Gene annotation

Bioplatforms Fungi Genomics Workshop 2024
Australian National University

Rita Tam
Zhenyan Luo

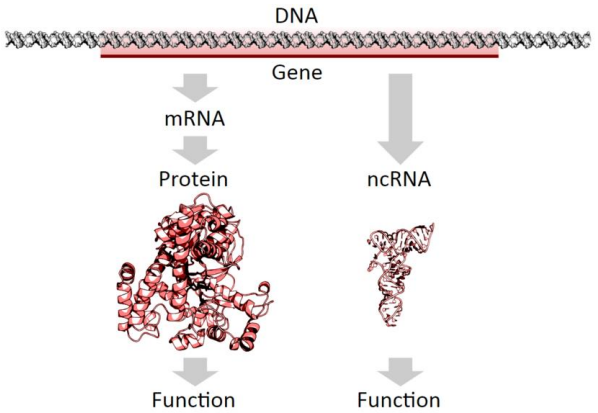
TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

“My genome assembly is quality-controlled, looking good! What now?”

Your assembly `.fasta` file contains nothing more than just nucleotides…

Find genes to give biological meaning to it!

What can genes tell you?



Changes in gene repertoires underlie **phenotypic evolution**

e.g. development, metabolism, reproduction, pathogenicity...

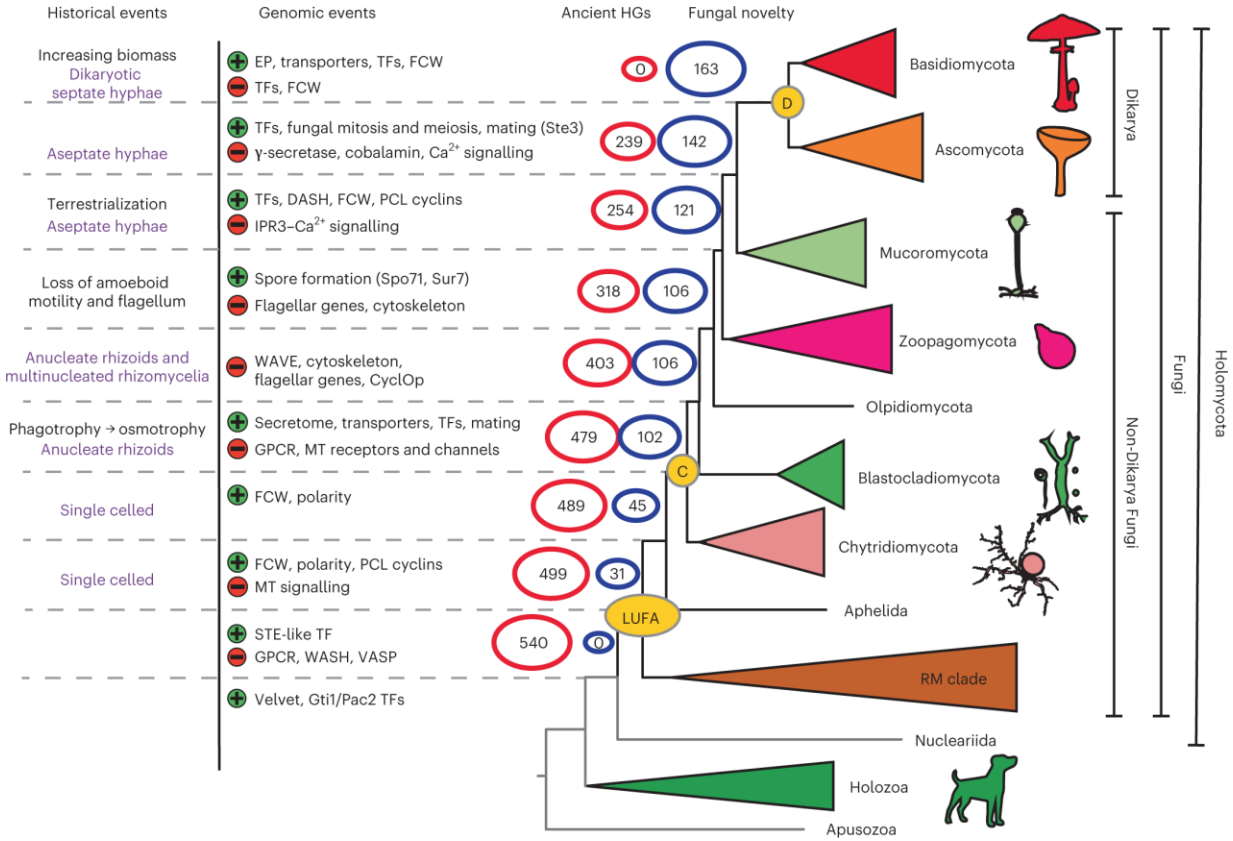
Gene family evolution enables **adaptation and speciation**

Gene can be viewed as **evolutionary units** that evolve independently from species tree (sometimes)

Gabaldón 2024; Thomas Shafee

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

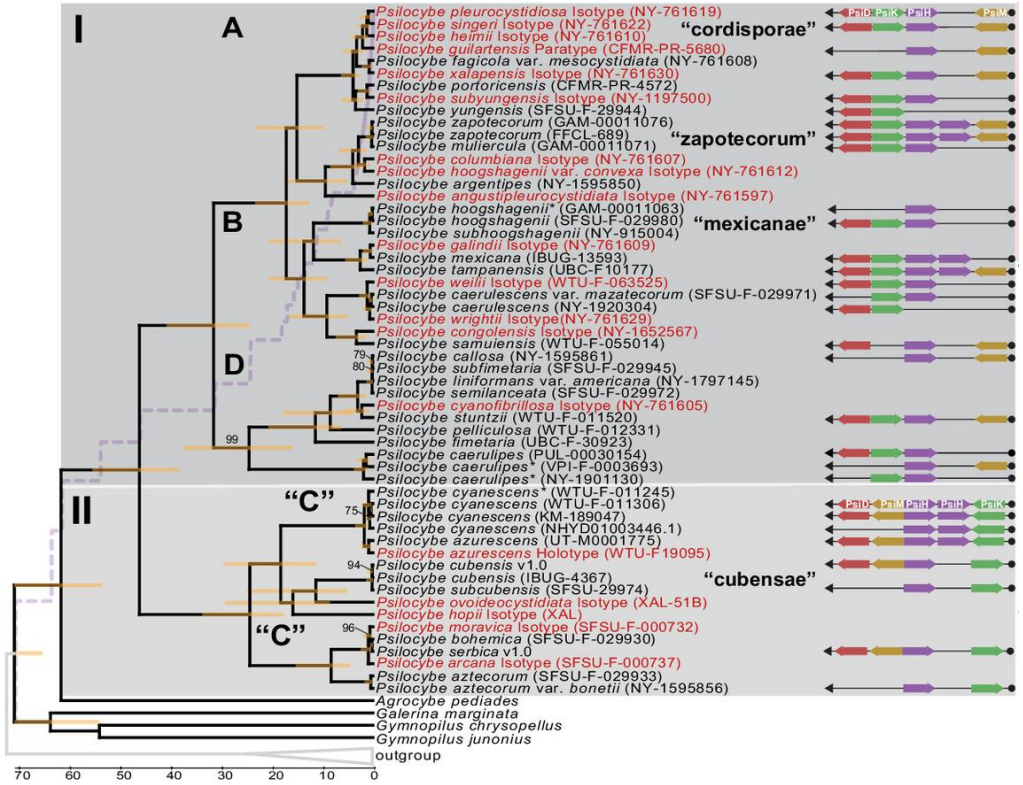
Genes retain footprints of historical evolutionary events



