University of Glasgow

Wellcome Trust Workshop

Companion

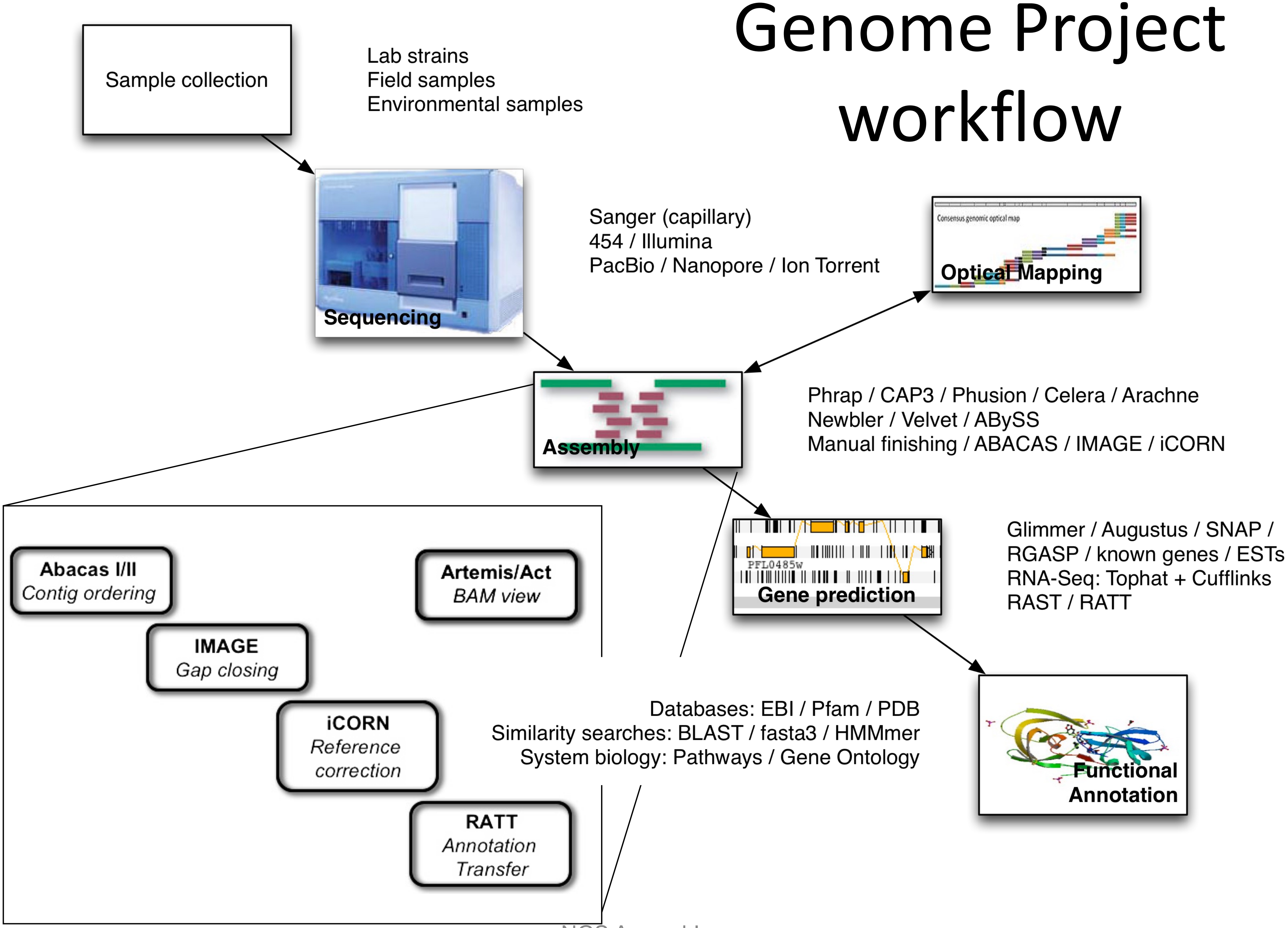
wellcome trust sanger institute

GeneDB

wellcome trust

EuPathDB
Eukaryotic Pathogen Database Resources
http://EuPathDB.org

Genome Project workflow



Sample collection

Lab strains
Field samples
Environmental samples

Sequencing

Sanger (capillary)
454 / Illumina
PacBio / Nanopore / Ion Torrent

Optical Mapping

Assembly

Phrap / CAP3 / Phusion / Celera / Arachne
Newbler / Velvet / ABySS
Manual finishing / ABACAS / IMAGE / iCORN

Gene prediction

Glimmer / Augustus / SNAP /
RGASP / known genes / ESTs
RNA-Seq: Tophat + Cufflinks
RAST / RATT

Functional Annotation

Databases: EBI / Pfam / PDB
Similarity searches: BLAST / fasta3 / HMMmer
System biology: Pathways / Gene Ontology

