Genome Resequencing analysis

2012/Apr/27

outline

- import data
- Mapping reads to reference genome
- SNP detect
- DIP (InDel) detect
  - deletion/insertion polymorphisms
  - A.k.a **In**sert **del**etion

3 files for 454 genome resequencing

- Ecoli.FLX.fna
  (Reads sequence in fasta format)

- Ecoli.FLX.qual
  (Reads quelity in fasta format)

- NC_010473.gbk
  (E. coli str. K-12 substr. DH10B, complete genome sequence in Genbank format)

3 files for 454 genome resequencing

Read sequence (Ecoli.FLX.fna)

```
>EECRH8001A0WUU
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGTAATGCCGTCGCCCGCCTGTCCGGTGAC
GATTTCCAGCGCGCCATCGCCACAGGCAATCAGCAGTGGCGCAACAGAAATCACGCTCCC
CGGCTGTGCTTTGCTGGCATGAGGATGAACACGCGACGACCAGACGGTGAATTTCTGATT
GCCAACATAGCTGAAGGCACCCGGCCACGGATCGGCAACGGCACGTACCATGTTGTGCAG

>EECRH8001DOWTE
GGCGTCTTTTATAAAGATGAGCCCATCAA
TTTCAGATTATCTGGCCACAAAACAGCGT
CGAATTTGCGGCGTGATTTTTGACTGGGA
AATCAGCTTAATGAATATCTCCCGCTTTA

>EECRH8001EBQ91
CCGTACGATCCGAATACCCAACGACGGGT
ATTAATGAGCGCTTTGCCCGCCATTTTCG
CCGGATATAACCGTCGGGGCCATCCGCAT
CCGGTGCCGACCAACCTGAACCTTATCCA
```

Read Quality (Ecoli.FLX.qual ***)***

```
>EECRH8001A0WUU
14 7 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 28 9 26 35 28
28 27 34 28 28 28 26 24 37 33 15 28 34 28 28 27 27 31 22 32 24 27 27 28 27
24 27 36 32 13 35 28 28 28 27 25 23 26 34 28 27 25 25 28 32 24 25 28 27 29
21 26 29 20 28 27 27 27 27 28 26 26 31 23 27 27 28 34 27 28 26 28 36 32 14
25 25 28 27 27 27 28 37 33 20 5 34 27 26 20 28 26 28 23 37 33 14 26 27 27
34 28 26 27 28 27 19 34 27 28 26 27 31 22 27 27 26 28 28 26 26 25 27 24 33
25 25 28 22 24 35 28 26 23 33 26 36 31 12 28 27 27 25 33 26 27 27 18 32 24
28 25 28 26 27 28 27 28 32 24 33 26 25 28 34 30 9 35 28 27 18 28 28 32 25
28 28 23 34 28 27 34 27 22 34 28 27 27 24 24 28 23 34 27 27 26 27 32 24 27
28 28 27 24 27

>EECRH8001DOWTE
31 12 28 28 28 28 37 33 20 5 26 27 34 30 10 27 28 28 28 28 27 36 32 13 28
28 28 37 32 14 28 34 27 28 27 34 27 28 28 27 27 33 25 27 28 27 27 33 26 27
33 26 27 27 28 34 28 34 28 27 37 33 16 27 27 28 28 35 28 27 28 28 28 34 26
33 25 28 28 37 33 20 6 28 27 27 27 27 34 27 27 28 36 31 12 27 28 27 27 37
33 14 36 32 13 27 27 28 28 27 27 27 28 27 35 28 36 32 13 27 27 28 33 25 36
32 13 25 28 32 24 27 28 27 27 27 38 34 24 14 4 28 27 27 28
```

Reference genome (NC_010473.gbk ***)***

```
LOCUS       NC_010473 4686137 bp DNA circular BCT 23-JAN-2012
DEFINITION  Escherichia coli str. K-12 substr. DH10B chromosome, complete genome.
ACCESSION   NC_010473 VERSION NC_010473.1 GI:170079663 DBLINK Project: 58979
KEYWORDS    .
SOURCE      Escherichia coli str. K-12 substr. DH10B
  ORGANISM  Escherichia coli str. K-12 substr. DH10B Bacteria; Proteobacteria; Gammaproteobacteria;
            Enterobacteriales; Enterobacteriaceae; Escherichia.
FEATURES    Location/Qualifiers
  source    1..4686137
            /organism="Escherichia coli str. K-12 substr. DH10B"
            /mol_type="genomic DNA"
            /strain="K-12"
            /sub_strain="DH10B"
            /db_xref="taxon:316385"
gene        190..255
            /gene="thrL" /locus_tag="ECDH10B_0001"
            /db_xref="GeneID:6058969"
CDS         190..255
            /gene="thrL"
            /locus_tag="ECDH10B_0001"
            /note="involved in threonine biosynthesis; controls the expression of the thrLABC operon"
            /codon_start=1
            /transl_table=11
            /product="thr operon leader peptide"
            /protein_id="YP_001728984.1"
            /db_xref="GI:170079664"
            /db_xref="ASAP:AEC-0000073"
            /db_xref="GeneID:6058969"
gene        337..2799
            /gene="thrA"
            /locus_tag="ECDH10B_0002"
            /db_xref="GeneID:6058347“
……
……
```

http://www.ncbi.nlm.nih.gov/nuccore/170079663

How many reads in a fasta file?

- Extract lines with “>” character

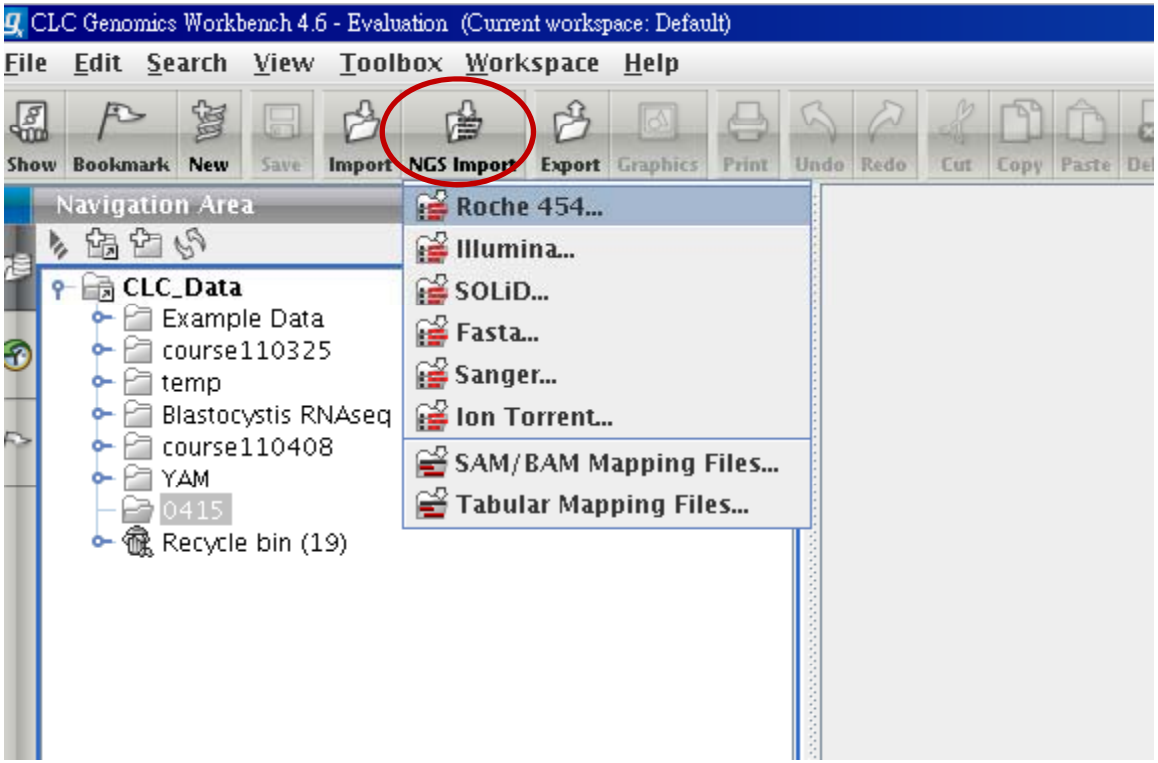
```
grep   “>”   Ecoli.FLX.fna
```

- And count it

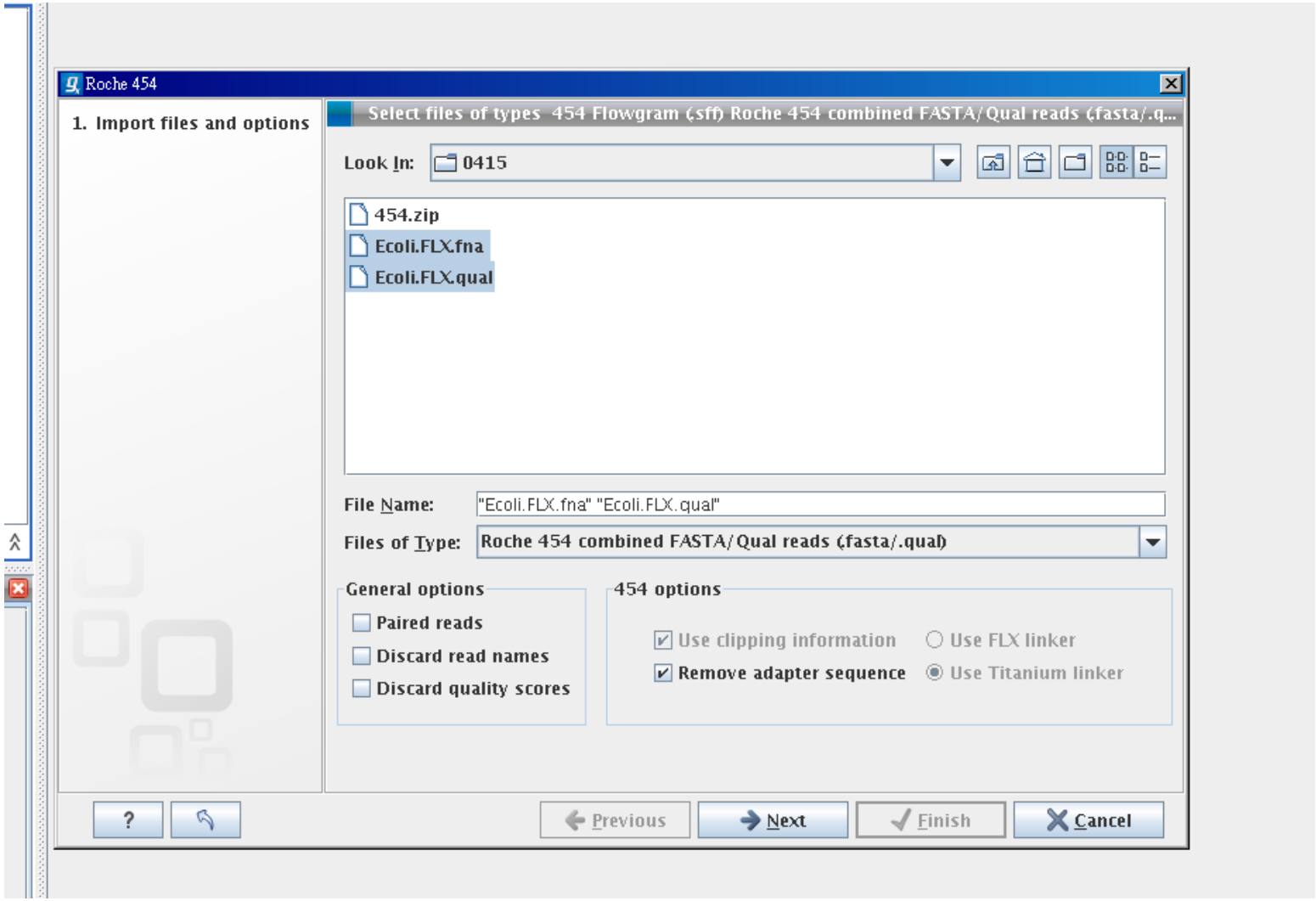
```
grep   -c  “>”   Ecoli.FLX.fna
```

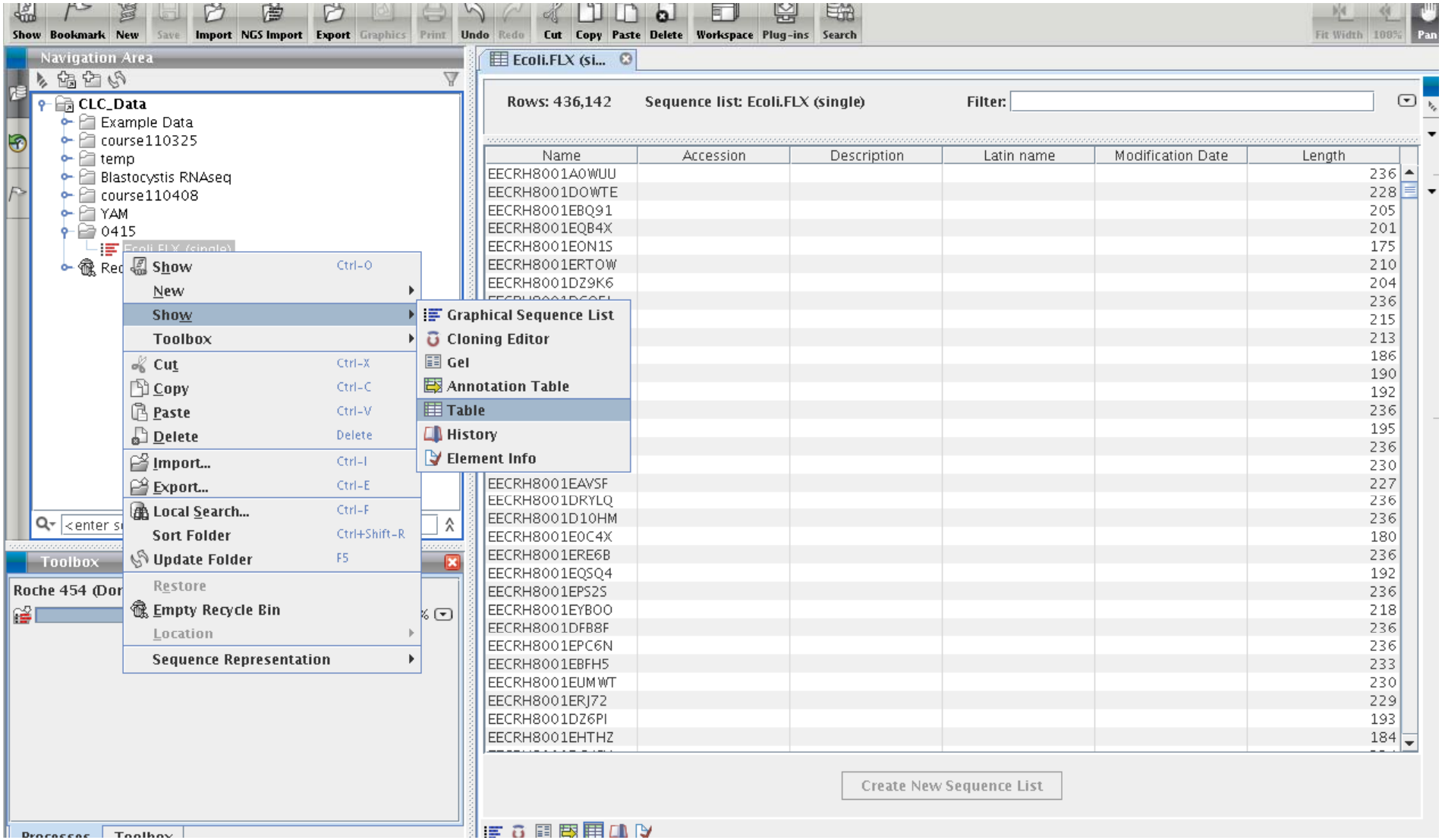
- How many reads in quality file?

