CTAG

Canadian Triticum Advancement through Genomics

Reference sequencing of bread wheat chromosome 1A

IWGSC Standards and Protocols Workshop

PAG XXIII January 12th 2015

GenomeCanada

GenomePrairie

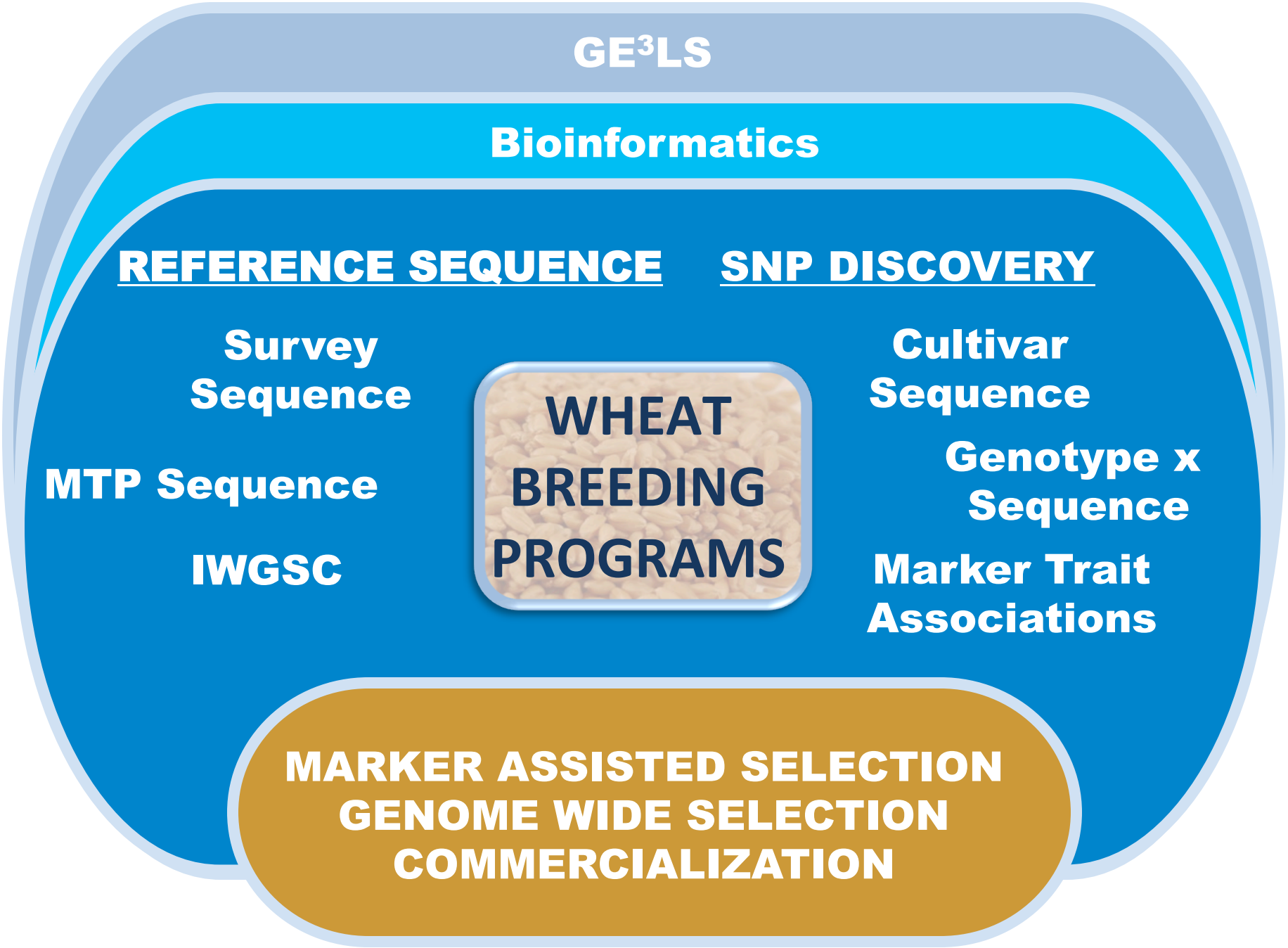
University of Saskatchewan

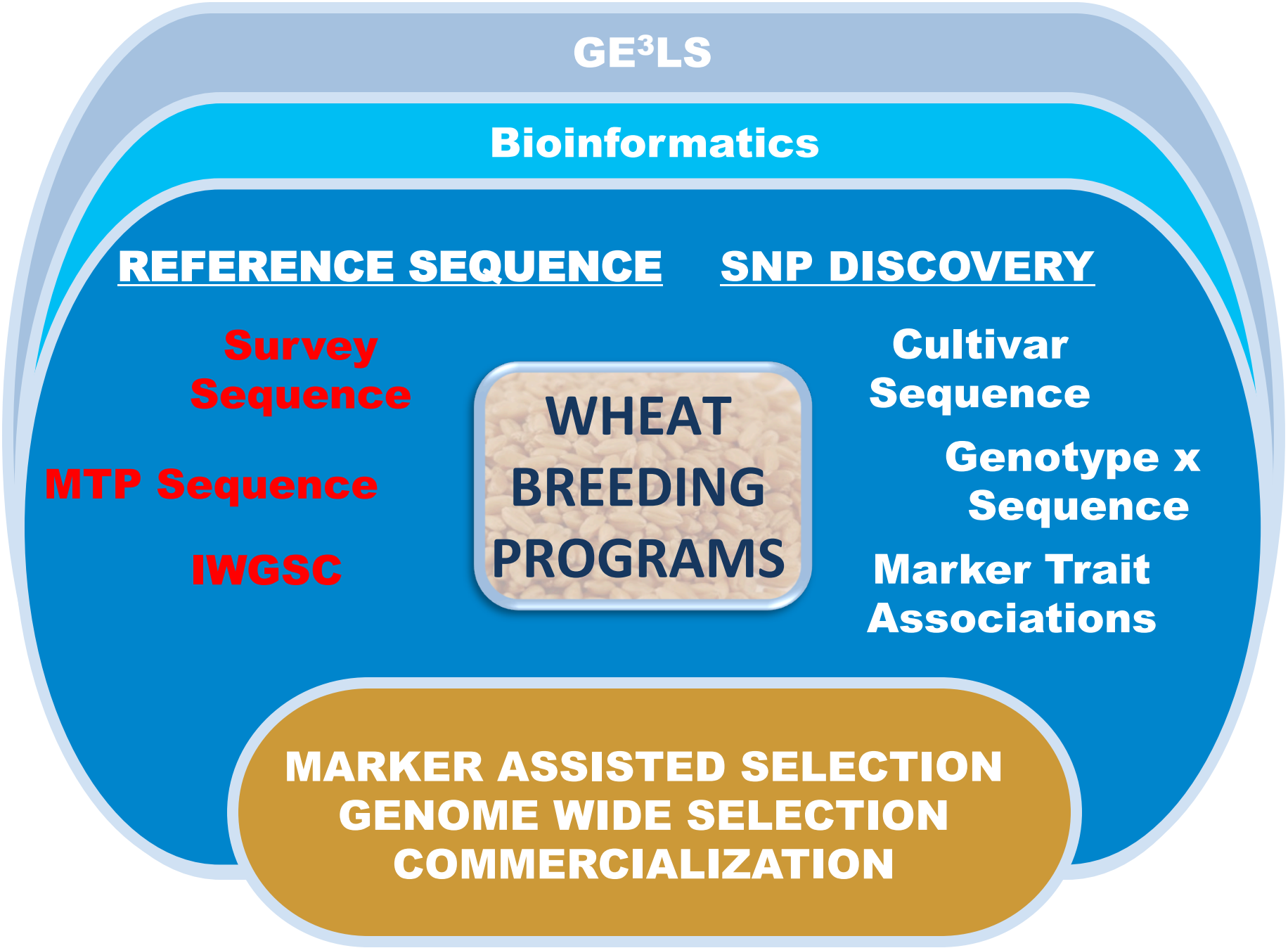
Agriculture and Agri-Food Canada

Agriculture et Agroalimentaire Canada

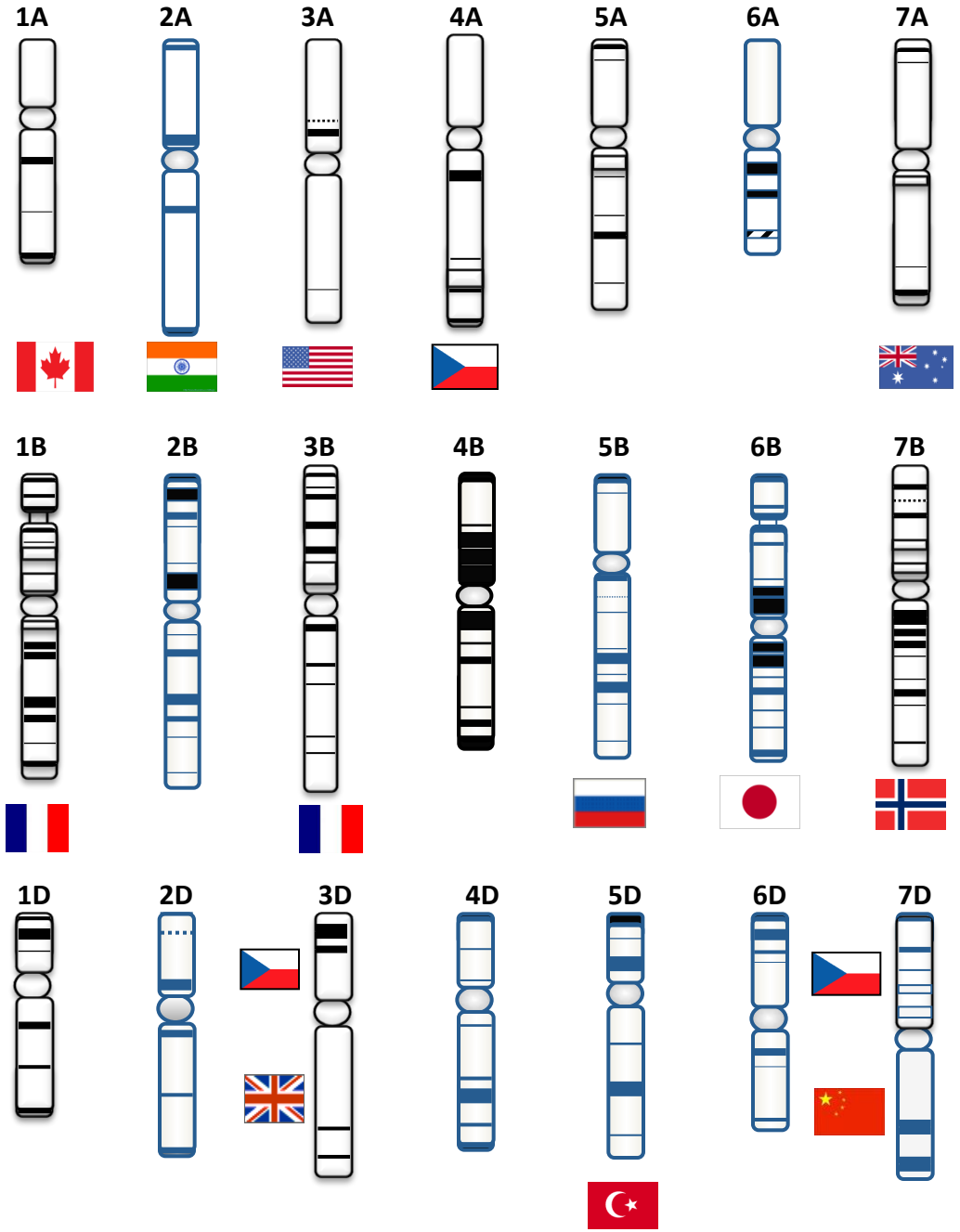
NRC·CNRC Canada

University of Regina


GE³LS
Bioinformatics
REFERENCE SEQUENCE
SNP DISCOVERY
Survey Sequence
MTP Sequence
IWGSC
WHEAT BREEDING PROGRAMS
Cultivar Sequence
Genotype x Sequence
Marker Trait Associations
MARKER ASSISTED SELECTION
GENOME WIDE SELECTION
COMMERCIALIZATION


GE³LS
Bioinformatics
REFERENCE SEQUENCE
SNP DISCOVERY
Survey Sequence
MTP Sequence
IWGSC
WHEAT BREEDING PROGRAMS
Cultivar Sequence
Genotype x Sequence
Marker Trait Associations
MARKER ASSISTED SELECTION
GENOME WIDE SELECTION
COMMERCIALIZATION

Reference Sequencing of Bread Wheat



T. aestivum

cv Chinese Spring

Chromosome 1A

CTAG
Canadian Triticum Advancement through Genomics

University of Saskatchewan

NRC·CNRC

Agriculture and Agri-Food Canada

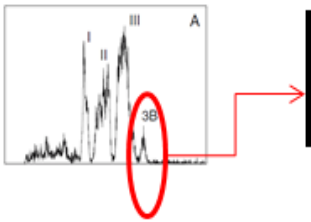
University of Zurich[UZH]

