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: <u>http://protozoacompanion.gla.ac.uk/</u>

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 <u>https://tinyurl.com/yxuyqszu</u>

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

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

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

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

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

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.

	DOD mics Reso	Release 26 14 Oct 15	23	the second	Gene ID:	PF3D7_	A EuPathDB Projec	t
- 12		TYPE	-			About	it PlasmoDB Help Login Register Contact Us 📧 😭	
Home New Search 👻 My S	Strategies	My Basket (0)	Tools 👻	Data Summary	Downloads	Communit	ity 👻 📩 My Favo	orites
Data Summary					Understanding [Downloads		
Data Summary		Applications now	open for the	annual EuPathDB workshop	Data Files		6, 2016 (application deadline February 16, 2016).	
News and Tweets					Sequence Retrie			
					Upload Commun			
 15 October 2015 PlasmoDB 26 Released 23 July 2015 PlasmoDB 25 Released 13 April 2015 PlasmoDB 24 		Identify (Genes by:	Iden	Download Comm EuPathDB Public	nunity Files cations	Tools:	

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	Name	Last modified Size	Description	
AlternativeGeneMode	Parent Directory	-		
Current_Release/	pathwayFiles/	14-Oct-2015 12:46 -		
doc/	Pberghei/			
pathwayF les/	PbergheiANKA/	Name	Last modified Size Description	
release-10.0/	Pchabaudi/	Provide Dispectory	-	
release-11 0/	Prostnevi/	> asta/		
release-11.1	PcoatneyiHackeri/	gat/ Name	Last modifie	ed Size Description
release-12.0/	Pcynomolgi/	trunscriptExpre	lory	
release-13.0/	PcynomolgiB/	data	/ Name	Last modified Size Description
Telease-24/	Pfalciparum/	14-Oct-2015 12:46		
release-25/	Pfalciparum3D7/	14-Oct-2015 12:46 -	PlasmoDB-26_PcoatneyiBackeri_Genome.fasta	14-Oct-2015 12:46 27M
release-26/	PfalciparumIT/	14-Oct-2015 12:46 -		14-Oct-2015 12:46 54M
release-4.4/	Pgallinaceum%	14-Oct-2015 12:46 -		
release-5.0/	Pknowlesi/	14-Oct-2015 12:46 -		
release-5.1/	PknowlesiH/	14-Oct-2015 12:46 -		
release-5.2/	D Plasmodium/	14-Oct-2015 12:46 -		
release-5.3/	Preichenowi/	14-Oct-2015 12:46 -		