Merényi et al, 2023

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Evolution of psilocybin biosynthetic gene clusters

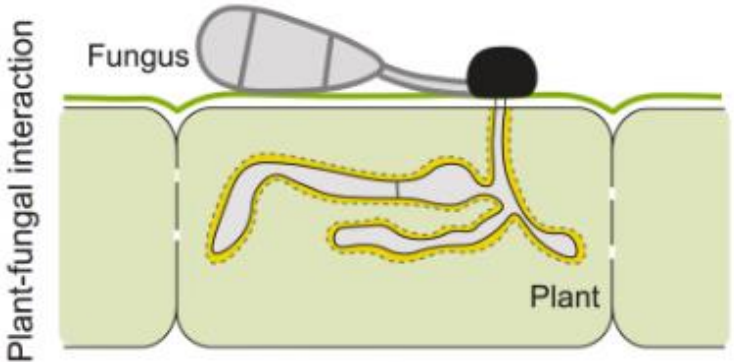


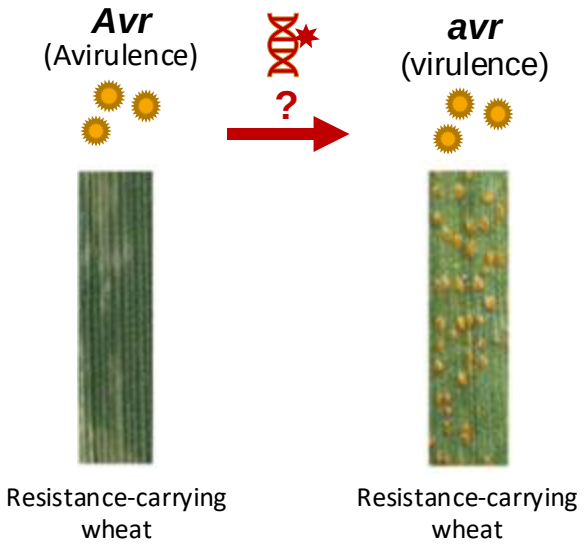
Psilocybe aka "magic mushroom"

Bradshaw et al. 2024

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Avirulence genes in plant fungal pathogens





- Rust fungi diversify avirulence proteins to manipulate & evade host immunity
- Almost no sequence similarity between avirulence proteins → hard to identify

- Start from finding signalling peptide in their N-terminus

Huang, Liu & Cook, 2023

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Now with a genome assembly,

1. How do I find genes?
2. How do I know their functions?

Two main steps of gene annotation

1. Gene prediction

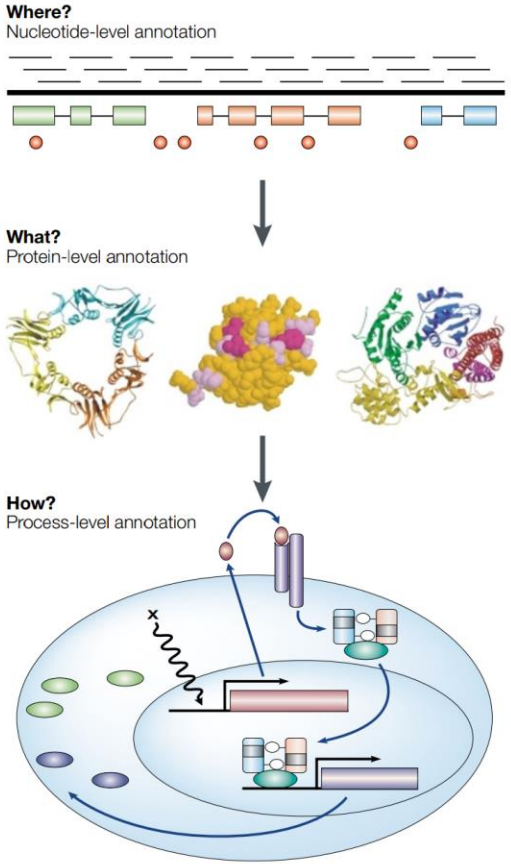
Protein-coding, tRNA, rRNA

UTRs, exons, introns

Alternative splice isoforms (optional)

2. Functional annotation

Assign biological functions to predicted genes



Stein, 2001

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Two main steps of gene annotation

1. Gene prediction

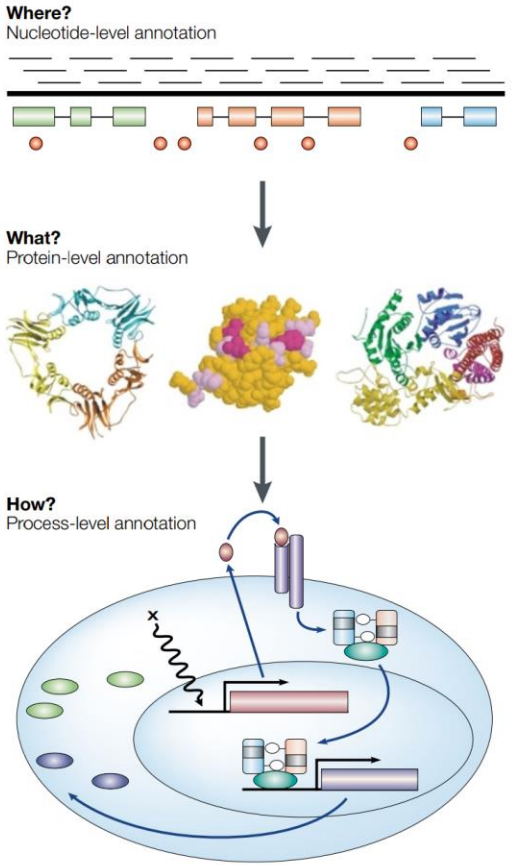
Protein-coding, tRNA, rRNA

UTRs, exons, introns

Alternative splice isoforms (optional)

2. Functional annotation

Assign biological functions to predicted genes



Stein, 2001

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Different formats of evidence for gene prediction

Protein evidence (Optional)

- Protein database: UniProtKb, Pfam
- Protein sequences of related species

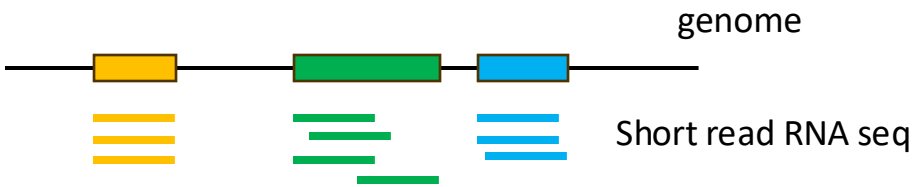
Transcript evidence

- RNA seq (fastq): Illumina single/pair-end RNA seq, ONT RNA seq
- RNA bam file: RNA seq mapped to genome
- Transcriptome assembly: assembled RNA seq

