TGAC
The Genome Analysis Centre™

BBSRC EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium

Bread Wheat Chromosome-based Survey Sequencing Initiative Update

Jane Rogers

IWGSC Workshop
International Triticeae Mapping Initiative - 2012

Sequencing Survey Initiative

TGAC
The Genome Analysis Centre™

Institute of Experimental Botany of the AS CR, v. v. i.

URGI

Graminor

HelmholtzZentrum münchen
German Research Center for Environmental Health

Amplified Sorted DNA (IEB)

~50X Survey sequence of all individual chromosomes

Assembly of gene catalog for each chromosome /arm (TGAC)

Comparative – "Genome Zipper" (MIPS)

Virtual Gene Order of the 21 Bread Wheat Chromosomes

INRA
Institut National de la Recherche Agronomique

NIAS
National Institute of Agrobiological Sciences

Department of Biotechnology
Ministry of Science & Technology, Government of India

The Research Council of Norway

NRC·CNRC Canada

biogemma

Together, we are
KANSAS WHEAT

The University Of Queensland
AUSTRALIA

FranceAgriMer

Bioplatforms Australia

ANR
Agence Nationale de la Recherche

ICARDA

BBSRC
bioscience for the future

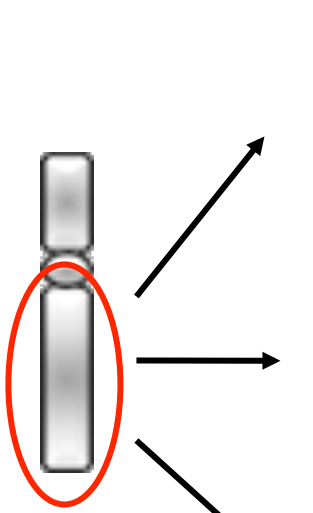
Australian Government
Department of Industry, Innovation, Science, Research and Tertiary Education

Australian Government
Australian Research Council

USDA NIFA
United States Department of Agriculture
National Institute of Food and Agriculture

