

***Advanced Genomics - Bioinformatics Workshop***

***Vamalwa***  
*RI Hub, Nairobi, Kenya*  
[hub.africabiosciences.org/](http://hub.africabiosciences.org/)  
[amalwa@cgiar.org](mailto:amalwa@cgiar.org)



7<sup>th</sup> – 18<sup>th</sup> September 2015

**biosciences**  
eastern and central afri

# What can be learned from gene annotation?

- The genome is an organismal blueprint
- Annotation of protein coding genes provides a “parts list” for an organism
- Analysis of individual genes can be highly informative
  - Predicted molecular function
  - Biological significance of multi copy genes (paralogs)
  - Gene sequence facilitates experimental wet work
    - Gene expression analysis (sex specificity, tissue specificity, stage specificity)
    - Antibody production
    - In vitro protein expression for protein function analysis
    - Gene knockdown
    - Ectopic expression


# Why compare genes/genomes?


- Historical record of evolutionary events
- Similarities and differences provide clues to speciation and differences in function
- Differences to watch for
  - Species specific genes
  - Sequence differences between orthologs are informative about
    - evolutionary relationships between species
    - evolutionary rate of a gene family can demonstrate purifying selection or positive selection
  - Functional differences between orthologous protein coding sequences
  - Gene family expansions/contractions (novel paralogs)
  - Gene localization relative to other genes (synteny) can indicate genomic rearrangements
- Correlate genetic differences to morphological, physiological and behavioral traits

# How do you compare genes/genomes?

- Compare general features of whole genomes between species
  - Genome size
  - Nucleotide composition - A/T to G/C ratios
  - Large scale features such as the presence of repetitive sequences/transposable elements
  - Number of gene orthologs shared between species
- Focus on genomic features associated with biological function
  - Development
  - Plant Immunity
  - Metabolism
  - Molecular signaling
  - Symbiosis
  - Reproduction
  - Chemosensation
  - Host seeking
  - Housekeeping

# Comparative Analysis Pipeline

- 
- Choose a gene of interest from Rice, *A. thaliana* or related characterised organism
  - Identify gene homologs/orthologs in other species by homology based BLAST analysis

- 
- Obtain protein and nucleotide sequences for putative gene homologs or orthologs
  - Generate alignments to determine similarity and identity between homologous sequences and identify orthologs and paralogs

- 
- Analyse genomic loci to compare gene structure (exon/intron structure), synteny of surrounding genes, correct gene models

# Choosing a gene of interest

- Choose gene of interest from related characterised organism (Rice, *A. thaliana*)
- Identify gene homologs/orthologs in other species by homology based BLAST analysis

- Obtain protein and nucleotide sequences for putative gene homologs or orthologs
- Generate alignments to determine similarity and identity between homologous sequences and identify orthologs and paralogs

- Analyse genomic loci to compare gene structure (exon/intron structure), synteny of surrounding genes, correct gene models

- Chosen based on characterized function in an organism as described in the literature
- Identified by experimental evidence (such as RNA-Seq)

Sequences for genes/proteins of interest can be obtained from a variety of locations

## The Rice Annotation Project

<http://rapdb.dna.affrc.go.jp/>

Genomic database for *Oryza sativa* ssp. *japonica* cv.

## Rice Genome Annotation Project

<http://rice.plantbiology.msu.edu/index.shtml>

Rice Genome Annotation Project Database and Resource

## National Center for Biotechnology Information NCBI

[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

US national resource for molecular biology information

## Finger Millet gene annotation tables

[Annotation Tables](#)

