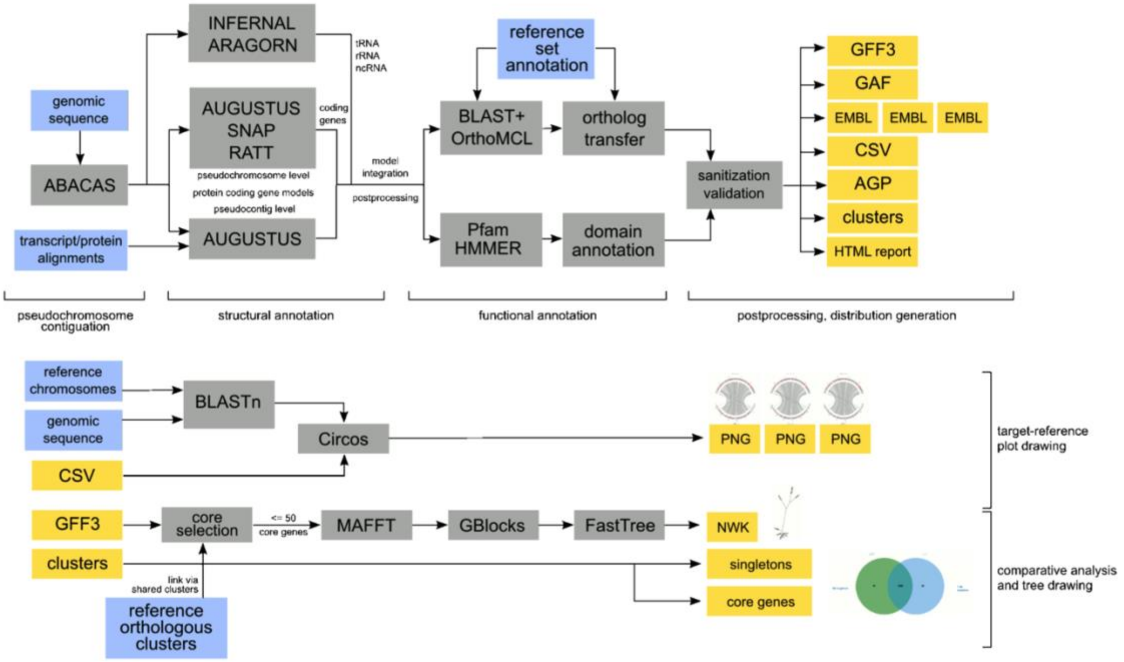
Genome Annotation with Companion (Part 1)

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.



For this exercise, we will start with an assembled genome that is unannotated. We will obtain the assembled FASTA files from EuPathDB sites. Companion can be accessed here: http://protozoacompanion.gla.ac.uk/

Each group will download one of the following genomes (the tinyURL links will initiate the download) and will use Companion to compare with the specified genome as reference.

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

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

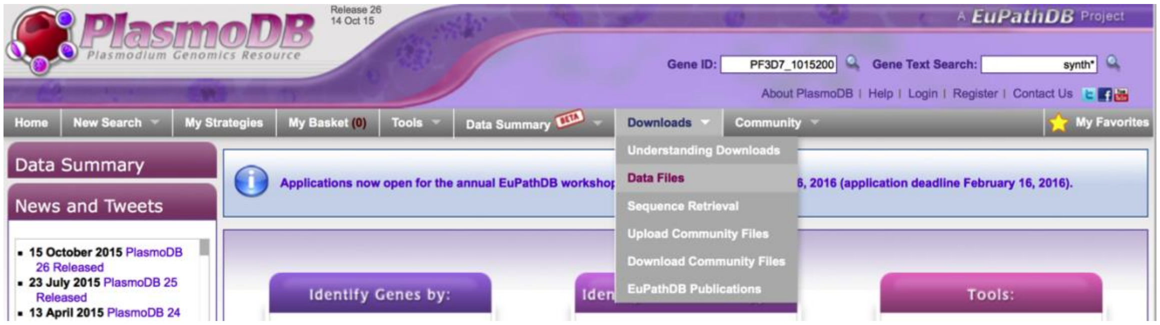
Group 3 – *Cryptosporidium meleagridis* using *Cryptosporidium parvum* as reference
https://tinyurl.com/yyqxgr5q

