

# **Intra-specific and interspecific protein content analysis**

Johann Joets

Institut Diversité, Evolution, Ecologie du Vivant

INRAE – Université Paris-Saclay

*École "Assemblage & Annotation" AVIESAN 2022 - Roscoff*



## Intra-specific and interspecific protein content analysis

Goals :

- Reduce redundancy (alternative CDS) => **compute representative sequences**
  - Build set of repeated sequences or highly conserved families(genes/proteins)
  - Compare gene/protein content between
    - assemblies versions
    - individuals of a single species
    - different species
- | => **Gene loss/gain (toward pangenomics)**



# Intra-specific and interspecific protein content analysis

## Goals :

- Reduce redundancy (alternative CDS) => **compute representative sequences**
  - Build set of repeated sequences or highly conserved families(genes/proteins)
  - Compare gene/protein content between
    - assemblies versions
    - individuals of a single species
    - different species
- | => **Gene loss/gain (toward pangenomics)**

## Methods / tools :

- Sequence similarity-based clustering : **CD-HIT** => intra-specific comparaison (*low seq divergence*)
- Phylogeny-based orthology classification : **OrthoFinder** => inter-specific comparaison (*higher seq divergence*)



# OrthoFinder. Our dataset:

fungi.prot.fasta

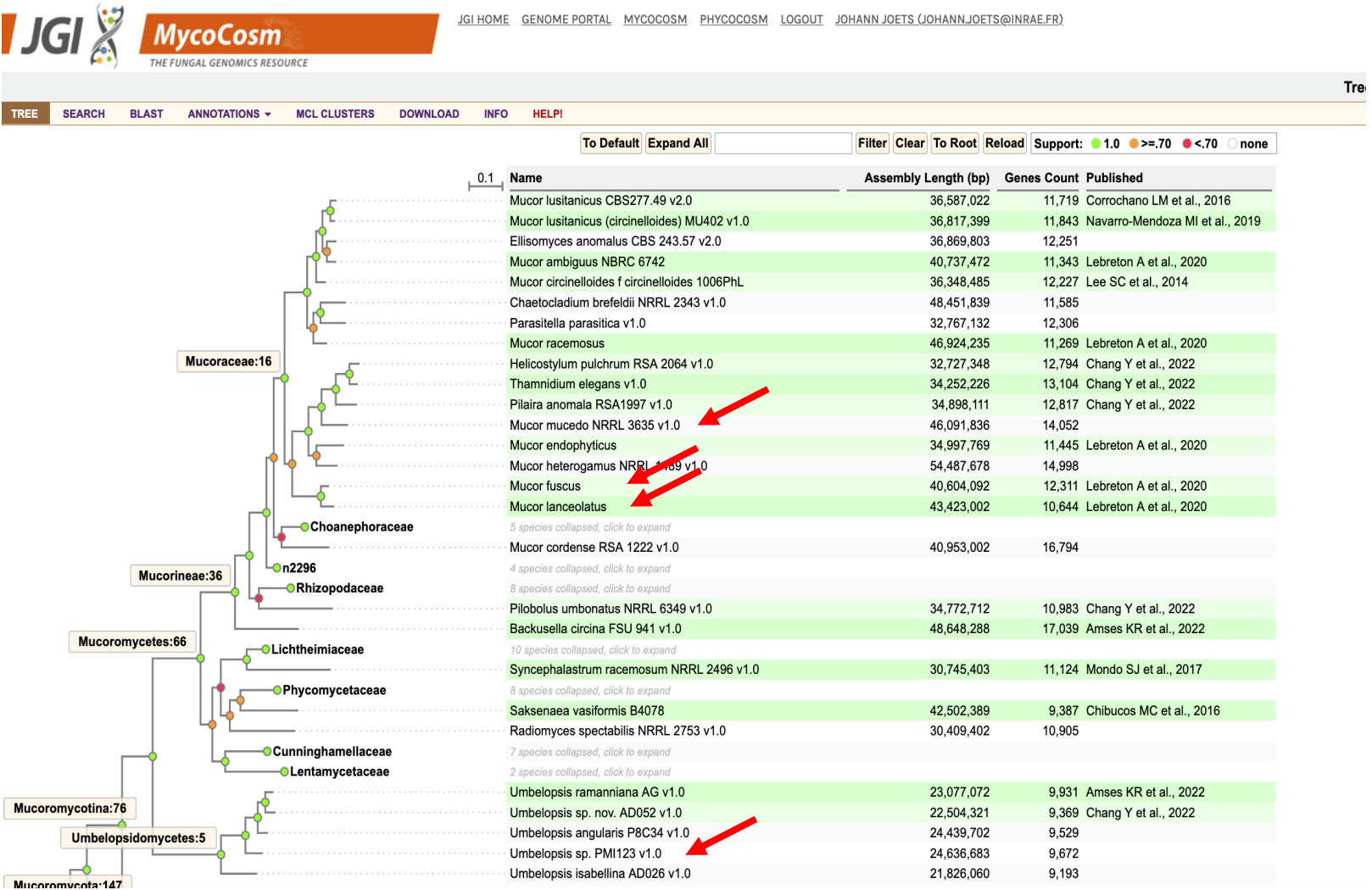
a list with 4 items

Mmucedo.aa.fasta

Umbsp.aa.fasta.gz

Mucfus1.aa.fasta.gz

Muclan1.aa.fasta.gz





# OrthoFinder. Our dataset:

Phylogenetic Status of Two Undescribed Zygomycete Species from Korea: *Actinomucor elegans* and *Mucor minutus*

Thuong T. T. Nguyen<sup>1</sup>, Hee-Young Jung<sup>2</sup>, Youn Su Lee<sup>3</sup>, Kerstin Voigt<sup>4</sup> and Hyang Burm Lee<sup>1\*</sup>

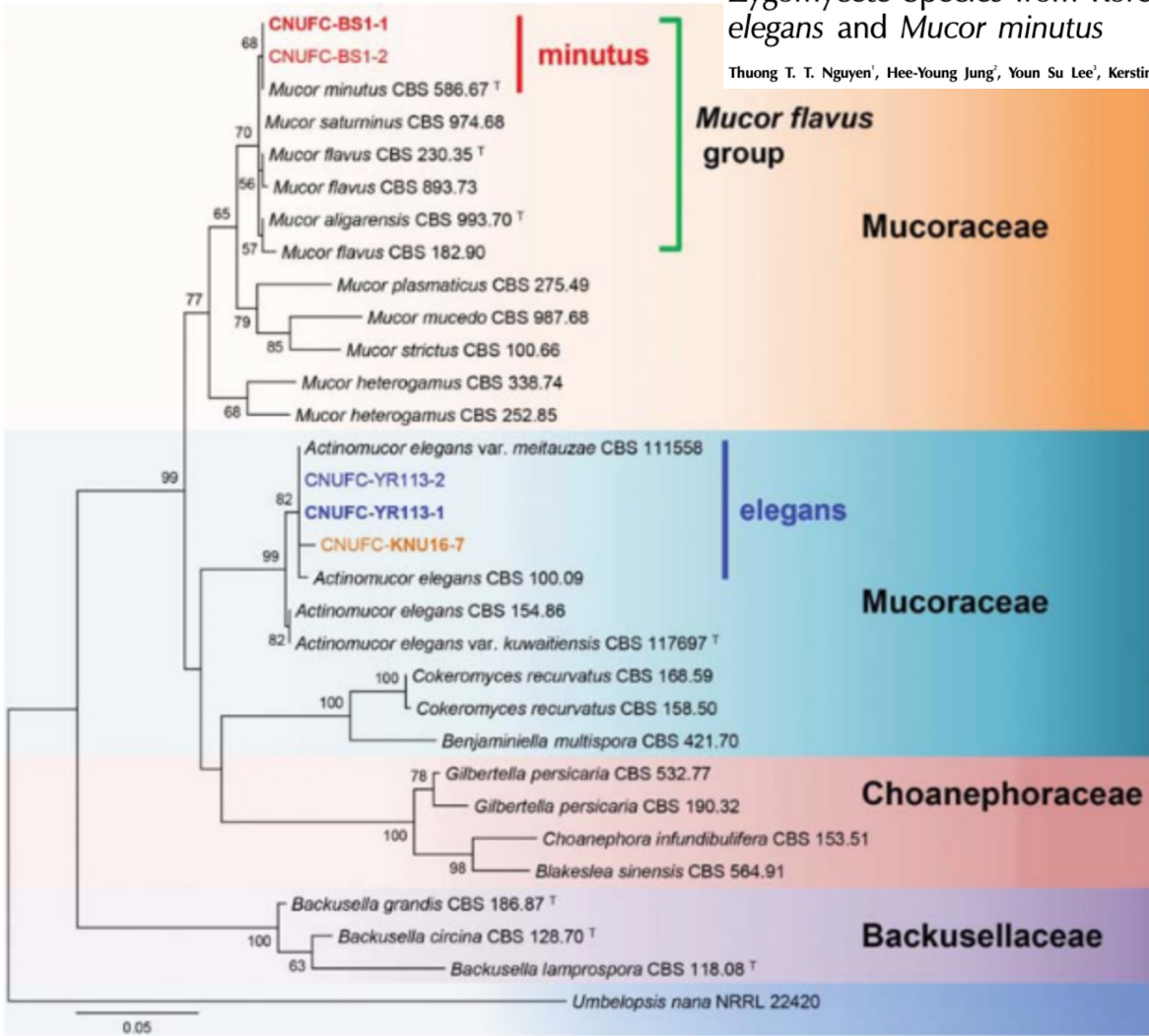
**fungi.prot.fasta**  
a list with 4 items

**Mmucedo.aa.fasta**

**Umbsp.aa.fasta.gz**

**Mucfus1.aa.fasta.gz**

**Muclan1.aa.fasta.gz**



**Fig. 2.** Phylogenetic tree based on maximum likelihood analysis of 28S rDNA sequences for *Actinomucor elegans* CNUFC-YR113-1, *A. elegans* CNUFC-YR113-2, *A. elegans* CNUFC-KNU16-7, *Mucor minutus* CNUFC-BS1-1, and *M. minutus* CNUFC-BS1-2. *Umbelopsis nana* was used as an outgroup. Bootstrap support values of  $\geq 50\%$  are indicated at the nodes. The bar indicates the number of substitutions per position.



# Datasets library

Galaxy France

Workflow

Visualize

Données partagées

Aide

Utilisateur

Using 12%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Search

+ Folder

+ Datasets

Export to History

Download

Libraries / EBAll A&A 2022

Bibliothèque de données

Histories

Workflows

Visualisations

Pages

	<div></div> Name	Description	Type	Size	Updated	State
<div></div>	<div></div> Assembly	Data for assembly	folder		2 days ago	
<div></div>	<div></div> Functional annotation	Protein sequence datasets	folder		2 hours ago	
<div></div>	<div></div> Polishing	Data for assembly polishing	folder		a day ago	
<div></div>	<div></div> Prokaryotic annotation	Data for prokaryotic annotation track	folder		7 days ago	

<<

<

1

>

>>

10

per page, 4 total














# Datasets library

[+ Folder](#)[+ Datasets ▾](#)[Add to History ▾](#)[Download](#)[Details](#)

[Libraries](#) / [EBI A&A 2022](#) / [Functional annotation](#)

	<input type="checkbox"/>	Name	◆	Description
	<input type="checkbox"/>	EA1197.default.HQ.prot.fa		uploaded fasta file
	<input type="checkbox"/>	EA1197.default.HQ.prot.fa		uploaded fasta file
	<input type="checkbox"/>	F331.default.HQ.prot.fa		uploaded fasta file
	<input type="checkbox"/>	F331.default.HQ.prot.fa		uploaded fasta file
	<input type="checkbox"/>	GASPE.default.HQ.prot.fa		uploaded fasta file
	<input type="checkbox"/>	GASPE.default.HQ.prot.fa		uploaded fasta file
	<input checked="" type="checkbox"/>	Mmucedo.aa.fasta		uploaded fasta file
	<input type="checkbox"/>	Mmucedo.prot.fasta		uploaded fasta file
	<input checked="" type="checkbox"/>	Mucfus1.aa.fasta.gz		uploaded fasta.gz file
	<input type="checkbox"/>	Mucfus1.aa.fasta.gz		uploaded fasta.gz file
	<input checked="" type="checkbox"/>	Muclan1.aa.fasta.gz		uploaded fasta.gz file
	<input type="checkbox"/>	Muclan1.aa.fasta.gz		uploaded fasta.gz file
	<input checked="" type="checkbox"/>	Umbasp.aa.fasta.gz		uploaded fasta.gz file
















<input type="checkbox"/>	Name	Description	Type	Size	Updated	State
	<input type="checkbox"/> EA1197.default.HQ.prot.fa	uploaded fasta file	fasta	20.5 MB	over 1 year ago	
	<input type="checkbox"/> EA1197.default.HQ.prot.fa	uploaded fasta file	fasta	20.5 MB	over 1 year ago	
	<input type="checkbox"/> F331.default.HQ.prot.fa	uploaded fasta file	fasta	20.8 MB	over 1 year ago	
	<input type="checkbox"/> F331.default.HQ.prot.fa	uploaded fasta file	fasta	20.8 MB	over 1 year ago	
	<input type="checkbox"/> GASPE.default.HQ.prot.fa	uploaded fasta file	fasta	20.9 MB	over 1 year ago	
	<input type="checkbox"/> GASPE.default.HQ.prot.fa	uploaded fasta file	fasta	20.9 MB	over 1 year ago	
	<input checked="" type="checkbox"/> Mmucedo.aa.fasta	uploaded fasta file	fasta	6 MB	over 1 year ago	
	<input type="checkbox"/> Mmucedo.prot.fasta	uploaded fasta file	fasta	6 MB	over 1 year ago	
	<input checked="" type="checkbox"/> Mucfus1.aa.fasta.gz	uploaded fasta.gz file	fasta.gz	3.4 MB	over 1 year ago	
	<input checked="" type="checkbox"/> Mucfus1.aa.fasta.gz	uploaded fasta.gz file	fasta.gz	3.4 MB	over 1 year ago	
	<input checked="" type="checkbox"/> Muclan1.aa.fasta.gz	uploaded fasta.gz file	fasta.gz	3 MB	over 1 year ago	

javascript:void(0)



as Datasets

as a Collection

<input type="checkbox"/>	Name	Description
	<input type="checkbox"/> EA1197.default.HQ.prot.fa	uploaded fasta file
	<input type="checkbox"/> EA1197.default.HQ.prot.fa	uploaded fasta file
	<input type="checkbox"/> F331.default.HQ.prot.fa	uploaded fasta file
	<input type="checkbox"/> F331.default.HQ.prot.fa	uploaded fasta file
	<input type="checkbox"/> GASPE.default.HQ.prot.fa	uploaded fasta file
	<input type="checkbox"/> GASPE.default.HQ.prot.fa	uploaded fasta file
	<input checked="" type="checkbox"/> Mmucedo.aa.fasta	uploaded fasta file
	<input type="checkbox"/> Mmucedo.prot.fasta	uploaded fasta file
	<input checked="" type="checkbox"/> Mucfus1.aa.fasta.gz	uploaded fasta.gz file
	<input type="checkbox"/> Mucfus1.aa.fasta.gz	uploaded fasta.gz file
	<input checked="" type="checkbox"/> Muclan1.aa.fasta.gz	uploaded fasta.gz file
	<input type="checkbox"/> Muclan1.aa.fasta.gz	uploaded fasta.gz file
	<input checked="" type="checkbox"/> Umbbsp.aa.fasta.gz	uploaded fasta.gz file



search

+ Folder

+ Datasets

Add to History

Download

Details

Create a collection from a list of datasets

Collections of datasets are permanent, ordered lists of datasets that can be passed to tools and workflows in order to have analyses done on each member of the entire group. This interface allows you...

↺

⌵

Mmucedo.aa.fasta

Discard

Mucfus1.aa.fasta.gz

Discard

Mucian1.aa.fasta.gz

Discard

Umbbsp.aa.fasta.gz

Discard

Hide original elements?

☒

Name:

fungi.prot.fasta

Cancel

Create collection

Mucfus1.aa.fasta.gz

uploaded fasta.gz file

fasta.gz

3.4 MB

over 1 year ago



Tools

orthof

Upload Data

Show Sections

OrthoFinder finds orthogroups in a set of proteomes

WORKFLOWS

All workflows

OrthoFinder

Tool Parameters

Orthofinder settings

From fasta file

OrthoFinder can also be used to find orthogroups from a previous analysis

Please provide the input files

Select input files

One fasta file

Input contains nucleotide or amino acid sequences?

Amino acid

Type to Search

Label	Details	Time
9: Umbbsp.aa.fasta.gz	fasta.gz	2024-06-04 16:39
8: Muclan1.aa.fasta.gz	fasta.gz	2024-06-04 16:39
7: Mucfus1.aa.fasta.gz	fasta.gz	2024-06-04 16:39
6: Mmucedo.aa.fasta	fasta	2024-06-04 16:39

Upload

Cancel Ok

Run Tool

History: Orthofinder

fungi.prot.fasta

a list with 4 datasets

Download

1: Mmucedo.aa.fasta

2: Mucfus1.aa.fasta.gz

3: Muclan1.aa.fasta.gz

4: Umbbsp.aa.fasta.gz



# OrthoFinder: Galaxy

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

Using 12%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

orthofinder

Upload Data

Show Sections

OrthoFinder finds orthogroups in a set of proteomes

WORKFLOWS

All workflows

Orthofinder starting point

From fasta files

OrthoFinder can be run in 2 steps. Choose 'From fasta files' to run OrthoFinder from scratch or 'From blast results' if you have all the blast results from a previous OrthoFinder run.

