

Exploring protein domains and clusters across species in Ensembl and MycoCosm

We're going to use the HMMER tool, which is embedded in Ensembl Fungi with the Ensembl Fungi species as reference proteomes and standard search parameters. You can use the full tool here: <https://www.ebi.ac.uk/Tools/hmmer/search/phmmer> if you want to work with other species and set your own thresholds. This tool uses protein sequences as input and finds domains within them and maps against all species, like a more powerful, domain oriented BLAST. We'll then explore this at a finer scale looking at a handful of closely related species in the MycoCosm Cluster tool.

In the BioMart demo, we found genes that were associated with reduced virulence in *Fusarium solani*, and which did not have an orthologue in *F. oxysporum*. We're going to explore one of these genes in more detail: *PEP2* or *NechaG73962*.

- a) Search *Fusarium solani* for *NechaG73962* at fungi.ensembl.org. Navigate to the **Transcript** tab and either export the protein sequence in FASTA format or highlight and copy it.
- b) Using the link in the Ensembl Fungi header navigate to the HMMER tool. Paste in your sequence and click **Submit**.
 - I. What is the PFAM domain identified in this sequence?
 - II. Hover over the domain image, what is the length of the aligned region with our submitted protein sequence?
- c) In the Significant Query Matches table at the bottom of the page, click on the black Customise button and add 'Phylum' to the table.
 - I. To which Phylum do the top hits belong to?
 - II. Which species is reported as the top hit? Is this the same as *Fusarium solani* (think about the reproductive cycle...)?
- d) We can explore the taxonomy more broadly elsewhere. Click on the Taxonomy tab just above the domain image.
 - I. How many hits were there in the Basidiomycota?
 - II. Click to expand the Agaricomycetes node by clicking on the arrow, and then the Agaricales. Which families are represented?

NOTE: You may need to click on the node name (e.g. Agaricales), to reposition the image.
- e) Let's explore this protein domain further using the JGI MycoCosm cluster tool. This will enable us to compare the presence genes containing this domain across a closely related group. Navigate to mycocosm.jgi.doe.gov and search for *Fusarium solani*. Select *Fusarium solani* FSSC 5 v1.0, then click on the **MCL clusters** option at the top of the page. Search for the protein domain we identified, *SnoaL_4*.
 - I. For the first cluster, 4,213, which species is missing any hits?
 - II. Click on the link for the domain (4213) to explore in more detail. We can see from the donut plots that only 4 species contain this SnoaL-like domain.
 - III. Which species have the most similar protein lengths, and contain the SnoaL-like domain?

- f) Click on Synteny in the final column.
- Are the clusters around this conserved between these species? Click on them to highlight the occurrence in the image.
 - Which other species has the most similar synteny to *Fusarium solani* (Fusso1) protein 495533? Why do you think this is?

Answers

Exploring protein domains and clusters across species in Ensembl and MycoCosm

We're going to use the HMMER tool, which is embedded in Ensembl Fungi with the Ensembl Fungi species as reference proteomes and standard search parameters. You can use the full tool here: <https://www.ebi.ac.uk/Tools/hmmer/search/phmmer> if you want to work with other species and set your own thresholds. This tool uses protein sequences as input and finds domains within them and maps against all species, like a more powerful, domain oriented BLAST. We'll then explore this at a finer scale looking at a handful of closely related species in the MycoCosm Cluster tool.

In the BioMart demo, we found genes that were associated with reduced virulence in *Fusarium solani*, and which did not have an orthologue in *F. oxysporum*. We're going to explore one of these genes in more detail: *PEP2* or NechaG73962.

- g) Search *Fusarium solani* for NechaG73962 at fungi.ensembl.org. Navigate to the Transcript tab and either export the protein sequence in FASTA format, or highlight and copy it.

Answer: Go to fungi.ensembl.org. From the homepage select *Fusarium solani* from the drop-down list and type in NechaG73962. Hit Go.



