De Novo Assembly of High-throughput Short Read Sequences

Chuming Chen

Center for Bioinformatics and Computational Biology (CBCB)
University of Delaware

NECC Third Skate Genome Annotation Workshop
May 23, 2011

Outline

- Genome assembly primer
- High-throughput short read sequencing (NGS) assembly pipeline
- Case study

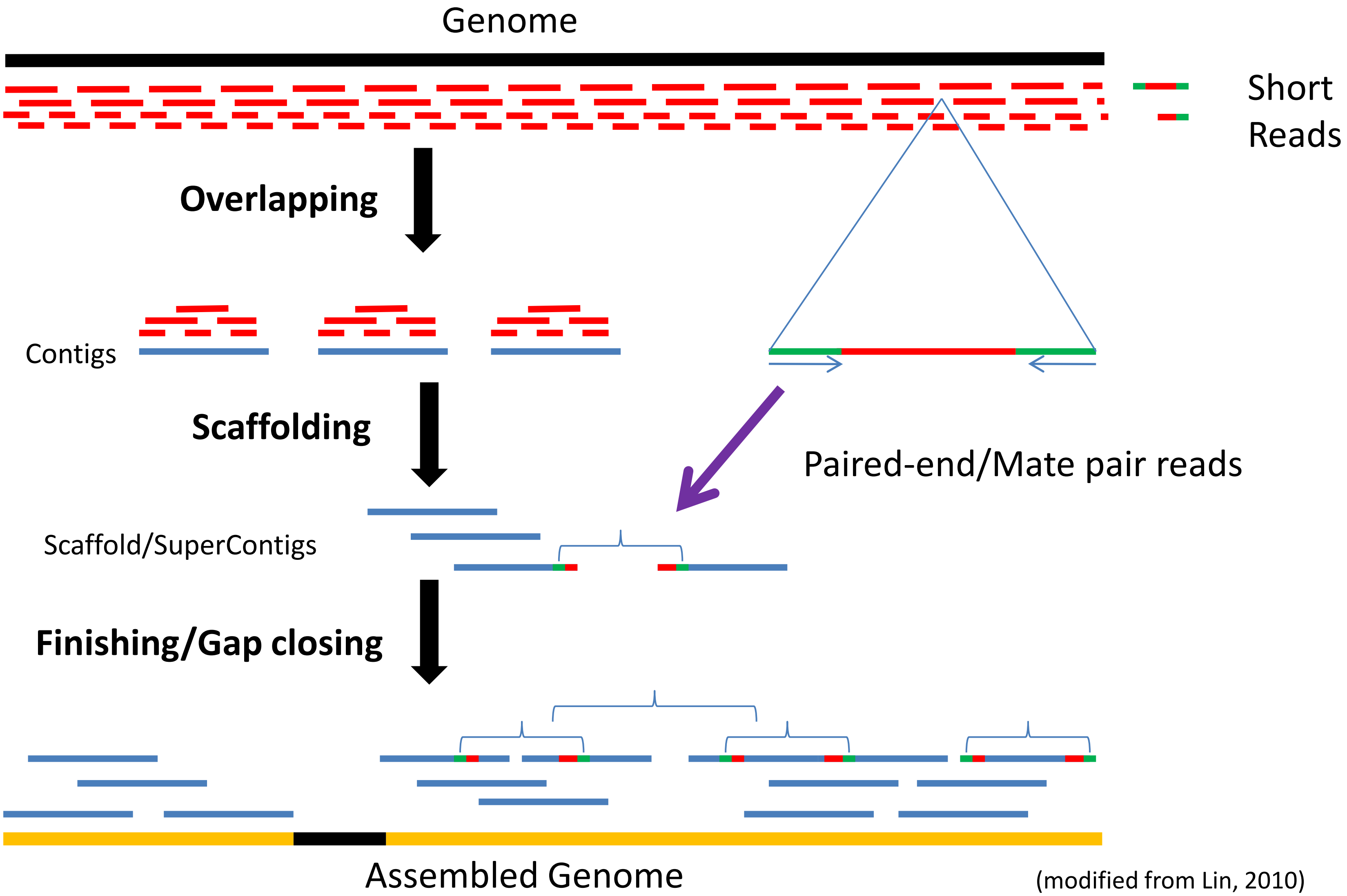
Genome sequencing and assembly

- Genome
  - Long stretches of contiguous DNA sequences (base pairs)
  - Different genome sizes (i. e. virus: 3.5k, Human: 3.3 billion)
- Genome sequencers (NHGRI, Feb. 4, 2011)
  - Sanger-based sequencing (500-600 bases)
  - 454 sequencing (300-400 bases)
  - Illumina and SOLiD sequencing (50-100 bases)
- Sequencing and assembly
  - A genome must be fragmented, sequenced piece by piece and then re-assembled to obtain the full contiguous sequence

Assembly approaches

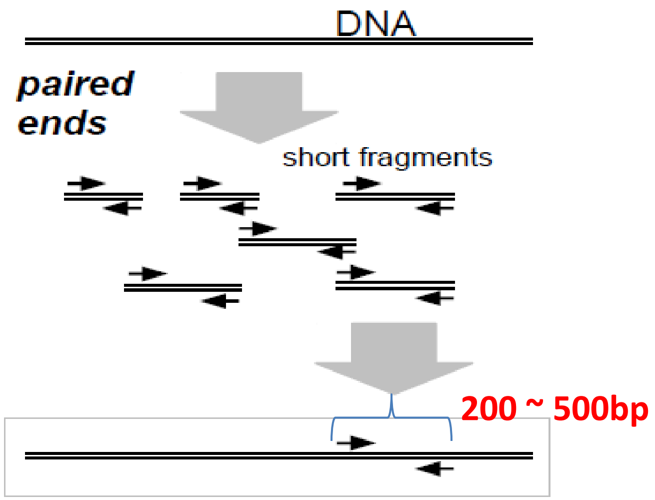
- Hierarchical assembly
  - Mapping the genome to a set of large insert clones
  - Reduce the assembly of the sequencing reads from the entire genome to a single clone, typically 40 - 200 Kb
  - The genome sequence is then assembled by aligning sequences of adjacent clones
- Whole genome shot-gun assembly
  - The entire genome is fragmented
  - The shotgun process takes reads from random positions along the chromosomes that make up one genome
  - The assembler then reconstructs the reads up to the chromosome length
  - Assembly is possible because the target is over-sampled by the shotgun reads, such that reads overlap

Whole genome shot-gun sequencing and assembly

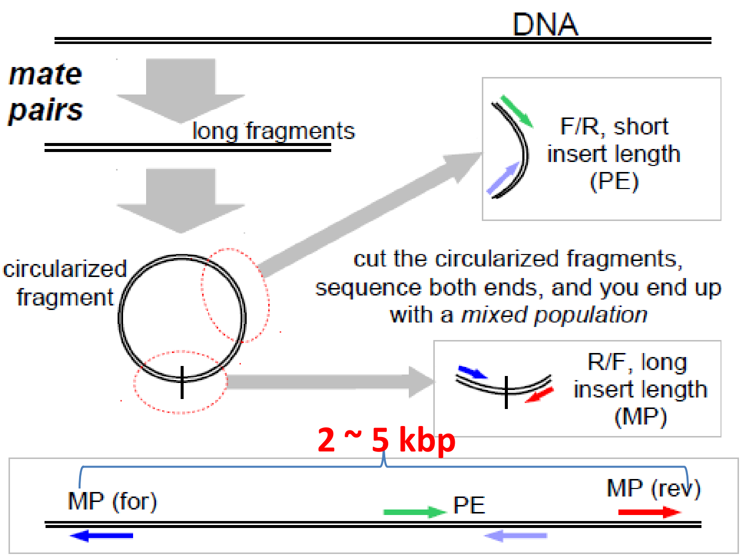


(modified from Lin, 2010)

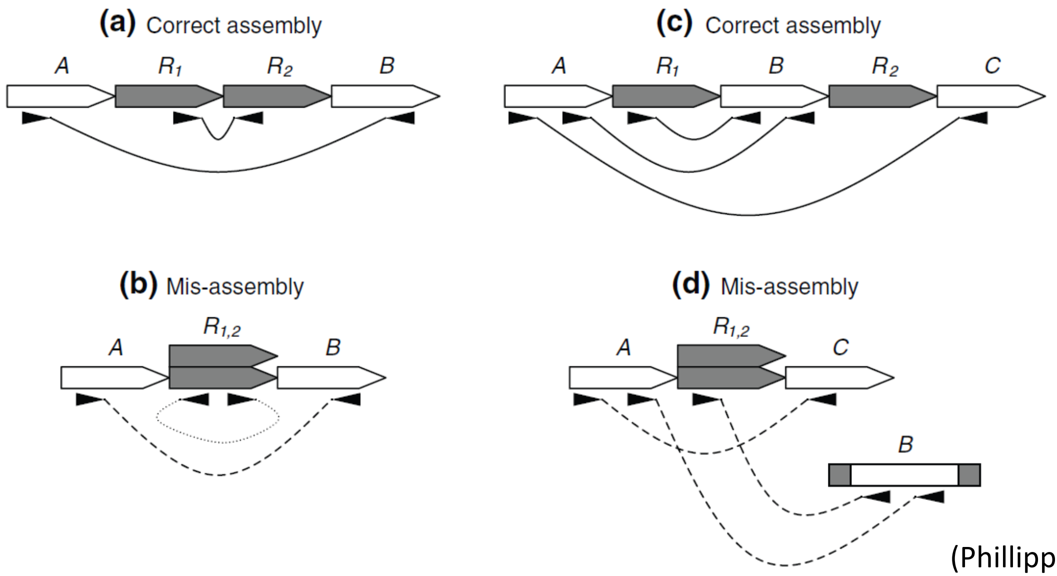
Paired-end vs. mate-pair reads



(Fass, 2010)

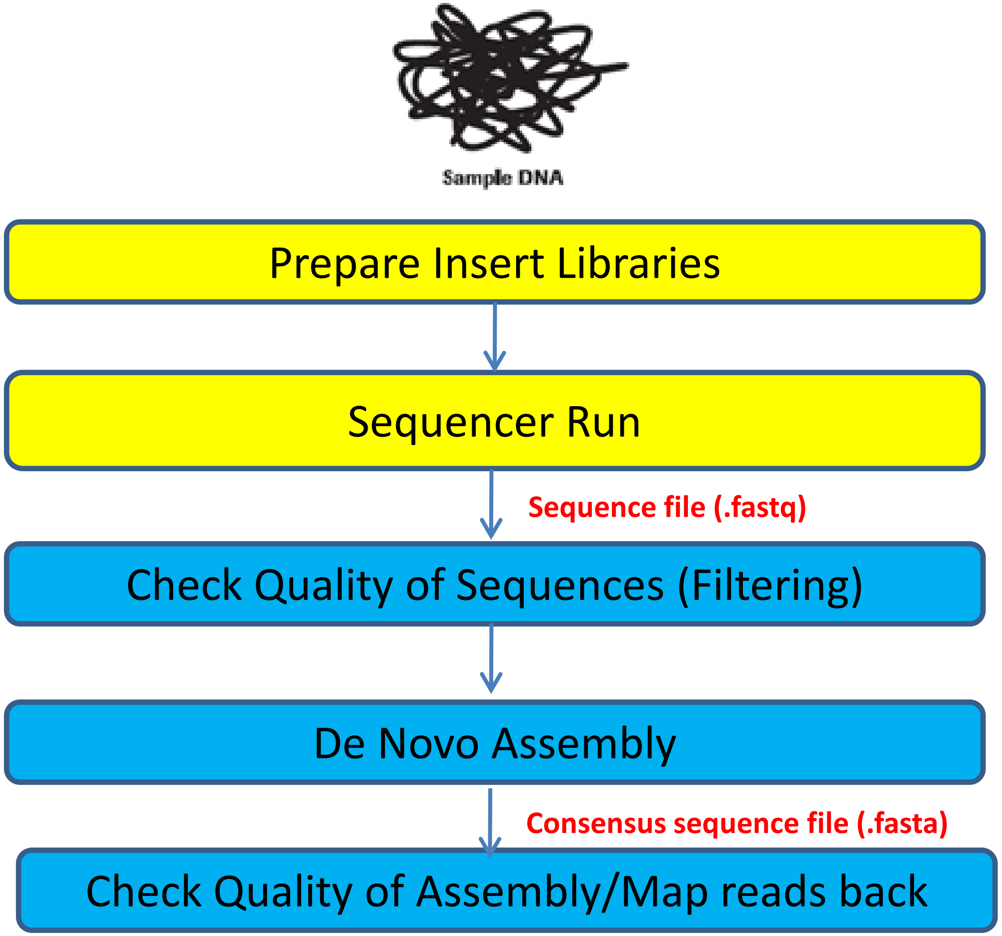

DNA
mate pairs
long fragments
F/R, short insert length (PE)
circularized fragment
cut the circularized fragments, sequence both ends, and you end up with a *mixed population*
R/F, long insert length (MP)
2 ~ 5 kbp
MP (for)
PE
MP (rev)