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

			Iù	entify Genomic S	Sequences b	based o	on Organism	
1.	expand all colla	apse all	0(Drganism				
Pc Ge	Find a search opset Isolate Sequences copy Number/Ploidy Genomic Sequence Organism BLAST	T 2 Cres / ID(s)		1 selected, out of 55 The first below				
> SN	NPs			 Trypanosoma congolense Trypanosoma congolense IL3000 	~ ~			
> ES	STS		3	 Trypanosoma cruzi Trypanosoma evansi 				
> OF	RFs etabolic Pathways		3	Trypanosoma grayi	1			
) Co	ompounds		3	Trypanosoma theileri Trypanosoma vivax				
1		1		select all clear all expand all collapse all				
(Sequences)	m Add Step				Strategy: Org	anism(4) * Rename Duplicate Save As Share	3	
2839 Sequen Step 1	Add Step					Delete		
2839 Genon Strategy: O	nic Sequences from Step 1 Irganism(4)	Revise						
all_sequence_ 2839	types chromosomes superconti 11 0	gs contigs 2828						
Genomic Seq	uence Results							
•	1 2 3 142	Rows per page: 2	0 0		Download Add to Basket	Add Columns		
	Sequence ID congo.pschr.11 Genomic Se	Organism equence Results	00	↓F Length O	Chromosome O			
	congo.pschr.10 congo.pschr.8 congo.pschr.7	1 2 3	142 🕨 Row	is per page: 20 ¢			Download	Add to Basket Add Columns
💮 T.	congo.pschr.9					17.00		
		Sequence ID		🔶 Organism 😧 😫		Length 🖸	Chromo	osome 😣
🕀 T.	congo.pschr.3	Sequence ID T.congo.pschr.11		 Organism ? S Trypanosoma congolense IL3000 		4,815,855	Chromo11	osome 😧
🕀 T.	congo.pschr.3	Sequence ID T.congo.pschr.11 T.congo.pschr.10		Organism Organism		4,815,855 3,840,014	 Chromo 11 10 	osome 오
<u>⊕</u> T.	congo.pschr.3	Sequence ID T.congo.pschr.11 T.congo.pschr.10 T.congo.pschr.8 T.congo.pschr.7		Organism Organism		Length 4,815,855 3,840,014 2,469,386 2,264,776	 Chromo 11 10 8 7 	osome 🛇
<u>⊖</u> T.	congo.pschr.3	Sequence ID T.congo.pschr.11 T.congo.pschr.10 T.congo.pschr.8 T.congo.pschr.7 T.congo.pschr.9		Organism Organism Organism Congolense IL3000 Trypanosoma congolense IL3000 Trypanosoma congolense IL3000 Trypanosoma congolense IL3000 Trypanosoma congolense IL3000		Length 4,815,855 3,840,014 2,469,386 2,264,776 2,203,460	Chromo 11 10 8 7 9	osome 🛇
<u>⊕</u> T.	congo.pschr.3	Sequence ID Congo.pschr.11 T.congo.pschr.10 T.congo.pschr.8 T.congo.pschr.7 T.congo.pschr.9 T.congo.pschr.3		Organism Organism Organism Congolense IL3000 Trypanosoma congolense IL3000		Length Q 4,815,855 3,840,014 2,469,386 2,264,776 2,203,460 1,480,841	Chrome 11 10 8 7 9 3	osome O
- T.	congo.pechr.3	Sequence ID T.congo.pschr.11 T.congo.pschr.8 T.congo.pschr.7 T.congo.pschr.9 T.congo.pschr.3		Organism Companism Company Company		Length Q 4,815,855 3,840,014 2,469,386 2,264,776 2,203,460 1,480,841	 Chromo 11 10 8 7 9 3 	osome O
My Stra	ttogies: New Or	Sequence ID Loongo.pschr.11 Loongo.pschr.10 Loongo.pschr.8 Loongo.pschr.7 T.congo.pschr.9 T.congo.pschr.3 Dened (1) All (7)	Basket Publ	Organism Organism		4,815,855 3,840,014 2,469,386 2,264,776 2,203,460 1,480,841	 Chromo 11 10 8 7 9 3 	ssome O
My Stra Genomic	ttegies: New Op : Sequence (20) Employ baskel	Sequence ID Loongo.pschr.11 T.congo.pschr.10 Loongo.pschr.8 T.congo.pschr.9 T.congo.pschr.9 T.congo.pschr.3 Dened (1) All (7)	Basket Publ	Organism Organism		Length 4,815,855 3,840,014 2,469,386 2,264,776 2,203,460 1,480,841	Chromo	In case of Fron: Pix Based
My Stra Genomic Refresh	tegies: New Op : Sequence (20) Empty basket	Sequence ID Loongo.pschr.11 T.congo.pschr.10 Loongo.pschr.8 T.congo.pschr.9 T.congo.pschr.3 Dened (1) All (7) a strategy	Basket Publ	Organism Organism		Length 4 4,815,855 3,840,014 2,269,366 2,264,776 2,203,460 1,480,841	Chrome Three Chrome Three	In case of Error; Fix Based so during or are refind. WHY of Dawn for Da in the baset.
My Stra Genomic Refreah 20 Genoni	tegies: New Op s sequence (20) Empty basket (Save basket to mic Sequences	Sequence ID Loongo.pschr.11 T.congo.pschr.10 Loongo.pschr.8 T.congo.pschr.9 T.congo.pschr.3 Deneed (1) All (7) a strategy	Basket Publ	Organism Organism		4.815,855 3,840,014 2,469,386 2,264,776 2,203,460 1,480,841	Chromo The C	In case of Error; Fix Basted to charge or are refind. Why? of Da will not be in the basted. ease download your Da now.
