IWGSC Workshop and Training Session
7-8 April, 2011

Wheat Bioinformatics activities at the CCG

Matt Bellgard
Director, Centre for Comparative Genomics

MURDOCH UNIVERSITY
PERTH, WESTERN AUSTRALIA

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Outline

- Contribution to the IWGSC
  - Two specific regions of chromosome 3B
  - Chromosome 7A
- Analytical environment for analysis
- Informatics issues

Outline

- Contribution to the IWGSC
  - Two specific regions of chromosome 3B
  - Chromosome 7A
- Analytical environment for analysis
- Informatics issues

Chromosome 3B

- **The ctg506 region** Selected for detailed analysis
  - Several cell wall invertase (IVR1) genes are located in this region.
  - These genes are often important in maintaining pollen viability during early development.

- **The ctg344 region** Selected for detailed analysis
  - Carries the gwm533 marker that is widely used to track disease resistance (Sr2) located on 3B
  - Potentially a region important in several disease resistances

Core Ideas for Assembly

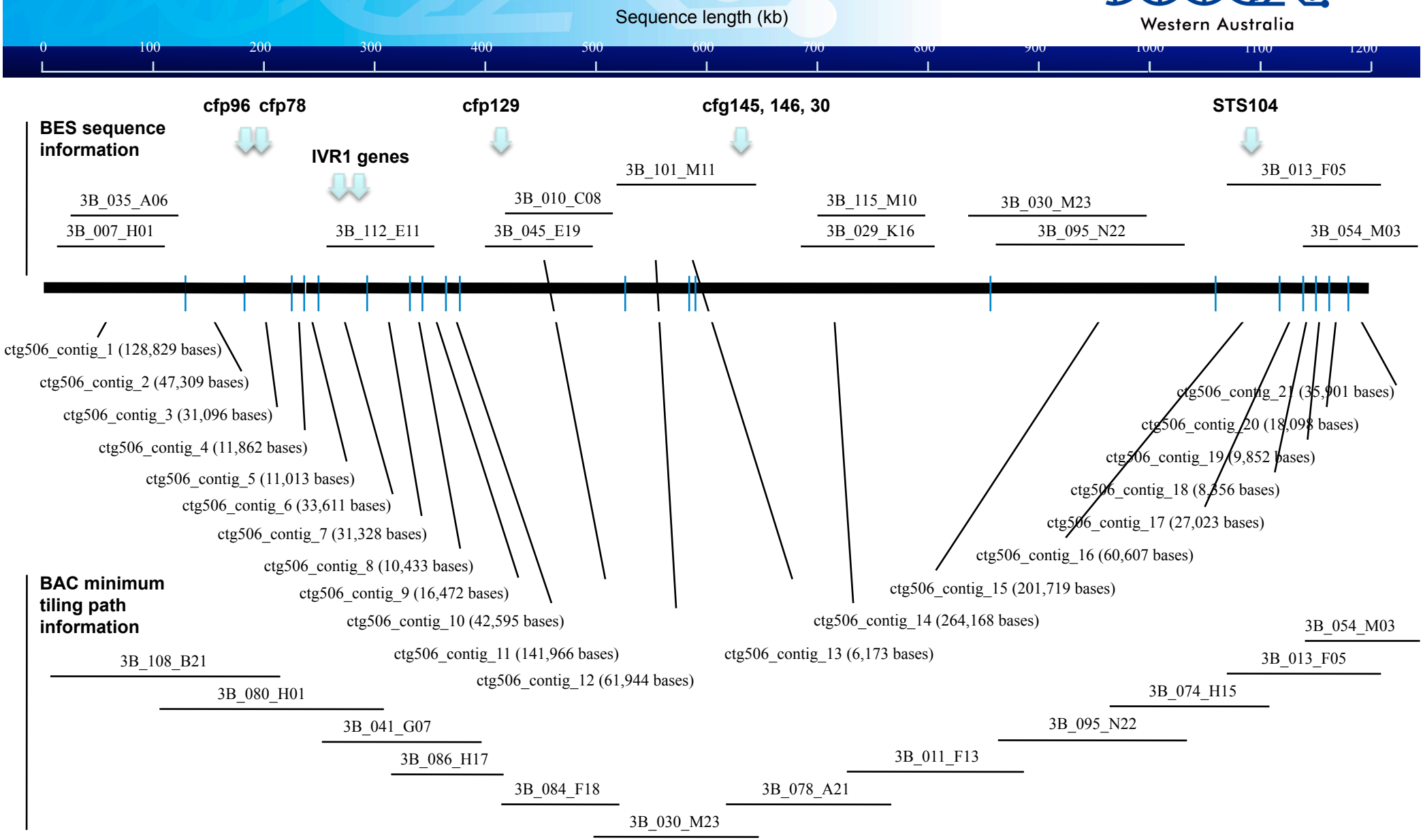
- We know wheat is “difficult” to assemble
  - mis-assemblies are and will be common
- If we see the same sequence assembled using different assemblers and data
  - this is more likely to be correct
- Points of divergence between different assemblers need to be analysed in detail
- Information from databases such as TREP need to be considered in compiling the detailed genome sequence.
- Genome sequence information needs to relate to a high quality molecular genetic map with traits included

Assembly approach

- Each BAC (Illumina short reads 70bp, paired-end, Sanger) assembled separately using Velvet
  - using multiple parameter combinations
- Different assemblies of the same BAC compared (mummer, freckle)
  - Potential mis-assemblies identified (eg: assemblies that disagree at certain points)
- Using the fingerprint assembly as a guide, identifying sequence present in overlapping BACs – take contigs that agree between different BACs to be “confident”. Start with large contigs and work down to shorter contigs.
- Look at paired-end alignments (BWA + genomeview, Hawkeye) and identify potential mis-assemblies
- Use paired-end information (illumina short reads 500bp insert + Sanger 4kb insert) to try to extend contigs (in general, this is difficult)
- Use LTR information from eg: TREP to order contigs