Import 454 reads

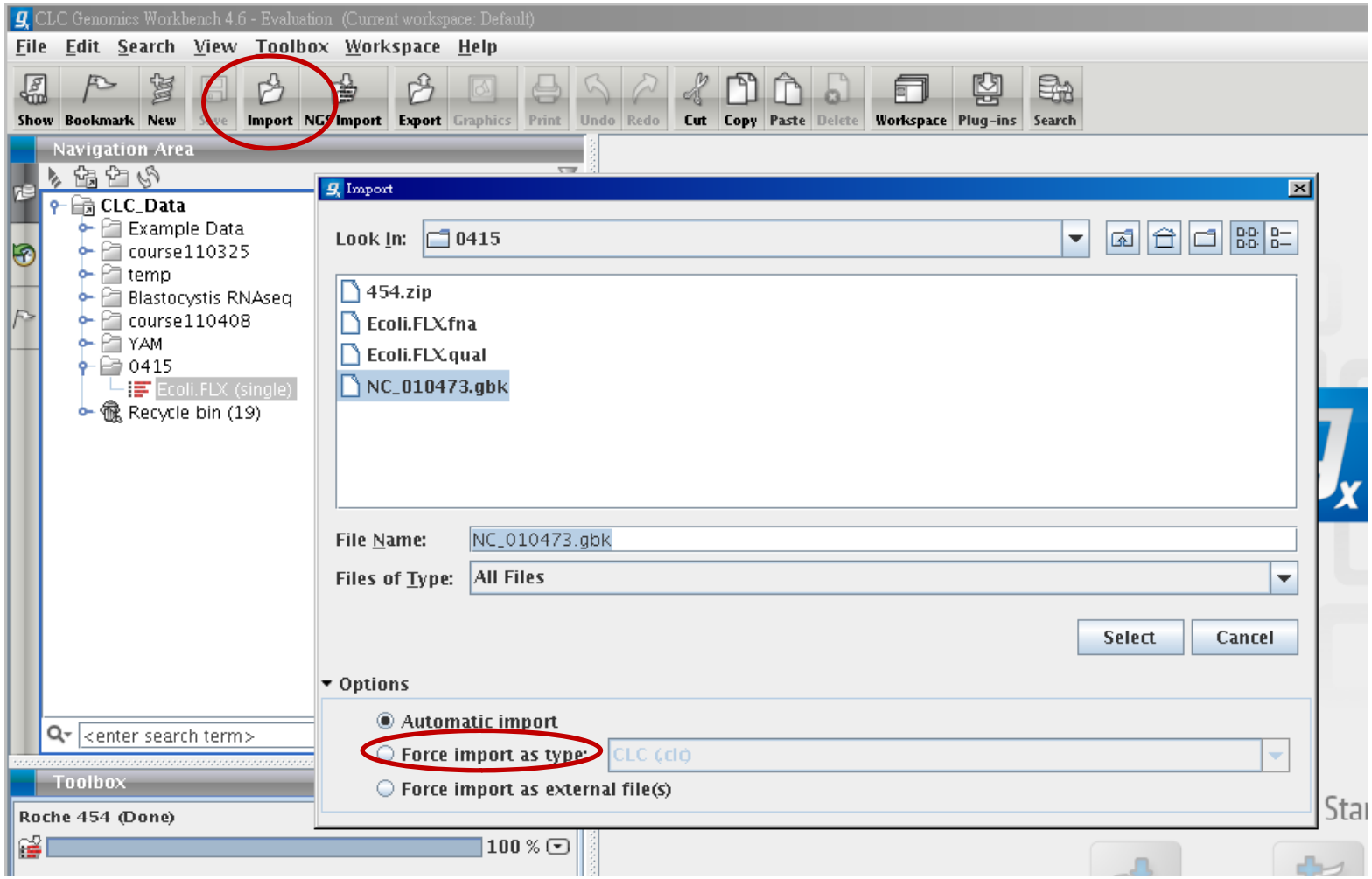


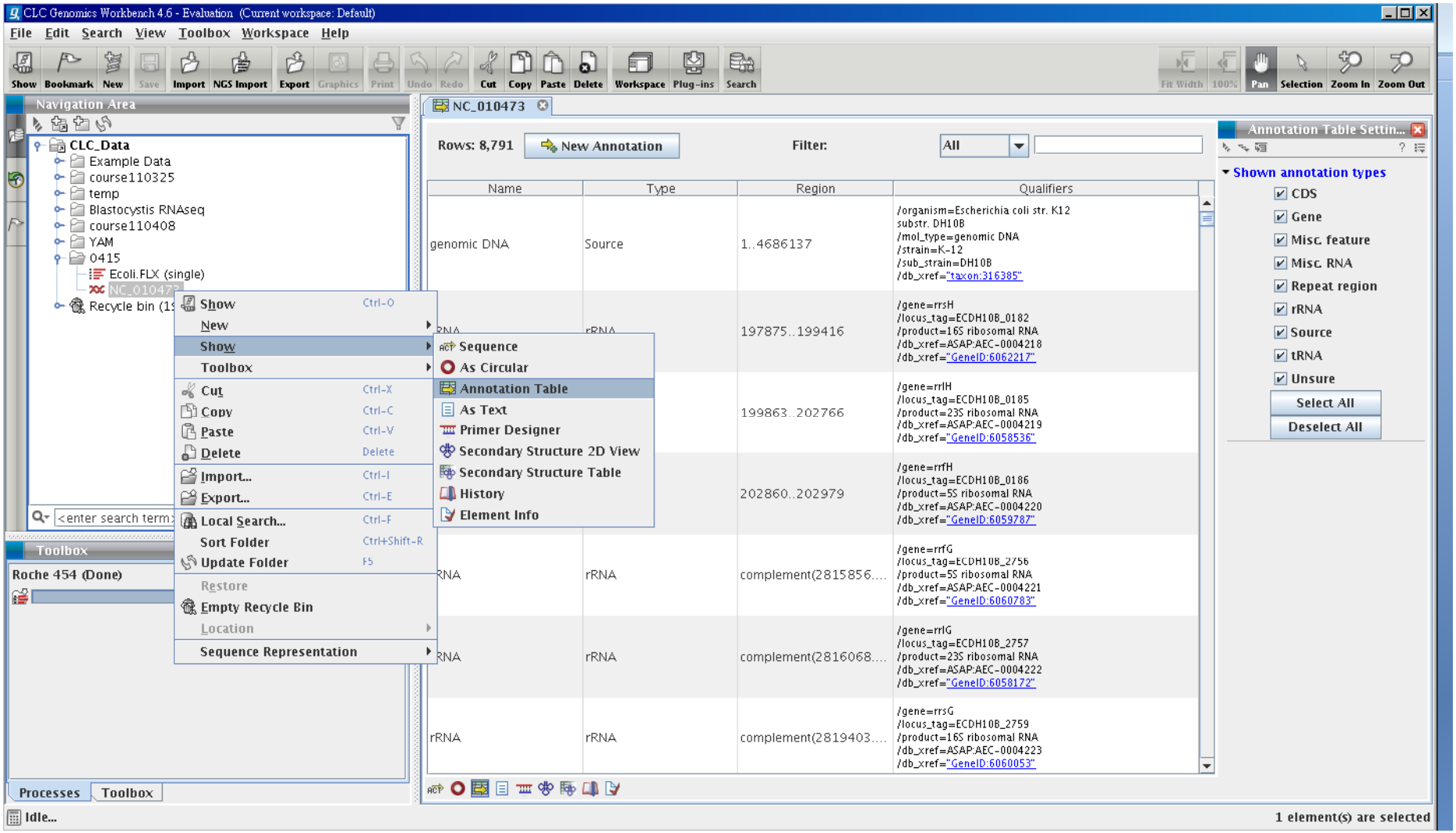
Select 2 files via **Ctrl** key



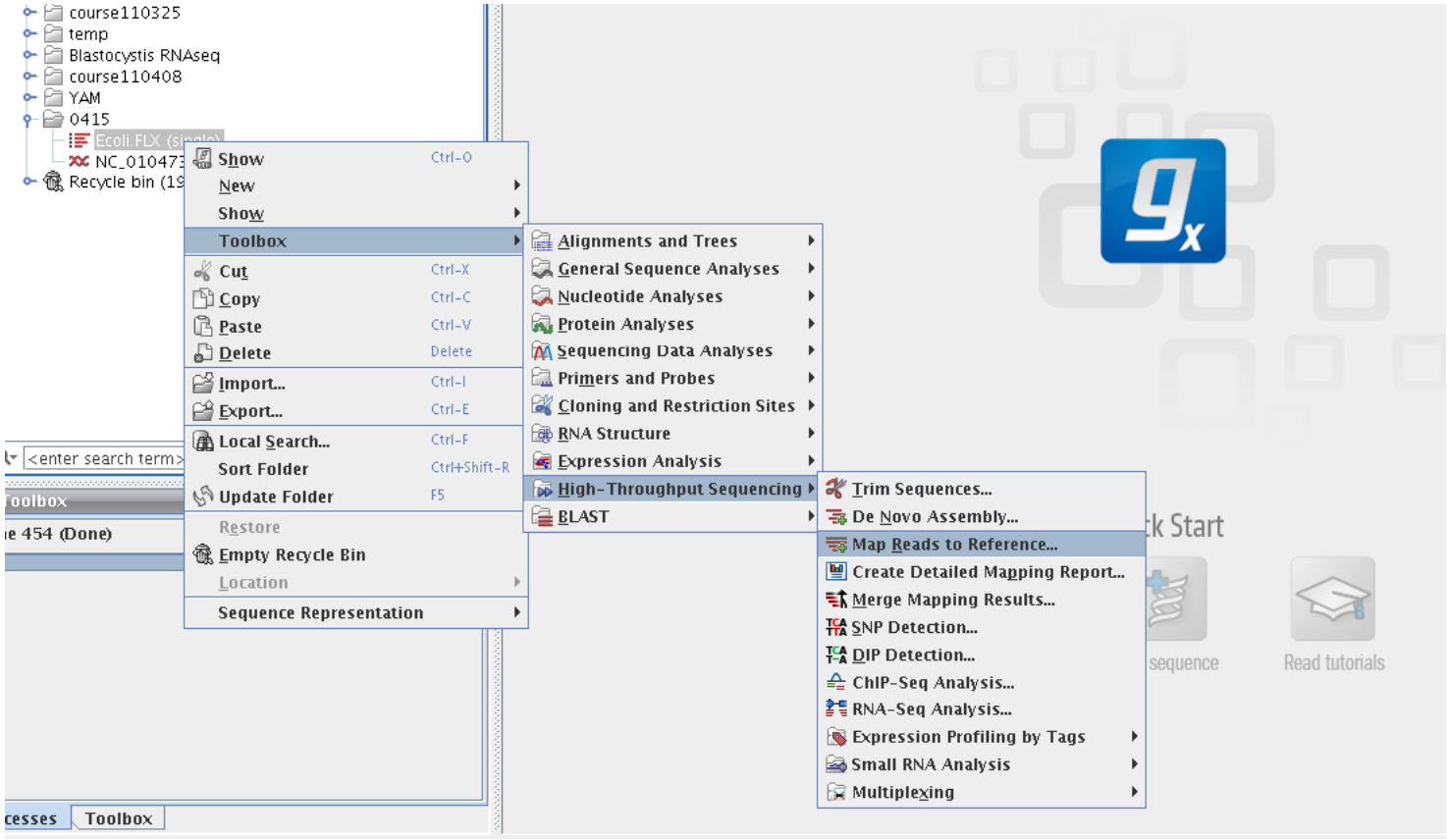

Show
Bookmark
New
Save
Import
NGS Import
Export
Graphics
Print
Undo
Redo
Cut
Copy
Paste
Delete
Workspace
Plug-ins
Search
Fit Width
100%
Pan
Navigation Area
CLC_Data
Example Data
course110325
temp
Blastocystis RNAseq
course110408
YAM
0415
Ecoli.FLX (single)
Show Ctrl-O
New
Show
Toolbox
Cut Ctrl-X
Copy Ctrl-C
Paste Ctrl-V
Delete Delete
Import... Ctrl-I
Export... Ctrl-E
Local Search... Ctrl-F
Sort Folder Ctrl+Shift-R
Update Folder F5
Restore
Empty Recycle Bin
Location
Sequence Representation
Graphical Sequence List
Cloning Editor
Gel
Annotation Table
Table
History
Element Info
Toolbox
Roche 454 (Dor
Processes
Toolbox
Ecoli.FLX (si...
Rows: 436,142
Sequence list: Ecoli.FLX (single)
Filter:
Name
Accession
Description
Latin name
Modification Date
Length
EECRH8001A0WUU 236
EECRH8001DOWTE 228
EECRH8001EBQ91 205
EECRH8001EQB4X 201
EECRH8001EON1S 175
EECRH8001ERTOW 210
EECRH8001DZ9K6 204
236
215
213
186
190
192
236
195
236
230
EECRH8001EAVSF 227
EECRH8001DRYLQ 236
EECRH8001D10HM 236
EECRH8001E0C4X 180
EECRH8001ERE6B 236
EECRH8001EQSQ4 192
EECRH8001EPS2S 236
EECRH8001EYBOO 218
EECRH8001DFB8F 236
EECRH8001EPC6N 236
EECRH8001EBFH5 233
EECRH8001EUMWT 230
EECRH8001ERJ72 229
EECRH8001DZ6PI 193
EECRH8001EHTHZ 184
Create New Sequence List

