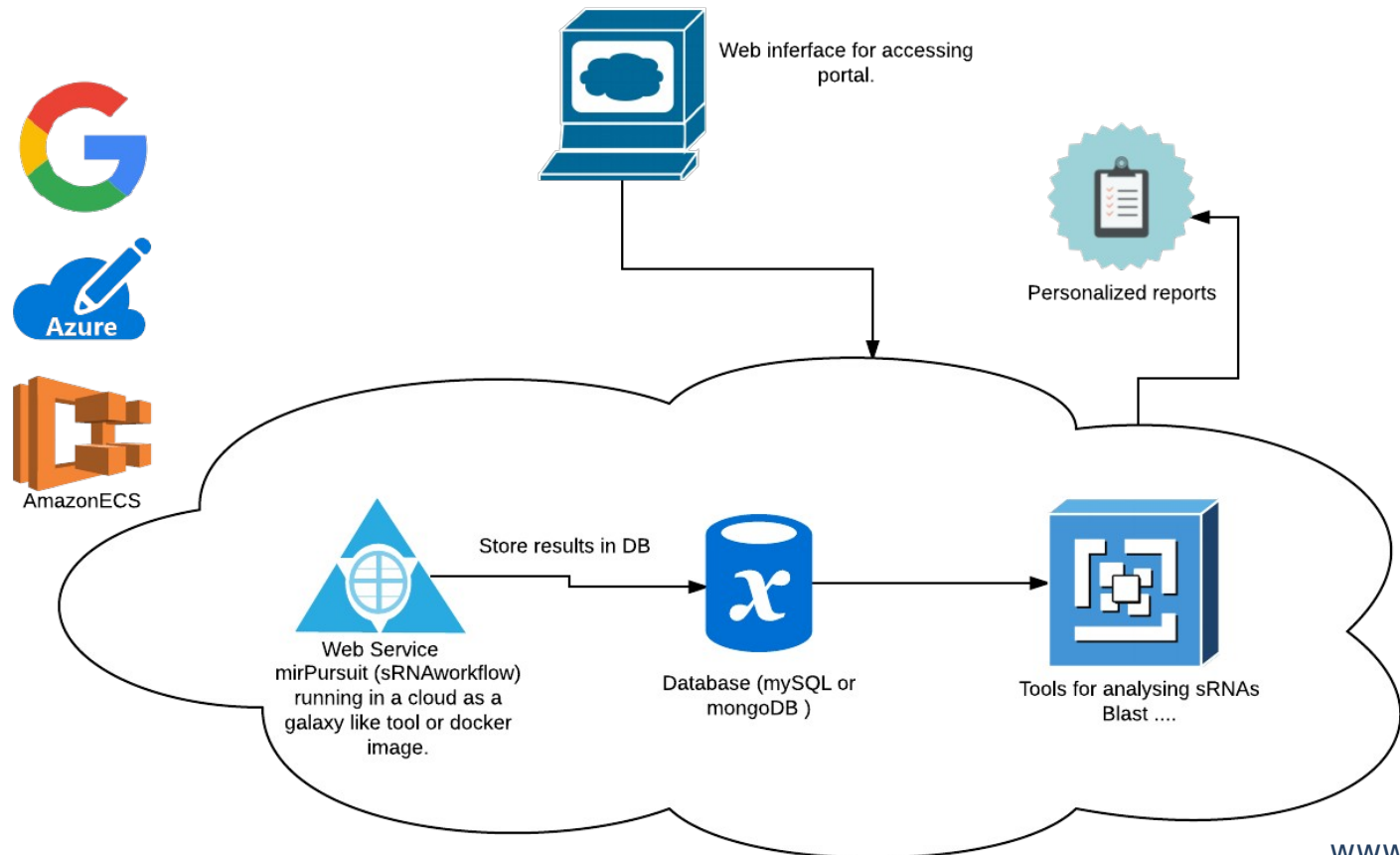# Plant small RNA portal

Bruno Costa

