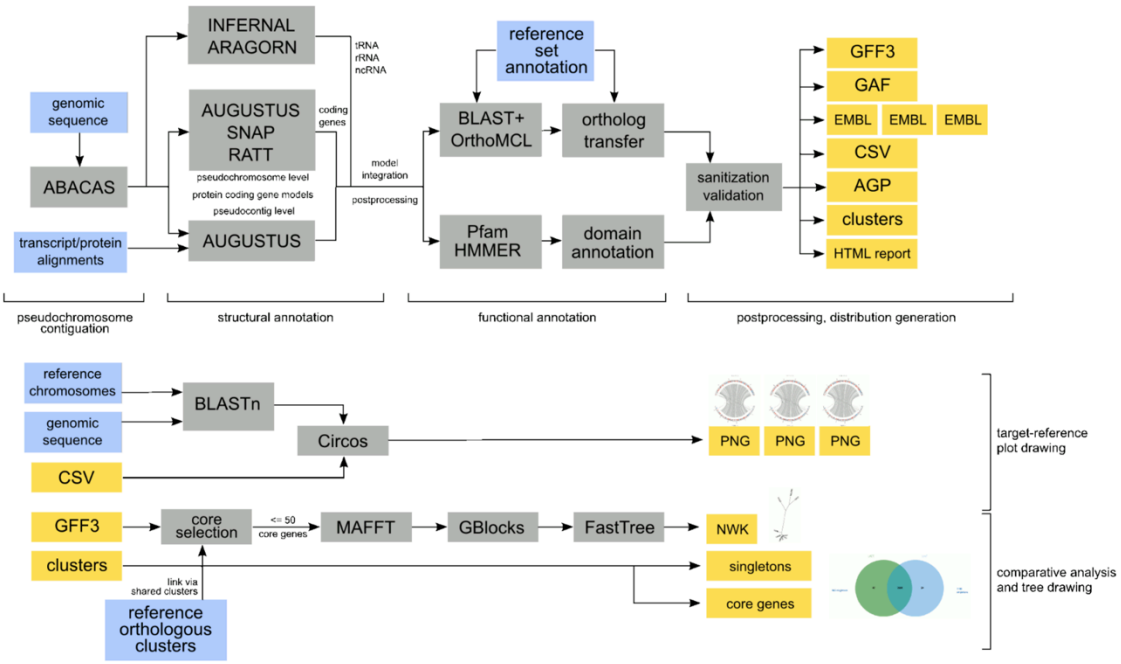
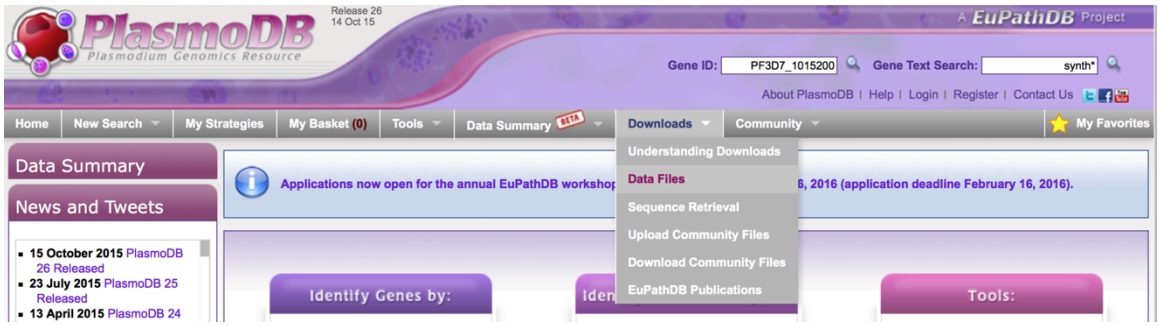
Genome Annotation with Companion (Part 1)

For this exercise we will start with an assembled genome that is unannotated.
We will obtain the assembled FASTA files from EuPathDB sites.
Companion is housed at Sanger and can be accessed here:
https://companion.sanger.ac.uk
Companion, is an online pipeline that employs different software to annotate and compare an assembled sequence to a reference-annotated genome.
The figure below illustrates the Companion pipeline, the software used and the expected output.

