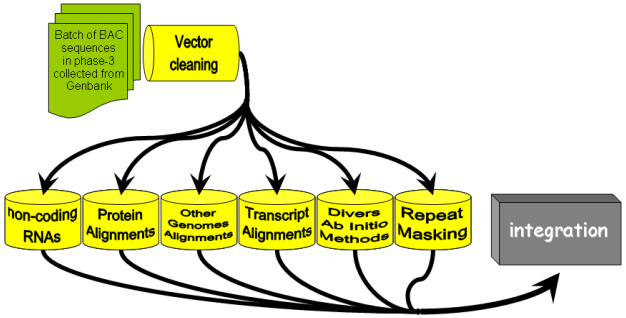
1st transPLANT user training workshop

November 13th, 2012

INRA Versailles



Annotating *Triticeae* sequences: the TriAnnot pipeline

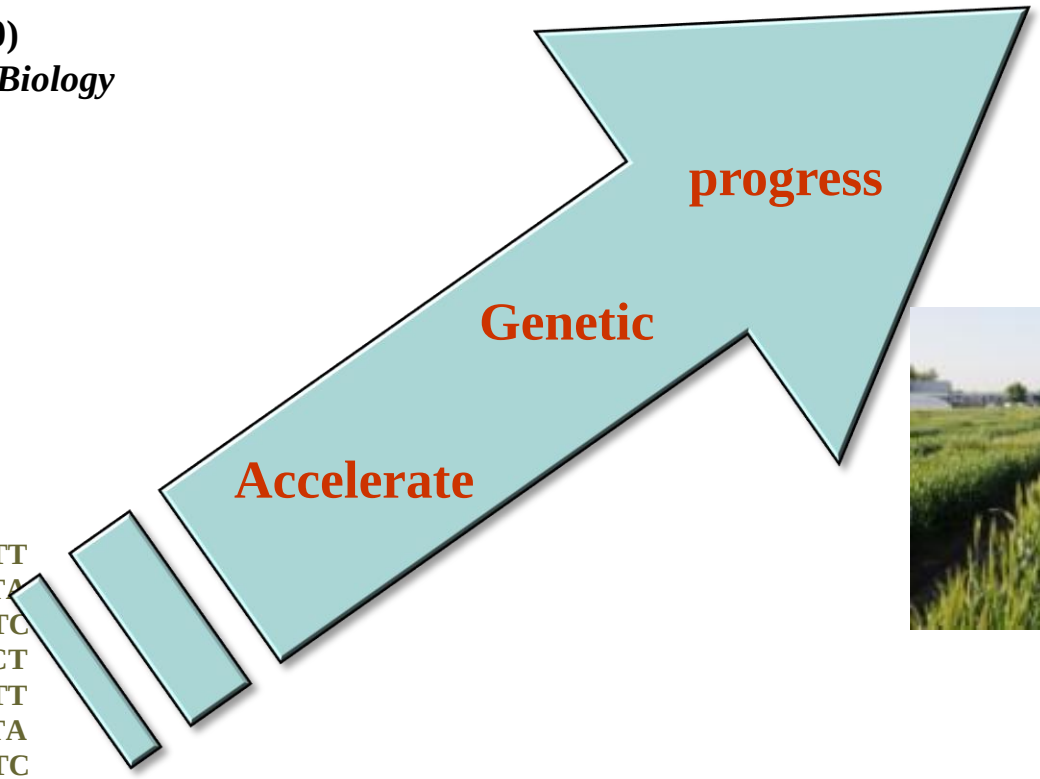
Leroy P, Guilhot N, Sakai H, Bernard A, Theil S, Choulet F, Reboux S, Viseux C, Amano N, Giacomoni F, Alaux M, Legeai F, Cerutti L,

Itoh T, Quesneville H, Feuillet C

Plant Genomics

“Today genomics is the primary driver for unification in biological science”

Cook and Varshney (2010)
Current Opinion in Plant Biology
13:115-118.



CCCATGCGCGAATCGATCGATTT
CTGACCATAGCTAGACTAGTCTA
GGCGCTAGATCGATCGATCGATC
GATTTCAGACTGGAAGTCATGCT
CCCATGCGCGAATCGATCGATTT
CTGACCATAGCTAGACTAGTCTA
GGCGCTAGATCGATCGATCGATC
GATTTCAGACTGGAAGTCATGCT

Genome Annotation

"**Annotation is one of the most difficult tasks in genome sequencing projects, yet it is essential for connecting genome sequence to biology**"

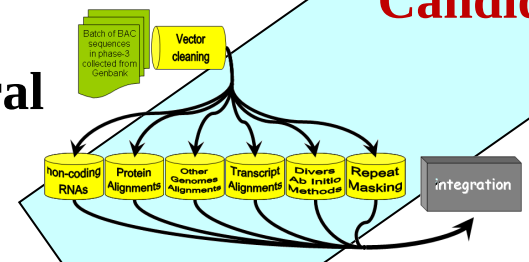
Elsik *et al*. 2006 Genome Research

Accelerate genetic progress

Markers

Gene Candidates

Structural



Functional

CCCATGCGCGAATCGATCGATTT
CTGACCATAGCTAGACTAGTCTA
GGCGCTAGATCGATCGATCGATC
GATTTCAGACTGGAAGTCATGCT
CCCATGCGCGAATCGATCGATTT
CTGACCATAGCTAGACTAGTCTA
GGCGCTAGATCGATCGATCGATC
GATTTCAGACTGGAAGTCATGCT

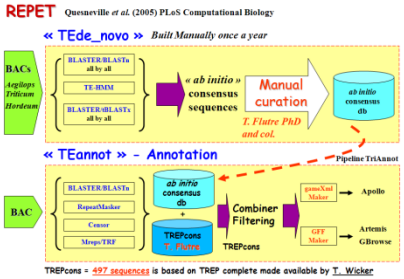
transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

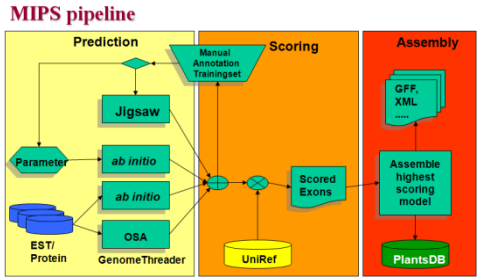
INRA

Pipelines

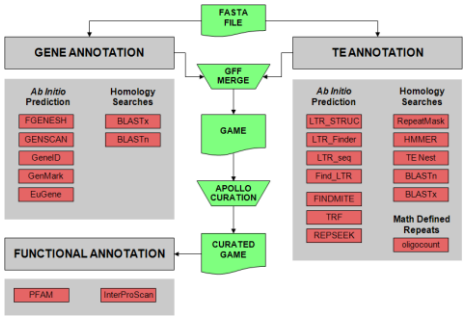
REPET



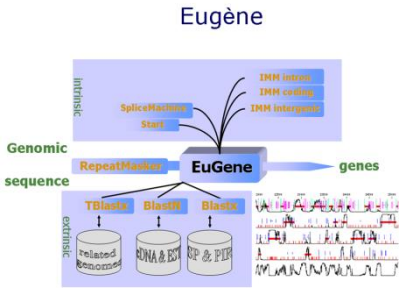
MIPS pipeline



DAWG-PAWs



Eugène

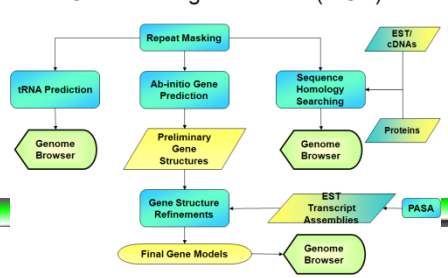


PASA

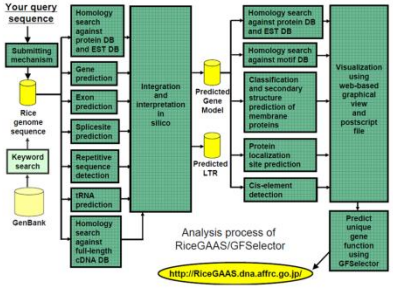
Gnomon

FPGP

TIGR



RiceGAAs



transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

FranceAgriMer

Funding from the
European Community's
FP7 Programme - GA
FP7-212019

transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

The International Wheat Genome Sequencing Consortium

www.wheatgenome.org

Launched in 2005 on the initiative of Kansas Growers

Executive director
K. Eversole

Cochairs
R. Appels
J. Dvorak
C. Feuillet
B. Gill
B. Keller
Y. Ogihara

Sponsors (23)

Coordinating Commitee (64)

General members (>500)

transPLANT

Leroy *et al.* (2012) Frontiers in Plant Science 3:1-14

frontiers in
PLANT SCIENCE

METHODS ARTICLE
published: 31 January 2012
doi: 10.3389/fpls.2012.00005

TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes

Philippe Leroy[1], Nicolas Guilhot[1], Hiroaki Sakai[2], Aurélien Bernard[1,3], Frédéric Choulet[1], Sébastien Theil[1], Sébastien Reboux[4], Naoki Amano[2,5], Timothée Flutre[4], Céline Pelegrin[1], Hajime Ohyanagi[6,7], Michael Seidel[8], Franck Giacomoni[9], Mathieu Reichstadt[10], Michael Alaux[4], Emmanuelle Gicquello[1], Fabrice Legeai[11], Lorenzo Cerutti[12], Hisataka Numa[2], Tsuyoshi Tanaka[2], Klaus Mayer[8], Takeshi Itoh[2], Hadi Quesneville[4] and Catherine Feuillet[1]**

TriAnnot Architecture
Modular

N. Guilhot
GDEC

Initial sequence

Panels

I — **Transposable Elements**

TEannot (TREPcons) — BLASTx (TREPprot) — k-mer frequency (CS 3B 2X)

RepeatMasker Masking

II — **Protein Coding Genes**

ab initio
FgeneSH, Augustus
GeneMarkHMM, geneid

Similarity
Exonerate / Gmap
Plant FL-cDNA, mRNA & ESTs
Exonerate
Plant proteomes

Gene Modeling
Transcripts
SIMsearch
ab initio & Transcripts
EuGene
ab initio
Augustus

Merge

Functional Annotation
known function
putative function
domain containing protein
expressed sequence
conserved unknown function
hypothetical protein
IWGSC guideline

InterProscan (Pfam, Prosite, Smart, GO)

Best Hit (plant proteomes)

Panel I & II masking

III — **Non Coding Sequences**

Conserved NCS
BLASTn: *A. thaliana*, *O. sativa*, *Z. mays*, *S. bicolor*, *B. distachyon*, plastids & mitochondria, EMBL plant wgs / tsa
BLASTx: SwissProt, TrEMBL, Uniref, nr, NCBI RefSeq proteins, Plant proteomes

ncRNAs
tRNAscan

IV — **Molecular Markers**

Microsatellites (SSRs)
Tandem Repeat Finder (TRF)

Update
112 databanks
21 programs

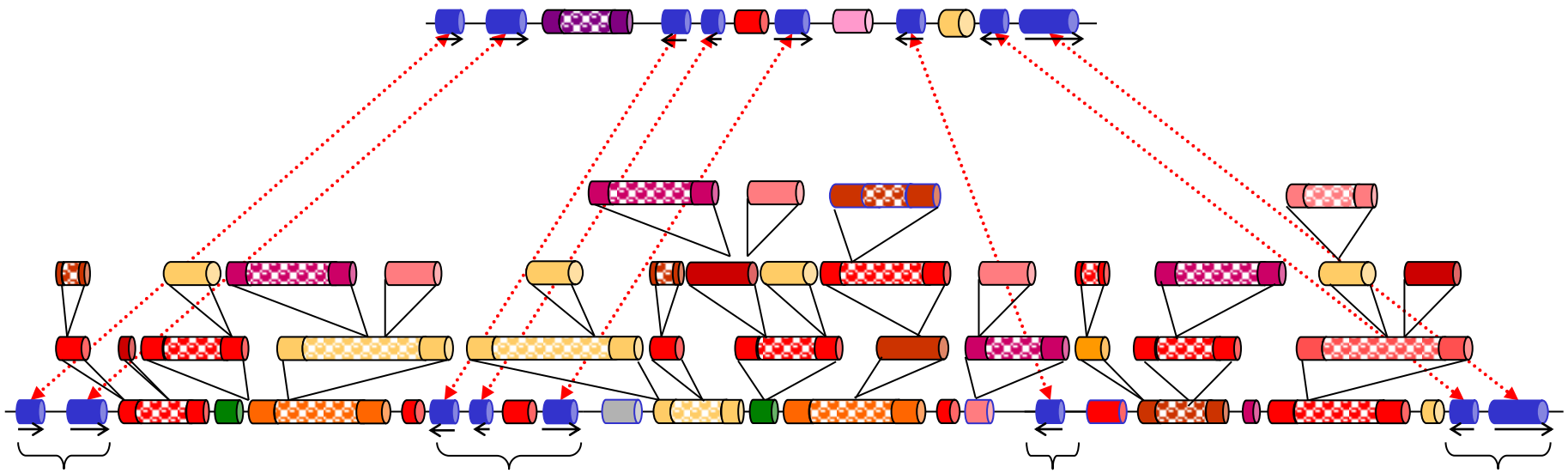
Wheat
Barley
Rice
Oak

transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

Transposable Elements



Erayman et al., 2004

To be masked

Transposable Elements “code”

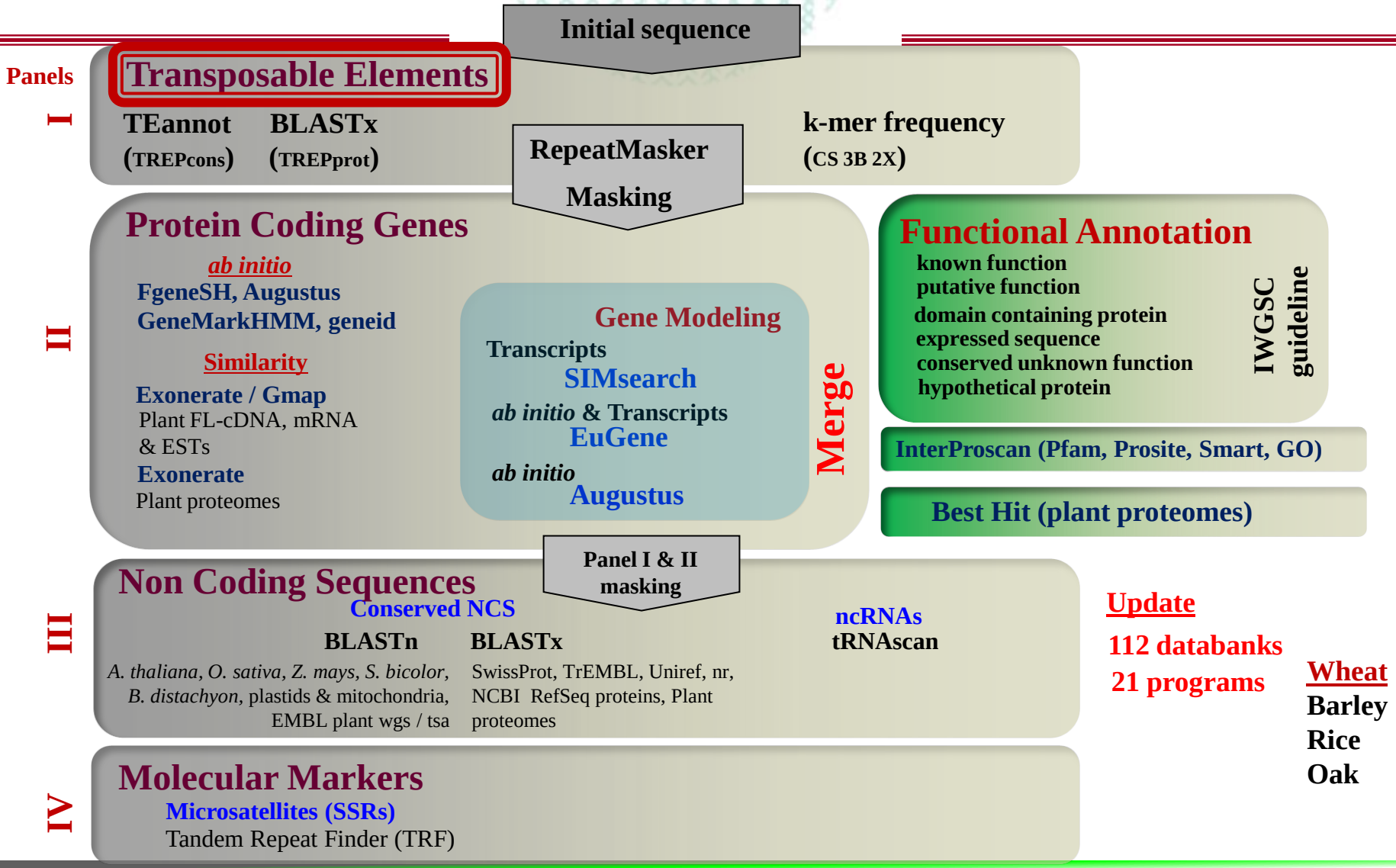
Wicker *et al.* (2007)

A unified classification system for eukaryotic transposable elements.