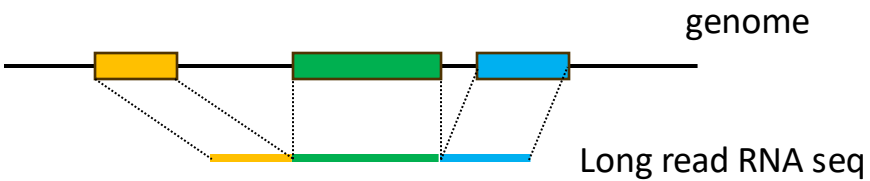
TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Mapping RNA seq to genome assembly

Hisat2 : Mapping short read RNA seq



Minimap2 : Mapping long read RNA seq with splice-aware mode

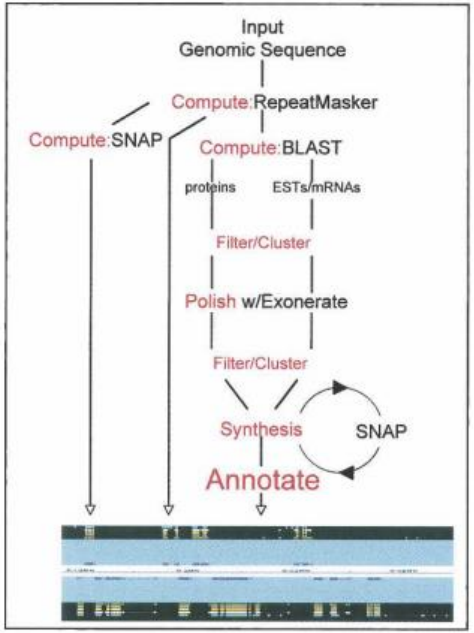


TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

MAKER for gene prediction

MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes

Brandi L. Cantarel[1], Ian Korf[2], Sofia M.C. Robb[3], Genis Parra[2], Eric Ross[4], Barry Moore[1], Carson Holt[1], Alejandro Sánchez Alvarado[3,4], and Mark Yandell[1,5]



Use RepeatMasker, BLAST, SNAP and exonerate as dependencies

Use Genome, ESTs/cDNA, protein as input

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

FGENESH for gene prediction

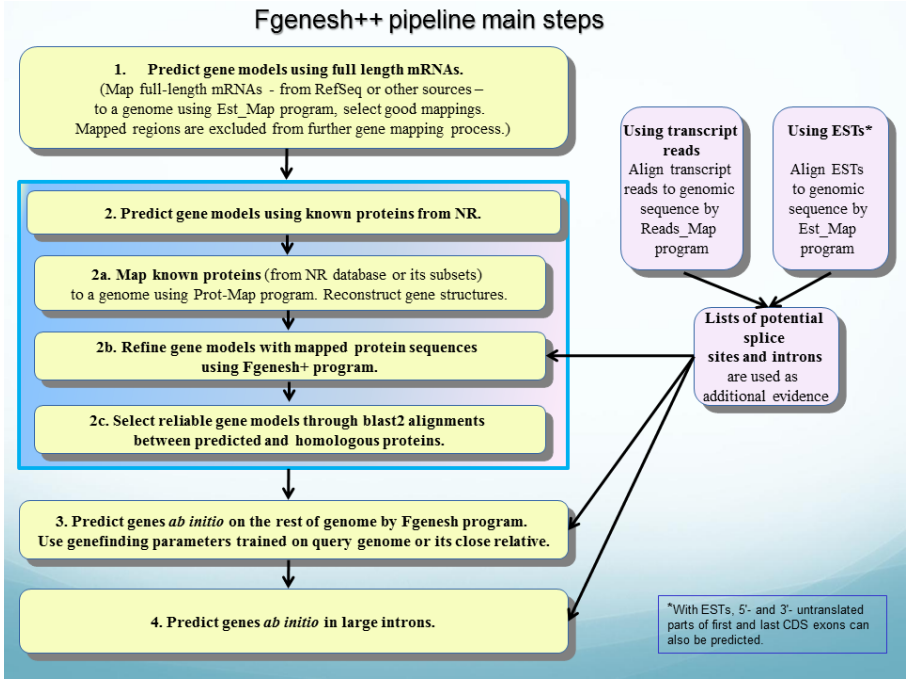
Research | Open access | Published: 07 August 2006

Automatic annotation of eukaryotic genes, pseudogenes and promoters

Victor Solovyev, Peter Kosarev, Igor Seledsov & Denis Vorobyev

Genome Biology **7**, Article number: S10 (2006) | Cite this article

15k Accesses | **569** Citations | Metrics



Use genome, mRNAs, protein as input

Web version of FGENESH can be used with parameters for genomes of human, mouse, *Drosophila*, nematode, dicot plants, monocot plants, yeast and *Neurospora*.

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

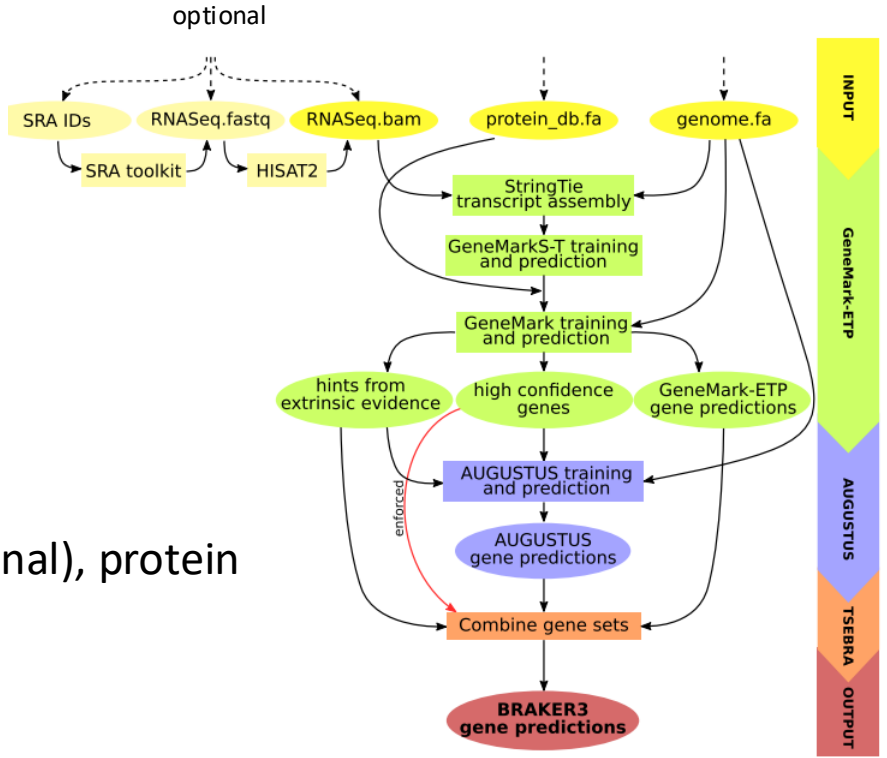
BRAKER3 for gene prediction

Method

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA

Lars Gabriel,[1,2] Tomáš Brůna,[3] Katharina J. Hoff,[1,2] Matthis Ebel,[1,2] Alexandre Lomsadze,[4] Mark Borodovsky,[4,5,6] and Mario Stanke[1,2,6]

