Towards a finished sequence for chromosome 7A: Building a high-quality pseudomolecule

Gabriel Keeble-Gagnere, Murdoch University

GRDC
Grains Research & Development Corporation

Bioplatforms Australia

Murdoch
UNIVERSITY

IWGSC
International Wheat Genome Sequencing Consortium

GYDLE

Acknowledgments

Funding
Grains Research Development Corporation
Bioplatforms Australia

ACCWI group
Rudi Appels, Hollie Webster, Shahidul Islam, Xueyan Chen, Yingjun Zhang, Johan Nystrom-Persson

Flow-sorting DNA/BAC library construction
Jaroslav Dolezel, Hana Simkova
Institute of Experimental Botany
Czech Republic

Fingerprinting BAC library
Mingcheng Luo group
UC Davis

Physical map assembly
Zeev Frenkel, Ambraham Korol
Haifa University

Genetic maps
MAGIC: Colin Cavanagh, Emma Huang, Jen Taylor (CSIRO)
MAGIC GBS: Matt Hayden (DEPI)
CSxRenan: Pierre Sourdille, Benoît Darrier (INRA)

***T. monoccocum* genetic map**
Population: Jorge Dubcovsky
90k chip: Matt Hayden, Kerrie Forrest

Manual assembly improvement/finishing, Gydle
Philippe Rigault

DNA sequencing
Matt Tinning
AGRF

Annotation
TriAnnot: Philippe Leroy, Aurelien Bernard (INRA)
geneID (CRG): Francisco Camara, Anna Vlasova (CRG, Spain), Juan Carlos Sanchez (ACPFG)
Storage proteins: Angela Juhasz (Hungary)
QTL mapping/Significant genome regions: Delphine Fleury (ACPFG)
Specific genes: Hui-xian Zhao (NW A&F Uni, China)

Pseudomolecule
Fred Choulet, Etienne Paux
INRA

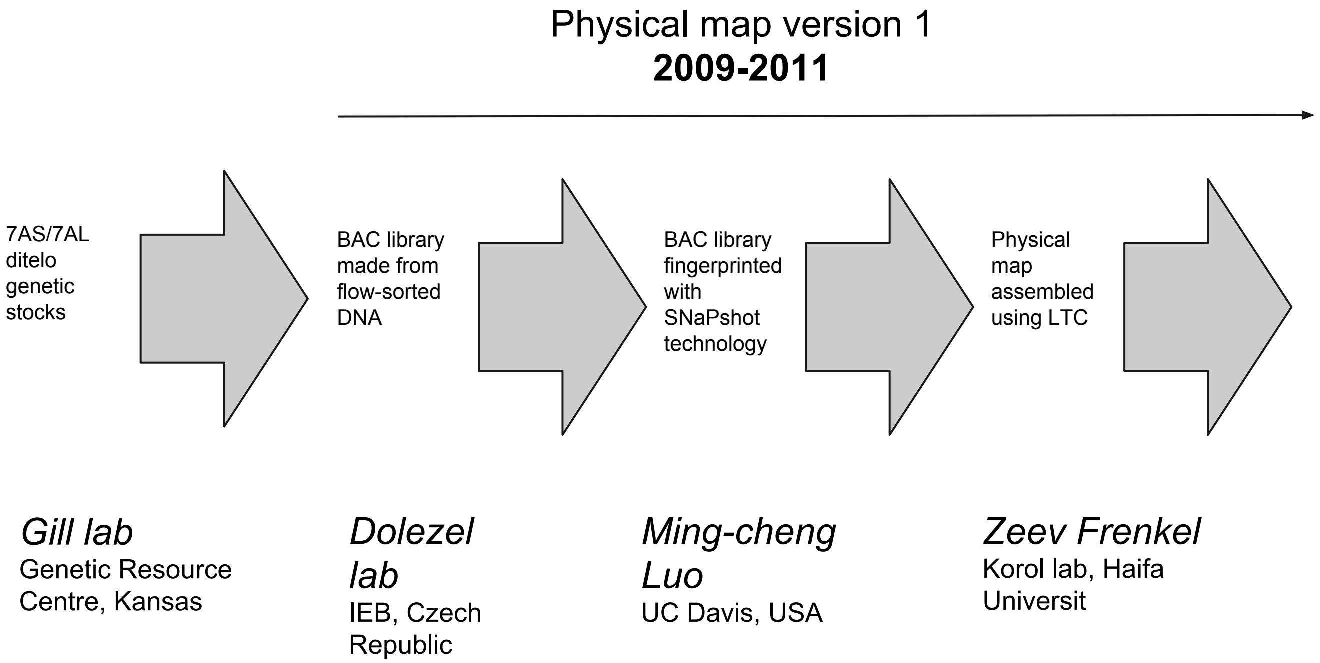
7A mate-pair sequencing of amplified DNA
Matt Hayden, Josquin Tibbits, Sami Hakim
DEPI

Whole-genome mate-pair data
Andy Sharpe, David Konkin, Curtis Pozniak
NRC, Canada

Bionano map
Jaroslav Dolezel, Hana Simkova, Mingcheng Luo

Supercomputing resources
iVEC/Pawsey Supercomputing Centre

Project timeline 2009-2011



Physical map version 1
2009-2011

7AS/7AL ditelo genetic stocks

BAC library made from flow-sorted DNA

BAC library fingerprinted with SNaPshot technology

Physical map assembled using LTC

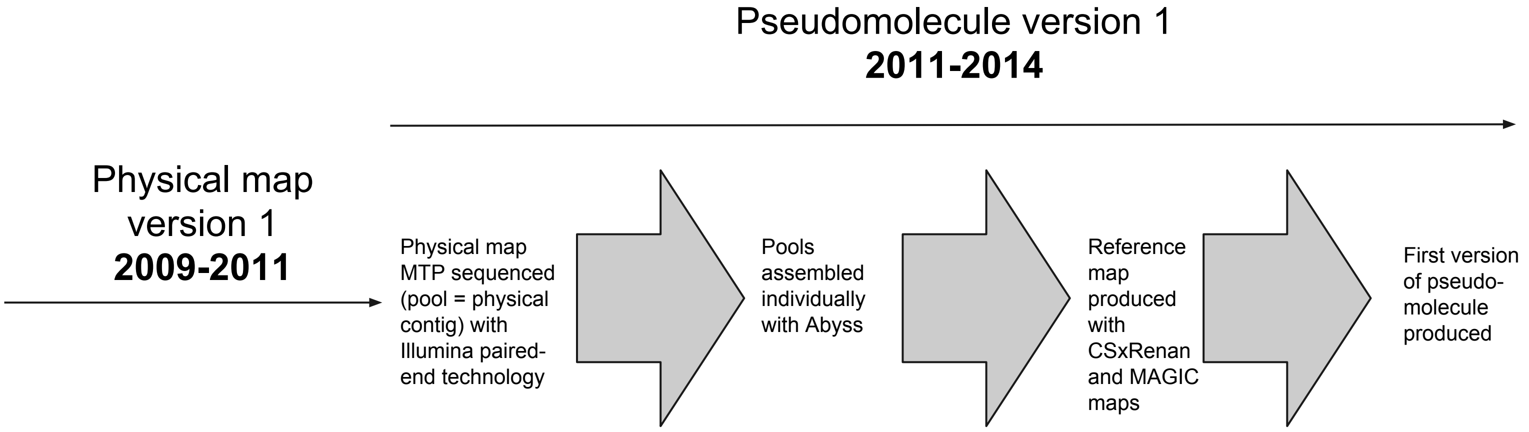
Gill lab
Genetic Resource Centre, Kansas

Dolezel lab
IEB, Czech Republic

Ming-cheng Luo
UC Davis, USA

Zeev Frenkel
Korol lab, Haifa Universit

Project timeline 2011-2014



Physical map
version 1
2009-2011

Pseudomolecule version 1
2011-2014

Physical map MTP sequenced (pool = physical contig) with Illumina paired-end technology

Pools assembled individually with Abyss

Reference map produced with CSxRenan and MAGIC maps

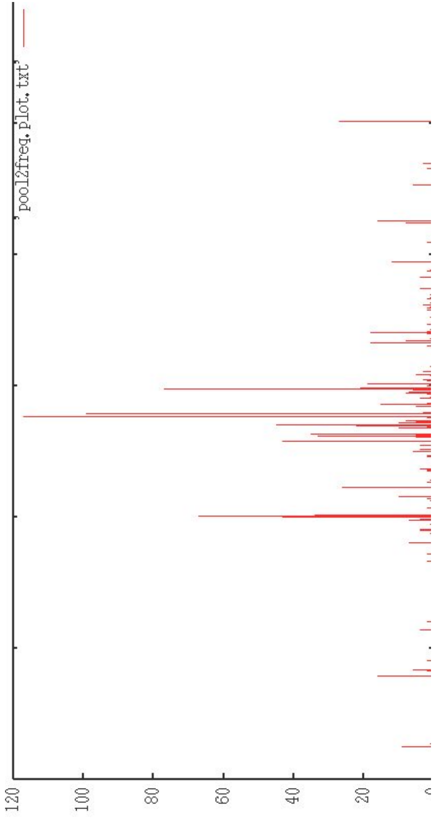
First version of pseudo-molecule produced

AGRF
Australia

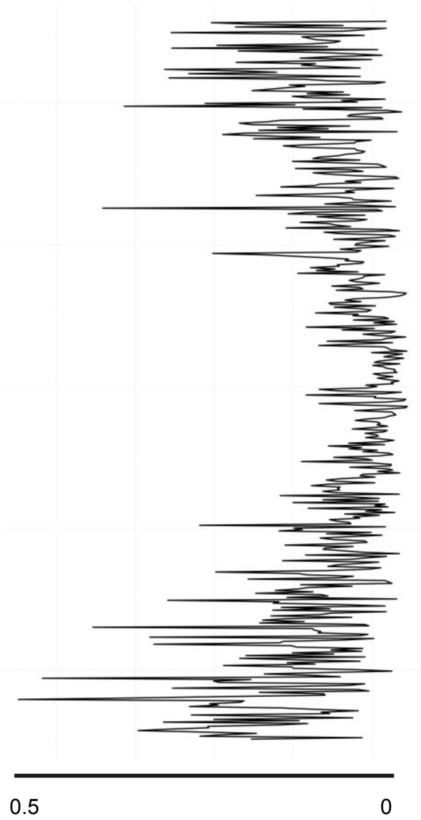
Fred Choulet,
Etienne Paux
INRA, France
Emma Huang,
Jen Taylor group
CSIRO, Australia