IWGSC

Project Overview



Flow sorted
amplified DNA
Jaroslav Dolezel

DCC (TGAC)
wheatdcc.tgac.bbsrc.ac.uk

Quality checking

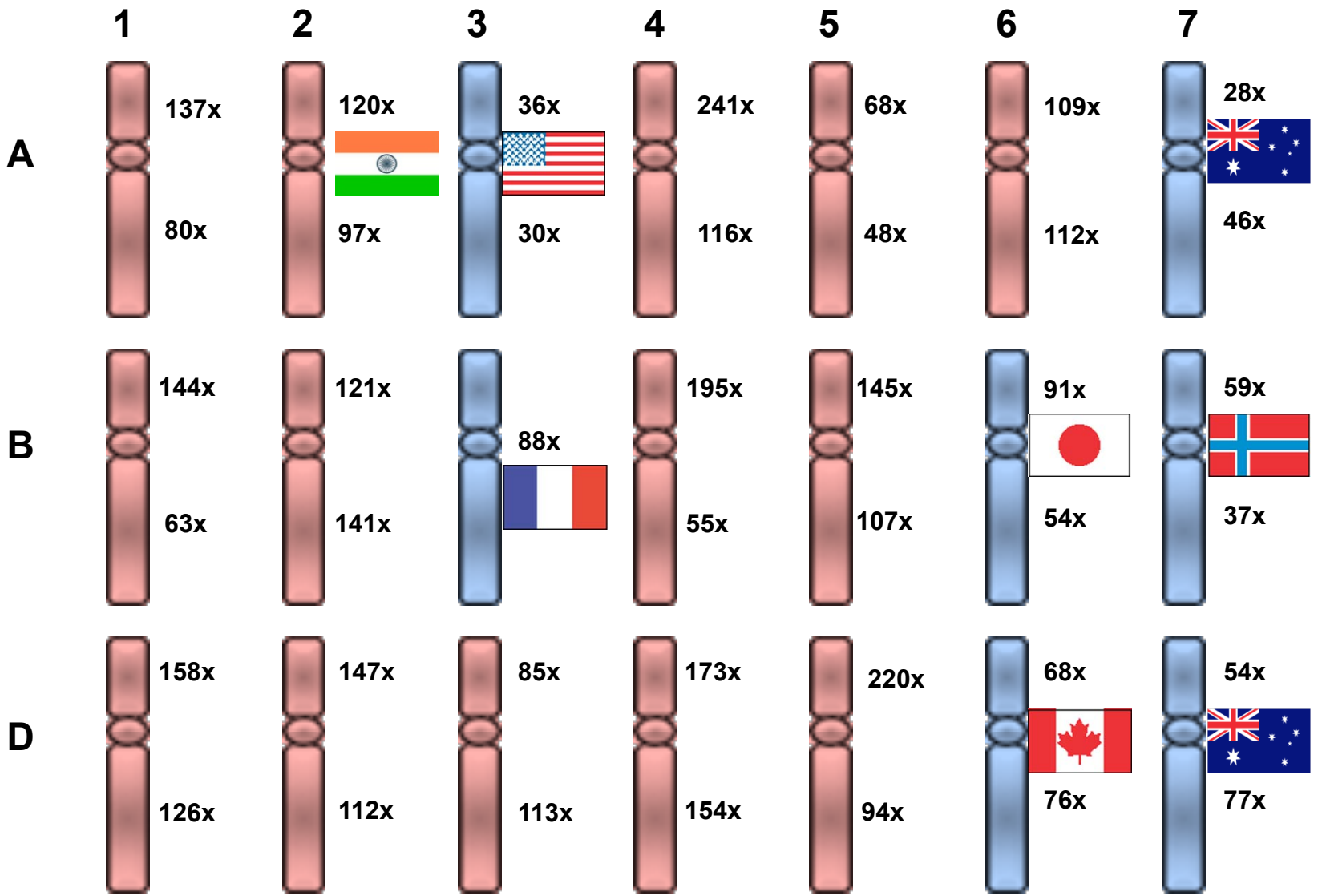
Assemblies

One Assembly per Chromosome Arm

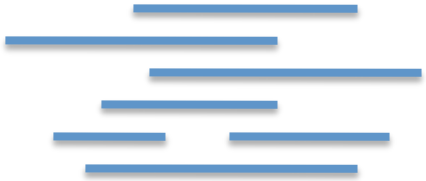
Genome Zipper

Klaus Mayer

Illumina Sequence Coverage



Stage 1: Sequence Assembly



Set of contigs

Remove contigs
< 200 bp

Assessment of assemblies

Resource

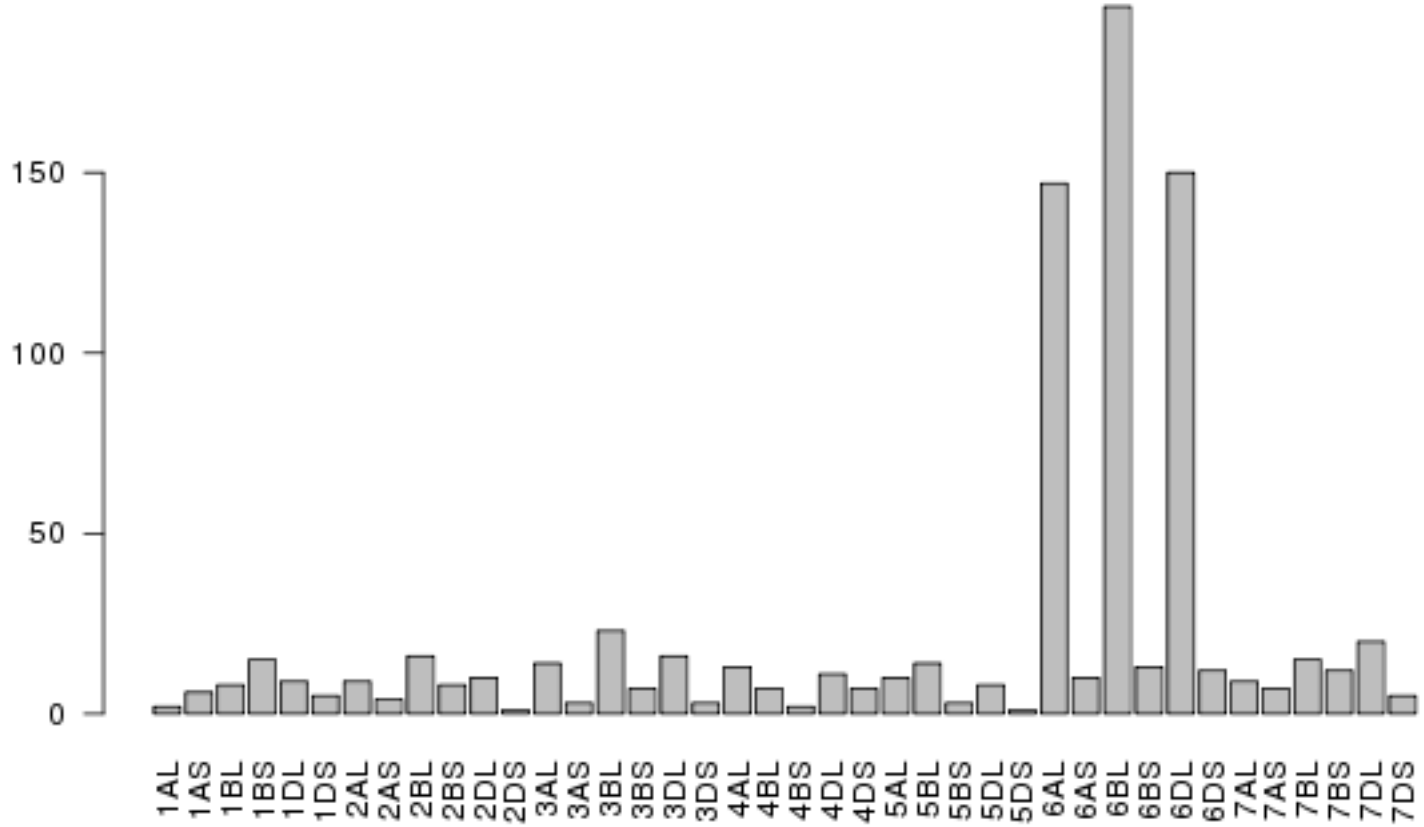
ABySS: A parallel assembler for short read sequence data

Jared T. Simpson,[1] Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol[2]

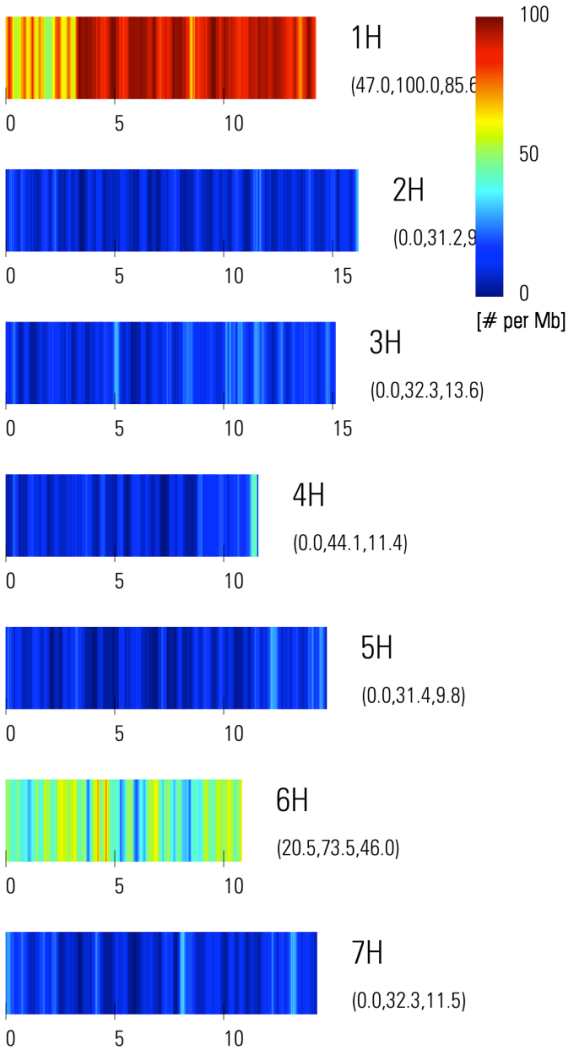
Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada

Assessing the assemblies for purity

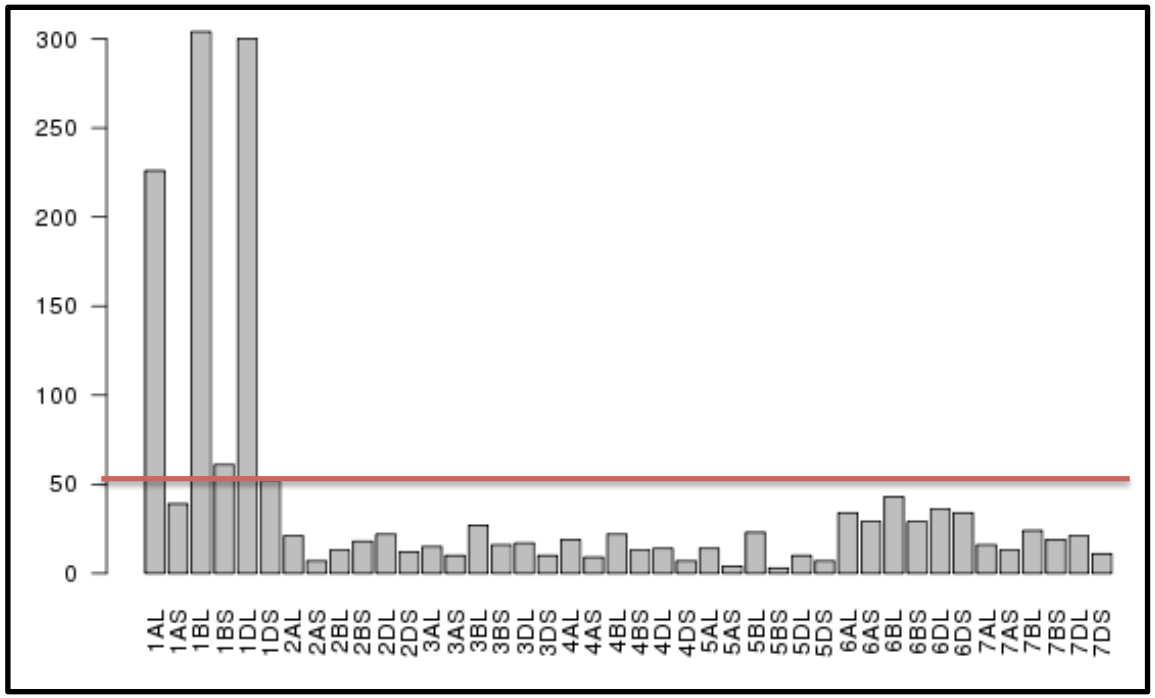
Alignment of bin-mapped wheat ESTs to the repeat masked assemblies (Qi *et al*, 2004)



Contamination in assemblies? (1BL)



Aligning contigs to the virtual barley genome
(Mihaela Martis, MIPS)



Aligning bin-mapped wheat ESTs to contigs (TGAC)

Contamination Assessment

11 assemblies contaminated

Action taken

- Regenerate flow-sorted chromosomes from wheat DNA
- Remake the libraries
- Resequence
- Reassemble

This improved 7 assemblies, leaving 4 problematic ones (1AL, 1BL, 5DL, 6BS)

TGAC developed a kmer-based cleaning approach to generate clean assemblies for these arms – see Jon Wright's presentation on Monday

Summary Statistics for whole genome

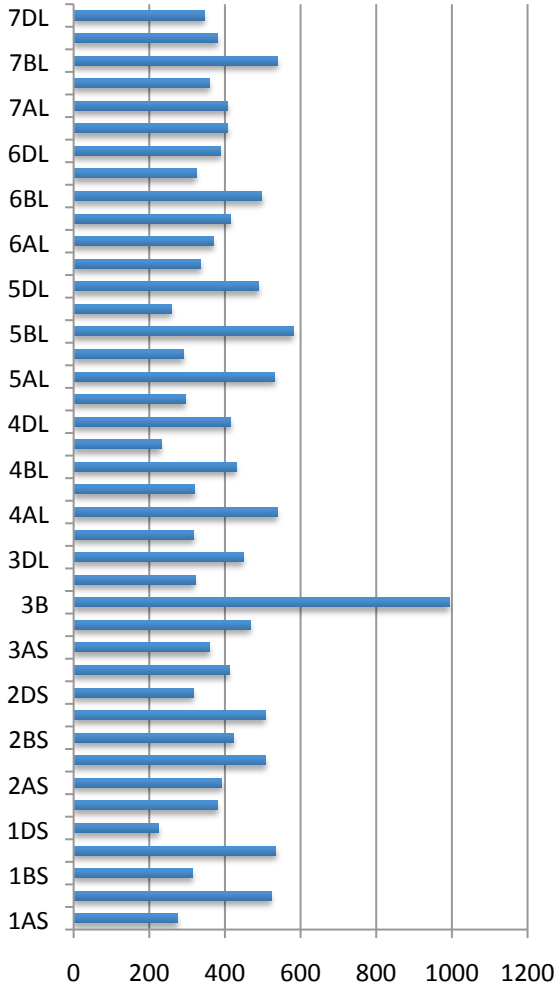
Assembly statistics

- Average GC content = **45.0%**
- N50 contig length (after filtering) = **2.4 kbp**

