

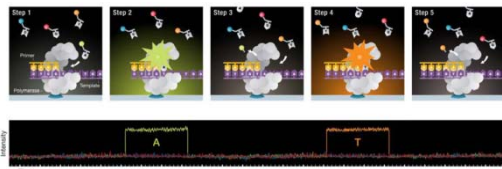


Next Generation Sequences & Chloroplast Assembly

Next Generation Sequencing (NGS) Technologies



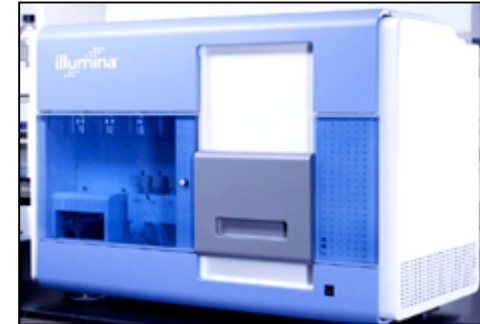
GS-Titanium; Roche 454



SMRT; Pacific Bioscience



SOLiD; ABI



Solexa; Illumina



Helicos; Helicos Bioscience

- Next (or Current) generation sequencing technologies have accelerated the speed of genome sequencing projects and have broaden application range of genome sequences.

NGS: 454 Technology

DNA Isolation

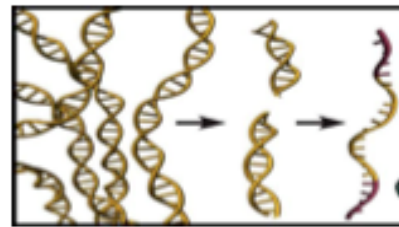
DNA Fragmentation

DNA Amplification

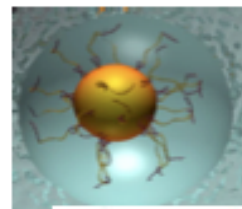
DNA Sequencing

Bioinformatics

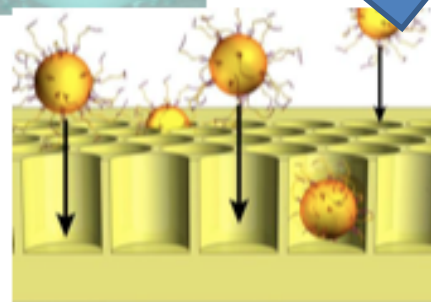
- Contigs are constructed



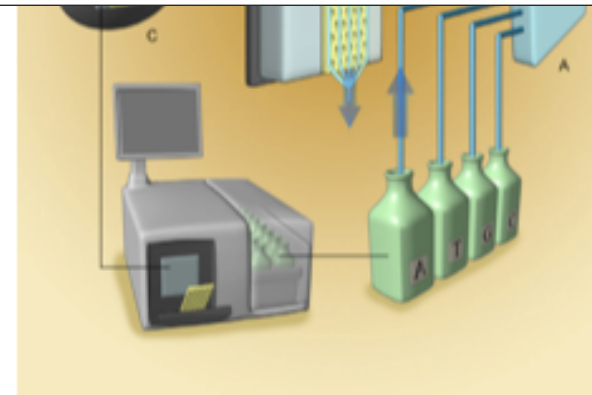
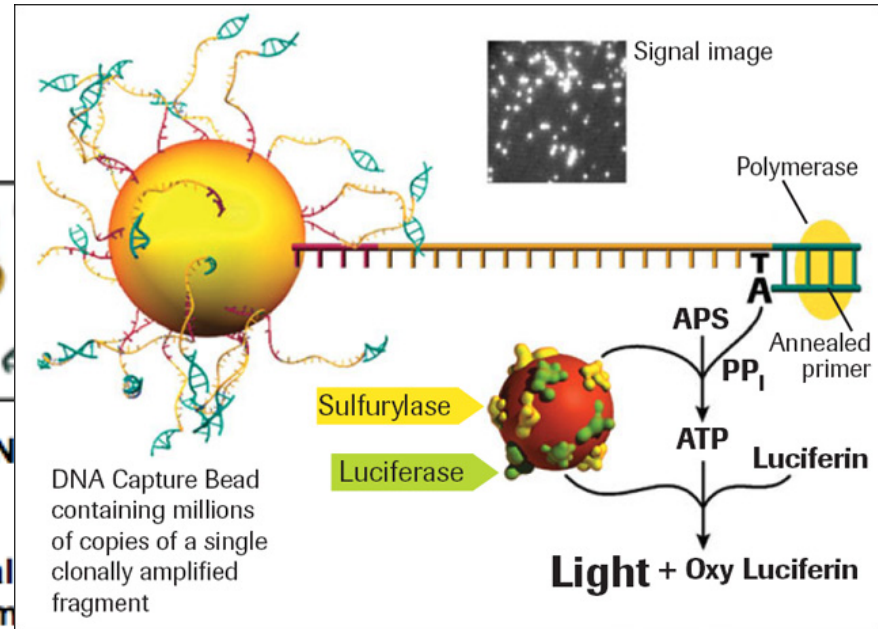
1) Prepare Adapter Ligated ssDNA



