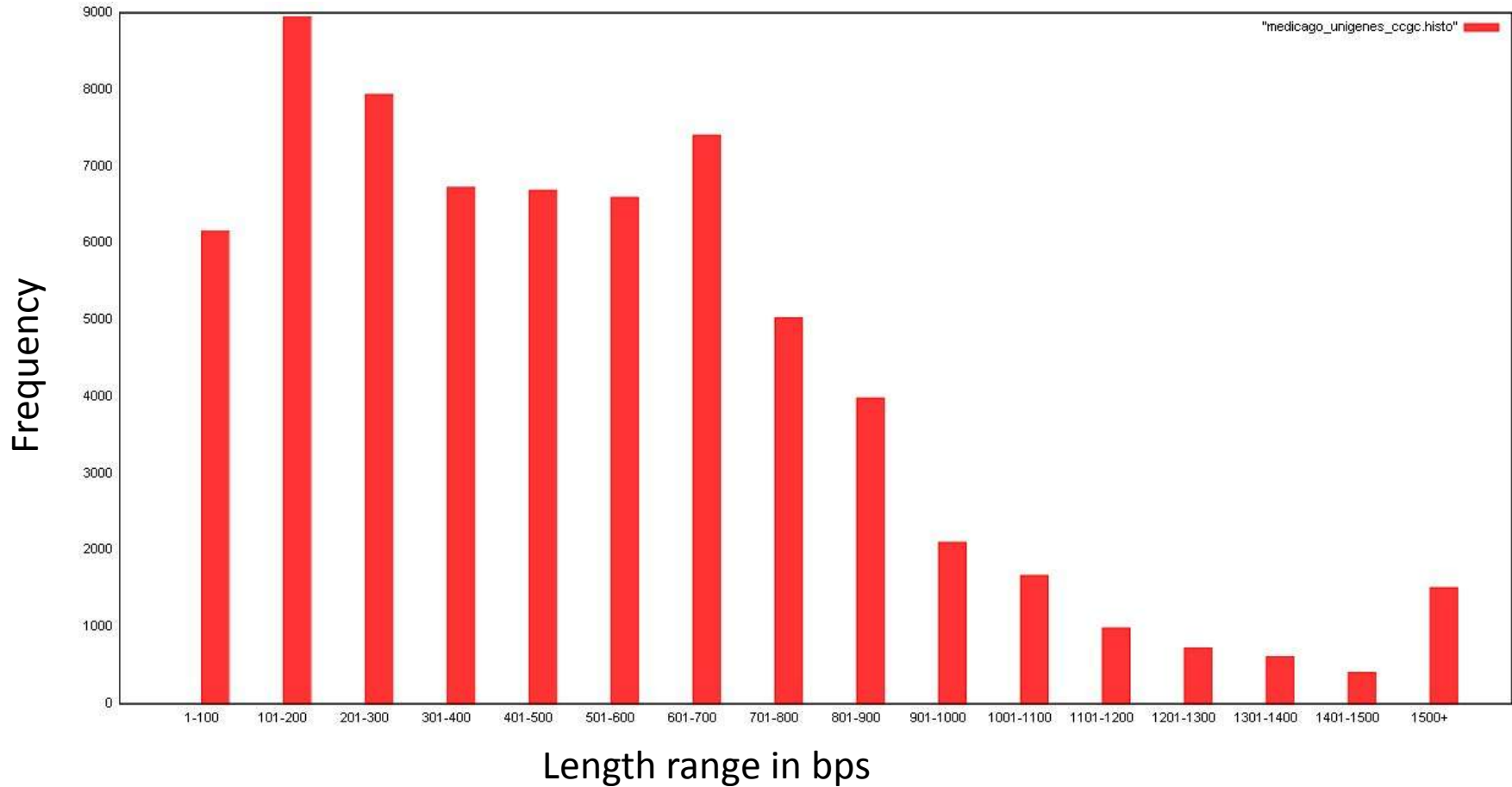
Shotgun or 3'-anchored 454 transcript profiling



Eveland, A. L., et al. *Plant Physiol.* 2008;146:32-44

Medicago in-silico digestion

– *Aci* I provides Titanium “ready” fragments



- *Aci* I (CCGC) – GC rich; longer fragments (non-palindromic)

Trial 3' 454 Sequencing in Pea

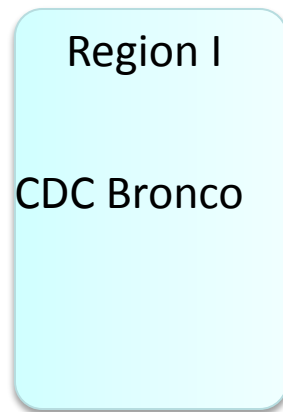
sstDNA Library Quality Assessment and Quantitation
qPCR (SYBR Green)