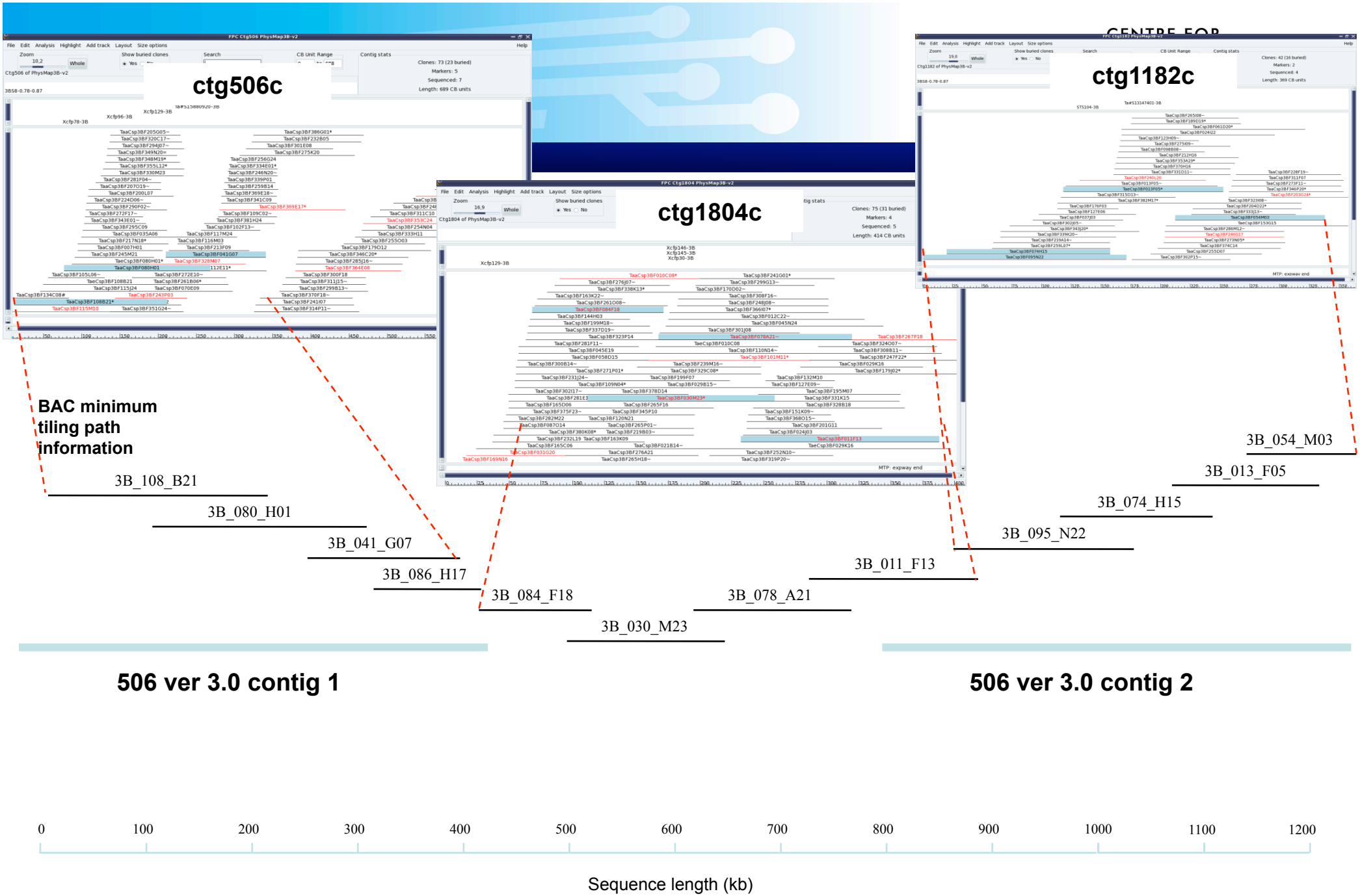
Original ctg506

CENTRE FOR COMPARATIVE GENOMICS
Western Australia

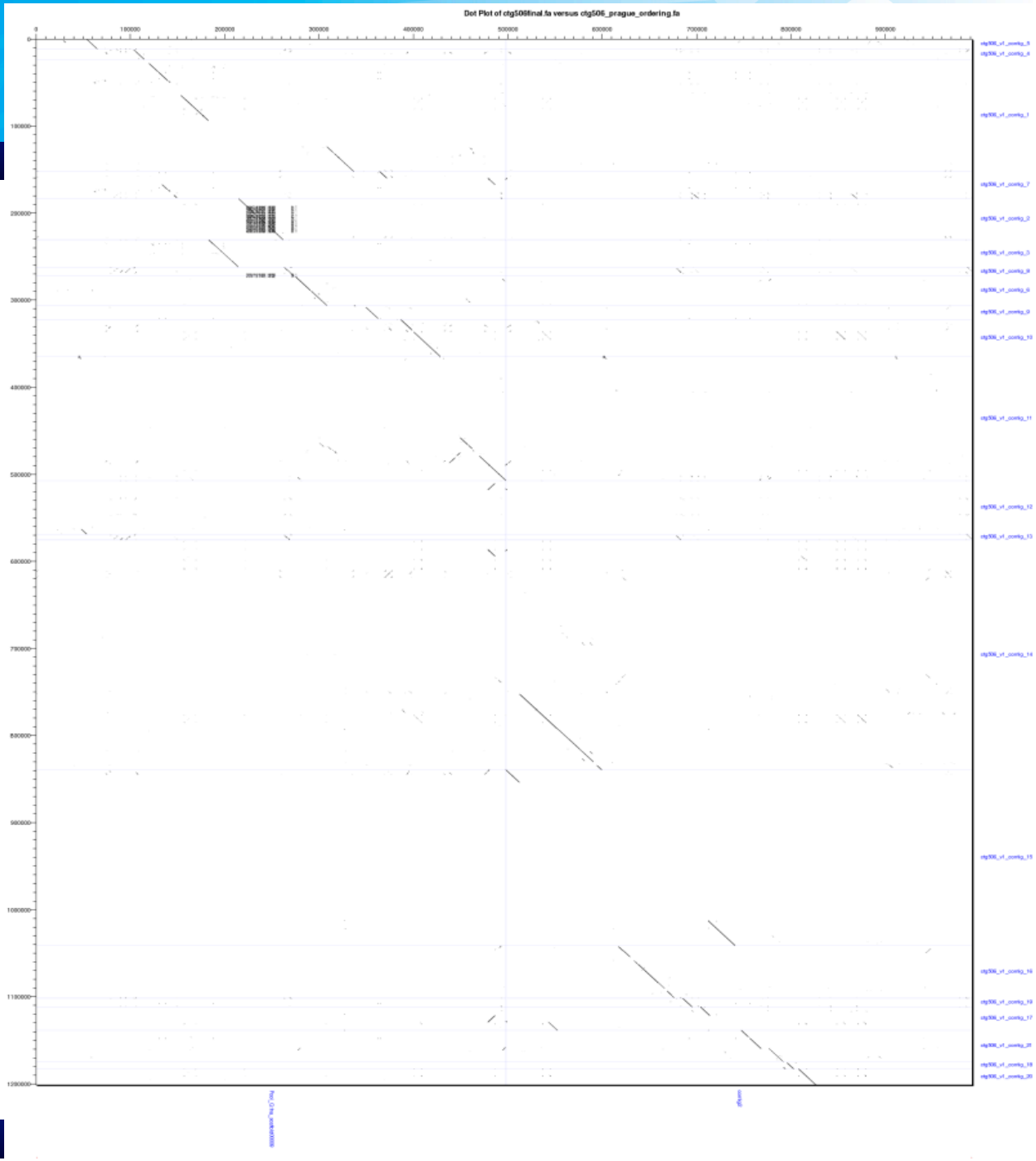


INRA 454/8kb of Ctg506

- CCG-illumina/Sanger sequencing compared to INRA 454/8kb mate pairs sequencing:
- The INRA scaffolds for ctg506
  - Closed 6 small gaps of Ns in CCG assembly
  - Improved the ordering of contigs in the CCG sequencing
  - Highlights miss-assemblies (at least 6)
- The CCG sequencing
  - Orient large INRA scaffolds
  - Closed 30/83 gaps in INRA sequencing assembly
    - Most of these gaps were in the introns of regions annotated as coding sequences.


