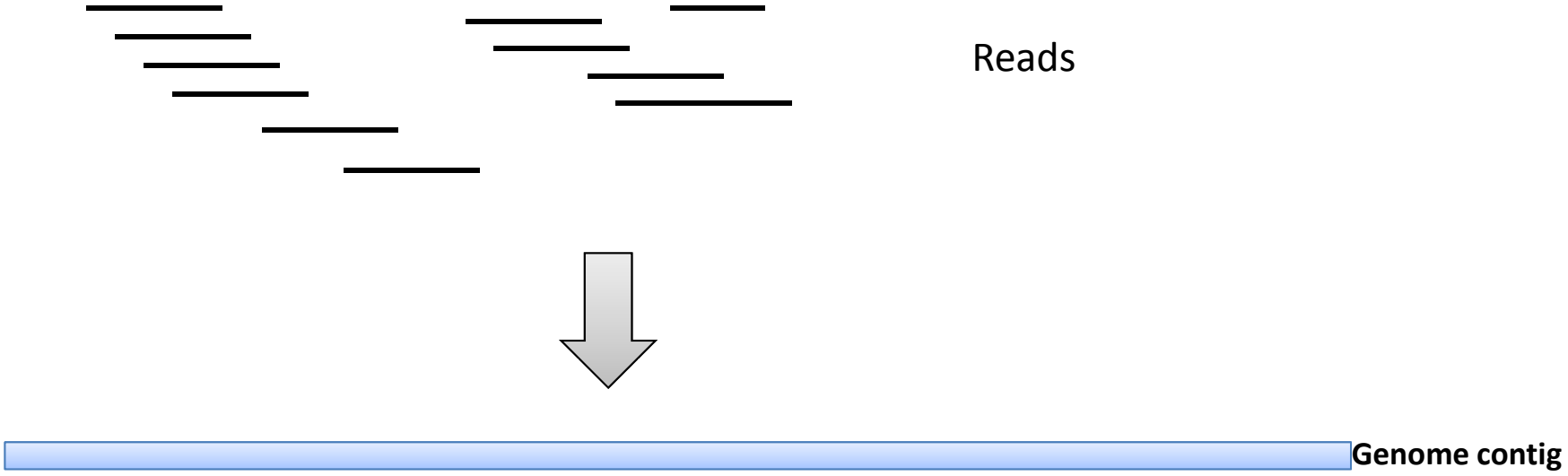
Denovo genome assembly and analysis

outline

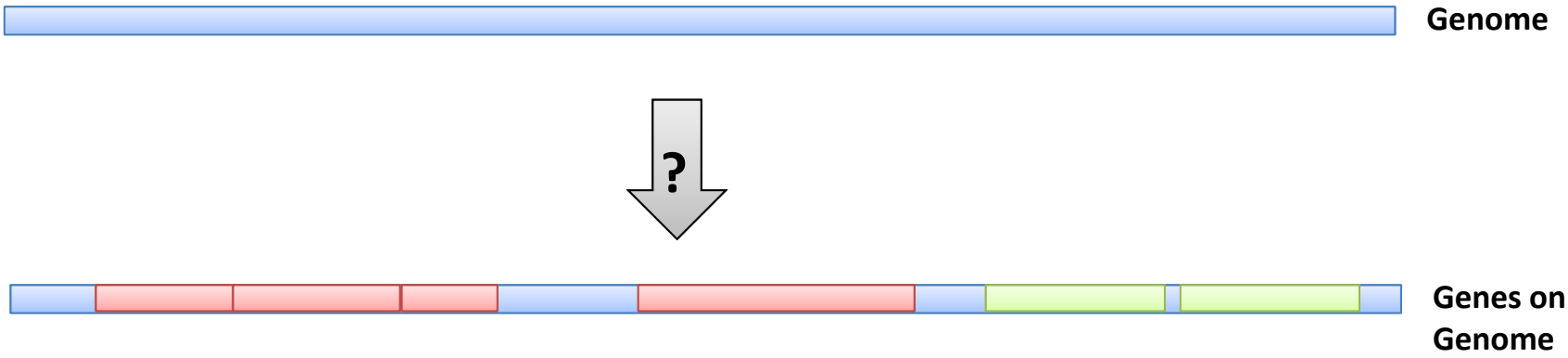
- De novo genome assembly
- Gene finding from assembled contigs
- Gene annotation