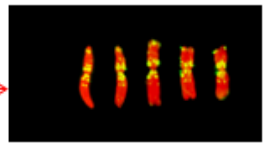
Sabancı Üniversitesi

CTAG
Canadian Triticum Advancement through Genomics

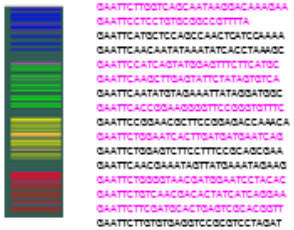
Chromosome 1A

IWGSC STRATEGY FOR OBTAINING A REFERENCE WHEAT GENOME SEQUENCE





1. BAC library construction

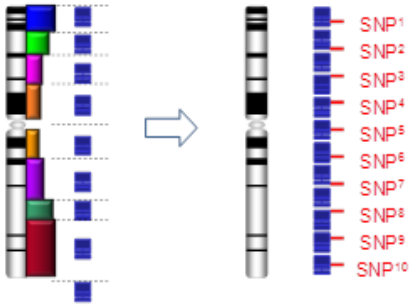


2. BAC fingerprinting (HICF/WGP)

3. Contig assembly by FPC/LTC

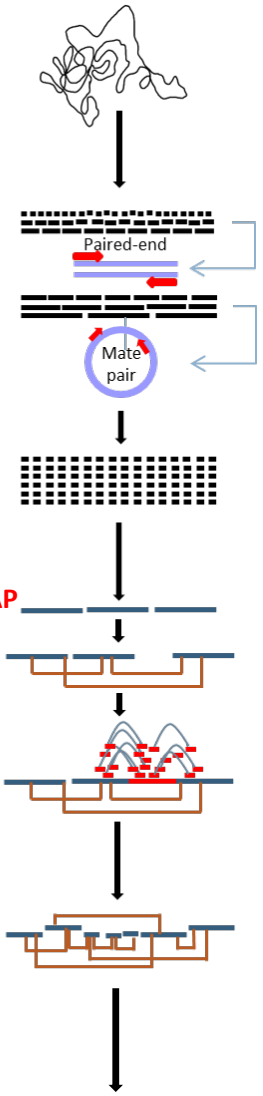
4. MTP sequencing /Scaffold assembly

5. Pseudomolecule construction
(meiotic/LD/RH mapping)



6. Automated and curated annotation

CTAG Strategy for assembly of 1AS BAC MTP

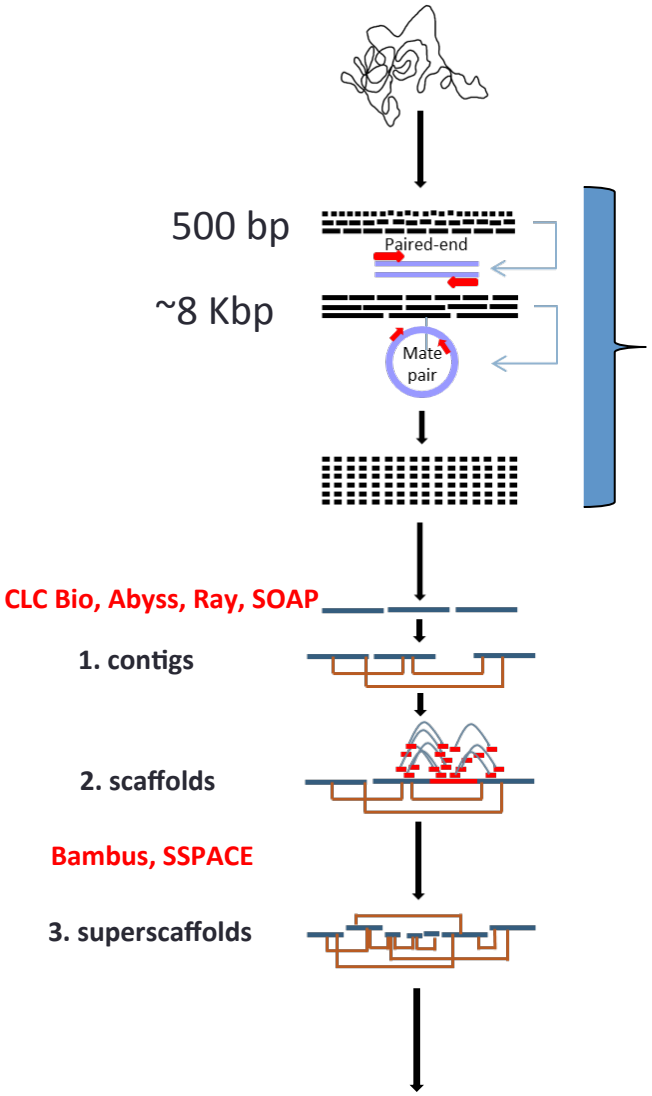


- BAC scaffolds (batches of 96)
- 1AS MTP Pseudomolecule
- Annotation – INRA / Triannot

CTAG
Canadian Triticum Advancement through Genomics

CTAG Strategy for assembly of 1AS BAC MTP

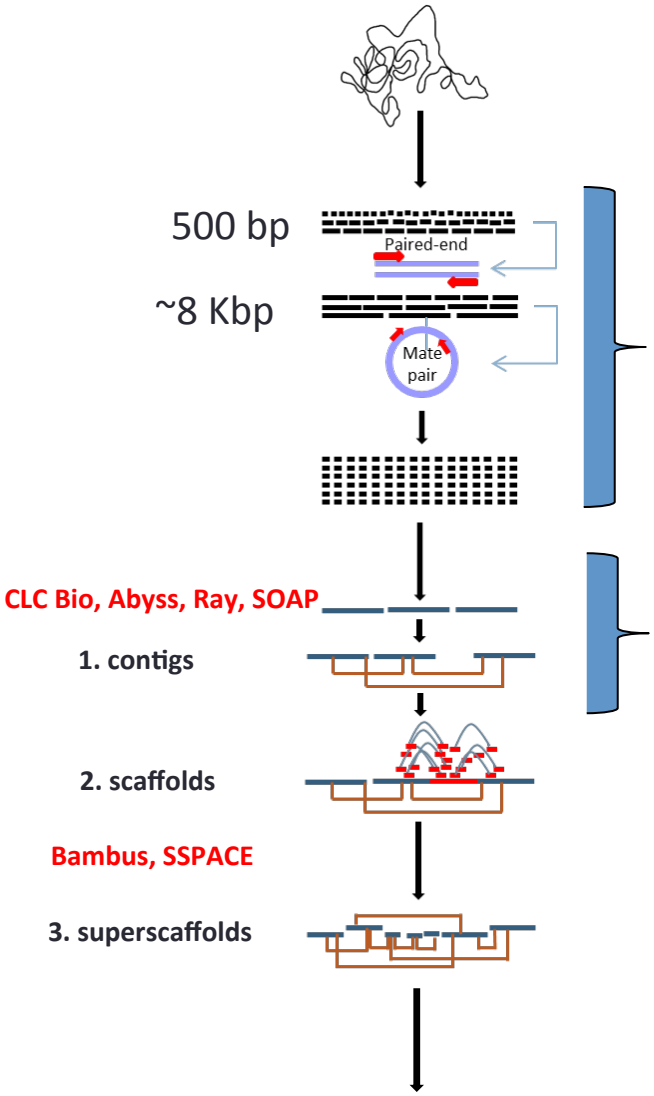
- 96 BAC preps (**4,134 total for 1AS**)
- 96 plex TruSeq PCR free library kits
- MiSeq 2 x 250 bp pair-end (PE) reads
- Mate-pair (MP) of BAC pools (96-384)



- **BAC scaffolds (batches of 96)**
- **1AS MTP Pseudomolecule**
- **Annotation – INRA / Triannot**

CTAG
Canadian Triticum Advancement through Genomics

CTAG Strategy for assembly of 1AS BAC MTP



- 96 BAC preps (**4,134 total for 1AS**)
- 96 plex TruSeq PCR free library kits
- MiSeq 2 x 250 bp pair-end (PE) reads
- Mate-pair (MP) of BAC pools (96-384)

- Trial MiSeq PE data (96 BACs)
- All BACs indexed / bar-coded for partitioning
- Trial Nextera MP library – 96 pooled BACs

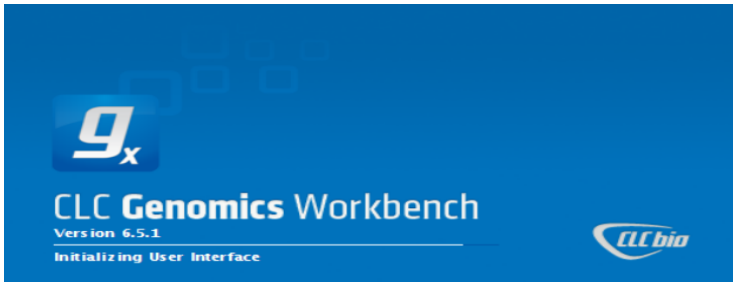
- **BAC scaffolds (batches of 96)**
- **1AS MTP Pseudomolecule**
- **Annotation – INRA / Triannot**

CTAG
Canadian Triticum Advancement through Genomics

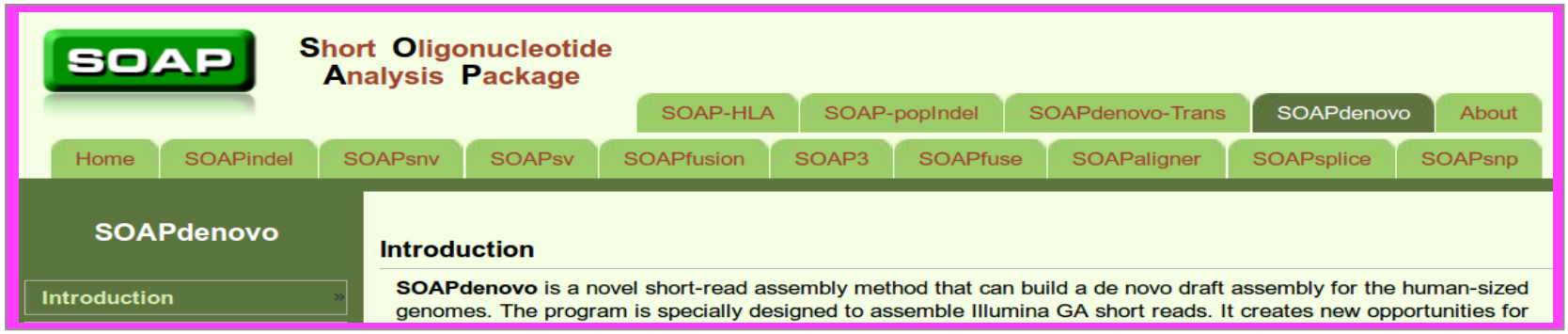
Selection of assembler

1. http://www.clcbio.com
2. http://soap.genomics.org.cn
3. http://www.bcgsc.ca/~/software/abyss
4. http://denovoassembler.sourceforge.net

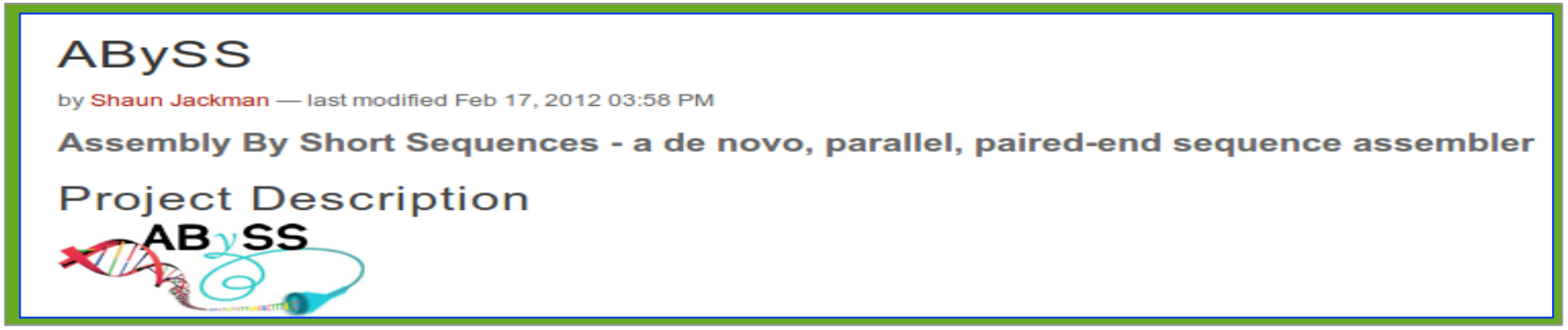
1.



2.



3.



4.

