QC and genome (or metagenome) assembly

Quality control in NGS reads

- Why not use raw reads?
  - PCR amplification bias
  - Sequencing errors (random)
  - Sequencer malfunction
  - Unscreened vectors/adaptors/primers
  - Contamination
  - Assembly programs are highly sensitive to the read error

Quality investigation using FastQC

Summary

- 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
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

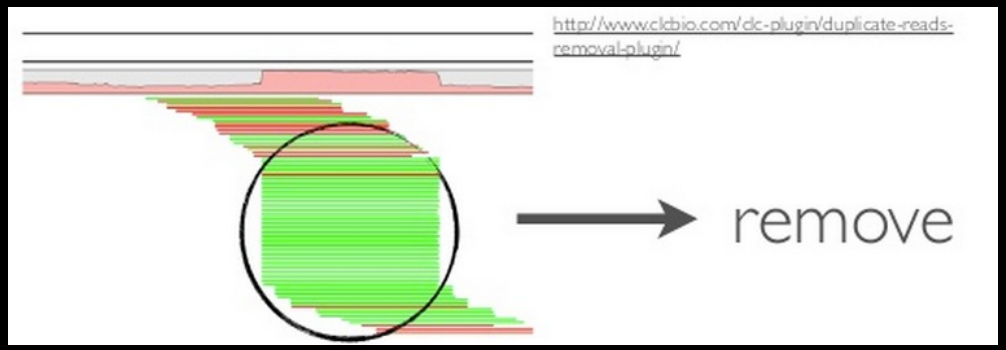
Basic Statistics

Measure	Value
Filename	mini_propreD0_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	250000
Filtered Sequences	0
Sequence length	101
%GC	44

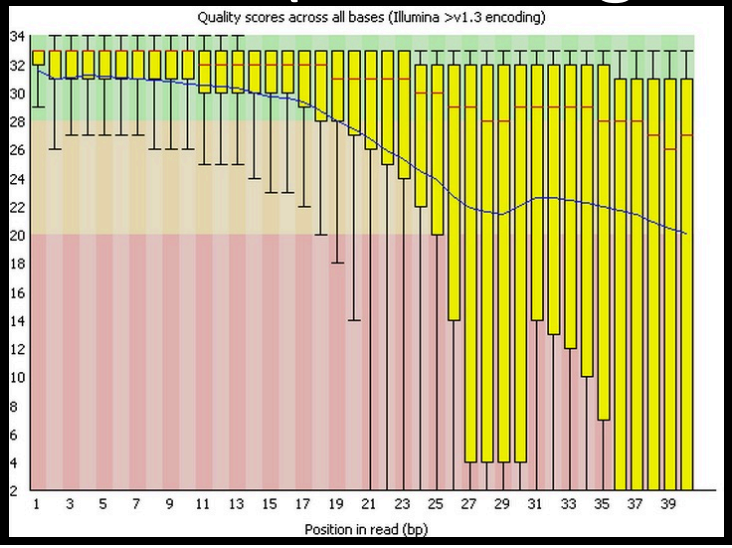
Per base sequence quality

Common criteria to filter low quality sequences

- Vector, adaptor or primer sequences
- PCR duplicates



- Low quality region from the end (including the whole reads)



Practice

- Use fastqc to review the quality in given raw read file

- Use in-house perl script to remove the adaptors, low quality regions and PCR duplicates, etc. for genomic data.