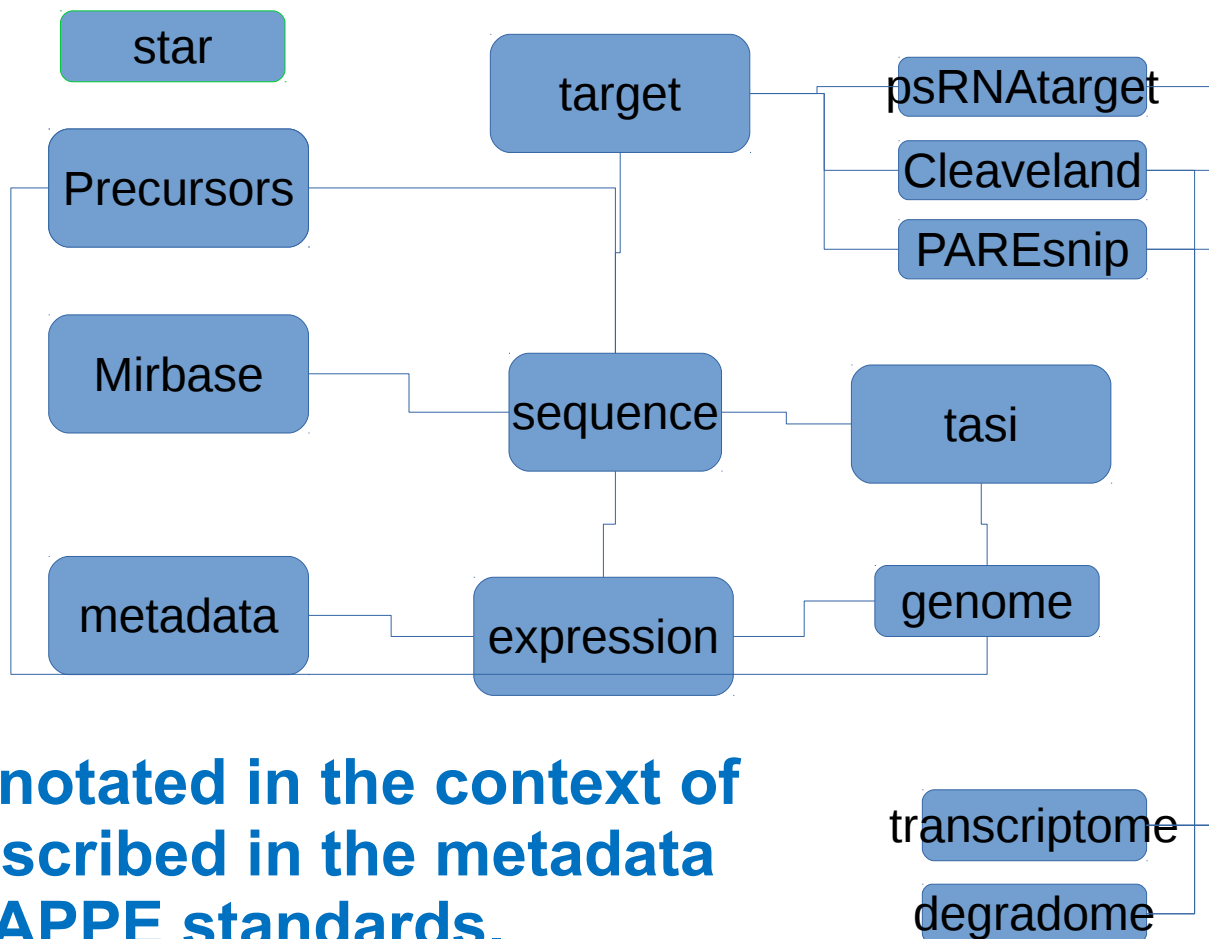*iBET – Instituto de Biologia Experimental e Tecnológica*