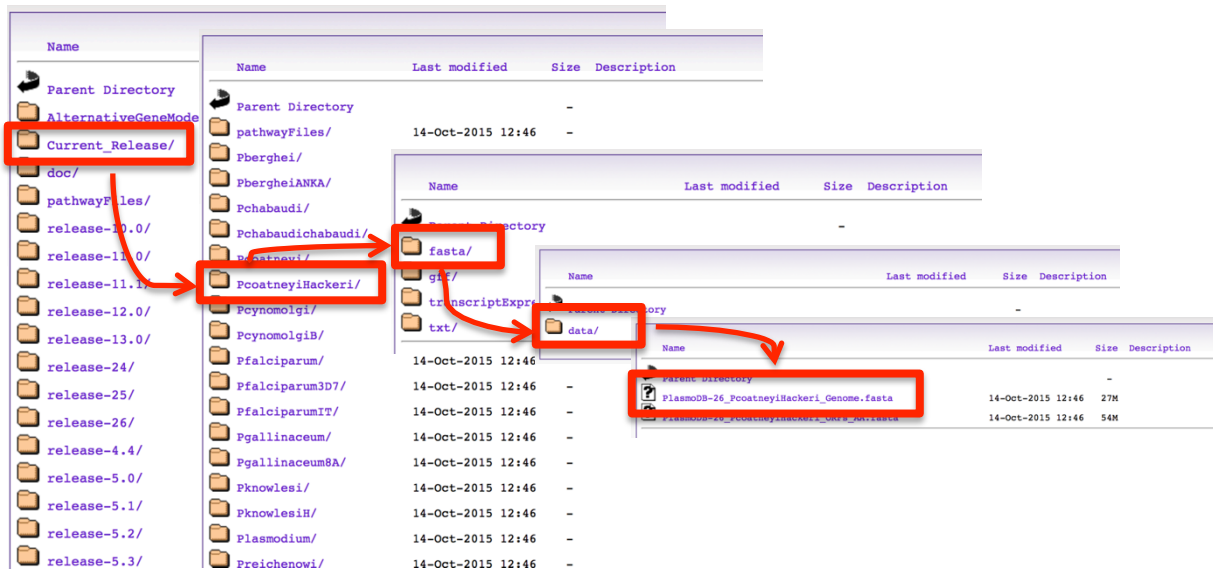


Each group will download one of the genomes as indicated below.
Group 1 – *Cryptosporidium baileyi* TAMU-09Q1
http://tinyurl.com/hgtddbz
Group 2 – *Cryptosporidium meleagridis* UKMEL1
http://tinyurl.com/zskwvf5
Group 3 – *Cryptosporidium hominis* UKH1
http://tinyurl.com/zoxy48u
Group 4 – *Plasmodium coatneyi* Hackeri
http://tinyurl.com/je4hftj
Group 5 – *Acanthamoeba palestinensis* Reich (only the largest 3000 contigs)
http://tinyurl.com/j2rd9lq

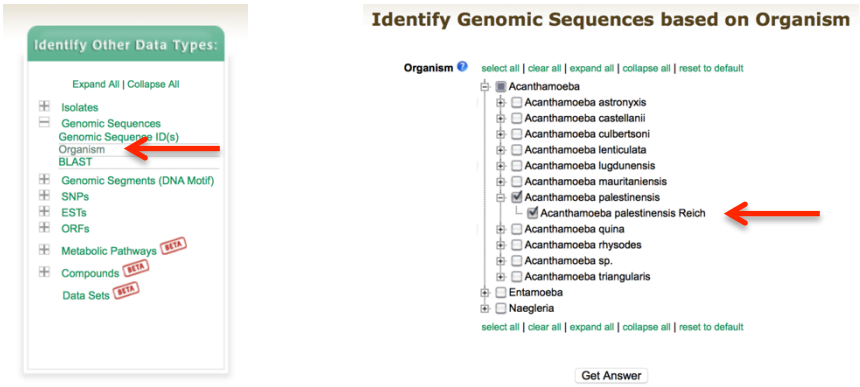
All genomes in EuPathDB sites are available for download form the “Data File” download section, which you can access from the Downloads menu in the gray tool bar.