Abacas I/II
Contig ordering

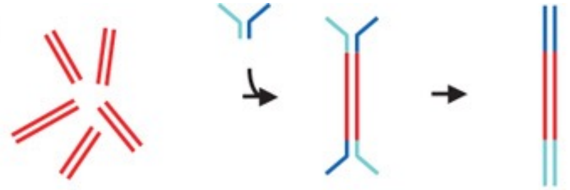
Artemis/Act
BAM view

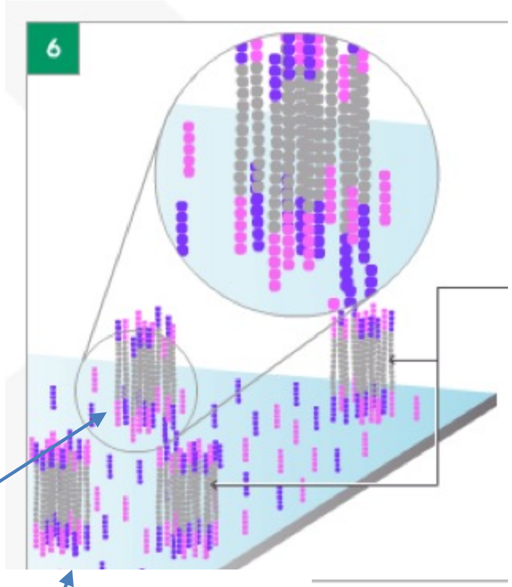
IMAGE
Gap closing

iCORN
Reference correction

RATT
Annotation Transfer

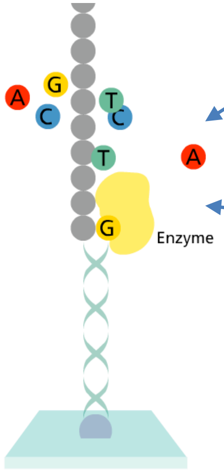
Illumina sequencing

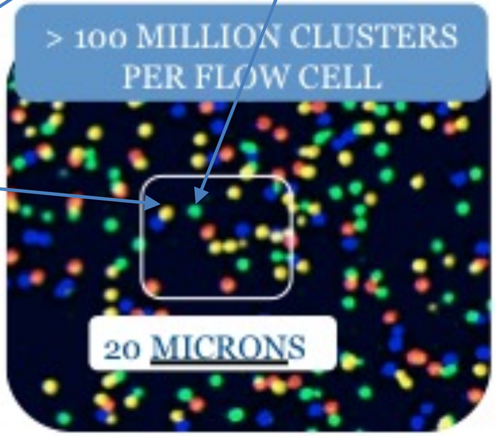




clusters

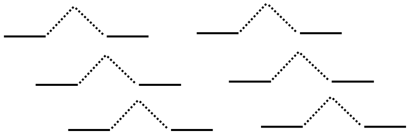
Cycle 1





A C
G T

Millions of reads pairs

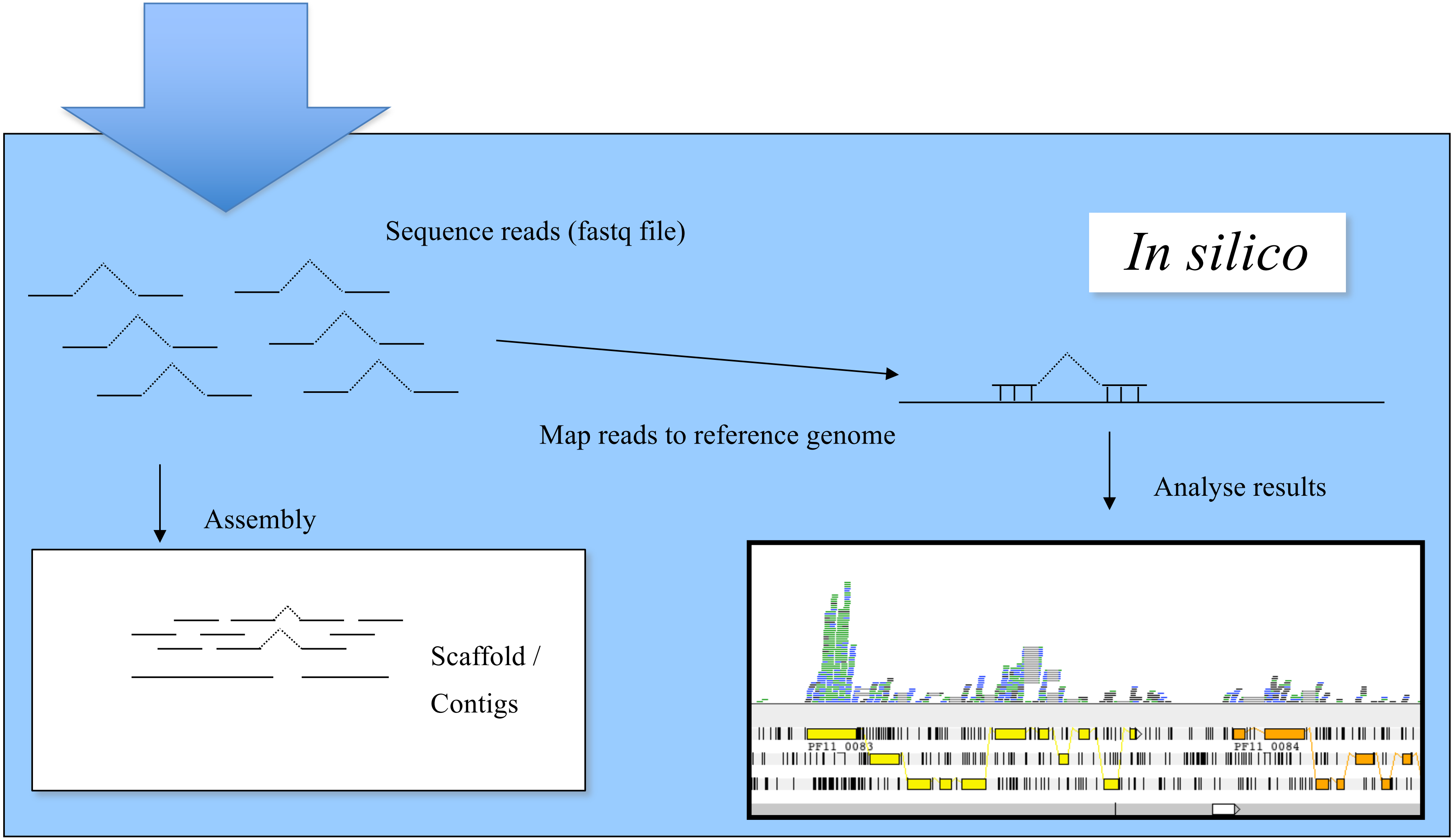


Many very short reads
Illumina X10 / HiSeq 4000

LSHTM'16

Sequencing

Not to scale



Sequence reads (fastq file)

