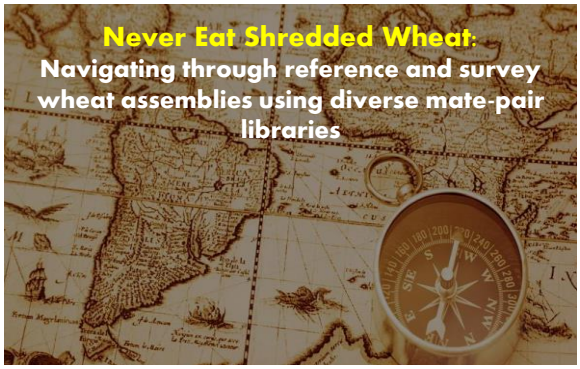
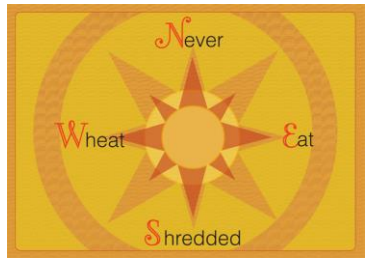
Never Eat Shredded Wheat:
Navigating through reference and survey wheat assemblies using diverse mate-pair libraries



Dr. David J. F. Konkin
National Research Council of Canada

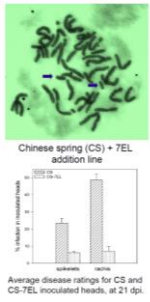


Outline

- 7EL assembly effort
- CS + 7EL mate pair library preparation
- IWGSC survey sequence improvements -> version 3
- 1A reference assembly update

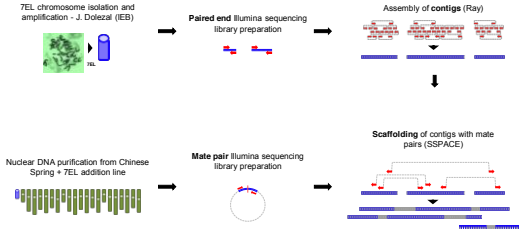
Thinopyrum elongatum chr 7EL harbours a source of fusarium head blight resistance

Thinopyrum elongatum (E genome, 2n=14)



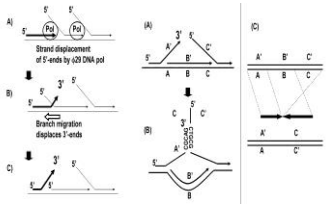
Average disease ratings for CS and CS-7EL inoculated heads, at 21 dpi.

7EL assembly overview



Unamplified nuclear DNA was used for mate pair libraries

- Multiple displacement amplification leads to chimera formation
  - ~1 per 22kb with E. coli DNA
  - Unamplified Nuclear DNA used rather than amplified DNA from isolated chromosomes

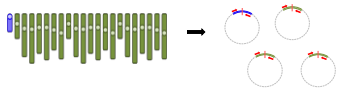


Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnology, 7(1), 19.

Mate pair library prep

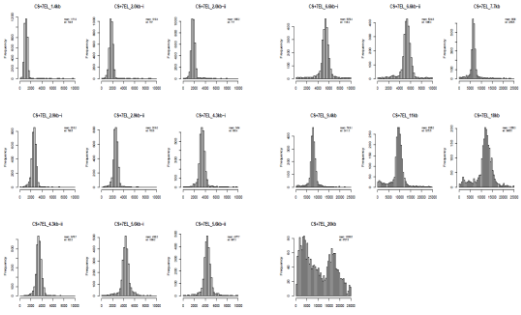
Nuclear DNA purification from Chinese Spring (green) + 7EL (blue) addition line

Mate pair library preparation and sequencing (Illumina Nextera and Lucigen NxSeq 40kb)

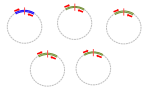


- 16 Nextera libraries (10 sizes), one Lucigen NxSeq 40kb
- Initial sequencing run used to estimate library diversity, subsequent focus on most diverse libraries
- 12.6 lanes of Hiseq rapid mode 2 x 150 bp reads, 1 lane Miseq 2 x 250 bp
- 591 Gbp raw sequence, 186 Gbp processed
- 113 bp average length after processing

