The wheat genome visualised in Apollo: variation in the clarity of defining and naming of gene models.

Rudi Appels, Murdoch University WA,
based at AgriBio and University of Melbourne, Victoria

The work is a part of the IWGSC developing an annotation of the gene models to engage the IWGSC network of collaborators for confirming and refining automated annotation of WGA ver 1.0 and integrate functional analyses

- Platforms
- Standards
- Training
- Implementation of IWGSC annotation outputs

IWGSC-wheat genome annotation proposal

Platforms

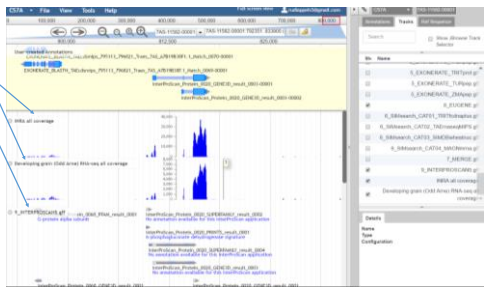
The annotation platform needs to link multiple tools, facilitate data sharing and analysis, and/or trace and record analysis pipelines while offering a clear, friendly user interface. The Apollo platform provides a dynamic environment to capture new annotations which ideally would interface with EnsemblPlants where the automated annotation of genome sequences are located.

Developing, maintaining and extending a platform such as Apollo for wheat is not a trivial process and current options include:

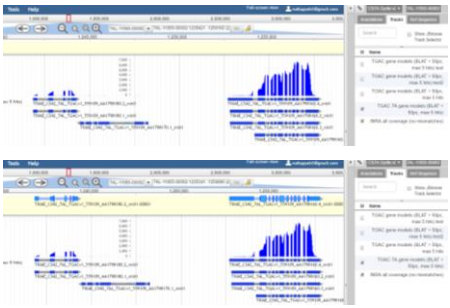
- EMBL-ABR (Melbourne, Australia)
- TGAC (Cambridge, UK)

Tracks in Apollo

- RNAseq evidence
- Capture automated annotations
- Pseudogenes
- Retrotransposable elements (Clarite)
- Proteome data (to be come)

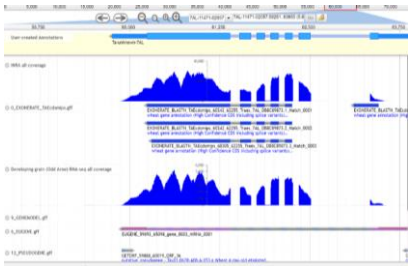


Validating gene models with available RNAseq data in Apollo



The curated gene model modified to include two 5' exons into a single gene unit.

RNAseq alignments (no mismatches) from all the tissues reported in Pingualt et al (2015) and from the grain development tissue reported in Pfeiffer et al (2014). Also shown are the gene structure predictions from several automated annotations

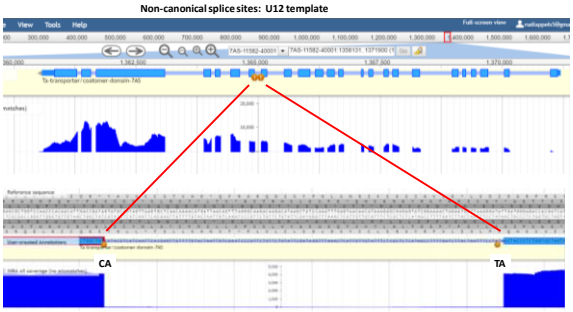


Non-canonical splice sites

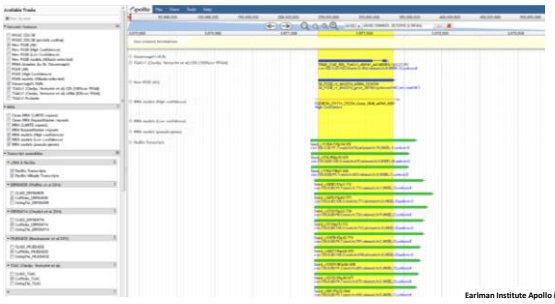
Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, International Wheat Genome Sequencing Consortium,* Klaus F. X. Mayer X, Olsen O-A (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345: DOI: 10.1126/science.1250091

Pingault L, Choulet F, Alberti A, Glover N, Wincker P, Feuillte C, Paux E (2015). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. Genome Biology 16:29 DOI 10.1186/s13059-015-0601-9.