Nature Reviews Genetics 8:973-982

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

TriAnnot Architecture



transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

INRA

Protein-coding gene models

Similarity based

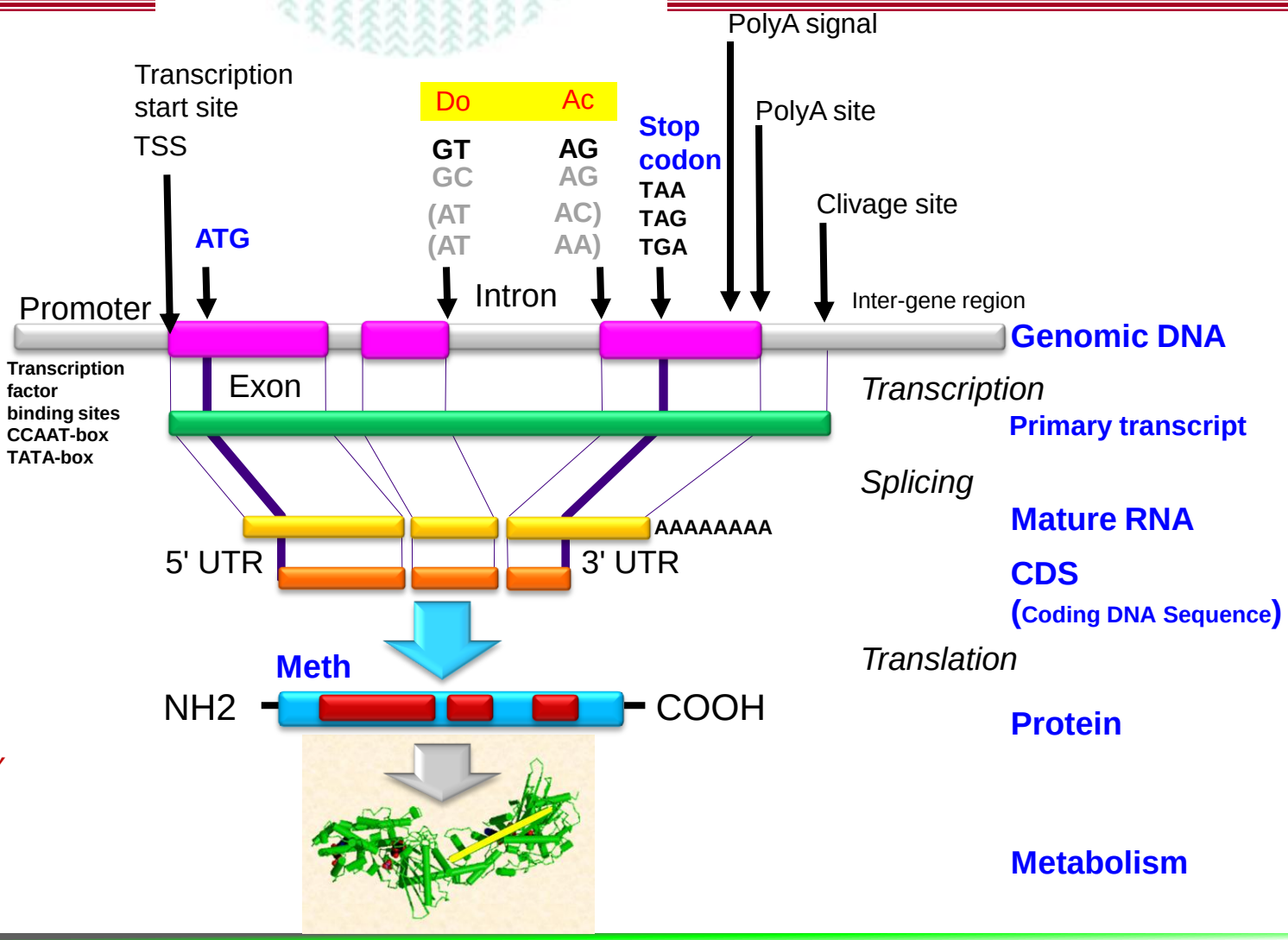
Biological evidences
"Extrinsic approach"

Combiner

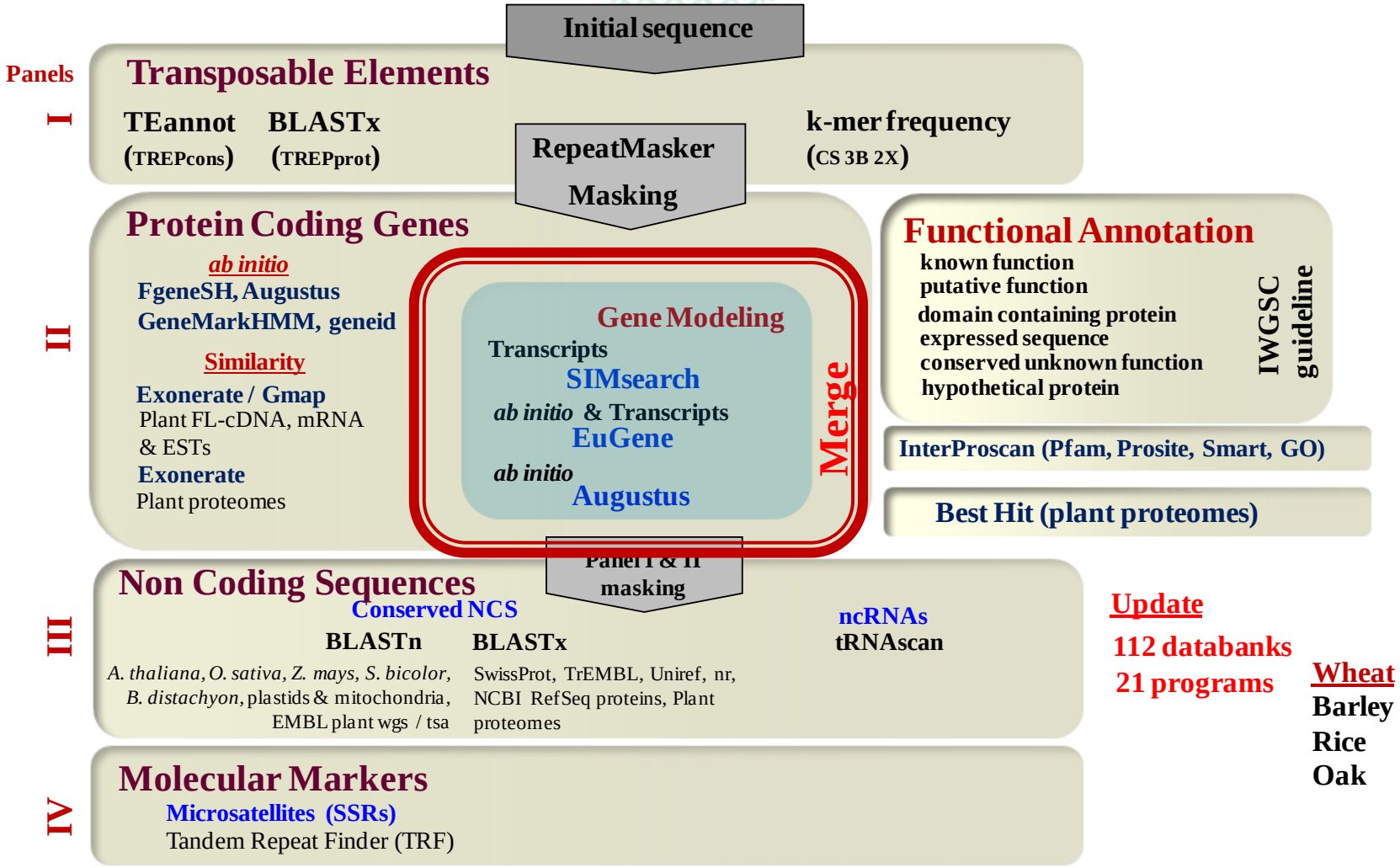
Biological evidences & training data set

ab initio

Training data set
"Intrinsic approach"



Use a color code system to assess the confidence of the structural annotation



Structural Annotation

5 Categories

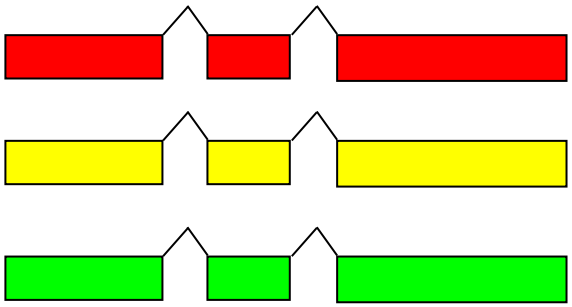
Similarity

SIMsearch

Cat 1

Cat 2

Cat 3



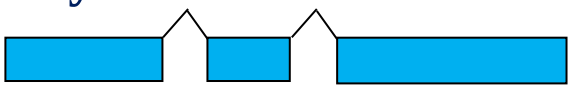
Wheat FL-cDNAs + wheat RNAseq

CDS-genes derived from reference genomes annotation (*Os-irgsp, Bd*)

