NEXT GENERATION SEQUENCING (NGS)

Andreas Gisel

International Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

a.gisel@cgiar.org

□ Short history bio-sequencing
□ NGS Sequencing technologies
□ NGS Sequence data
□ NGS Applications
□ NGS Data Analysis

OVERVIEW

□ First fully sequenced bio-sequence

HISTORY

- First fully sequenced bio-sequence
  - amino acid of insulin (51aa) 1955

HISTORY

□ First fully sequenced bio-sequence

- amini acid of insulin (51aa) 1955

□ First fully sequence nucleic acid

- tRNA (75nt) 1965

□ First DNA

- Bacteriophage (5375nt) 1977

□ DNA sequencing

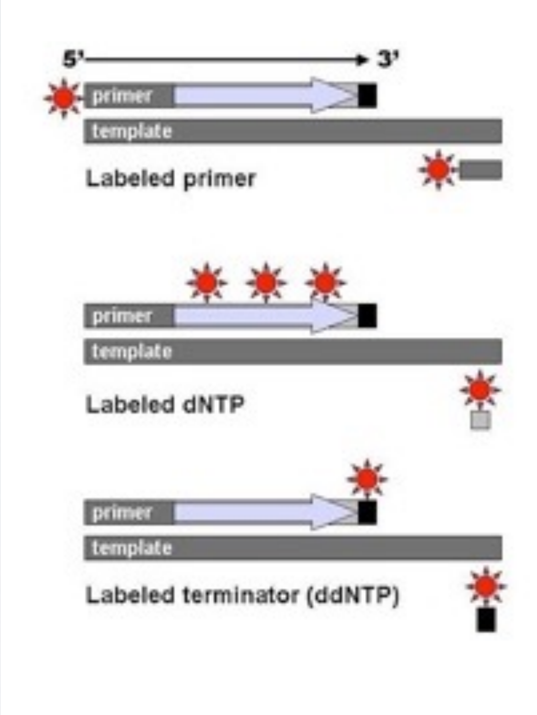
- Sanger sequencing technology (1975)

HISTORY

- sequencing by chain-termination method

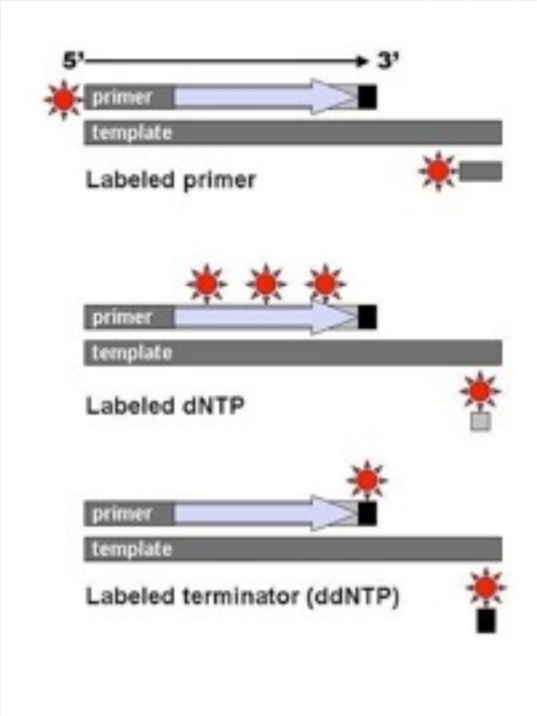
SANGER TECHNOLOGY

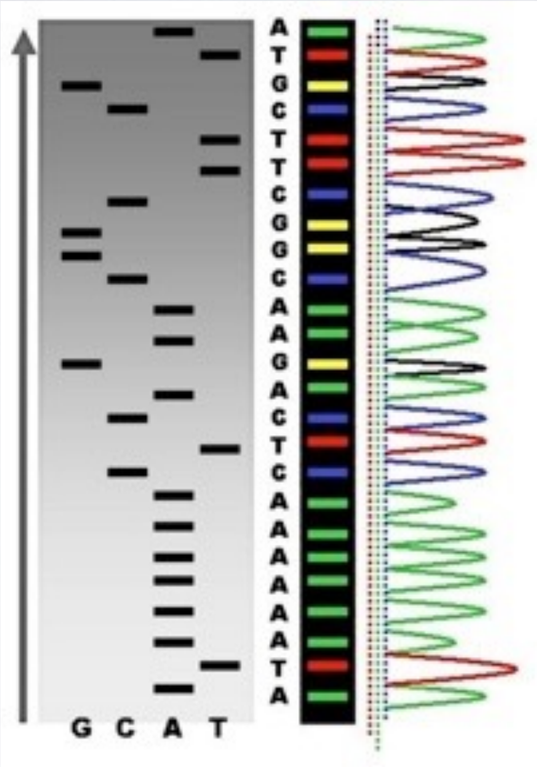
- sequencing by chain-termination method



SANGER TECHNOLOGY

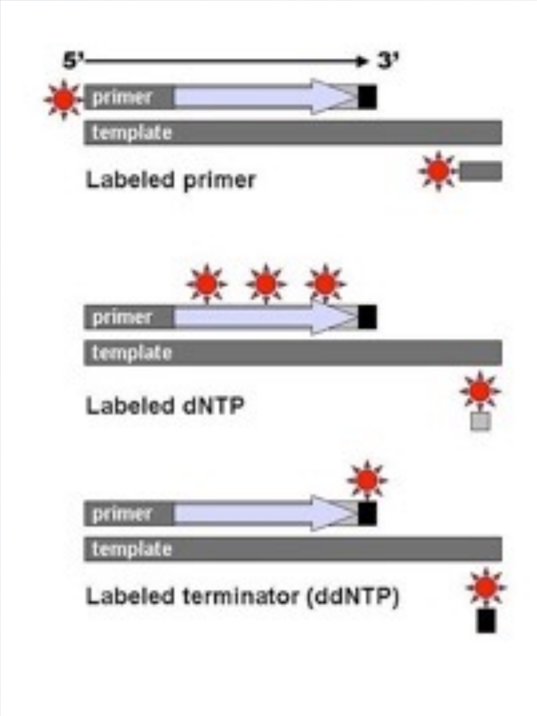
- sequencing by chain-termination method

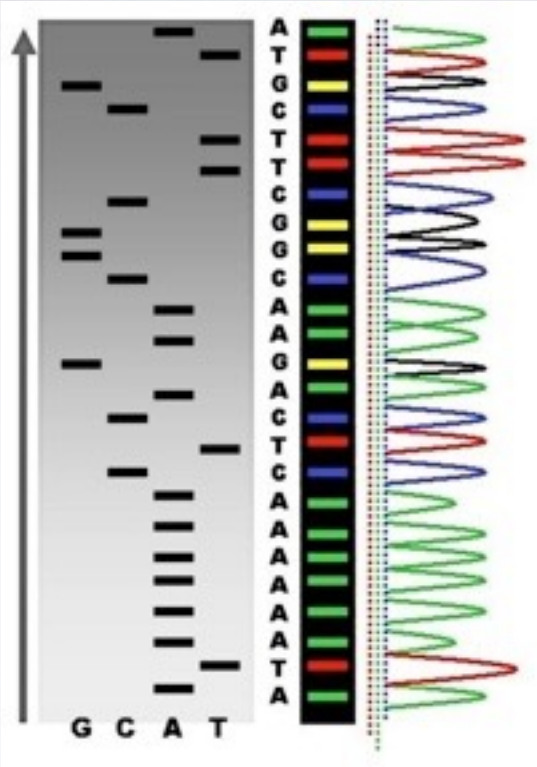


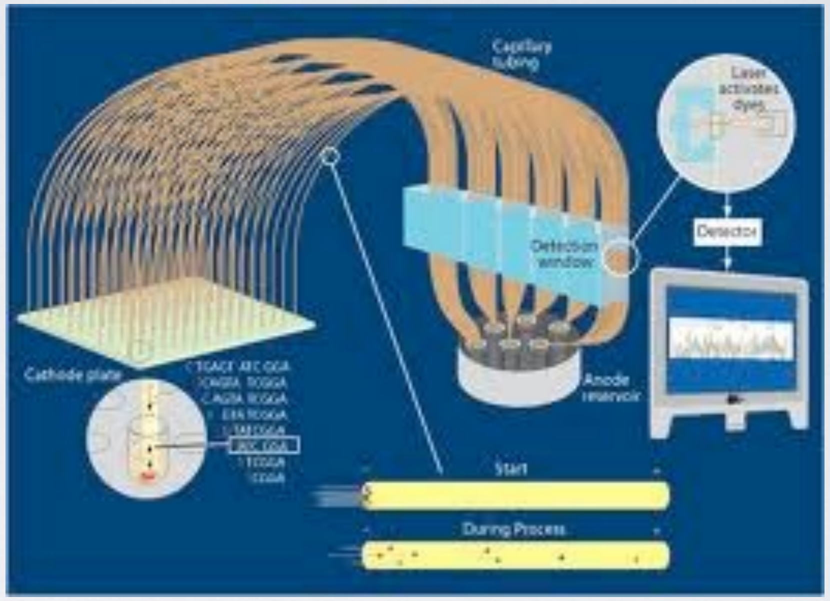


SANGER TECHNOLOGY

- sequencing by chain-termination method

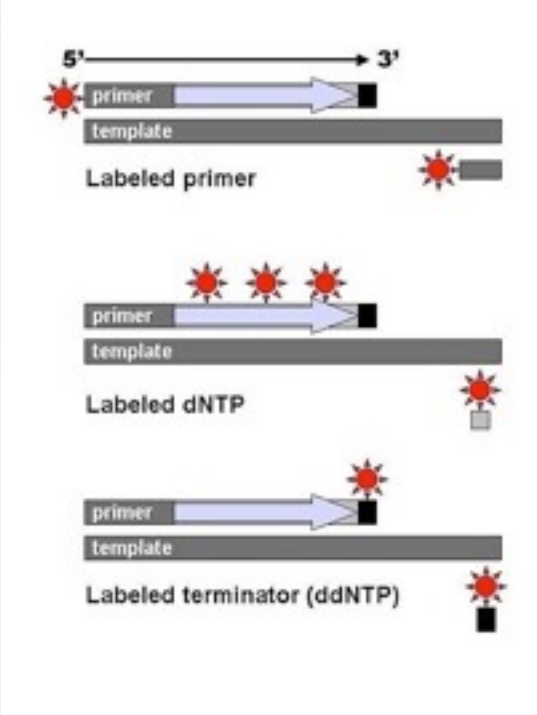


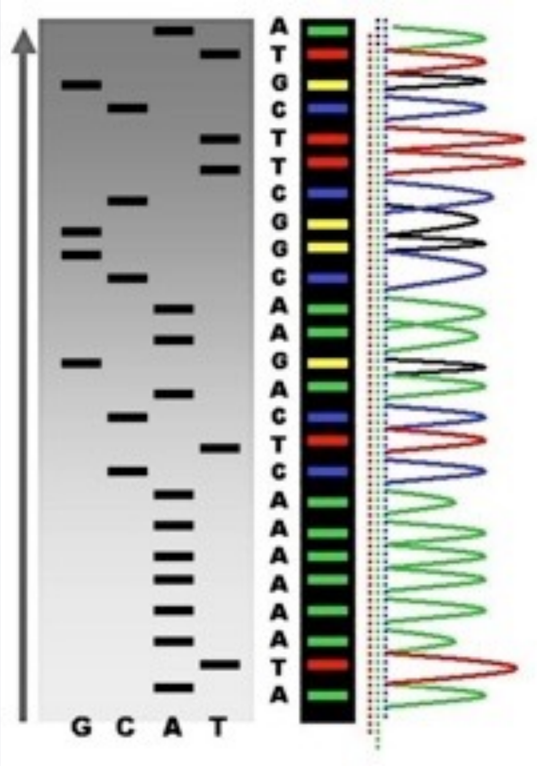


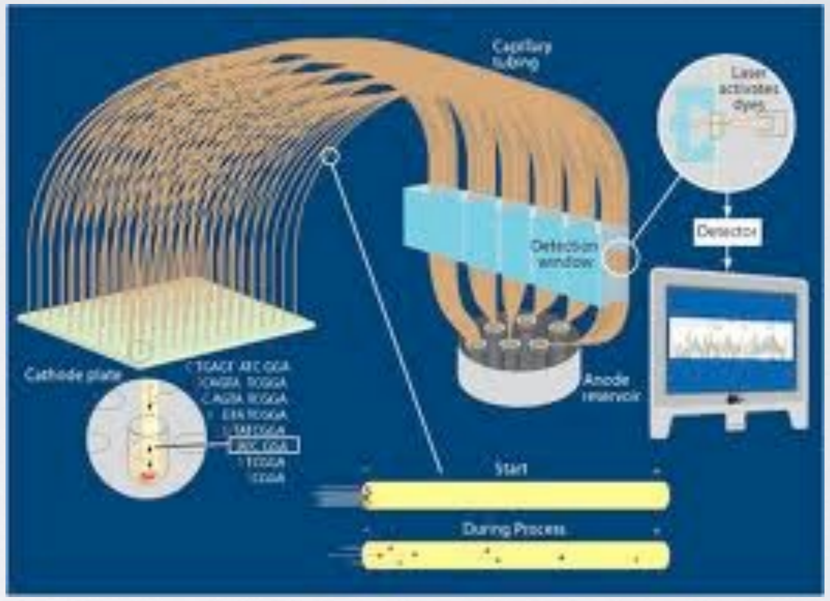


SANGER TECHNOLOGY

- sequencing by chain-termination method

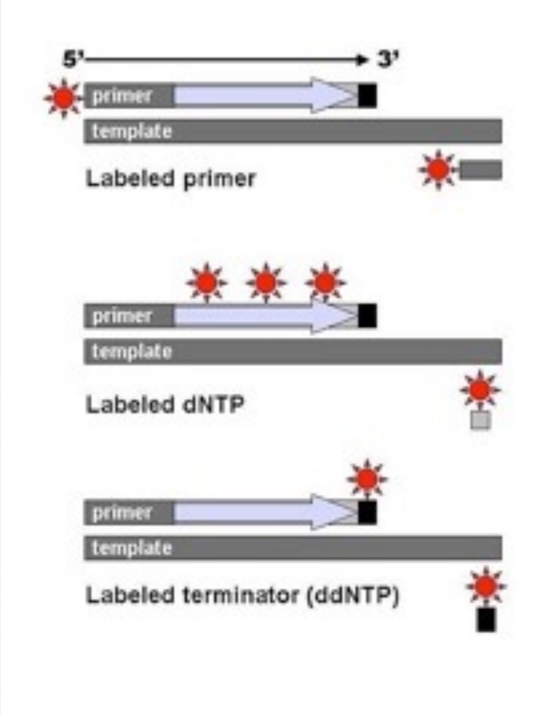


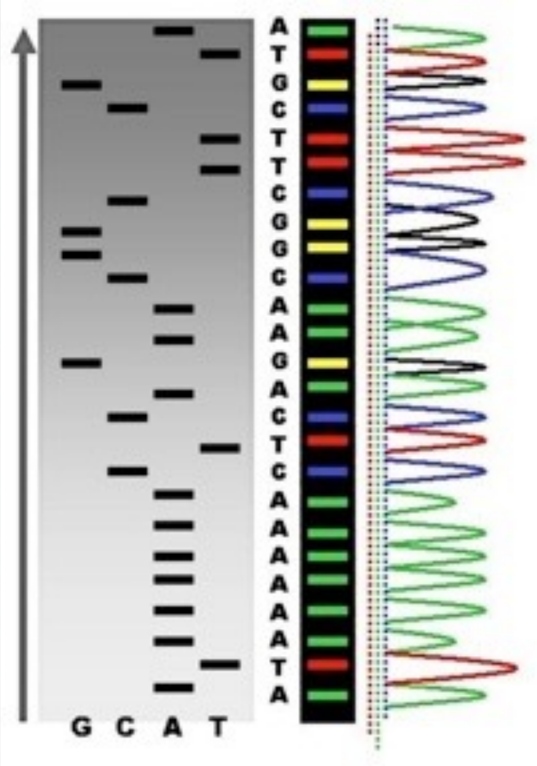


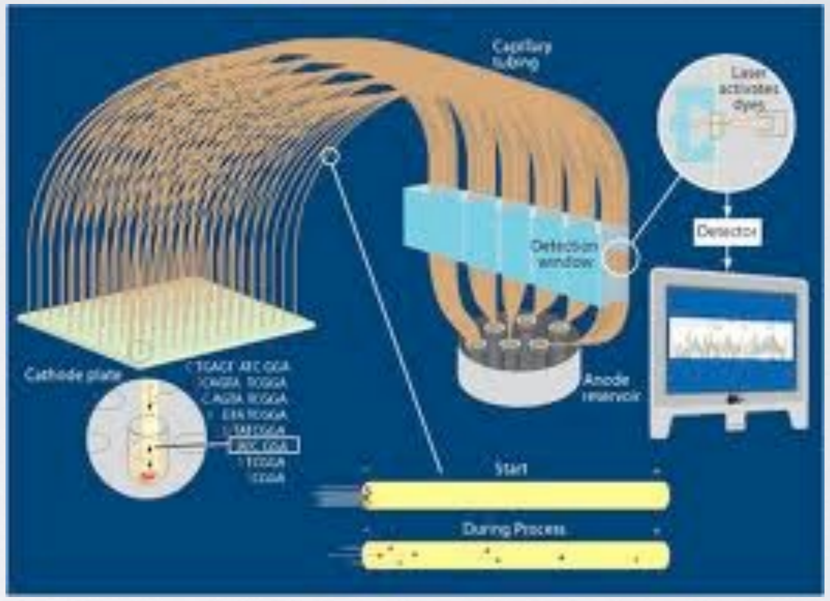


SANGER TECHNOLOGY

- sequencing by chain-termination method







- DNA sequencing by capillary electrophoresis
- 384 reactions in parallel
- sequences up to 1000nt

SANGER TECHNOLOGY

NEXT GENERATION SEQUENCING

NEXT GENERATION SEQUENCING

□ Sequencing by synthesis

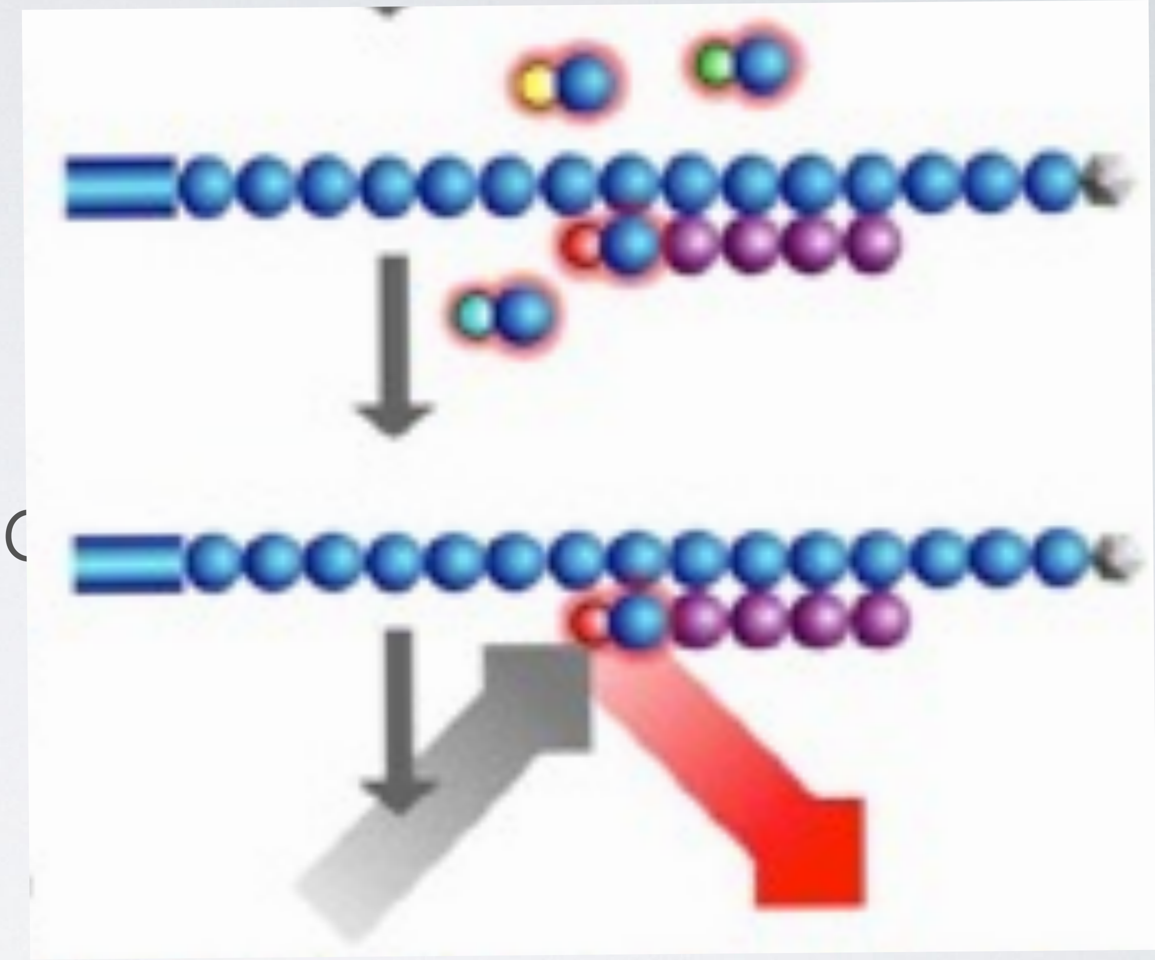
□ highly parallelized sequencing

□ Paired-end sequencing

□ Sequencing by synthesis

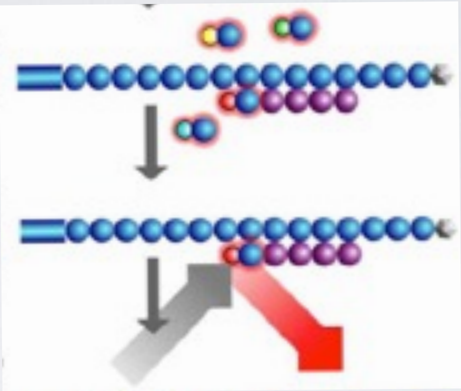
□ highly parallelized sequenc

□ Paired-end sequencing



NEXT GENERATION SEQUENCING

□ Sequencing by synthesis

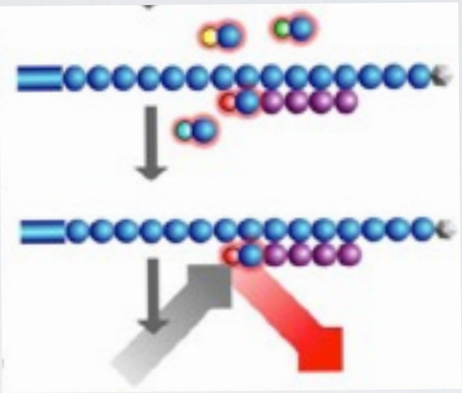


□ highly parallelized sequencing

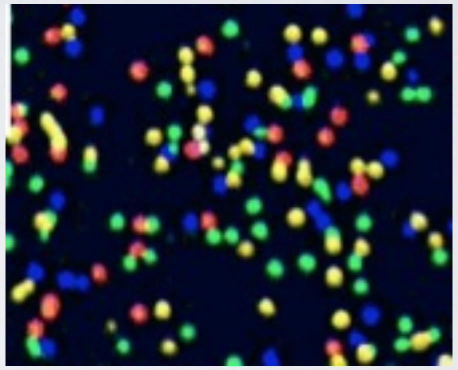
□ Paired-end sequencing

NEXT GENERATION SEQUENCING

- □ Sequencing by synthesis



- □ highly parallelized sequencing

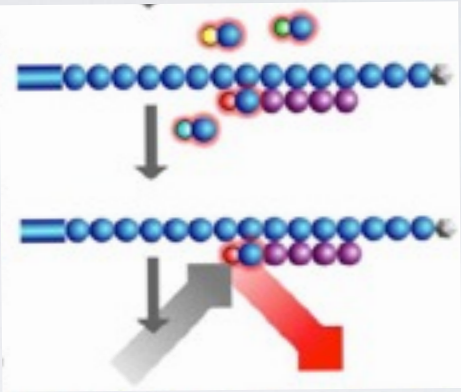


- □ Paired-end sequencing

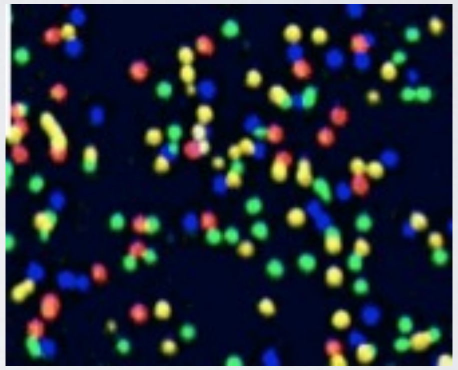
NEXT GENERATION SEQUENCING

NEXT GENERATION SEQUENCING

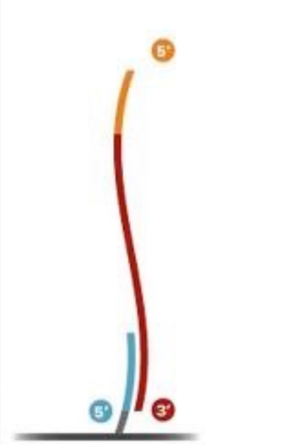
- □ Sequencing by synthesis



- □ highly parallelized sequencing



- □ Paired-end sequencing



TECHNOLOGIES

Amplification steps

□ 454 Roche

□ Solexa Illumina

□ SOLiD Applied Biosystems

Single molecule

□ Pacific BioSciences

□ Ion Torrent

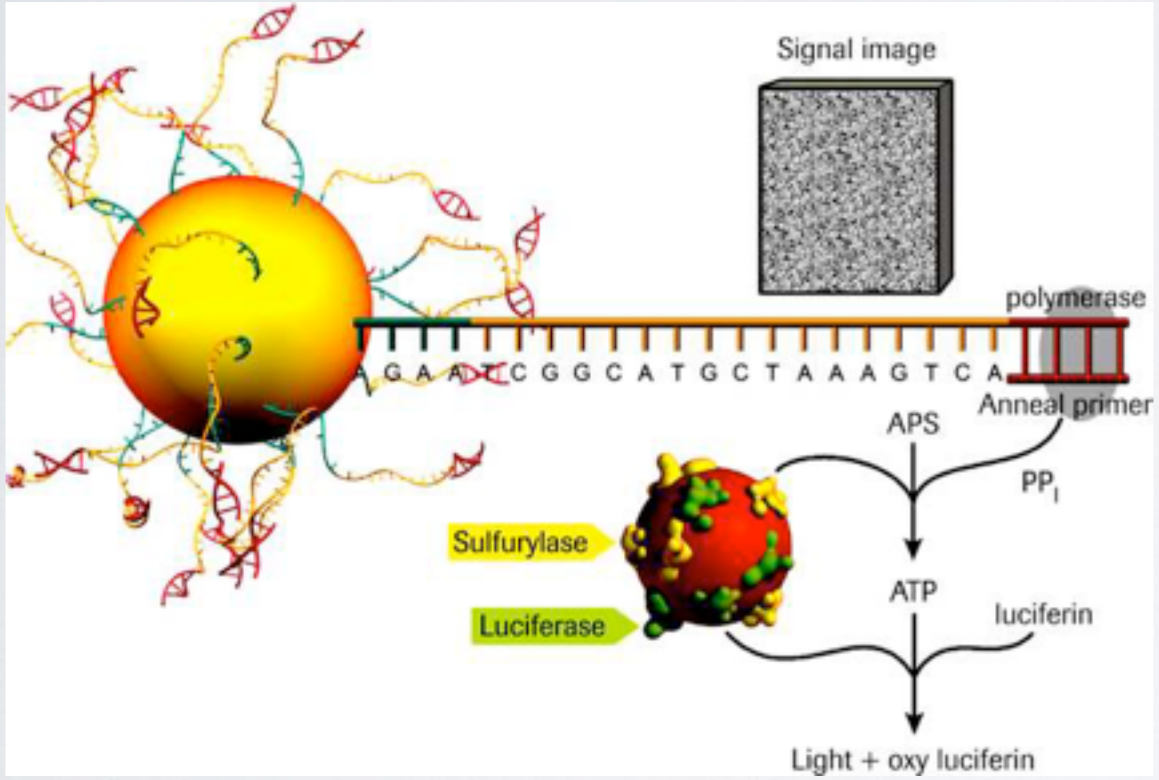
□ Nanopore

454 ROCHE

□ DNA template immobilized to nano-beads
□ Emulsion PCR
□ Pyro-Sequencing in nano wells (1.6M reads)
□ Sequencing by synthesis

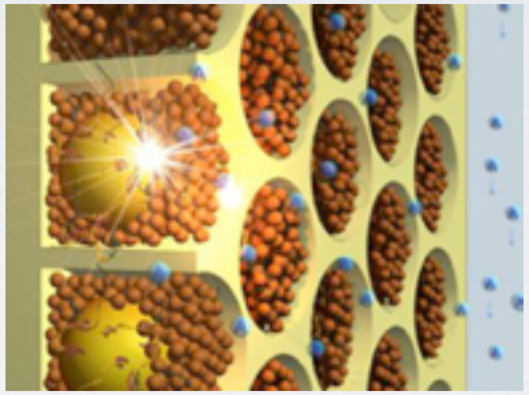
454 ROCHE

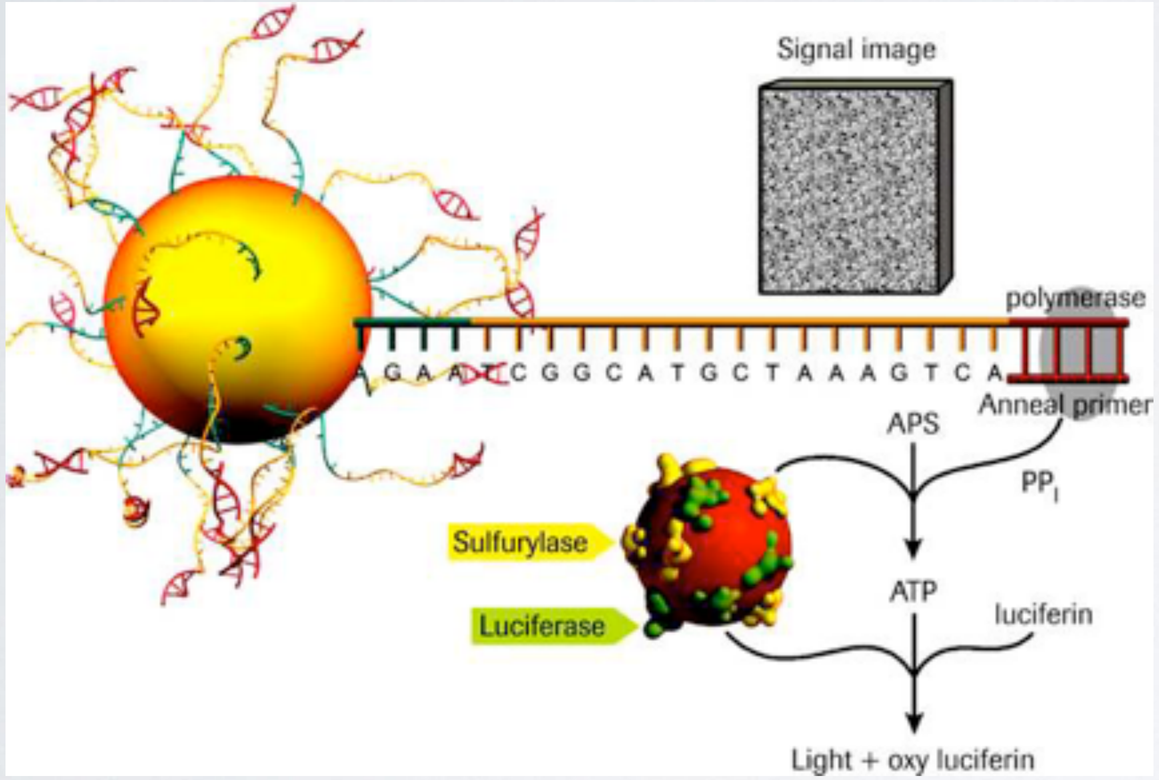
- DNA template immobilized to nano-beads
- Emulsion PCR
- Pyro-Sequencing in nano wells (1.6M reads)
- Sequencing by synthesis



