Science For A Better Life

Wheat Genome Structural Annotation Using a Modular and Evidence-combined Annotation Pipeline

Xi Wang

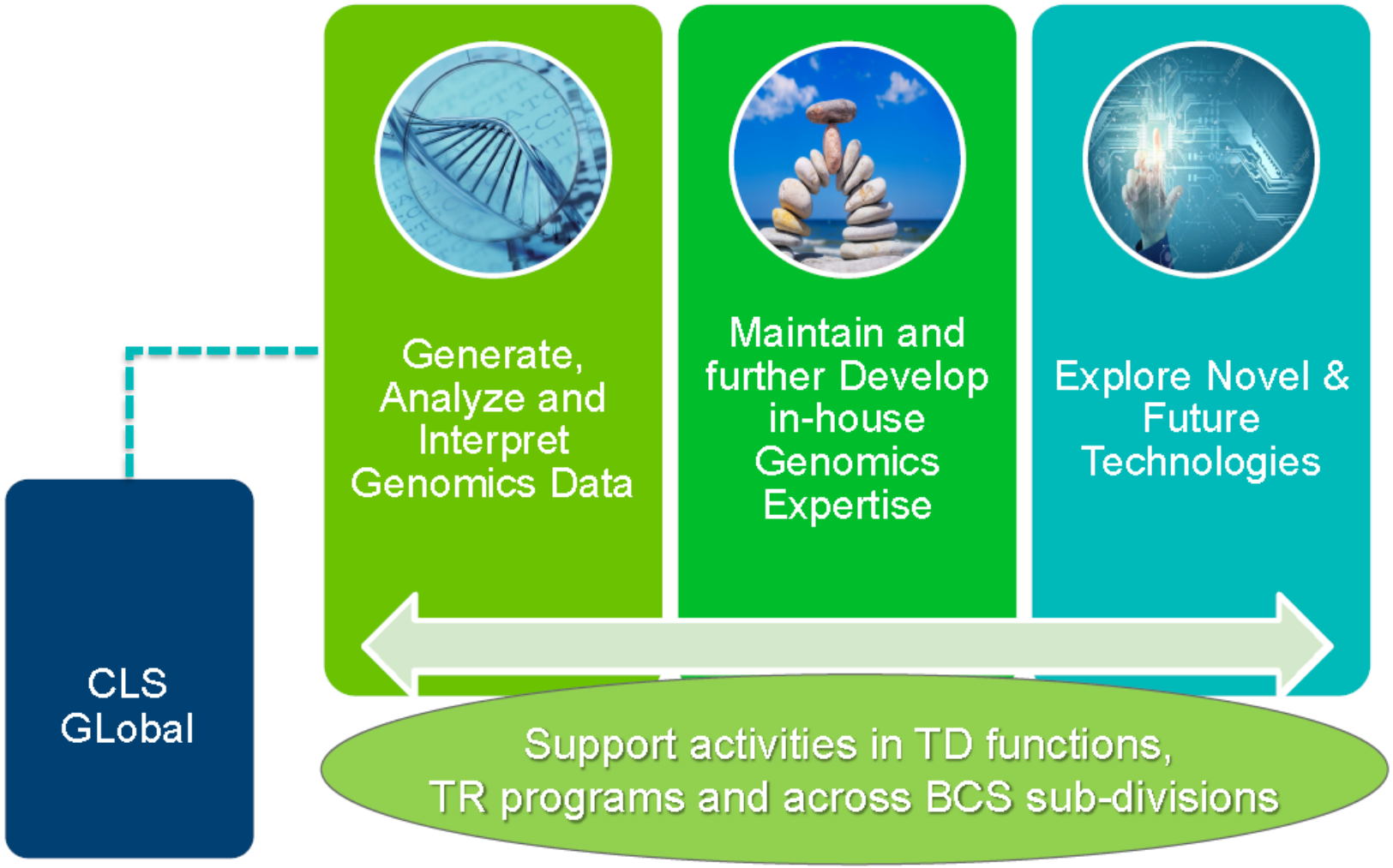
Bioinformatics Scientist

Computational Life Science

17/01/2017

Genomics

Group Mission Statement



Genomics

Resources consisting of different data types

Generate, Analyze and Interpret Genomics Data

Generation of genomics data…

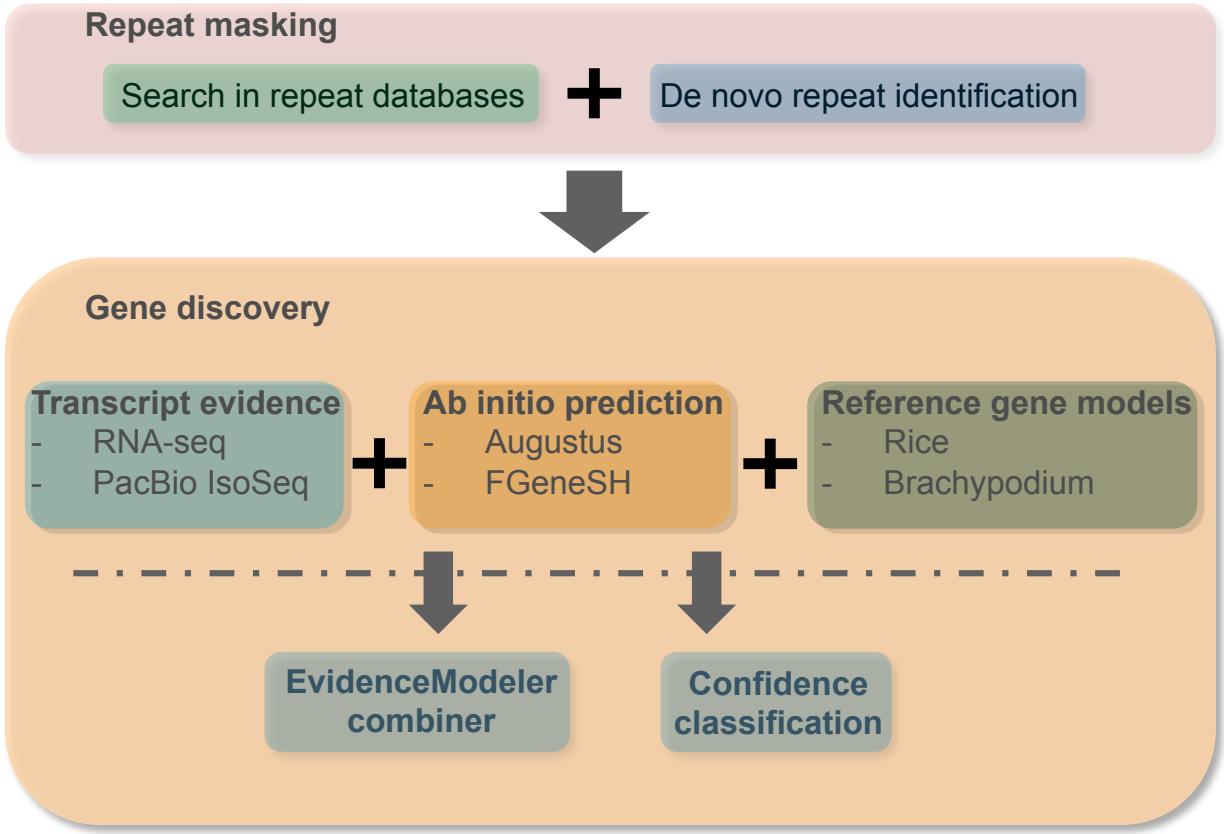
ATGGCTTCCT CTATGTTCTC CTCCACCGCT GTGGTTACCT
CCCCGGCTCA AGCCACCATG GTCGCTCCAT TCACCGGCTT
GAAGTCATCC GCTTCTTTCC CGGTCACCCG CAAGGCCAAC
AACGACATTA CTTCCATCAC AAGCAACGGA

…. is more than just about the Sequence of DNA

From candidate genes to open ended gene discovery