Import Reference genome

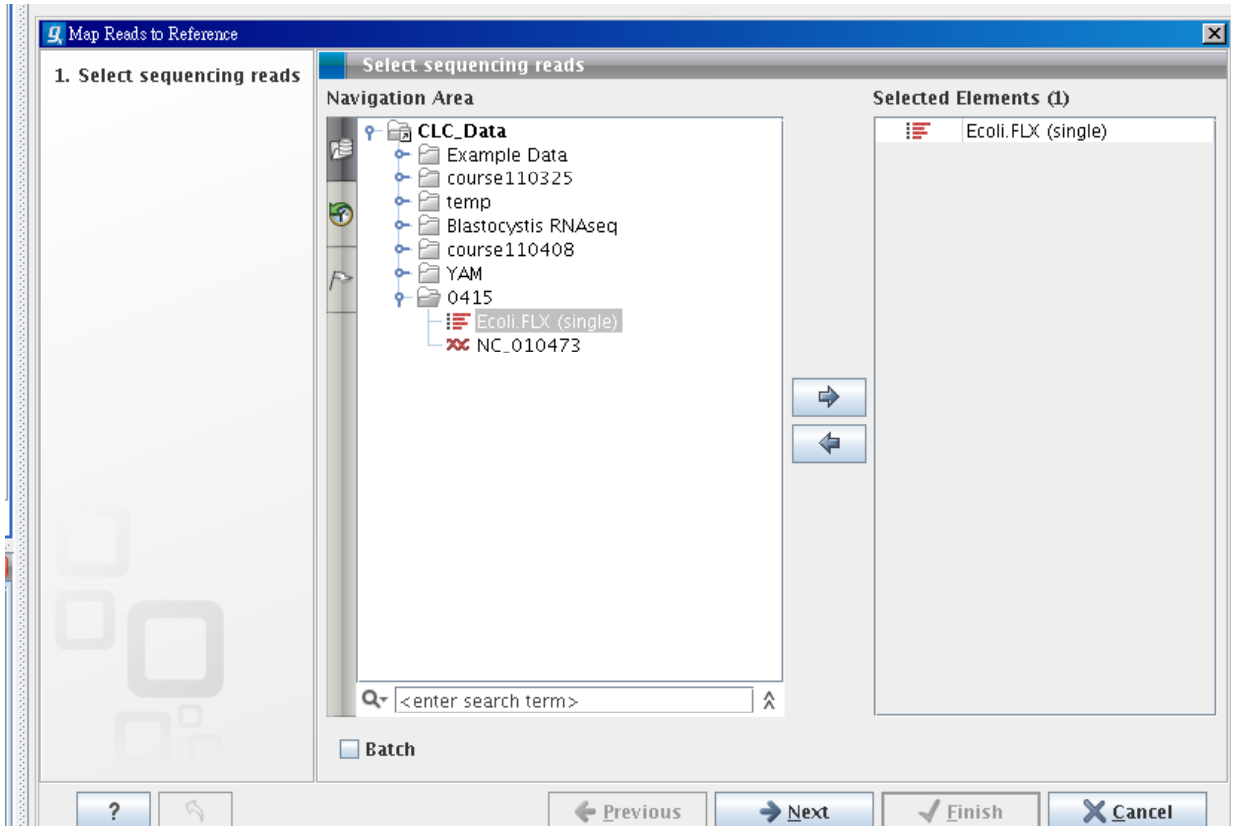



CLC Genomics Workbench 4.6 - Evaluation (Current workspace: Default)
File Edit Search View Toolbox Workspace Help
Show Bookmark New Save Import NGS Import Export Graphics Print Undo Redo Cut Copy Paste Delete Workspace Plug-ins Search
Fit Width 100% Pan Selection Zoom In Zoom Out
Navigation Area
CLC_Data
Example Data
course110325
temp
Blastocystis RNAseq
course110408
YAM
0415
Ecoli.FLX (single)
Recycle bin
Show Ctrl-O
New
Show
Toolbox
Cut Ctrl-X
Copy Ctrl-C
Paste Ctrl-V
Delete Delete
Import... Ctrl-I
Export... Ctrl-E
Local Search... Ctrl-F
Sort Folder Ctrl+Shift-R
Update Folder F5
Restore
Empty Recycle Bin
Location
Sequence Representation
Sequence
As Circular
Annotation Table
As Text
Primer Designer
Secondary Structure 2D View
Secondary Structure Table
History
Element Info
Toolbox
Roche 454 (Done)
Processes Toolbox
Idle...
NC_010473
Rows: 8,791
New Annotation
Filter:
All
Name Type Region Qualifiers
genomic DNA Source 1..4686137 /organism=Escherichia coli str. K12 substr. DH10B /mol_type=genomic DNA /strain=K-12 /sub_strain=DH10B /db_xref="taxon:316385"
197875..199416 /gene=rrsH /locus_tag=ECDH10B_0182 /product=16S ribosomal RNA /db_xref=ASAP:AEC-0004218 /db_xref="GeneID:6062217"
199863..202766 /gene=rrlH /locus_tag=ECDH10B_0185 /product=23S ribosomal RNA /db_xref=ASAP:AEC-0004219 /db_xref="GeneID:6058536"
202860..202979 /gene=rrfH /locus_tag=ECDH10B_0186 /product=5S ribosomal RNA /db_xref=ASAP:AEC-0004220 /db_xref="GeneID:6059787"
rRNA rRNA complement(2815856.... /gene=rrfG /locus_tag=ECDH10B_2756 /product=5S ribosomal RNA /db_xref=ASAP:AEC-0004221 /db_xref="GeneID:6060783"
rRNA rRNA complement(2816068.... /gene=rrlG /locus_tag=ECDH10B_2757 /product=23S ribosomal RNA /db_xref=ASAP:AEC-0004222 /db_xref="GeneID:6058172"
rRNA rRNA complement(2819403.... /gene=rrsG /locus_tag=ECDH10B_2759 /product=16S ribosomal RNA /db_xref=ASAP:AEC-0004223 /db_xref="GeneID:6060053"
Annotation Table Settin...
Shown annotation types
CDS
Gene
Misc. feature
Misc. RNA
Repeat region
rRNA
Source
tRNA
Unsure
Select All
Deselect All
1 element(s) are selected