[1]*Institute of Mathematics and Computer Science, University of Greifswald, 17489 Greifswald, Germany;* [2]*Center for Functional Genomics of Microbes, University of Greifswald, 17489 Greifswald, Germany;* [3]*U.S. Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA;* [4]*Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA;* [5]*School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*



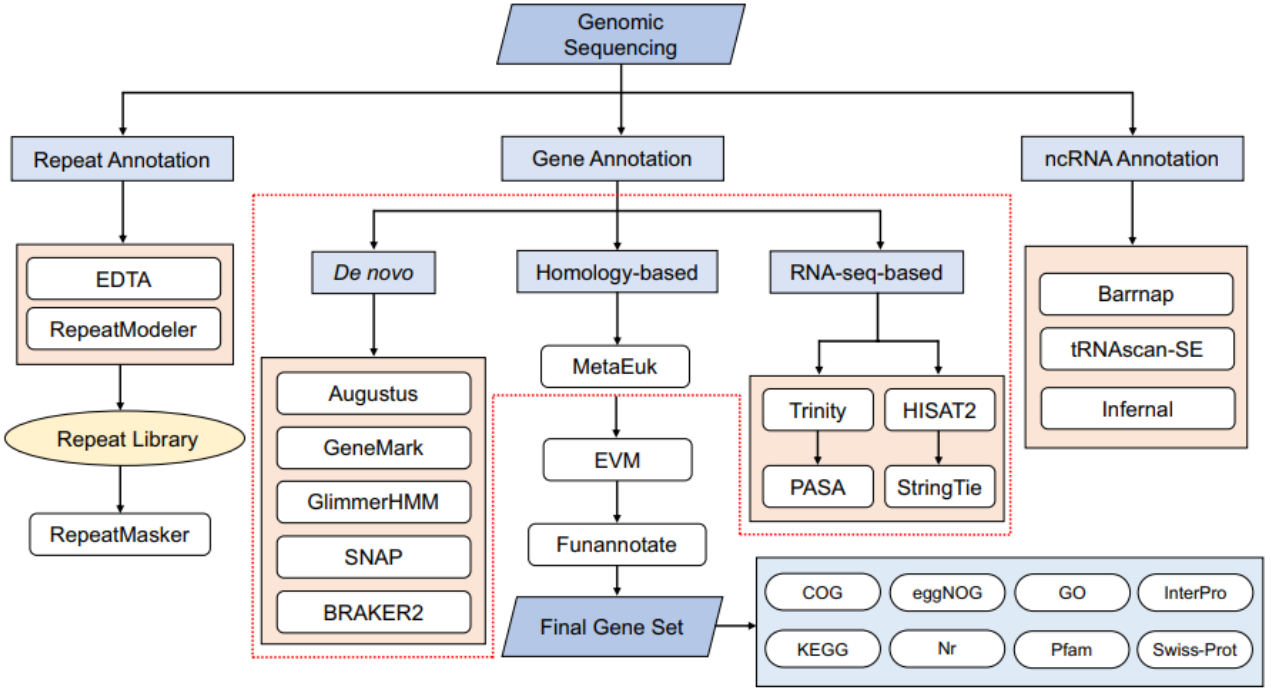
Use soft-masked genome, ESTs/cDNA (optional), protein as input

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Funannotate for gene predicti

Use genome, ESTs/cDNA, protein as input

Merge results from different gene prediction tools



clean –> sort –> mask –> train –> predict –> update –> annotate –> compare

Wang et al. 2023

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Helixer for gene prediction

JOURNAL ARTICLE

Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning

Felix Stiehler, Marvin Steinborn, Stephan Scholz, Daniela Dey, Andreas P M Weber, Alisandra K Denton

Bioinformatics, Volume 36, Issue 22-23, December 2020, Pages 5291–5298,
https://doi.org/10.1093/bioinformatics/btaa1044
Published: 16 December 2020 **Article history** ▾

Model is trained to differentiate Intergenic, untranslated, coding and intron

Using Only DNA sequence as input

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Main output of gene prediction

.proteins.fa	Multi-fasta file of protein coding genes
.cds-transcripts.fa	Multi-fasta file of transcripts (mRNA)
.scaffolds.fa	Genome file same as the input
.gff3	Genome annotation in GFF3 format
.gbk	Annotated Genome in GenBank Flat File format

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Main output of gene prediction

.gff3 gene annotation information

Seq id	Source	Type	Start	End		Strand		Features
AU3_HapA_CHR01	funannotate	gene	119247	119936	.	-	.	ID=MK676_000001;
AU3_HapA_CHR01	funannotate	mRNA	119247	119936	.	-	.	ID=MK676_000001-T1;Parent=MK676_000001;product=hypothetical protein;
AU3_HapA_CHR01	funannotate	exon	119757	119936	.	-	.	ID=MK676_000001-T1.exon1;Parent=MK676_000001-T1;
AU3_HapA_CHR01	funannotate	exon	119247	119726	.	-	.	ID=MK676_000001-T1.exon2;Parent=MK676_000001-T1;
AU3_HapA_CHR01	funannotate	CDS	119757	119936	.	-	0	ID=MK676_000001-T1.cds;Parent=MK676_000001-T1;
AU3_HapA_CHR01	funannotate	CDS	119247	119726	.	-	0	ID=MK676_000001-T1.cds;Parent=MK676_000001-T1;

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Main output of gene prediction

.gbk Genebank format of annotation

```
gene            complement(374922..375452)
                /locus_tag="MK676_012878"
mRNA            complement(374922..375452)
                /locus_tag="MK676_012878"
                /product="hypothetical protein"
CDS             complement(374922..375452)
                /locus_tag="MK676_012878"
                /codon_start=1
                /product="hypothetical protein"
                /protein_id="ncbi:MK676_012878-T1"
                /translation="MGHSEADNTILPSKGADTATTSLRGHIQSQPQCIQQCTSVQGLP
                DPCRTVEEWHKFLPECEKIRGESQYLQIAQWMASIHGEQKHDSIDTRMEEKQPSTTQT
                SAKNSPSGQQRKFQCEKAATSSKKGKGKAPAPKPYNQGYRIPKIQQDAIENVFRMART
                MMELQKKGEARLKYQK"
```