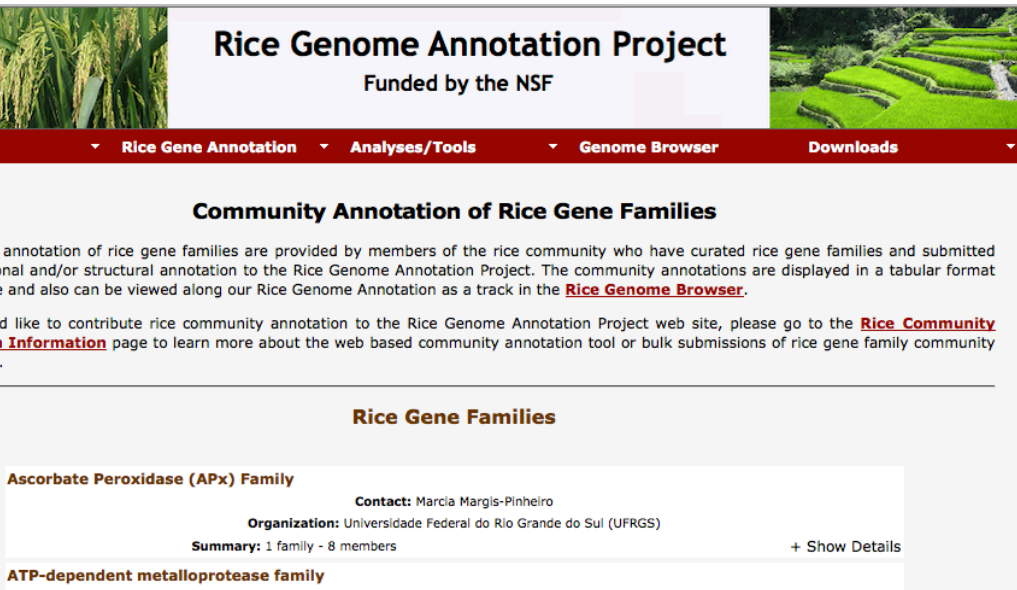
Tables containing the identifiers and annotations for genes annotated in the *Finger Millet* genome.

## Uniprot/TrEMBL

Curated protein database at European Molecular Biology Labs

# Choosing a gene of interest

Search by Functional category at RGAP



**Rice Genome Annotation Project**  
Funded by the NSF

Rice Gene Annotation | Analyses/Tools | Genome Browser | Downloads

### Community Annotation of Rice Gene Families

Community annotation of rice gene families are provided by members of the rice community who have curated rice gene families and submitted functional and/or structural annotation to the Rice Genome Annotation Project. The community annotations are displayed in a tabular format and also can be viewed along our Rice Genome Annotation as a track in the [Rice Genome Browser](#).

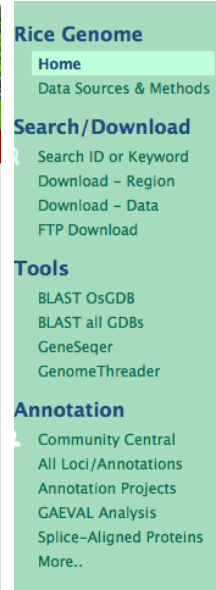
If you would like to contribute rice community annotation to the Rice Genome Annotation Project web site, please go to the [Rice Community Annotation Information](#) page to learn more about the web based community annotation tool or bulk submissions of rice gene family community annotations.

#### Rice Gene Families

**Ascorbate Peroxidase (APx) Family**

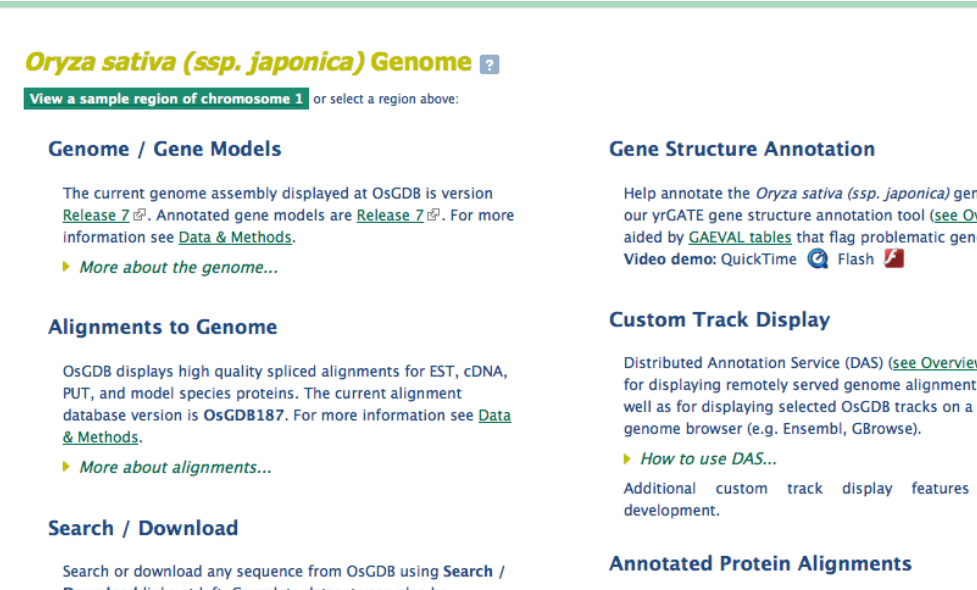
**Contact:** Marcia Margis-Pinheiro  
**Organization:** Universidade Federal do Rio Grande do Sul (UFRGS)  
**Summary:** 1 family - 8 members [+ Show Details](#)