Poaceae FL-cDNAs

ab initio & Similarity

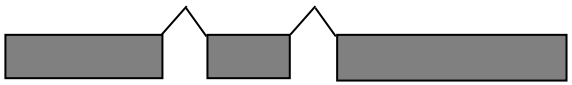
Cat 4



EuGene Combiner :
Augustus + wheat-ESTs + SIMnuc + SIMprot

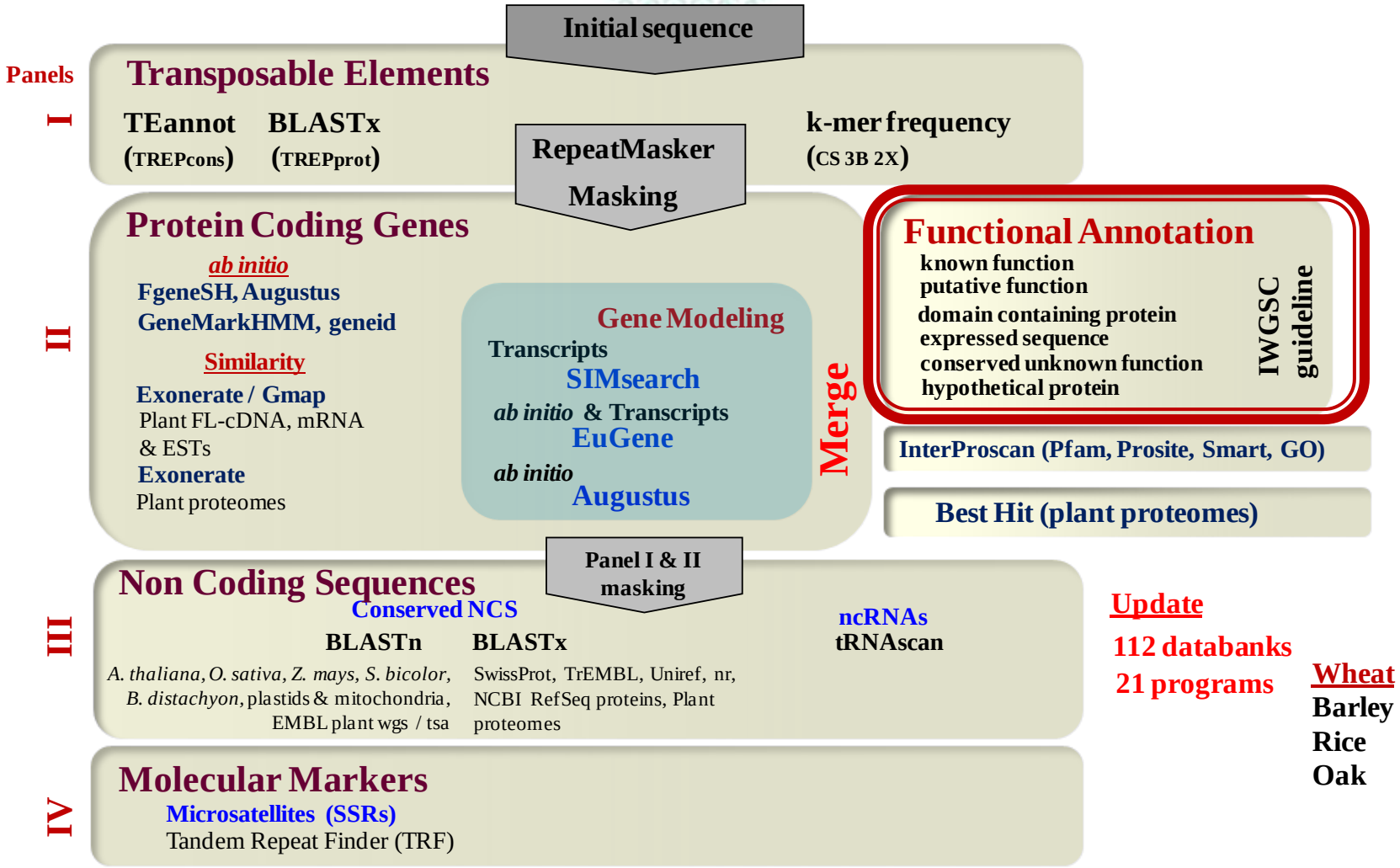
ab initio

Cat 5



Augustus
ab initio gene prediction only

Use a color code system to assess the confidence of the structural annotation

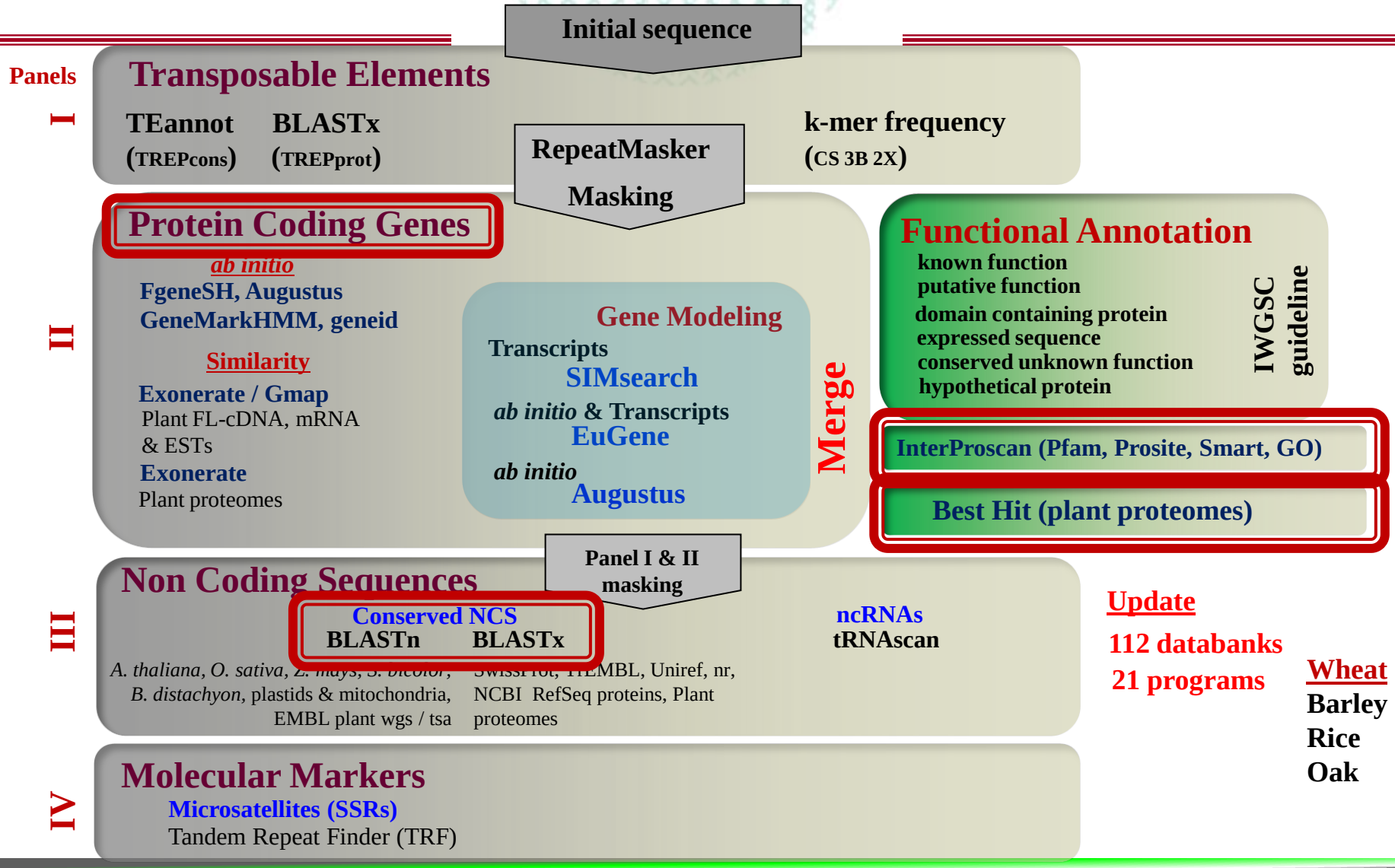

Initial sequence
Panels
I
Transposable Elements
TEannot (TREPcons)
BLASTx (TREPprot)
RepeatMasker Masking
k-mer frequency (CS 3B 2X)
II
Protein Coding Genes
ab initio
FgeneSH, Augustus
GeneMarkHMM, geneid
Similarity
Exonerate / Gmap
Plant FL-cDNA, mRNA & ESTs
Exonerate
Plant proteomes
Gene Modeling
Transcripts
SIMsearch
ab initio & Transcripts
EuGene
ab initio
Augustus
Merge
Functional Annotation
known function
putative function
domain containing protein
expressed sequence
conserved unknown function
hypothetical protein
IWGSC guideline
InterProscan (Pfam, Prosite, Smart, GO)
Best Hit (plant proteomes)
Panel I & II masking
III
Non Coding Sequences
Conserved NCS
BLASTn
A. thaliana, O. sativa, Z. mays, S. bicolor, B. distachyon, plastids & mitochondria, EMBL plant wgs / tsa
BLASTx
SwissProt, TrEMBL, Uniref, nr, NCBI RefSeq proteins, Plant proteomes
ncRNAs
tRNAscan
IV
Molecular Markers
Microsatellites (SSRs)
Tandem Repeat Finder (TRF)
Update
112 databanks
21 programs
Wheat
Barley
Rice
Oak

Functional Annotation

6 Classes

known-function	**>80% identity** **>80% coverage**	**UniProtKB/Swiss-Prot**
putative-function	**>45% identity** **>50% coverage** ***Not annotated as putative or hypothetical***	**UniProtKB/Swiss-Prot** **UniProtKB/TrEMBL**
domain-containing-protein		**Pfam**
expressed-sequence	**>45% identity** **>50% coverage**	**ESTs databanks**
conserved-unknown-function	**>45% identity** **>50% coverage** ***annotated as putative or hypothetical***	**UniProtKB/Swiss-Prot** **UniProtKB/TrEMBL**
hypothetical-protein	**no known function**	

TriAnnot Architecture



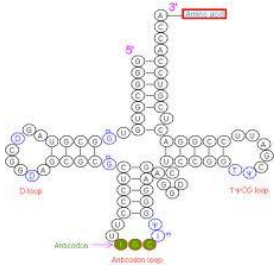
transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

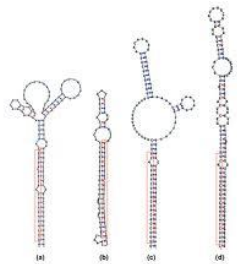
INRA

Non-coding RNAs

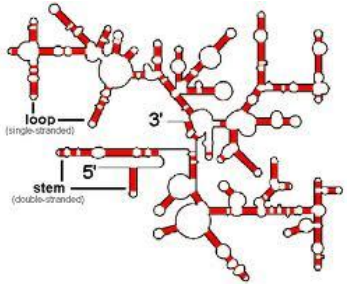
* tRNAs



*miRNA



* rRNA



Rfam

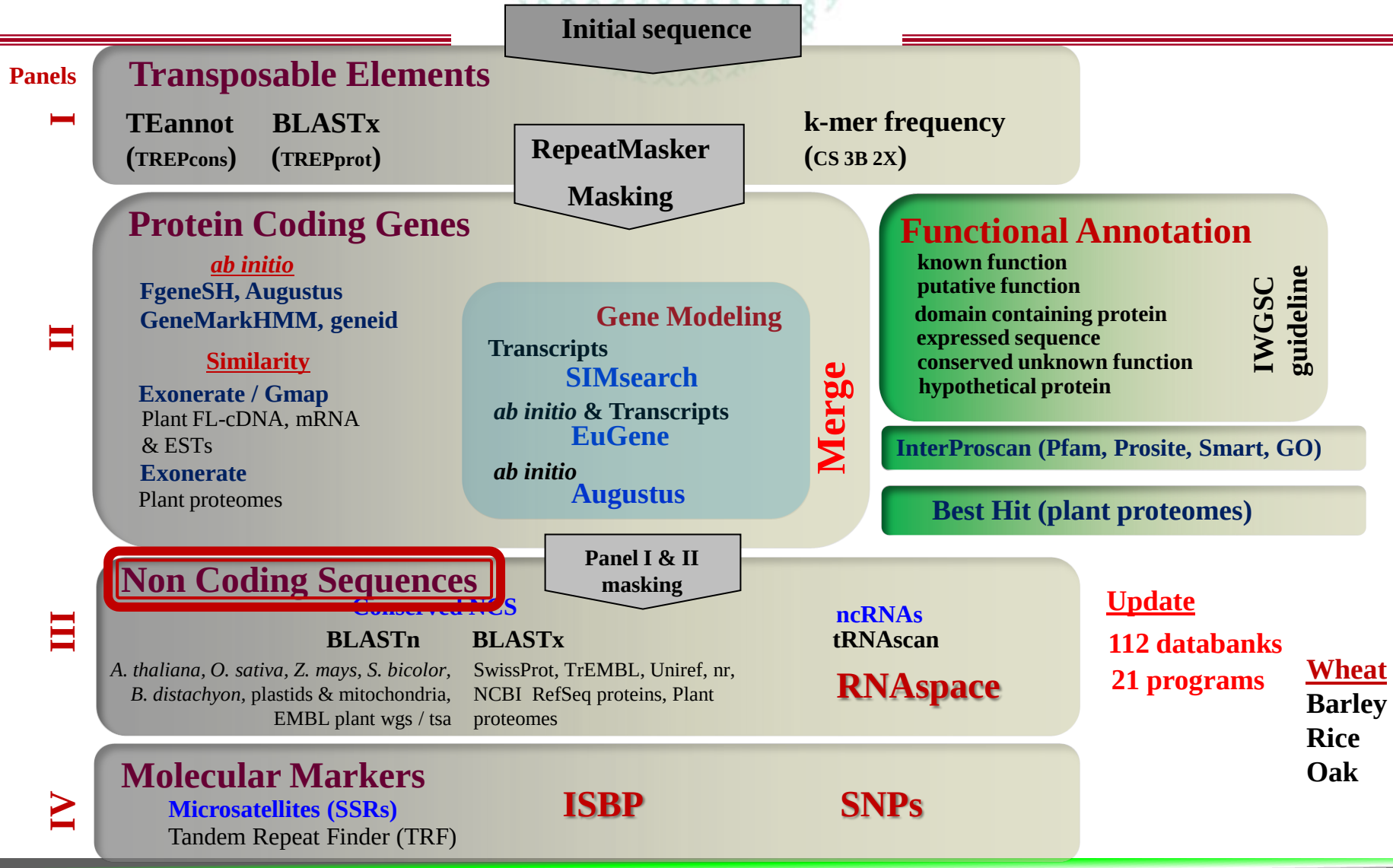
Rfam 11.0 (August 2012, 2208 families)

*snoRNA

➔ **many other families**

piRNA, siRNA, paRNA, uaRNA

TriAnnot Architecture



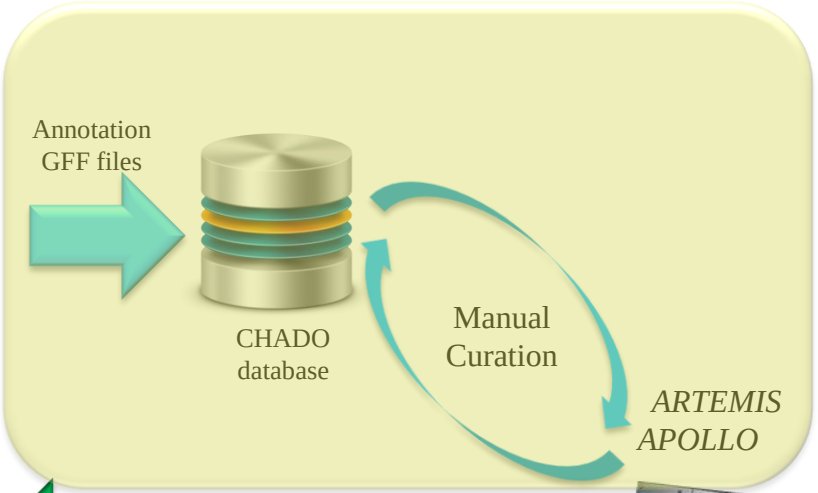
How to use the TriAnnot pipeline ?

~900 cores – 8.5 Tfop

Web submission

Email
GBrowse link
Download

Small Scale analysis



Automated sequence upload

Automated annotation download

User
Repository

Email

Large Scale analysis

transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

TriAnnot v3.6

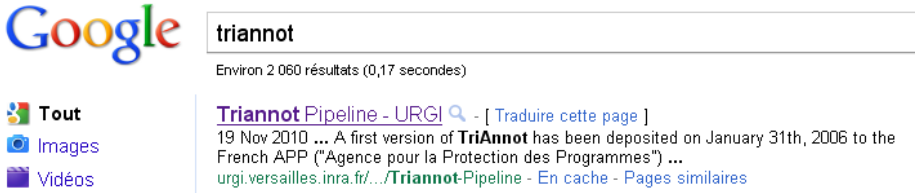
Friendly web interface

Login/Password

~ 125 users registered since the beginning
~ 59 users at present
~ 33 actively using the pipeline
➔ 726 analysis since January 2012