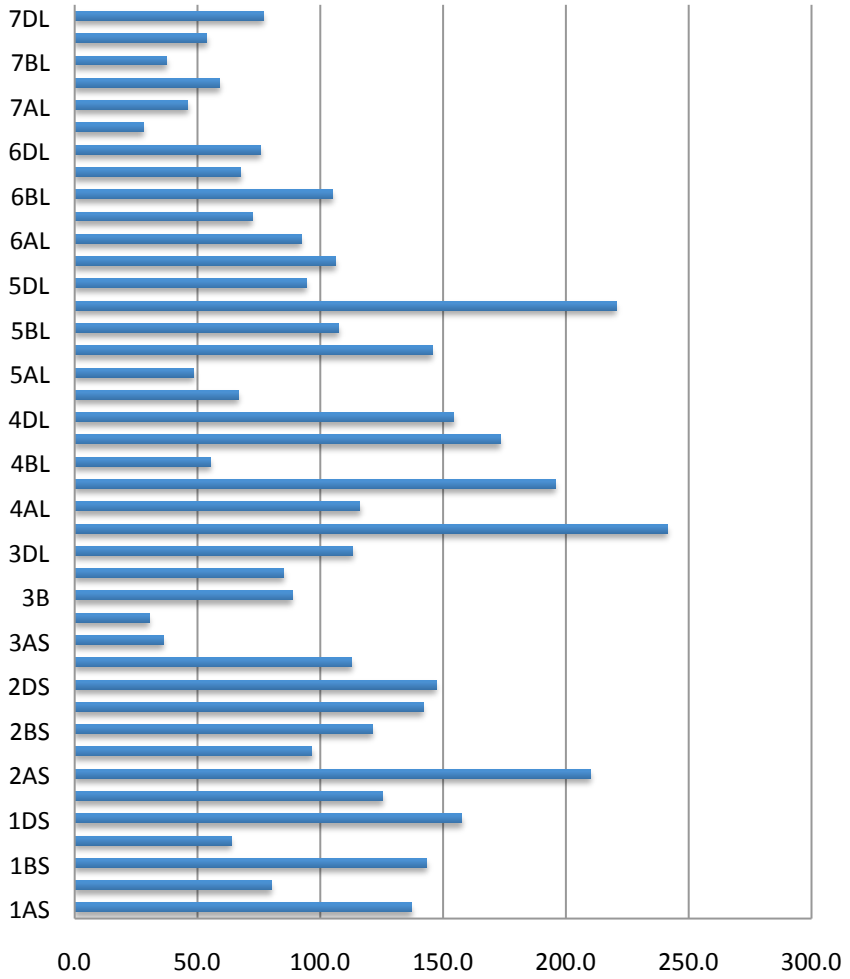
- Estimated gene count (based on hits to a barley gene set provided by IBGSC)
  - **1,526** (average per short arm)
  - **2,460** (average per long arm)
  - Total **83,977**