Non-canonical splice sites: U12 template

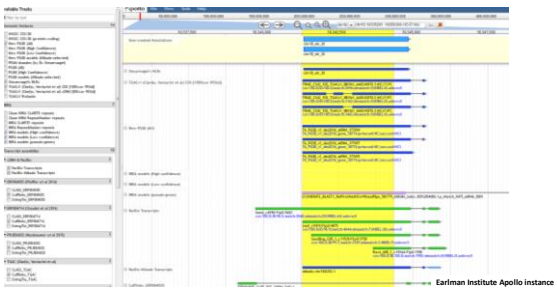


Apollo search with CDS for Grain Softness Protein (GSP in Ha locus on 5D) identifies best gene model



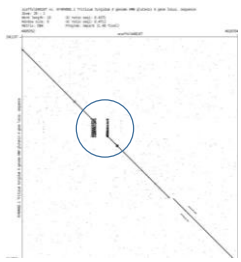
Earlman Institute Apollo instance

Sr45 CDS search in chr 1D finds nlr30



Earlman Institute Apollo instance

The Apollo process highlighted a problem in the gene structure of a well known HMW-glutenin gene. Comparison of IWGSC HMW glutenin region and published sequences indicated a problem in the assembly



Short repeat sequences in the HMW glutenin gene sequences are missing from the IWGSC-refv1 assembly

Alignments to published Chinese Spring HMW GS sequences (O Anderson 2009; Thompson et al 1985)

Standards

International efforts addressing data heterogeneity challenges are taking place and the wheat community needs to make use of developments such as the Interoperability Platform, drawing together a group of experts drawn from across Europe, in ELIXIR. The Interoperability Platform is guided by the FAIR data principles, which state that data must be Findable, Accessible, Interoperable, and Re-usable. As presented by ELIXIR, these principles mean:

- Findable: data must be easy to find by both humans and computer systems. For this to happen we need to describe the data with metadata that includes a unique, persistent identifier, and make the data available in a searchable resource. An agreed naming of gene models in the advanced drafts of the wheat genome by the wheat community is critical.
- Accessible: data must be put in long-term storage in such a way that either the data itself or its metadata can be accessed easily. This access can either be open or with a well-defined license.
- Interoperable: datasets can be combined by humans as well as computer systems. Data formats use shared vocabularies and/or ontologies.
- Re-usable: data can be used for future research and to be processed further by computer programs. Metadata identifies the provenance of the data.

Note: Data Commons, ELIXIR set of core resources, ELIXIR Tools and Data Services Registry and BioSharing.

Gene catalogue for annotation and locations of loci

CATALOGUE OF GENE SYMBOLS FOR WHEAT: 2013-2014 SUPPLEMENT R.A. McIntosh, J. Dubcovsky, W.J. Rogers, C. Morris, R. Appels and X.C. Xia

The University of Sydney, Plant Breeding Institute Cobbitty, PMB 4011, Narellen, N.S.W. 2570, Australia.
robert.mcintosh@sydney.edu.au

Wheat Initiative Expert Working Group (EWG)

This EWG is aimed at maintaining and improving wheat quality and safety under varying environmental conditions. Our expert group will focus on wheat quality and safety in the broad sense, including seed proteins, allergens, carbohydrates, and nutrition quality including micronutrients, grain processing, food safety, genetic resources and gene nomenclature. We will also share genetic resources and unify gene nomenclature related to grain quality

http://www.wheatinitiative.org/sites/default/files/attached_file/ewg_improving_wheat_quality_annual_report_2015.pdf

17. Dormancy (Seed)
17.1. Vivipary

Vp-A1g [[11047]].	*Vp-A1b* [11047].	v:	Kayamona [11047]; Sonalika [11047]; Yaqui 50 [11047]; Yecora Rojo 76 [11047].
		e:	GenBank Gu385899 [11047].
Vp-A1b [[11047]].	*Vp-A1d* [11047].	v:	Attila [11047]; Glenlea [11047]; Tanori F71 [11047].
		e:	GenBank Gu385901 [11047].
Vp-A1i [[11047]].	*Vp-A1f* [11047].	v:	Debeira [11047]; Kancahn [11047]; Rayon F89 [11047].
		e:	GenBank Gu385903 [11047].
Vp-1Be [10998].	v:	Fulingkemai [10999]; Hongmangchan [10998]; Wangshuibai [10999].	
Vp-1Bf [10998].	v:	Wanxanbaimaizi [10998].	