In silico

Map reads to reference genome

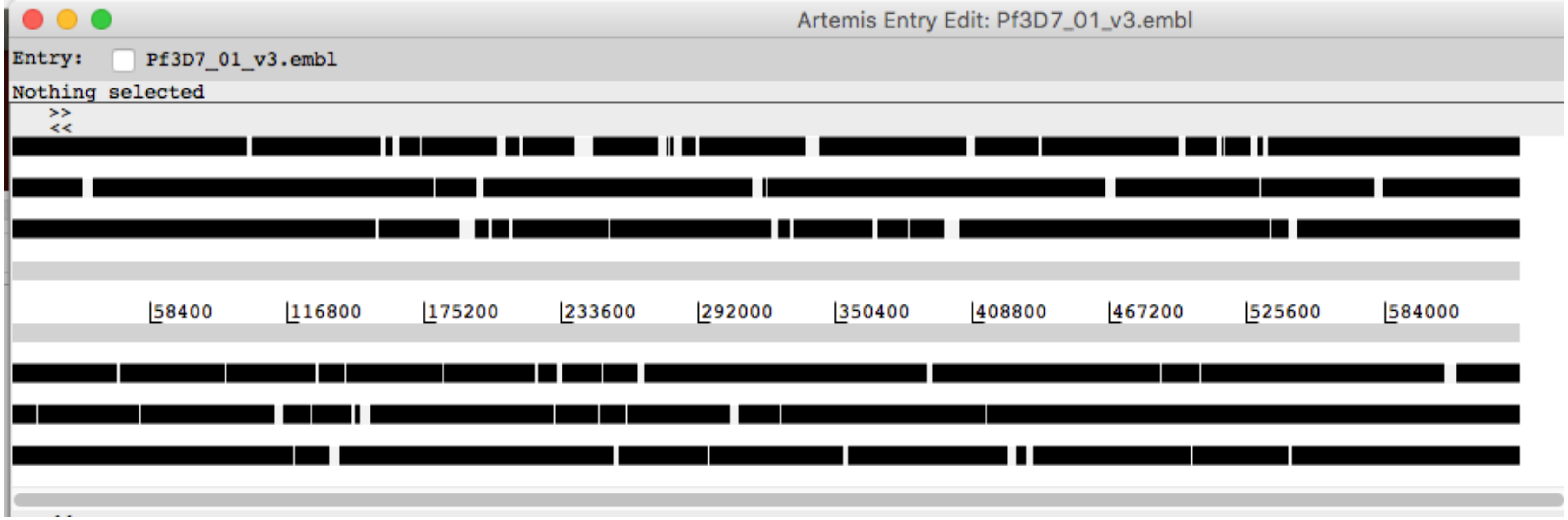
Analyse results

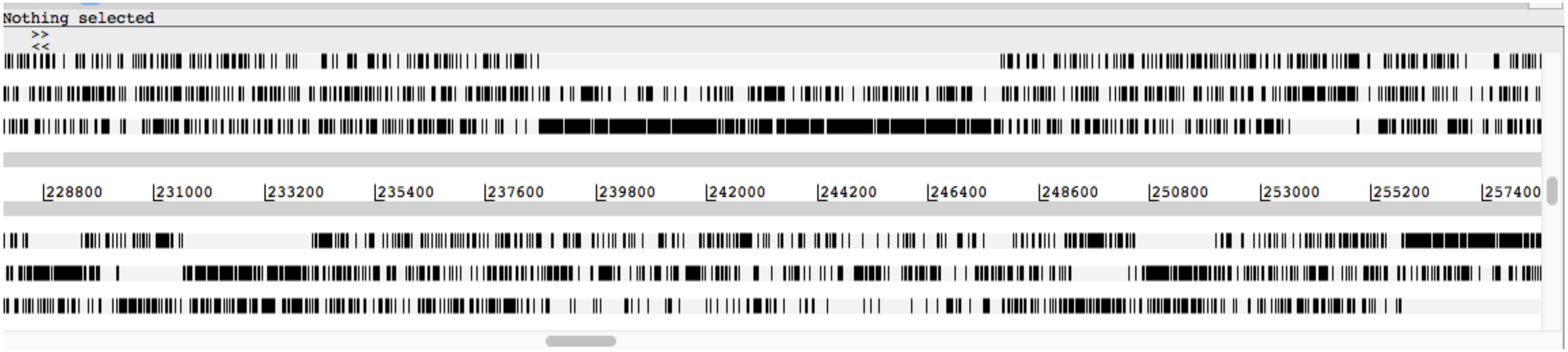
Assembly

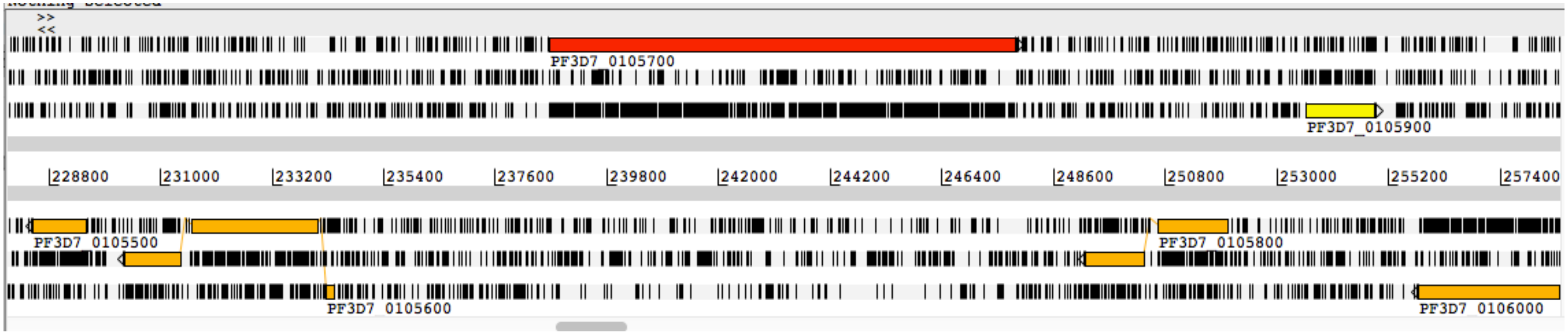
Scaffold /

Contigs

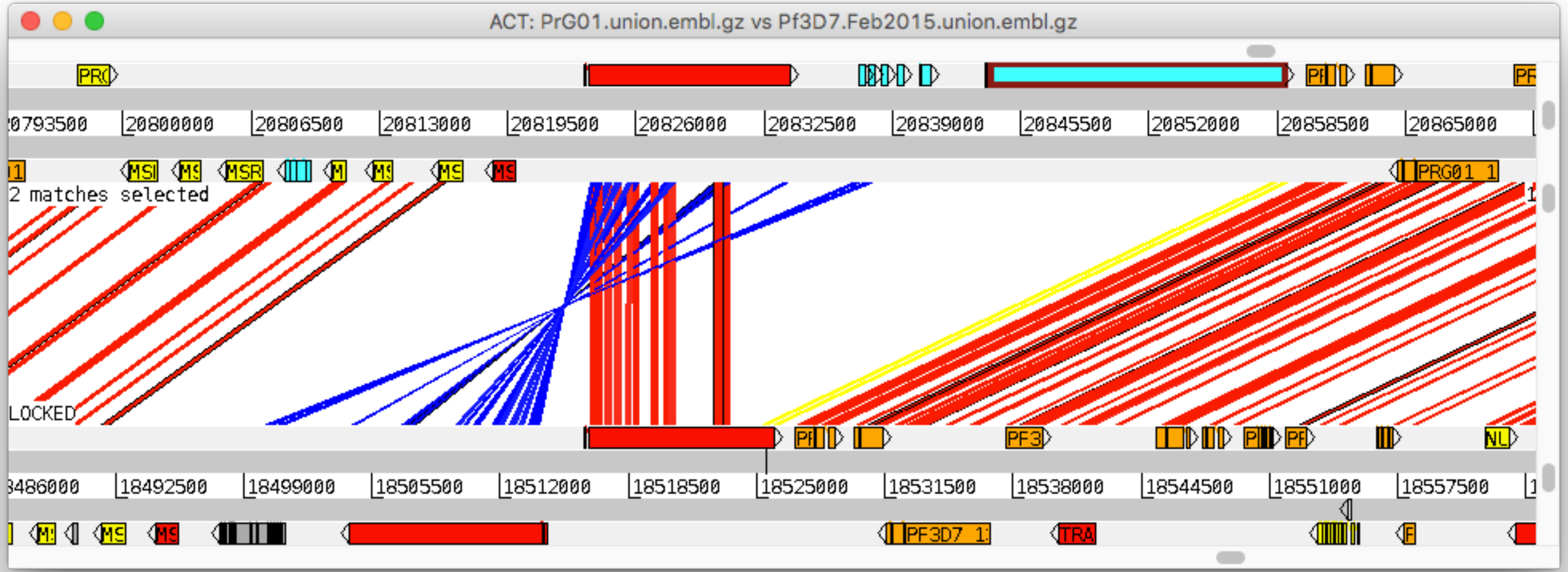
Why annotate a genome sequence?