Denovo genome assembly

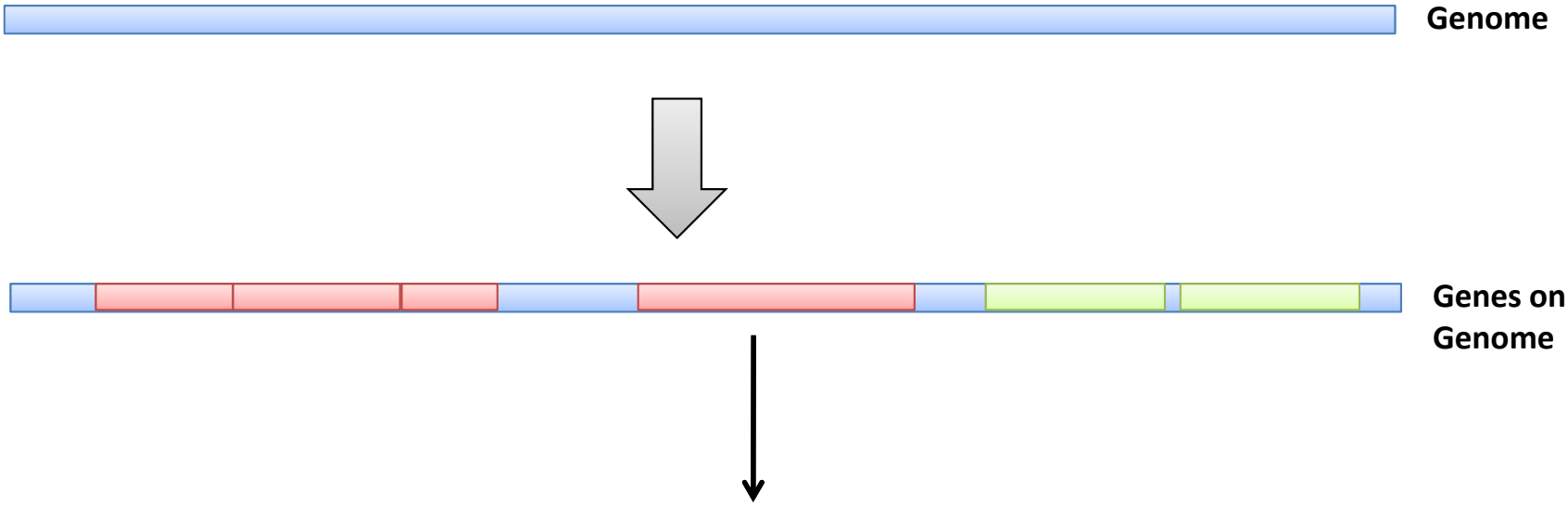


Gene finding

- To find out coding region on genome sequence



Gene Annotation



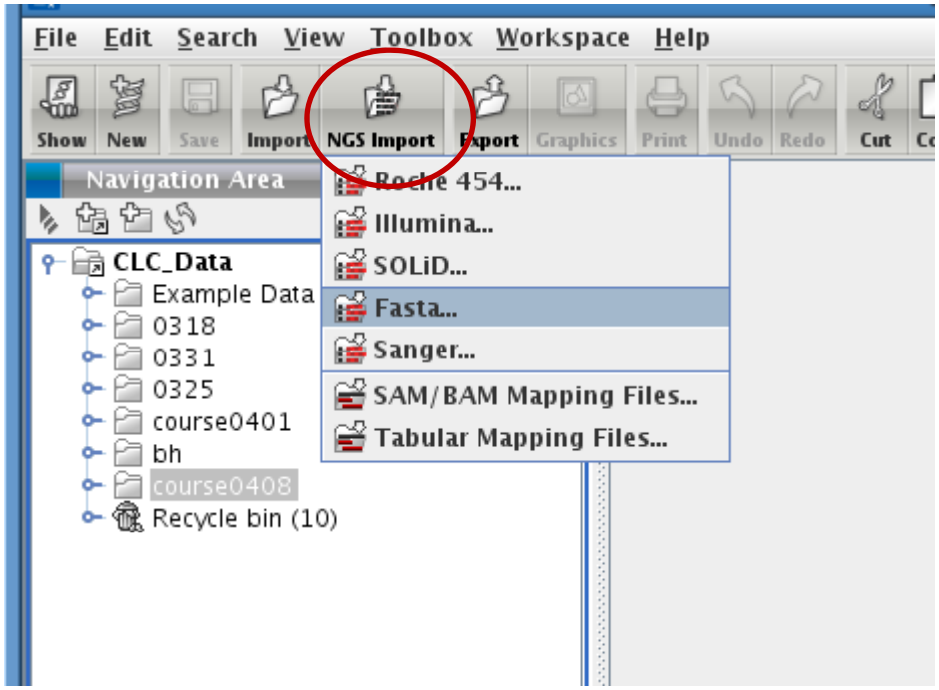
- For each gene….
  - Conserved?
  - Domain?
  - Function?

