

# TriAnnot: a user friendly web interface for structural and functional automatic annotation of plant genomes.

Leroy P<sup>1</sup>, Guilhot N<sup>1</sup>, Sakai H<sup>3</sup>, Bernard A<sup>1</sup>, Choulet F<sup>1</sup>, Pelegrin C<sup>1</sup>, Reboux S<sup>2</sup>, Flutre T<sup>2</sup>, Amano N<sup>3</sup>, Seidel M<sup>4</sup>, Ohyanagi H<sup>3</sup>, Alaux M<sup>2</sup>, Numa H<sup>3</sup>, Tanaka T<sup>3</sup>, Mayer K<sup>4</sup>, Itoh T<sup>3</sup>, Quesneville H<sup>2</sup>, Feuillet C<sup>1</sup>

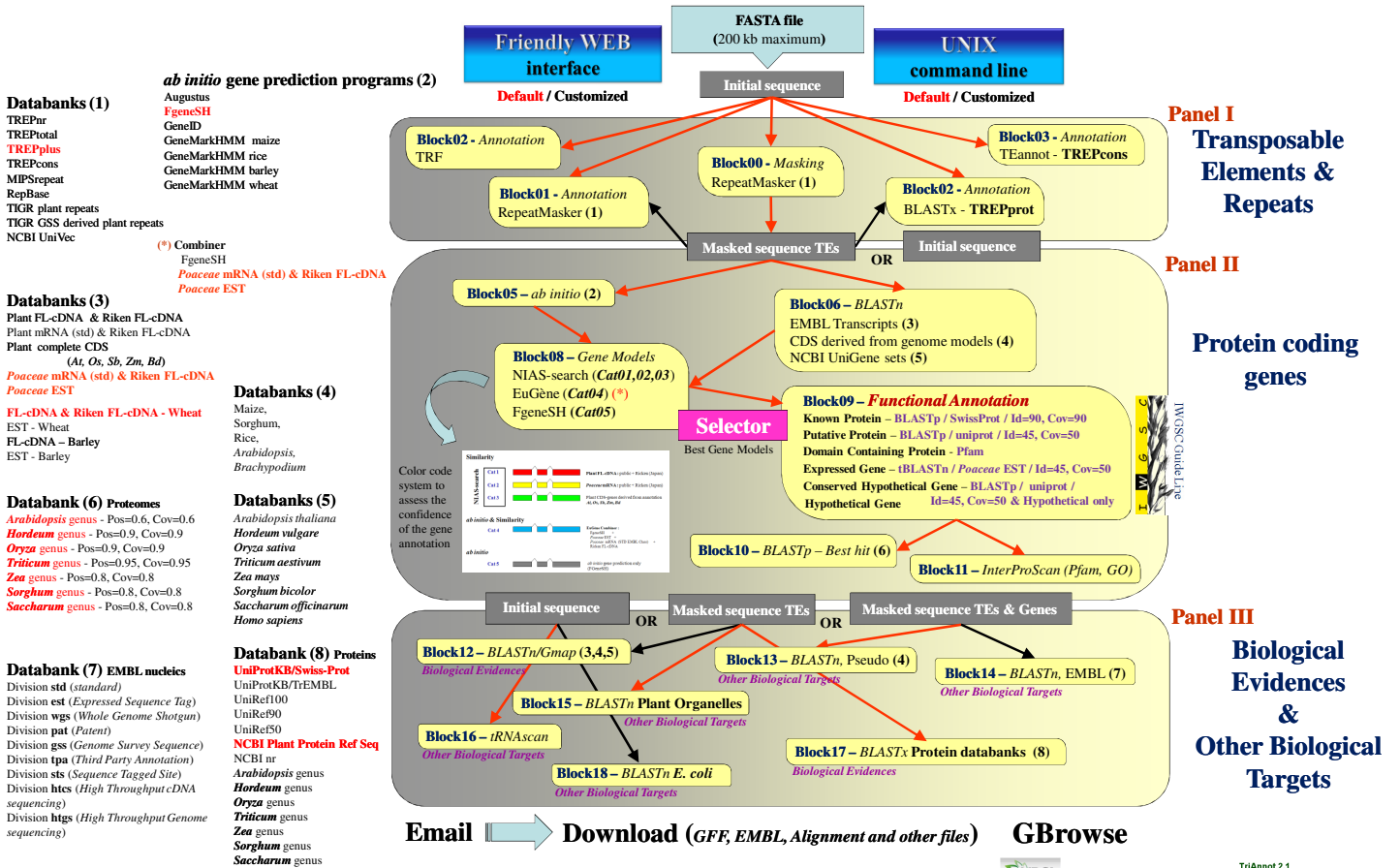
1. INRA-UBP, UMR 1095, 234 Avenue du Brézat, F-63100 Clermont-Ferrand, France.  
 2. INRA-URGI, Route de Saint Cyr, F-78000, Versailles, France  
 3. National Institut of Agrobiology Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan.  
 4. MIPS/IBIS Institute of Bioinformatics and System Biology, Helmholtz Center Munich, D-85764 Neuherberg, Germany.