- Existing pipelines for genome structural annotation
  - Not suitable for complex genome, resulting in high false positive prediction rate
  - High quality, but intensive, lack of reproducibility and extensive manual work
  - → Requiring pipeline that is fast and efficient with competitive output

MEGAP – A Modular and Evidence-combined Genome Annotation Pipeline



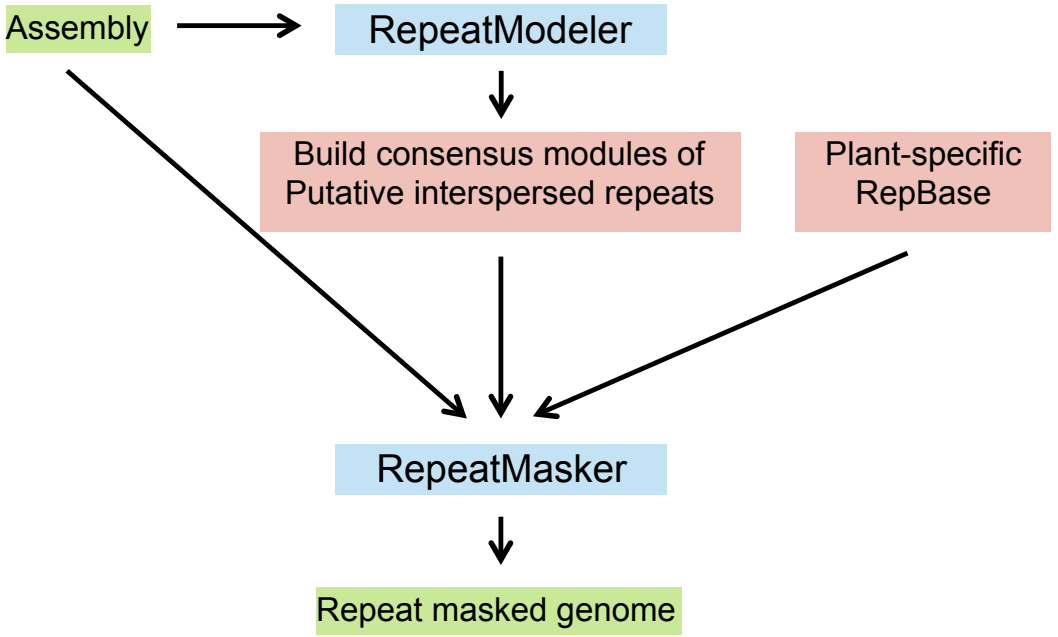
- Leveraging on strength from diverse evidence resources
  - Robustness of transcript evidence
  - Precision of exon-intron structure using computational tools
  - Support from high quality, curated reference data sets
- Confidence level classification
- efficiency
- Sustainable and reproducible
- Modularity

