

Comparative phage genomics

Marie-Agnès PETIT

Phage team of the MICALIS dept

INRAE, Jouy en Josas, FRANCE



Once all ORF are called and functions are predicted..

We have a **Genbank file**

- We would like to view globally the genetic map of this phage
- And to compare it to other, similar phages, to try assign it to a species and a genus

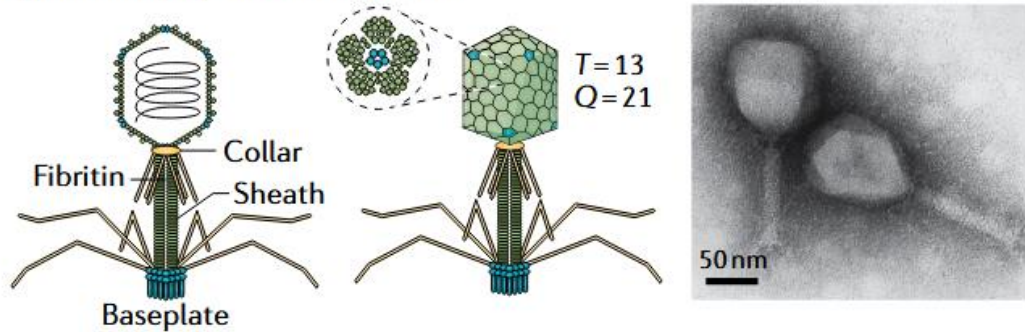
```
LOCUS       Lam4C10                49168 bp    DNA     linear   PHG
DEFINITION  Escherichia virus Lambda4C10, complete genome
ACCESSION   Lam4C10
KEYWORDS    .
SOURCE      Escherichia virus Lam4C10
  ORGANISM  Escherichia virus Lam4C10

FEATURES             Location/Qualifiers
     source            1..49168
                       /mol_type="genomic DNA"
                       /db_xref="taxon:66666666"
                       /genome_md5=""
                       /project="mapetit_66666666"
                       /genome_id="66666666.317980"
     CDS               35..451
                       /db_xref="SEED:fig|66666666.317980.peg.1"
                       /translation="MAIQWYAQRETDIENEKLRKELDDLRAAAESDLQPGTIDYERYR
LTKAQADAQELKNAREDGWLETELFTFILQRVAQEISGILVVRVPLTLQRKYPDISPS
HLDVVKTEIAKASNVAAGAGENVGGWIDDFRAEGS"
                       /product="Terminase small subunit"
                       /colour=3
                       /transl_table=11
     CDS               423..2351
                       /db_xref="SEED:fig|66666666.317980.peg.2"
                       /translation="MISDAQKAANAAGAIATGLLSLIIPVPLTTVQWANKHYL PKES
SYTPGRWETLPFQVGMNMGNDLIRTVNLIKSARVGYTKMLLGVEAYFIEHKSRNSL
LFQPTDAAEDFMKSHVEPTIRDVPALLELAPWFGRKHRDNTLTLKRFSSGVGFWCLG
GAAAKNYREKSVDVVVCYDELSSFEPDVEKEGSPTLLGDKRIEGSVWPKSIRGSTPKIK
GSCQIEKAANESAHFMRFYVPCPHCGEEQYLKFGDDASPFGLKWEKNKPESVFYLCEH
HGCVIHQSELDQSNRWICENTGMWTRDGLMFFSARGDEIPPPRSITFHIWTAYSPFT
TWVQIVYDWLDALKDPNGLKTFVNTTLGETWEEAVGEKLDHQVLMDKVRYTAAVPAR
VWYLTAGIDSQRNRFEMYVWGWAPGEEAFLVDKIIIMGRPDEEETLLRVDAAINKKYR
HADGTEMTISRVCWDIGGIDGEIVYQRSKKHGVRVLPVKGASVYGKPVITMPKTRNQ
```

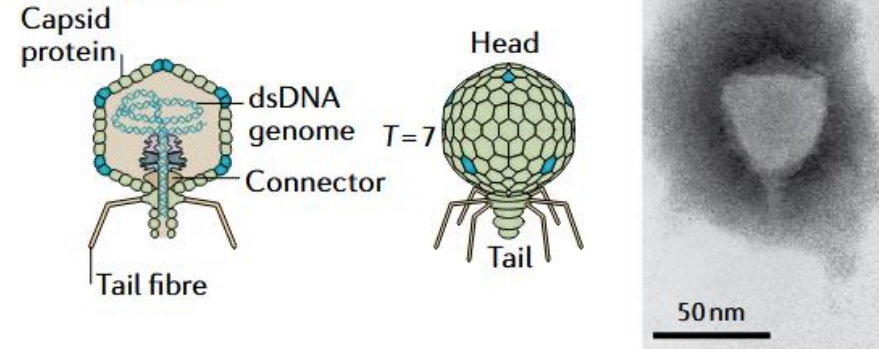
The main functions encoded by a phage genome

Tailed

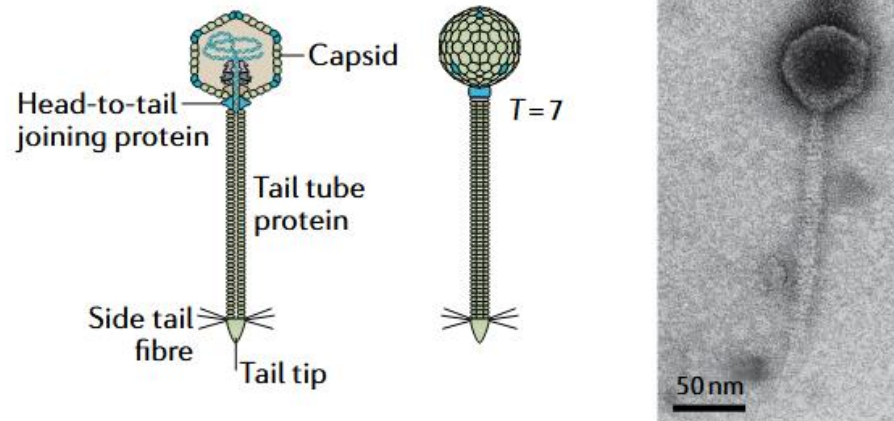
Myoviridae (T4) and Herelleviridae



Podoviridae (T7)



