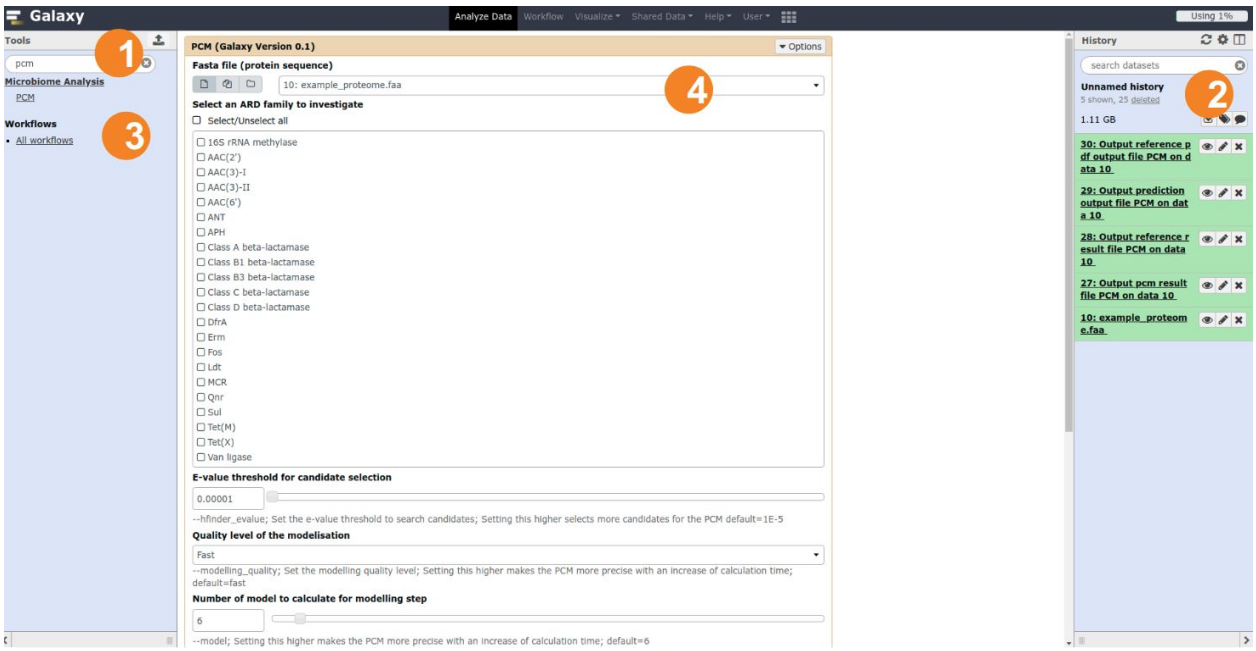
Usage of PCM in galaxy

PCM is deployed on galaxy.pasteur.fr in section microbiome_analysis/PCM (3) as follow:



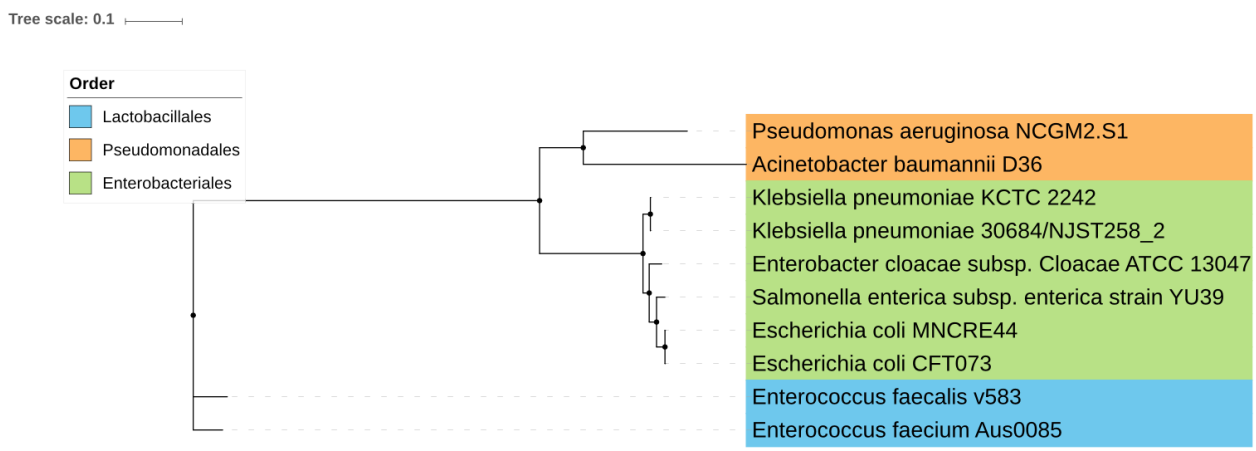
First, a fasta file with protein sequence need to be loaded in an history to be available for the workflow (1).