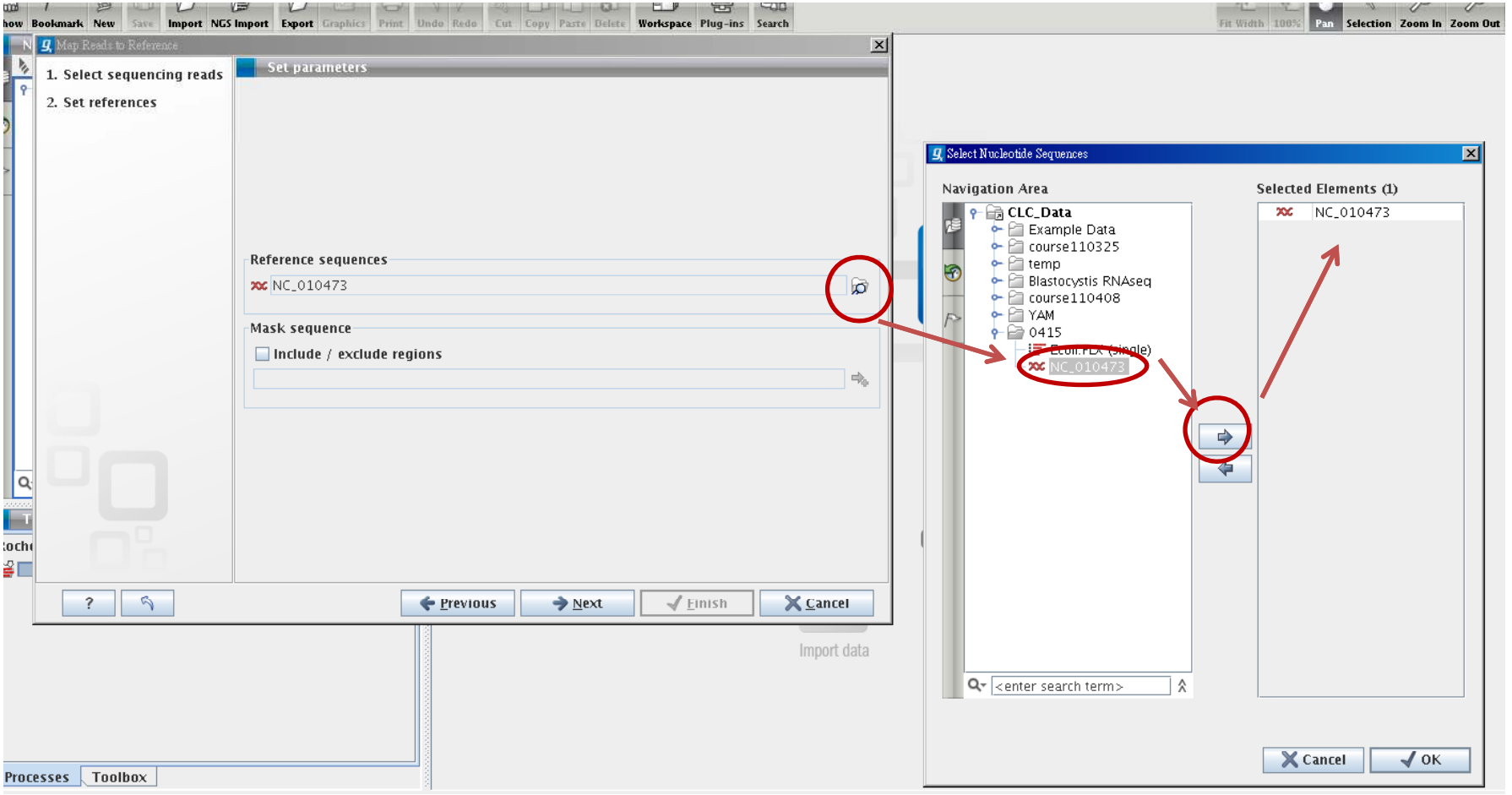
Mapping reads to Reference

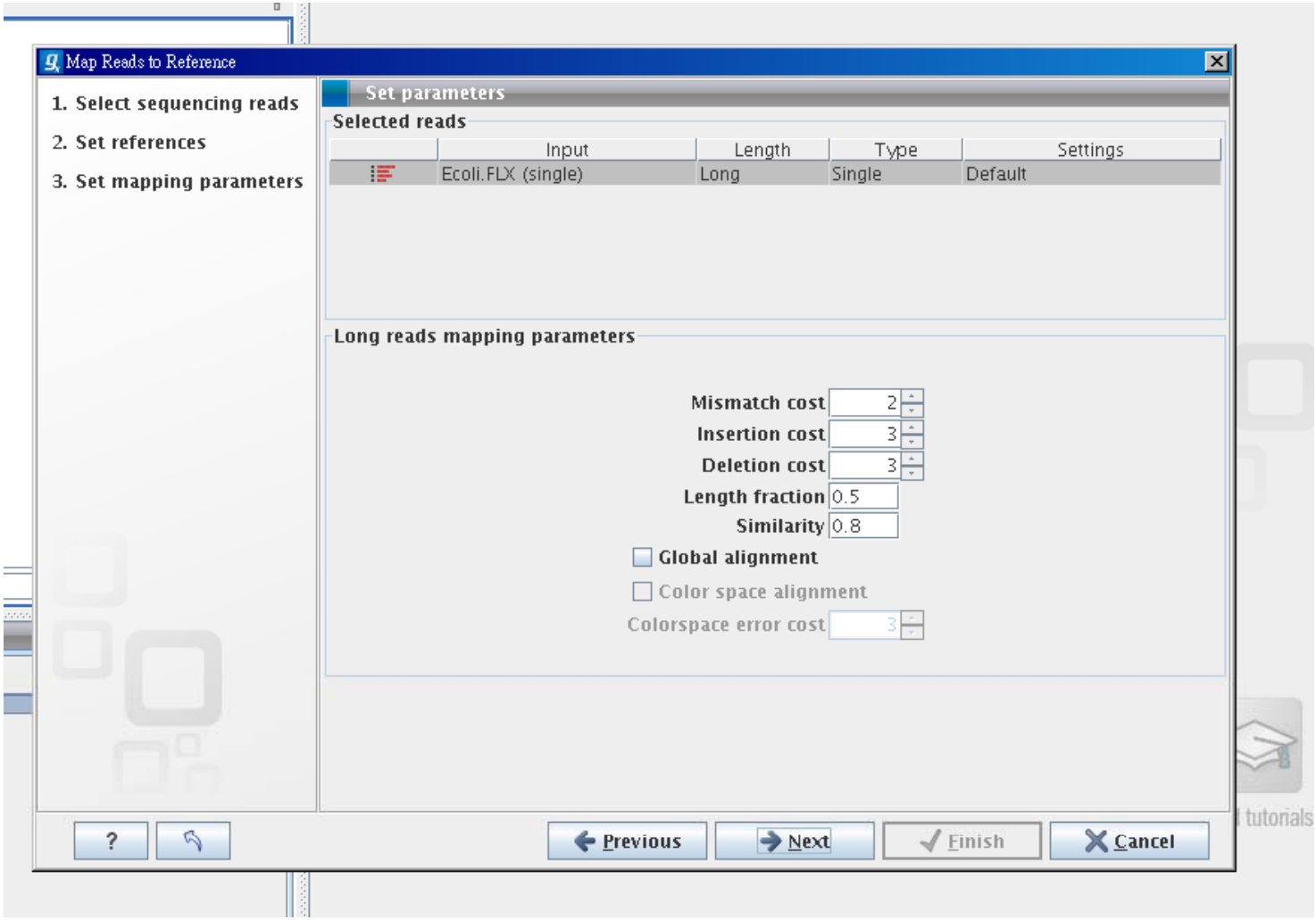


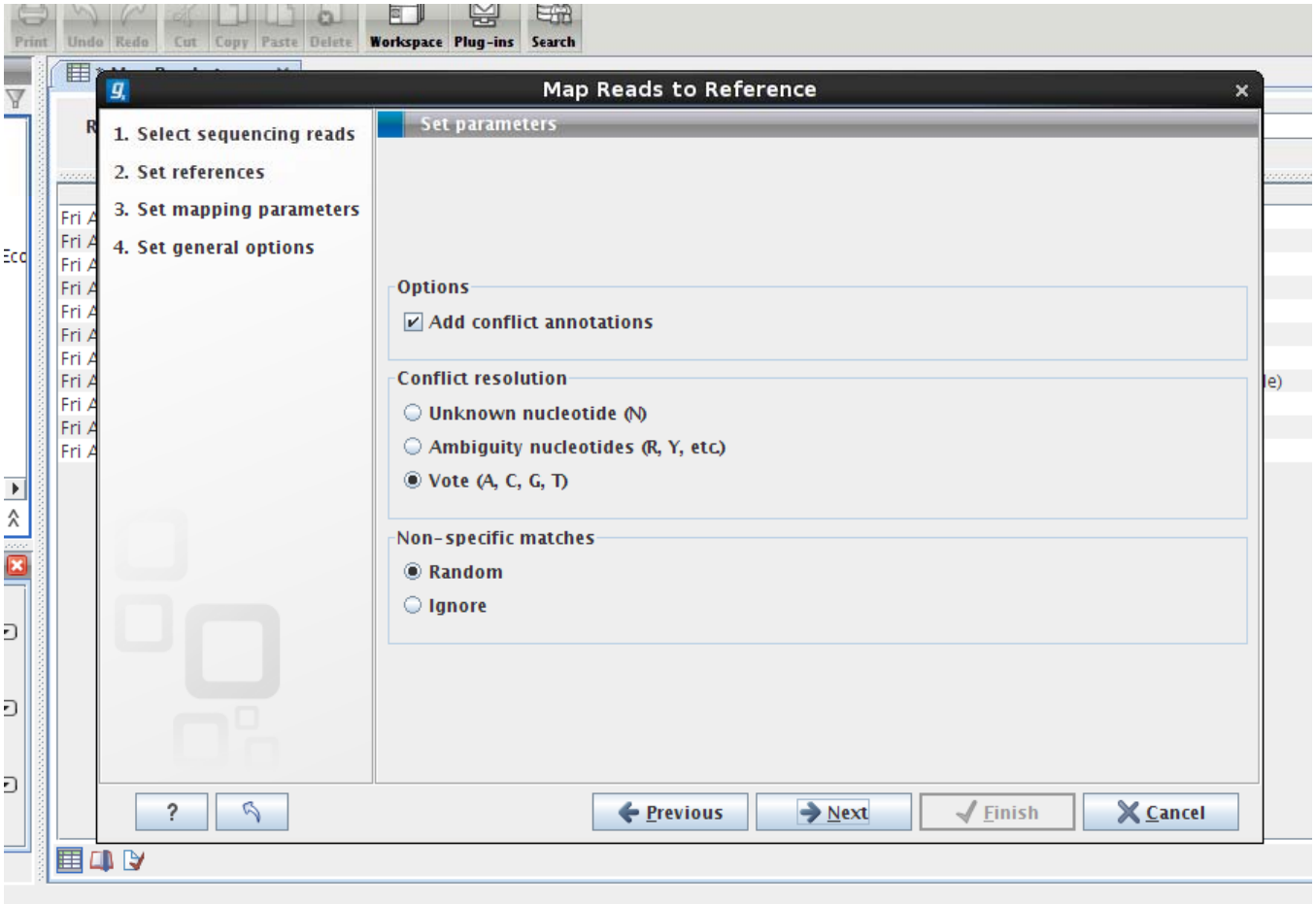
Select reads



Select NC_010473 as reference



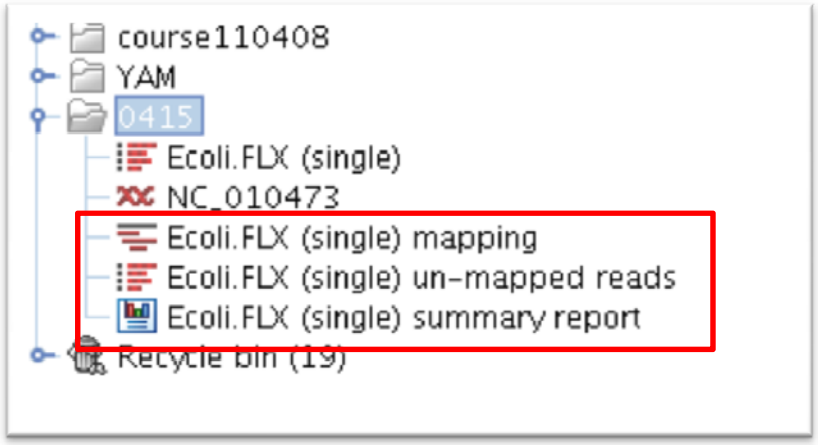

Map Reads to Reference
1. Select sequencing reads
2. Set references
3. Set mapping parameters
Set parameters
Selected reads
Input
Length
Type
Settings
Ecoli.FLX (single)
Long
Single
Default
Long reads mapping parameters
Mismatch cost
2
Insertion cost
3
Deletion cost
3
Length fraction
0.5
Similarity
0.8
Global alignment
Color space alignment
Colorspace error cost
3
?
Previous
Next
Finish
Cancel
tutorials


Print
Undo
Redo
Cut
Copy
Paste
Delete
Workspace
Plug-ins
Search
Map Reads to Reference
1. Select sequencing reads
2. Set references
3. Set mapping parameters
4. Set general 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
?
Previous
Next
Finish
Cancel

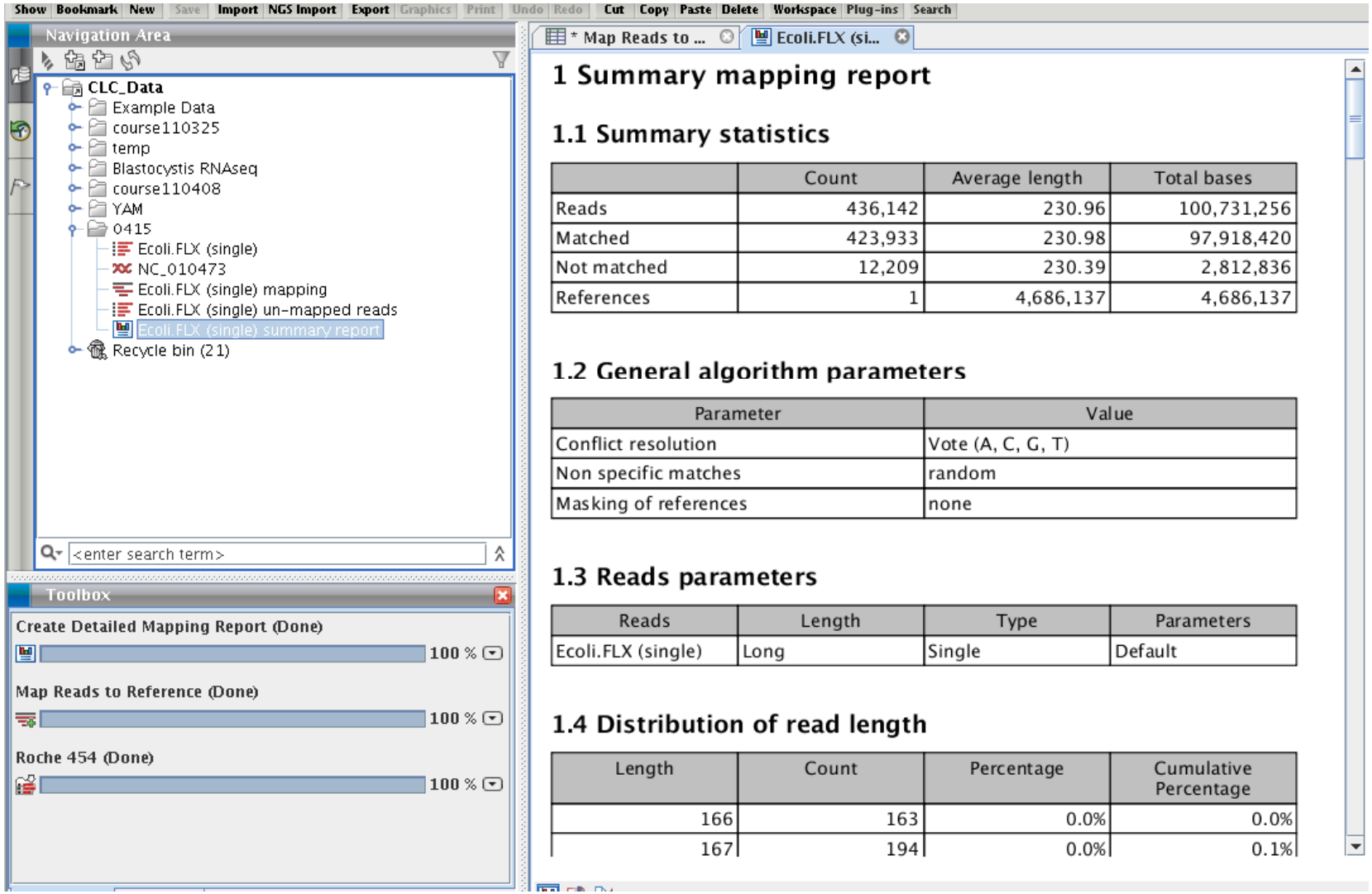
IUPAC nucleotide code

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Mapping result



Mapping report



1 Summary mapping report

1.1 Summary statistics

	Count	Average length	Total bases
Reads	436,142	230.96	100,731,256
Matched	423,933	230.98	97,918,420
Not matched	12,209	230.39	2,812,836
References	1	4,686,137	4,686,137

1.2 General algorithm parameters

Parameter	Value
Conflict resolution	Vote (A, C, G, T)
Non specific matches	random
Masking of references	none

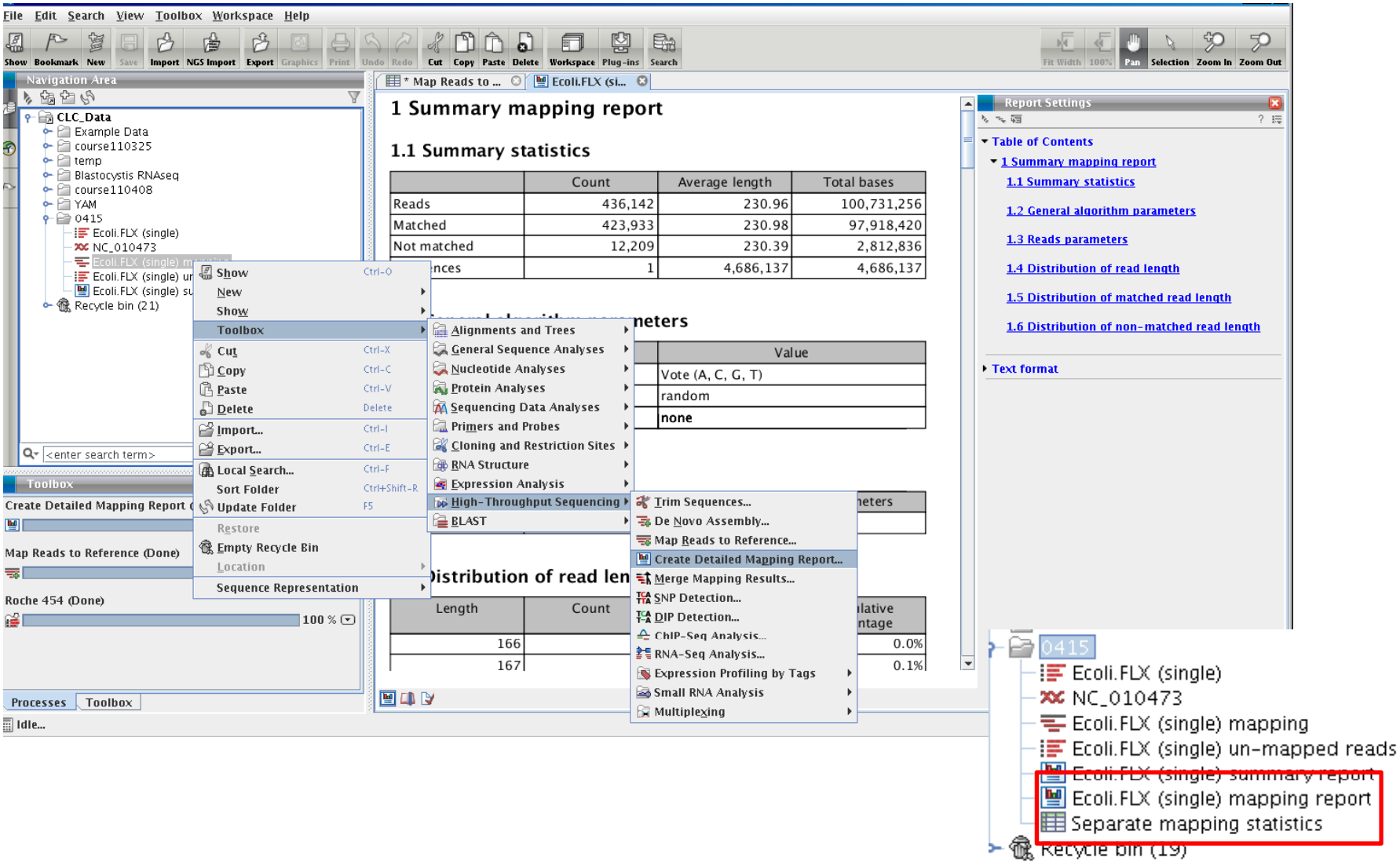
1.3 Reads parameters

Reads	Length	Type	Parameters
Ecoli.FLX (single)	Long	Single	Default

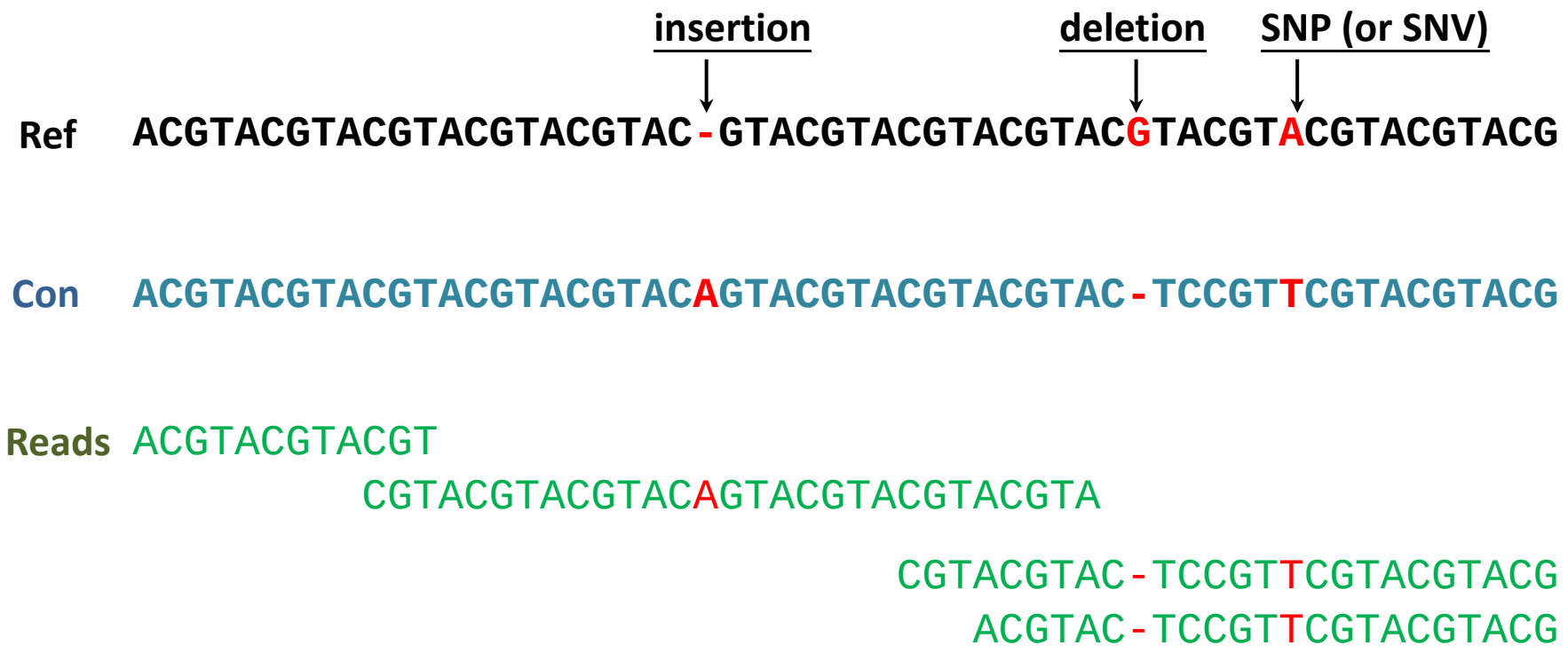
1.4 Distribution of read length

Length	Count	Percentage	Cumulative Percentage
166	163	0.0%	0.0%
167	194	0.0%	0.1%