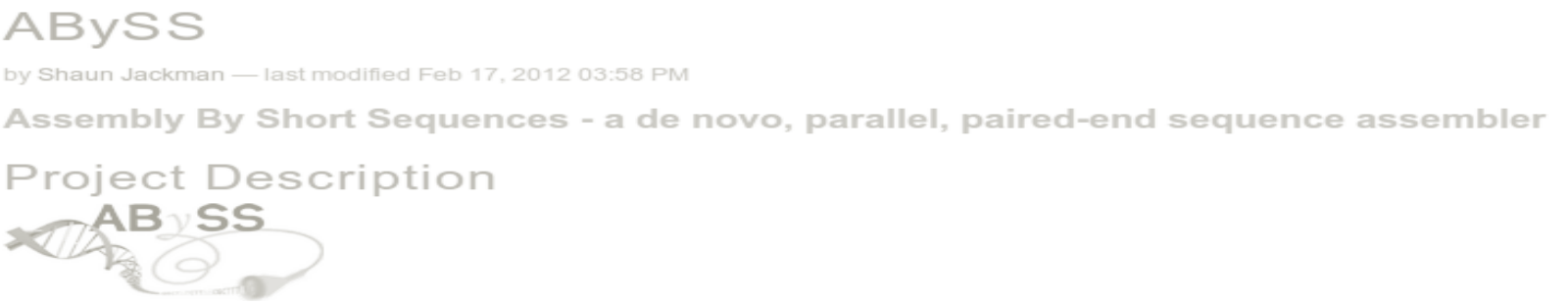
Selection of assembler

1. http://www.clcbio.com
2. http://soap.genomics.org.cn
3. http://www.bcgsc.ca/~/software/abyss
4. http://denovoassembler.sourceforge.net

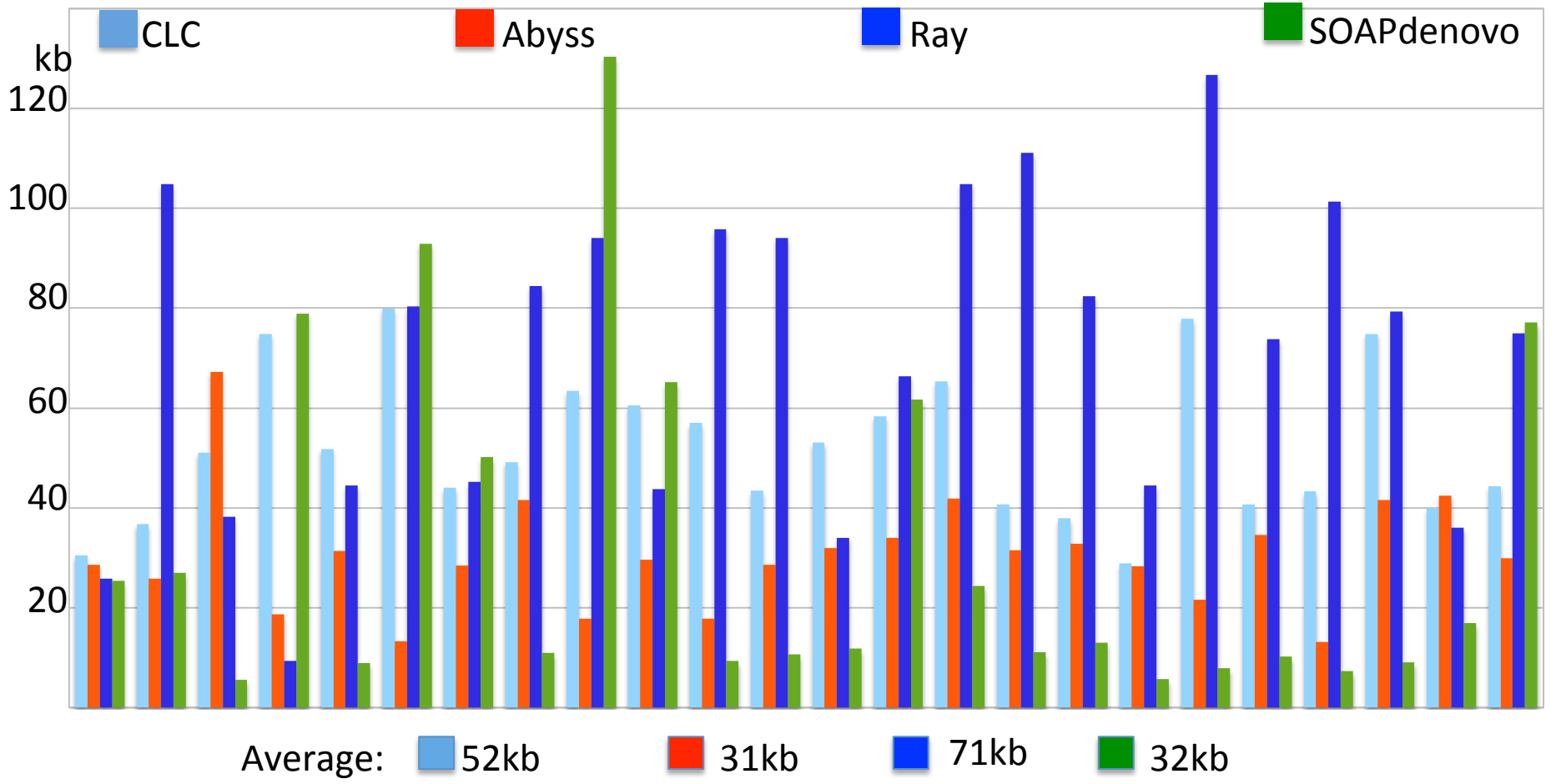
1.

2.

3.



4.

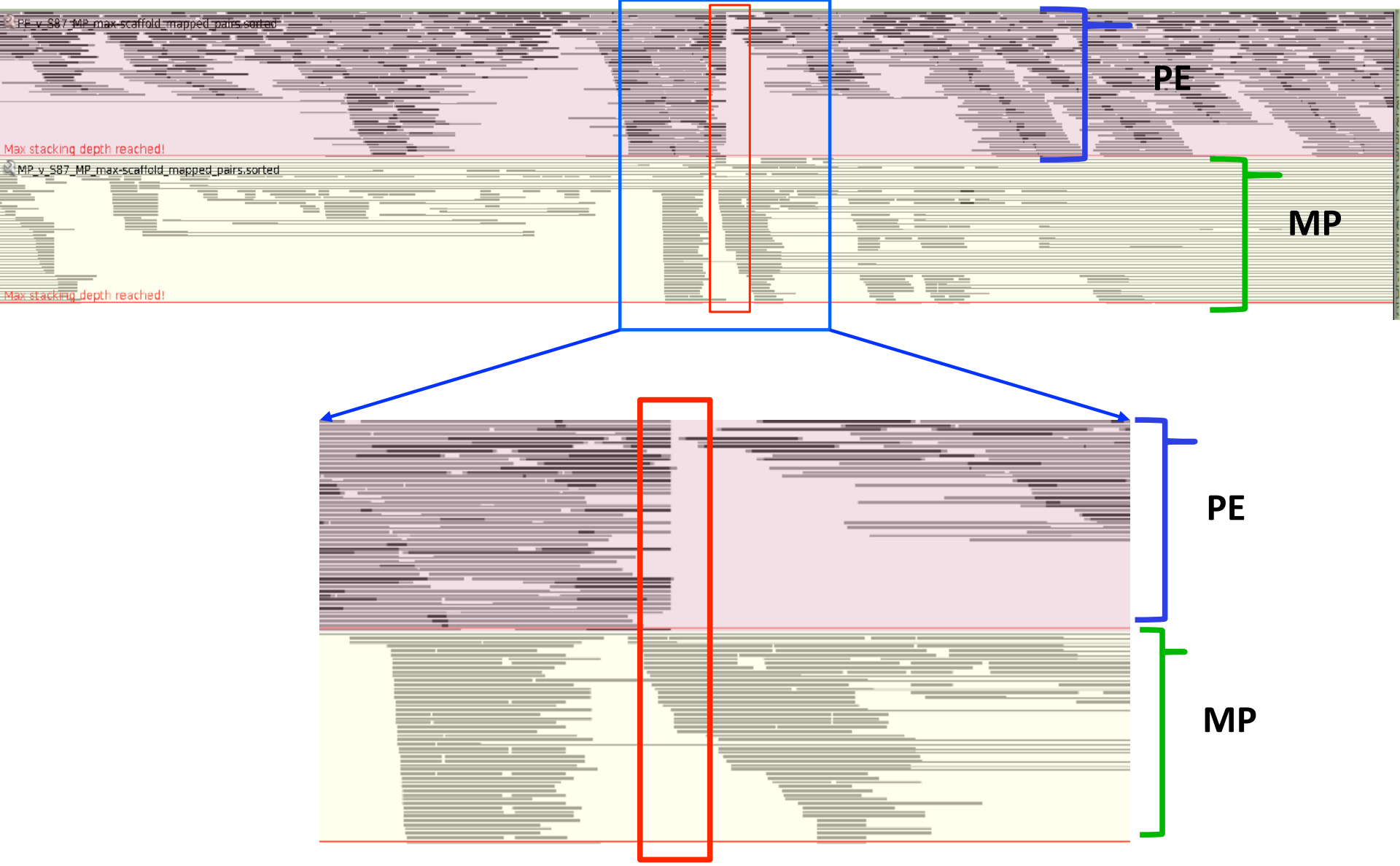


Comparison of longest contig of 24 trial BACs with different assemblers (PE reads only)

Assembly – PE read assembly with MP reads

Mapping of the reads back to the assembled scaffold (BAC87) with IGV

CTAG
Canadian Triticum Advancement through Genomics



An example of gap-closing for separate contigs from paired-end (PE) reads with mate-paired (MP) reads (BAC87, k=21)

CTAG
Canadian Triticum Advancement through Genomics

Overlaps between neighbouring BACs in MTP

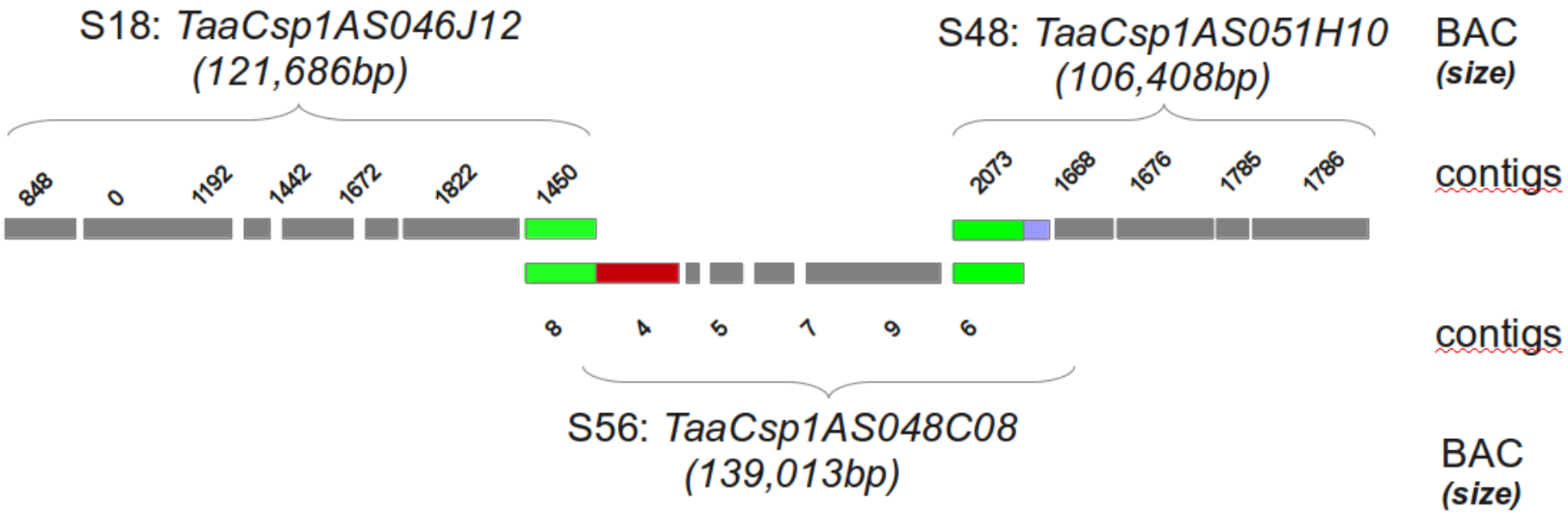
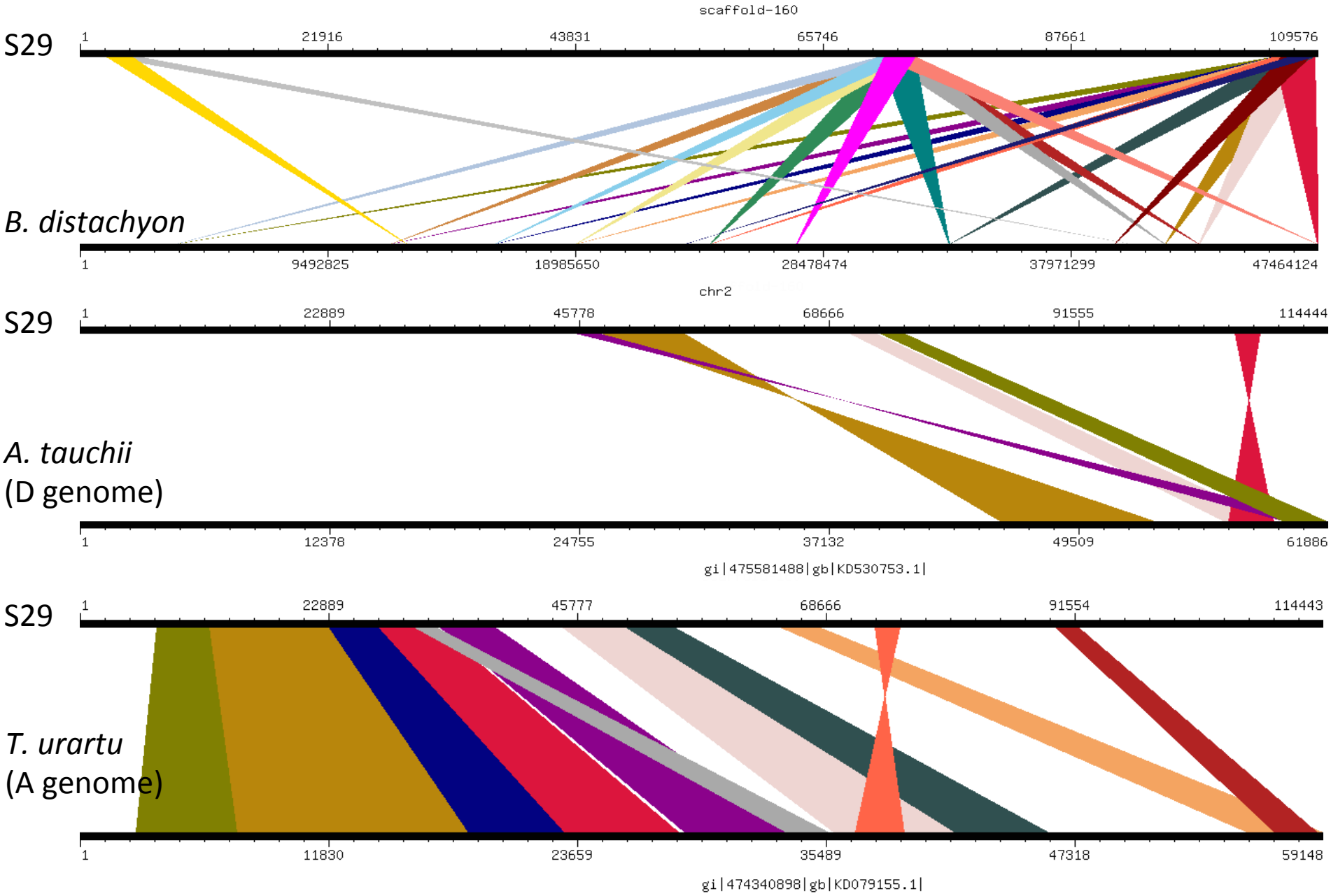


Diagram of the orientation of the three BAC in the MTP contig ltc4425.