(Fass, 2010)



(Phillippy, 2008)

Assembly pipeline



Prepare Insert Libraries

Sequencer Run

Sequence file (.fastq)

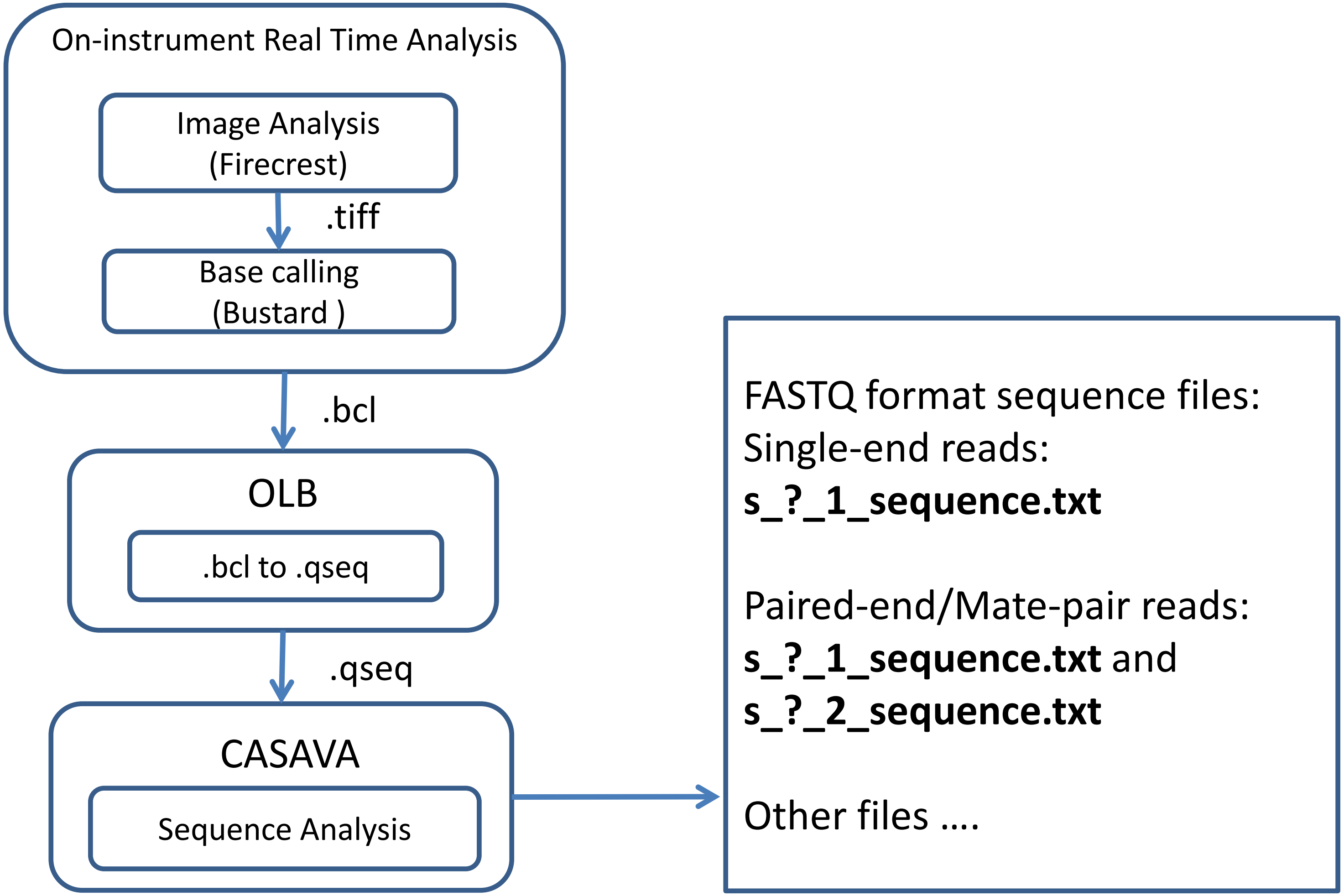
Check Quality of Sequences (Filtering)

De Novo Assembly

Consensus sequence file (.fasta)

Check Quality of Assembly/Map reads back

HiSeq2000 sequence data processing pipeline



On-instrument Real Time Analysis

Image Analysis (Firecrest)

.tiff

Base calling (Bustard)

.bcl

OLB

.bcl to .qseq

.qseq

CASAVA

Sequence Analysis

FASTQ format sequence files:
Single-end reads:
s_?_1_sequence.txt

Paired-end/Mate-pair reads:
s_?_1_sequence.txt and
s_?_2_sequence.txt

Other files ….

FASTQ format sequence files

@Sequence ID

Sequence (A, C, G, T, N)

+Sequence ID

ASCII encoded quality score

s_?_1_sequence.txt

```
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAATACCATTTAGTTCAAGGGAAAGCCAGCAAAGTTCGATCAAGGATAGACCGAGG
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfffffgggfafffcdfcfffbfdddeaegfgfgafaffW^a]a`b^d[dbdbbcccccb[`dW\a`a`^caa[ffcffafaccfcd_fcfc_cb
@HWI-ST741_0085:1:1101:1417:1939#0/1
GCTCAGTACTGGGACCCCAGCTATTTACAAAATATATTAATGATCTGGATGAGGGATTTGAAGGCAATATCTCCAAGTTTGCGGATGACATTAAGCTGGGG
+HWI-ST741_0085:1:1101:1417:1939#0/1
hhhhghhhhhhhhhhhhhhhhhhhhhehhhhhhhghhhhhhhhhhhhhghhhghhghhhhahfghhhfhhhhhggaffffhhhdghg[hghhffhhhageh
...
```

s_?_2_sequence.txt

```
@HWI-ST741_0085:1:1101:1444:1939#0/2
CCCAGCTTATCCTTGCAACTCTTCTTAAATAGAGGCACAACATTAATCACCCTCCCTTCTATGGACTCCTTTTATACCTCACCCTTCCTCGGCAGGGCCAG
+HWI-ST741_0085:1:1101:1444:1939#0/2
edeaadffffcaffcdaeaeffdfdecfefaceccfdffdfddfffffdffdfffdcaffffcffgaggggggg`ggggggfggggggggd]eedaa_^BB
@HWI-ST741_0085:1:1101:1417:1939#0/2
ACCATTCGCTTTCTTTACTGCCTGCTGCACCTGCATGCCTACCTTCAATGACTGGTGTACCATGACACCCAGGTCTCGCTGCATCTCCCCCTTTCCCAATC
+HWI-ST741_0085:1:1101:1417:1939#0/2
hhhhghhhhhhhhhhhhhhhhhghhhdhhghhhhhhhhhchhhhhhhhhhfghhhhfhhdhfhhdhhahhhfefhhfdfffedhhhehhhhhehhcfffffd
...
```

A closer look at sequence file

```
       1        2   3    4    5  6 7
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAAT
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfffffgggfafffcdfcfffbfdddeaegfgfgafaffW^a]
```

```
       1        2   3    4    5  6 7
@HWI-ST741_0085:1:1101:1444:1939#0/2
CCCAGCTTATCCTTGCAACTCTTCTTAAATAGAGGCACAACATTAATCA
+HWI-ST741_0085:1:1101:1444:1939#0/2
Edeaadffffcaffcdaeaeffdfdecfefaceccfdffdfddfffffd
```

1. the unique instrument name
2. flowcell lane
3. tile number within the flowcell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. index number for a multiplexed sample (0 for no multiplexing)
7. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

(Wikipedia.org)

Quality scores

- Phred quality scores *Q* are defined as a property which is logarithmically related to the base-calling error probabilities *P*
- A Phred score of a base is: $Q_{phred} = -10 \log_{10} P$, where *P* is the estimated probability of a base being wrong

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

- Phred scoring scheme, encoded as an ASCII character by adding 64 to the Phred value

```
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAAT
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfffffgggfafffcdfcfffbfdddeaegfgfgafaffW^a]
```

HiSeq score

```
39 33 38 33 38 38 23 30 33 29
```

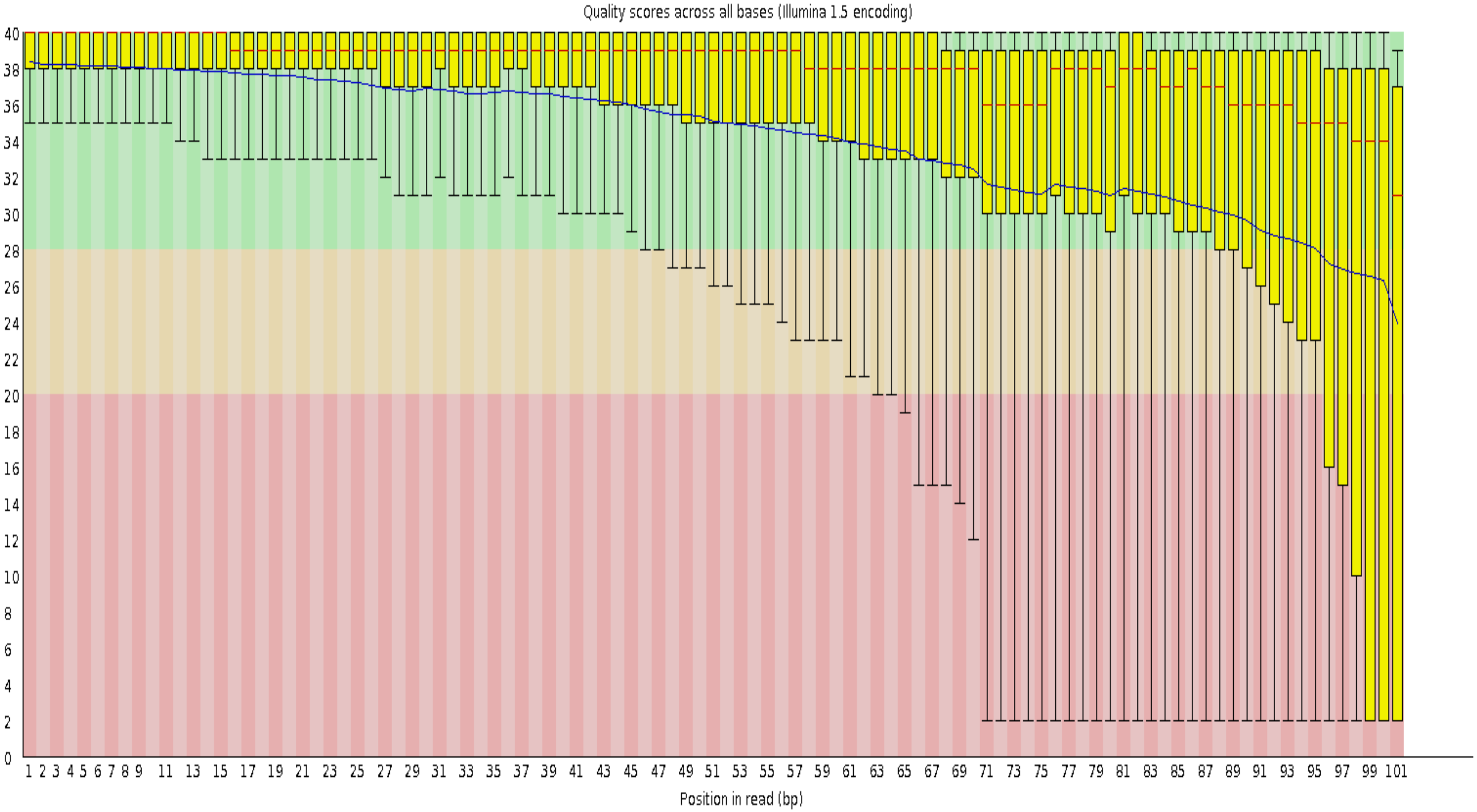
Phred score

(Wikipedia.org)

Sequence reads quality assessment

- **FastQC** (Baraham Bioinformatics, UK)
  - Basic Statistics
  - Per Base Sequence Quality
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Base GC Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Duplicate Sequences
  - Overrepresented Sequences
  - Overrepresented Kmers