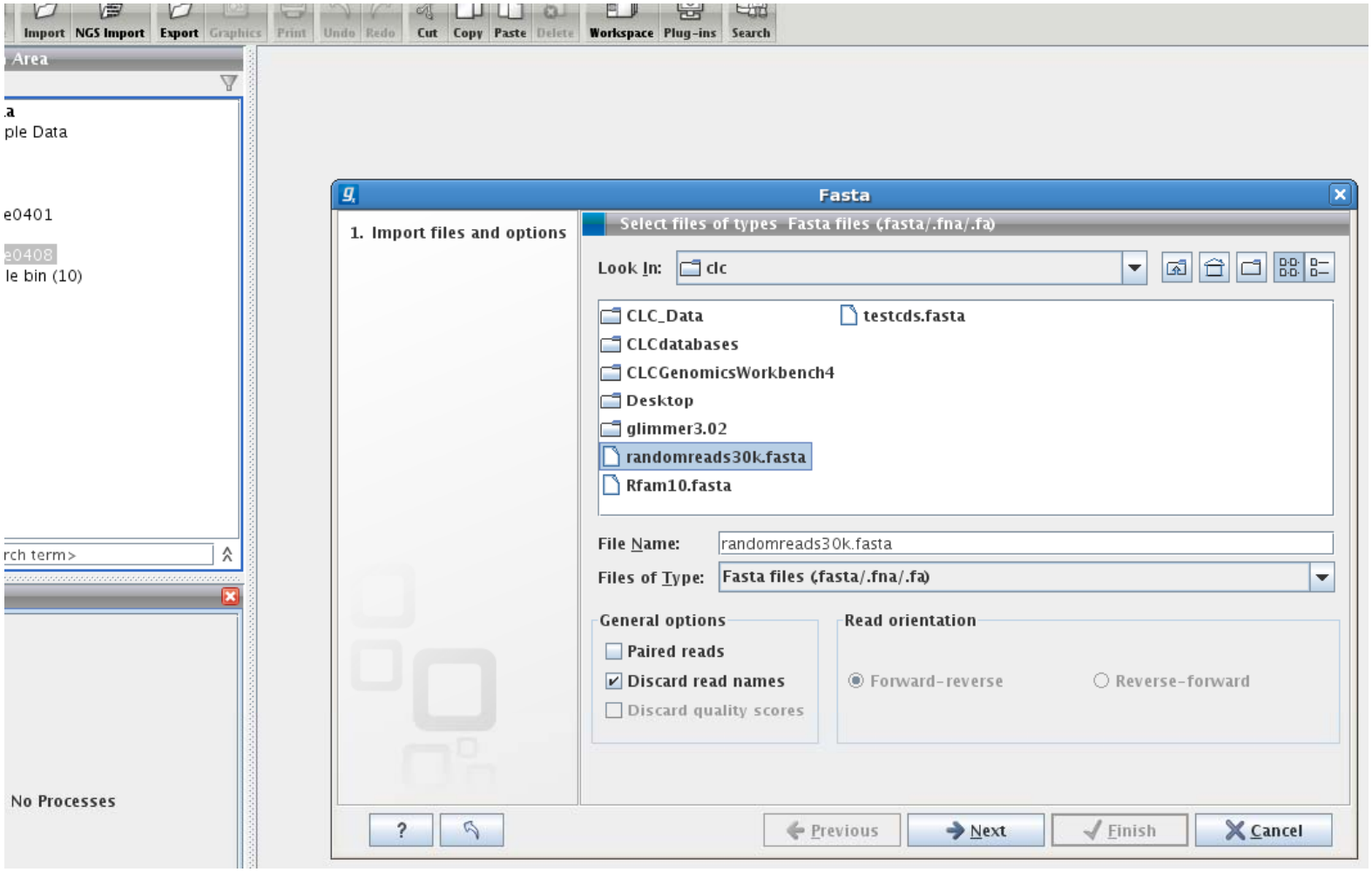
get reads file

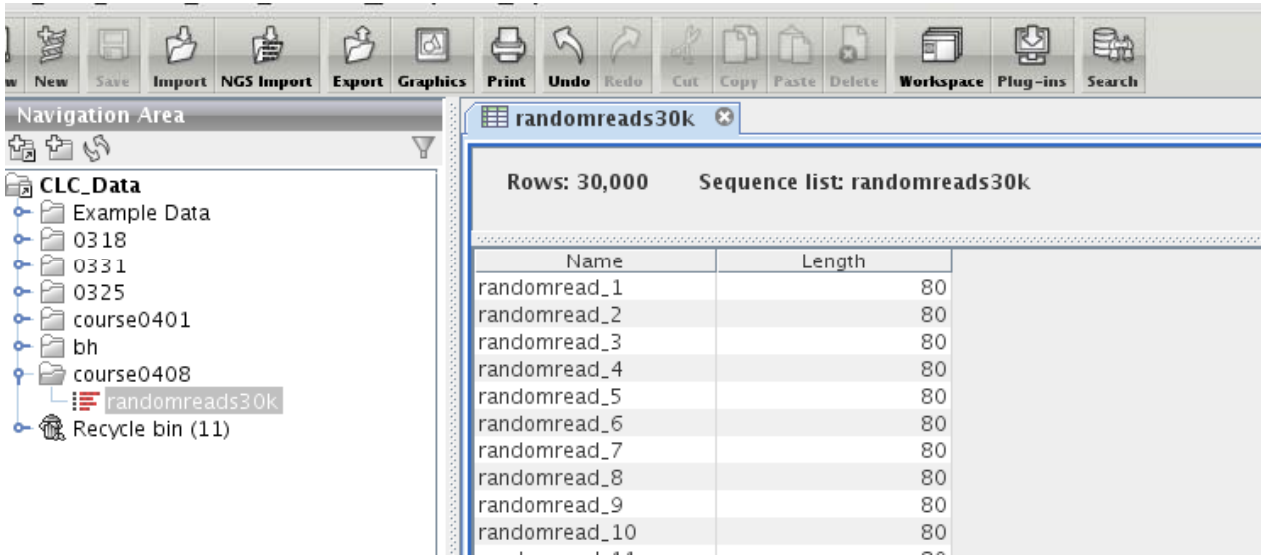
- download a random generated reads file
  - http://163.25.92.61/course/randomreads30k.fasta