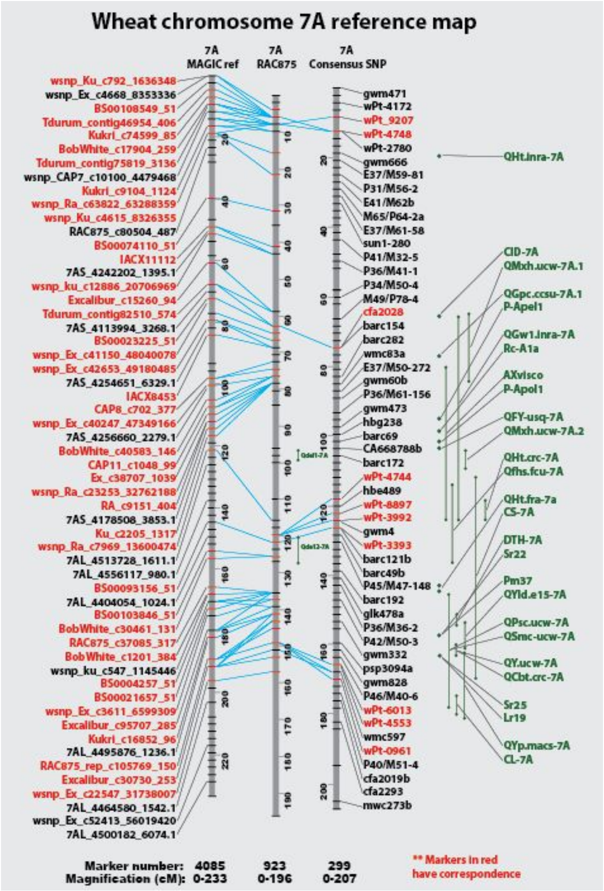
Pseudomolecule first version 2014

Centromere repeat frequency



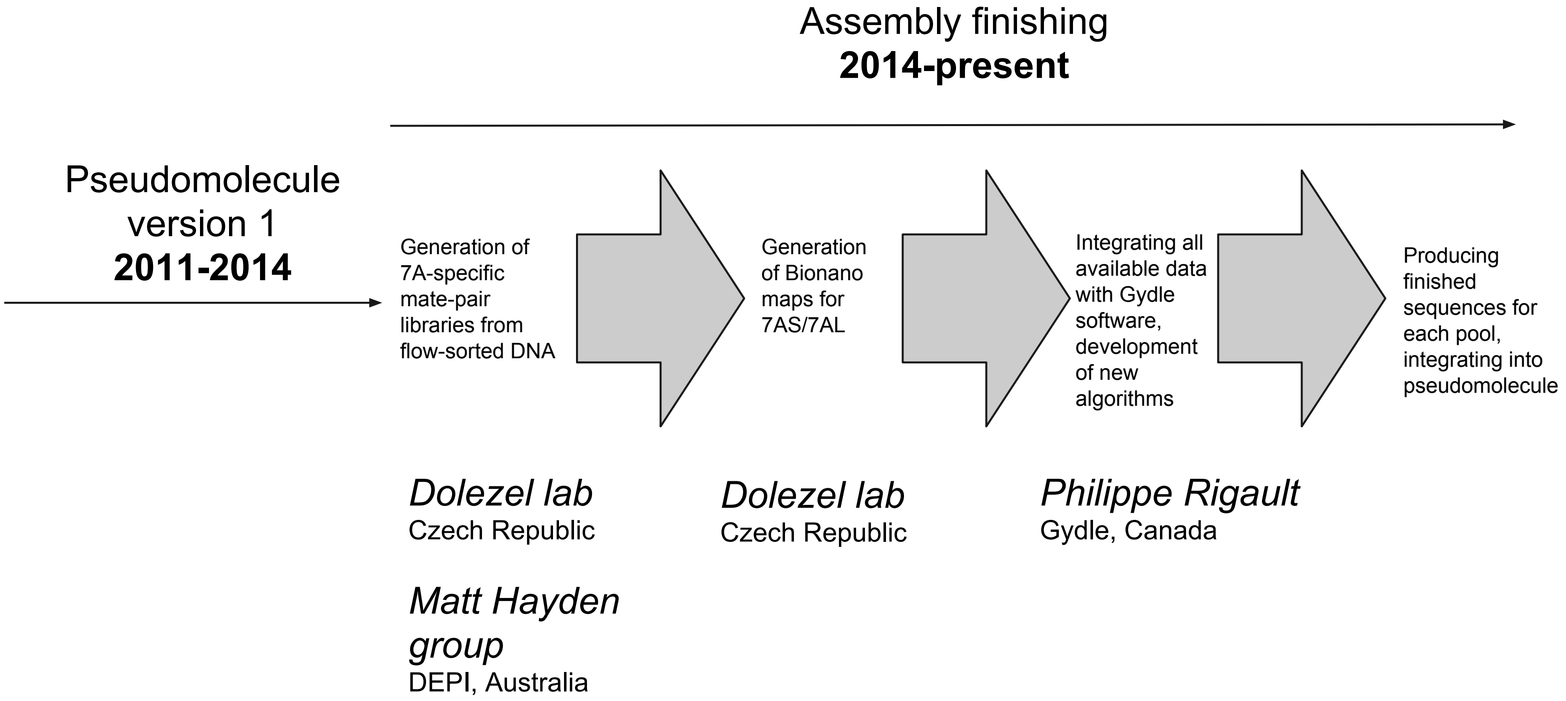
Gene density per 10kb (TriAnnot annotation)





Wheat chromosome 7A reference map

Project timeline 2014-2016



Assembly finishing
2014-present

Pseudomolecule version 1
2011-2014

Generation of 7A-specific mate-pair libraries from flow-sorted DNA

Generation of Bionano maps for 7AS/7AL

Integrating all available data with Gydle software, development of new algorithms

Producing finished sequences for each pool, integrating into pseudomolecule

Dolezel lab
Czech Republic

Matt Hayden group
DEPI, Australia

Dolezel lab
Czech Republic

Philippe Rigault
Gydle, Canada

Summary of raw data

BAC fingerprints for every BAC in physical library

Illumina PE sequencing of physical map MTP BACs in pools

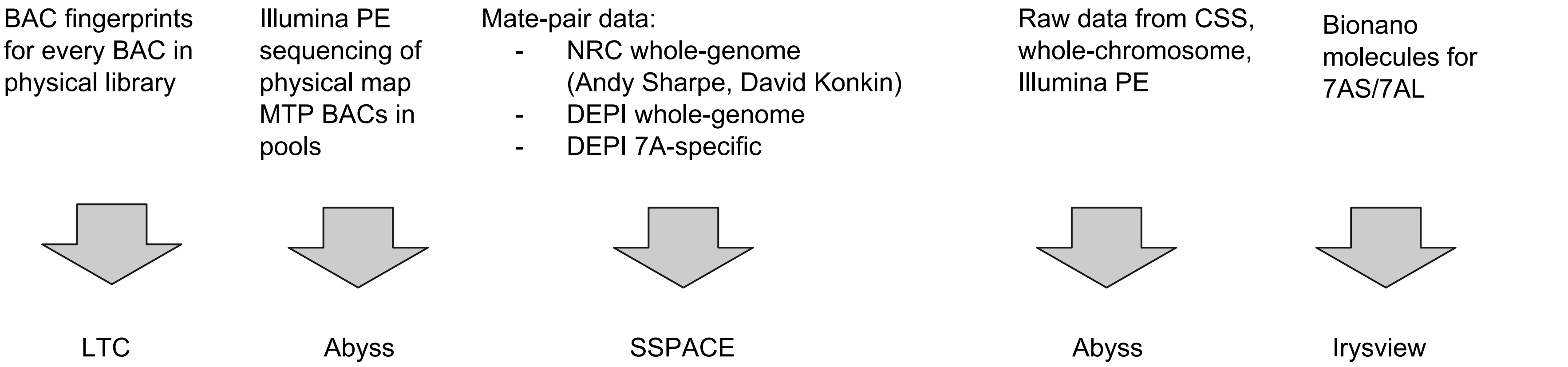
Mate-pair data:

- NRC whole-genome (Andy Sharpe, David Konkin)
- DEPI whole-genome
- DEPI 7A-specific

Raw data from CSS, whole-chromosome, Illumina PE

Bionano molecules for 7AS/7AL

Summary of raw data



BAC fingerprints for every BAC in physical library

LTC

Illumina PE sequencing of physical map MTP BACs in pools

Abyss

Mate-pair data:

- NRC whole-genome (Andy Sharpe, David Konkin)
- DEPI whole-genome
- DEPI 7A-specific

SSPACE

Raw data from CSS, whole-chromosome, Illumina PE

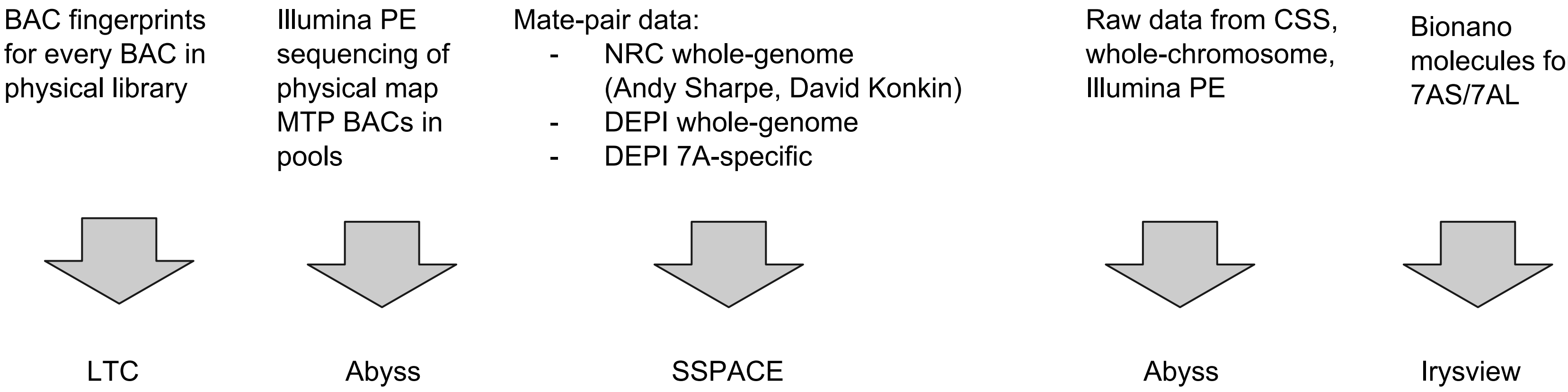
Abyss

Bionano molecules for 7AS/7AL

Irysview

Traditionally, use different tools to exploit each dataset

Summary of raw data



BAC fingerprints for every BAC in physical library

LTC

Illumina PE sequencing of physical map MTP BACs in pools

Abyss

Mate-pair data:

- NRC whole-genome (Andy Sharpe, David Konkin)
- DEPI whole-genome
- DEPI 7A-specific

SSPACE

Raw data from CSS, whole-chromosome, Illumina PE

Abyss

Bionano molecules for 7AS/7AL

Irysview

The difficulty is in integrating all available data together in a *consistent* way that is cross-validated against each data source

Summary of raw data

BAC fingerprints for every BAC in physical library

Illumina PE sequencing of physical map MTP BACs in pools

Mate-pair data:

- NRC whole-genome (Andy Sharpe, David Konkin)
- DEPI whole-genome
- DEPI 7A-specific

Raw data from CSS, whole-chromosome, Illumina PE

Bionano molecules for 7AS/7AL

With Gydle, we are able to integrate all available data simultaneously to produce high-quality sequence.