Reference:

Probiotics modulated gut microbiota suppresses hepatocellular carcinoma growth in mice.
Li J, Sung CY, Lee N, **Ni Y,** Pihlajamäki J, **Panagiotou G**, El-Nezami H.Proc Natl Acad Sci U S A. 2016 Mar 1;113(9):

Format of the in-house QC result

```
Progrom,Parameters,Input and Output information:
/usr/bin/pipelineForQC.pl all -f mini_propreD0_1.fq -r mini_propreD0_2.fq -c 20 -q 33 -a ATCGGAAGAGC -l 30 -p -o mini_propreD0_qc
Input file:mini_propreD0_1.fq and mini_propreD0_2.fq
Low quality threshold:20
Input file format:33
Adapter sequence: ATCGGAAGAGC
Output file: mini_propreD0_qc.20_1.fastq and mini_propreD0_qc.20_2.fastq
Output stat file: mini_propreD0_qc.stat.xls

Remove adpter information:
Name    Total_Reads     Valid_Data      Percentage
mini_propreD0_1.fq      250000  248999  99.59%
mini_propreD0_2.fq      250000  248999  99.59%

Remove lower bases information:
Name    Total_Reads     Total_bases     High_quility_reads      High_quility_bases      Single_Reads    Single_bases
mini_propreD0_1.fq      250000  25250000        244361  24382157(96.56%)        3886    378082(1.49%)
mini_propreD0_2.fq      250000  25250000        244361  24265052(96.09%)        222     20744(0.08%)

Duplication information:
Total reads: 244361
Duplicate reads: 503
Duplication: 0.2%

All completed successfully!

Time consumption is 0.26 min
```