Per base sequence quality



Yellow box: 25-75 quartile; Black whisker: min-max ; Red line: median; Blue line: mean

Trim Sequences

- Quality trimming
  - Based on quality scores
- Ambiguity trimming
  - Remove stretches of Ns
- Adapter sequence trimming
  - Remove sequence adapters
- Base trim
  - Remove a specified number of bases at either 3' or 5' end of the reads
- Length trimming
  - Remove reads shorter or longer than a specified threshold

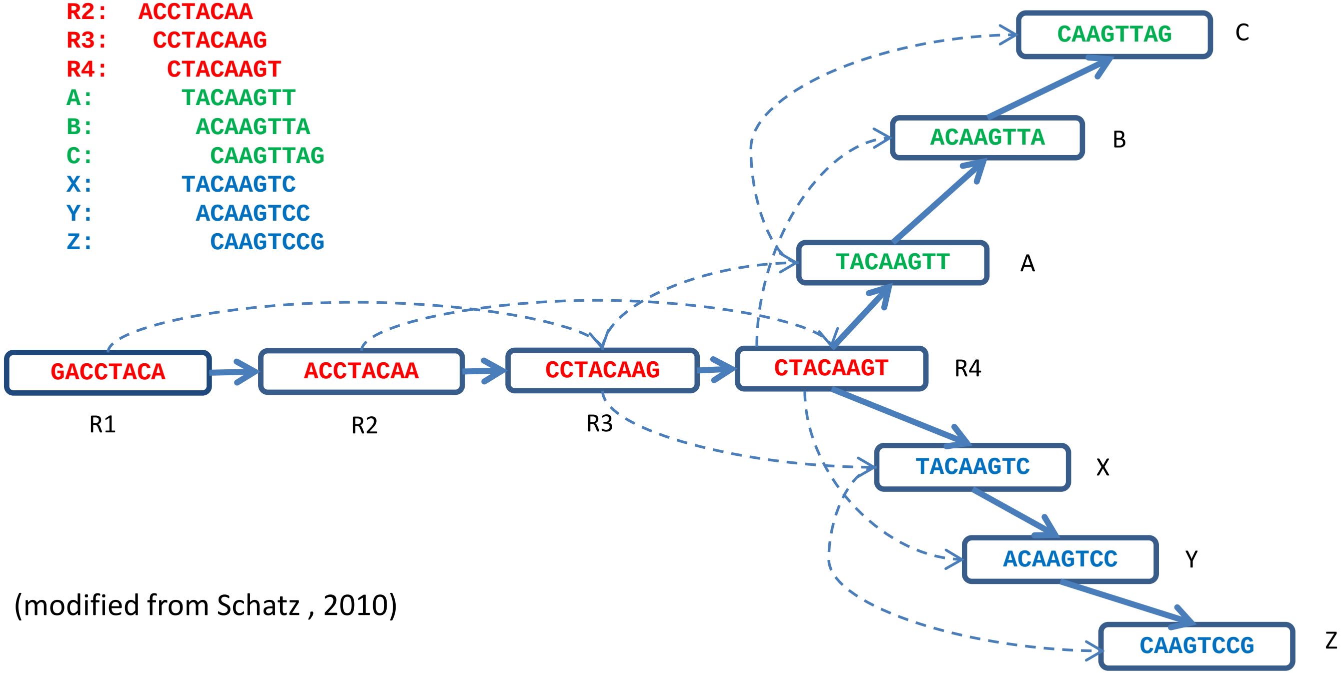
De Novo Assembly algorithms

- Overlap/Layout/Consensus Graph
- de Brujin Graph

Overlap/Layout/Consensus graphs

- A node corresponds to a read, an edge denotes an overlap between two reads.
- The overlap graph is used to compute a layout of reads and consensus sequence of contigs by pair-wise sequence alignment.
- Good for sequences with limited number of reads but significant overlap. Computational intensive for short reads (short and high error rate).
- Example assemblers: Celera Assembler, Arachne, CAP and PCAP

```
R1: GACCTACA
R2:  ACCTACAA
R3:   CCTACAAG
R4:    CTACAAGT
A:      TACAAGTT
B:       ACAAGTTA
C:        CAAGTTAG
X:      TACAAGTC
Y:       ACAAGTCC
Z:        CAAGTCCG
```

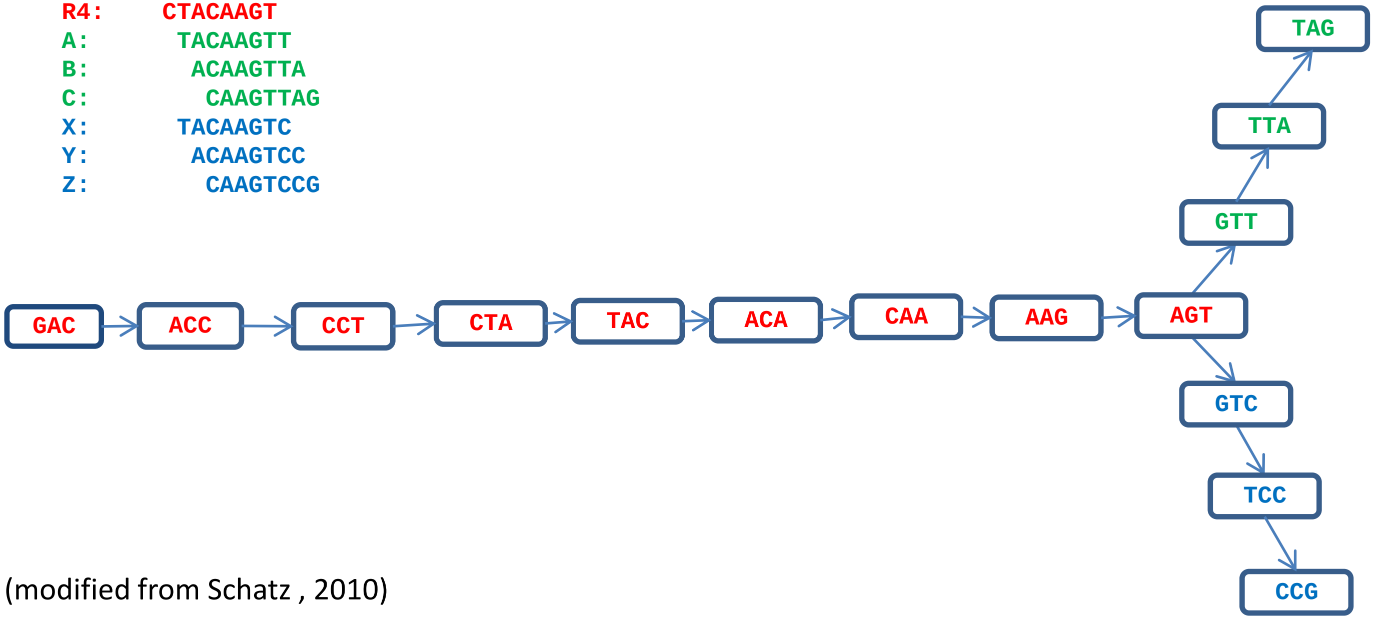


(modified from Schatz , 2010)

de Brujin graphs

- No need for all against all overlap discovery.
- Break reads into smaller sequences of DNA (K-mers, K denotes the length in bases of these sequences).
- Captures overlaps of length K-1 between these K-mers.
- More sensitive to repeats and sequencing errors.
- By construction, the graph contains a path corresponding to the original sequence.
- Example assemblers: Euler, Velvet, ABySS, AllPaths, SOAPdenovo, CLC Bio

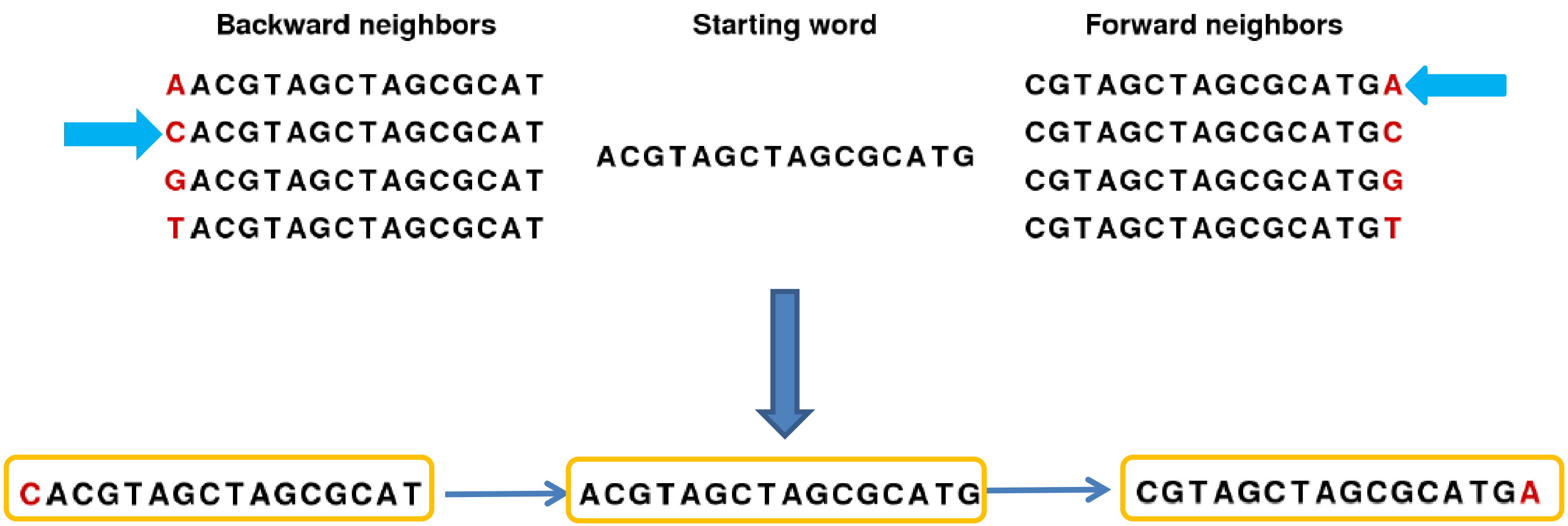
```
R1: GACCTACA
R2:  ACCTACAA
R3:   CCTACAAG
R4:    CTACAAGT
A:       TACAAGTT
B:        ACAAGTTA
C:         CAAGTTAG
X:       TACAAGTC
Y:        ACAAGTCC
Z:         CAAGTCCG
```



(modified from Schatz , 2010)

CLC Bio De Novo assembly

- Make a table of the words (K-mers) seen in the reads.
- Build de Bruijn graph from the word table.
- Use the reads to resolve the repeats.
- Use the information from paired reads to resolve larger repeats.
- Output resulting contigs based on the paths.



Backward neighbors	Starting word	Forward neighbors
AACGTAGCTAGCGCAT	ACGTAGCTAGCGCATG	CGTAGCTAGCGCATGA
CACGTAGCTAGCGCAT		CGTAGCTAGCGCATGC
GACGTAGCTAGCGCAT		CGTAGCTAGCGCATGG
TACGTAGCTAGCGCAT		CGTAGCTAGCGCATGT

CACGTAGCTAGCGCAT → ACGTAGCTAGCGCATG → CGTAGCTAGCGCATGA

(modified from CLC Bio, 2011)