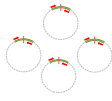
Library size distributions



Mate pair filtering



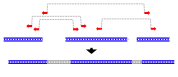
Removal of mate pairs that **map to 7EL sequences** (3% removed)



Removal of mate pairs that **map to multiple chromosome arms** (75 % of reads removed)

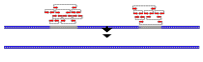
Scaffolding and gapfilling -> IWGSC version 3

Scaffolding of contigs with mate pairs (SSPACE)



- SSPACE v3.0 scaffolder
- Bowtie for mapping
- 32 bp minimum read length
- No mismatches allowed
- Minimum 3 unique connections required

Gapfilling (gapcloser)



- Original chromosome-arm specific paired-end reads
- Default settings

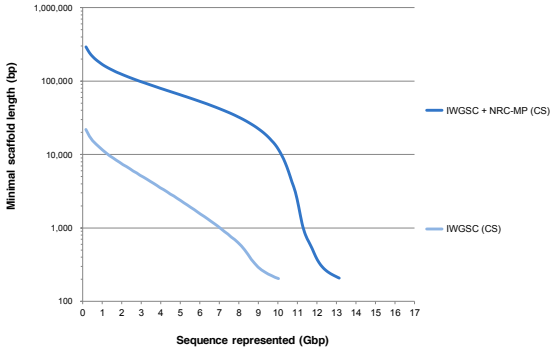
Improvements – assembly metrics

	IWGSC version 2	IWGSC version 3 (scaffolded and gapfilled)
Total contig Length (Gbp)	10.0	11.5
Total scaffold length (Gbp)	10.1	13.3
Number of contigs (millions)	11.7	8.6
Number of scaffolds (millions)	10.8	7.2
Contig N50 Length (Kbp)	2.2	9.6
Scaffold N50 length (Kbp)	2.4	47.4
Largest contig (Kbp)	71	193
Largest scaffold (Kbp)	71	755

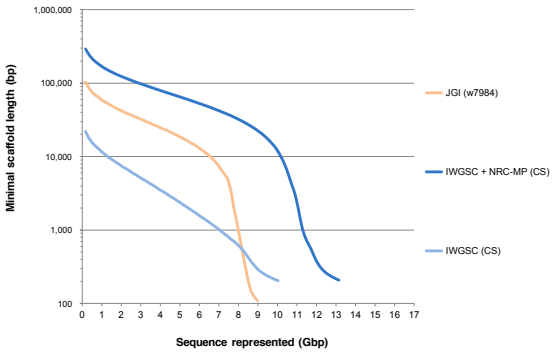
Improvements - usefulness

	IWGSC version 2	IWGSC version 3 (scaffolded and gapfilled)
Sequence anchored by POPSEQ (Gbp)	4.4	7.1
% 90K markers anchored	95.0	96.6
% full-length ESTs (98% ident, 70% cov)	51.3	76.8
% high-confidence transcripts (99% ident, 90% cov)	92.7	94.4
% beta capture probes anchored	75.6	87.8