www.ibet.pt

Plant sRNA Portal

The plant small RNA (sRNA) portal a web interface for a sRNA database and related tools for producing analyzing and classifying sRNAs
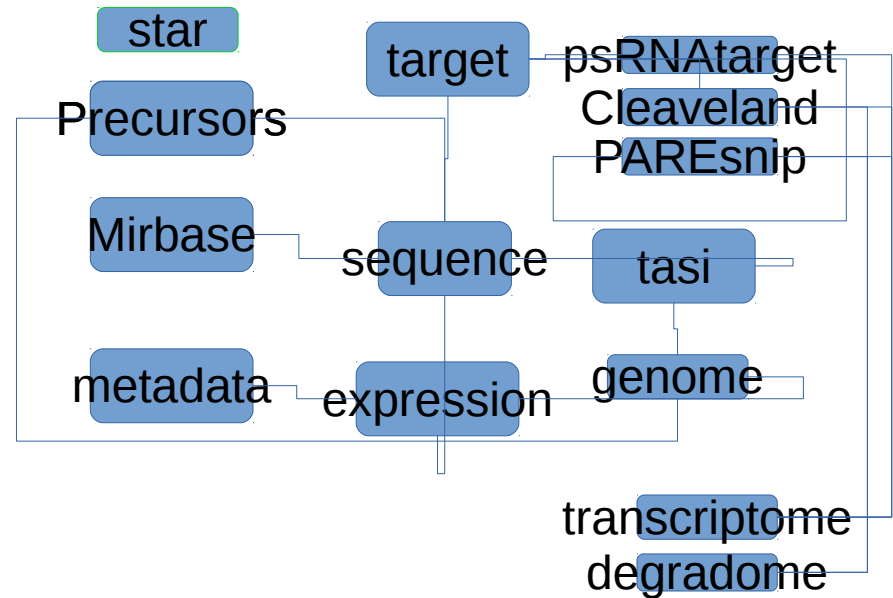
Centered around each sequence

Sequences are annotated in the context of the experiment described in the metadata table based on MIAPPE standards.

Importing data to database

Web interface that annotates data based on MIAPPE standards. Datasets are identified by the metadata but the granularity of the database is the sRNA sequence.

miRPursuit
Raw data can be processed and imported directly to the database where more precise metadata will be kept on the annotation process.

Customizable tool for sRNA analysis

Import raw data to be processed on the cloud and have your data saved automatically in the database.
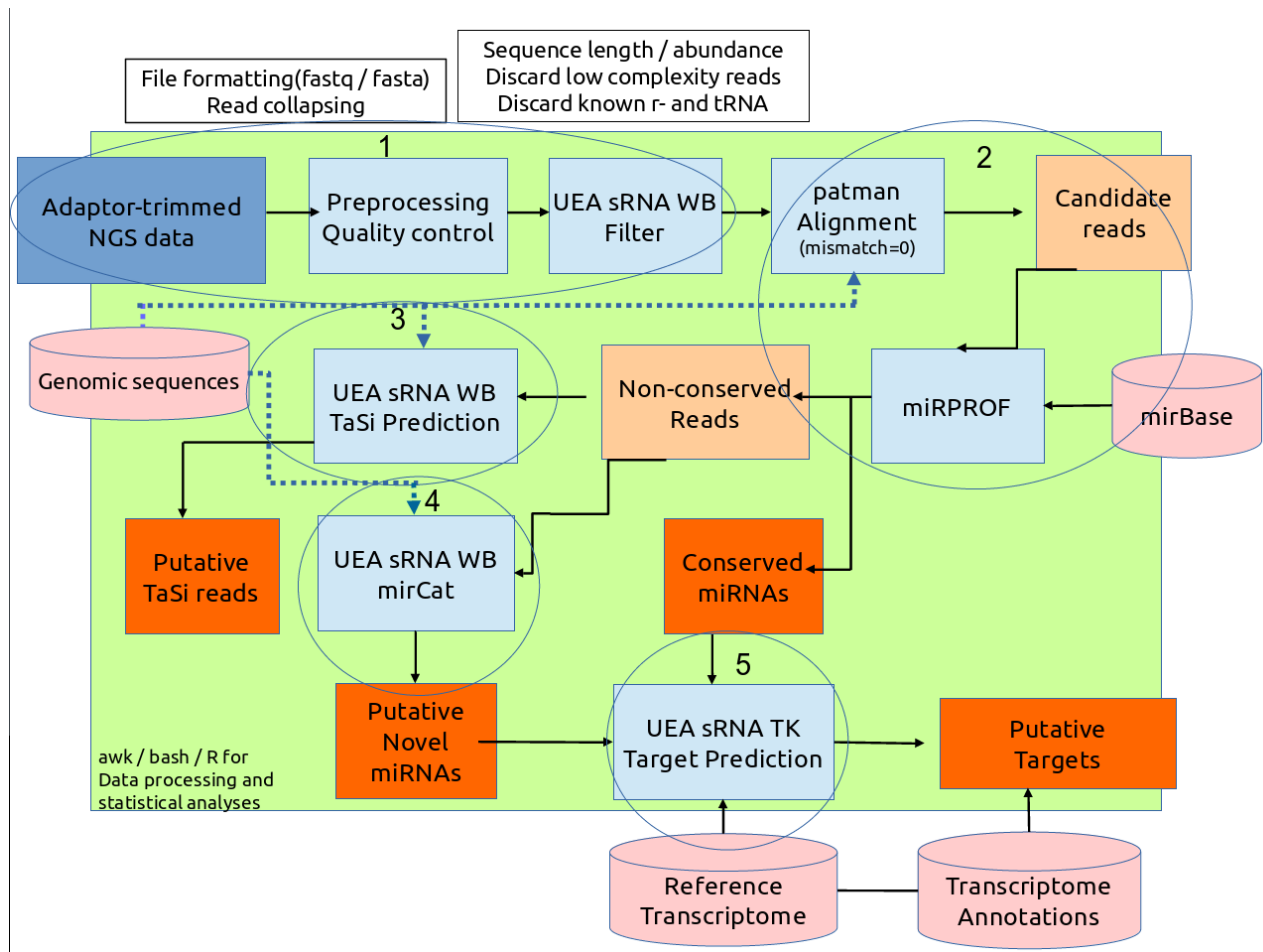
Confidentiality options