Search: for

e.g. NAT2 or alcohol*

Click on the gene name hyperlink on the results page, this will take you to the gene tab. Click on the transcript tab [Transcript: NechaT73962](#) to go to the transcript tab.

Transcript tab

Transcript: NechaT73962

Description
Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:[C7ZC16.6](#)]

Location
[Chromosome 14: 1,141,191-1,142,037](#) reverse strand.

About this transcript
This transcript has [2 exons](#) and is annotated with [4 domains and features](#).

Gene
This transcript is a product of gene [NechaG73962](#) [Show transcript table](#)

On the left-hand navigation panel there is a link for [Protein](#) under the Sequence header. Highlight the protein sequence and copy it.

View protein sequence

Transcript: NechaT73962

About this transcript
Gene
This transcript is a product of gene [NechaG73962](#) [Show transcript table](#)

Protein sequence

[Download sequence](#) [BLAST this sequence](#)

Exons An exon Another exon Residue overlaps splice site

Markup loaded

• Variants are filtered by consequence type

Copy protein sequence

MVNLHSLPQGSRPNAAIRNNGPDSLALERLKLRELAEGWPSYRDSCEMENFESIFHPGAY
VYTTWGRVAYQDFIAASKAGMDKGAFIGHRCHGSSDINVDGTRAVTKLKATITQRFEV
GGSEFDVEADCRFCFYFEKINGSWGARLVKHWEKDRMIPVNPAPFPQVDEDKLKAYPPG
YKYLAYWQETAMGIRVLLDMPGHRHVGTVNLEKHDELYNLAKRMLEGEQIEV

Using the link in the Ensembl Fungi header navigate to the HMMER tool. Paste in your sequence and click [Submit](#).

EnsemblFungi | HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog | Login/Register

Search Ensembl Fungi...

Shortcut to HMMER

HMMER

phmmer

protein sequence vs protein sequence database

Paste in your sequence or use the [example](#)

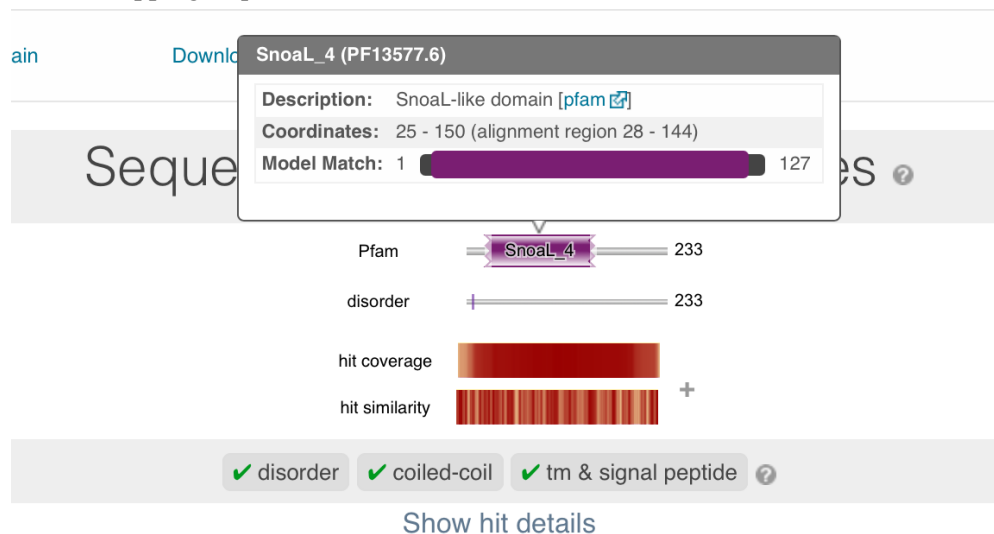
```

HVNGLSLPQGRPHAAIRNNGPDLALERLKLRLAGWPTDSCENEFESIFHNGAY
VYTTWGRVATGDFIAASRAGHDEGAFIRHCHGSDTINVESTAVTEKATITURFEV
GGEFQVEADCRFCFTEFIRNGHARLVKINTYERDHIIVSPPAFQVSDLEKATFFS
YRTLATWGTAMGIRYLLQHPGRIRFQTVLEKIDELTWLAKNGEGQIEY
  
```




Submit Clear

- I. What is the PFAM domain identified in this sequence?
- II. Hover over the domain image, what is the length of the aligned region with our submitted protein sequence?