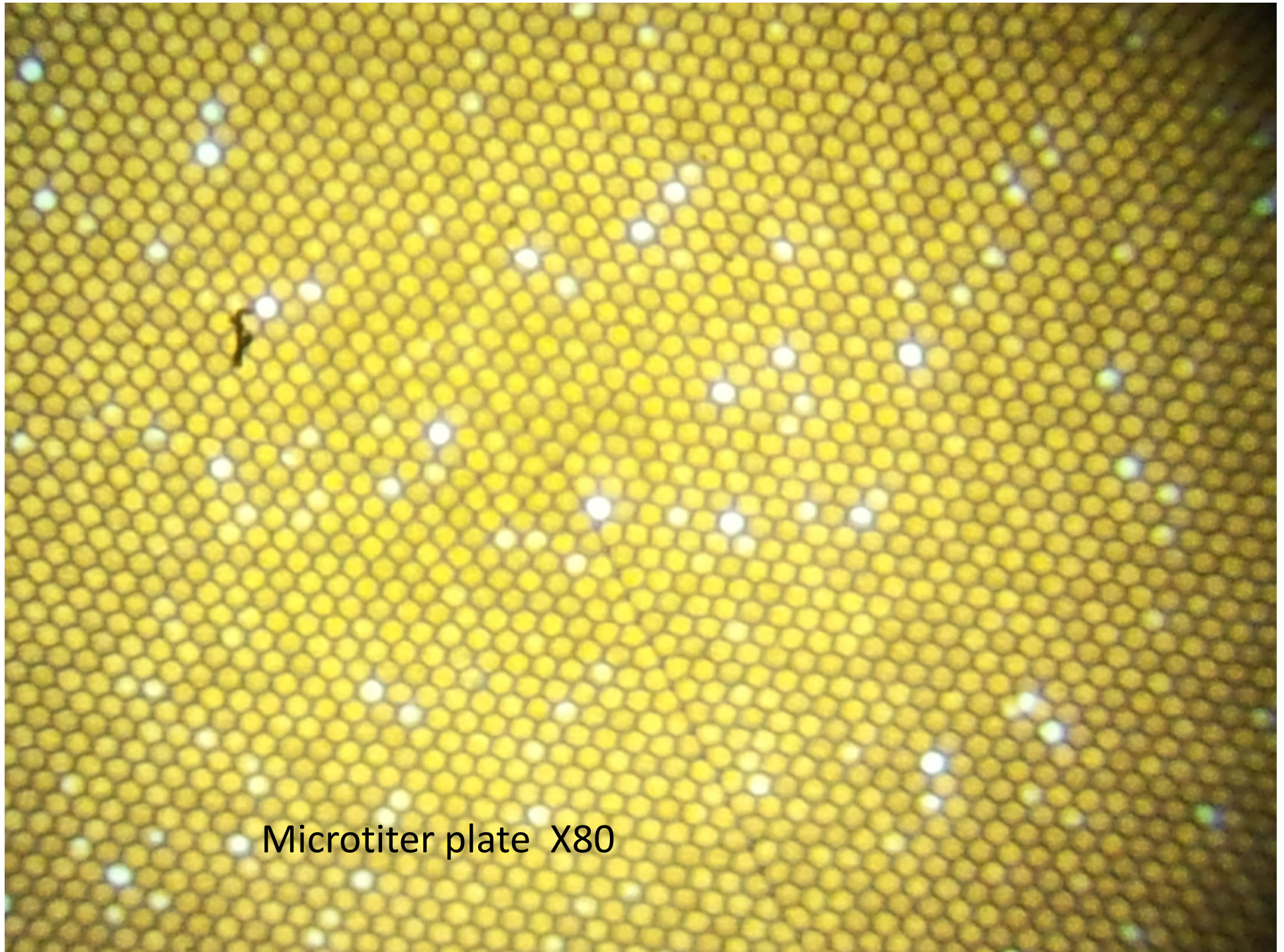
2) Clonal amplification on 28 μm



3) Load beads and enzymes in PicoTiter Plate™

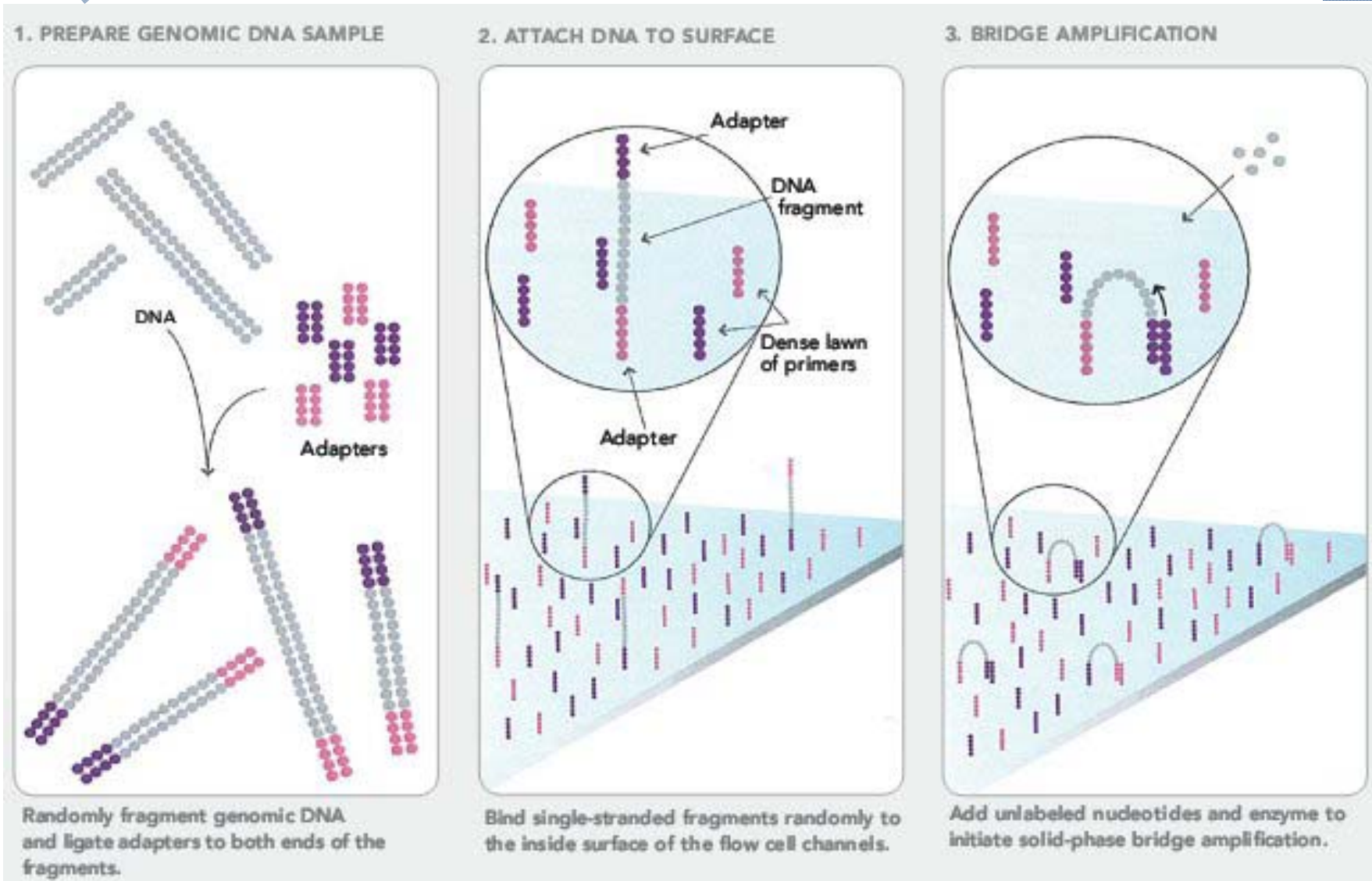


4) Perform Sequencing by synthesis on the 454 Instrument



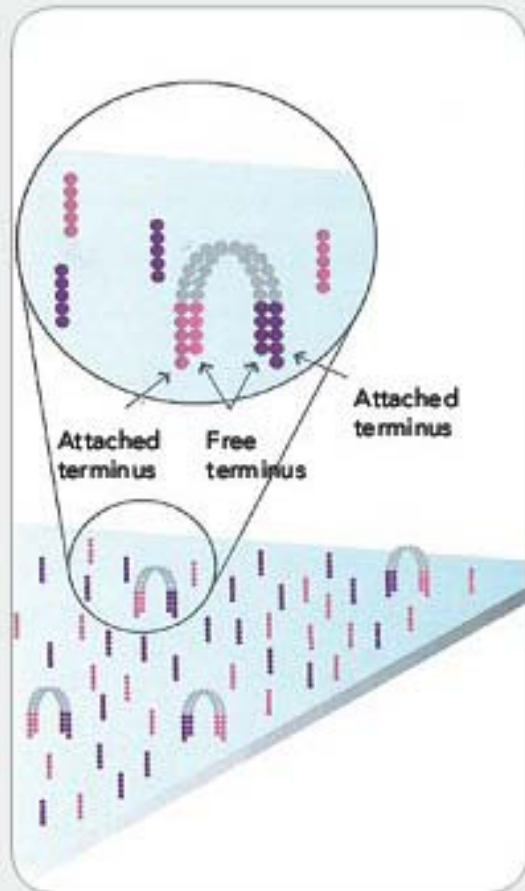
Microtiter plate X80

NGS: Solexa Technology



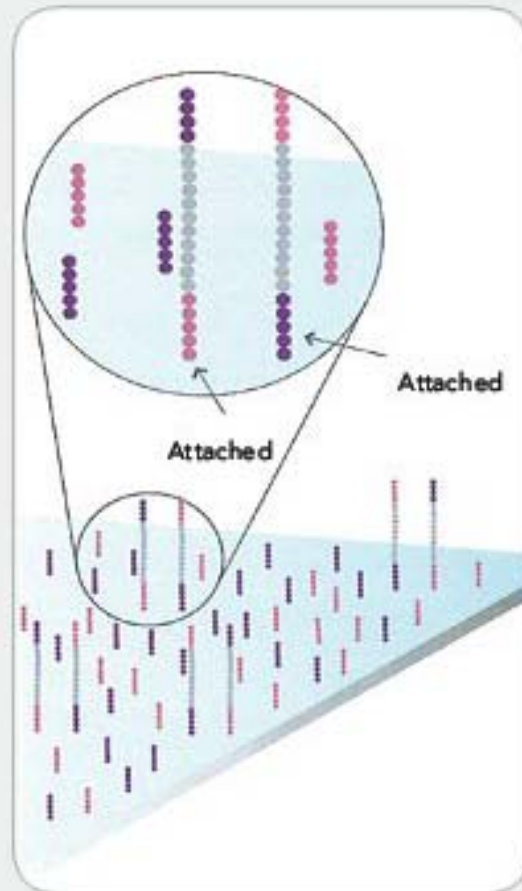
NGS: Solexa Technology

4. FRAGMENTS BECOME DOUBLE STRANDED