CSS Assembly Analysis

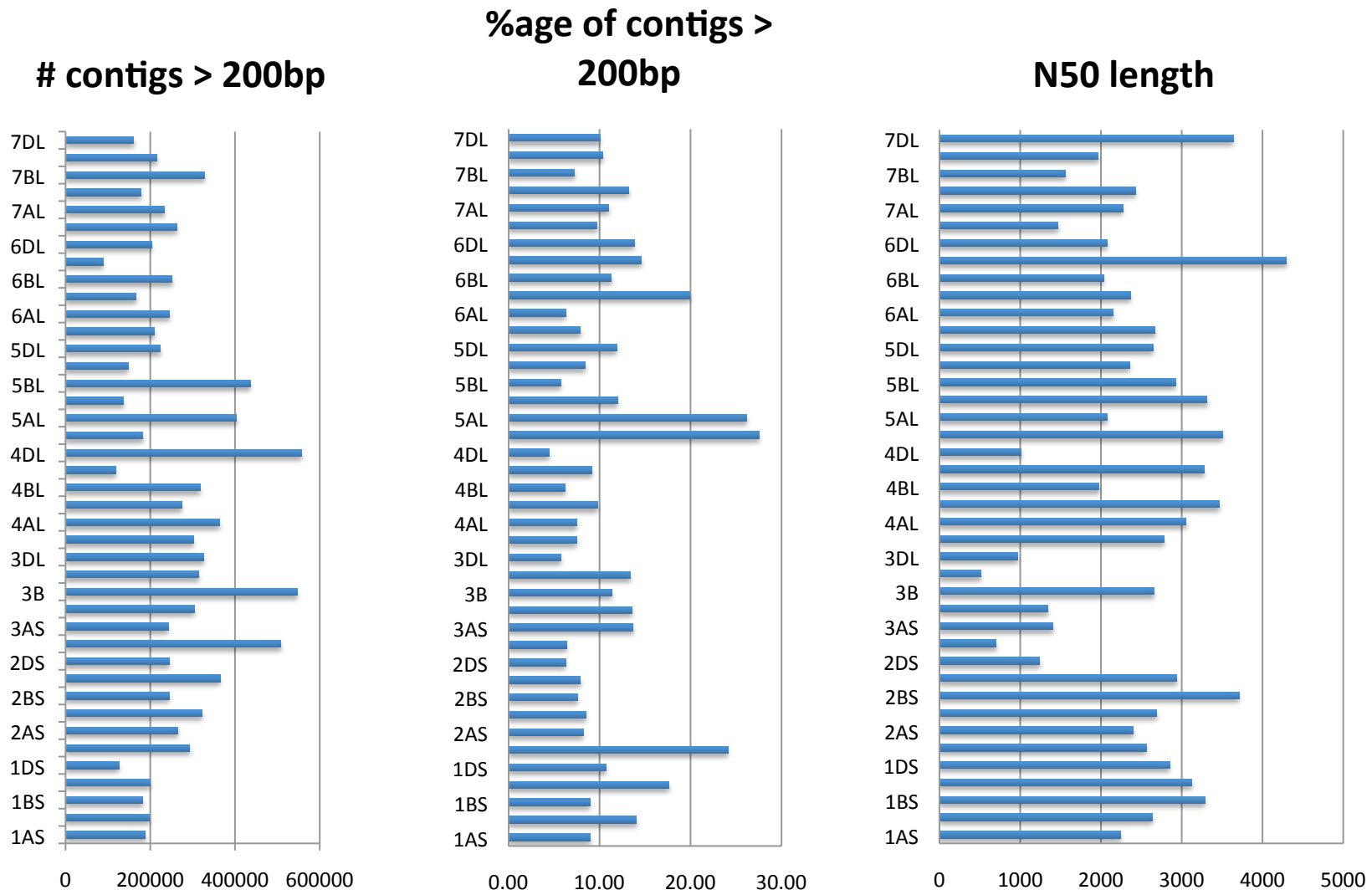
Chromosome size



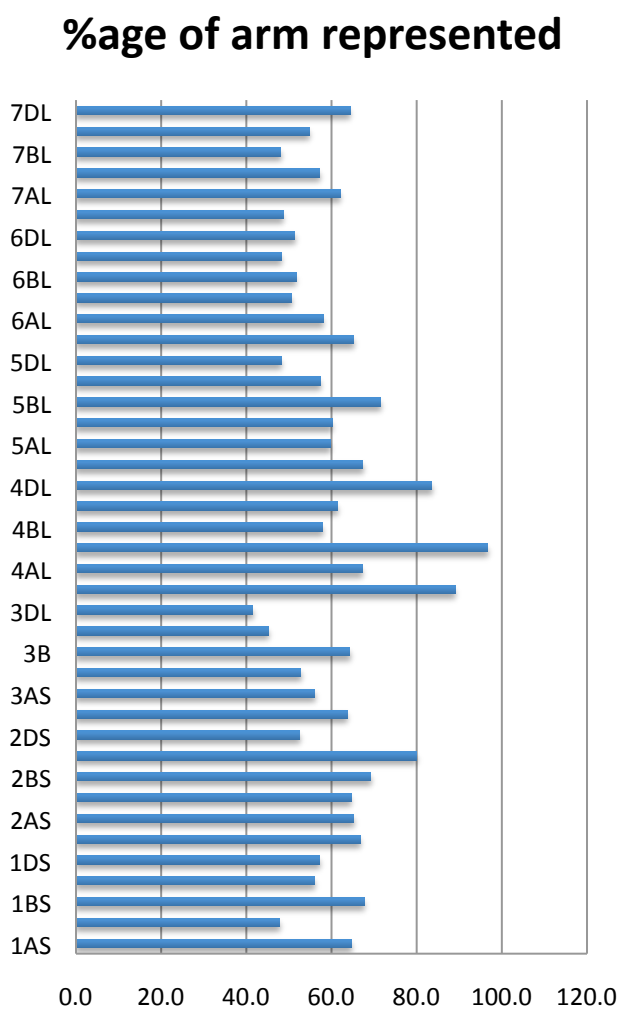
Chromosome arm coverage