Percentage of high quality bases

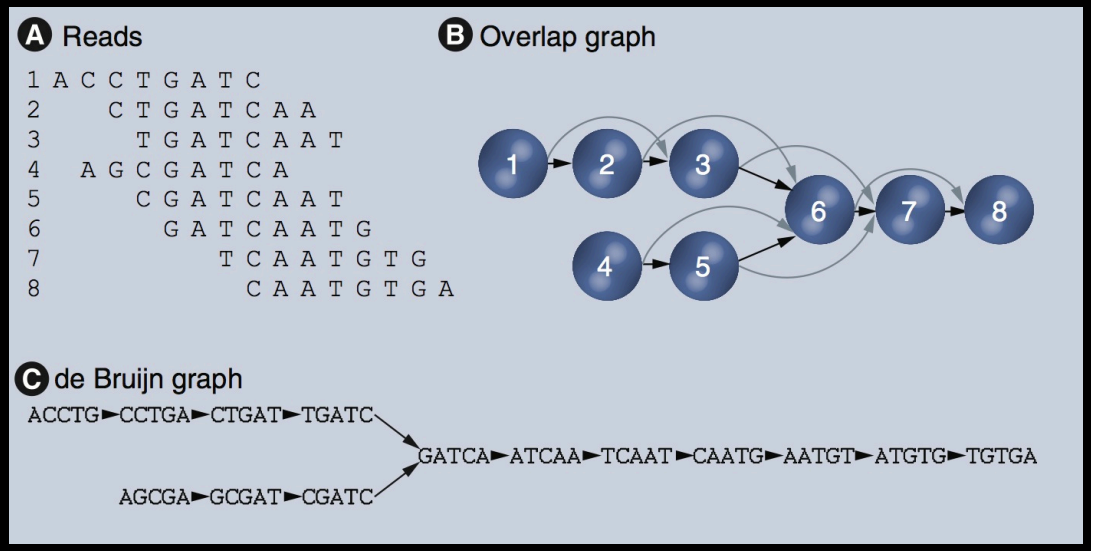
Percentage of PCR duplicates

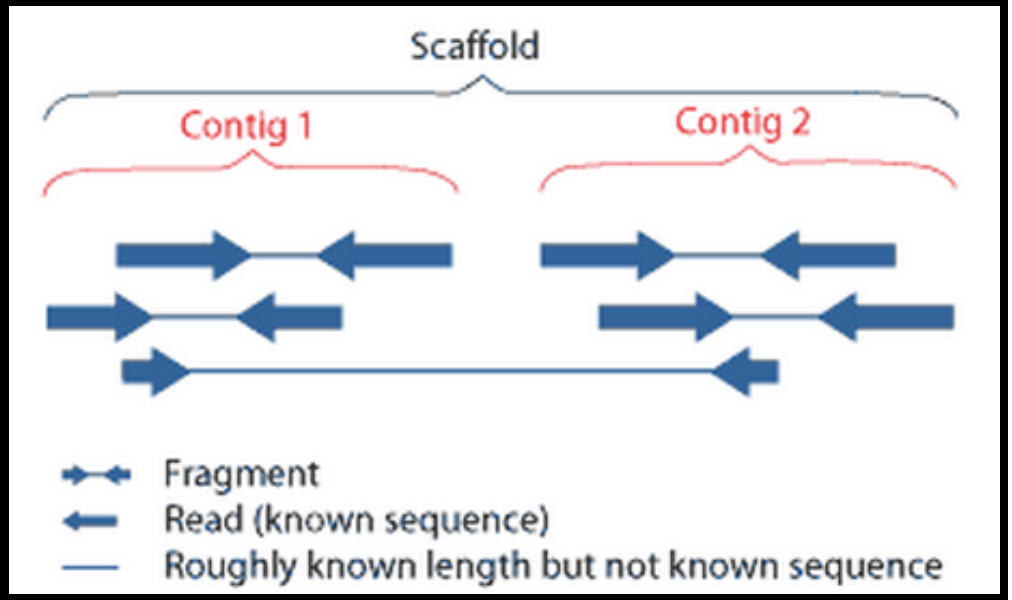
Genome assembling

- Why assemble?
  - Get full-length genes and better annotation
  - Achieve genome structure
  - Raw materials for in-depth comparative genomics and evolutionary study