Word (K-mer) size

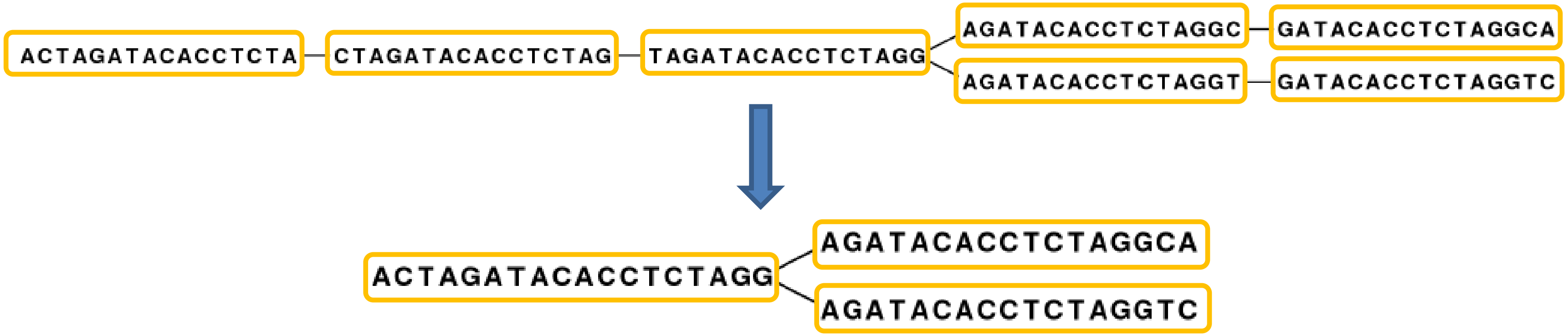
To strike a balance, CLC bio’s de novo assembler chooses a word length based on the amount of input data: the more data, the longer the word length. It is based on the following:

word size 12: 0 bp - 30000 bp
word size 13: 30001 bp - 90002 bp
word size 14: 90003 bp - 270008 bp
word size 15: 270009 bp - 810026 bp
word size 16: 810027 bp - 2430080 bp
word size 17: 2430081 bp - 7290242 bp
word size 18: 7290243 bp - 21870728 bp
word size 19: 21870729 bp - 65612186 bp
word size 20: 65612187 bp - 196836560 bp
word size 21: 196836561 bp - 590509682 bp
word size 22: 590509683 bp - 1771529048 bp
word size 23: 1771529049 bp - 5314587146 bp
word size 24: 5314587147 bp - 15943761440 bp
word size 25: 15943761441 bp - 47831284322 bp
word size 26: 47831284323 bp - 143493852968 bp
word size 27: 143493852969 bp - 430481558906 bp
word size 28: 430481558907 bp - 1291444676720 bp
word size 29: 1291444676721 bp - 3874334030162 bp
word size 30: 3874334030163 bp - 11623002090488 bp
word size 31: 11623002090489 bp and up

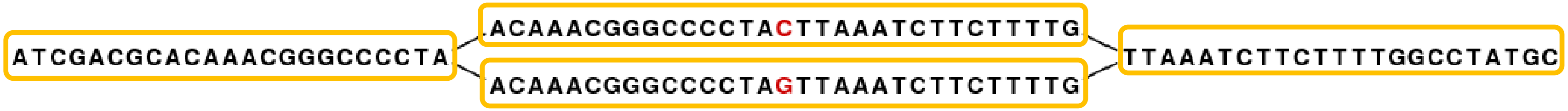
(CLC Bio, 2011)

Repeats or sequencing errors

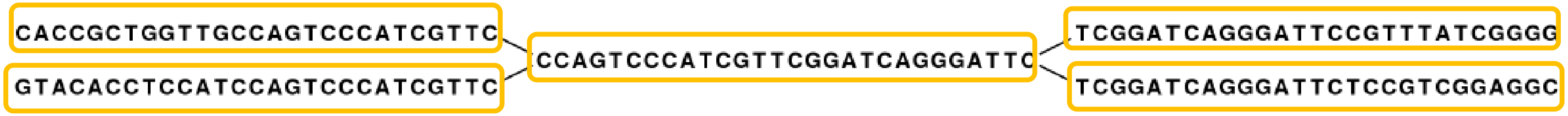
Graph Reduction



SNP or Sequencing Error


ATCGACGCACAAACGGGCCCCTA
ACAAACGGGCCCCTACTTAAATCTTCTTTTG
ACAAACGGGCCCCTAGTTAAATCTTCTTTTG
TTAAATCTTCTTTTGGCCTATGC

Repeat Sequence


CACCGCTGGTTGCCAGTCCCATCGTTC
GTACACCTCCATCCAGTCCCATCGTTC
CCAGTCCCATCGTTCGGATCAGGGATTC
TCGGATCAGGGATTCCGTTTATCGGGG
TCGGATCAGGGATTCTCCGTCGGAGGC

(modified from CLC Bio, 2011)

Assembly quality assessment

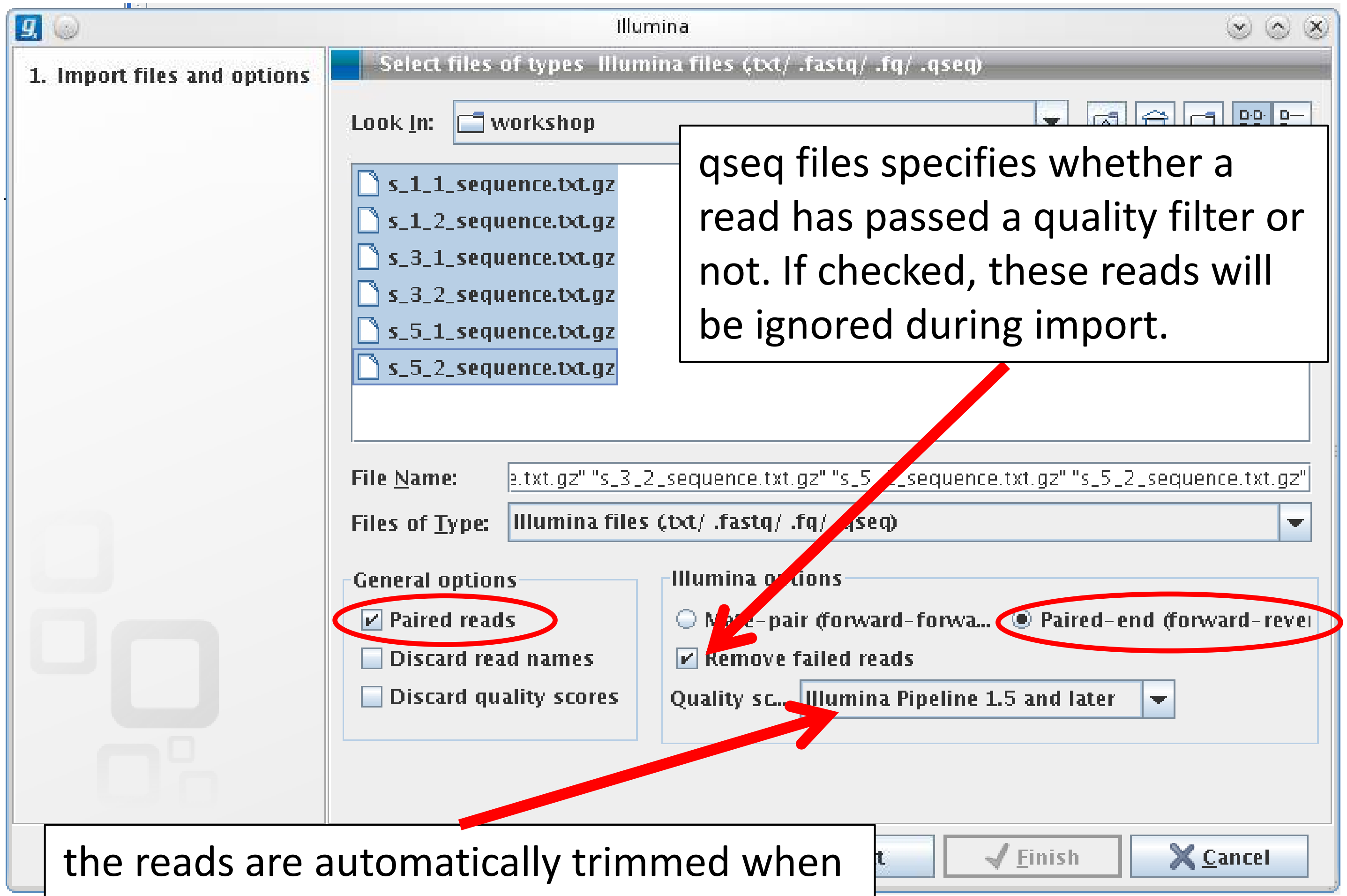
- Continuity
  - Lengths distribution of contigs/scaffolds.
  - Average length, minimum and maximum lengths, combined total lengths.
  - **N50** captures how much of the assembly is covered by relatively large contigs.
  - The N50 is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly.
  - Compute N50
    - first ordering all contigs (or scaffolds) by length,
    - then summing up their lengths (starting with the longest) until the sum exceeds 50% of the total length of all contigs.
    - the corresponding contig length is N50.

- Accuracy or "Correctness"
  - **Base accuracy** – the frequency of calling the correct nucleotide at a given position in the assembly.
  - **Mis-assembly rate** – the frequency of rearrangements, significant insertions, deletions and inversions.

Case study

- Show the basic steps involved in De Novo assembly of high-throughput short read sequences
- Data
  - 3 lanes of Illumina HiSeq 2000 short read sequences for the little skate (a couple weeks ago)
- Assembler
  - CLC Bio Genomics Workbench 4.6

Import sequence data



Select files of types Illumina files (.txt/ .fastq/ .fq/ .qseq)

Look In: workshop

- s_1_1_sequence.txt.gz
- s_1_2_sequence.txt.gz
- s_3_1_sequence.txt.gz
- s_3_2_sequence.txt.gz
- s_5_1_sequence.txt.gz
- s_5_2_sequence.txt.gz

File Name: e.txt.gz" "s_3_2_sequence.txt.gz" "s_5_1_sequence.txt.gz" "s_5_2_sequence.txt.gz"

Files of Type: Illumina files (.txt/ .fastq/ .fq/ .qseq)

General options

- ☑ Paired reads
- □ Discard read names
- □ Discard quality scores

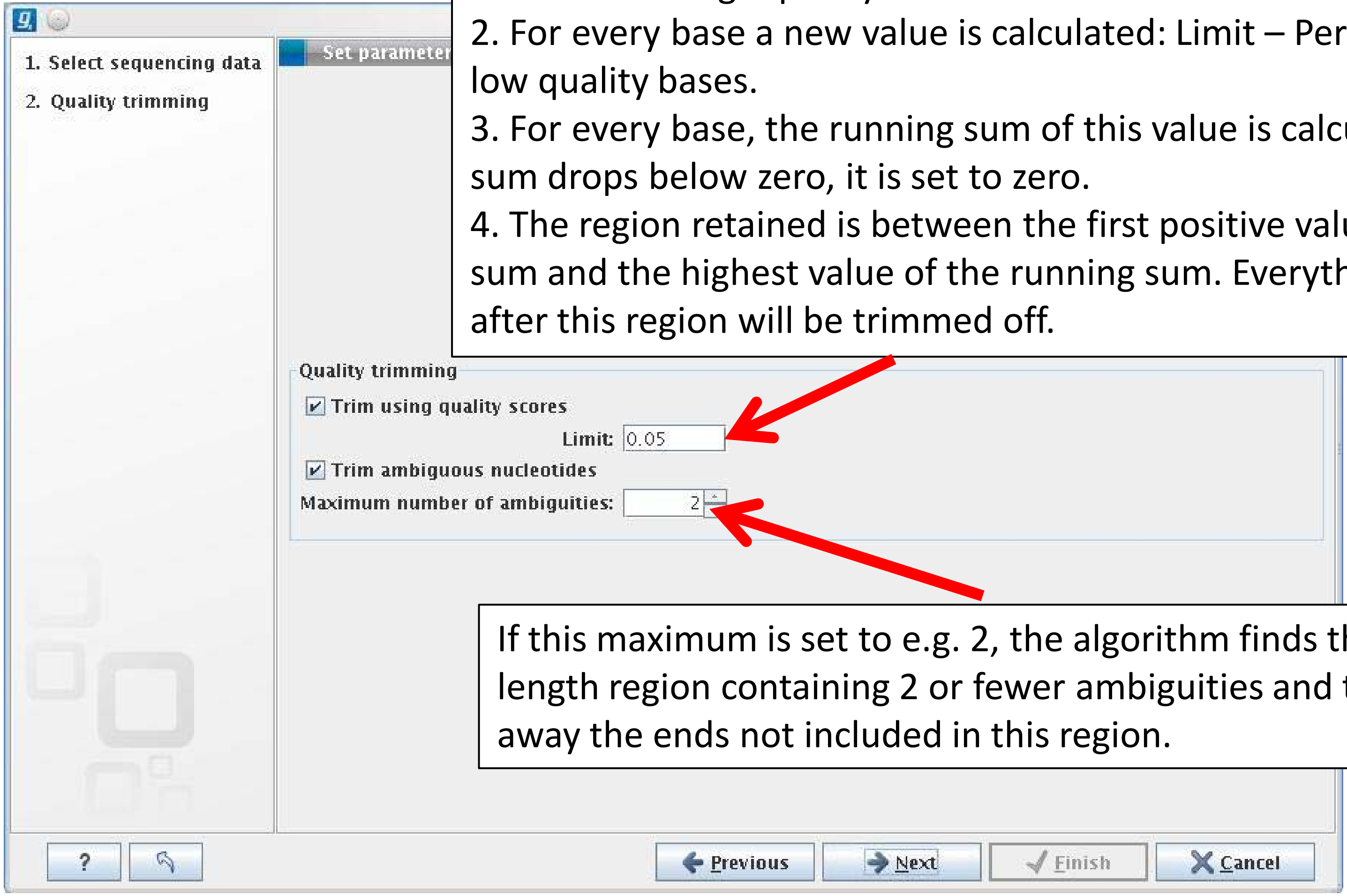
Illumina options

- ○ Mate-pair (forward-forwa... ● Paired-end (forward-rever
- ☑ Remove failed reads
- Quality sc... Illumina Pipeline 1.5 and later

qseq files specifies whether a read has passed a quality filter or not. If checked, these reads will be ignored during import.

