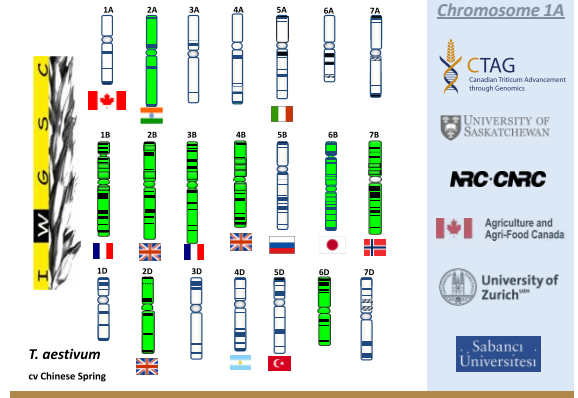
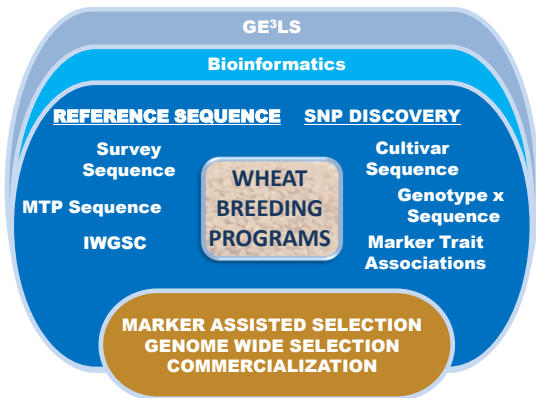
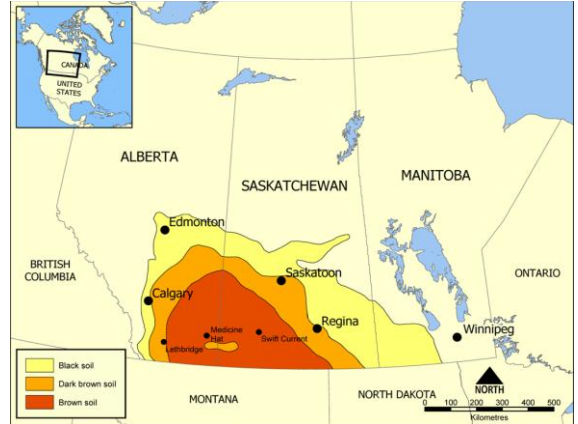




Reference sequencing of bread wheat chromosome 1A

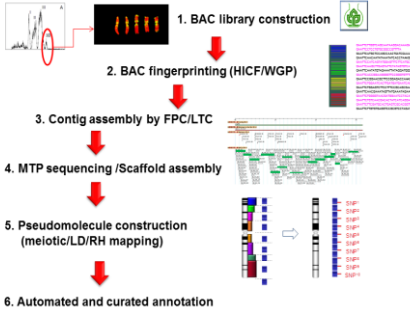
IWGSC Standards and Protocols Workshop

PAG XXII January 14<sup>th</sup> 2014

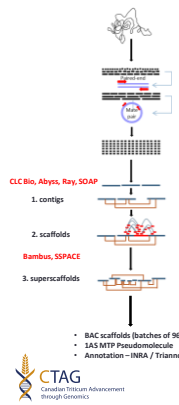


MTP Sequence: Chromosome 1A

IWGSC STRATEGY FOR OBTAINING A REFERENCE WHEAT GENOME SEQUENCE

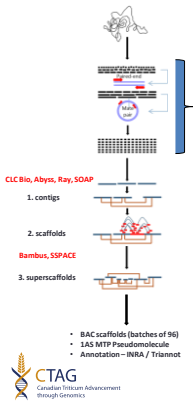


CTAG Strategy for assembly of 1A BAC MTP



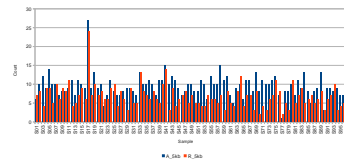
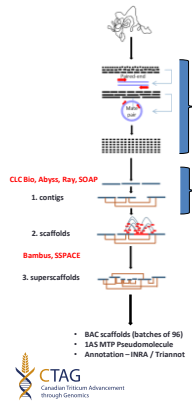
### CTAG Strategy for assembly of 1A BAC MTP