ctg506c
ctg1804c
ctg1182c
BAC minimum tiling path information
3B_108_B21
3B_080_H01
3B_041_G07
3B_086_H17
3B_084_F18
3B_030_M23
3B_078_A21
3B_011_F13
3B_095_N22
3B_074_H15
3B_013_F05
3B_054_M03
506 ver 3.0 contig 1
506 ver 3.0 contig 2
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
Sequence length (kb)

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

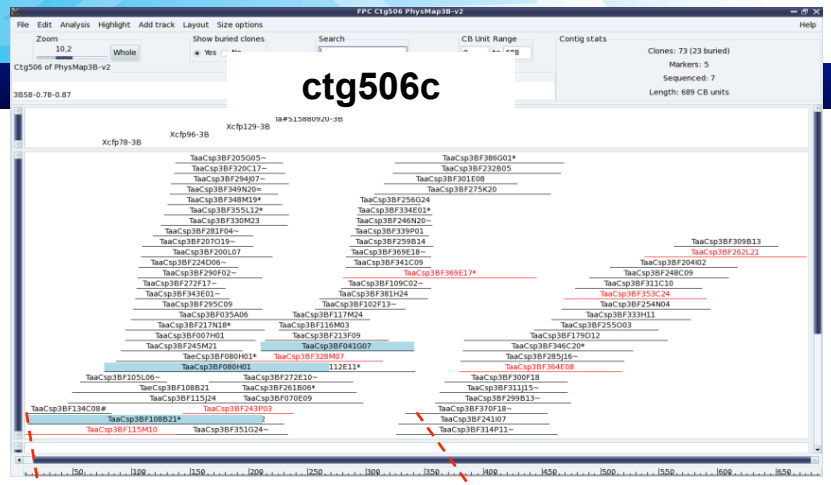


Alignment of CCG short read + Sanger assembly (y-axis) against version 3 (combined CCG + INRA) (x-axis)

The CCG contigs have been re-arranged in an attempt to fit the order revealed by INRA scaffolds