Position of genes, mRNA, CDS and corresponding protein sequences are included

Can be used as input of funannotate annotate

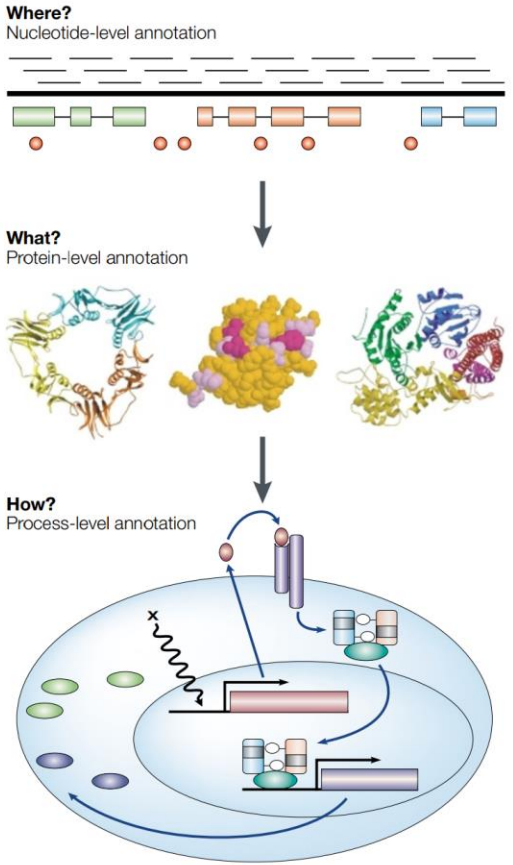
TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Two main steps of gene annotation

1. Gene prediction
   - Protein-coding, tRNA, rRNA
   - UTRs, exons, introns
   - Alternative splice isoforms (optional)

2. Functional annotation

Assign biological functions to predicted genes



Stein, 2001

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

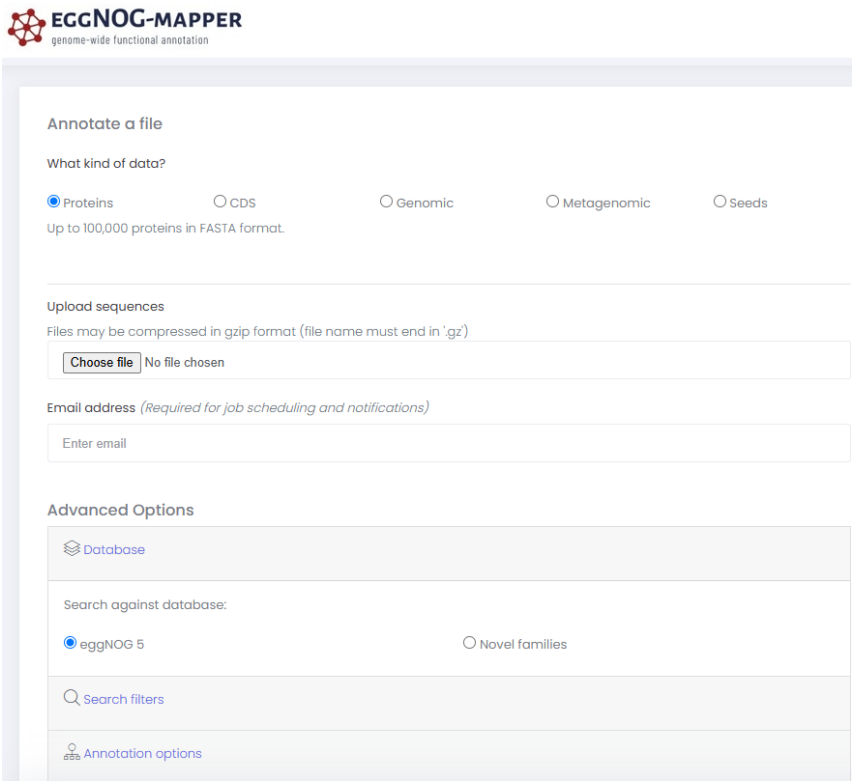
Functional annotation with eggNOG-mapper

eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale

Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork ✉, Jaime Huerta-Cepas ✉

Molecular Biology and Evolution, Volume 38, Issue 12, December 2021, Pages 5825–5829, https://doi.org/10.1093/molbev/msab293
Published: 01 October 2021

Use precomputed orthologous groups and phylogenies from the eggnog database to transfer functional information from fine-grained orthologs only



TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Functional annotation with antiSMASH

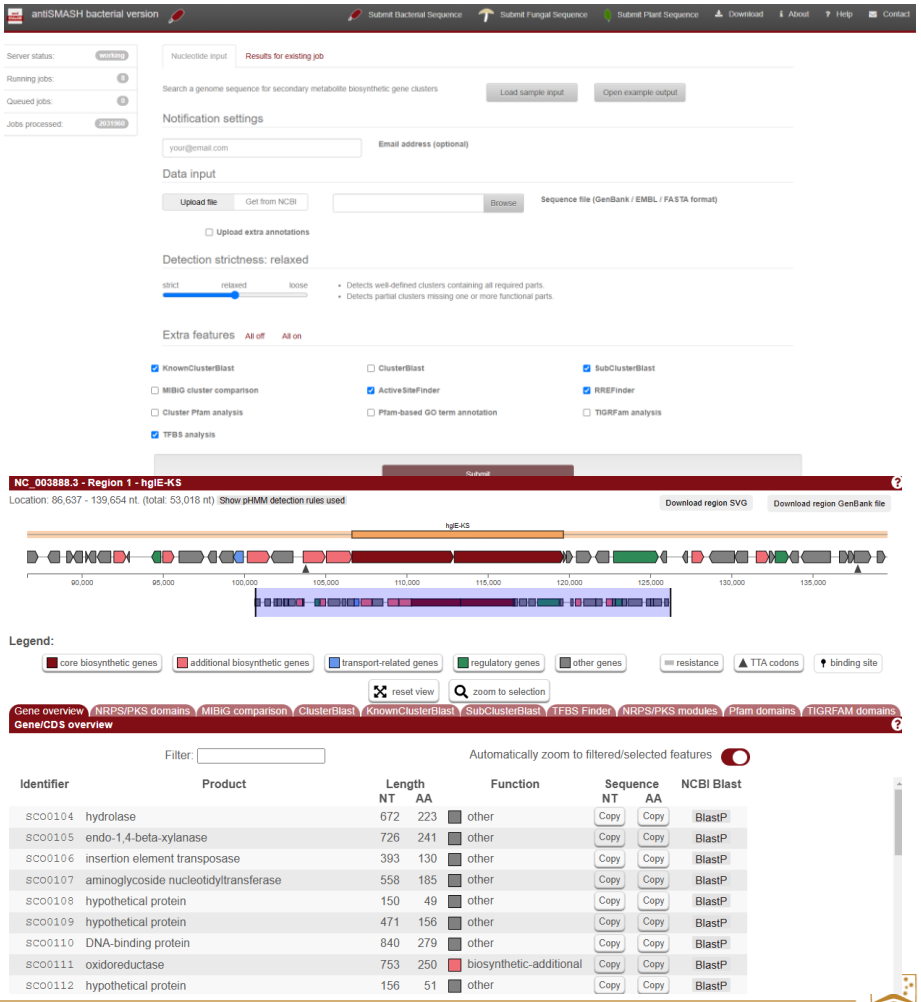
JOURNAL ARTICLE

antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation

Kai Blin ✉, Simon Shaw, Hannah E Augustijn, Zachary L Reitz, Friederike Biermann, Mohammad Alanjary, Artem Fetter, Barbara R Terlouw, William W Metcalf, Eric J N Helfrich ... Show more