Private data should have confidentiality options in order to only be visible to certain people while not published.

Most data in DB is visible to unauthenticated users.
But private data should require an authentication.

**For greater security private data should be in different database with stricter permissions but same structure.**

6
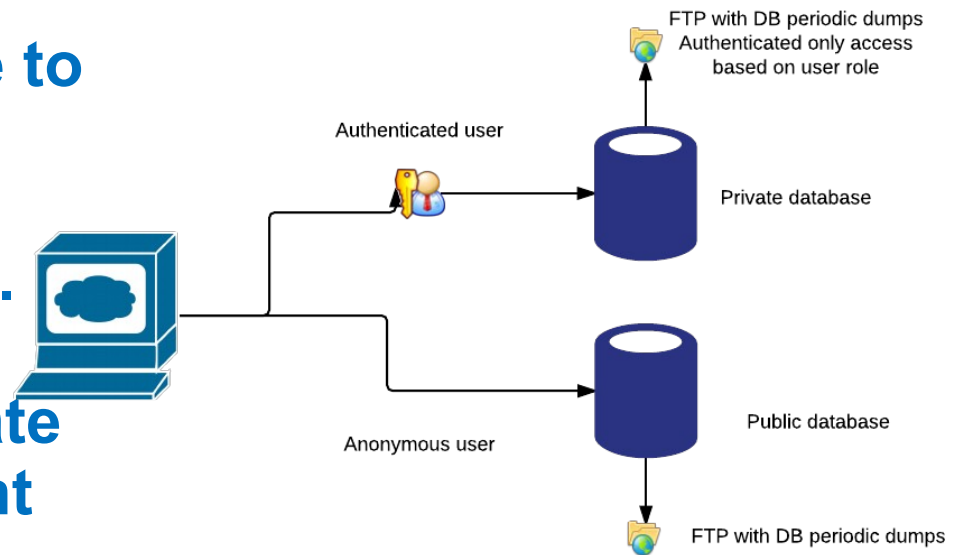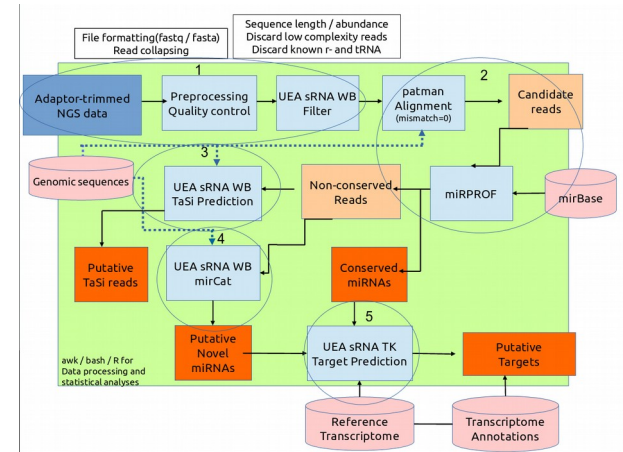
Resources Genomes

Should have all publicly available plant genomes on the server.