Group 4 *Cryptosporidium baileyi* using *Cryptosporidium parvum* as reference
https://tinyurl.com/yyffffrd

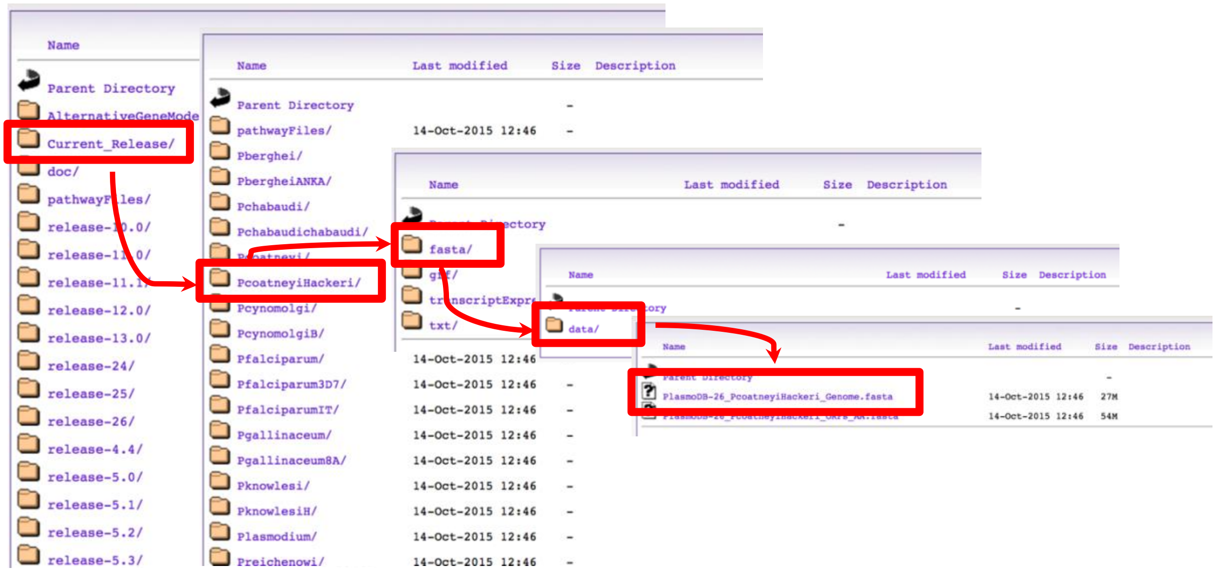
Group 5 *Leishmania amazonensis* using *Leishmania major* as reference
https://tinyurl.com/yy5wkymk

Group 6 *Trypanosoma congolense* using *Trypanosoma brucei* 927 as reference.
https://tinyurl.com/y6yqys7w

A word about downloads:
TinyUrls above 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.


```
Name

Parent Directory
AlternativeGeneMode
Current_Release/
doc/
pathwayFiles/
release-10.0/
release-11.0/
release-11.1/
release-12.0/
release-13.0/
release-24/
release-25/
release-26/
release-4.4/
release-5.0/
release-5.1/
release-5.2/
release-5.3/
```

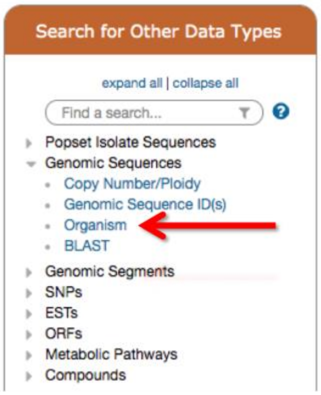
```
Name                          Last modified      Size  Description

Parent Directory                                  -
pathwayFiles/                 14-Oct-2015 12:46   -
Pberghei/
PbergheiANKA/
Pchabaudi/
Pchabaudichabaudi/
Pcoatneyi/
PcoatneyiHackeri/
Pcynomolgi/
PcynomolgiB/
Pfalciparum/                  14-Oct-2015 12:46
Pfalciparum3D7/               14-Oct-2015 12:46   -
PfalciparumIT/                14-Oct-2015 12:46   -
Pgallinaceum/                 14-Oct-2015 12:46   -
Pgallinaceum8A/               14-Oct-2015 12:46   -
Pknowlesi/                    14-Oct-2015 12:46   -
PknowlesiH/                   14-Oct-2015 12:46   -
Plasmodium/                   14-Oct-2015 12:46   -
Preichenowi/                  14-Oct-2015 12:46   -
```