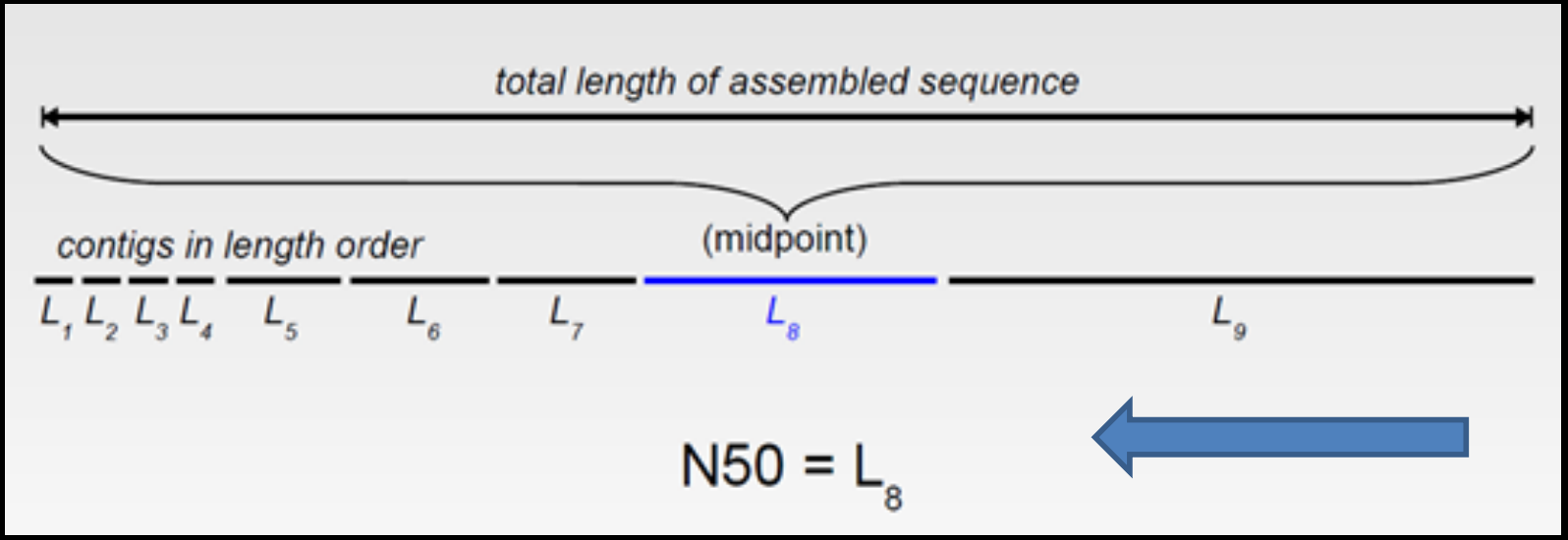
Important concept in genome assembly

- K-mer
  - Length k substring
- Overlap and de Bruijn graph
- Contig, scaffold and N50, L50 size
- Singleton
  - Reads have no overlap with other reads





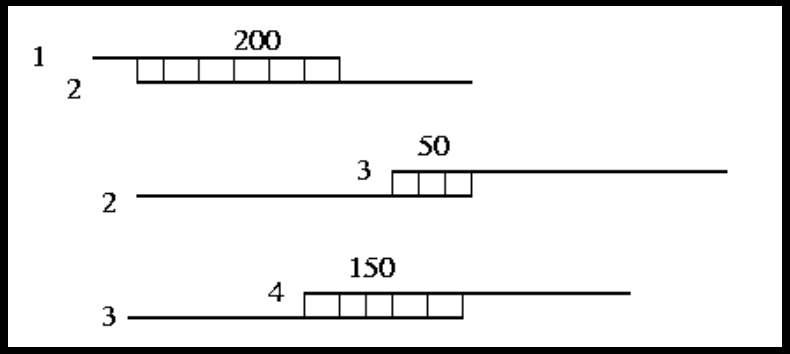
How to calculate N50

