

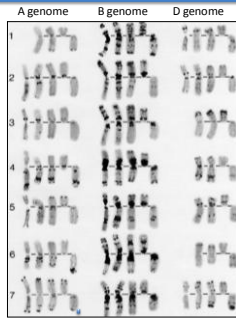
Integration of Whole Genome Assemblies and Genetic Information in Hexaploid Wheat

Mario Caccamo
mario.caccamo@tgac.ac.uk
@mcaccamo

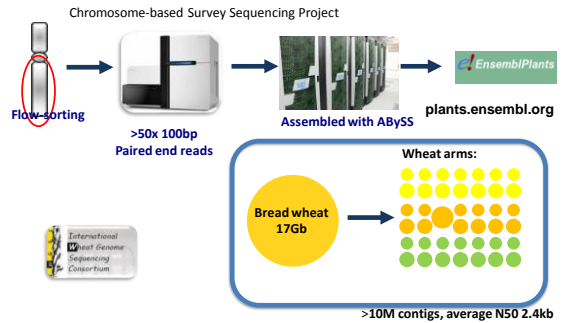


Wheat genome architecture

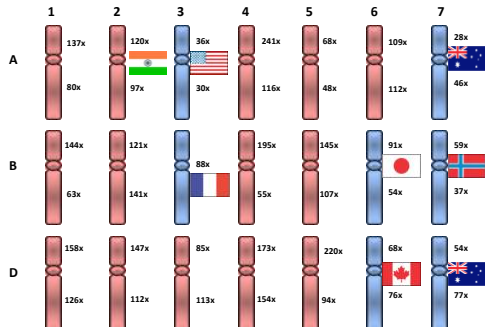
- Large size (17,000 Mb)
- Allohexaploid genome: A, B, D genomes (3 sets of 7 homeologous chromosomes)
- Homologous vs Homeologous features
- High repeat content (80-85 %)



IWGSC CSS Project



Sequencing Coverage



Assemblies in plants.ensembl.org



What we can do ...



- Annotate genes within contigs (intron-exon structure)
- Coding variants
- Link features to chromosomes (within sub-genomes)
- Implement localised synteny studies
- Obtain estimations for
 - coding genes
 - "pseudogenes"
 - lineage specific genes
 - comparative analysis of homoeologous genes**