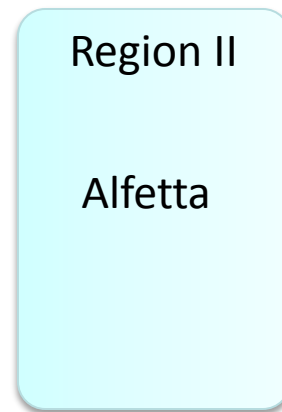
Dilute 2 sstDNA Libraries (1×10^6 molecules/ μ l)



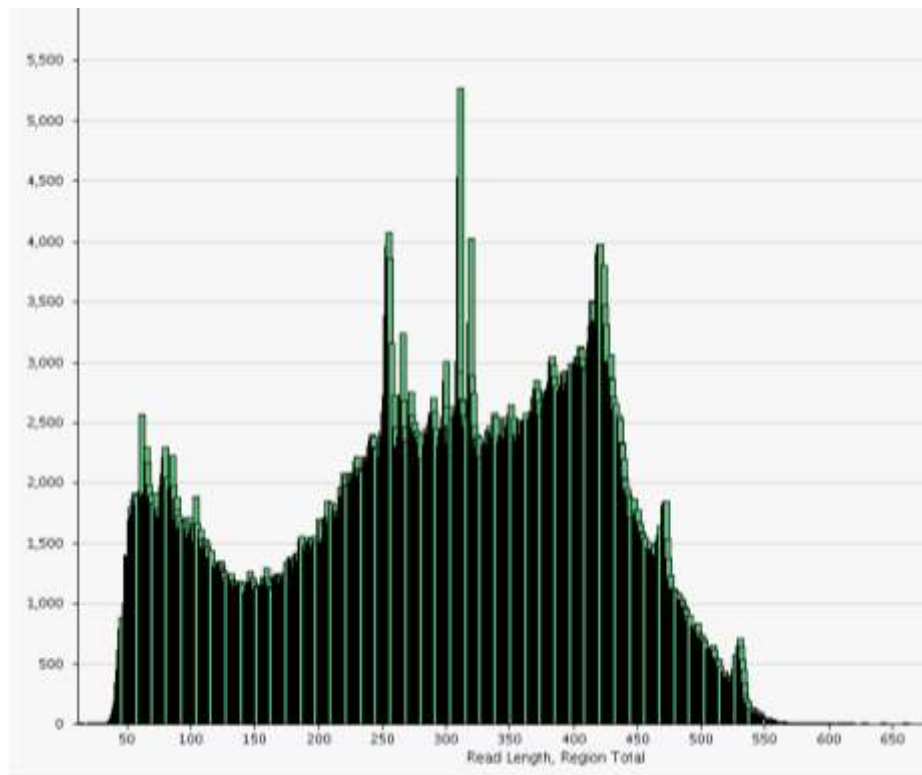
Sequence each sstDNA library / region on a full 454 run



~450K reads



~450K reads



TCAG (Library)	Region		Total
	1	2	
Raw Wells	925,128	819,690	1,744,818
Key Pass Wells	901,755	800,841	1,702,596
Passed Filter Wells	472,899	483,140	956,039
Total Bases	139,787,293	138,732,243	278,519,536
Length Average	295.60	287.15	291.33
Length Std Deviation	129.06	119.99	
Longest Reads Length	617	761	761
Shortest Reads Length	21	12	12
Median Reads Length	314.0	295.0	304.0