- open CLC to assemble contigs from reads

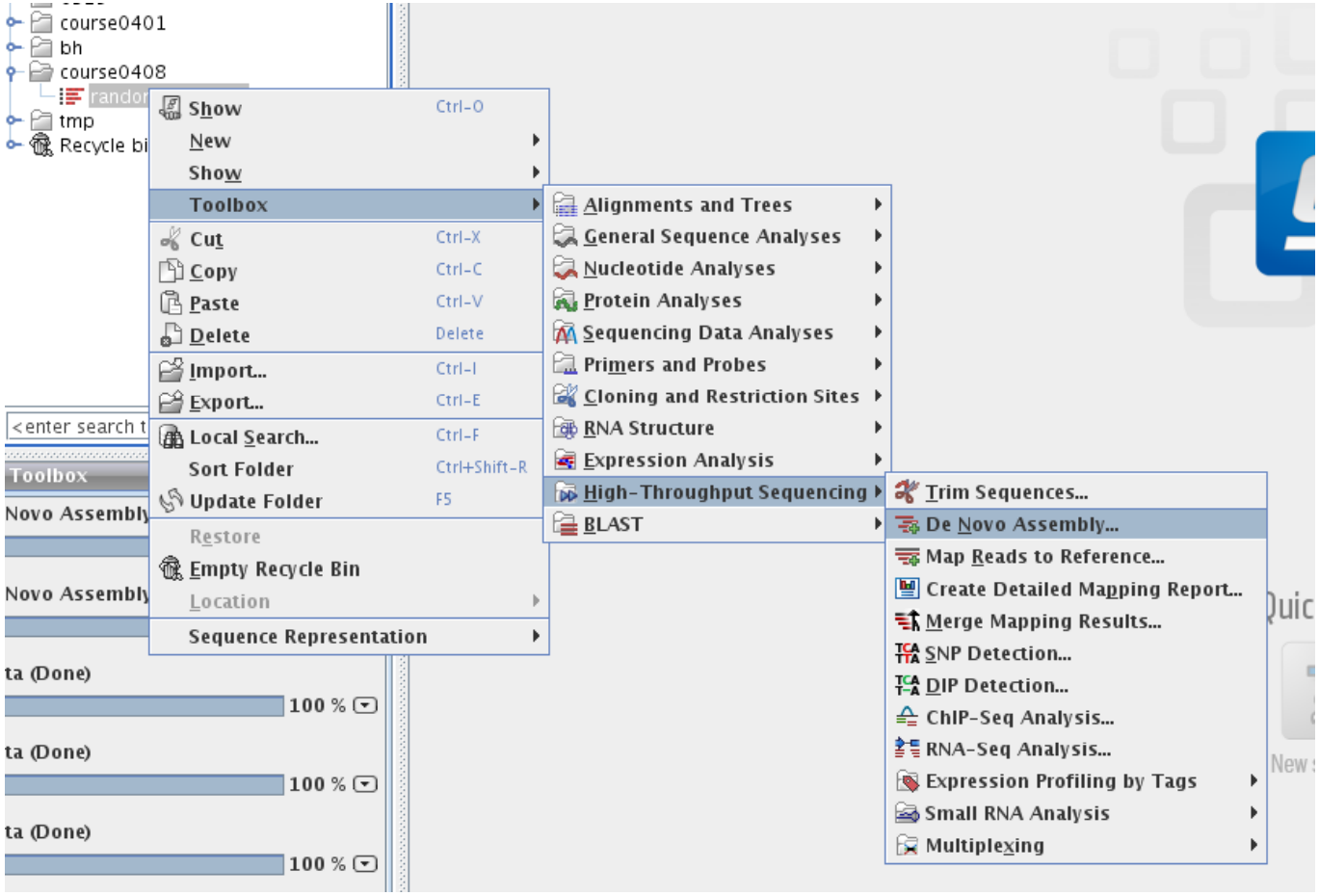
NGS import the reads file

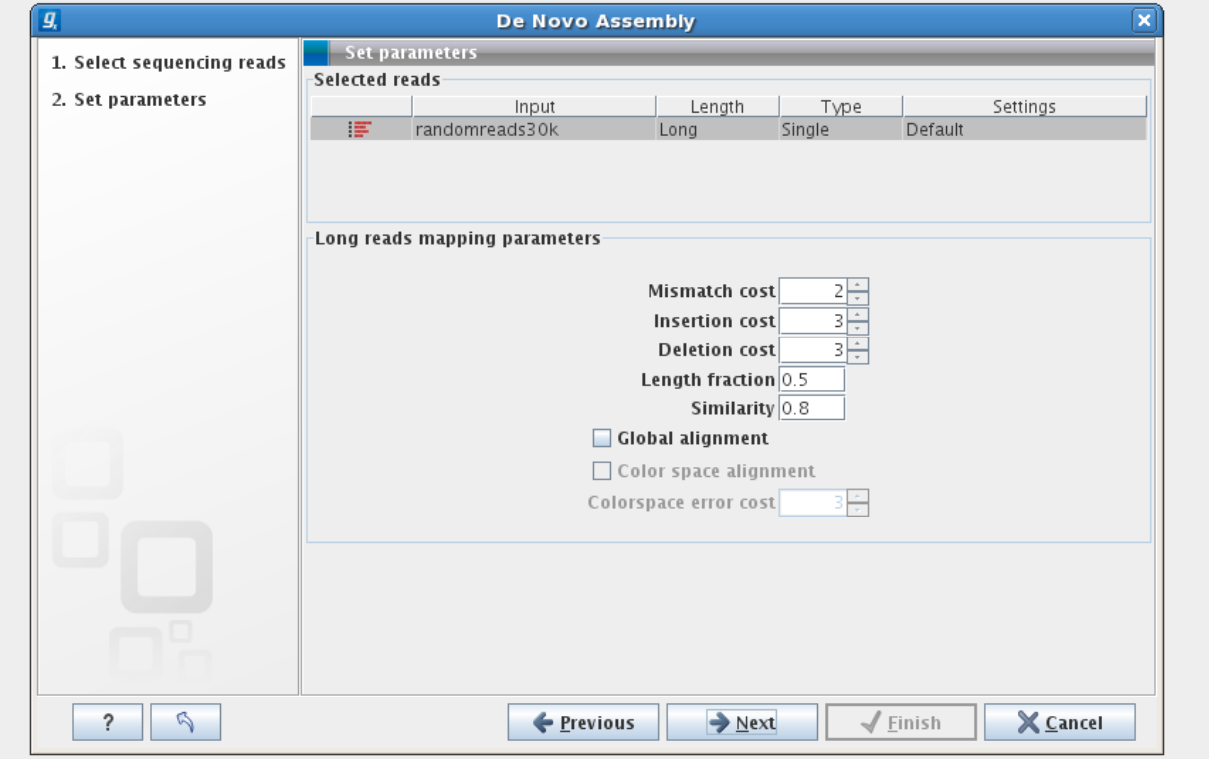


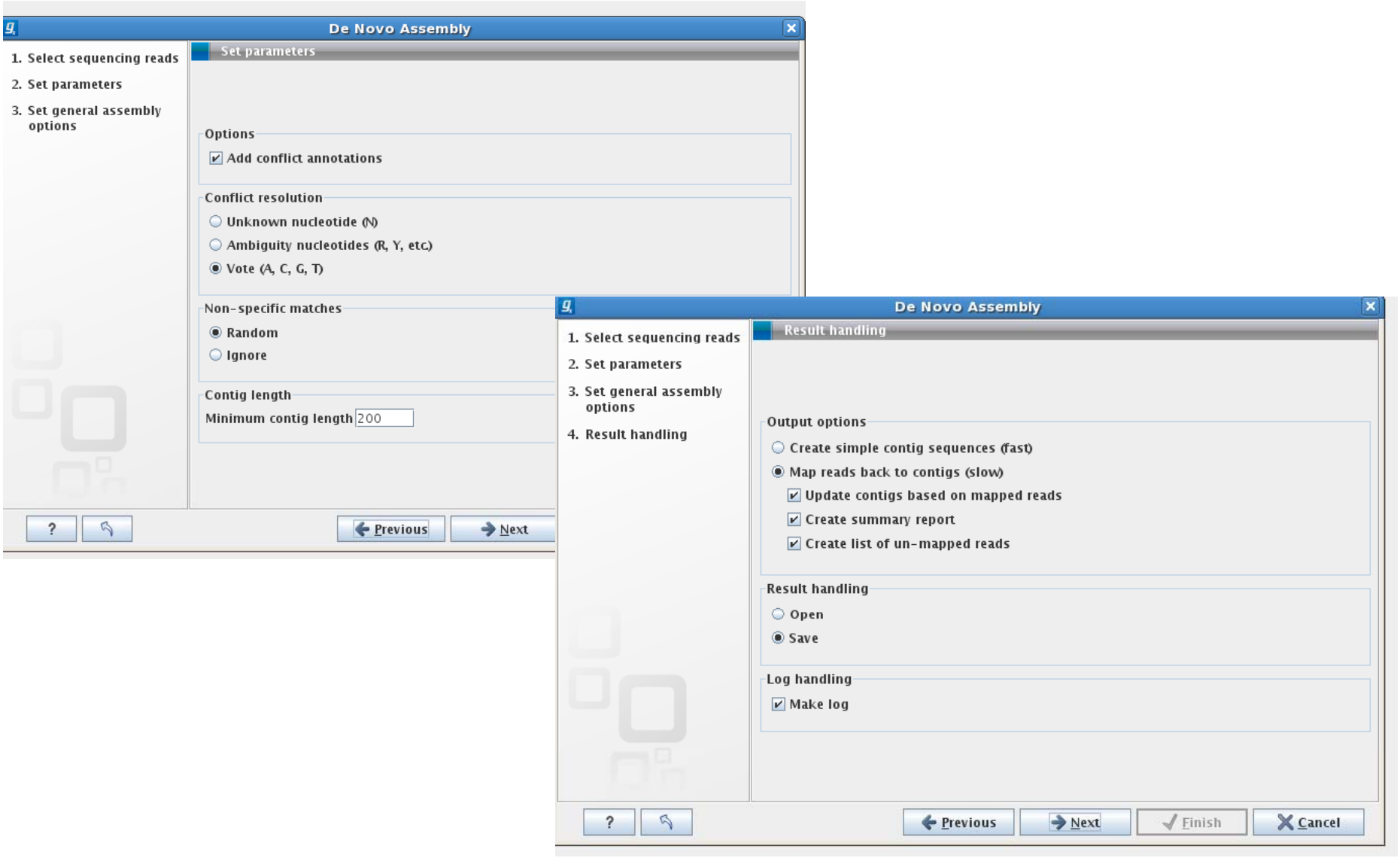

Import
NGS Import
Export
Graphics
Print
Undo
Redo
Cut
Copy
Paste
Delete
Workspace
Plug-ins
Search
Area
ple Data
e0401
e0408
le bin (10)
rch term>
No Processes
Fasta
1. Import files and options
Select files of types Fasta files (.fasta/.fna/.fa)
Look In:
clc
CLC_Data
CLCdatabases
CLCGenomicsWorkbench4
Desktop
glimmer3.02
randomreads30k.fasta
Rfam10.fasta
testcds.fasta
File Name:
randomreads30k.fasta
Files of Type:
Fasta files (.fasta/.fna/.fa)
General options
Paired reads
Discard read names
Discard quality scores
Read orientation
Forward-reverse
Reverse-forward
Previous
Next
Finish
Cancel


New
Save
Import
NGS Import
Export
Graphics
Print
Undo
Redo
Cut
Copy
Paste
Delete
Workspace
Plug-ins
Search
Navigation Area
CLC_Data
Example Data
0318
0331
0325
course0401
bh
course0408
randomreads30k
Recycle bin (11)
randomreads30k
Rows: 30,000
Sequence list: randomreads30k
Name
Length
randomread_1 80
randomread_2 80
randomread_3 80
randomread_4 80
randomread_5 80
randomread_6 80
randomread_7 80
randomread_8 80
randomread_9 80
randomread_10 80