454 ROCHE

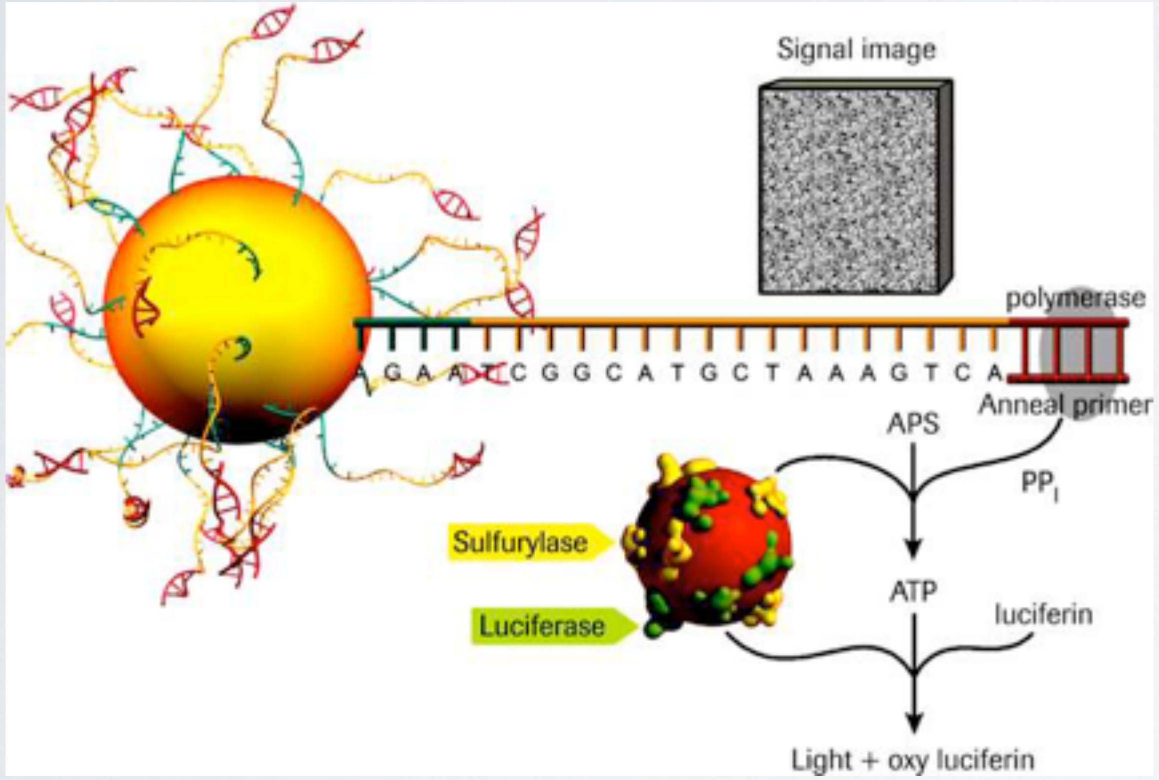
- DNA template immobilized to nano-beads
- Emulsion PCR
- Pyro-Sequencing in nano wells (1.6M reads)
- Sequencing by synthesis

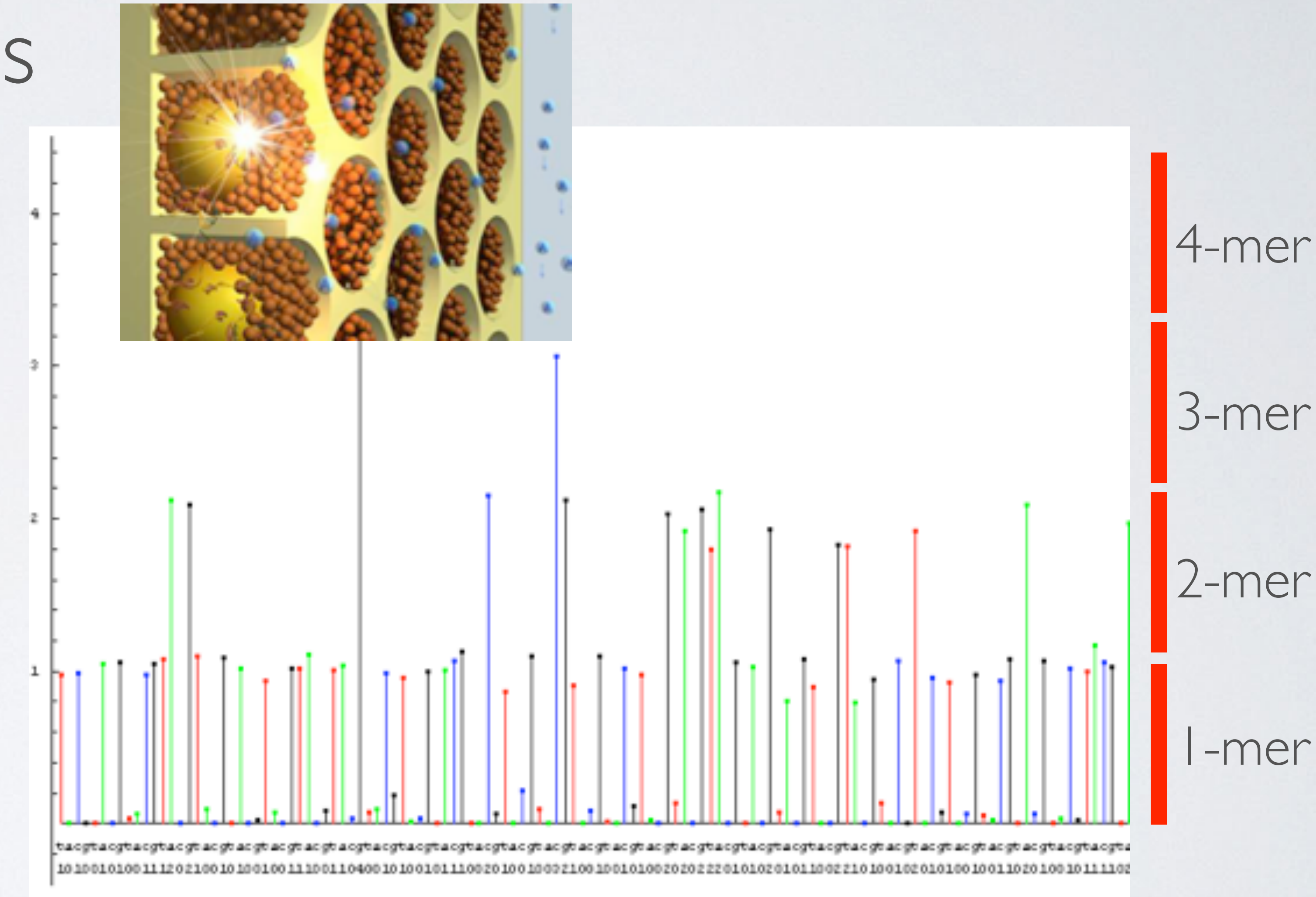




454 ROCHE

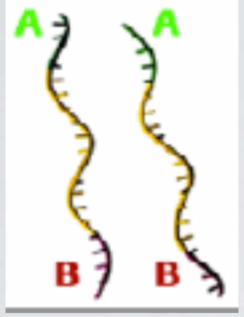
- □ DNA template immobilized to nano-beads
- □ Emulsion PCR
- □ Pyro-Sequencing in nano wells (1.6M reads)
- □ Sequencing by synthesis



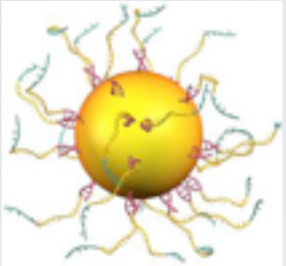


454 ROCHE

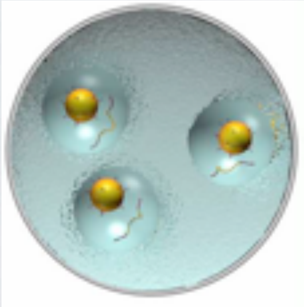
Sequence fragmentation



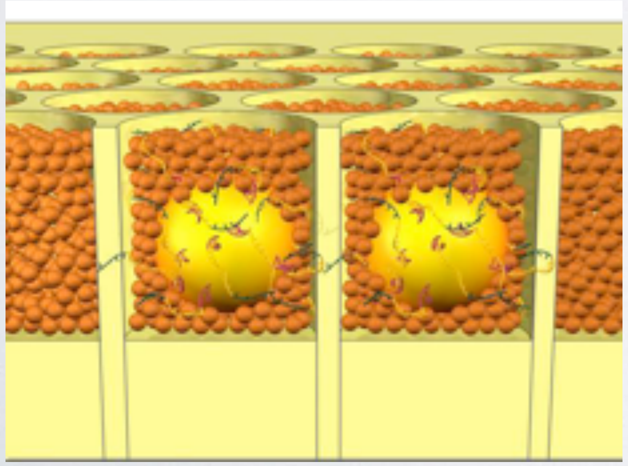
Ligation of adaptors



Sequence immobilization



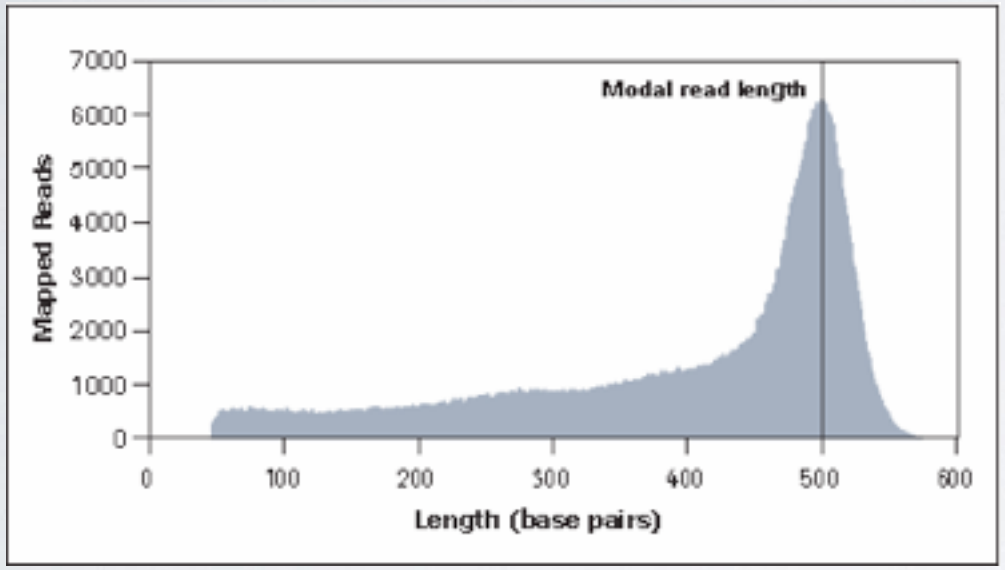
emulsion PCR

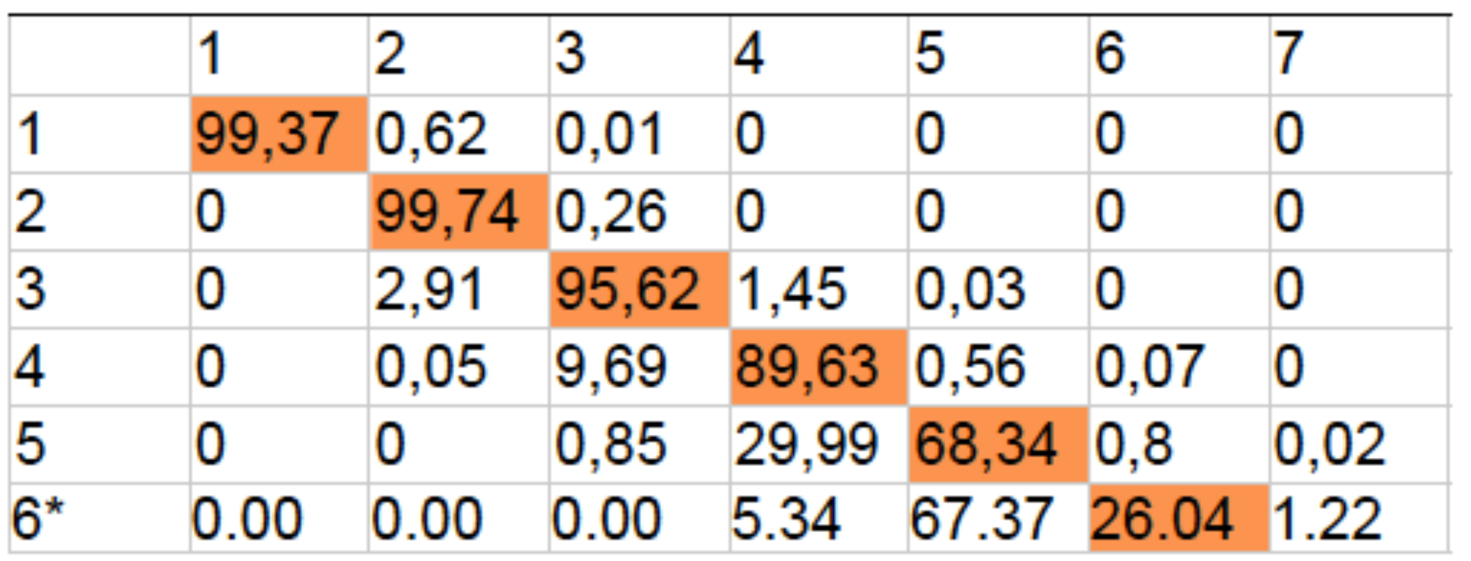


distribution in nano wells

454 ROCHE

- sequencing length up to 1000nt (800nt)
- up to 1.2M reads
- 600 - 800Mb per run
- problems with homo polymers





	1	2	3	4	5	6	7
1	99,37	0,62	0,01	0	0	0	0
2	0	99,74	0,26	0	0	0	0
3	0	2,91	95,62	1,45	0,03	0	0
4	0	0,05	9,69	89,63	0,56	0,07	0
5	0	0	0,85	29,99	68,34	0,8	0,02
6*	0.00	0.00	0.00	5.34	67.37	26.04	1.22

Vicarico et al pers comm.

454 ROCHE

ILLUMINA

□ DNA template immobilized to a flow cell

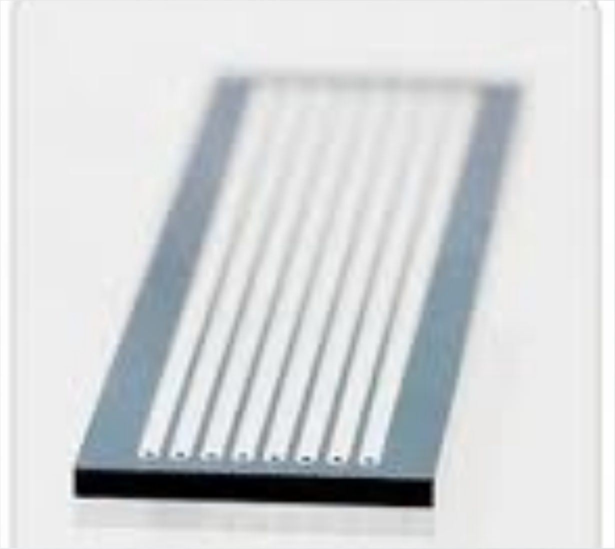
□ Cluster formation by Bridge PCR

□ Sequencing on flow cell (3000M reads)

□ Sequencing by synthesis (protected nts)

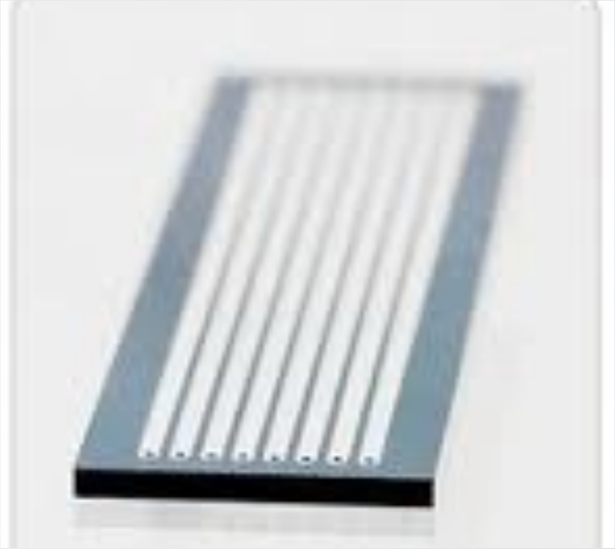
ILLUMINA

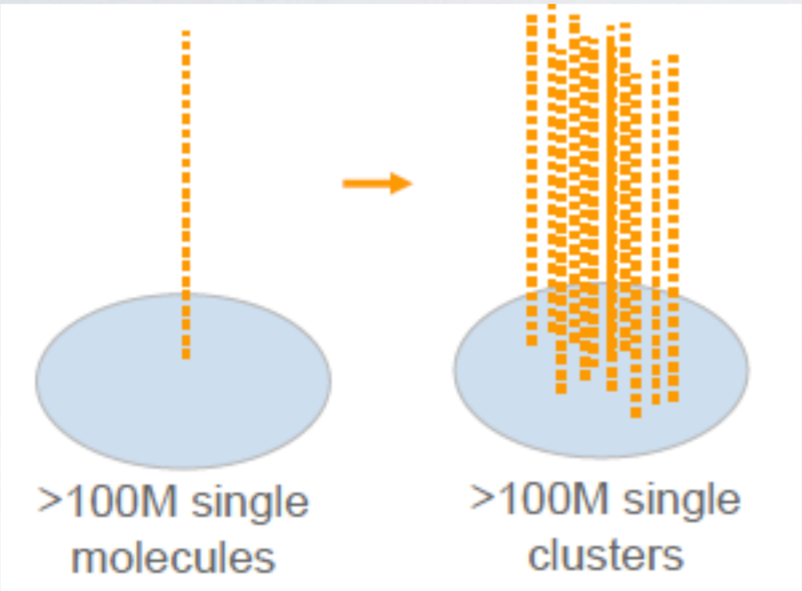
- □ DNA template immobilized to a flow cell
- □ Cluster formation by Bridge PCR
- □ Sequencing on flow cell (3000M reads)
- □ Sequencing by synthesis (protected nts)



ILLUMINA

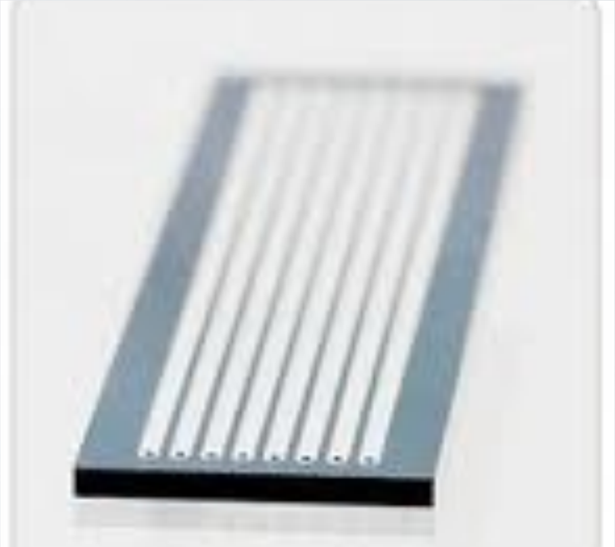
- □ DNA template immobilized to a flow cell
- □ Cluster formation by Bridge PCR
- □ Sequencing on flow cell (3000M reads)
- □ Sequencing by synthesis (protected nts)

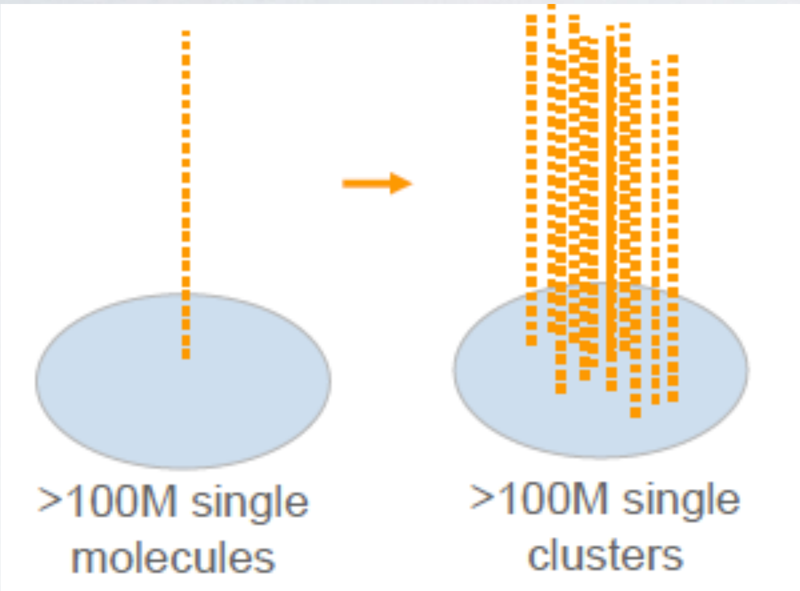


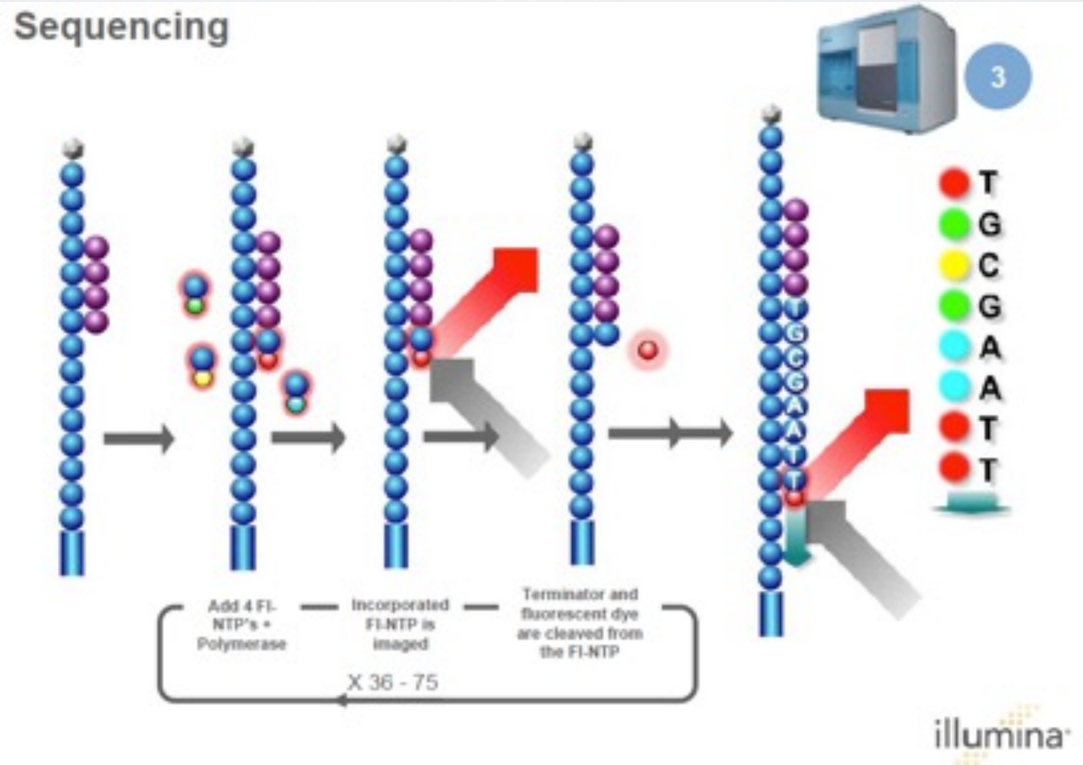


ILLUMINA

- □ DNA template immobilized to a flow cell
- □ Cluster formation by Bridge PCR
- □ Sequencing on flow cell (3000M reads)
- □ Sequencing by synthesis (protected nts)

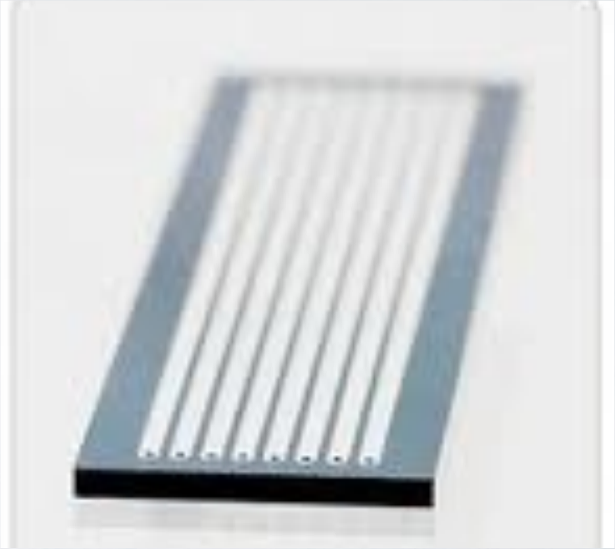


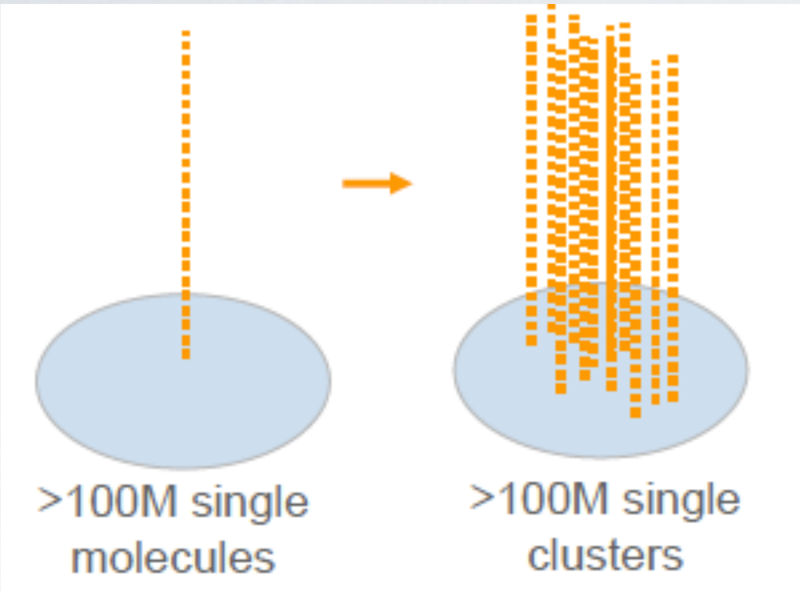


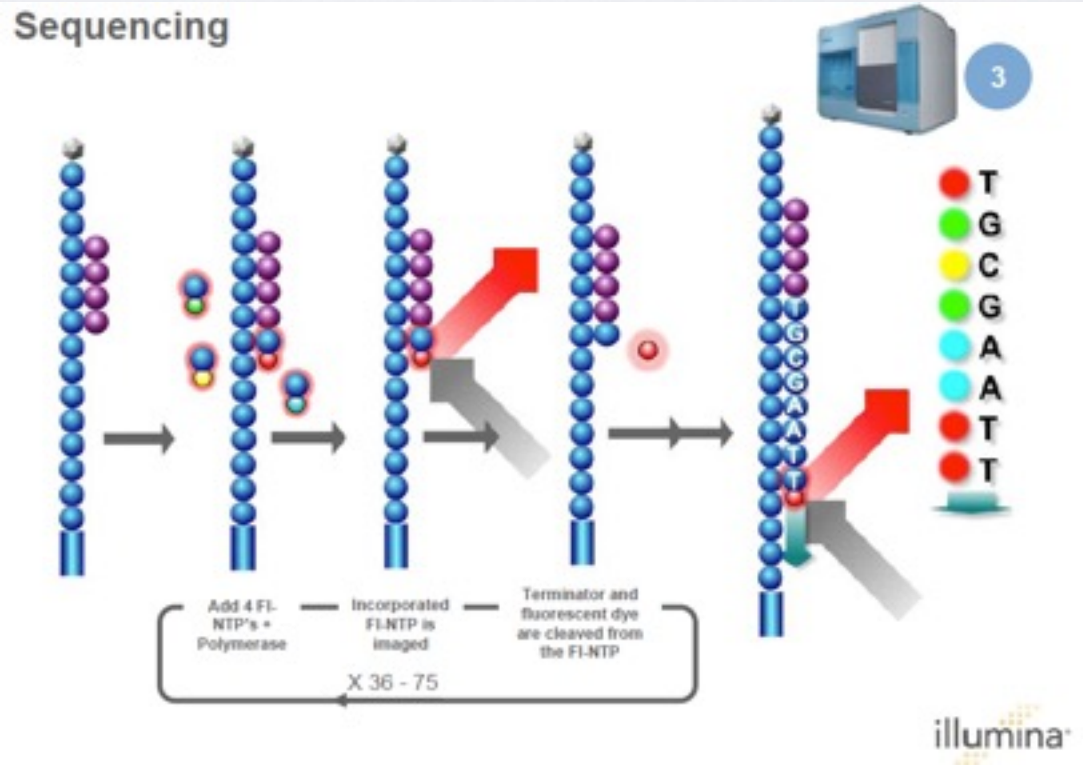

Sequencing
3
T
G
C
G
A
A
T
T
Add 4 Fl-NTP's + Polymerase
Incorporated Fl-NTP is imaged
Terminator and fluorescent dye are cleaved from the Fl-NTP
X 36 - 75
illumina