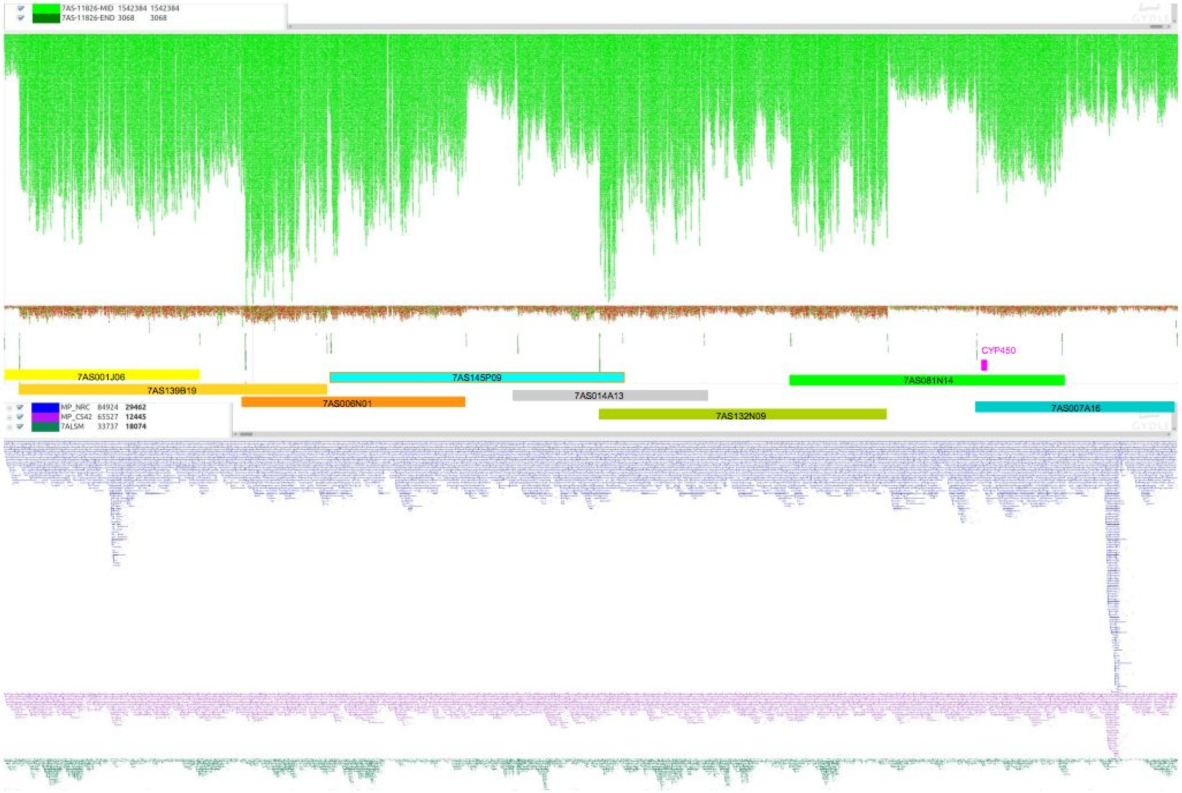
Finished, cross-validated sequence

We consider a given sequence *finished* when:

1. We have a single contig
2. Paired-end and mate-pair data is consistent across the entire sequence
3. The physical map BACs can be precisely ordered along the contig
4. The sequence aligns to Bionano consistently

In other words, the sequence is cross-validated by the raw sequence data, the physical map, and Bionano.

Finished sequence (7AS-11826)

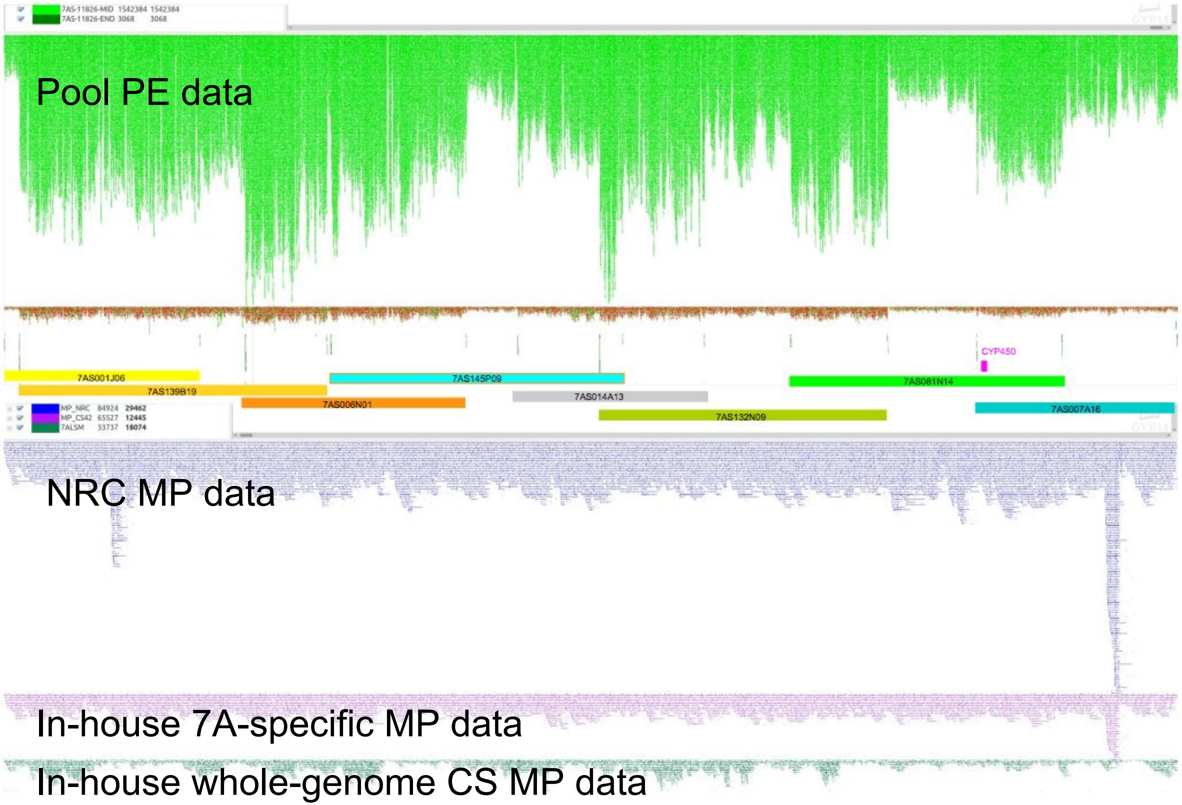


☑	🟩 7AS-11826-MID	1542384	1542384
☑	🟩 7AS-11826-END	3068	3068

☑	🟦 MP_NRC	84924	**29462**
☑	🟪 MP_CS42	65527	**12445**
☑	🟩 7ALSM	33737	**18074**

1. Single contig

Finished sequence (7AS-11826)



Pool PE data

NRC MP data

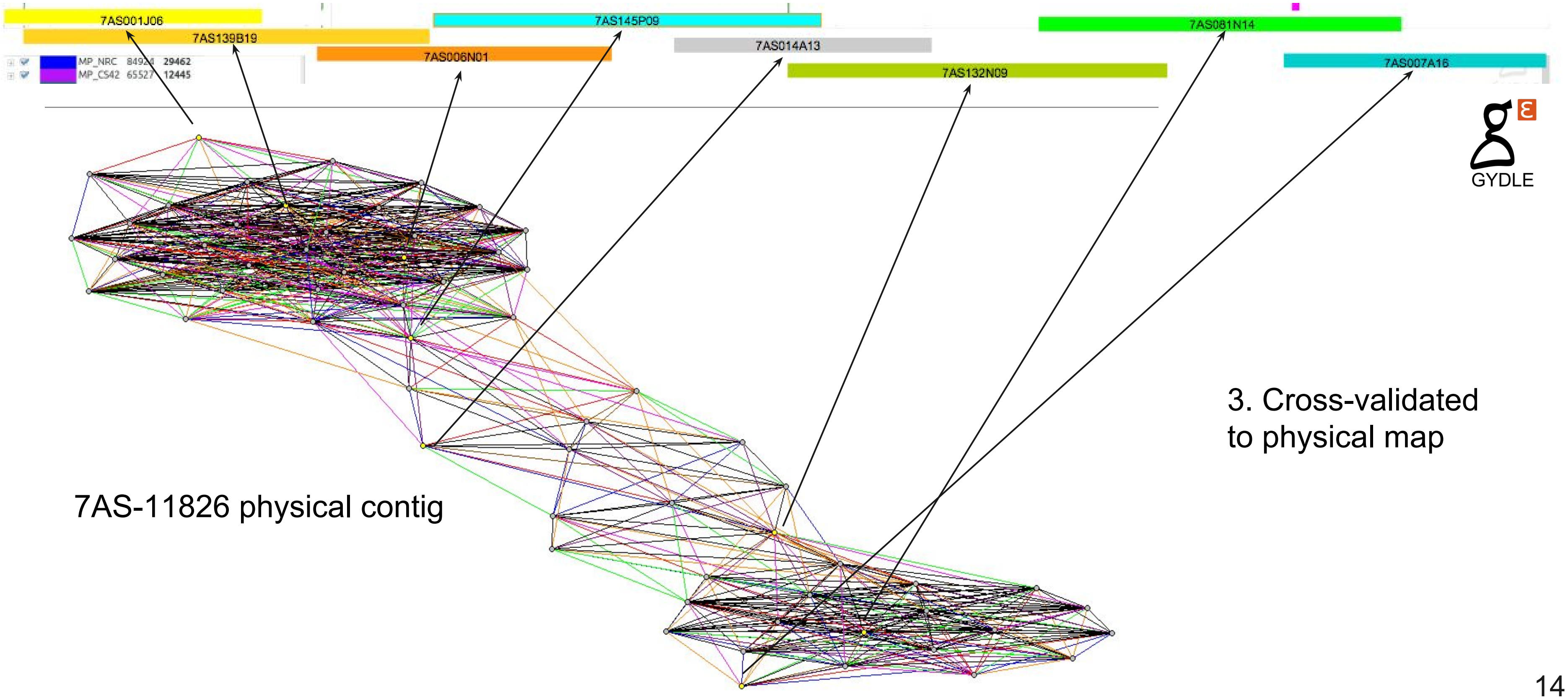
In-house 7A-specific MP data

In-house whole-genome CS MP data

2. Raw data is consistent

GYDLE

Finished sequence (7AS-11826)