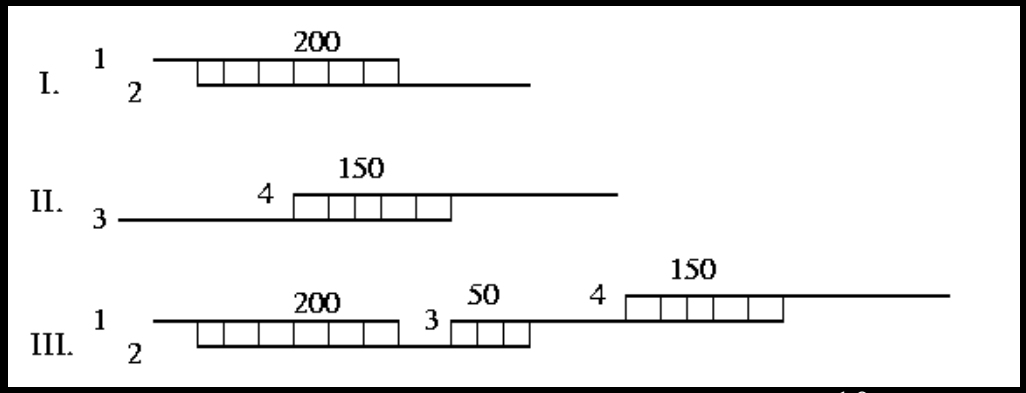


TIGR Assembler/phrap

Greedy algorithm

- Build a rough map of fragment overlaps
- Pick the largest scoring overlap
- Merge the two fragments
- Repeat until no more merges can be done

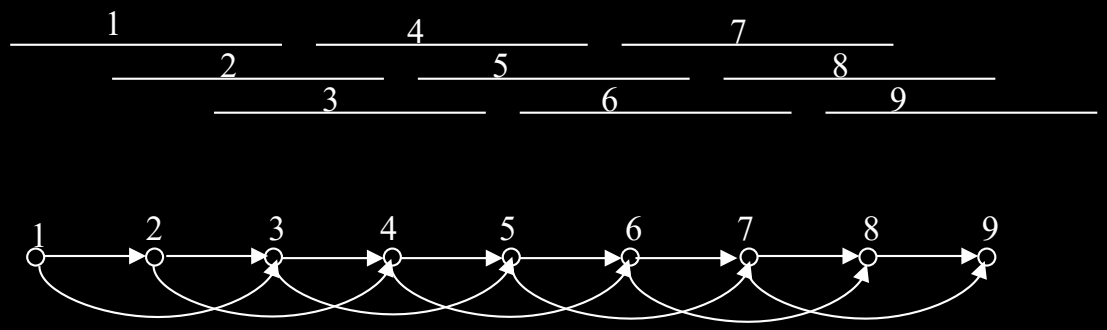




Overlap-layout-consensus (Newbler)