Select input fasta files

10: fungi.prot.fasta (with implicit datatype conversion)

8: maize.prot.fasta

One fasta file per species; species and sequence names in the results will remain the same than in the input files.

Input contains nucleotide or amino acid sequences?

Amino acid

Sequence search program

Diamond (faster)

Orthofinder run mode

Full run (including gene trees)

Method for gene tree inference

Dendroblast (faster)

Inflation parameter

1,5

Modify inflation parameter for MCL. Not recommended. (-l)

Generate output about gene duplication events

Yes

Email notification

No

Send an email notification when the job completes.

Execute

History

Back to data

fungi.prot.fasta

a list with 4 items

Mmucedo.aa.fasta


Umbasp.aa.fasta.gz









Mucfus1.aa.fasta.gz

Muclan1.aa.fasta.gz






# OrthoFinder: Galaxy


 **Galaxy France**


 Workflow  Visualize  Données partagées ▾  Aide ▾  Utilisateur ▾   

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

**Tools**  

orthofinder 


 Upload Data

 Show Sections

**OrthoFinder** finds orthogroups in a set of proteomes

**WORKFLOWS**

All workflows

 Executed **OrthoFinder** and successfully added 1 job to the queue.

The tool uses this input:

- **10: fungi.prot.fasta (with implicit datatype conversion)**

It produces 12 outputs:

- **20: OrthoFinder on data 5, data 6, and others: orthogroups (txt)**
- **21: OrthoFinder on data 5, data 6, and others: orthogroups (tsv)**
- **22: OrthoFinder on data 5, data 6, and others: species overlaps**
- **23: OrthoFinder on data 5, data 6, and others: unassigned genes**
- **24: OrthoFinder on data 5, data 6, and others: overall comparative genomics statistics**
- **25: OrthoFinder on data 5, data 6, and others: per species comparative genomics statistics**
- **26: OrthoFinder on data 5, data 6, and others: species tree**
- **27: OrthoFinder on data 5, data 6, and others: species tree with node labels**
- **28: OrthoFinder on data 5, data 6, and others: species tree with duplication events**
- **29: OrthoFinder on data 5, data 6, and others: duplication events**
- **30: OrthoFinder on data 5, data 6, and others: duplications per orthogroup**
- **31: OrthoFinder on data 5, data 6, and others: duplications per species tree node**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.







# CD-HIT a sequence clustering program

Was originally developed to remove redundancy and select representative sequences at very high speed

CD-HIT, use incremental clustering algorithm method:

1/ Sequences are sorted in order of decreasing length.

2/ The longest one becomes the representative of the first cluster.


3/ Each remaining sequence is compared to the representatives of existing clusters.

If the similarity with any representative is above a given threshold, it is grouped into that cluster.

Otherwise, a new cluster is defined with that sequence as the representative.





Try to select a representative sequence per gene.

 Galaxy France
Workflow Visualize Données partagées Aide

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

### Tools

 Upload Data

 Show Sections

**cd-hit** Cluster or compare biological sequence datasets

---

**WORKFLOWS**

All workflows

```
>Zm00064aa000022_T001
MAMAEKVKEKMLMLRSSDNQEFVKE SVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHAE
PTGPGDAAGTTNRSAEDELNIFDADFVNVEHSTLLDLILAANYLDIKGLLNLARQTITDLINGKMPEEVC
KTNIKNDLTIPSTSALATTMPSSERQM EARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPPL EIVK
YPALLERAWGWDSILPYCPVNSWPLYKTYLQEYYERN SREVL AHQVEANAASGP NLNGADENDLATLFNL
CISREVQL L NLL KRRDKKHNVDEISLNEKL TNCARQIT FVEYAGFPFPSVAL KCIMAESDLL CMLLMCGT
GMISQVNICSRIRKVAFRFM TYKGGPCFAAAATMMATTKEAKLSG LLRNRCRENNGPFSRSVFIRKWTI
AAMFRICEECSPVEESTGGYTATKLILDGSDDKNPCDEL VNKDNL LQRNKINKKYQGLFKGKTTKCRHP
VQKQEPGDGATPASPS PANPISTVVQKLW
>Zm00064aa000022_T002
MAMAEKVKEKMLMLRSSDNQEFVKE SVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHAE
PTGPGDAAGTTNRSAEDELNIFDADFVNVEHSTLLDLILAANYLDIKGLLNLARQTITDLINGKMPEEVC
KTNIKNDLTIPSTSALATTMPSSERQM EARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPPL EIVK
YPALLERAWGWDSILPYCPVNSWPLYKTYLQEYYERN SREVL AHQVEANAASGP NLNGADENDLATLFNL
CISREVQL L NLL KRRDKKHNVDEISLNEKL TNCARQIT FVEYAGFPFPSVAL KCIMAESDLL CMLLMCGT
GMISQVNICSRIRKVAFRFM TYKGGPCFAAAATMMATTKEAKLSG LLRNRCRENNGPFSRSVFIRKWTI
AAMFRICEECSPVEESTGGYTATKLILDGSDDKNPCDEL VNKDNL LQRNKINKKYQGRGFS
>Zm00064aa000022_T003
MAMAEKVKEKMLMLRSSDNQEFVKE SVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHAE
PTGPGDAAGTTNRSAEDELNIFDADFVNVEHSTLLDLILAANYLDIKGLLNLARQTITDLINGKMPEEVC
KTNIKNDLTIPSTSALATTMPSSERQM EARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPPL EIVK
YPESMVNVRHNHHWR
>Zm00064aa000022_T004
LFYLCCSVCCSIPSRLSFSFYFFVTILNDNYRK HGERPPQPPL EIVKY PALLERAWGWDSILPYCPVNS
WPLYKTYLQEYYERN SREVL AHQVEANAASGP NLNGADENDLATLFNL CISREVQL L NLL KRRDKKHNV D
EISLNEKL TNCARQIT FVEYAGFPFPSVAL KCIMAESDLL CMLLMCGT GMISQVNICSRIRKVAFRFM TY
KGGPCFAAAATMMATTKEAKLSG LLRNRCRENNGPFSRSVFIRKWTIAAMFRICEECSPVEESTGGYTA
TKLILDGSDDKNPCDEL VNKDNL LQRNKINKKYQGLFKGKTTKCRHPVQKQEPGDGATPASPS PANPI S
TVVQKLW
>Zm00064aa000023_T001
MAAAAATFGFLHPPIRKPAPVPLYILRLPTQPHSKTHPRSPPLL FLLL GRRRG GP IAAFNTTS SSTNAP
ASPTYDVREA EA AVADLL REGGASADDAAS IAARAPAYAAMLADGVREL DEL GL WASWSSGARARL GLSG
VMEMEMGRLGFR RKVYLMGRSKPDHGVP LLES LGMRLSSAKL IAPYVAAAGL TVLIDRVKFLK EM LFSS
SDYAILIGRNAKRMMTYLSIPADDALQSTLSFFEKNVLF FREMGVDKKT TGKIL CRSPEIFASNVDN TLK
KKIDFLT NF GVSKHHL PRIIRKYPELLLL DINCTLL PRMN YLLEMGL SKKDLC SMIFRF SP LLGYSIELV
MKPKLEFLLRTMKKPLKAVVEYPRFKKKL
```



CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit Cluster or compare biological sequence datasets (Galaxy Version 4.8.1+galaxy0)

Run Tool

Tool Parameters

Sequences to cluster/compare \*

6264: Gasperep.fasta

(-i)

Cluster / Compare (i.e. call cd-hit[-est] / cd-hit[-est]-2d)?

Cluster sequences

Sequence type?

Protein

For nucleotides the -est variant of cd-hit is called

Sequence identity threshold \*

0.9

Global sequence identity: number of identical alignment positions divided by the full length of the shorter sequence (-c)

Word size \*

5

Suggested word size: 5 for thresholds 0.7 ~ 1.0 4 for thresholds 0.6 ~ 0.7 3 for thresholds 0.5 ~ 0.6 2 for thresholds 0.4 ~ 0.5 (-n)

Tolerance for redundancy \*

2

(-t)

Advanced options

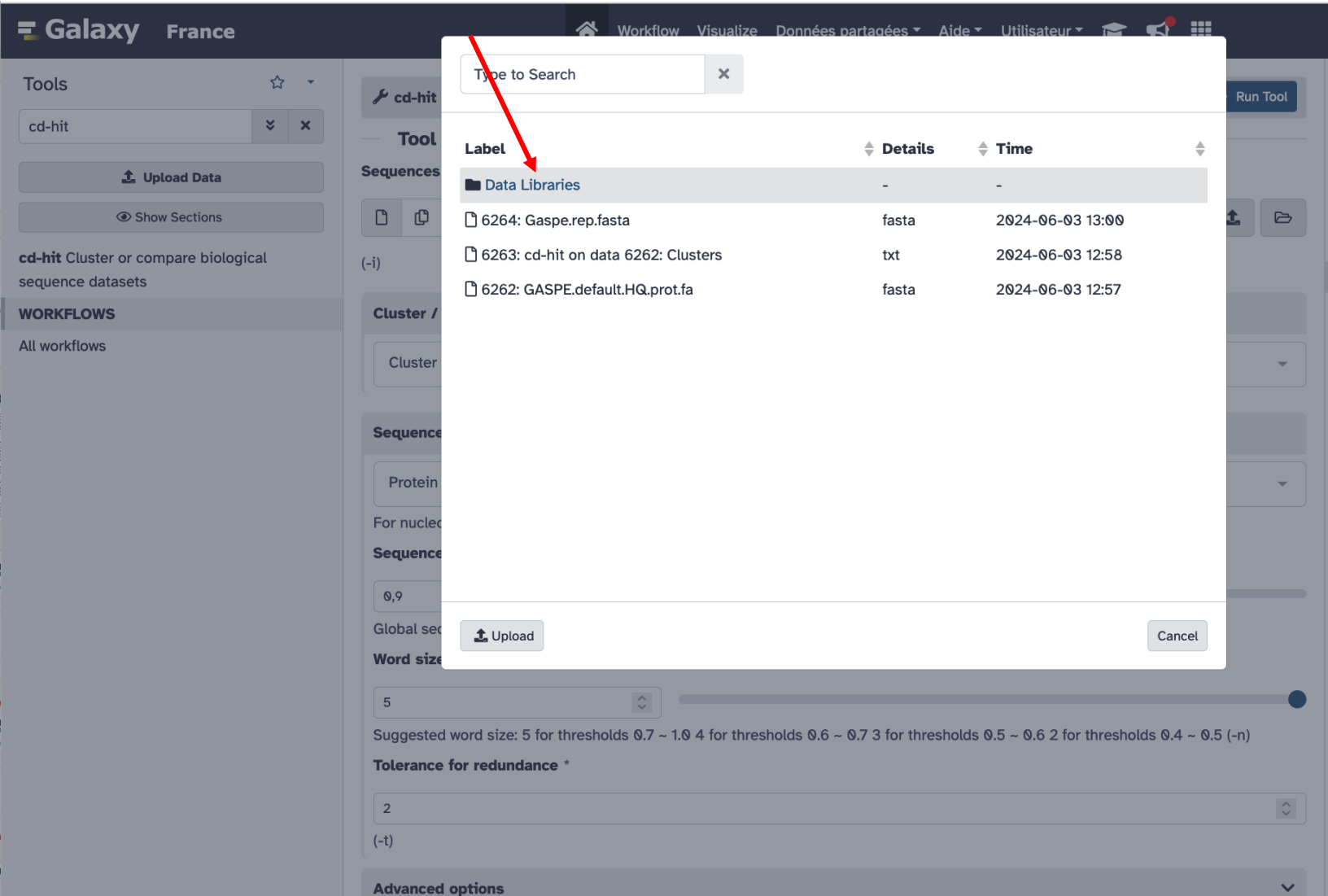


# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.





# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit

Sequences

Cluster /

Sequence

Protein

For nucle

Sequence

0,9

Global seq

Word size

5

Suggested word size: 5 for thresholds 0.7 ~ 1.0 4 for thresholds 0.6 ~ 0.7 3 for thresholds 0.5 ~ 0.6 2 for thresholds 0.4 ~ 0.5 (-n)

Tolerance for redundancy \*

2

(-t)

Type to Search

Label	Details	Time
workflow4metabolomics	Workflow4Metabolomics referenced histories	2020-06-29 09:40
Roscoff 2021	Data for Assembly and Annotation training, Roscoff 2021	2021-12-01 10:24
RepeatMasker libraries	RepeatMasker libraries	2023-09-05 09:30
ProteoRE	ProteoRE datasets	2020-05-07 13:13
GTN - Material	Galaxy Training Network Material	2020-07-10 08:23
Formation sRNA 2022	Data for Formation sRNA 2022	2022-11-22 14:24
Formation REaCTION	Formation du réseau INRAE REaCTION	2023-10-09 15:23
Formation M2 Agro Rennes	Formation M2 Agro Rennes	2023-10-10 10:07
Formation BIGOMICS	BIGOMICS, Génomique comparative, Montpellier	2023-10-11 19:59
EBAII A&A 2022	Ecole EBAII Assemblage & Annotation septembre	2022-09-19 09:09

Back Upload

Cancel



Try to select a representative sequence per gene.

Galaxy France

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit

Tool Sequences

(-i)

Cluster /

Cluster

Sequence

Protein

For nucle

Sequence

0,9

Global se

Word size

5

Suggested word size: 5 for thresholds 0.7 ~ 1.0 4 for thresholds 0.6 ~ 0.7 3 for thresholds 0.5 ~ 0.6 2 for thresholds 0.4 ~ 0.5 (-n)

Tolerance for redundancy \*

2

Workflow Visualize Données partagées Aide Utilisateur

Function

/Functional annotation/Umbasp.aa.fasta.gz

-

-

/Functional annotation/Mucan1.aa.fasta.gz

-

-

/Functional annotation/Mucan1.aa.fasta.gz

-

-

/Functional annotation/Mucfus1.aa.fasta.gz

-

-

/Functional annotation/Mucfus1.aa.fasta.gz

-

-

/Functional annotation/Mmucedo.prot.fasta

-

-

/Functional annotation/Mmucedo.aa.fasta

-

-

/Functional annotation/GASPE.default.HQ.prot.fa

-

-

/Functional annotation/GASPE.default.HQ.prot.fa

-

-

/Functional annotation/F331.default.HQ.prot.fa

-

-

/Functional annotation/F331.default.HQ.prot.fa

-

-

/Functional annotation/EA1197.default.HQ.prot.fa

-

-

/Functional annotation/EA1197.default.HQ.prot.fa

-

-

Back

Upload

Cancel



# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Galaxy

France

Home

Workflow

Visualize

Données partagées

Aide

Utilisateur

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit Cluster or compare biological sequence datasets (Galaxy Version 4.8.1+galaxy0)

Run Tool

Tool Parameters

Sequences to cluster/compare \*

Selected: /Functional annotation/F331.default.HQ.prot.fa

(-i)

Cluster / Compare (i.e. call cd-hit[-est] / cd-hit[-est]-2d)?

Cluster sequences

Sequence type?

Protein

For nucleotides the -est variant of cd-hit is called

Sequence identity threshold \*

0,9

Global sequence identity: number of identical alignment positions divided by the full length of the shorter sequence (-c)

Word size \*

5

Suggested word size: 5 for thresholds 0.7 ~ 1.0 4 for thresholds 0.6 ~ 0.7 3 for thresholds 0.5 ~ 0.6 2 for thresholds 0.4 ~ 0.5 (-n)

Tolerance for redundance \*

2

(-t)

Advanced options



# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Galaxy

France

Workflow

Visualize

Données partagées

Aide

Utilisateur

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit Cluster or compare biological sequence datasets (Galaxy Version 4.8.1+galaxy0)

Run Tool

Tool Parameters

Sequences to cluster/compare \*

Selected: /Functional annotation/F331.default.HQ.prot.fa

(-i)

Cluster / Compare (i.e. call cd-hit[-est] / cd-hit[-est]-2d)?

Cluster sequences

Sequence type?

Protein

For nucleotides the -est variant of cd-hit is called

Sequence identity threshold \*

0.99

Global sequence identity: number of identical alignment positions divided by the full length of the shorter sequence (-c)

Word size \*

5

Suggested word size: 5 for thresholds 0.7 ~ 1.0 4 for thresholds 0.6 ~ 0.7 3 for thresholds 0.5 ~ 0.6 2 for thresholds 0.4 ~ 0.5 (-n)

Tolerance for redundancy \*

2

(-t)

Advanced options

0.99



# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

Using 12%

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

Modifier les attributs du jeu de données

Attributes Datatypes Permissions

name

F331.rep.fasta

Info

Annotation - optional

Database/Build - optional

unspecified (?)

Save Auto-detect

History

Rechercher des données

data

113 MB 6 6264

6267: cd-hit on data 6265: Representative sequences

48,521 sequences

format fasta, génome de référence ?