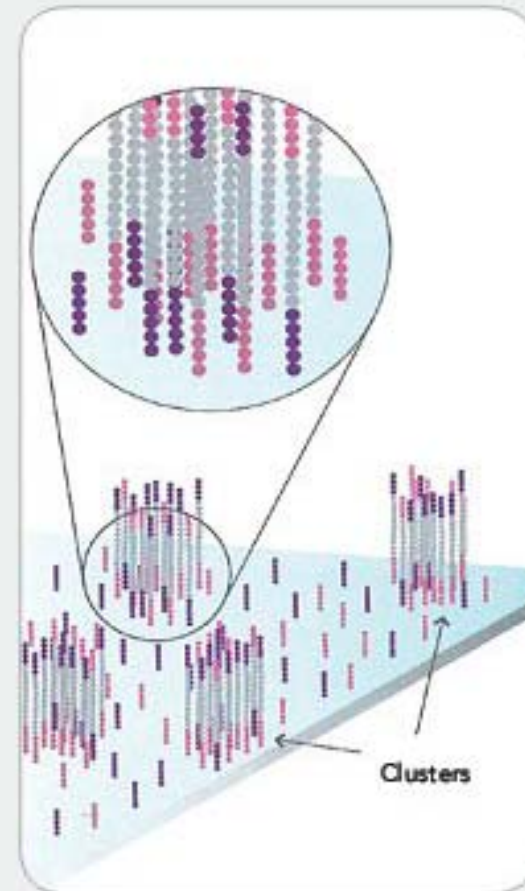
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



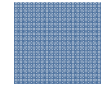
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

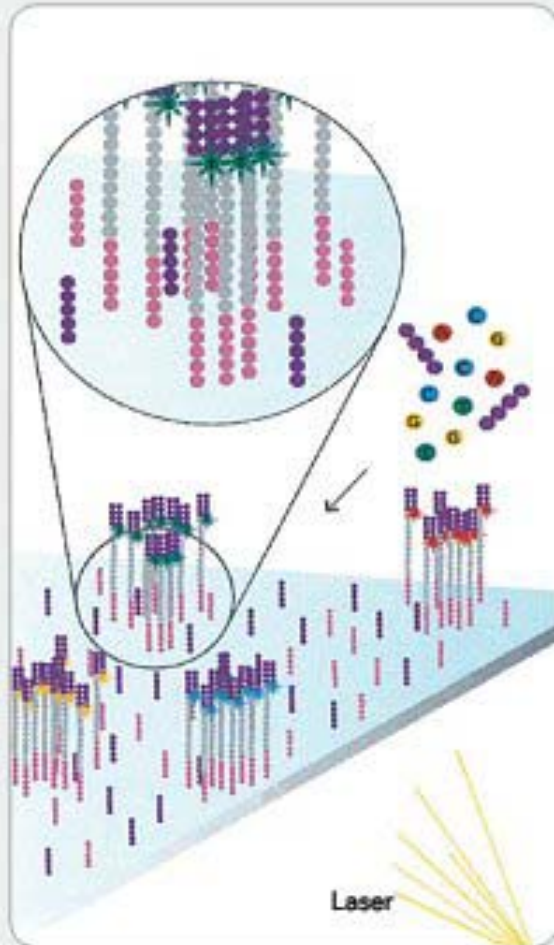


Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

NGS: Solexa Technology



7. DETERMINE FIRST BASE



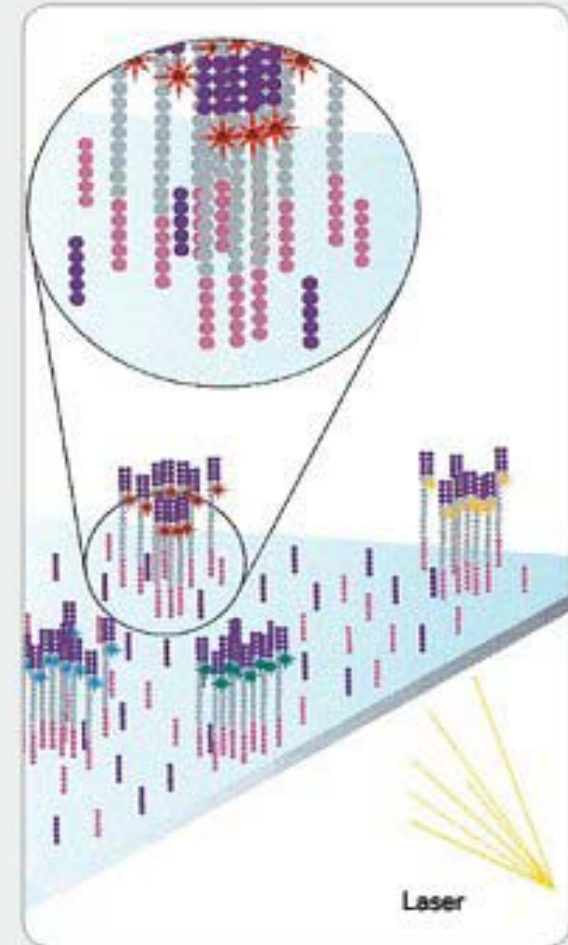
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

