TGAC
The Genome Analysis Centre

Building Excellence in Genomics and Computational Bioscience

TGAC
The Genome Analysis Centre™
BBSRC

Integration of Whole Genome Assemblies and Genetic Information in Hexaploid Wheat

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

Wheat genome architecture

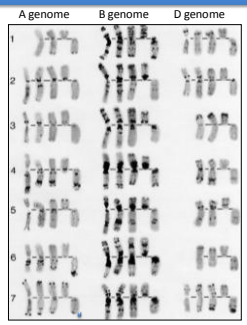
TGAC
The Genome Analysis Centre™

Large size (17,000 Mb)

Allohexaploid genome: A, B, D genomes
(3 sets of 7 homeologous chromosomes)

Homologous vs Homeologous features

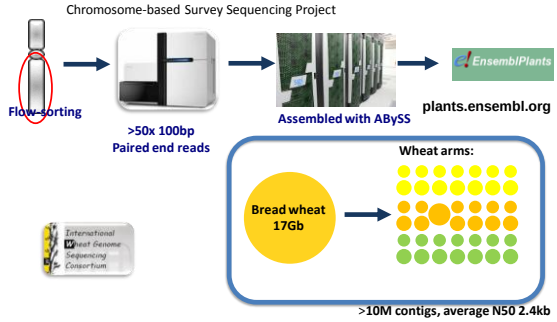
High repeat content (80-85 %)



IWGSC CSS Project

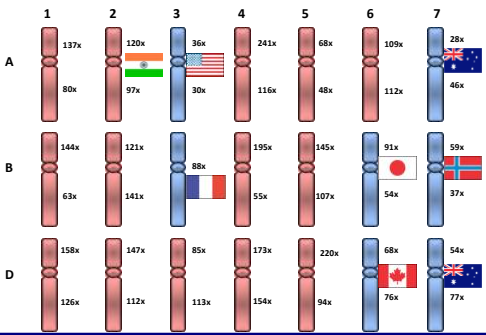
TGAC
The Genome Analysis Centre™

Chromosome-based Survey Sequencing Project



Sequencing Coverage

TGAC
The Genome Analysis Centre™



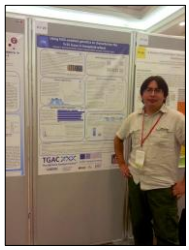
Assemblies in plants.ensembl.org

TGAC
The Genome Analysis Centre™

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

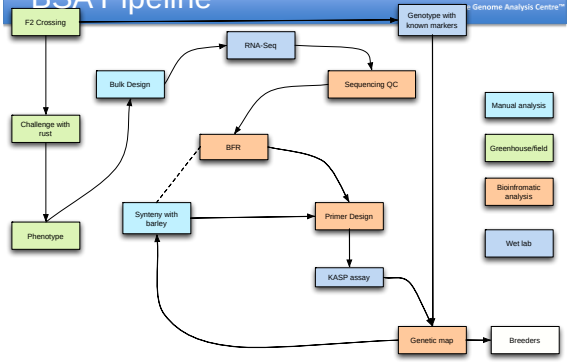


Ricardo Ramirez-Gonzalez

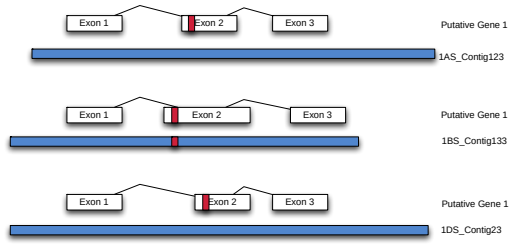
@CristobalUauy

John Innes Centre

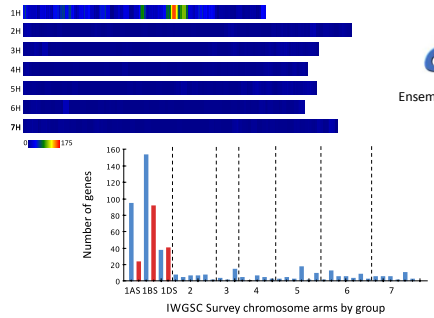
BSA Pipeline



Gene expresion in hexaploid wheat

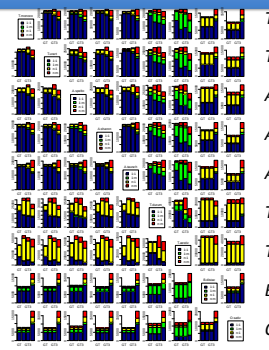


Validation and candidate selection



Ensembl DB & API

Triticeae comparative analysis



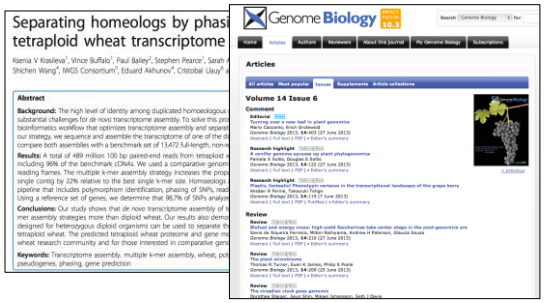
T . monococcum
T. urartu
Ae. speltoides
Ae. sharonensis
Ae. tauschii
T. durum
T. aestivum
B. distacchyon
O. sativa