the reads are automatically trimmed when a quality score B is encountered in the input file

Quality trimming



1. Select sequencing data

2. Quality trimming

Set parameter

Quality trimming

☑ Trim using quality scores

Limit: 0.05

☑ Trim ambiguous nucleotides

Maximum number of ambiguities: 2

1. Convert quality score (Q) to error probability: Perror = $10^{Q/-10}$, low values are high quality bases.

2. For every base a new value is calculated: Limit – Perror, negative for low quality bases.

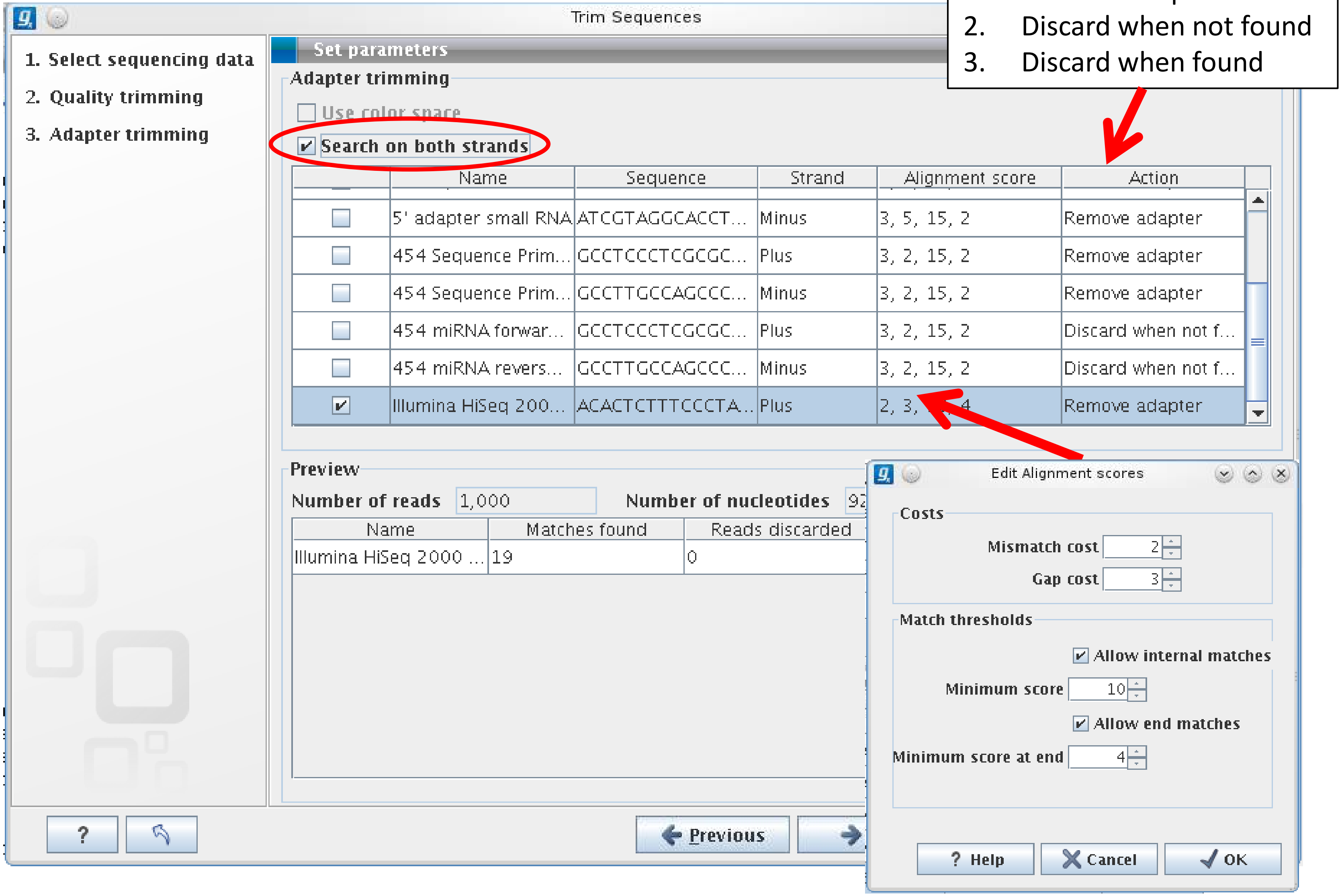
3. For every base, the running sum of this value is calculated. If the sum drops below zero, it is set to zero.

4. The region retained is between the first positive value of the running sum and the highest value of the running sum. Everything before and after this region will be trimmed off.

If this maximum is set to e.g. 2, the algorithm finds the maximum length region containing 2 or fewer ambiguities and then trims away the ends not included in this region.

Adapter trimming

1. Remove adapter
2. Discard when not found
3. Discard when found



1. Select sequencing data
2. Quality trimming
3. Adapter trimming

Set parameters

Adapter trimming

□ Use color space

☑ Search on both strands

	Name	Sequence	Strand	Alignment score	Action
□	5' adapter small RNA	ATCGTAGGCACCT...	Minus	3, 5, 15, 2	Remove adapter
□	454 Sequence Prim...	GCCTCCCTCGCGC...	Plus	3, 2, 15, 2	Remove adapter
□	454 Sequence Prim...	GCCTTGCCAGCCC...	Minus	3, 2, 15, 2	Remove adapter
□	454 miRNA forwar...	GCCTCCCTCGCGC...	Plus	3, 2, 15, 2	Discard when not f...
□	454 miRNA revers...	GCCTTGCCAGCCC...	Minus	3, 2, 15, 2	Discard when not f...
☑	Illumina HiSeq 200...	ACACTCTTTCCCTA...	Plus	2, 3, [illegible], 4	Remove adapter