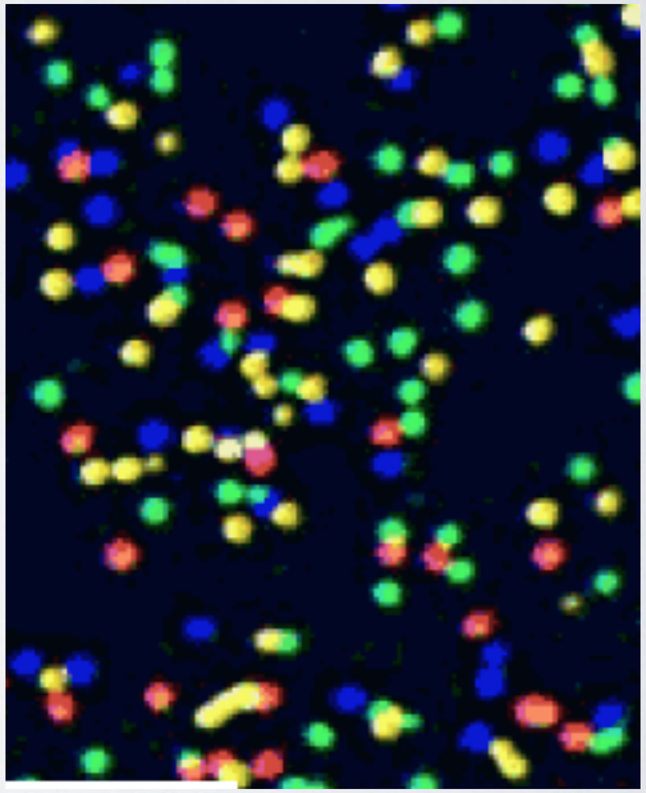
ILLUMINA

- □ DNA template immobilized to a flow cell
- □ Cluster formation by Bridge PCR
- □ Sequencing on flow cell (3000M reads)
- □ Sequencing by synthesis (protected nts)

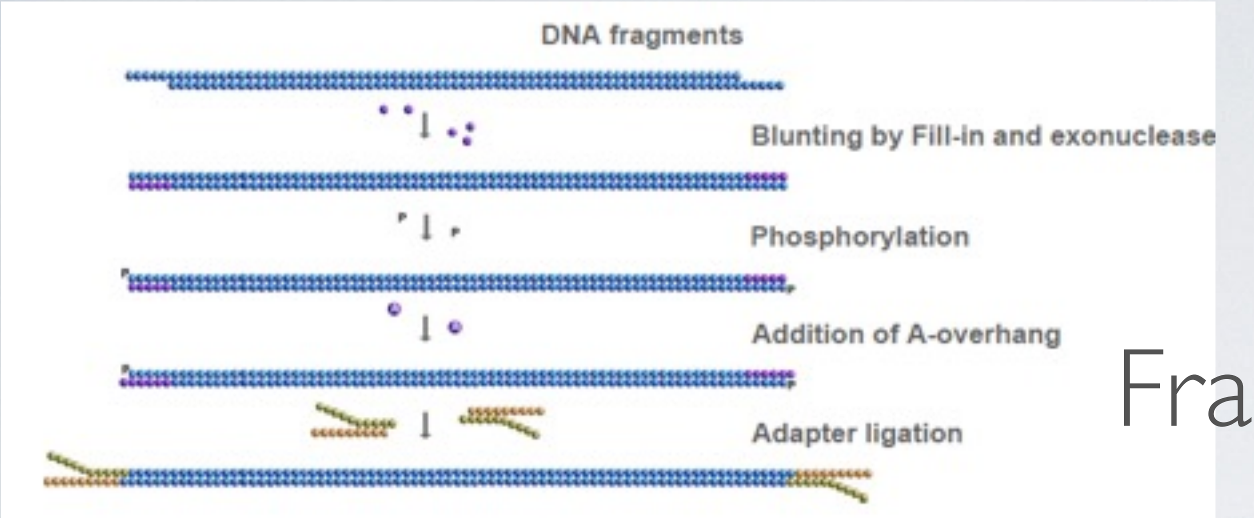




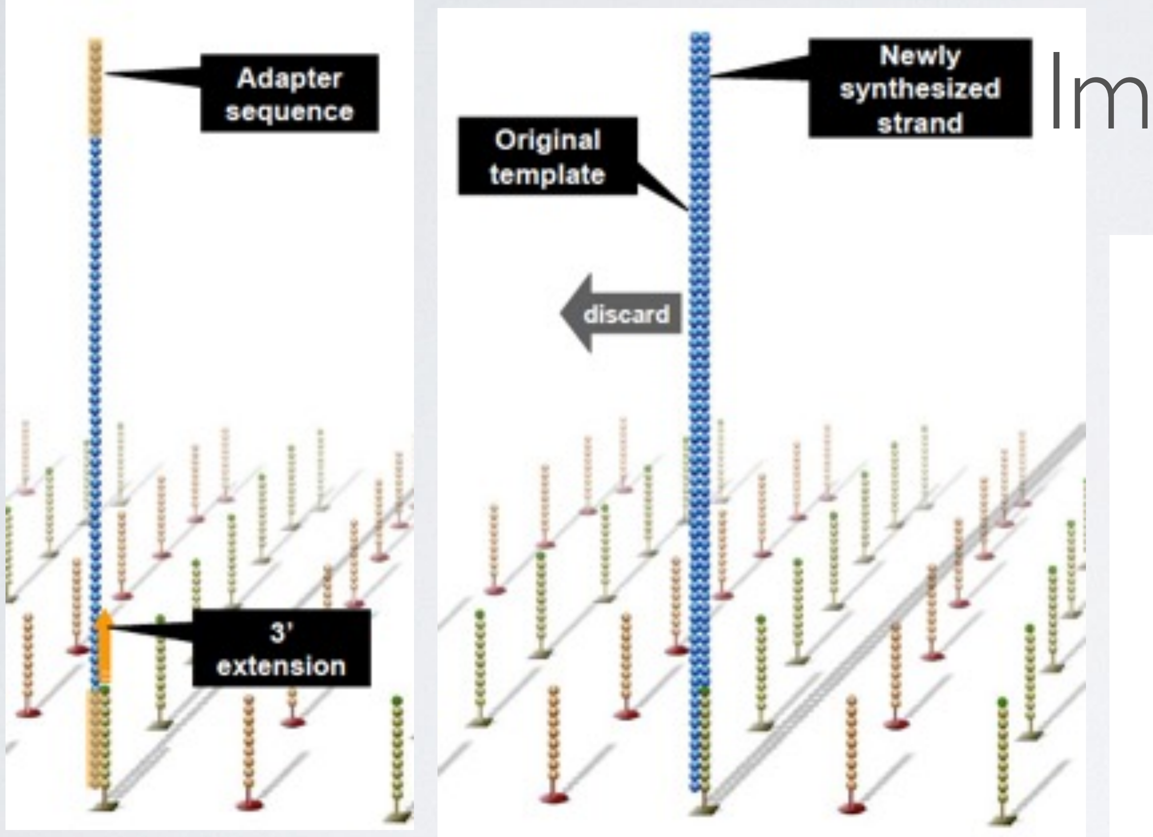

Sequencing
3
Add 4 Fl-NTP's + Polymerase
Incorporated Fl-NTP is imaged
Terminator and fluorescent dye are cleaved from the Fl-NTP
X 36 - 75
T
G
C
G
A
A
T
T
illumina



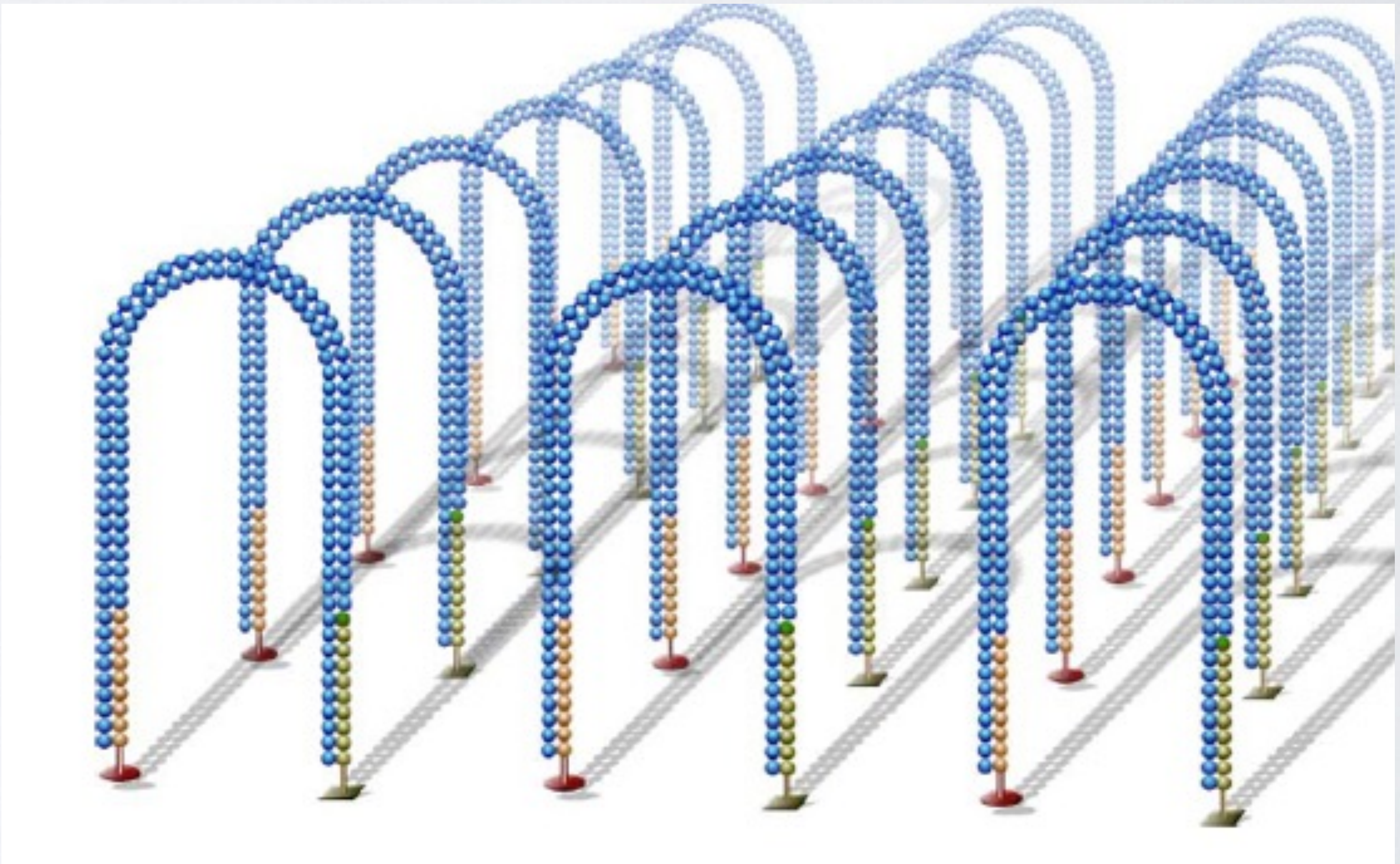
ILLUMINA



Fragmentation and adaptor ligation



Immobilization and strand synthesis



Bridge PCR to form clusters

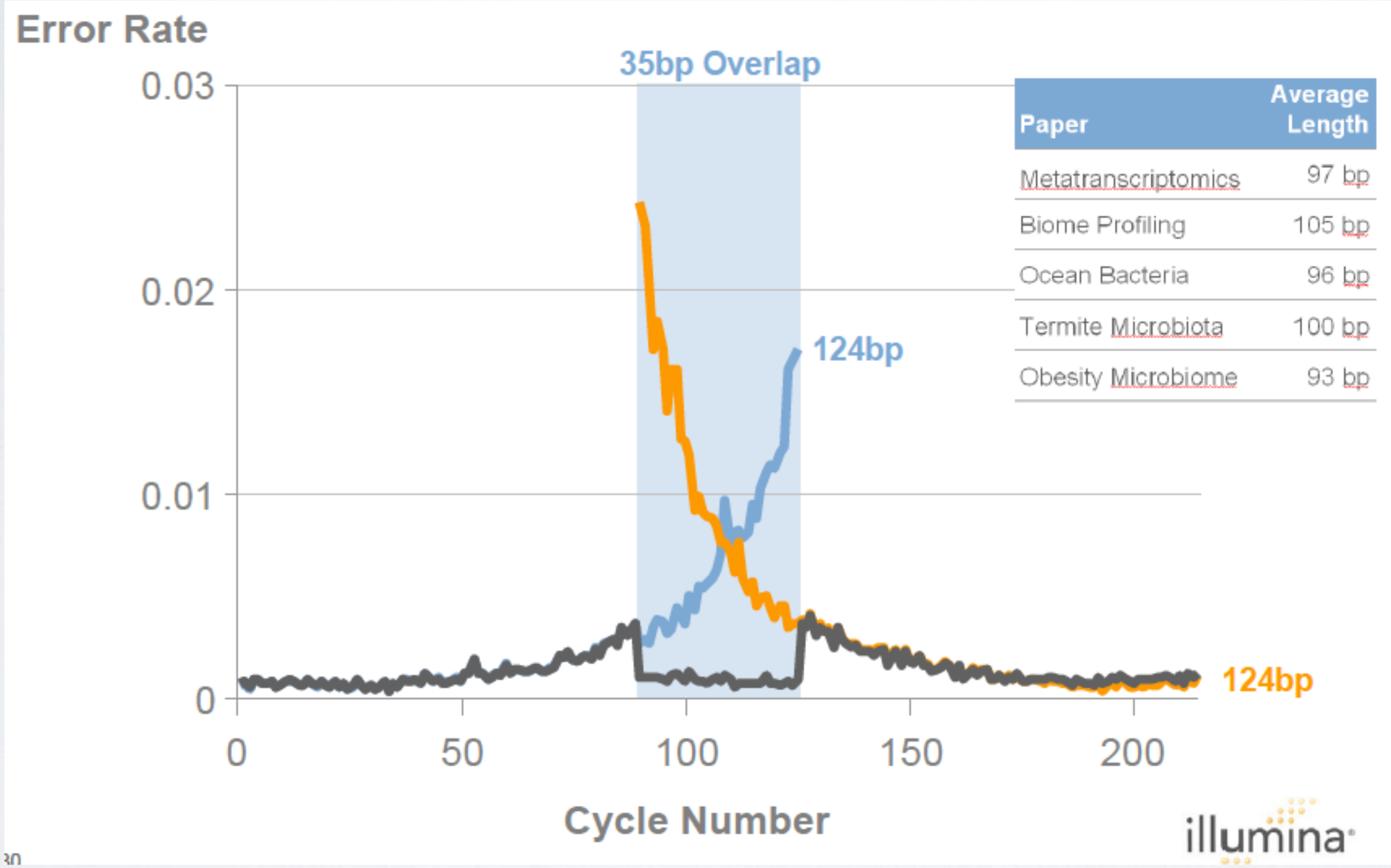
ILLUMINA

□ sequencing length up to 250nt

□ up to 3000M sequences - high coverage

□ 400 - 600Gb per run

□ sequence size limitation



Paper	Average Length
Metatranscriptomics	97 bp
Biome Profiling	105 bp
Ocean Bacteria	96 bp
Termite Microbiota	100 bp
Obesity Microbiome	93 bp

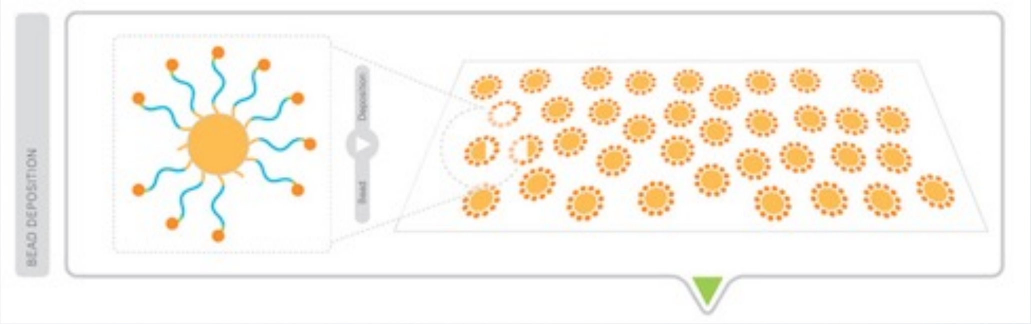
ILLUMINA

SOLID

- □ DNA template immobilized to a nano bead
- □ Cluster formation by emulsion PCR
- □ Sequencing on flow cell (4800M reads)
- □ Sequencing by ligation

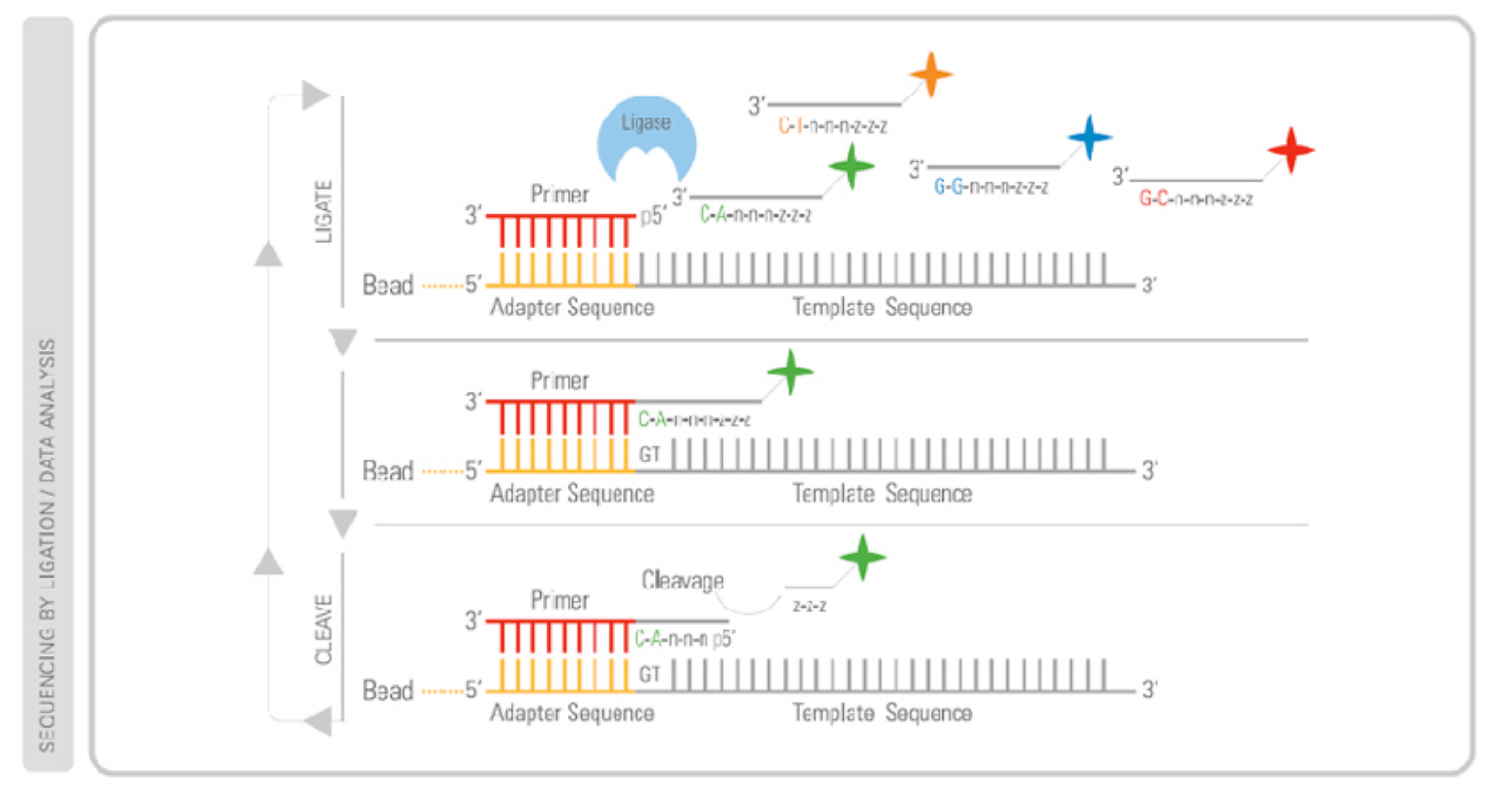
SOLID

- □ DNA template immobilized to a nano bead
- □ Cluster formation by emulsion PCR
- □ Sequencing on flow cell (4800M reads)
- □ Sequencing by ligation



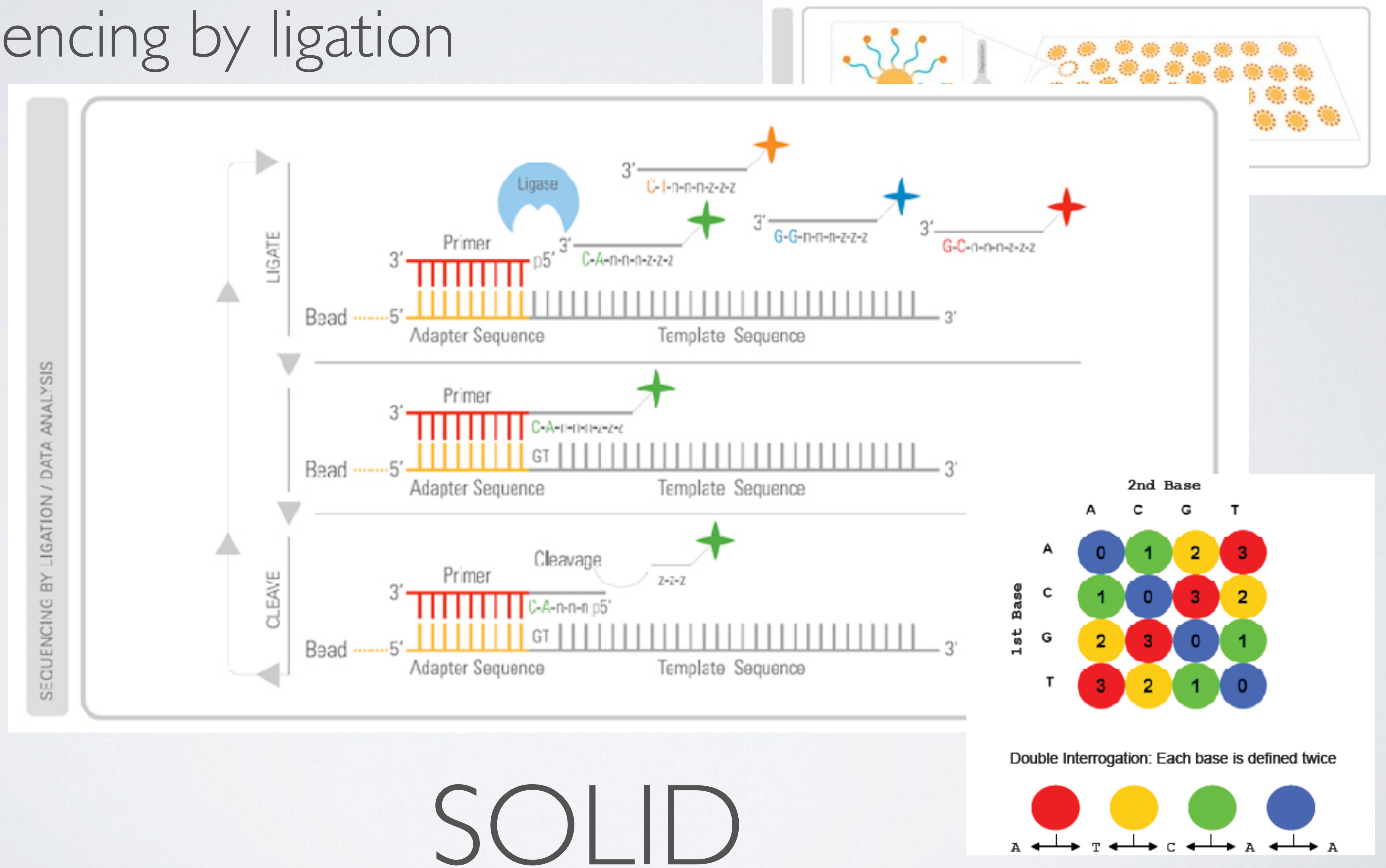
SOLID

- □ DNA template immobilized to a nano bead
- □ Cluster formation by emulsion PCR
- □ Sequencing on flow cell (4800M reads)
- □ Sequencing by ligation



SOLID

- □ DNA template immobilized to a nano bead
- □ Cluster formation by emulsion PCR
- □ Sequencing on flow cell (4800M reads)
- □ Sequencing by ligation

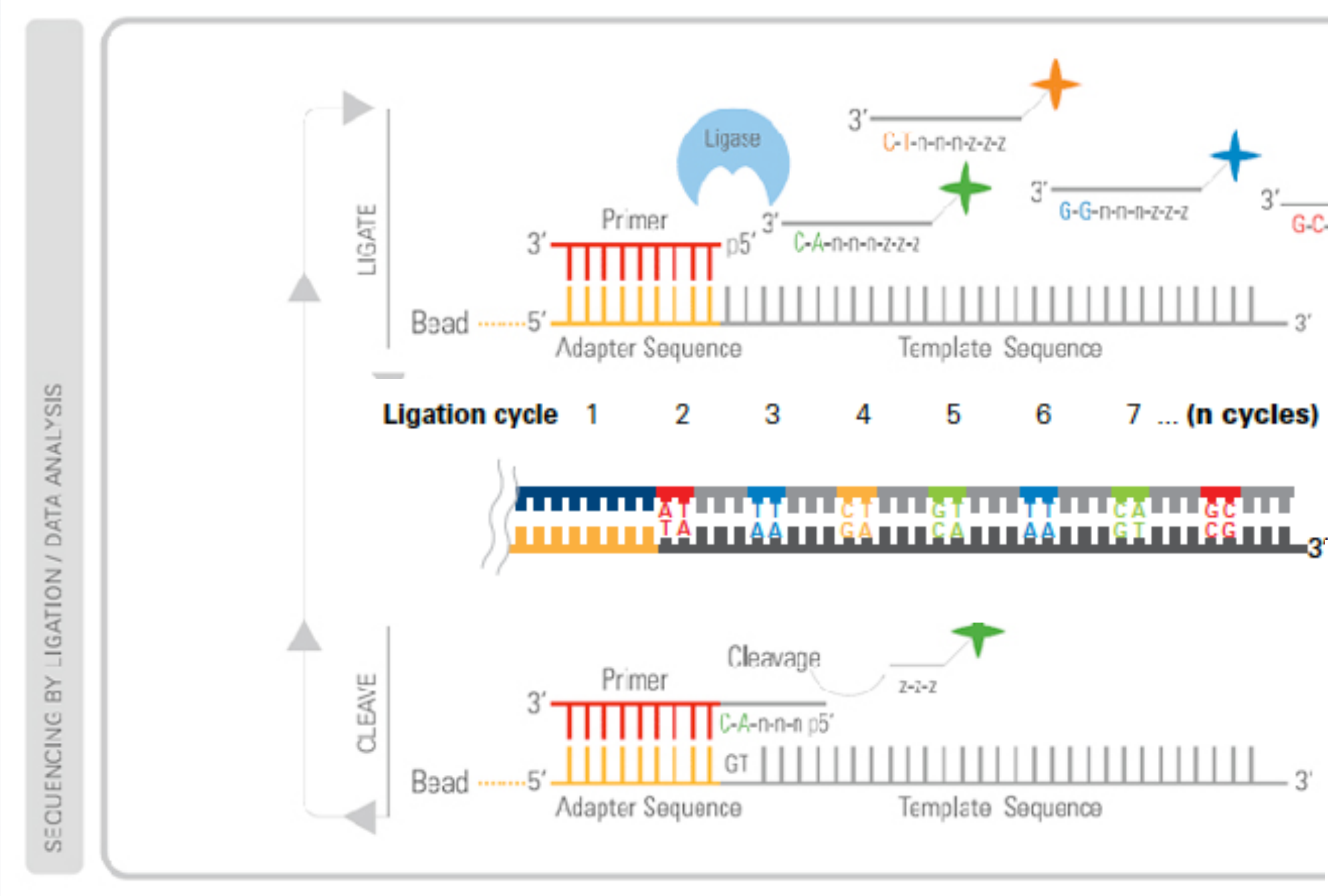


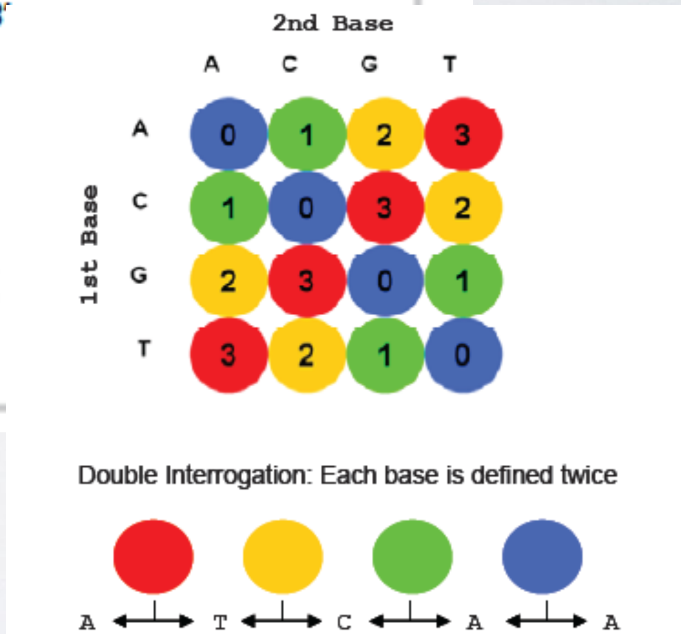
		2nd Base			
		A	C	G	T
1st Base	A	0	1	2	3
	C	1	0	3	2
	G	2	3	0	1
	T	3	2	1	0

Double Interrogation: Each base is defined twice

SOLID

□ DNA template immobilized to a nano bead
□ Cluster formation by emulsion PCR
□ Sequencing on flow cell (4800M reads)
□ Sequencing by ligation