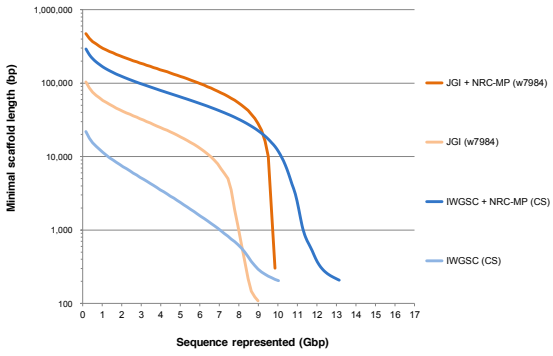
Genome assembly comparison



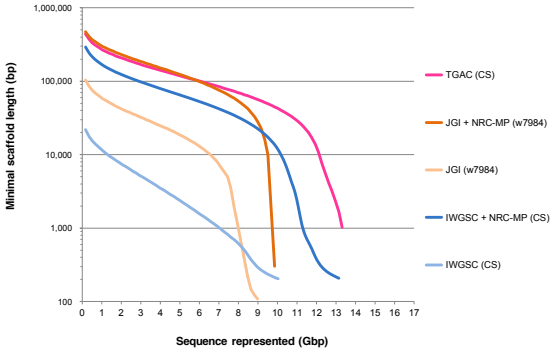
Genome assembly comparison



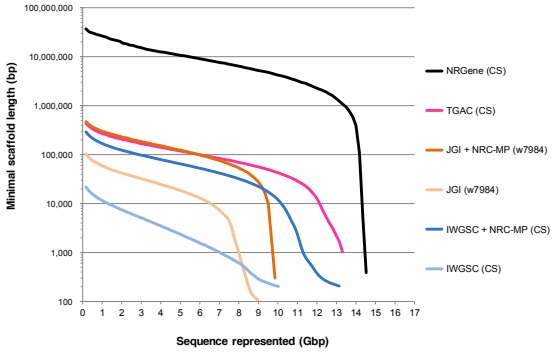
Genome assembly comparison



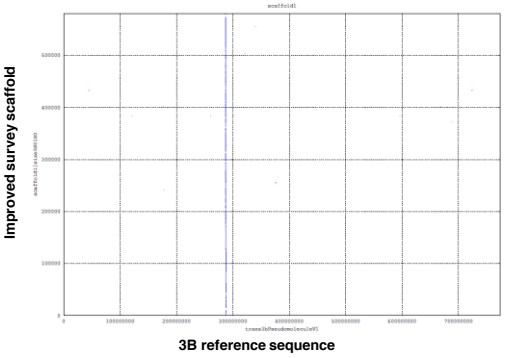
Genome assembly comparison



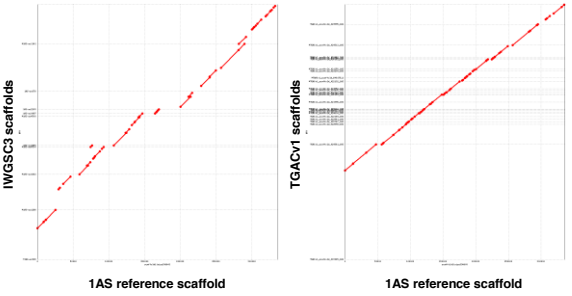
Genome assembly comparison



Comparison of IWGSC CSS v3 with 3B reference



IWGSC CSS v3 vs TGACv1



IWGSC CSS v3 – available to IWGSC membership

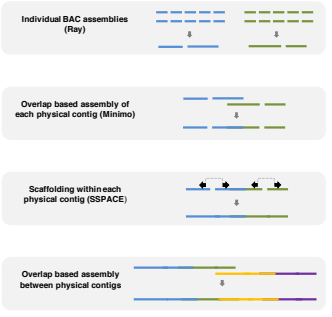
- News:
  http://wheat-urgi.versailles.inra.fr/Seq-Repository/News

  Assembly details:
  http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies

  Direct link to download:
  https://urgi.versailles.inra.fr/download/wheat/survey_v3/

  Direct link to the browser:
  https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_iwgsc_survey_sequence_v3/

1A reference map-based assembly strategy



Individual BAC assemblies

- 96 indexed BACs per Miseq run (2 x 250 bp)
- Individual BAC assemblies with Ray (kmer 51)
  - Ray > SOAP2denovo, Abyss and CLCbio
  - Ray ~ SPades and Discovar
- Of ~4100 BACSs :
  - 29 BACs failed during sequencing
  - 277 BACs removed because assembly size was greater than 50% larger than estimated
- 16709 contigs, 60.2 Kbp L50, 476 Mbp total size

Overlaps between BAC within a physical contig

- Minimo used to join overlaps within physical contigs
  - % 98 identity over 1000 bp
- Differences between adjacent assemblies a problem
  - different estimate of gap or ~indel
  - ~ 20 Mb of overlaps remaining
- 7160 contigs, 80.6 Kbp L50, 265 Mbp total size