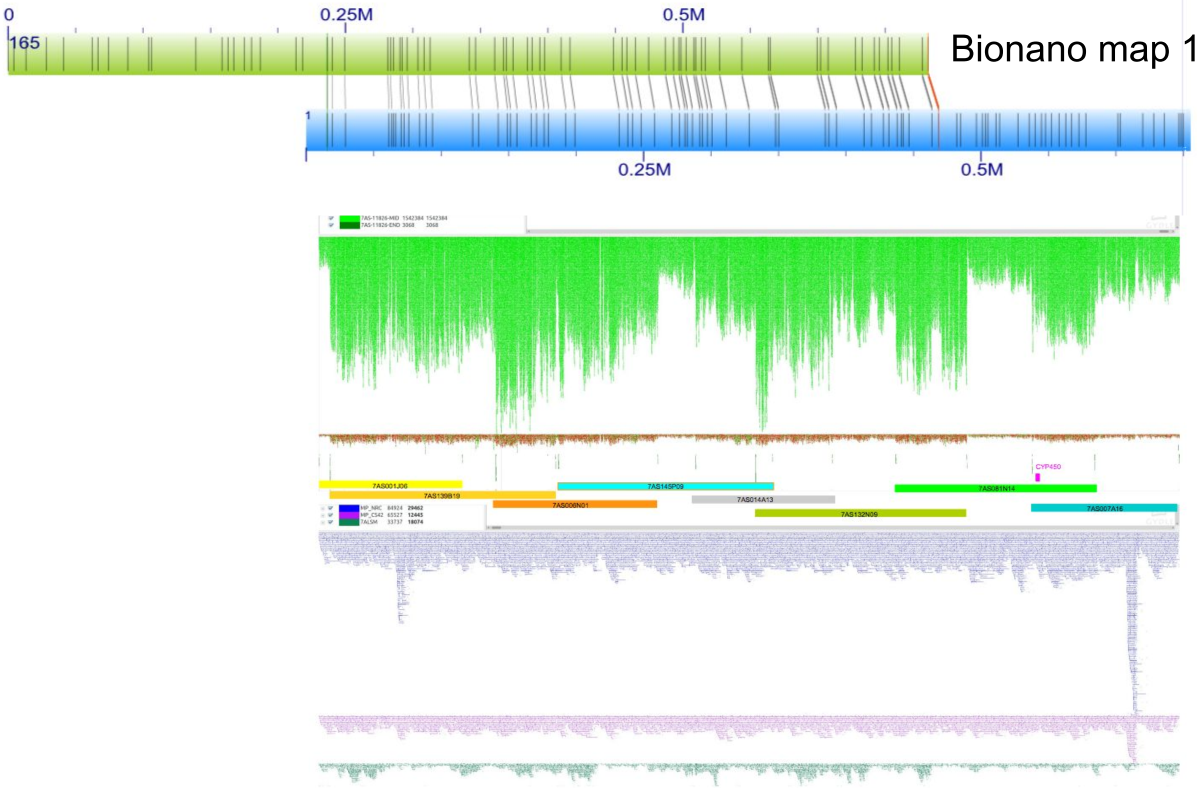


7AS-11826 physical contig

3. Cross-validated
to physical map

GYDLE

Finished sequence (7AS-11826)



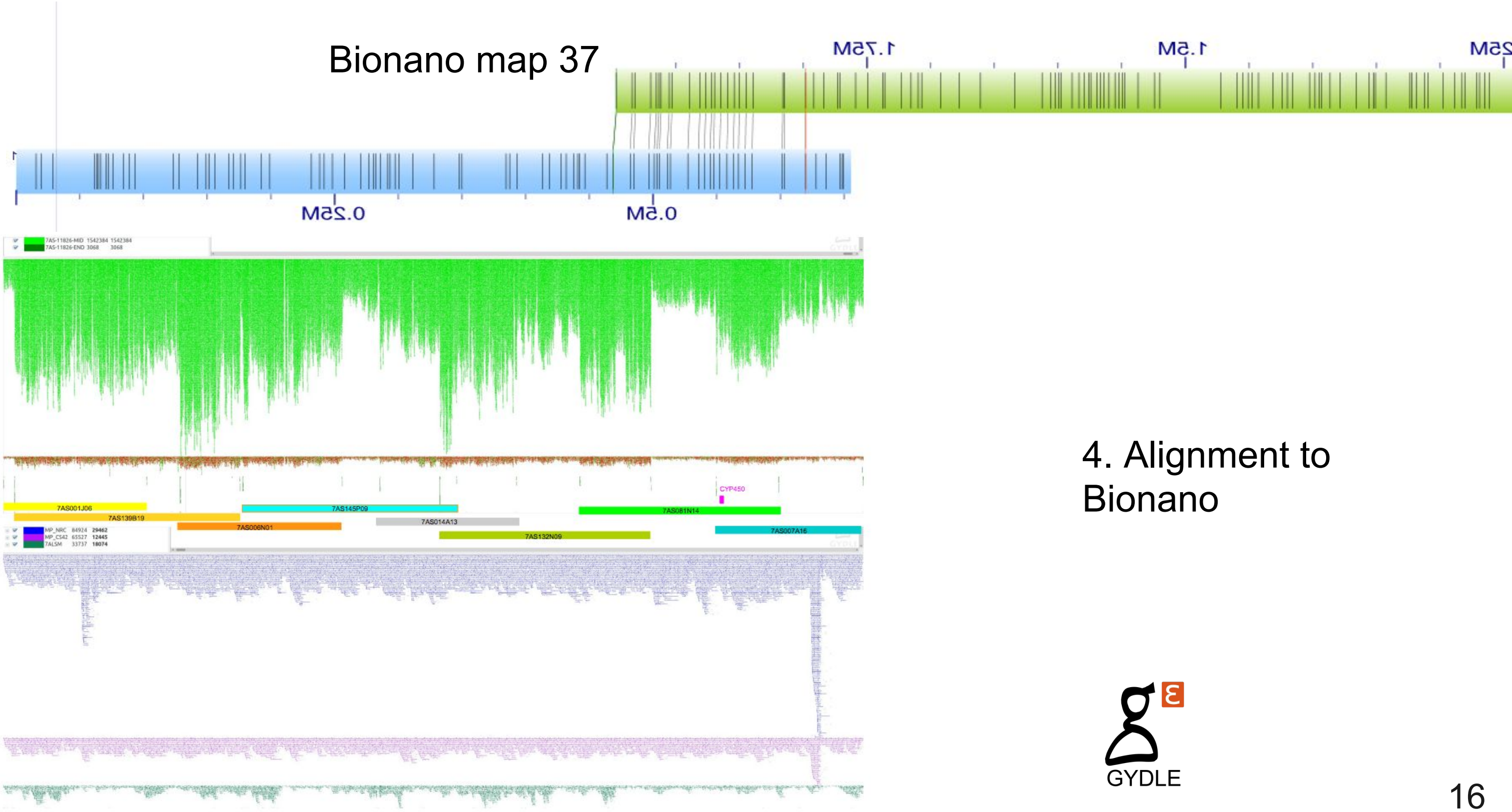
Bionano map 165

7AS-11826 sequence

4. Alignment to Bionano

GYDLE

Finished sequence (7AS-11826)