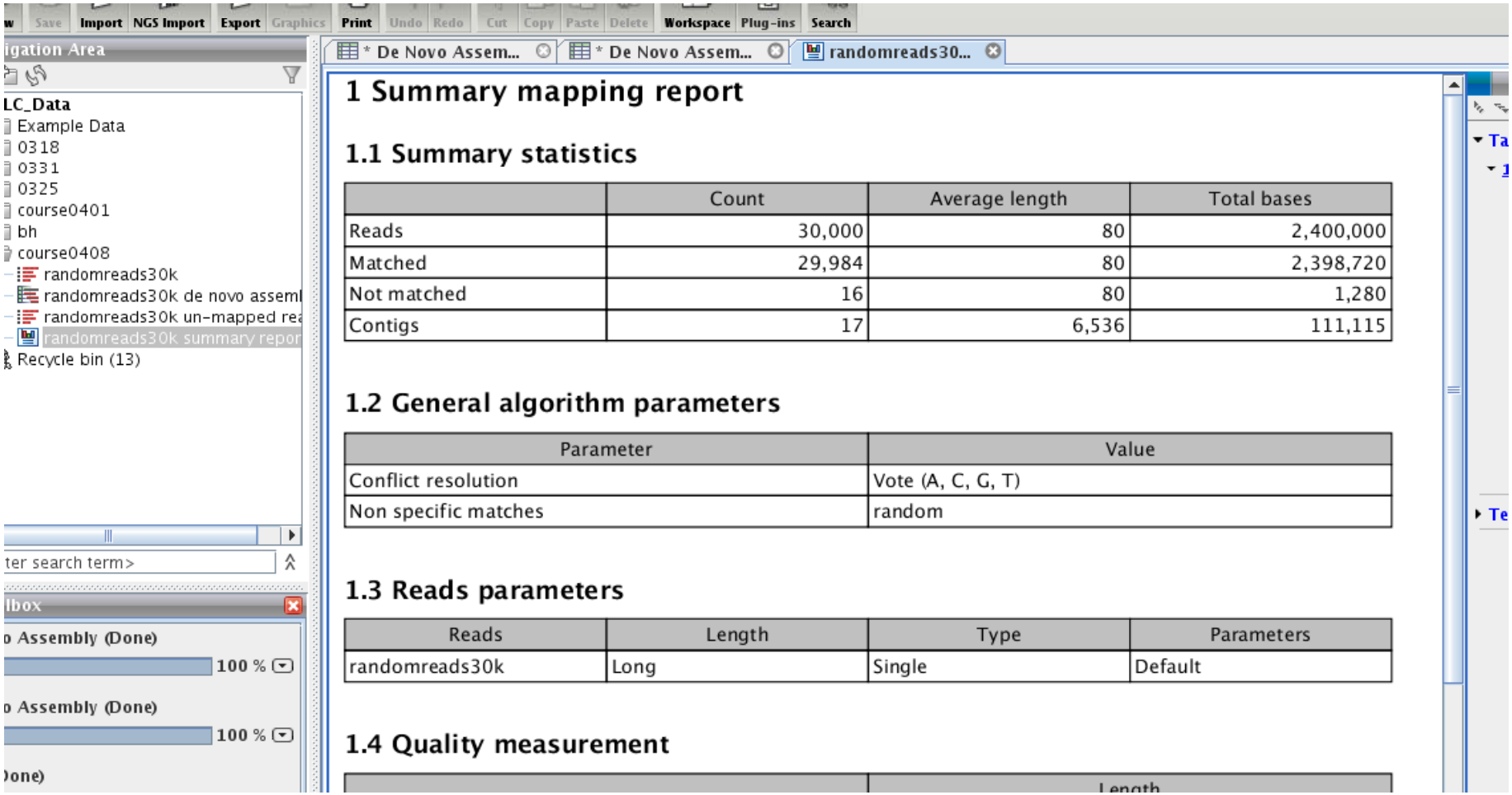
Denovo assembly




De Novo Assembly
1. Select sequencing reads
2. Set parameters
Set parameters
Selected reads
Input
Length
Type
Settings
randomreads30k
Long
Single
Default
Long reads mapping parameters
Mismatch cost
2
Insertion cost
3
Deletion cost
3
Length fraction
0.5
Similarity
0.8
Global alignment
Color space alignment
Colorspace error cost
3
?
Previous
Next
Finish
Cancel


De Novo Assembly
1. Select sequencing reads
2. Set parameters
3. Set general assembly options
Set parameters
Options
Add conflict annotations
Conflict resolution
Unknown nucleotide (N)
Ambiguity nucleotides (R, Y, etc.)
Vote (A, C, G, T)
Non-specific matches
Random
Ignore
Contig length
Minimum contig length 200
?
Previous
Next
De Novo Assembly
1. Select sequencing reads
2. Set parameters
3. Set general assembly options
4. Result handling
Result handling
Output options
Create simple contig sequences (fast)
Map reads back to contigs (slow)
Update contigs based on mapped reads
Create summary report
Create list of un-mapped reads
Result handling
Open
Save
Log handling
Make log
?
Previous
Next
Finish
Cancel