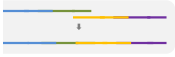
Scaffolding within each physical contig

- All mate pairs mapped against complete chromosome to reduce volume of data and remove multimapping mates
- SSPACE scaffolder
  - Iterative scaffolding with overlapping sets of mate pair libraries
  - Five connections required per join
- 2013 Scaffolds (~average 2.8 scaffolds per physical contig)
- 268 Mbp total sequence (1.3% gap)
- Scaffold L50 length: 273.5 Kbp

1A reference map-based assembly strategy

	# scaffolds	Scaffold L50 (Kbp)	Total length (Mbp)
Individual BAC assemblies (Ray)	16709	60.2	476
Overlap based assembly of each physical contig (Minimo)	7160	80.6	265
Scaffolding within each physical contig (SSPACE)	2013	276.5	268
Overlap based assembly between physical contigs	TBD	TBD	TBD

Overlap based assembly between physical contigs

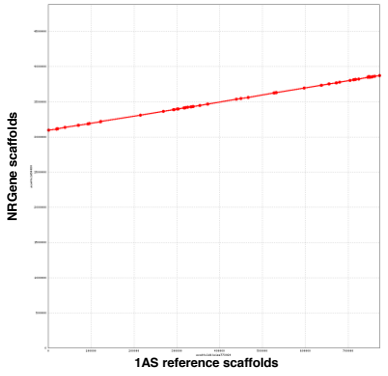


- Exploring different options for merging overlaps
  - Megamerge, cabog, custom software – what are others using?
  - ~ 200 overlaps > 20 kb

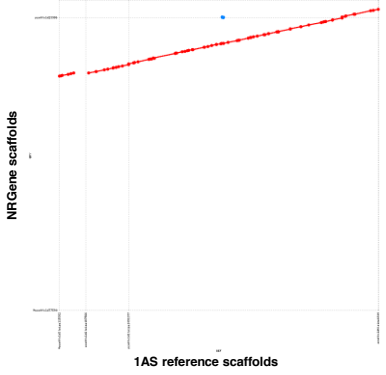
1AS summary statistics

	1AS physical contig assembly summary stats
Scaffold Total	2185
Contig Total	8792
Scaffold Sequence Total (Mbp)	280.0
Contig Sequence Total (Mbp)	273.2
Scaffold L50 (Kbp)	265.5
Scaffold N50	282
Contig L50 (Kbp)	80.1
Contig N50	1014
Max Scaffold Length (Kbp)	2232
Max Contig Length (Kbp)	514
% Gap	2.43

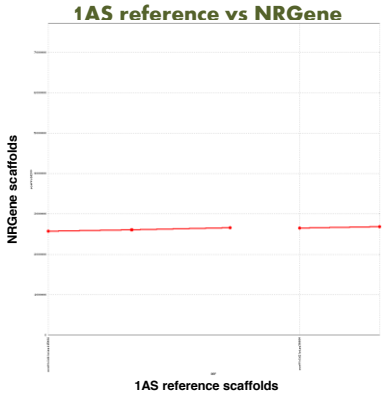
1AS reference vs NRGene



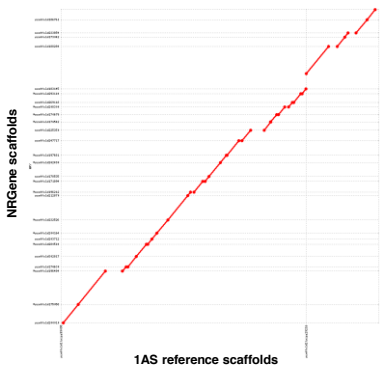
1AS reference vs NRGene



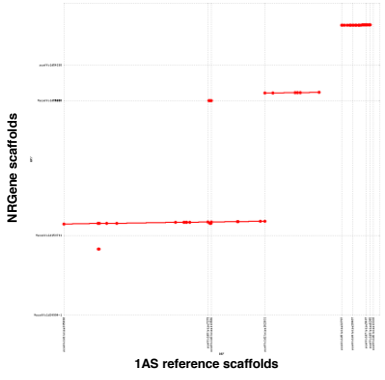
1AS reference vs NRGene



1AS reference vs NRGene



1AS reference vs NRGene



The road forward - Augment the NRGene assembly

- Identify/correct missassemblies and missing data using:
  - reference assemblies
  - optical maps
  - WGP-based physical maps
  - HiC maps
  - genetic maps
  - mate pairs
- Focus further BAC sequencing on low confidence or missing regions