The overlapping ends (green boxes) have been detected from the assemblies, and grey boxes are contigs whose relative positions and orientation are unknown

CTAG
Canadian Triticum Advancement through Genomics

Synteny with other genomes



Synteny between BAC S29 and the corresponding regions of three genomes (GSV)

CTAG
Canadian Triticum Advancement through Genomics

1AS MTP – status mid-2014

- All 1AS BACs PE sequenced – PCR free libraries
- 1/3 assemble to single scaffold with PE data **alone**
- Trial BAC pool MP data – should enable another at least another 1/3 of BACs to single scaffold
- MP libraries for all 4,134 clones – 96 or 384 non-indexed pools?

OR

- Whole genome MP for Chinese Spring?

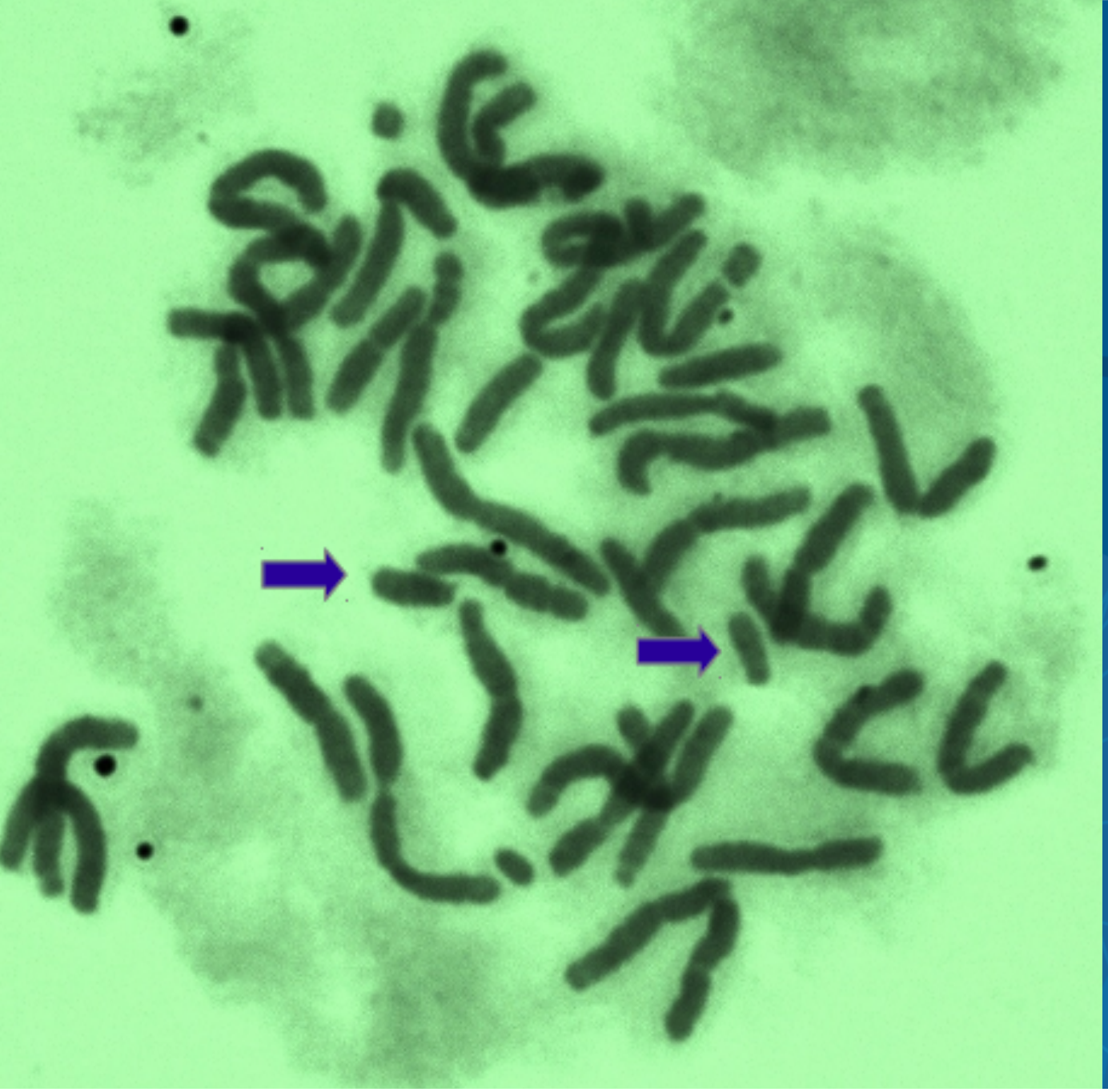
NRC Wheat Improvement Flagship Genomics Assisted Breeding (GAB)

Activity 2: Generating Strategic Genome Sequence Data for Wheat

Sequencing of wild “alien” addition chromosomes in Chinese Spring

Tall wheat grass (*Thinopyrum elongatum*) – FHB resistance (7EL)

Rye (*Secale cereale*) – cold (Puma 5RL)



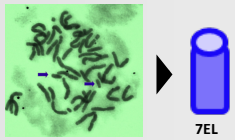
E genome – 7EL carries FHB resistance gene

E genome – very different at sequence level - difficult to induce recombination with 7A, 7B and 7D.

Ditelocentric addition line in Chinese Spring (arrowed)

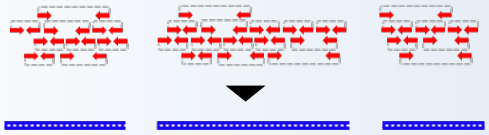
7EL assembly overview

7EL chromosome isolation and amplification - J. Dolezal (IEB)



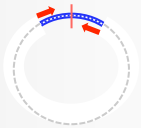
Paired end Illumina sequencing library preparation

Assembly of **contigs** (Ray)

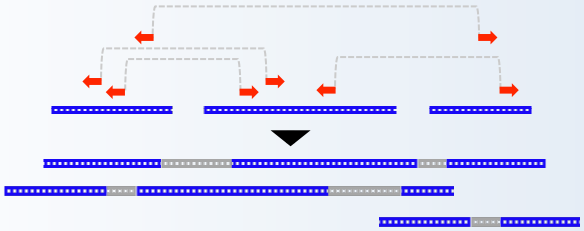


Nuclear DNA purification from Chinese Spring + 7EL addition line

Mate pair Illumina sequencing library preparation

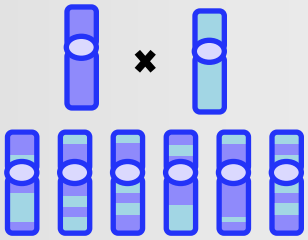


Scaffolding of contigs with mate pairs (SSPACE)

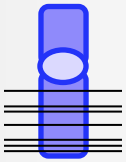


Annotation

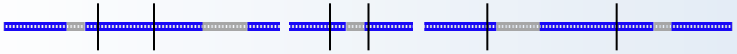
Mapping population generation



POPSEQ and/or GBS-based mapping



Integration of mapping data for ordering and orienting scaffolds



7EL assembly overview – MP data

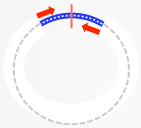
7EL chromosome isolation and amplification - J. Dolezal (IEB)

Paired end Illumina sequencing library preparation

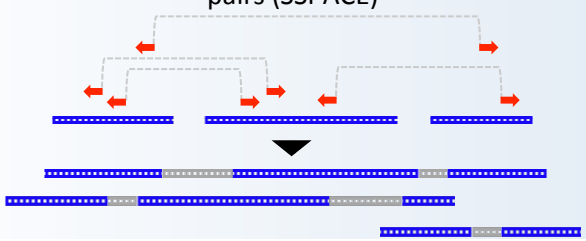
Assembly of **contigs** (Ray)

Nuclear DNA purification from Chinese Spring (green) + 7EL (blue) addition line

Mate pair Illumina sequencing library preparation



Scaffolding of contigs with mate pairs (SSPACE)



- ✓ 0.5 mg of DNA purified
- ✓ 16 Nextera mate-pair libraries (1.4 -20 Kbp)
- ✓ 12.6 lanes of Illumina HiSeq Rapid mode:
  - 2 × 150 bp
  - 584 Gbp raw data
  - 192 Gbp cleaned and filtered data
- ✓ 11x base coverage / 180x physical coverage
- ✓ Preliminary (60% of data) scaffolding results:

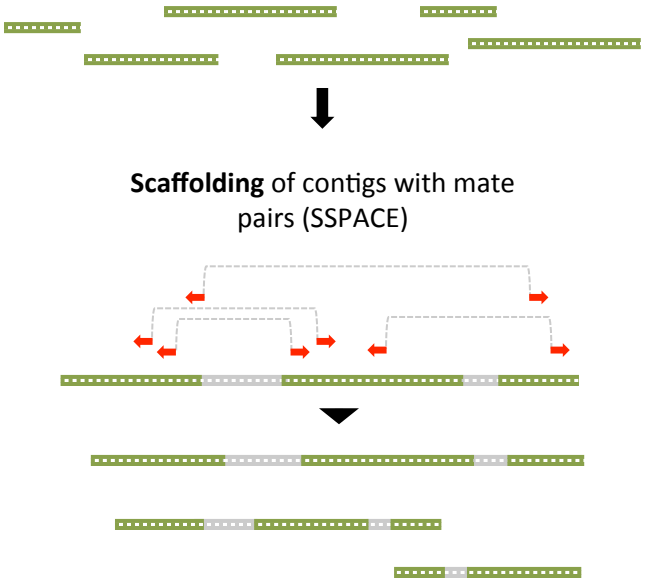
k-mer = 61 assembly	Contig	Scaffold
Total length	205 M	**241 M**
N50	7,997	**40,596**
Longest contig/ scaffold:	96,035	**471,944**

Chinese Spring 7EL mate pairs can also be used for scaffolding of **Chinese Spring reference chromosome and survey sequences**

Nuclear DNA purification from Chinese Spring (green) + 7EL (blue) addition line

Mate pair Illumina sequencing library preparation

Chinese Spring contigs



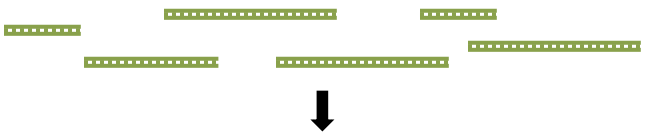
Scaffolding of contigs with mate pairs (SSPACE)

Chinese Spring 7EL mate pairs – scaffolding 1AS MTP

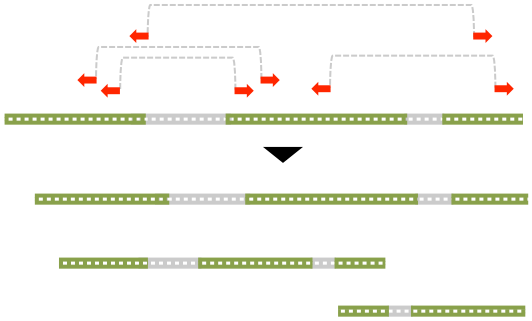
Nuclear DNA purification from Chinese Spring (green) + 7EL (blue) addition line

Mate pair Illumina sequencing library preparation

Chinese Spring contigs

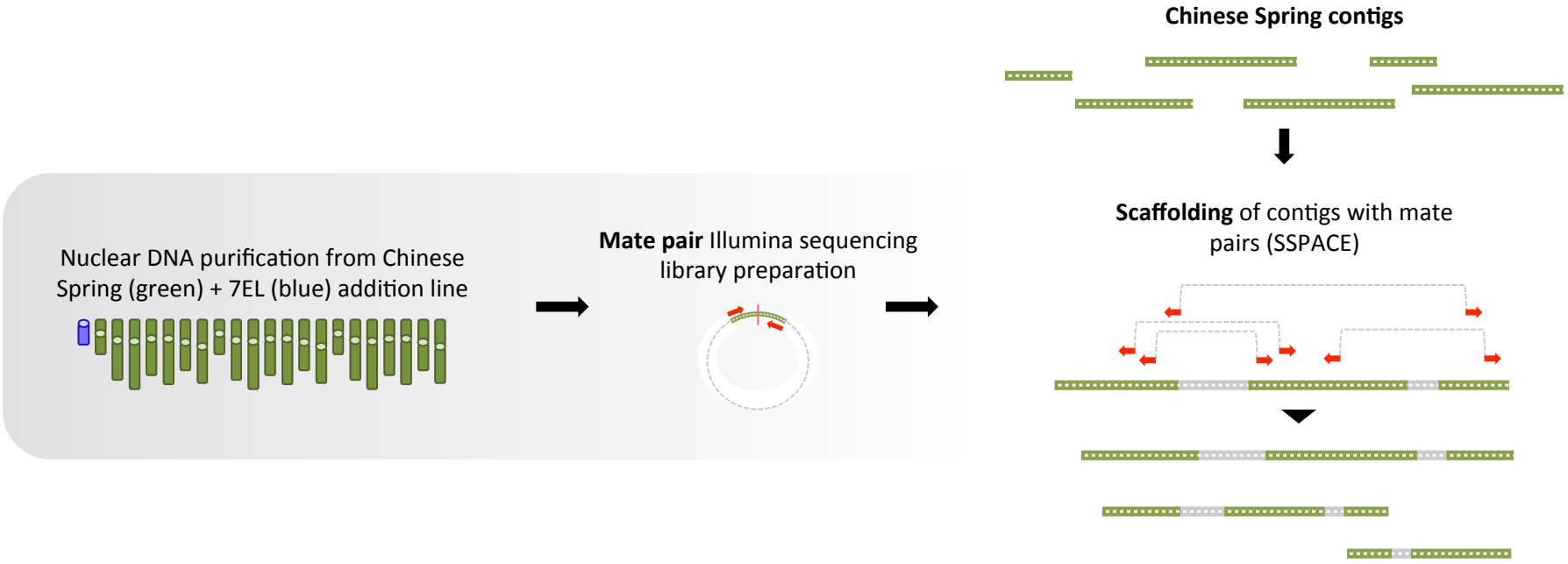


Scaffolding of contigs with mate pairs (SSPACE)