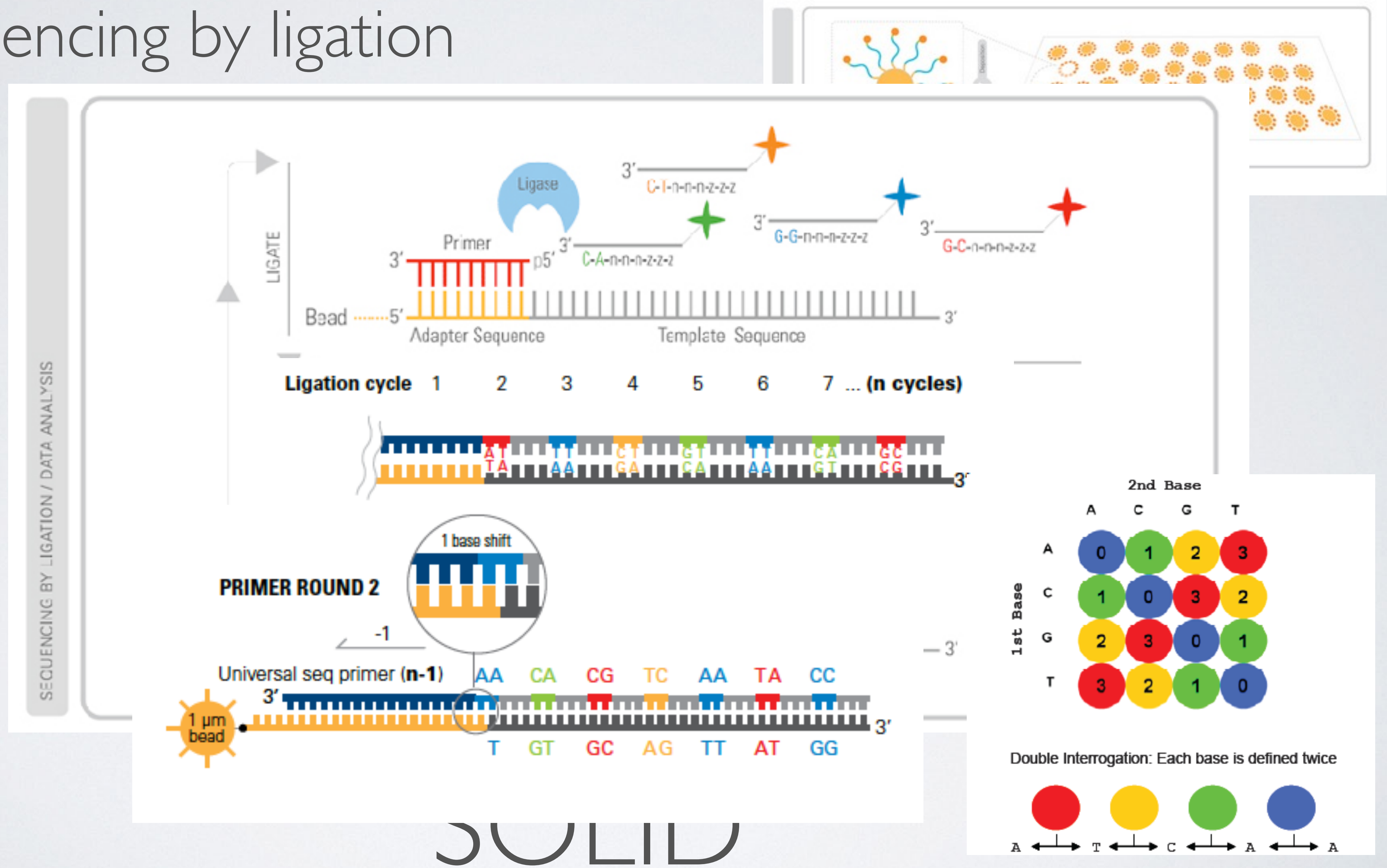


	2nd Base			
1st Base	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

Double Interrogation: Each base is defined twice

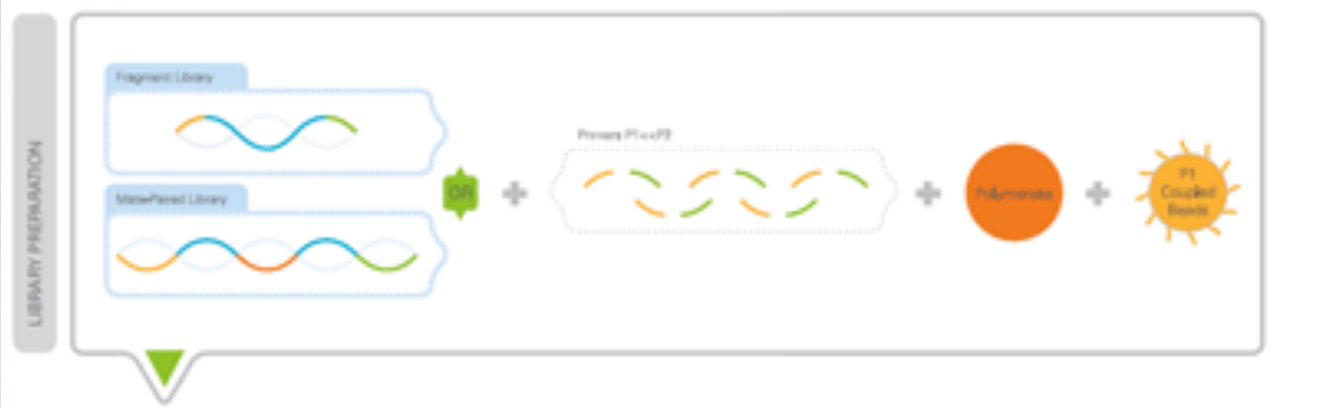
SOLID

- ❑ DNA template immobilized to a nano bead
- ❑ Cluster formation by emulsion PCR
- ❑ Sequencing on flow cell (4800M reads)
- ❑ Sequencing by ligation

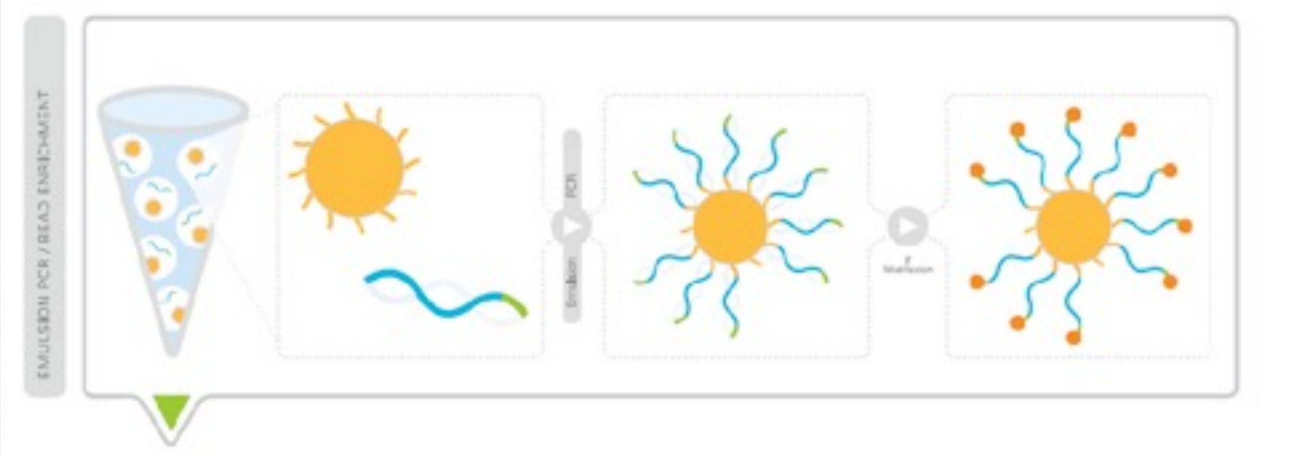


	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

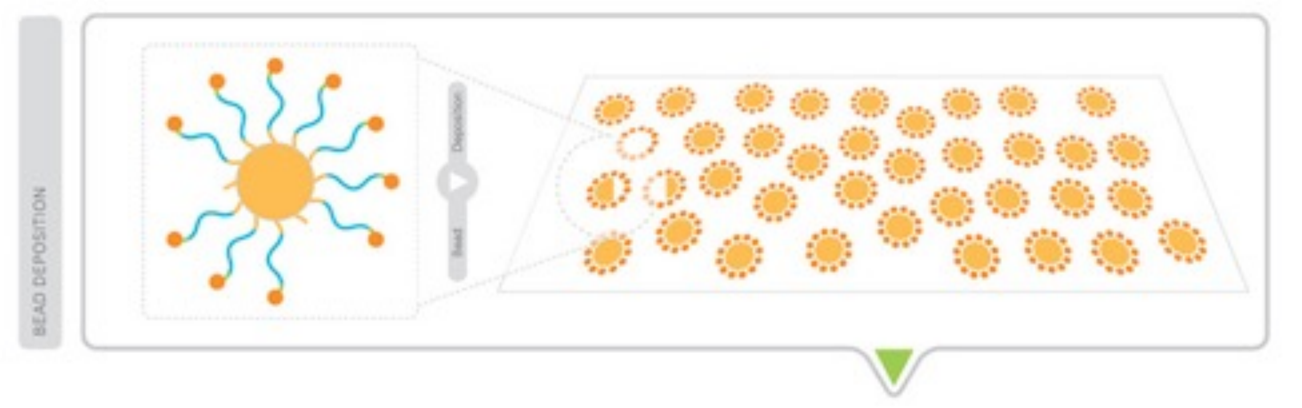
SOLID



Fragmentation, adaptors and immobilization



Emulsion PCR and bead separation



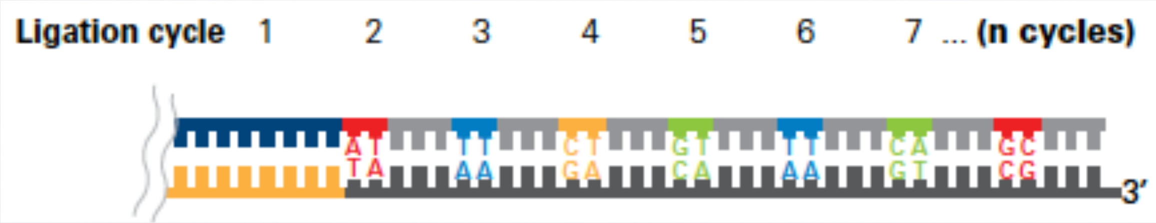
Bead deposition

SOLID

- sequencing length up to 75nt
- up to 4800M sequences - high coverage
- - 300 Gb per run
- sequence size limitation
- several sequencing rounds
- Every base is called twice

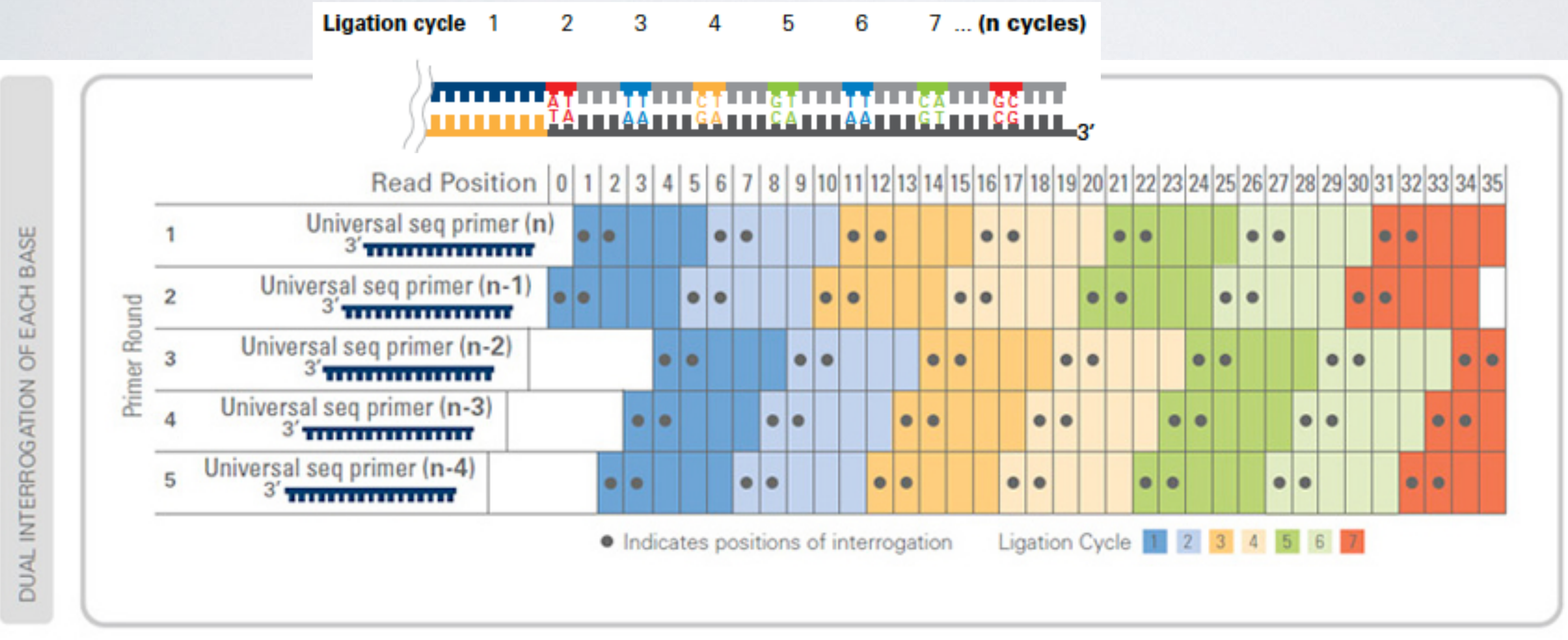
SOLID

- □ sequencing length up to 75nt
- □ up to 4800M sequences - high coverage
- □ - 300 Gb per run
- □ sequence size limitation
- □ several sequencing rounds
- □ Every base is called twice



SOLID

- sequencing length up to 75nt
- up to 4800M sequences - high coverage
- - 300 Gb per run
- sequence size limitation
- several sequencing rounds
- Every base is called twice


Ligation cycle 1 2 3 4 5 6 7 ... (n cycles)
3'
Read Position 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer Round
1 Universal seq primer (n) 3'
2 Universal seq primer (n-1) 3'
3 Universal seq primer (n-2) 3'
4 Universal seq primer (n-3) 3'
5 Universal seq primer (n-4) 3'
● Indicates positions of interrogation
Ligation Cycle 1 2 3 4 5 6 7
DUAL INTERROGATION OF EACH BASE

SOLID

- □ Sequence fragmentation
- □ Adaptor ligation
- □ Sequence immobilization
- □ PCR amplification (emulsion or bridge PCR)
- □ Real-time sequencing (by synthesis or ligation)
- □ huge amount of short sequence reads
- □ high coverage
- □ difficulties with assembling

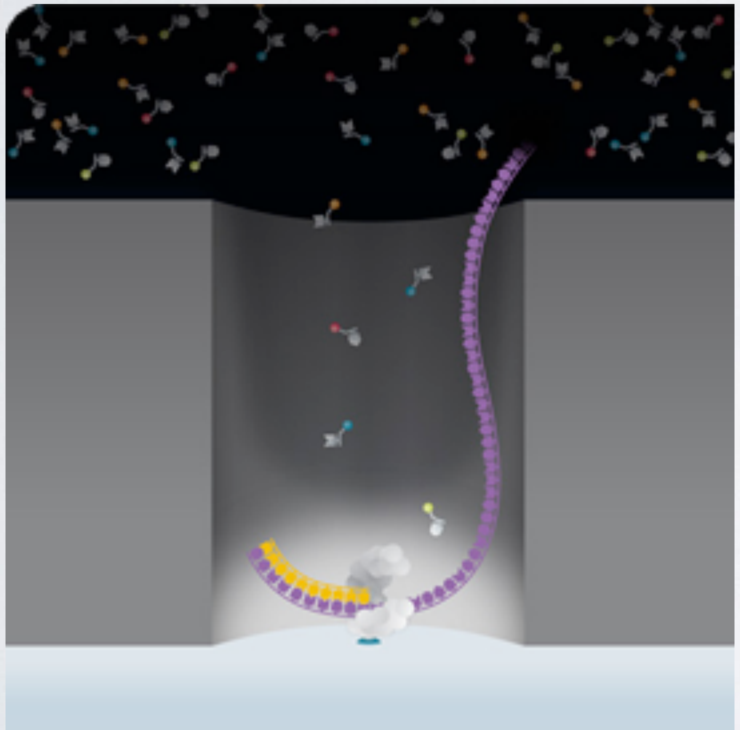
SUMMARY

PACIFIC BIOSCIENCES

- □ Polymerase immobilized in a nano well
- □ NO amplification (true single molecule sequencing)
- □ Sequencing on flow cell (75K reads)
- □ Sequencing by synthesis (fluorescence)
- □ Read length up to 10000nt average >1000
- □ Fast sample preparation and sequencing (8h)

PACIFIC BIOSCIENCES

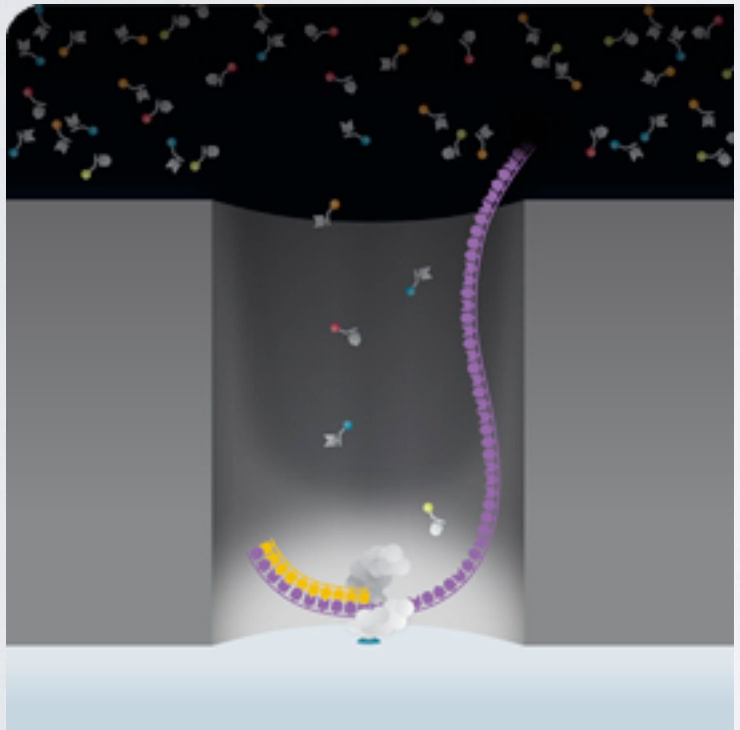
- Polymerase immobilized in a nano well
- NO amplification (true single molecule sequencing)
- Sequencing on flow cell (75K reads)
- Sequencing by synthesis (fluorescence)
- Read length up to 10000nt average >1000
- Fast sample preparation and sequencing (8h)



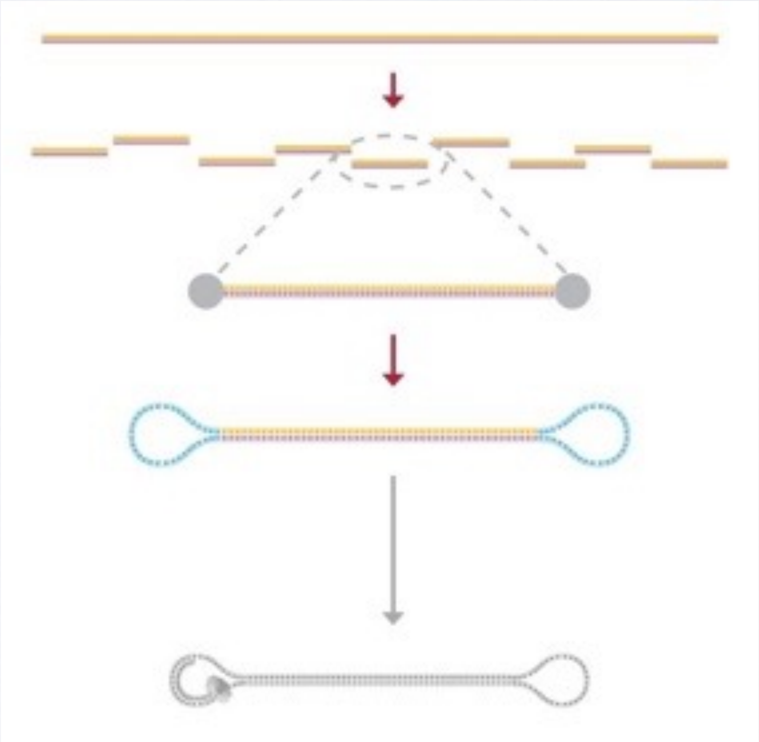
Zero-mode waveguide

PACIFIC BIOSCIENCES

- □ Polymerase immobilized in a nano well
- □ NO amplification (true single molecule sequencing)
- □ Sequencing on flow cell (75K reads)
- □ Sequencing by synthesis (fluorescence)
- □ Read length up to 10000nt average >1000
- □ Fast sample preparation and sequencing (8h)



Zero-mode waveguide



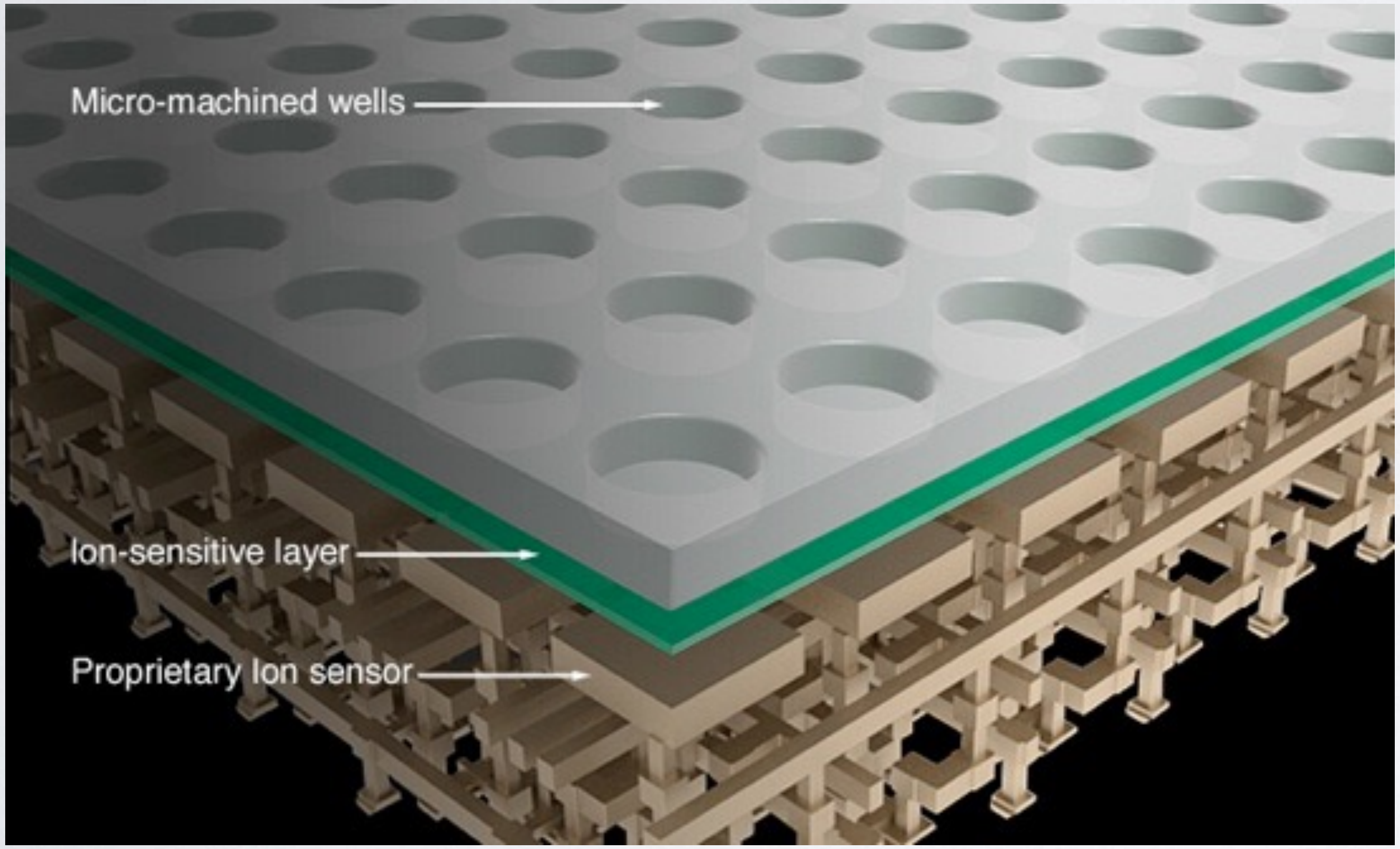
PACIFIC BIOSCIENCES

ION TORRENT

□ Immobilized in a nano well (semiconductor)
□ NO amplification (true single molecule sequencing)
□ Sequencing on flow cell (1M reads)
□ Sequencing by synthesis (H+ release)
□ Read length up to 10000nt average >400
□ Fast sample preparation and sequencing (8h)

ION TORRENT

- Immobilized in a nano well (semiconductor)
- NO amplification (true single molecule sequencing)
- Sequencing on flow cell (1M reads)
- Sequencing by synthesis (H+ release)
- Read length up to 10000nt average >400
- Fast sample preparation and sequencing (8h)



ION TORRENT

□ Immobilized in a nano well (semiconductor)

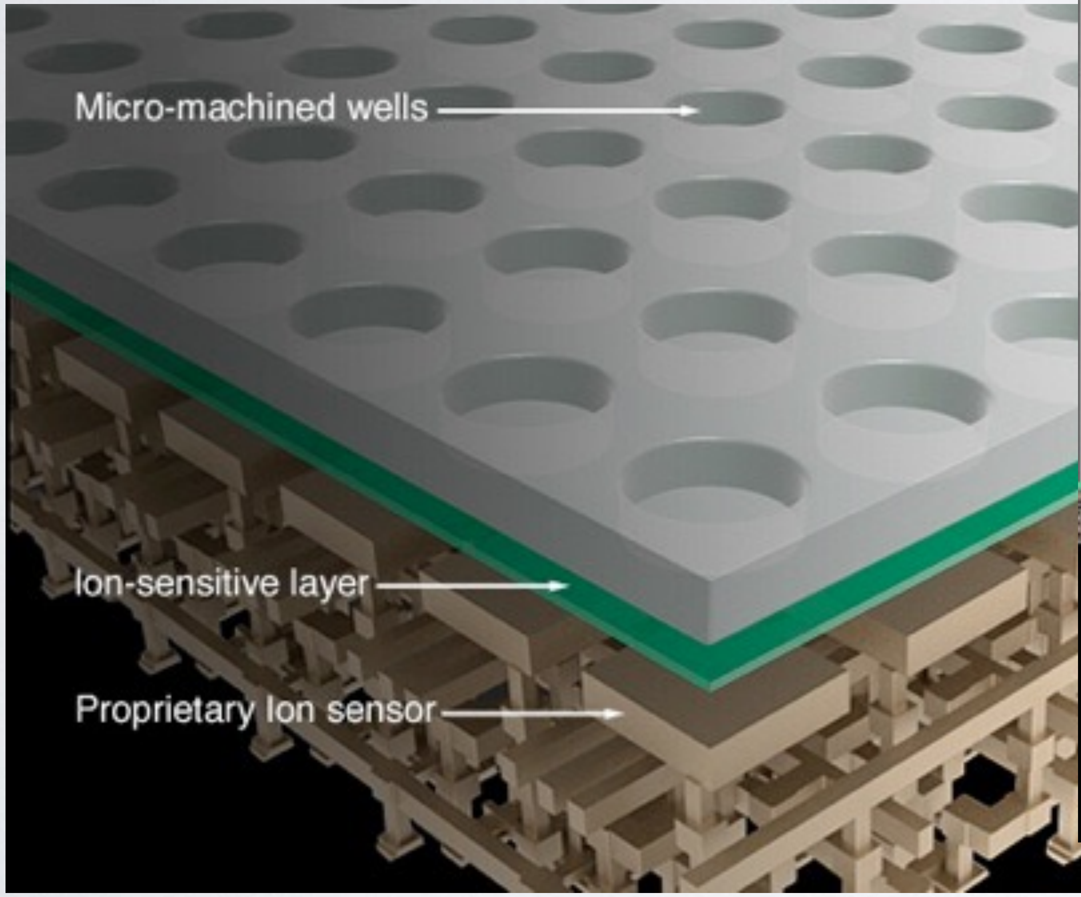
□ NO amplification (true single molecule sequencing)

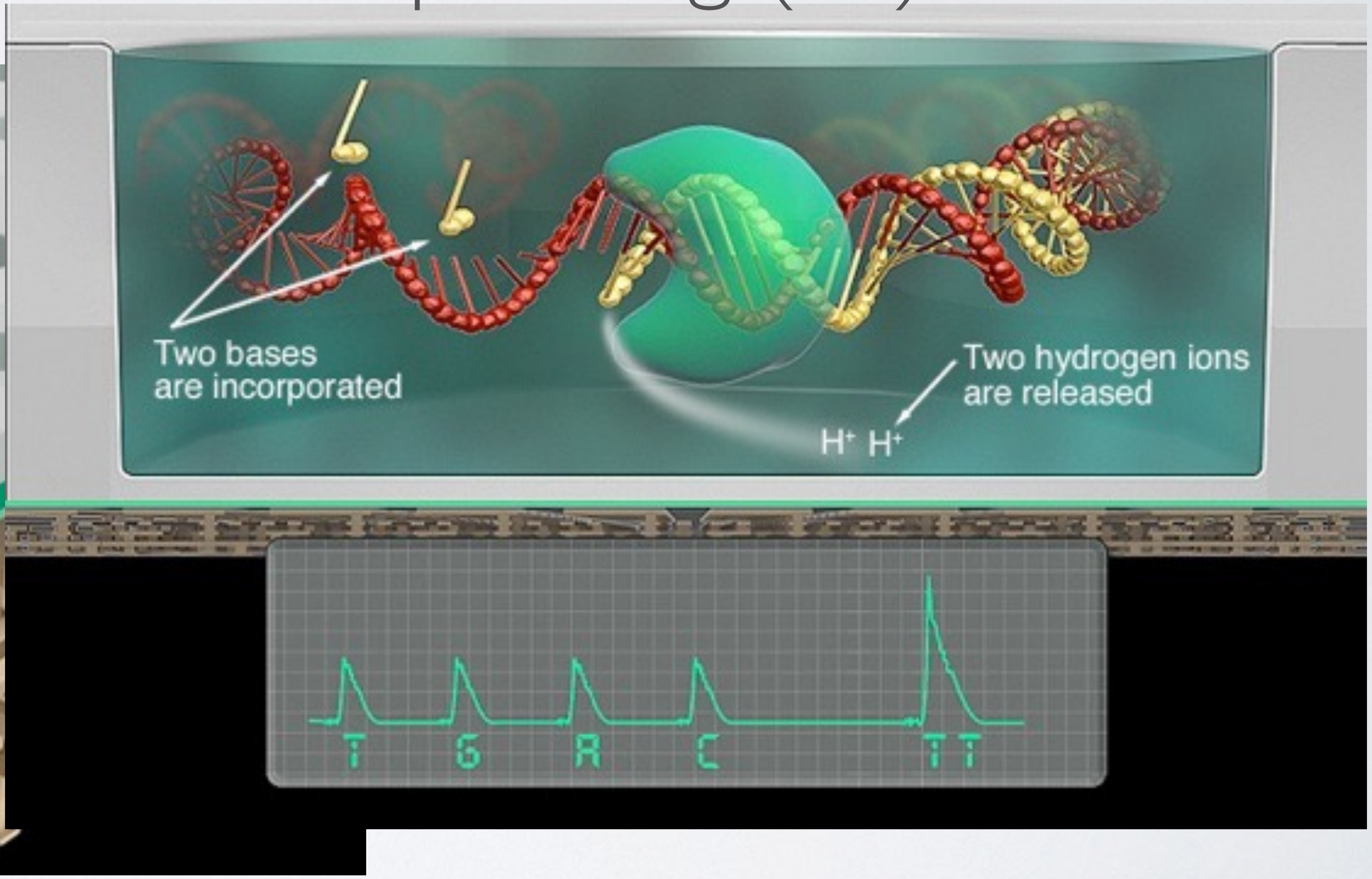
□ Sequencing on flow cell (1M reads)