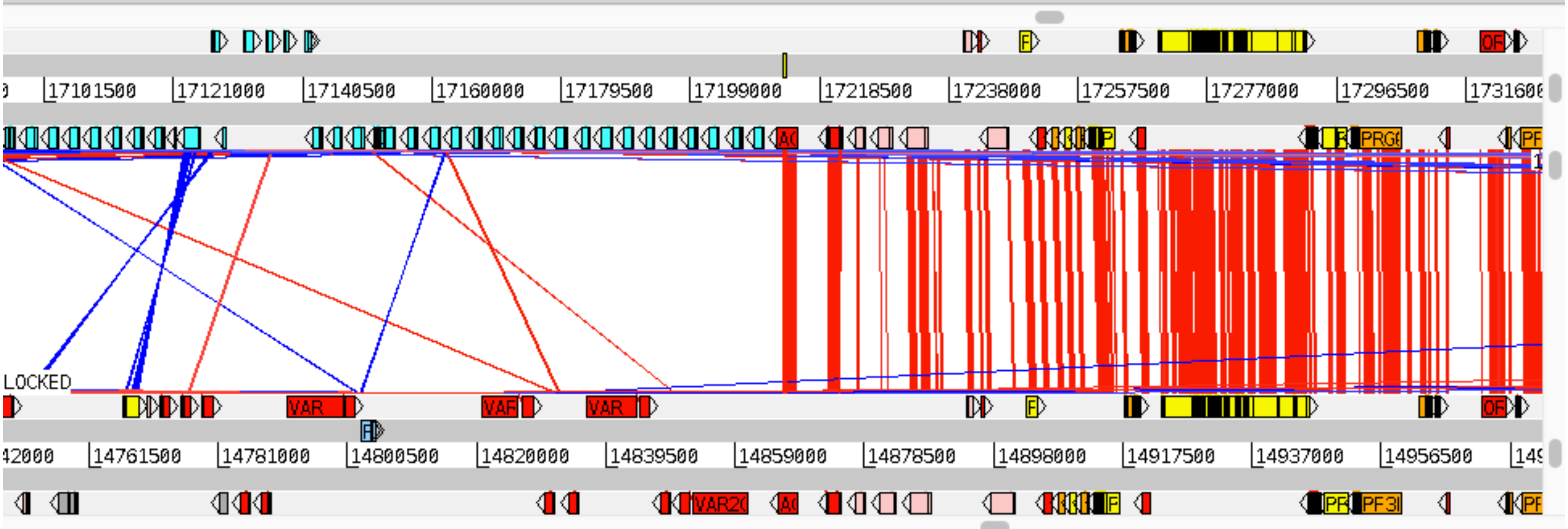


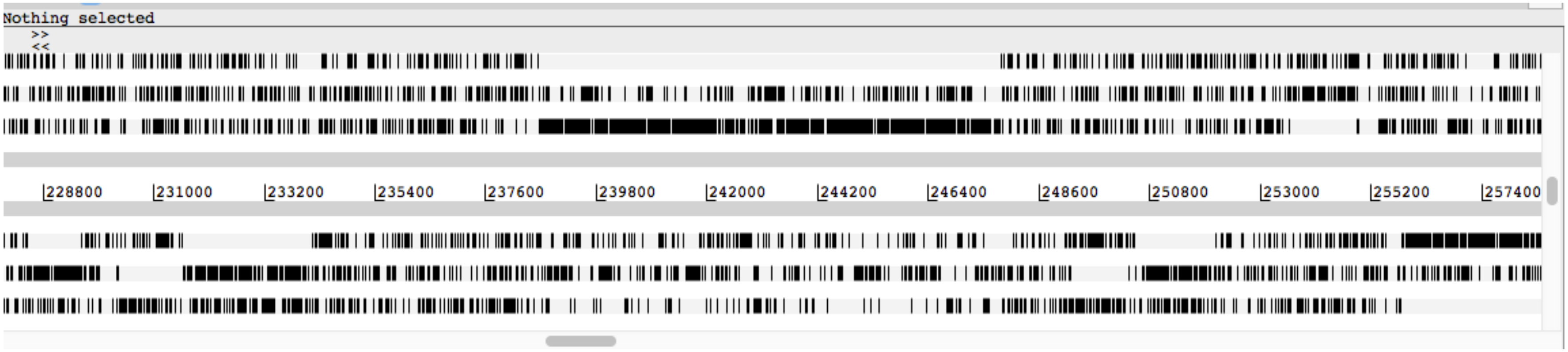
The aim to find differences on more diverse genes

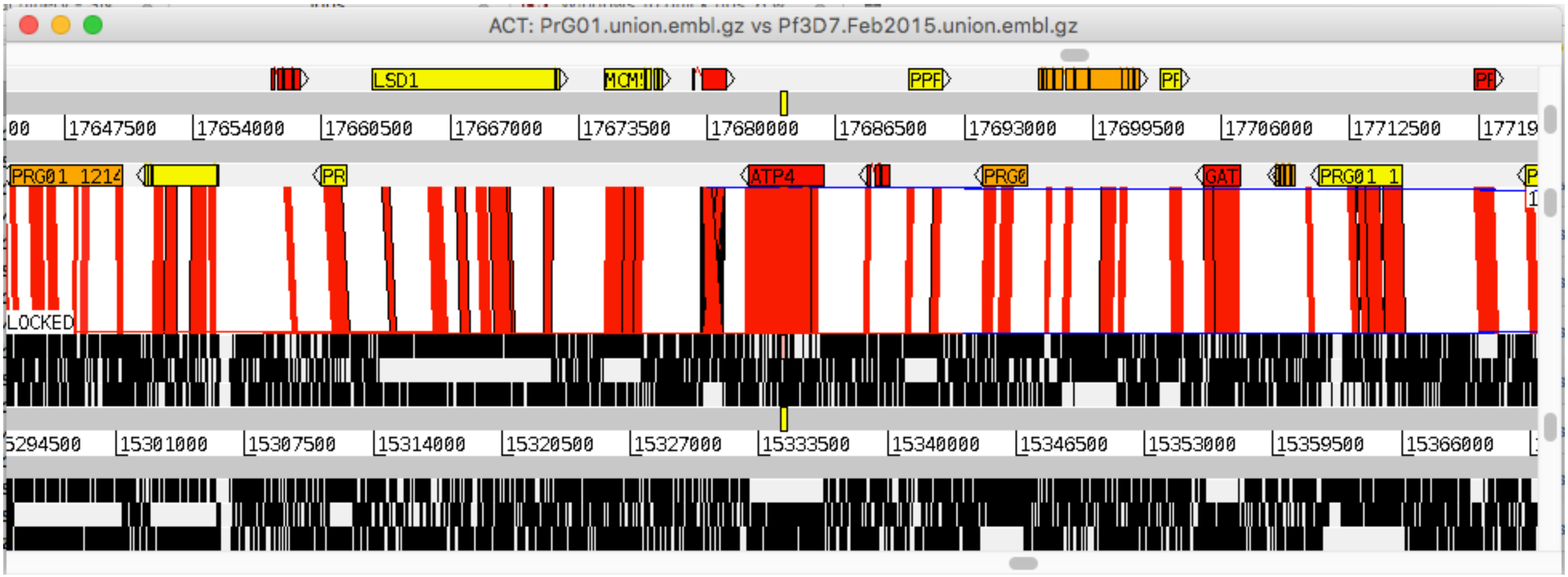


ACT: PrG01.union.embl.gz vs Pf3D7.Feb2015.union.embl.gz

2 matches selected




Nothing selected
>>
<<
228800
231000
233200
235400
237600
239800
242000
244200
246400
248600
250800
253000
255200
257400