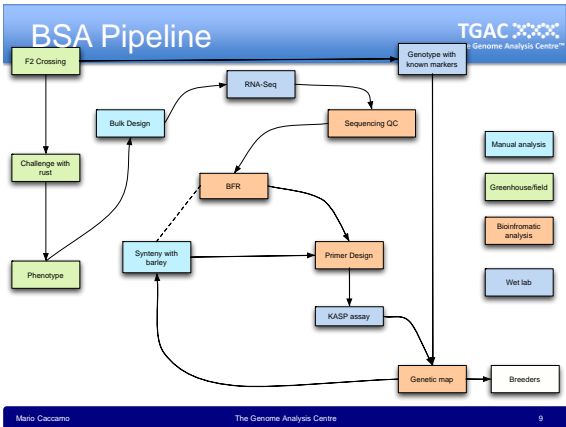
Bulk Segregant Analysis



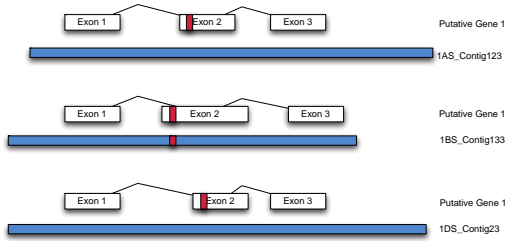
Ricardo Ramirez-Gonzalez



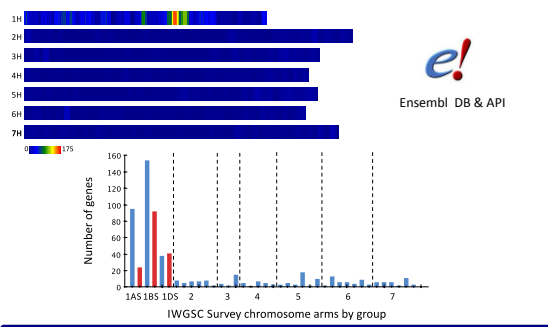
@CristobalUauy



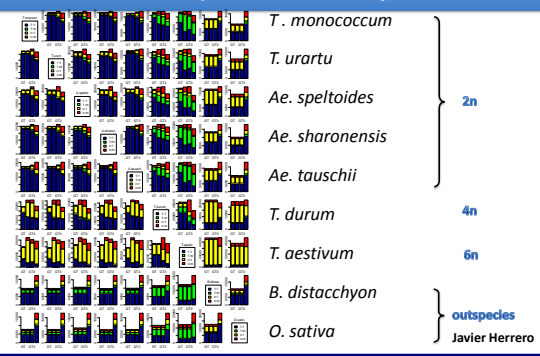
Gene expression in hexaploid wheat



Validation and candidate selection



Triticeae comparative analysis



Exome Capture

TGAC The Genome Analysis Centre

genomebiology.com/content/14/6

Separating homeologs by phase tetraploid wheat transcriptome

Genome Biology 14:122 (2013)

Abstract
Background: The high level of identity among duplicated homeologous substantial challenges for *de novo* transcriptome assembly. To solve this problem we developed a workflow that optimizes transcriptome assembly and separates our strategy, we sequence and assemble the transcriptome of one of the diploid components with a benchmark set of 13422 full-length, non-redundant transcripts.
Results: A total of 489 million 100 bp paired-end reads from tetraploid wheat including both the benchmark cDNAs. We used a comparative genomic approach to identify the multiple k-mer assembly strategy increases the proportion of single contigs by 22% relative to the best single k-mer size. Homology-guided assembly that includes polymorphism identification, phasing of SNPs, and using a reference set of genes, we determine that 96.7% of SNPs analyzed.
Conclusions: Our study shows that *de novo* transcriptome assembly of tetraploid wheat is possible. We demonstrate that a multiple k-mer assembly strategy is more effective than single k-mer assembly. Our results also demonstrate that a homology-guided approach can be used to separate the transcriptome of tetraploid wheat. The predicted tetraploid wheat genome and gene models will be made available to the wheat research community and to those interested in comparative genomics.
Keywords: Transcriptome assembly, multiple k-mer assembly, wheat, polyploidization, phasing, gene prediction

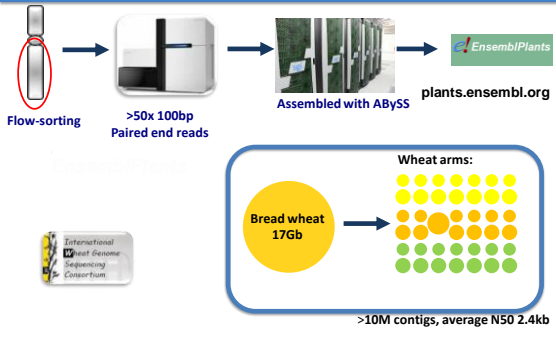
What we CAN'T do ...

TGAC The Genome Analysis Centre

- Obtain a complete and finished genome
- Detect large re-arrangements
- Annotate regulatory elements
- "Properly" develop genome-wide studies (such as GWAS)
- Characterise CNVs, and other structural variants