Annotation is one of the most difficult tasks in genome sequencing projects, yet it is essential for connecting genome sequence to biology. Structural and Functional annotations consist in determining the position and structure of genes as well as of other features such as transposable elements and non coding RNA, and inferring their putative function in the genome. This requires a complex successive combination (pipeline or workflow) of software, algorithms and methods. Automation of such a pipeline is necessary to manage large amounts of data released by genome sequencing projects. To achieve a systematic and comprehensive annotation of the wheat genome sequence (17 Gb), a pipeline called TriAnnot V2.1 has been developed by INRA Clermont-Ferrand (GDEC) and Versailles (URGI) in partnership with NIAS (Japan), under the umbrella of the IWGSC (*International Wheat Genome Sequencing Consortium* - <http://www.wheatgenome.org>). The objective of TriAnnot is to provide the international scientific community with an online user friendly, fast and as complete as possible annotation tool in view of the sequencing of the wheat genome. As it is the case for every workflow, the TriAnnot pipeline should minimize manual expertise which is slow and labor-intensive, and maximize relevant automatic annotation which is a relatively rapid process that allows frequent updates to accommodate new data. TriAnnot V2.1 will be a major tool within the ANR/FranceAgriMer 3BSEQ flagship project which objectives are to obtain, annotate and utilize the whole wheat chromosome 3B (1Gb) sequence.

~ <http://www.clermont.inra.fr/triannot/> ~

January 4th, 2011

## TriAnnot Workflow



- Databanks (1)**
- TREPnr
  - TREPtotal
  - TREPplus
  - TREPcons
  - MIPRepeat
  - RepBase
  - TIGR plant repeats
  - TIGR GSS derived plant repeats
  - NCBI UniVec

- ab initio gene prediction programs (2)**
- Augustus
  - FgeneSH
  - GeneID
  - GeneMarkHMM maize
  - GeneMarkHMM rice
  - GeneMarkHMM barley
  - GeneMarkHMM wheat

- Databanks (3)**
- Plant FL-cDNA & Riken FL-cDNA
  - Plant mRNA (std) & Riken FL-cDNA
  - Plant complete CDS (At, Os, Sb, Zn, Bd)
  - Poaceae mRNA (std) & Riken FL-cDNA
  - Poaceae EST
  - FL-cDNA & Riken FL-cDNA - Wheat
  - EST - Wheat
  - FL-cDNA - Barley
  - EST - Barley

- (\*) Combiner**
- FgeneSH
  - Poaceae mRNA (std) & Riken FL-cDNA
  - Poaceae EST

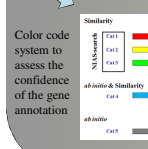
- Databanks (4)**
- Maize, Sorghum, Rice, Arabidopsis, Brachypodium

- Databank (6) Proteomes**
- Arabidopsis genus - Pos=0.6, Cov=0.6
  - Hordeum genus - Pos=0.9, Cov=0.9
  - Oryza genus - Pos=0.9, Cov=0.9
  - Triticum genus - Pos=0.95, Cov=0.95
  - Zea genus - Pos=0.8, Cov=0.8
  - Sorghum genus - Pos=0.8, Cov=0.8
  - Saccharum genus - Pos=0.8, Cov=0.8