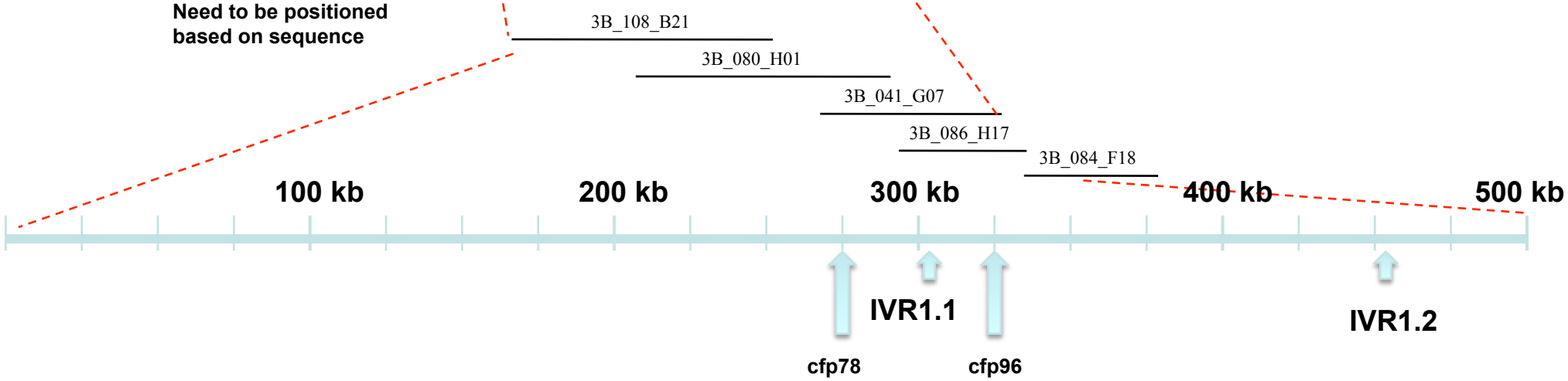
Detailed analysis of IVR gene region

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia



BAC minimum tiling path information

Need to be positioned based on sequence

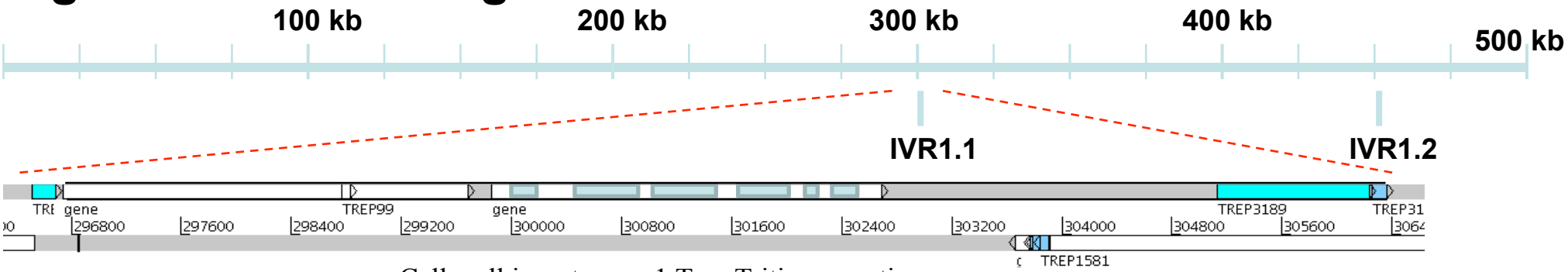


Hollie Webster PhD thesis

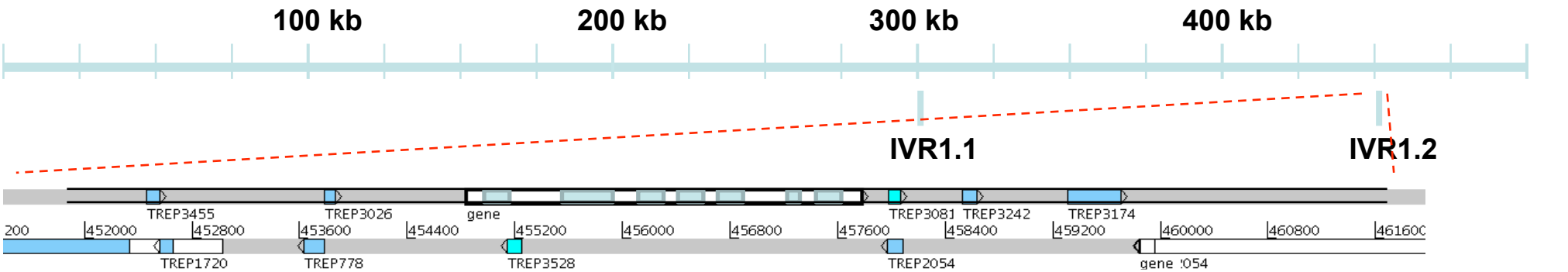
Detailed analysis of IVR gene region

CENTRE FOR COMPARATIVE GENOMICS
Western Australia

ctg506c ver3.0 contig1



Cell wall invertase n=1 Tax=Triticum aestivum RepID=O81118_WHEAT"; Evalue "0.0"; Identity "81% (476/586)";



Cell wall invertase n=1 Tax=Triticum aestivum RepID=O81118_WHEAT"; Evalue "1e-114"; Identity "54% (235/432)"

Hollie Webster PhD thesis

Summary – Chromosome 3B

- Have a good appreciation of sequence assembly issues for wheat genome
- Comparative genomics of a repetitive protein kinase locus in ctg344 and biological studies of cell wall invertase genes on ctg506
- Waiting for ctg1804c and ctg344 scaffolds from INRA

Outline

- Contribution to the IWGSC
  - Two specific regions of chromosome 3B
  - Chromosome 7A
- Analytical environment for analysis
- Informatics issues

Chromosome 7A

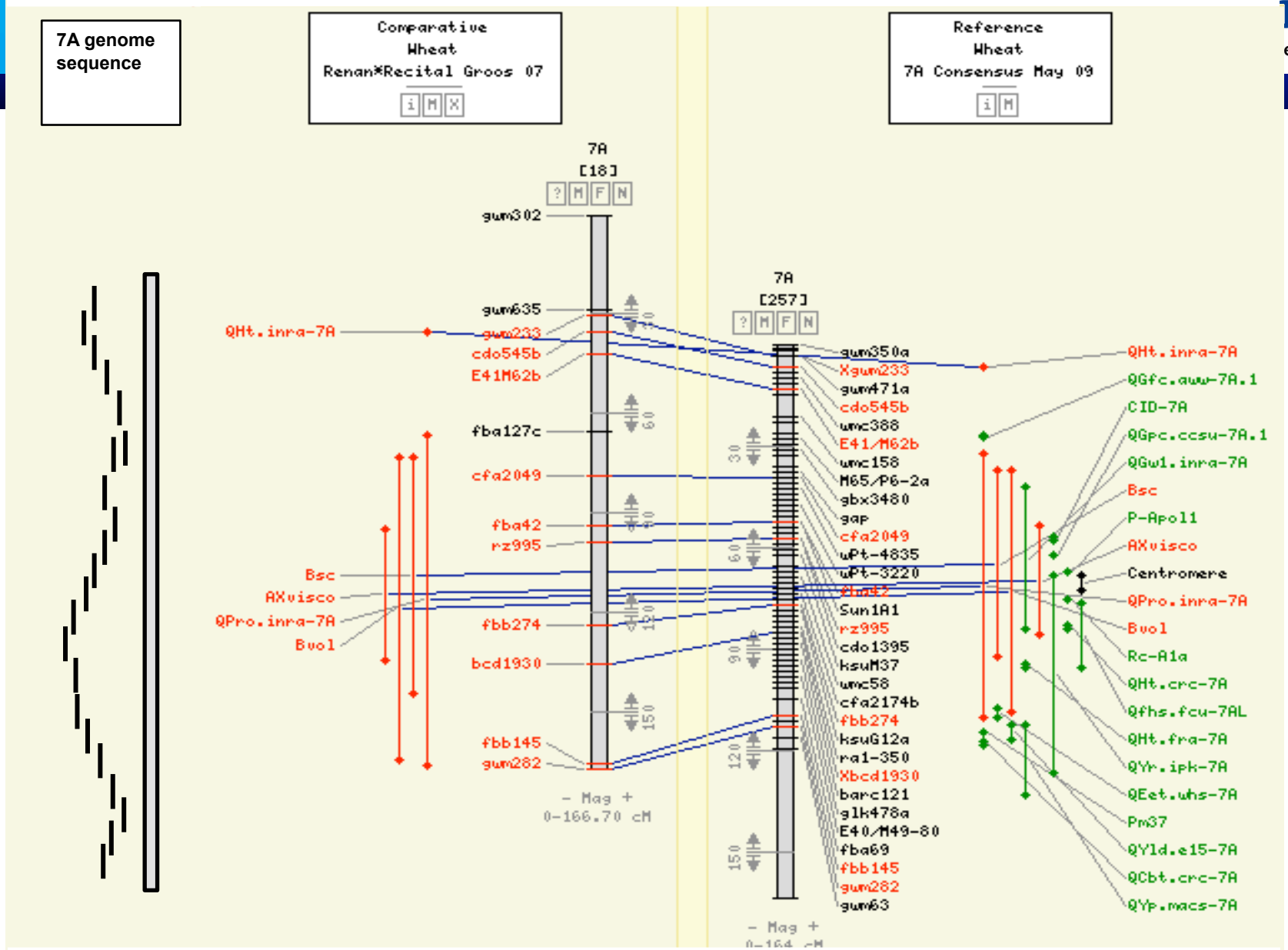
Overall aims

- Detailed analysis of QTL regions that are important to Australian agriculture
- Trait information linked directly to genome sequence
- ISBP identification of new markers for wheat breeding and selection for traits of interest