The file must be in the following format:

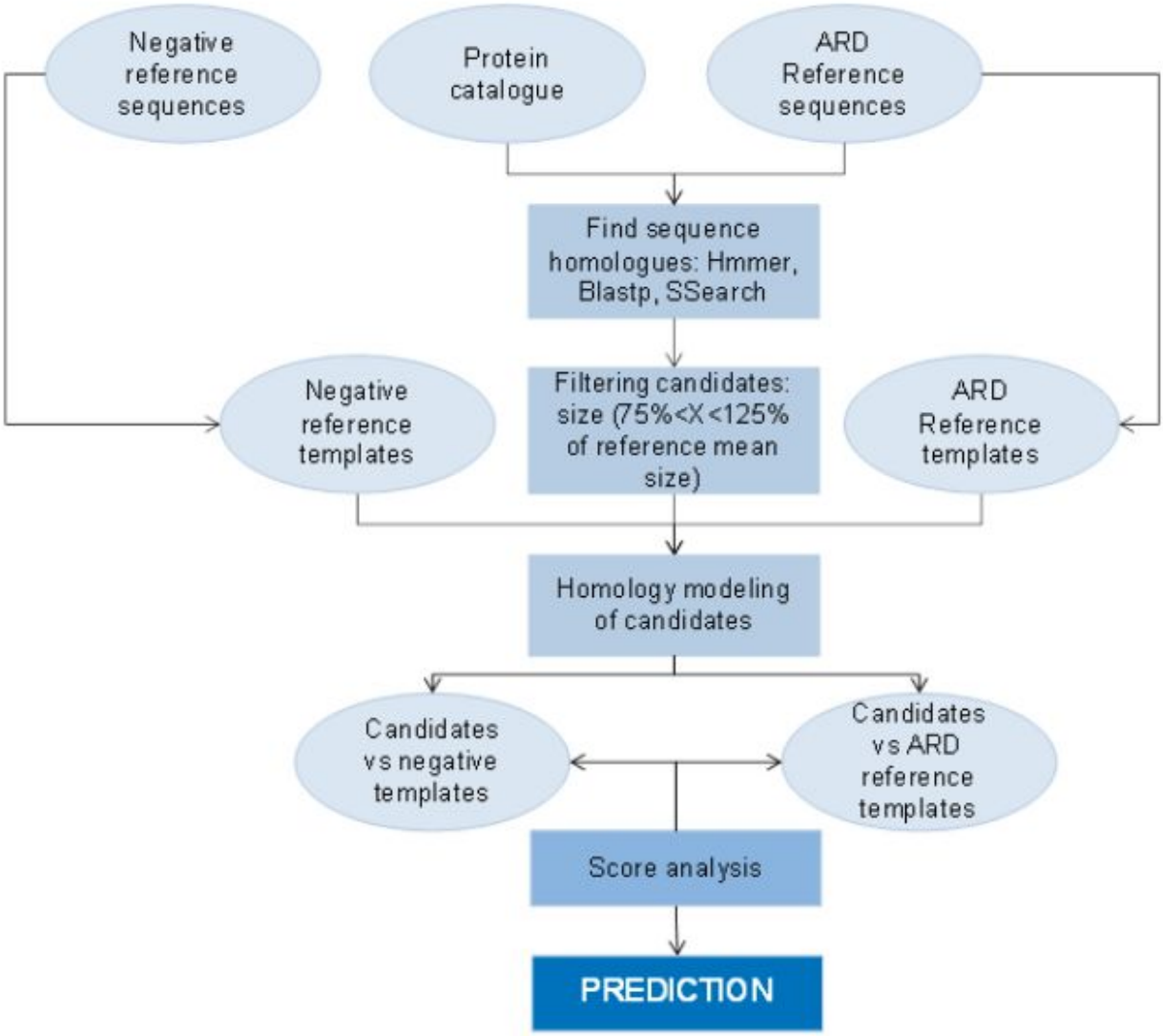
```
>ID1
MNTFGQIHNNMPYLFLLAFIMNFYDQFNNSISGQEMCYEVESI
FNNHQVDIIGAPAAAFKPLELQKGLGTKGAIVNYPILQVTGNI
>ID2
MNTFGQIHNNMPYLFLLAFIMNFYDQFNNSISGQEMCYEVESI
FNNHQVDIIGAPAAAFKPLELQKGLGTKGAIVNYPILQVTGNI
```