Answers: The image shown in the centre middle of the page shows the domain (or domains) matched in your sequence. Hovering over the domain will give you some summary information, including the length of the overlapping sequence.



- h) In the Significant Query Matches table at the bottom of the page, click on the black Customise button and add 'Phylum' to the table.

Customise		
Species	Cross-references	E-value
Nectria haematococca (strain 77-13-4 / ATCC MYA-4622 / FGSC 9596 / MPVI)	  	2.6e-163

Customise Results

Select Visible Columns

- ☐ Row Count
- ☐ Secondary Accessions and Ids
- ☒ Description
- ☒ Species
- ☒ Cross-references
- ☐ Kingdom
- ☒ Phylum
- ☐ Known Structure
- ☐ Identical Seqs
- ☐ Number of Hits
- ☐ Number of Significant Hits
- ☐ Bit Score
- ☐ Hit Positions

Rows Per Page

- ☒ 50
- ☐ 100
- ☐ 250
- ☐ 1000
- ☐ 2500












Update
Restore Defaults

I. To which Phylum do the top hits belong to?

Answer: We can see that the column of the first hits are all listed as ‘Ascomycota’

II. Which species is reported as the top hit? Is this the same as *Fusarium solani* (think about the reproductive cycle...)?

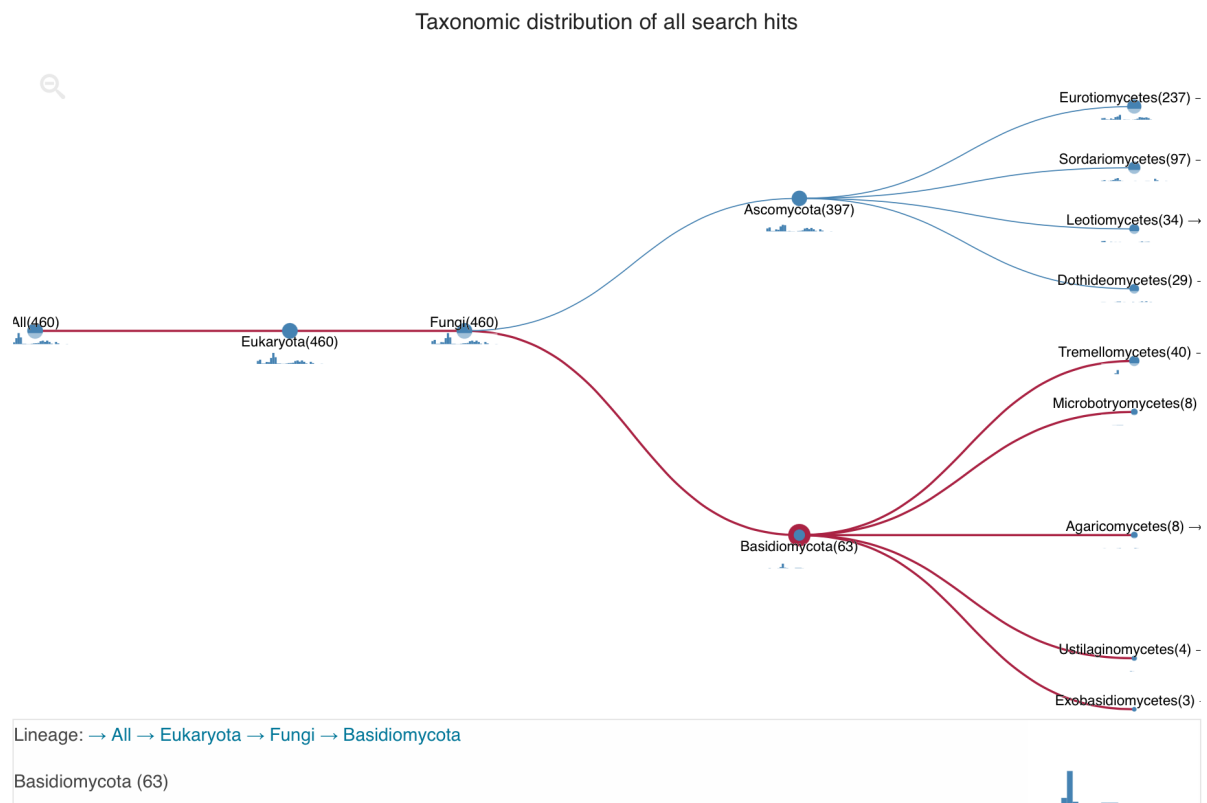
Answer: The sexual form (teleomorph) of *Fusarium solani* (the anamorph) is *Nectria haematococca*.