Create detail report



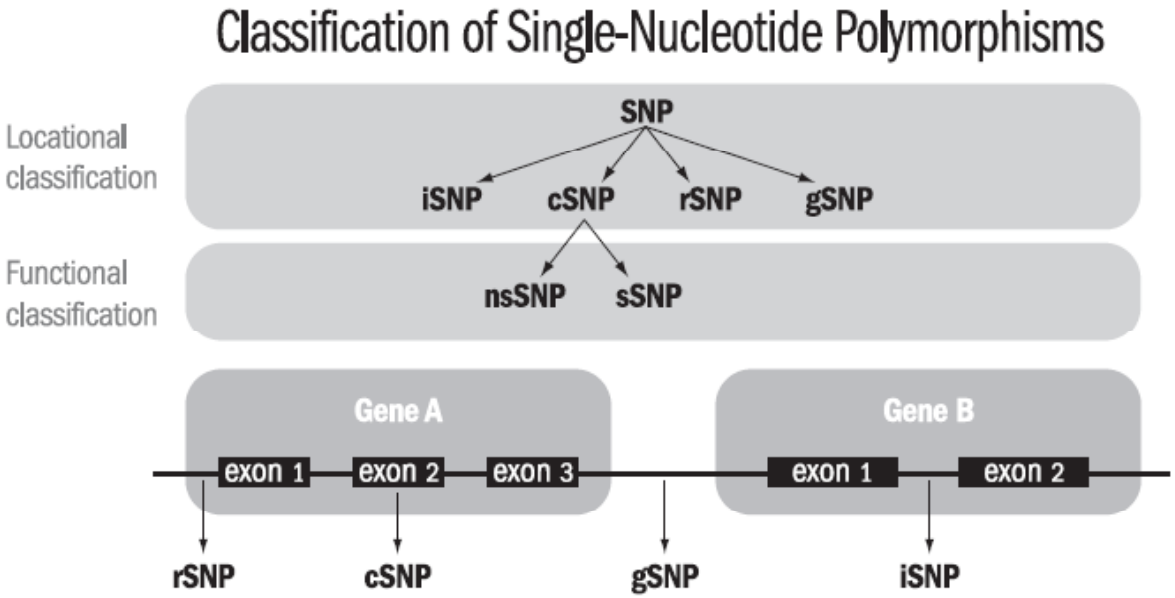
Reference and consensus sequence



SNP

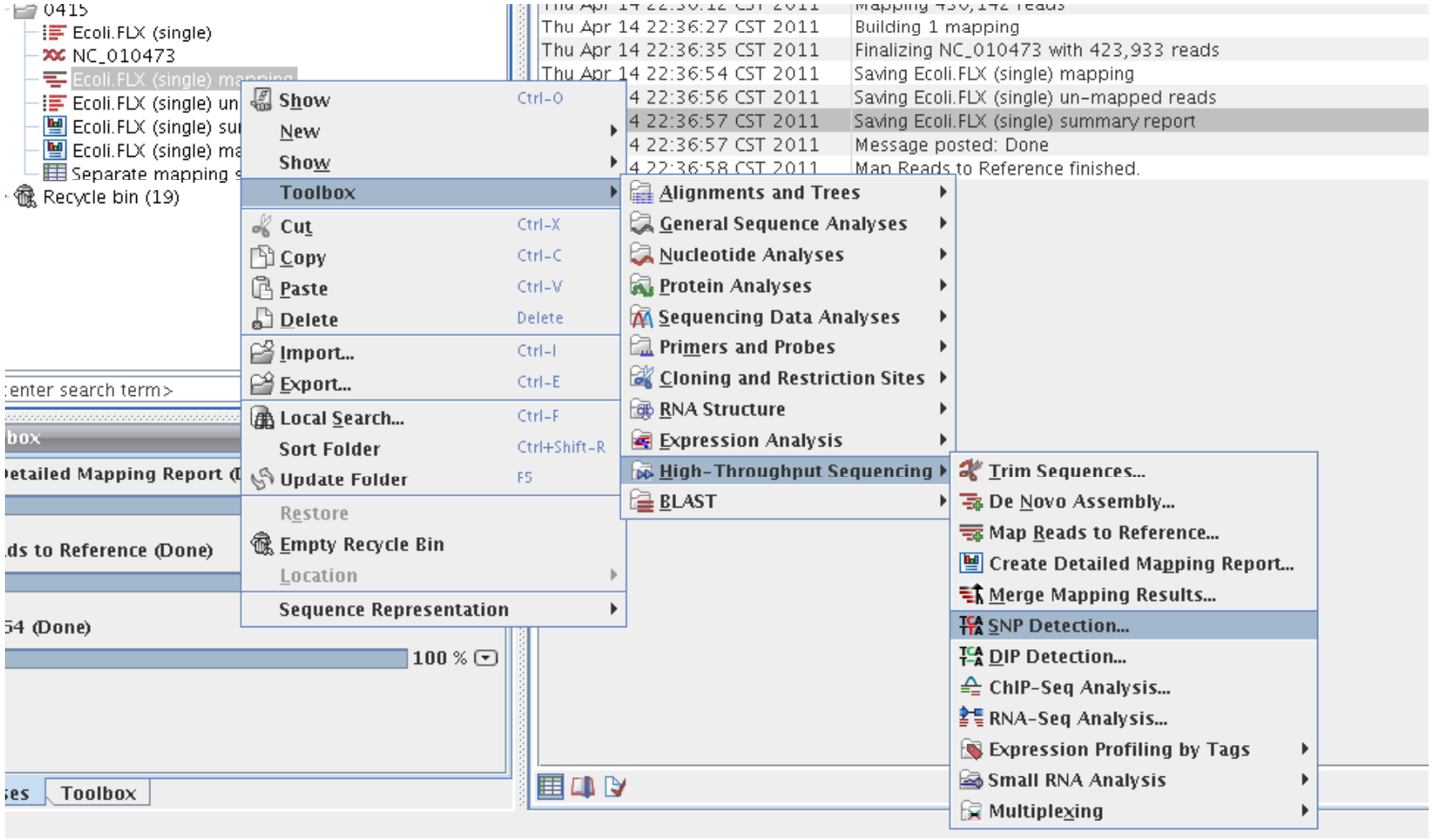
- gSNPs
  - intergenic regions.
- rSNPs
  - regulatory regions,
- iSNPs
  - intronic regions
- cSNPs
  - coding regions (exons)

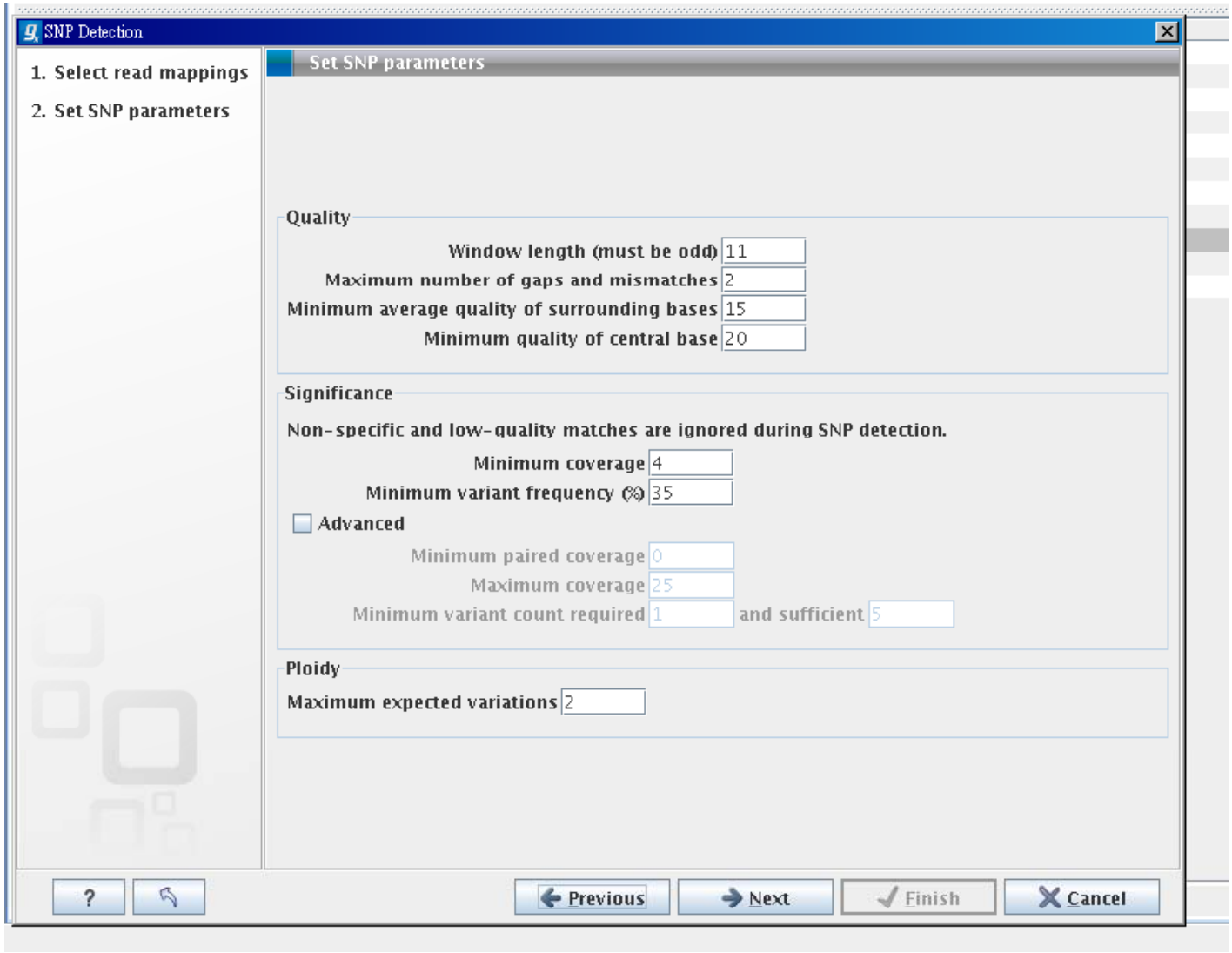
- sSNP
  - synonymous SNP
- nsSNP
  - Non-synonymous SNP



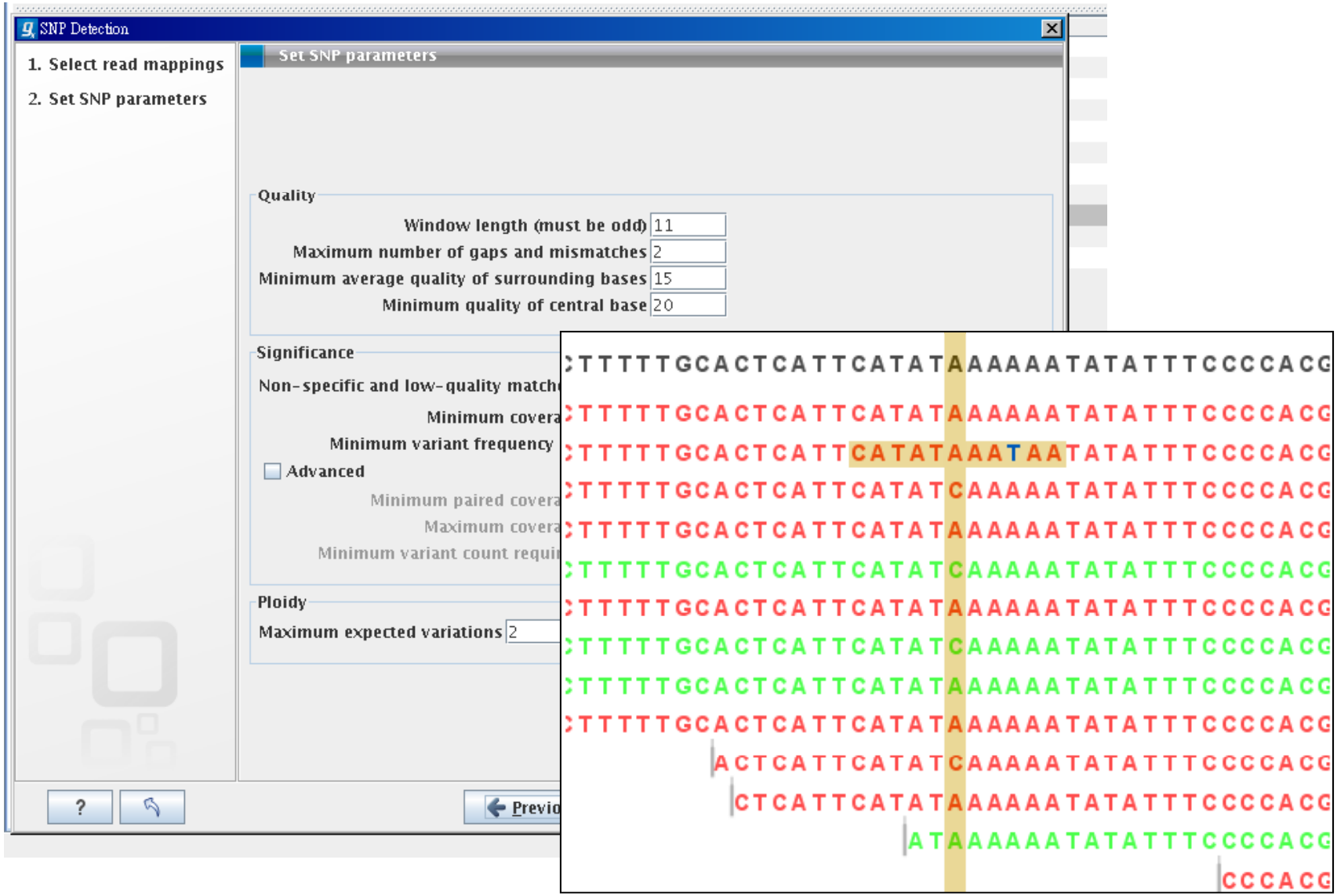
http://www.clcbio.com/sciencearticles/BE-SNP.pdf