The ID is >name (without any space or points).

The example data appears in the history (2). We provided for the example: example_proteome (https://github.com/aghozlane/pcm/blob/master/example/example_proteome.faa).
This dataset correspond to the proteome of the following species:



For this example we will search resistance genes in the proteome with the workflow of PCM:



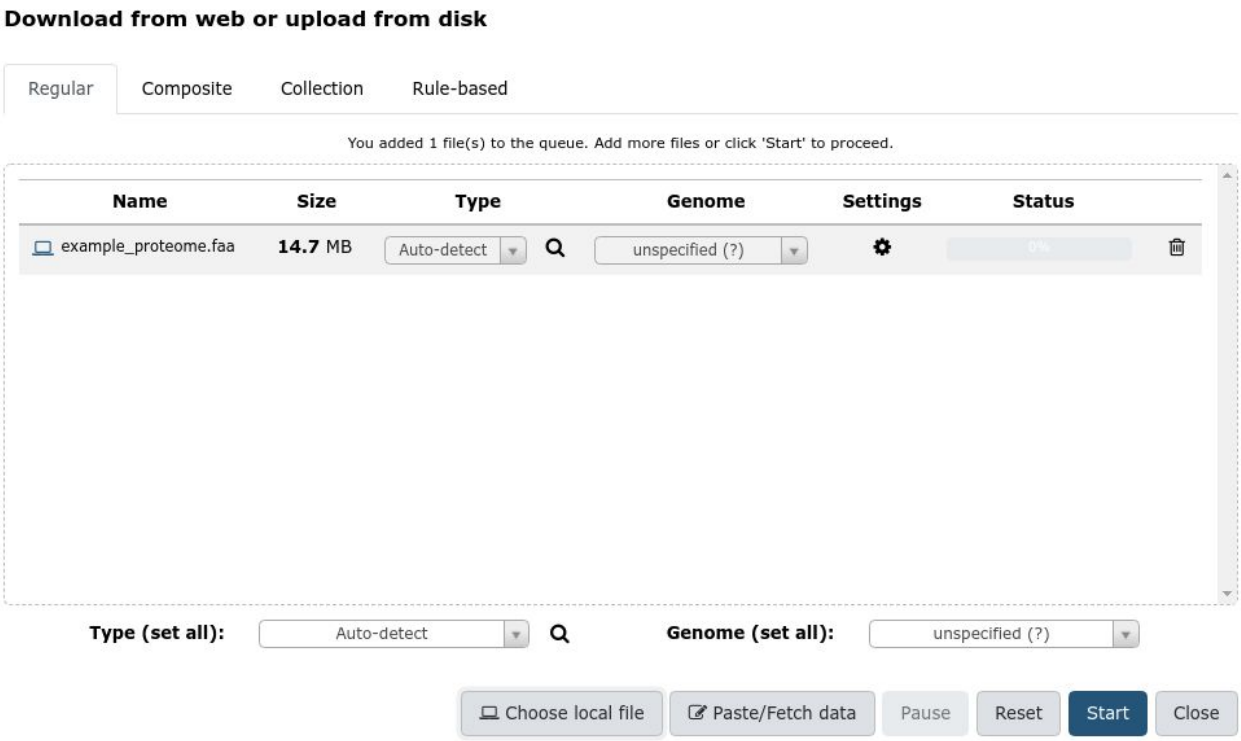
Several parameters can be modified:

- One or several ARD families can be investigated. By default, PCM allows to screen the ARD families of clinical interest.
- Set the E-value threshold for candidate selection. This threshold impacts the candidates selection. A high e-value will select more candidates for PCM.
- Select the quality level of the modelisation. A high level of modelling quality will make PCM to spend more time to improve the protein model. By default, the mode fast is enough to produce a carbon-alpha model to predict an ARD.
- Select the number of model to calculate for modelling step. The homology modelling is a heuristic approach where the starting point is crucial. A high number of model corresponds to a high number of different start possible. By default, the number of model is set to 6 which is enough to screen the set of possibilities.
- Select the number of template to consider for modelling step. Setting this parameter higher makes the PCM more precise with a risk of overfitting. However some ARDs families do not have more than 3 ARD templates (notably MCR).
- Select the number of bootstrap to calculate for candidate classification. Setting this higher makes the classification more precise with an increase of calculation time.

Step by step:

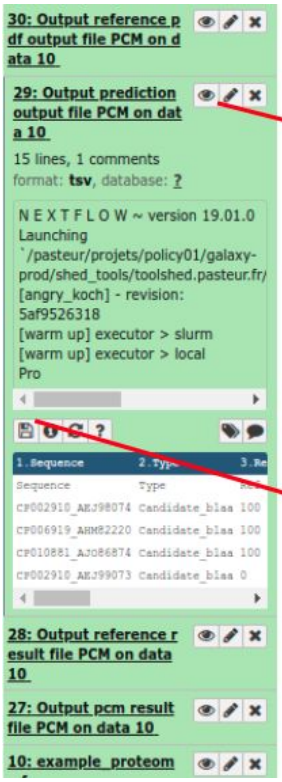
1. Load the data:

Click on upload symbol  , then choose local files and finally start upload



2. Adjust your parameters set

3. Execute
4. Results are available when the right panel turn green. They can be visualised inside galaxy or downloaded (one by one).



Sequence	Type	Ref	tneg	Adjusted p.value of being a Ref	TM.score_TMalign_ref	TM.score_TMalign_tneg	Prediction
Sequence	Type	Ref	tneg	Adjusted p.value of being a Ref	TM.score_TMalign_ref	TM.score_TMalign_tneg	Prediction
CP002910_AEJ98074	Candidate_blaa	100	0	3.46895869915534e-23	0.99744	0.81893	Very_likely_Ref
CP006919_AHM82220	Candidate_blaa	100	0	3.46895869915534e-23	0.98559	0.69116	Very_likely_Ref
CP010881_AJO86874	Candidate_blaa	100	0	3.46895869915534e-23	0.98686	0.86876	Very_likely_Ref
CP002910_AEJ99073	Candidate_blaa	0	100	1	0.91786	0.88424	Not_a_Ref
CP006918_AHM78534	Candidate_blaa	0	100	1	0.91917	0.88344	Not_a_Ref
CP010881_AJO86812	Candidate_blaa	100	0	3.46895869915534e-23	0.99862	0.88749	Very_likely_Ref
CP006918_AHM79651	Candidate_blaa	100	0	3.46895869915534e-23	0.99345	0.81105	Very_likely_Ref
AP012280_BAK88510	Candidate_blaa	0	100	1	0.88029	0.90219	Not_a_Ref
AE014075_AAN81121	Candidate_blaa	0	100	1	0.81273	0.87495	Not_a_Ref
CP010882_AJO86990	Candidate_blaa	100	0	3.46895869915534e-23	0.99862	0.88749	Very_likely_Ref
AE016830_AAO81293	Candidate_blaa	100	0	3.46895869915534e-23	0.85035	0.61575	Likely_Ref
CP010876_AJO84151	Candidate_blaa	0	100	1	0.83077	0.88323	Not_a_Ref
CP010881_AJO86866	Candidate_blaa	100	0	3.46895869915534e-23	0.99831	0.86279	Very_likely_Ref
CP001918_ADF62979	Candidate_blaa	0	100	1	0.84524	0.86563	Not_a_Ref
CP010881_AJO86821	Candidate_blaa	100	0	3.46895869915534e-23	0.98686	0.86876	Very_likely_Ref