Aegilops tuschii (bread wheat)	**4.50Gb**
Arabidopsis thaliana	**0.14Gb**
Medicago truncatula	**0.42Gb**
Oriza sativa (Rice)	**0.38Gb**
Picea Abies	**12.0Gb**
Pinus Taeda	**18.0Gb**
Populus trichocarpa	**0.42Gb**
Vitis vinifera (Grape vine)	**0.49Gb**
Zea mays (Maize)	**3.24Gb**

Should have the possibility to use a custom genome (stored for a limited time).

7

Resources Transcriptomes
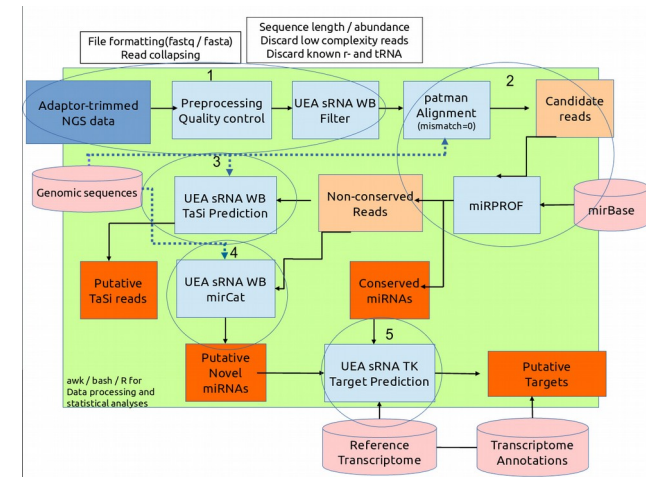
Should have all publicly available plant transcriptomes on the server.
Around the 100Mb – 400Mb each

Should have the possibility to use a custom transcriptome (stored for a limited time).

Resources - running

miRPursuit can be very memory intensive when searching for new sRNAs will need arround 32Gb of RAM for each run with a 6Gb genome.

Processor-wise it isn't extremely intensive (except for target prediction) 4 cores of a Intel Xeon E5-2620 should process a dataset with 2 libraries using a 6Gb genome in a day.

9

Resources – Target prediction
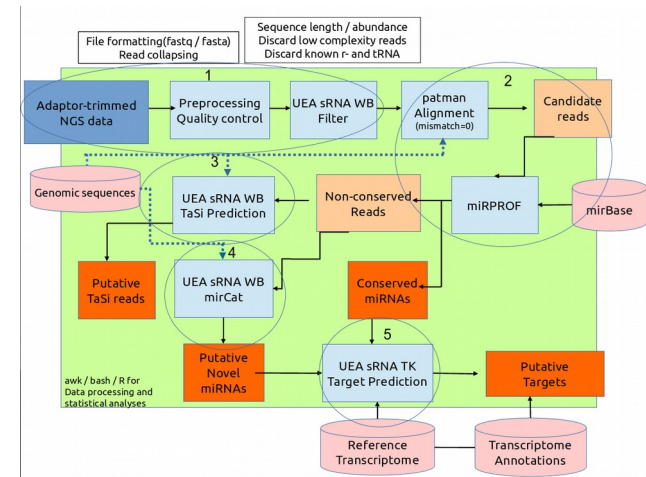
Target prediction of 1100 sRNAs with Cleaveland 4 running on 40 threads of a Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz **takes around 48hours.**