Wheat annotation using MEGAP – repeat masking

CS IWGSC WGA v0.4

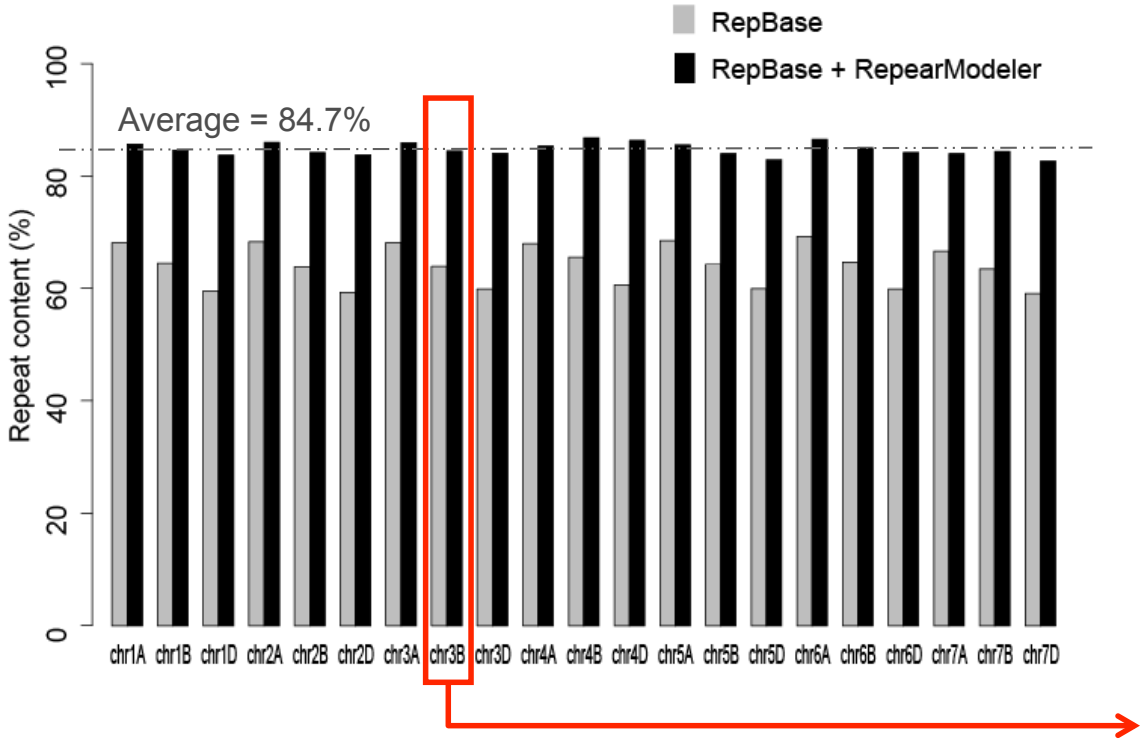
- Released on June, 2016
- Based on NRGene
- Anchoring/ordering using POPSEQ + HiC
- 21 pseudochromosomes, 14Gb

Repeat masking



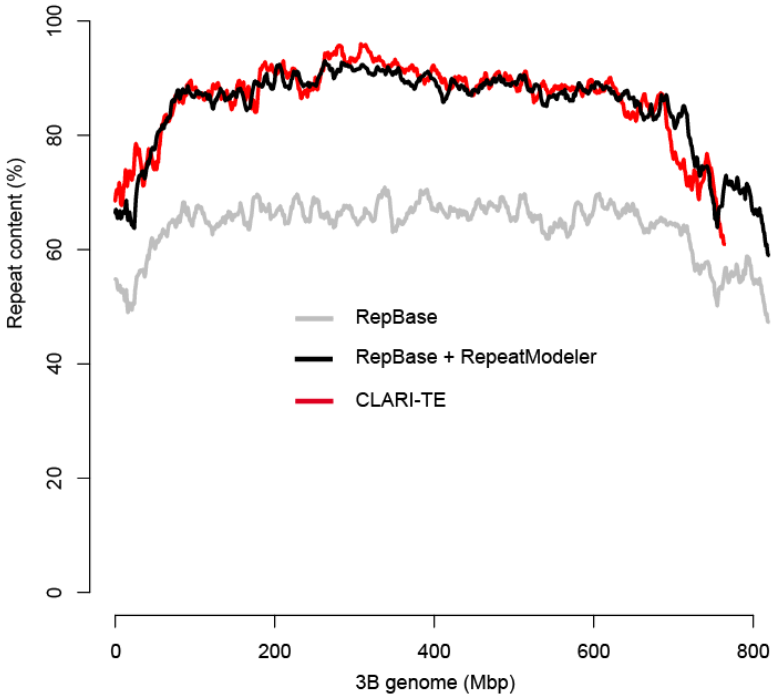
Wheat annotation using MEGAP – repeat masking