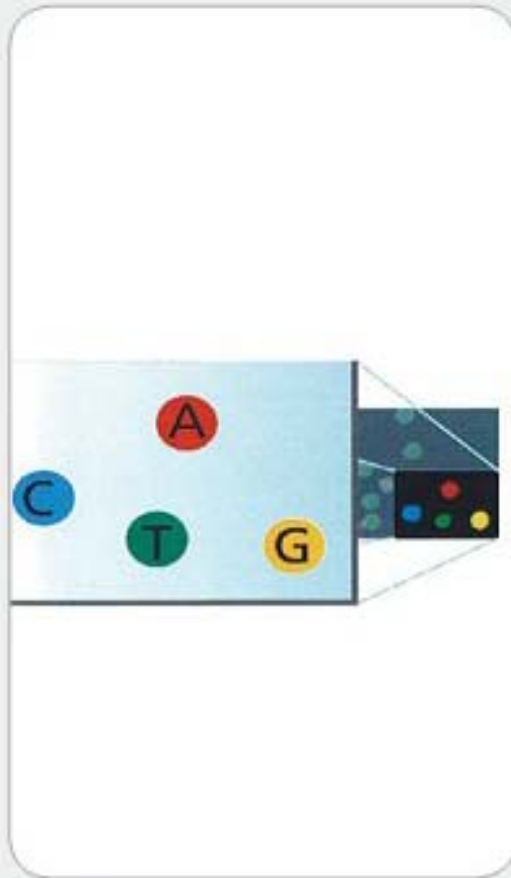
9. DETERMINE SECOND BASE



Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

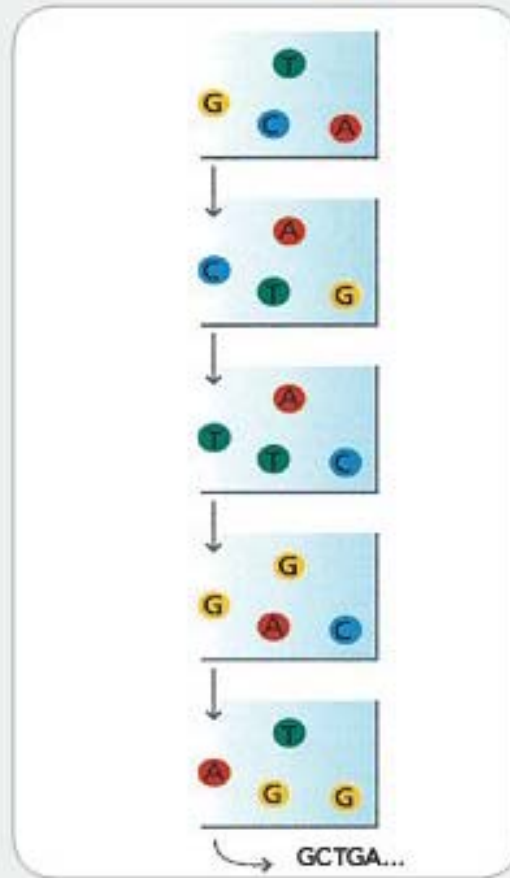
NGS: Solexa Technology

10. IMAGE SECOND CHEMISTRY CYCLE



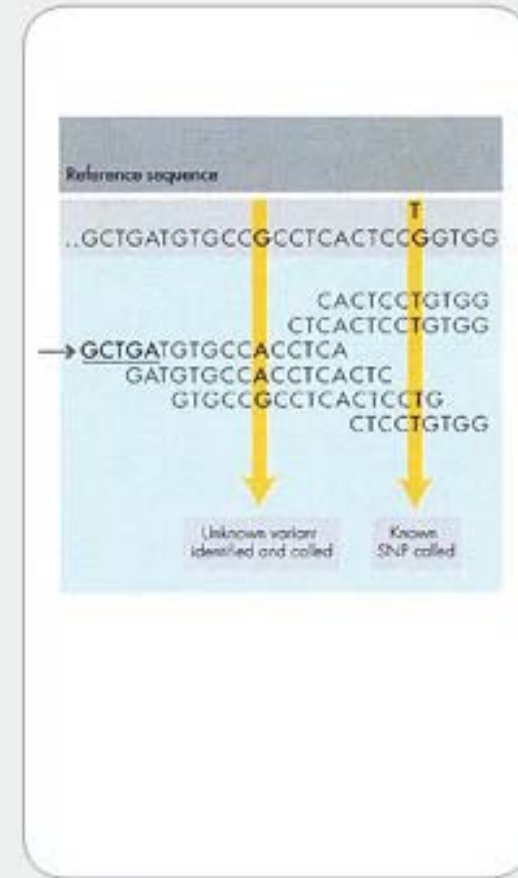
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

Capacities of Next Generation Sequencers



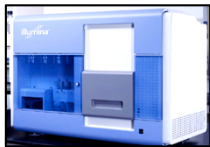
ABI 3730; ABI

$384 \times 700 \text{ bp} = 268,800 \text{ bp} = \mathbf{269Kb}$ (per one reaction / 1 hr)



GS-Titanium; Roche 454

$950,000 \times 450 \text{ bp} = 405,000,000 \text{ bp} = \mathbf{405Mb}$ (per one reaction / 2-3 days)



Solexa GA2; Illumina

$35,000,000 \times 7 \times (151 \times 2) \text{ bp} = 73,990,000,000 \text{ bp} = \mathbf{74.0 Gb}$ (per one reaction / 12 days)



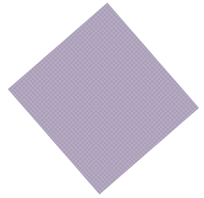
HiSeq2000; Illumina

$70,000,000 \times 14 \times (101 \times 2) \text{ bp} = 197,960,000,000 \text{ bp} = \mathbf{198.0 Gb}$ (per one reaction / 8 days)

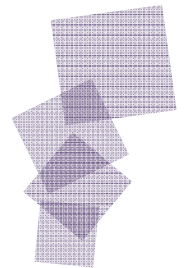
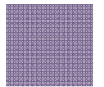


SOLiD 4; ABI

$1,400,000,000 \times 75 \text{ bp} (50+25) = 105,500,000,000 \text{ bp} = \mathbf{105.5Gb}$ (per one reaction / 11 days)