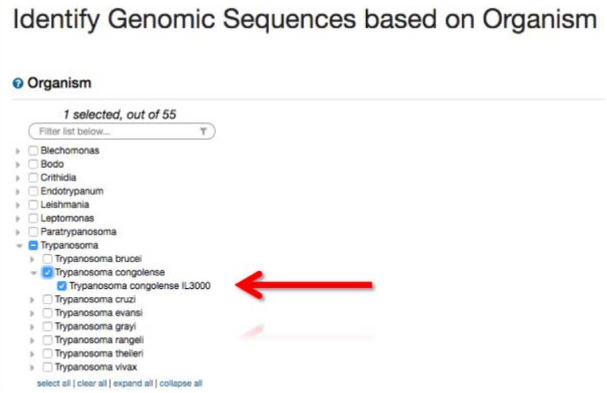
```
Name                          Last modified      Size  Description

Parent Directory                                  -
fasta/
gff/
transcriptExpr
txt/
```

```
Name                          Last modified      Size  Description

Parent Directory                                  -
data/
```

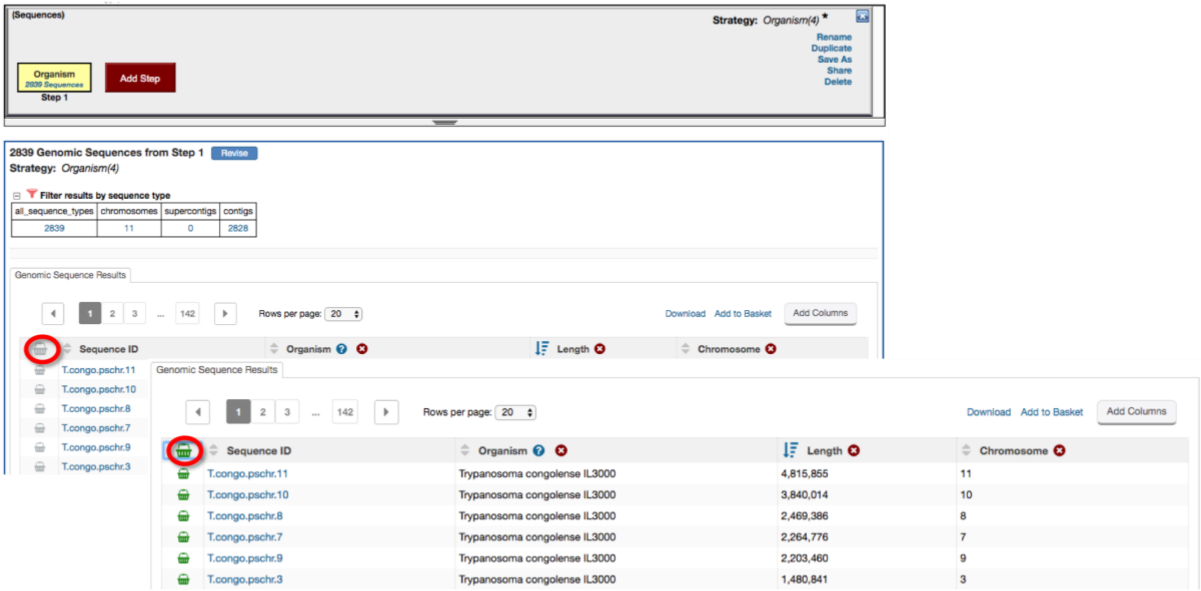
```
Name                                              Last modified      Size  Description

Parent Directory                                                      -
PlasmoDB-26_PcoatneyiHackeri_Genome.fasta         14-Oct-2015 12:46   27M
PlasmoDB-26_PcoatneyiHackeri_ORFs_AA.fasta        14-Oct-2015 12:46   54M
```