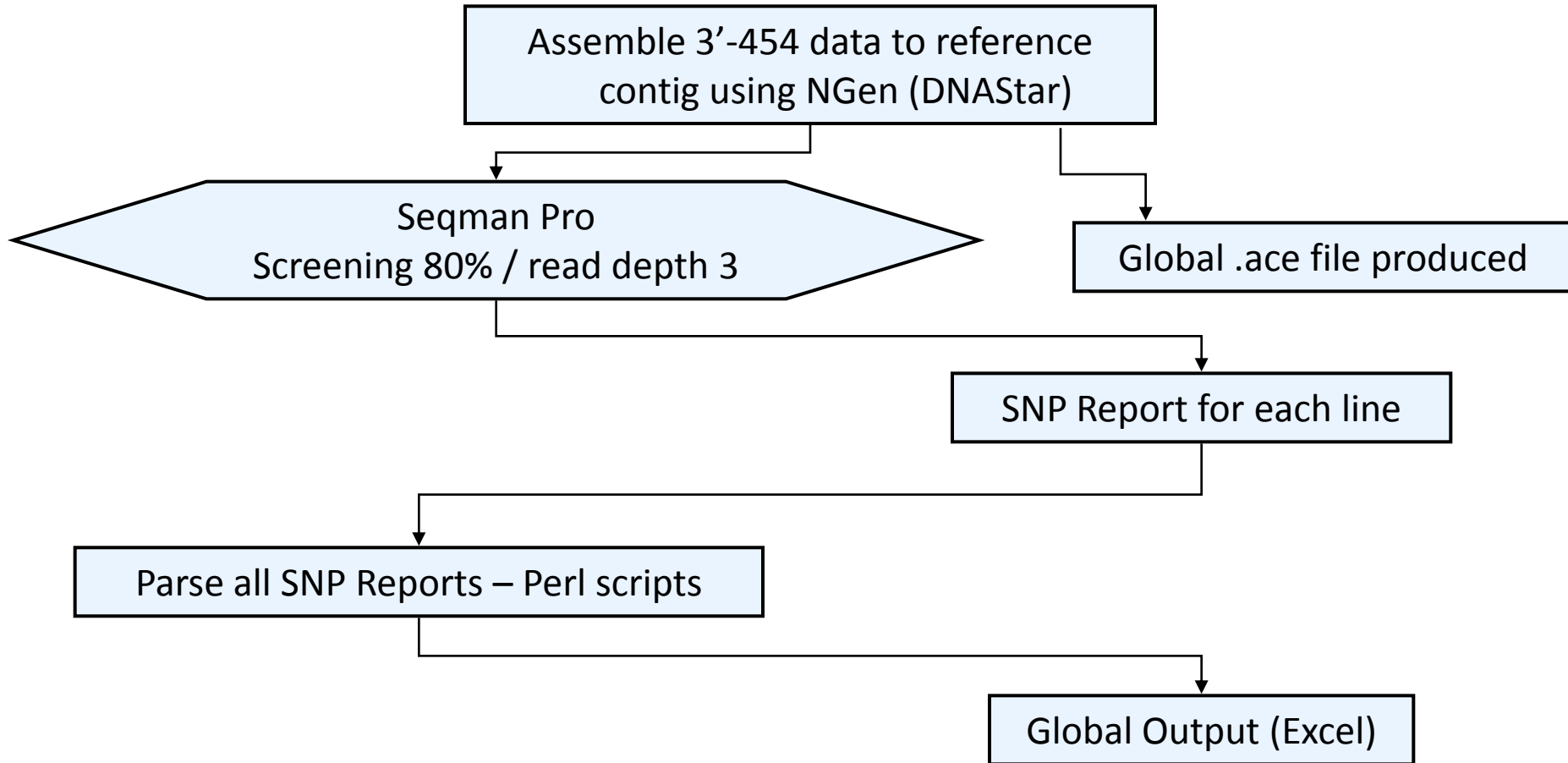
Lentil Genotypes for SNP Discovery

- 11 different lentil genotypes
- Five tissues – leaf, stem, etiolated seedling, flower, seed

Plant species	Genotype (line)	454 Library Name	Passed Filter Reads	454 Sequencing Run Files	SFF Files
Lentil (<i>Lens culinaris</i>)	CDC Redberry	L1	535,835; 498,396	Rong30jan09, Region 1; 19mar09, Region 1	FP3I4WR01, FSTOUBP01
Lentil (<i>Lens culinaris</i>)	Robin	L2	394,707	Rong30Mar09, Region 1 and Lentil2_6run2, Region 1	FTEALA301, FT9S7HC01
Lentil (<i>Lens culinaris</i>)	964a-46	L3	370,621	L3_L5_28may09, Region 1	FWK606101
Lentil (<i>Lens culinaris</i>)	Eston-A	L4	317,613	L4_L9_may1, Region 1	FU6SZER01
Lentil (<i>Lens culinaris</i>)	PI 320937	L5	343,486	L3_L5_28may09, Region 2	FWK606102

- **CDC Redberry** – reference assembly >1M reads – NGen
- Data assembled per genotype against reference – NGen
- Common contigs identified between genotypes

SNP Analysis Pipeline



Wayne Clarke – Bioinfo session

Lentil SNP Discovery

Line	Contigs with SNPs	Total SNPs	Average Read Depth
Robin	2177	4727	12
964A	1967	4138	10
Eston-A	1240	2690	17
PI_320937	2003	4276	11
LC8602303T	2762	6371	13
ILL5588	1684	3740	8
CDC_Milestone	1569	3390	31
ILL-8006	4304	9797	15
L01-827A	3264	10793	10
PI72815	5847	19946	10