- 96 BAC preps (Amplicon Express)
- 96 TruSeq / Nextera kits (Illumina)
- MiSeq 2 x 250 pair-end (PE) reads (1 run, 8Gb, 96 BACs)
- Mate-pair (MP) of BAC pools (96-384)
- ~9000 BACs for 1AS and 1AL



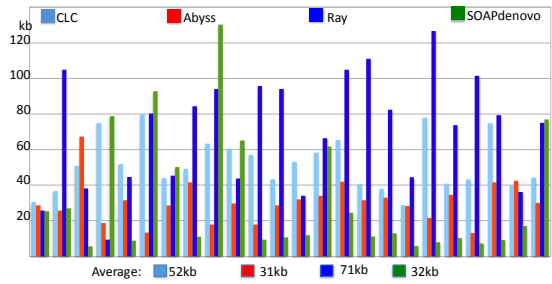
### CTAG Strategy for assembly of 1A BAC MTP

- 96 BAC preps (Amplicon Express)
- 96 TruSeq / Nextera kits (Illumina)
- MiSeq 2 x 250 pair-end (PE) reads (1 run, 8Gb, 96 BACs)
- Mate-pair (MP) of BAC pools (96-384)
- ~9000 BACs for 1AS and 1AL

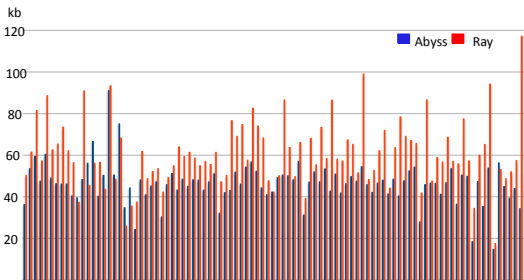


### Selection of assembler

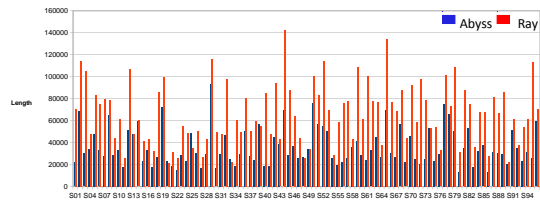
1. <http://www.clcbio.com>
2. <http://soap.genomics.org.cn>
3. <http://www.bcgsc.ca/~software/abyss>
4. <http://denovoassembler.sourceforge.net>



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

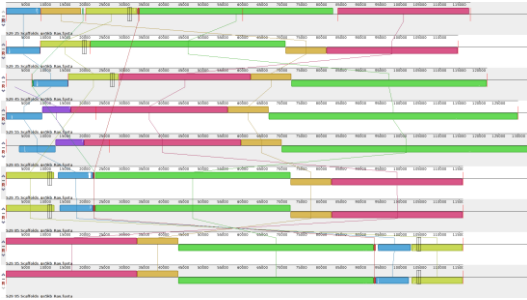


Total assembly length of each BACs for contigs longer than 5kb.



Longest contig size of each 96 1AS BACs - among all (PE reads only)

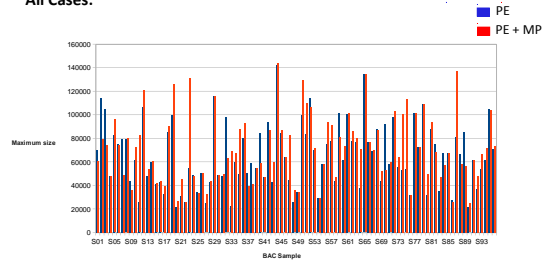
### Optimization of kmer selection with Ray



Alignment of PE read assemblies with different k-mers with Ray for BAC S29 (k=25, 35, 45, 55, 65, 75, 85 and 95)

### Ray assembly improvement with MP reads

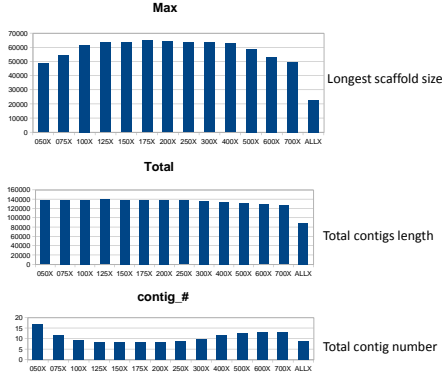
All Cases:



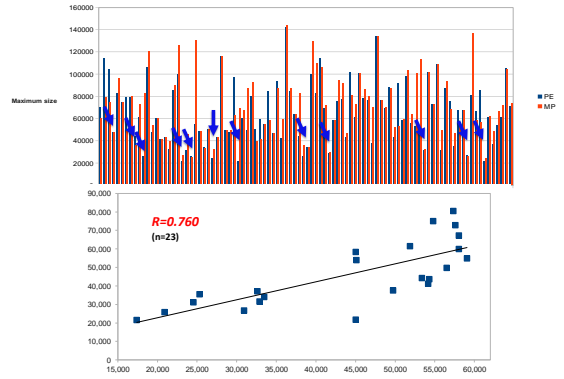
Comparison of maximum scaffold length without vs. with MP reads (Ray)

\*Sébastien Boisvert, François Lavolette, and Jacques Corbell. Journal of Computational Biology. November 2010, 17(11): 1519-1533. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies.

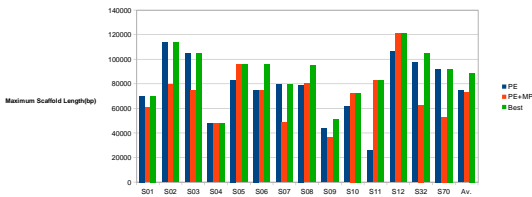
### Optimization of read coverage



Comparison of three metrics of Ray assemblies with different coverage

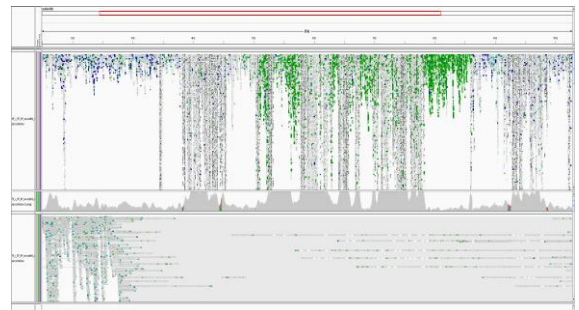


### Assembly gap-closing with MP reads

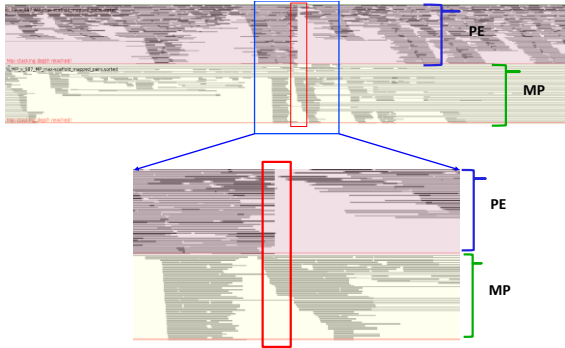


Improvement of assembly with PE & MP reads of 100x~250x coverage (based on ~150kb insert size) and full data set

Note: The best scaffold is the longest one from all assemblies with different kmers (15~97) with different coverage (100x, 150x, 200x, 250x and full data sets)



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



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



An example: closing broken scaffolds from paired-end (PE) reads with mate-paired (MP) reads (BAC75, k=57)

**Overlaps between neighbouring BACs in MTP**

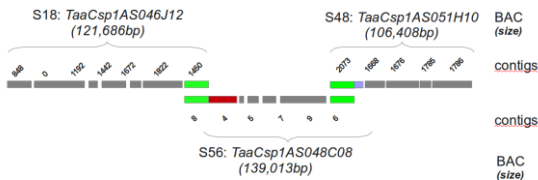
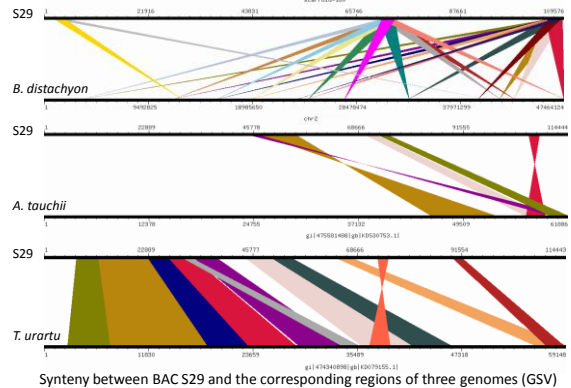


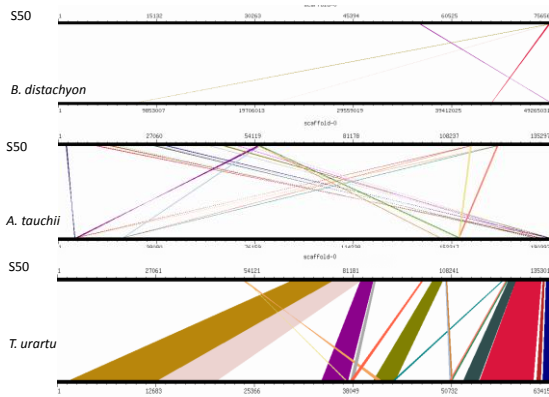
Diagram of the orientation of the three BAC in the MTP contig Itc4425.

