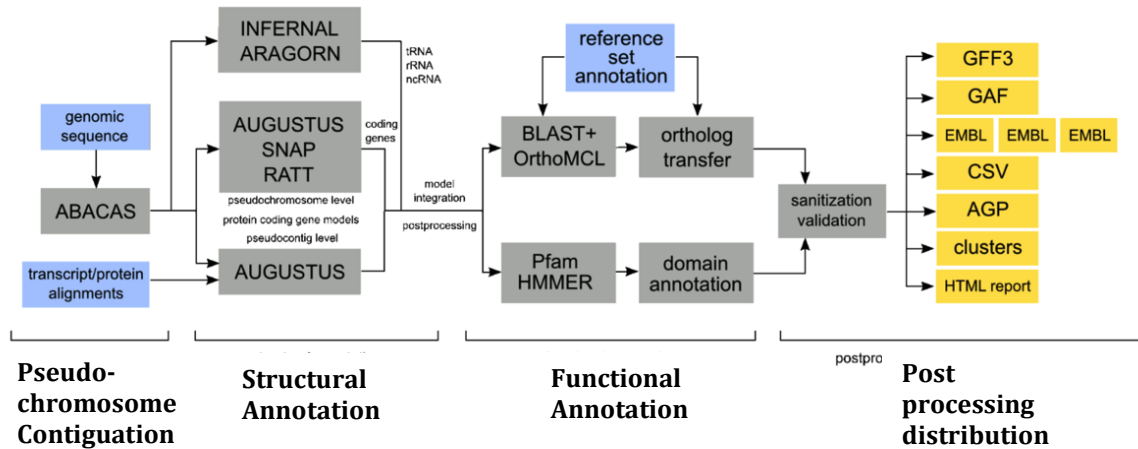
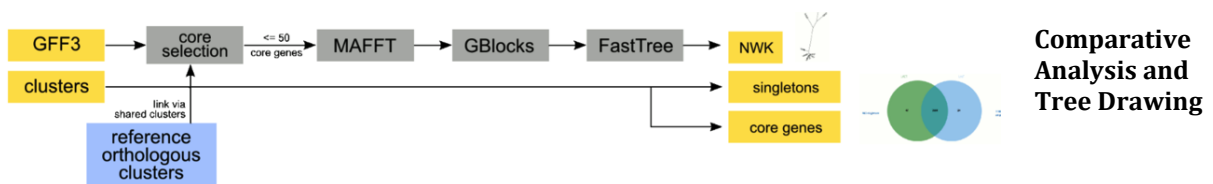


Genome Annotation with Companion (Part 1)

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.



- GFF3 – (General Feature Format) Structural annotation - one line per feature, each containing 9 columns of data, plus optional track definition lines
- EMBL – One sequence entry starts with an identifier line ("ID"), followed by further annotation lines.
- AGP – (A Golden Path) Describes the assembly of a larger sequence object from smaller objects.
- GAF – (GO Annotation File) – Functional annotation - GAFs are tab-delimited plain text files, where each line in the file represents a single association between a gene product and a GO term, with an evidence code, the reference to support the link between them, and other information.



For this exercise, we will start with an assembled genome that is unannotated. We will obtain the assembled FASTA files from EuPathDB sites and submit the genome to the Companion pipeline for annotation. Companion is housed at Sanger and can be accessed here: <https://companion.sanger.ac.uk> Each group will download one of the following genomes. The tinyURL links will initiate the download.

Group 1 – *Plasmodium coatneyi* Hackeri using *Plasmodium knowlesi* as reference
<https://tinyurl.com/y6keezu3>

Group 2 - *Plasmodium coatneyi* Hackeri using *Plasmodium falciparum* as reference
<https://tinyurl.com/y6keezu3>

Group 3 – *Cryptosporidium meleagridis* using *Cryptosporidium parvum* as reference
<https://tinyurl.com/yy8m6o57>

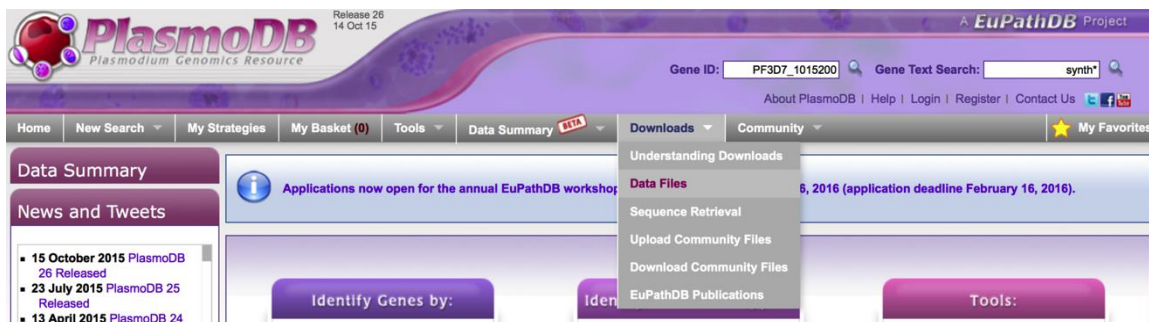
Group 4 *Cryptosporidium baileyi* using *Cryptosporidium parvum* as reference
<https://tinyurl.com/yxi9z3ea>

Group 5 *Leishmania amazonensis* using *Leishmania major* as reference
<https://tinyurl.com/y5tlewhj>

Group 6 *Trypanosoma congolense* using *Trypanosoma brucei* 927 as reference.
<https://tinyurl.com/y3hla6qm>

A word about downloads:

TinyURLs above for are direct links to our genome FASTA files in the corresponding EuPathDB site downloads section. All genomes in EuPathDB sites are available for download from 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.