Main entity: read

Relationship between reads: overlap



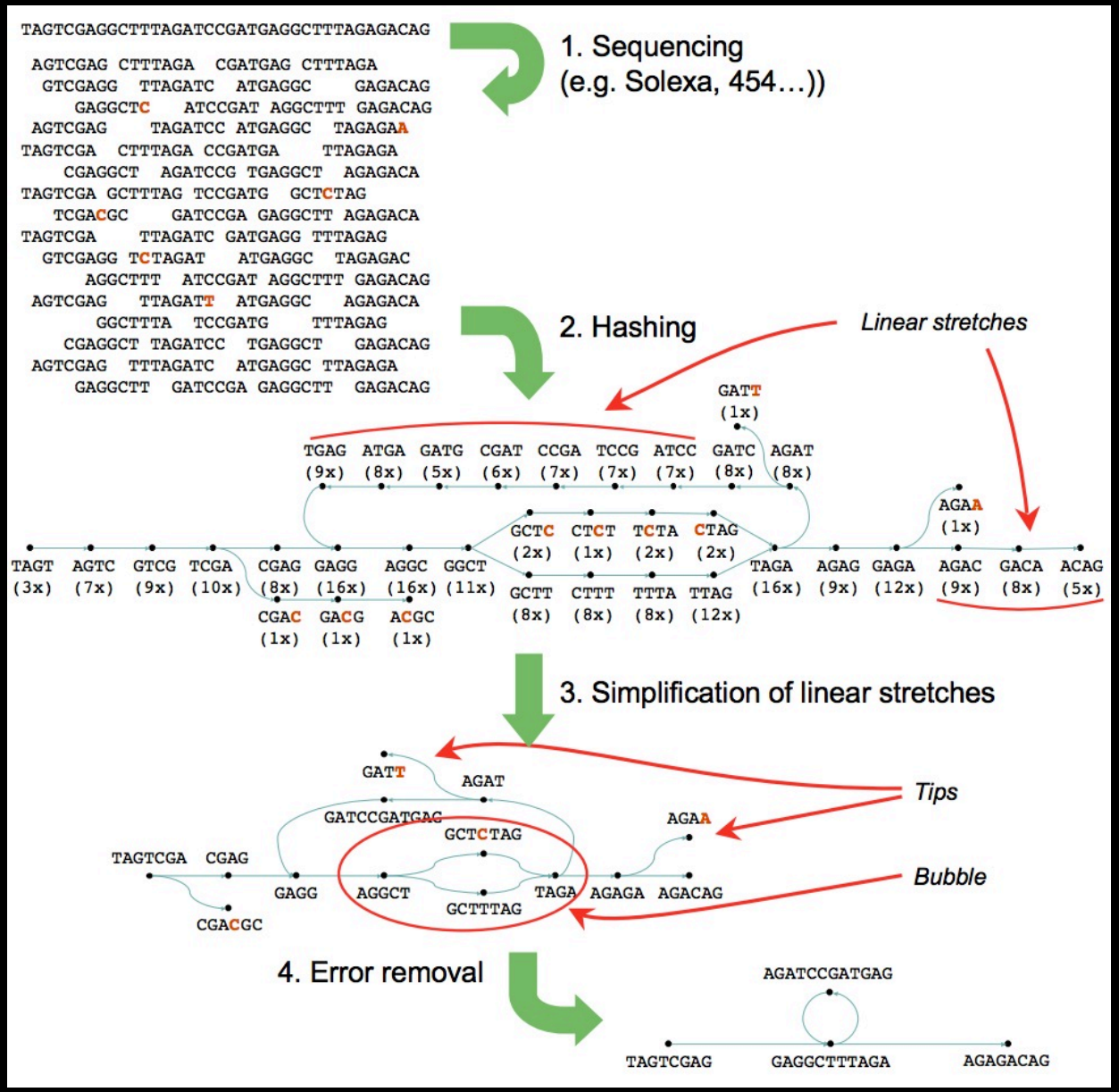
ACCTGA

ACCTGA

AGCTGA

ACCAGA

De Bruijn graph assembler (IDBA, Velvet, etc)