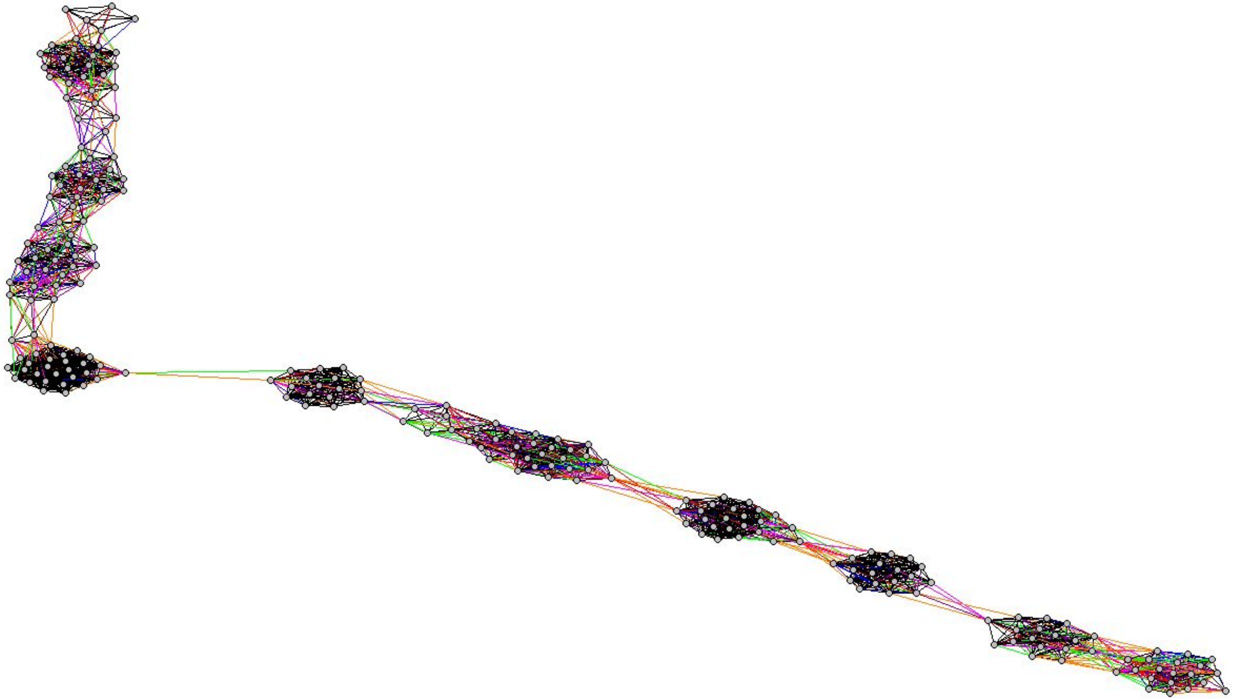
Bionano map 37

7AS-11826 sequence

4. Alignment to Bionano

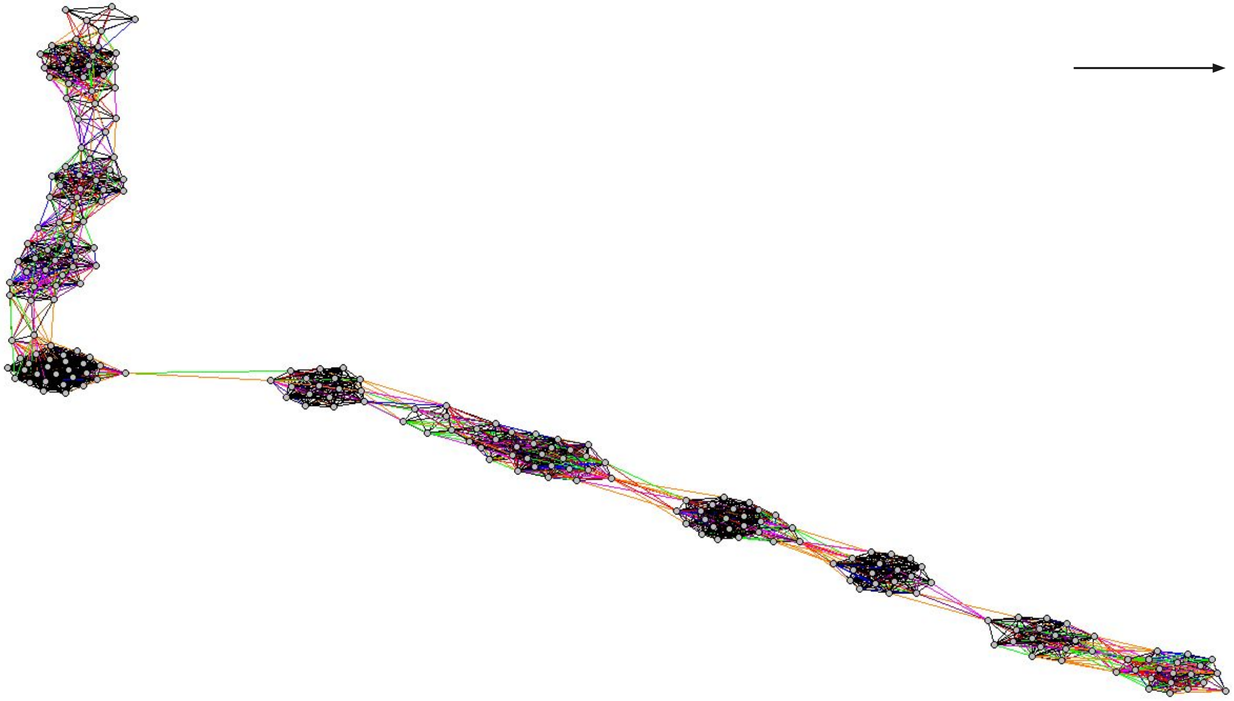
A case study: 7AS-11582 physical contig

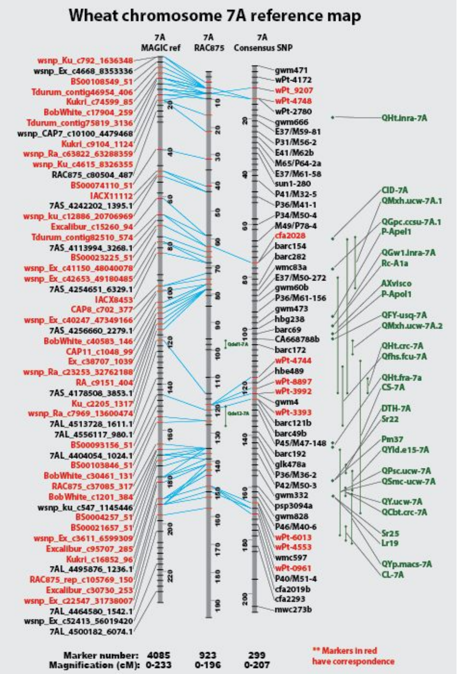
- 2Mb physical contig containing 224 clones (29 in MTP)



A case study: 7AS-11582 physical contig

- 2Mb physical contig containing 224 clones (29 in MTP)





Target genes

Before sequencing began, we had a list of target genes of interest that were known to be on chromosome 7A.

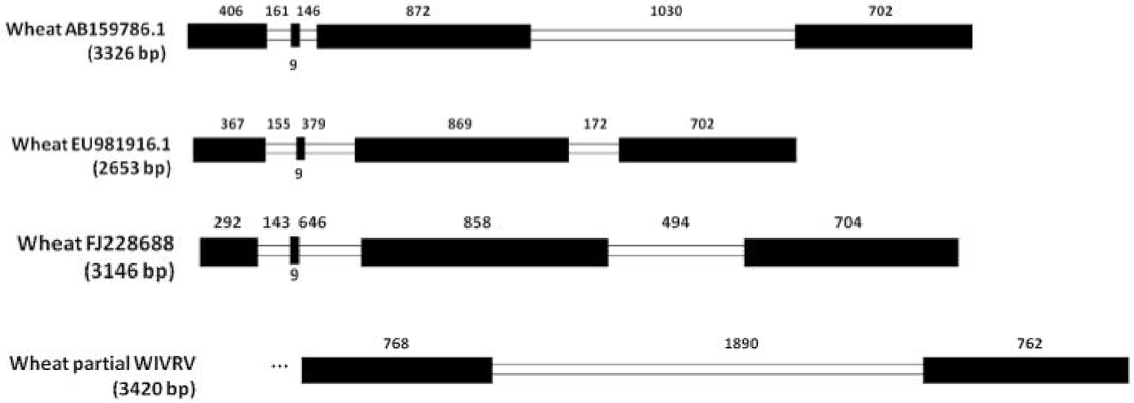
This included a set of fructan biosynthesis genes reported on in 2012 by ACPFG (Huynh et al., Plant Mol Biol, 2012):

1-SST

1-FFT

6-SFT

WIVRV

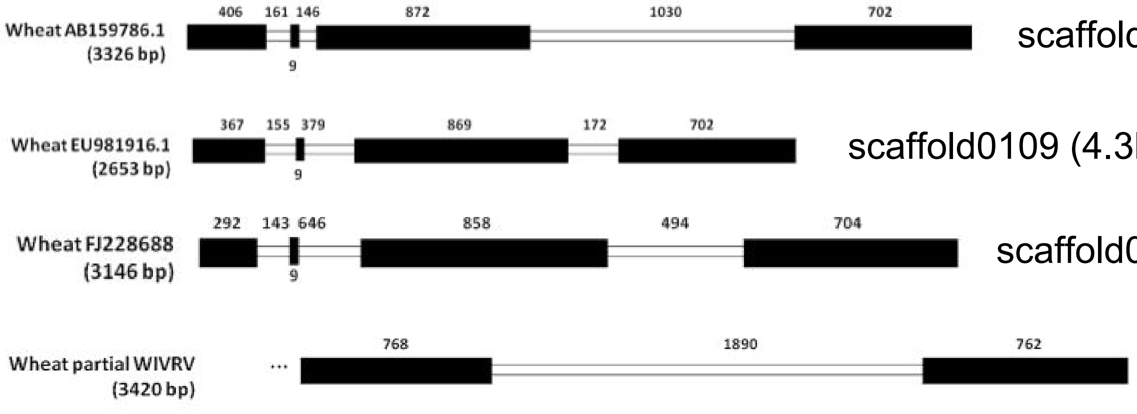


Genes appear in first sequencing batch

Stats of Abyss assembly of paired-end sequencing of BAC pool:

- 273 scaffolds
- 25.6kb N50
- Total length 2.42Mb

Four target genes on four separate scaffolds:



1-SST scaffold0260 (104kb)

1-FFT scaffold0109 (4.3kb)

6-SFT scaffold0150 (13.8kb)

WIVRV scaffold0077 (43kb)

Some genes appear in another pool

As the sequencing of BAC pools progressed, we found another pool (7AS-11832) also contained this set of genes.

- Initially, thought this might be a real perfect copy, but:
  - Copies of the genes were identical
  - When we looked for divergent sequence between the two “regions” (in order to establish they are distinct copies), we found that the surrounding sequence context for the genes was also identical

When we looked at perfect duplications across the assembly, we found:

- 76Mb perfectly duplicated sequences on 7AL
- 45Mb perfectly duplicated sequences on 7AS