Tasks

- BAC-based physical map assembly of 7AS and 7AL
- BAC end sequencing
- Survey sequencing of 7A
- In-depth sequencing of QTL regions
- Anchor to genetic map

CENTRE FOR
COMPARATIVE GENOMICS
estern Australia



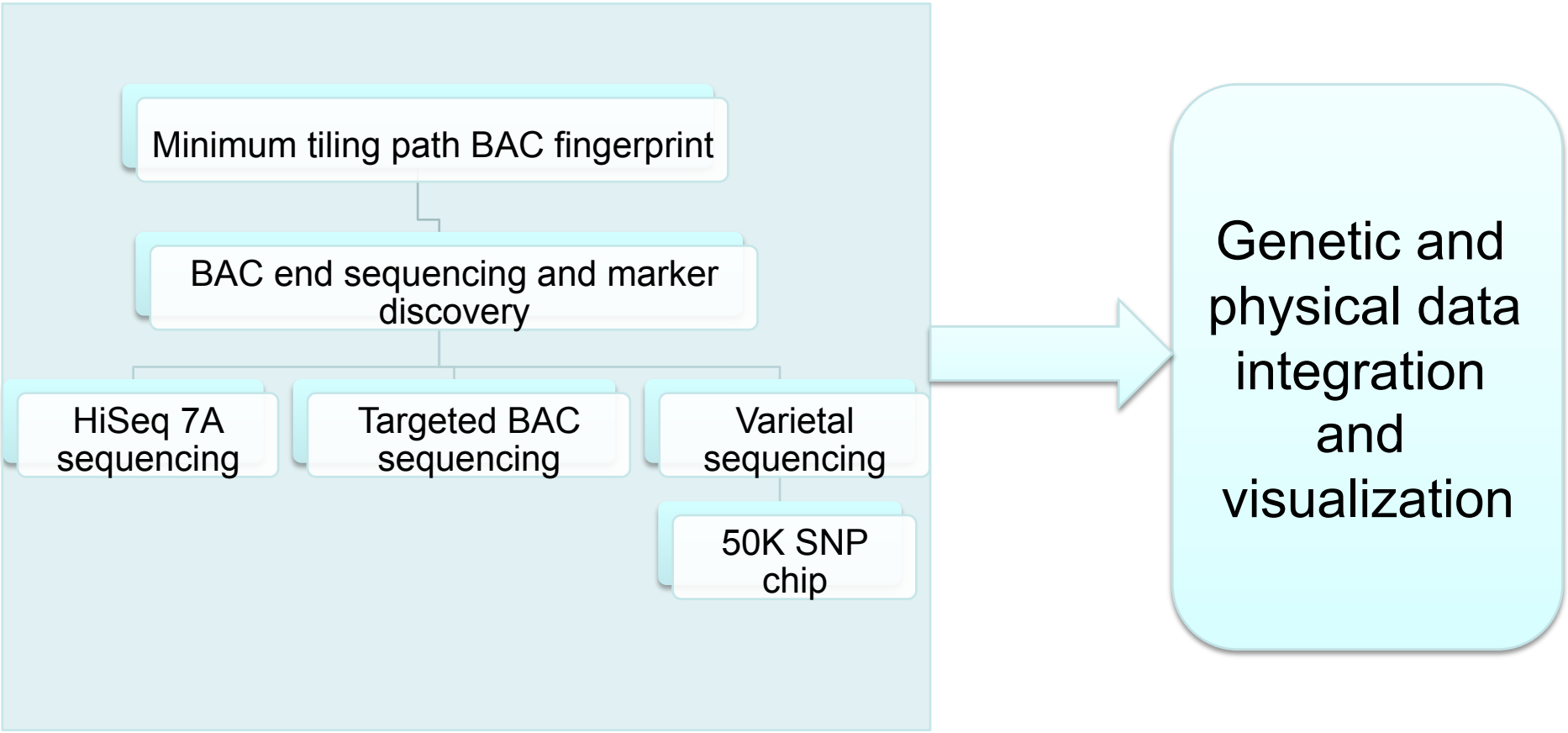
- Cmap (at CCG) compiles published QTL/trait data onto composite map.
- The QTL/trait data includes the information from the Wheat Gene Catalogue (McIntosh et al).
- The composite map is built from sections of published maps that share common markers to allow their integration into a master map. Trait data is incorporated based on flanking molecular markers. The molecular markers allow projection on to the genomic DNA sequence

Genome sequence | Molecular genetic maps | Traits

7A timeline

- **2010** Funding finalised (GRDC and BioPlatforms Australia)
  - 7A physical map, support to the IWGSC, survey sequencing, QTL region sequencing
- **March 2011** 7AS BAC library (58,000 BAC clones) produced by Dolezel lab
  - UC Davis for DNA fingerprinting (Mingcheng Luo)
- **May 2011** LTC/FPC 7AS fingerprint assembly
- **June 2011** compile BAC contigs and define minimum tiling paths (MTP) for physical map
- **May/June 2011** 7AL library from Dolezel lab (underway) shipped to UC Davis for DNA fingerprinting
- **June** commence BAC end sequencing (BACs from MTPs)
- **2011** Targeted BAC sequencing
- **2012** anchoring of BAC contigs to genetic maps

CCG: Wheat Bioinformatics 7A (Data integration)



Outline

- Contribution to the IWGSC
  - Two specific regions of chromosome 3B
  - Chromosome 7A
- Analytical environment for analysis
- Informatics issues