ACT: PrG01.union.embl.gz vs Pf3D7.Feb2015.union.embl.gz
LSD1
MCM
PPF
PF
PF
00
17647500
17654000
17660500
17667000
17673500
17680000
17686500
17693000
17699500
17706000
17712500
17719
PRG01_1214
PR
ATP4
PRG0
GAT
PRG01_1
P
LOCKED
5294500
15301000
15307500
15314000
15320500
15327000
15333500
15340000
15346500
15353000
15359500
15366000

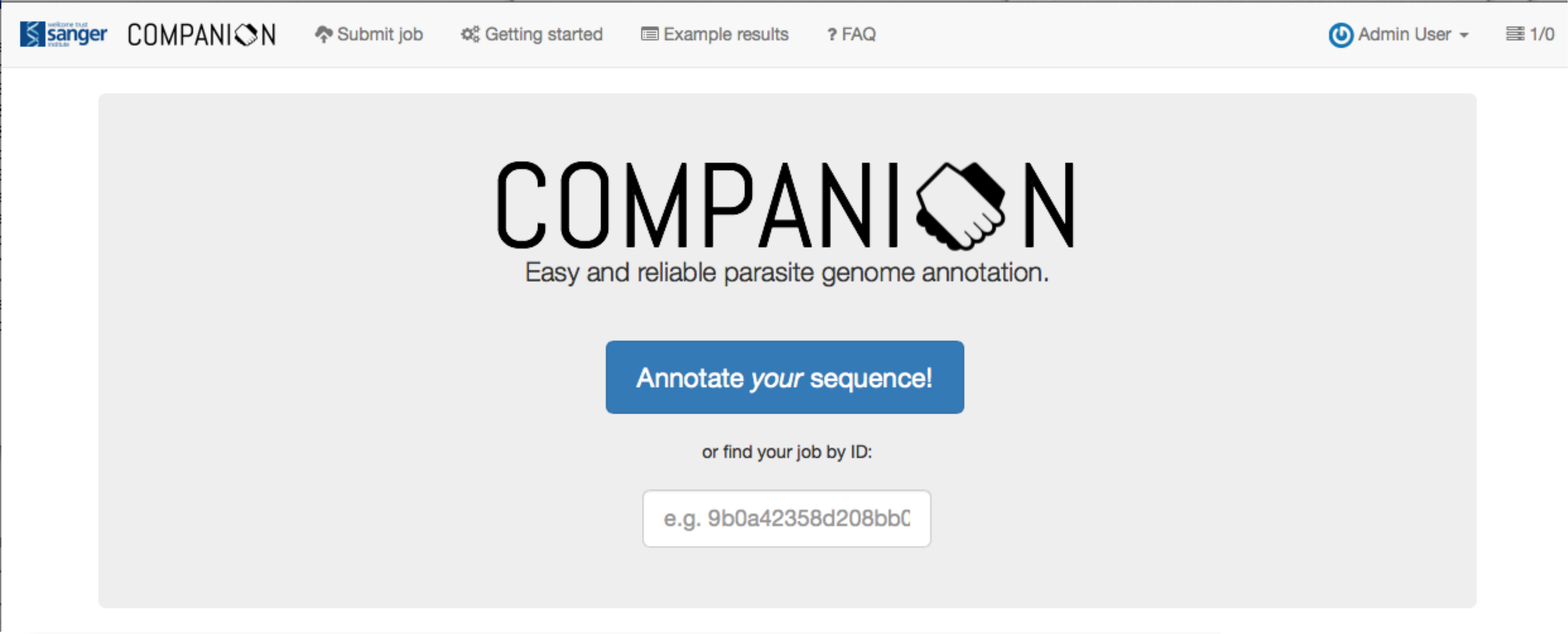
Annotation of genomes:

- To be good, it takes time
- RNA-Seq can help
- Functional annotation
- Submission to databases is an issue

There are many parasites that were unannotated, > 19 Leishmania

Idea:

Good annotation of a reference genome for a species class
Inherit the annotation
Well defined pipeline to overcome exchange issues



COMPANION

Easy and reliable parasite genome annotation.

Annotate *your* sequence!

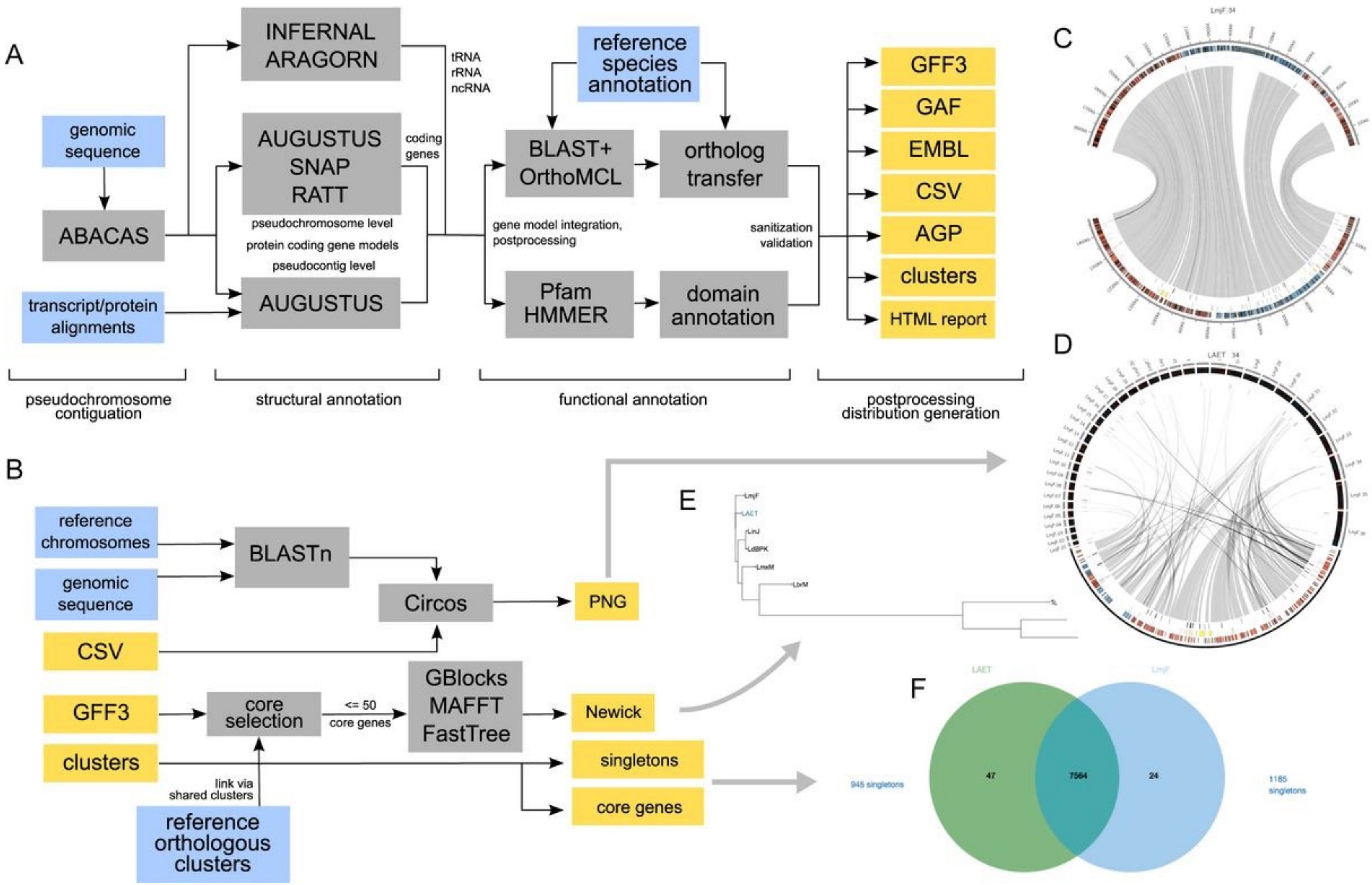
or find your job by ID:

e.g. 9b0a42358d208bb0

Companion: a web server for annotation and analysis of parasite genomes

Sascha Steinbiss[1,*], Fatima Silva-Franco[2], Brian Brunk[3], Bernardo Foth[1], Christiane Hertz-Fowler[2], Matthew Berriman[1] and Thomas D. Otto[1]

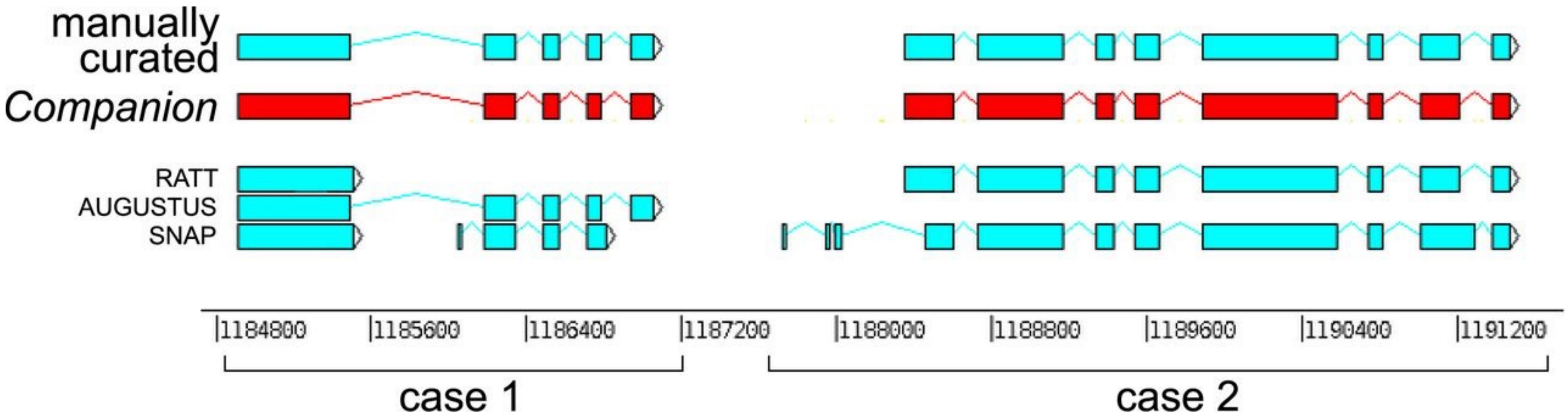
Schematic overview of the Companion workflows.



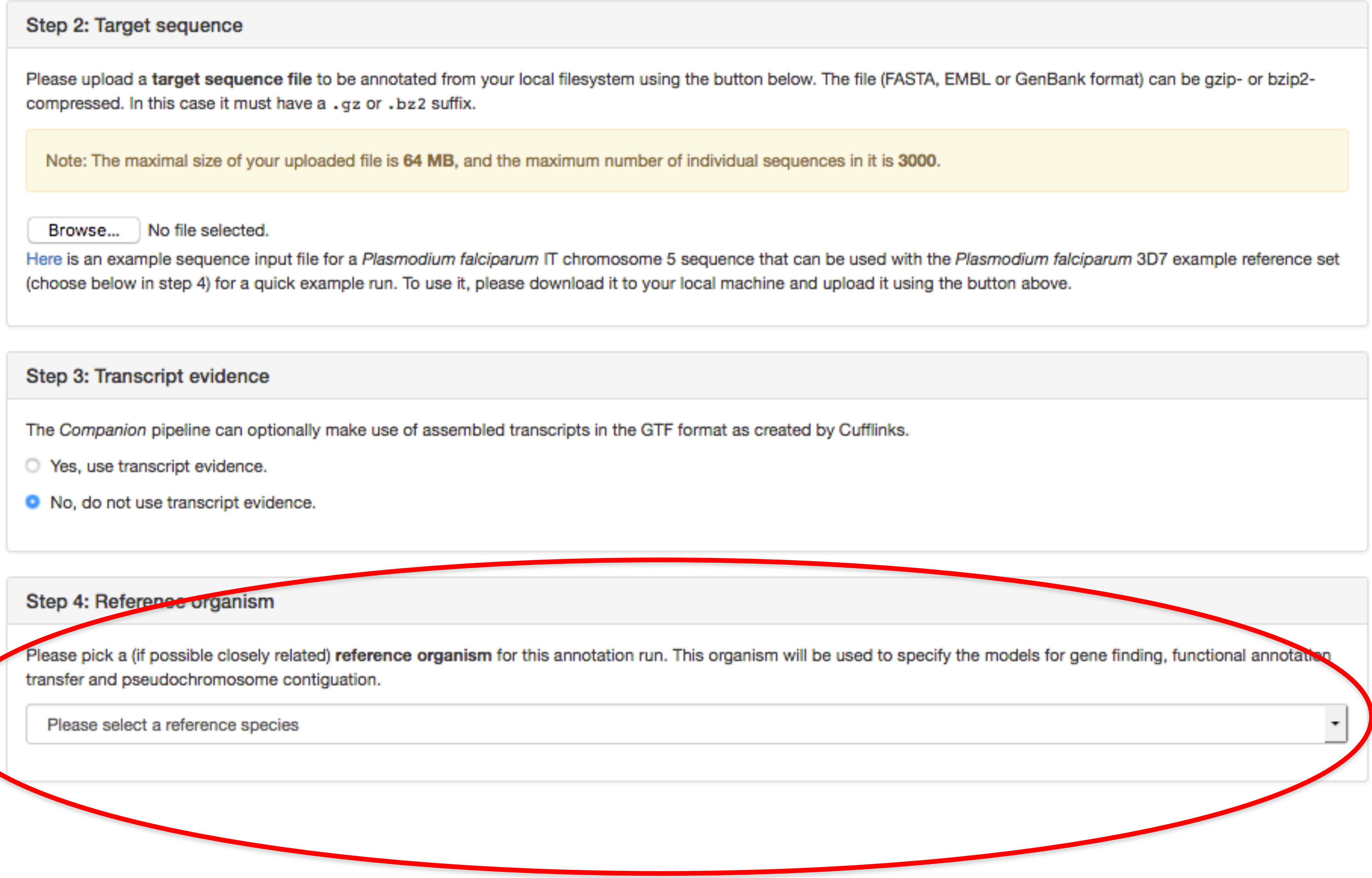
Sascha Steinbiss et al. Nucl. Acids Res. 2016;nar.gkw292

Nucleic Acids Research

Example of gene model integration across different gene finders.