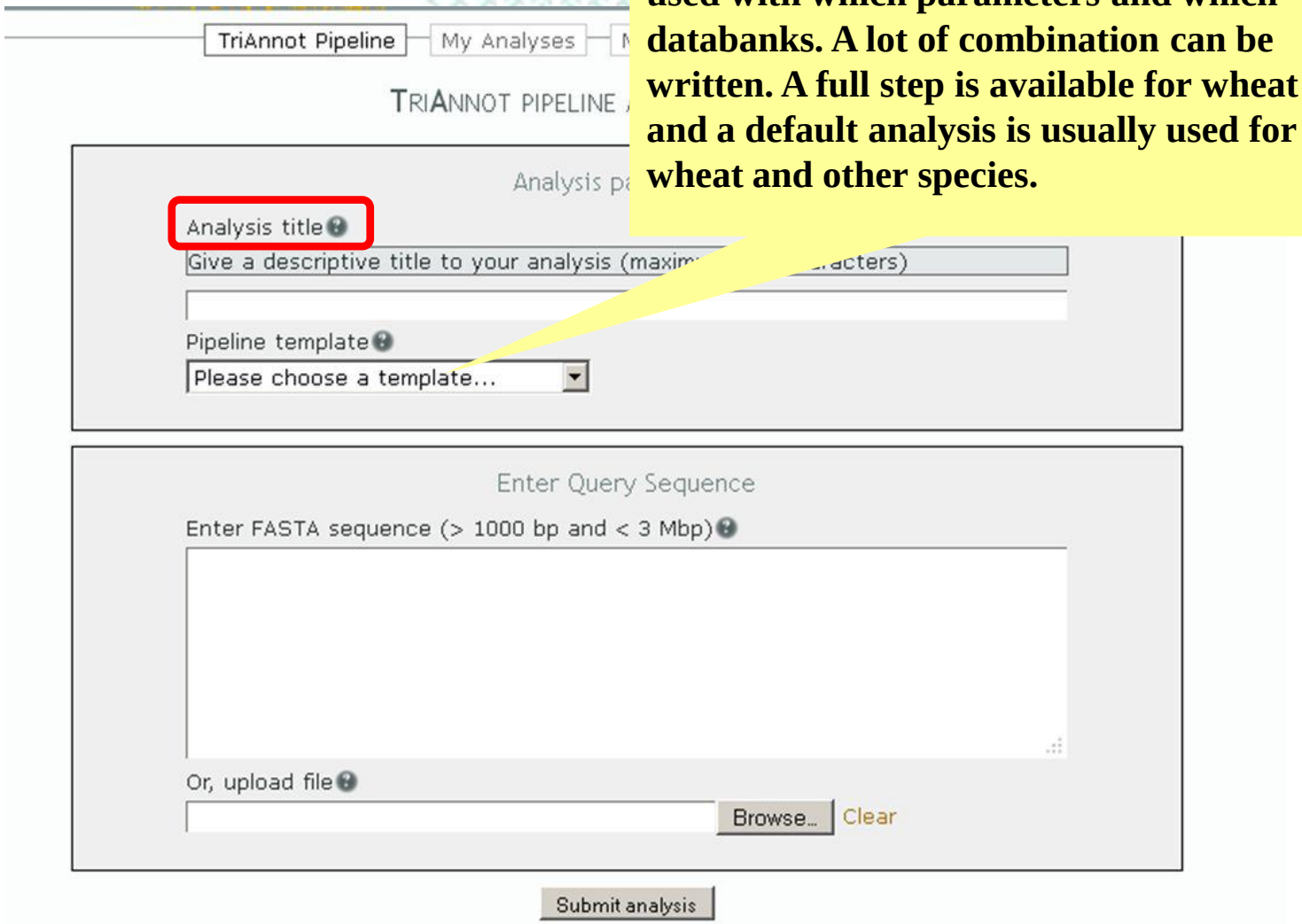
http://www.clermont.inra.fr/triannot

Submission
Automatic annotation
Graphical Viewing (Gbrowse)
Download data
EMBL, GFF, align

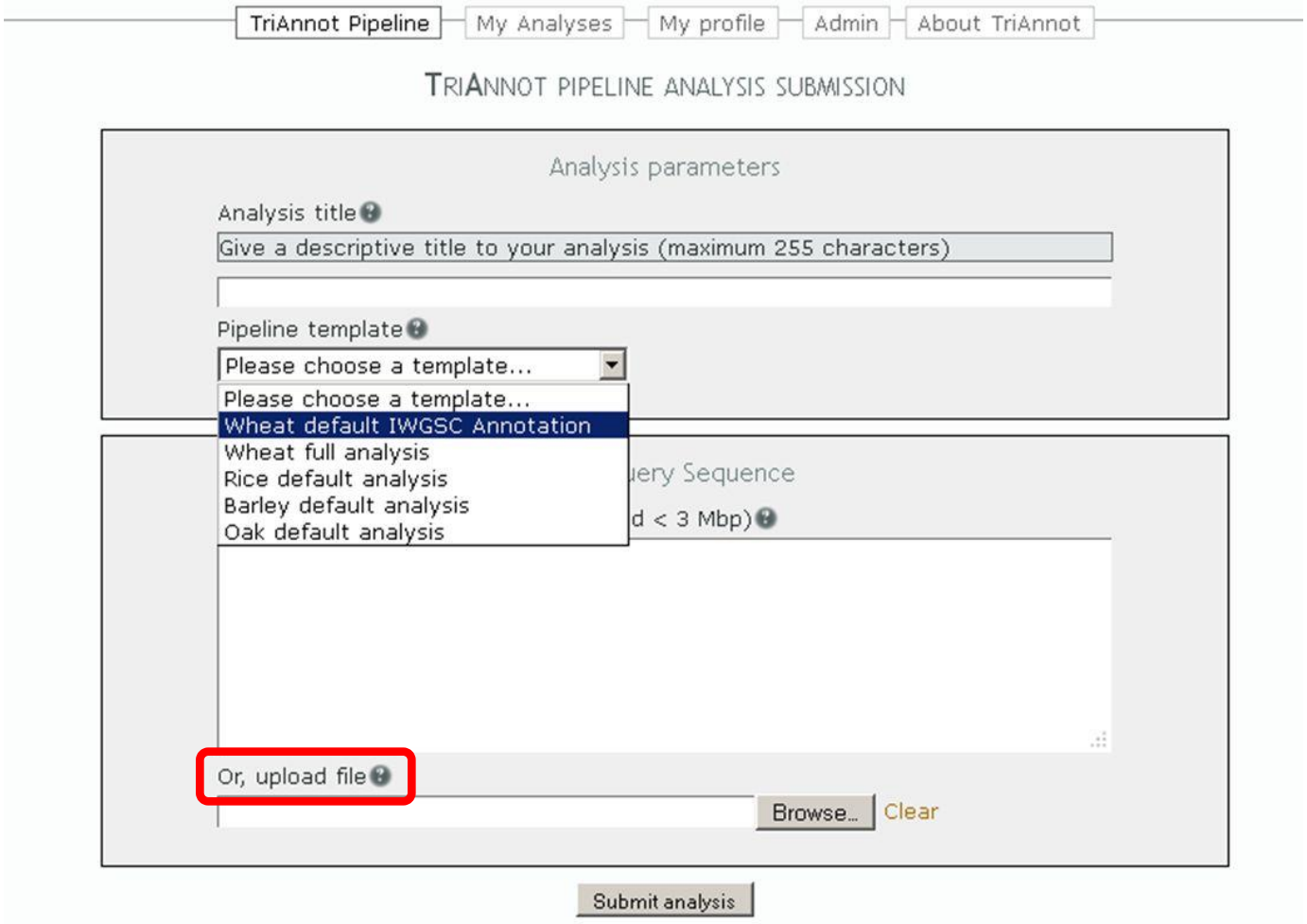


"*TriAnnot Pipeline*" TAB

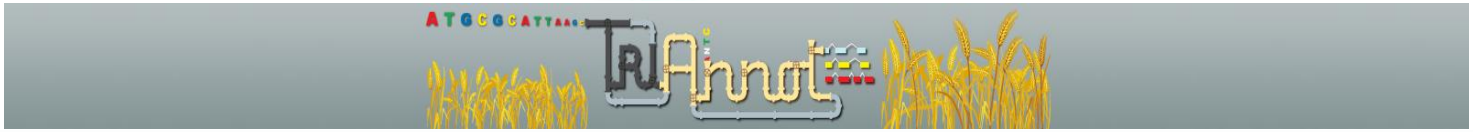
The template is the "menu" : the step.xml which defines the type of programs to be used with which parameters and which databanks. A lot of combination can be written. A full step is available for wheat and a default analysis is usually used for wheat and other species.



Several possibilities for analysis



Status



TriAnnot Pipeline | My Analyses | My profile | Admin | About TriAnnot

My TriAnnot pipeline analyses

Pending (0) | Running (0) | Finished (4) | Failed (2) | All (6)

Search: Show 15 entries

Status	Submitted	Progress	Title	Sequence	Sequence Length	Started	Finished	Pipeline Template	TriAnnot version
✗	12-01-12 10:33	0%	Just a test on a short sequence	ANF4514	4367	12-01-12 10:35	12-01-12 10:45	Wheat IWGSC Annotation	3.5
✗	12-01-12 10:33	0%	Another test	ANF4514	1257	12-01-12 10:35	12-01-12 10:45	Wheat IWGSC Annotation	3.5
✓ ⬇	12-01-03 11:18	100%	ctg0464b_000000001_000200000	ctg0464b_000000001_000200000	200000	12-01-03 11:20	12-01-03 12:02	Wheat IWGSC Annotation	3.5
✓ ⬇	11-12-31 12:14	100%	BACsynth12	Synth12	127860	11-12-31 12:15	11-12-31 13:33	Wheat IWGSC Annotation	3.5
✓ ⬇	11-12-20 12:18	100%	Scaffolds A20	A20	117015	11-12-20 12:20	11-12-20 13:09	Wheat IWGSC Annotation	3.5
✓ ⬇	11-12-07 11:51	100%	Test_3Mb	ctg954_3Mb	2999940	11-12-23 18:15	11-12-24 11:39	Wheat IWGSC Annotation	3.5

Showing 1 to 6 of 6 entries

With selected rows: Download Results | Delete

- Analyses is waiting in the queue for a free slot
- Analyses is in progress
- Analyses has been processed successfully
- Analyses has failed

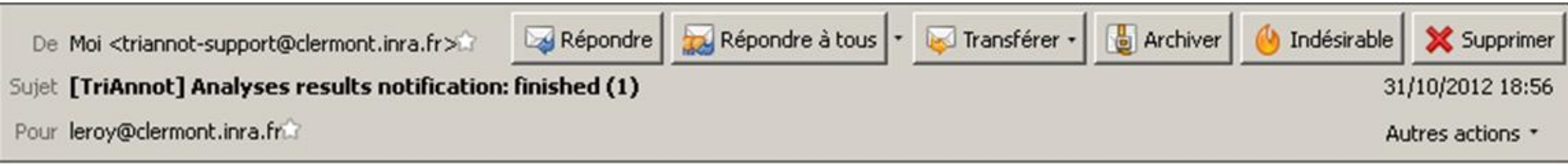
- View results with GBrowse
- Download results file

transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

Email

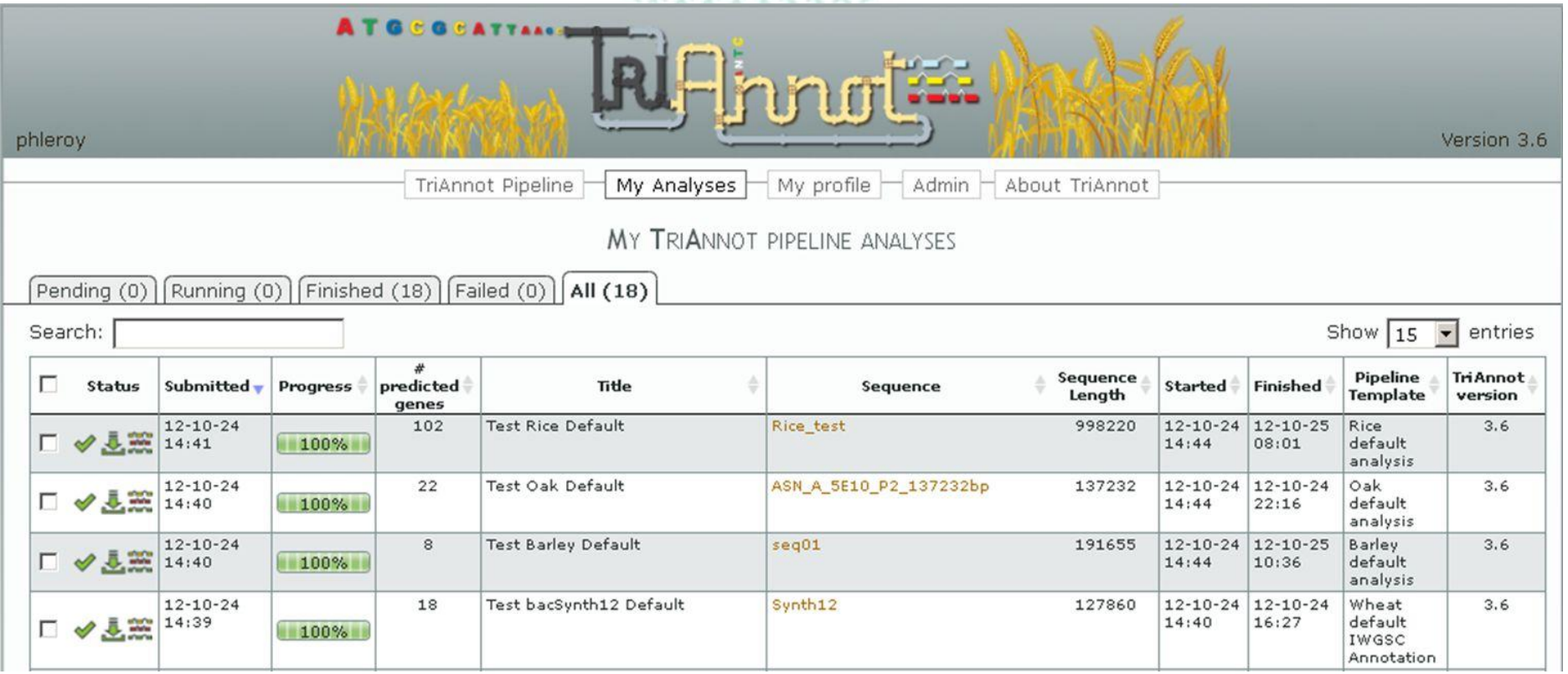


Dear Philippe Leroy,

Results for the following TriAnnot analysis are available for download:
739 - "case study transPLANT" submitted on "2012-10-31 17:46:51"
(5 genes predicted)

You can download results files from "My Analyses" section on TriAnnot pipeline web interface.
http://urgi.versailles.inra.fr/triannot/?results

GBrowse

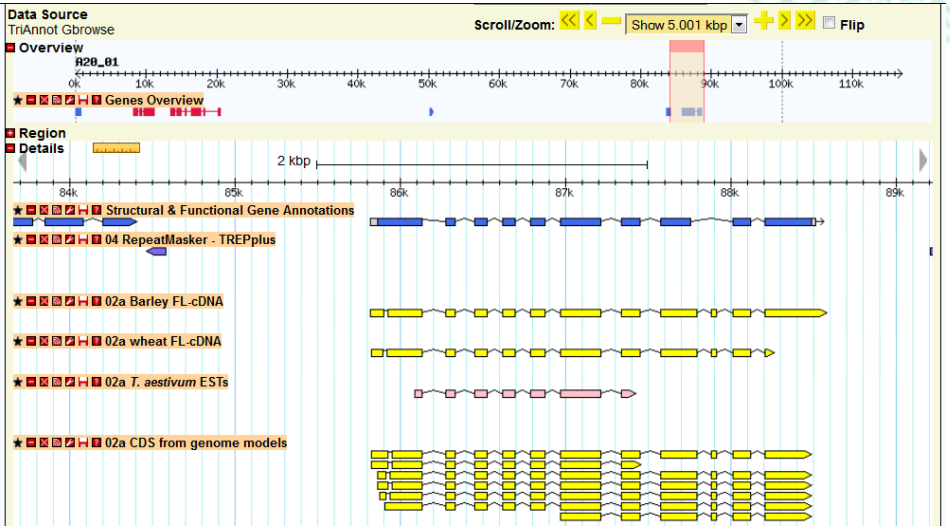


Status	Submitted	Progress	# predicted genes	Title	Sequence	Sequence Length	Started	Finished	Pipeline Template	TriAnnot version
	12-10-24 14:41	100%	102	Test Rice Default	Rice_test	998220	12-10-24 14:44	12-10-25 08:01	Rice default analysis	3.6
	12-10-24 14:40	100%	22	Test Oak Default	ASN_A_5E10_P2_137232bp	137232	12-10-24 14:44	12-10-24 22:16	Oak default analysis	3.6
	12-10-24 14:40	100%	8	Test Barley Default	seq01	191655	12-10-24 14:44	12-10-25 10:36	Barley default analysis	3.6
	12-10-24 14:39	100%	18	Test bacSynth12 Default	Synth12	127860	12-10-24 14:40	12-10-24 16:27	Wheat default IWGSC Annotation	3.6