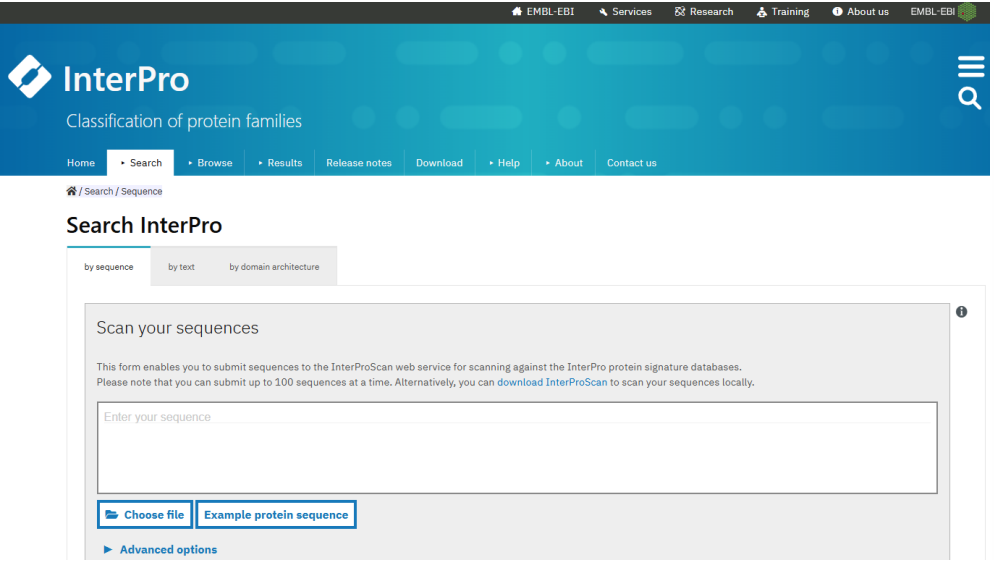
Nucleic Acids Research, Volume 51, Issue W1, 5 July 2023, Pages W46–W50,
https://doi.org/10.1093/nar/gkad344
Published: 04 May 2023 **Article history** ▾

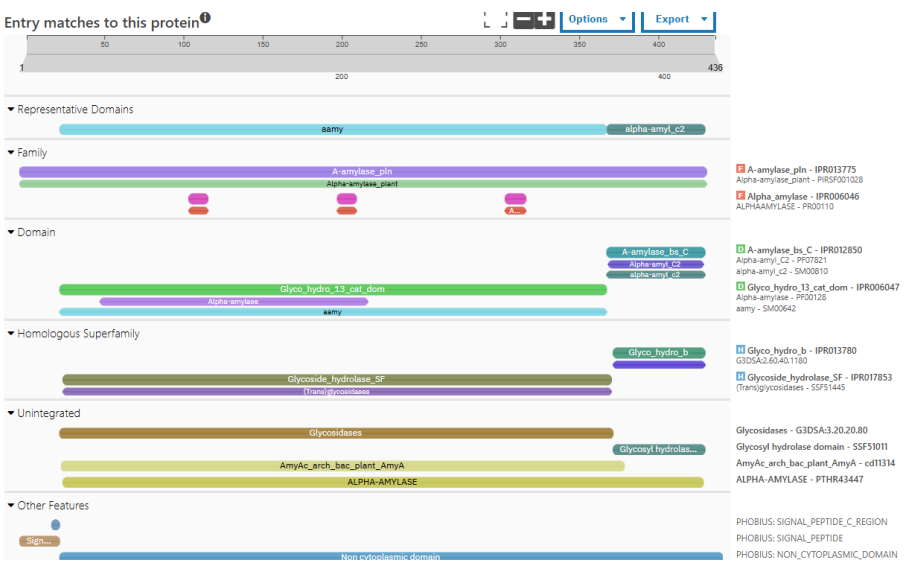
Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences



TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Functional annotation with InterProScan5





InterProScan predicts multiple aspects of protein features and functions

Stop codons are not allowed within protein sequences

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

GFF3 output of InterProScan5

Name of protein	Source	Type	Start	End	E-value	Strand		Description
Pca12NC29_STE3.2	.	polypeptide	1	385	.	+	.	ID=Pca12NC29_STE3.2;md5=3181c601b7150e9202c97020f5efcd8f
Pca12NC29_STE3.2	PANTHER	protein_match	3	335	1.3E-90	+	.	date=30-10-2024;Target=Pca12NC29_STE3.2 3 335;Ontology_term="GO:0004932","GO:0007186","GO:0016020";ID=match$1_3_335;Name=PTHR28097;status=T;Dbxref="InterPro:IPR0014
Pca12NC29_STE3.2	CDD	protein_match	7	228	4.09343E-59	+	.	date=30-10-2024;Target=Pca12NC29_STE3.2 7 228;ID=match$2_7_228;signature_desc=7tmD_STE3;Name=cd14966;status=T
Pca12NC29_STE3.2	PRINTS	protein_match	268	282	3.5E-27	+	.	date=30-10-2024;Target=Pca12NC29_STE3.2 268 282;Ontology_term="GO:0004932","GO:0007186","GO:0016020";ID=match$3_268_282;signature_desc=Fungal pheromone STE3 GPCR signature;Name=PR00899;status=T;Dbxref="InterPro:IPR001499"
Pca12NC29_STE3.2	PRINTS	protein_match	88	101	3.5E-27	+	.	date=30-10-2024;Target=Pca12NC29_STE3.2 88 101;Ontology_term="GO:0004932","GO:0007186","GO:0016020";ID=match$3_88_101;signature_desc=Fungal pheromone STE3 GPCR signature;Name=PR00899;status=T;Dbxref="InterPro:IPR001499"

Protein ID

IPR001499	GPCR fungal pheromone mating factor, STE3	INTERPRO	G protein-coupled receptors (GPCRs) constitute a vast protein family that encompasses a wide range of functions, including various autocrine, paracrine and endocrine processes. They show considerable ...