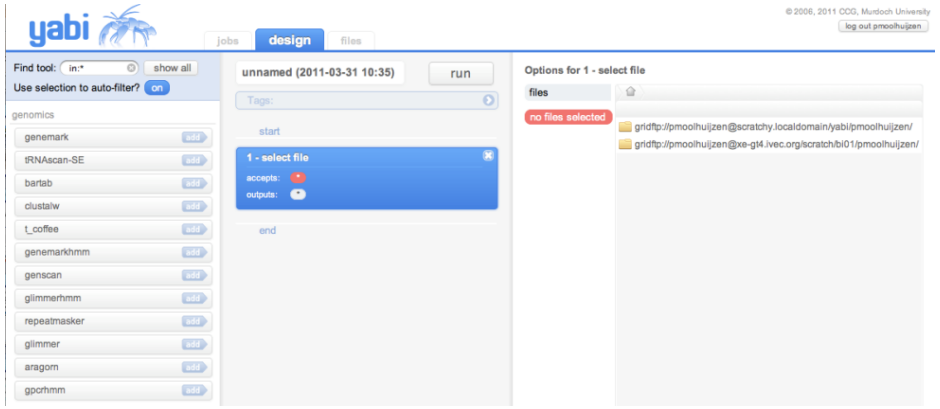
CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

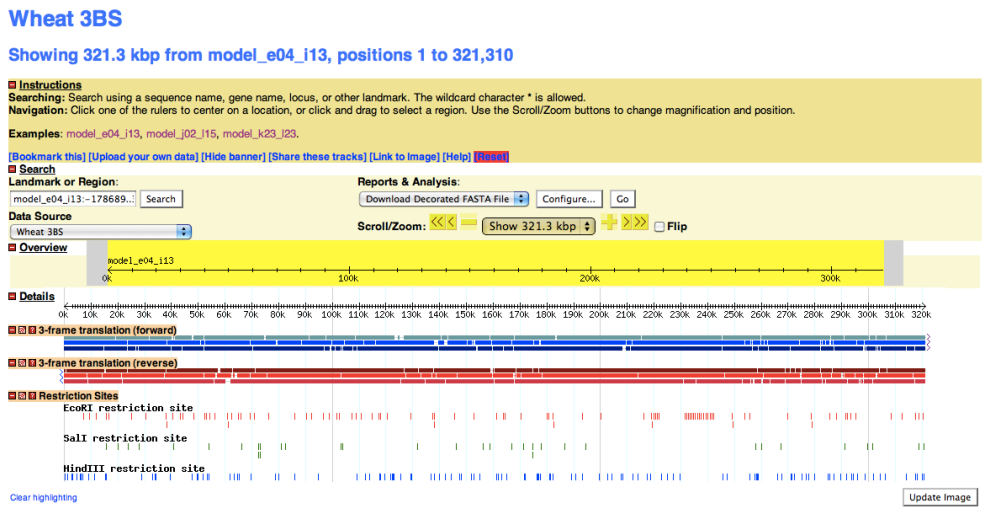
CCG: Wheat Bioinformatics activities 3B

HTP Assembly (Ctg344 and Ctg506)

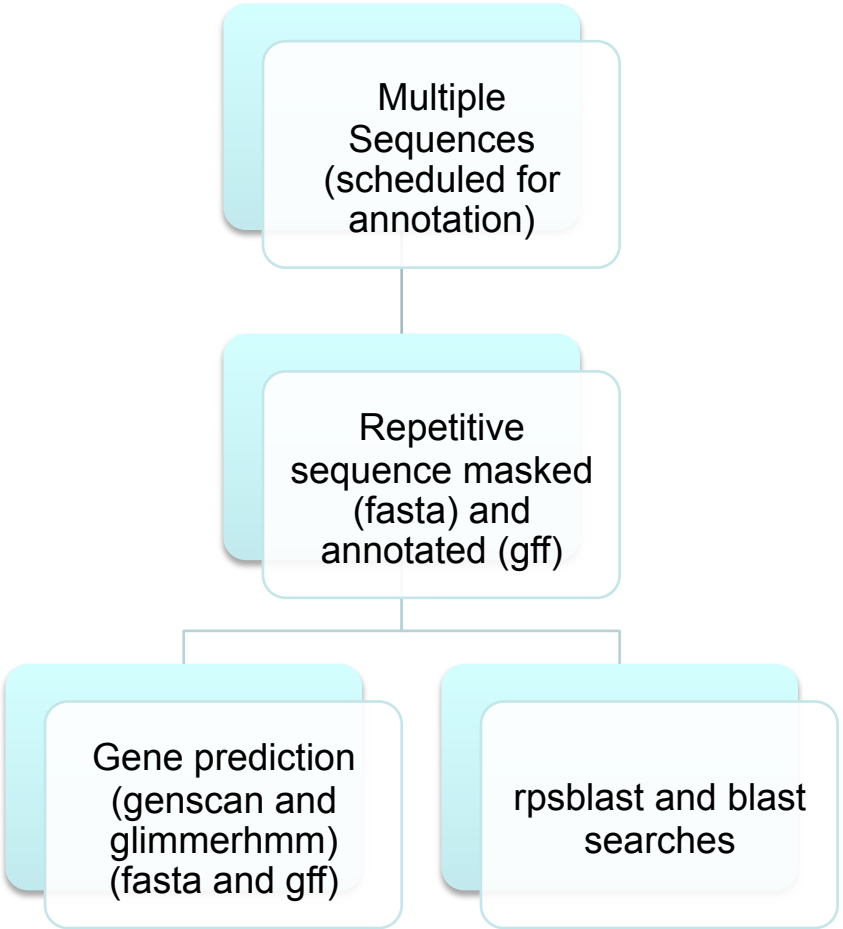
Annotation (HTP pipeline)

Visualization (Gbrowse)