2n
4n
6n
outspecies

Javier Herrero

Exome Capture

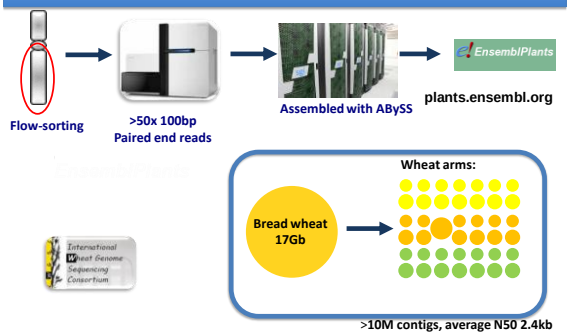
genomebiology.com/content/14/6



What we CAN'T do …

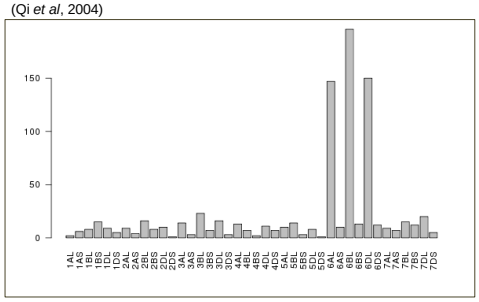
- Obtain a complete and finished genome
- Detect large re-arrangements
- Annotate regulatory elements
- "Properly" develop genome-wide studies (such as GWAS)
- Characterise CNVs, and other structural variants

Divide & conquer approach

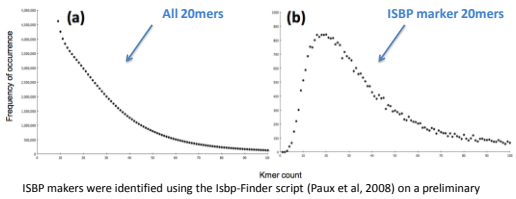


Flow-sorting purity

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



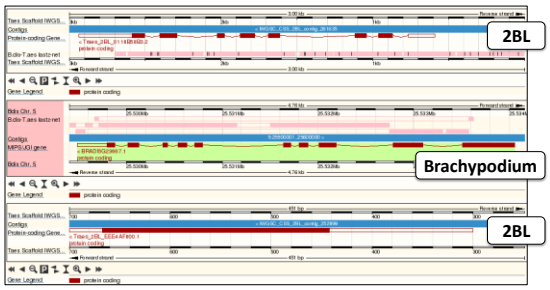
What about DNA amplification?



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



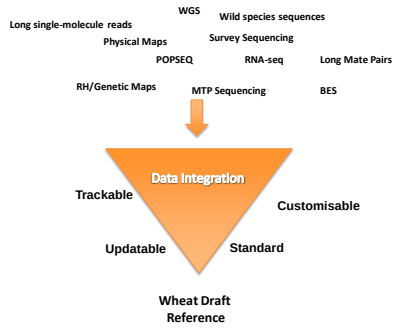
Whole genome datasets

Read Type	Sequencer	Read Length	Fragment Size (bp)	# bases (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*)

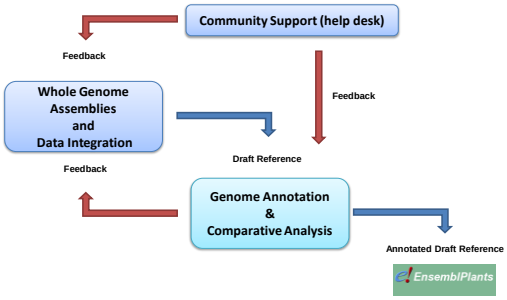
* Fragment coverage

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

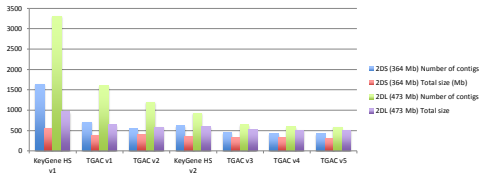


Genome Reference Releases



plants.ensembl.org

WGP 2D* Map



Keygene high stringency maps

TGAC v1 + v3: filtered using CSS
TGAC v4: filtered on <5 BACs
TGAC v2 + v5: filtered using CSS + <5 BACs

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

Acknowledgements

Computational Genomics

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

Matt Clark, Leah Clissold, Kirsten McLay and Genomics team