- Analyses is waiting in the queue for a free slot
- Analyses is in progress
- Analyses has been processed successfully
- Analyses has failed

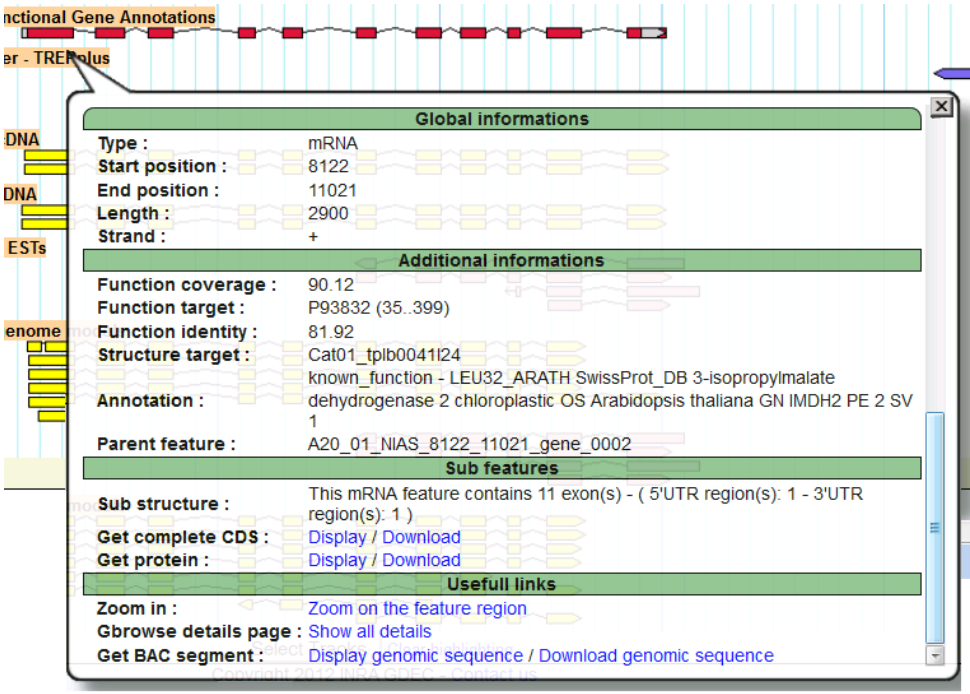
- View results with GBrowse
- Download results file

Gbrowse - v2.33

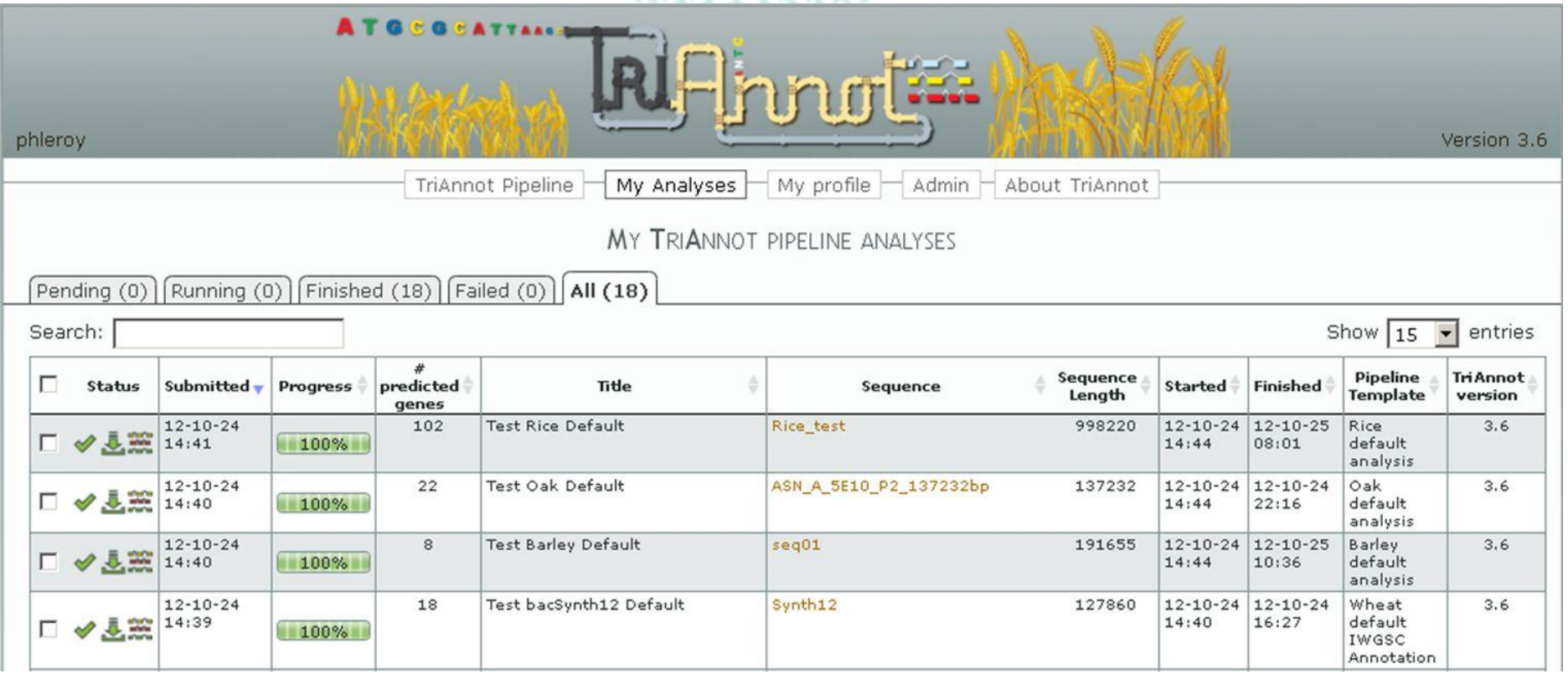


A quick graphical overview of gene models with a color code system to assess the confidence of structural annotation and biological evidences.

Popup window to retrieve gene structural & functional annotation information, and download sequences (mRNA, CDS, protein).



Download data

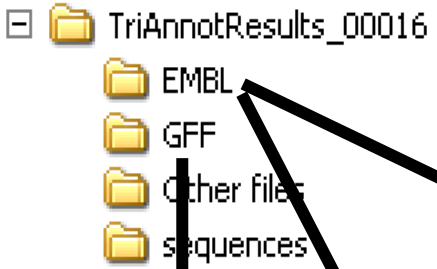


Status	Submitted	Progress	# predicted genes	Title	Sequence	Sequence Length	Started	Finished	Pipeline Template	TriAnnot version
	12-10-24 14:41	100%	102	Test Rice Default	Rice_test	998220	12-10-24 14:44	12-10-25 08:01	Rice default analysis	3.6
	12-10-24 14:40	100%	22	Test Oak Default	ASN_A_5E10_P2_137232bp	137232	12-10-24 14:44	12-10-24 22:16	Oak default analysis	3.6
	12-10-24 14:40	100%	8	Test Barley Default	seq01	191655	12-10-24 14:44	12-10-25 10:36	Barley default analysis	3.6
	12-10-24 14:39	100%	18	Test bacSynth12 Default	Synth12	127860	12-10-24 14:40	12-10-24 16:27	Wheat default IWGSC Annotation	3.6

- Analyses is waiting in the queue for a free slot
- Analyses is in progress
- Analyses has been processed successfully
- Analyses has failed

- View results with GBrowse
- Download results file

Manual curation



GenomeView (1838)

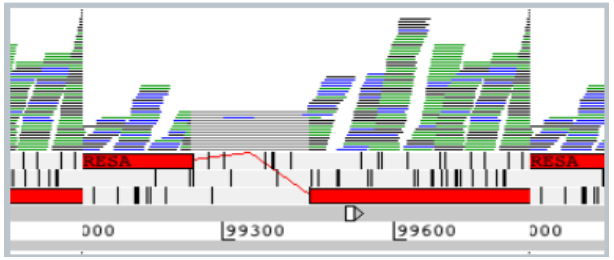
http://genomeview.org/

Abeel *et al.* (2011) **Nucleic Acids Res**. doi: 10.1093/nar/gkr995

Artemis (13.2.0)

http://www.sanger.ac.uk/resources/software/artemis/

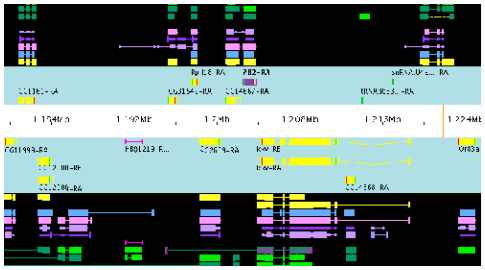
Carver *et al.* (2008) **Bioinformatics** 24, 2672-2676



Apollo (1.11.6)

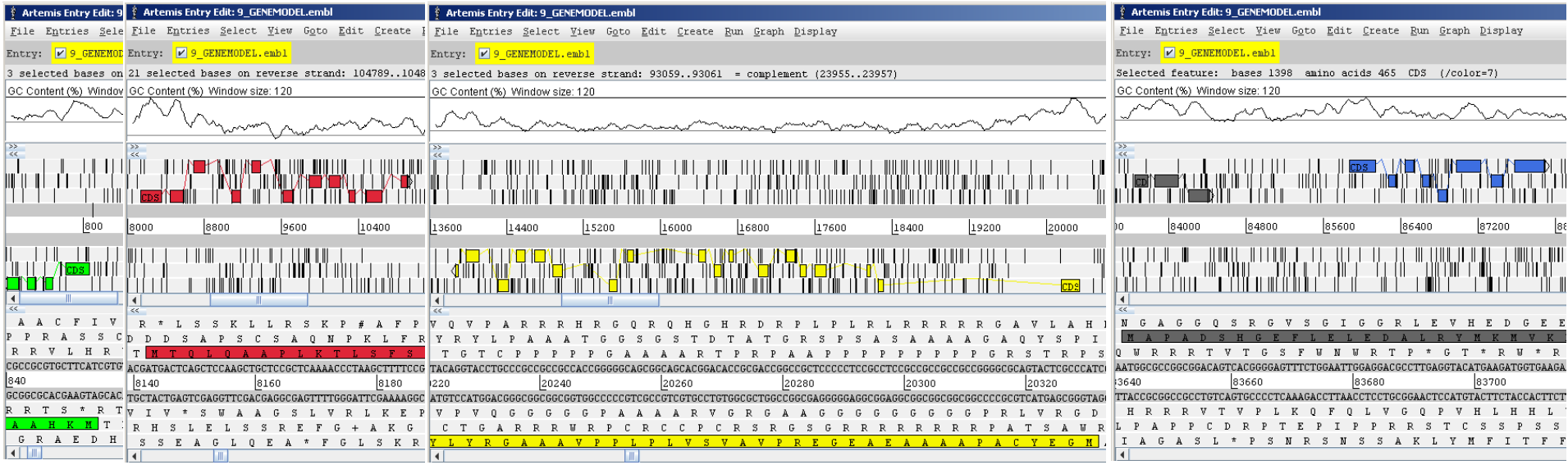
http://apollo.berkeleybop.org/current/index.html

Lewis *et al.* (2002) **Genome Biology** 3, research0082



Artemis

Rutherford *et al. (2000)*
Bioinformatics 16:944-945



Cat 3 Cat 1 Cat 2 Cat 5 Cat 4

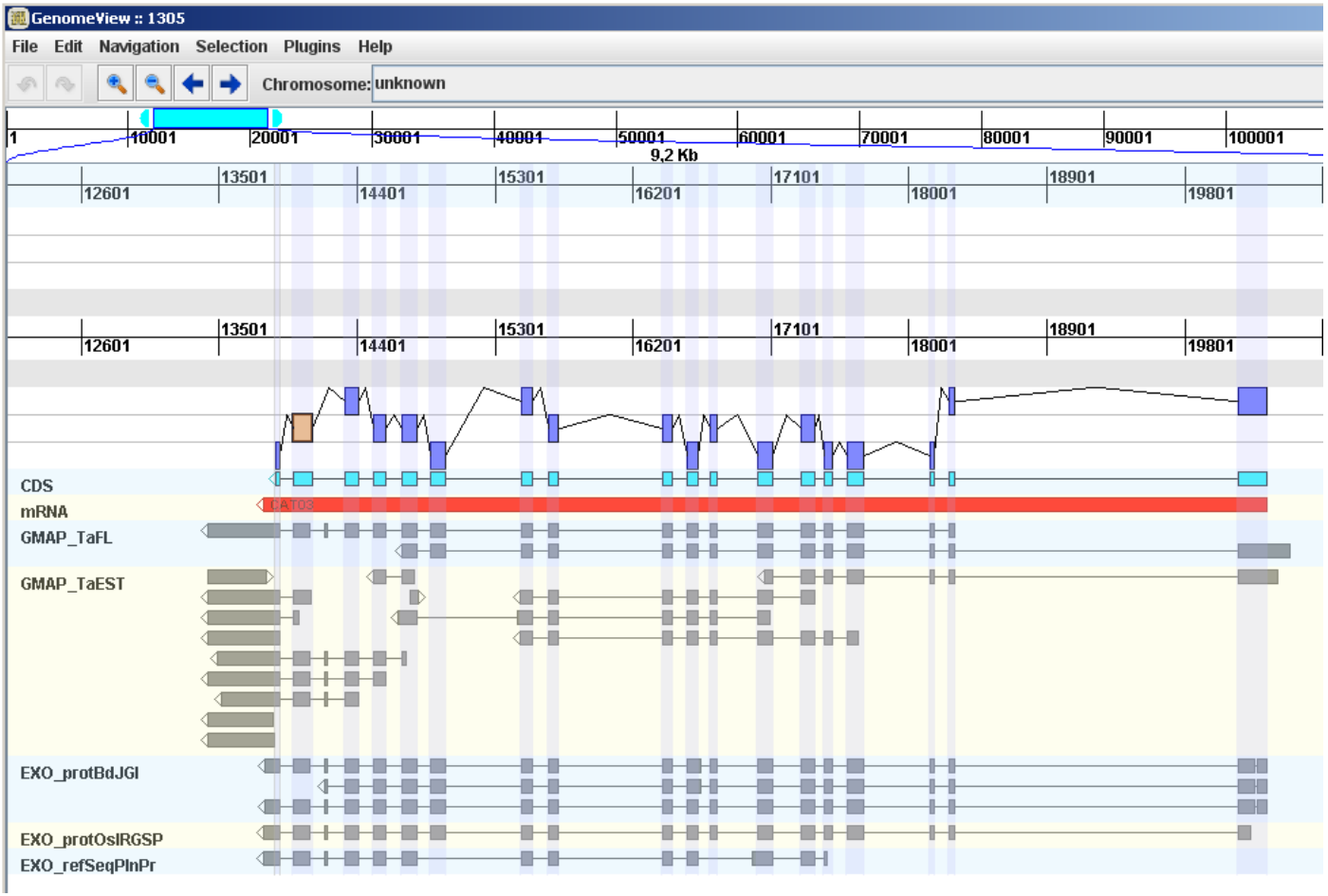
transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