- Mate pairs mapping to 1AS assembly identified
- SSPACE v3 using Bowtie for mapping
  - no mismatches allowed, requirement of at least 5 unique connections per join

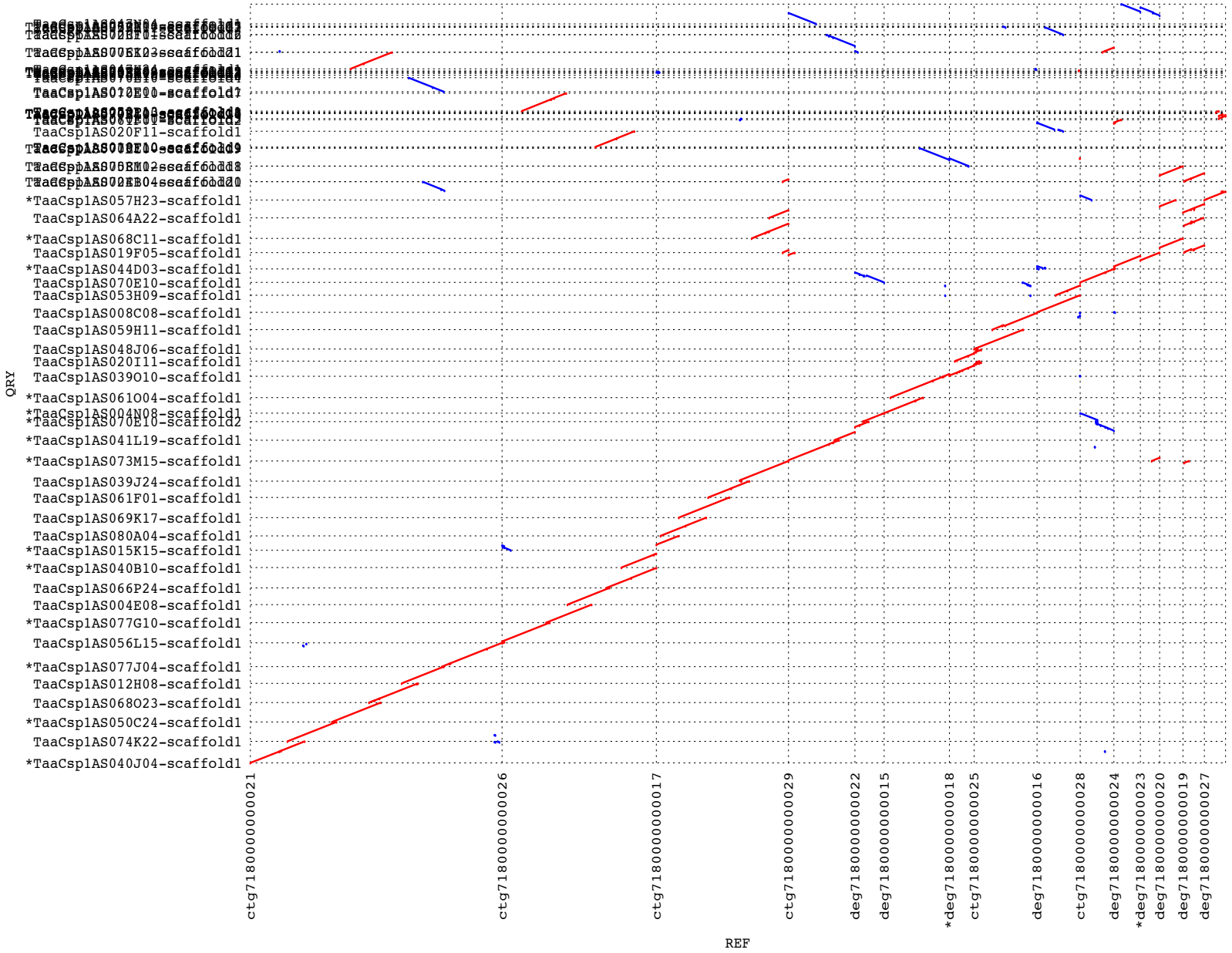
Chinese Spring 7EL mate pairs – scaffolding 1AS MTP



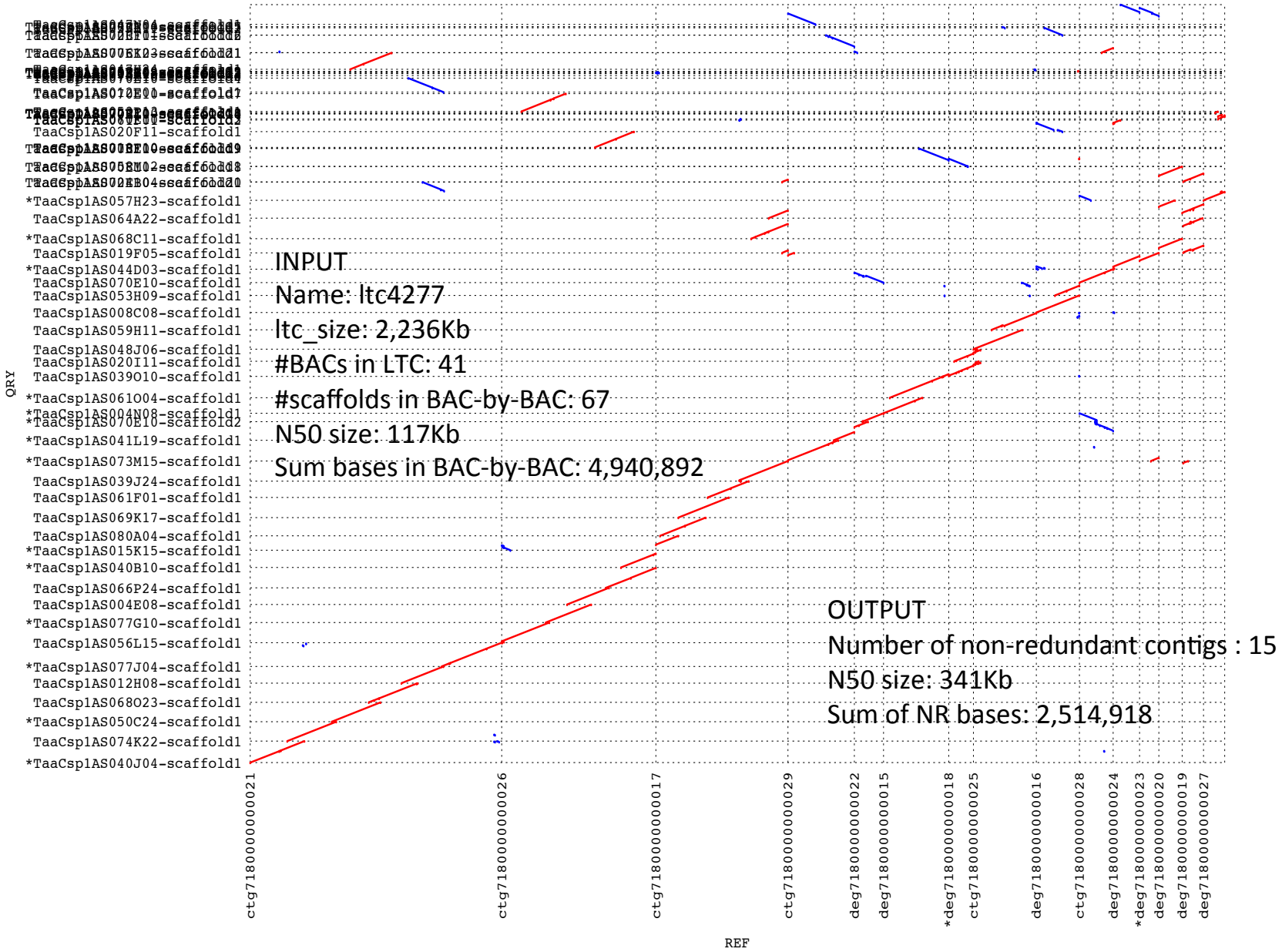
SUMMARY STATS FOR 1AS MTP	#scaffolds in	#scaffolds out
Mean	3.8	1.6
# of BACs down to 1 scaffold	1,749	3,110
# of BACs down to 2 scaffolds	800	608
# of BACs down to 1 or 2 scaffolds	2,549	3,718
# of BACs with > 2 scaffolds remaining	1,584	415
Mean reduction in # scaffs (for BACs with > 1 scaff in)		3.8
Totals	15,725	6,627

Note – 220 BACs (~5%) in 2 or more scaffolds represent 2 BACs (i.e. likely contamination)

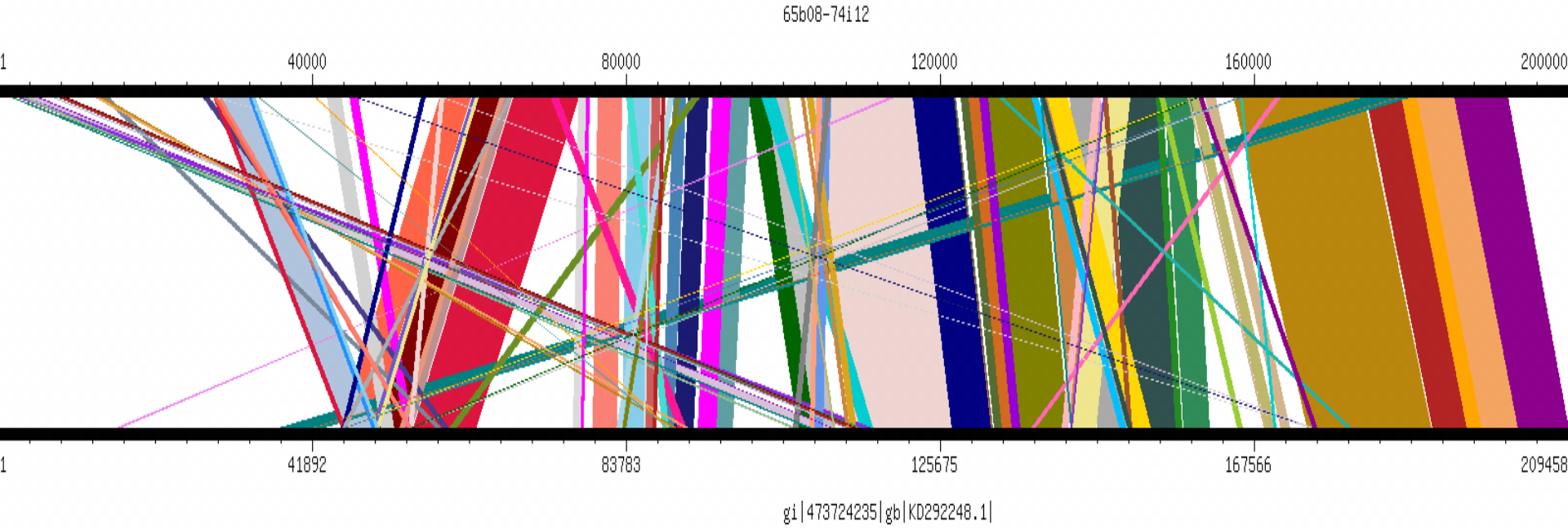
1AS physical contig (41 BACs) v. CABOG consensus



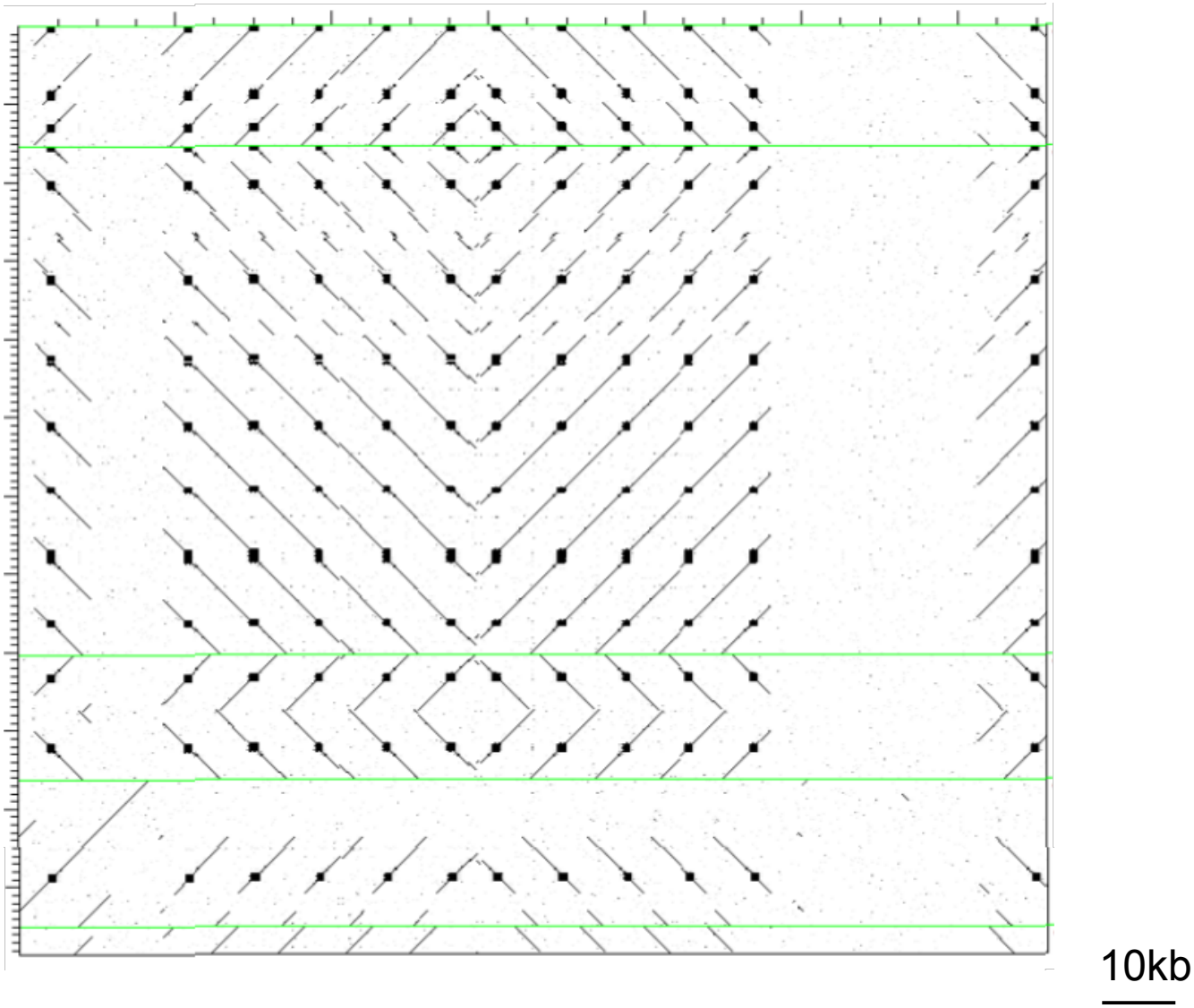
1AS physical contig (41 BACs) v. CABOG consensus



Synteny with other genomes



Alignment of sequenced physical contig ltc14 against a *T. urartu* scaffold (KD292248).