Program: CD-HIT, V4.8.1 (+OpenMP), Aug 07 2022, 06:48:22

>Zm00062aa000001\_T001

SHANMPWIMHAGTEQRHAACGASLLWSSLQPSTVVMAAAAATFGF

THPRSPPLLFLLGRRRGPIAAFPNTTSSSTNAPASPTYDVREA

PAYAAMLADGVREDELGLWASWSSGARARLGLSGVEMEMGRLG

MRLSSAKLIAPYVAAAGLTVLIDRVKFLKEMLFSSSDYAILIGN

6266: cd-hit on data 6265: Clusters

6265: F331.default.HQ.prot.fasta

6264: Gaspe.rep.fasta

6263: cd-hit on data 6262: Clusters



# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

WorkflowVisualizeDonnées partagéesAideUtilisateur

Using 12%

<Cluster 20  
0 2474aa, >Zm00062aa005629\_T00... at 100.00%  
1 2651aa, >Zm00062aa005629\_T00... \*  
2 2477aa, >Zm00062aa005629\_T00... at 100.00%  
>Cluster 21  
0 2626aa, >Zm00062aa028241\_T00... \*  
>Cluster 22  
0 2621aa, >Zm00062aa023112\_T00... \*  
>Cluster 23  
0 2615aa, >Zm00062aa000100\_T00... \*  
>Cluster 24  
0 2609aa, >Zm00062aa007116\_T00... \*  
>Cluster 25  
0 2579aa, >Zm00062aa004846\_T00... \*  
>Cluster 26  
0 2574aa, >Zm00062aa002475\_T00... \*  
>Cluster 27  
0 2565aa, >Zm00062aa039121\_T00... \*  
>Cluster 28  
0 2563aa, >Zm00062aa029684\_T00... \*  
>Cluster 29  
0 110aa, >Zm00062aa026574\_T00... at 99.09%  
1 2552aa, >Zm00062aa033122\_T00... \*  
>Cluster 30  
0 2538aa, >Zm00062aa015588\_T00... \*  
>Cluster 31  
0 2478aa, >Zm00062aa013952\_T00... \*  
>Cluster 32  
0 2449aa, >Zm00062aa035739\_T00... \*  
>Cluster 33  
0 2447aa, >Zm00062aa035984\_T00... \*  
1 1604aa, >Zm00062aa035984\_T00... at 99.31%  
>Cluster 34  
0 2441aa, >Zm00062aa042116\_T00... \*  
>Cluster 35  
0 2439aa, >Zm00062aa028764\_T00... \*  
>Cluster 36  
0 2410aa, >Zm00062aa008109\_T00... at 100.00%  
1 2410aa, >Zm00062aa008109\_T00... at 100.00%  
2 2410aa, >Zm00062aa008109\_T00... at 100.00%  
3 2410aa, >Zm00062aa008109\_T00... at 100.00%  
4 2432aa, >Zm00062aa008109\_T00... \*  
>Cluster 37

History

Rechercher des données

data

113 MB6 6264

6267: F331.rep.fasta

6266: cd-hit on data 6265: Clusters

Add Tags

106,018 lines

format txt, génome de référence ?

Program: CD-HIT, V4.8.1 (+OpenMP), Aug 07 2022, 06:48:22

>Cluster 0  
0 5422aa, >Zm00062aa029250\_T00... at 100.00%  
1 5425aa, >Zm00062aa029250\_T00... \*  
>Cluster 1  
0 5066aa, >Zm00062aa031031\_T00... \*

6265: F331.default.HQ.prot.fa

6264: Gaspe.rep.fasta

6263: cd-hit on data 6262: Clusters

(\*) Sequence selected as representative



# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Workflow

Visualize

Données partagées

Aide

Utilisateur

Using 12%

This dataset is large and only the first megabyte is shown below.

Show all | Save

>Zm00062aa000001\_T001

SHANMPWIMHAGTEQRHAACGASLLWSSLQPSTVVMAAAATFGFLHPPIRKPAVPPLYILRLPTKPHSK

THPRSPPLLFLLLGRRRGPIAAFPNTTSSSTNAPASPTYDVREAEEAADLLREGGASADDAASIAARA

PAYAAMLADGVREDELGLWASWSSGARARLGLSGVEMEMGRLGFRKRVYLMGRSKPDHGVVPLESLG

MRLSSAKLIAPYVAAAGLTVLIDRVKFLKEMLFSSSDYAILIGRNAKRMITYLSIPADDALQSTLSFFEK

MEARYGGVSMLGHDVSFPYLIESFPMLLCSEDNHLKPLVDFLEHIGIPKPIASVLLLFPPIILSDVE

NDIKPRIREWEGAGIEQDYVSRMLLKYPWILSTSIVIENYSQMLLFFNQKRISSTVLAIAVKSWPBILGSS

SKRMNSVLELHVLFHVLGISKKMVVPVITSSPQLLLRKPDQFMQNVLFREMGVDKTTGKILCRSPEIFASN

VDNTLKKKIDFLTNFVSKHHLPRIIRKYPELLLLDINCTLPRMNYLLEMLSKKDLCSMIFRFSPLL

YSIELVMKPKLEFLLRTMKKPLKAVVEYPRYFSYSLGKIKPRFWLQSRNIDCTLEMLAKNDELFAEE

YLGGLLEKPLQSSIGS

>Zm00062aa000002\_T001

LFYLCSSVCCSIPSRSLSFSFYFFVTILNDNYRKHGERPPQPLEIVKYPALLERAWGWSILPYCPVNS

WPLYKTYLQEYYERNREVLHQVEANAASGPNLNGADENDLATLNLNLCISREVQLNLLKRRDQKHNV

EISLNEKLTNCARQITFVEYAGFPFSPVALKCIMAESDLLCMLMCGTGMSIQVNICSRIRKVAFRFMTY

KGPGCFAAATMMATTKEAKLMSGLLRNRCRENNGPFSRSVFIRKWTIAAMFRICEECSPEVESTGGYTA

TKLILDGSDDKNPCDKELVNKDNLQRNKINQYQGLFKGKTTKCRHPVQKQEPGDGATPASPSPANPIS

TVVQKLW

>Zm00062aa000002\_T002

MAMAEKVKEKMLMLRSSDNQEFVVKESVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHA

PTGPGDAAGTTNRSADDELNIFDADFVNVEHSTLLDLILAANYLDIKGLNLNARQTITDLINGKMP

ETNINNDLTIPTSSALATTMPSSERKQMEARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPLEIVK

YPALLERAWGWSILPYCPVNSWPLYKTYLQEYYERNREVLHQVEANAASGPNLNGADENDLATLNL

CISREVQLNLLKRRDQKHNVDEISLNEKLTNCARQITFVEYAGFPFSPVALKCIMAESDLLCMLMCGT

GMSIQVNICSRIRKVAFRFMTYKGPGCFAAATMMATTKEAKLMSGLLRNRCRENNGPFSRSVFIRKWTI

AAMFRICEECSPEVESTGGYTATKLILDGSDDKNPCDKELVNKDNLQRNKINQYQGLFKGKTTKCRHP

VQKQEPGDGATPASPSPANPIS TVVQKLW

>Zm00062aa000002\_T004

MAMAEKVKEKMLMLRSSDNQEFVVKESVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHA

PTGPGDAAGTTNRSADDELNIFDADFVNVEHSTLLDLILAANYLDIKGLNLNARQTITDLINGKMP

ETNINNDLTIPTSSALATTMPSSERKQMEARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPLEIVK

YPESMVNRHNNHWR

>Zm00062aa000003\_T001

MPSALRRLFATILVYCEPSDVAVLWQKHLDA SEDYQRRSQSKTHVEQMVLIDIRNMLQSMGDKITFPL

PPIIDAYDDAIGTAREVYEEESIEPAAGDVALKDSLNEEQRAAYDKILSAVDTDQGGFFVDGPGGTEKT

YLVRVPLTTLRSQGKIAVATATSGVAASIMPGGRTAHSRFKIPLTIDDGAVLPVVRKGSRAQVVASLWM

SYLWESMSHLKLVSNMRKNDPWFAEYLLRVGGGTEVTNSDGDIRLPDEVCVPYSGSDSDLDNLIDFVFP

NLNENMSDSTYITSRAILSTRNDWVDINAKMIDRFQGEHTVYHSFDSAMDPPHNYPPFELNTLTPNGL

...

History

Rechercher des données

data

86.3 MB

6

6264

6267: F331.rep.fasta

6266: cd-hit on data 6265: Clusters

6265: F331.default.HQ.prot.fasta

6264: Gaspere.rep.fasta

6263: cd-hit on data 6262: Clusters

6262: GASPE.default.HQ.prot.fasta



# CD-HIT Use case 1

Alternatives transcripts

Painfull for gene content comparison between genomes.

Try to select a representative sequence per gene.

Workflow

Visualize

Données partagées

Aide

Utilisateur

Using 12%

This dataset is large and only the first megabyte is shown below.

Show all | Save

>Zm00062aa000001\_T001

SHANMPWIMHAGTEQRHAACGASLLWSSLQPSTVVMAAAATFGFLHPPIRKPAVPPPLYILRLPTKPHSK

THPRSPPLLFLLLGRRRGPIAAFPNTTSSSTNAPASPTYDVREAEEAADLLREGGASADDAASIAARA

PAYAAMLADGVREDELGLWASWSSGARARLGLSGVEMEMGRLGFRKRVYLMGRSKPDHGVVPLLESGL

MRLSSAKLIAPYVAAAGLTVLIDRVKFLKEMLFSSSDYAILIGRNAKRMITYLSIPADDALQSTLSFFEK

MEARYGGVSMLGHDVSFPYLIESFPMLLCSEDNHLKPLVDFLEHIGIPKPIASVLLLFPPIILSDVE

NDIKPRIREWEGAGIEQDYVSRMLLKYPWILSTSIVIENYSQMLLFFNQKRISSTVLAIAVKSWPBILGSS

SKRMNSVLELHVLFHVLGISKKMVVPVITSSPQLLLRKPDQFMQNVLFREMGVDKTTGKILCRSPEIFASN

VDNTLKKKIDFLTNFVSKHHLPRIIRKYPELLLLDINCTLPRMNYLLEMLSKKDLCSMIFRFSPLL

YSIELVMKPKLEFLLRTMKKPLKAVVEYPRYFSYSLGKIKPRFWLQSRNIDCTLTEMLAKNDELFAEE

YLGGLLEKPLQSSIGS

>Zm00062aa000002\_T001

LFYLCSSVCCSIPSRSLFSFYFFVTILNDNYRKHGERPPQPLEIVKYPALLERAWGWSILPYCPVNS

WPLYKTYLQEYYERNSREVLHQVEANAASGPNLNGADENDLATLNLNLCISREVQLNLLKRRDQKHNV

EISLNEKLTNCARQITFVEYAGFPFSPVALKCIMAESDLLCMLMCGTGMSIQVNICSRIRKVAFRFMTY

KGPGCFAAATMMATTKEAKLMSGLLRNRCRENNGPFSRVFIRKWTIAAMFRICEECSPEVESTGGYTA

TKLILDGSDDKNPCDKELVNKDNLLQRNKINQYQGLFKGKTTKCRHPVQKQEPGDGATPASPSPANPIS

TVVQKLW

>Zm00062aa000002\_T002

MAMAEKVKEKMLMLRSSDNQEFEVKESVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHA

PTGPGDAAGTTNRSADDELNIFDADFVNVEHSTLLDLILAANYLDIKGLNLNARQTITDLINGKMPPEVC

KTNIKNDLTIPSTALATTMPSSERKQMEARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPLEIVK

YPALLERAWGWSILPYCPVNSWPLYKTYLQEYYERNSREVLHQVEANAASGPNLNGADENDLATLNL

CISREVQLNLLKRRDQKHNVDEISLNEKLTNCARQITFVEYAGFPFSPVALKCIMAESDLLCMLMCGT

GMSIQVNICSRIRKVAFRFMTYKGPGCFAAATMMATTKEAKLMSGLLRNRCRENNGPFSRVFIRKWTI

AAMFRICEECSPEVESTGGYTATKLILDGSDDKNPCDKELVNKDNLLQRNKINQYQGLFKGKTTKCRHP

VQKQEPGDGATPASPSPANPIS TVVQKLW

>Zm00062aa000002\_T004

MAMAEKVKEKMLMLRSSDNQEFEVKESVAMQSMTLKKMVEDGCADKGIPLPNVTSHILVKVIEYCNKHA

PTGPGDAAGTTNRSADDELNIFDADFVNVEHSTLLDLILAANYLDIKGLNLNARQTITDLINGKMPPEVC

KTNIKNDLTIPSTALATTMPSSERKQMEARFIAYMQETAERMTGFAVRGEKYRKHGERPPQPLEIVK

YPESMVNRHNNHWR

>Zm00062aa000003\_T001

MPSALRRLFATILVYCEPSDVAVLWQKHLDA SEDYQRRSQSKTHVEQMVLI DIRNMLQSMGDKITFPL

PPIIDAYDDAIGTAREVYEEESIEPAAGDVALKDSLNEEQRAAYDKILSAVDTDQGGFFVDGPGGTEKT

YLVRVPLTTLRSQGKI AVATATSGVAASIMPGGRTAHSRFKIPLTIDDGAVLPVVRKGSRAQV VASSLWM

SYLWESMSHLKLVSNMRKNDPWFAEYLLRVGGGTEVTNSDGDIRLPDEVCVPYSGSDSDLDNLIDFVFP

NLNENMSDSTYITSRAILSTRNDWVDINAKMIDRFQGEHTVYHSFDSAMDPPHNYPPFELNTLTPNGL

SRNKLKQSGDRLNTRNMLCSTRLNRSRKTALNKLKQVHGLKESGSDNLSQVHFGG

History

Rechercher des données

data

86.3 MB

6

6264

6267: F331.rep.fasta

6266: cd-hit on data 6265: Clusters

6265: F331.default.HQ.prot.fasta

6264: Gaspere.rep.fasta

6263: cd-hit on data 6262: Clusters

6262: GASPE.default.HQ.prot.fasta



# CD-HIT Use case 2

Compare 2 gene/protein dataset

From 2 annotation versions

From 2 individuals (same species)

Input : outputs from use-case 1

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

Using 12%

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit Cluster or compare biological sequence datasets (Galaxy Version 4.8.1+galaxy0)

Run Tool

Tool Parameters

Sequences to cluster/compare \*

(-i)

Cluster / Compare (i.e. call cd-hit[-est] / cd-hit[-est]-2d)?

Compare with 2nd sequence data set

Other sequences to cluster/compare \*

(-i2)

Sequence type?

Protein

For nucleotides the -est variant of cd-hit is called

Sequence identity threshold \*

0,9

Global sequence identity: number of identical alignment positions divided by the full length of the shorter sequence (-c)

Word size \*

5

Suggested word size: 5 for thresholds 0.7 ~ 1.0 4 for thresholds 0.6 ~ 0.7 3 for thresholds 0.5 ~ 0.6 2 for thresholds 0.4 ~ 0.5 (-n)

Tolerance for redundancy \*

2

(-t)

History

Rechercher des données

data

120 MB 6 6267

6270: EA1197.rep.fasta

6268: EA1197.default.HQ.prot.fa

6267: F331.rep.fasta

6265: F331.default.HQ.prot.fa

6264: Gaspé.rep.fasta

6262: GASPE.default.HQ.prot.fa



CD-HIT Use case 2

Compare 2 gene/protein dataset

From 2 annotation versions

From 2 individuals (same species)

Zm00063 : EA1197

Zm00062 : F331

Workflow

Visualize

Données partagées

Aide

Utilisateur

Using 12%

This dataset is large and only the first megabyte is shown below.

Show all | Save

>Cluster 0

05425aa, >Zm00063aa018611\_T00... \*

15425aa, >Zm00062aa029250\_T00... at 99.83%

>Cluster 1

05066aa, >Zm00063aa032449\_T00... \*

15066aa, >Zm00062aa031031\_T00... at 99.98%

>Cluster 2

04863aa, >Zm00063aa019422\_T00... \*

14863aa, >Zm00062aa006742\_T00... at 99.40%

>Cluster 3

04210aa, >Zm00063aa003366\_T00... \*

14210aa, >Zm00062aa023384\_T00... at 99.95%

>Cluster 4

03882aa, >Zm00063aa034457\_T00... \*

13877aa, >Zm00062aa033117\_T00... at 99.95%

>Cluster 5

03691aa, >Zm00063aa030400\_T00... \*

13691aa, >Zm00062aa005734\_T00... at 99.78%

>Cluster 6

03650aa, >Zm00063aa035322\_T00... \*

13631aa, >Zm00062aa013354\_T00... at 92.29%

23642aa, >Zm00062aa041101\_T00... at 99.84%

>Cluster 7

03631aa, >Zm00063aa007786\_T00... \*

>Cluster 8

03581aa, >Zm00063aa030440\_T00... \*

13581aa, >Zm00062aa005775\_T00... at 99.83%

>Cluster 9

03287aa, >Zm00063aa002573\_T00... \*

>Cluster 10

03212aa, >Zm00063aa015366\_T00... \*

>Cluster 11

03007aa, >Zm00063aa008396\_T00... \*

>Cluster 12

02872aa, >Zm00063aa020205\_T00... \*

History

Rechercher des données

data

130 MB86267

6272: cd-hit on data 6267 and data 6270: Representative sequences

6271: cd-hit on data 6267 and data 6270: Clusters

6270: EA1197.rep.fasta

6268: EA1197.default.HQ.prot.fasta

6267: F331.rep.fasta

6265: F331.default.HQ.prot.fasta

6264: Gaspé.rep.fasta

6262: GASPE.default.HQ.prot.fasta



CD-HIT Use case 2

Compare 2 gene/protein dataset

From 2 annotation versions

From 2 individuals (same species)

Zm00063 : EA1197

Zm00062 : F331

Workflow

Visualize

Données partagées

Aide

Utilisateur

Using 12%

This dataset is large and only the first megabyte is shown below.