report

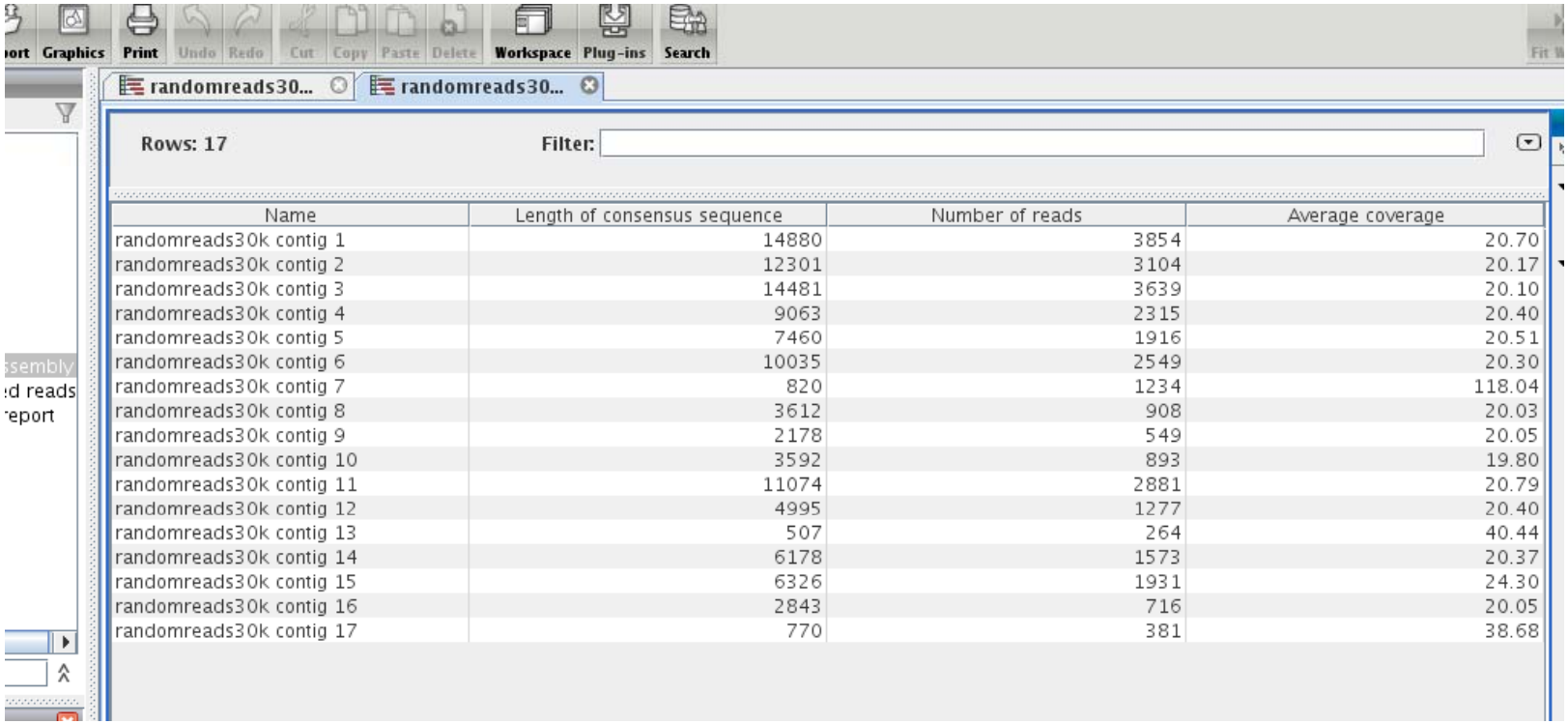


	Count	Average length	Total bases
Reads	30,000	80	2,400,000
Matched	29,984	80	2,398,720
Not matched	16	80	1,280
Contigs	17	6,536	111,115

Parameter	Value
Conflict resolution	Vote (A, C, G, T)
Non specific matches	random

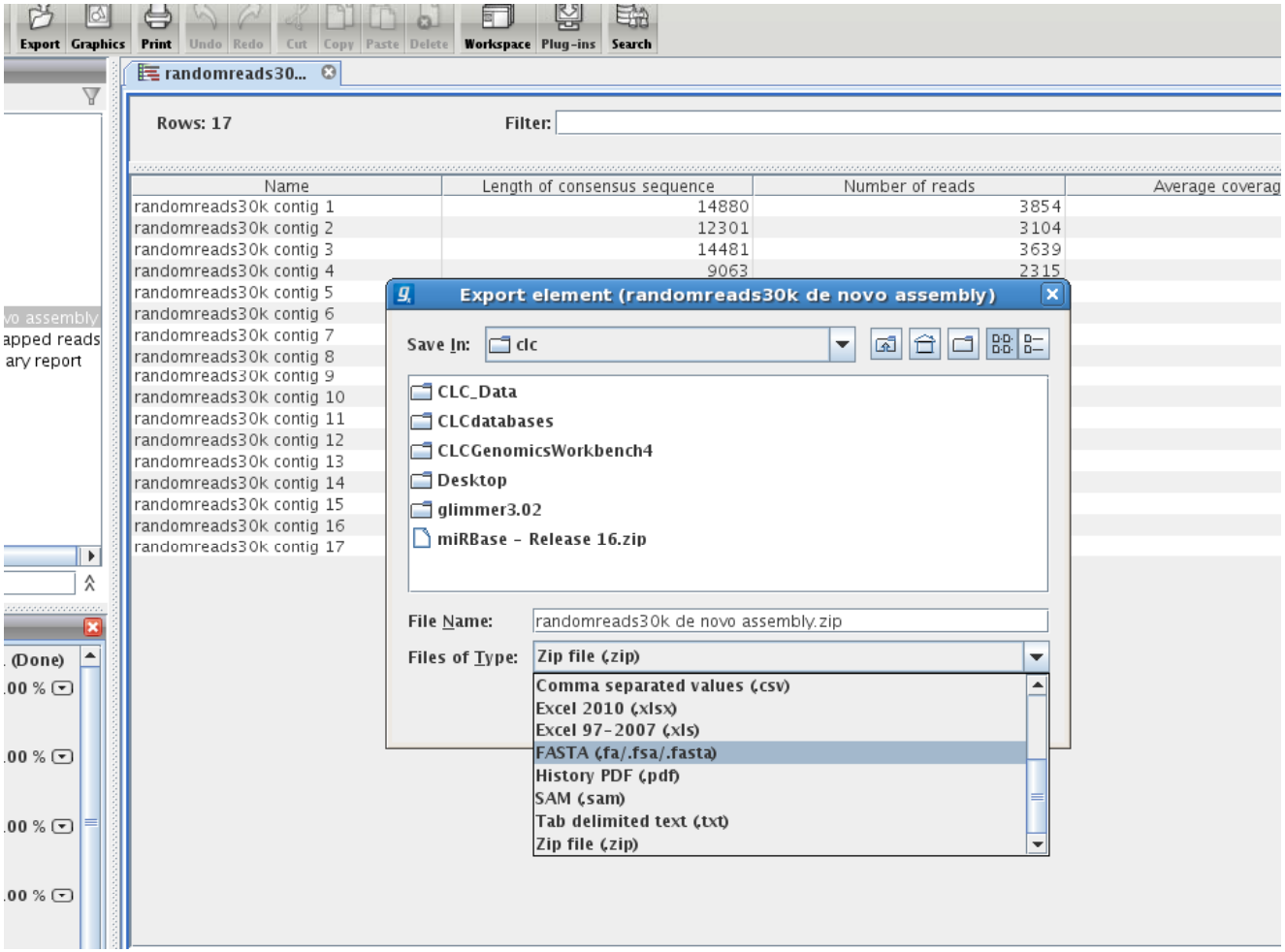
Reads	Length	Type	Parameters
randomreads30k	Long	Single	Default

assembled contigs



Name	Length of consensus sequence	Number of reads	Average coverage
randomreads30k contig 1	14880	3854	20.70
randomreads30k contig 2	12301	3104	20.17
randomreads30k contig 3	14481	3639	20.10
randomreads30k contig 4	9063	2315	20.40
randomreads30k contig 5	7460	1916	20.51
randomreads30k contig 6	10035	2549	20.30
randomreads30k contig 7	820	1234	118.04
randomreads30k contig 8	3612	908	20.03
randomreads30k contig 9	2178	549	20.05
randomreads30k contig 10	3592	893	19.80
randomreads30k contig 11	11074	2881	20.79
randomreads30k contig 12	4995	1277	20.40
randomreads30k contig 13	507	264	40.44
randomreads30k contig 14	6178	1573	20.37
randomreads30k contig 15	6326	1931	24.30
randomreads30k contig 16	2843	716	20.05
randomreads30k contig 17	770	381	38.68