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

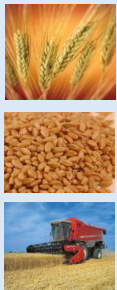
**Synteny with other genomes**



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




Synteny between BAC S50 and the corresponding regions of three genomes (GSV)



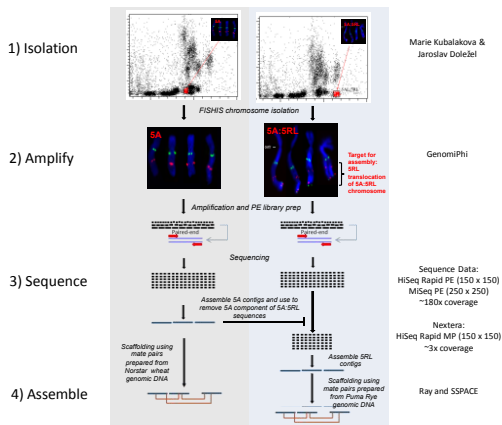
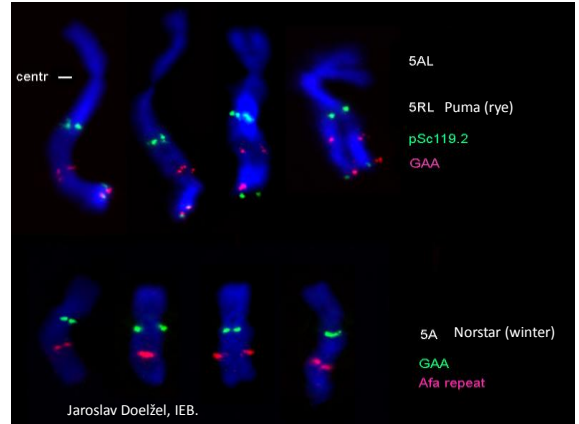
**Timeline for sequencing of 1A MTP**

- Begin sequencing of 1AS
  - SAB approval Dec. 2012
- Completion of 1A sequencing
  - October 2014
  - Currently 800 BACs sequenced
  - 3,200 BACs for 1AS delivered in Dec 2013
- Completion of Assembly and Annotation
  - July 2015



### Key Points

- **MiSeq data performs well for 1A MTP BAC sequencing**
  - even better with increased read length (2 x 300 reads)
  - optimize MP libraries (e.g. 384 pools)
- **Assembly**
  - Ray performs well with these data; very fast
  - synteny with *T. urartu* valuable for validation of assemblies
  - assess super-scaffolding (Bambus)
  - build 1A MTP pseudomolecule / annotate
- **Assess whole chromosome sequencing**
  - HiSeq / MiSeq PE sequencing on amplified flow sorted chromosomes
  - HiSeq MP sequencing on total nuclear DNA
  - strategy to complement 1A BAC MTP assembly
  - trialing with introgressions (rye and tall wheatgrass) in another project (B. Fowler and T. Ouellet)



Summary of paired-end sequence data obtained for Norstar wheat chromosome 5A.

Sequencing technology	Insert size (bp)	Total raw sequence length (Gb)	Total quality filtered sequence length (Gb)	Number of quality filtered paired-reads (millions)	Average read length after filtering
Illumina HiSeq rapid mode (2 x 150 bp, 3 lanes total)	300	34.3	30.4	104.7	145
	425	52.2	39.7	141.7	140
	550	48.1	36.4	129.3	141
	<b>subtotal</b>	<b>135.6</b>	<b>106.5</b>	<b>375.7</b>	<b>142</b>
Illumina MiSeq (2 x 250 bp, 8 lanes total)	300	12.0	11.6	47.8	243
	425	22.9	21.6	90.8	238
	550	20.5	18.8	81.1	232
	<b>subtotal</b>	<b>55.4</b>	<b>52.1</b>	<b>219.8</b>	<b>237</b>
<b>Total</b>		<b>191.0</b>	<b>158.6 (~187x coverage)</b>		