□ Sequencing by synthesis (H+ release)

□ Read length up to 10000nt average >400

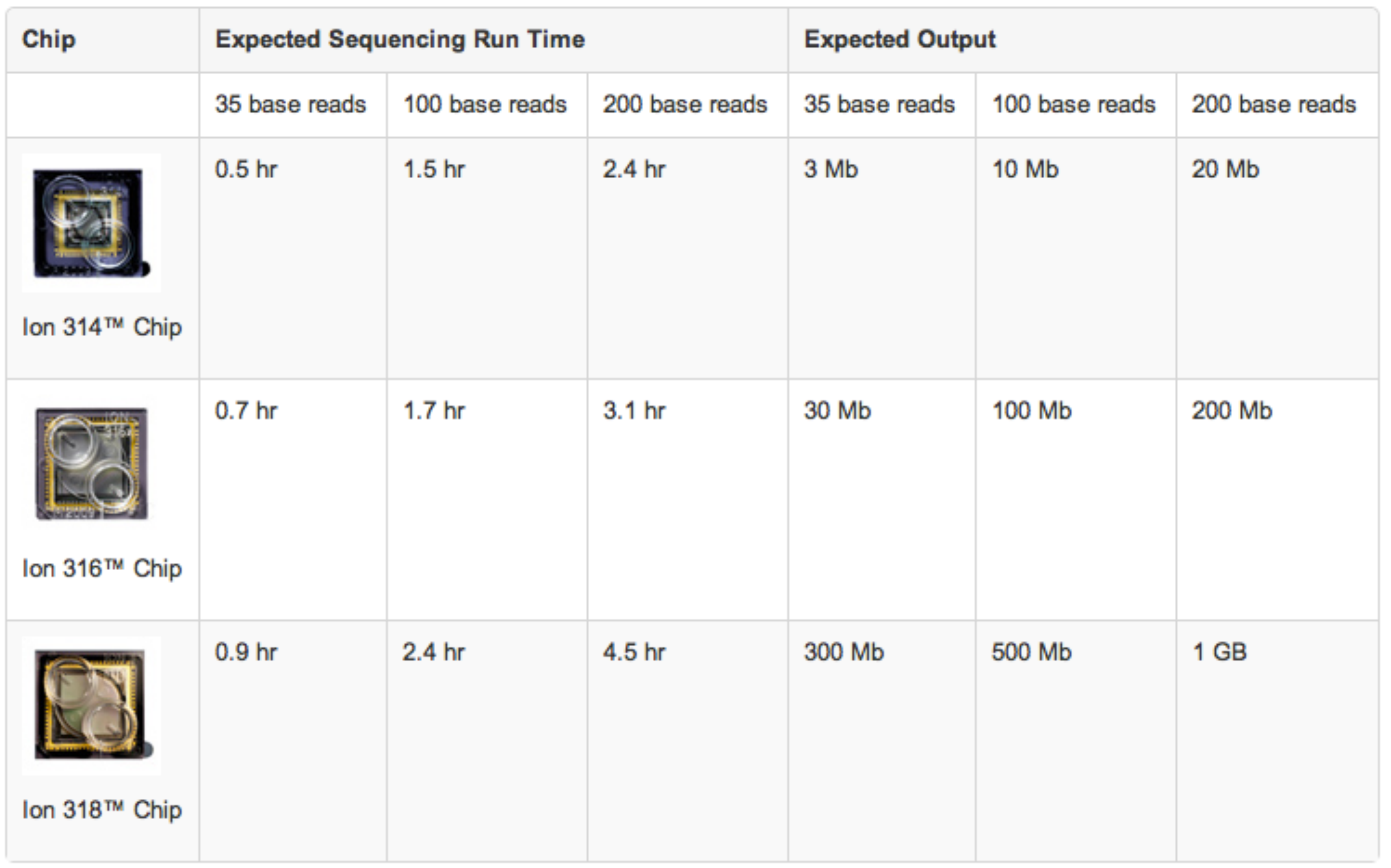
□ Fast sample preparation and sequencing (8h)





ION TORRENT

- Immobilized in a nano well (semiconductor)
- NO amplification (true single molecule sequencing)
- Sequencing on flow cell (5M reads)
- Sequencing by synthesis (H+ release)
- 00
- g (8h)



Chip	Expected Sequencing Run Time			Expected Output		
	35 base reads	100 base reads	200 base reads	35 base reads	100 base reads	200 base reads
Ion 314™ Chip	0.5 hr	1.5 hr	2.4 hr	3 Mb	10 Mb	20 Mb
Ion 316™ Chip	0.7 hr	1.7 hr	3.1 hr	30 Mb	100 Mb	200 Mb
Ion 318™ Chip	0.9 hr	2.4 hr	4.5 hr	300 Mb	500 Mb	1 GB

ION TORRENT

NANOPORE

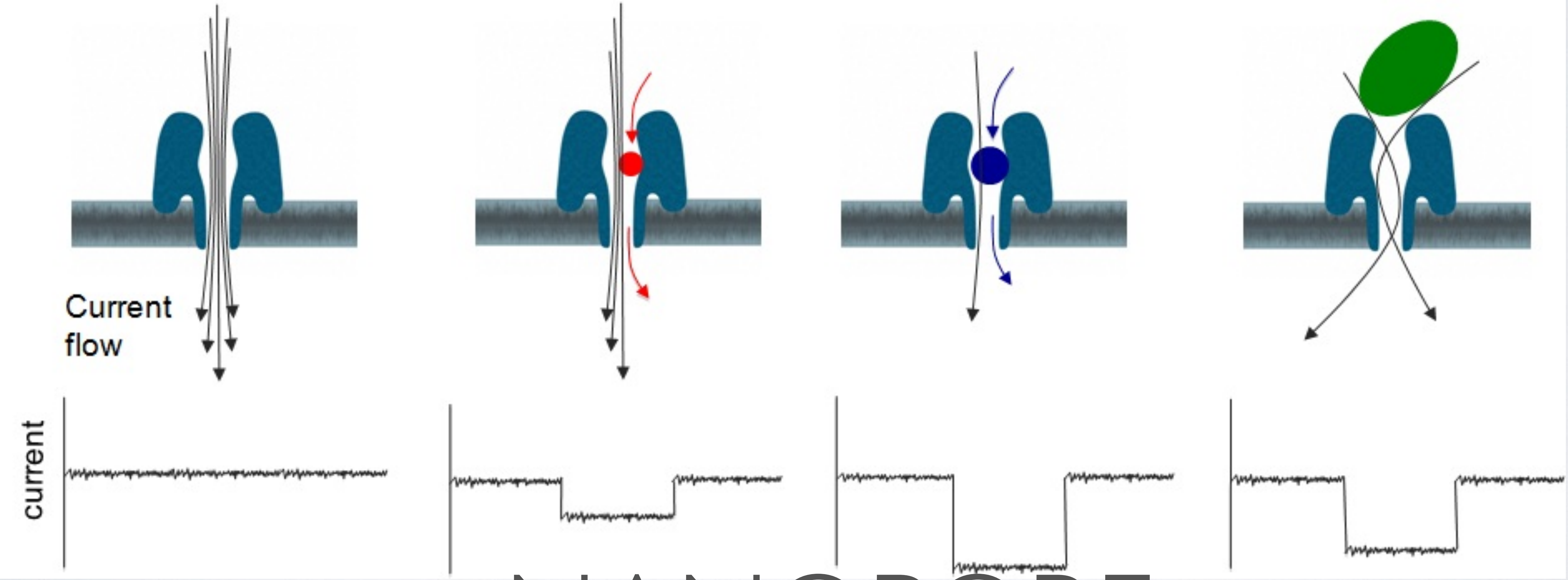
doi:10.1038/nature.2012.10051

□ NO immobilization
□ NO amplification (true single molecule sequencing)
□ Sequencing through solid-state nanopore
□ Sequencing by current disruption (8000 pores)
□ Read length up to 100000nt
□ in future 20 pores sequence human genome in 15min

NANOPORE

doi:10.1038/nature.2012.10051

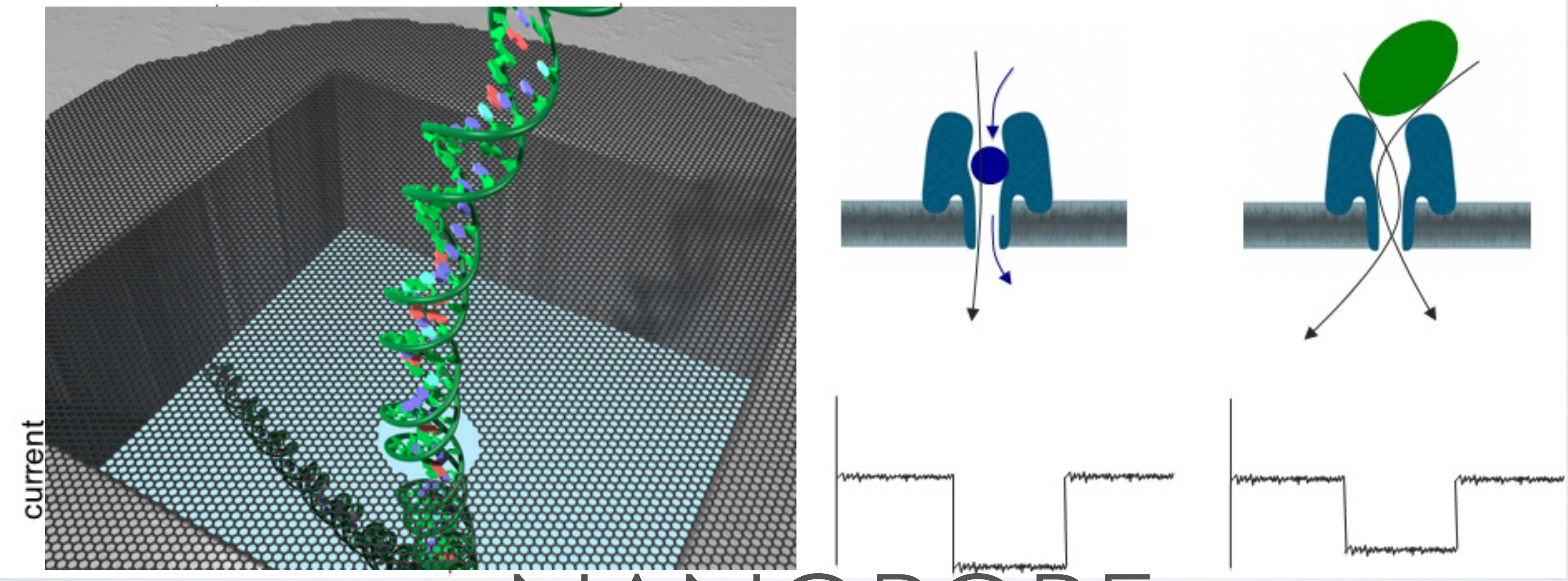
- □ NO immobilization
- □ NO amplification (true single molecule sequencing)
- □ Sequencing through solid-state nanopore
- □ Sequencing by current disruption (8000 pores)
- □ Read length up to 100000nt
- □ in future 20 pores sequence human genome in 15min



NANOPORE

doi:10.1038/nature.2012.10051

- □ NO immobilization
- □ NO amplification (true single molecule sequencing)
- □ Sequencing through solid-state nanopore
- □ Sequencing by current disruption (8000 pores)
- □ Read length up to 100000nt
- □ in future 20 pores sequence human genome in 15min



NANOPORE

doi:10.1038/nature.2012.10051

□ NO immobilization

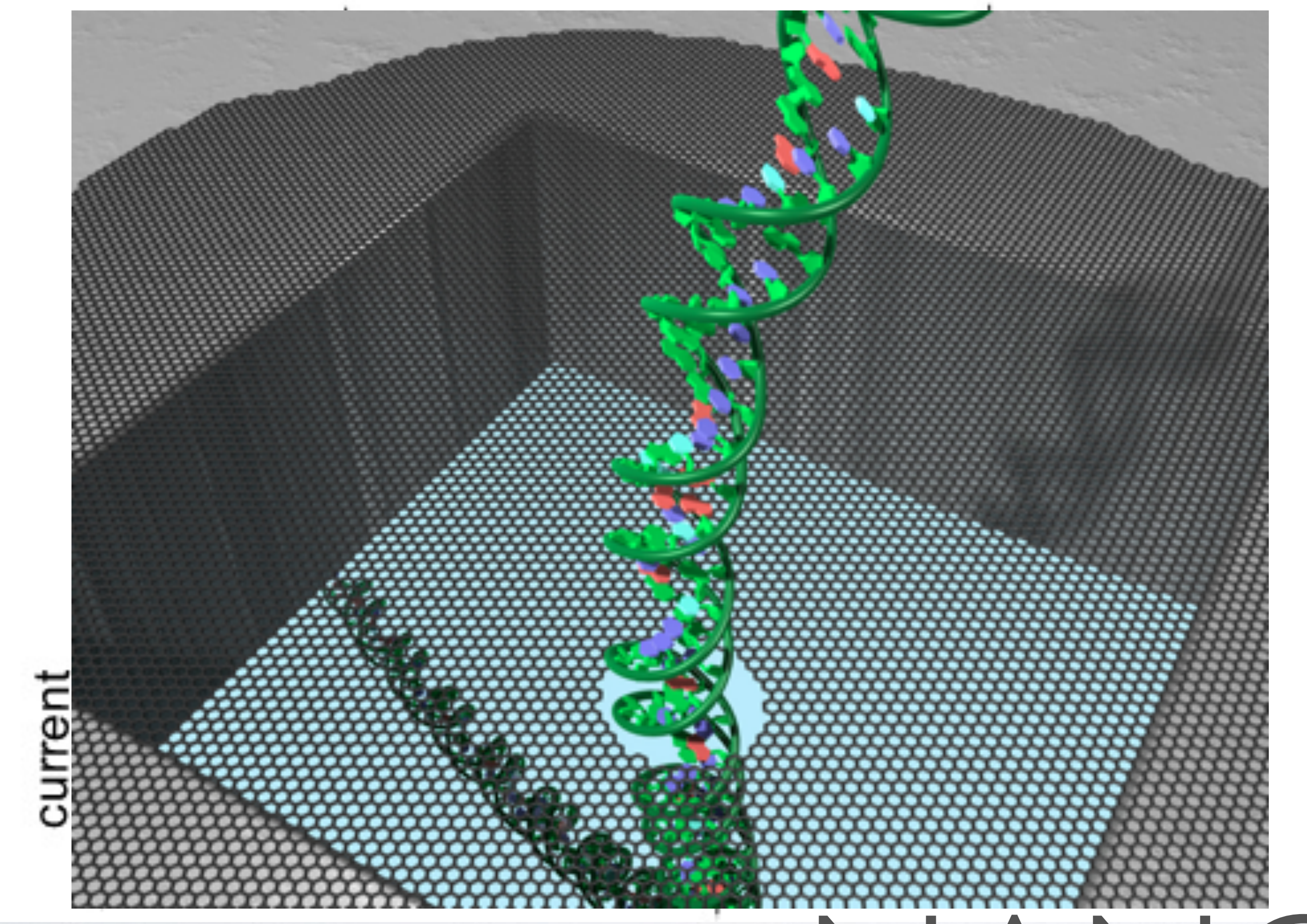
□ NO amplification (true single molecule sequencing)

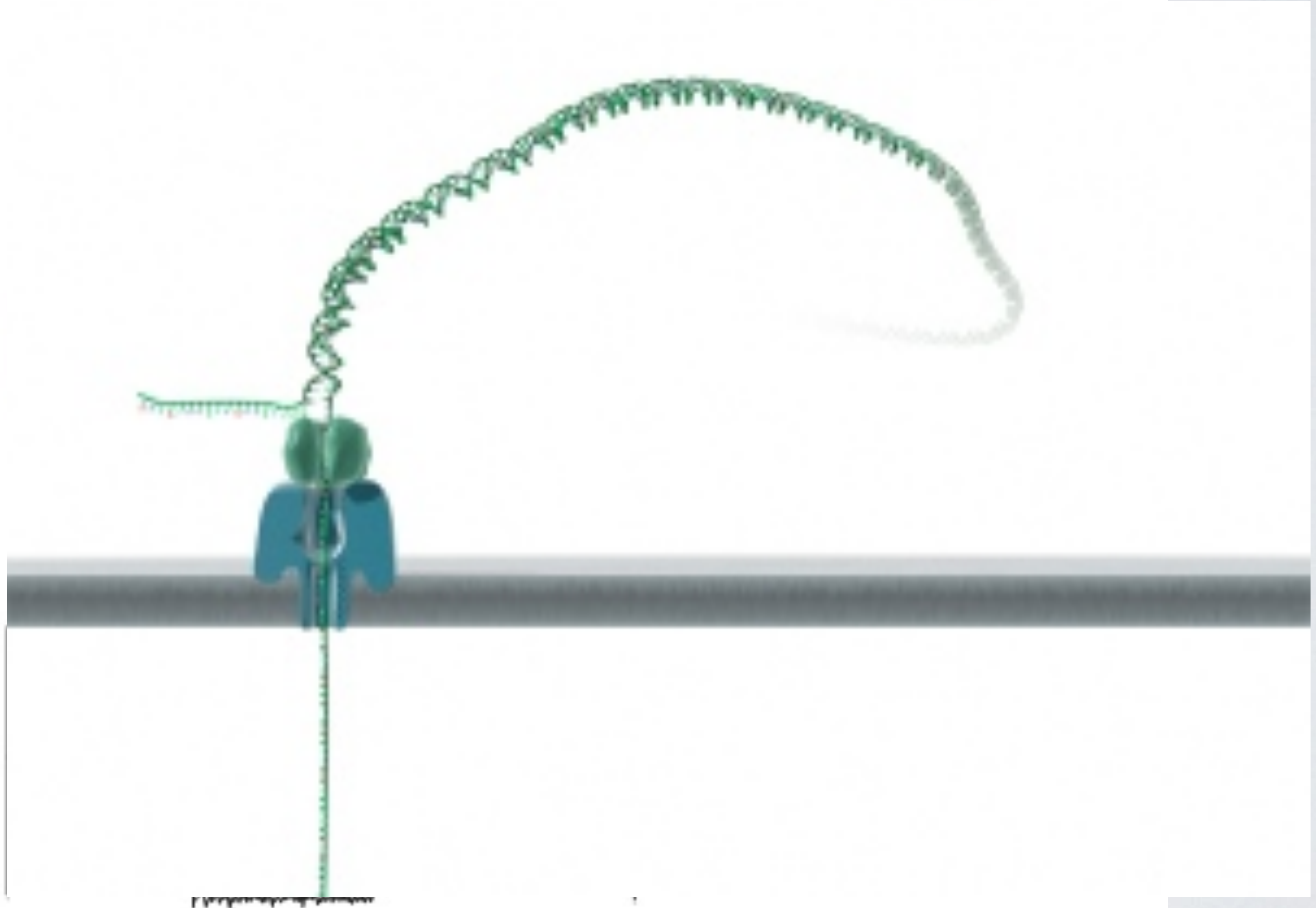
□ Sequencing through solid-state nanopore

□ Sequencing by current disruption (8000 pores)

□ Read length up to 100000nt

□ in future 20 pores sequence human genome in 15min





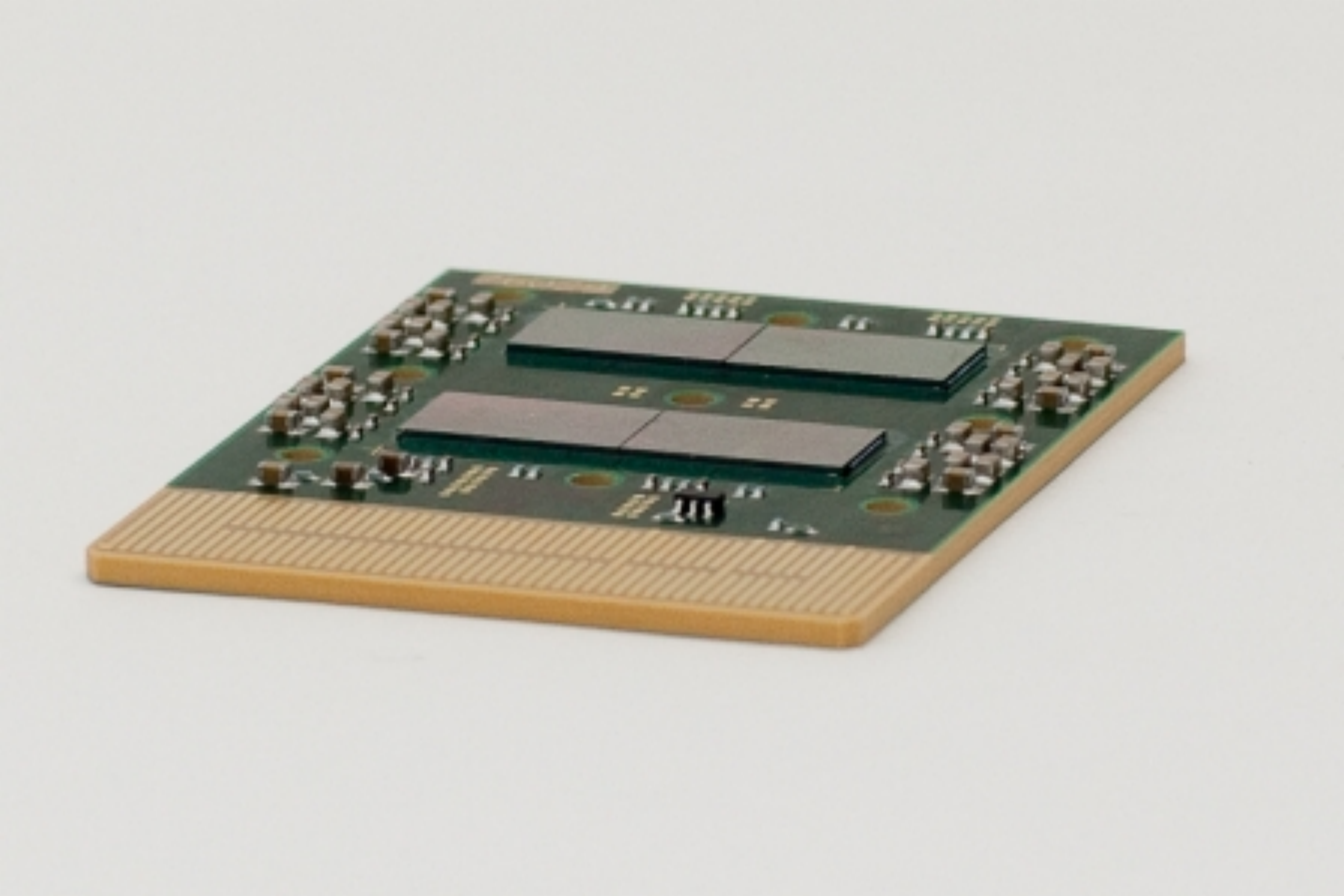
NANOPORE

doi:10.1038/nature.2012.10051

□ NO immobilization
□ NO amplification (true single molecule sequencing)
□ Sequencing through solid-state nanopore
□ Sequencing by current disruption (8000 pores)
□ Read length up to 100000nt
□ in future 20 pores sequence human genome in 15min

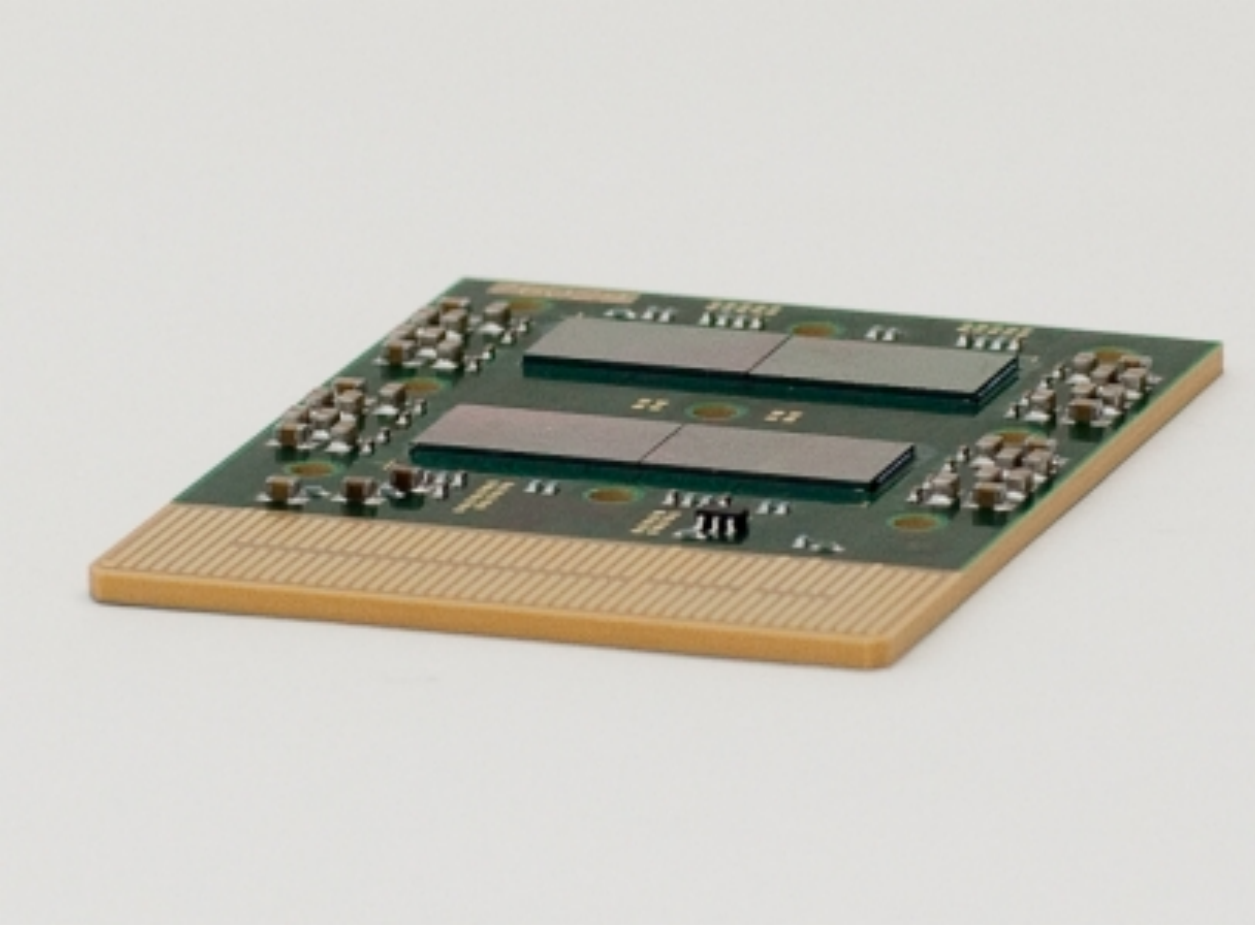
NANOPORE

- □ NO immobilization
- □ NO amplification (true single molecule sequencing)
- □ Sequencing through solid-state nanopore
- □ Sequencing by current disruption (8000 pores)
- □ Read length up to 100000nt
- □ in future 20 pores sequence human genome in 15min



NANOPORE

□ NO immobilization
□ NO amplification (true single molecule sequencing)
□ Sequencing through solid-state nanopore
□ Sequencing by current disruption (8000 pores)
□ Read length up to 100000nt
□ in future 20 pores sequence human genome in 15min



NANOPORE

DATA FORMAT

DATA FORMAT

□ Illumina

- SCARF (s_*_sequence.txt): Solexa Compact ASCII Read Format

```
HWI-EAS255_4_FC2010Y
1:43:110:790:TTAATCTACAGAATAGATAGCTAGCATATATTT:IIIIIIIIIIIIIIIIAIIIIIIIII&;II&,I
HWI-EAS255_4_FC2010Y:
1:43:122:836:GATCGGAAGGCTCGTATGCCGTCTTCTTCTTTT:IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

- FASTQ (*.fastq)

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+HWI-EAS255_4_FC2010Y_1_43_110_790
IIIIIIIIIIIIIIIIAIIIIIIIII&;II&,I
```

DATA FORMAT

SOLiD

- CSFASTA (xxxx.csfasta): Color Space FASTA

```
>1_51_64_F3
T10301031230333233203333000021122223
>1_51_127_F3
T20103232332031323101101002003103102
```

- QUAL (xxxx_.QV.qual):

```
>1_51_64_F3
12 7 21 16 6 2 25 5 25 26 6 7 2 8 5 2 3 2 6 21 5 2 3 9 4 2 2 2 17 6 2 2 2 5 3
>1_51_127_F3
3 18 15 4 11 2 6 4 4 6 2 7 2 9 4 3 2 6 18 2 2 4 3 2 2 2 2 2 2 4 2 3 4 4 2
```

DATA FORMAT

454 Roche

- Roche 454 SFF Standard Flowgram Format (*.sff)

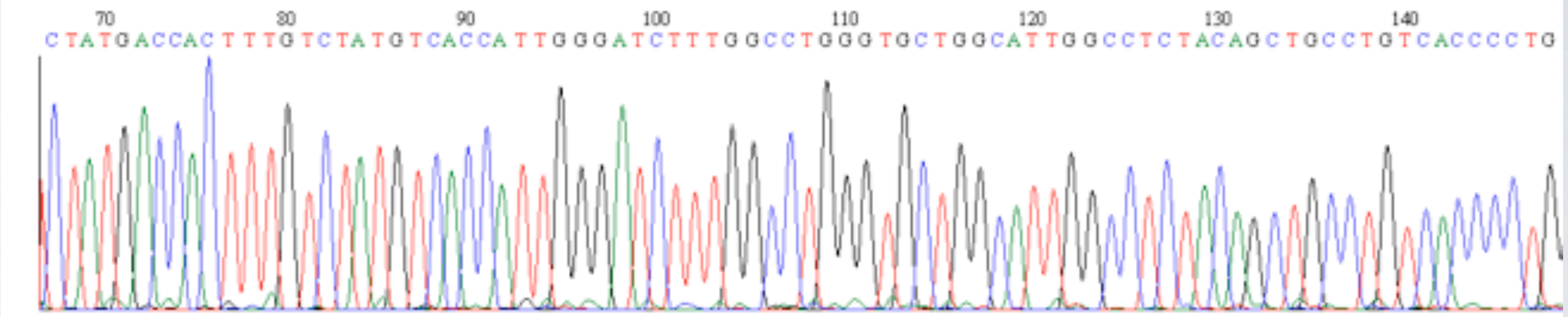
- FASTA (*.fna)

```
>E6PIHNP01B74B0
AACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAG
AAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGC
```