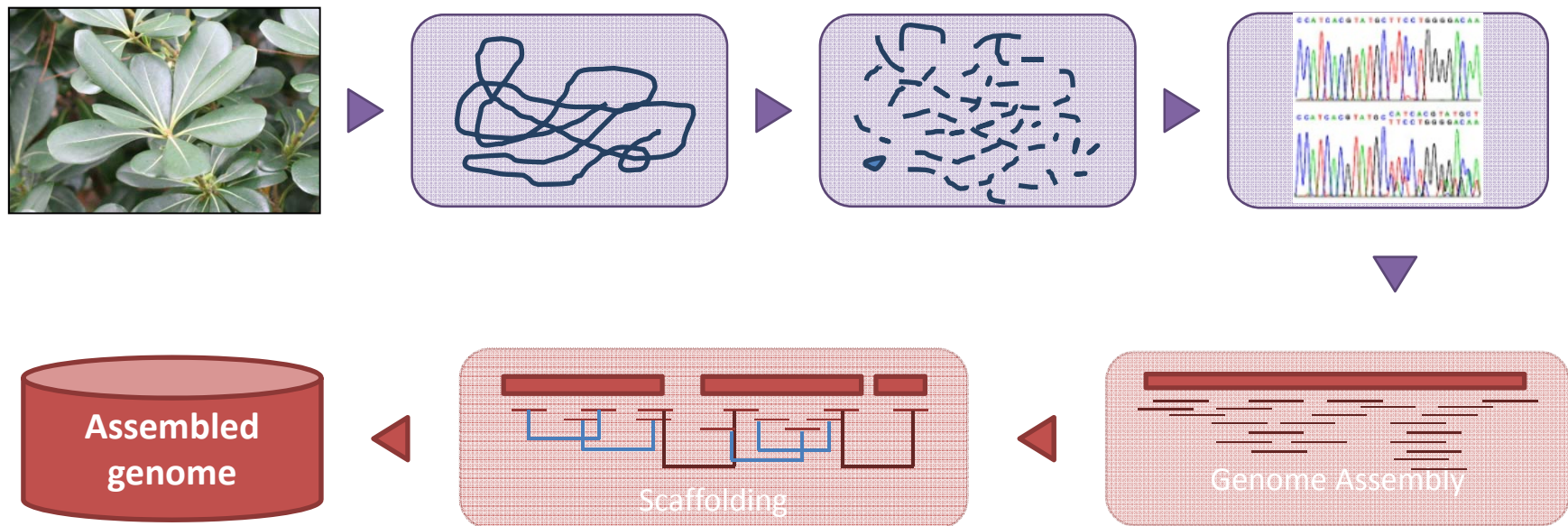


Genome Assembly Processes With NGS Sequences



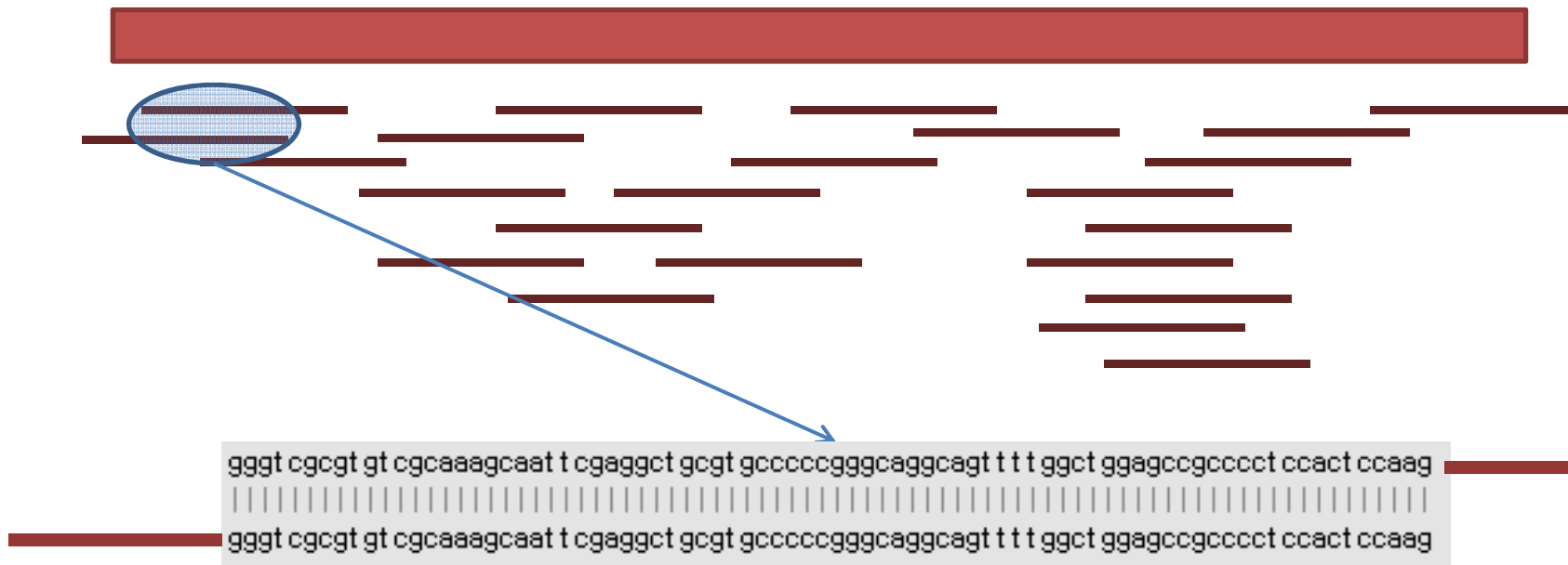
Whole Shotgun Sequence Strategy

- Assembly process is essential for genome project because read length of each sequence is less than 1 kb.
- Assembly process was conducted by several popular programs, such as phrap and PCAP3, for Sanger sequences.



Genome Assembly Process

- We can perform genome assembly manually!



23 sequences should be **compared** with each other!

$${}^{23}C_2 = 23 \times 22 / 2 = 253 \text{ comparison!}$$

How to Find Overlapped Sequences?

- Using **dynamic algorithm**, we can make the program for finding similar sequences.
- Complexity of this algorithm is $O(n^3)$.

GAATTCAGTTA (sequence #1)
GGATCGA (sequence #2)

	G	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1
A	0	1	1	1	1	1	1	1	1	1
T	0	1	1	1	1	1	1	1	1	1
C	0	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1
A	0	1	1	1	1	1	1	1	1	1

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2	2	3

/mismatch in the diagonal),
sequence #1),
sequence #2)]

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G		1									
A			1								
T				2	2						
C						3					
G							4	4			
A									5	5	5
											6

G _ A A T T C A G T T A
| | | | | | |
G G _ A _ T C _ G _ _ A

Famous Bioinformatics Tools for Alignments

- Global alignment : ClustalW, T-coffee, and MUSCLE
- Local alignment : FASTA, and BLAST (Basic Local Alignment Searching Tools) provided by NCBI.
- Pair-wise alignment : BLAST and FASTA
- Multiple sequence alignment : ClustalW, T-coffee, MUSCLE, and etc.

The screenshot displays the EMBL-EBI ClustalW2 web interface. The top navigation bar includes links for Databases, Tools, Research, Training, Industry, About Us, and Help. The main content area is titled "ClustalW2 - Multiple Sequence Alignment" and provides a description of the tool. It guides users through three steps: 1. Enter your input sequences, 2. Set your Pairwise Alignment Options, and 3. Set your Multiple Sequence Alignment Options. Each step includes a "More options..." link for detailed configuration. The left sidebar contains various utility links such as Help, FAQ, Jalview, and related applications.

Example of Genome Assembly: *Vitis Vinifera*

Vol 449 | 27 September 2007 | doi:10.1038/nature06148

nature

LETTERS

The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla

The French–Italian Public Consortium for Grapevine Genome Characterization*

A total of 6.2 million end-reads were produced by our consortium, representing an 8.4-fold coverage of the genome. Within the assembly, performed with Arachne¹², 316 supercontigs represent allelic haplotypes that constitute 11.6 million bases (Mb)

Pair-wise comparison of 6,200,000 reads

$$6,200,000 \text{C}_2 = 19,219,996,900,000 \text{ comparisons}$$

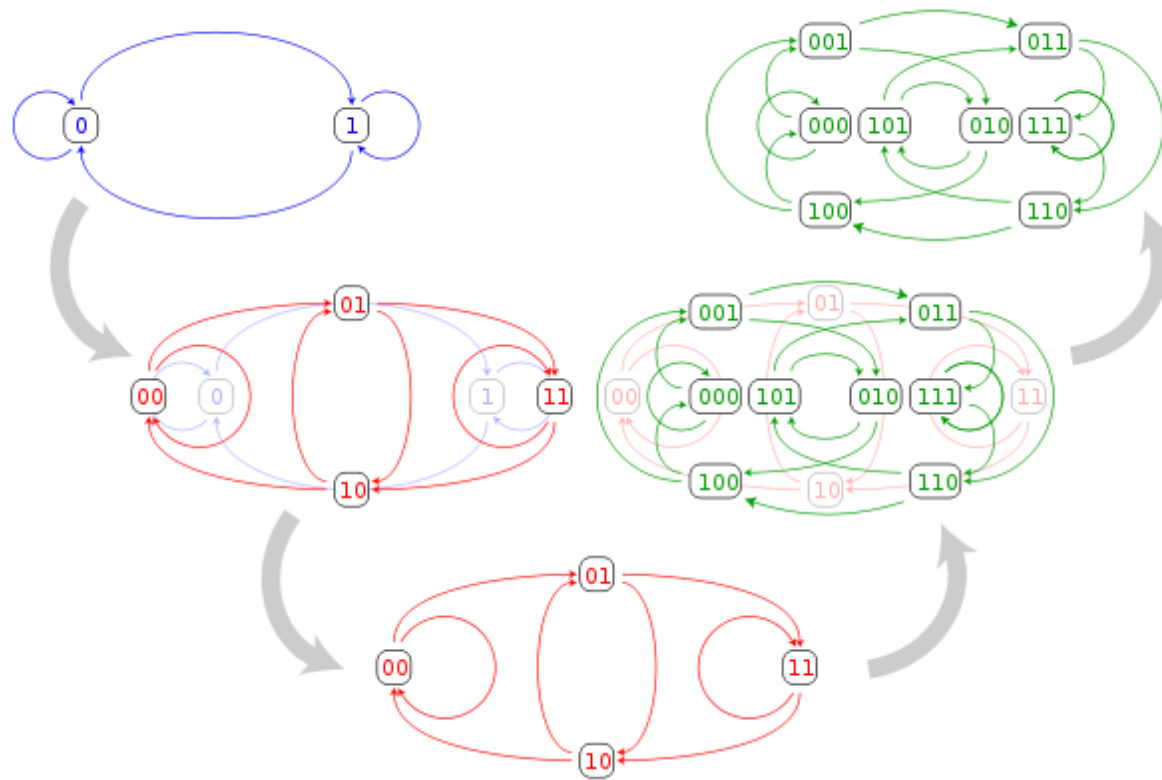


De brujin Graph: Alternative Method For Alignment

$$S := \{s_1, \dots, s_m\}$$

$$V = S^m = \{(s_1, \dots, s_1, s_1), (s_1, \dots, s_1, s_2), \dots, (s_1, \dots, s_1, s_m), (s_1, \dots, s_2, s_1), \dots, (s_m, \dots, s_m, s_m)\}.$$

$$E = \{((v_1, v_2, \dots, v_n), (w_1, w_2, \dots, w_n)) : v_2 = w_1, v_3 = w_2, \dots, v_n = w_{n-1}\}.$$





De brujin Graph Algorithm For Alignment (1)