export fasta file



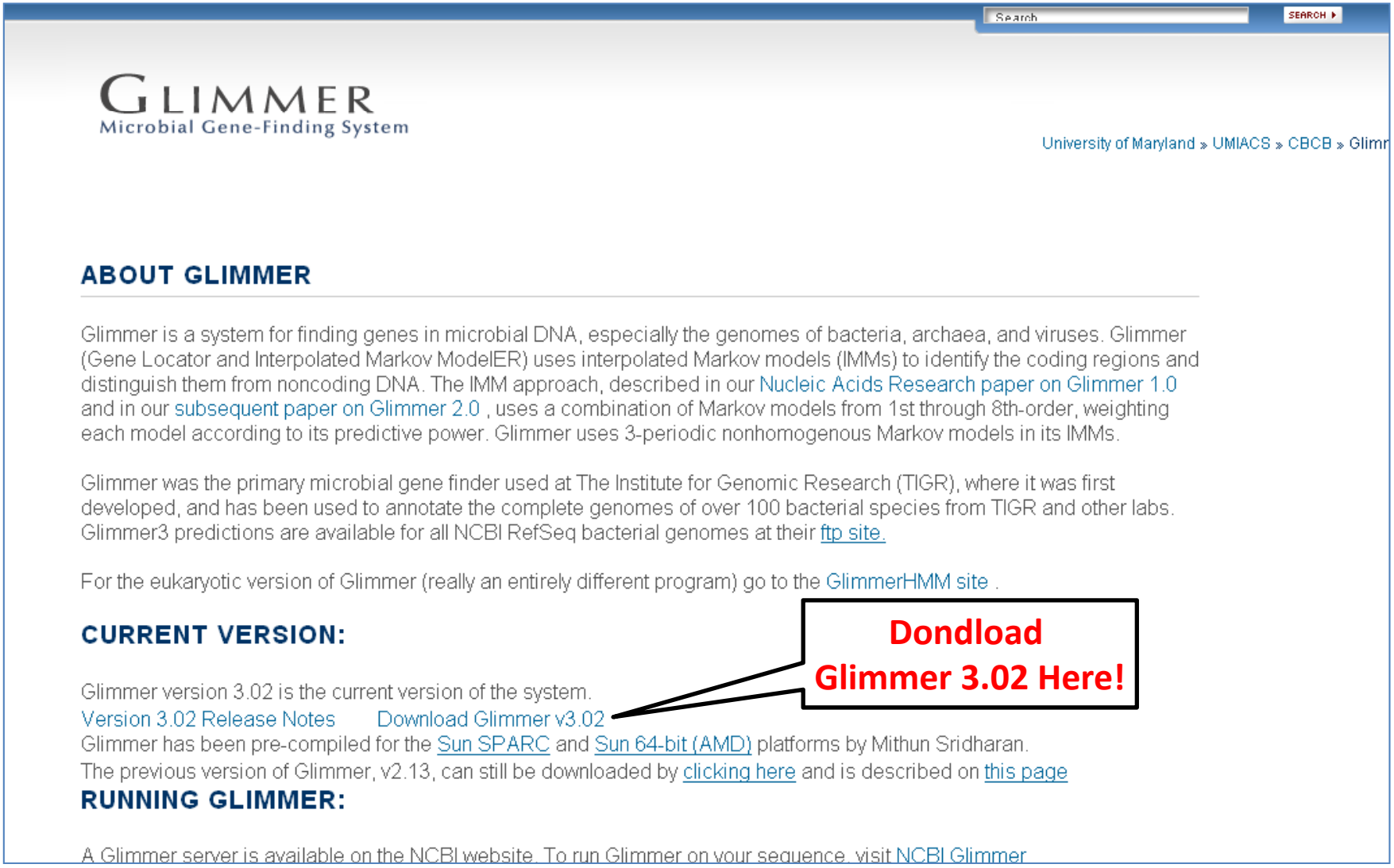
Glimmer

- Glimmer is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses.
  - (**G**ene **L**ocator and **I**nterpolated **M**arkov **M**odel**ER**)

- http://www.cbcb.umd.edu/software/glimmer/
- Center for Bioinformatics & Computational Biology, University of Maryland

- Paper about Glimmer 1.0
  - S. Salzberg, A. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models, *Nucleic Acids Research* 26:2 (1998), 544-548.
- Glimmer2.0
  - A.L. Delcher, D. Harmon, S. Kasif, O. White, and S.L. Salzberg. Improved microbial gene identification with GLIMMER, *Nucleic Acids Research* 27:23 (1999), 4636-4641.
- Glimmer 3.0
  - A.L. Delcher, K.A. Bratke, E.C. Powers, and S.L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23:6 (2007), 673-679.

http://www.cbcb.umd.edu/software/glimmer/



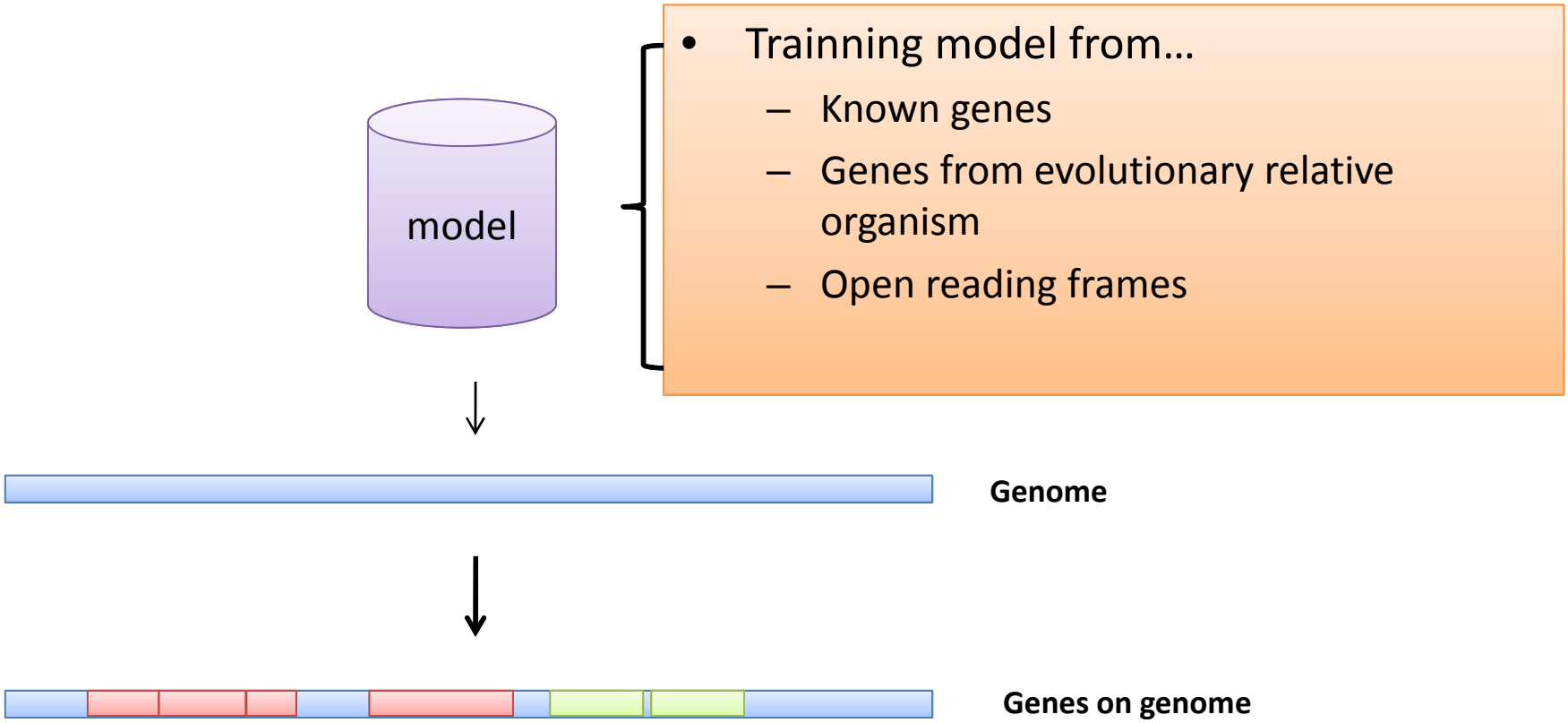
Or download glimmer from here

- wget http://163.25.92.61/course/glimmer302.tar.gz

Glimmer install

- extract
  - tar zxvf glimmer302.tar.gz
  - tree -d glimmer3.02/
- go into directory of glimmer's source code
  - cd glimmer3.02/src/
  - pwd
- compile the binary code
  - make
- executable binary will be located in
  - (glimmer3.02/bin/)