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

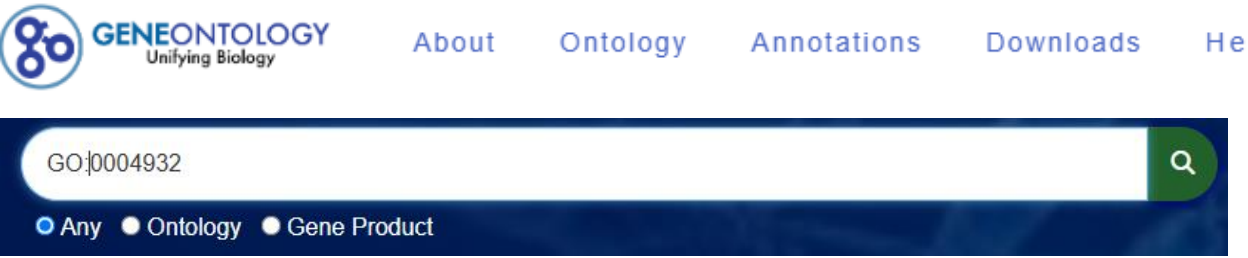
TSV output of InterProScan5

Name of protein	md5	Length of protein	Source	ID	Hits name	Start	End	E-value		Date	Interpro ID	Interpro description	Gene Ontology
Pca12NC29_STE3.2	3181c601b7150e9202c97020f5efcd8f	385	PANTHER	PTHR28097	PHEROMONE A FACTOR RECEPTOR	3	335	1.3E-90	T	30-10-2024	IPR001499	GPCR fungal pheromone mating factor, STE3	GO:0004932|GO:0007186|GO:0016(
Pca12NC29_STE3.2	3181c601b7150e9202c97020f5efcd8f	385	CDD	cd14966	7tmD_STE3	7	228	4.09343E-59	T	30-10-2024	-	-	
Pca12NC29_STE3.2	3181c601b7150e9202c97020f5efcd8f	385	PRINTS	PR00899	Fungal pheromone STE3 GPCR signature	268	282	3.5E-27	T	30-10-2024	IPR001499	GPCR fungal pheromone mating factor, STE3	GO:0004932|GO:0007186|GO:0016(
Pca12NC29_STE3.2	3181c601b7150e9202c97020f5efcd8f	385	PRINTS	PR00899	Fungal pheromone STE3 GPCR signature	88	101	3.5E-27	T	30-10-2024	IPR001499	GPCR fungal pheromone mating factor, STE3	GO:0004932|GO:0007186|GO:0016(

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

TSV output of InterProScan5

Name of protein	md5	Length of protein	Source	ID	Hits name	Start	End	E-value		Date	Interpro ID	Interpro description	Gene Ontology
Pca12NC29_STE3.2	3181c601b7150e9202c97020f5efcd8f	385	PANTHER	PTHR28097	PHEROMONE A FACTOR RECEPTOR	3	335	1.3E-90	T	30-10-2024	IPR001499	GPCR fungal pheromone mating factor, STE3	GO:0004932|GO:0007186|GO:0016(



Term	Definition	Ontology source	space	Synonyms	Alt ID
mating-type factor pheromone receptor activity	Combining with a mating-type factor pheromone to initiate a change in cell activity.	molecular_function	GO		

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Functional annotation with BUSCO

JOURNAL ARTICLE

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs FREE

Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov ✉ Author Notes

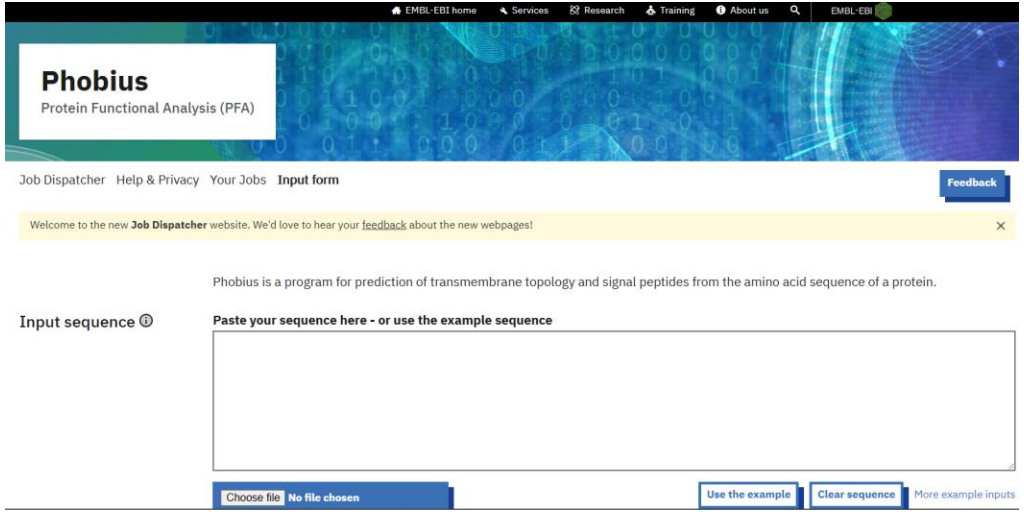
Bioinformatics, Volume 31, Issue 19, October 2015, Pages 3210–3212,
https://doi.org/10.1093/bioinformatics/btv351
Published: 09 June 2015 **Article history** ▾

BUSCO measures completeness of genome assembly, gene set and transcriptome completeness

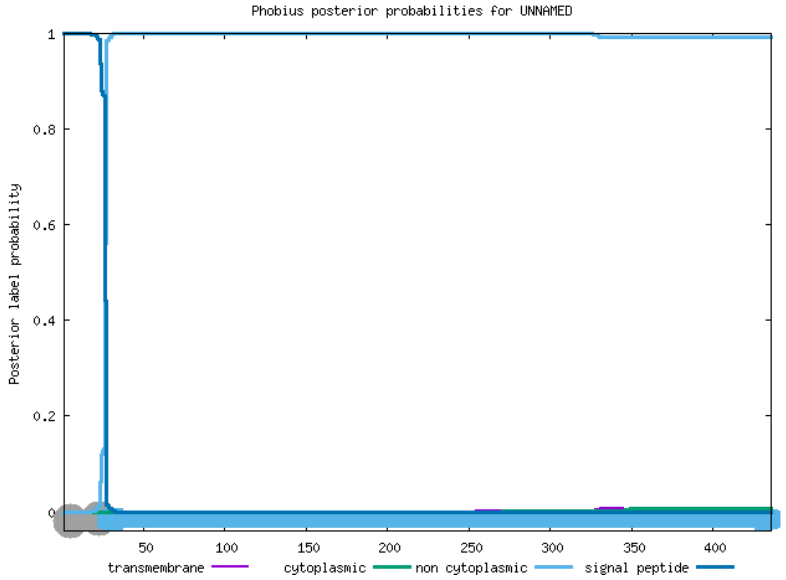
BUSCO also outputs genes identified as conserved single-copy orthologs

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Functional annotation with Phobius




```
ID   UNNAMED
FT   SIGNAL       1     25
FT   REGION       1      6       N-REGION.
FT   REGION       7     20       H-REGION.
FT   REGION      21     25       C-REGION.
FT   TOPO_DOM     26    436       NON CYTOPLASMIC.
//
```



Phobius predicts transmembrane topology and signal peptides from the amino acid sequence of proteins

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Functional annotation with SignalP and EffectorP

Brief Communication | Open access | Published: 03 January 2022

SignalP 6.0 predicts all five types of signal peptides using protein language models

Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne & Henrik Nielsen

Nature Biotechnology **40**, 1023–1025 (2022) | Cite this article

Submit data

Sequence submission: paste the sequence(s) *and/or* upload a local file

Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 1000.
The long output format might timeout for more than 100 entries.

Mirror Use SignalP 6.0 on BioLib if this server is heavily loaded.