Preview

Number of reads: 1,000

Number of nucleotides: 92

Name	Matches found	Reads discarded
Illumina HiSeq 2000 ...	19	0

Costs

Mismatch cost: 2

Gap cost: 3

Match thresholds

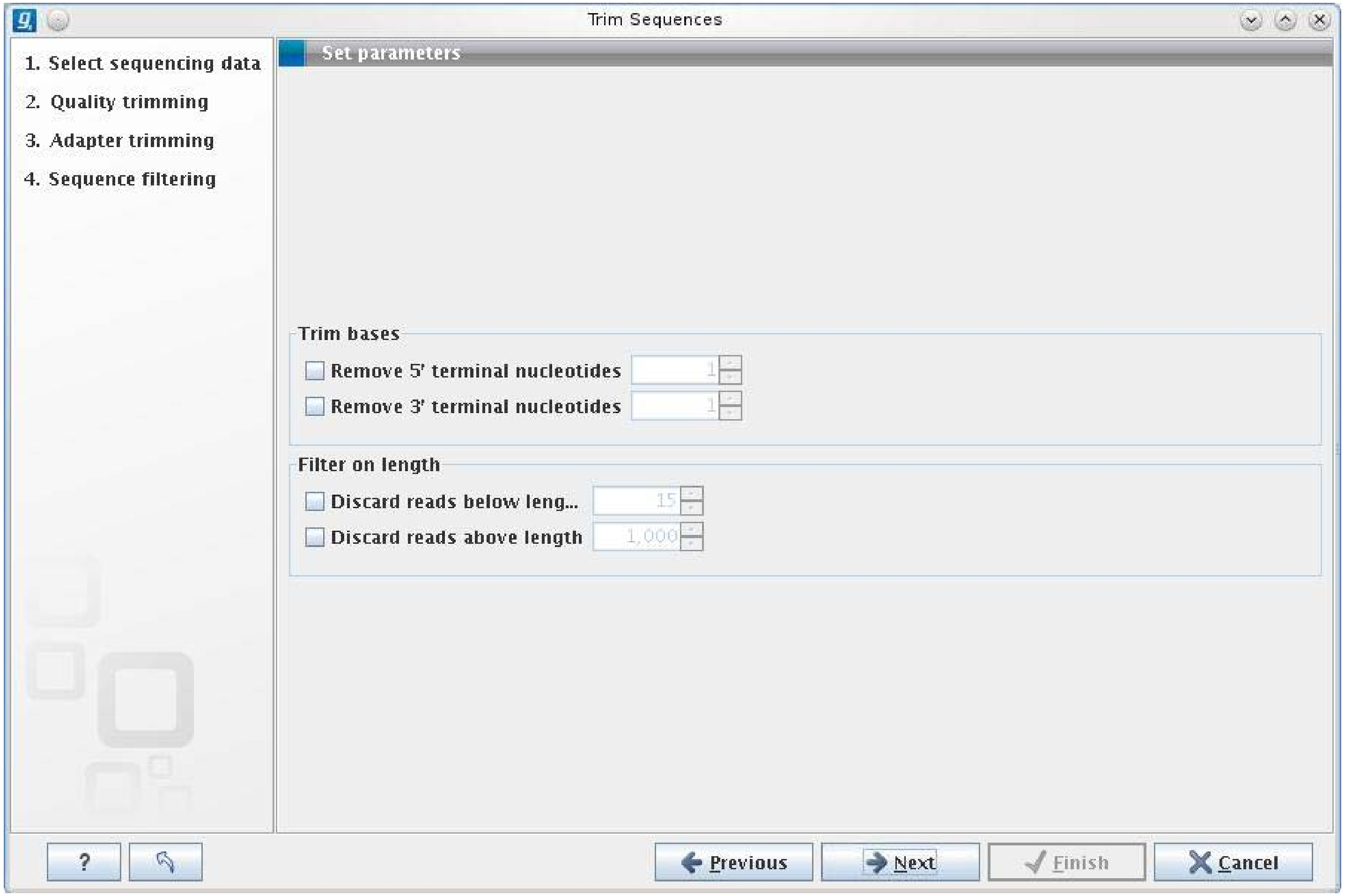
☑ Allow internal matches

Minimum score: 10

☑ Allow end matches

Minimum score at end: 4

Sequence filtering



Trim Sequences

1. Select sequencing data
2. Quality trimming
3. Adapter trimming
4. Sequence filtering

Set parameters

Trim bases

□ Remove 5' terminal nucleotides 1

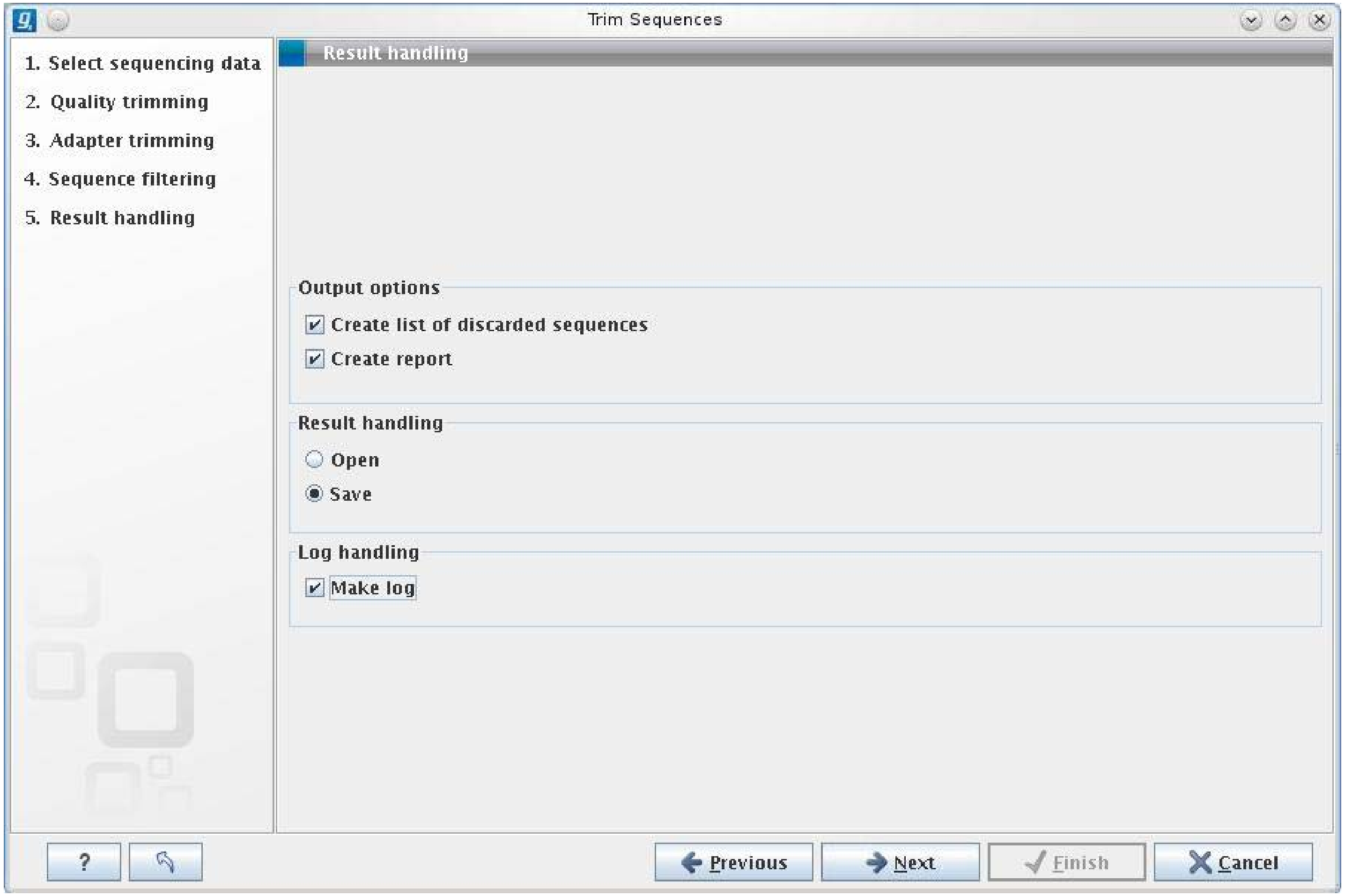
□ Remove 3' terminal nucleotides 1

Filter on length

□ Discard reads below leng... 15

□ Discard reads above length 1,000

Trimming result handling



Output options

☑ Create list of discarded sequences

☑ Create report

Result handling

○ Open

● Save

Log handling

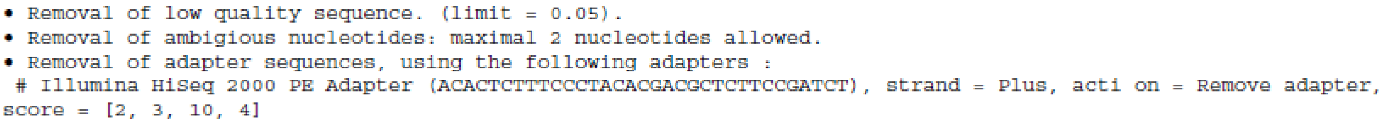
☑ Make log

Trimming report

Summary

Name	Number of reads	Avg. length	Number of reads after trim	Percentage retained	Avg. length after trim
S_1_1_sequence (paired)	190,790,624	93.4	190,784,343	~100%	93.1
S_1_1_sequence	1,655,955	63.9	1,654,977	99.94%	63.0
S_3_1_sequence (paired)	209,140,424	92.9	209,131,515	~100%	92.6
S_3_1_sequence	2,016,294	62.4	2,015,199	99.95%	61.8
S_5_1_sequence (paired)	223,212,034	92.4	223,201,524	~100%	92.1
S_5_1_sequence	2,373,091	62.2	2,371,693	99.98%	61.5

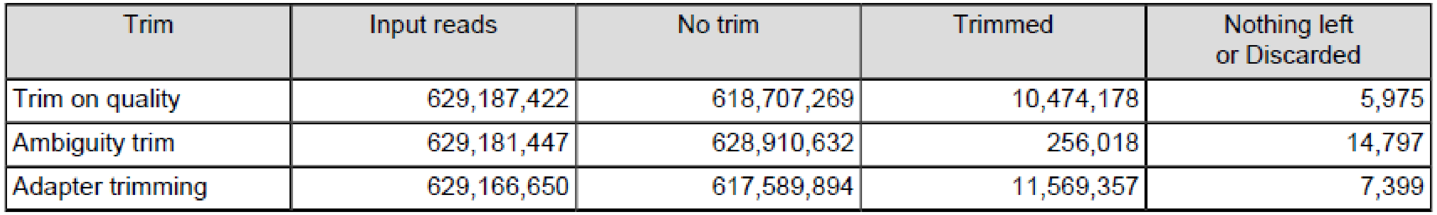
Trim settings



- Removal of low quality sequence. (limit = 0.05).
- Removal of ambigious nucleotides: maximal 2 nucleotides allowed.
- Removal of adapter sequences, using the following adapters :

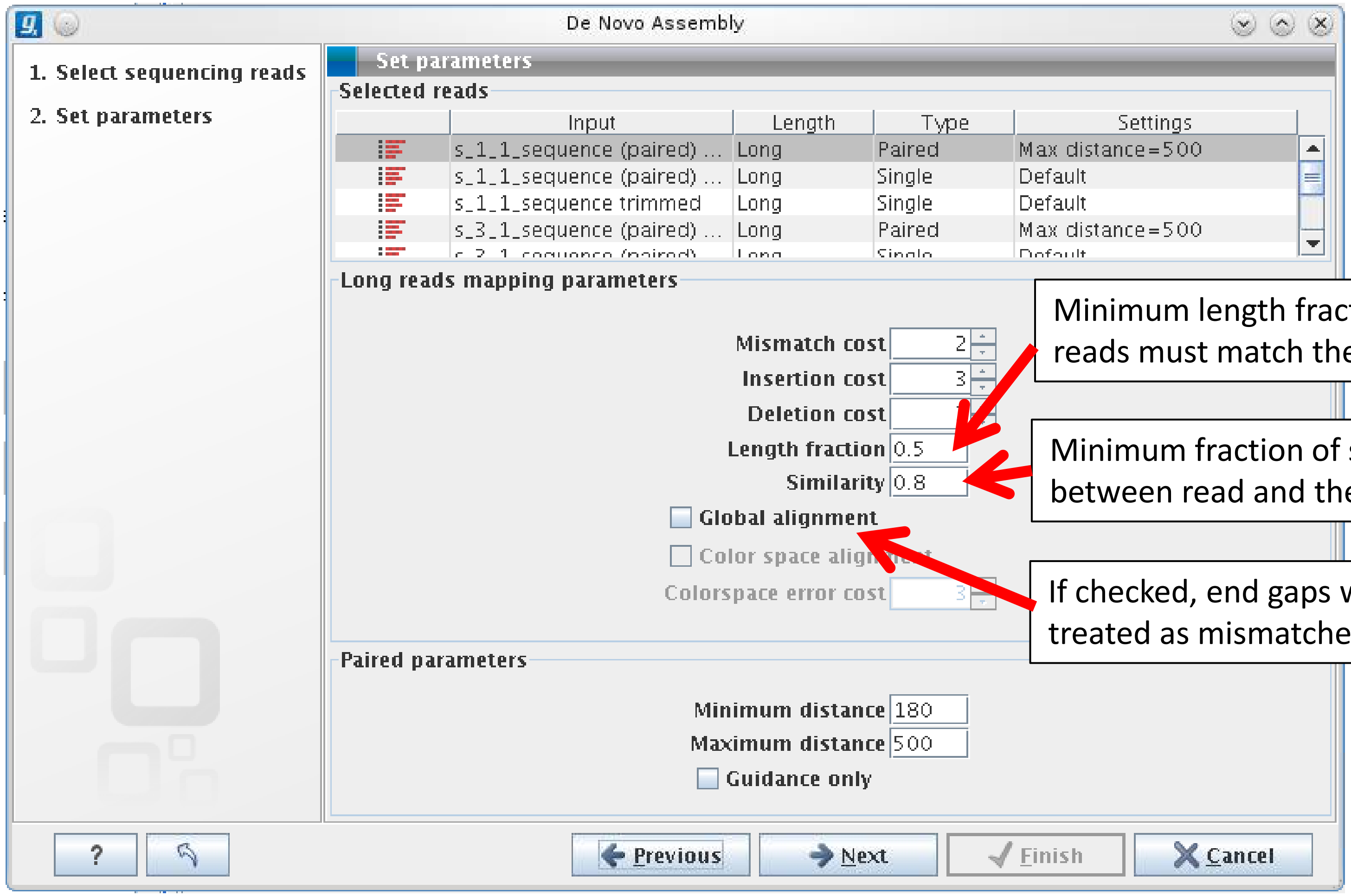
 # Illumina HiSeq 2000 PE Adapter (ACACTCTTTCCCTACACGACGCTCTTCCGATCT), strand = Plus, acti on = Remove adapter, score = [2, 3, 10, 4]

Detailed results



Trim	Input reads	No trim	Trimmed	Nothing left or Discarded
Trim on quality	629,187,422	618,707,269	10,474,178	5,975
Ambiguity trim	629,181,447	628,910,632	256,018	14,797
Adapter trimming	629,166,650	617,589,894	11,569,357	7,399

Assembly parameters



Set parameters

Selected reads

Input	Length	Type	Settings
s_1_1_sequence (paired) ...	Long	Paired	Max distance=500
s_1_1_sequence (paired) ...	Long	Single	Default
s_1_1_sequence trimmed	Long	Single	Default
s_3_1_sequence (paired) ...	Long	Paired	Max distance=500
s_3_1_sequence (paired)	Long	Single	Default

Long reads mapping parameters

Mismatch cost: 2

Insertion cost: 3

Deletion cost:

Length fraction: 0.5

Similarity: 0.8

□ Global alignment

□ Color space alignment

Colorspace error cost: 3

Paired parameters

Minimum distance: 180

Maximum distance: 500

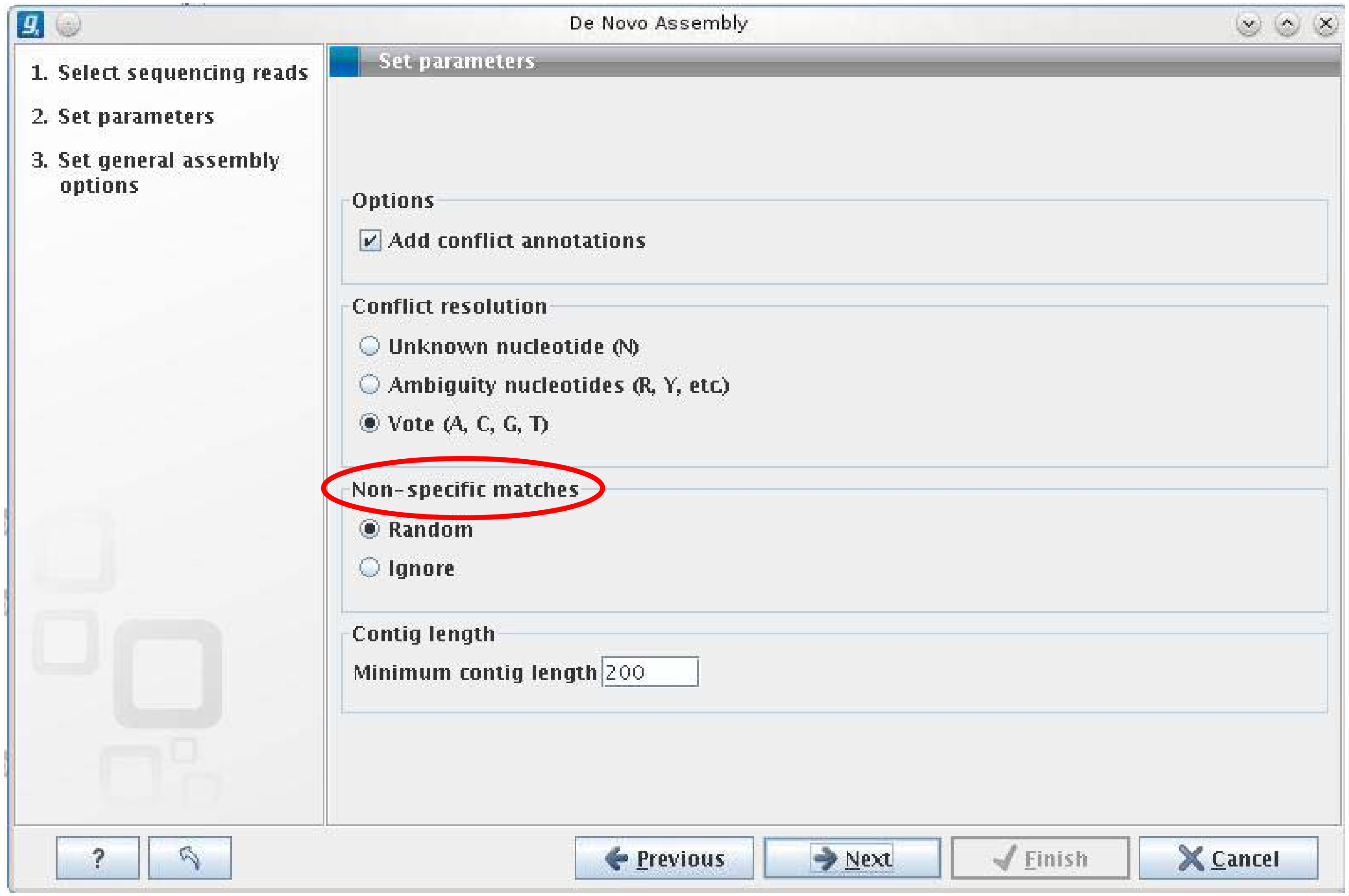
□ Guidance only

Minimum length fraction of reads must match the reference

Minimum fraction of similarity between read and the reference

If checked, end gaps will be treated as mismatches

General assembly options



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

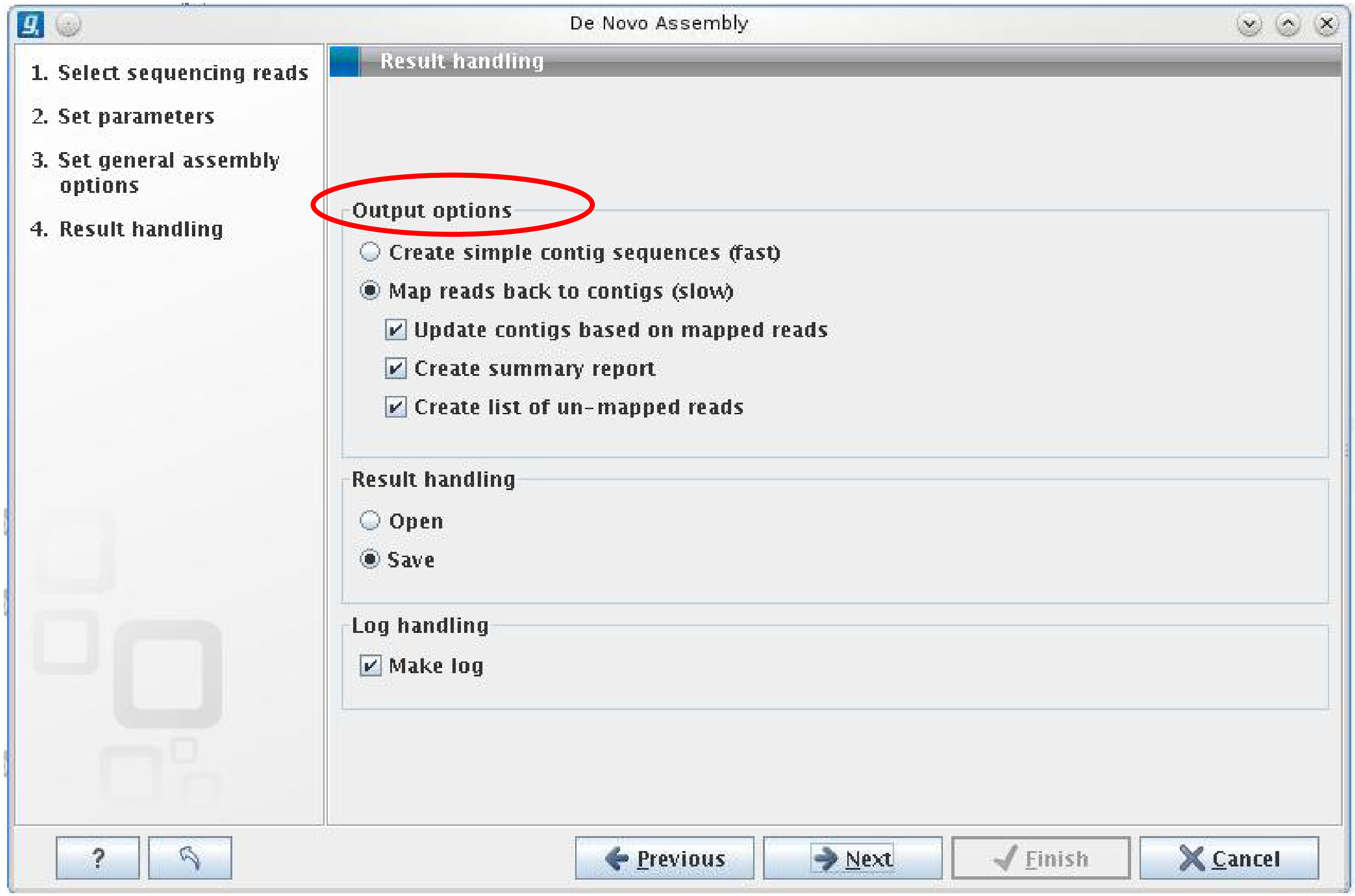
● Random

○ Ignore

Contig length

Minimum contig length 200

Assembly 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

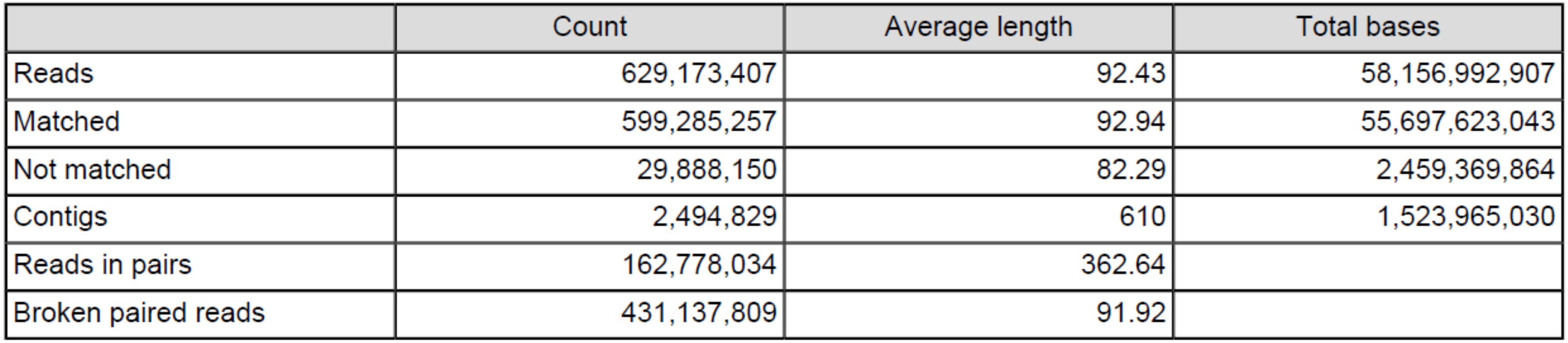
● Save

Log handling

☑ Make log

Assembly report

Summary statistics

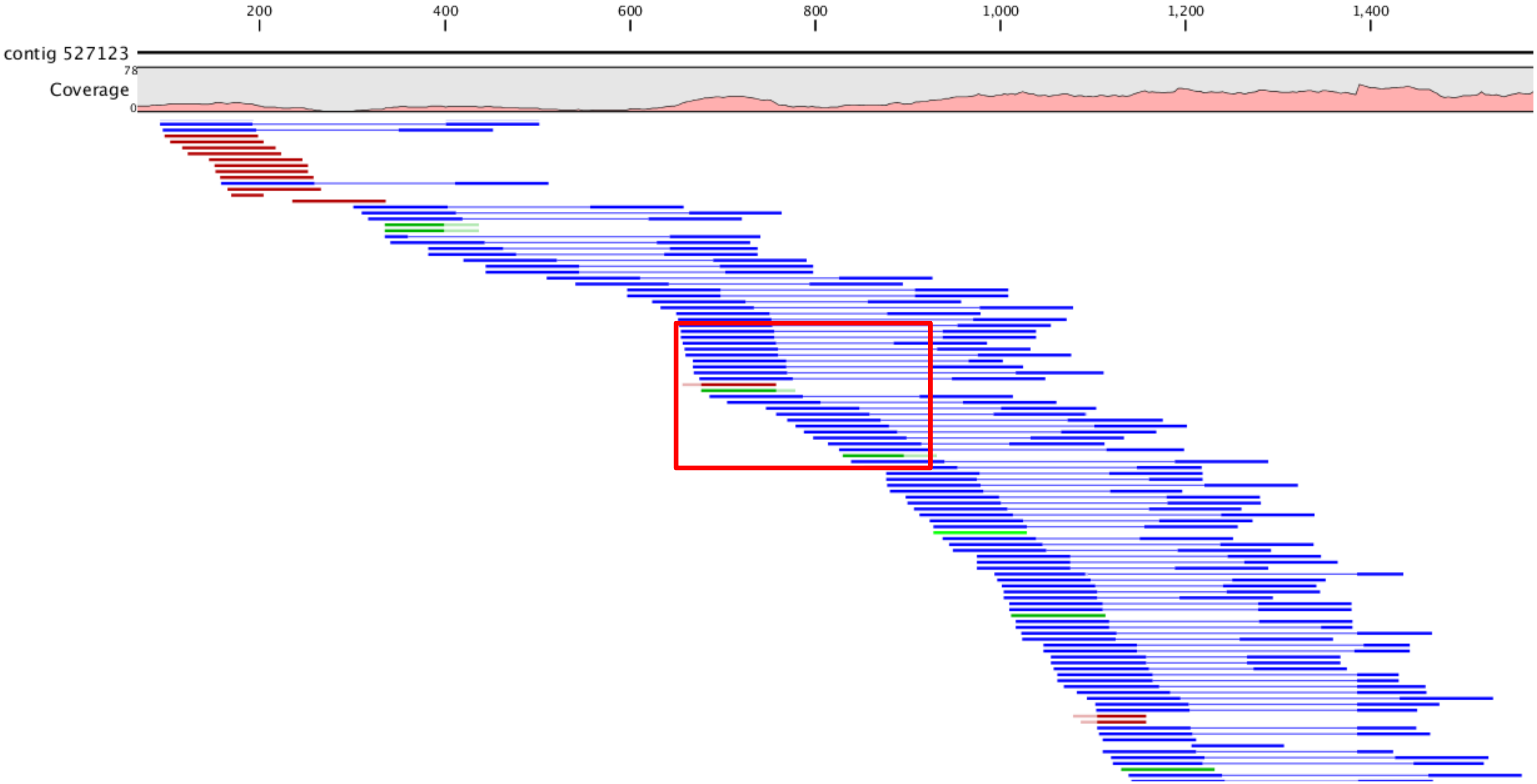


	Count	Average length	Total bases
Reads	629,173,407	92.43	58,156,992,907
Matched	599,285,257	92.94	55,697,623,043
Not matched	29,888,150	82.29	2,459,369,864
Contigs	2,494,829	610	1,523,965,030
Reads in pairs	162,778,034	362.64	
Broken paired reads	431,137,809	91.92	