Sascha Steinbiss et al. Nucl. Acids Res. 2016;nar.gkw292



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.

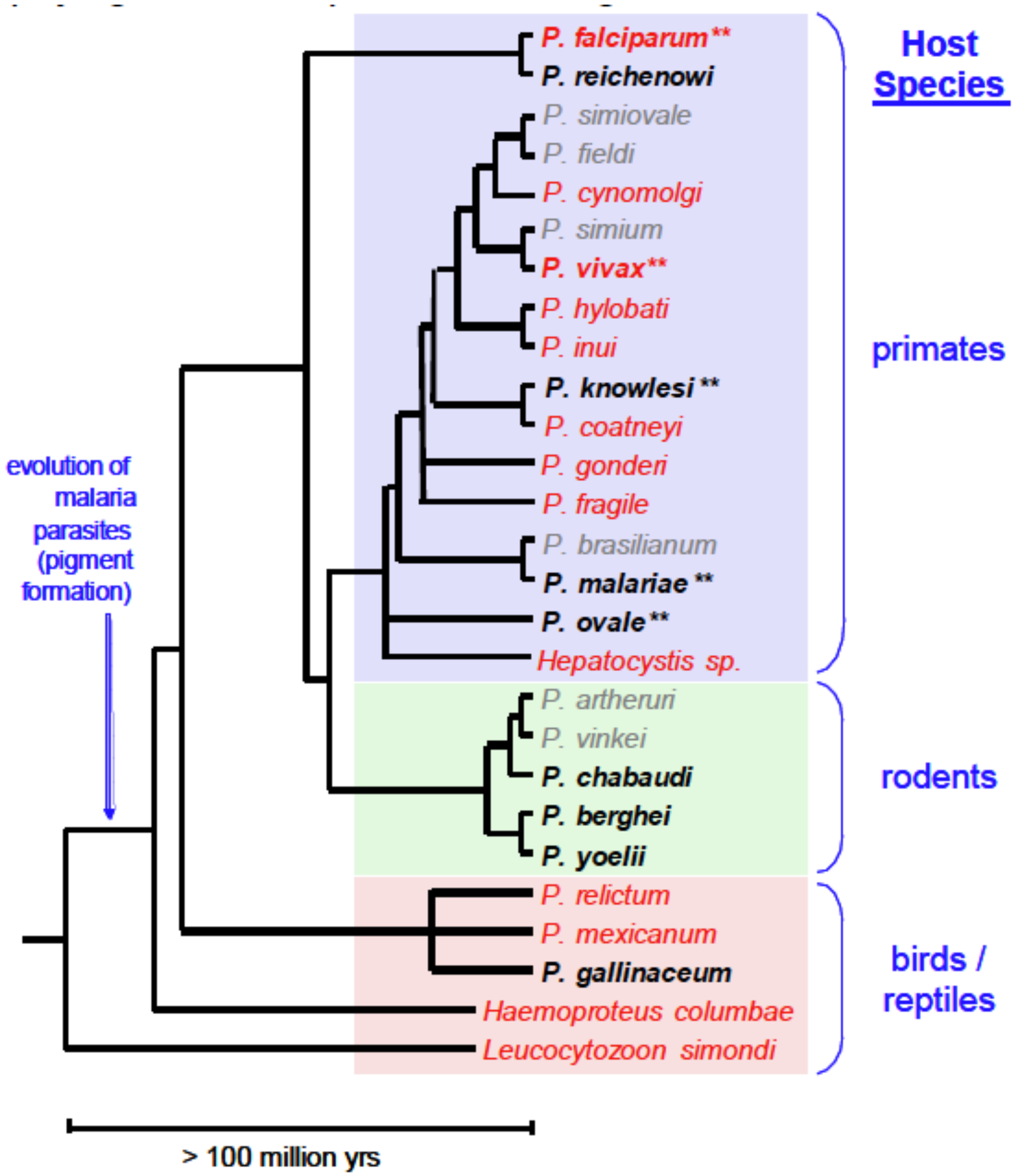
● No, do not use transcript evidence.