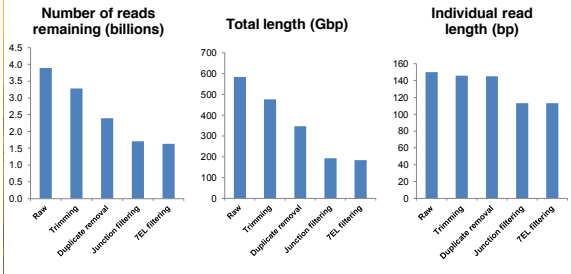
The Canadian Wheat Alliance

Acknowledgements

Andrew Sharpe
Yifang Tang
Kevin Koh
Janet Condie
Carling Clarke
Dustin Cram
Larissa Ramsay
Chad Matsalla
Curtis Pozniak
Jennifer Ens
Krystalee Wiebe
Ron MacLachlan
Therese Ouellet
George Fedak
Jaroslav Doležel
Marie Kubalakova

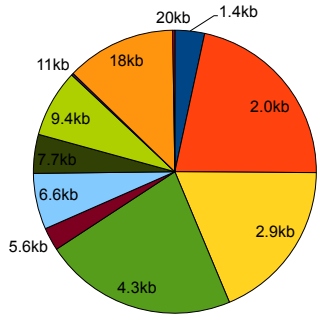
Government of Saskatchewan
University of Saskatchewan
Canada

Nextera library processing



	IWGSC version 3	Improvement
Sequence anchored by POPSEQ (Gbp)	7.1	2.7
% 90K markers anchored	97	1.6
% full-length ESTs (98% ident, 70% cov)	77	26
% high-confidence transcripts (99% ident, 90% cov)	94	1.7
% beta capture probes anchored	88	12

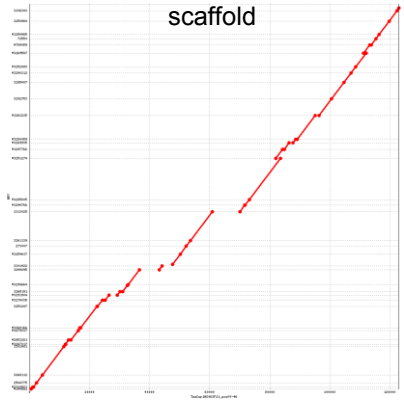
Contribution of each library to total # reads



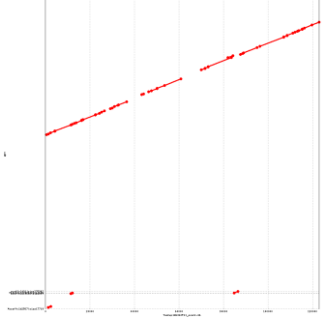
Keys points to mention about NRGene vs reference

- NRGene contig N50 is 51 kb, reference is 102kb and climbing
- Discrepancies are bound to happen but focusing on discrepancies is more productive than generating a new assembly that will inevitably still have problems

1AS survey contigs v. 1AS BAC sequence scaffold



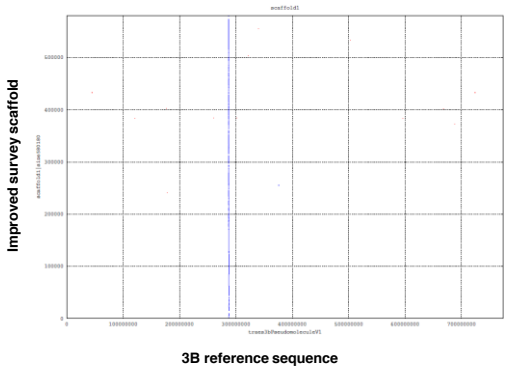
1AS survey scaffold v. 1AS BAC sequence scaffold