Show all | Save

>Cluster 0

05425aa, >Zm00063aa018611\_T00... \*

15425aa, >Zm00062aa029250\_T00... at 99.83%

>Cluster 1

05066aa, >Zm00063aa032449\_T00... \*

15066aa, >Zm00062aa031031\_T00... at 99.98%

>Cluster 2

04863aa, >Zm00063aa019422\_T00... \*

14863aa, >Zm00062aa006742\_T00... at 99.40%

>Cluster 3

04210aa, >Zm00063aa003366\_T00... \*

14210aa, >Zm00062aa023384\_T00... at 99.95%

>Cluster 4

03882aa, >Zm00063aa034457\_T00... \*

13877aa, >Zm00062aa033117\_T00... at 99.95%

>Cluster 5

03691aa, >Zm00063aa030400\_T00... \*

13691aa, >Zm00062aa005734\_T00... at 99.78%

>Cluster 6

03650aa, >Zm00063aa035322\_T00... \*

13631aa, >Zm00062aa013354\_T00... at 92.29%

23642aa, >Zm00062aa041101\_T00... at 99.84%

>Cluster 7

03631aa, >Zm00063aa007786\_T00... \*

>Cluster 8

03581aa, >Zm00063aa030440\_T00... \*

13581aa, >Zm00062aa005775\_T00... at 99.83%

>Cluster 9

03287aa, >Zm00063aa002573\_T00... \*

>Cluster 10

03212aa, >Zm00063aa015366\_T00... \*

>Cluster 11

03007aa, >Zm00063aa008396\_T00... \*

>Cluster 12

02872aa, >Zm00063aa020205\_T00... \*

History

Rechercher des données

data

130 MB86267

6272: cd-hit on data 6267 and data 6270: Representative sequences

6271: cd-hit on data 6267 and data 6270: Clusters

6270: EA1197.rep.fasta

6268: EA1197.default.HQ.prot.fasta

6267: F331.rep.fasta

6265: F331.default.HQ.prot.fasta

6264: Gaspé.rep.fasta

6262: GASPE.default.HQ.prot.fasta

Orthologous gene to 13354 is missing in EA1197 ?

Orthologous genes missing in F331 ?



# CD-HIT Use case 3 : extand previous use-case to more than 2 sequences sets

Compare n gene/protein dataset

From n annotation versions

From n individuals (same species)

Input : outputs from UC1

Step 1 files concatenation

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

Tools

conca

Upload Data

Show Sections

Concatenate datasets tail-to-head (cat)

Concatenate datasets tail-to-head

Concatenate FASTA alignment by species

bcftools concat Concatenate or combine VCF/BCF files

Concatenate two BED files

hamronize summarize: Concatenate and summarize AMR detection reports

Rename.seqs Rename sequences by concatenating the group name

FASTA Merge Files and Filter Unique Sequences Concatenate FASTA database files together

WORKFLOWS

All workflows

Concatenate datasets tail-to-head (cat) (Galaxy Version 9.3+galaxy1)

Run Tool

Tool Parameters

Datasets to concatenate \*

6272: cd-hit on data 6267 and data 6270: Representative sequences

6271: cd-hit on data 6267 and data 6270: Clusters

6270: EA1197.rep.fasta

6268: EA1197.default.HQ.prot.fa

6267: F331.rep.fasta

6265: F331.default.HQ.prot.fa

Dataset

1: Dataset

Select \*

6272: cd-hit on data 6267 and data 6270: Representative sequences

6271: cd-hit on data 6267 and data 6270: Clusters

6270: EA1197.rep.fasta

6268: EA1197.default.HQ.prot.fa

6267: F331.rep.fasta

6265: F331.default.HQ.prot.fa

2: Dataset

Select \*

6270: EA1197.rep.fasta

6268: EA1197.default.HQ.prot.fa

6267: F331.rep.fasta

6265: F331.default.HQ.prot.fa

6264: Gaspere.rep.fasta

6262: GASPE.default.HQ.prot.fa

+ Insert Dataset



## CD-HIT Use case 3 : extend previous use-case to more than 2 sequences sets

Compare n gene/protein dataset

From n annotation versions

From n individuals (same species)

Input : outputs from UC1

Step 1 files concatenation

Step 2 cluster seq with CD-Hit

Galaxy France

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

cd-hit Cluster or compare biological sequence datasets (Galaxy Version 4.8.1+galaxy0)

Run Tool

Tool Parameters

Sequences to cluster/compare \*

Selected: /Comparative genomics/F331.Gaspe.EA1197.rep.fasta

(-i)

Cluster / Compare (i.e. call cd-hit[-est] / cd-hit[-est]-2d)?

Cluster sequences

Sequence type?

Protein

For nucleotides the -est variant of cd-hit is called

Sequence identity threshold \*

0,97

Global sequence identity: number of identical alignment positions divided by the full length of the shorter sequence (-c)

Word size \*

0.97



## CD-HIT Use case 3 : extend previous use-case to more than 2 sequences sets

Compare n gene/protein dataset

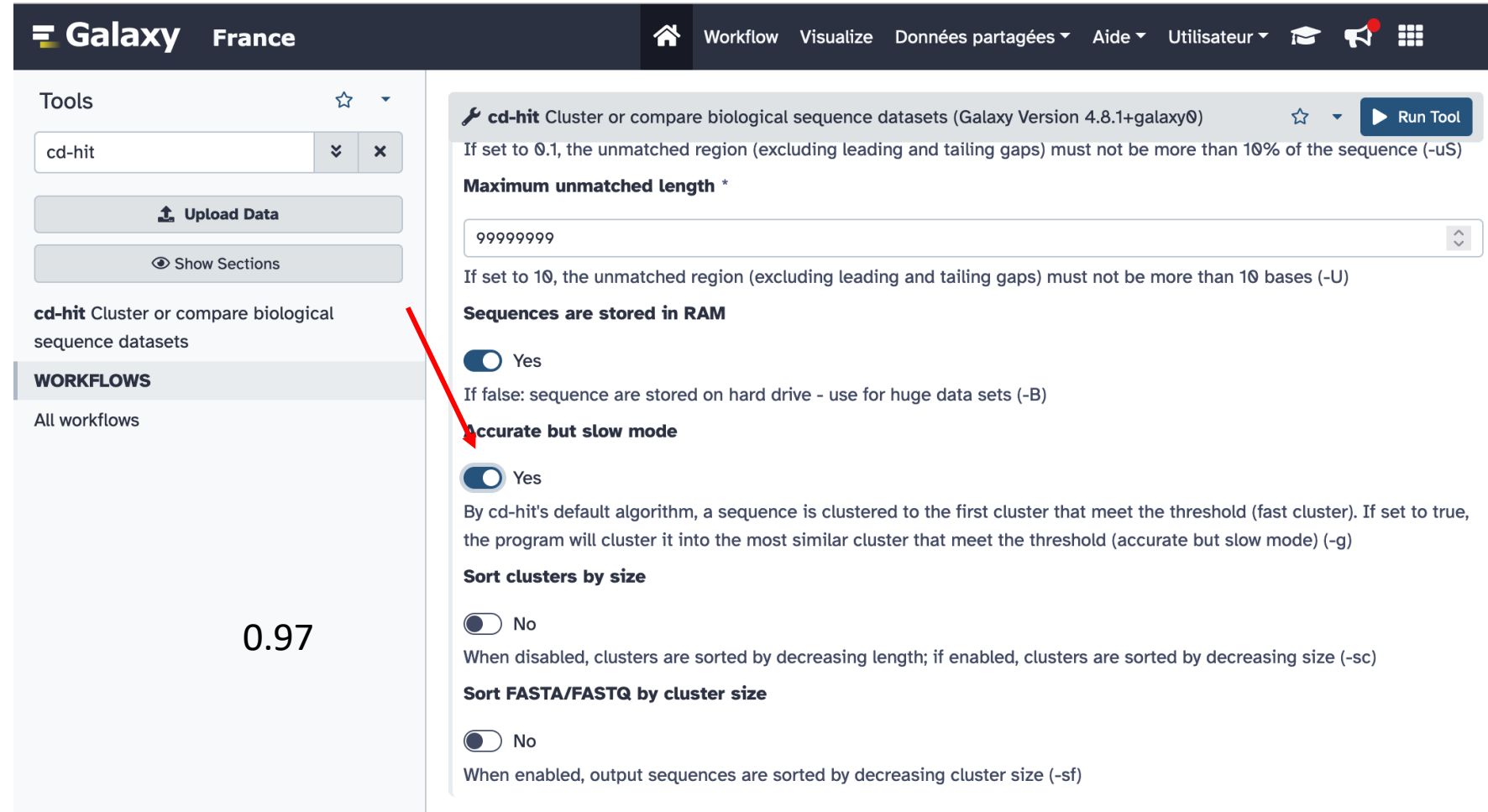
From n annotation versions

From n individuals (same species)

Input : outputs from UC1

Step 1 files concatenation

Step 2 cluster seq with CD-Hit



**Galaxy France**

Workflow Visualize Données partagées Aide Utilisateur

**Tools**

cd-hit

Upload Data

Show Sections

**cd-hit** Cluster or compare biological sequence datasets

**WORKFLOWS**

All workflows

**cd-hit** Cluster or compare biological sequence datasets (Galaxy Version 4.8.1+galaxy0)

If set to 0.1, the unmatched region (excluding leading and trailing gaps) must not be more than 10% of the sequence (-uS)

**Maximum unmatched length \***

99999999

If set to 10, the unmatched region (excluding leading and trailing gaps) must not be more than 10 bases (-U)

**Sequences are stored in RAM**

☒ Yes

If false: sequence are stored on hard drive - use for huge data sets (-B)

**Accurate but slow mode**

☒ Yes

By cd-hit's default algorithm, a sequence is clustered to the first cluster that meet the threshold (fast cluster). If set to true, the program will cluster it into the most similar cluster that meet the threshold (accurate but slow mode) (-g)

**Sort clusters by size**

☐ No

When disabled, clusters are sorted by decreasing length; if enabled, clusters are sorted by decreasing size (-sc)

**Sort FASTA/FASTQ by cluster size**

☐ No

When enabled, output sequences are sorted by decreasing cluster size (-sf)

0.97



CD-HIT Use case 3 : ex

Compare n gene/prot

From n annotation ve

From n individuals (sa

Input : outputs from L

Step 1 files concatena

Step 2 cluster seq with

- Zm00064 : Gaspe
- Zm00063 : EA1197
- Zm00062 : F331

Galaxy

France

Workflow

Visualize

Données partagées

Aide

Utilisateur

Tools

cd-hit

Upload Data

Show Sections

cd-hit Cluster or compare biological sequence datasets

WORKFLOWS

All workflows

0130aa>Zm00063aa043711\_T00... \*>Cluster 600010130aa>Zm00063aa044627\_T00... \*1130aa>Zm00062aa040294\_T00... at 100.00%>Cluster 600020130aa>Zm00063aa044846\_T00... \*>Cluster 600030130aa>Zm00063aa044897\_T00... \*1130aa>Zm00062aa040567\_T00... at 100.00%2130aa>Zm00064aa040891\_T00... at 100.00%>Cluster 600040130aa>Zm00062aa000798\_T00... \*>Cluster 600050130aa>Zm00062aa001058\_T00... \*1130aa>Zm00064aa001104\_T00... at 100.00%>Cluster 600060130aa>Zm00062aa002134\_T00... \*>Cluster 600070130aa>Zm00062aa004172\_T00... \*1130aa>Zm00064aa004266\_T00... at 100.00%>Cluster 600080130aa>Zm00062aa004378\_T00... \*1130aa>Zm00064aa004472\_T00... at 100.00%>Cluster 600090130aa>Zm00062aa005951\_T00... \*1130aa>Zm00064aa006122\_T00... at 99.23%>Cluster 600100130aa>Zm00062aa006214\_T00... \*1107aa>Zm00064aa006390\_T00... at 100.00%>Cluster 600110130aa>Zm00062aa006445\_T00... \*>Cluster 600120130aa>Zm00062aa006531\_T00... \*>Cluster 600130130aa>Zm00062aa007889\_T00... \*

Cluster 60007

Tout surligner

Respecter la casse

Respecter les accents et diacritiques

Mots entiers

Occurrence 1 sur 1



**CD-HIT Use case 3 : extand previous use-case to more than 2 sequences sets**

Compare n gene/protein dataset

From n annotation versions  
From n individuals (same species)

Input : outputs from UC1

Step 1 files concatenation

Step 2 cluster seq with CD-Hit

```
>Cluster 60004
0      130aa, >Zm00062aa000798_T00... *
>Cluster 60005
0      130aa, >Zm00062aa001058_T00... *
1      130aa, >Zm00064aa001104_T00... at 100.00%
>Cluster 60006
0      130aa, >Zm00062aa002134_T00... *
>Cluster 60007
0      130aa, >Zm00062aa004172_T00... *
1      130aa, >Zm00064aa004266_T00... at 100.00%
>Cluster 60008
0      130aa, >Zm00062aa004378_T00... *
1      130aa, >Zm00064aa004472_T00... at 100.00%
```

Zm00064 : Gaspe  
Zm00063 : EA1197  
Zm00062 : F331



## Tools

filter fasta

Upload Data

Show Sections

**Filter FASTA** on the headers and/or the sequences

**Filter fasta** to remove sequences based on input criteria (filter\_fasta)

## WORKFLOWS

All workflows

**Filter FASTA** on the headers and/or the sequences (Galaxy Version 2.3)



Run Tool

## Tool Parameters

## FASTA sequences \*



1: F331.Gaspe.EA1197.rep.fasta



## Criteria for filtering on the headers

Regular expression on the headers

## Regular expression pattern the header should match \*

Zm00064aa004266

Use the Python regular expression syntax as specified in <https://docs.python.org/3/library/re.html>

## Criteria for filtering on the sequences

No filtering

## Remove duplicate sequences



No





## Tools



miniprot



Upload Data

Show Sections

**Miniprot index** build a genome index for miniprot

**Miniprot align** align a protein sequence against a genome with affine gap penalty, splicing and frameshift

## WORKFLOWS

All workflows

## Miniprot align



Run Tool

align a protein sequence against a genome with affine gap penalty, splicing and frameshift  
(Galaxy Version 0.13+galaxy0)

## Tool Parameters

## Database type

Pre-indexed



Build an index from FASTA or use a pre-indexed database

## Pre-indexed genomic database \*



4: Zm000064aa004266 (as twobit)



A pre-indexed database built by miniprot

## Protein sequence (FASTA) \*



4: Zm000064aa004266



Protein sequences to be aligned in FASTA format

## Output format \*

GFF3





Tools

miniprot

Upload Data

Show Sections

**Miniprot index** build a genome index for miniprot

**Miniprot align** align a protein sequence against a genome with affine gap penalty, splicing and frameshift

WORKFLOWS

All workflows

compa

/Comparative genomics/Gaspe.genome.fasta.gz	-	-
/Comparative genomics/F331.rep.fasta	-	-
/Comparative genomics/F331.genome.fasta.gz	-	-
/Comparative genomics/F331.Gaspe.EA1197.rep.fasta	-	-
/Comparative genomics/EA1197Sub.fasta	-	-
/Comparative genomics/EA1197.rep.fasta	-	-
/Comparative genomics/EA1197int.bed	-	-
/Comparative genomics/EA1197.genome.mpi	-	-
/Comparative genomics/EA1197.genome.fasta.gz	-	-

Back Upload

Cancel

Protein sequences to be aligned in FASTA format

Output format \*

GFF3



## Tools



miniprot



Upload Data

Show Sections

**Miniprot index** build a genome index for miniprot

**Miniprot align** align a protein sequence against a genome with affine gap penalty, splicing and frameshift

## WORKFLOWS

All workflows

## Miniprot align



Run Tool

align a protein sequence against a genome with affine gap penalty, splicing and frameshift  
(Galaxy Version 0.13+galaxy0)

## Tool Parameters

## Database type

Pre-indexed



Build an index from FASTA or use a pre-indexed database

## Pre-indexed genomic database \*



Selected: /Comparative genomics/EA1197.genome.mpi



A pre-indexed database built by miniprot

## Protein sequence (FASTA) \*



4: Zm00064aa004266



Protein sequences to be aligned in FASTA format

## Output format \*

GFF3





[Workflow](#)[Visualize](#)[Données partagées ▾](#)[Aide ▾](#)[Utilisateur ▾](#)

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
##PAF	Zm00064aa004266_T001	130	0	130	-	chr8	193330780	31199353 31199915 366 390 0 AS:i:536 ms:i:616 np:i:124 fs:i:0 st:i:0 da:i:0 do:i:0 cg:Z:44M83U53M92N32M cs:Z::14*tacC:29*gG~gt80ag-gc:2*ggcS:35*aagT:2*gagk8*cgC*cagK~gt92ag:18*cgC:13
chr8	miniprot	mRNA	31199351	31199915	616	-	.	ID=MP0000001;Rank=1;Identity=0.9385;Positive=0.9538;Target=Zm00064aa004266_T001 1 130
chr8	miniprot	CDS	31199783	31199915	210	-	0	Parent=MP0000001;Rank=1;Identity=0.9778;Target=Zm00064aa004266_T001 1 4
chr8	miniprot	CDS	31199542	31199702	245	-	2	Parent=MP0000001;Rank=1;Identity=0.8868;Target=Zm00064aa004266_T001 45
chr8	miniprot	CDS	31199351	31199449	161	-	0	Parent=MP0000001;Rank=1;Identity=0.9688;Target=Zm00064aa004266_T001 99
chr8	miniprot	stop_codon	31199351	31199353	0	-	0	Parent=MP0000001;Rank=1



[Workflow](#)[Visualize](#)[Données partagées ▾](#)[Aide ▾](#)[Utilisateur ▾](#)