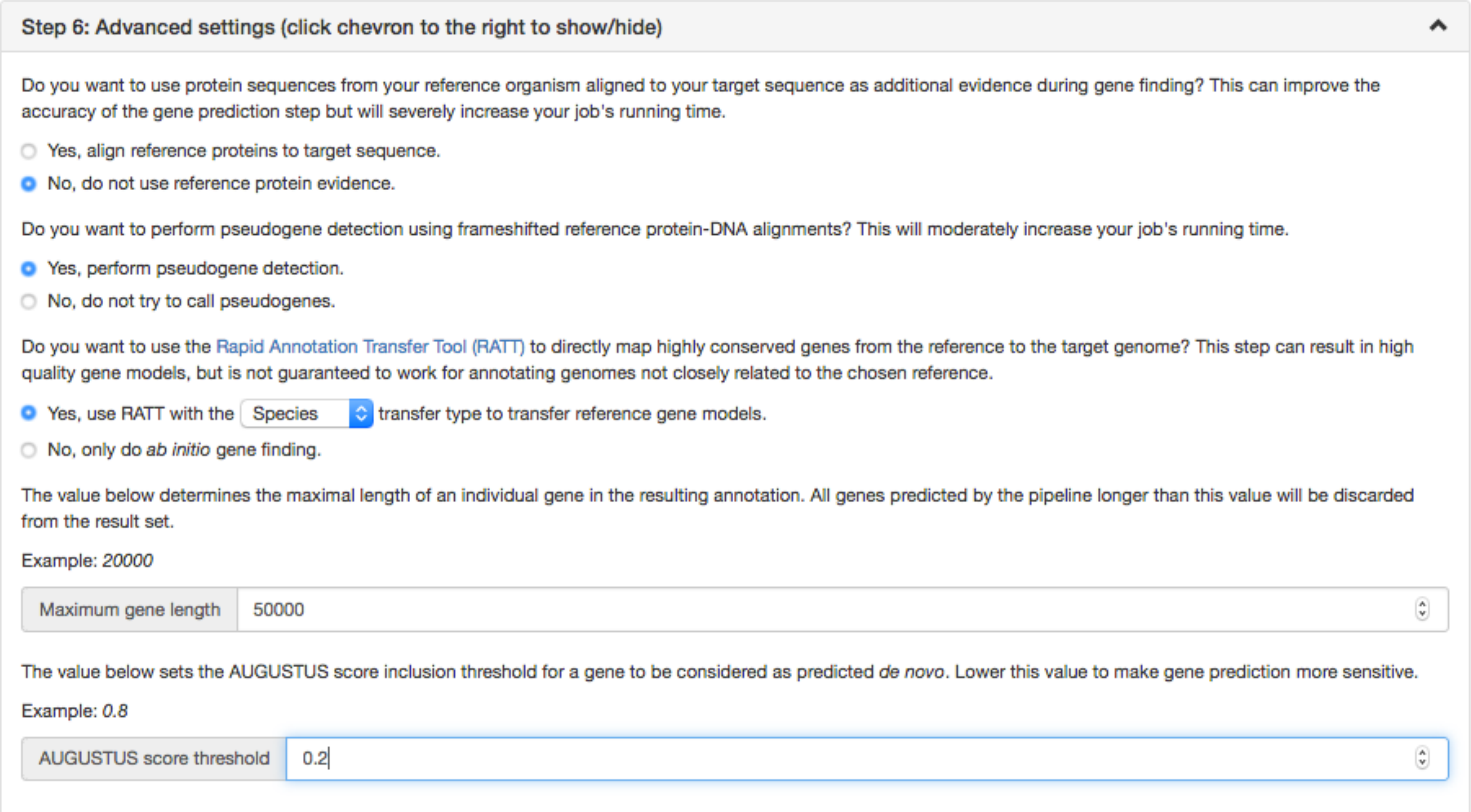
Step 4: Reference organism

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

What is a good reference?





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

Do you want to use protein sequences from your reference organism aligned to your target sequence as additional evidence during gene finding? This can improve the accuracy of the gene prediction step but will severely increase your job's running time.

○ Yes, align reference proteins to target sequence.

● No, do not use reference protein evidence.

Do you want to perform pseudogene detection using frameshifted reference protein-DNA alignments? This will moderately increase your job's running time.

● Yes, perform pseudogene detection.

○ No, do not try to call pseudogenes.

Do you want to use the Rapid Annotation Transfer Tool (RATT) to directly map highly conserved genes from the reference to the target genome? This step can result in high quality gene models, but is not guaranteed to work for annotating genomes not closely related to the chosen reference.

● Yes, use RATT with the Species transfer type to transfer reference gene models.

○ No, only do *ab initio* gene finding.

The value below determines the maximal length of an individual gene in the resulting annotation. All genes predicted by the pipeline longer than this value will be discarded from the result set.

Example: *20000*

Maximum gene length: 50000

The value below sets the AUGUSTUS score inclusion threshold for a gene to be considered as predicted *de novo*. Lower this value to make gene prediction more sensitive.

Example: *0.8*

AUGUSTUS score threshold: 0.2

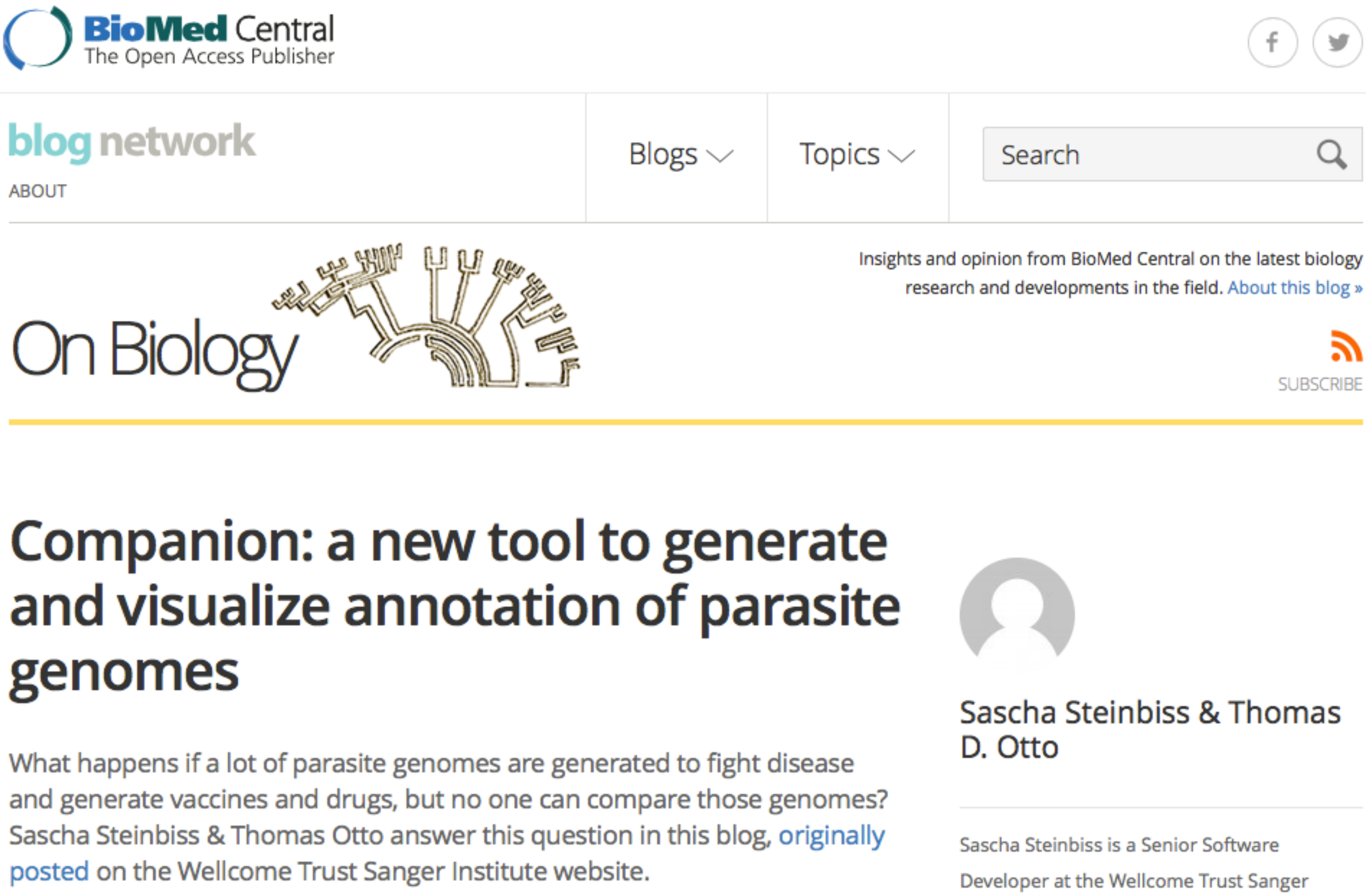
off you go… do it yourself.

google: companion sanger

This is group task

Omar kindly downloaded already the genome sequences

Go to page 62 in the manual



Insights and opinion from BioMed Central on the latest biology research and developments in the field. About this blog »

On Biology

Companion: a new tool to generate and visualize annotation of parasite genomes

What happens if a lot of parasite genomes are generated to fight disease and generate vaccines and drugs, but no one can compare those genomes? Sascha Steinbiss & Thomas Otto answer this question in this blog, originally posted on the Wellcome Trust Sanger Institute website.

Sascha Steinbiss & Thomas D. Otto

Sascha Steinbiss is a Senior Software Developer at the Wellcome Trust Sanger