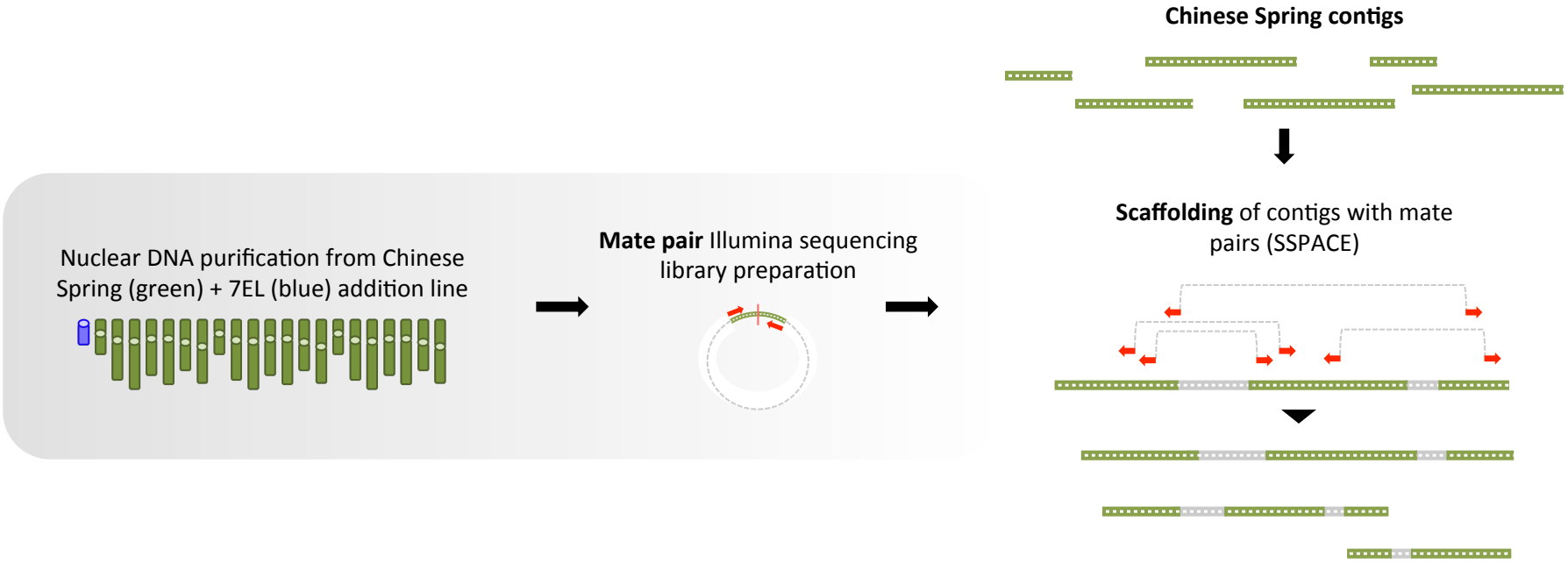
Dotplot of two BACs showing high repetitive regions within each sequence.
X-axis is a single scaffold of ~130kb, and Y-axis is 5 contigs (~120kb in total)

CTAG
Canadian Triticum Advancement through Genomics

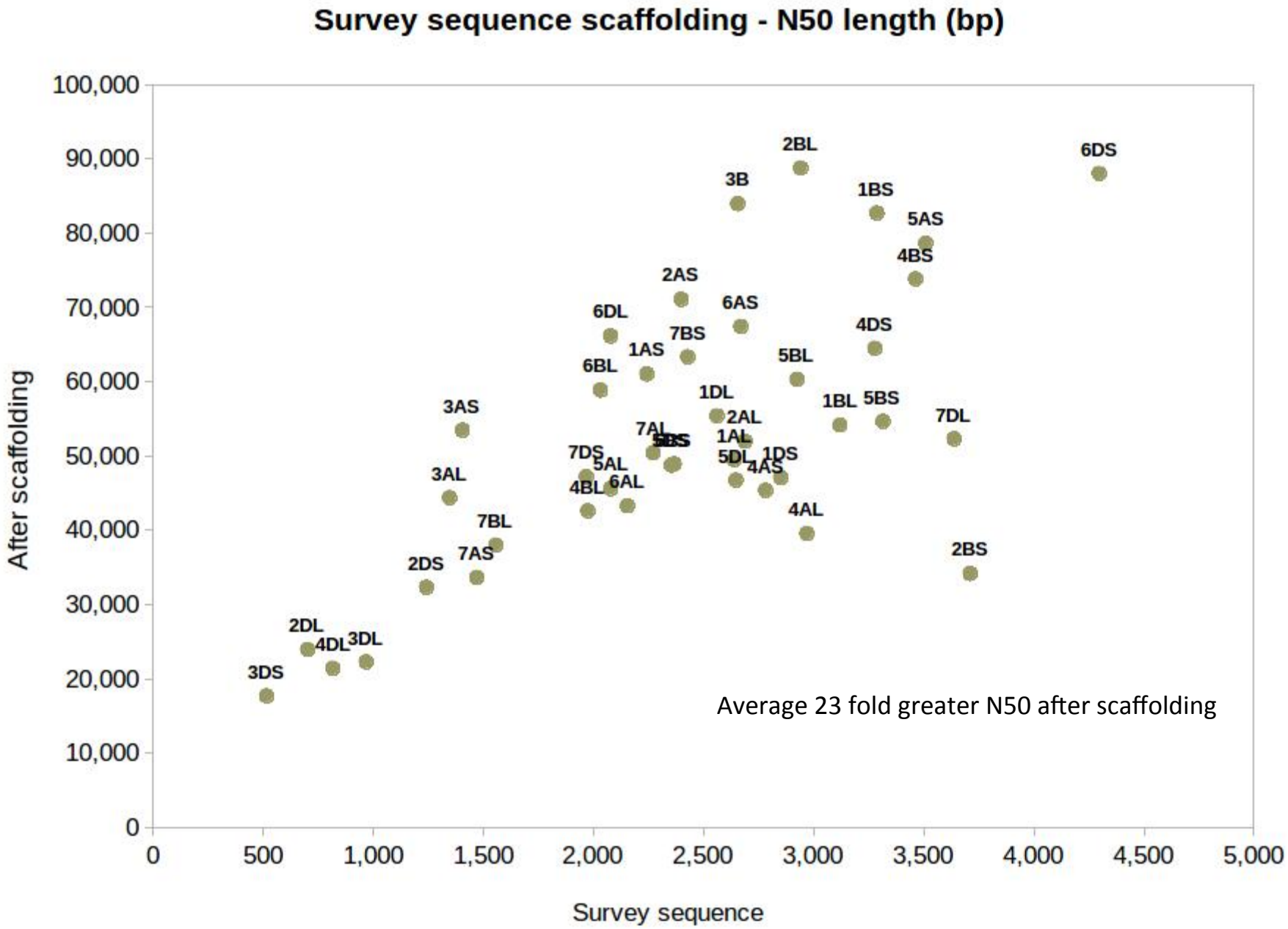
Summary status for sequencing 1A MTP

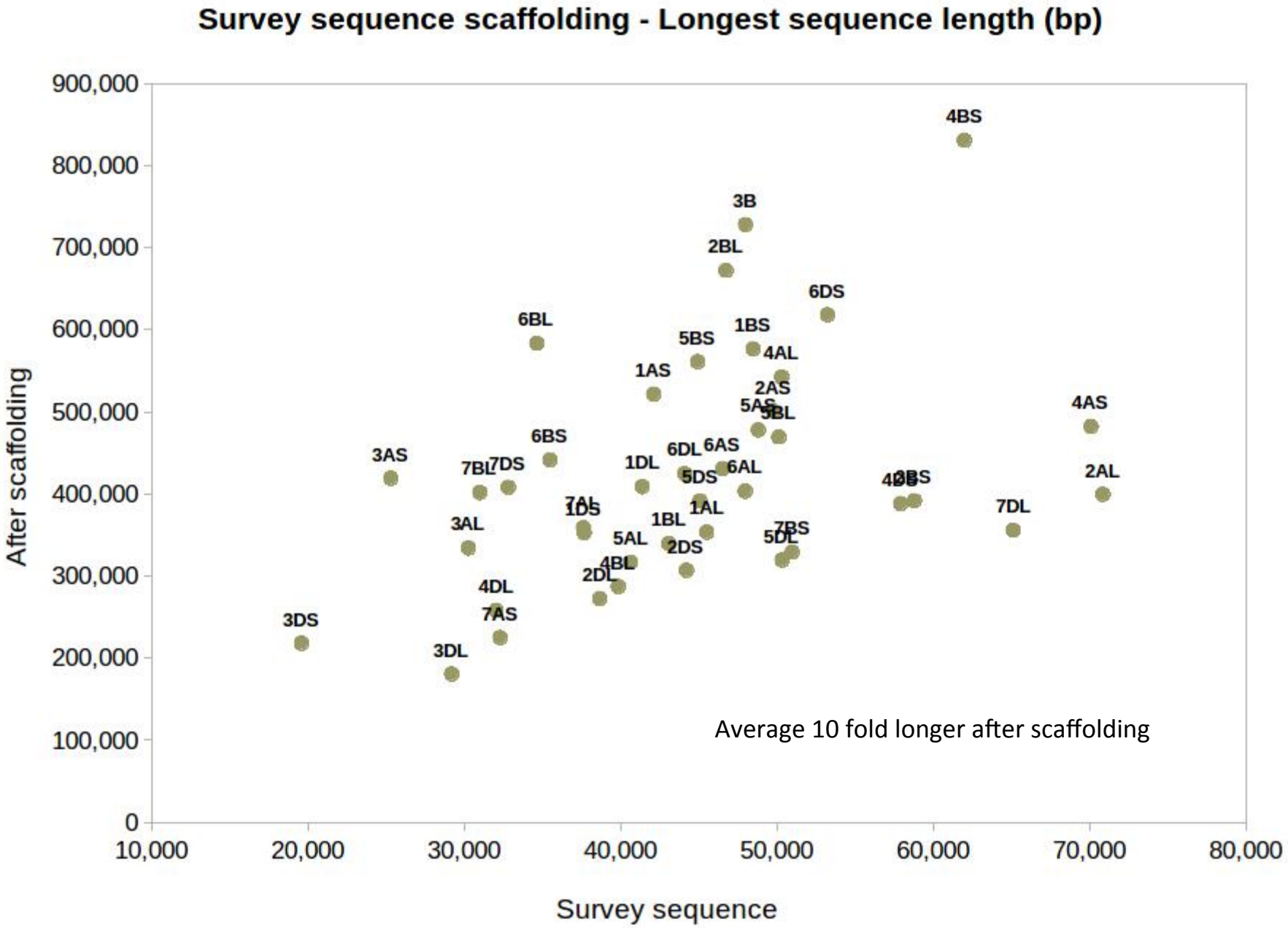
- Timeline for 1A sequencing
  - 4,134 1AS MTP BACs sequenced and assembled
  - 6,291 1AL MTP BACs delivered in Nov 2014 from INRA stock centre
  - Autogen DNA preparation (4 x 96 per run)
  - 1AL sequencing and assembly completed by April 2015
- Completion of Assembly and Annotation for 1AS and 1AL
  - July 2015