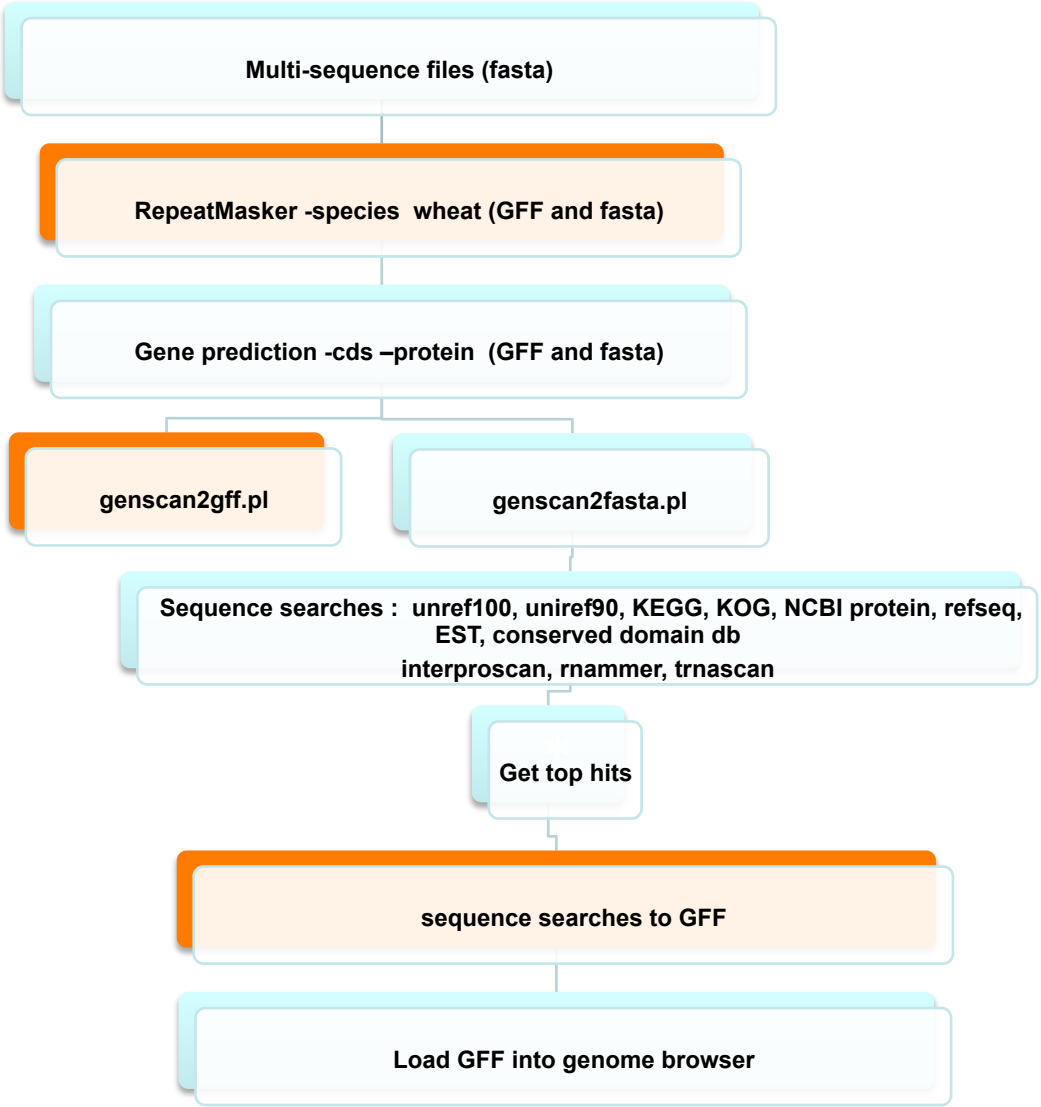


Bioinformatics annotation workflows URGI (triannot) and YABI



Bioinformatics workflows - Annotate sequences (YABI and Triannot)

CENTRE FOR COMPARATIVE GENOMICS
Western Australia



YABI - Front-end

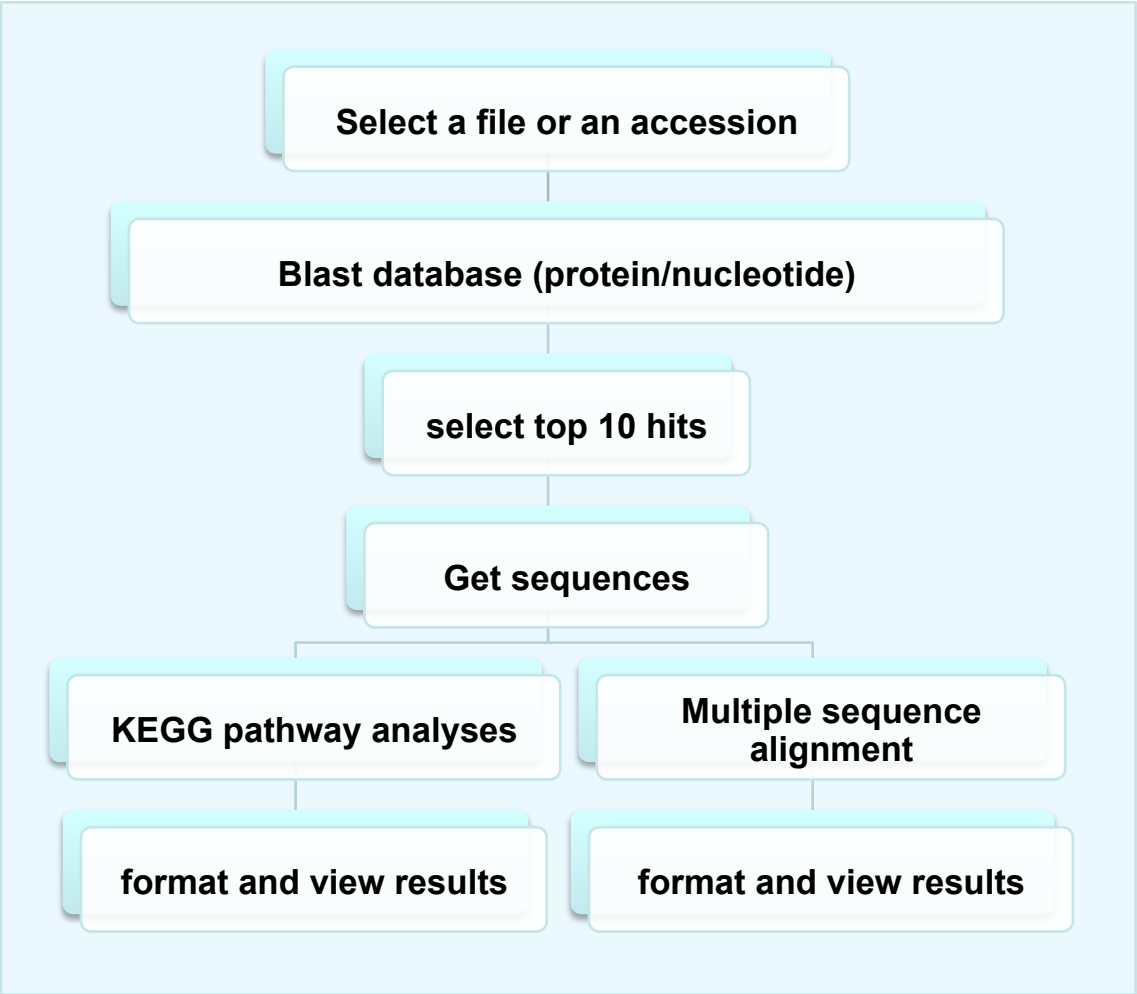
- Designed for a non-technical users
  - jobs tab, design tab, files tab
- Simple and easy to use as possible
- Reuse workflows
- User access
  - When you log in, user sees what they have access to
  - Allow scientists to work together
- Drag and drop tools