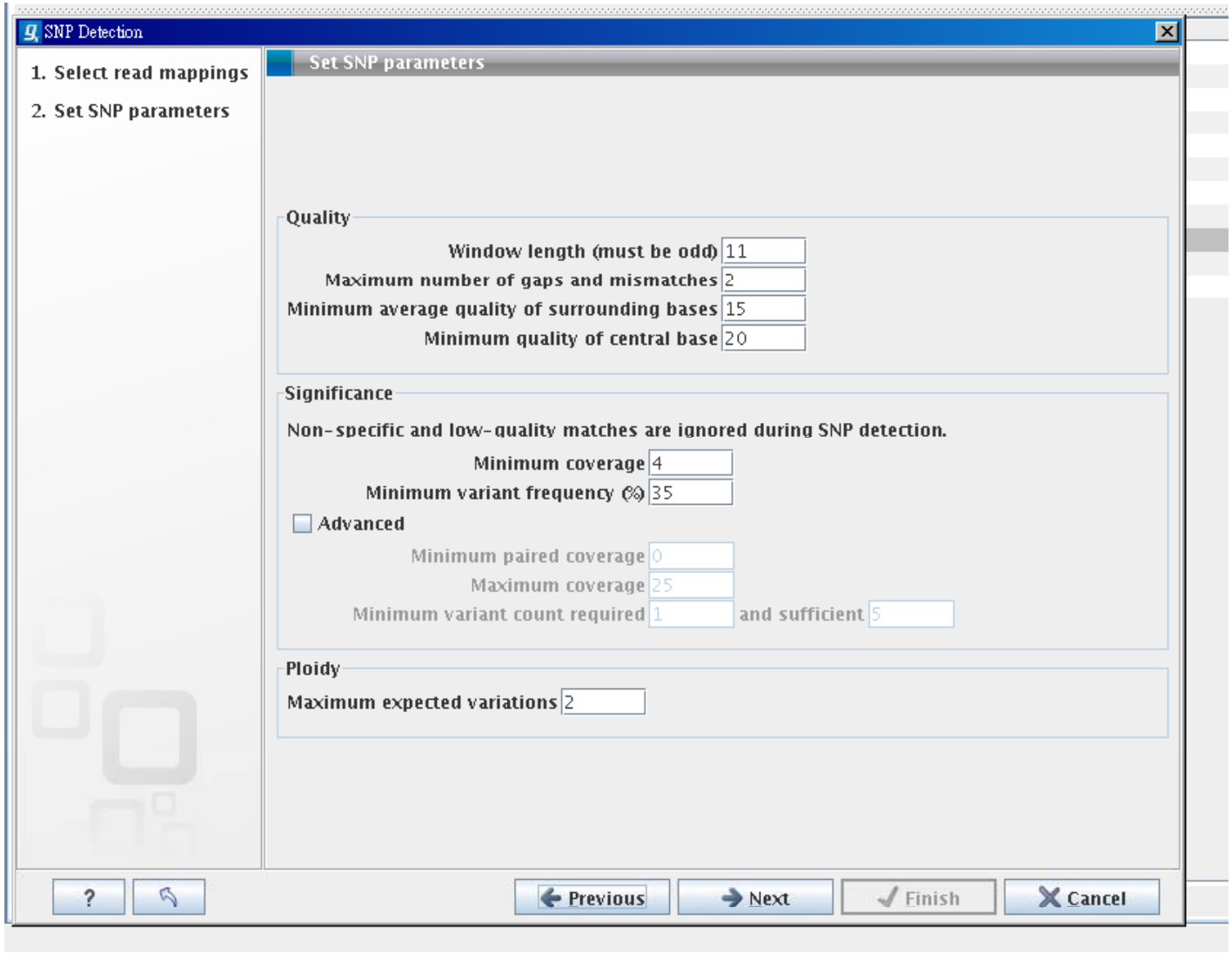
SNP detection

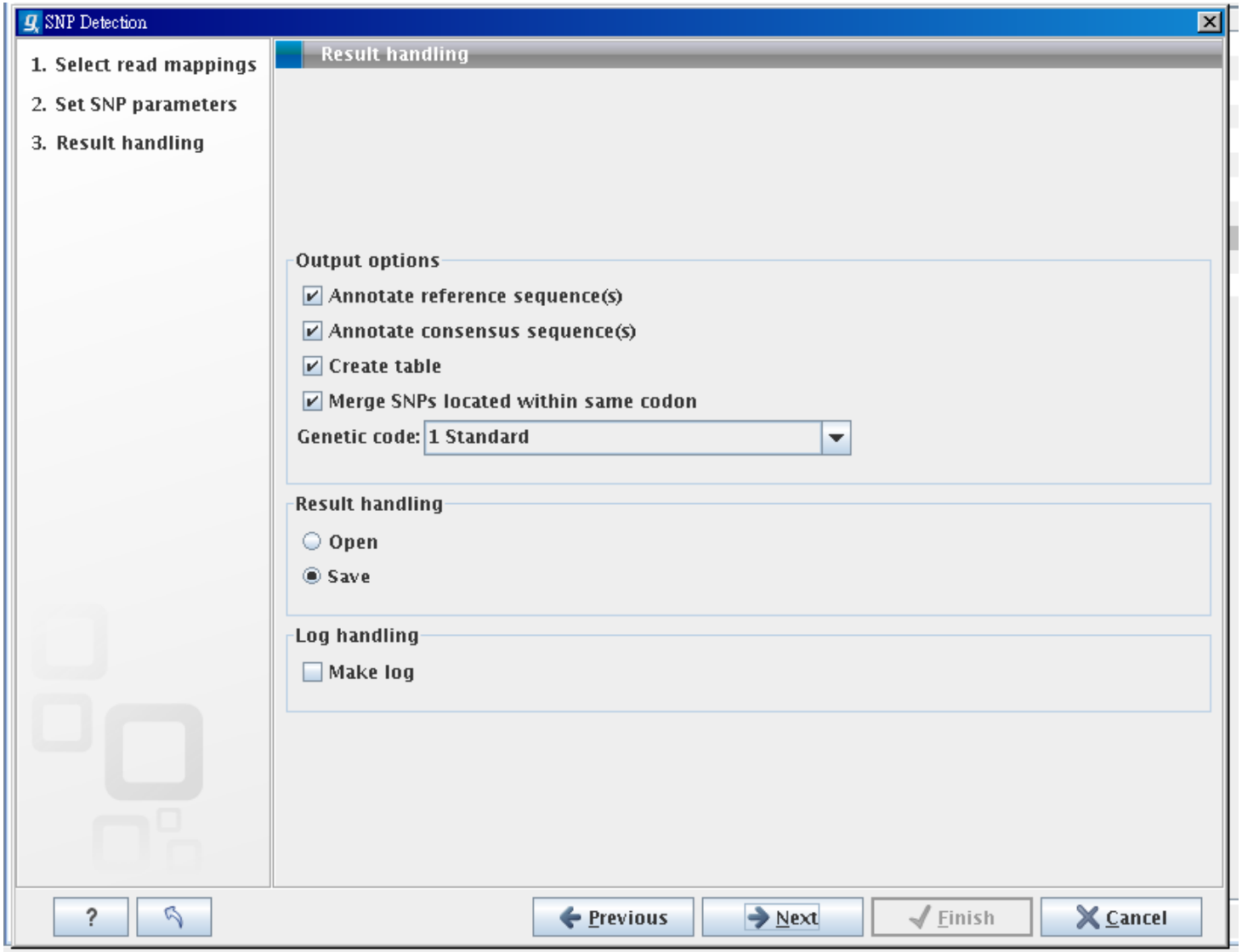



SNP Detection
1. Select read mappings
2. Set SNP parameters
Set SNP parameters
Quality
Window length (must be odd) 11
Maximum number of gaps and mismatches 2
Minimum average quality of surrounding bases 15
Minimum quality of central base 20
Significance
Non-specific and low-quality matches are ignored during SNP detection.
Minimum coverage 4
Minimum variant frequency (%) 35
Advanced
Minimum paired coverage 0
Maximum coverage 25
Minimum variant count required 1 and sufficient 5
Ploidy
Maximum expected variations 2
?
Previous
Next
Finish
Cancel

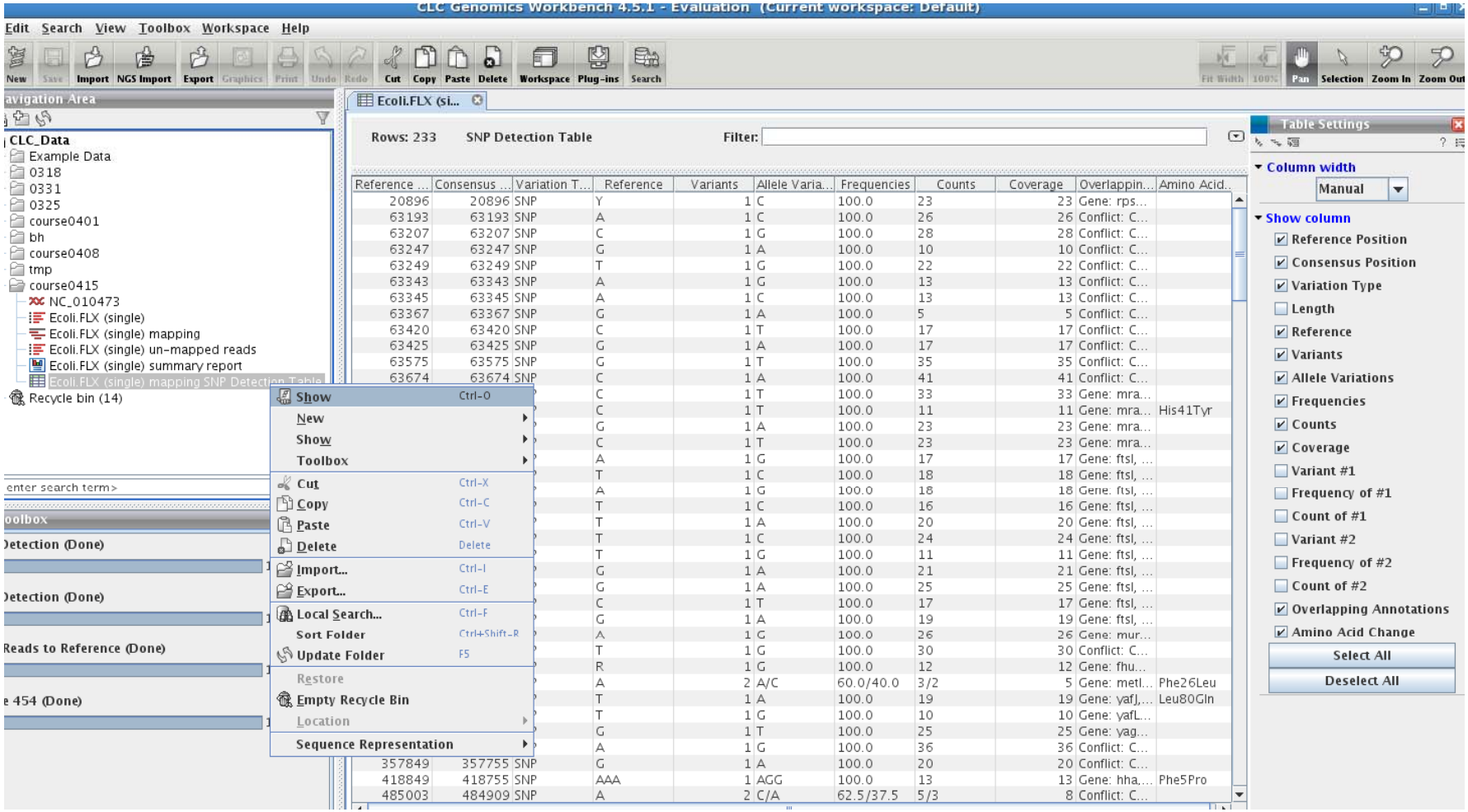
filter of quality



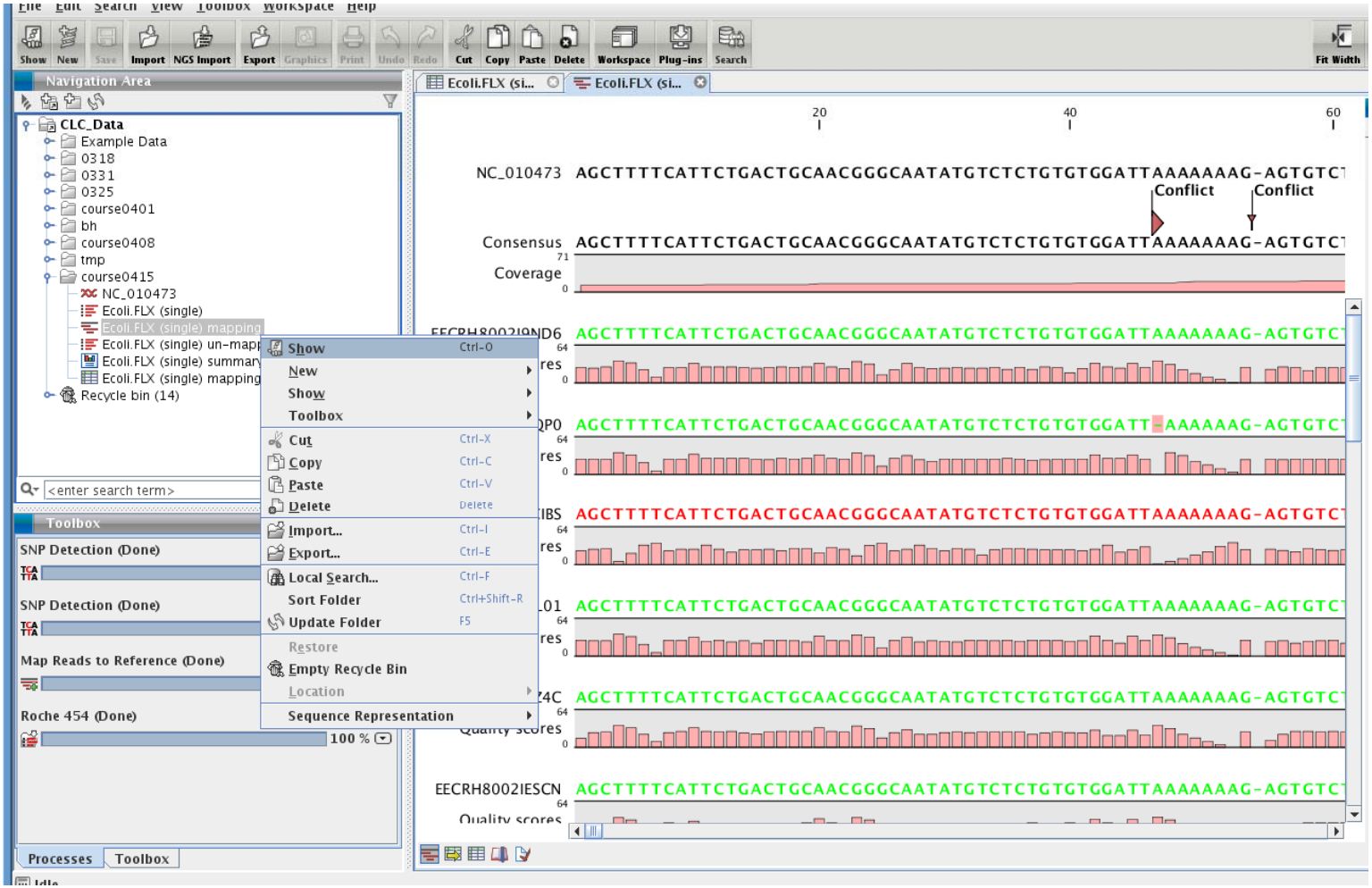

SNP Detection
1. Select read mappings
2. Set SNP parameters
Set SNP parameters
Quality
Window length (must be odd) 11
Maximum number of gaps and mismatches 2
Minimum average quality of surrounding bases 15
Minimum quality of central base 20
Significance
Non-specific and low-quality matches are ignored during SNP detection.
Minimum coverage 4
Minimum variant frequency (%) 35
Advanced
Minimum paired coverage 0
Maximum coverage 25
Minimum variant count required 1 and sufficient 5
Ploidy
Maximum expected variations 2
?
Previous
Next
Finish
Cancel