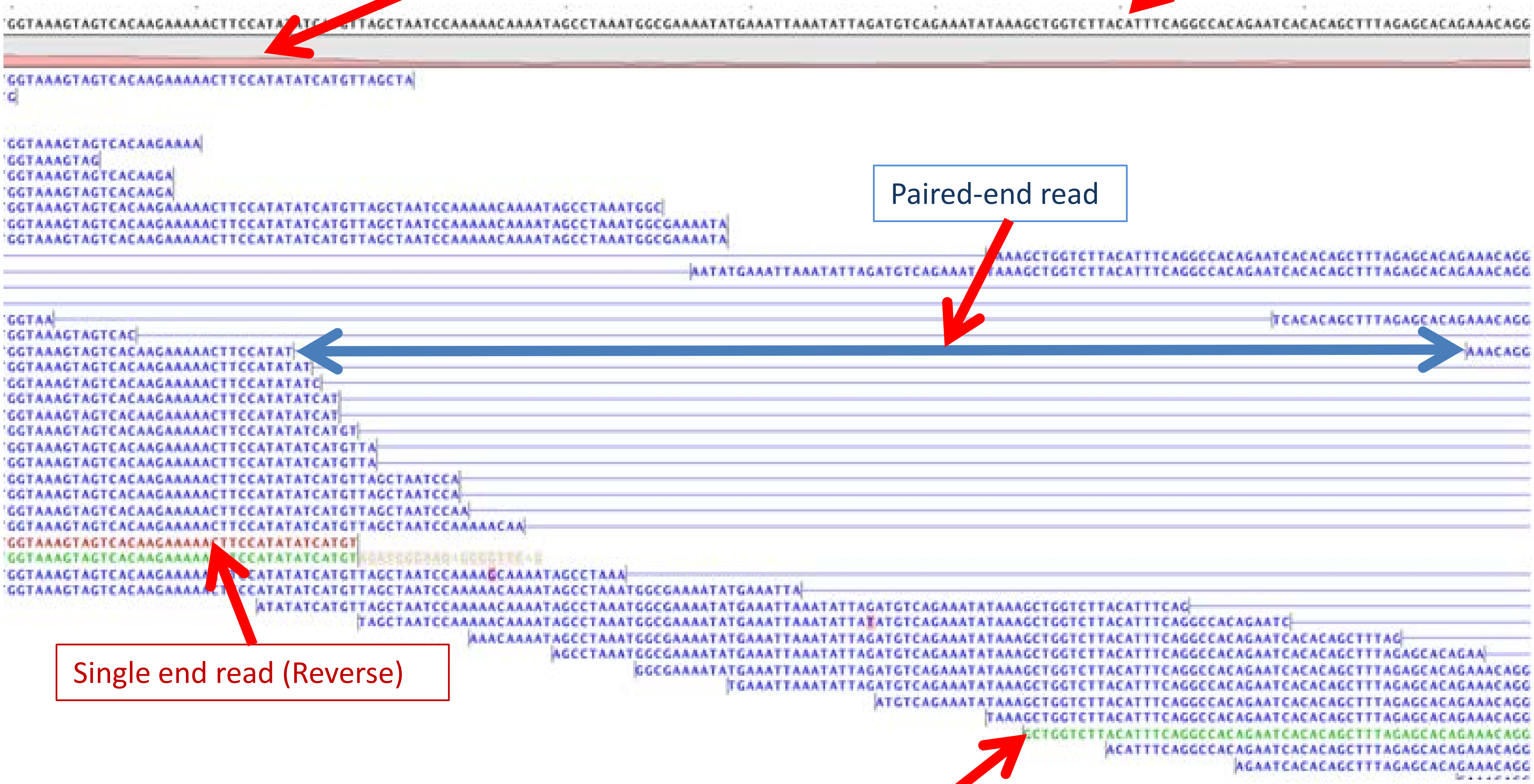
Quality assessment

Total length of sequences (bp): 1,523,965,030
Total number of contigs: 2,494,829
Max contig length (bp): 22,049
Mean contig length (bp): 610.85
Median contig length (bp): 371
Min contig length (bp): 200
N25: 1720
N50: 891
N75: 435
N90: 251
Total GC count (bp): 650,659,197
GC (%): 42.70

View assembled contig



Base level view



Coverage

Consensus sequence

Paired-end read

Single end read (Reverse)

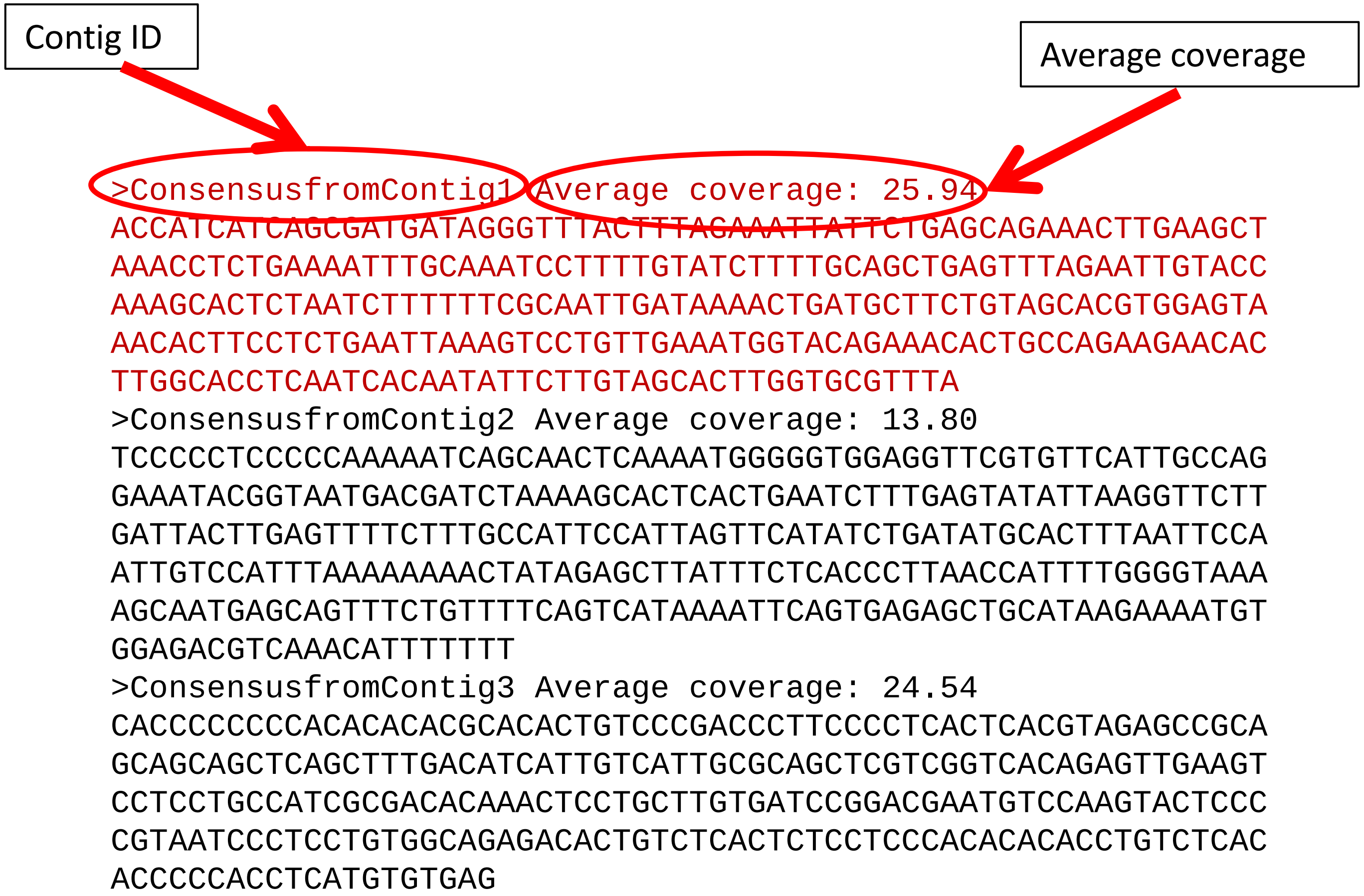
Single end read (Forward)

```
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCAAAAACAAAATAGCCTAAATGGCGAAAATATGAAATTAAATATTAGATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTA
G

GGTAAAGTAGTCACAAGAAAA
GGTAAAGTAG
GGTAAAGTAGTCACAAGA
GGTAAAGTAGTCACAAGA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCAAAAACAAAATAGCCTAAATGGC
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCAAAAACAAAATAGCCTAAATGGCGAAAATA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCAAAAACAAAATAGCCTAAATGGCGAAAATA
                                                                                                           AAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                          AATATGAAATTAAATATTAGATGTCAGAAAT TAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG


GGTAA                                                                                                                                    TCACACAGCTTTAGAGCACAGAAACAGG
GGTAAAGTAGTCAC
GGTAAAGTAGTCACAAGAAAAACTTCCATAT                                                                                                                               AAACAGG
GGTAAAGTAGTCACAAGAAAAACTTCCATATAT
GGTAAAGTAGTCACAAGAAAAACTTCCATATATC
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCAT
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCAT
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGT
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCAA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGTTAGCTAATCCAAAAACAA
GGTAAAGTAGTCACAAGAAAAACTTCCATATATCATGT
GGTAAAGTAGTCACAAGAAAA    CCATATATCATGT
GGTAAAGTAGTCACAAGAAAA     CATATATCATGTTAGCTAATCCAAAAGCAAAATAGCCTAAA
GGTAAAGTAGTCACAAGAAAA   TCCATATATCATGTTAGCTAATCCAAAAACAAAATAGCCTAAATGGCGAAAATATGAAATTA
                           ATATATCATGTTAGCTAATCCAAAAACAAAATAGCCTAAATGGCGAAAATATGAAATTAAATATTAGATGTCAGAAATATAAAGCTGGTCTTACATTTCAG
                                      TAGCTAATCCAAAAACAAAATAGCCTAAATGGCGAAAATATGAAATTAAATATTATATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATC
                                                  AAACAAAATAGCCTAAATGGCGAAAATATGAAATTAAATATTAGATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAG
                                                           AGCCTAAATGGCGAAAATATGAAATTAAATATTAGATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAA
                                                                    GGCGAAAATATGAAATTAAATATTAGATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                              TGAAATTAAATATTAGATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                                              ATGTCAGAAATATAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                                                          TAAAGCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                                                              GCTGGTCTTACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                                                                       ACATTTCAGGCCACAGAATCACACAGCTTTAGAGCACAGAAACAGG
                                                                                                                                     AGAATCACACAGCTTTAGAGCACAGAAACAGG
```

Assembled contigs



Contig ID

Average coverage

```
>ConsensusfromContig1 Average coverage: 25.94
ACCATCATCAGCGATGATAGGGTTTACTTTAGAAATTATTCTGAGCAGAAACTTGAAGCT
AAACCTCTGAAAATTTGCAAATCCTTTTGTATCTTTTGCAGCTGAGTTTAGAATTGTACC
AAAGCACTCTAATCTTTTTTCGCAATTGATAAAACTGATGCTTCTGTAGCACGTGGAGTA
AACACTTCCTCTGAATTAAAGTCCTGTTGAAATGGTACAGAAACACTGCCAGAAGAACAC
TTGGCACCTCAATCACAATATTCTTGTAGCACTTGGTGCGTTTA
>ConsensusfromContig2 Average coverage: 13.80
TCCCCCTCCCCCAAAAATCAGCAACTCAAAATGGGGGTGGAGGTTCGTGTTCATTGCCAG
GAAATACGGTAATGACGATCTAAAAGCACTCACTGAATCTTTGAGTATATTAAGGTTCTT
GATTACTTGAGTTTTCTTTGCCATTCCATTAGTTCATATCTGATATGCACTTTAATTCCA
ATTGTCCATTTAAAAAAAACTATAGAGCTTATTTCTCACCCTTAACCATTTTGGGGTAAA
AGCAATGAGCAGTTTCTGTTTTCAGTCATAAAATTCAGTGAGAGCTGCATAAGAAAATGT
GGAGACGTCAAACATTTTTTT
>ConsensusfromContig3 Average coverage: 24.54
CACCCCCCCCACACACACGCACACTGTCCCGACCCTTCCCCTCACTCACGTAGAGCCGCA
GCAGCAGCTCAGCTTTGACATCATTGTCATTGCGCAGCTCGTCGGTCACAGAGTTGAAGT
CCTCCTGCCATCGCGACACAAACTCCTGCTTGTGATCCGGACGAATGTCCAAGTACTCCC
CGTAATCCCTCCTGTGGCAGAGACACTGTCTCACTCTCCTCCCACACACACCTGTCTCAC
ACCCCCACCTCATGTGTGAG
```

Summary

- Genome sequencing and assembly problem
- Short read sequence assembly pipeline
  - Sequence data format (FASTQ)
  - Read quality assessment
  - Sequence trimming
  - De Novo assembly algorithms and tools
  - Assembly quality assessment
- Case study
  - Little Skate Illumina HiSeq 2000 short read sequences
  - CLC Bio Genomics Workbench

References

- Dawei Lin, Short Read Assembly, UC Davis 2010 Bioinformatics Short Course, September 13, 2010, UC Davis
- Joe Fass, Genome Sequence Assembly in Action!, UC Davis 2010 Bioinformatics Short Course, September 13, 2010, UC Davis
- NHGRI, http://www.genome.gov/sequencingcosts/, Feb. 4, 2011
- Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 2008;9(3):R55. Epub 2008 Mar 14.
- Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010 Sep;20(9):1165-73. Epub 2010 May 27.
- CLC Bio. Manual for CLC Genomics Workbench 4.6 Windows, Mac OS X and Linux April 1, 2011
- CLC Bio. Manual for CLC Genomics Workbench 4.7 Windows, Mac OS X and Linux
- May 14, 2011
- Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010 Jun;95(6):315-27. Epub 2010 Mar 6.

Thank You!

Questions???