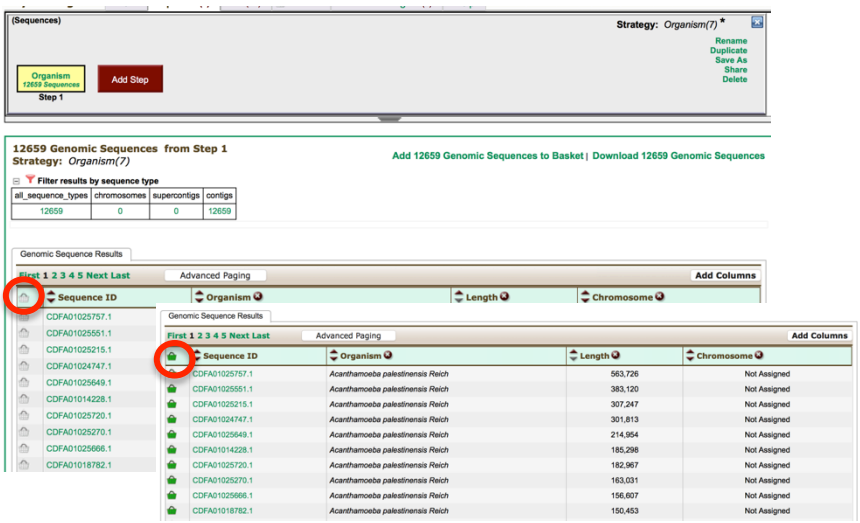


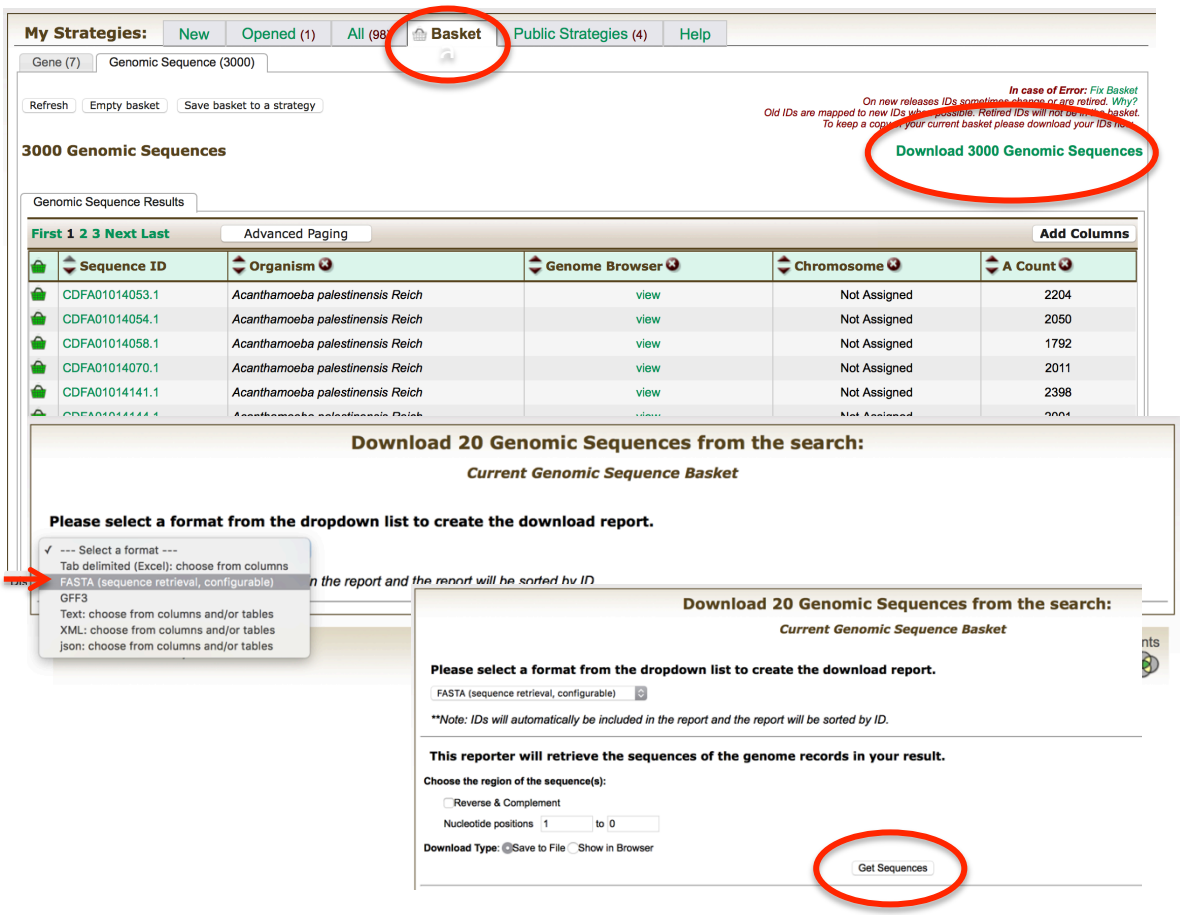
Selecting the Data Files option takes you to the download directories where you can navigate to the genome and data type you are looking for.