The image shows a multi-level directory view. The left pane shows a tree structure with 'Current_Release/' highlighted. The middle pane shows a list of sub-directories including 'PcoatneyiHackeri/'. The right pane shows a list of files within 'PcoatneyiHackeri/' with 'data/' highlighted. A further zoomed-in view of the 'data/' folder shows a file named 'PlasmidB-26_PcoatneyiHackeri_Genome.fasta' highlighted.

To download specific contigs/scaffolds/chromosomes instead of entire genomes, use a genomic sequence search and place the desired sequences into your basket.

The 'Identify Other Data Types' menu is shown with 'Genomic Sequences' expanded. Under 'Genomic Sequences', 'Organism' is selected with a red arrow pointing to it.

Identify Genomic Sequences based on Organism

The search results for 'Organism' are displayed. Under the 'Acanthamoeba' category, 'Acanthamoeba palestiniensis Reich' is selected with a red arrow pointing to it.

The screenshot shows a search results page for 'Organism(7)'. It displays 12659 genomic sequences. A table below the search bar shows the results: all_sequences_types: 12659, chromosomes: 0, supercontigs: 0, contigs: 12659. The main table lists genomic sequence results with columns for Sequence ID, Organism, Length, and Chromosome. The first few rows are highlighted, showing IDs like CDFAI025757.1 and CDFAI025551.1, all belonging to Acanthamoeba palestiniensis Reich.

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

the report and the report will be sorted by ID.

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

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
On new releases IDs sometimes change or are retired. Why?
Old IDs are mapped to new IDs where possible. Retired IDs will not be in the basket.
To keep a copy of your current basket please download your IDs.

Download 3000 Genomic Sequences

Genomic Sequence Results

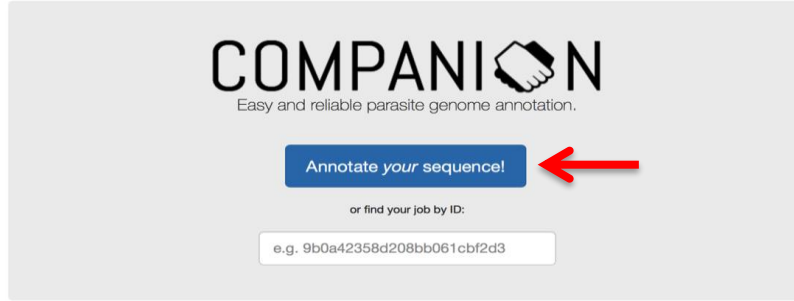
First 1 2 3 Next Last Advanced Paging Add Columns

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
CDFA01014144.1	Acanthamoeba palestinensis Reich	view	Not Assigned	2001
CDFA01014148.1	Acanthamoeba palestinensis Reich	view	Not Assigned	4170
CDFA01014170.1	Acanthamoeba palestinensis Reich	view	Not Assigned	2482
CDFA01014191.1	Acanthamoeba palestinensis Reich	view	Not Assigned	1551
CDFA01014193.1	Acanthamoeba palestinensis Reich	view	Not Assigned	1460
CDFA01014199.1	Acanthamoeba palestinensis Reich	view	Not Assigned	2328
CDFA01014208.1	Acanthamoeba palestinensis Reich	view	Not Assigned	4282

-**Back to the Annotation:** Once you have downloaded your sequence file, go to the Companion site:

<https://companion.sanger.ac.uk>

- Click on the “Annotate your sequence” link.



Easy.

Annotation of a new genome could be as easy as uploading your scaffold sequences (FASTA, EMBL, GenBank), choosing a reference (from our set of 62 species) and pushing a button!

Full-stack.

The pipeline spans many aspects of new genome production, from pseudochromosome contiguation, structural and functional gene annotation over comparative analyses to visualization.

-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).

Submit a new annotation job

Step 1: Basic job properties

First of all, please specify a free-text **name** for your new job. It should reflect the purpose of your job, and should probably include the organism you are annotating.

Example: *My new species annotation*

Job name

Please also give a short **species prefix** that will be used to name entities (such as genes, pseudochromosomes, etc.) generated during the annotation run. It should not contain spaces or special characters.

Example: *LDON*

Species prefix

Finally, please provide a **species name** that describes the target species you are annotating.

Example: *Leishmania donovani*

Species name

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.

<http://tolweb.org/Cryptosporidium/124803>

<http://tolweb.org/Plasmodium/68071>

Leishmania phylogenetic tree -

<https://journals.plos.org/plosntds/article/figure?id=10.1371/journal.pntd.0003339.g005>

Trypanosoma Tree

https://projects.exeter.ac.uk/meeg/sites/default/files/pictures/tryp_tree.jpg

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**.

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 are a few more parameters you may want to examine. For the purpose of our exercise we will keep these at the default values.

Step 5: Pseudochromosome contiguation

The contiguation step will try to orientate the sequences in your input file to align with the chromosomal sequences of the reference organism to build pseudochromosomes, which will then be used as the target sequences for gene annotation. This step is optional; if it is not desired then no modifications will be made to the input sequences.

- Yes, contiguate pseudochromosomes.
 No, do not modify my input sequences.

Select minimum required match length for contig placement: 500 bp

200  20000

Select minimum required match similarity for contig placement: 85 %

30  100

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.

Step 6: Advanced settings (click chevron to the right to show/hide)

Your contact information (optional)

You can leave your email address if you want to be notified when your job starts and finishes. This is absolutely optional, if you choose not to share your email address, you can always manually check the status of your job using a private link provided by us after submission.

Email

To protect the service from automated bots, please prove that you are a human by ticking the box below.

I'm not a robot



Submit job