IWGSC-wheat genome annotation proposal

Project needs to interface with platforms used by other genome projects

- CyVerse provides life scientists with computational infrastructure to handle datasets and complex analyses, thus enabling data-driven discovery.
- Ensembl Plants is a genome-centric portal for plant species of scientific interest. It is developed by EMBL-EBI and is powered by the Ensembl software system for the analysis and visualisation of genomic data
- EMBL-ABR – wheat project focused on Apollo platform

Training

An interactive environment such as Apollo requires users to be registered and participate in an induction process that includes a naming of genes schema that is agreed to by the community.

The training process would promote and proactively support the exchange of information between the centres and other national and international efforts and entities, in academia as well as industry. The process provides the means to coordinate the development of standards, and common processes and procedures, including best practices when it comes to documentation and traceability of methods and workflows.

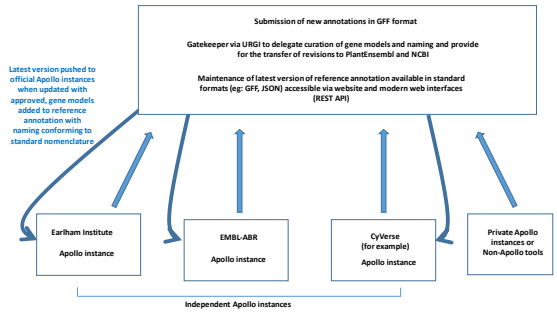
Local expertise centres distributed across the IWGSC network would facilitate the implementation of the full data life cycle, from data discovery, through storage, processing, analysis, interpretation, and visualisation to publication. These centres would generally be known for their level of contribution and resources in the analysis of specific aspects of the wheat genome.

In terms of researcher skills, the findings from the UK BBSRC on People and Skills showed that there is a need for increase in skills *"across all career stages when it comes to basic skills in scripting, coding and bioinformatics. Applied skills in mathematical modelling, applied statistics (experimental design) and data management, including data visualisation are required by all researchers and should be the focus of efforts."*

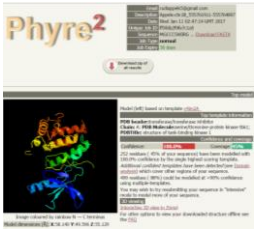
IWGSC-wheat genome annotation proposal

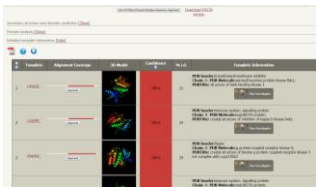
Implementation of an IWGSC annotation network

- Advancements and developments occurring in bioinformatics (e.g. ELIXIR, BD2K, CyVERSE, Corbel).
- Ensure the links with industry are developed and future partnership opportunities between academia and industry are viable
- NCBI, EMBL-EBI for making community annotation (automated and manual) broadly available
- Dedicated resource to manage and lead the network.

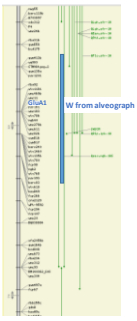


Large-scale functional annotation using Phyre2



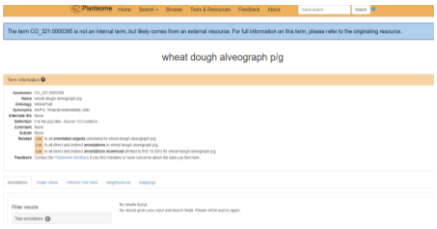


Lawrence A. Kelley,Structural Bioinformatics Group,Department of LifeSciences,Imperial College London



Planteome framework to describe traits for functional annotation

http://planteome.org/node/106; contact Pankaj Jaiswal



Steps in annotation:

- **Automated annotation – utilized outputs from several systems**
- **Confirm annotations manually using RNAseq alignments (missing exons, accuracy of intron-exon junctions, start of 3'-UTR and 5'UTR)**
- **Functional annotations (community experts, Phyre2, InterProScan)**