Significant Query Matches (460) in <i>ensemblgenomes</i> (v.44)						Customise
Target	Description	Phylum	Species	Cross-references	E-value	
> NechaG73962	Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:C7ZC16]	Ascomycota	Nectria haematococca (strain 77-13-4 / ATCC MYA-4622 / FGSC 9596 / MPVI)	  	2.6e-163	
> LW93_4799	Uncharacterized protein	Ascomycota	Gibberella fujikuroi	 	1.6e-137	
> FFB14_04603	Pea pathogenicity protein 2	Ascomycota	Fusarium fujikuroi (GCA_900096505)	 	2.1e-137	
> AU210_001920	hypothetical protein	Ascomycota	Fusarium oxysporum f. sp. radicis-cucumerinum	 	6.2e-137	
> FOWG_10080	pea pathogenicity protein 2	Ascomycota	Fusarium oxysporum f. sp. lycopersici MN25 (GCA_000259975)	 	6.2e-137	

i) We can explore the taxonomy more broadly elsewhere. Click on the Taxonomy tab just above the domain image.

I. How many hits were there in the Basidiomycota?

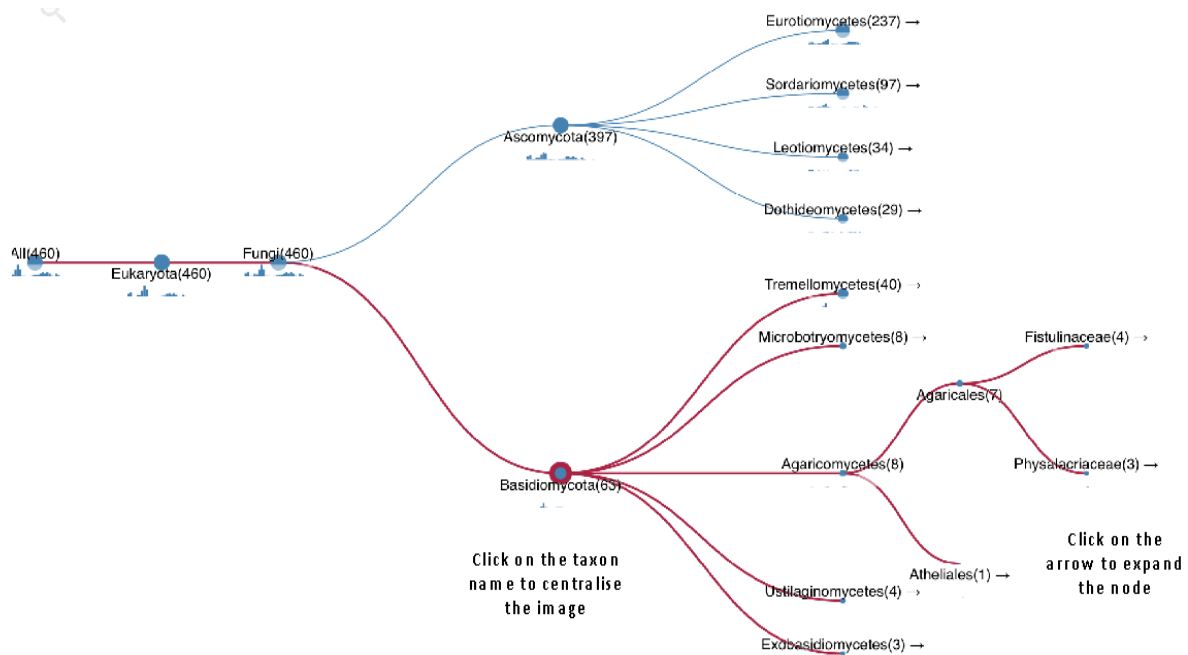
Answer: We can see from the number in the parentheses that there are 63 hits.