- QUAL (*.qual)

```
>E6PIHNP01B74B0
34 27 28 26 34 28 28 35 28 25 28 28 28 28 28 27 28 28 28 32 25 28 28 25 27 27
27 31 22 28 31 24 28 27 27 27 27 27 25 28 27 28 34 26 27 32 25 27 31 22 25 24
28 20 27 31 23 33 25 27 32 25 22 28 28 27 34 27 27 24 27 25 25 25 25 25 27 31
24 27 26 17 23 15 28 25 28 36 32 13 34 28 22 26 26 27 28 27 27 27 17 20 28 27
28 27 27 24 34 28 27 32 27 28 26 33 27 27 34 28 35 28 28 34 27 39 35 24 14 4
27 25 24 34 28 35 28 26 27 27 18 18 17 31 26 27 25 28 27 18 29 21 28
```

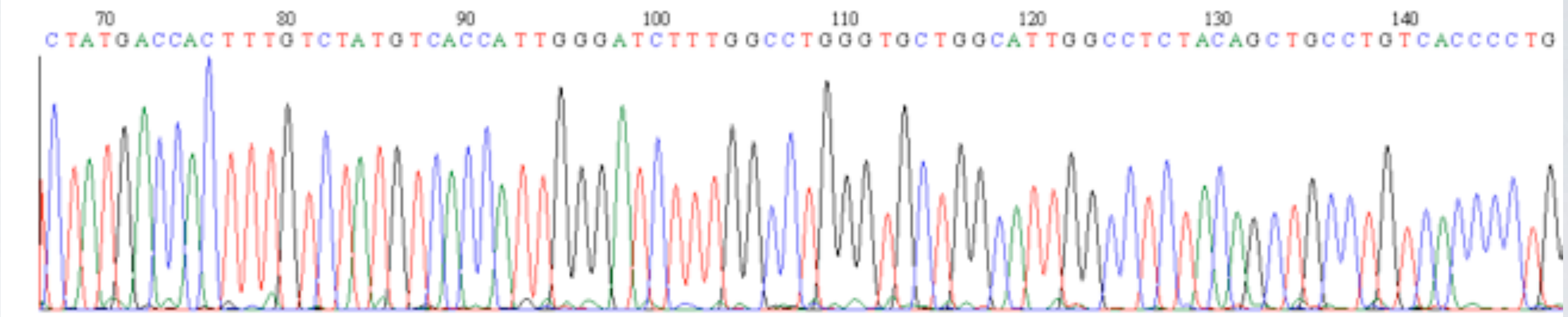
DATA FORMAT



Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

QUALITY SCORE

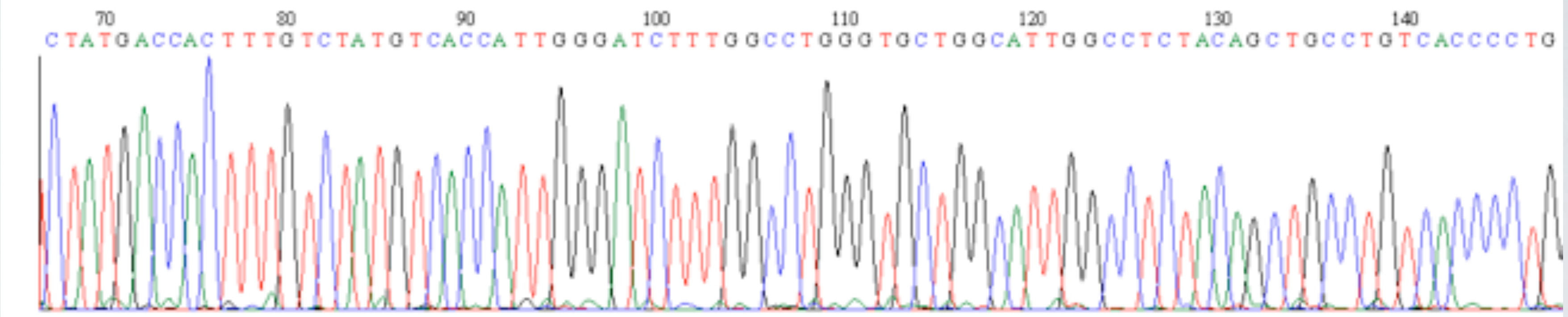


Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

```
@HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
TAAAGAAACTAAGAATAAGCAGATTATCTCGTAT
+HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
ffffffdaadKccaccfffefdceffefefe`b`
```

QUALITY SCORE



Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

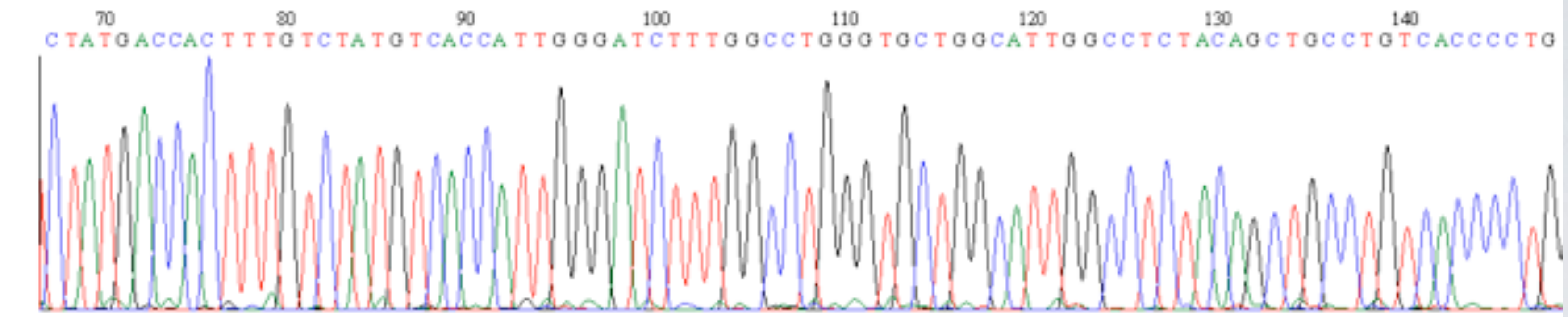
They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

```
@HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
TAAAGAAACTAAGAATAAGCAGATTATCTCGTAT
+HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
ffffffdaadKccaccfffefdceffefefe`b`
```

Sanger quality code (Phred): ASCII character code = phred quality value + 33

Illumina quality code: ASCII character code = phred quality value + 64

QUALITY SCORE



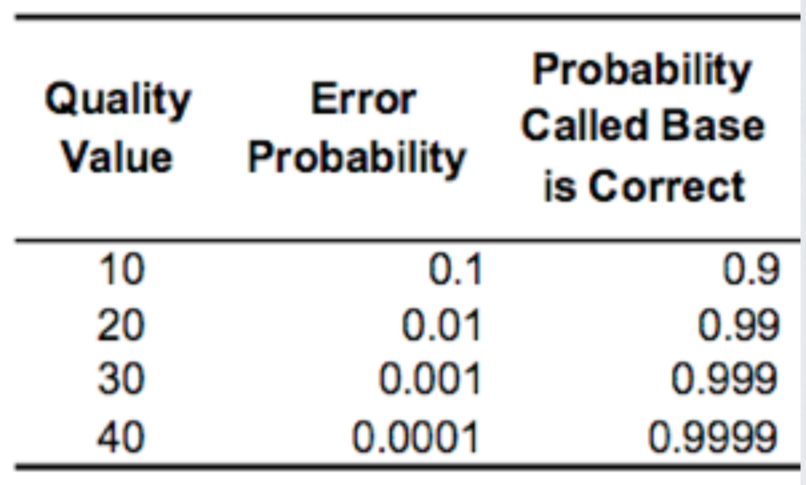
Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

```
@HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
TAAAGAAACTAAGAATAAGCAGATTATCTCGTAT
+HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
ffffffdaadKccaccfffefdceffefefe`b`
```

Sanger quality code (Phred): ASCII character code = phred

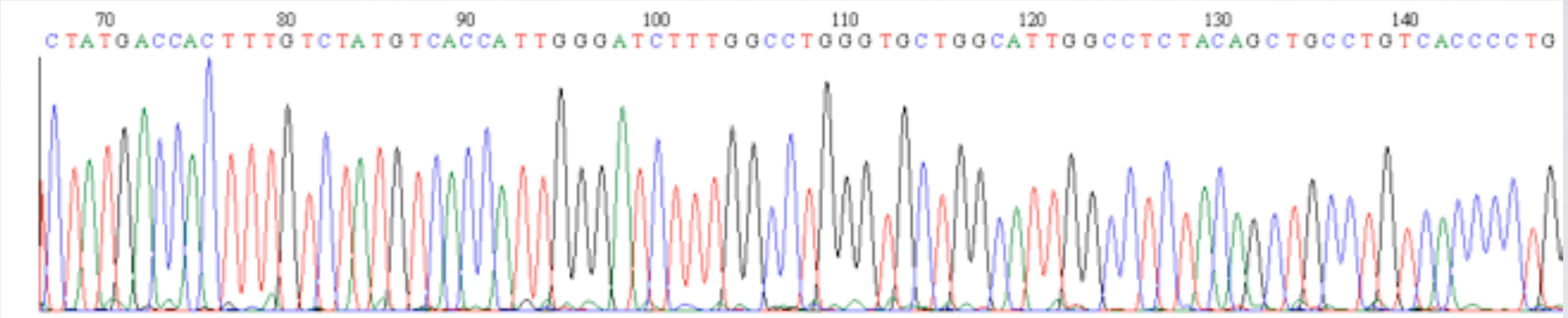
Illumina quality code: ASCII character code = phred quality



Quality Value	Error Probability	Probability Called Base is Correct
10	0.1	0.9
20	0.01	0.99
30	0.001	0.999
40	0.0001	0.9999

$q=-10\log_{10}(p)$

QUALITY SCORE



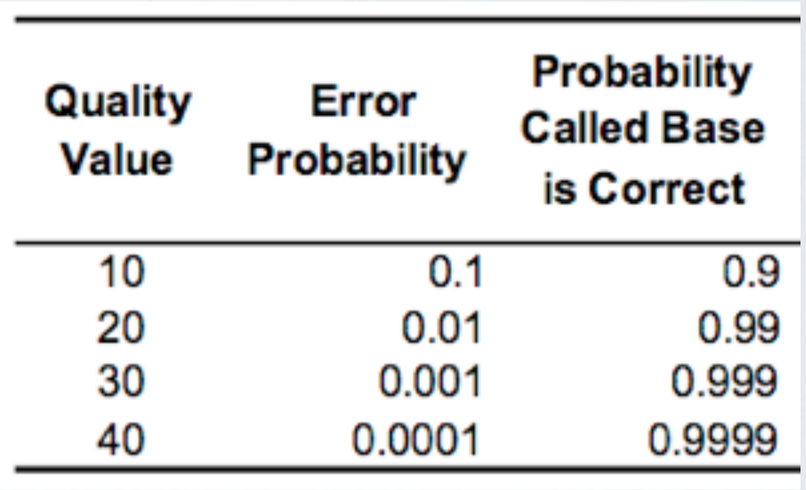
Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

```
@HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
TAAAGAAACTAAGAATAAGCAGATTATCTCGTAT
+HWIEAS210R_0008:6:1:1600:1545#NNCANC/1
ffffffdaadKccaccfffefdceffefefe`b`
```

Sanger quality code (Phred): ASCII character code = phred

Illumina quality code: ASCII character code = phred quality



Quality Value	Error Probability	Probability Called Base is Correct
10	0.1	0.9
20	0.01	0.99
30	0.001	0.999
40	0.0001	0.9999

Illumina: `f` (ASCII 102) => 102 - 64 = 38

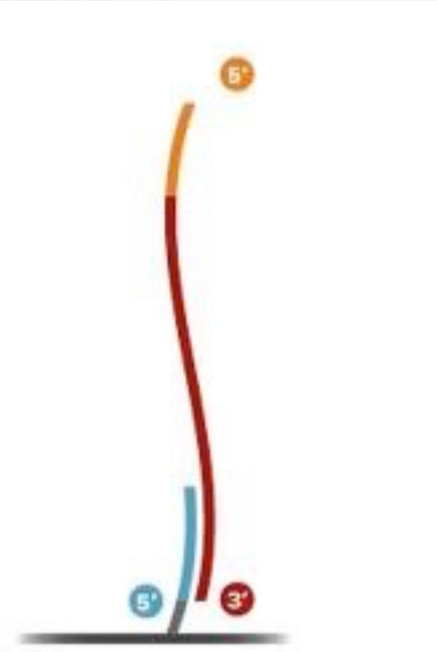
Phred: `f` (ASCII 102) => 102 - 33 = 69

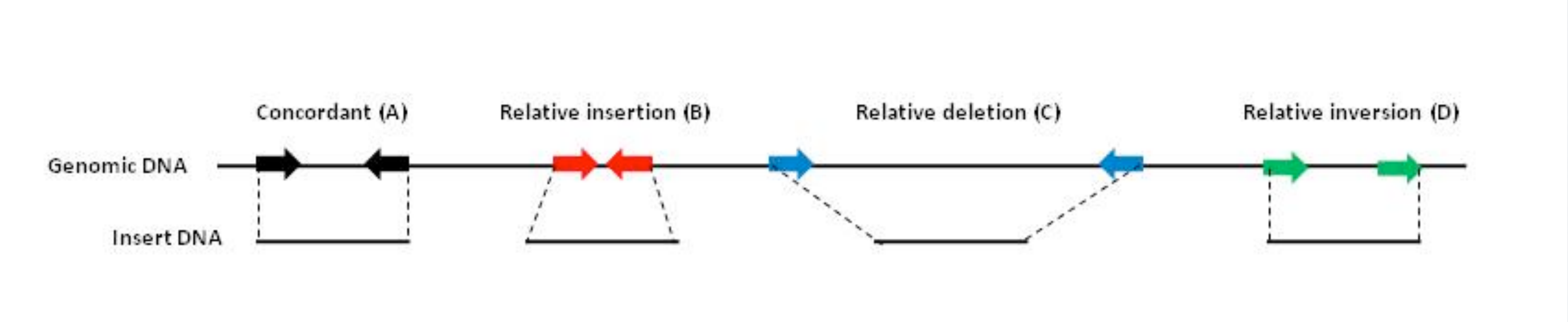
Illumina: ` (ASCII 96) => 96 - 64 = 32

$q = -10\log_{10}(p)$

QUALITY SCORE

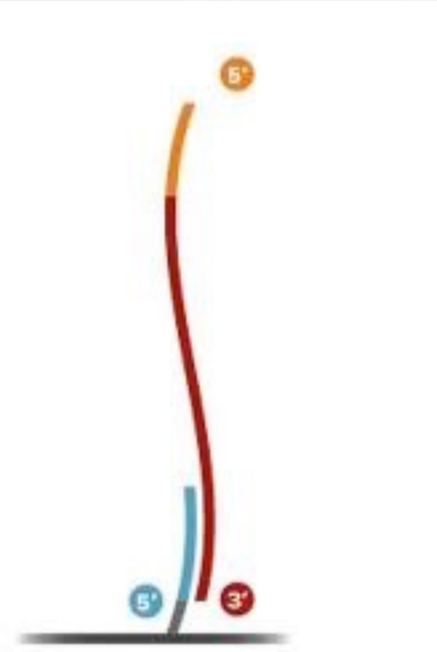
Sequencing of the two end of the same DNA fragment

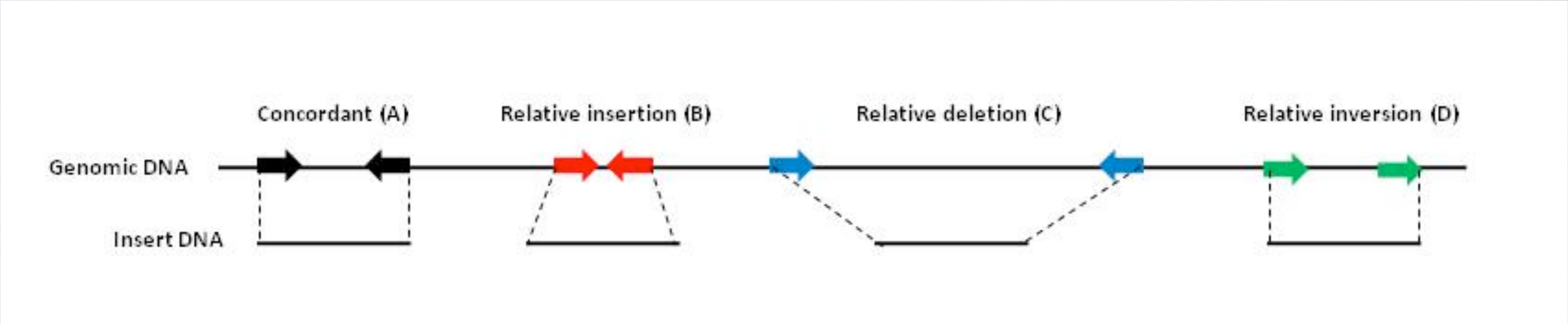


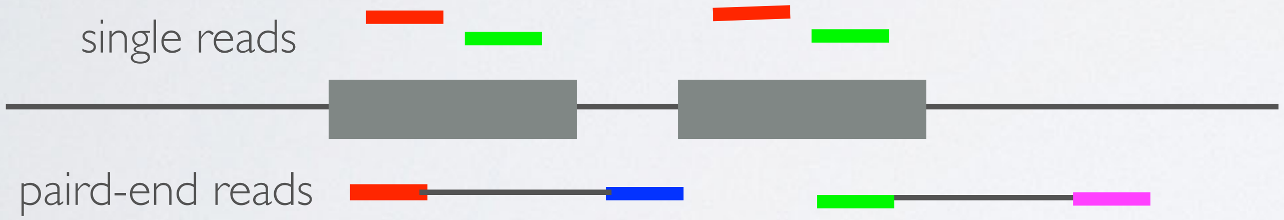


PAIRED-END

Sequencing of the two end of the same DNA fragment



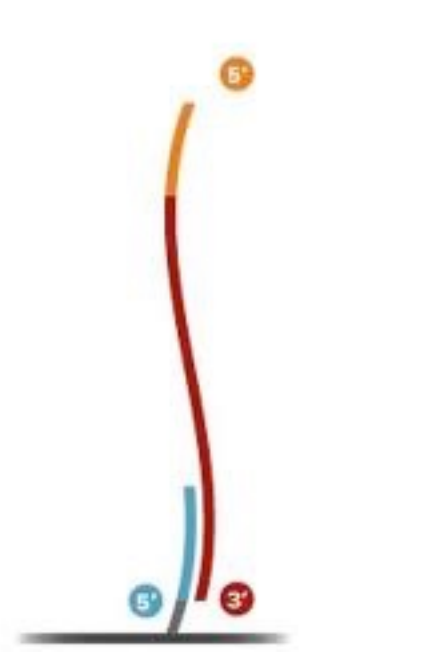


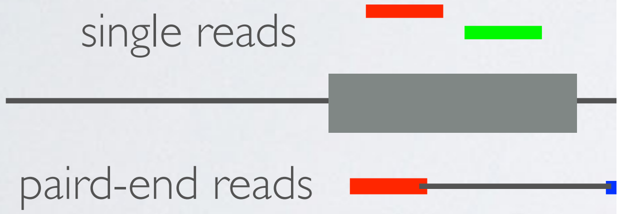


Sequence repeates

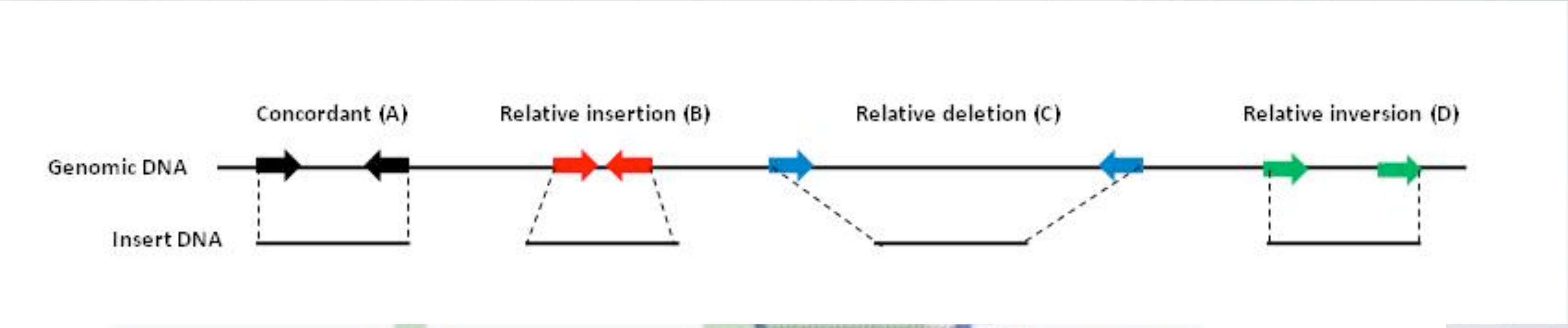
PAIRED-END

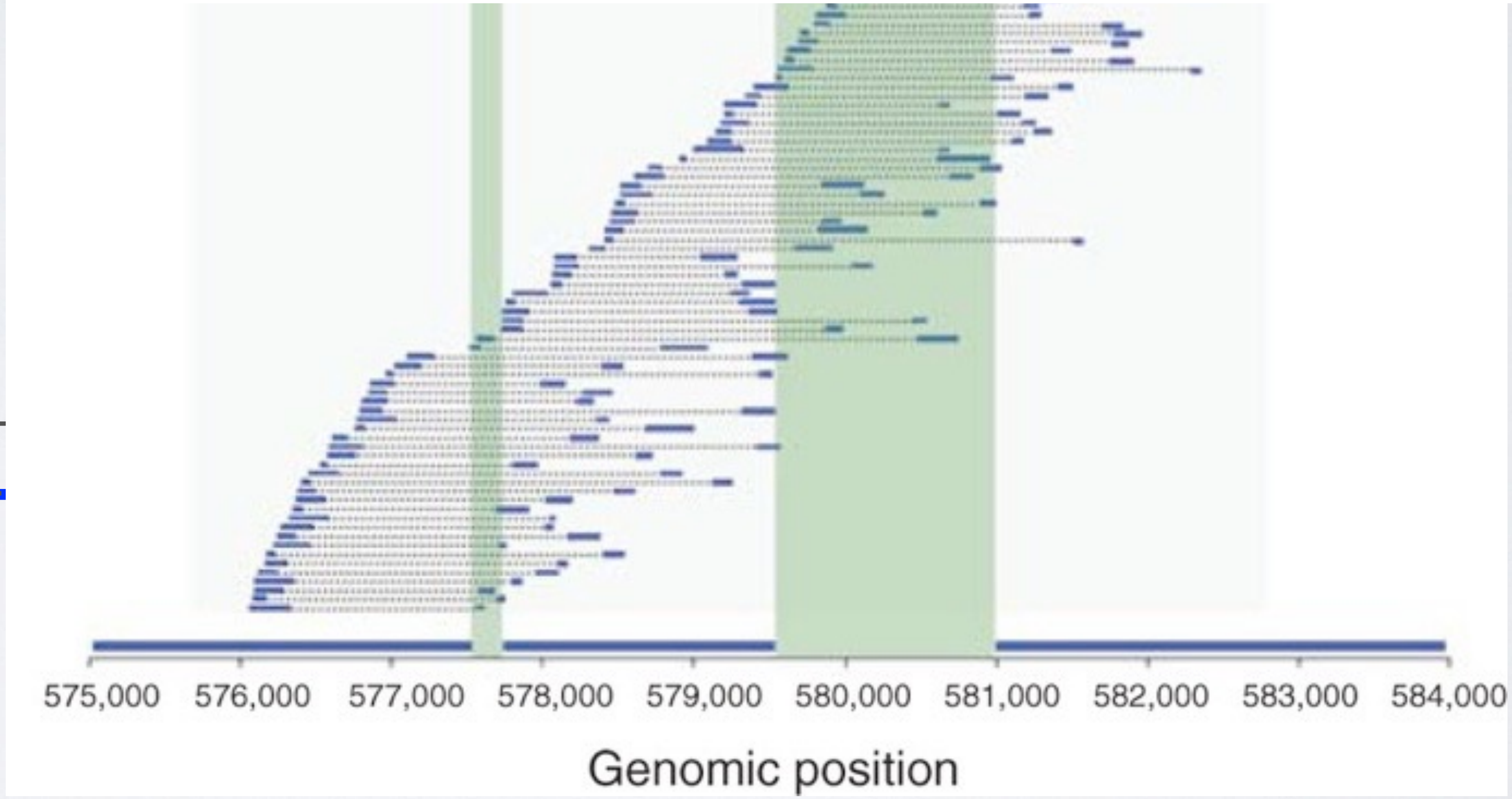
Sequencing of the two end of the same DNA fragment



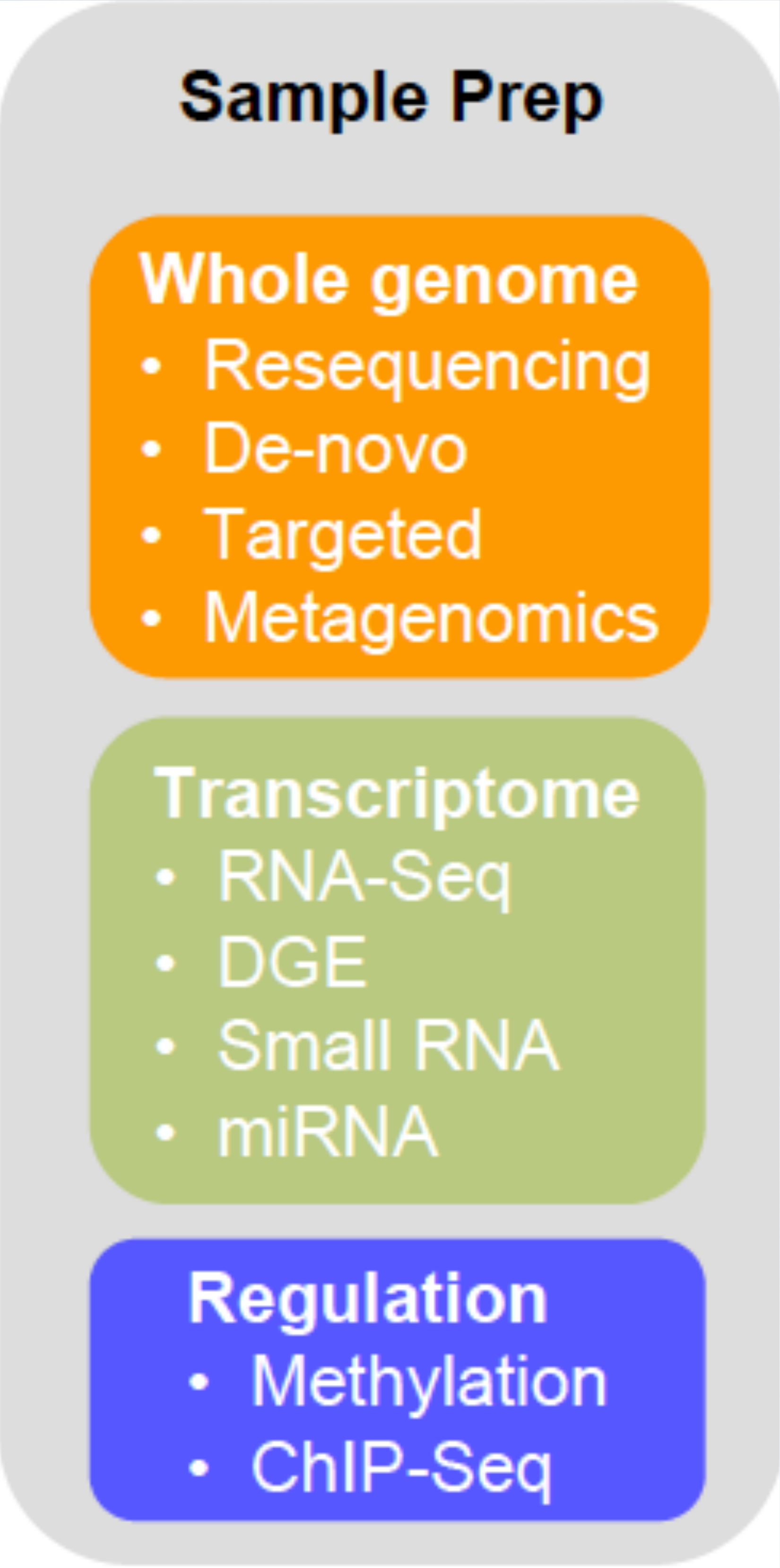


Sequence





PAIRED-END



Sample Prep

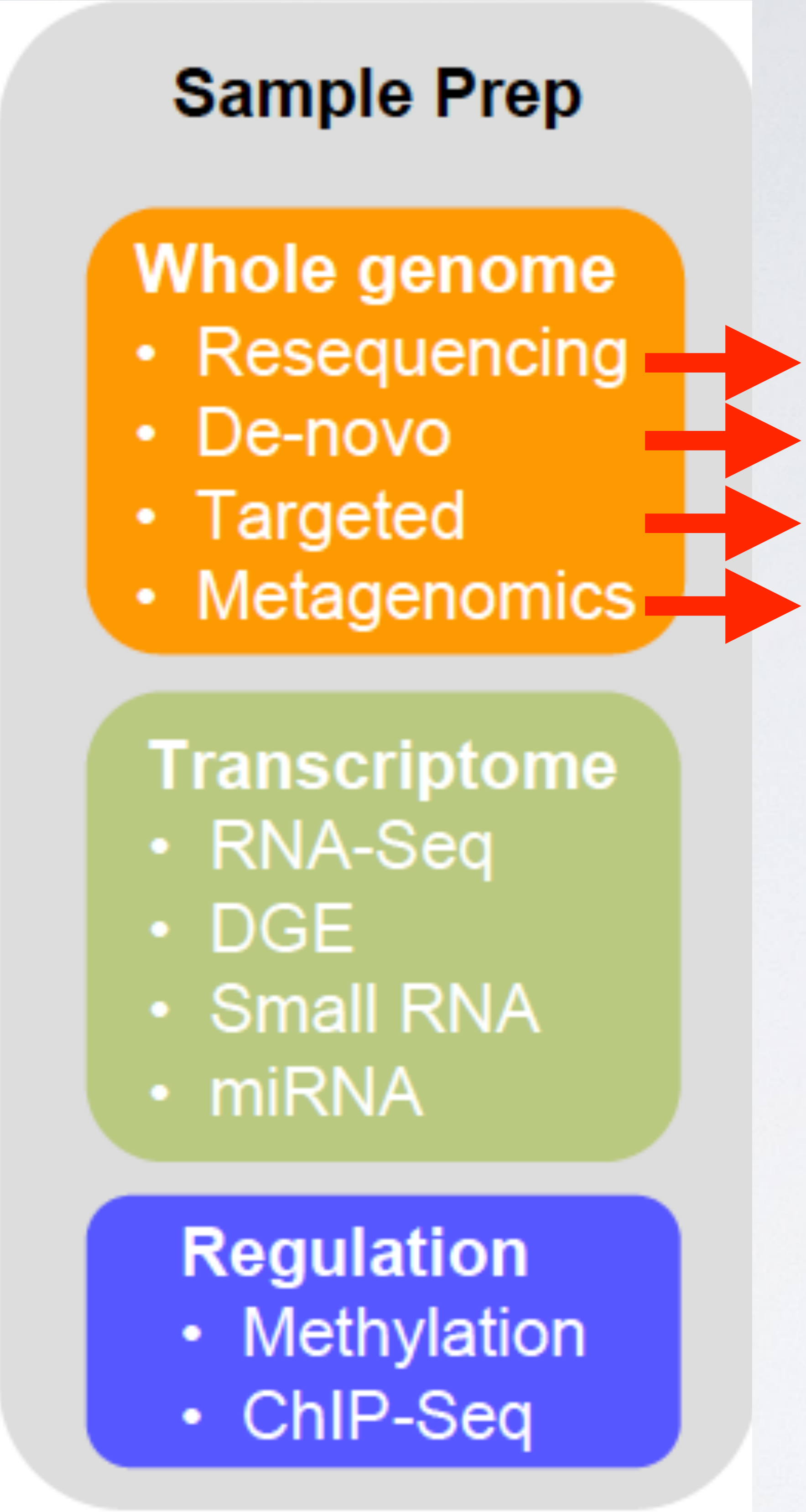
Whole genome
- Resequencing
- De-novo
- Targeted
- Metagenomics