YABI - Front-end (2)

- Usability
  - Users warned if tools dragged out of place
  - Anticipate file extensions require for a given tool
    - Tools filtered
  - Tags – capture meta data
  - Errors trapped by system
- File manager
  - Drag and drop files
  - File copying via streaming
    - not via front-end
- Command-line
  - Power users

YABI demo

Simple Blast search and further analyses

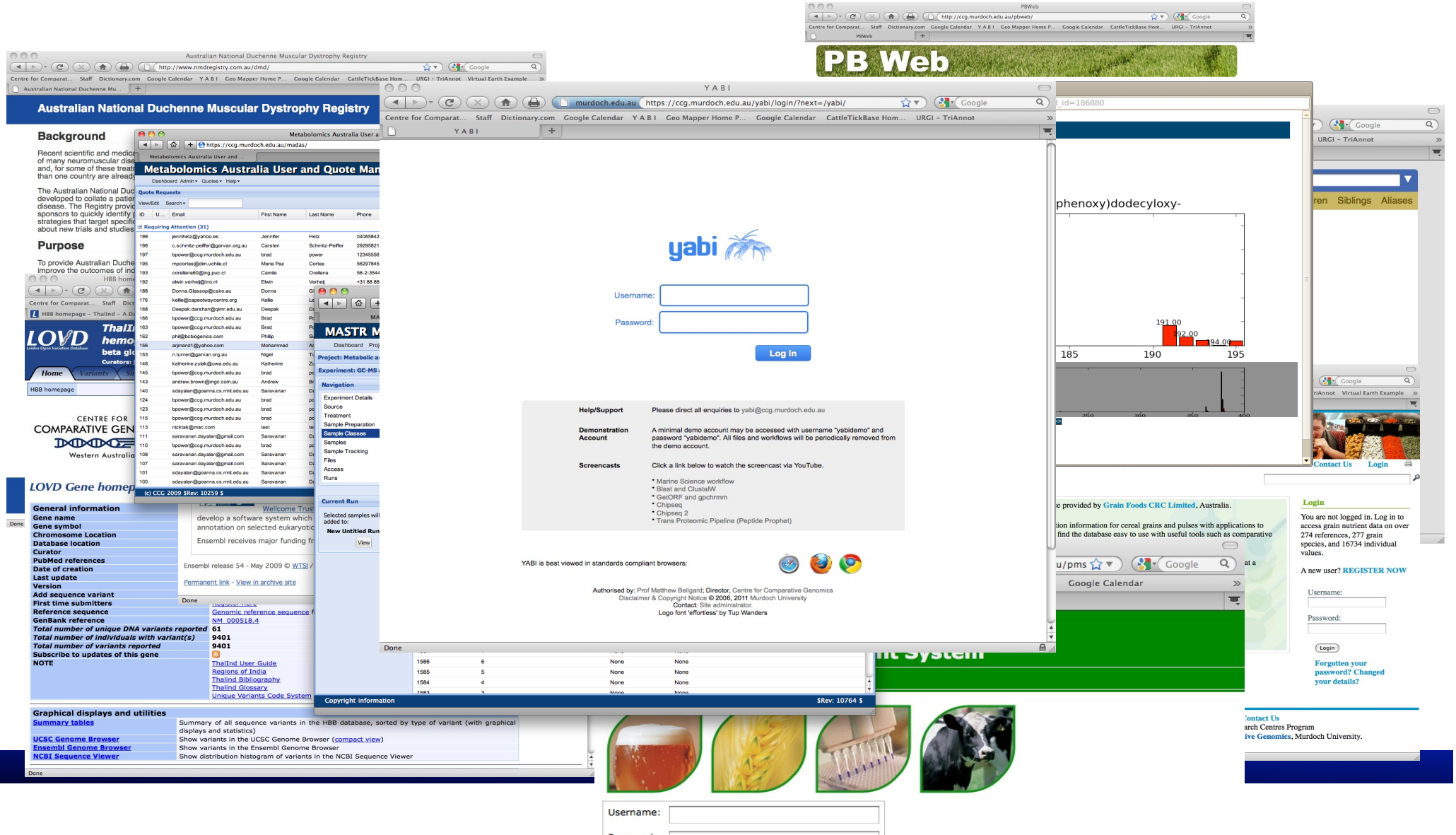


CCG: Hardware

- Stage 1A Pawsey Centre
- Ranked 87 in the world
- 9600 cores

CCG: Software development

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia



Next generation sequencing analysis projects

- De novo genome/transcriptome assembly and annotation
  - Wheat, barley, rat mutant, Wine yeast, cane toad, Rhizopertha, Campylobacter, Euphorbia, Cattle tick, dog tick, rhizobium, spirochetes
- Transcriptomics/epigenomics/metagenomics
  - microRNAs (human, Arabidopsis, cattle tick)
  - Epigenomics in Arabidopsis
  - Metagenomics of environmental samples (ancient DNA)
- Disease association
  - LPK rat mutant, human disorders, diagnosis assays

Outline

- Contribution to the IWGSC
  - Two specific regions of chromosome 3B
  - Chromosome 7A
- Analytical environment for analysis
- Informatics issues

Informatics considerations

- Process/timing to integrate 3B CCG results back to IWGSC
- Protocols for BAC physical assembly
- Process/timing to integrate 7A results back to IWGSC
  - Linkages to CCG resources
- CCG could assist in scoping LTC software porting requirements

Acknowledgements

CCG Team

- Rudi Appels
- Gabriel Keeble
- Adam Hunter
- Andrew McGreggor
- Paula Moolhuijzen

Collaborators

- Catherine Feuillet, Etienne Paux and Frederic Choulet (INRA, France)
- Jaroslav Dolezel (Czech Republic)
- Mingcheng Luo (UC Davis)

Funding

- GRDC
- BioPlatforms Australia