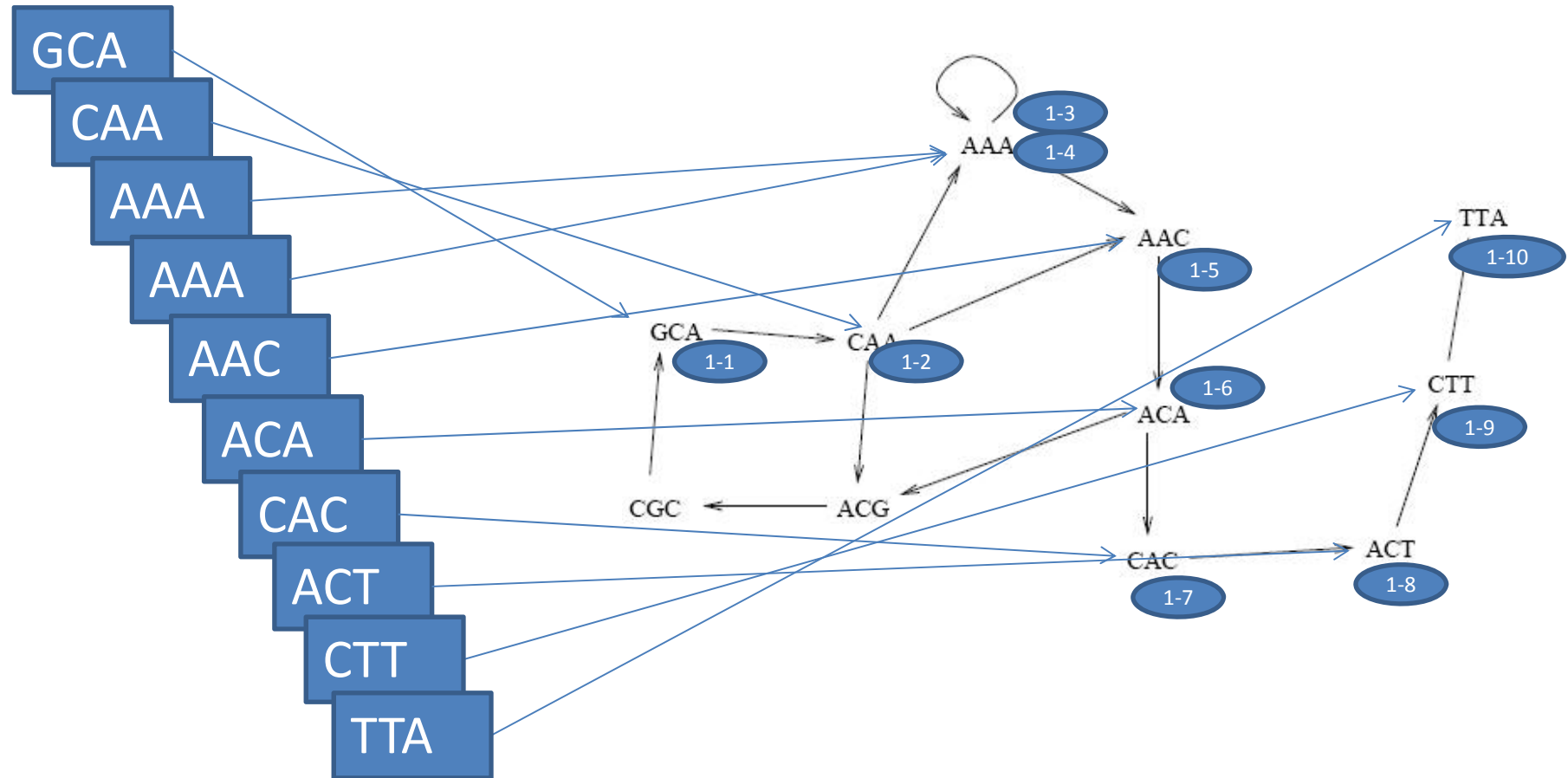


- This algorithm has been utilized for finding overlapped short-read sequences quickly.
- This algorithm consists of three parts:
 - i) Generating k-mer sequences**
 - ii) Constructing de brujin graph**
 - iii) Resolving the graph with generating sequences**

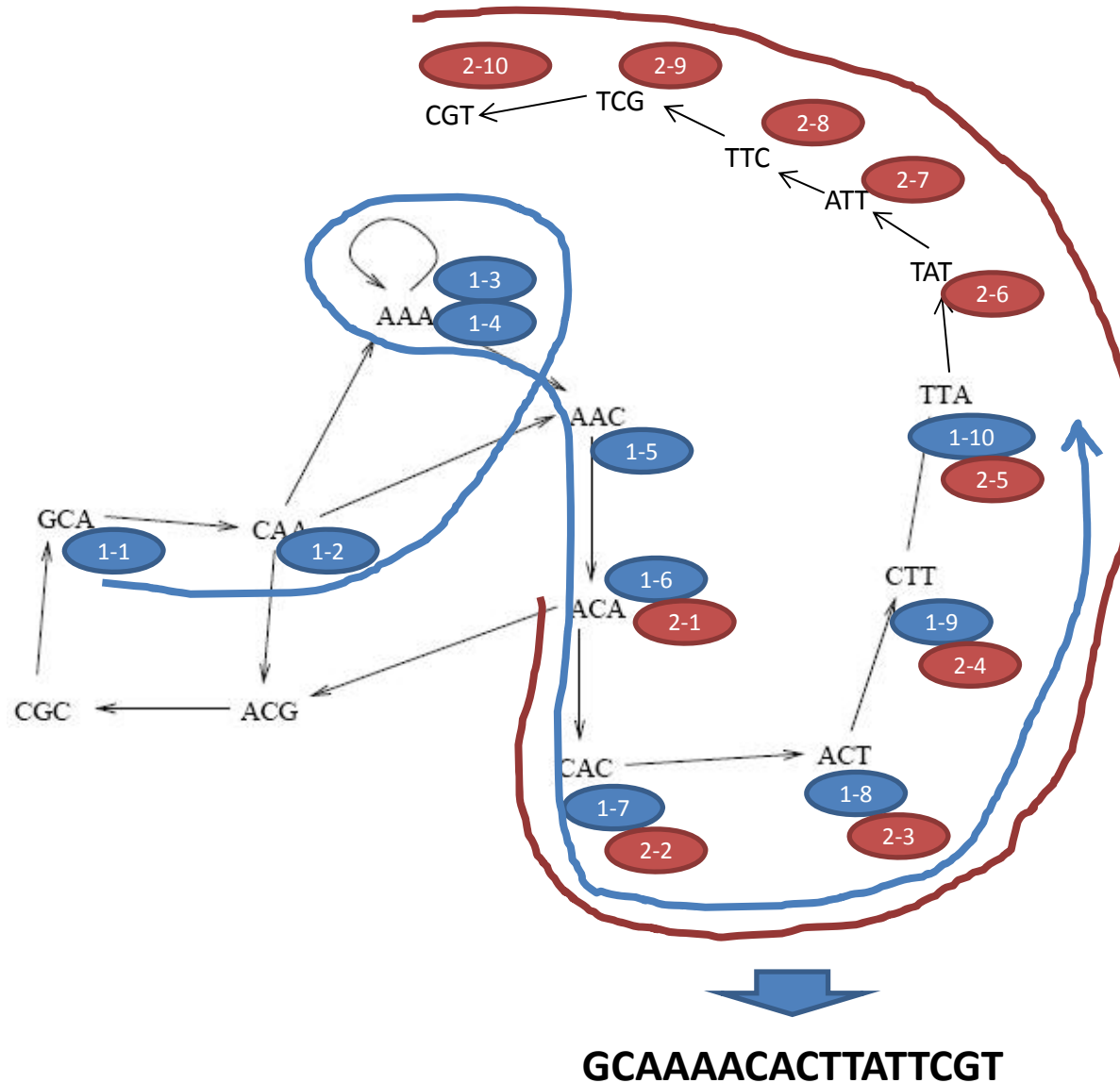
De brujin Graph Algorithm For Alignment (2)

GCAAAACACTTA...