Transcriptome
- RNA-Seq
- DGE
- Small RNA
- miRNA

Regulation
- Methylation
- ChIP-Seq

APPLICATIONS

APPLICATIONS



Sample Prep

Whole genome

- Resequencing
- De-novo
- Targeted
- Metagenomics

Transcriptome

- RNA-Seq
- DGE
- Small RNA
- miRNA

Regulation

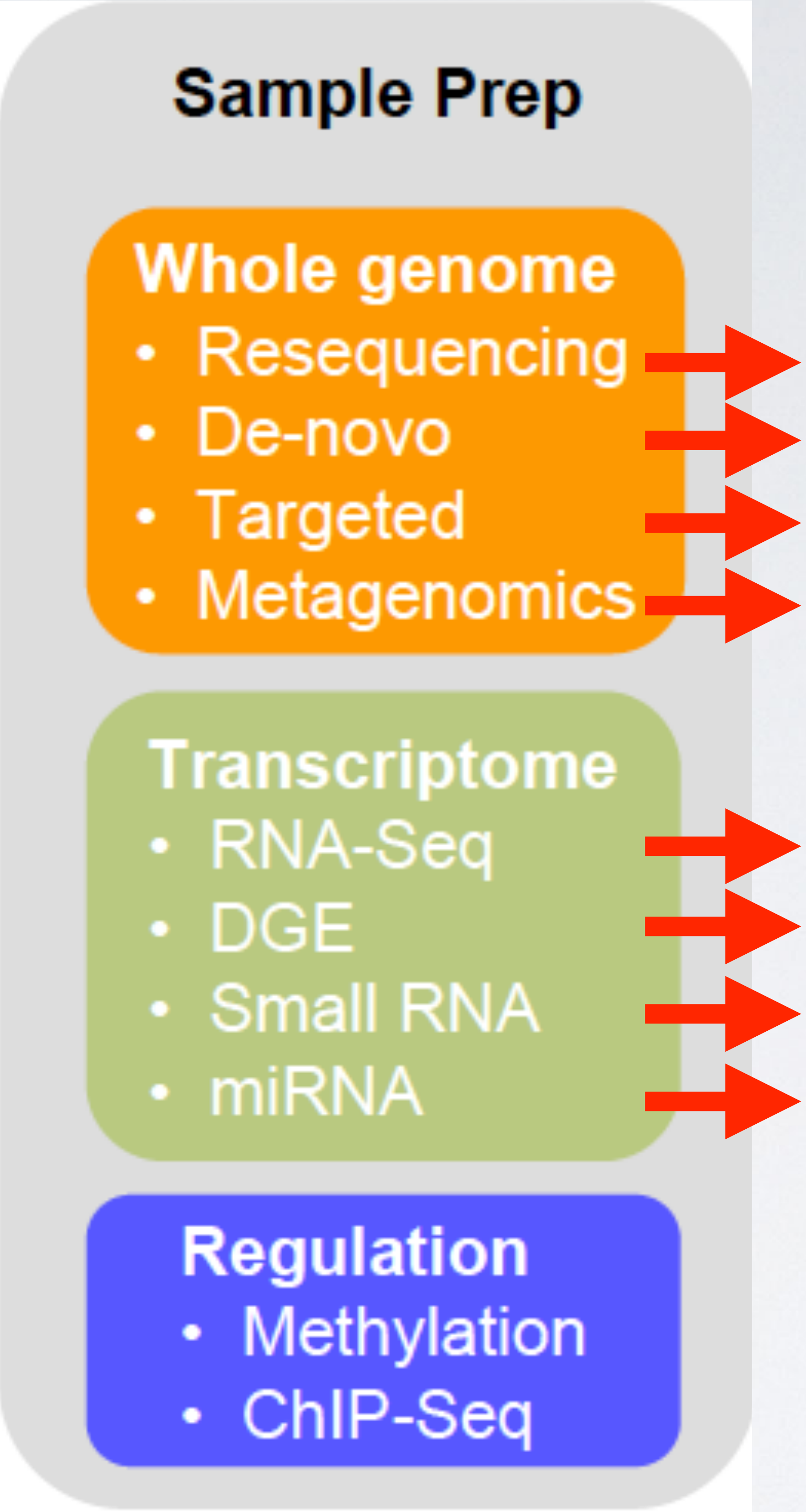
- Methylation
- ChIP-Seq

sequence variations such as SNP, CNV, inserts, deletions, reversions

new genomes

sequence variations with higher coverage

environmental studies, community studies



Sample Prep

Whole genome
- Resequencing
- De-novo
- Targeted
- Metagenomics

Transcriptome
- RNA-Seq
- DGE
- Small RNA
- miRNA

Regulation
- Methylation
- ChIP-Seq

sequence variations such as SNP, CNV, inserts, deletions, reversions

new genomes

sequence variations with higher coverage

environmental studies, community studies

transcriptomics, splicing variants,

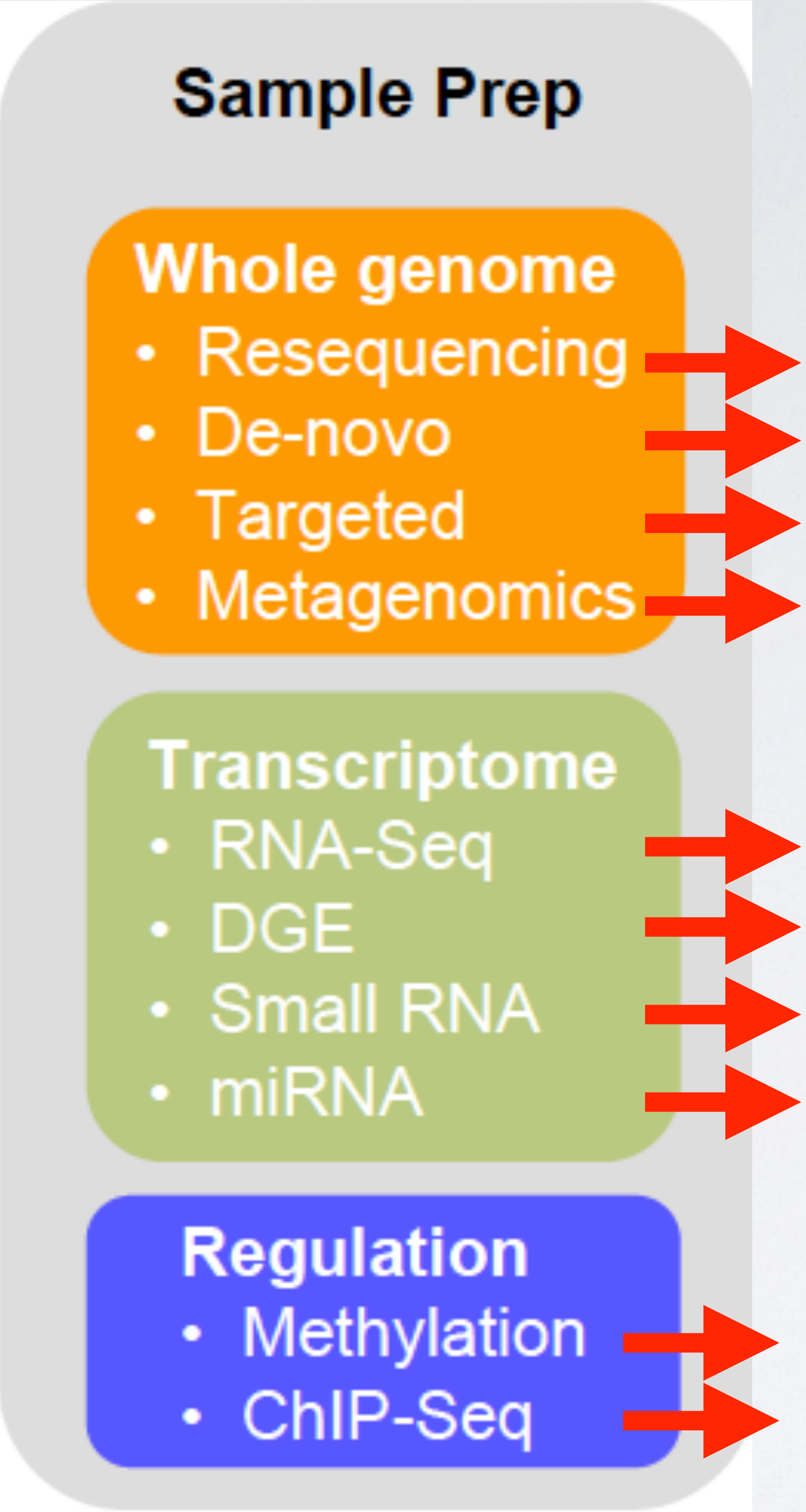
digital gene expression

non-coding RNA research

miRNA induced regulations

APPLICATIONS

APPLICATIONS



Sample Prep

Whole genome
- Resequencing
- De-novo
- Targeted
- Metagenomics

Transcriptome
- RNA-Seq
- DGE
- Small RNA
- miRNA

Regulation
- Methylation
- ChIP-Seq

sequence variations such as SNP, CNV, inserts, deletions, reversions

new genomes

sequence variations with higher coverage

environmental studies, community studies

transcriptomics, splicing variants,

digital gene expression

non-coding RNA research

miRNA induced regulations

epigenetics, methylation induced regulations

protein DNA interactions such as TFBS, histon, polymerase

□ Quality control

□ Mapping
□ Assembly
□ Digital gene expression

DATA ANALYSIS

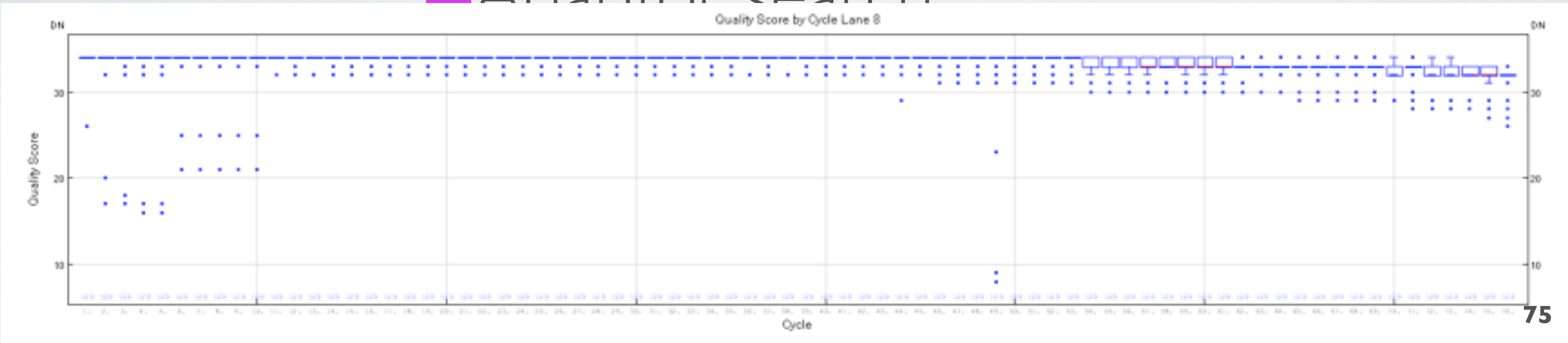
- □ Quality control
  - Quality score distribution
  - Sequence size distribution
  - Sequence coverage
  - Adaptor search

- □ Mapping
- □ Assembly
- □ Digital gene expression

DATA ANALYSIS

□ Quality control

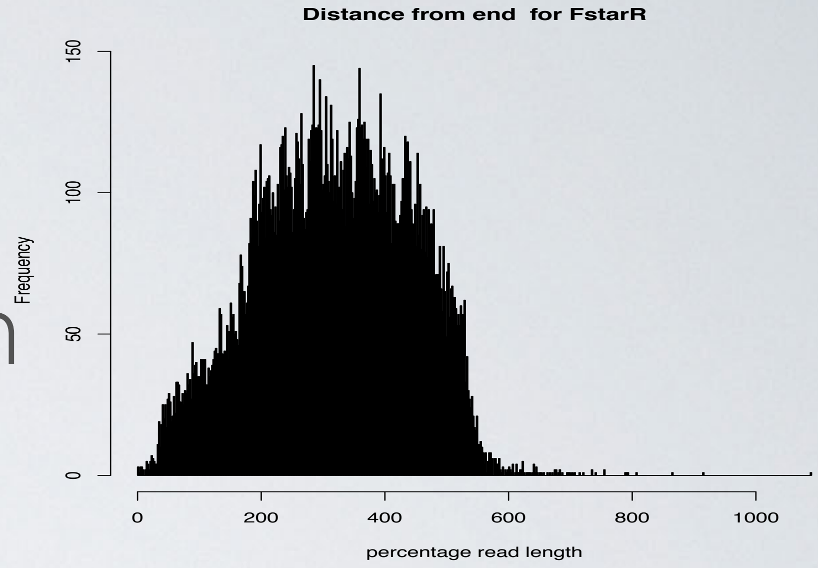
- Quality score distribution
- Sequence size distribution
- Sequence coverage
- Adaptor search

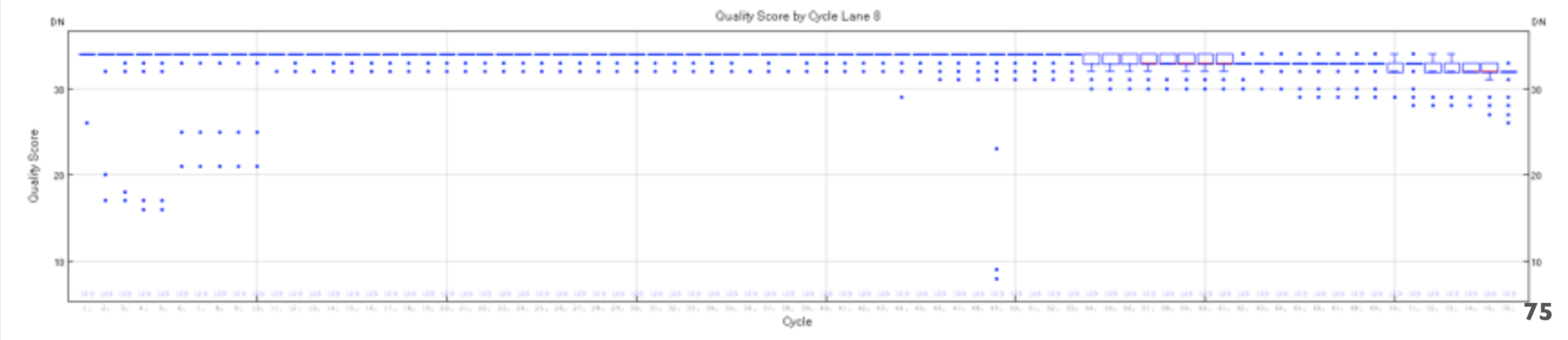


□ Digital gene expression

DATA ANALYSIS

- Quality control
  - Quality score distribution
  - Sequence size distribution
  - Sequence coverage
  - Adaptor search

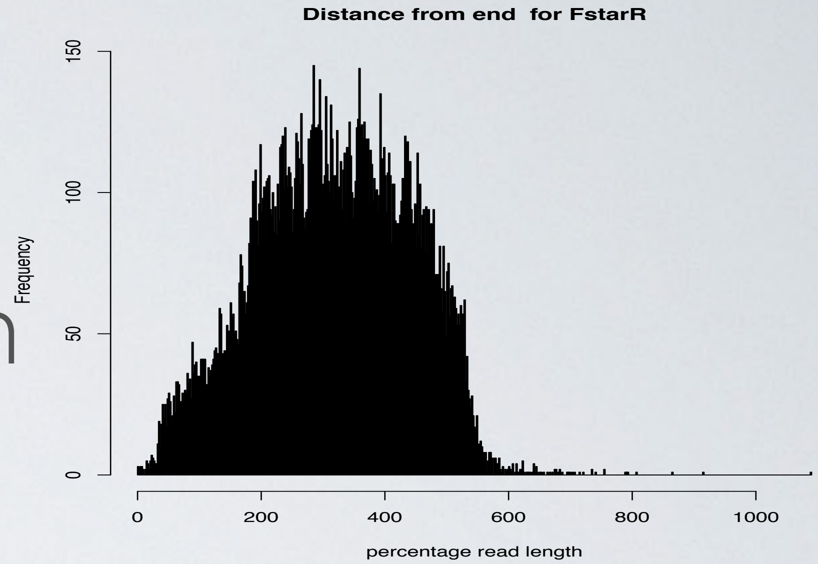


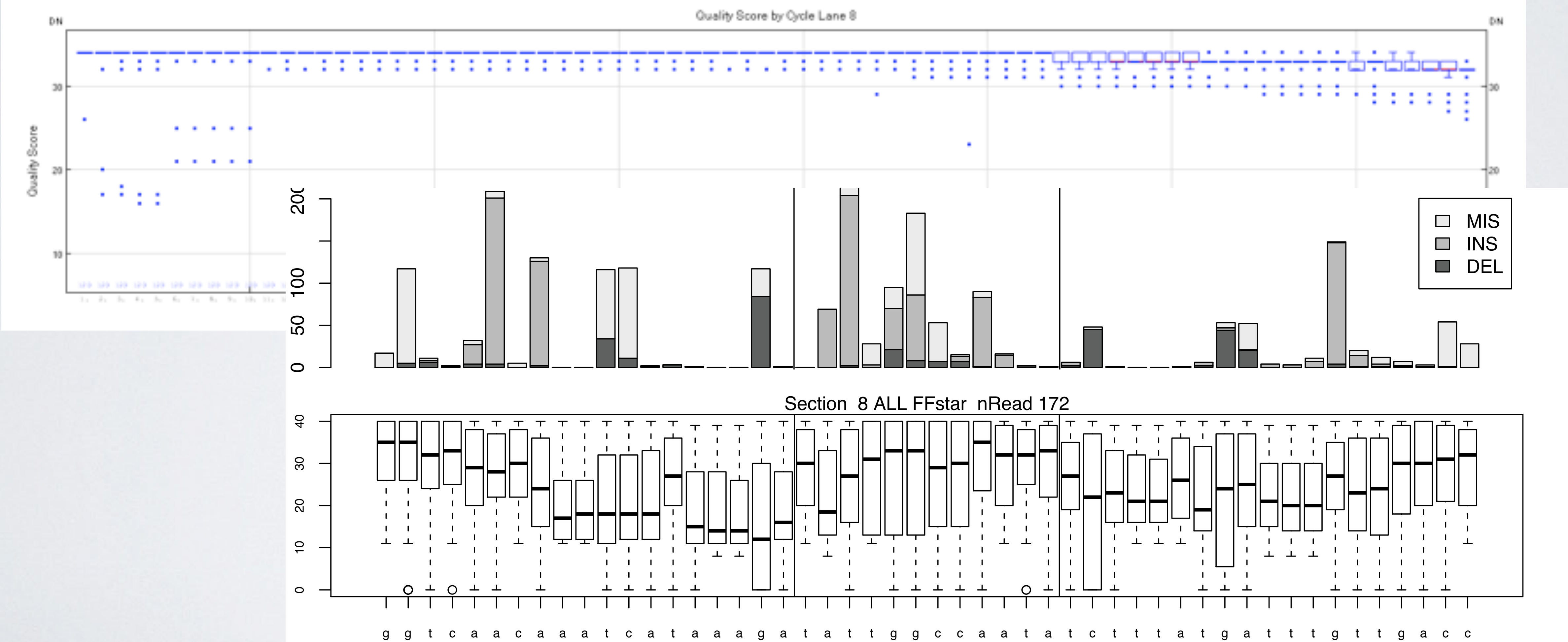


- Digital gene expression

DATA ANALYSIS

- Quality control
  - Quality score distribution
  - Sequence size distribution
  - Sequence coverage
  - Adaptor search





DATA ANALYSIS

- □ Quality control

- □ Mapping
  - Aligning short sequences on a reference
- □ Assembly
- □ Digital gene expression

DATA ANALYSIS

- □ Quality control
- □ Mapping
  - Aligning short sequenc
- □ Assembly
- □ Digital gene expression

DATA ANALYSIS

□ Quality control

□ Mapping
□ Assembly
□ Digital gene expression
□ Visualisation

DATA ANALYSIS

□ Quality control

□ Mapping

□ Assembly

- Assemble sequencing reads to recover the sequence in investigation

□ Digital gene expression

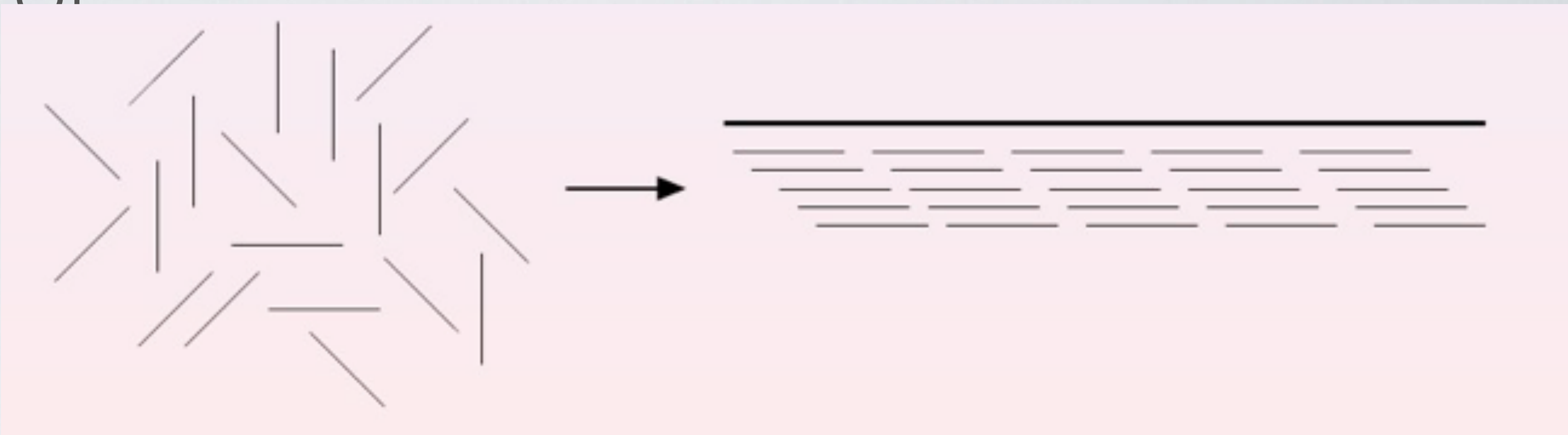
□ Visualisation

DATA ANALYSIS

□ Quality control

□ Mapping

□ Assembly



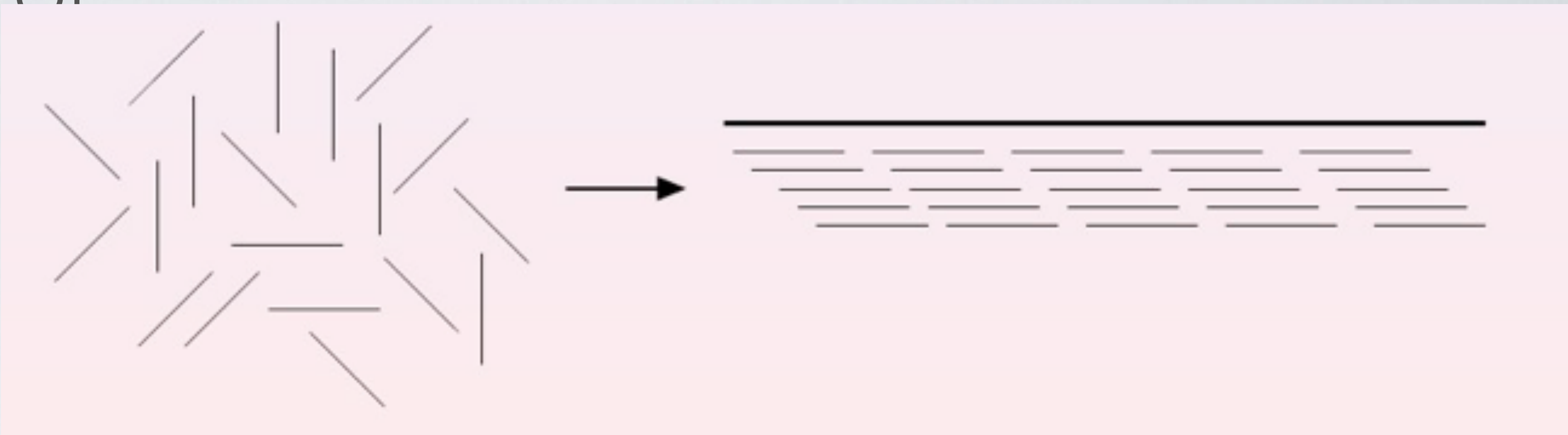
- Assemble sequencing reads to recover the sequence in investigation

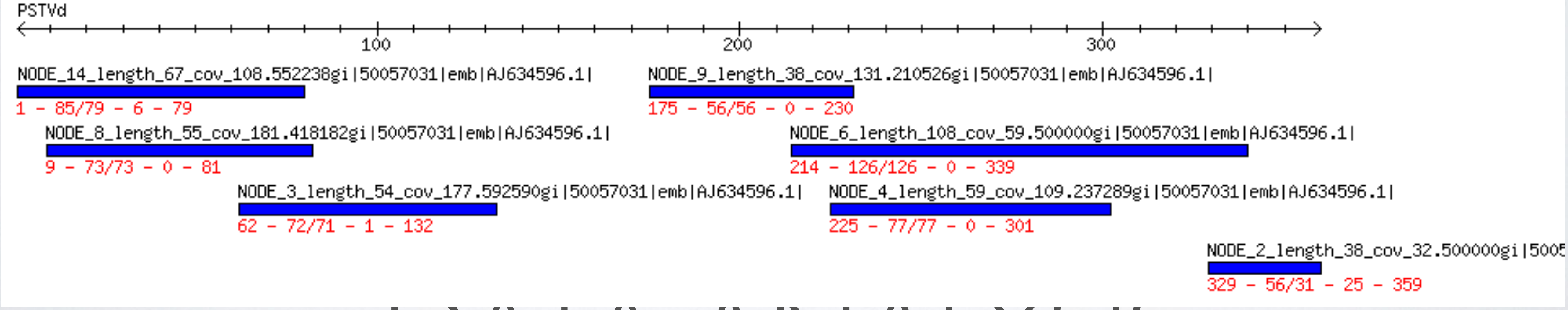
□ Digital gene expression

□ Visualisation

DATA ANALYSIS

- □ Quality control
- □ Mapping
- □ Assembly
  - Assemble sequencing reads to recover the sequence in investigation
- □ Digital gene expression
- □ Visualisation





PSTVd

100 200 300

NODE_14_length_67_cov_108.552238gi|50057031|emb|AJ634596.1|
1 - 85/79 - 6 - 79

NODE_9_length_38_cov_131.210526gi|50057031|emb|AJ634596.1|
175 - 56/56 - 0 - 230

NODE_8_length_55_cov_181.418182gi|50057031|emb|AJ634596.1|
9 - 73/73 - 0 - 81

NODE_6_length_108_cov_59.500000gi|50057031|emb|AJ634596.1|
214 - 126/126 - 0 - 339

NODE_3_length_54_cov_177.592590gi|50057031|emb|AJ634596.1|
62 - 72/71 - 1 - 132

NODE_4_length_59_cov_109.237289gi|50057031|emb|AJ634596.1|
225 - 77/77 - 0 - 301

NODE_2_length_38_cov_32.500000gi|5005
329 - 56/31 - 25 - 359

DATA ANALYSIS

- □ Quality control

- □ Mapping
- □ Alignment
- □ Digital gene expression
  - Compare differentially expressed sequences using there frequency witin the data set
- □ Visualisation