**ATP-dependent metalloprotease family**



**Rice Genome**

- Home
- Data Sources & Methods
- Search/Download
  - Search ID or Keyword
  - Download - Region
  - Download - Data
  - FTP Download
- Tools
  - BLAST OsGDB
  - BLAST all GDBs
  - GeneSeqer
  - GenomeThreader
- Annotation
  - Community Central
  - All Loci/Annotations
  - Annotation Projects
  - GAEVAL Analysis
  - Splice-Aligned Proteins
  - More..
- Other Resources



## *Oryza sativa (ssp. japonica)* Genome ?

[View a sample region of chromosome 1](#) or select a region above:

### Genome / Gene Models

The current genome assembly displayed at OsGDB is version [Release 7](#). Annotated gene models are [Release 7](#). For more information see [Data & Methods](#).

[More about the genome...](#)

### Alignments to Genome

OsGDB displays high quality spliced alignments for EST, cDNA, PUT, and model species proteins. The current alignment database version is **OsGDB187**. For more information see [Data & Methods](#).

[More about alignments...](#)

### Search / Download

Search or download any sequence from OsGDB using [Search / Download](#) links at left. Complete datasets can also be

### Gene Structure Annotation

Help annotate the *Oryza sativa (ssp. japonica)* genome using our yrGATE gene structure annotation tool ([see Overview](#)) aided by [GAEVAL tables](#) that flag problematic gene models. [Video demo:](#) QuickTime [Flash](#)

### Custom Track Display

Distributed Annotation Service (DAS) ([see Overview](#)) for displaying remotely served genome alignments as well as for displaying selected OsGDB tracks on a genome browser (e.g. Ensembl, GBrowse).

[How to use DAS...](#)

Additional custom track display features under development.

### Annotated Protein Alignments

**Rice Genome Annotation Project**  
[http://rice.plantbiology.msu.edu/annotation\\_community\\_families.shtml](http://rice.plantbiology.msu.edu/annotation_community_families.shtml)

Database of *Rice* genes and genomes

Gene of interest identified by differential expression analysis (RNA-Seq)



# Homologous/Orthologous Gene Identification

1.

- Choose gene of interest from related characterised organism (Rice, *A. thaliana*)
- Identify gene homologs/orthologs in other species by homology based BLAST analysis

2.

- Obtain protein and nucleotide sequences for putative gene homologs or orthologs
- Generate alignments to determine similarity and identity between homologous sequences and identify orthologs and paralogs

3.

- Analyse genomic loci to compare gene structure (exon/intron structure), synteny of surrounding genes, correct gene models

Sequences for BLAST analysis can be downloaded from your database of choice such as RAP-DB, <http://rice.plantbiology.msu.edu> or obtained from local files ([Finger Millet predicted proteome](#) + [Finger millet predicted transcriptome](#))



```
>my gene of interest pep:NOVEL_protein_coding  
MTLLDLVHERNQLTMKLCIIFTVLAVAANITTALRAFAVIKNMLDCHERLGINEEDLMVIQDLSDIKAASEYTPGQQCSIYCQSEAYGFTRRGQ  
KWFMRKQPRIAQKYNLDKVFNCK RYATDTCDGPIHLAQCAQQYPLQAGDRNP
```



BLAST Results Reveal Multiple Hits Across the *Glossina* Genomes (No significant *Musca* hits)

### *Ricei*

GAUT029311-PA  
GAUT029310-PA  
GAUT029308-PA

### *Sorghum Bicolor*

GBRI010920-PA  
GBRI010919-PA  
GBRI010924-PA  
GBRI010929-PA

### *A. thaliana*

GMOY005874-PA  
GMOY005875-PA  
GMOY005876-PA



# Obtain Sequences of Interest

1.