II. Click to expand the Agaricomycetes node by clicking on the arrow, and then the Agaricales. Which families are represented?

Answer: Fistulinaceae and Physalacriaceae families are shown here with 4 and 3 members respectively.

NOTE: You may need to click on the node name (e.g. Agaricales), to reposition the image.



- j) Let's explore this protein domain further using the JGI MycoCosm cluster tool. This will enable us to compare the presence genes containing this domain across a closely related group. Navigate to mycocosm.jgi.doe.gov and search for *Fusarium solani*. Select *Fusarium solani* FSSC 5 v1.0, then click on the MCL clusters option at the top of the page. Search for the protein domain we identified, SnoaL_4.

[JGI HOME](#)
[GENOME PORTAL](#)
[MYCOCOSM](#)
[LOGIN](#)

[SEARCH](#)
[BLAST](#)
[BROWSE](#)
[ANNOTATIONS](#)
[MCL CLUSTERS](#)
[SYNTENY](#)
[DOWNLOAD](#)
[INFO](#)
[HOME](#)
[STATUS](#)
[HELP](#)

Run **Fusso1 comparative clustering.2371**

Multigene clusters: **15,016**
Average multigene cluster size: **10.00**
Created at: **30-Mar-2018**

Models in multigene clusters: **150,173**
Singletons: **6,116**
Tracks: **9**

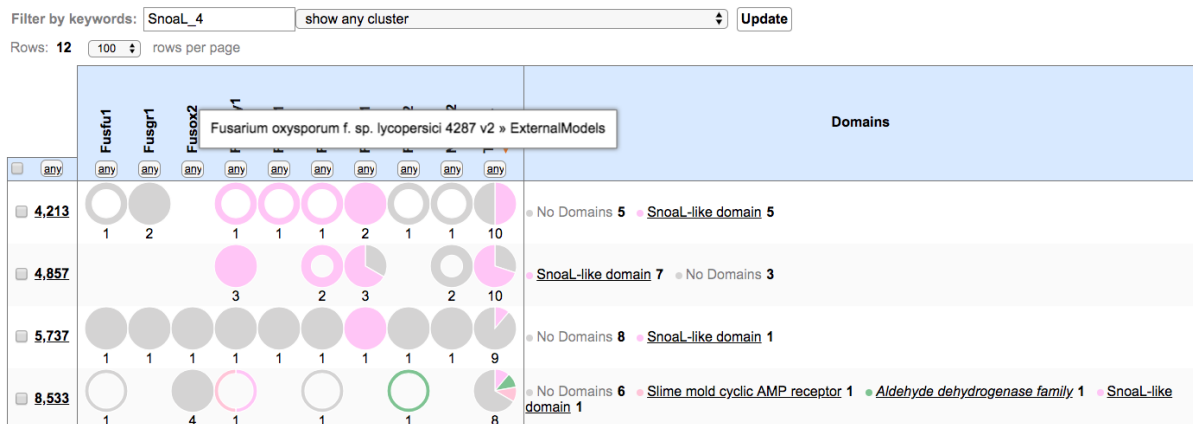
Show Charts: ☒
Show Counters: ☒
Show Domains: ☒