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





Identify Genomic Sequences based on Organism



Strategy: *Organism(4)* *

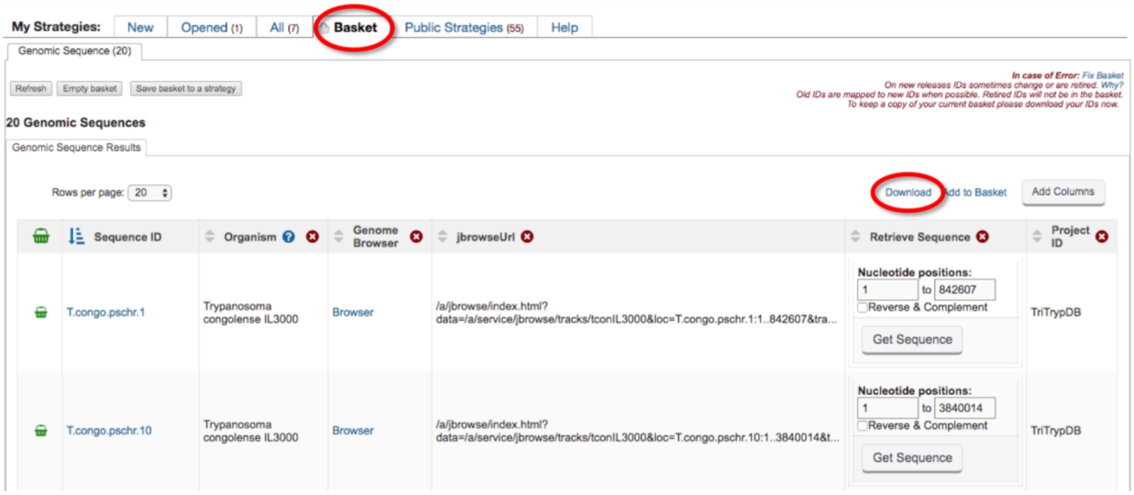
2839 Genomic Sequences from Step 1

Strategy: *Organism(4)*

Filter results by sequence type

all_sequence_types	chromosomes	supercontigs	contigs
2839	11	0	2828

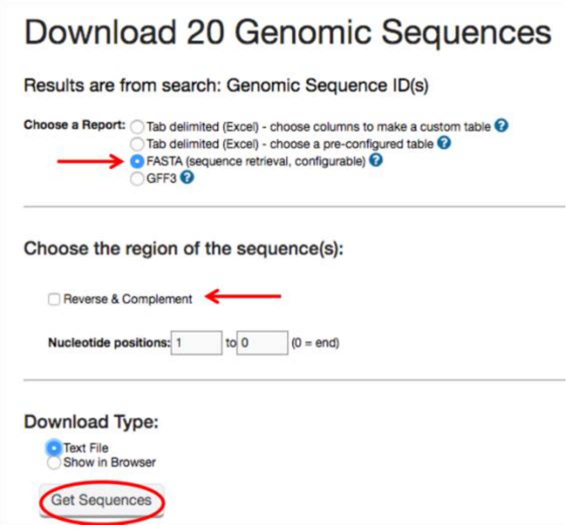
Sequence ID	Organism	Length	Chromosome
T.congo.pschr.11	Trypanosoma congolense IL3000	4,815,855	11
T.congo.pschr.10	Trypanosoma congolense IL3000	3,840,014	10
T.congo.pschr.8	Trypanosoma congolense IL3000	2,469,386	8
T.congo.pschr.7	Trypanosoma congolense IL3000	2,264,776	7
T.congo.pschr.9	Trypanosoma congolense IL3000	2,203,460	9
T.congo.pschr.3	Trypanosoma congolense IL3000	1,480,841	3



In case of Error: Fix Basket
On new releases IDs sometimes change or are retired. Why?
Old IDs are mapped to new IDs when possible. Retired IDs will not be in the basket.
To keep a copy of your current basket please download your IDs now.