- Choose gene of interest from related characterised organism (Rice, *A. thaliana*)
- Identify gene homologs/orthologs in other species by homology based BLAST analysis

2.

- Obtain protein and nucleotide sequences for putative gene homologs or orthologs
- Generate alignments to determine similarity and identity between homologous sequences and identify orthologs and paralogs

3.

- Analyse genomic loci to compare gene structure (exon/intron structure), synteny of surrounding genes, correct gene models

Sequences can be downloaded from your database of choice such as RAP-DB, <http://rice.plantbiology.msu.edu> or obtained from local files ([Finger Millet predicted proteome](#) + [Finger millet predicted transcriptome](#))

# Best hits against RCOY005874

>GMOY005874:GMOY005874-RA peptide: GMOY005874-PA pep:NOVEL\_protein\_coding  
MTLLDLVHERNQLTMKLCIIFTVLAVAANITTALRAFAVIKNMLDCHERLGINEEDLMVIQDLSDIKA  
ASEYTPGQQCSIYQCSEAYGFTRRGQLKKWFMRKQPRIAQKYNLDKVFQNCK  
RYATDTCDGPIHLAQCAQQYPLQAGDRNP

*aliana*

>GAUT029311:GAUT029311-RA peptide: GAUT029311-PA pep:\_protein\_coding  
MKLFIILTVLAVAANIASALRAFAVIKNMLDCHERLGISEEDLMVVQDLSDIKSASEYTP  
GQQCSIYQCSEAYGFTRRGQLKKWFMRKQPRIAQKYNLDKVFQNCKRYATDTCDGPIHLA  
ICAQQYPLHAGERNL

*um bicolor*

>GBRI010920:GBRI010920-RA peptide: GBRI010920-PA pep:\_protein\_coding  
MKFCIILIVLVAAANTASAIRAFAVIKNMLNCHERLGISEDDLTVVQDLSDVKAPSEYTA  
GQKCSIYQCSEAYGFTRKRGQLKKWFMRKQPRIAHRYNLDKAFSHCQEYATDTCDGPIQLA  
RCVQQFPMHA

# Generate Alignments to determine Identity/Similarity

1.

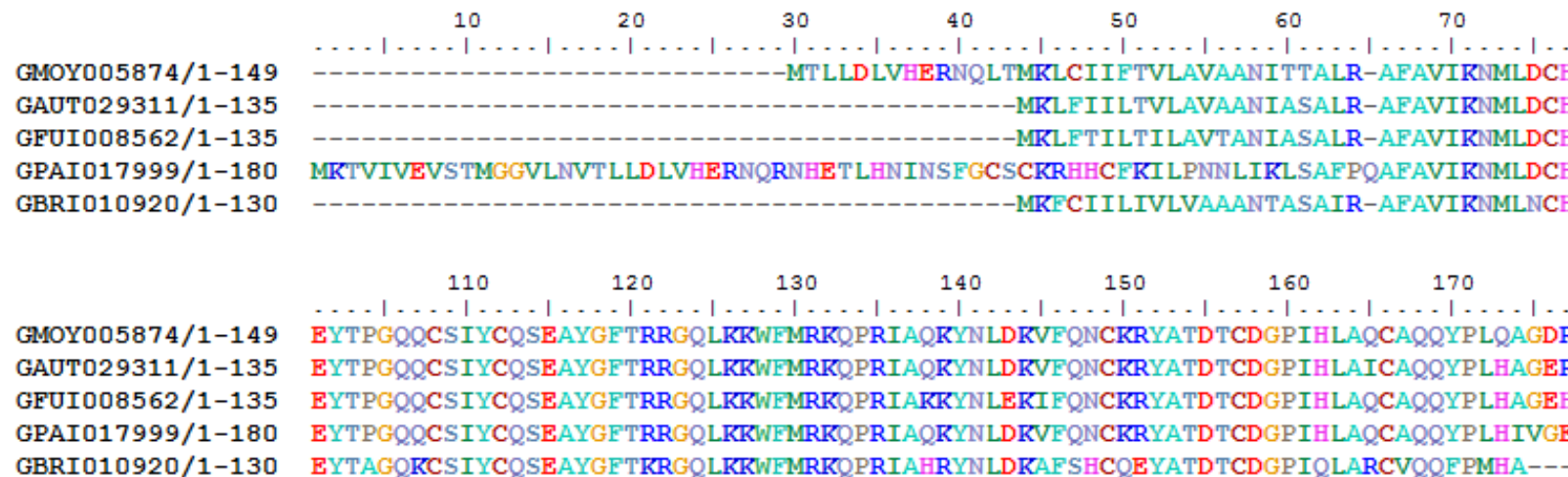
- Choose gene of interest from related characterised organism (Rice, *A. thaliana*)
- Identify gene homologs/orthologs in other species by homology based BLAST analysis