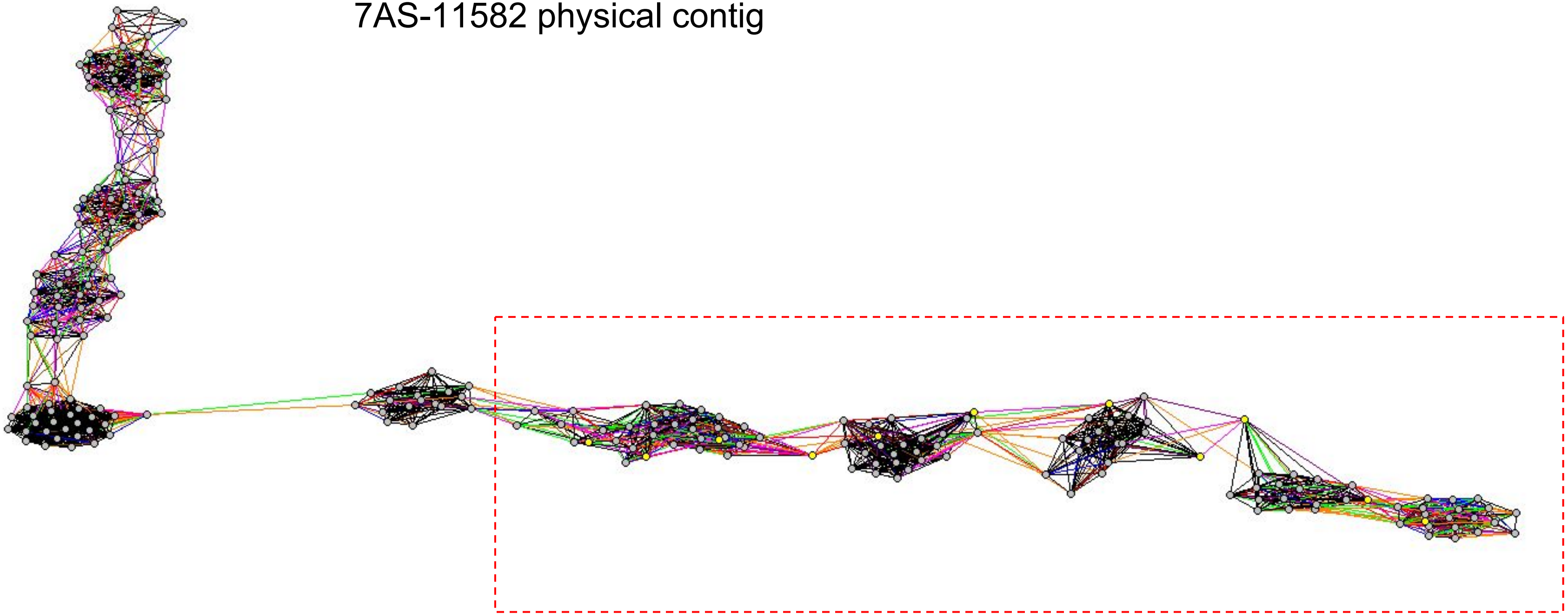
This led us to suspect there may be a contamination problem.

Contamination from 7AS-11582

- No contaminating BACs in 7AS-11582
- Evidence that BACs from 7AS-11582 contaminate 2 other pools

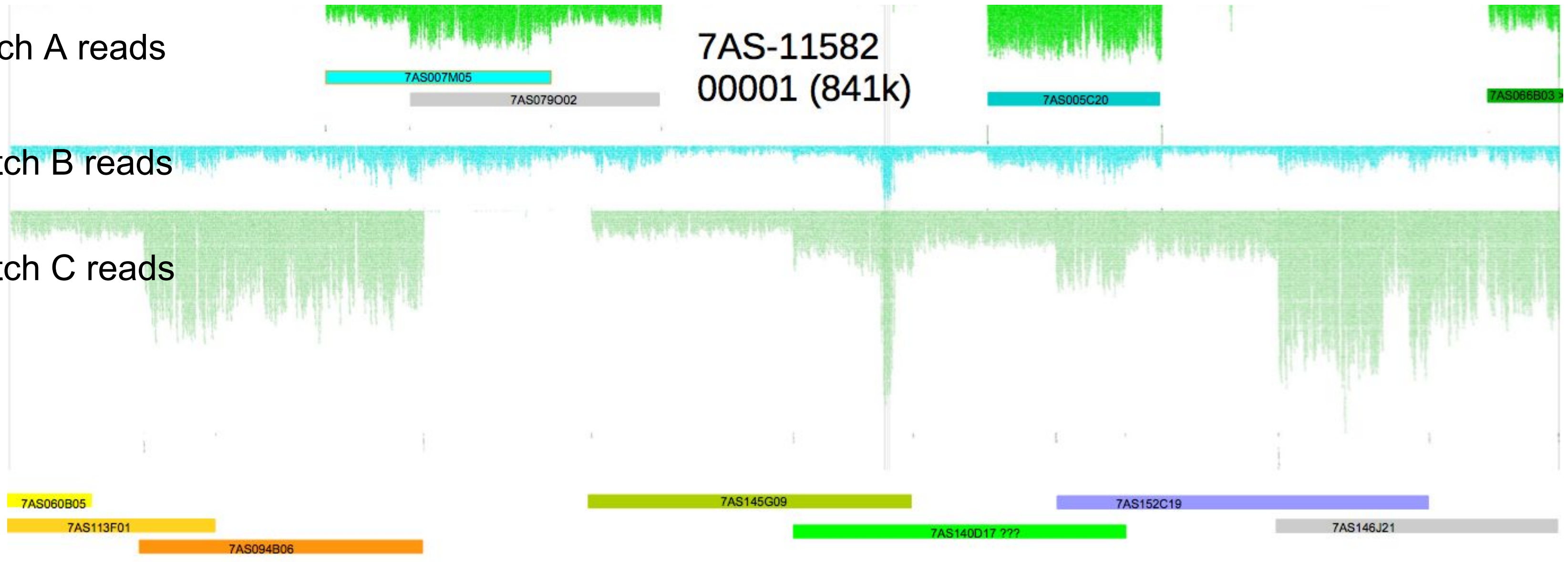
Linking sequence to physical map

7AS-11582 physical contig



* Physical map networks from LTC (Frenkel et al.)

Linking sequence to physical map

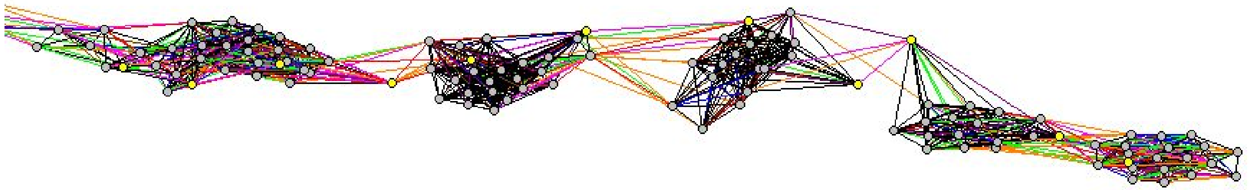


Batch A reads

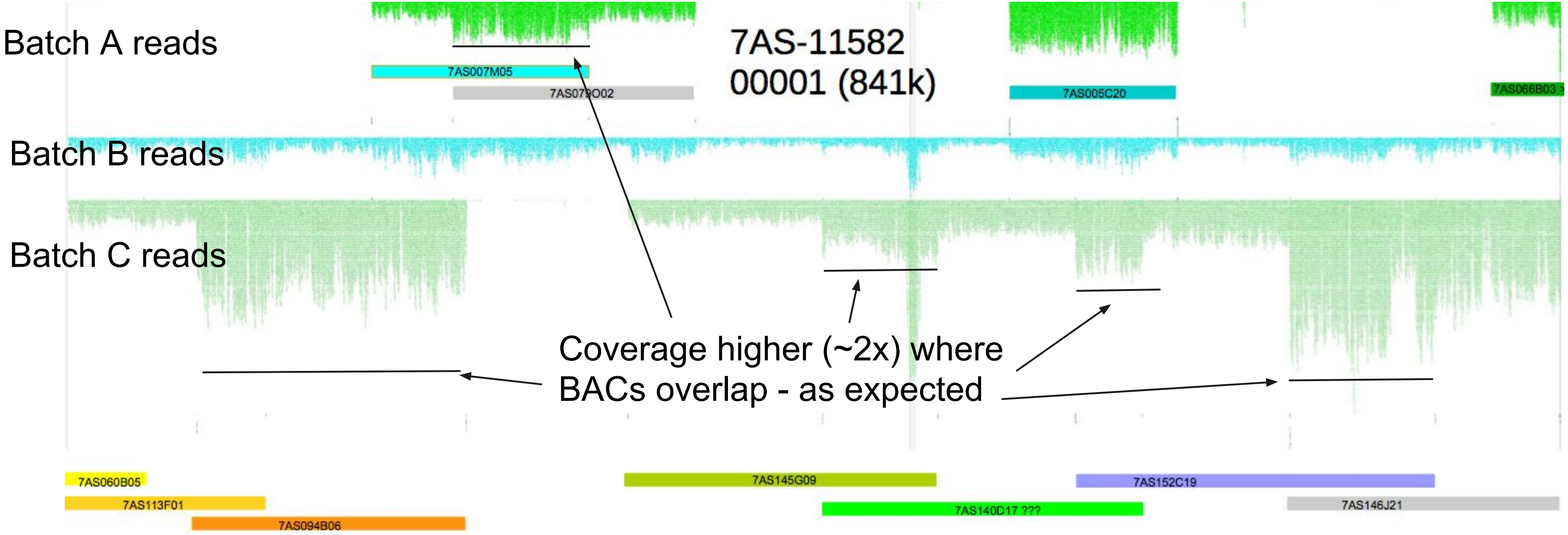
Batch B reads

Batch C reads

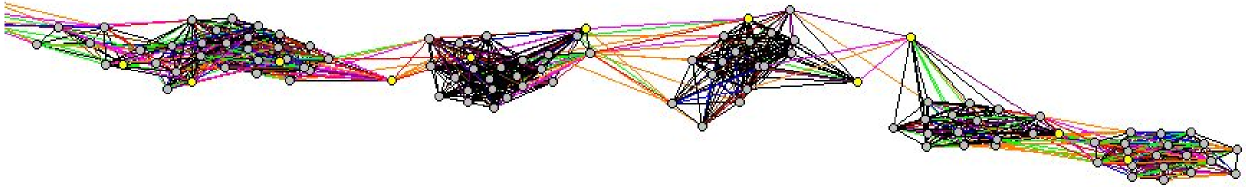
7AS-11582
00001 (841k)