To download specific contigs/scaffolds/chromosomes you can use a genomic sequence search and place the desired sequences into your basket.

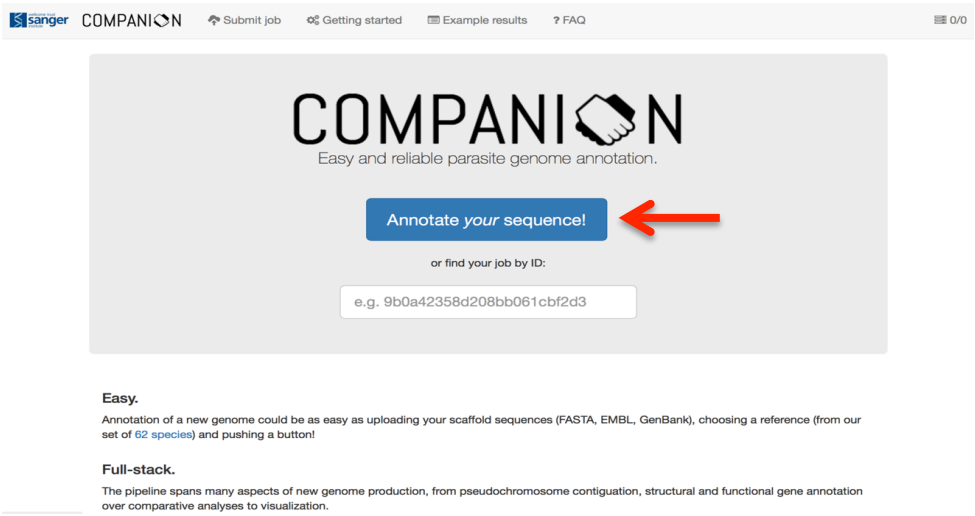



(Sequences)
Strategy: Organism(7) *
Rename
Duplicate
Save As
Share
Delete
Organism
12659 Sequences
Step 1
Add Step
12659 Genomic Sequences from Step 1
Strategy: Organism(7)
Add 12659 Genomic Sequences to Basket | Download 12659 Genomic Sequences
Filter results by sequence type
all_sequence_types | chromosomes | supercontigs | contigs
12659 | 0 | 0 | 12659
Genomic Sequence Results
First 1 2 3 4 5 Next Last
Advanced Paging
Add Columns
Sequence ID
Organism
Length
Chromosome
CDFA01025757.1 Acanthamoeba palestinensis Reich 563,726 Not Assigned
CDFA01025551.1 Acanthamoeba palestinensis Reich 383,120 Not Assigned
CDFA01025215.1 Acanthamoeba palestinensis Reich 307,247 Not Assigned
CDFA01024747.1 Acanthamoeba palestinensis Reich 301,813 Not Assigned
CDFA01025649.1 Acanthamoeba palestinensis Reich 214,954 Not Assigned
CDFA01014228.1 Acanthamoeba palestinensis Reich 185,298 Not Assigned
CDFA01025720.1 Acanthamoeba palestinensis Reich 182,967 Not Assigned
CDFA01025270.1 Acanthamoeba palestinensis Reich 163,031 Not Assigned
CDFA01025666.1 Acanthamoeba palestinensis Reich 156,607 Not Assigned
CDFA01018782.1 Acanthamoeba palestinensis Reich 150,453 Not Assigned