2.

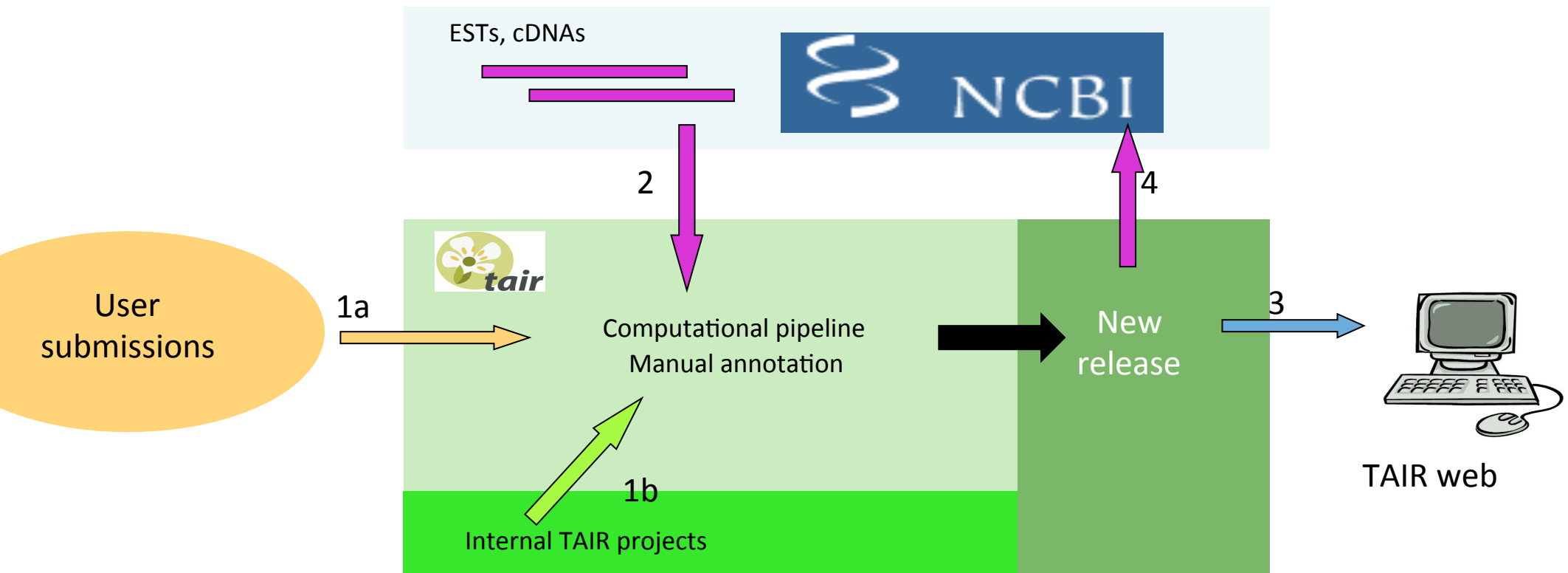
- Obtain protein and nucleotide sequences for putative gene homologs or orthologs
- Generate alignments to determine similarity and identity between homologous sequences and identify orthologs and paralogs

3.