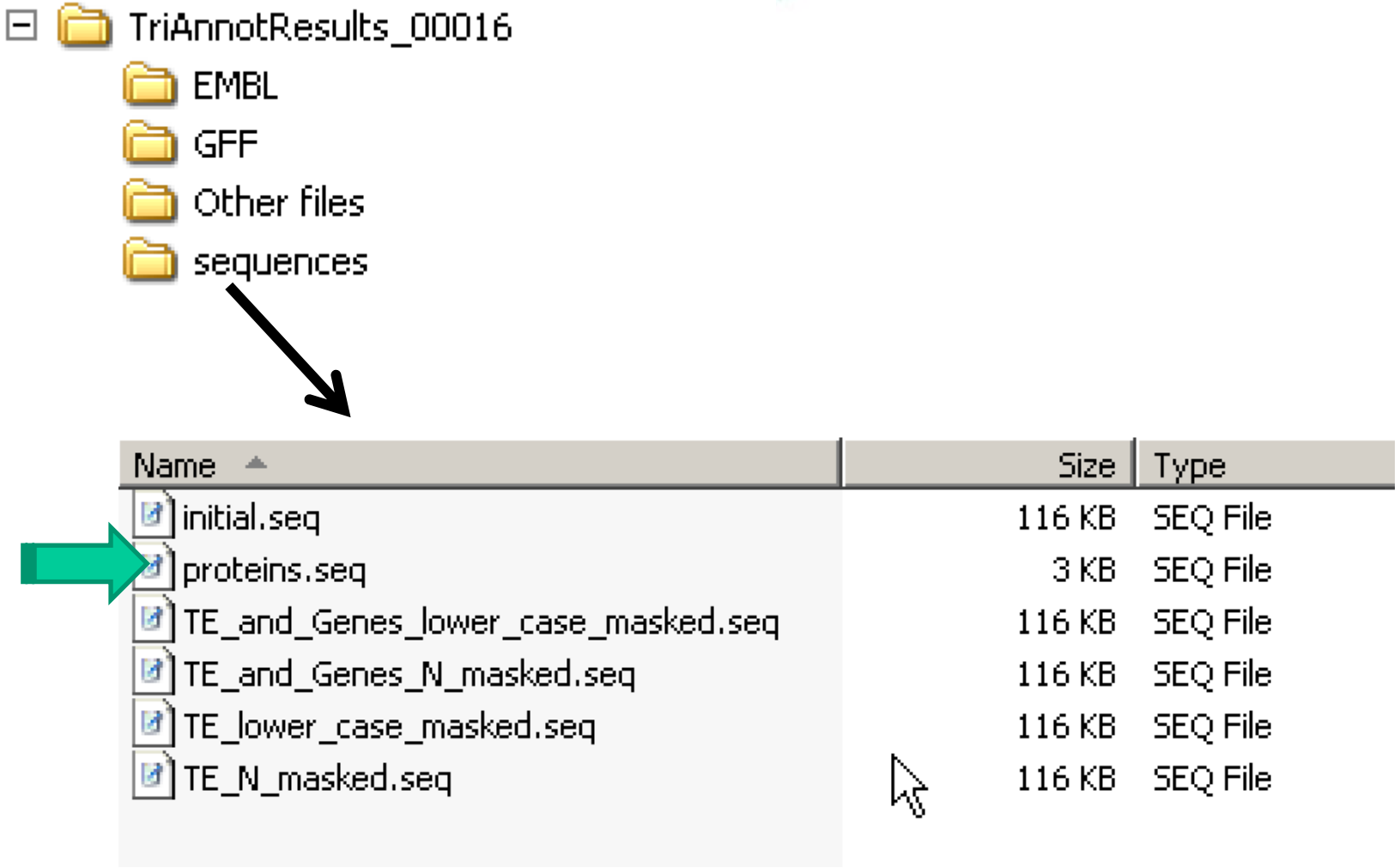
GenomeView

Thomas Abeel
Broad Institute



A case study

Sequences


TriAnnotResults_00016
EMBL
GFF
Other files
sequences
Name
Size
Type
initial.seq
116 KB
SEQ File
proteins.seq
3 KB
SEQ File
TE_and_Genes_lower_case_masked.seq
116 KB
SEQ File
TE_and_Genes_N_masked.seq
116 KB
SEQ File
TE_lower_case_masked.seq
116 KB
SEQ File
TE_N_masked.seq
116 KB
SEQ File

proteins.seq

>A20_**SIMsearch**_85867_87288_gene_0001_mRNA_0001
MKHAAALLLILAAAALVLLILAPAAHVLLPSAQYHWLNPEIHDAGGLSRQ
GFPEGFVFGTATSAYQVEGMAEQGGRGPSIWDAFIKIPGTIPGNGTANVA
VDEYHRYKEDVDIMKNMGFDAYRFSISWSRIFPDGTGKVNWEGVGYYNRL
IDYMLQQGITPYANLYQYDLPLALHQQYLGWLSPRIVGAFADYADFCFKV
FGDRVKNWFTFNEPRCVAALGYDIGLHAPGRCSQCSAGGDSTTEPYLAAH
HLILSHAAAVRRYRDKYQVRLTEDYLQVISQFKNHIFNGSA*
>A20_SIMsearch_13417_20382_gene_0002_mRNA_0001
MGEYCAPAAAAEAEGERPVAVSVLPLPPVAAAGRYLYGDYDRCSTKQVFD
NLHGNISLDPLAREFVDTEEFQRLRDLKQLGLTYLVFPGAVHTRFEHSLG
VYRLAGEAMNNLQKYQGNELGIDRIDVQTVKLAGLLHDIGHGPFSHLFEH
EFLPRVHPGSTWSHEHMSALLLDSIVDKHAIDIEPDYLKVIKEMIVASSD
VSTAEGVKEKRFLYDIVANGRNGIDVDKFDYIDRDCRACGIGSNFQHWRL
LEGMRVMGDEICYPAKDYLSIHKLFTTRADLHRTVYTHAKVKAVELMLVD
ALVEANEYLGISLHADDPEDFWKLDDTIVKSIETAPNDELKKAKEIIQRI
RRRELYKFCNQYSVPKDKLDHFKNITAQDIVCSQITSKVLLKEEDVAVSN
VKIDLTRGKDNPLESSVVMPGFRMRREVPDNGRAREPPAARVQPGQDREG
VRQEAGAGGGGVGGV*
>A20_SIMsearch_8122_11021_gene_0003_mRNA_0001
MTQLQAAPLKTLSFSGRRSAVAAARPRHATATFRCSDAARSYIITLLPGD
GIGPEVVAVAKDVLSAAGAKEGVELRFREMLMGGAALDAVGVPLPDETLA
AAQASDAILLGAIGGYKWDSNEKHLKPETGLLNIRAGLGVFANLRPATVL
PQLVDASTLKKEVAEGVDIMVVRELTGGIYFGQPRGFGTNDMGEEIGFNT
EIYSVSEIDRIARVAFEVARKRGGKLCSVDKANVLEASMLWRKRVTAIAS
EFPDVELSHMYVDNASMQLVRNPKQFDTIVTNNIFGDILSDEASMITGSI
GMLPSASVGESGPGLFEPIHGSAPDIAGQDKANPLATILSAAMLLKYGLG
AETAAKRIETAVTETLDNGFRTGDIYSPGTTLVGCKRMGEEVLKALESQK
*
>A20_EUGENE_16_854_gene_0001_mRNA_0001
MKHAAALLLILAAAAHVLLPWAQCHRLNPEIHDAGGLSRQGFPEGFVFGT
AASAYQVEGMAEQGGRGPSIWDAFIKIPGTIAGNGTADVAVDEYHRYKVC
IEDVDIMKNMGFDAYRFSISWSRIFPDGTGKVNWEGVDYYNRLIDYMLQQ
GNWNFYTRQEA*
>A20_EUGENE_50225_50590_gene_0004_mRNA_0001
METQAPPASLDLSLALATMPQPLPPPAAAAPPLSLQAAGDAVSSAVAGAG
WKVFSCLFYEKKFLKSQALGGHQNAHRKDRGAAGWNASLYLPAADRPWPP
TTATSHPEIGDENQLDLSLKL*

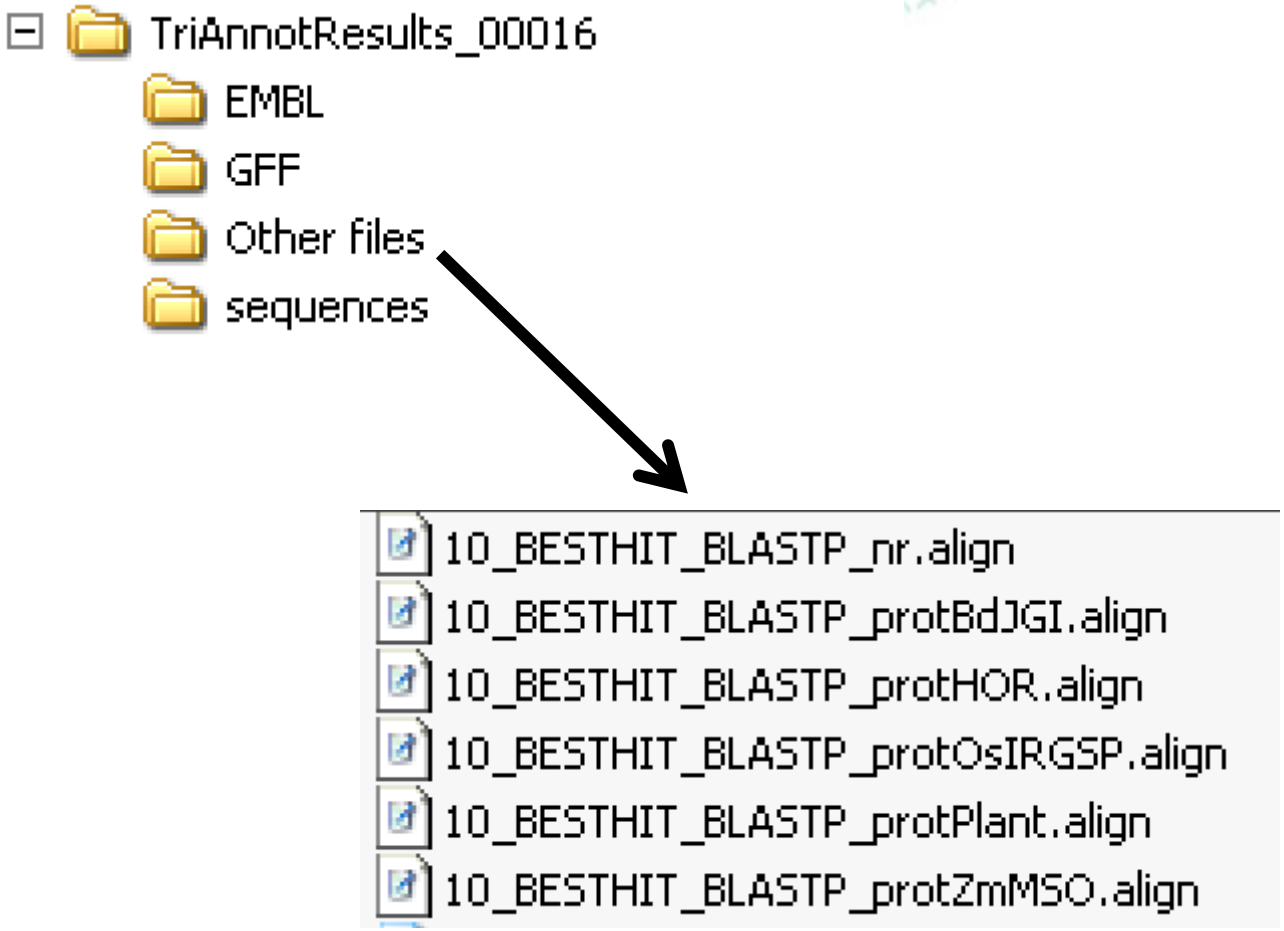
Rules of GFF & EMBL files naming

http://urgi.versailles.inra.fr/Species/Wheat/Triannot-Pipeline/Usage

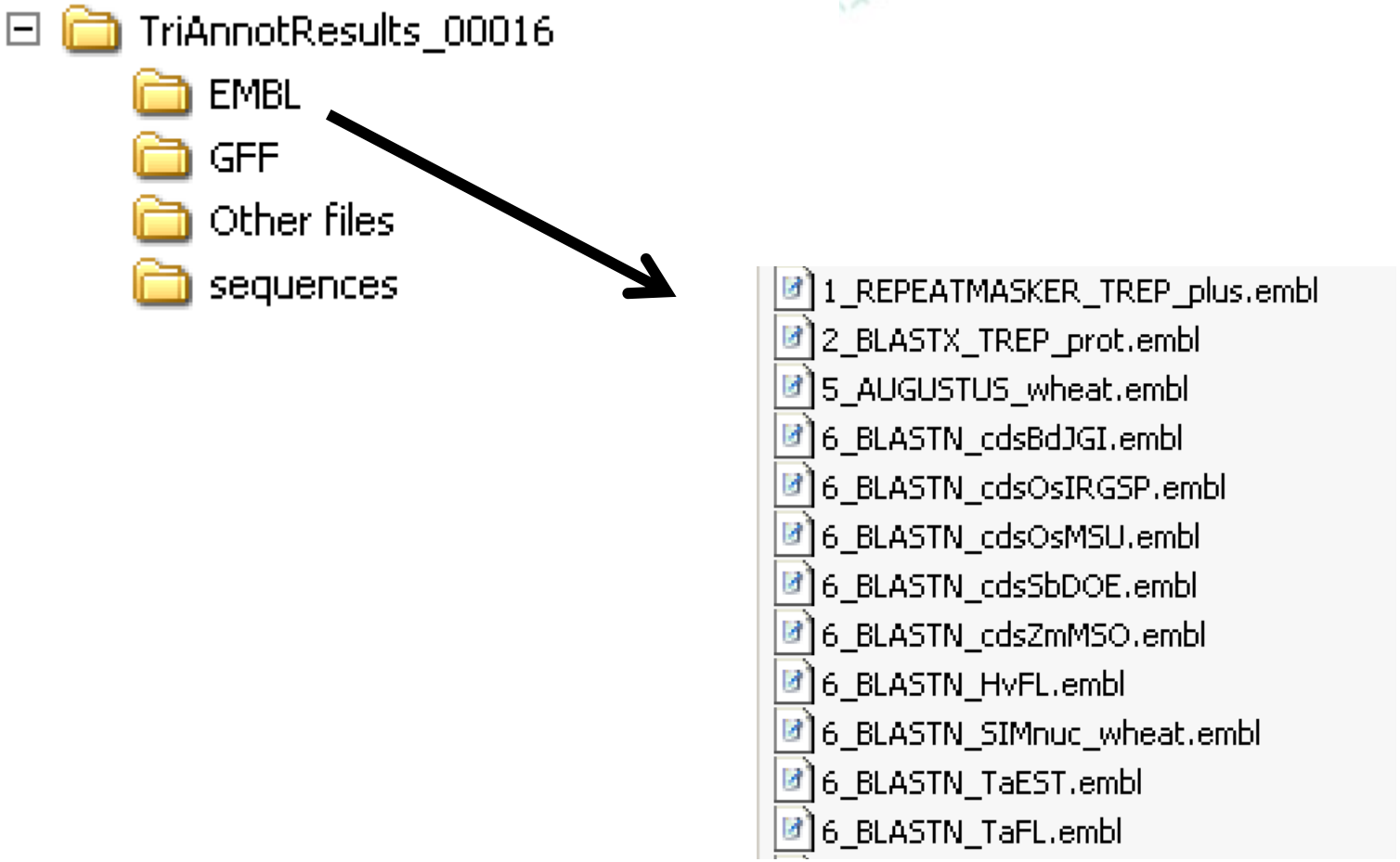
- **EMBL** folder
  - In each folder files are tagged and follows the following rules:
    - A number related to the step number
    - The type of programs *i.e.* REPATMASKER; AUGUSTUS; BLASTX; BLASTN;BLASTP; EXONERATE; EUGENE; GENEMODEL; BESTHIT; TRNASCAN-SE; TRF.
    - The databank used (see databanks). When no databank is used, just the type of program is displayed *i.e.* TRF. For *ab initio* gene prediction programs the matrix used is displayed *i.e.* AUGUSTUS_wheat.
    - Extension **.embl**
  - **Few examples**
    - Step1 - ***Transposable Elements annotation & masking***
      - 1_REPEATMASKER_TREP_plus.embl
    - Step2 - ***BLASTx against TREPprot***
      - 2_BLASTX_TREP_prot.embl
    - Step5 - ***ab initio gene prediction***
      - 5_AUGUSTUS_wheat.embl
    - Step6 - ***BLASTn / Exonerate***
      - 6_BLASTN_cdsBdJGI.embl
      - 6_BLASTN_cdsOsIRGSP.embl
      - *etc.*