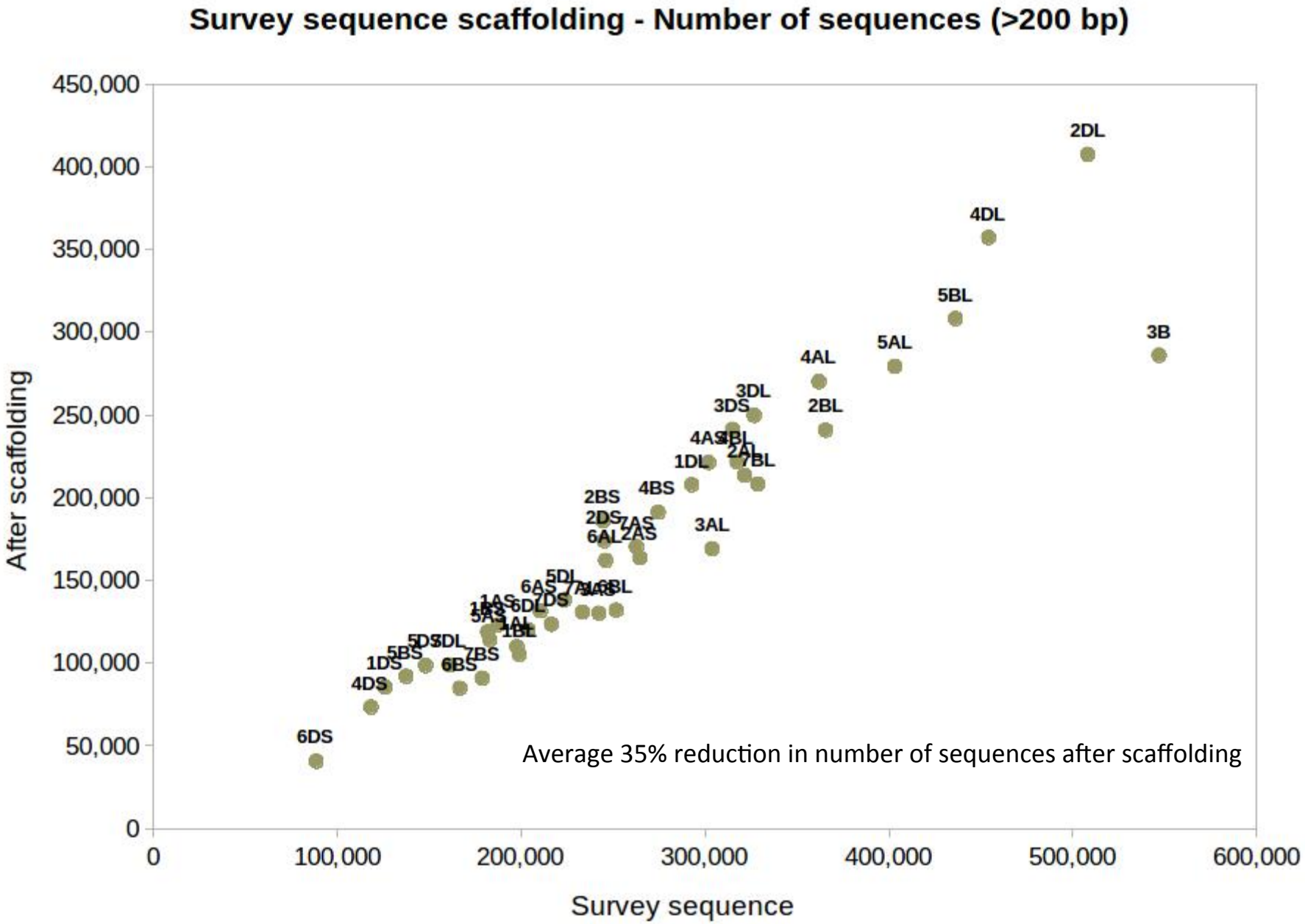
Chinese Spring 7EL mate pairs – scaffolding survey sequence

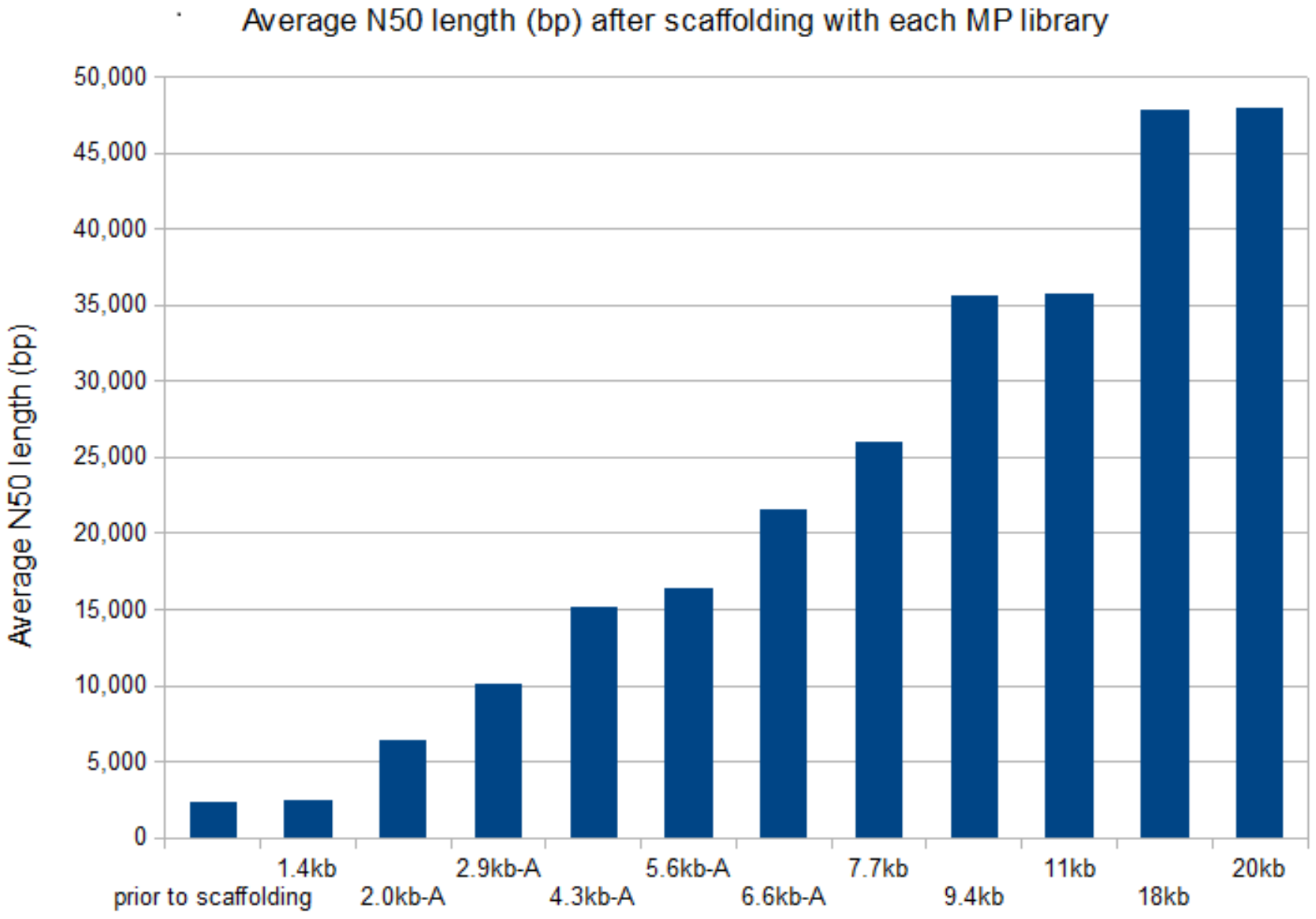


- Mate pairs mapping to 7EL assembly removed
- Mate pairs mapping to > 1 survey sequence chromosome arm removed
- SSPACE v3 using Bowtie for mapping
  - no mismatches allowed, requirement of at least 3 unique connections per join

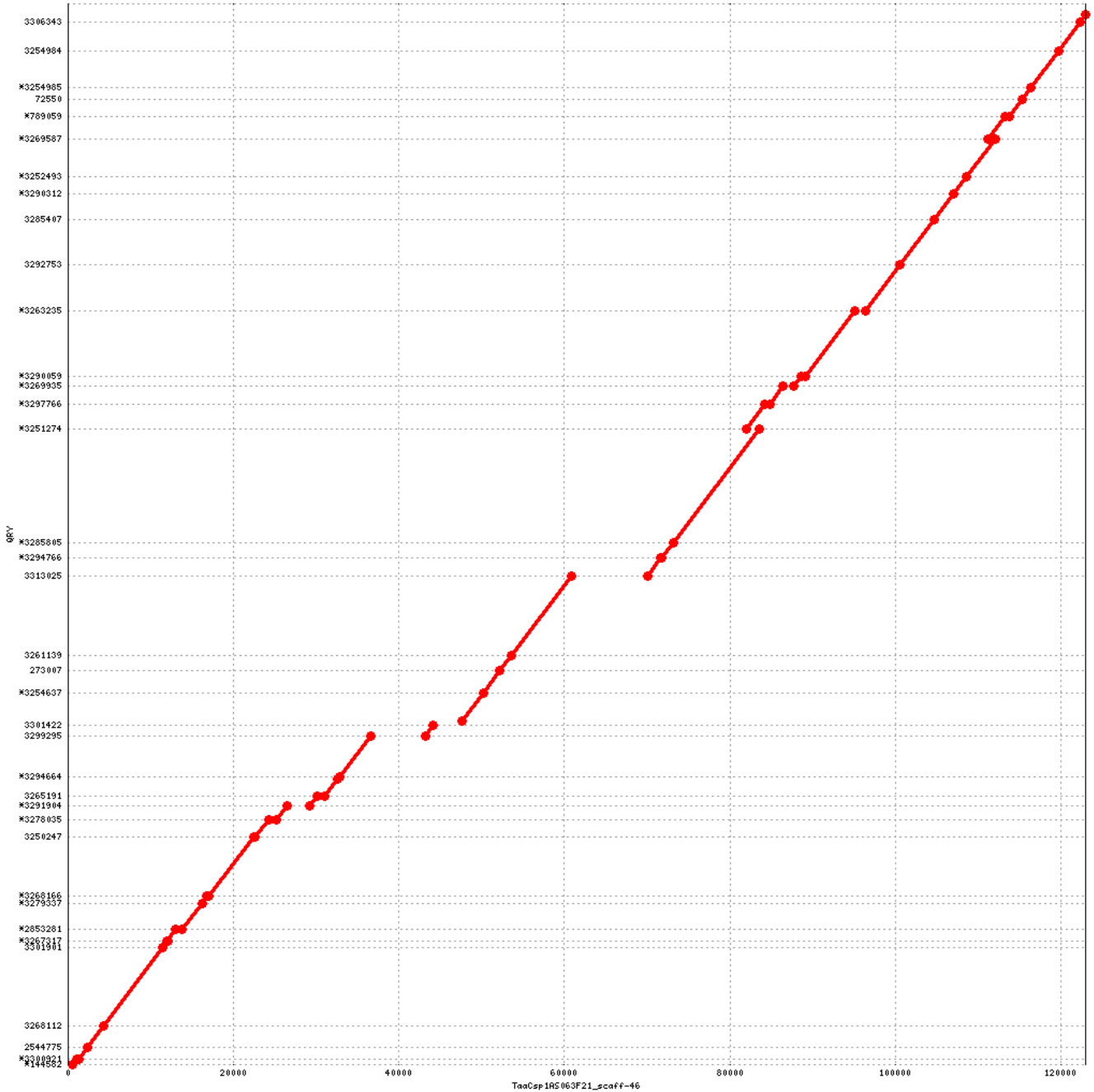

Survey sequence scaffolding - N50 length (bp)
After scaffolding
Survey sequence
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
2BL
6DS
3B
1BS
5AS
4BS
2AS
6AS
6DL
4DS
7BS
1AS
5BL
6BL
1DL
1BL
5BS
7DL
3AS
2AL
7AL
1AL
7DS
5AL
5DL
1DS
4AS
3AL
6AL
4BL
4AL
7BL
2BS
7AS
2DS
2DL
4DL
3DL
3DS
Average 23 fold greater N50 after scaffolding


Survey sequence scaffolding - Longest sequence length (bp)
After scaffolding
Survey sequence
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
10,000
20,000
30,000
40,000
50,000
60,000
70,000
80,000
4BS
3B
2BL
6DS
6BL
1BS
5BS
4AL
1AS
2AS
5AS
3BL
4AS
6BS
6DL
6AS
3AS
7BL
7DS
1DL
6AL
5DS
2AL
7AL
1DS
1AL
7DL
3AL
1BL
7BS
5AL
2DS
5DL
4BL
2DL
4DL
7AS
3DS
3DL
Average 10 fold longer after scaffolding