- SNPs identified in >80% of reads and depth of at least 3 reads
- Transition / transversions only – no indels (homopolymers)
- *L. ervoides* - more diversity
- Non-redundant set 11,379 contigs out of 27,921 with SNPs (41%)

SNP Haplotype Structure – Contig00020 all genotype assemblies

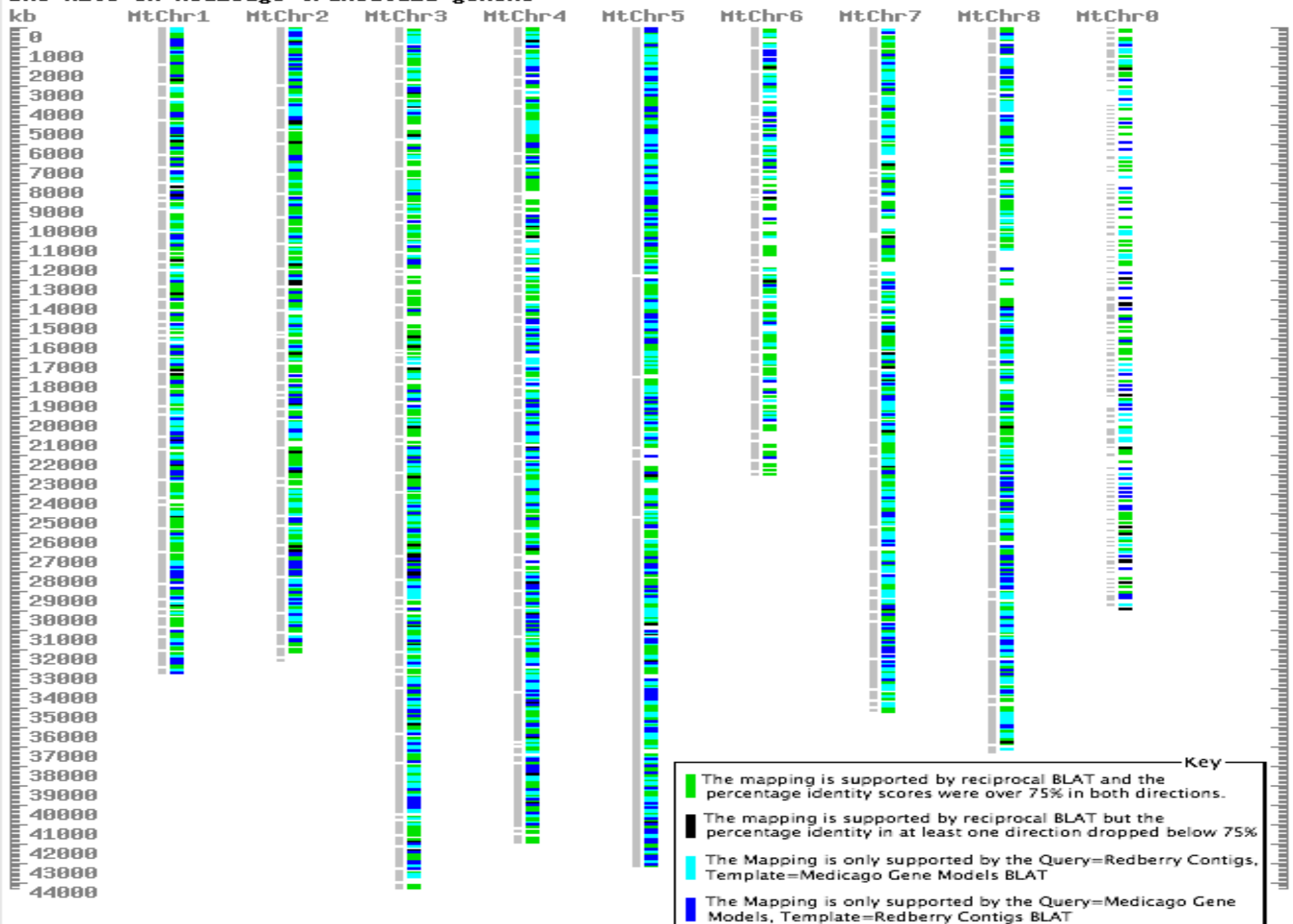
RB Contig	SNP Position	Mt Genes	Gene Location	Type of Mapping	BLAT Scores	Left Flanking Sequence	Right Flanking Sequence	Redberry (Ref)	CDC_Mil 964A	Estone -A	ILL-8006	ILL55 88	L01-827A	LC860 2303T	PI728 15	PI_32 0937 Robin
Contig 00020	92	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	92	633	T	0	0	0	0	C	0	100	0
Contig 00020	266	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	266	459	C	0	0	0	0	T	0	0	0
Contig 00020	495	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	495	230	A	G	G	0	0	G	G	0	11.1
Contig 00020	508	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	508	217	G	0	0	0	0	A	0	0	0