Siphoviridae (lambda)

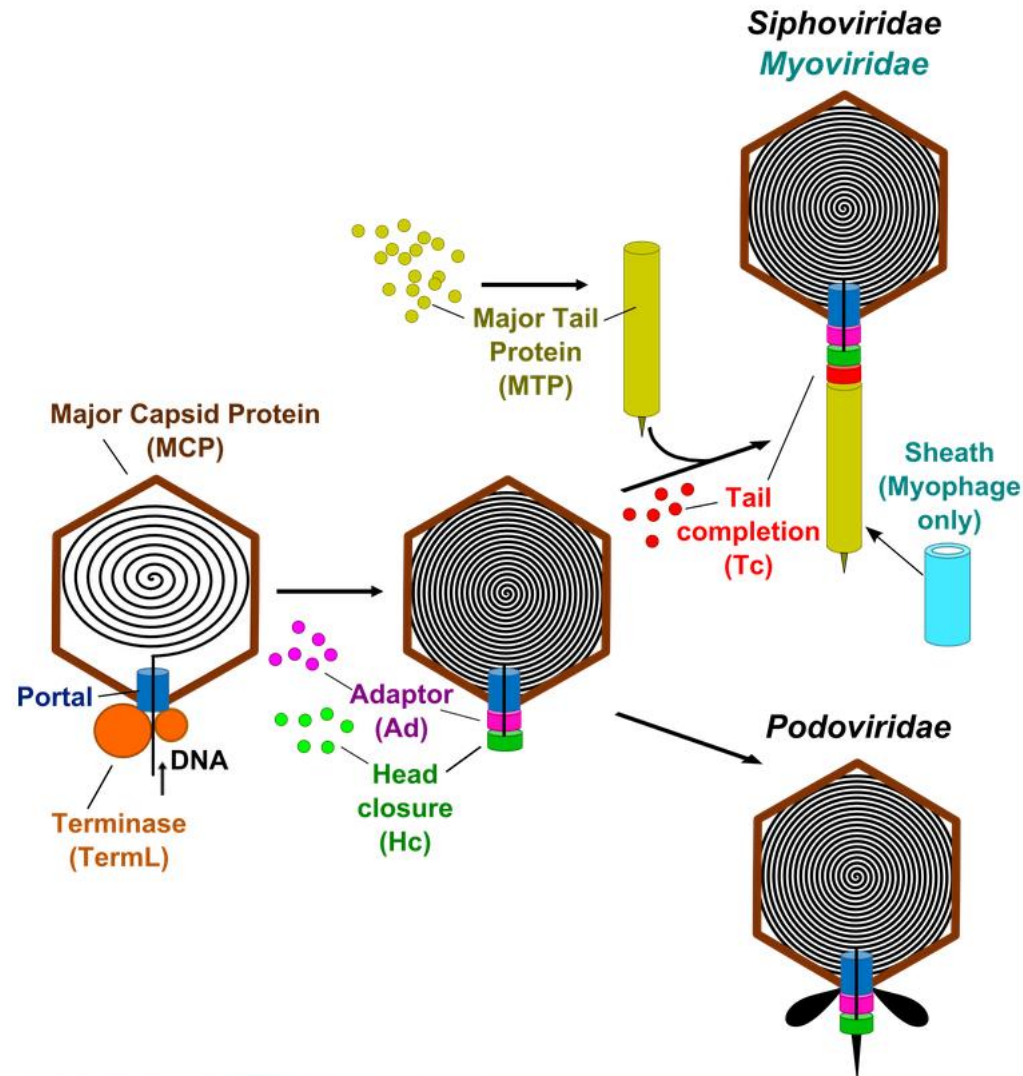


Dion et al., Nature Reviews Microbiol. 2020

3 categories of structural components for tailed phages:
Head (or capsid)
Connector
Tail

Of note: Particle shape no longer used for taxonomy, we now say myovirus/siphovirus/podovirus

Conserved functions inside each category



Head:

MCP, decoration, minor head proteins

Portal, Terminase L and S

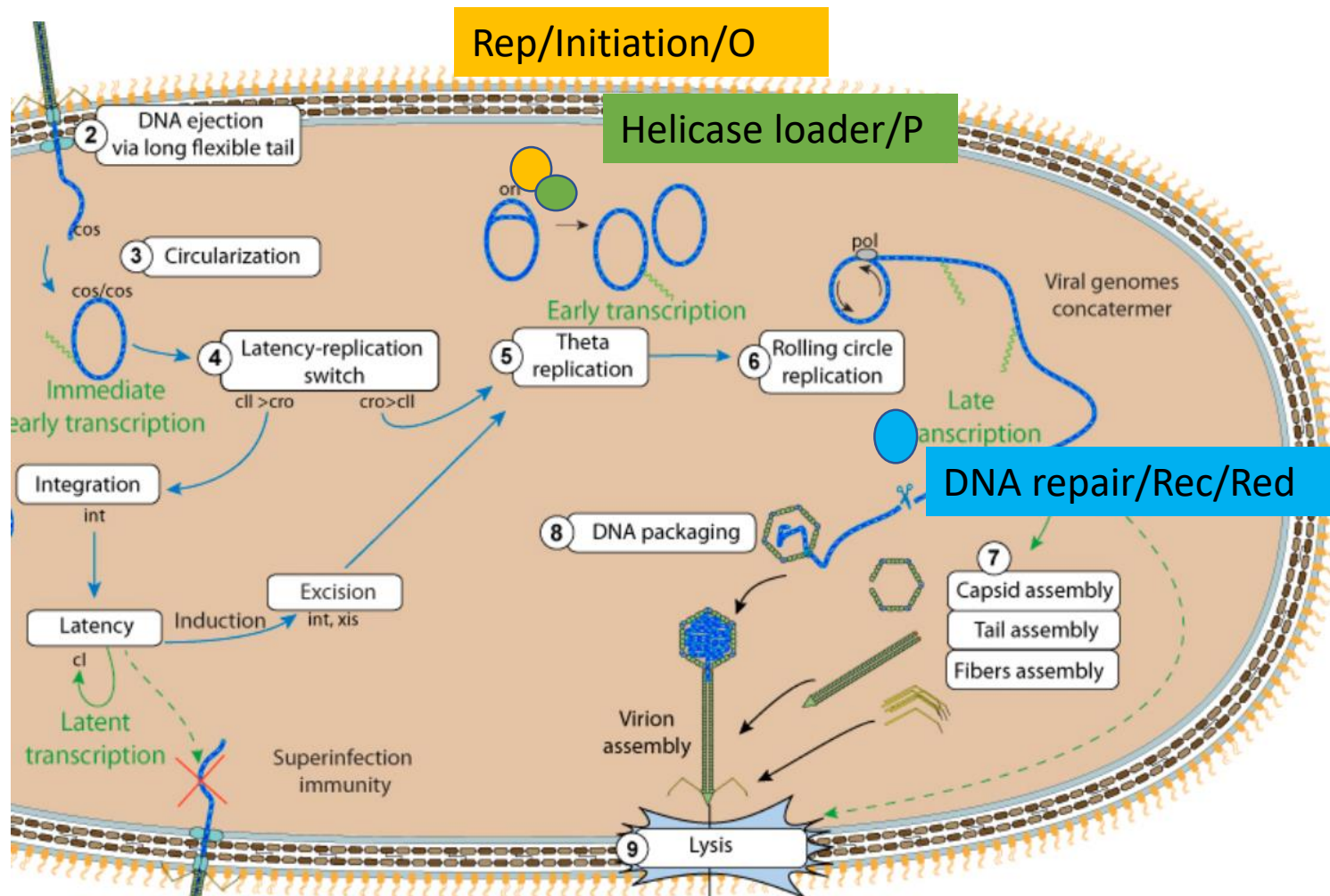
Connector:

Adaptor, head closure, Tail completion (tail terminator/neck)

Tail:

MTP (or tail tube), Sheath, Tail length tape measure, base plate, tail fiber, receptor binding RBP

The replication module



- Replication initiation

- DNA repair

Often taken from the host:

DNA polymerase

Replicative DNA helicase

RNA polymerase

Ribosomes

...

Once all ORF are called and functions are predicted..

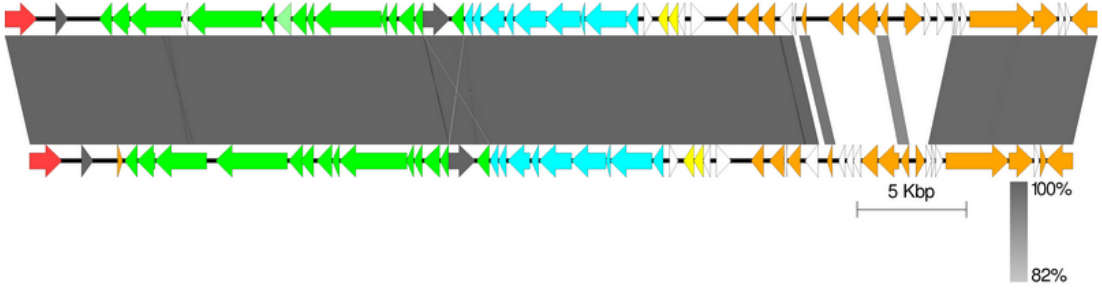
1

- We have a Genbank file
- Global genetic map

→ Use Easyfig

Takes Genbank files as entry

Other softwares: Genoplots, clinker



Easyfig
Easy genome comparison figures.

Example files

Example 1
Comparison of two ~50 Kb prophage regions from different *Salmonella* genomes. - [Easyfig_example1_files.zip](#)

Example 2
Creating a comparison figure of three *Escherichia coli* genomes. - [Easyfig_example2_files.zip](#)

Windows
Easyfig - [Easyfig_2.2.5_win.zip](#)

OSX
Easyfig - [Easyfig_2.2.2_OSX.zip](#)

Linux
Easyfig - [Easyfig_2.2.2_linux.tar.gz](#)

Preparing the Genbank file

- Create a copy of the gbk file
- Next to each gene with a function predicted, insert a field colour