- Analyse genomic loci to compare gene structure (exon/intron structure), synteny of surrounding genes, correct gene models

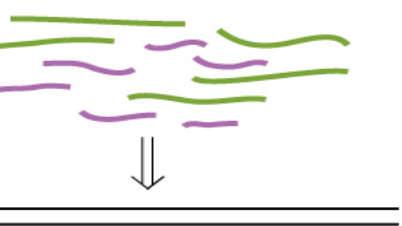


# Web server Example: Generating & curating gene models at TAIR

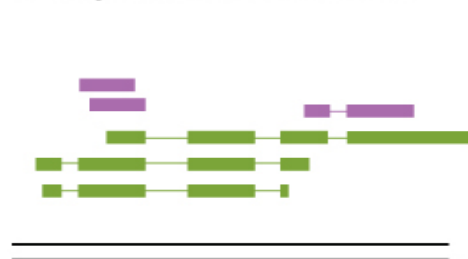


# Generating & curating gene models at TAIR

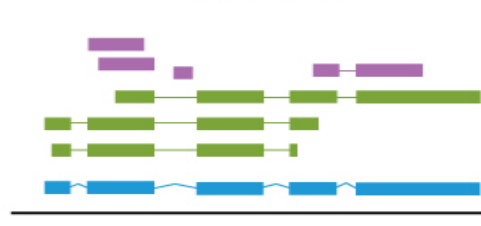
1 Align cDNAs and ESTs to genome using Sim4 and Blat