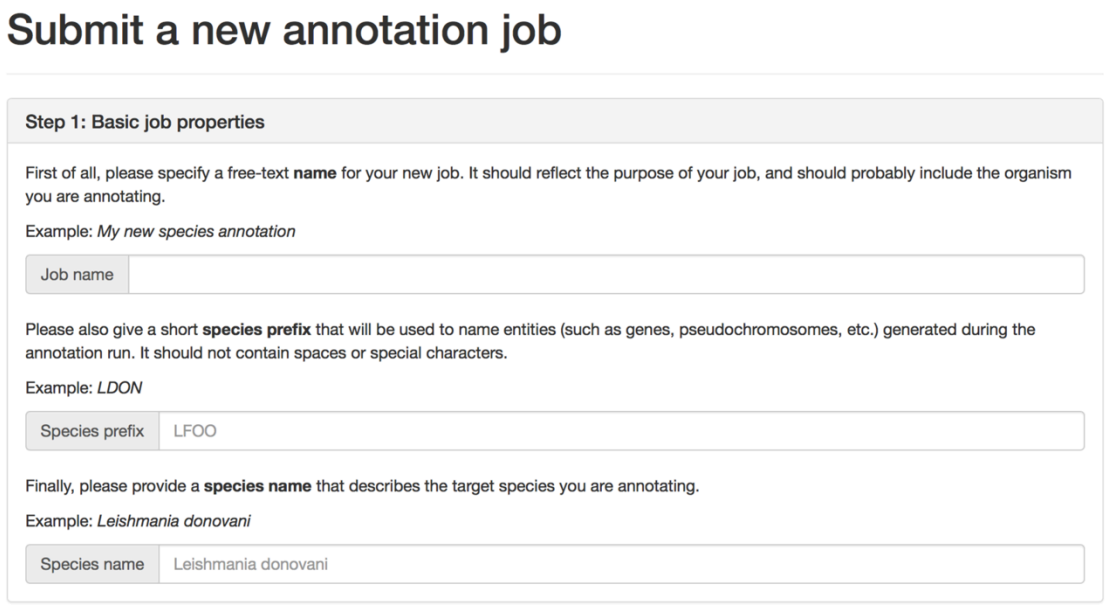

My Strategies:
New
Opened (1)
All (98
Basket
Public Strategies (4)
Help
Gene (7)
Genomic Sequence (3000)
Refresh
Empty basket
Save basket to a strategy
In case of Error: Fix Basket
3000 Genomic Sequences
Download 3000 Genomic Sequences
Genomic Sequence Results
First 1 2 3 Next Last
Advanced Paging
Add Columns
Sequence ID
Organism
Genome Browser
Chromosome
A Count
CDFA01014053.1 Acanthamoeba palestinensis Reich view Not Assigned 2204
CDFA01014054.1 Acanthamoeba palestinensis Reich view Not Assigned 2050
CDFA01014058.1 Acanthamoeba palestinensis Reich view Not Assigned 1792
CDFA01014070.1 Acanthamoeba palestinensis Reich view Not Assigned 2011
CDFA01014141.1 Acanthamoeba palestinensis Reich view Not Assigned 2398
Download 20 Genomic Sequences from the search:
Current Genomic Sequence Basket
Please select a format from the dropdown list to create the download report.
--- Select a format ---
Tab delimited (Excel): choose from columns
FASTA (sequence retrieval, configurable)
GFF3
Text: choose from columns and/or tables
XML: choose from columns and/or tables
json: choose from columns and/or tables
Download 20 Genomic Sequences from the search:
Current Genomic Sequence Basket
Please select a format from the dropdown list to create the download report.
FASTA (sequence retrieval, configurable)
**Note: IDs will automatically be included in the report and the report will be sorted by ID.
This reporter will retrieve the sequences of the genome records in your result.
Choose the region of the sequence(s):
Reverse & Complement
Nucleotide positions 1 to 0
Download Type: Save to File Show in Browser
Get Sequences

- Once you have downloaded your sequence file, go to the Companion site: https://companion.sanger.ac.uk

- Click on the “Annotate your sequence” link.