Repeat content MEGAP

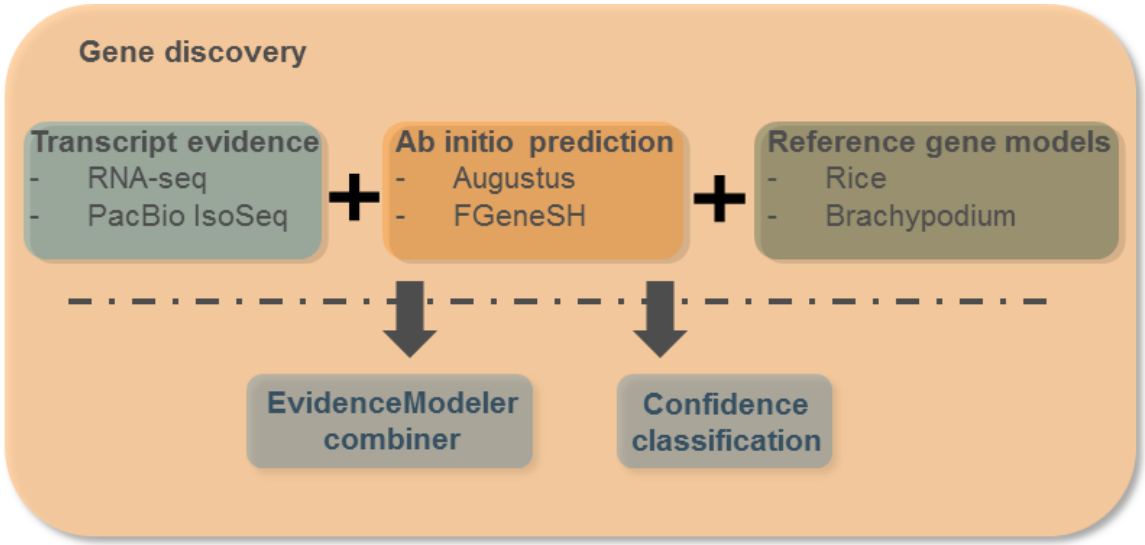


3B repeat distribution

CLARI-TE → 3B: 85%



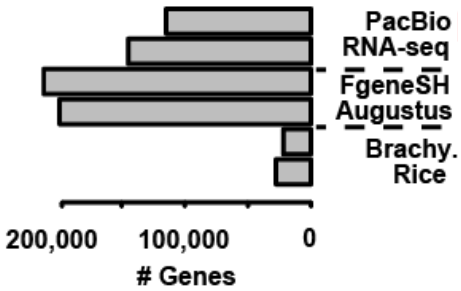
Wheat annotation using MEGAP – gene annotation



Datasets & evidence types

- **Transcriptome data:**
  - Public and in-house RNA-seq for CS and internal cultivars
  - 92 tissues and developmental stages
  - Pooled, normalized and size-selected in-house PacBio IsoSeq Libraries
- **Ab initio prediction**
  - Augustus
  - FgeneSH
- **Close related species**
  - Rice
  - Brachypodium

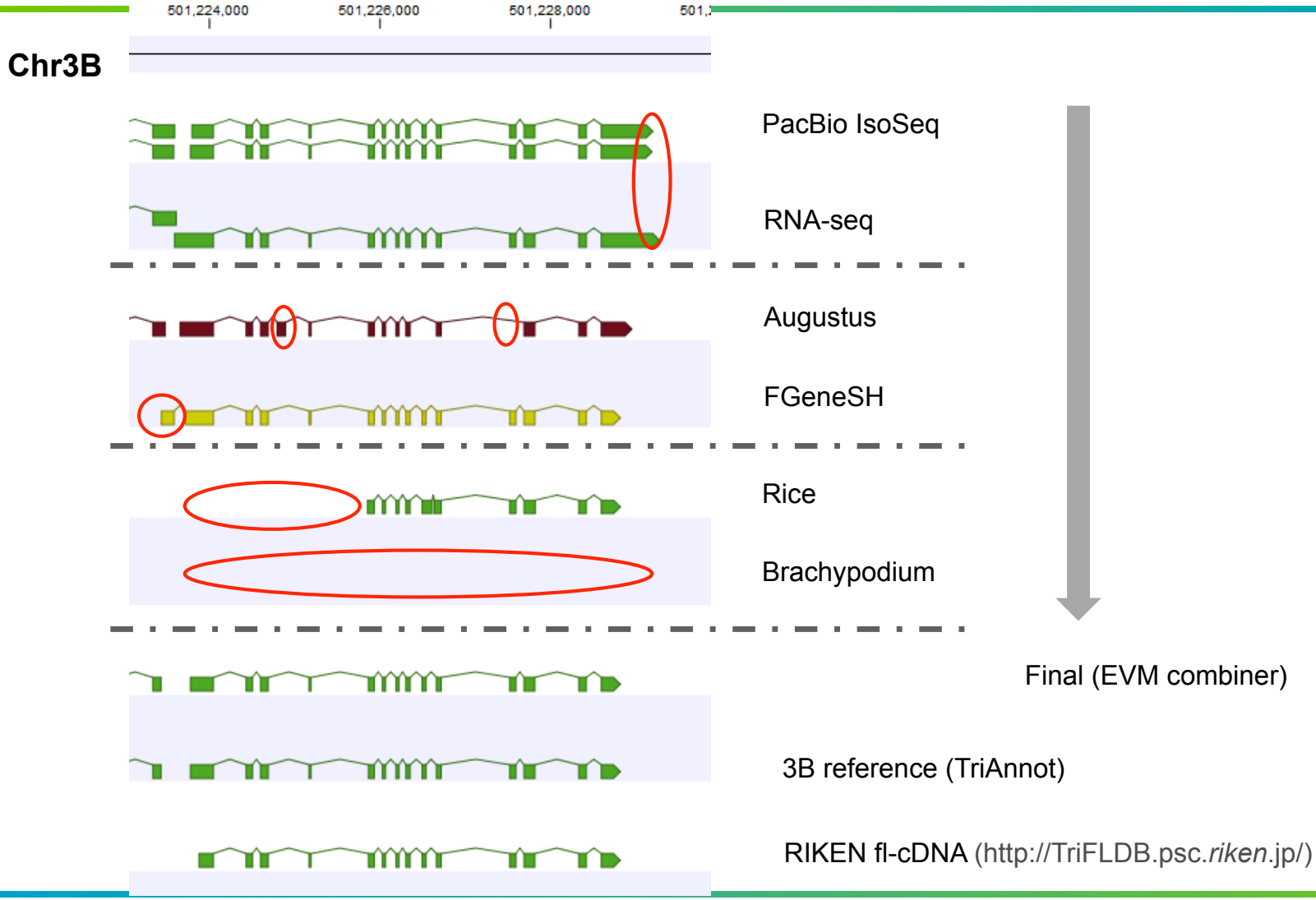
MEGAP gene annotation – evidence types, confidence levels and evidence combiner