Percentage of reads with variant nucleotide
(100 = <3 reads)

Summary of Lentil SNP discovery

- Total Contigs: 27,921
- Contigs with SNPs: 11,379 (41%)
- Total SNPs (80% cutoff + depth of 3): 44,942
- Transitions: 27,227 (61%)
- Transversions: 17,715 (39%)
- Frequency: 0.21 for all lines
- Contigs mapped to Medicago (BLAT)
- SNP genotyping
 - Illumina 1,536 GG arrays for dense maps
 - KASPar – single-plex, validation & MAS

BAC hits on *Medicago truncatula* genome



KASPar Validation Assays

Contig	SNP position	SNP detected by Allele											SNP detected by Allele specific 1 (blue)	SNP detected by Allele specific 2 (red)	
		Redberry	964a-46	CDC_Milestone	Eston-A	ILL-8006	ILL5588	L01-827A	LC860230-3T	PI72815	PI_3209-37	Robin			
00012	343	T	C	C	C	T	C	C	T	T	T	T	25	C	T
00020	495	A	G	G	A	A	G	A	G	A	G	G	11.1	A	G
00030	350	A	G	A	A	A	G	A	A	A	A	A		A	G
00056	456	A	C	C	C	C	A	A	C	A	C	C		A	C
00067	332	C	G	G	C	C	C	C	C	C	C	C		C	G
00093	289	G	A	G	A	A	A	G	A	G	A	G		A	G

RB Contig	SNP Position	Mt Genes	Gene Location	Type of Mapping	BLAT Scores	Left Flanking Sequence	Right Flanking Sequence	Redberry (Ref)	CDC_Mil 964A	Eston estone -A	ILL-8006	ILL5588	L01-827A	LC860230-3T	PI72815	PI_3209-37	Robin
Contig 00020	92	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	92	633	T	0	0	0	0	C	0	100	0	0
Contig 00020	266	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	266	459	C	0	0	0	0	T	0	0	0	0
Contig 00020	495	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	495	230	A	G	G	0	0	G	0	G	0	G
Contig 00020	508	IMGA Medtr8-g037310.1	8-8051860-8052744	1	85/86	508	217	G	0	0	0	0	A	0	0	0	0

- 150 KASPar markers now tested in mapping population approx. 85% success rate

Future Prospects

- **New Sequencing Platforms**

- Ion Torrent – now
- PacBio later in 2011 – long reads (>1kb) and “strobe” reads
- Nanopore sequencers?

- **New Methods**

- faster / automation
- higher throughput = higher multiplexing / pooling

- **New Bioinfo Tools**

- Developing as fast as platforms – analysis and viewing
- AllpathsLG – Broad Institute January 2011
 - assembly of large genomes with short reads



Acknowledgements

CanSeq Project:

AAFC-SRC

Isobel Parkin
Matthew Links
Rob Wood
Brittany Polley

NRC-PBI

Andrew Sharpe
Faouzi Bekkaoui
Kevin Koh
Jacek Nowak
Carrie Haimanot
Carling Tallon

TO1000 Team:

U of Missouri-Colombia

Chris Pires

JCVI

Chris Town
Jason Miller

Warwick HRI

Guy Barker

INRA

Boulos Chalhoub

Pulse SNP Discovery:

NRC-PBI:

Raju Datla / Raj Selvaraj
Kishore Gali
Rong Li
Janet Condie
Christine Sidebottom
Shuqing Qiu

AAFC-Saskatoon

Isobel Parkin
Wayne Clarke
Matt Links
Brent Mooney

U. of Sask. CDC:

Bert Vandenberg
Kirstin Bett
Buyamin Taran
Tom Warkentin
Lacey Sanderson
Emily Barlow
Perumal Vijayan

Funding:

Saskatchewan ADF
AAFC ABIP

CanSeq Funding:

Industry Partners, AAFC MII, NRC-PBI, Genome
Alberta / Prairie



National Research
Council Canada

Conseil national
de recherches Canada

Canada