-Follow the instructions as described on the Companion website:
1. Provide basic information about the job you are about to submit. This includes a job name, species prefix (usually the first letter of the genus and the first three letters of the species: *Acanthamoeba palestinensis* = Apal).



2. In step 2, choose the assembly file that you downloaded.
3. In step 3, indicate if you will be using RNAseq evidence to guide the annotation – in this exercise we will **not** use any RNAseq data.
4. In step 4, select the reference sequence you would like to use to transfer the annotation and to compare your sequence to. Typically you would like to use a reference that is closely related, so a phylogenetic tree might be useful to look at. Here are examples for *Plasmodium* and *Cryptosporidium*. There is only one reference for *Acanthamoeba*.

http://tolweb.org/Plasmodium/68071

http://tolweb.org/Cryptosporidium/124803

Step 2: Target sequence

Please upload a **target sequence file** to be annotated from your local filesystem using the button below. The file (FASTA, EMBL or GenBank format) can be gzip- or bzip2-compressed. In this case it must have a `.gz` or `.bz2` suffix.

Note: The maximal size of your uploaded file is **64 MB**, and the maximum number of individual sequences in it is **3000**.

Choose File no file selected

Here is an example sequence input file for a *Plasmodium falciparum* IT chromosome 5 sequence that can be used with the *Plasmodium falciparum* 3D7 example reference set (choose below in step 4) for a quick example run. To use it, please download it to your local machine and upload it using the button above.

Step 3: Transcript evidence

The *Companion* pipeline can optionally make use of assembled transcripts in the GTF format as created by Cufflinks.

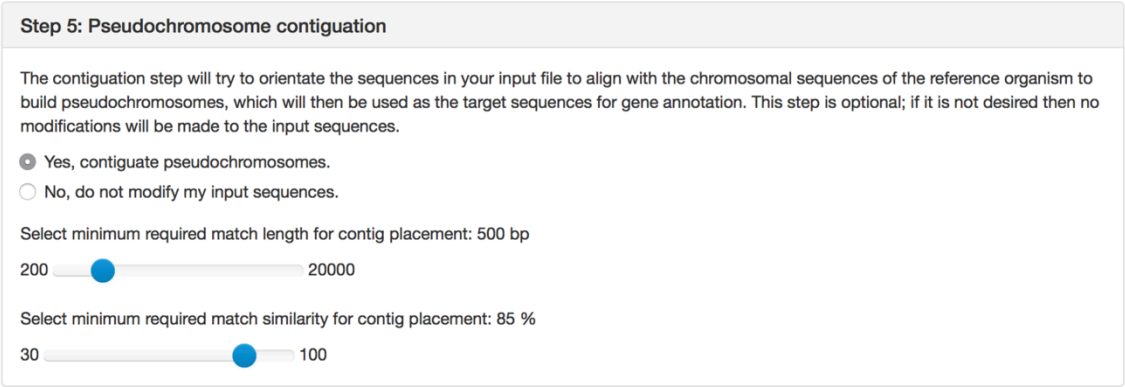
- ○ Yes, use transcript evidence.
- ◉ No, do not use transcript evidence.

Step 4: Reference organism

Please pick a (if possible closely related) **reference organism** for this annotation run. This organism will be used to specify the models for gene finding, functional annotation transfer and pseudochromosome contiguation.

Please select a reference species

5. In step 5, there a few more parameters you may one to examine. For the purpose of our exercise we will keep these at the default values.



6. Enter your email address to get an update when your job starts running and when it is complete. Next, click on the “I’m not a robot” captcha (Completely Automated Public Turing test to tell Computers and Humans Apart). Finally, click on the “Submit Job” link.