PHROG functional category

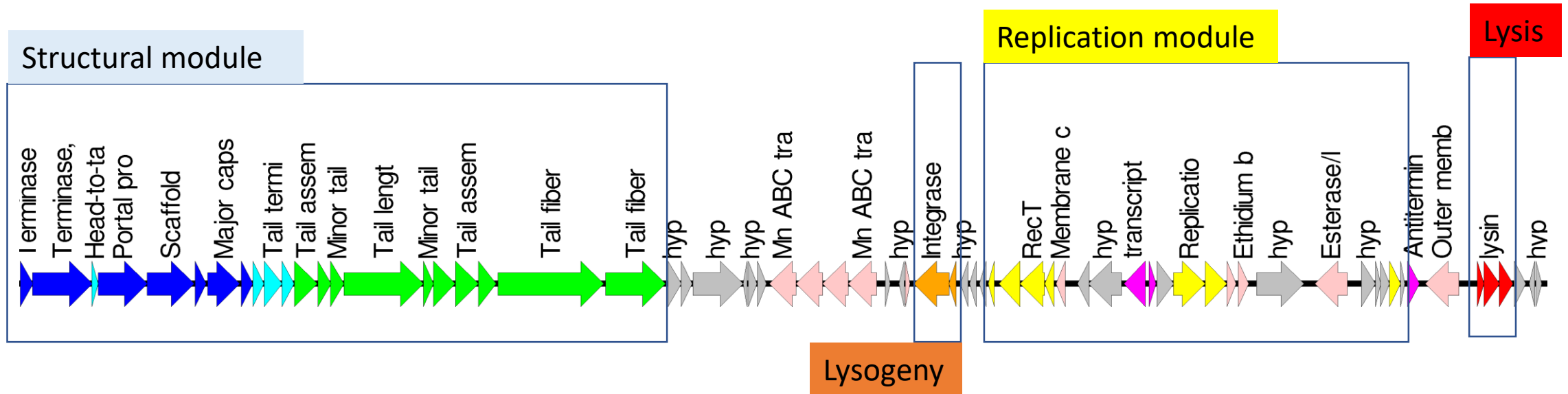
4	Head
5	Connector
3	Tail
7	DNA, RNA and nt metabolism
2	Lysis (holin, endolysin,...)
10	Integration/excision
6	Transcriptional regulator
12	Morons, AMG, host take-over
	Unknown function (default colour)

```
LOCUS      Lam4C10                49168 bp    DNA    linear    PHG
DEFINITION Escherichia virus Lambda4C10, complete genome
ACCESSION  Lam4C10
KEYWORDS   .
SOURCE     Escherichia virus Lam4C10
           ORGANISM  Escherichia virus Lam4C10

FEATURES             Location/Qualifiers
     source            1..49168
                       /mol_type="genomic DNA"
                       /db_xref="taxon:6666666"
                       /genome_md5=""
                       /project="mapetit_6666666"
                       /genome_id="6666666.317980"
     CDS               35..451
                       /db_xref="SEED:fig|6666666.317980.peg.1"
                       /translation="MAIQWYAQRETDIENEKLRKELDDLRAAAESDLQPGTIDYERYR
LTKAQADAQELKNAREDGWVLETELFTFILQRVAQEISGILVRVPLTLQRKYPDISPS
HLDVVKTEIAKASNVAAKAGENVGGWIDDFRAEGS"
                       /product="Terminase small subunit"
                       /colour=3
                       /transl_table=11
     CDS               423..2351
                       /db_xref="SEED:fig|6666666.317980.peg.2"
                       /translation="MISDAQKAANAAGAIATGLLSLIIPVPLTTVQWANKHYLKPES
SYTPGRWETLPFQVGIMNCMGNDLIRTVNLIK SARVGYTKMLLGVEAYFIEHKSRSNL
LFQPTDSAAEDFMKSHVEPTIRDVPALLELAPWFGKRHRDNTLLTKRFSSGVGFWCLG
GAAAKNYREKSVDVWCYDELSSFEPDVEKEGSPTLLGDKRIEGSVWPKSIRGSTPKIK
GSCQIEKAANESAHFMRFYVPCPHCGEEQYLKFGDDASPFGLKWEKNKPESV FYLCEH
HGCVIHQSELDQSNRWICENTGMWTRDGLMFFSARGDEIPPPRSITFHIWTAYSPFT
TWVQIVYDWDALKDPNGLKTFVNTT LGETWEEAVGEKLDHQVLMDKVVR YTA AVPAR
VWYLTAGIDSQRNRFEMYVWG WAPGEEAFLVDKIIIMGRPDEEETLLRVDAAINKKYR
HADGTEMTISRVCWDIGGIDG EIVYQRSKKGHGVFRVLPVKGASVYGKPVITMPKTRNQ
```

Easyfig result

Lam4C10



Once all ORF are called and functions are predicted..

2

- We have a Genbank file
- We would like to contemplate the genetic map of this phage
- And to compare it to other, similar phages, to try assign it to a species and a genus

```
LOCUS      Lam4C10                49168 bp    DNA    linear    PHG
DEFINITION Escherichia virus Lambda4C10, complete genome
ACCESSION  Lam4C10
KEYWORDS   .
SOURCE     Escherichia virus Lam4C10
ORGANISM   Escherichia virus Lam4C10
```

TAXONOMY?

```
FEATURES             Location/Qualifiers
     source            1..49168
                        /mol_type="genomic DNA"
                        /db_xref="taxon:6666666"
                        /genome_md5=""
                        /project="mapetit_6666666"
                        /genome_id="6666666.317980"
     CDS               35..451
                        /db_xref="SEED:fig|6666666.317980.peg.1"
                        /translation="MAIQWYAQRETDIENEKLRKELDDLRAAAESDLQPGTIDYERYR
                        LTKAQADAQELKNAREDGWVLETELFTFILQRVAQEISGILVRVPLTLQRKYPDISPS
                        HLDVVKTEIAKASNVAAKAGENVGGWIDDFRRAEGS"
                        /product="Terminase small subunit"
                        /colour=3
                        /transl_table=11
     CDS               423..2351
                        /db_xref="SEED:fig|6666666.317980.peg.2"
                        /translation="MISDAQKAANAAGAIATGLLSLIIPVPLTTVQWANKHYLKPES
                        SYTPGRWETLPFQVGIMNCMGNDLIRTVNLIKSARVGYTKMLLGVEAYFIEHKSRSNL
                        LFQPTDSAAEDFMKSHVEPTIRDVPALLELAPWFGRKHRDNTLLTKRFSSGVGFWCLG
                        GAAAKNYREKSVDVWCYDELSSFEPDVEKEGSPTLLGDKRIEGSVWPKSIRGSTPKIK
                        GSCQIEKAANESAHFMRFYVPCPHCGEEQYLKFGDDASPFGLKWEKNKPESVFYLCEH
                        HGCVIHQSELDQSNRWICENTGMWTRDGLMFFSARGDEIPPPRSITFHIWTAYSPFT
                        TWVQIVYDWDLDALKDPNGLKTFVNTTLGETWEEAVGEKLDHQVLMDKVVRYTAAVPAR
                        VVYLTAGIDSQRNRFEMYVWGWPGEAEFLVDKIIIMGRPDEEETLLRVDAAINKKYR
                        HADGTEMTISRVCWDIGGIDGEIVYQRSKKGHGVFRVLPVKGASVYGKPVITMPKTRNQ
```

A simple and efficient first trial: whole genome alignment to the NCBI viral database

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

Mon, 21 Jul 2025

Here are a few highlights in our latest BLAST+ release:

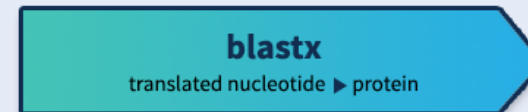
Download BLAST+ 2.17.0 now!