20 Genomic Sequences

Sequence ID	Organism	Genome Browser	jbrowseUrl	Retrieve Sequence	Project ID
T.congo.pschr.1	Trypanosoma congolense IL3000	Browser	/a/jbrowse/index.html?data=/a/service/jbrowse/tracks/tconIL3000&loc=T.congo.pschr.1:1..842607&tra...	Nucleotide positions: 1 to 842607 □ Reverse & Complement Get Sequence	TriTrypDB
T.congo.pschr.10	Trypanosoma congolense IL3000	Browser	/a/jbrowse/index.html?data=/a/service/jbrowse/tracks/tconIL3000&loc=T.congo.pschr.10:1..3840014&t...	Nucleotide positions: 1 to 3840014 □ Reverse & Complement Get Sequence	TriTrypDB



Download 20 Genomic Sequences

Results are from search: Genomic Sequence ID(s)

Choose a Report:
○ Tab delimited (Excel) - choose columns to make a custom table
○ Tab delimited (Excel) - choose a pre-configured table
● FASTA (sequence retrieval, configurable)
○ GFF3

Choose the region of the sequence(s):

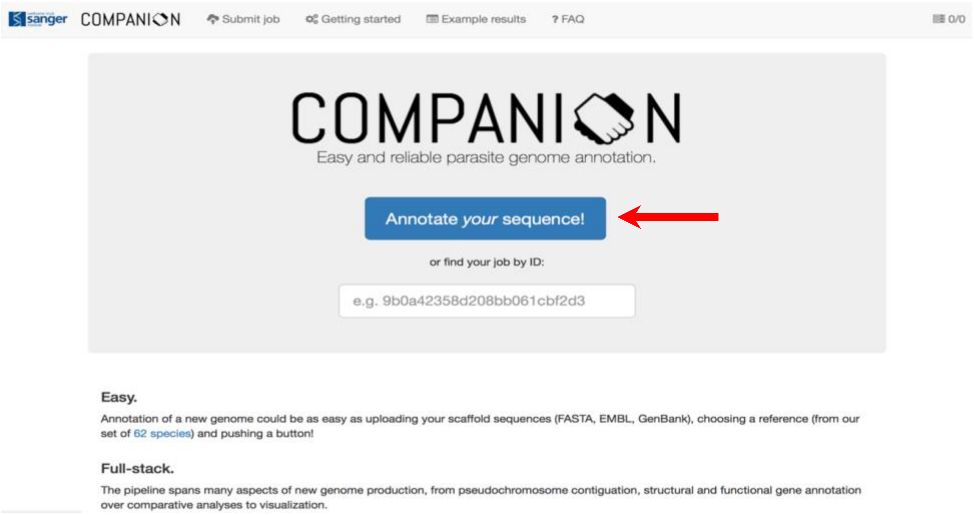
□ Reverse & Complement

Nucleotide positions: 1 to 0 (0 = end)

Download Type:
● Text File
○ Show in Browser

-**Back to the Annotation**: Once you have downloaded your sequence file, go to the Companion site:
http://protozoacompanion.gla.ac.uk/

- Click on the “Annotate your sequence” link.



Easy.