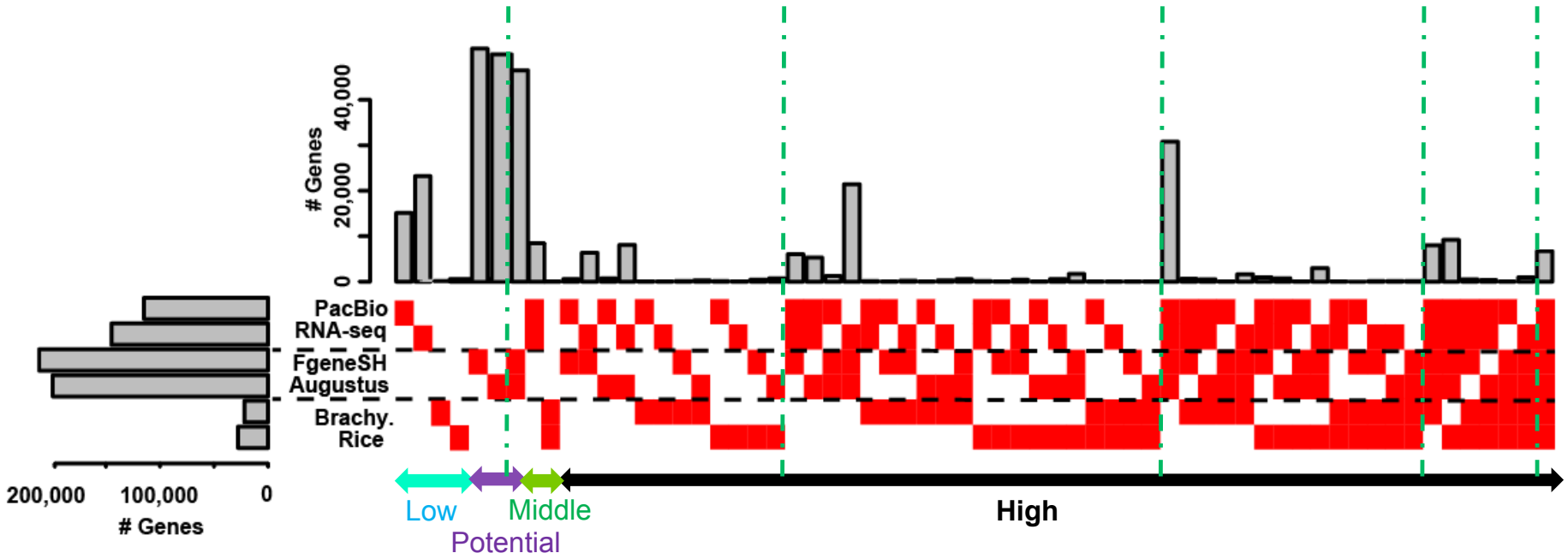
EvidenceModeler (EVM) compares and merges evidences

- Robustness of transcriptome evidence
- Precise prediction in exon-intron boundary, gene start and stop
- Evidence weighting score (transcriptome > related species > ab initio)