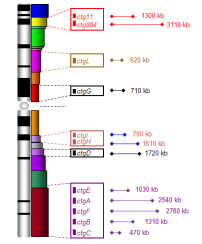
- Databanks (5)**
- Arabidopsis thaliana
  - Hordeum vulgare
  - Oryza sativa
  - Triticum aestivum
  - Zea mays
  - Sorghum bicolor
  - Saccharum officinarum
  - Homo sapiens

- Databank (7) EMBL nucleics**
- Division std (standard)
  - Division est (Expressed Sequence Tag)
  - Division wgs (Whole Genome Shotgun)
  - Division pat (Patent)
  - Division gss (Genome Survey Sequence)
  - Division tpa (Third Party Annotation)
  - Division sts (Sequence Tagged Site)
  - Division htes (High Throughput cDNA sequencing)
  - Division htgs (High Throughput Genome sequencing)

- Databank (8) Proteins**
- UniProtKB/Swiss-Prot
  - UniProtKB/TrEMBL
  - UniRef100
  - UniRef90
  - UniRef50
  - NCBI Plant Protein Ref Seq
  - NCBI nr
  - Arabidopsis genus
  - Hordeum genus
  - Oryza genus
  - Triticum genus
  - Zea genus
  - Sorghum genus
  - Saccharum genus



## Performance Evaluation

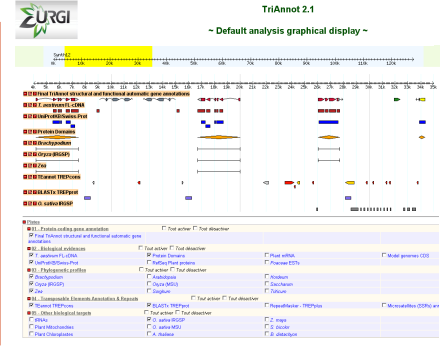


Data set manually expertized  
 Wheat 3B chromosome (1Gb)  
 12 contigs - 18.2 Mb - 148 genes  
 Choulet et al. (2010)  
 Plant Cell 22:1686-1701

**Fitness calculated with Eval**  
 Keibler & Brent (2003)  
 BMC Bioinformatics 4:50

Pipeline	Predicted Genes	TP (%)	Gene		Exon		Fitness (%)
			Sr	Sp	Sr	Sp	
FPGP	304	69	46.6	22.7	71.3	58.3	45.8
MIPS	215	53	35.1	24.2	61.1	50.8	40.3
RiceGAAS	848	52	35.1	6.1	70.2	18.0	22.9
TriAnnot Full	371	75	50.7	20.2	67.6	45.5	42.1
TriAnnot without FgeneSH	265	65	43.9	24.5	65.1	53.5	44.0
TriAnnot without TrEgine and FgeneSH	105	61	41.2	58.1	60.7	85.8	59.4

(1) TP = number of True Positive genes  
 (2) Fitness = (SnG\*SpG\*SnE\*SpE)\*22



## Conclusion

TriAnnot V2.1 tries to combine the best features of well known international pipelines. It provides the best compromise in term of fitness, Sensitivity (Sn) and Specificity (Sp). TriAnnot proposes a color code system to assess the confidence of the structural automatic gene model annotation. It gives a functional annotation with Pfam domain protein & Gene Ontology, as well as alignment of best hits against proteome databanks. Additional functionalities are under development (EuGene, RNAseq, miRNA).

Pipeline	On Web	Repeat Mask	Gene Model	TEs annotation	Customize analysis	Data Download	email	Graphical display	Job monitoring
DNA Subway	X	X	-	X	x	gff (copy/paste)	-	X	X
RiceGAAS	X	-	X	X	-	-	-	X	-
FPGP	X	-	-	-	-	gff	X	-	-
MIPS	-	X	X	X	-	-	-	-	-
TriAnnot	X	X	X	X	X	gff/emb	X	X	planned