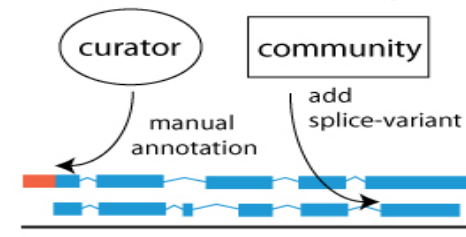
2 Aligned cDNAs and ESTs



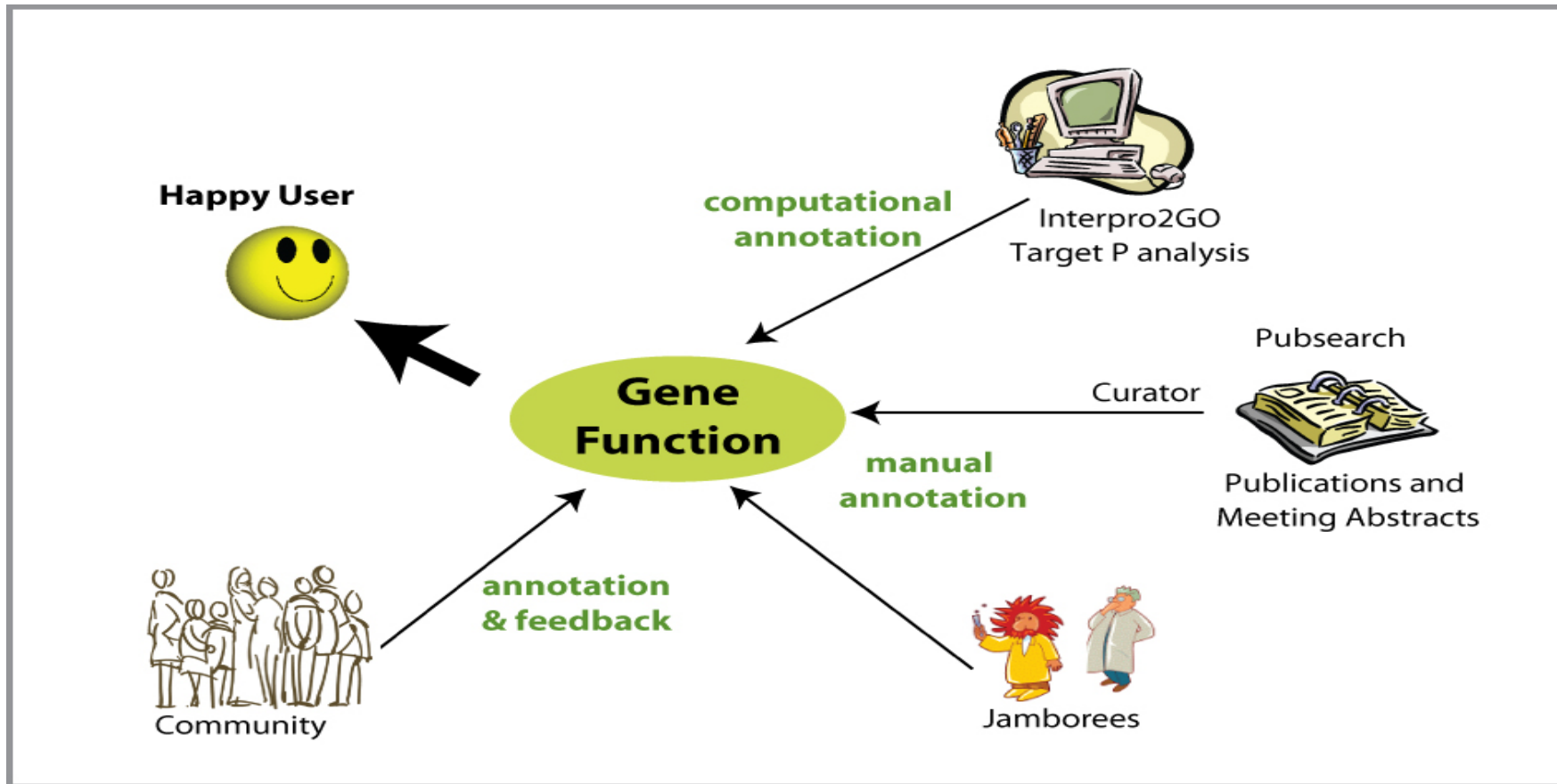
3 Use PASA to generate gene models



4 Curation by curator and community



# Functional gene annotation



1.

- Choose gene of interest from related characterised organism (Rice, *A. thaliana*)
- Identify gene homologs/ orthologs in other species by homology based BLAST analysis

2.

- Obtain protein and nucleotide sequences for putative gene homologs or orthologs
- Generate alignments to determine similarity and identity between homologous sequences and identify orthologs and paralogs

3.

- Analyse genomic loci to compare gene structure (exon/ intron structure), synteny of surrounding genes, correct gene models

# Use WebApollo for gene structure analysis and gene model correction

## [WebApollo Tutorial](#)

The WebApollo instances for the genomes can be accessed from the following links.

[Finger Millet](#)

[Rice](#)

[Maize](#)

[A. \*Thaliana\*](#)

- WebApollo can be used to examine gene models, exon/intron structure
- Can also be used to edit gene models if they are erroneous



# Tools Provided for Local Analysis

- We have provided the installation files for programs that can be used to extract, manipulate and analyze data from the sequence databases
- The installation files can be found [here](#)  
<http://hpc.ilri.cgiar.org/beca/training/FingerMillet2015/>
- The programs include
  - Bioedit/UGENE
    - PC based sequence viewing and analysis program
    - Can perform alignments and local BLAST analyses
    - All tsetse data has been provided as BLAST databases
    - Instructions for setting up BioEdit with WINE on a Mac can be found [here](#).
  - Integrated Genome Browser
    - Sequence/Genome viewing software for PC, Mac and Unix
    - Can display large scaffolds with the annotation data from Vectorbase
  - MEGA 6.0
    - Popular alignment and phylogenetic analysis software for PC and Mac

# Annotation Goals + Groups

- Follow your interests/passion!
- Group up with people interested in the same topic
- Identify, annotate and compare genes of interest
- Summarize and explain interesting findings to the group

# Instructions for Metadata Collection Template

## Course Website:

<http://hpc.ilri.cgiar.org/beca/training/FingerMillet2015/>  
[http://hpc.ilri.cgiar.org/beca/training/FingerMillet2015/  
program.html](http://hpc.ilri.cgiar.org/beca/training/FingerMillet2015/program.html)