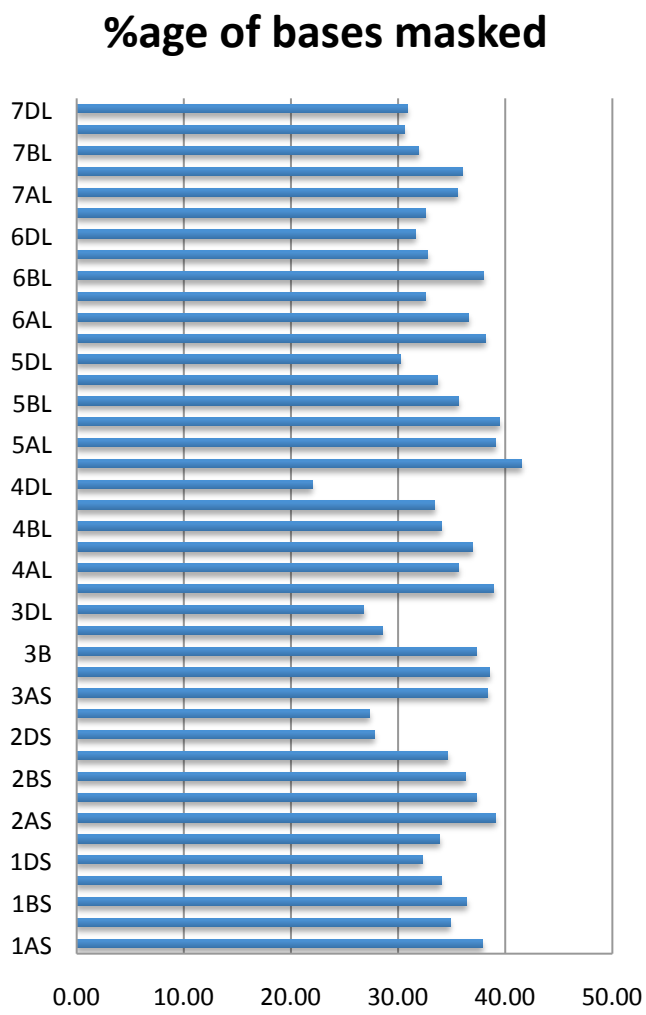


CSS Assembly Analysis – assembled contigs

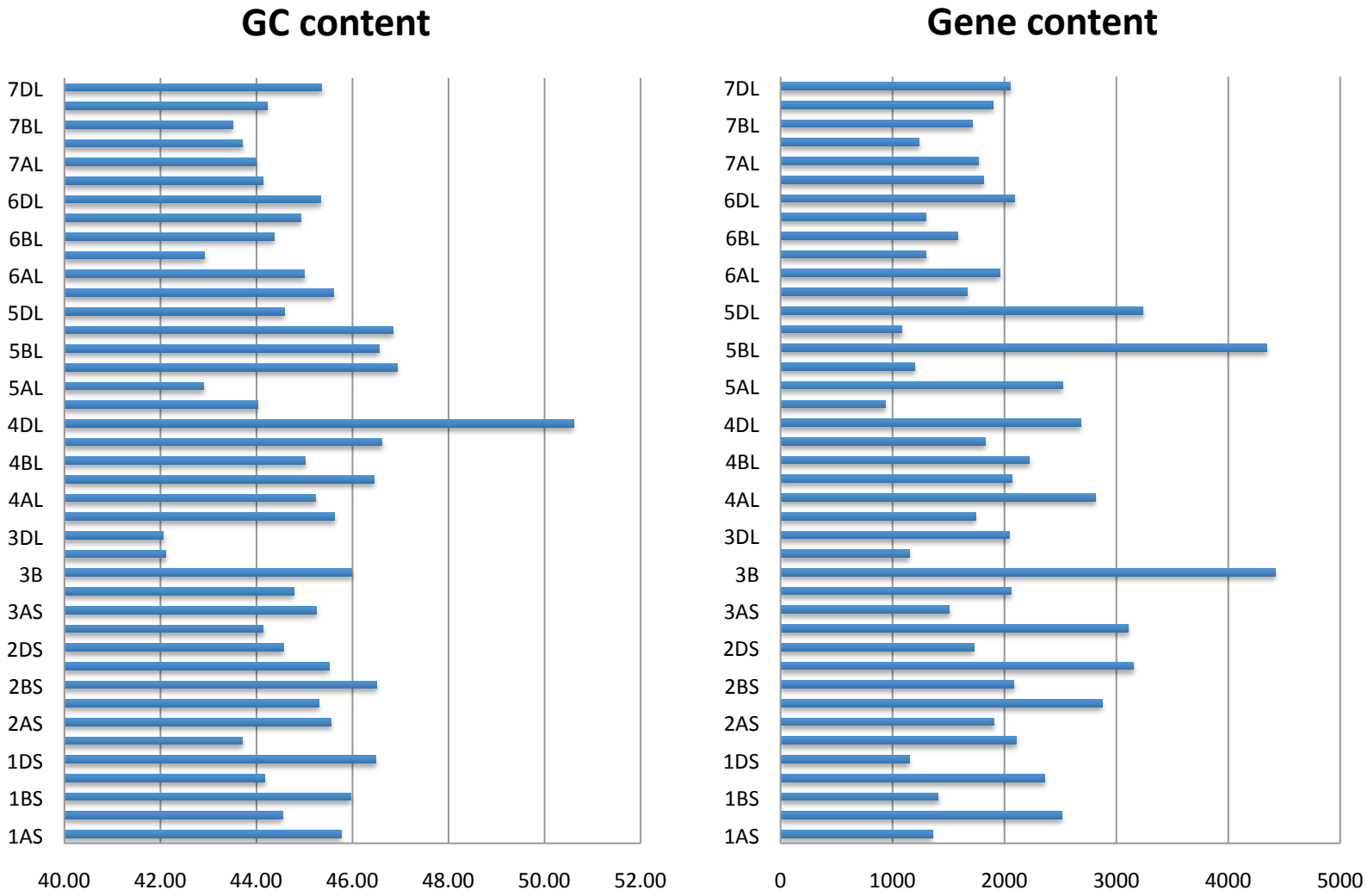


CSS Assembly Analysis
- chromosome coverage estimates