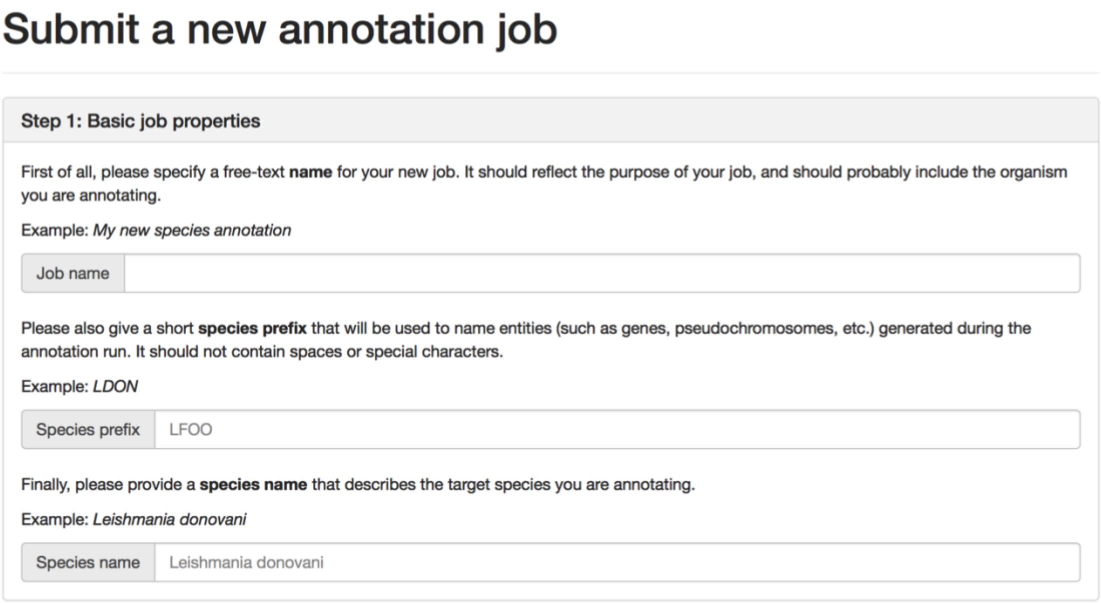
Annotation of a new genome could be as easy as uploading your scaffold sequences (FASTA, EMBL, GenBank), choosing a reference (from our set of 62 species) and pushing a button!

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: *Cryptosporidium parvum* = Cpar).



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: LFOO

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

Example: *Leishmania donovani*

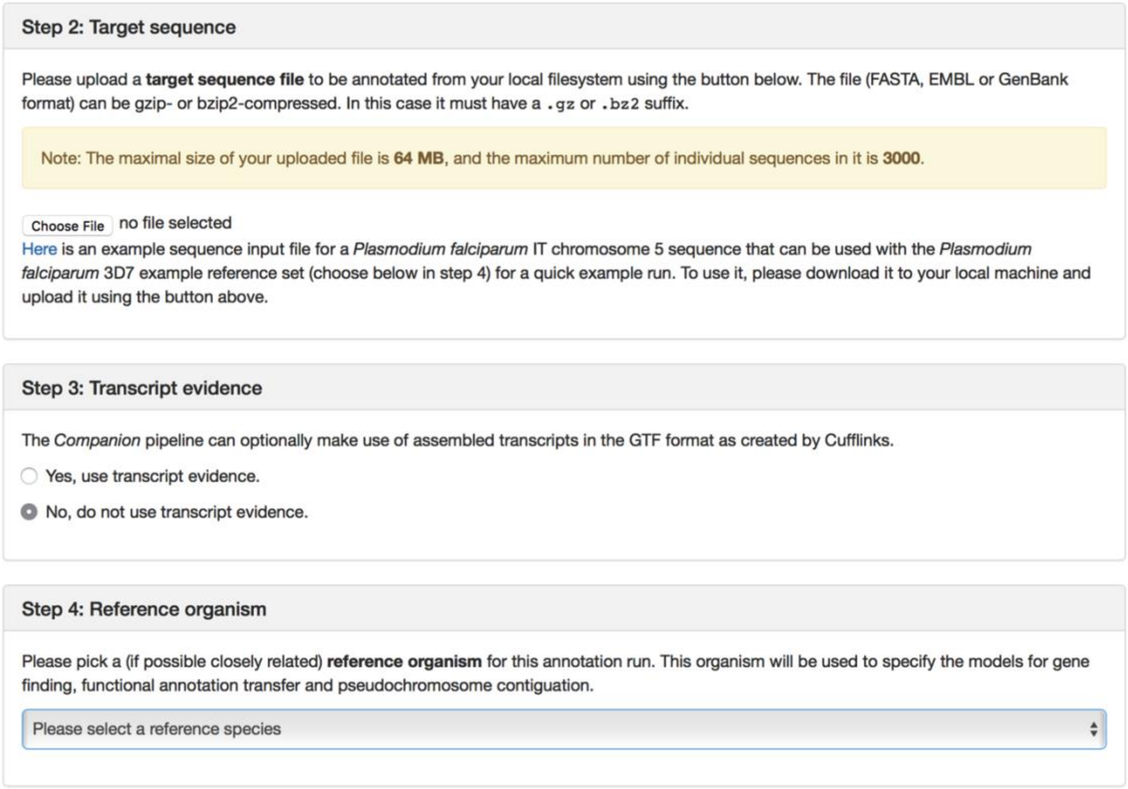
Species name: Leishmania donovani

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 of phylogenies for *Plasmodium* and *Cryptosporidium*.