Velvet

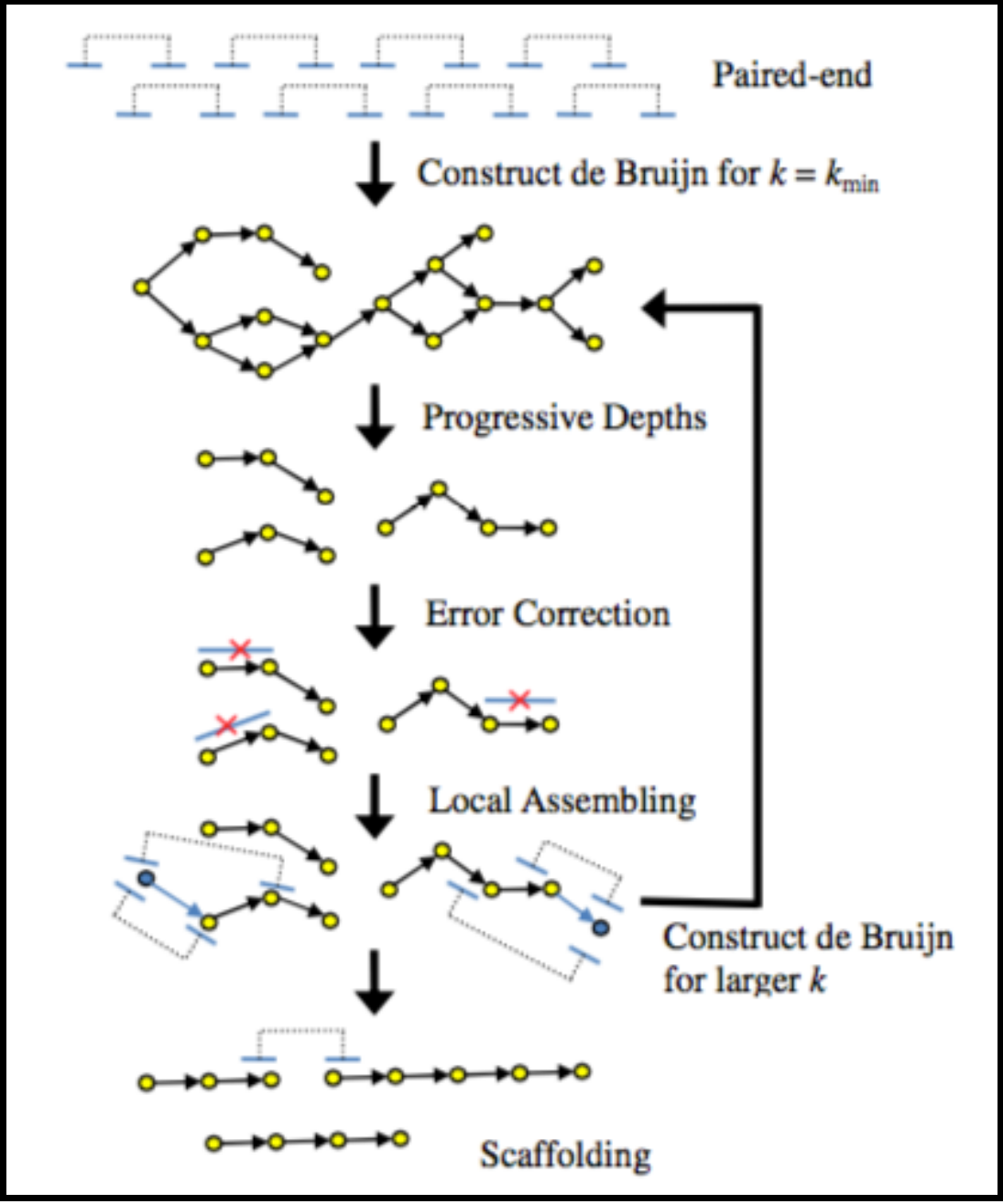
Software for NGS de novo assembly

- **Velvet**
  - special solution for repeat reconstruction
- **Newbler**
  - **overlap-layout-consensus method**
- **IDBA**
  - de Bruijn assembler with iterative k-mer size
- Soapdenovo
  - General assembler
- ALLPATHs-LG
  - Base quality information adopted
- Abyss
  - Parallel computing and speed
- ...

Metagenome assembly

- Why assembling instead of directly mapping?
  - Better functional annotation
  - Achieve normal genes, genomic segments or full genomes
  - Distinguish the genomes highly similar species (still very hard)
  - Study the dynamics of genome structure, including HGT
- Softwares
  - MetaVelvet
  - IDBA_ud
  - SOAPdenovo

Principles in IDBA_UD



Output (summary) of IDBA_UD

```
kmer 80
kmers 260644 260249
merge bubble 14
contigs: 394 n50: 710 max: 6000 mean: 725 total length: 285690 n80: 492
aligned 13660 reads
confirmed bases: 137591 correct reads: 4308 bases: 0
distance mean 192.373 sd 70.5041
seed contigs 394 local contigs 788
kmer 90
kmers 251810 251430
merge bubble 13
contigs: 387 n50: 719 max: 6000 mean: 731 total length: 283011 n80: 495
aligned 13567 reads
confirmed bases: 136480 correct reads: 4299 bases: 0
distance mean 192.488 sd 70.5873
seed contigs 387 local contigs 774
kmer 100
kmers 245808 245433
merge bubble 11
contigs: 372 n50: 742 max: 6000 mean: 744 total length: 277125 n80: 506
reads 62128
aligned 13386 reads
distance mean 192.541 sd 70.8117
expected coverage 0.0557924
edgs 4
contigs: 368 n50: 743 max: 6000 mean: 751 total length: 276510 n80: 507
(END)
```

Evaluation of genome assembly

- Criteria
  - Uneven coverage
  - Escalated SNP rates
  - Mis-orientated PE reads
  - Improper insert length of PE