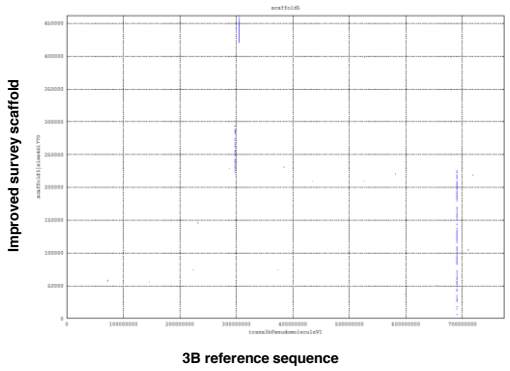
Marker alignment

Type	Source	Number of markers	WSS v1	WSS v3	Improvement
SNPs	Bristol	7,228	4,904	5,127	223
	90K	91,829	87,218	88,741	1,523
	9K	8,632	8,287	8,389	102
ESTs	NSF	12,185	10,429	10,941	512
	mapped wheat	2,926	2,399	2,678	279
	Sourdille	6,596	5,661	5,975	314
Other	beta exome capture design	107,969	81,659	94,852	13,193
	Dart-GBS	29,375	18,063	19,480	1,417
	Dart-public	2,000	1,346	1,419	73
	Dart-ver3	5,552	3,711	3,939	228

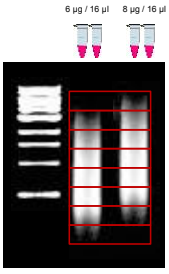
Comparison of improved survey scaffolds with 3B reference



Comparison of improved survey scaffolds with 3B reference



How to make many diverse Nextera libraries in parallel



- 4 reactions with 2 different ratios of input DNA to tagment enzyme
- FIGE gel electrophoresis to separate fragments
- Cut out bands combining both lanes of the same size
- Zymo gel extraction
- Combine/divide fractions as needed to provide appropriate input for circularization
- Minimize PCR cycles to preserve diversity

Wheat Genome assemblies – Chapman 2015


Genome Biology

METHOD Open Access

A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome

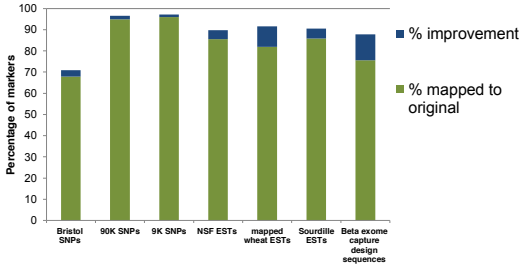
- Whole genome shotgun sequencing of hexaploid synthetic wheat line W7984
- 3 paired end libraries (250, 500 and 800 bp), 2 mate pair libraries (1kb and 4kb)
- Meraculous assembler, kmer 51
- 9.1 Gbp total sequence length
- 21.2 kb scaffold N50 length

Wheat WGS W7984 Scaffolding Summary

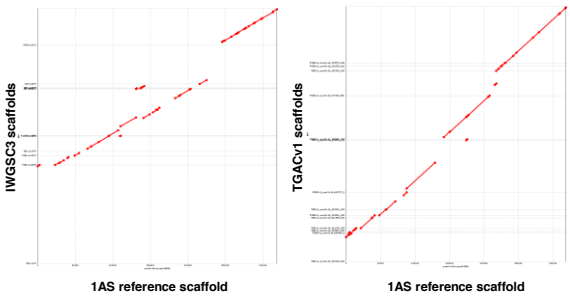
	Total Scaffold length (Gbp)	Scaffold # (millions)	Scaffold N50 length (kbp)	Largest Scaffold (kbp)
Mascher et al. 2015	8.2	0.96	24	267
With improved scaffolding	9.8	0.27	127	964

(500 bp cutoff for stats)

Marker Alignment



IWGSC3 vs TGACv1 nucmer plots



Overlay BAC reference sequence on NRGene assembly

Sequence BACs with WGP fingerprints matching poorly assembled NRGene scaffolds

Sequence BACs in region identified as low confidence based on optical mapping

Integrate optical mapping,

- Use reference assemblies, optical maps, WGP-based physical maps, HiC map, mate pairs and genetic data to identify/correct missassemblies and missing data
  - Mate pairs useful for defining scission point?
- Focus on integrating above resources and only sequences BACs in low confidence non-assembled regions