2_BLASTX_TREP_prot.embl

Protein alignments



EMBL files


TriAnnotResults_00016
EMBL
GFF
Other files
sequences
1_REPEATMASKER_TREP_plus.embl
2_BLASTX_TREP_prot.embl
5_AUGUSTUS_wheat.embl
6_BLASTN_cdsBdJGI.embl
6_BLASTN_cdsOsIRGSP.embl
6_BLASTN_cdsOsMSU.embl
6_BLASTN_cdsSbDOE.embl
6_BLASTN_cdsZmMSO.embl
6_BLASTN_HvFL.embl
6_BLASTN_SIMnuc_wheat.embl
6_BLASTN_TaEST.embl
6_BLASTN_TaFL.embl

1_REPEATMASKER_TREP_plus.embl
2_BLASTX_TREP_prot.embl
5_AUGUSTUS_wheat.embl
6_BLASTN_cdsBdJGI.embl
6_BLASTN_cdsOsIRGSP.embl
6_BLASTN_cdsOsMSU.embl
6_BLASTN_cdsSbDOE.embl
6_BLASTN_cdsZmMSO.embl
6_BLASTN_HvFL.embl
6_BLASTN_SIMnuc_wheat.embl
6_BLASTN_TaEST.embl
6_BLASTN_TaFL.embl
6_EXONERATE_cdsBdJGI.embl
6_EXONERATE_cdsOsIRGSP.embl
6_EXONERATE_cdsOsMSU.embl
6_EXONERATE_cdsSbDOE.embl
6_EXONERATE_cdsZmMSO.embl
6_EXONERATE_HvFL.embl
6_EXONERATE_SIMnuc_wheat.embl
6_EXONERATE_TaEST.embl
6_EXONERATE_TaFL.embl
7_BLASTX_protBdJGI.embl
7_BLASTX_protHOR.embl
7_BLASTX_protOsIRGSP.embl
7_BLASTX_protOsMSU.embl
7_BLASTX_protSbDOE.embl
7_BLASTX_protTRI.embl
7_BLASTX_protZmMSO.embl
7_BLASTX_refSeqPlnProt.embl
7_BLASTX_SIMprot_wheat.embl
7_BLASTX_uniprot_sprot.embl
7_EXONERATE_protBdJGI.embl
7_EXONERATE_protHOR.embl
7_EXONERATE_protOsIRGSP.embl
7_EXONERATE_protOsMSU.embl
7_EXONERATE_protSbDOE.embl
7_EXONERATE_protTRI.embl
7_EXONERATE_protZmMSO.embl
7_EXONERATE_refSeqPlnProt.embl
7_EXONERATE_SIMprot_wheat.embl
7_EXONERATE_uniprot_sprot.embl
8_EUGENE.embl
9_GENEMODEL.embl
10_BESTHIT_BLASTP_nr.embl
10_BESTHIT_BLASTP_protAtTAIR.embl
10_BESTHIT_BLASTP_protBdJGI.embl
10_BESTHIT_BLASTP_protHOR.embl
10_BESTHIT_BLASTP_protOsIRGSP.embl
10_BESTHIT_BLASTP_protOsMSU.embl
10_BESTHIT_BLASTP_protPlant.embl
10_BESTHIT_BLASTP_protSAC.embl
10_BESTHIT_BLASTP_protSbDOE.embl
10_BESTHIT_BLASTP_protTRI.embl
10_BESTHIT_BLASTP_protZmMSO.embl
10_EXONERATE_nr.embl
10_EXONERATE_protAtTAIR.embl
10_EXONERATE_protBdJGI.embl
10_EXONERATE_protHOR.embl
10_EXONERATE_protOsIRGSP.embl
10_EXONERATE_protOsMSU.embl
10_EXONERATE_protPlant.embl
10_EXONERATE_protSAC.embl
10_EXONERATE_protSbDOE.embl
10_EXONERATE_protTRI.embl
10_EXONERATE_protZmMSO.embl
11_INTERPROSCAN.embl
12_BLASTN_genoAtTAIR.embl
12_BLASTN_genoBdJGI.embl
12_BLASTN_genoOsIRGSP.embl
12_BLASTN_genoOsMSU.embl
12_BLASTN_genoSbDOE.embl
12_BLASTN_genoZmMSO.embl
12_BLASTN_refSeq_chloro.embl
12_BLASTN_refSeq_mito.embl
13_BLASTX_protAtTAIR.embl
13_BLASTX_protBdJGI.embl
13_BLASTX_protOsIRGSP.embl
13_BLASTX_protOsMSU.embl
13_BLASTX_protSbDOE.embl
13_BLASTX_protZmMSO.embl
14_TRNASCAN-SE.embl
15_TRF.embl

Gene Annotation

Avoid to open empty files with GenomeView!

INRA

9_GENEMODEL.embl

```
ID   unknown; SV 1; linear; unassigned DNA; STD; UNC; 117015 BP.
XX
AC   unknown;
XX
XX
FH   Key             Location/Qualifiers
FH
FT   gene            complement(13417..20382)
FT                   /locus_tag="A20_CAT01_1"
FT                   /color="223 38 56"
FT                   /label=CAT01
FT                   /id="A20_NIAS_13417_20382_gene_0001"
FT   mRNA            complement(13417..20382)
FT                   /locus_tag="A20_CAT01_2"
FT                   /color="223 38 56"
FT                   /label=CAT01
FT                   /id="A20_NIAS_13417_20382_gene_0001_mRNA_0001"
FT   exon            complement(13417..13642)
FT                   /locus_tag="A20_CAT01_3"
FT                   /color="223 38 56"
FT                   /label=CAT01
FT                   /id="A20_NIAS_13417_13642_gene_0001_mRNA_0001_exon_0001"
FT                   /note="Ontology_term: SO:0000202"
FT   exon            complement(13660..13894)
FT                   /locus_tag="A20_CAT01_4"
FT                   /color="223 38 56"
FT                   /label=CAT01
FT                   /id="A20_NIAS_13660_13894_gene_0001_mRNA_0001_exon_0002"
FT                   /note="Ontology_term: SO:0000004"
FT   exon            complement(13984..14111)
```

Functional annotation

```
FT   exon            complement(20152..20382)
FT                   /locus_tag="A20_CAT01_21"
FT                   /color="223 38 56"
FT                   /label=CAT01
FT                   /id="A20_NIAS_20152_20382_gene_0001_mRNA_0001_exon_0019"
FT                   /note="Ontology_term: SO:0000200"
FT   CDS             complement(join(13869..13894,13984..14111,14318..14409,
FT                   14507..14587,14693..14794,14883..14975,15469..15542,
FT                   15652..15707,16392..16454,16555..16622,16706..16744,
FT                   17016..17106,17298..17382,17450..17501,17598..17704,
FT                   18140..18162,18260..18297,20152..20331))
FT                   /locus_tag="A20_CAT01_22"
FT                   /color="223 38 56"
FT                   /label=CAT01
FT                   /id="A20_NIAS_13417_20382_gene_0001_mRNA_0001_joinedCDS"
FT                   /note="Similar_to: HMMScan - domain_containing_protein -->
FT                   PF01966.15 - HD domain (1.6e-13)"
FT                   /note="Structure_target: Cat01_tplb0006o12"
FT   gene            8122..11021
```

V3.6

This is not a tutorial about how to annotate a sequence ☺

GenomeView (1838)

http://genomeview.org/

Abeel *et al.* (2011) **Nucleic Acids Res**. doi: 10.1093/nar/gkr995

Check list 1

- ✓ Build a curation repertory without empty files / BLAST files / SIMsearch files
- ✓ File / Load data… / local file/**9_GENEMODEL.embl**
  - Zoom and move (4 ways - the best = up/down/right/left keys) => **2nd gene**
  - Track list

With GenomeView the user needs the correspondence between the embl file name and the feature name of the track keys (wheat full analysis- TriAnnot 3.6**).**

1_REPEATMASKER_TREP_plus ➔ **repeat_region**
1_REPEATMASKER_TREP_nr ➔ **repeat_region**
1_REPEATMASKER_TREP_total ➔ **repeat_region**
1_REPEATMASKER_ALL_Repbase ➔ **repeat_region**
1_REPEATMASKER_MIPS_repeat ➔ **repeat_region**
1_REPEATMASKER_TIGR_Fam_Repeats ➔ **repeat_region**
1_REPEATMASKER_TIGR_GSS_Repeats ➔ **repeat_region**

1_REPEATMASKER_univec ➔ **repeat_region**
1_REPEATMASKER_Ecoli ➔ **repeat_region**

2_BLASTX_TREP_prot ➔ **BLASTX_TREP_pro**

5_AUGUSTUS_wheat ➔ **AUGUSTUS_wheat**
5_FGENESH ➔ **FGeneSH**
5_GENEID ➔ **GeneID**
5_GMHMM_wheat ➔ **GMHMM_wheat**

8_EUGENE ➔ **Eugene**

9_GENEMODEL ➔ gene / mRNA / exon / CDS

6_EXONERATE_AtFL ➔ **EXO_N_AtFL**
6_EXONERATE_HvFL ➔ **EXO_N_HvFL**
6_EXONERATE_OsFL ➔ **EXO_N_OsFL**
6_EXONERATE_PlnFL ➔ **EXO_N_PlnFL**
6_EXONERATE_PoaFL ➔ **EXO_N_PoaFL**
6_EXONERATE_PpFL ➔ **EXO_N_PpFL**
6_EXONERATE_PtFL ➔ **EXO_N_PtFL**
6_EXONERATE_RosiFL ➔ **EXO_N_RosiFL**
6_EXONERATE_TaFL ➔ **EXO_N_TaFL**
6_EXONERATE_ZmFL ➔ **EXO_N_ZmFL**

6_EXONERATE_HvEST ➔ **EXO_N_HvEST**
6_EXONERATE_OsEST ➔ **EXO_N_OsEST**
6_EXONERATE_QuerEST ➔ **EXO_N_QuerEST**
6_EXONERATE_SoEST ➔ **EXO_N_SoEST**
6_EXONERATE_TaEST ➔ **EXO_N_TaEST**
6_EXONERATE_ZmEST ➔ **EXO_N_ZmEST**

6_EXONERATE_rnaSeq_wheat ➔ **EXO_N_rnaSeq_wh**

6_EXONERATE_AT_unigene ➔ **EXO_N_AT_unigen**
6_EXONERATE_BD_unigene ➔ **EXO_N_BD_unigen**
6_EXONERATE_HV_unigene ➔ **EXO_N_HV_unigen**
6_EXONERATE_OS_unigene ➔ **EXO_N_OS_unigen**
6_EXONERATE_PP_unigene ➔ **EXO_N_PP_unigen**
6_EXONERATE_PT_unigene ➔ **EXO_N_PT_unigen**
6_EXONERATE_QR_unigene ➔ **EXO_N_QR_unigen**
6_EXONERATE_SB_unigene ➔ **EXO_N_SB_unigen**
6_EXONERATE_SO_unigene ➔ **EXO_N_SO_unigen**
6_EXONERATE_TA_unigene ➔ **EXO_N_TA_unigen**
6_EXONERATE_VV_unigene ➔ **EXO_N_VV_unigen**
6_EXONERATE_ZM_unigene ➔ **EXO_N_ZM_unigen**

6_EXONERATE_cdsAtTAIR ➔ **EXO_N_cdsAtTAIR**
6_EXONERATE_cdsBdJGI ➔ **EXO_N_cdsBdJGI**
6_EXONERATE_cdsOsIRGSP ➔ **EXO_N_cdsOsIRGS**
6_EXONERATE_cdsOsMSU ➔ **EXO_N_cdsOsMSU**
6_EXONERATE_cdsPp ➔ **EXO_N_cdsPp**
6_EXONERATE_cdsPt ➔ **EXO_N_cdsPt**
6_EXONERATE_cdsSbDOE ➔ **EXO_N_cdsSbDOE**
6_EXONERATE_cdsVv ➔ **EXO_N_cdsVv**
6_EXONERATE_cdsZmMSO ➔ **EXO_N_cdsZmMSO**

6_EXONERATE_PlnCpltCDS ➔ **EXO_N_PlnCpltCD**

6_EXONERATE_PlnMrnaSTD ➔ **EXO_N_PlnMrnaST**
6_EXONERATE_PoaMrnaSTD ➔ **EXO_N_PoaMrnaST**

7_EXONERATE_pepATH ➔ **EXO_X_pepATH**
7_EXONERATE_pepBDI ➔ **EXO_X_pepBDI**
7_EXONERATE_pepOSA_IRGSP ➔ **EXO_X_pepOSA_IR**
7_EXONERATE_pepOSA_IMSU ➔ **EXO_X_pepOSA_MS**
7_EXONERATE_pepPPE ➔ **EXO_X_pepPPE**
7_EXONERATE_pepPTR ➔ **EXO_X_pepPTR**
7_EXONERATE_pepSBI ➔ **EXO_X_pepSBI**
7_EXONERATE_pepVVI ➔ **EXO_X_pepVVI**
7_EXONERATE_pepZMA ➔ **EXO_X_pepZMA**