K = 3



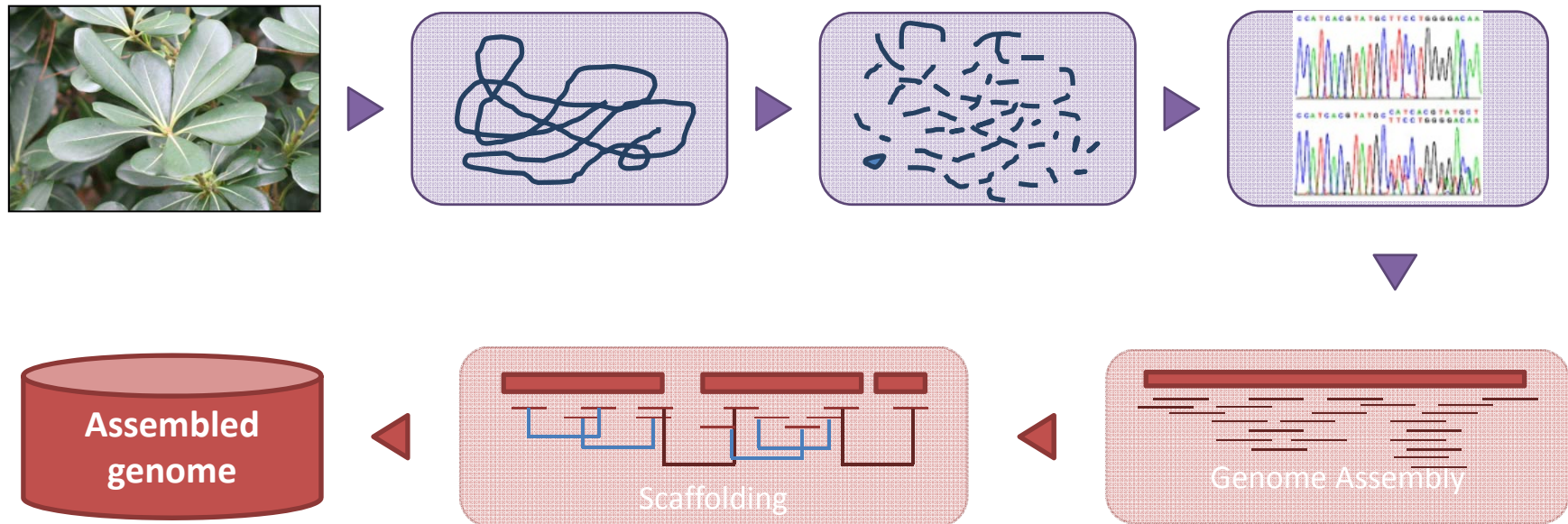
De brujin Graph Algorithm For Alignment (4)



How to Assembly Chloroplast Sequences

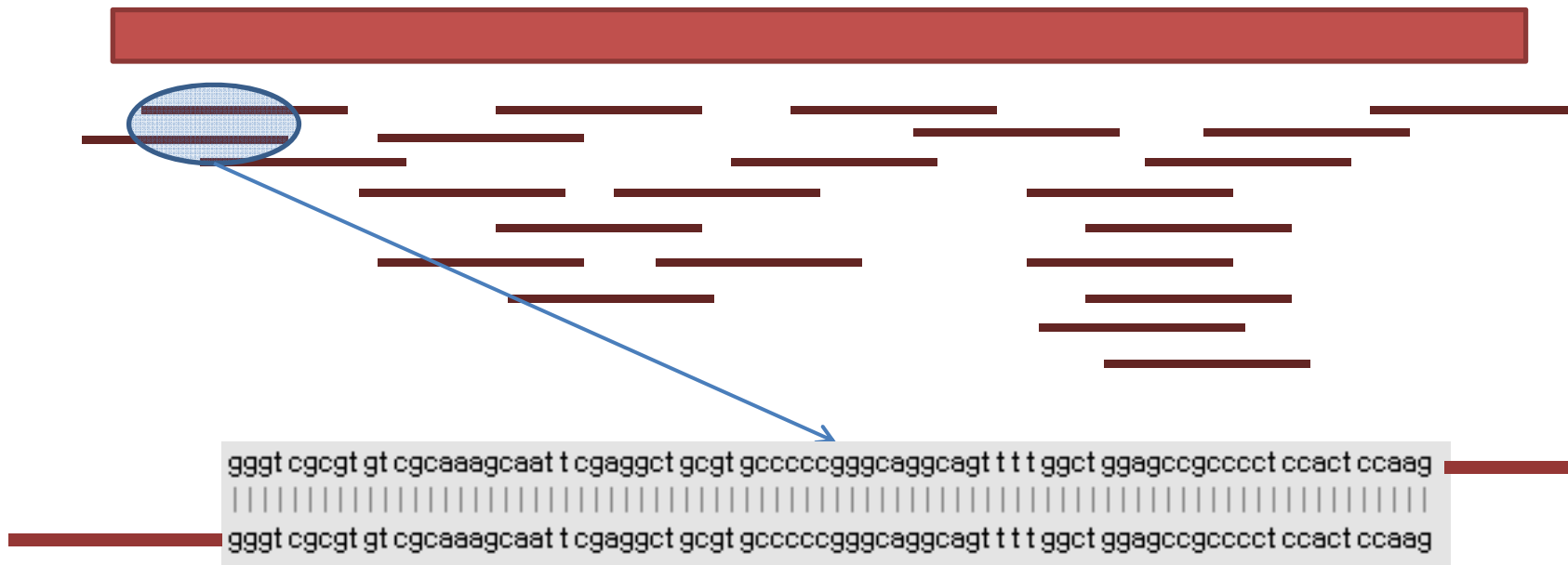
Revisit: Whole Shotgun Sequence Strategy

- Assembly process is essential for genome project because read length of each sequence is less than 1 kb.
- Assembly process was conducted by several popular programs, such as phrap and PCAP3, for Sanger sequences.



Revisit: Genome Assembly Process

- We can perform genome assembly manually!



23 sequences should be **compared** with each other!

$${}^{23}C_2 = 23 * 22 / 2 = 253 \text{ comparison!}$$

Genome Assemblers For NGS Sequences (1)

Resource

Velvet: Algorithms for de novo short read assembly using de Bruijn graphs

Daniel R. Zerbino and Ewan Birney¹

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

We have developed a new set of algorithms, collectively called “Velvet,” to manipulate de Bruijn graphs for genomic sequence assembly. A de Bruijn graph is a compact representation based on short words (k -mers) that is ideal for high coverage, very short read (25–50 bp) data sets. Applying Velvet to very short reads and paired-ends information only, one can produce contigs of significant length, up to 50-kb N50 length in simulations of prokaryotic data and 3-kb N50 on simulated mammalian BACs. When applied to real Solexa data sets without read pairs, Velvet generated contigs of ~8 kb in a prokaryote and 2 kb in a mammalian BAC, in close agreement with our simulated results without read-pair information. Velvet represents a new approach to assembly that can leverage very short reads in combination with read pairs to produce useful assemblies.

[Supplemental material is available online at www.genome.org. The code for Velvet is freely available, under the GNU Public License, at <http://www.ebi.ac.uk/~zerbino/velvet>.]