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
##PAF	Zm00064aa004266_T001	130	0	130	-	chr8	193330780	31199353 31199915 366 390 0 AS:i:536 ms:i:616 np:i:124 fs:i:0 st:i:0 da:i:0 do:i:0 cg:Z:44M83U53M92N32M cs:Z::14*tacC:29*gG~gt80ag-gc:2*ggcS:35*aagT:2*gagk8*cgC*cagK~gt92ag:18*cgC:13
chr8	miniprot	mRNA	31199351	31199915	616	-	.	ID=MP0000001;Rank=1;Identity=0.9385;Positive=0.9538;Target=Zm00064aa004266_T001 1 130
chr8	miniprot	CDS	31199783	31199915	210	-	0	Parent=MP0000001;Rank=1;Identity=0.9778;Target=Zm00064aa004266_T001 1 4
chr8	miniprot	CDS	31199542	31199702	245	-	2	Parent=MP0000001;Rank=1;Identity=0.8868;Target=Zm00064aa004266_T001 45
chr8	miniprot	CDS	31199351	31199449	161	-	0	Parent=MP0000001;Rank=1;Identity=0.9688;Target=Zm00064aa004266_T001 99
chr8	miniprot	stop_codon	31199351	31199353	0	-	0	Parent=MP0000001;Rank=1



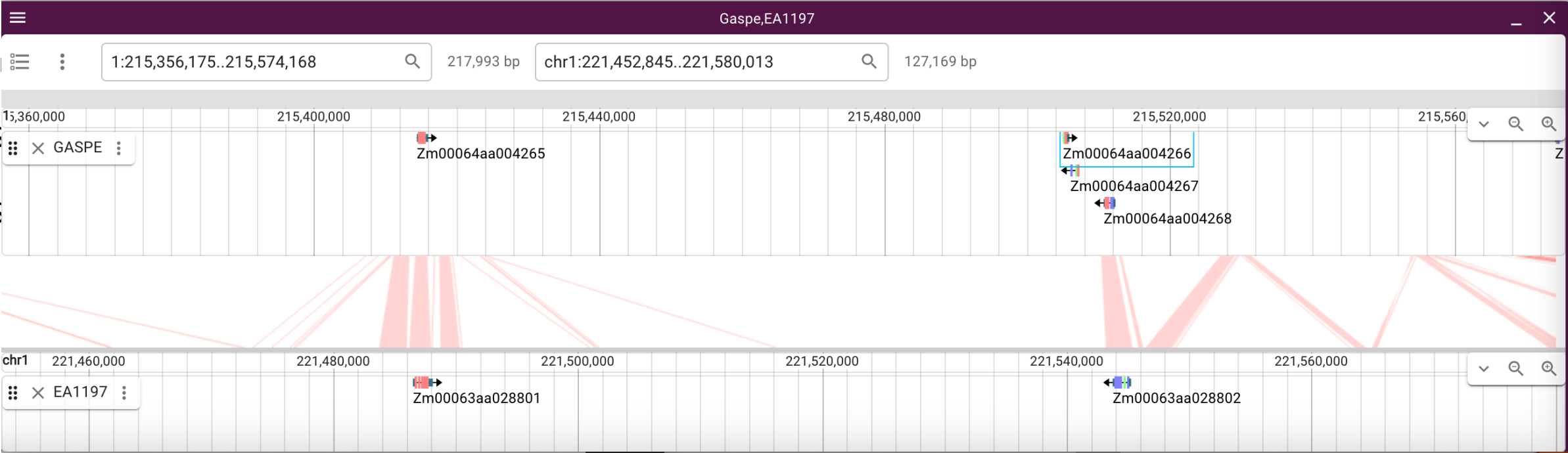
# CD-HIT Use case 3 : extand previous use-case to more than 2 sequences sets

Compare n gene/protein dataset

From n annotation versions

From n individuals (same species)

```
>Cluster 60007
0      130aa, >Zm00062aa004172_T00... *
1      130aa, >Zm00064aa004266_T00... at 100.00%
```







## Tools



getfasta



Upload Data

Show Sections

**bedtools getfasta** use intervals to extract sequences from a FASTA file

## WORKFLOWS

All workflows

**bedtools getfasta** use intervals to extract sequences from a FASTA file (Galaxy Version 2.31.1+galaxy0)

Run Tool

## Tool Parameters

BED/bedGraph/GFF/VCF/EncodePeak file \*



Selected: /Comparative genomics/gaspeInt.bed



(-bed)

## Choose the source for the FASTA file

History



FASTA file \*



Selected: /Comparative genomics/Gaspe.genome.fasta.gz



(-fi)

Use the 'name' column in the BED file and the coordinates for the FASTA headers in the output FASTA file

☒ No

(-name)

Use the 'name' column in the BED file for the FASTA headers in the output FASTA file

☒ No

(-nameOnly)

Report extract sequences in a tab-delimited format instead of in FASTA format

☒ No

(-tab)

Force strandedness

☒ No

If the feature occupies the antisense strand, the sequence will be reverse complemented (-s)



## Tools



getfasta



Upload Data

Show Sections

**bedtools getfasta** use intervals to extract sequences from a FASTA file

## WORKFLOWS

All workflows

**bedtools getfasta** use intervals to extract sequences from a FASTA file (Galaxy Version 2.31.1+galaxy0)

Run Tool

## Tool Parameters

BED/bedGraph/GFF/VCF/EncodePeak file \*



Selected: /Comparative genomics/EA1197int.bed



(-bed)

## Choose the source for the FASTA file

History



FASTA file \*



Selected: /Comparative genomics/EA1197.genome.fasta.gz



(-fi)

Use the 'name' column in the BED file and the coordinates for the FASTA headers in the output FASTA file

☒ No

(-name)

Use the 'name' column in the BED file for the FASTA headers in the output FASTA file

☒ No

(-nameOnly)

Report extract sequences in a tab-delimited format instead of in FASTA format

☒ No

(-tab)

Force strandedness



## Tools



chrom



Upload Data

Show Sections

**Chromeister** ultra-fast pairwise genome comparisons

**SnEff chromosome-info:** list chromosome names/lengths

**xcms findChromPeaks (xcmsSet)**  
Chromatographic peak detection

**xcms plot chromatogram** Plots base peak intensity chromatogram (BPI) and total ion current chromatogram (TIC) from MSnbase or xcms experiment(s)

**xcms findChromPeaks Merger** Merge xcms findChromPeaks RData into a unique file to be used by group

**xcms groupChromPeaks (group)** Perform the correspondence, the grouping of chromatographic peaks within and between samples.

**xcms refineChromPeaks (refine)** Remove or merge chromatographic peaks based on specific criteria.

**xcms fillChromPeaks (fillPeaks)** Integrate

**Chromeister** ultra-fast pairwise genome comparisons (Galaxy Version 1.5.a+galaxy1)

## Tool Parameters

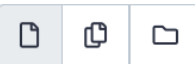
## Query sequence \*



12: EA1197\_cluster60007

Query sequence file in fasta format

## Reference sequence \*



9: gaspe\_cluster60007

Reference sequence file in fasta format

## Output dotplot size \*

500



Use around 1000 for chromosome-sized sequences and around 2000 for complete genomes

## K-mer seed size \*

32

Use 32 as default, and 16 in case no similarities are found

## Diffuse value \*

1



Level of the heuristic subsampling employed. Change to 1 or 2 if no similarity is found

## Add grid to plot for multi-fasta data sets



Yes



## Tools



chrom



Upload Data

Show Sections

**Chromeister** ultra-fast pairwise genome comparisons

**SnEff chromosome-info:** list chromosome names/lengths

**xcms findChromPeaks (xcmsSet)**

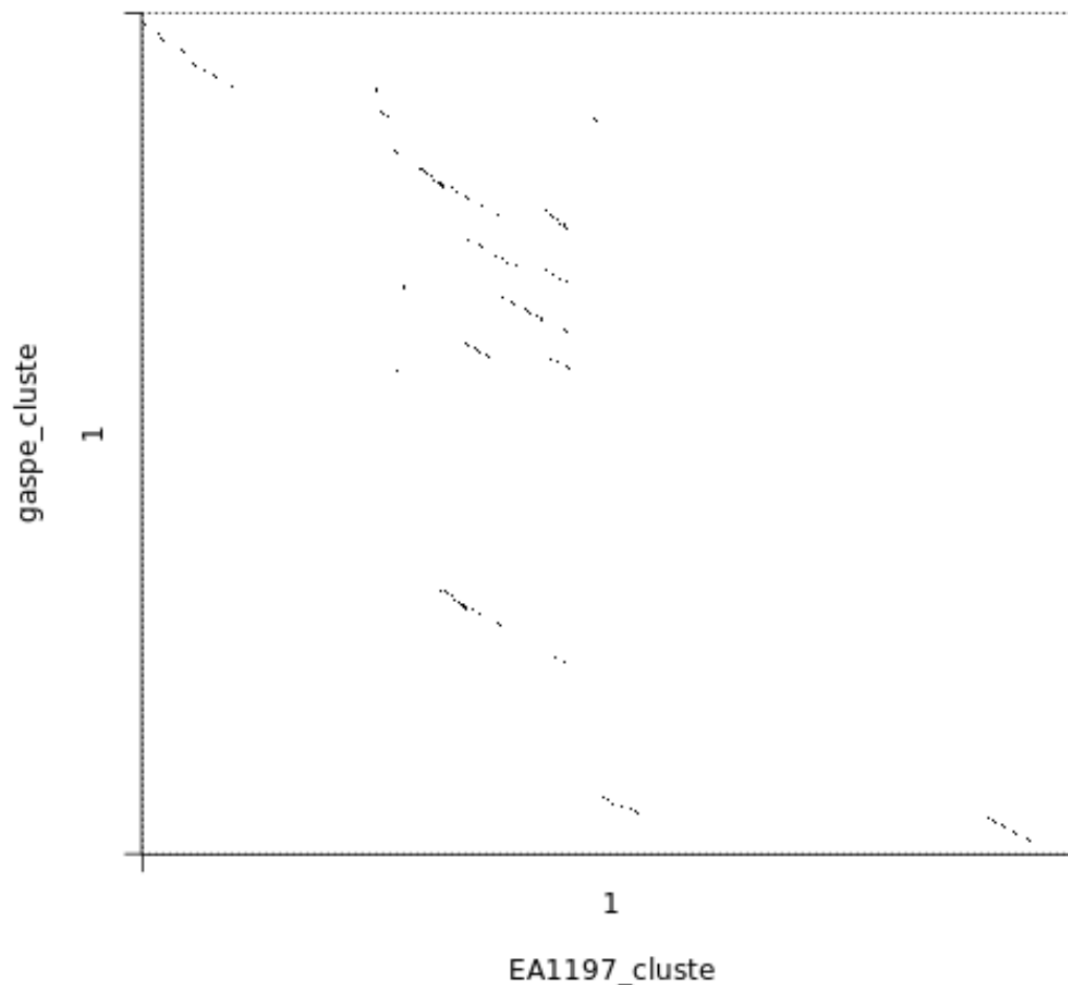
Chromatographic peak detection

**xcms plot chromatogram** Plots base peak intensity chromatogram (BPI) and total ion current chromatogram (TIC) from MSnbase or xcms experiment(s)

**xcms findChromPeaks Merger** Merge xcms findChromPeaks RData into a unique file to be used by group

**xcms groupChromPeaks (group)** Perform

**EA1197\_cluster60007-gaspe\_cluster60007.mat filt. score= 0.!**





Tools

chrom

Upload Data

Show Sections

- Chromeister** ultra-fast pairwise genome comparisons
- SnEff chromosome-info:** list chromosome names/lengths
- xcms findChromPeaks (xcmsSet)**  
Chromatographic peak detection
- xcms plot chromatogram** Plots base peak intensity chromatogram (BPI) and total ion current chromatogram (TIC) from MSnbase or xcms experiment(s)
- xcms findChromPeaks Merger** Merge xcms findChromPeaks RData into a unique file to be used by group
- xcms groupChromPeaks (group)** Perform the correspondence, the grouping of chromatographic peaks within and between samples.
- xcms refineChromPeaks (refine)** Remove or merge chromatographic peaks based on specific criteria.
- xcms fillChromPeaks (fillPeaks)** Integrate areas of missing peaks

Chromeister ultra-fast pairwise

Tool Parameters

Query sequence \*

4: Zm00064aa

Query sequence file in fasta format

Reference sequence \*

4: Zm00064aa

Reference sequence file in fasta format

Output dotplot size \*

1000

Use around 1000 for chromosome-size

K-mer seed size \*

32

Use 32 as default, and 16 in case no

Diffuse value \*

4

Level of the heuristic subsampling employed. Change to 1 or 2 if no similarity is found

Add grid to plot for multi-fasta data sets

Yes

Do not use grid if your multi-fasta contains more than a hundred sequences (approximately)

compa

Label	Details	Time
/Comparative genomics/gaspeSub.fasta	-	-
/Comparative genomics/Gaspe.rep.fasta	-	-
/Comparative genomics/gaspeInt.bed	-	-
/Comparative genomics/Gaspe.genome.fasta.gz	-	-
/Comparative genomics/F331.rep.fasta	-	-
/Comparative genomics/F331.genome.fasta.gz	-	-
/Comparative genomics/F331.Gaspe.EA1197.rep.fasta	-	-
/Comparative genomics/EA1197Sub.fasta	-	-
/Comparative genomics/EA1197.rep.fasta	-	-
/Comparative genomics/EA1197int.bed	-	-
/Comparative genomics/EA1197.genome.mpi	-	-
/Comparative genomics/EA1197.genome.fasta.gz	-	-

Back Upload Cancel



CD-HIT Use case 3 : extand previous use-c

Compare n gene/protein dataset

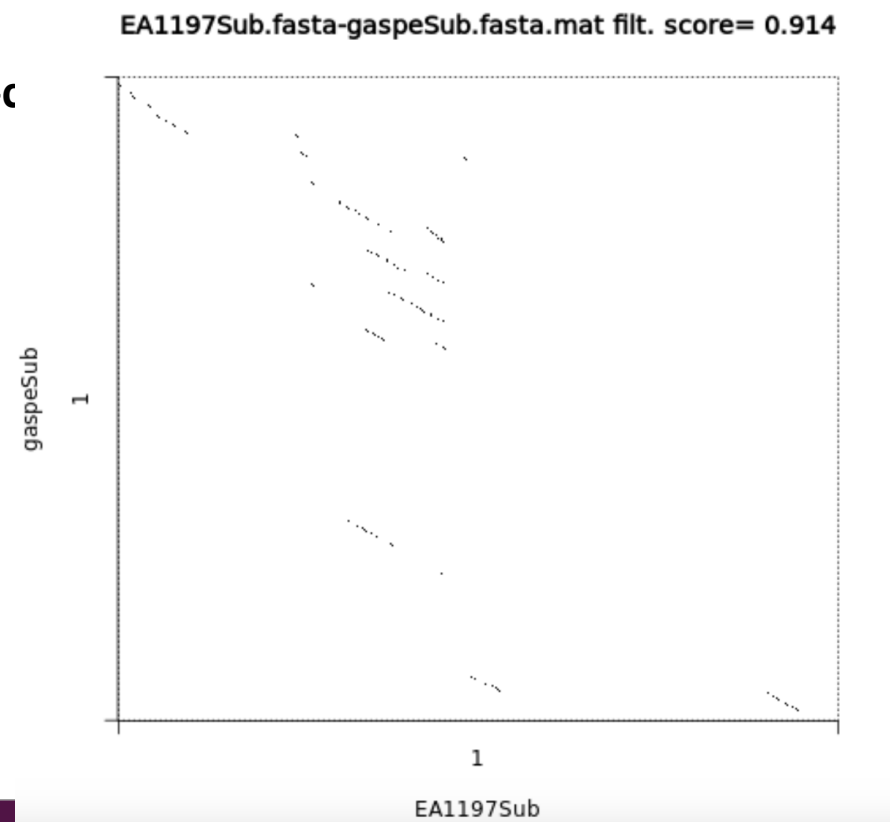
From n annotation versions

From n individuals (same species)

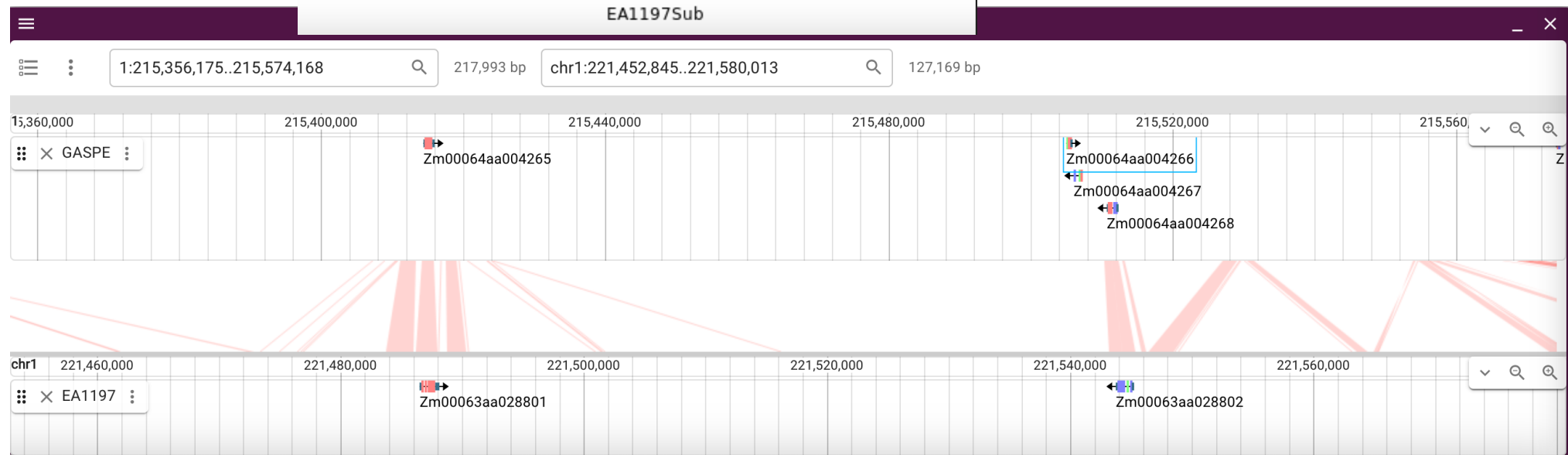
Input : outputs from UC1

Step 1 files concatenation

Step 2 cluster seq with CD-Hit



Zm00064 : Gaspe  
Zm00063 : EA1197  
Zm00062 : F331





## Tools



chrom



Upload Data

Show Sections

**Chromeister** ultra-fast pairwise genome comparisons

**SnpEff chromosome-info:** list chromosome names/lengths

**xcms findChromPeaks (xcmsSet)**

Chromatographic peak detection

**xcms plot chromatogram** Plots base peak intensity chromatogram (BPI) and total ion current chromatogram (TIC) from MSnbase or xcms experiment(s)

**xcms findChromPeaks Merger** Merge xcms findChromPeaks RData into a unique file to be used by group

**xcms groupChromPeaks (group)** Perform the correspondence, the grouping of chromatographic peaks within and between samples.