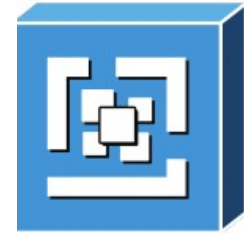**The server should query the DB to check if sequences already have targets predicted with similar parameters**
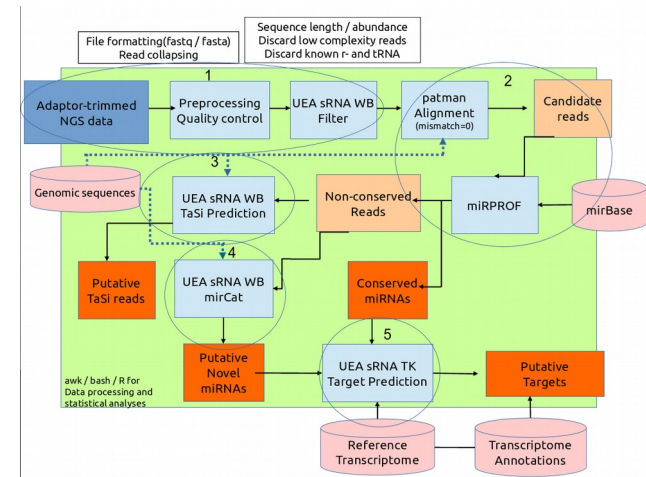
Analysis

Search

The main service should be the ability to query the database.
Search by sequence, related sequence or identifier. The ontologies in the meta-data table should allow for more complex searches based on experimental conditions.

Blast

Use blast on precursores and targets.
No need for blast to be running on our servers. It either redirects to NCBI/EBI blast with the sequence or runs blast locally against the NCBI servers.

Differential Expression analysis

Select datasets from DB and chose libraries to be compared using R based packages

11

Analysis – AWS EC2 instances

Memory optimized

Either through a long term payment for one instance or with multiple instances during short periods not paying for down time.

DYES	Year	3 Year	Hour	Type of instance
224	16 001,00 $	31 805,00 $	2,96 $	r3.8xlarge: 244 GiB of memory, 32 vCPUs, 2 x 320 GB of SSD-based instance storage, 64-bit platform, 10 Gigabit Ethernet
268	2 000,00 $	3 975,00 $	0,31 $	r3.xlarge: 30.5 GiB of memory, 4 vCPUs, 1 x 80 GB of SSD-based instance storage, 64-bit platform

DYES – Days till year upfront makes economic sense

Analysis – AWS EC2 instances

Cumpute optimized

Either through a long term payment for one instance or with multiple instances during short periods not paying for down time.

AmazonECS

DYES	Year	3 Year	Hour	
225	10 786,00 $	21 669,00 $	1,91 $	c3.8xlarge: 60 GiB of memory, 32 vCPUs, 640 GB of SSD-based local instance storage, 64-bit platform Intel Xeon E5-2680 v2 ("Ivy Bridge")
225	10 671,00 $	22 694,00 $	1,91 $	c4.8xlarge: 60 GiB of memory, 36 vCPUs, 64-bit platform E5-2666 v3 ("Haswell")

DYES – Days till year upfront makes economic sense

Deployment strategy

Local instance
A local instance for down time usage. During heavy traffic dedicated instances could be spun up to handle traffic.

AmazonECS

Local Server

Spin up multiple instances

Cloud instances

14
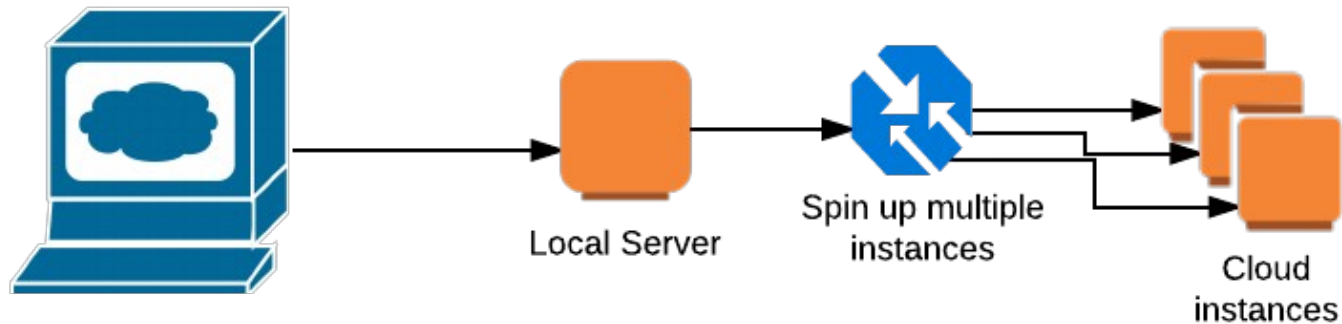
Thank you

www.ibet.pt

Bruno Costa