[More BLAST news...](#)

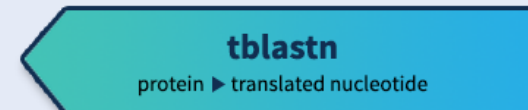
Web BLAST



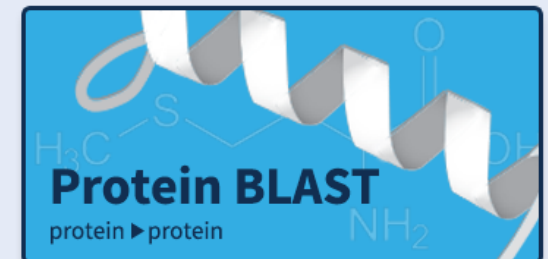
Nucleotide BLAST
nucleotide ▶ nucleotide



blastx
translated nucleotide ▶ protein



tblastn
protein ▶ translated nucleotide



Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human

Mouse

Rat

Microbes

How to run a nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide database

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file Aucun fichier sélectionné. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Core nucleotide database (core_nt) [?](#)

Organism Optional exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)

1. Upload a **fasta file** of your genome:

```
>Lam4C10
TAGTAAGGGCATTGAATCTGTATTTGATACTGCCATGGCAATTCAGTGGT
ATGCGCAGAGGGGAAACTGATATCGAAAAACGAAAAGCTCCGCAAAGAAGT
GACGATTTGCGTGC GGCCAGCGGAGTCAGATTTACAACCCGGCCACCATTGA
CTATGAACGCTACCGGCTCACAAAAGCGCAGGCAGATGCGCAGGAACTGA
AAAATGCCCGTGAAGACGGAGTAGTGCTGGAAAAGTGAAGTGTTCACCTTC
ATTCTGCAACGTGTGGCACAGGAGATTTCCGGGGATACTTGTGCGTGTGCC
GTTGACATTACAGCGTAAATATCCGGACATTTACCATCACACCTTGATG
TGGTGAAAAGTGAAGTTCGCGAAAAGCCTCCAATGTTGCAGCTAAGGCCGGT
GAAAACGTGGGCGGGTGGATCGATGATTTTCAGACGCGCAGAAGGCAGCTA
```

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide database

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file Aucun fichier sélectionné. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Core nucleotide database (core_nt) [?](#)

Organism Optional exclude [?](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)

1. Upload a fasta file of your genome:

```
>Lam4C10
TAGTAAGGGCATTGAATCTGTATTTGATACTGCCATGGCAATTCAGTGGT
ATGCGCAGAGGGAAAACGATATCGAAAAACGAAAAGCTCCGCAAAGAAGT
GACGATTTGCGTGCGGCAGCGGAGTCAGATTTACAACCCGGCACCATTGA
CTATGAACGCTACCGGCTCACAAAAGCGCAGGCAGATGCGCAGGAACTGA
AAAATGCCCGTGAAGACGGAGTAGTGCTGGAAAACGAACTGTTTACCTTC
ATTCTGCAACGTGTGGCACAGGAGATTTCCGGGGATACTTGTGCGTGTGCC
GTTGACATTACAGCGTAAATATCCGGACATTTACCATCACACCTTGATG
TGGTGAAAACGTAATCGCGAAAAGCCTCCAATGTTGCAGCTAAGGCCGGT
GAAAACGTGGGCGGGTGGATCGATGATTTTCAGACGCGCAGAAGGCAGCTA
```

2. Align it against all sequenced viral genomes

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide database

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file Aucun fichier sélectionné. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Core nucleotide database (core_nt) [?](#)

Organism Optional exclude [?](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast) [?](#)

1. Upload a fasta file of your genome:

```
>Lam4C10
TAGTAAGGGCATTGAATCTGTATTTGATACTGCCATGGCAATTCAGTGGT
ATGCGCAGAGGGAAAACGATATCGAAAACGAAAAGCTCCGCAAAGAAGTGA
GACGATTTGCGTGC CGGCAGCGGAGTCAGATTTACAACCCGGCACCATTGA
CTATGAACGCTACCGGCTCACAAAAGCGCAGGCAGATGCGCAGGAACTGA
AAAATGCCCGTGAAGACGGAGTAGTGCTGGAAAACGAACTGTTTACCTTC
ATTCTGCAACGTGTGGCACAGGAGATTTCCGGGGATACTTGTGCGTGTGCC
GTTGACATTACAGCGTAAATATCCGGACATTTACCATCACACCTTGATG
TGGTGAAAACGTAATCGCGAAAAGCCTCCAATGTTGCAGCTAAGGCCGGT
GAAAACGTGGGCGGGTGGATCGATGATTTTCAGACGCGCAGAAGGCAGCTA
```