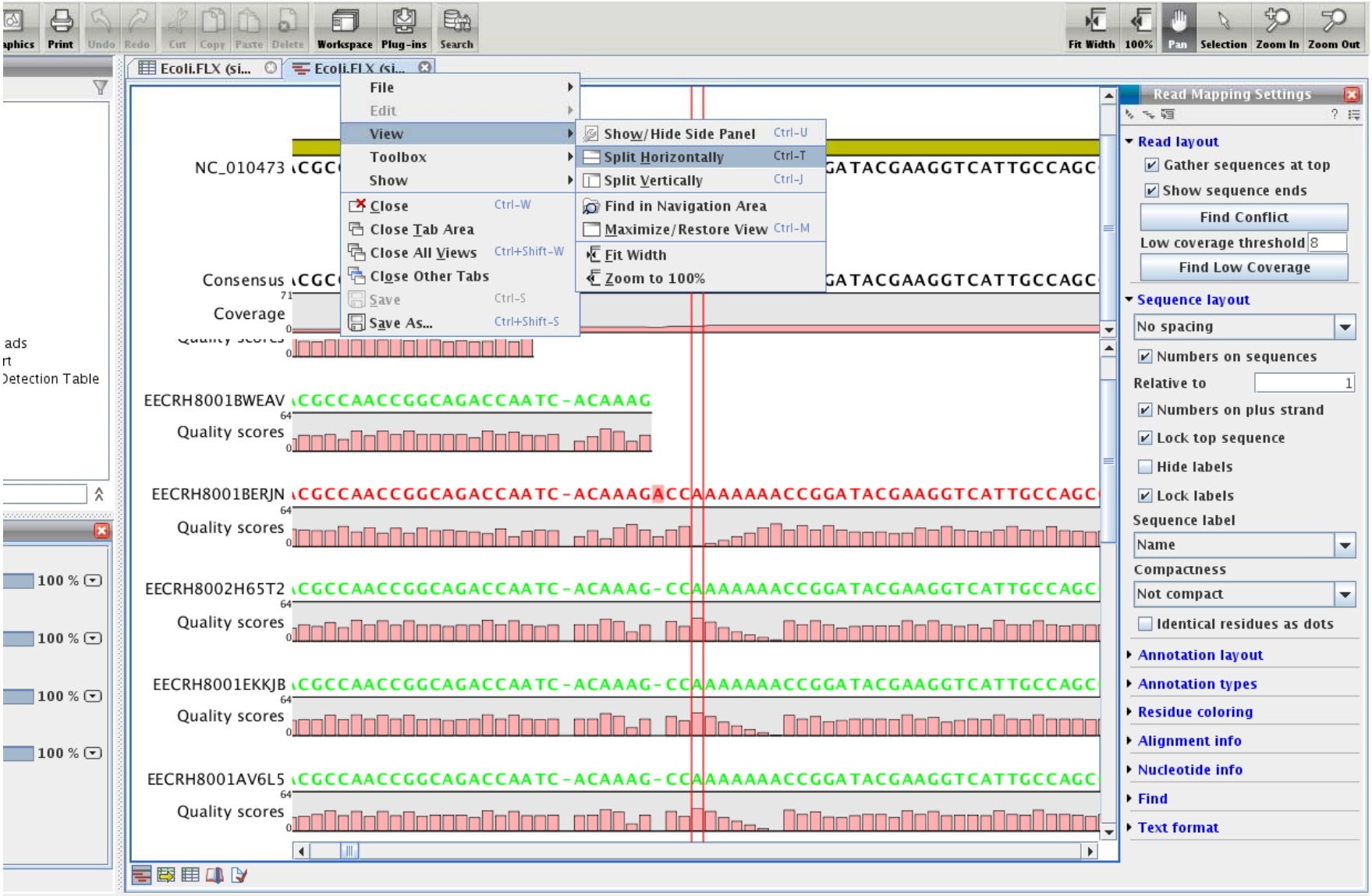

SNP Detection
1. Select read mappings
2. Set SNP parameters
3. Result handling
Result handling
Output options
Annotate reference sequence(s)
Annotate consensus sequence(s)
Create table
Merge SNPs located within same codon
Genetic code: 1 Standard
Result handling
Open
Save
Log handling
Make log
?
Previous
Next
Finish
Cancel