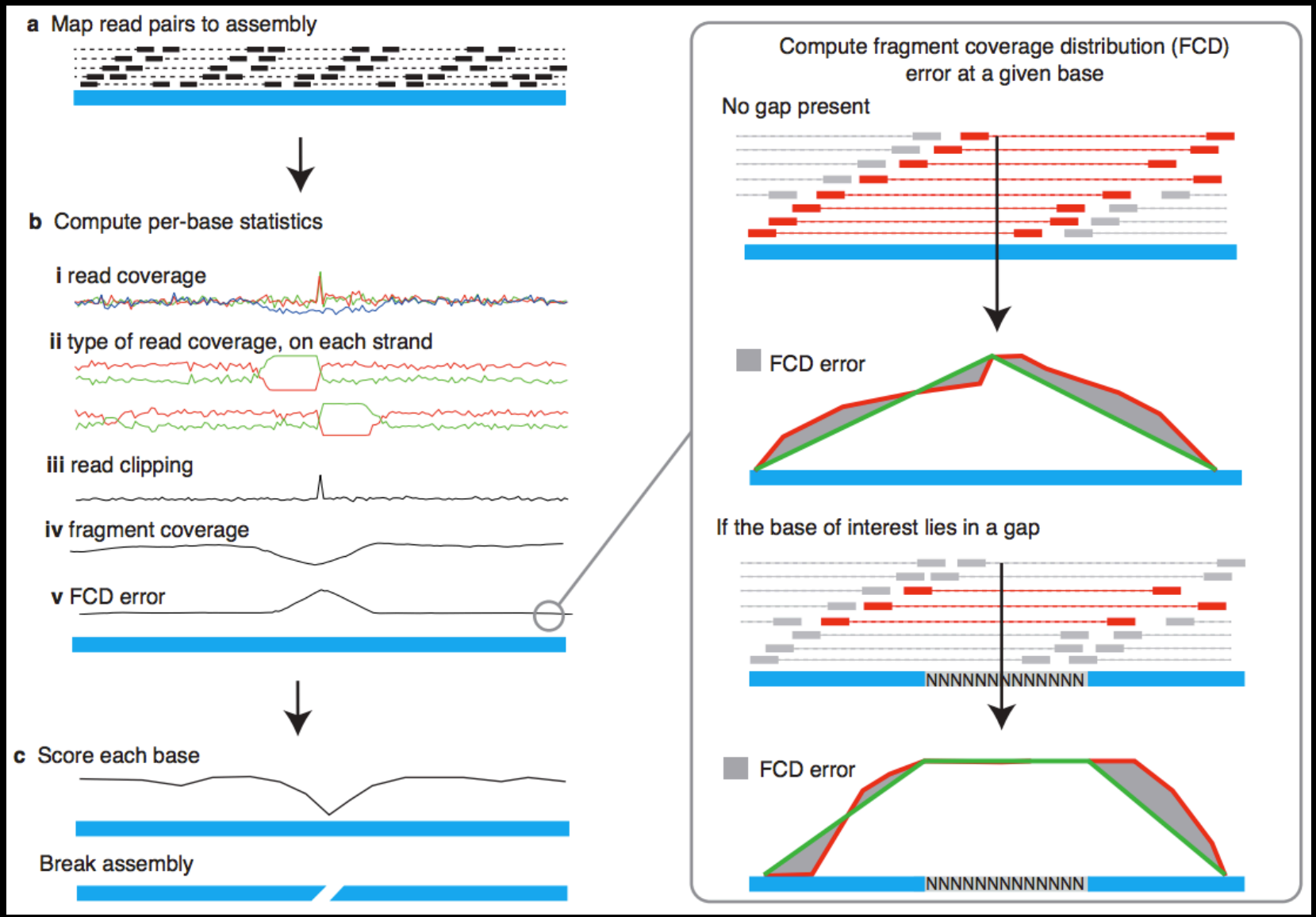
- Softwares
  - Gage
  - Assemblathon
  - Amosvalidate
  - CGAL
  - ALE
  - **REAPR**
  - **GENOVO**

] Reference genome required

Important concepts in REAPR

- FCD errors
  - Fragment Coverage Distribution
  - Difference between the theoretical (context based) and observed FCD
- Error free bases
  - at least 5X perfect and unique coverage
  - the FCD at that base is OK.

Principles of REAPR



Genovo, a principle based genome evaluator

$$\sum_i \text{score}^i_{\text{READ}} - \log(|\mathcal{B}|)L + \log(|\mathcal{B}|)V_0 S$$

S, the number of contigs
L, the total length of all the contigs
B, the DNA letter alphabet
V_0, minimal overlap bases between two consecutive reads

- Frist item
  - Evaluate whether the read occurred on their right contigs
  - Any difference between read and contig will confer penalty
- Second item
  - Penalize the total contig size
  - Incur penalty if contig size is too big (inflating the read mapping score)
- Third item
  - The total overlap length between reads

Practice

- Assemble genome using IDBA, Velvet and Newbler
- Evaluate the genome assembly and pick up the best one