Download as clusters CSV compressed by Gzip

Filter by keywords: SnoaL_4 show any cluster Update

I. For the first cluster, 4,213, which species is missing any hits?

Answer: There is no 'donut' in the first row for the species Fusox2. Hover over the name or look at the list below the table to see what this species/assembly full name is, it is *Fusarium oxysporum* f. sp. lycopersici 4287 v2 ExternalModels.



II. Click on the link for the domain (4213) to explore in more detail. We can see from the donut plots that only 4 species contain this SnoL-like domain.

Answer: The pink colour corresponds to the SnoL-like domain.

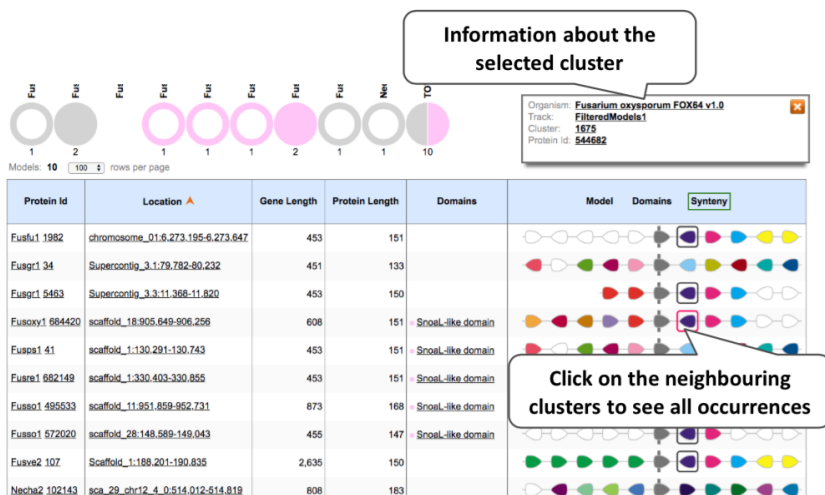
III. Which species have the most similar protein lengths, and contain the SnoL-like domain?

Protein Id	Location	Gene Length	Protein Length	Domains	Model	Domains	Synteny
Fusfu1 1982	chromosome_01:6,273.195-6,273.647	453	151				
Fusgr1 34	Supercontig_3:1:79,782-80,232	451	133				
Fusgr1 5463	Supercontig_3:3:11,368-11,820	453	150				
Fusox1 684420	scaffold_18:905,649-906,256	608	151	SnoL-like domain			
Fusps1 41	scaffold_1:130,291-130,743	453	151	SnoL-like domain			
Fusre1 682149	scaffold_1:330,403-330,855	453	151	SnoL-like domain			
Fusso1 495533	scaffold_11:951,859-952,731	873	168	SnoL-like domain			
Fusso1 572020	scaffold_28:148,589-149,043	455	147	SnoL-like domain			
Fusve2 107	Scaffold_1:188,201-190,835	2,635	150				
Necha2 102143	sca_29_chr12_4_0:514,012-514,819	808	183				

These three have the same protein length

k) Click on Synteny in the final column.

I. Are the clusters around this conserved between these species? Click on them to highlight the occurrence in the image.



II. Which other species has the most similar synteny to *Fusarium solani* (Fusso1) protein 495533? Why do you think this is?

Answer: *Nectria haematococca* v2.0 FilteredModels1. We know this to be the sexual form of *F. solani* so this is expected.