My Stra Genomic 20 Genomic Genomic	tegies: New Op s sequence (20) Empty basket Sequence Results Sequence Results	Sequence ID Loongo.pschr.11 Loongo.pschr.13 Loongo.pschr.8 T.congo.pschr.9 T.congo.pschr.3 Deened (1) All (7) a strategy	Basket Publ	Organism Organism		IP Length • 4,815,855 3,340,014 • 2,469,386 2,264,776 • 2,203,460 1,480,841 •	Chrome Throme	In case of Error; Fir Balant to damp or over referd. WY? of Da will not be in the based. Bear download your Da now.
My Stra Genomic 20 Genonic Genomic I R	ttegies: New Or Sequence (20) Empty basket I Serve basket to mic Sequence Results Sequence Results tows per page: 20 \$	Sequence ID Loongo.pach.11 Loongo.pach.11 Loongo.pach.8 Loongo.pach.8 Loongo.pach.9 Loongo.pach.9 Loongo.pach.3 Denned (1) All (7) a strategy	Basket Publ	Organism Organism		IP Length • 4,815,855 3,340,014 • 2,469,386 2,264,776 • • 2,203,460 1 • • • 0H /Ds are m •	Chromo In	In case of Bror; Tri Basini to charge of a service without they to charge of a service without they? of Da will not be in the basistic ease downbad your Da now.
My Stra Genomic 20 Genomic Genomic R	ttegies: New Or sequence (20) Empty basket I Sequence Results tows per page: 20 \$ Life Sequence ID	Sequence ID Loongo.pschr.11 Loongo.pschr.11 Loongo.pschr.8 Loongo.pschr.9 Loongo.pschr.9 Loongo.pschr.3 a strategy Organism ② ②	Basket Publ	Organism Organism <porganism <p="">Organism <porganism <p="">Organism <porganism <p="">Organism <porganism <p="" <porganism="">Organism <porganism <="" <porganism="" p=""> Organism <porganism <="" <porganism="" p=""> Organism</porganism></porganism></porganism></porganism></porganism></porganism>		IP Length Q 4,815,855 3,340,014 2,469,386 2,264,776 2,264,776 2,203,460 1,480,841 1	Chromod	In case of Bror, Pic Baland to despe or any whited May? of Da will not be in the based. Add Columns Add Columns Project O
My Stra Genomic 20 Genomic Genomic Rafreah Genomic Genomic Genomic	tegies: New O; Sequence (20) Empty basket Save basket to mic Sequences Sequence Results Kows per page: 20 ¢ 1 Sequence ID T.congo.pschr.1	Sequence ID Loongo.psch.11 T.congo.psch.11 Loongo.psch.8 T.congo.psch.8 T.congo.psch.9 T.congo.psch.7 astrategy Composition	Basket Publ	Organism Organism <porganism <p="">Organism Organism Organism Organism <porganism <p="">Organism <porganism <="" <porganism="" p=""> Organism <porganism <="" <porganism="" p=""> Organis</porganism></porganism></porganism></porganism>	008kioc=T.congo.pschr.1:1.	IP Length O 4,815,855 3,340,014 2,469,386 2,264,776 2,203,460 1 2,203,460 1 Old /Da are m 6 8426078tra 8	Chromo II II IO B 7 9 3 On new releases ICI sometime apped to new file when possible. Releve to new releases ICI sometime apped to new file when possible. Releve to new releases ICI sometime apped to new file when possible. Releve Control of the possible of the possibl	In case of Front Pic Based is charge or are refind. Why? of Da will not be in the based of Da will not be in the based. Add Columns Add Columns Project O 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):
Nucleotide positions: 1 to 0 (0 = end)
Download Type: Text File Show in Browser Get Sequences

-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.



-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 jo	b 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 an even we first will be used to name estimation (such as games, people because and) approved during the				
annotation run. It : Example: LDON	should not contain spaces or special characters.				
annotation run. It Example: <i>LDON</i> Species prefix	LFOO				
Example: LDON Species prefix Finally, please pro Example: Leishma	LFOO vide a species name that describes the target species you are annotating. nia donovani				
Finals also give a annotation run. It : Example: <i>LDON</i> Species prefix Finally, please pro Example: <i>Leishma</i> Species name	LFOO vide a species name that describes the target species you are annotating. nia donovani 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



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 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 %

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.



Submit job

×