Linking sequence to physical map

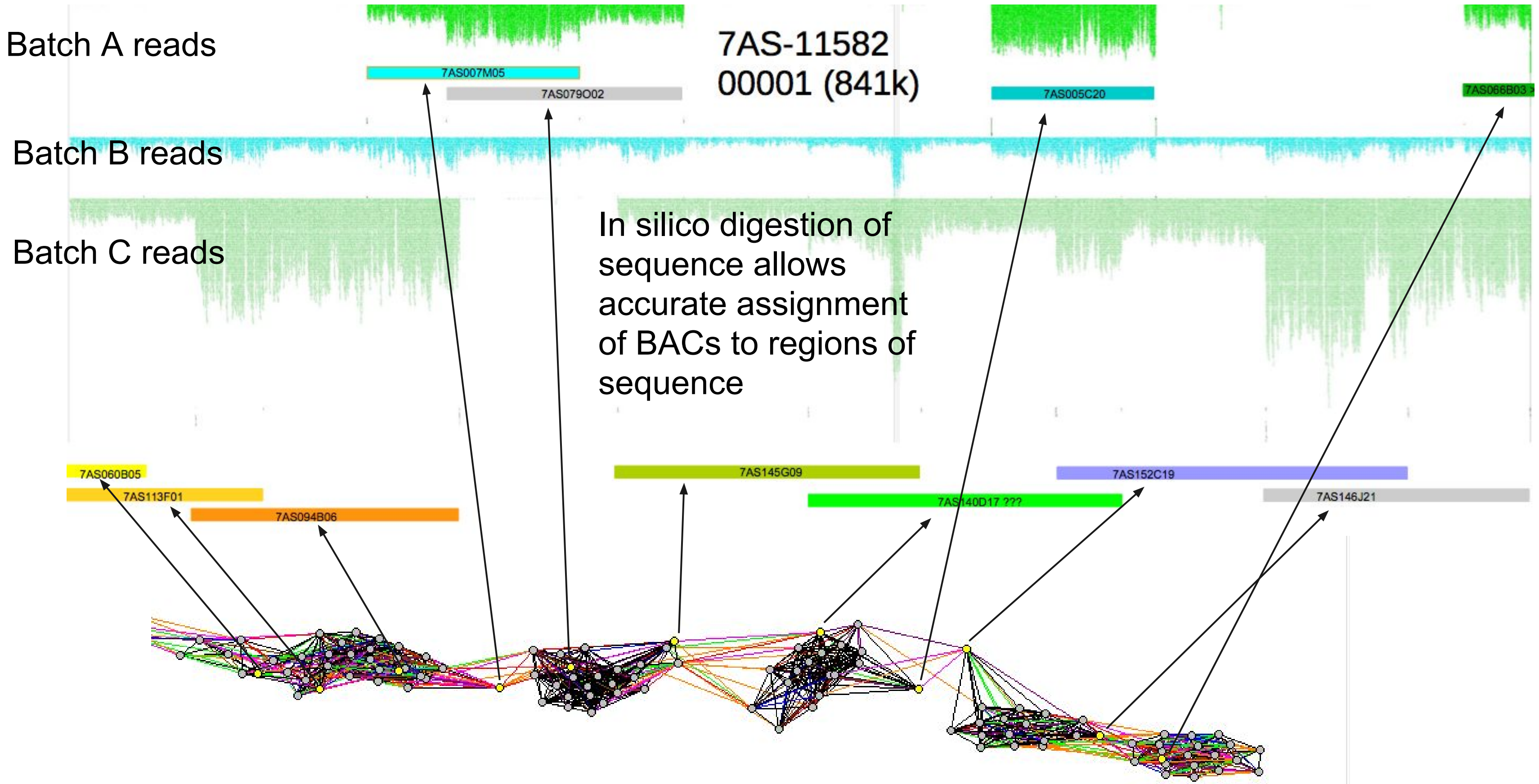


Coverage higher (~2x) where BACs overlap - as expected



GYDLE

Linking sequence to physical map



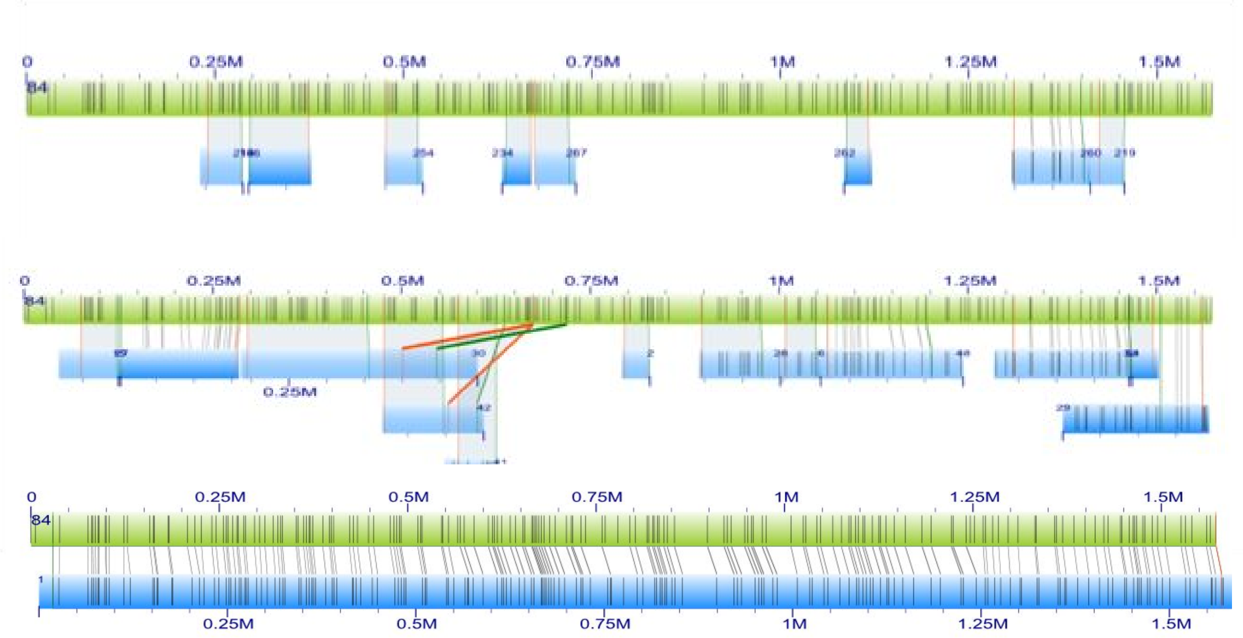
7AS-11582 00001 (841k)

In silico digestion of sequence allows accurate assignment of BACs to regions of sequence

GYDLE

Bionano

Bionano 7AS map 84 shown in green;
7AS-11582 scaffolds in blue.



Initial pool assembly from paired-end reads only (Abyss)

↓

After mate-pair data, before Bionano

↓

Finished sequence

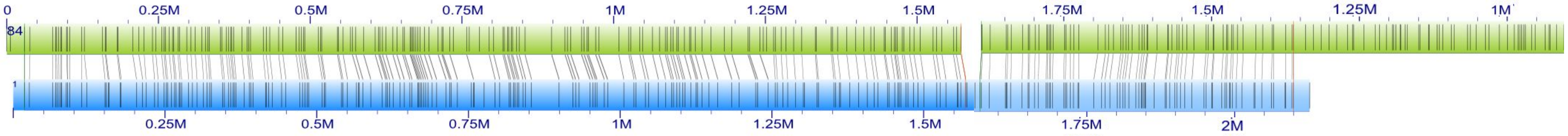
* Alignments in IrisView

Bionano for 7AS-11582

Finished sequence bridges Bionano maps

Bionano map 84

Bionano map 49



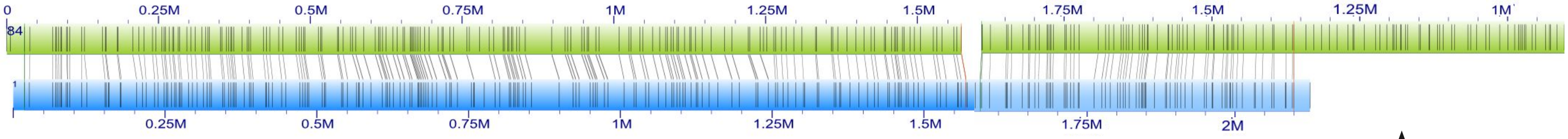
7AS-11582 finished sequence

Bionano for 7AS-11582

Finished sequence bridges Bionano maps

Bionano map 84

Bionano map 49



7AS-11582 finished sequence

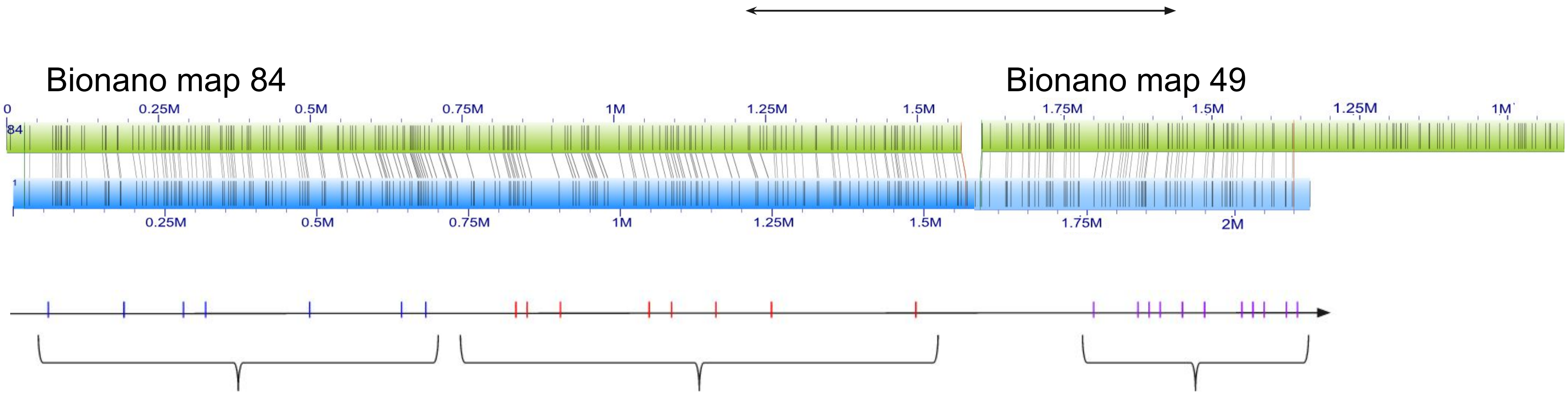
Bionano will identify adjacent pool

Genes clusters on 7AS-11582

Finished sequence bridges Bionano maps

Bionano map 84

Bionano map 49



Cluster of 7 NB-ARC domain genes

Cluster of 8 sugar metabolism genes including the 1-FFT, 1-SST, 6-SFT, WIRVR genes mentioned earlier

Cluster of more than 10 protein-kinase domain-containing genes and LRR-repeat receptor containing genes

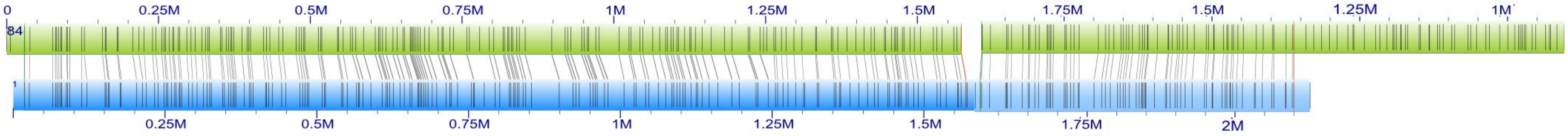
Based on TriAnnot (Philippe Leroy, INRA) annotation - needs manual curation.