**xcms refineChromPeaks (refine)** Remove or merge chromatographic peaks based on specific criteria.

**xcms fillChromPeaks (fillPeaks)** Integrate areas of missing peaks

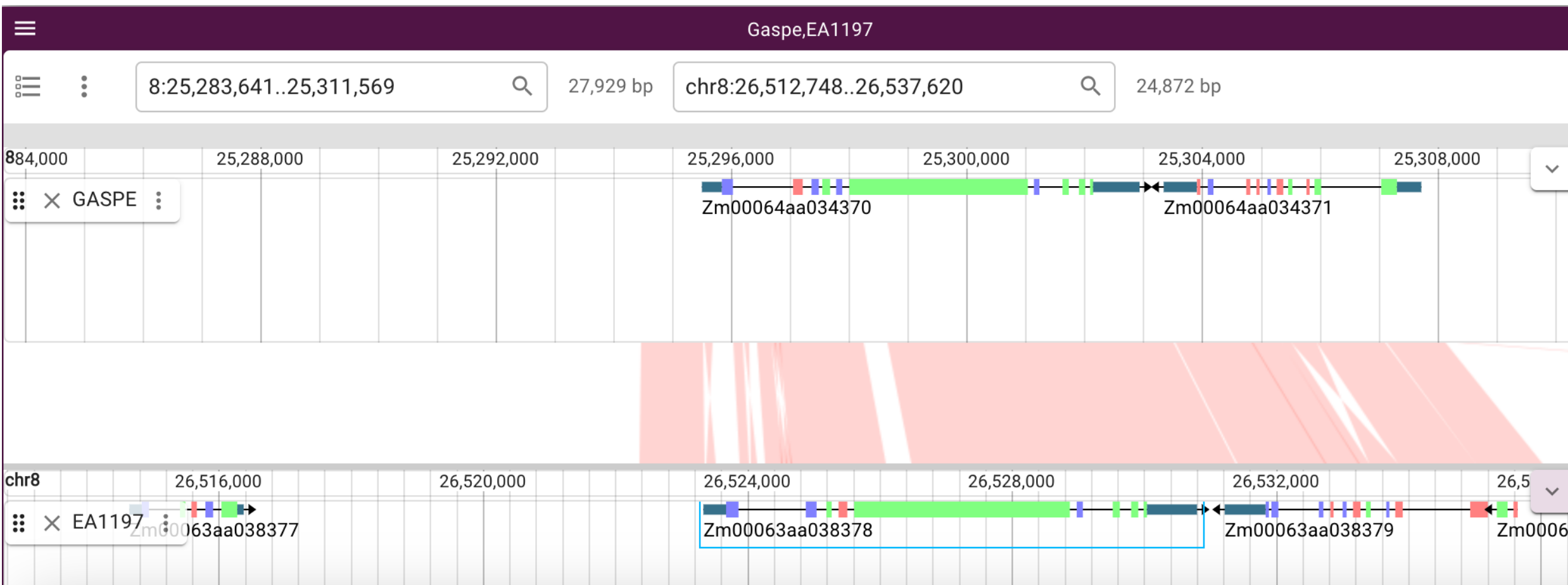
**bcftools color-chrs** plugin Color shared chromosomal segments, requires phased

CTs



```
>Cluster 608
0 1394aa, >Zm00063aa013167_T00... at 98.78%
1 1396aa, >Zm00062aa018836_T00... *
2 1396aa, >Zm00064aa019116_T00... at 99.14%
>Cluster 609
0 1396aa, >Zm00064aa026404_T00... *
>Cluster 610
0 1396aa, >Zm00064aa040375_T00... *
>Cluster 611
0 1392aa, >Zm00063aa036892_T00... at 98.20%
1 1262aa, >Zm00062aa042738_T00... at 98.81%
2 1396aa, >Zm00064aa043042_T00... *
>Cluster 612
0 1395aa, >Zm00063aa038378_T00... *
1 1395aa, >Zm00062aa034030_T00... at 99.93%
>Cluster 613
0 1395aa, >Zm00062aa021699_T00... *
>Cluster 614
0 1394aa, >Zm00063aa020662_T00... *
1 1394aa, >Zm00062aa007996_T00... at 99.93%
2 1394aa, >Zm00064aa008223_T00... at 99.93%
>Cluster 615
0 1394aa, >Zm00062aa006093_T00... *
1 147aa, >Zm00062aa020836_T00... at 97.96%
2 147aa, >Zm00064aa021133_T00... at 97.96%
>Cluster 616
0 1392aa, >Zm00063aa014460_T00... *
1 1392aa, >Zm00062aa020118_T00... at 99.93%
2 1392aa, >Zm00064aa020356_T00... at 99.93%
>Cluster 617
0 1392aa, >Zm00063aa031317_T00... *
1 1234aa, >Zm00062aa006621_T00... at 97.97%
>Cluster 618
0 1391aa, >Zm00063aa022626_T00... *
1 1303aa, >Zm00062aa009876_T00... at 98.54%
2 1302aa, >Zm00064aa010183_T00... at 99.77%
>Cluster 619
0 1391aa, >Zm00064aa041061_T00... *
>Cluster 620
0 1146aa, >Zm00063aa013849_T00... at 98.17%
1 1390aa, >Zm00063aa013849_T00... *
2 1039aa, >Zm00063aa013849_T00... at 98.85%
>Cluster 621
0 1390aa, >Zm00063aa044387_T00... *
>Cluster 622
0 1390aa, >Zm00064aa010602_T00... *
```







## Tools



chrom



Upload Data

Show Sections

**Chromeister** ultra-fast pairwise genome comparisons

**SnpEff chromosome-info:** list chromosome names/lengths

**xcms findChromPeaks (xcmsSet)**

Chromatographic peak detection

**xcms plot chromatogram** Plots base peak intensity chromatogram (BPI) and total ion current chromatogram (TIC) from MSnbase or xcms experiment(s)

**xcms findChromPeaks Merger** Merge xcms findChromPeaks RData into a unique file to be used by group

**xcms groupChromPeaks (group)** Perform the correspondence, the grouping of chromatographic peaks within and between samples.

**xcms refineChromPeaks (refine)** Remove or merge chromatographic peaks based on specific criteria.

**xcms fillChromPeaks (fillPeaks)** Integrate areas of missing peaks

**bcftools color-chrs** plugin Color shared chromosomal segments, requires phased

CTC



```
>Cluster 608
0      1394aa, >Zm00063aa013167_T00... at 98.78%
1      1396aa, >Zm00062aa018836_T00... *
2      1396aa, >Zm00064aa019116_T00... at 99.14%
>Cluster 609
0      1396aa, >Zm00064aa026404_T00... *
>Cluster 610
0      1396aa, >Zm00064aa040375_T00... *
>Cluster 611
0      1392aa, >Zm00063aa036892_T00... at 98.20%
1      1262aa, >Zm00062aa042738_T00... at 98.81%
2      1396aa, >Zm00064aa043042_T00... *
>Cluster 612
0      1395aa, >Zm00063aa038378_T00... *
1      1395aa, >Zm00062aa034030_T00... at 99.93%
>Cluster 613
0      1395aa, >Zm00062aa021699_T00... *
>Cluster 614
0      1394aa, >Zm00063aa020662_T00... *
1      1394aa, >Zm00062aa007996_T00... at 99.93%
2      1394aa, >Zm00064aa008223_T00... at 99.93%
2      1368aa, >Zm00064aa012957_T00... at 99.85%
>Cluster 660
0      1368aa, >Zm00064aa034370_T00... *
>Cluster 661
0      1367aa, >Zm00062aa038519_T00... *
2      1392aa, >Zm00064aa020356_T00... at 99.93%
>Cluster 617
0      1392aa, >Zm00063aa031317_T00... *
1      1234aa, >Zm00062aa006621_T00... at 97.97%
>Cluster 618
0      1391aa, >Zm00063aa022626_T00... *
1      1303aa, >Zm00062aa009876_T00... at 98.54%
2      1302aa, >Zm00064aa010183_T00... at 99.77%
>Cluster 619
0      1391aa, >Zm00064aa041061_T00... *
>Cluster 620
0      1146aa, >Zm00063aa013849_T00... at 98.17%
1      1390aa, >Zm00063aa013849_T00... *
2      1039aa, >Zm00063aa013849_T00... at 98.85%
>Cluster 621
0      1390aa, >Zm00063aa044387_T00... *
>Cluster 622
0      1390aa, >Zm00064aa010602_T00... *
```



## Tools





 Upload Data

 Show Sections

**needle** Needleman-Wunsch global alignment

**RAMClustR define experiment** Definition of experimental design used for record keeping and writing spectra data.

**Interactive CellXgene Environment**

**MAGeCK mle** - perform maximum-likelihood estimation of gene essentiality scores


**featureCounts** Measure gene expression in RNA-Seq experiments from SAM or BAM files

**xcms get a sampleMetadata file** which need to be filled with extra information

**xcms get a sampleMetadata file** which need to be filled with extra information

## WORKFLOWS

All workflows

 **needle** Needleman-Wunsch global alignment (Galaxy Version 5.0.0.1)



 Run Tool

## Tool Parameters

### Sequence 1 \*










### Sequence 2 \*










### Gap open penalty \*



### Gap extension penalty \*



### Brief identity and similarity \*



### Output alignment file format \*



## Additional Options

### Email notification

☐ No

Send an email notification when the job completes.



Galaxy

France

Workflow

Visualize

Données partagées

Aide

Ut

Tools

nee

Upload Data

Show Sections

needle

Needleman-Wunsch global alignment

RAMClustR

define experiment

Definition of experimental design used for record keeping and writing spectra data.

Interactive CellXgene

Environment

MAGECK

mle

- perform maximum-likelihood estimation of gene essentiality scores

featureCounts

Measure gene expression in RNA-Seq experiments from SAM or BAM files

xcms

get a sampleMetadata file

which need to be filled with extra information

xcms

get a sampleMetadata file

which need to be filled with extra information

WORKFLOWS

All workflows

#=====

#

# Aligned\_sequences: 2

# 1: Zm00063aa038378\_T001

# 2: Zm00064aa034370\_T001

# Matrix: EBL0SUM62

# Gap\_penalty: 10.0

# Extend\_penalty: 0.5

#

# Length: 1410

# Identity: 1348/1410 (95.6%)

# Similarity: 1351/1410 (95.8%)

# Gaps: 57/1410 ( 4.0%)

# Score: 6944.5

#

#

#=====

Zm00063aa0383

1

MPLVRFVEVRNEVGLGDPGLYGGGAAAAAATAAAGEEKPALLEGVAVAG

50

Zm00064aa0343

1

MPLVRFVEVRNEVGLGDPGLYGGGAAAAAATAAAGEEKPALLEGVAVAG

50

Zm00063aa0383

51

LVGILRQLGDLAEFAADVFDLHEQVIATSARGRKVLTRVQNI EAALPSL

100

Zm00064aa0343

51

LVGILRQLGDLAEFAADVFDLHEQVIATSARGRKVLTRVQNI EAALPSL

100

Zm00063aa0383

101

EKAVKNQKSHIHFAYVPGSDWHTQLQNEQNHLLATDLPRFMMS-----

144

Zm00064aa0343

101

EKAVKNQKSHIHFAYVPGSDWHTQLQNEQNHLLATDLPRFMMSYEECRD

150

Zm00063aa0383

145

-----FDNAGAGACVKRYSDPSYFKKAWDTMRADKNANVQREKRSQ

185

Zm00064aa0343

151

PPRLYLDDKFDNAGAGACVKRYSDPSYFKKAWDTMRADKNANVQREKRSQ

200

Zm00063aa0383

186

KIKRKGSRLEPYHGQAMPRHRGTEFQRSLTAGQLVNRCTSHTLRYNNF

235

Zm00064aa0343

201

KIKRKGSRLEPYHGQAMPRHRGTEFQRSLTAGQLVN-----

237

Zm00063aa0383

236

VSVIILYGCSLLIFEWHIQNSIFNSVFSSRQFASPSNDGGGISEHKSTPD

285

Zm00064aa0343

238

-----RQFASPSNDGGGISEHKSTPD

258

Zm00063aa0383

286

ARSNPDNMSRSSSLSSKTLSSVEQAIDTKPSAVSHENGHGKSLDTKLHK

335

Zm00064aa0343

259

ARSNPDNMSRSSSLSSKTLSSVEQAIDTKPSAVSHENGHGKSLDTKLHK

308







# OrthoFinder: phylogenetic orthology inference for comparative genomics

Emms and Kelly *Genome Biology* (2019) 20:238  
<https://doi.org/10.1186/s13059-019-1832-y>

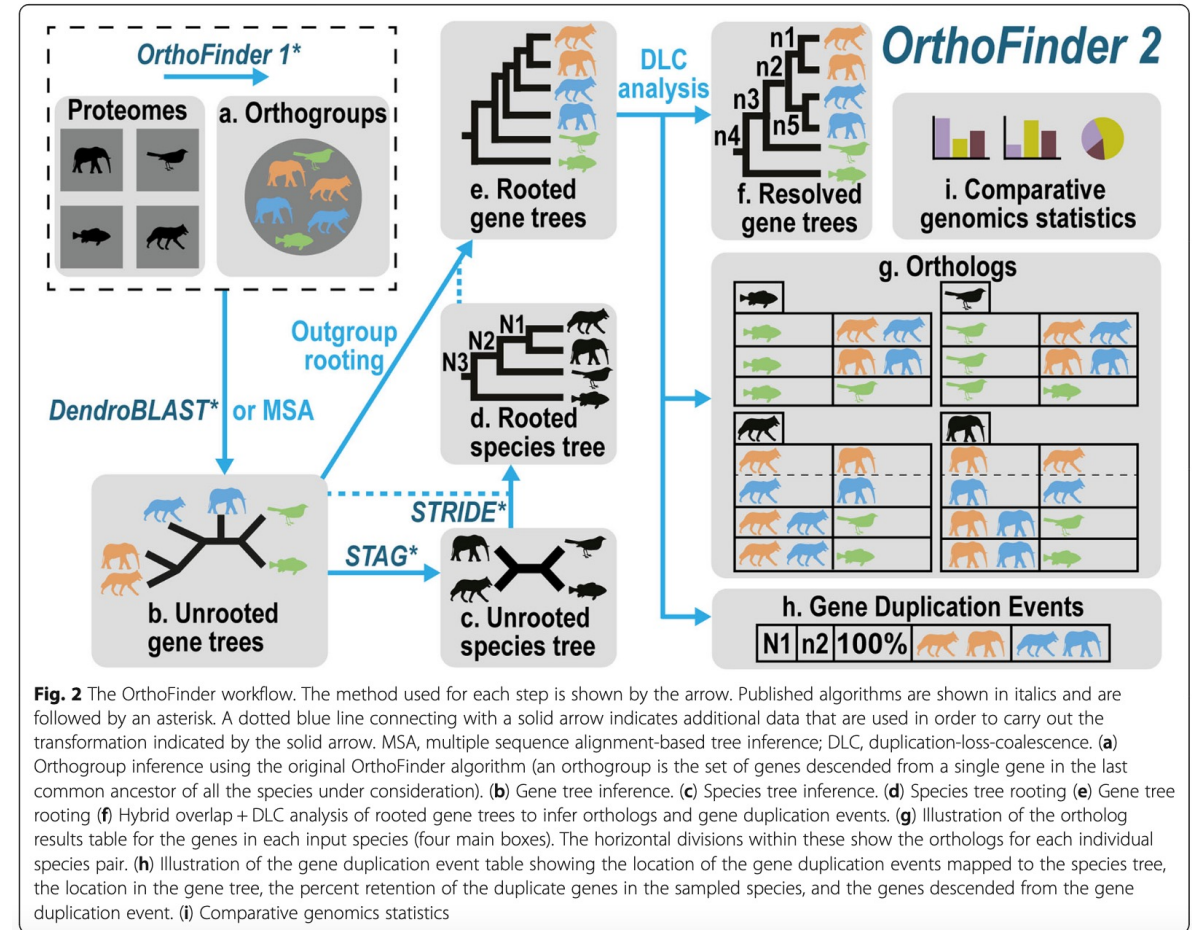
Genome Biology

SOFTWARE

Open Access

## OrthoFinder: phylogenetic orthology inference for comparative genomics

David M. Emms and Steven Kelly\*





## OrthoFinder best practices : Selecting which species to include

The first question to ask yourself is what species should you include. The answer to this probably depends on what kind of analysis you want to perform.

Three standard analyses are:

- Performing a comparative analysis across a clade of species
- Identifying orthologs between a pair, or among a small number, of species
- Investigating changes at a particular point in evolutionary history

In the first case just get the proteomes for all the species in your clade that you can. Generally, you don't need to include an outgroup for your clade of interest—in fact, this will push back the point in evolutionary history at which your orthogroups are defined ([Orthogroups, Orthologs & Paralogs](#)) and so it's usually better not to since your orthogroups will have lower resolution.