Velvet Assembler (1)

```
lecture@AMBORELLA:~/M_kobus/1
[lecture@AMBORELLA 1]$ ~/exeman/velvet_1.1.03/velveth k31 31 -shortPaired -fastq 90000.0-1.fastq 90000.0-2.fastq
[0.000000] Reading FastQ file 90000.0-1.fastq
[0.193801] 90064 reads found.
[0.193810] Done
[0.193821] Reading FastQ file 90000.0-2.fastq
[0.388996] 90053 reads found.
[0.389005] Done
[0.389037] Reading read set file k31/Sequences;
[0.427236] 180117 sequences found
[0.595392] Done
[0.595403] 180117 sequences in total.
[0.595502] Writing into roadmap file k31/Roadmaps...
[0.673878] Inputting sequences...
[0.673890] Inputting sequence 0 / 180117
[3.079403] === Sequences loaded in 2.405528 s
[3.079434] Done inputting sequences
[3.079438] Destroying splay table
[3.091199] Splay table destroyed
[lecture@AMBORELLA 1]$
```

```
~ex lecture@AMBORELLA:~/M_kobus/1
[0.595392] Done
[0.595403] 180117 sequences in total.
[0.595502] Writing into roadmap file k31/Roadmaps...
[0.673878] Inputting sequences...
[0.673890] Inputting sequence 0 / 180117
[3.079403] === Sequences loaded in 2.405528 s
[3.079434] Done inputting sequences
[3.079438] Destroying splay table
[3.091199] Splay table destroyed
[lecture@AMBORELLA 1]$ ~/exeman/velvet_1.1.03/velvetg k31 -ins_length 500 -cov_cutoff auto -exp_cov auto
[0.000000] Reading roadmap file k31/Roadmaps
[0.288275] 180117 roadmaps read
[0.288487] Creating insertion markers
[0.309452] Ordering insertion markers
[0.470370] Counting preNodes
[0.487079] 147937 preNodes counted, creating them now
[1.053024] Adjusting marker info...
[1.071163] Connecting preNodes
[lecture@AMBORELLA 1]$
```

~/exeman/velvet_1.1.03/velvetg k31 -ins_length 500 -cov_cutoff auto -exp_cov auto

Velvet Assembler (2)



```
lecture@AMBORELLA:~/M_kobus/1/k31
>NODE_1_length_1569_cov_299.312927
AAGTGCTTTTCTTGCGTATTCTTGATGAACCAGCGTTTATATATAGATGTGGGAGGATTT
GTTTGGGAAGTAATAAGCCCCTTTGACATCTCTTCATCTGCAAGAATTCTCGACGTGAA
AACACAGAGACAAAGGGCTGATCTTTGAATAGGAAAAAGAATGGATCTGCAGGGTCCCAA
ACGAATTGGCTTATTCGAAAAAGCCTTGTTCTTTGGAAGATCTATCTCGTGTCTGGTAC
TGCATGGTTCCTACTCTGCAAGAACTCCGAATCATTCTCTTGAAGCTCATCCTCTTTATGA
TAAATGATCCGCTTGCCCCGAAATGACCTGGCCCCAATAGGGAAATCCCAATTCATTGGTC
CTTTCGATACAATCAAATAGATTGCCACAAGGGCGCCATATTCTAGGAGCCCAACTATG
TGATTGAATAAATCCTCCTCTATCTGTTGCGGGTTCGAGGGCTCCTTCTCCTTCCCCTTCT
TCAAACCTCCGATTTCGTATTTTTTCATATAGAAATCTCTGATCAACGATAGAACAAGATCCA
TTTTGCATCATATCTAACTGATTCTTGTTTCGGACCGAAGAAGCAATGTCACTCGATCA
TTATCAAACCTGACTGCAATCTTTTTCTGTCCGTGAGGATCCCACCAGAGCGCCTTCTACT
TCTAATTCTAATAGGCCATGAAGTAGATCAGAATCATTCTCAACGAATCCATAAGAAGGG
ATCCAATTTTTTTCATCGGGTCCGGGTGGAGACCAAGATCTTGAGCGACCAATCCGGCA
GAACAACCTCAAAGATAAAGAAGTATCGTTAATTTCTTCATGCTCGTTCCAAGTTCGAAG
TACCATTTGTACAAATAAGAATCCCCTTCCTTACATGATTTCTTCTTCATATAGATAGAT
ATAGGATCTATGGGGCAATTACTTAGAAGTACATTTTGTGCAACAGCCCTTCCTATCTGA
TAGAAAAGGATCCCATGATCCTGAACCGATCTTACCTGGGATCGCAAATCCCAAGTTTGT
CTATGAAGAGCTGATCTAATTGTATTAGTGTCTATAATGGATTTCTTCTGTGTAATACTA
ATTGATAGGGCCTCATTGGTAAGTGCTACAAGATCTCGTGCAATTGGAACCCATGGTCATG
GACCCGAATCCGTTAGTATGGAACATTTTCTTTTCCAAGTGAAATCCCCTAGTATATGAA
AGAATGAAAAAGTGCTTTTCGTTGTTGTGGAATAAGAAGCCTTCGTATCTTAATGCATGTA
TTTAATTTATTTCGGAGCTATTAGAGCGGGATCCACTTTTTTGGGAATATGAGTCGAAGCA
ATAACAAGAATATTTCTAGTGGAACATCTTTCACAATCTCTGGAGAGATAGTTCATAAT
AGACCGAGGGGATAAGTAATTCGACTCATTACATACAGATCATGAATGTTTGGAATCCAT
ATTATGCAAGGGGACATTGCTTTTGCTAATTCTAATTGAGGGGTGATATCAAATCGGTCT
ATTTTCGGCGTCATATACATAGTTAGCACATTCGTCATAGTTAGCAGCTCCGTATCAAGG
TCATCATCAATATCGTCACTATCATCAATATCGATATCG
>NODE_4_length_2412_cov_363.436157
TCGTCACTATCATCAATATCGATATCGTCAATAAGATAACCTTTAGGCTTGTCATCCAGG
```

1,1

Top

Velvet Assembler (3)



```
lecture@AMBORELLA:~/M_kobus/1/k31
drwxrwxr-x. 2 lecture lecture 4096 May 11 02:20 k31
[lecture@AMBORELLA 1]$ cd k31/
[lecture@AMBORELLA k31]$ ls -al
total 54276
drwxrwxr-x. 2 lecture lecture 4096 May 11 02:20 .
drwxrwxr-x. 3 lecture lecture 4096 May 11 02:20 ..
-rw-rw-r--. 1 lecture lecture 125181 May 11 02:20 contigs.fa
-rw-rw-r--. 1 lecture lecture 6935269 May 11 02:20 Graph2
-rw-rw-r--. 1 lecture lecture 2691126 May 11 02:20 LastGraph
-rw-rw-r--. 1 lecture lecture 360 May 11 02:20 Log
-rw-rw-r--. 1 lecture lecture 2300235 May 11 02:20 PreGraph
-rw-rw-r--. 1 lecture lecture 9478851 May 11 02:20 Roadmaps
-rw-rw-r--. 1 lecture lecture 34015339 May 11 02:20 Sequences
-rw-rw-r--. 1 lecture lecture 1855 May 11 02:20 stats.txt
[lecture@AMBORELLA k31]$ perl ~/stat_fasta.pl contigs.fa
# of contigs : 20 ea
Total length : 122429 bp
Average length : 6121.45 bp
N20 : 40329
N50 : 15323
N90 : 3928
[lecture@AMBORELLA k31]$
```

Statistics of assembled sequences

N50

- Sort all contigs from largest to smallest
- Determining the minimum set of contigs whose sizes total 50% of the entire genome
- The N50 is distinct from the mean and the median!
- 3 3 4 6 7 8 8 9 9 9 10 11 13 25
- **Mean** = $125/14 = 8.93$ (sum=125)
- **Median** = $(3+25)/2 = 14$
- $125/2 = 62.5$ sum of contig lengths reach to 62.5
(from the largest to the smallest)
- $25+13+11+10=59$
- $25+13+11+10+9=68$
- **Therefore, N50 = 9**

N50 계산의 예

Contig 번호	각각의 contig 길이	누적값		
1	33407	33407		
7	15243	48650		
6	14172	62822	52860.5	N50= 14172
8	9250	72072		
2	8275	80347		
10	5714	86061		
3	5406	91467		
13	5227	96694		
4	3683	100377		
11	2849	103226		
9	1251	104477		
14	663	105140		
5	442	105582		
12	136	105718		
16	2	105720		
15	1	105721		
합계	105721			
합계 /2	52860.5			