Bionano for 7AS-11582

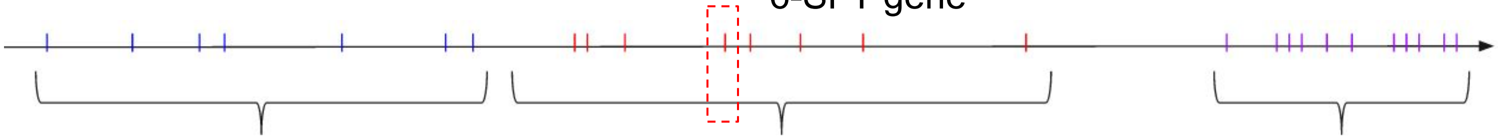
Finished sequence bridges Bionano maps

Bionano map 84

Bionano map 49



6-SFT gene



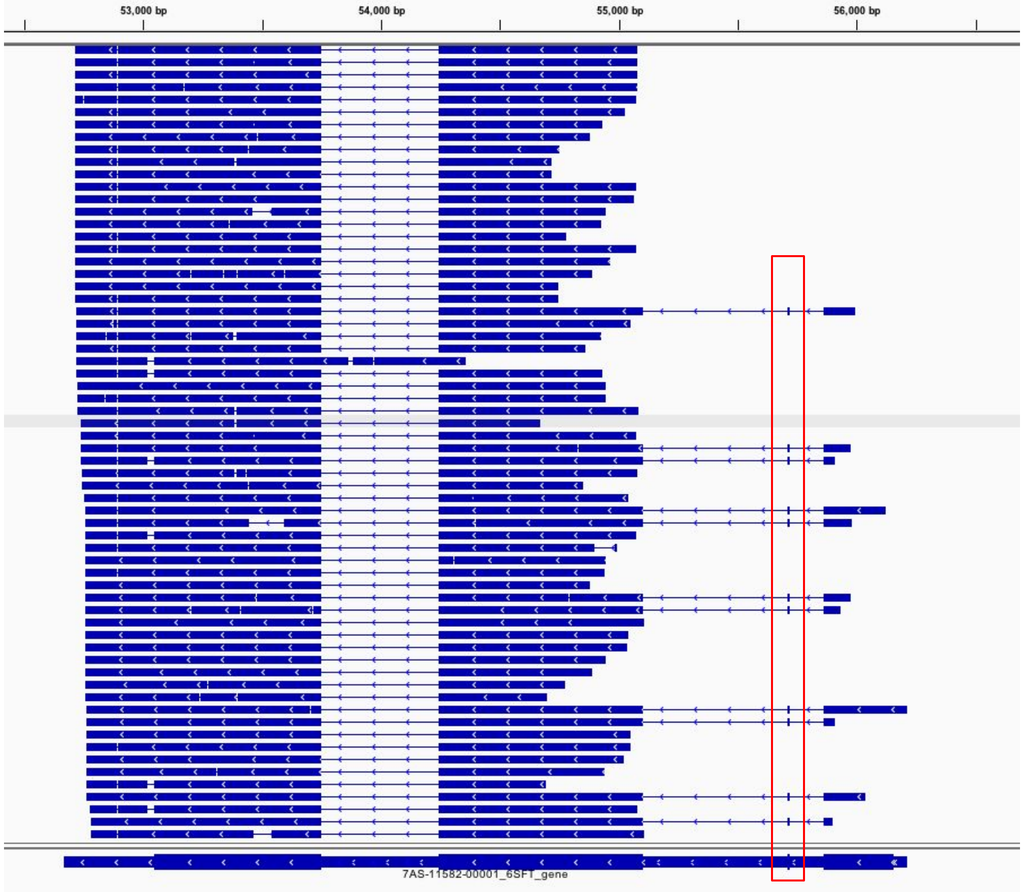
Cluster of 7 NB-ARC domain genes

Cluster of 8 sugar metabolism genes including the 1-FFT, 1-SST, 6-SFT, WIRVR genes mentioned earlier

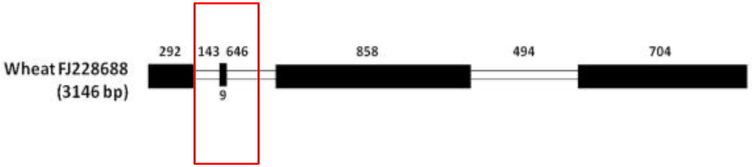
Cluster of more than 10 protein-kinase domain-containing genes and LRR-repeat receptor containing genes

Based on TriAnnot (Philippe Leroy, INRA) annotation - needs manual curation.

Manual curation of 6-SFT gene



Raw Pacbio FLNC RNA reads from *Dong et al. 2015* (kindly provided in advance of publication)



* alignments in IGV

Next steps

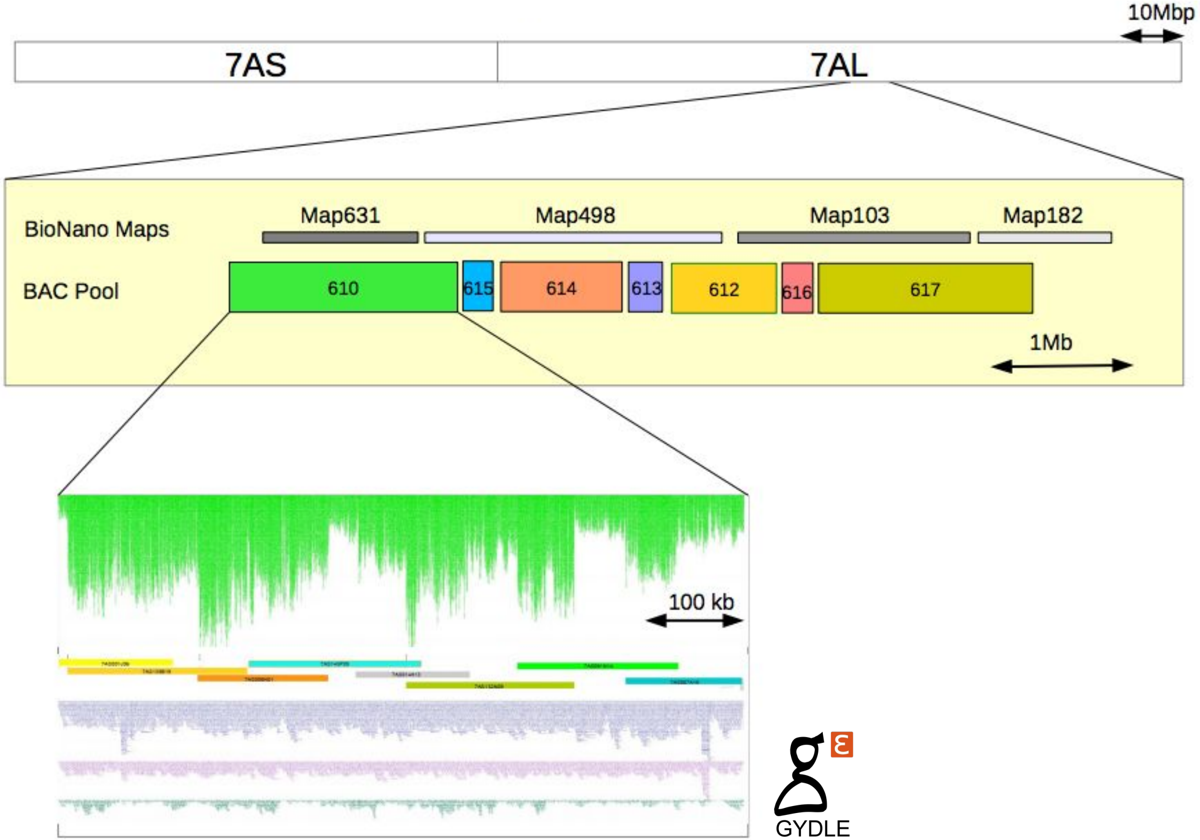
We are producing finished sequences for all physical contigs.

This will result in around 732 finished “pools”.

Challenge is then to fill in space between each pool. For this, we will have:

- Keygene tags for all MTP BACs plus ~1100 BACs that were not sequenced but which we think are between pools (based on scaffolded physical map)
- CSS PE reads as well as our own 7A-specific MP data covering the intra-pool space
- Bionano maps to assist in joining pools
- NRgene assembly of Chinese Spring 7A (completed December 2015) will provide an advanced reference to assist in validating and extending our assembly through regions not covered by BACs

Towards a finished sequence



GYDLE

Acknowledgments

Funding
Grains Research Development Corporation
Bioplatforms Australia

ACCWI group
Rudi Appels, Hollie Webster, Shahidul Islam, Xueyan Chen, Yingjun Zhang, Johan Nystrom-Persson

Flow-sorting DNA/BAC library construction
Jaroslav Dolezel, Hana Simkova
Institute of Experimental Botany
Czech Republic

Fingerprinting BAC library
Mingcheng Luo group
UC Davis

Physical map assembly
Zeev Frenkel, Ambraham Korol
Haifa University

Genetic maps
MAGIC: Colin Cavanagh, Emma Huang, Jen Taylor (CSIRO)
MAGIC GBS: Matt Hayden (DEPI)
CSxRenan: Pierre Sourdille, Benoît Darrier (INRA)

***T. monoccocum* genetic map**
Population: Jorge Dubcovsky
90k chip: Matt Hayden, Kerrie Forrest

Manual assembly improvement/finishing, Gydle
Philippe Rigault

DNA sequencing
Matt Tinning
AGRF

Annotation
TriAnnot: Philippe Leroy, Aurelien Bernard (INRA)
geneID (CRG): Francisco Camara, Anna Vlasova (CRG, Spain), Juan Carlos Sanchez (ACPFG)
Storage proteins: Angela Juhasz (Hungary)
QTL mapping/Significant genome regions: Delphine Fleury (ACPFG)
Specific genes: Hui-xian Zhao (NW A&F Uni, China)

Pseudomolecule
Fred Choulet, Etienne Paux
INRA

7A mate-pair sequencing of amplified DNA
Matt Hayden, Josquin Tibbits, Sami Hakim
DEPI

Whole-genome mate-pair data
Andy Sharpe, David Konkin, Curtis Pozniak
NRC, Canada

Bionano map
Jaroslav Dolezel, Hana Simkova, Mingcheng Luo

Supercomputing resources
iVEC/Pawsey Supercomputing Centre