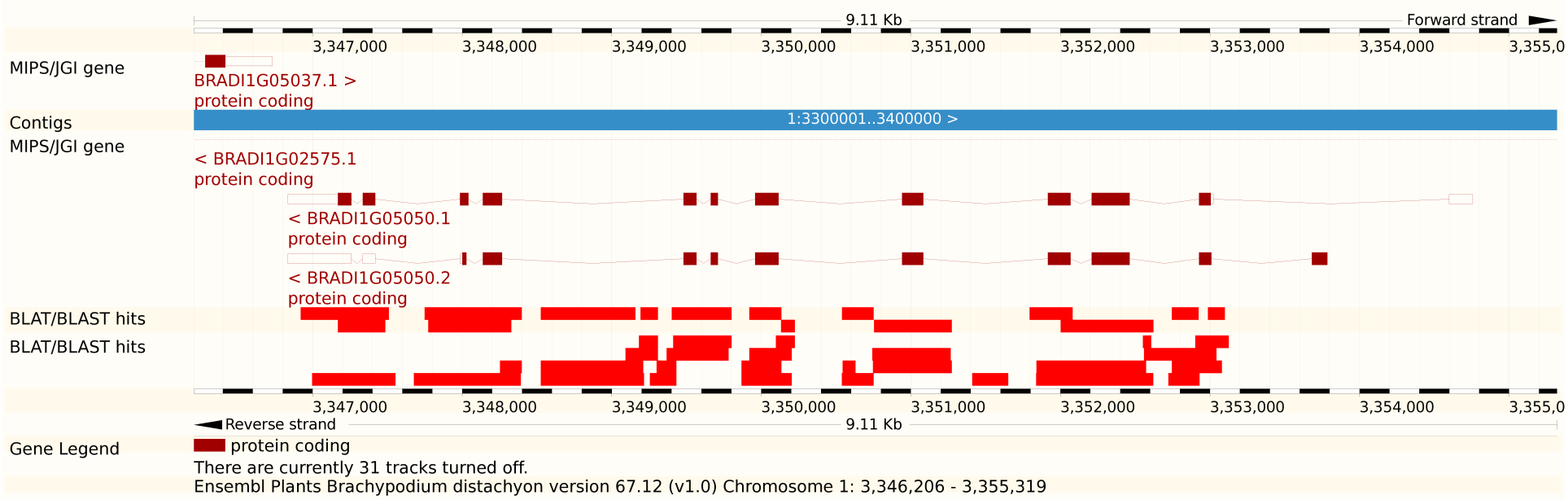




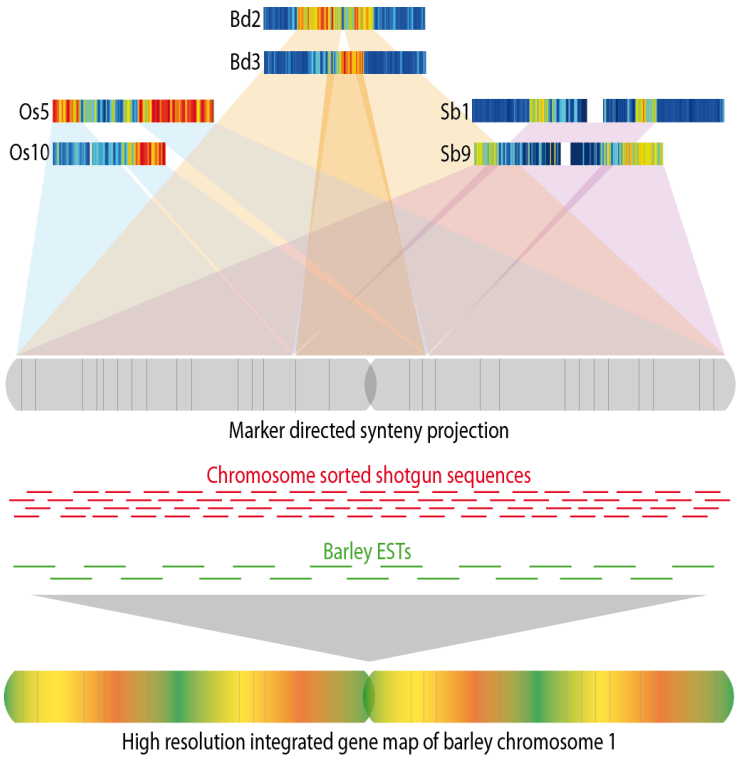
CSS Assembly Analysis
– base composition and gene content