Summary of Nextera mate-pair sequence data obtained for Norstar wheat genome + statistics for mapping to Norstar 5A PE assembly

Mean insert size based on mapping / based on bioanalyzer trace	Number of high confidence pairs*	Number of low confidence pairs*	Percentage of high confidence pairs* that both map to contigs	Percentage of high confidence pairs* mapped within contigs with proper orientation and distance
3.1 Kb / 4.1 Kb	13,494,481	4,792,712	2.1 % (282,814)	97% (134,713/138,829)
4.5 Kb / 6.0 Kb	348,601	45,565	1.6 % (5,649)	94% (2,334/2,476)
7.0 Kb / 8.5 Kb	15,138,295	6,546,418	2.0 % (295,580)	91% (95,606/105,419)

\*Mate pairs for which the mate pair junction could be identified were considered high confidence

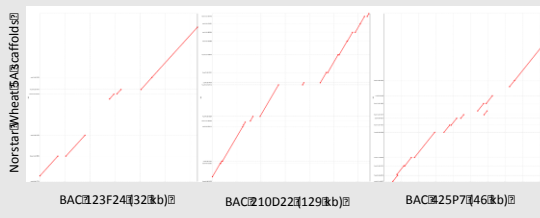
Note – 5A mate pairs reads ~1x physical coverage (3.2kb and 7kb, respectively) – will aim for 10x each for scaffolding proper

Statistics for Norstar 5A assembly using Ray with kmer = 85 (min 500bp)

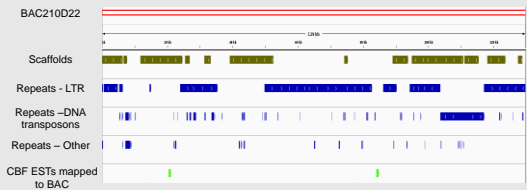
Assembly details	Number	Total length (Mb)	Largest	N50	Average
PE data only (Contigs)	382,079	933	100,677	5,025	2,442
PE data only (Scaffolds)	368,127	937	100,677	5,448	2,544
+ 2.5 kb mate pairs (Scaffolds)	361,532	945	130,605	6,058	2,614
+ 4.5 kb mate pairs (Scaffolds)	361,524	945	130,605	6,058	2,614
+ 8.5 kb mate pairs (Scaffolds)	347,131	989	135,381	8,751	2,848
+ all LQ <sup>a</sup> mate pairs (Scaffolds)	345,400	994	167,780	9,252	2,877

<sup>a</sup>Mate pairs for which the mate pair junction could not be identified were considered low confidence (LQ)

### Mummer alignments of current Norstar 5A scaffold vs 3 different Norstar 5A BACs assembled from Sanger sequencing



### Comparison of the regions of Norstar wheat BAC 210D22 to which scaffolds from the Norstar 5A assembly map to



Mummer alignment (red) with the repeat content (blue) and expressed sequence tags (ESTs) of C-repeat/dehydration-responsive element binding factor (CBF)

### Statistics for Puma rye 5RL preliminary assembly<sup>a</sup> using Ray with kmer = 79

Assembly details	Number	Total length (Mb)	Largest	N50	Average
PE data only (Contigs)	160,913	388	58,208	4,104	2,414
PE data only (Scaffolds)	151,279	391	58,208	4,370	2,584
+ 2.5 kb mate pairs (Scaffolds)	151,279	396	59,329	5,119	2,676
+ 4.5 kb mate pairs (Scaffolds)	147,995	401	65,605	5,468	2,747
+ 8.5 kb mate pairs (Scaffolds)	145,913	417	107,294	6,671	2,955
+ all LQ <sup>c</sup> mate pairs (Scaffolds)	138,251	426	110,878	7,699	3,079

<sup>a</sup> assembly using only ~80% of acquired sequence, ~150x coverage

<sup>c</sup> Mate pairs for which the mate pair junction could not be identified were considered low confidence (LQ)



DNA Technologies & Bioinformatics  
Labs, NRC Saskatoon  
Dr. Yifang Tang  
Janet Condie  
Kevin Koh  
David Konkin (5A / 5A-5RL)  
- see Poster 30

## Acknowledgements



Dr. Curtis Pozniak, CDC U. of Saskatchewan  
Dr. Brian Fowler (5A / 5A-5RL)  
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 Doelžel, IEB, Czech Republic  
Marie Kubalaková