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 phylogenetic 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.

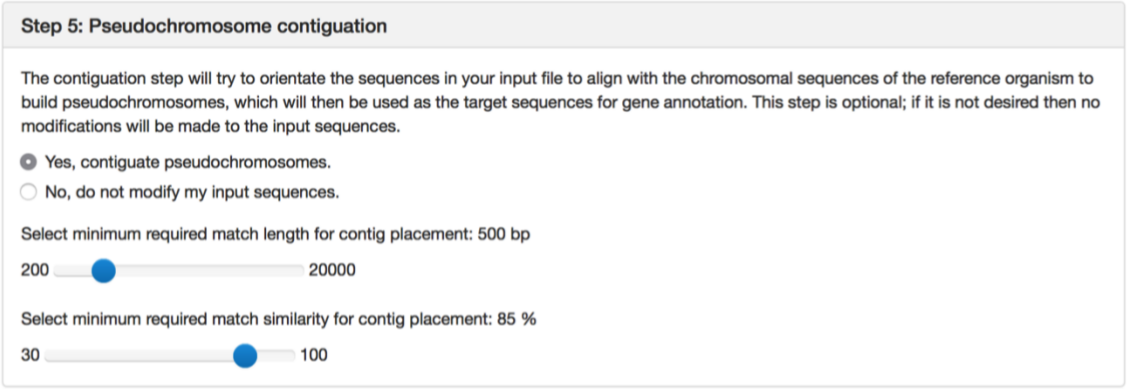
○ Yes, use transcript evidence.

● 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.

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.

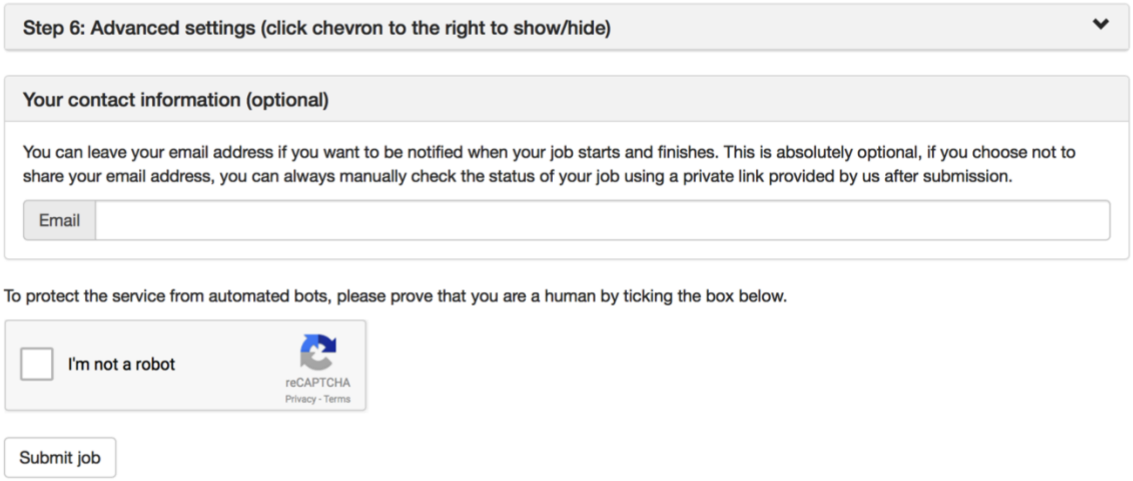
● Yes, contiguate pseudochromosomes.

○ No, do not modify my input sequences.

Select minimum required match length for contig placement: 500 bp

Select minimum required match similarity for contig placement: 85 %

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.