EVM combines evidences into final gene structure



MEGAP gene annotation – evidence types, confidence levels and evidence combiner

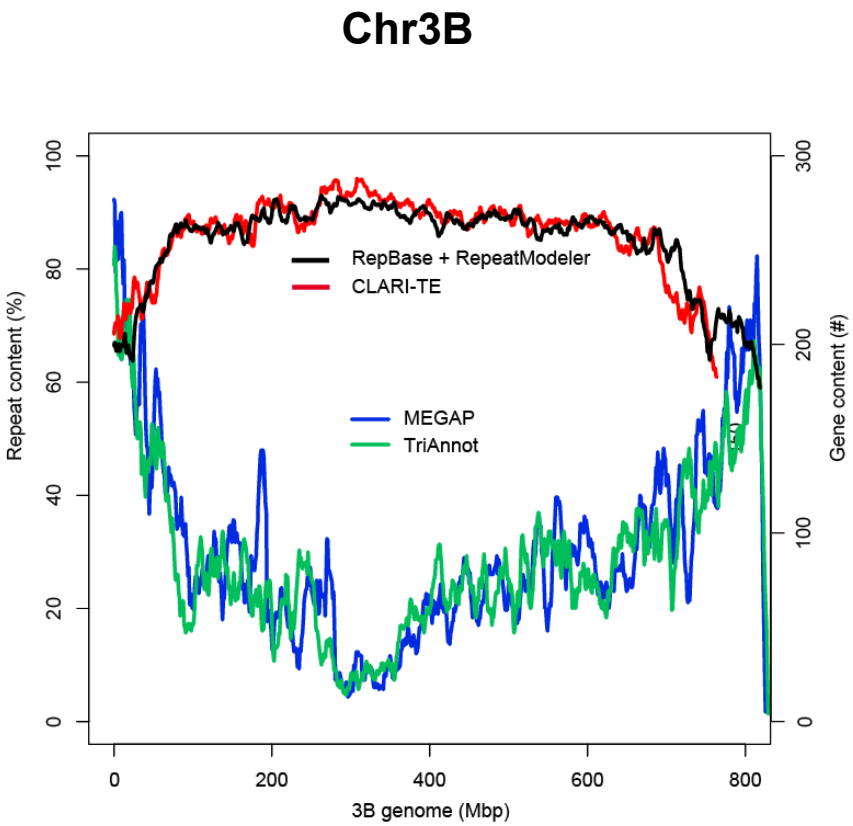


EvidenceModeler (EVM) compares and merges evidences

Confidence level	# Genes
Total supported	**164,785**
High	120,329
Middle	8,820
Low	35,636
Potential (predicted only)	125,295

Comparison of MEGAP with existing gene data sets

	CS NRGene	CS IWGSC	CS TGAC
Assembly	NRGene v0.4	Survey contigs	TGAC scaffolds
Annotation			
Provider	MEGAP	IWGSC	TGAC
Whole genome			
# Gene	129,149	103,274	104,352
mRNA length (bp; mean/ median)	1,271/1,027	1,209/977	3,146/1,912
Coding content (Mb)	164	124	328
# Predicted ORFs	113,573	99,354	103,504
3B chromosome			
# Protein coding gene	7,252	7,264	6,361
mRNA length (bp; mean/ median)	1,232/998	1,094/909	3,111/1,872
Coding content (Mb)	9	8	18



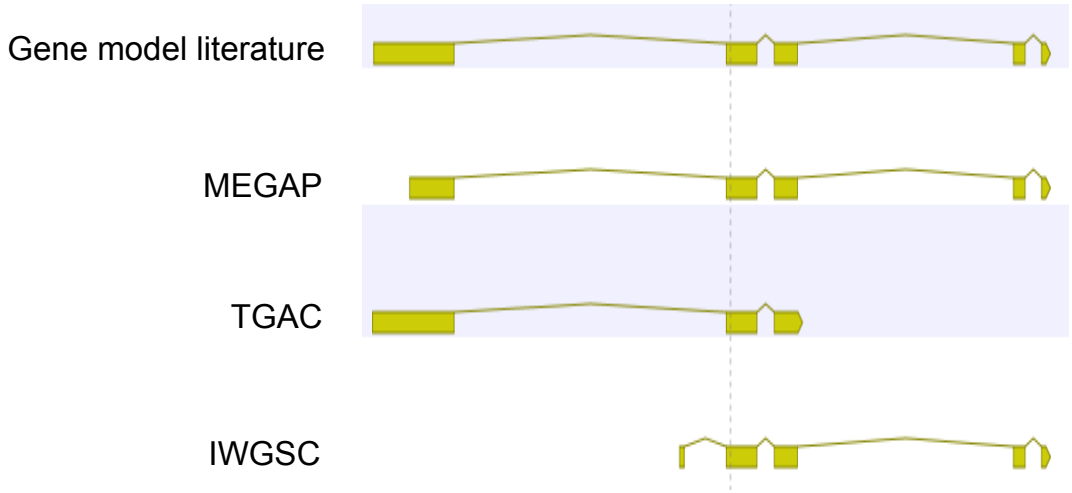
MEGAP shows higher sensitivity and accuracy in gene annotation

RIKEN fl-cDNA sequence library

(http://TriFLDB.psc.*riken*.jp/)

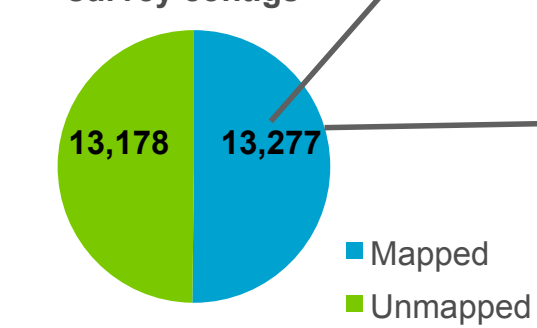
Total	2,393
# mapped to IWGSC	1,671
# mapped to TGAC	1,926
# mapped to MEGAP	**2,040**