In the second case, it is good to ensure you have sufficient species sampling so as to get the best results. The same rule applies as for inferring a good phylogenetic tree: you should break up long branches with intermediate species. **You want an absolute minimum of 4 species and somewhere between 6-10 is probably optimal.**

If you're interested in what happened on a particular branch of the species tree, then you should likewise ensure good species sampling—ideally at least two species below the branch, at least two species on the closest branch above and two or more species in the outgroup.



# OrthoFinder. Our dataset:

fungi.prot.fasta

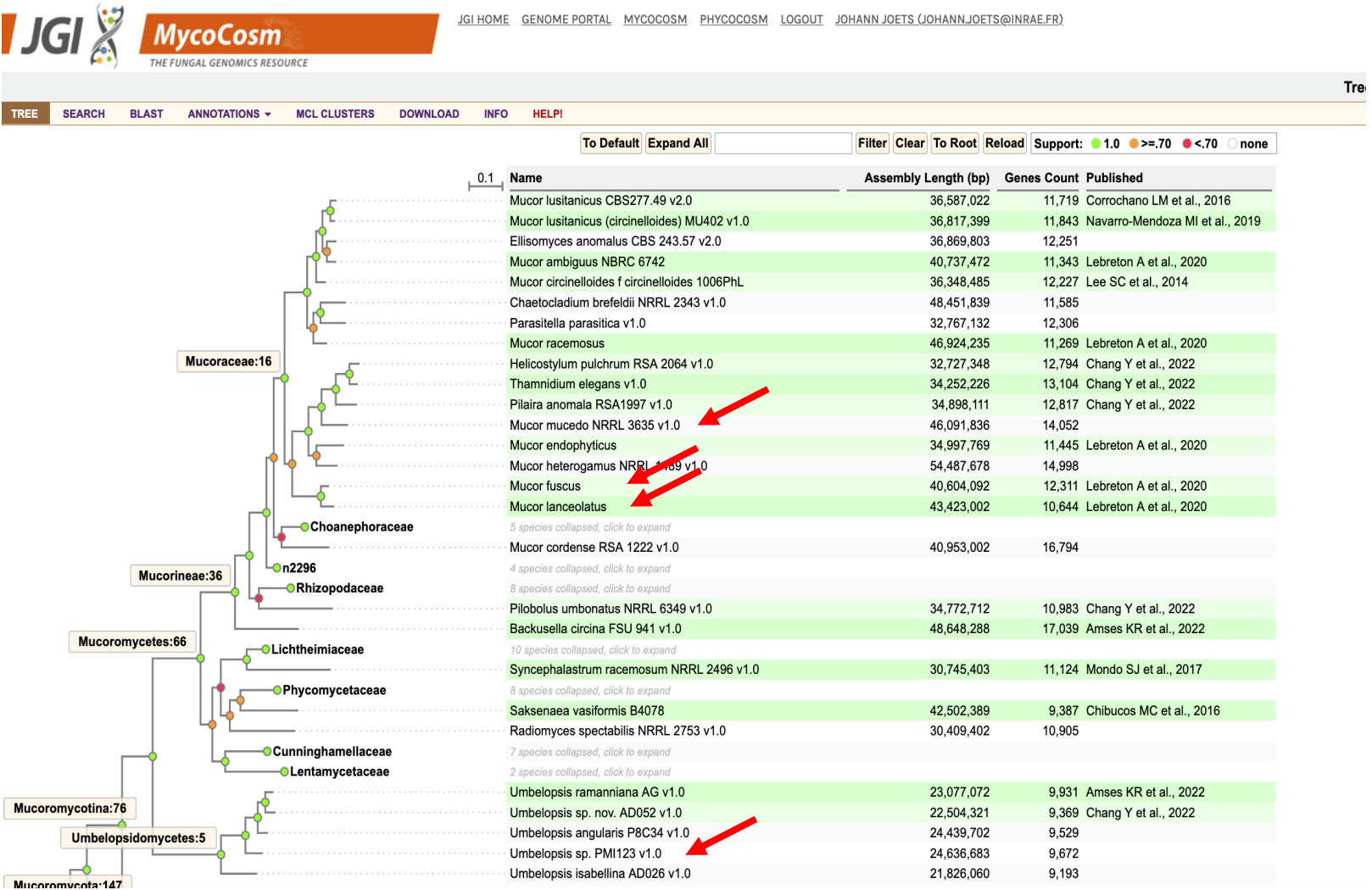
a list with 4 items

Mmucedo.aa.fasta

Umbsp.aa.fasta.gz


Mucfus1.aa.fasta.gz









Muclan1.aa.fasta.gz








# OrthoFinder: Galaxy


 **Galaxy France**


 Workflow  Visualize  Données partagées ▾  Aide ▾  Utilisateur ▾   

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

**Tools**  

orthofinder 


 Upload Data

 Show Sections

**OrthoFinder** finds orthogroups in a set of proteomes

**WORKFLOWS**

All workflows

 Executed **OrthoFinder** and successfully added 1 job to the queue.

The tool uses this input:

- **10: fungi.prot.fasta (with implicit datatype conversion)**

It produces 12 outputs:

- **20: OrthoFinder on data 5, data 6, and others: orthogroups (txt)**
- **21: OrthoFinder on data 5, data 6, and others: orthogroups (tsv)**
- **22: OrthoFinder on data 5, data 6, and others: species overlaps**
- **23: OrthoFinder on data 5, data 6, and others: unassigned genes**
- **24: OrthoFinder on data 5, data 6, and others: overall comparative genomics statistics**
- **25: OrthoFinder on data 5, data 6, and others: per species comparative genomics statistics**
- **26: OrthoFinder on data 5, data 6, and others: species tree**
- **27: OrthoFinder on data 5, data 6, and others: species tree with node labels**
- **28: OrthoFinder on data 5, data 6, and others: species tree with duplication events**
- **29: OrthoFinder on data 5, data 6, and others: duplication events**
- **30: OrthoFinder on data 5, data 6, and others: duplications per orthogroup**
- **31: OrthoFinder on data 5, data 6, and others: duplications per species tree node**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.



# OrthoFinder: Galaxy

27: OrthoFinder on data  
5, data 6, and others: spe  
cies tree with node labels



1 line

format: **newick**, génome de référence: ?

OrthoFinder version 2.5.4 Copyright  
(C) 2014 David Emms

2022-09-28 04:21:27 : Starting  
OrthoFinder 2.5.4

1 thread(s) for highly parallel tasks  
(BLAST searches etc.)

1 thread(s) for OrthoFinder algorithm

Checking required programs are  
installed


-----











(Umbasp.aa.fasta.gz:0.326289,(Mmucedo.aa.fa






# OrthoFinder: Galaxy


 **Galaxy France**

 Workflow  Visualize  Données partagées  Aide  Utilisateur   

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

**Tools**  



 Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation


Filter and Sort

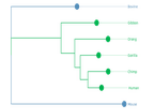
Join, Subtract and Group



GENOMIC FILE MANIPULATION

Convert Formats

FASTA/FASTQ


 **Editor**  
Manually edit text









 **Phylogenetic Tree Visualization**  
A performant, reusable, and extensible tree visualisation library for the web hosted at: <http://biojs.io/d/phylocanvas>.

 **Phyloviz**  
Phylogenetic data analysis from multiple data sources. 






# OrthoFinder: Galaxy


 **Galaxy France**

 Workflow  Visualize  Données partagées  Aide  Utilisateur   

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

**Tools**  



 Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation


Filter and Sort

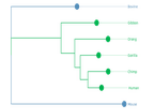
Join, Subtract and Group



GENOMIC FILE MANIPULATION

Convert Formats

FASTA/FASTQ





 **Editor**  
Manually edit text

 **Phylogenetic Tree Visualization**  
A performant, reusable, and extensible tree visualisation library for the web hosted at: <http://biojs.io/d/phylocanvas>.

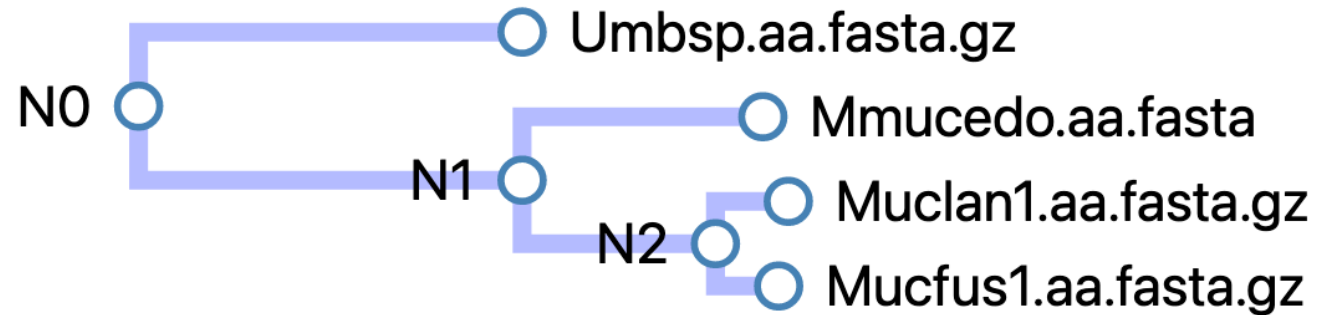
 **Phyloviz**  
Phylogenetic data analysis from multiple data sources. 



# OrthoFinder: Galaxy

**Galaxy France** [Workflow](#) [Visualize](#) [Données partagées](#) [Aide](#) [Utilisateur](#)    

Phylogenetic Tree from OrthoFinder on data 5, data 6, and others: species tree with node labels | Alt+click to select nodes





# OrthoFinder: Galaxy



MycoCosm

THE FUNGAL GENOMICS RESOURCE

[JGI HOME](#)

[GENOME PORTAL](#)

[MYCOCOSM](#)

[PHYCOCOSM](#)

[LOGOUT](#)

[JOHANN JOETS \(JOHANN.JOETS@INRAE.FR\)](#)

TREE

SEARCH

BLAST

ANNOTATIONS

MCL CLUSTERS

DOWNLOAD

INFO

HELP

To Default

Expand All

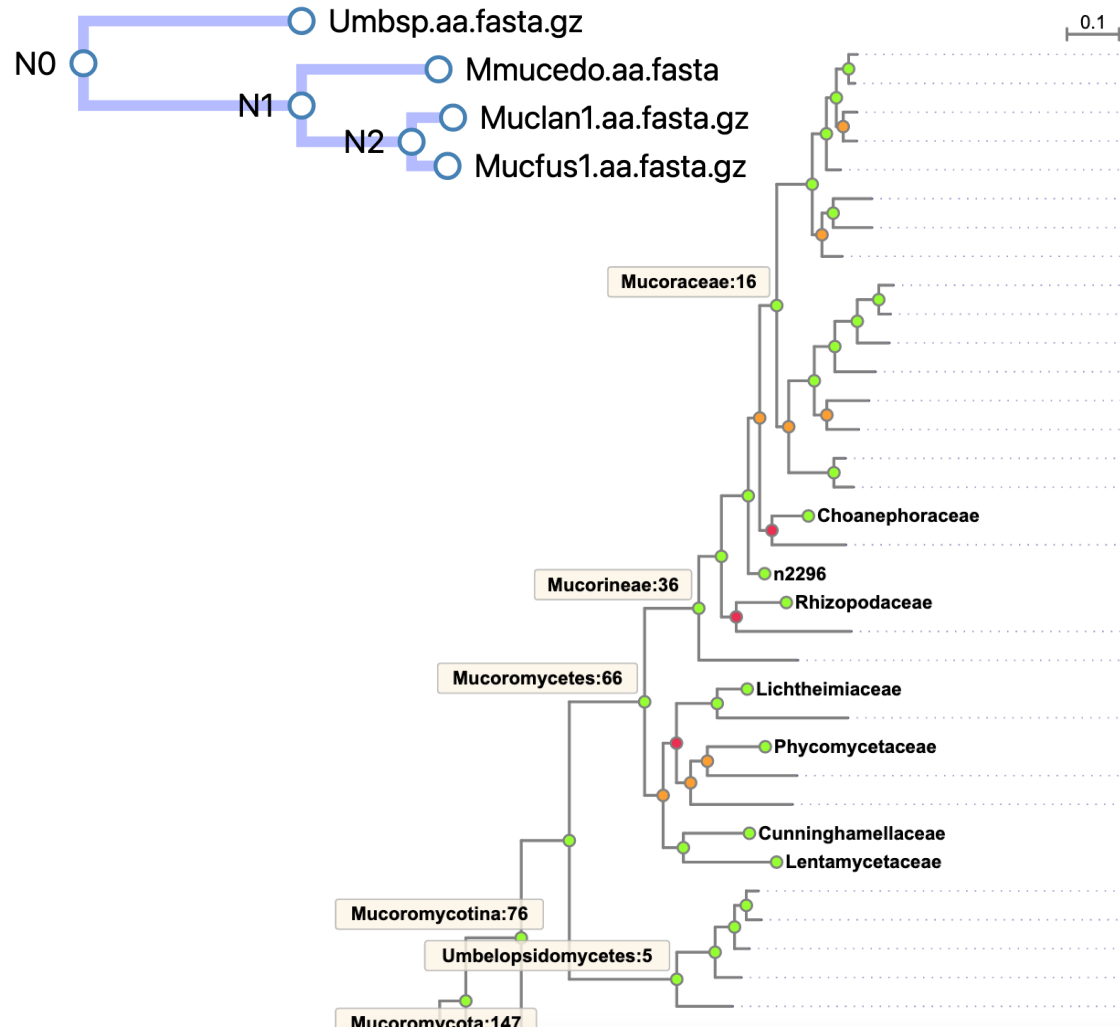
Filter

Clear

To Root

Reload

Support: 1.0 >= .70 < .70 none



Name	Assembly Length (bp)	Genes Count	Published
Mucor lusitanicus CBS277.49 v2.0	36,587,022	11,719	Corrochano LM et al., 2016
Mucor lusitanicus (circinelloides) MU402 v1.0	36,817,399	11,843	Navarro-Mendoza MI et al., 2019
Ellisomyces anomalus CBS 243.57 v2.0	36,869,803	12,251	
Mucor ambiguus NBRC 6742	40,737,472	11,343	Lebreton A et al., 2020
Mucor circinelloides f. circinelloides 1006PhL	36,348,485	12,227	Lee SC et al., 2014
Chaetocladium brefeldii NRRL 2343 v1.0	48,451,839	11,585	
Parasitella parasitica v1.0	32,767,132	12,306	
Mucor racemosus	46,924,235	11,269	Lebreton A et al., 2020
Helicostylum pulchrum RSA 2064 v1.0	32,727,348	12,794	Chang Y et al., 2022
Thamnidium elegans v1.0	34,252,226	13,104	Chang Y et al., 2022
Pilaira anomala RSA1997 v1.0	34,898,111	12,817	Chang Y et al., 2022
Mucor mucedo NRRL 3635 v1.0	46,091,836	14,052	
Mucor endophyticus	34,997,769	11,445	Lebreton A et al., 2020
Mucor heterogamus NRRL 1489 v1.0	54,487,678	14,998	
Mucor fuscus	40,604,092	12,311	Lebreton A et al., 2020
Mucor lanceolatus	43,423,002	10,644	Lebreton A et al., 2020
5 species collapsed, click to expand			
Mucor cordense RSA 1222 v1.0	40,953,002	16,794	
4 species collapsed, click to expand			
8 species collapsed, click to expand			
Pilobolus umbonatus NRRL 6349 v1.0	34,772,712	10,983	Chang Y et al., 2022
Backusella circina FSU 941 v1.0	48,648,288	17,039	Amses KR et al., 2022
10 species collapsed, click to expand			
Syncephalastrum racemosum NRRL 2496 v1.0	30,745,403	11,124	Mondo SJ et al., 2017
8 species collapsed, click to expand			
Saksenaea vasiformis B4078	42,502,389	9,387	Chibucos MC et al., 2016
Radiomyces spectabilis NRRL 2753 v1.0	30,409,402	10,905	
7 species collapsed, click to expand			
2 species collapsed, click to expand			
Umbelopsis ramanniana AG v1.0	23,077,072	9,931	Amses KR et al., 2022
Umbelopsis sp. nov. AD052 v1.0	22,504,321	9,369	Chang Y et al., 2022
Umbelopsis angularis P8C34 v1.0	24,439,702	9,529	
Umbelopsis sp. PM1123 v1.0	24,636,683	9,672	
Umbelopsis isabellina AD026 v1.0	21,826,060	9,193	



# OrthoFinder: Galaxy

Number of species		4
Number of species		4
Number of genes		45727
Number of genes in orthogroups		42977
Number of unassigned genes		2750
Percentage of genes in orthogroups		94.0
Percentage of unassigned genes		6.0
Number of orthogroups		9227
Number of species-specific orthogroups		707
Number of genes in species-specific orthogroups		3857
Percentage of genes in species-specific orthogroups		8.4
Mean orthogroup size		4.7
Median orthogroup size		4.0
G50 (assigned genes)		4
G50 (all genes)		4
O50 (assigned genes)		2851
O50 (all genes)		3195
Number of orthogroups with all species present		5270
Number of single-copy orthogroups		3448
Date		2022-09-28
Orthogroups file		Orthogroups.tsv
Unassigned genes file		Orthogroups_UnassignedGenes.tsv
Per-species statistics		Statistics_PerSpecies.tsv
Overall statistics		Statistics_Overall.tsv
Orthogroups shared between species		Orthogroups_SpeciesOverlaps.tsv
Average number of genes per-species in orthogroup	Number of orthogroups	Percentage of orthogroups
Number of genes	Percentage of genes	
<1	3049	33.0
'1	5328	57.7
'2	573	6.2
'3	141	1.5
'4	63	0.7
'5	20	0.2