7_EXONERATE_protATH➔ **EXO_X_protATH**
7_EXONERATE_protBDI ➔ **EXO_X_protBDI**
7_EXONERATE_protHOR ➔ **EXO_X_protHOR**
7_EXONERATE_protOSA ➔ **EXO_X_protOSA**
7_EXONERATE_protPPE ➔ **EXO_X_protPPE**
7_EXONERATE_protPTR ➔ **EXO_X_protPTR**
7_EXONERATE_protSAC ➔ **EXO_X_protSAC**
7_EXONERATE_protSBI ➔ **EXO_X_protSBI**
7_EXONERATE_protTRI ➔ **EXO_X_protTRI**
7_EXONERATE_protVVI ➔ **EXO_X_protVVI**
7_EXONERATE_protZMA ➔ **EXO_X_protZMA**

7_EXONERATE_protPlant ➔ **EXO_X_protPlant**
7_EXONERATE_protPoa ➔ **EXO_X_protPoa**
7_EXONERATE_protRos ➔ **EXO_X_protRos**

7_EXONERATE_refSeqPlnProt ➔ **EXO_X_refSeqPln**
7_EXONERATE_uniprot_sprot ➔ **EXO_X_uniprot_s**
7_EXONERATE_uniprot_trembl ➔ **EXO_X_uniprot_t**

10_EXONERATE_pepATH ➔ **EXO_P_pepATH**
10_EXONERATE_pepBDI ➔ **EXO_P_pepBDI**
10_EXONERATE_pepOSA_IRGSP ➔ **EXO_P_pepOSA_IR**
10_EXONERATE_pepOSA_IMSU ➔ **EXO_P_pepOSA_MS**
10_EXONERATE_pepPPE ➔ **EXO_P_pepPPE**
10_EXONERATE_pepPTR ➔ **EXO_P_pepPTR**
10_EXONERATE_pepSBI ➔ **EXO_P_pepSBI**
10_EXONERATE_pepVVI ➔ **EXO_P_pepVVI**
10_EXONERATE_pepZMA ➔ **EXO_P_pepZMA**

10_EXONERATE_protHOR ➔ **EXO_P_protHOR**
10_EXONERATE_protSAC ➔ **EXO_P_protSAC**
10_EXONERATE_protTRI ➔ **EXO_P_protTRI**

10_EXONERATE_protPlant ➔ **EXO_P_protPlant**
10_EXONERATE_protPoa ➔ **EXO_P_protPoa**
10_EXONERATE_protRos ➔ **EXO_P_protRos**

10_EXONERATE_nr ➔ **EXO_P_nr**
10_EXONERATE_refSeqPlnProt ➔ **EXO_P_refSeqPln**
10_EXONERATE_uniprot_sprot ➔ **EXO_P_uniprot_s**
10_EXONERATE_uniprot_trembl ➔ **EXO_P_uniprot_t**

12_BLASTN_embl_pln ➔ **BLASTN_embl_pln**
12_BLASTN_embl_est_pln ➔ **BLASTN_embl_est**
12_BLASTN_embl_gss_pln ➔ **BLASTN_embl_gss**
12_BLASTN_embl_htc_pln ➔ **BLASTN_embl_htc**
12_BLASTN_embl_htg_pln ➔ **BLASTN_embl_htg**
12_BLASTN_embl_pat_pln ➔ **BLASTN_embl_pat**
12_BLASTN_embl_sts_pln ➔ **BLASTN_embl_sts**
12_BLASTN_embl_tsa_pln ➔ **BLASTN_embl_tsa**
12_BLASTN_embl_wgs_pln ➔ **BLASTN_embl_wgs**

12_BLASTN_genoAtTAIR ➔ **BLASTN_genoAtTA**
12_BLASTN_genoBdJGI ➔ **BLASTN_genoBdJG**
12_BLASTN_genoOsIRGSP ➔ **BLASTN_genoOsIR**
12_BLASTN_genoOsMSU ➔ **BLASTN_genoOsMS**
12_BLASTN_genoPpe ➔ **BLASTN_genoPpe**
12_BLASTN_genoPtr ➔ **BLASTN_genoPtr**
12_BLASTN_genoSbDOE ➔ **BLASTN_genoSbDO**
12_BLASTN_genoVvi ➔ **BLASTN_genoVvi**
12_BLASTN_genoZmMSO ➔ **BLASTN_genoZmMS**

12_BLASTN_refSeq_chloro ➔ **BLASTN_refSeq_c**
12_BLASTN_refSeq_mito ➔ **BLASTN_refSeq_m**

13_EXONERATE_pepATH ➔ **BLASTX_pepATH**
13_EXONERATE_pepBDI ➔ **BLASTX_pepBDI**
13_EXONERATE_pepOSA_IRGSP ➔ **BLASTX_pepOSA_IR**
13_EXONERATE_pepOSA_IMSU ➔ **BLASTX_pepOSA_MS**
13_EXONERATE_pepPPE ➔ **BLASTX_pepPPE**
13_EXONERATE_pepPTR ➔ **BLASTX_pepPTR**
13_EXONERATE_pepSBI ➔ **BLASTX_pepSBI**
13_EXONERATE_pepVVI ➔ **BLASTX_pepVVI**
13_EXONERATE_pepZMA ➔ **EBLASTX_pepZMA**

11_INTERPROSCAN ➔ **misc_feature**

14_TRNASCAN-SE ➔ **tRNA**

15_TRF ➔ **tandem_repeat**

Check list 1

- ✓ Build a curation repertory without empty files / BLAST files / SIMsearch files
- ✓ File / Load data… / local file/**9_GENEMODEL.embl**
  - Zoom and move (4 ways - the best = up/down/right/left keys) => **2nd gene**
  - Track list
  - Exon/CDS/Gene Structure (hidden) – keep exon
    - Clic on exon / CDS - shadow bands (last exon)
    - Details on selected items (windows on the right)
    - CTRL E on exon (label=Category / functional Annotation)
      - Change label=3-isopropylmalate dehydrogenase 2 chloroplastic1
      - **problem with save**
  - See Query: **Google, NCBI** (Ensembl, Plaza)
  - Add Track exon, first exon ➔ Start / Stop / Intron-Exon junctions (problem first Do/Ac)
- ✓ Biological Evidences (BLASTn; BLASTx; BLASTp – best hits)
  - EXO_N / EXO_X / EXO_P ➔ move tracks / compact tracks
  - Correction of Do/Ac first 2 junctions – expand last exon
    - PoaFL / HvFL / CDS

Check list 2

- ✓ Biological Evidences (BLASTn; BLASTx; BLASTp – best hits)
  - EXO_X
  - EXO_P (case of empty file ➔ see files on repertory)
    - See "Other files" ➔ multiple alignment **pepBDI & protHor**
- ✓ Case of the last Gene (wrong gene model as seen on the Gbrowse)
  - Open all the following evidences
    - Wheat RNA-seq / FL-cDNA (keep FL-cDNA)
    - Ta_unigene (delete an item)
    - Hv_FL (keep FL-cDNA)
    - CDS / EXO_X (compact) / EXO_P
    - InterProScan

Check list 3

- An example with **Augustus** Gene Model
  - **Clone** feature – change type to CDS
  - Select 2 exons ➔ **split** (CTRL/Caps)
  - Take the first Met
  - Clone wheat FL-cDNA and keep the missing exon => change type to CDS
  - **Merge** the tows CDS
  - Check Expasy (*Viridiplantaea*, no filter, 50, 50, 0,0001)
  - Navigation/**Search** aa sequence
  - Keep this gene (label=pseudo)

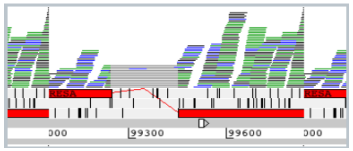
✓ Conserved Non-coding Sequences – CNSs : 12_; 13_; tRNAs

✓ SSRs

✓ TEs (Wicker's code)

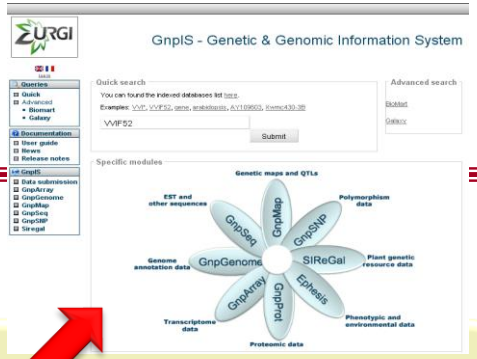
Check list 4

- ✓ Save annotation with an other name (gene/CDS/exon/mRNA)
  - See the file with wordpad
- ✓ Save Session
- ✓ Re open a previous analysis
- ✓ Recover previous annotation file ➔ trick
- ✓ See the saved annotation with ARTEMIS
  - ✓ Sub sequence function



10 last minutes

How to use the TriAnnot pipeline ?



v2.1 - 16,136 Scaff ~ 26 hrs

Annotation

CHADO database

Manual Curation

ARTEMIS
APOLLO

Web submission

Email
GBrowse link
Download

Automated sequence upload

Automated annotation download

User Repository

Email

~900 cores – 8.5 Tfop

Small Scale analysis

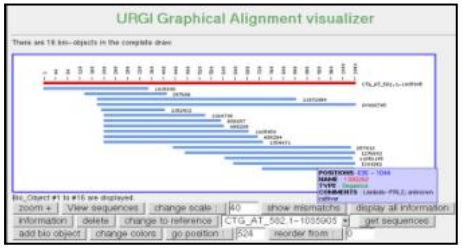
Large Scale analysis

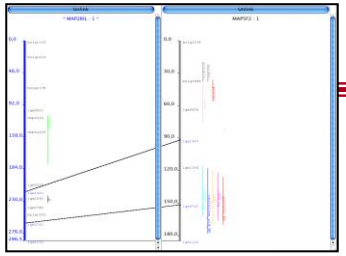
transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

INRA

Integrative Databases for interoperability

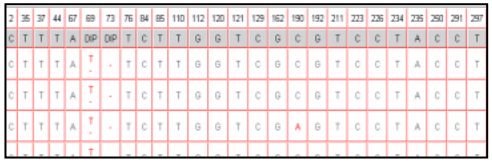


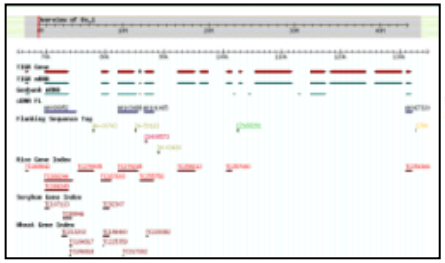


Genetic maps and QTLs

EST and other sequences

Polymorphism data





Genome annotation data

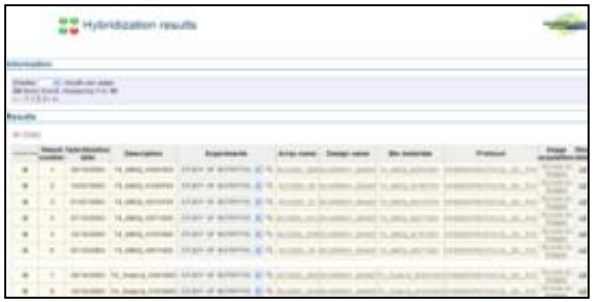
GnpSeq
GnpMap
GnpSNP
GnpGenome
SIReGal
GnpArray
GnpProt
Ephesis

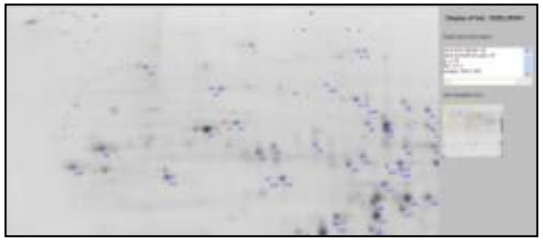
Plant genetic resource data

Transcriptome data

Phenotypic and environmental data

Proteomic data





transPLANT

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

INRA

Perspectives

New Merge Model

New **MergeGeneModel.pm**

- Multiple input files
  - ✓ Integrate manually curated genes
- Filtering predictions
  - ✓ No overlapping evidences AND no BLAST hit against plants proteomes **& Transposases**
- Validation against evidences
- Scoring
  - **Overlap with best blastp hit**
  - **% of identity**
  - **Validation score : nValid²/nAll**
  - **Penalty for non canonical splice sites**

- **Choice of best gene model**

S. Theil
GDEC

F. Choulet
GDEC

Two new independent code systems

Status

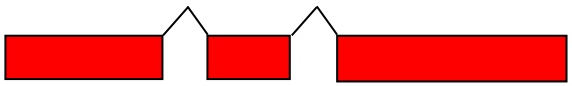
- **'Full'**
  - ✓ hit_coverage > **70%**
- **Pseudo**
  - ✓ **50%** ≤ hit_coverage ≤ **70%**
- **Fragment**
  - ✓ hit_coverage < **50%**

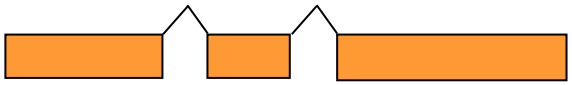
Confidence

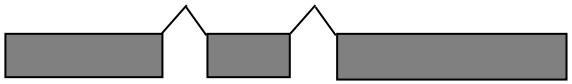
- **High confidence (clear evidences: start/stop/intron-exon junctions)**
- **Low confidence (no clear evidences : start/stop/intron-exon junctions)**

New color code system to assess the confidence of the structural annotation *

3 Categories

Cat 1  **Full** – *High Confidence* **(1)**

Cat 2  **Full** – *Low Confidence*

Cat 3  **Pseudo & fragments**

(1) This gene models do not need manual expertise

Wheat chromosome 3B annotation validation

Annotation on assembly V4.2 (expected overlap) – **8 789** Protein-coding gene models

After filtering

	high_confidence	low_confidence	
Full	**4 136**	2 477	**6 613**
Pseudo	769	692	**1 461**
Fragment	101	614	**715**
	56%	**44%**	

Manually expertised

Seq_problem : Ns | wrong splice site | frame shift

	Full	Pseudo	Fragment	
validated	**3 029**	470	25	
not_validated	622	305	184	
validated_seq_problem	19	212	3	
not_validated_seq_problem	30	87	9	
	3 700	**1 074**	**221**	**4 995**

Acknowledgements

P. Leroy
N. Guilhot
(S. Theil)
F. Choulet
C. Feuillet

C. Viseux
M. Alaux
H. Quesneville

H. Sakai
T. Itoh

C. Pozniack

C. Caron

L. Sterck

A. Mahul

D. Hill

Previous Collaborators

P. Dufour
L. Cerutti
B. Laubin
F. Sabot
F. Legeai
E. Gicquelo
F. Giacomoni
C. Caron
I. Blanc-Lenfle
A. Mahul
A. Claude
M. Liauzu
T. Flutre
I. Luyten
C. Pelegrin
A. Bernard
O. Inizan
M. Reichstadt
S. Reboux
N. Amano
H. Numa
T. Tanaka
M. Seidel
K. Mayer
H. Ohyanagi

Funding from the European Community's FP7 Programme - GA FP7-212019

Thank You very much

Massif du Sancy
Joël Damase

Murol Castle

http://www.clermont.inra.fr/triannot/

triannot-support@clermont.inra.fr