Concept of glimmer



4 steps to run the glimmer

1. long-orfs
   - This program identifies long, non-overlapping open reading frames (orfs) in a DNA sequence file.
2. extract
   - This program reads a genome sequence and a list of coordinates for it and outputs a multifasta file of the regions specified by the coordinates
3. build-icm
   - This program constructs an interpolated context model (ICM) from an input set of sequences.
4. glimmer3

g3-from-scartch.csh

- glimmer3.02/scripts/

- `g3-from-scratch.csh genome.fasta mygenome`
- The script would then run the commands:
  - `long-orfs -n -t 1.15 genome.fasta mygenome.longorfs`
  - `extract -t genome.fasta mygenome.longorfs > mygenome.train`
  - `build-icm -r mygenome.icm < mygenome.train`
  - `glimmer3 -o50 -g110 -t30 genom.seq mygenome.icm mygenome`

Output of glimmer (xxx.predict)

```
>gi|15638995|ref|NC_000919.1| Treponema pallidum subsp. pallidum str. Nichols, complete genome
orf00001         4      1398   +1      6.22
orf00003      1641      2756   +3      2.89
orf00004      2776      3834   +1      5.47
orf00005      3863      4264   +2      2.77
orf00006      4391      6832   +2      7.08
orf00007      6832      7074   +1      0.25
orf00008      7317      7967   +3      6.92
orf00009      7997      8260   +2      2.91
orf00010      9515      8340   -3      2.80
orf00011      9838      9984   +1      0.10
orf00013     10237     10362   +1      6.02
orf00014     10396     12378   +1      3.77
orf00015     12545     13210   +2      8.04
```

ID | Start & stop position | frame | score

Modification of the script g3-from-scartch.csh

vi ../scripts/g3-from-scartch.csh

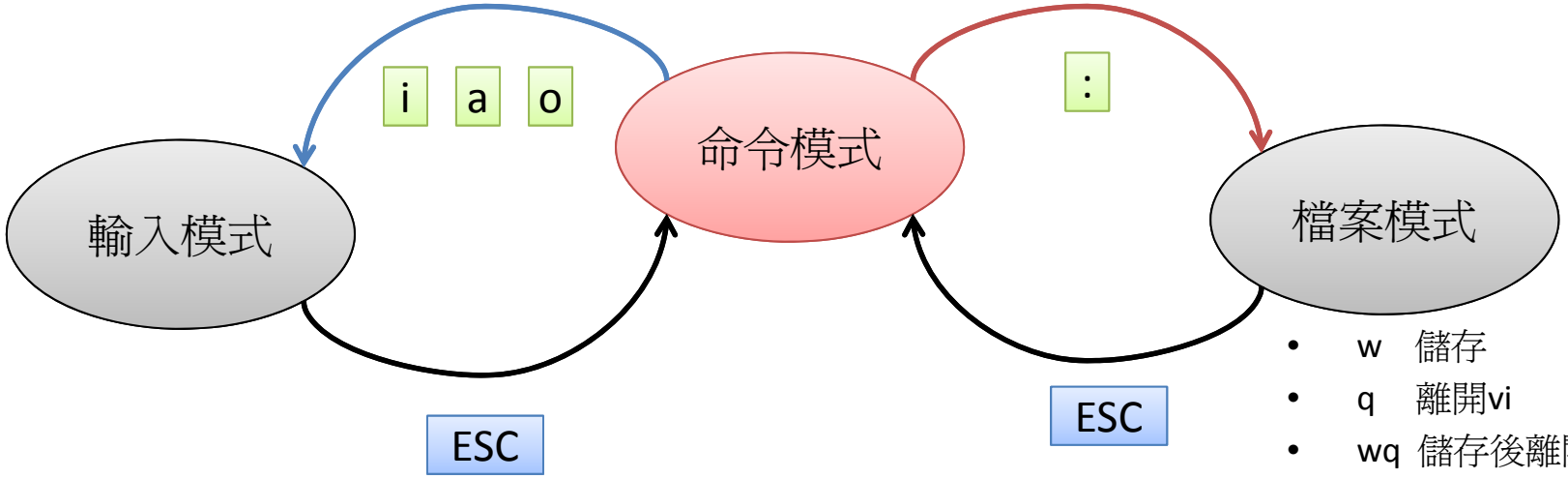
```
set awkpath = /fs/szgenefinding/Glimmer3/scripts
set glimmerpath = /fs/szgenefinding/Glimmer3/bin
```



```
set awkpath = ~/glimmer3.02/scripts
set glimmerpath = ~/glimmer3.02/bin
```

vi 編輯器:

vi filename



- w 儲存
- q 離開vi
- wq 儲存後離開
- q! 不儲存就離開

Convert coordinate file into fasta format (single fasta file)

- extract
  - Usage:

    extract genome_file coord_file > fasta_file

for multiple fasta file coordinate convert

- use home-made script to re-format coordinate file
  - http://163.25.92.61/course/multipredict.pl

- multi-extract
  - Usage:

    multi-extract genome_file coord_file > fasta_file

NetBlast

- The BLAST client, or blastcl3, bypasses the web browser and interacts directly with the NCBI BLAST server that powers the NCBI web BLAST service
- ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/

- But you can download here…
- cd ~ (go back to your home directory)
- wget http://163.25.92.61/course/netblast-2.2.25-ia32-linux.tar.gz

- extract
  - tar zxvf netblast-2.2.20-ia32-linux.tar.gz

blastcl3

- netblast-2.2.25/bin/

- ./blastcl3 -p program -i input_sequence -d dbname -o output_file

  -p (blastn, blastx, blastp, tbastn tblastx)

  -i (query file, predice genes here)

  -d (database name)

  nr, NCBI non-redundant database

  -o (output file)

Blast programs

-p program	-i Query sequence	-d database sequence
blastn	nucleotide	nucleotide
blastp	amino acid	amino acid
blastx	translated nucleotide	amino acid
tblastn	amino acid	translated nucleotide
tblastx	translated nucleotide	translated nucleotide

- ./blastcl3
  - -p blastn
  - -i mygene.fasta
  - -d nt
  - -o mygeneblast.html
  - -m 2
  - -K 1
  - -T T