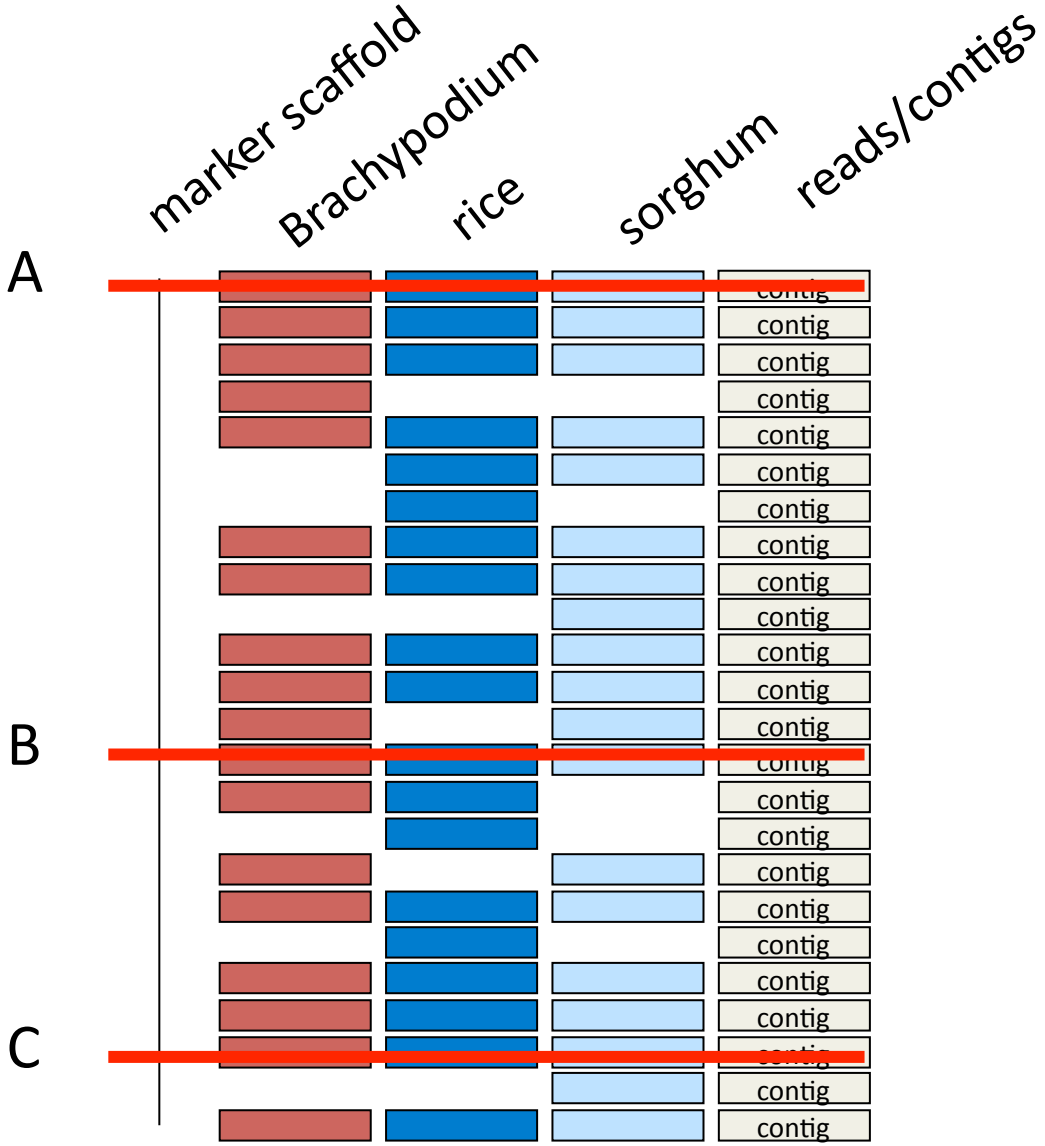


Alignment of assembled contigs with Brachypodium – an Ensembl browser view



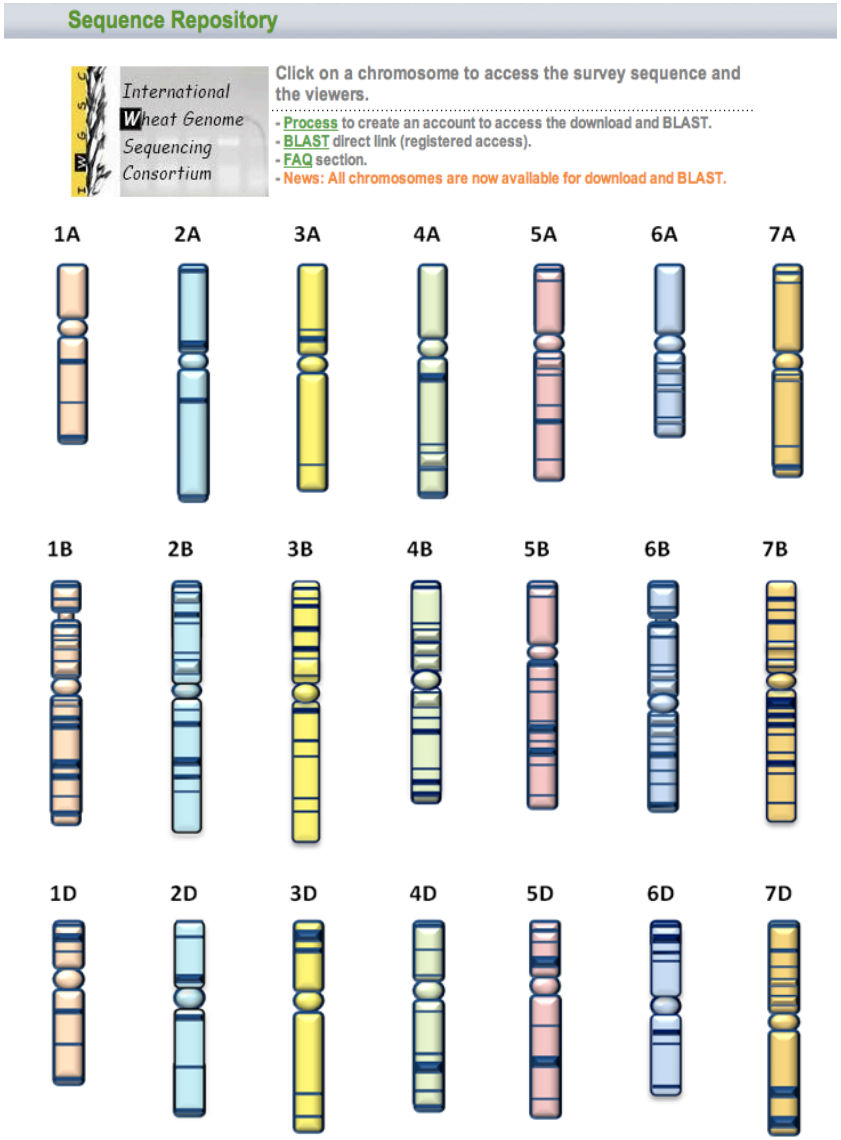
Stage 2: GenomeZipper + Virtual Gene Map: Syntenic Integration





Klaus Mayer
Mihaela Martis
MIPS

Stage 3: Survey Sequence Repository at URGI



Click on a chromosome to have access to the survey sequence with **download**, **blast search** and **viewers**.

WSS leaders can **download** and **blast** and non-members can blast and download the contigs matching hits once they agree to a data release policy.

All assemblies are available in the repository

http://urgi.versailles.inra.fr/Species/Wheat/Sequence-Repository

Applications of the survey sequence

The survey sequence provides a very fragmented view of the sequences of individual chromosome arms. It does not provide a true representation of the structure of the chromosomes but it does enable:

- Annotation of genes within contigs (intron-exon structure)
- Some limited annotation of 'pseudogenes'
- Analysis of coding variants
- *In silico* mapping of markers (e.g. genetic markers, trait markers) to chromosomes within sub-genomes
- Implement localised synteny studies
- Obtain estimations for
  - coding genes
  - lineage specific genes
  - comparative analysis of homoeologous genes

Next Steps

- Full analysis of data for publication – target submission date autumn 2012
- On publication read data and assemblies will be available from GenBank / EBI / DDBJ repositories in addition to sites with added value, e.g. URGI
- Assembly improvement options include:

  incorporation of 454 data (chr arm and whole genome)

  incorporation of BAC end data, where available

  addition of new data sets , e.g. chr arm mate pair data
- Improving the utility of data, e.g. visualisation options include providing synteny alignments of contigs with other genomes, e.g. Brachypodium and barley

What does the community need?

Acknowledgments

TGAC

Jon Wright

Mario Caccamo

Sarah Ayling

Kirsten McLay

Melanie Febrer

MIPS

Klaus Mayer

Mihaela Martis

URGI

Michael Alaux

IEB ASCR

Jaroslav Dolezel

Hana Simkova

IWGSC leadership

Kellye Eversole

Catherine Feuillet