2. Align it against all sequenced viral genomes

3. Megablast is the default, rapid version of Blast

Result

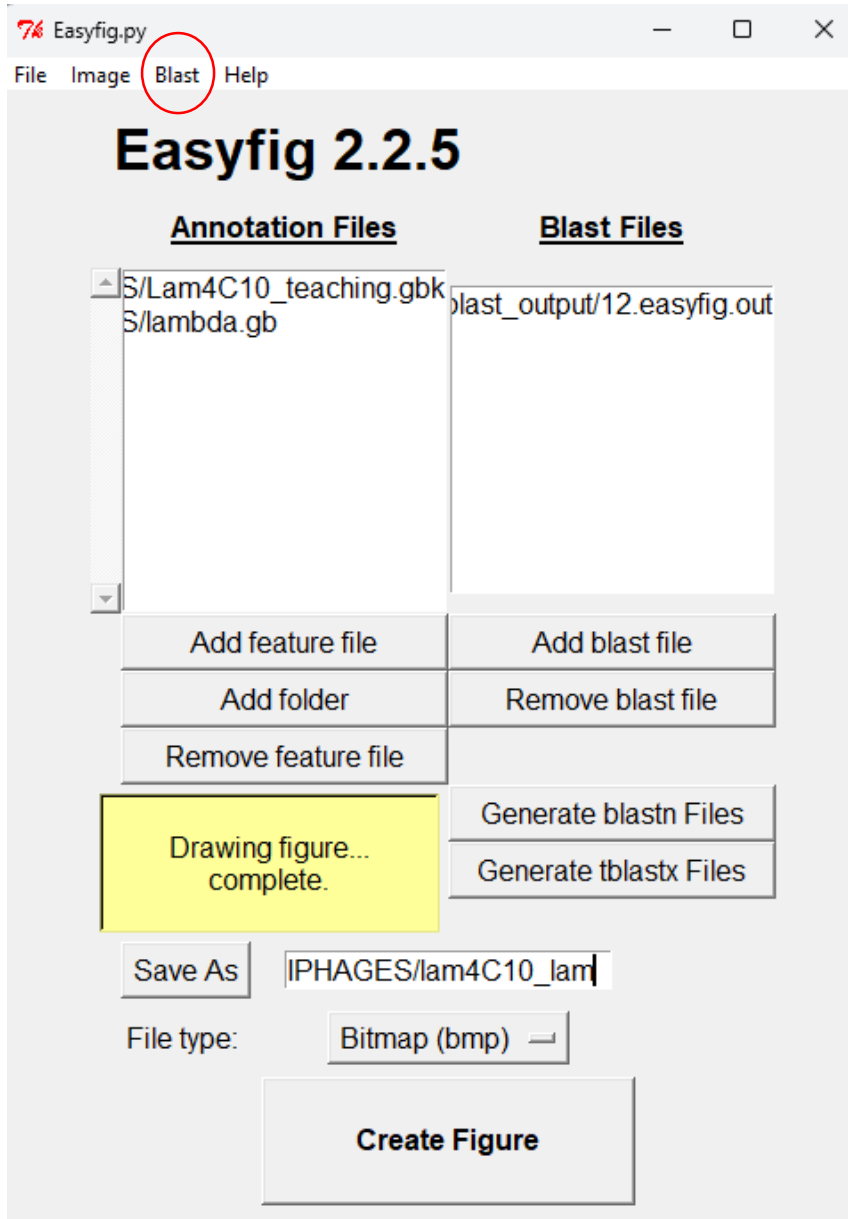
<input checked="" type="checkbox"/>	Escherichia phage PhiR63_2, complete genome	Escherichia phage PhiR63_2	30548	61144	73%	0.0	97.47%	48138	PV340574.1
<input checked="" type="checkbox"/>	Escherichia phage 21, complete genome	Escherichia phage 21	30531	55753	67%	0.0	97.45%	42931	OL657228.1
<input checked="" type="checkbox"/>	Escherichia phage Lambda_ev017 genome assembly, chromosome: 1	Escherichia phage ev017	30348	60912	73%	0.0	97.26%	50126	NC_049948.1
<input checked="" type="checkbox"/>	Escherichia phage PhiR57_1, complete genome	Escherichia phage PhiR57_1	30326	60206	72%	0.0	97.24%	49482	PV340572.1
<input checked="" type="checkbox"/>	MAG TPA_asm: Bacteriophage sp. isolate ctxLv2, partial genome	Bacteriophage sp.	22432	22811	27%	0.0	97.38%	37726	BK024793.1
<input checked="" type="checkbox"/>	Escherichia phage Perceval genome assembly, complete genome: monopartite	Escherichia phage Perceval	21272	60675	71%	0.0	99.84%	50143	OV696612.1
<input checked="" type="checkbox"/>	MAG TPA_asm: Caudoviricetes sp. isolate ctm1A1, partial genome	Caudoviricetes sp.	19699	30111	36%	0.0	98.17%	32059	BK032296.1
<input checked="" type="checkbox"/>	Escherichia virus Lambda_2G7b genome assembly, chromosome: 1	Escherichia phage 2G7b	17571	33944	43%	0.0	96.35%	47142	NC_049954.1
<input checked="" type="checkbox"/>	Escherichia phage PhiR30_1, complete genome	Escherichia phage PhiR30_1	17547	36913	48%	0.0	95.17%	47390	PV340558.1
<input checked="" type="checkbox"/>	Escherichia phage PhiR11_1, complete genome	Escherichia phage PhiR11_1	17455	32509	41%	0.0	96.32%	49016	PV340547.1
<input checked="" type="checkbox"/>	Escherichia phage PhiR11_1_star, complete genome	Escherichia phage PhiR11_1_star	17455	32509	41%	0.0	96.32%	50509	PV340548.1
<input checked="" type="checkbox"/>	Escherichia phage Lambda_ev207 genome assembly, chromosome: 1	Escherichia phage ev207	17350	45613	56%	0.0	96.14%	46685	NC_049949.1
<input checked="" type="checkbox"/>	Escherichia phage 434, complete genome	Escherichia phage 434	17341	26291	34%	0.0	96.13%	47993	OL657226.1
<input checked="" type="checkbox"/>	Mutant Escherichia phage 434 isolate 434B, complete genome	Escherichia phage 434	17335	29459	39%	0.0	96.12%	47075	OL657227.1
<input checked="" type="checkbox"/>	Escherichia phage Turner, complete genome	Escherichia phage Turner	17309	39939	53%	0.0	94.79%	48223	PP925837.1
<input checked="" type="checkbox"/>	Escherichia virus Lambda_2H10 genome assembly, chromosome: 1	Escherichia phage 2H10	17237	18020	24%	0.0	94.69%	46288	NC_049950.1
<input checked="" type="checkbox"/>	Escherichia phage Lambda_h434 imm21, complete genome	Escherichia phage Lambda_h434 imm21	17152	24846	32%	0.0	95.81%	43452	OM418627.1
<input checked="" type="checkbox"/>	Lambdavirus lambda_Lambda_AGU DNA, complete genome	Lambdavirus lambda	17108	24378	31%	0.0	95.80%	48503	LC730321.1
<input checked="" type="checkbox"/>	Escherichia phage Lambda_imm21, complete genome	Escherichia phage Lambda_imm21	17108	24400	31%	0.0	95.80%	46148	OM418625.1
<input checked="" type="checkbox"/>	Escherichia phage E26-Berryhill, partial genome	Escherichia phage E26-Berryhill	17102	24605	31%	0.0	95.79%	48850	OM475435.1
<input checked="" type="checkbox"/>	Escherichia phage Lambda_imm434, complete genome	Escherichia phage Lambda_imm434	17102	24367	31%	0.0	95.79%	47326	OM418626.1
<input checked="" type="checkbox"/>	Enterobacteria phage HK629, complete genome	Escherichia phage HK629	17102	17485	23%	0.0	95.79%	47288	NC_019711.1
<input checked="" type="checkbox"/>	Escherichia phage TPL-02, complete genome	Escherichia phage TPL-02	17102	17102	22%	0.0	95.79%	170231	PQ338057.1
<input checked="" type="checkbox"/>	Enterobacteria phage O276, complete genome	Enterobacteria phage O276	17102	24411	31%	0.0	95.79%	41969	NC_049951.1
<input checked="" type="checkbox"/>	Escherichia phage 4W, partial genome	Escherichia phage 4W	17102	24604	31%	0.0	95.79%	48634	OM475432.1
<input checked="" type="checkbox"/>	MAG: Caudoviricetes sp. isolate 61, complete genome	Caudoviricetes sp.	17102	17102	22%	0.0	95.79%	28800	MN855675.1
<input checked="" type="checkbox"/>	Escherichia phage Lambda, complete genome	Escherichia phage Lambda	17102	24486	31%	0.0	95.79%	48507	PV340536.1