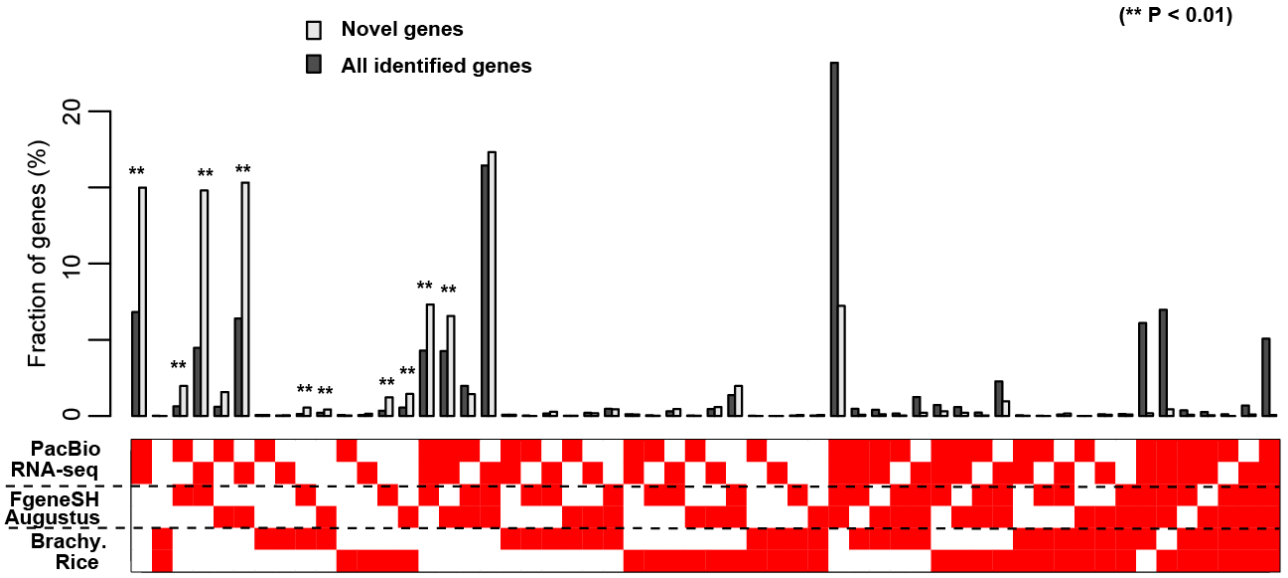
Disease resistant gene candidate on 3B

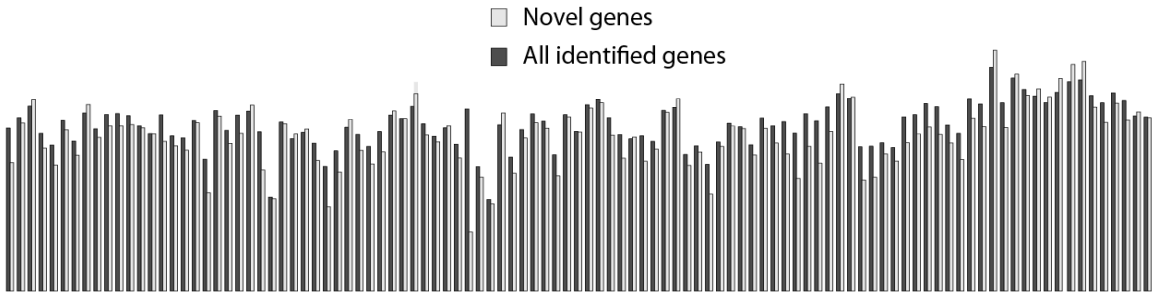


Novel genes annotated using improved assembly and combination of evidence types

	Total	Novel
IWGSC	103,274	6,454
MEGAP	129,149	26,455



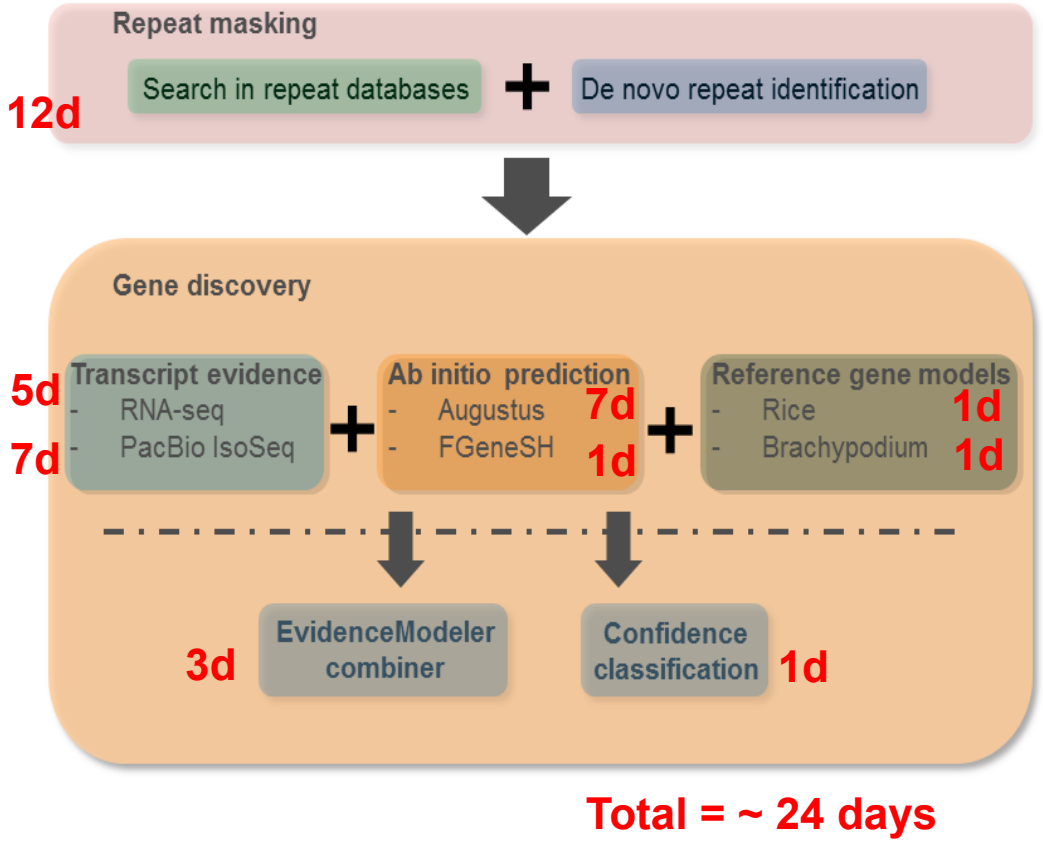




Summary

MEGAP

- Sustainable and reproducible
- Modularity – easy to extend and incorporate additional evidence, both for repeat masking and gene discovery
- Fast and efficient with competitive output
- Taking advantage of combination of
  - Robustness of transcript evidence
  - Precision of exon-intron structure using computational tools
  - Support from high quality, curated reference data sets
- Confidence level classification



Total = ~ 24 days

Summary

Key metrics wheat genome annotation using MEGAP

- Repeat content 84.7%
- Transcript loci: 164,785 of which 129,149 H/M confidence
- 1,271 bp mean length

Discussion and Perspectives

In collaboration with consortium and academia groups

- Alternative genome annotation pipelines:
  - MAKER2, FGeneSH++, BRAKER, Augustus
- Additional wheat EST data

- Annotation on wheat CS v1.0 assembly
  - Assembly was improved using BAC-related data sponsored by BAYER
  - By end of January
  - Openness of data sharing with consortium and collaboration

- Other genetic elements:
  - Isoforms, non-coding RNA, regulatory elements, promoter and enhancers

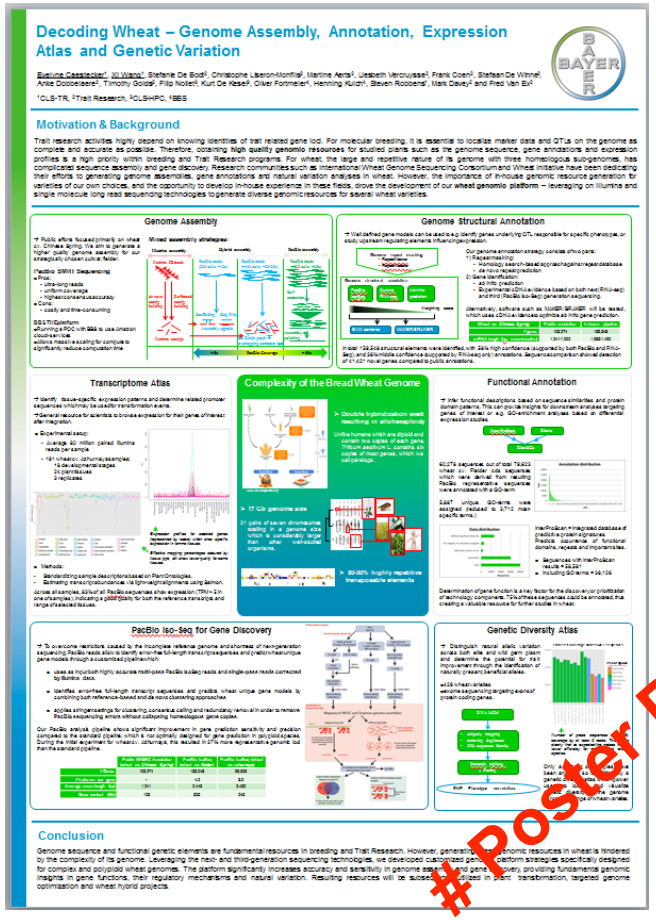
- Functional annotation:
  - Integrative approach combining homology –based inference and additional information such as QTL, expression data, genotyping-phenotyping association, etc

Acknowledgment

- **Computational Life Science group**
  - Evelyne Caestecker
  - Ken Heydrickx
  - Steven Robbens
- **Genomics & Genetics**
  - Lisbeth Vercruysse
  - Frank Coen
  - Fred van Ex
  - Mark Davey
- **HPC**
  - Filip Nollet

Decoding Wheat

- Highlights:
  - Wheat PacBio genome assembly
  - De novo gene discovery using PacBio IsoSeq data for wheat
  - Transcriptome atlas
  - Genetic diversity study using exome-capture sequencing



#Poster P0842

Science For A Better Life

Thank you!