Enter protein sequence(s) in fasta format...

For example proteins Click here
Format directly from your local disk: Choose file No file chosen

Organism
- Eukarya
- Other

"Eukarya" only predicts Sec/SPI SPs.

Output format:
- Long output
- Short output (no figures)

Model mode:
- Fast
- Slow

The slow mode takes 6x longer to compute. Use when accurate region borders are needed.

Submit Clear fields

SignalP uses fasta sequence file as input and predicts signal peptides

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Functional annotation with SignalP and EffectorP

› Mol Plant Microbe Interact. 2022 Feb;35(2):146-156. doi: 10.1094/MPMI-08-21-0201-R. Epub 2022 Feb 1.

EffectorP 3.0: Prediction of Apoplastic and Cytoplasmic Effectors in Fungi and Oomycetes

Jana Sperschneider [1], Peter N Dodds [2]

Affiliations + expand

PMID: 34698534 DOI: 10.1094/MPMI-08-21-0201-R

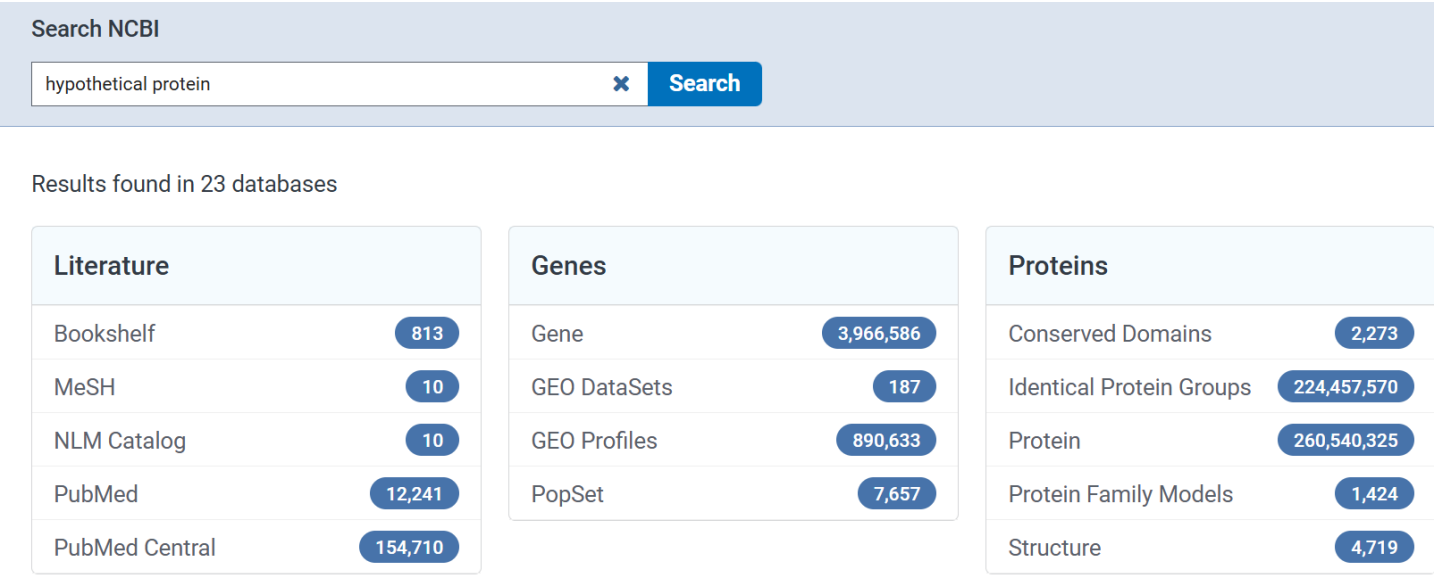
Table 3. Test sets for assessing the false-positive prediction rates for cytoplasmic and apoplastic effector prediction by EffectorP 3.0[a]

		EffectorP 3.0		EffectorP-fungi 3.0	
Test set	**Proteins (*n*)**	**Cytoplasmic**	**Apoplastic**	**Cytoplasmic**	**Apoplastic**
Sets depleted in cytoplasmic proteins					
Plant proteins with annotated apoplastic localization	362	37 (10.2%)	75 (20.7%)	39 (10.8%)	74 (20.4%)
Apoplastic proteome of *Magnaporthe oryzae* (Kim et al. 2013)	155	12 (7.7%)	32 (20.6%)	5 (3.2%)	32 (20.6%)
Apoplastic proteome of *Oryza sativa* (Kim et al. 2013)	94	1 (1.1%)	23 (24.5%)	0 (0%)	23 (24.5%)
Leaf apoplast proteome of *Brassica napus* var. *napus* infected with *Verticillium longisporum* (Floerl et al. 2008)	8	0 (0%)	5 (62.5%)	0 (0%)	5 (62.5%)
Leaf apoplast proteome of *Arabidopsis thaliana* infected with *V. longisporum* (Floerl et al. 2012)	27	1 (3.7%)	7 (25.9%)	0 (0%)	8 (29.6%)
Apoplastic proteome of *Nicotiana benthamiana* leaves (Goulet et al. 2010)	16	2 (12.5%)	10 (62.5%)	0 (0%)	11 (68.8%)
Fungal CAZYs (UniProt, reviewed entries)	1,164	115 (9.9%)	232 (19.9%)	68 (5.8%)	231 (19.8%)
Fungal saprophyte secreted proteins	24,432	2,018 (8.3%)	3,635 (14.9%)	1,625 (6.7%)	3,597 (14.7%)
False-positive rate for cytoplasmic effector prediction	26,258	2,186 (8.3%)	–	1,737 (6.6%)	–
Sets depleted in apoplastic proteins					
Plant proteins with annotated cytoplasmic localization	3,843	1,970 (51.3%)	84 (2.2%)	1,355 (37.5%)	86 (2.2%)
Fungal endoplasmic reticulum-localized proteins	71	23 (38%)	4 (5.6%)	16 (22.5%)	4 (5.6%)
Fungal Golgi-localized proteins	19	4 (21.1%)	1 (5.3%)	3 (15.8%)	1 (5.3%)
Fungal vacuole proteins	15	4 (26.7%)	0 (0%)	4 (26.7%)	0 (0%)
Human proteins	20,238	7,314 (36.1%)	513 (2.5%)	5,790 (28.6%)	480 (2.4%)
Bacterial type-III effectors (T3Enc) (Hui et al. 2020)	519	263 (50.7%)	17 (3.3%)	210 (40.5%)	17 (3.3%)
RxLR effector candidates (Win et al. 2007)	358	331 (92.5%)	0 (0%)	249 (69.6%)	0 (0%)
False-positive rate for apoplastic effector prediction	24,705	–	618 (2.5%)	–	587 (2.4%)

EffectorP uses predicted signal peptides as input and predicts apoplastic and cytoplasmic effectors in fungi and oomycetes

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Protein annotated as hypothetical protein without known function



Hypothetical proteins are those that are predicted to be expressed in an organism, but no evidence of them exist in gene banks

Proteins with unknown functions still may have crucial roles

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

THANK YOU

Australian
National
University