Divide & conquer approach

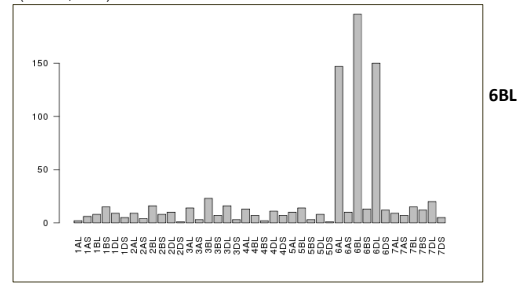
TGAC The Genome Analysis Centre



Flow-sorting purity

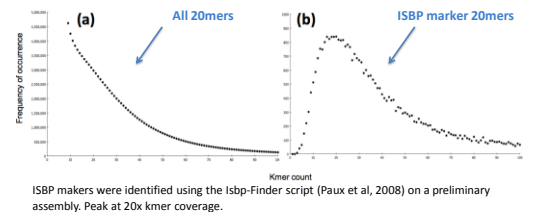
TGAC The Genome Analysis Centre

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



What about DNA amplification

TGAC The Genome Analysis Centre

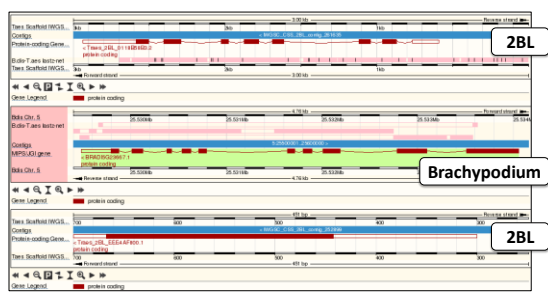


ISBP makers were identified using the Isbp-Finder script (Paux et al, 2008) on a preliminary assembly. Peak at 20x kmer coverage.

Assembly algorithms expect uniform coverage and "well-behaved" kmer distribution

Fragmented genes

TGAC The Genome Analysis Centre



Whole genome datasets

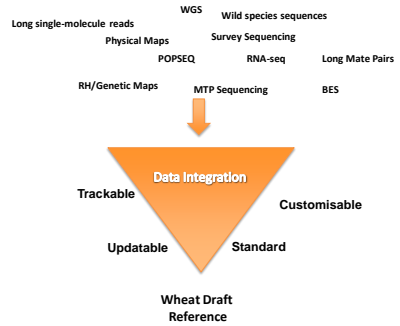


Read Type	Sequencer	Read Length	Fragment # bases Size (bp)	Fragment # bases Size (Gbps)	Coverage
Pair end	MiSeq	2 x 250 bp	600 - 800	12.9	0.76x
Pair end	HiSeq 2500	2 x 150 bp	600 - 800	131.7	7.74x
Pair end	HiSeq 2000	2 x 100 bp	600 - 800	335.3	19.73x
Mate pair	MiSeq	2 x 250 bp	6500	12.7	0.75x (9.67x*)
Mate pair	HiSeq (BGI)	2 x 100 bp	5000	6.1	0.36x (8.93x*)

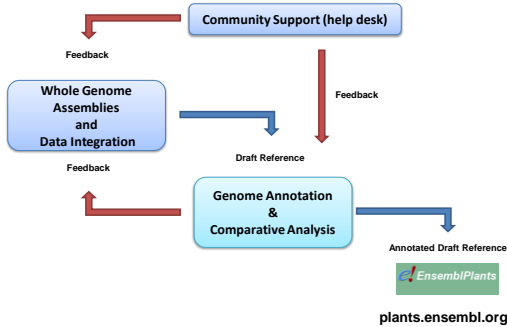
* Fragment coverage

Test assemblies	N50 (bp)	Assembly size
PE ABYSS k71 scaffolds	1,729	6.79 Gb
PE ABYSS k71 + BGI LMP (Soap) contigs		

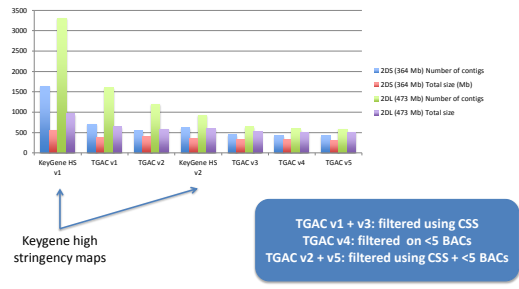
Integration of Data



Genome Reference Releases



WGP 2D* Map



2D* MTP



MTP according to the TriticeaeGenome guidelines for FPC

	2DS	2DL
Arm size (Mb)	364	473
Number of BAC clones in FPC contigs	37,634	50,687
Est. map size	364,392	590,666
Number of MTP clones	3,545	5,326
Number of MTP clones from TGAC v5*	3,180	4,737

* Removed MTP clones from excluded contigs



Fellowship Programme in Computational Biology

- Bioinformatics and Computational Biology.
- Competitive salary and research support grant.
- Fast application process, call open until posts are filled.
- Up to five years: for early career scientists who want to become scientific leaders in a dynamic research environment!

www.tgac.ac.uk/fellowship

Apply now!!!



Acknowledgements

Computational Genomics

- Sarah Ayling
- Paul Bailey
- Bernardo Clavijo
- Jon Wright

Matt Clark, Leah Clissold, Kirsten McLay and Genomics team

Mario Caetano The Genome Analysis Centre 26