- Hit to a well known temperate phage: the phage lambda
- Let us compare lam4C10 to lambda with Easyfig



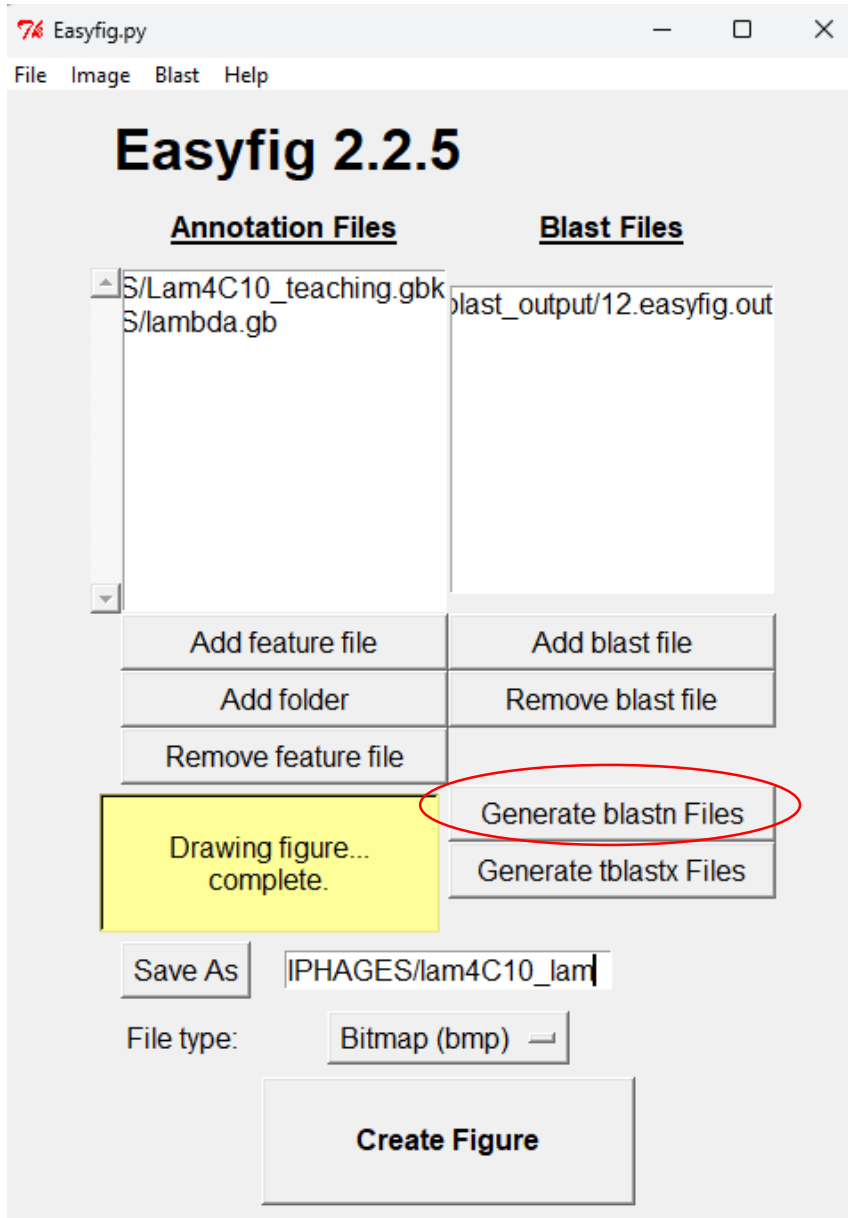
Comparing the two genomes with Easyfig

1. Download locally the Blast package



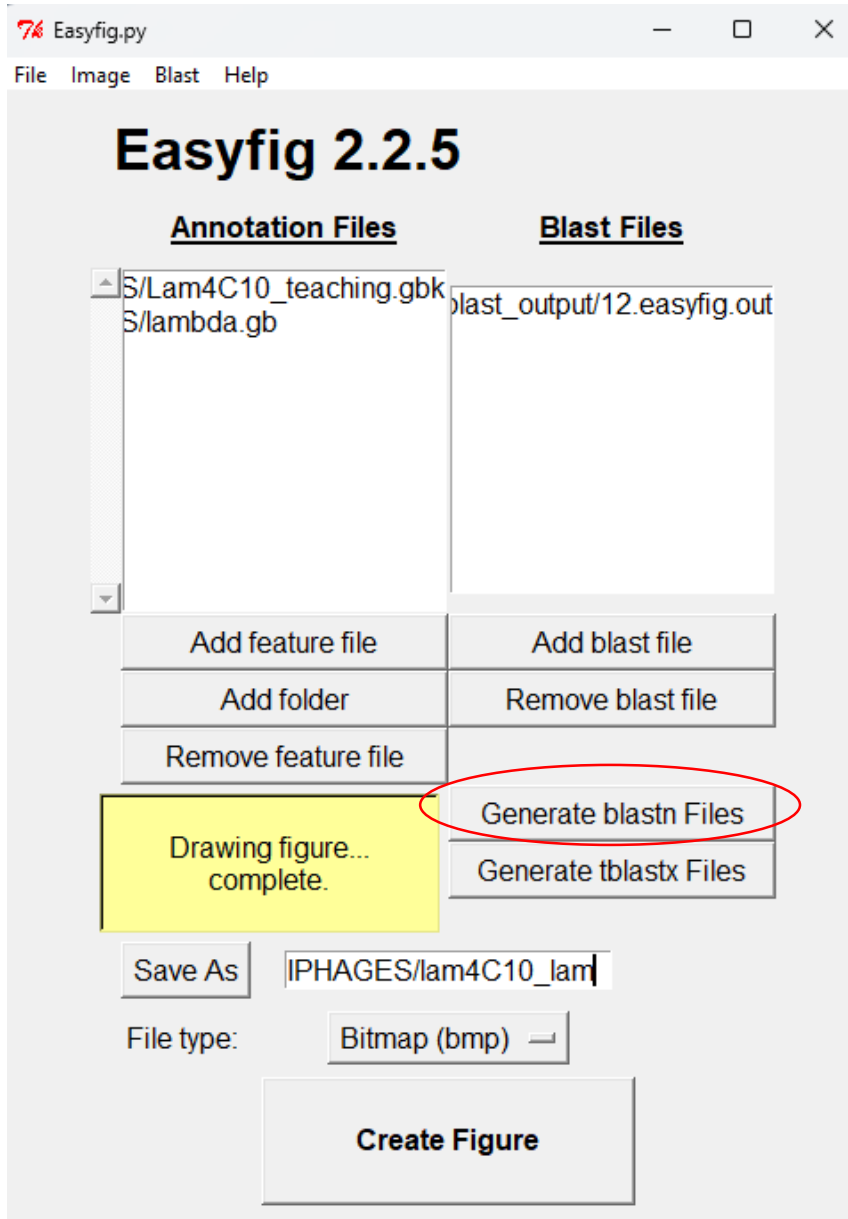
Comparing the two genomes with Easyfig

1. Download locally the Blast package
2. Indicate where to store the blastn result, and generate blastn files

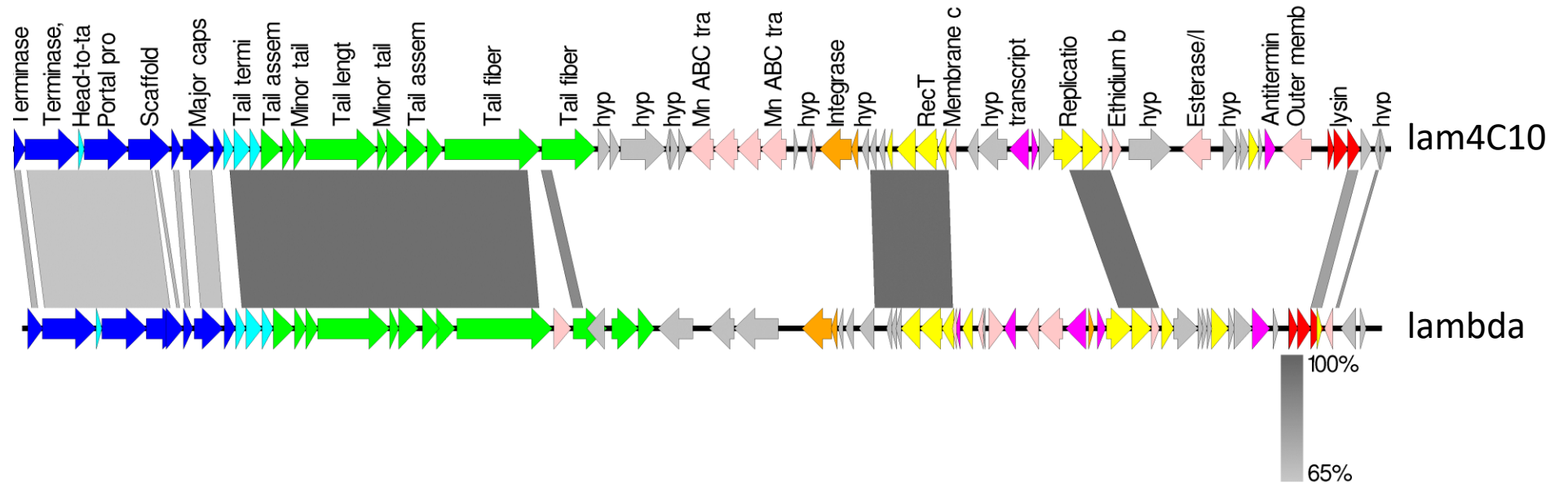


Comparing the two genomes with Easyfig

1. Download locally the Blast package
2. Indicate where to store the blastn result, and generate blastn files
3. Push the Create figure button



Result



Conclusions:

Few regions shared: our genome is not the same species/genus as lambda

Definitions:

→ 2 genomes belong to the same **species** if, over more than 80% of their genomes, they share **>95%** identity
→ ' ' ' ' **genus** ' ' ' ' **>70%**

lam4C10 defines a new genus, named **Bievrevirus**, and gave its name to the species: **bv4C10**

All new genomes sharing >95% identity with lam4C10 will be of the species « *Bievrevirus bv4C10* »

This is the classical bi-nominal way to name species, we say *Escherichia coli* or *Ratus norvegicus*

To conclude

- Starting from a phage genome in a fasta file format, it is nowadays easy to define
 1. Its open reading frames and protein function predictions. This will be practised this afternoon, using the genome of the phage Ecoha you tested this week, and Pharokka
 2. Its functional modules, and its overall genomic similarity to other phage genomes, to propose a taxonomic classification at the genus and species level. If we have time, this will also be done this afternoon for Ecoha.