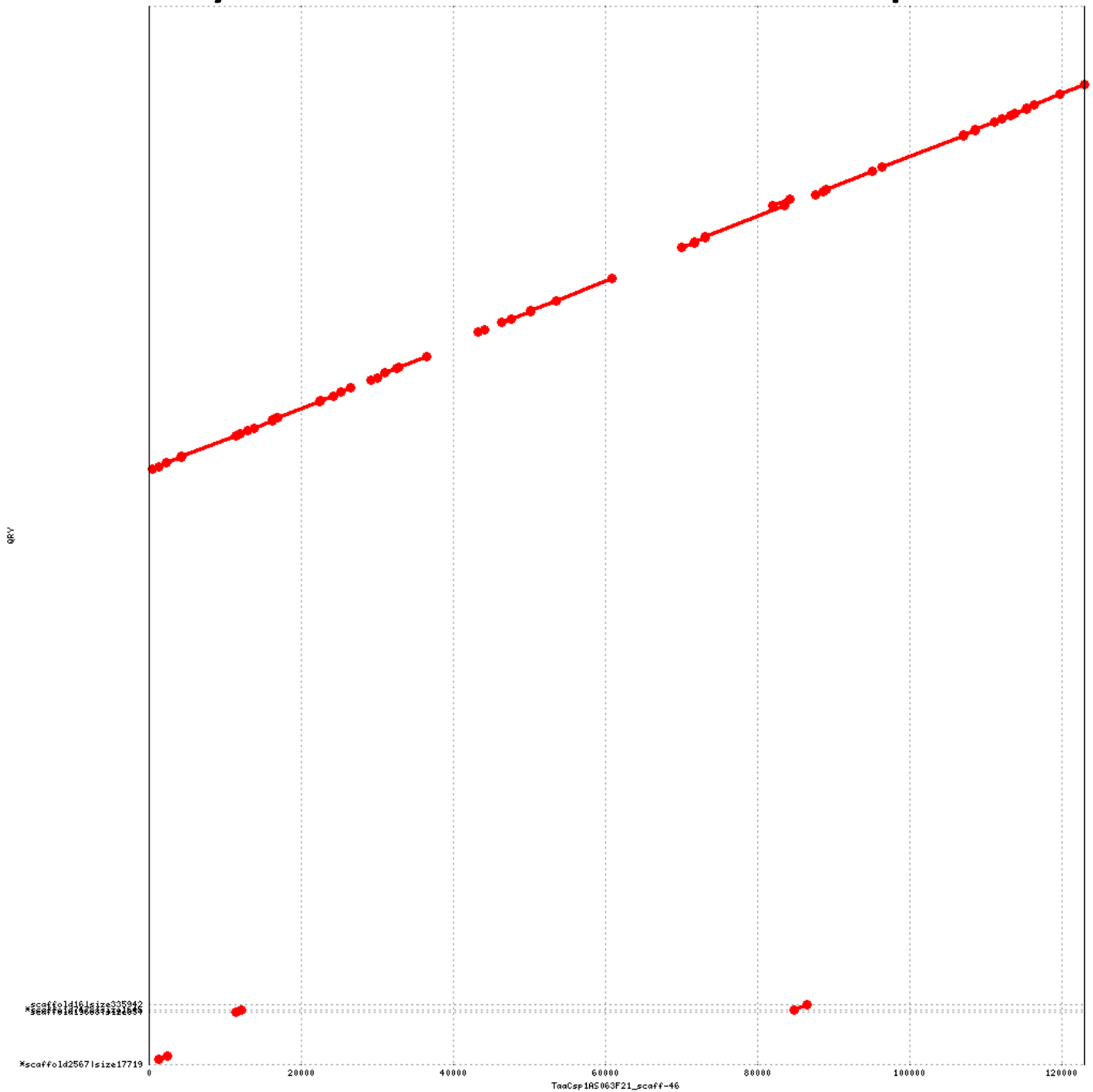

Survey sequence scaffolding - Number of sequences (>200 bp)
After scaffolding
Survey sequence
0
50,000
100,000
150,000
200,000
250,000
300,000
350,000
400,000
450,000
0
100,000
200,000
300,000
400,000
500,000
600,000
2DL
4DL
5BL
3B
5AL
4AL
3DL
3DS
2BL
1DL
7BL
4BS
2BS
2DS
7AS
6AL
3AL
5DL
6DL
7DS
1DS
5BS
4DS
6BS
7BS
6DS
Average 35% reduction in number of sequences after scaffolding


Average N50 length (bp) after scaffolding with each MP library
Average N50 length (bp)
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
prior to scaffolding
1.4kb
2.0kb-A
2.9kb-A
4.3kb-A
5.6kb-A
6.6kb-A
7.7kb
9.4kb
11kb
18kb
20kb

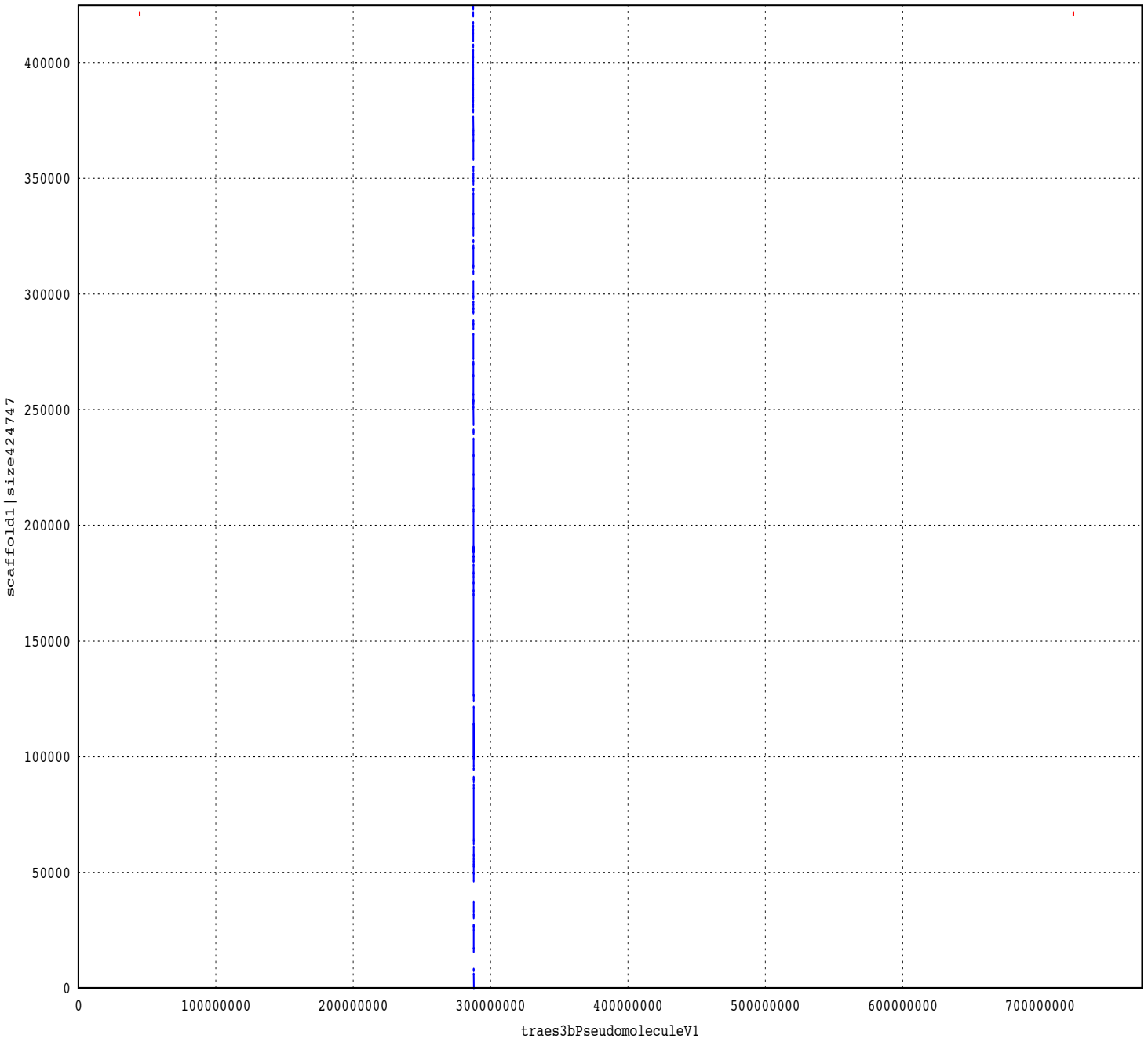
1AS survey contigs v. 1AS BAC sequence scaffold



1AS survey scaffold v. 1AS BAC sequence scaffold



3B survey scaffold v. 3B Psuedomolecule



Survey scaffold anchor stats for all CS chromosomes

CHR	Number of Contigs*	Total Assembly Length	#Contig positioned by POPSeq	#bps positioned by POPSeq	%assembly positioned	#Contigs positioned and orientated by Scaffolding +POPSeq	#bps positioned and orientated by Scaffolding +POPSeq	%assembly positioned after scaffolding
1AL	197,674	249,963,486	45,649	136,641,776	55%	91,542	194,267,076	78%
1AS	187,490	178,113,755	38,940	103,916,956	58%	67,916	129,655,488	73%
1BL	198,968	299,439,079	50,219	174,811,531	58%	99,517	241,865,633	81%
1BS	181,801	212,754,150	31,038	112,861,676	53%	64,050	158,401,963	74%
1DL	292,785	254,359,011	24,149	97,626,747	38%	82,277	177,835,638	70%
1DS	126,156	128,205,126	7,686	35,209,293	27%	36,023	80,238,998	63%
2AL	321,517	328,194,942	32,941	131,526,159	40%	106,827	233,645,461	71%
2AS	264,555	255,212,590	34,853	118,240,159	46%	100,896	193,923,883	76%
2BL	365,563	404,464,437	54,522	193,972,514	48%	124,744	297,676,517	74%
2BS	244,668	292,013,766	33,603	145,028,183	50%	65,884	196,456,657	67%
2DL	508,239	261,619,365	31,359	79,635,624	30%	96,866	134,427,841	51%
2DS	245,107	166,030,309	24,652	65,619,587	40%	71,922	107,231,792	65%
3AL	303,844	247,216,137	49,586	103,093,558	42%	133,828	179,517,605	73%
3AS	242,308	201,800,511	31,094	73,959,401	37%	110,413	150,853,649	75%
3B	546,922	638,625,269	99,341	321,122,800	50%	260,180	510,626,137	80%
3DL	326,758	186,457,895	37,874	77,891,770	42%	81,502	110,701,150	59%
3DS	314,944	145,374,274	26,447	40,173,749	28%	71,339	69,618,348	48%
4AL	362,010	355,934,257	27,248	116,236,207	33%	88,538	215,584,670	61%
4AS	301,954	282,335,959	25,068	96,669,909	34%	75,670	172,258,110	61%
4BL	317,294	248,664,471	41,135	119,021,671	48%	98,655	174,066,744	70%
4BS	274,504	308,205,293	50,927	189,567,586	62%	86,223	225,301,149	73%
4DL	454,216	254,434,292	23,842	58,249,889	23%	89,591	118,119,302	46%
4DS	118,290	142,105,148	14,198	59,531,441	42%	42,858	101,698,231	72%
5AL	403,265	318,135,065	35,333	113,348,448	36%	109,820	199,823,817	63%
5AS	182,938	198,819,005	5,578	31,169,678	16%	56,522	133,473,454	67%
5BL	436,173	415,170,143	49,140	187,476,431	45%	123,720	287,526,042	69%
5BS	137,380	174,467,817	19,794	80,497,001	46%	47,596	127,200,443	73%
5DL	223,456	236,791,573	29,604	107,788,181	46%	84,877	176,681,864	75%
5DS	148,048	148,047,159	14,458	49,826,323	34%	48,189	99,065,824	67%
6AL	245,867	214,378,543	30,828	98,984,004	46%	84,640	153,146,690	71%
6AS	210,388	219,217,215	28,234	104,098,636	47%	79,170	165,963,161	76%
6BL	251,706	257,411,872	38,064	105,791,844	41%	116,960	195,933,198	76%
6BS	166,632	210,183,634	30,962	93,388,287	44%	78,269	155,445,482	74%
6DL	203,805	199,800,717	23,763	85,201,598	43%	82,416	150,443,373	75%
6DS	88,542	156,583,290	18,701	87,679,813	56%	48,883	134,181,781	86%
7AL	233,306	252,444,836	32,435	100,690,814	40%	100,081	186,926,907	74%
7AS	262,653	198,002,837	31,628	75,350,273	38%	87,324	129,190,519	65%
7BL	328,725	259,605,653	50,397	112,699,366	43%	117,359	176,459,605	68%
7BS	178,789	206,115,020	48,514	129,914,644	63%	91,623	169,635,397	82%
7DL	161,061	222,868,133	31,832	126,294,969	57%	66,282	173,716,159	78%
7DS	216,406	209,134,978	41,796	104,614,864	50%	94,339	157,332,178	75%

Survey scaffold anchor stats for 1A, 5A and 6D

CHR	Number of Contigs*	Total Assembly Length	#Contig positioned by POPSeq	#bps positioned by POPSeq	%assembly positioned	#Contigs positioned and orientated by Scaffolding+POPSeq	#bps positioned and orientated by Scaffolding+POPSeq	%assembly positioned after scaffolding
1AL	197,674	249,963,486	45,649	136,641,776	55%	91,542	194,267,076	78%
1AS	187,490	178,113,755	38,940	103,916,956	58%	67,916	129,655,488	73%
5AL	403,265	318,135,065	35,333	113,348,448	36%	109,820	199,823,817	63%
5AS	182,938	198,819,005	5,578	31,169,678	16%	56,522	133,473,454	67%
6DL	203,805	199,800,717	23,763	85,201,598	43%	82,416	150,443,373	75%
6DS	88,542	156,583,290	18,701	87,679,813	56%	48,883	134,181,781	86%

* Contigs >200bp

Key Points

- **MiSeq performs well for 1A MTP BAC sequencing**
  - PE 2 x 250bp; even better with increased read length (2 x 300 reads)
  - whole genome MP libraries on HiSeq suitable for scaffolding – avoids BAC pool libraries
  - whole genome MP data – work for any MTP chromosome (e.g. 7A)
- **Assembly**
  - Ray performs well with these data; very fast (others – Discobar?)
  - synteny with *T. urartu* valuable for validation of assemblies
  - Consensus sequence for physical contigs (CABOG)
  - 40 Kbp large insert jump library for super-scaffolding (done - 16 x physical coverage)
  - assess super-scaffolding (SSPACE and Bambus) across and between physical contigs
  - build 1A MTP pseudomolecule / annotate
- **Enhanced CS Survey Sequence**
  - whole genome MP data – enabled robust scaffolding of all CS survey sequence
  - add 40 Kbp MP data and integrate with POPSEQ anchoring
  - new scaffold version of CS survey sequence for mid 2015

Acknowledgements

DNA Technologies & Bioinformatics Labs, NRC Saskatoon

Dr. David Konkin

Dr. Yifang Tang

Kevin Koh (Poster #582 – CS survey)

Janet Condie

Dr. Curtis Pozniak, CDC U. of Saskatchewan

Dr. Jennifer Ens

Krystalee Wiebe

Ron MacLachlan

Prof. Beat Keller, Ins. of Biology, U. of Zurich.

Dr. Thomas Wicker

Dr. Hikmet Budak, Sabanci University, Istanbul

Dr. Hélène Berges, INRA Plant Genomic Resource Center, Toulouse

Prof. Jaroslav Doležel, IEB, Czech Republic

Marie Kubalakova