History

Rechercher des données
?
x

data
30 shown, 1 deleted, 6202 hidden
352.53 MB

5, data 6, and others: species tree

25: OrthoFinder on data 5, data 6, and others: per species comparative genomics statistics

24: OrthoFinder on data 5, data 6, and others: overall comparative genomics statistics
51 lines, 1 comments
format: tsv, génome de référence: ?
OrthoFinder version 2.5.4 Copyright (C) 2014 David Emms
2022-09-28 04:21:27 : Starting OrthoFinder 2.5.4
1 thread(s) for highly parallel tasks (BLAST searches etc.)
1 thread(s) for OrthoFinder algorithm
Checking required programs are installed
-----

1.Number of species
2.4



# OrthoFinder: Galaxy

	Mmucedo.aa.fasta	Mucfus1.aa.fasta.gz	Muclan1.aa.fasta.gz	Umbasp.aa.fasta.gz
	Mmucedo.aa.fasta	Mucfus1.aa.fasta.gz	Muclan1.aa.fasta.gz	Umbasp.aa.fasta.gz
Number of genes	13403	12311	10644	9369
Number of genes in orthogroups	12626	11987	10268	8096
Number of unassigned genes	777	324	376	1273
Percentage of genes in orthogroups	94.2	97.4	96.5	86.4
Percentage of unassigned genes	5.8	2.6	3.5	13.6
Number of orthogroups containing species	7814	8339	8153	6236
Percentage of orthogroups containing species	84.7	90.4	88.4	67.6
Number of species-specific orthogroups	330	130	58	189
Number of genes in species-specific orthogroups	2454	567	198	638
Percentage of genes in species-specific orthogroups	18.3	4.6	1.9	6.8
Number of genes per-species in orthogroup	Number of orthogroups	Number of orthogroups	Number of orthogroups	Number of orthogroups
'0	1413	888	1074	2991
'1	6076	6705	6686	5002
'2	1119	1110	1114	907
'3	269	241	225	202
'4	97	93	68	66
'5	59	55	22	29
'6	42	34	19	12
'7	22	12	10	3
'8	14	15	2	3
'9	18	6	1	1
'10	6	7	2	3
11-15	35	33	2	5
16-20	24	10	1	3
21-50	27	17	1	0
51-100	5	1	0	0
101-150	1	0	0	0
151-200	0	0	0	0
201-500	0	0	0	0
501-1000	0	0	0	0
'1001+	0	0	0	0
Number of genes per-species in orthogroup	Percentage of orthogroups	Percentage of orthogroups	Percentage of orthogroups	Percentage of orthogroups
'0	15.3	9.6	11.6	32.4
'1	65.9	72.7	72.5	54.2
'2	12.1	12.0	12.1	9.8
'3	2.9	2.6	2.4	2.2
'4	1.1	1.0	0.7	0.7

**History**

Rechercher des données

**data**

30 shown, 1 deleted, 6202 hidden

352.53 MB

**25: OrthoFinder on data 5, data 6, and others: per species comparative genomics statistics**

99 lines, 1 comments  
format: **tsv**, génome de référence: ?

OrthoFinder version 2.5.4 Copyright (C) 2014 David Emms  
  
2022-09-28 04:21:27 : Starting OrthoFinder 2.5.4  
1 thread(s) for highly parallel tasks (BLAST searches etc.)  
1 thread(s) for OrthoFinder algorithm  
  
Checking required programs are installed  
-----

1.	2. Mmucedo
Number of genes	13403
Number of genes in orthogroups	12626
Number of unassigned genes	777
Percentage of genes in orthogroups	94.2

**24: OrthoFinder on data 5, data 6, and others: overall comparative genomics statistics**

**23: OrthoFinder on data 5, data 6, and others: unassigned genes**

**22: OrthoFinder on data 5, data 6, and others: species-specific orthogroups**



# OrthoFinder: Galaxy

Number of genes per-species in orthogroup	Percentage of genes	Percentage of genes	Percentage of genes	Percentage of genes
'0	0.0	0.0	0.0	0.0
'1	45.3	54.5	62.8	53.4
'2	16.7	18.0	20.9	19.4
'3	6.0	5.9	6.3	6.5
'4	2.9	3.0	2.6	2.8
'5	2.2	2.2	1.0	1.5
'6	1.9	1.7	1.1	0.8
'7	1.1	0.7	0.7	0.2
'8	0.8	1.0	0.2	0.3
'9	1.2	0.4	0.1	0.1
'10	0.4	0.6	0.2	0.3
11-15	3.2	3.4	0.2	0.7
16-20	3.2	1.4	0.2	0.5

**25: OrthoFinder on data 5, data 6, and others: perform species comparative genomics statistics**

99 lines, 1 comments

format: **tsv**, génome de référence: ?

OrthoFinder version 2.5.4 Copyright  
(C) 2014 David Emms

2022-09-28 04:21:27 : Starting

OrthoFinder 2.5.4

1 thread(s) for highly parallel tasks  
(BLAST searches etc.)

1 thread(s) for OrthoFinder algorithm



# OrthoFinder: Galaxy

21: OrthoFinder on data 5, data 6, and others: orthogroups (tsv)

9,227 lines, 1 comments  
format: **tsv**, génome de référence: ?

OrthoFinder version 2.5.4 Copyright (C) 2014 David Emms

2022-09-28 04:21:27 : Starting OrthoFinder 2.5.4  
1 thread(s) for highly parallel tasks (BLAST searches etc.)  
1 thread(s) for OrthoFinder algorithm

Checking required programs are installed



1.Orthogroup 2.Mmucedo.aa.fasta	
Orthogroup	Mmucedo.aa.fasta
OG00000000	FUN_000055-T1, FUN_000056-T1
OG00000001	FUN_000167-T1, FUN_000491-T1
OG00000002	FUN_000685-T1, FUN_000919-T1
OG00000003	FUN_002416-T1, FUN_002419-T1

312247 estExt_Genewise1Plus.0	312247 estExt_Genewise1Plus.0
361134 gm1.5473_g	361134 gm1.5473_g
325631 fgenes1_kg.18_#_70_#	325631 fgenes1_kg.18_#_70_#
313267 fgenes1_kg.1_#_570_#	313267 fgenes1_kg.1_#_570_#

Classeur3				
Ouvrir des classeurs récupérés ? Vos modifications récentes ont été enregistrées. Voulez-vous continuer à travailler là où vous vous étiez arrêté ?				
A6042	X	✓	fx	OG0006040
A	B	C	D	
5818 OG0005816	FUN_011186-T1	jgi Mucrus1 369267 Mt11816.m_SAM5U-IIike	jgi Muclan1 1369 SAM5U-IIike	jgi Umbasp_AD052
5819 OG0005817	FUN_011187-T1	jgi Mucfus1 358331 Mf86363.m_Unknown	jgi Muclan1 2915 MI92955.m_Unknown	jgi Umbasp_AD052
5820 OG0005818	FUN_011190-T1	jgi Mucfus1 358334 Guanosine-diphosphatase	jgi Muclan1 2918 Guanosine-diphosphatase	jgi Umbasp_AD052
5821 OG0005819	FUN_011197-T1	jgi Mucfus1 357467 Pirin	jgi Muclan1 3538 Pirin	jgi Umbasp_AD052
5822 OG0005820	FUN_011202-T1	jgi Mucfus1 357463 Signal	jgi Muclan1 3534 Signal	jgi Umbasp_AD052
5823 OG0005821	FUN_011209-T1	jgi Mucfus1 358899 Bloom, jgi Mucfus1 360777 ATP-dependent,	jgi Mucfus1 368146 Mf43602.m_ATP-dependent	
5824 OG0005822	FUN_011213-T1	jgi Mucfus1 368705 Mf96070.m_SH3	jgi Muclan1 7256 MI07161.m_SH3	jgi Umbasp_AD052
5825 OG0005823	FUN_011216-T1	jgi Mucfus1 358055 Gamma-secretase	jgi Muclan1 2952 Gamma-secretase	jgi Umbasp_AD052
5826 OG0005824	FUN_011221-T1	jgi Mucfus1 365074 Mf10441.m_Unknown	jgi Muclan1 7773 MI56041.m_Unknown	jgi Umbasp_AD052
5827 OG0005825	FUN_011222-T1	jgi Mucfus1 366442 Multiprotein-bridging	jgi Muclan1 9299 Multiprotein-bridging	jgi Umbasp_AD052
5828 OG0005826	FUN_011258-T1	jgi Mucfus1 357049 Mf96233.m_Unknown	jgi Muclan1 997 MI64269.m_Unknown	jgi Umbasp_AD052
5829 OG0005827	FUN_011314-T1	jgi Mucfus1 359380 Mf62724.m_Unknown	jgi Muclan1 3091 MI44294.m_Unknown	jgi Umbasp_AD052
5830 OG0005828	FUN_011315-T1	jgi Mucfus1 359378 Poly_rC_-binding	jgi Muclan1 3090 Poly_rC_-binding	jgi Umbasp_AD052
5831 OG0005829	FUN_011316-T1	jgi Mucfus1 366560 Mf17297.m_Unknown	jgi Muclan1 3089 MI49330.m_Unknown	jgi Umbasp_AD052
5832 OG0005830	FUN_011318-T1	jgi Mucfus1 359384 L-lactate	jgi Muclan1 1368 L-lactate	jgi Umbasp_AD052
5833 OG0005831	FUN_011322-T1	jgi Mucfus1 367884 Uracil-regulated	jgi Muclan1 1627 Uracil-regulated	jgi Umbasp_AD052
5834 OG0005832	FUN_011323-T1	jgi Mucfus1 367885 Uracil	jgi Muclan1 1626 Uracil	jgi Umbasp_AD052
5835 OG0005833	FUN_011324-T1	jgi Mucfus1 367887 Mf08238.m_Unknown	jgi Muclan1 1625 MI25463.m_Unknown	jgi Umbasp_AD052
5836 OG0005834	FUN_011331-T1	jgi Mucfus1 367888 Mf25101.m_ATP-dependent	jgi Muclan1 1621 MI24741.m_ATP-dependent	jgi Umbasp_AD052
5837 OG0005835	FUN_011344-T1	jgi Mucfus1 367677 Ubiquitin-conjugating	jgi Muclan1 4994 Ubiquitin-conjugating	jgi Umbasp_AD052
5838 OG0005836	FUN_011346-T1	jgi Mucfus1 360540 Ribosomal	jgi Muclan1 10260 Ribosomal	jgi Umbasp_AD052
5839 OG0005837	FUN_011347-T1	jgi Mucfus1 357212 Cell	jgi Muclan1 4975 Septin	jgi Umbasp_AD052
5840 OG0005838	FUN_011353-T1	jgi Mucfus1 359874 Mf15169.m_Unknown	jgi Muclan1 318 MI56381.m_Unknown	jgi Umbasp_AD052
5841 OG0005839	FUN_011354-T1	jgi Mucfus1 367680 Box	jgi Muclan1 320 Box	jgi Umbasp_AD052
5842 OG0005840	FUN_011356-T1	jgi Mucfus1 367678 Mf26248.m_Unknown	jgi Muclan1 322 MI09600.m_Unknown	jgi Umbasp_AD052
5843 OG0005841	FUN_011384-T1	jgi Mucfus1 365938 Haloacid	jgi Muclan1 5842 Haloacid	jgi Umbasp_AD052
5844 OG0005842	FUN_011386-T1	jgi Mucfus1 365936 Carbamoyl-phosphate	jgi Muclan1 5840 Carbamoyl-phosphate	jgi Umbasp_AD052
5845 OG0005843	FUN_011390-T1	jgi Mucfus1 366173 FKBP12-associated	jgi Muclan1 4245 FKBP12-associated	jgi Umbasp_AD052
5846 OG0005844	FUN_011392-T1	jgi Mucfus1 367157 Cytoplasmic	jgi Muclan1 6507 Cytoplasmic	jgi Umbasp_AD052_1 312247 estExt_Genewise1Plus.0
5847 OG0005845	FUN_011394-T1	jgi Mucfus1 365803 Transcription	jgi Muclan1 7889 Transcription	jgi Umbasp_AD052_1 361134 gm1.5473_g
5848 OG0005846	FUN_011396-T1	jgi Mucfus1 365804 Mf89779.m_Unknown	jgi Muclan1 7890 MI16212.m_Unknown	jgi Umbasp_AD052_1 325631 fgenes1_kg.18_#_70_#
5849 OG0005847	FUN_011442-T1	jgi Mucfus1 359948 NudC	jgi Muclan1 4879 NudC	jgi Umbasp_AD052_1 313267 fgenes1_kg.1_#_570_#



# OrthoFinder: Galaxy



SEARCH BLAST BROWSE ANNOTATIONS ▾ MCL CLUSTERS SYNTENY DOWNLOAD INFO HOME **HELP!**

Search By: Across: Terms:

Protein Id ▾

Default ▾

exact - fast ▾

Download as CSV ▾ compressed by Gzip ▾

Total genes found: 1

Gene	Gene Ontology	Annotations
<p>Portal: <a href="#">Mucfus1</a></p> <p>Portal Name: <b>Mucor fuscus</b></p> <p>Protein Id: <b>362841</b></p> <p>Transcript Id: <a href="#">363093</a></p> <p>Location: <a href="#">scaffold_291:32315-34090 (+)</a></p> <p>Model Name: <b>RNA polymerase II-associated protein 3</b></p> <p>Track: <b>ExternalModels</b></p>	<p><a href="#">GO:0005515</a> • protein binding</p>	<p><b>KOG4648</b> • Uncharacterized conserved protein, contains LRR repeats</p> <p><a href="#">PF07719</a> • Tetratricopeptide repeat</p> <p><a href="#">PF13877</a> • Potential Monad-binding region of RPAP3</p> <p><a href="#">PF13414</a> • TPR repeat</p> <p><a href="#">IPR019734</a> •</p> <p><a href="#">IPR025986</a> •</p> <p><a href="#">IPR013105</a> •</p> <p><a href="#">IPR013026</a> •</p>



History ↺ + 🗂 ⚙

◀ Back to data

**OrthoFinder on data 5, data 6, and others: resolved gene trees**  
a list with 6178 items

🚨 displaying only 1000 of 6178 items



**OG0000000\_tree**



1 line

format: **newick**, génome de référence: ?



```
(Mmucedo_aa_fasta_FUN_011173-T1:0.138041,(C266)n7:0.0582738)n5:0.0208431,((Mmucedo_aa_013360-T1:0.220884,Mmucedo_aa_fasta_FUN_00Mmucedo_aa_fasta_FUN_013109-T1:0.320795,Mmfastra_FUN_013196-T1:0.253951)n23:0.0476717
```

**OG0000001\_tree**



**InterPro** - Member

Classification of protein families

Home ▶ Search ▶ **Browse** ▶ Results ▶ Release notes ▶ Download ▶ Help ▶ About

🏠 / Browse / By Entry / Panther / PTHR46423 / Overview



**RNA POLYMERASE II-ASSOCIATED PROTEIN 3**

PTHR46423

PANTHER entry

**Overview**

Proteins 4k  
Taxonomy 6k  
Proteomes 1k  
Structures 10  
Signature

Member database

PANTHER

PANTHER type

Family

✎ Add your annotation

External Links

[View PTHR46423 in PANTHER](#)



# OrthoFinder: Galaxy

October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Orthogroup	Mmucedo.aa.fasta	Mucfus1.aa.fasta.gz	Muclan1.aa.fasta.gz
Orthogroup	Mmucedo.aa.fasta	Mucfus1.aa.fasta.gz	Muclan1.aa.fasta.gz
OG0009227	FUN_000003-T1		
OG0009228	FUN_000045-T1		
OG0009229	FUN_000085-T1		
OG0009230	FUN_000089-T1		
OG0009231	FUN_000101-T1		
OG0009232	FUN_000124-T1		
OG0009233	FUN_000125-T1		
OG0009234	FUN_000126-T1		
OG0009235	FUN_000131-T1		
OG0009236	FUN_000132-T1		
OG0009237	FUN_000133-T1		
OG0009238	FUN_000134-T1		
OG0009239	FUN_000135-T1		
OG0009240	FUN_000152-T1		
OG0009241	FUN_000157-T1		
OG0009242	FUN_000159-T1		
OG0009243	FUN_000176-T1		
OG0009244	FUN_000180-T1		
OG0009245	FUN_000187-T1		
OG0009246	FUN_000243-T1		
OG0009247	FUN_000254-T1		
OG0009248	FUN_000290-T1		
OG0009249	FUN_000292-T1		
OG0009250	FUN_000326-T1		
OG0009251	FUN_000337-T1		
OG0009252	FUN_000340-T1		
OG0009253	FUN_000450-T1		
OG0009254	FUN_000474-T1		
OG0009255	FUN_000498-T1		
OG0009256	FUN_000515-T1		

## History



Rechercher des données ? x

### data

30 shown, 1 deleted, 6202 hidden

352.53 MB



25: OrthoFinder on data 5, data 6, and others: per species comparative genomics statistics



24: OrthoFinder on data 5, data 6, and others: overall comparative genomics statistics



23: OrthoFinder on data 5, data 6, and others: unassigned genes



2,750 lines, 1 comments

format: tsv, génome de référence: ?

OrthoFinder version 2.5.4 Copyright (C) 2014 David Emms

2022-09-28 04:21:27 : Starting OrthoFinder 2.5.4  
1 thread(s) for highly parallel tasks (BLAST searches etc.)  
1 thread(s) for OrthoFinder algorithm

Checking required programs are installed