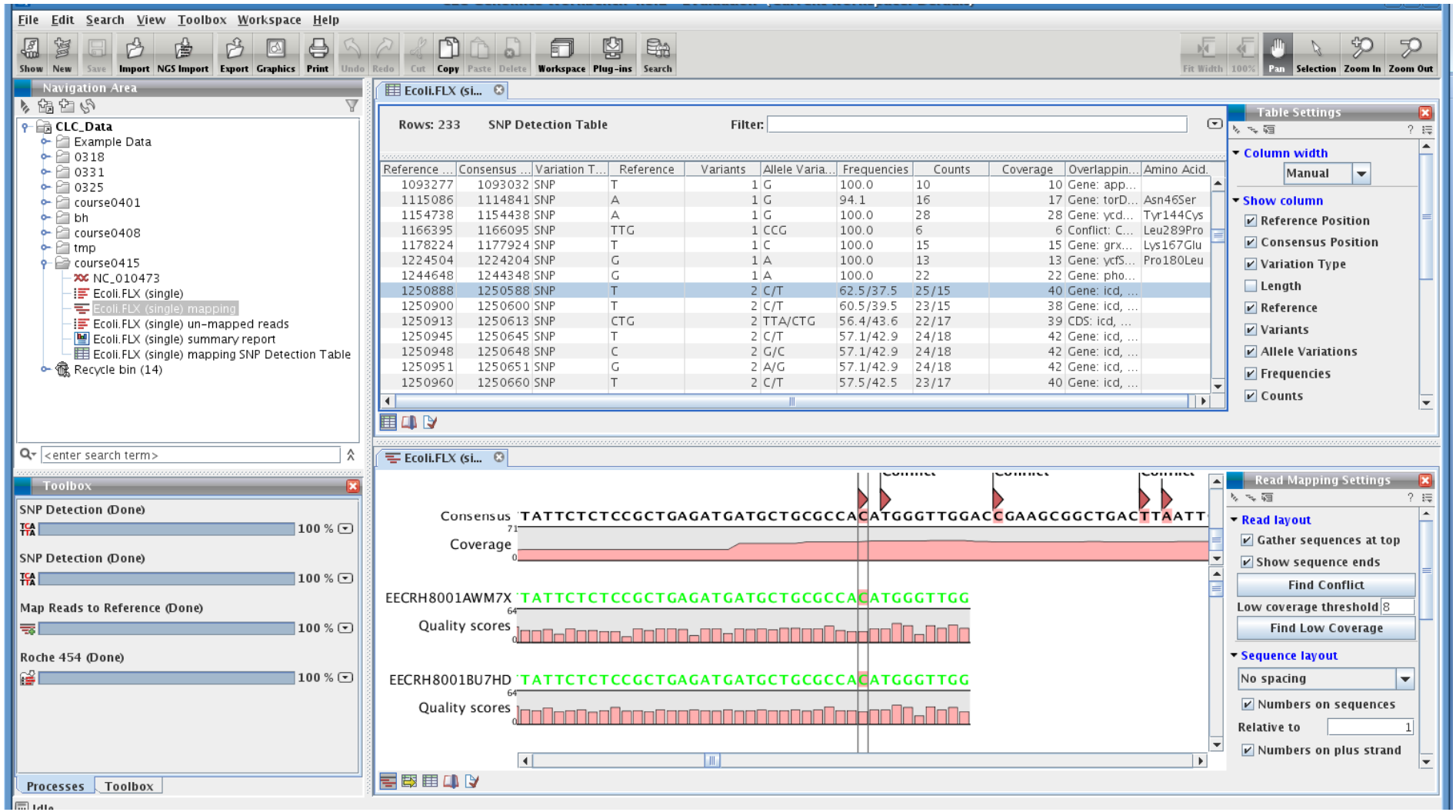
show SNP table



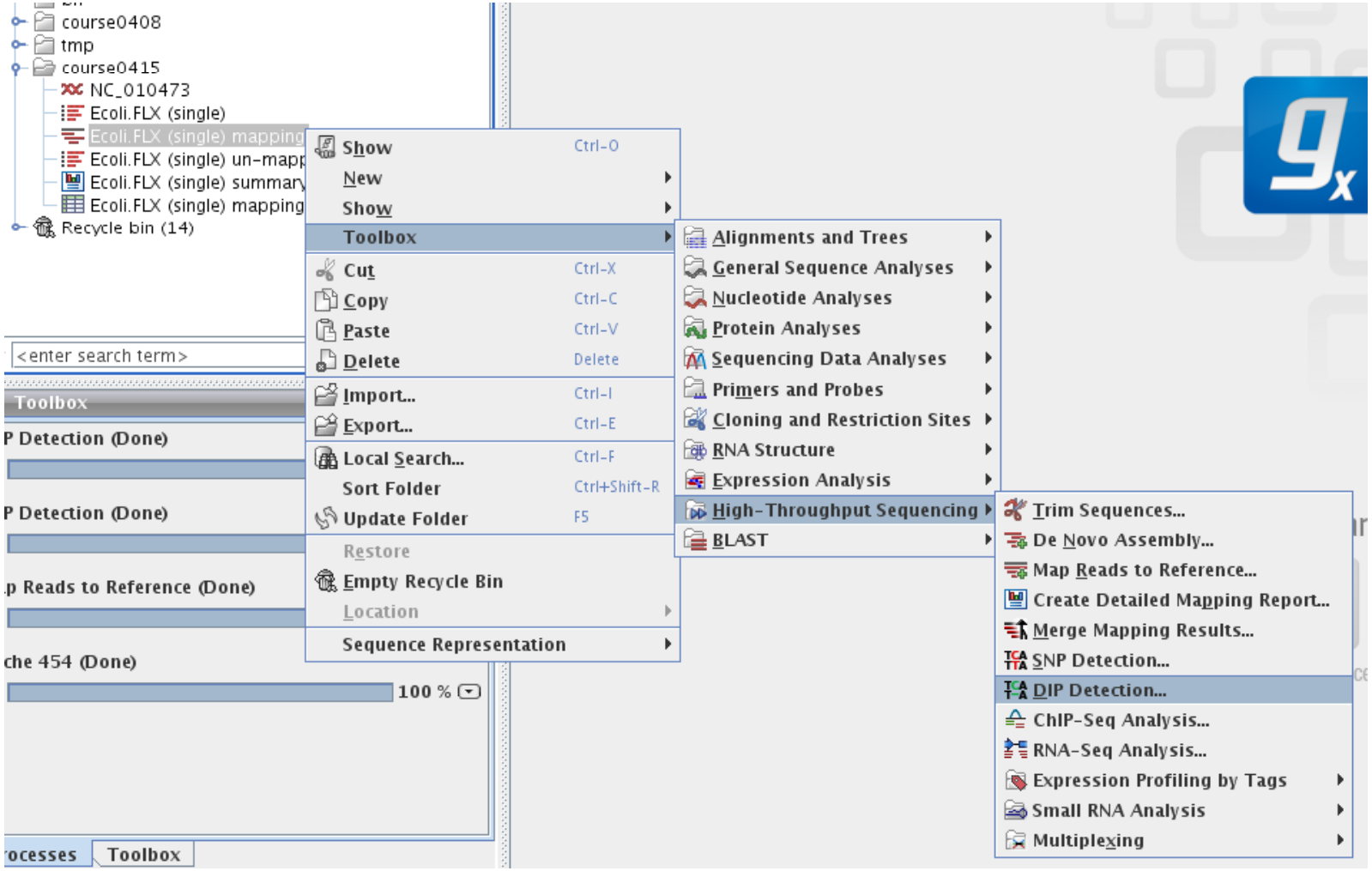
show genome/reads mapping

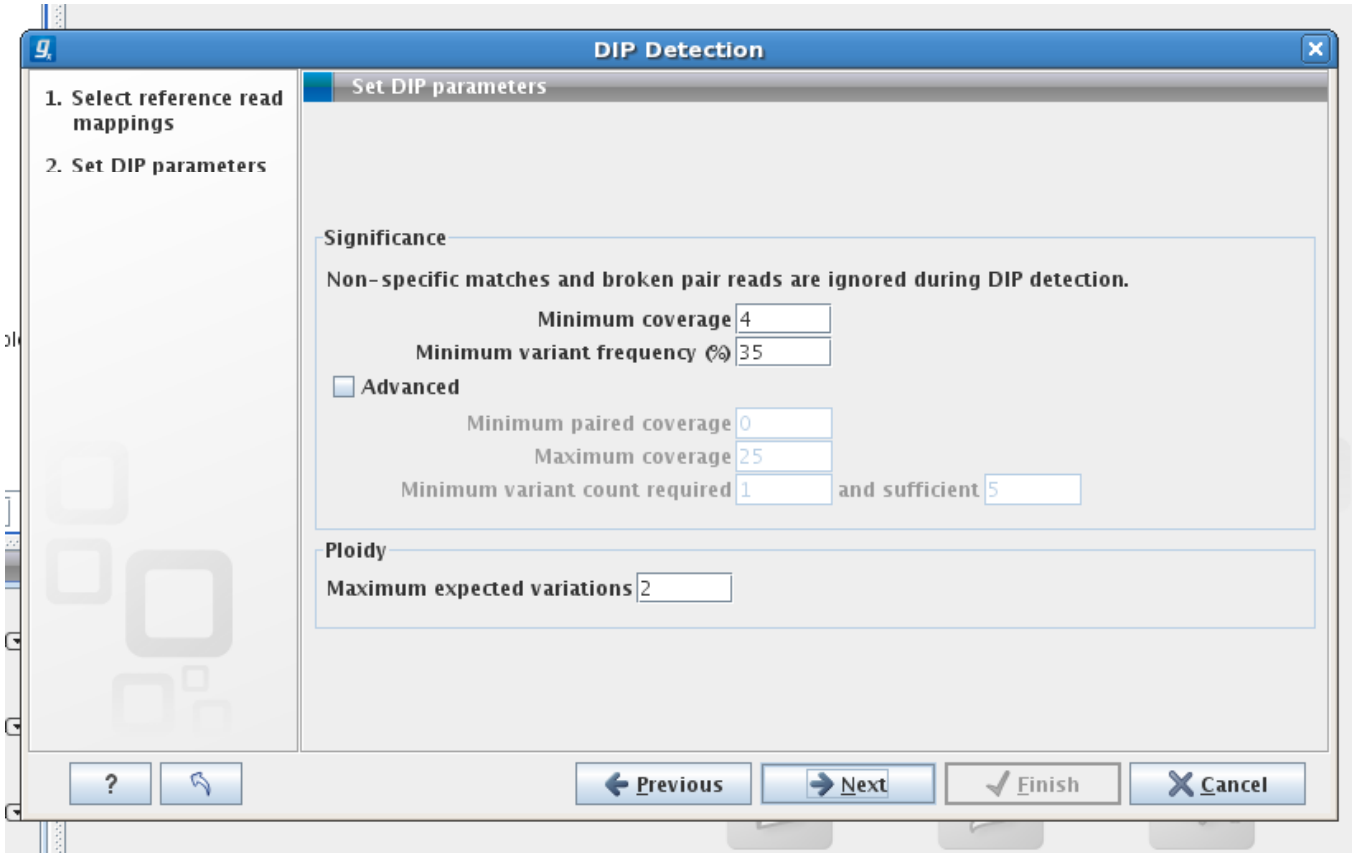


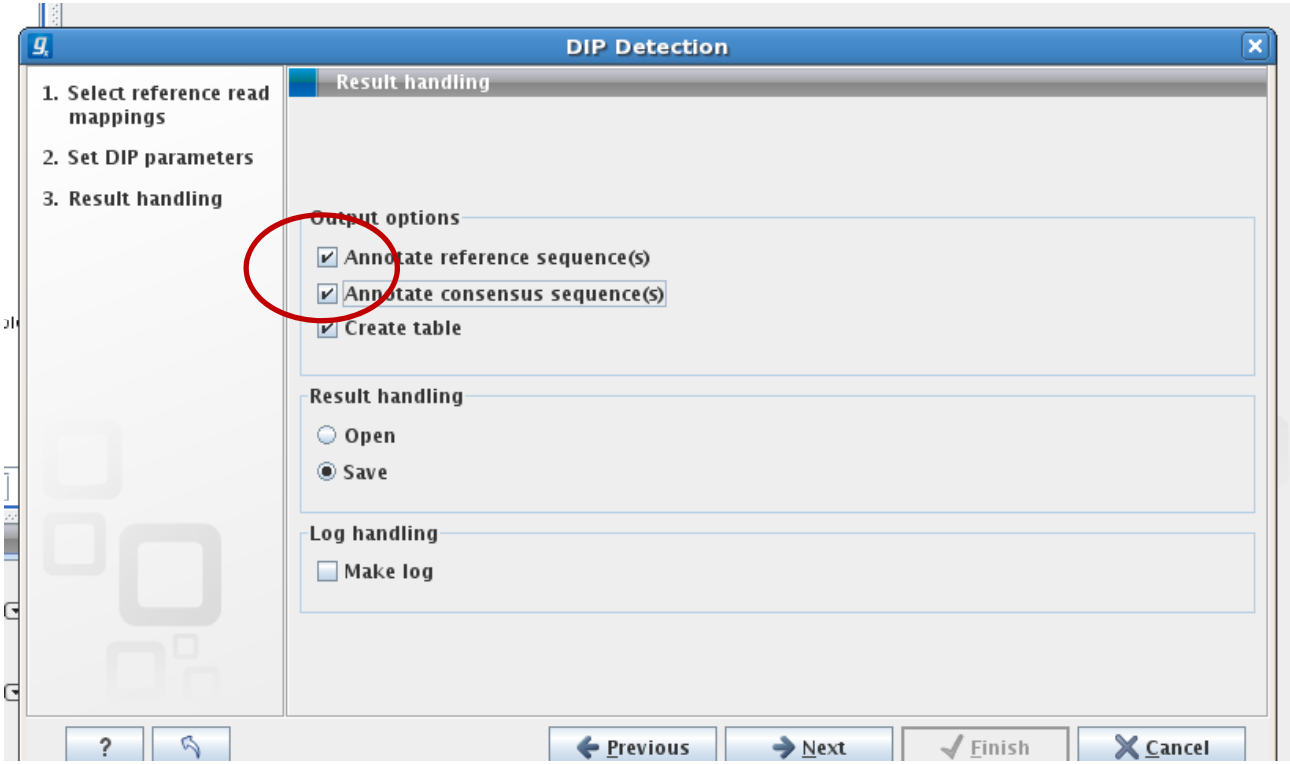

Print
Undo
Redo
Cut
Copy
Paste
Delete
Workspace
Plug-ins
Search
Fit Width
100%
Pan
Selection
Zoom In
Zoom Out
Ecoli.FLX (si...
Ecoli.FLX (si...
File
Edit
View
Toolbox
Show
Close Ctrl-W
Close Tab Area
Close All Views Ctrl+Shift-W
Close Other Tabs
Save Ctrl-S
Save As... Ctrl+Shift-S
Show/Hide Side Panel Ctrl-U
Split Horizontally Ctrl-T
Split Vertically Ctrl-J
Find in Navigation Area
Maximize/Restore View Ctrl-M
Fit Width
Zoom to 100%
NC_010473
Consensus
Coverage
EECRH8001BWEAV
EECRH8001BERJN
EECRH8002H65T2
EECRH8001EKKJB
EECRH8001AV6L5
Quality scores
Read Mapping Settings
Read layout
Gather sequences at top
Show sequence ends
Find Conflict
Low coverage threshold 8
Find Low Coverage
Sequence layout
No spacing
Numbers on sequences
Relative to 1
Numbers on plus strand
Lock top sequence
Hide labels
Lock labels
Sequence label
Name
Compactness
Not compact
Identical residues as dots
Annotation layout
Annotation types
Residue coloring
Alignment info
Nucleotide info
Find
Text format

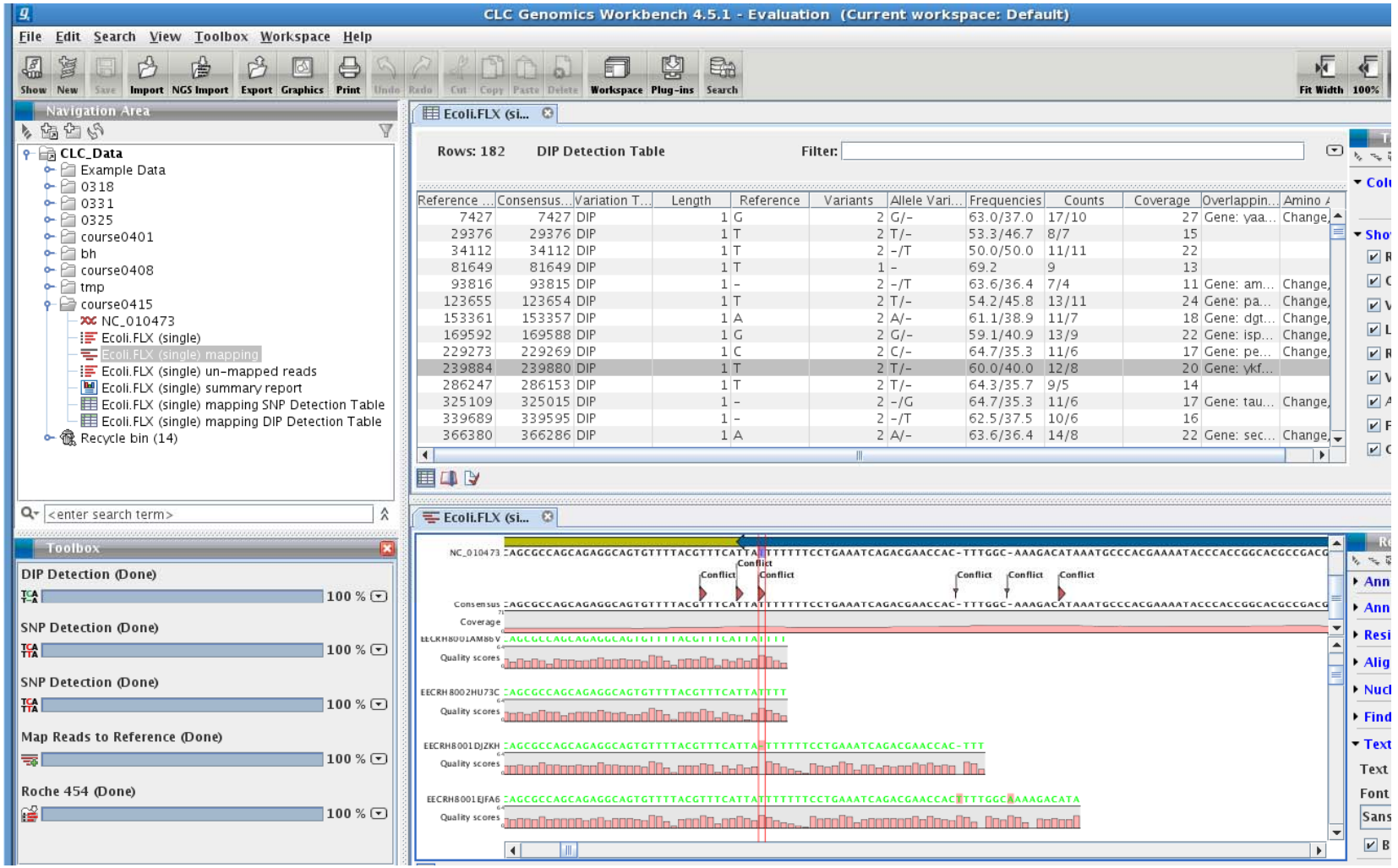

File Edit Search View Toolbox Workspace Help
Show New Save Import NGS Import Export Graphics Print Undo Redo Cut Copy Paste Delete Workspace Plug-ins Search
Fit Width 100% Pan Selection Zoom In Zoom Out
Navigation Area
CLC_Data
Example Data
0318
0331
0325
course0401
bh
course0408
tmp
course0415
NC_010473
Ecoli.FLX (single)
Ecoli.FLX (single) mapping
Ecoli.FLX (single) un-mapped reads
Ecoli.FLX (single) summary report
Ecoli.FLX (single) mapping SNP Detection Table
Recycle bin (14)
<enter search term>
Toolbox
SNP Detection (Done) 100 %
SNP Detection (Done) 100 %
Map Reads to Reference (Done) 100 %
Roche 454 (Done) 100 %
Processes Toolbox
Idle
Ecoli.FLX (si...
Rows: 233 SNP Detection Table Filter:
Reference ... Consensus ... Variation T... Reference Variants Allele Varia... Frequencies Counts Coverage Overlappin... Amino Acid.
1093277 1093032 SNP T 1 G 100.0 10 10 Gene: app...
1115086 1114841 SNP A 1 G 94.1 16 17 Gene: torD... Asn46Ser
1154738 1154438 SNP A 1 G 100.0 28 28 Gene: ycd... Tyr144Cys
1166395 1166095 SNP TTG 1 CCG 100.0 6 6 Conflict: C... Leu289Pro
1178224 1177924 SNP T 1 C 100.0 15 15 Gene: grx... Lys167Glu
1224504 1224204 SNP G 1 A 100.0 13 13 Gene: ycfS... Pro180Leu
1244648 1244348 SNP G 1 A 100.0 22 22 Gene: pho...
1250888 1250588 SNP T 2 C/T 62.5/37.5 25/15 40 Gene: icd, ...
1250900 1250600 SNP T 2 C/T 60.5/39.5 23/15 38 Gene: icd, ...
1250913 1250613 SNP CTG 2 TTA/CTG 56.4/43.6 22/17 39 CDS: icd, ...
1250945 1250645 SNP T 2 C/T 57.1/42.9 24/18 42 Gene: icd, ...
1250948 1250648 SNP C 2 G/C 57.1/42.9 24/18 42 Gene: icd, ...
1250951 1250651 SNP G 2 A/G 57.1/42.9 24/18 42 Gene: icd, ...
1250960 1250660 SNP T 2 C/T 57.5/42.5 23/17 40 Gene: icd, ...
Table Settings
Column width
Manual
Show column
Reference Position
Consensus Position
Variation Type
Length
Reference
Variants
Allele Variations
Frequencies
Counts
Ecoli.FLX (si...
Consensus TATTCTCTCCGCTGAGATGATGCTGCGCCACATGGGTTGGACCGAAGCGGCTGACTTAATT
Coverage
EECRH8001AWM7X TATTCTCTCCGCTGAGATGATGCTGCGCCACATGGGTTGG
Quality scores
EECRH8001BU7HD TATTCTCTCCGCTGAGATGATGCTGCGCCACATGGGTTGG
Quality scores
Read Mapping Settings
Read layout
Gather sequences at top
Show sequence ends
Find Conflict
Low coverage threshold 8
Find Low Coverage
Sequence layout
No spacing
Numbers on sequences
Relative to 1
Numbers on plus strand

DIP detection




DIP Detection
1. Select reference read mappings
2. Set DIP parameters
Set DIP parameters
Significance
Non-specific matches and broken pair reads are ignored during DIP detection.
Minimum coverage 4
Minimum variant frequency (%) 35
Advanced
Minimum paired coverage 0
Maximum coverage 25
Minimum variant count required 1 and sufficient 5
Ploidy
Maximum expected variations 2
?
Previous
Next
Finish
Cancel


DIP Detection
1. Select reference read mappings
2. Set DIP parameters
3. Result handling
Result handling
Output options
Annotate reference sequence(s)
Annotate consensus sequence(s)
Create table
Result handling
Open
Save
Log handling
Make log
Previous
Next
Finish
Cancel


CLC Genomics Workbench 4.5.1 - Evaluation (Current workspace: Default)
File Edit Search View Toolbox Workspace Help
Show New Save Import NGS Import Export Graphics Print Undo Redo Cut Copy Paste Delete Workspace Plug-ins Search Fit Width 100%
Navigation Area
CLC_Data
Example Data
0318
0331
0325
course0401
bh
course0408
tmp
course0415
NC_010473
Ecoli.FLX (single)
Ecoli.FLX (single) mapping
Ecoli.FLX (single) un-mapped reads
Ecoli.FLX (single) summary report
Ecoli.FLX (single) mapping SNP Detection Table
Ecoli.FLX (single) mapping DIP Detection Table
Recycle bin (14)
<enter search term>
Toolbox
DIP Detection (Done) 100 %
SNP Detection (Done) 100 %
SNP Detection (Done) 100 %
Map Reads to Reference (Done) 100 %
Roche 454 (Done) 100 %
Ecoli.FLX (si...
Rows: 182 DIP Detection Table Filter:
Reference ... Consensus... Variation T... Length Reference Variants Allele Vari... Frequencies Counts Coverage Overlappin... Amino A
7427 7427 DIP 1 G 2 G/- 63.0/37.0 17/10 27 Gene: yaa... Change
29376 29376 DIP 1 T 2 T/- 53.3/46.7 8/7 15
34112 34112 DIP 1 T 2 -/T 50.0/50.0 11/11 22
81649 81649 DIP 1 T 1 - 69.2 9 13
93816 93815 DIP 1 - 2 -/T 63.6/36.4 7/4 11 Gene: am... Change
123655 123654 DIP 1 T 2 T/- 54.2/45.8 13/11 24 Gene: pa... Change
153361 153357 DIP 1 A 2 A/- 61.1/38.9 11/7 18 Gene: dgt... Change
169592 169588 DIP 1 G 2 G/- 59.1/40.9 13/9 22 Gene: isp... Change
229273 229269 DIP 1 C 2 C/- 64.7/35.3 11/6 17 Gene: pe... Change
239884 239880 DIP 1 T 2 T/- 60.0/40.0 12/8 20 Gene: ykf...
286247 286153 DIP 1 T 2 T/- 64.3/35.7 9/5 14
325109 325015 DIP 1 - 2 -/G 64.7/35.3 11/6 17 Gene: tau... Change
339689 339595 DIP 1 - 2 -/T 62.5/37.5 10/6 16
366380 366286 DIP 1 A 2 A/- 63.6/36.4 14/8 22 Gene: sec... Change
Ecoli.FLX (si...
NC_010473
Conflict
Consensus
Coverage
Quality scores
Ann
Ann
Resi
Alig
Nucl
Find
Text
Text
Font
Sans

exercise

- solexa_ecoli_1.fq
- solexa_ecoli_2.fq