DATA ANALYSIS

□ Qua

□ Map

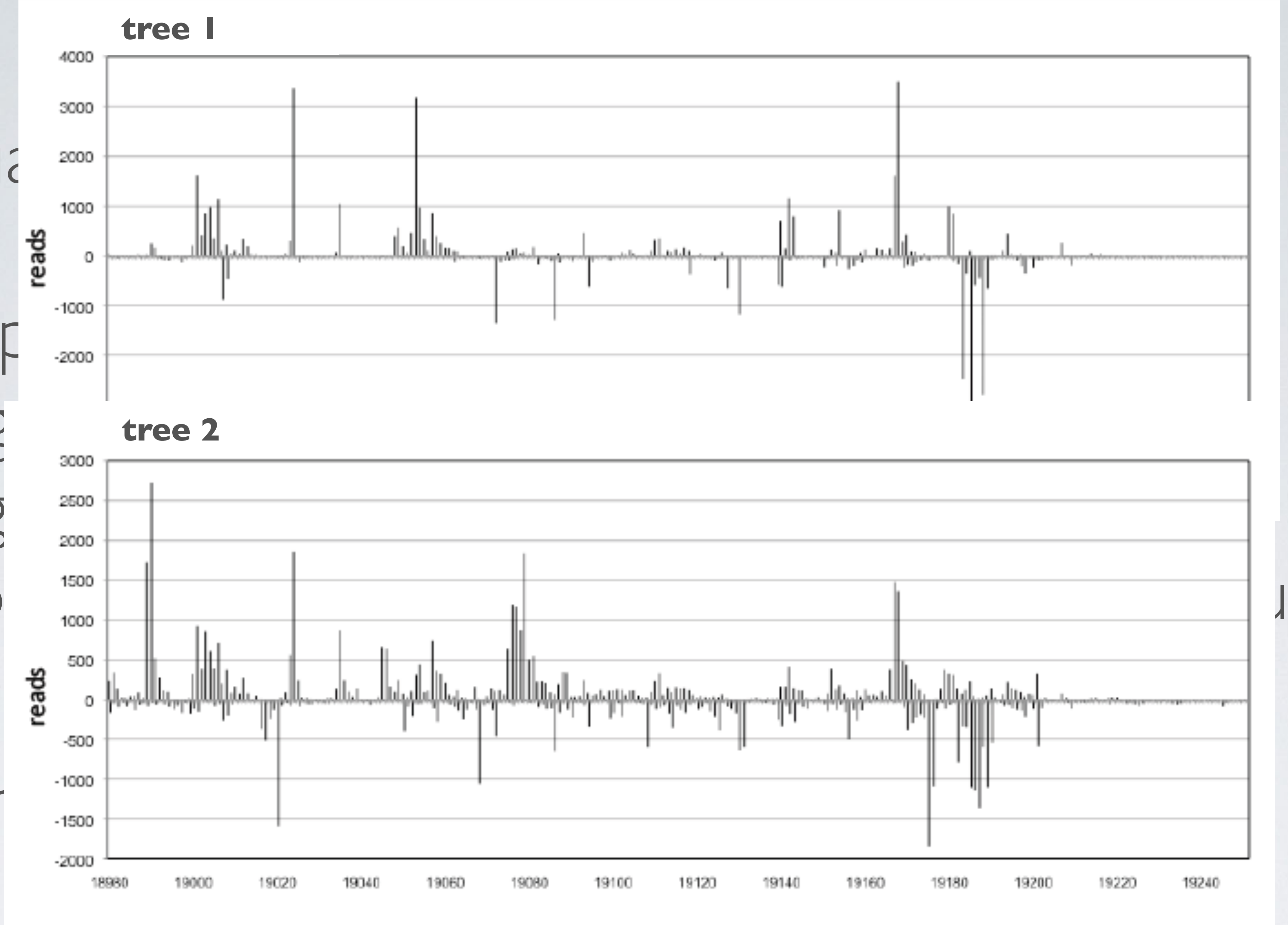
□ Alig

□ Dig

- Co using

there

□ Visu



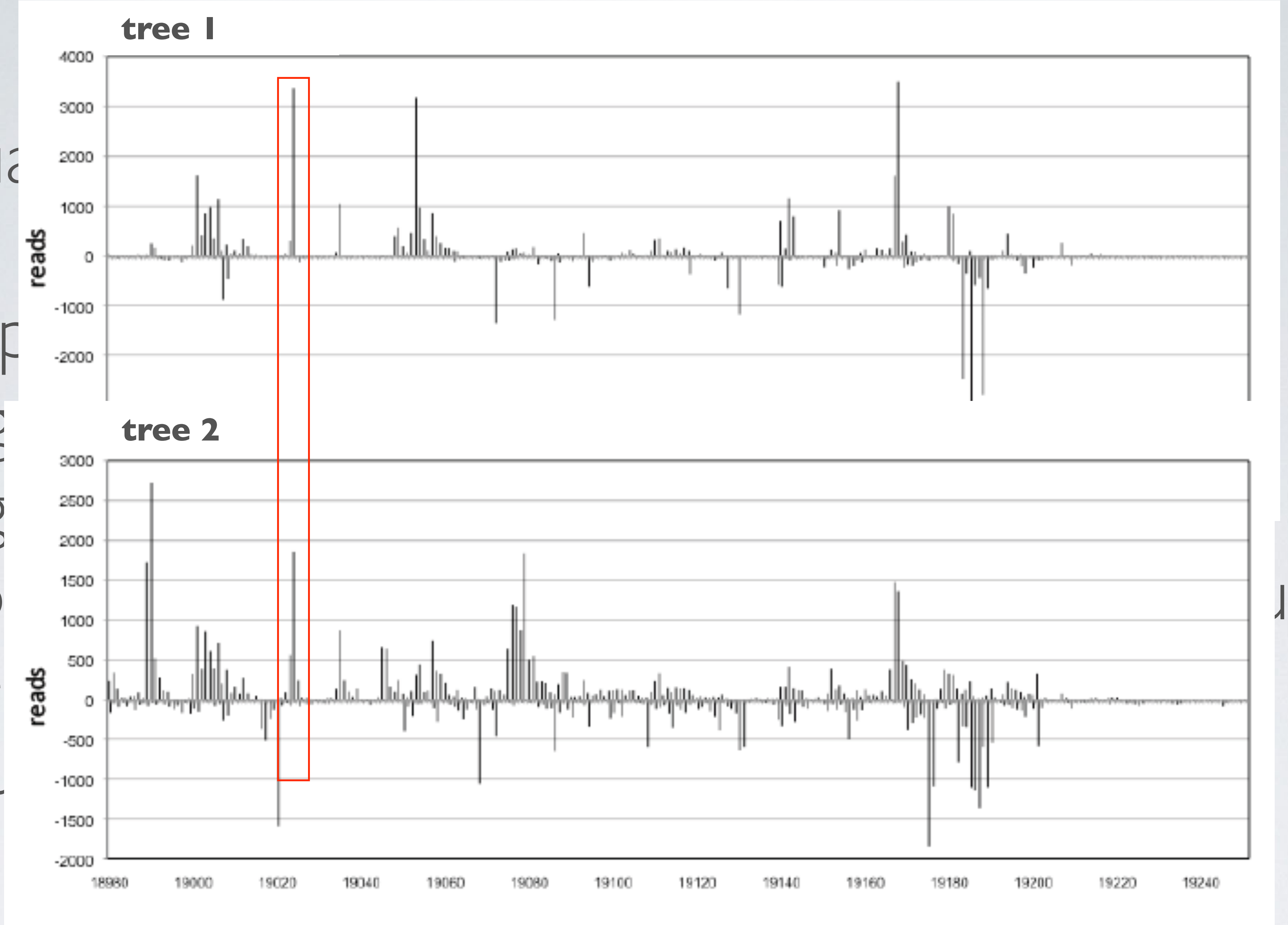
DATA ANALYSIS

DATA ANALYSIS

□ Qua

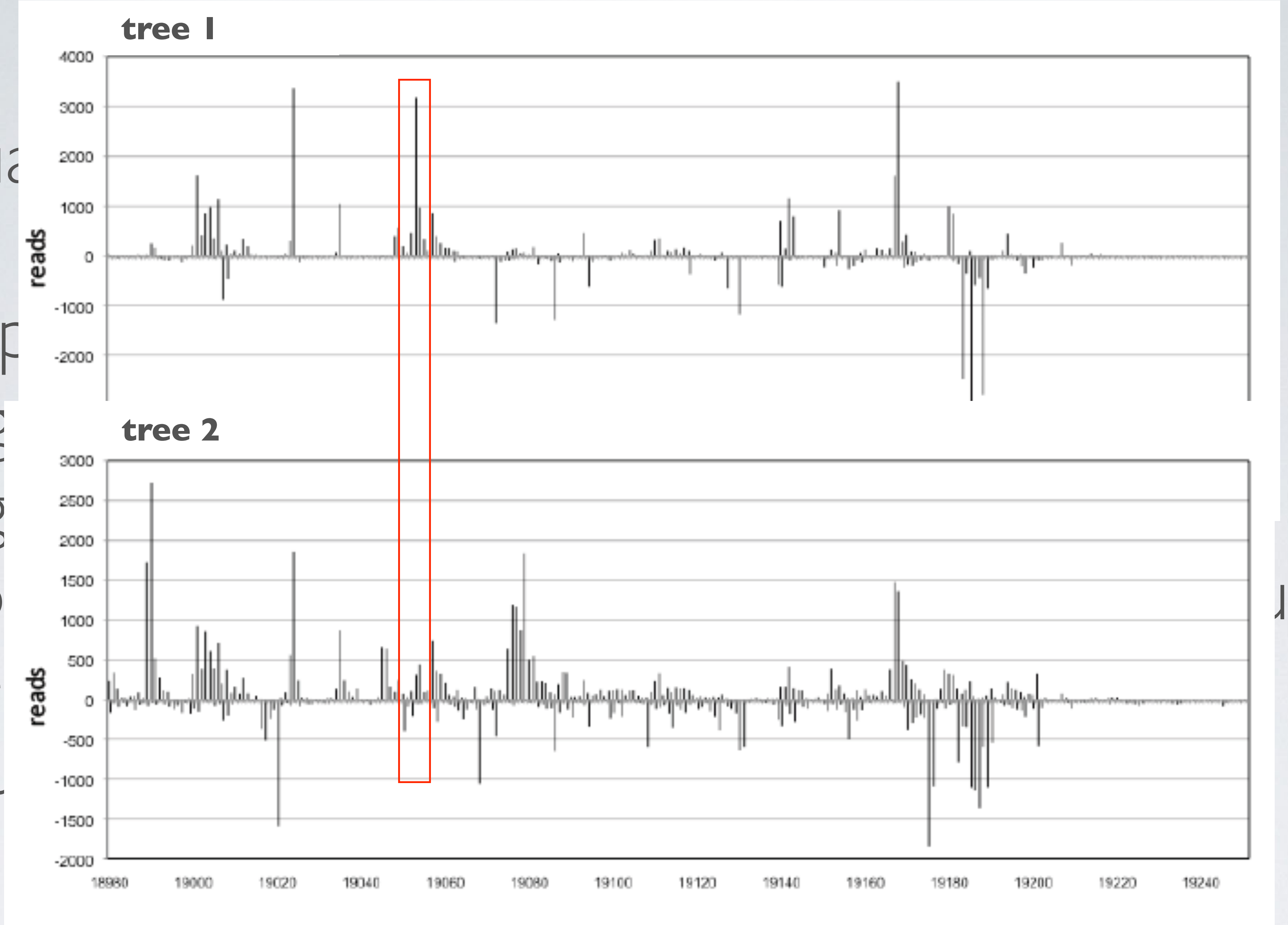
□ Map
□ Alig
□ Dig
- Co   using there

□ Visu



□ Qua

□ Map
□ Alig
□ Dig
- Co using
there
□ Visu



DATA ANALYSIS

□ Qua

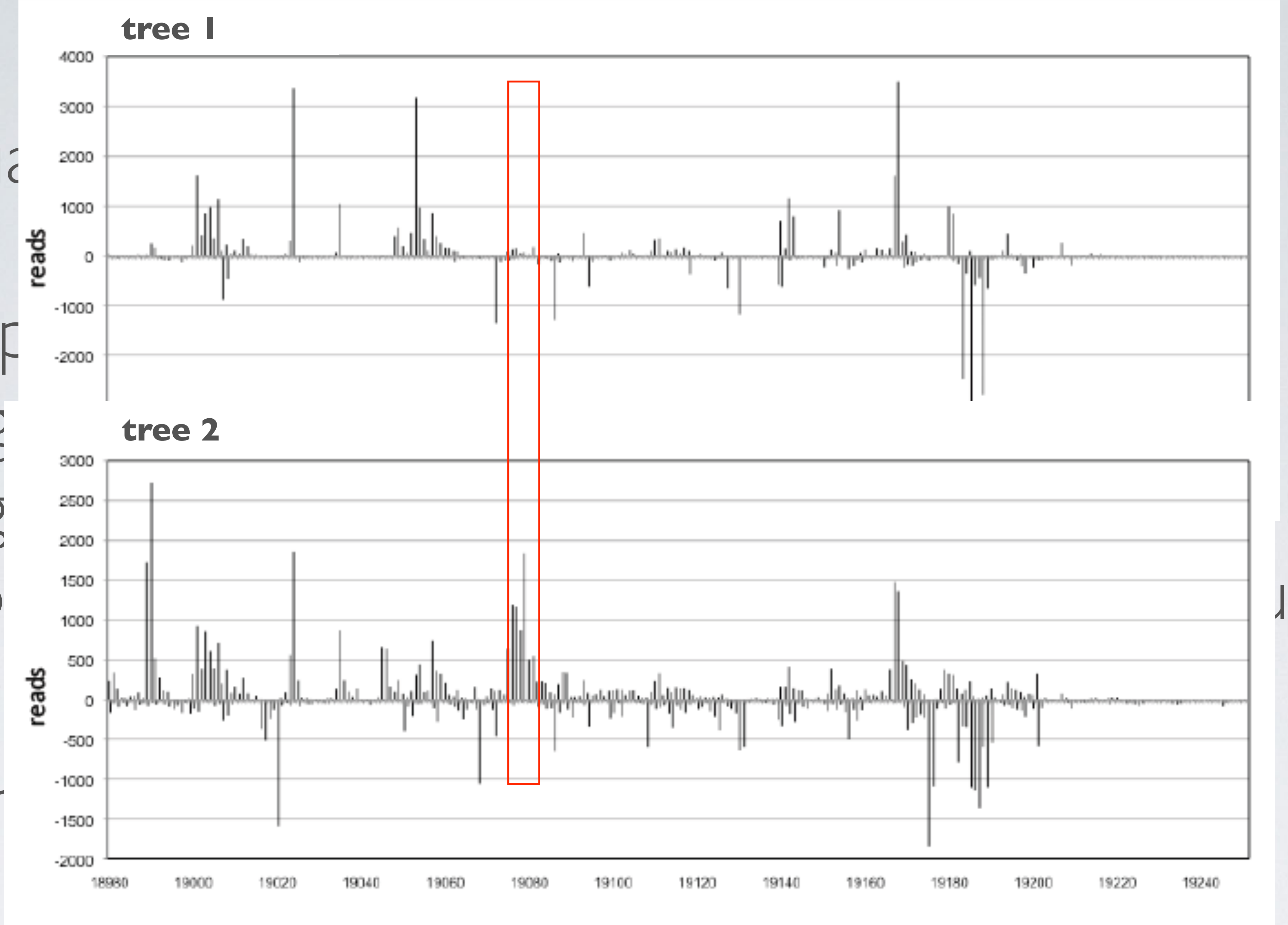
□ Map

□ Alig

□ Dig

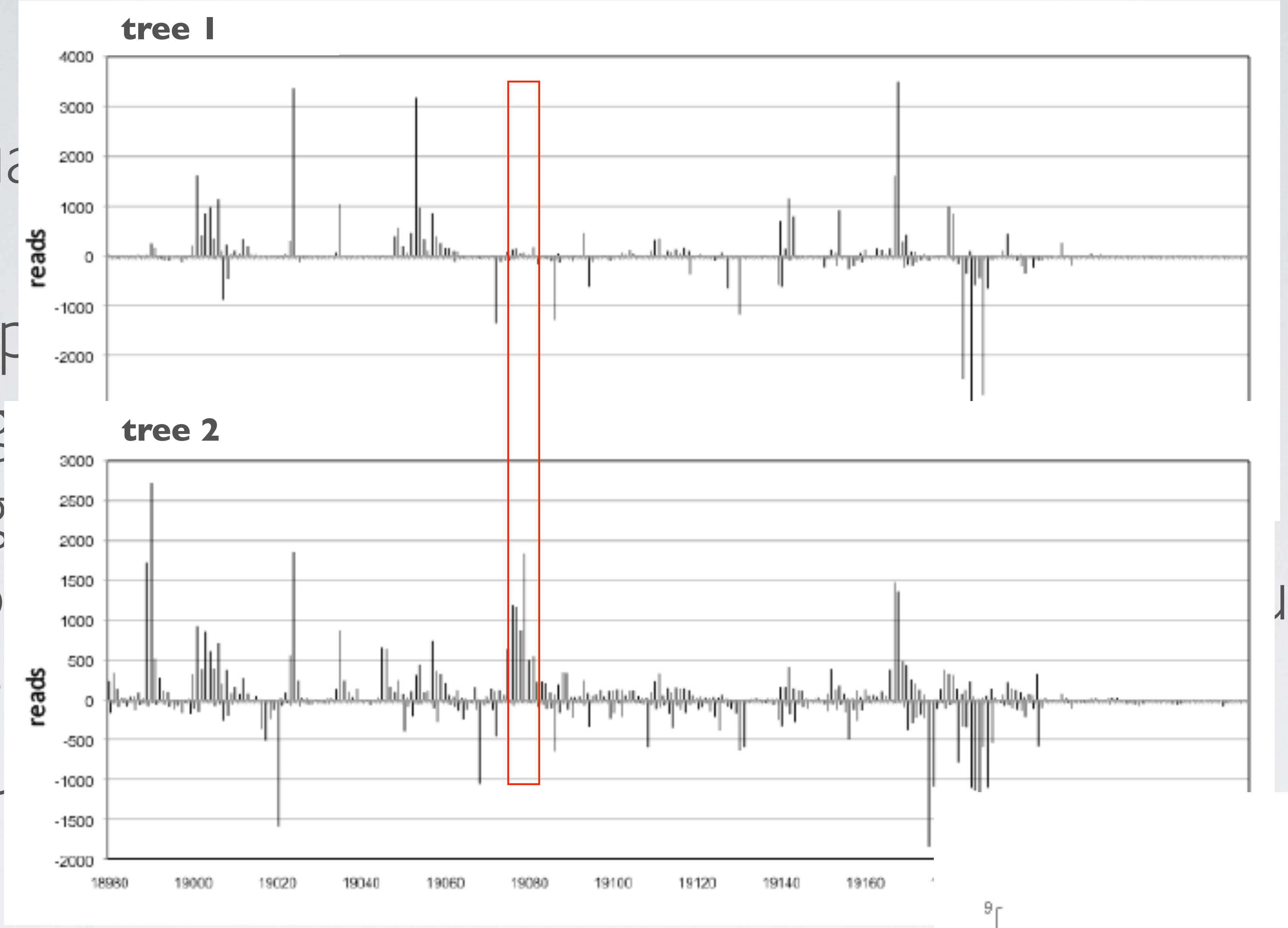
- Co using there

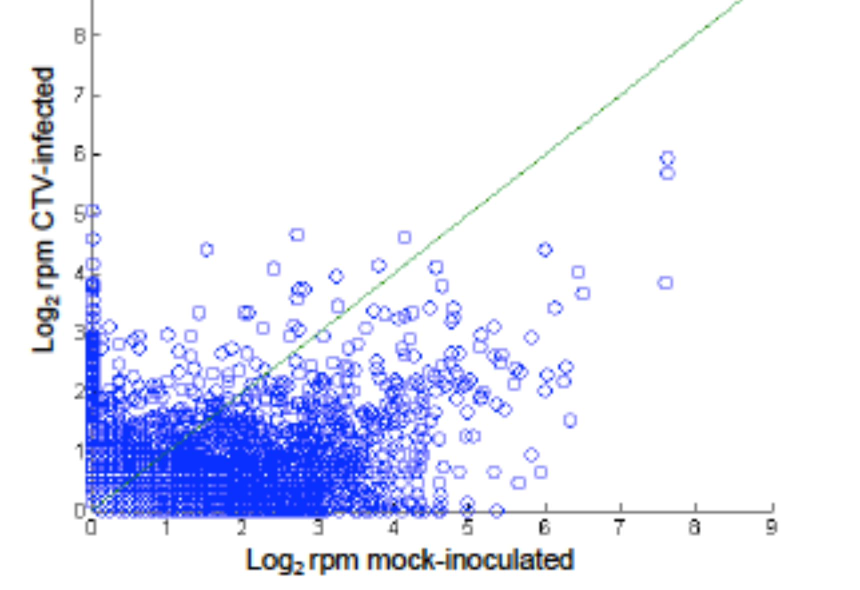
□ Visu



DATA ANALYSIS

- Qua
- Map
- Alig
- Dig
  - Co … using there
- Visu





DATA ANALY

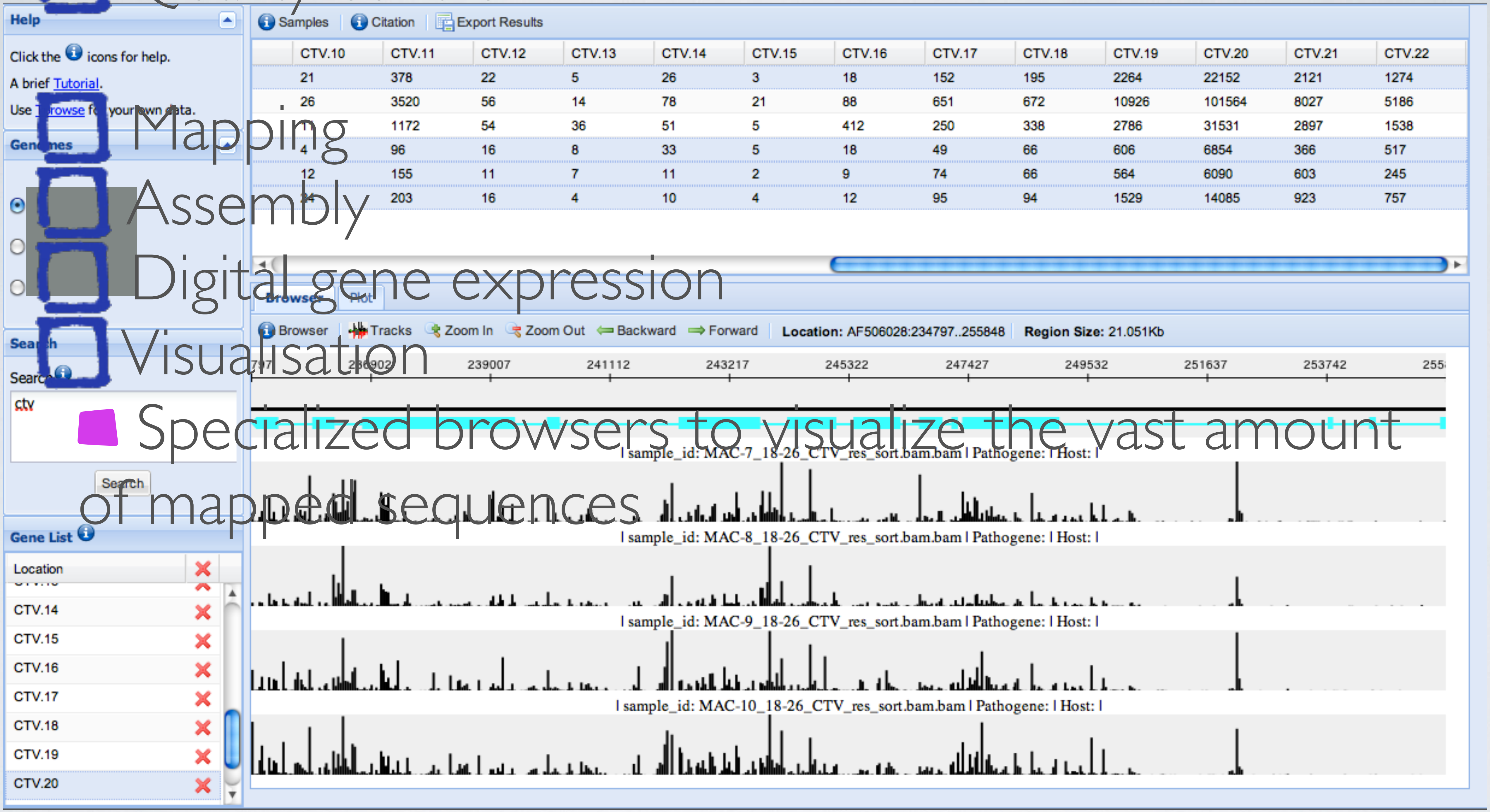
- □ Quality control

- □ Mapping
- □ Assembly
- □ Digital gene expression
- □ Visualisation
  - Specialized browsers to visualize the vast amount of mapped sequences

DATA ANALYSIS

DATA ANALYSIS

- □ Quality control
- □ Mapping
- □ Assembly
- □ Digital gene expression
- □ Visualisation
  - Specialized browsers to visualize the vast amount of mapped sequences



Help

Click the icons for help.

A brief Tutorial.

Use Browse for your own data.

Genomes

Search

Search

ctv

Search

Gene List

Location
CTV.14
CTV.15
CTV.16
CTV.17
CTV.18
CTV.19
CTV.20

Samples | Citation | Export Results

	CTV.10	CTV.11	CTV.12	CTV.13	CTV.14	CTV.15	CTV.16	CTV.17	CTV.18	CTV.19	CTV.20	CTV.21	CTV.22
	21	378	22	5	26	3	18	152	195	2264	22152	2121	1274
	26	3520	56	14	78	21	88	651	672	10926	101564	8027	5186
	11	1172	54	36	51	5	412	250	338	2786	31531	2897	1538
	4	96	16	8	33	5	18	49	66	606	6854	366	517
	12	155	11	7	11	2	9	74	66	564	6090	603	245
	24	203	16	4	10	4	12	95	94	1529	14085	923	757

Browser | Plot

Browser | Tracks | Zoom In | Zoom Out | Backward | Forward | **Location:** AF506028:234797..255848 | **Region Size:** 21.051Kb

| sample_id: MAC-7_18-26_CTV_res_sort.bam.bam | Pathogene: | Host: |

| sample_id: MAC-8_18-26_CTV_res_sort.bam.bam | Pathogene: | Host: |

| sample_id: MAC-9_18-26_CTV_res_sort.bam.bam | Pathogene: | Host: |

| sample_id: MAC-10_18-26_CTV_res_sort.bam.bam | Pathogene: | Host: |

TOOLS

□ Quality control

□ Mapping
□ Assembly
□ Digital gene expression
□ Visualisation

TOOLS

TOOLS

- □ Quality control
  - in most cases incorporated in sequencing platform software
  - GALAXY

- □ Mapping
- □ Assembly
- □ Digital gene expression
- □ Visualisation

TOOLS

□ Quality control

□ Mapping
- read indexing with hash table
- genome indexing with hash table
- genome indexing with suffix array
- SAM/BAM format
- http://lh3lh3.users.sourceforge.net/NGSalign.shtml

□ Assembly

□ Digital gene expression

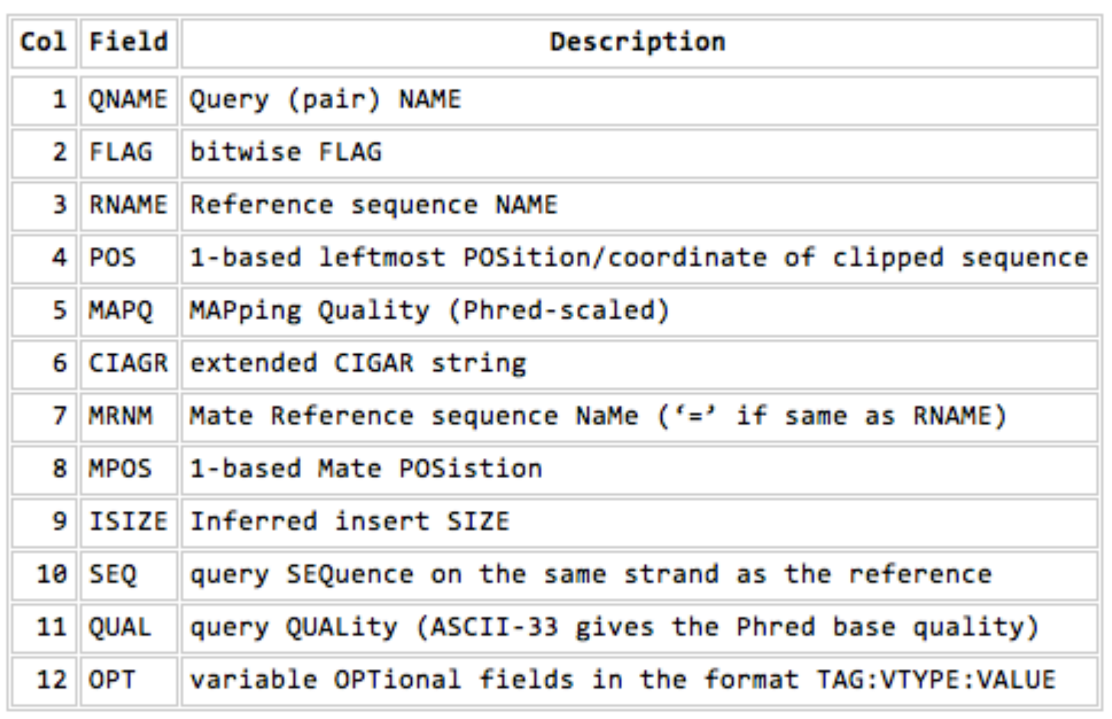
□ Visualisation

TOOLS

SAMTools (http://samtools.sourceforge.net/samtools.shtml)

- SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.
- SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

```
HWIEAS210R_0008:6:1:1118:15625#NNTANG/1_18 0    scaffold_2    19205786 255  18M  *    0    0
AGACCGGTAGACTTGAAC    d\ddd^a``^G_\bT_dd    XA:i:0   MD:Z:18  NM:i:0
```



Col	Field	Description
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

```
samtools view -bt ref_list.txt -o aln.bam aln.sam.gz
samtools sort aln.bam aln.sorted
samtools index aln.sorted.bam
samtools idxstats aln.sorted.bam
samtools view aln.sorted.bam chr2:20,100,000-20,200,000
samtools merge out.bam in1.bam in2.bam in3.bam
samtools faidx ref.fasta
samtools pileup -vcf ref.fasta aln.sorted.bam
samtools mpileup -C50 -gf ref.fasta -r chr3:1,000-2,000 in1.bam in2.bam
samtools tview aln.sorted.bam ref.fasta
```

- BAM Binary version of SAM

□ Quality control

□ Mapping

□ Assembly

- Greedy
- Overlap Layout Consensus (OLC)
- de Bruijn graph based

□ Digital gene expression

□ Visualisation

TOOLS

Assembler

- ### Greedy

  The greedy algorithms apply one basic operation: given any read or contig, add one more contig. The basic operation is repeated until no more operations are possible. Each operation uses the next highest-scoring overlap to make the next join.

- ### Overlap Layout Consensus (OLC)

  step 1 overlap discovery
  step 2 build and use the overlap graph
  step 3 multiple sequence alignment

- ### de Bruijn graph bases

  The de Bruijn graph approach circumvents the problems of overlap consensus assembly. Rather than using the reads 'as is' and trying to link them, the k-mers (all subsequences of length k within the reads) are computed and the reads are represented as a path through the k-mers. Such a paradigm handles redundancy better than the overlap consensus approach and makes the computation of paths more tractable.

TOOLS

Assembler

Greedy

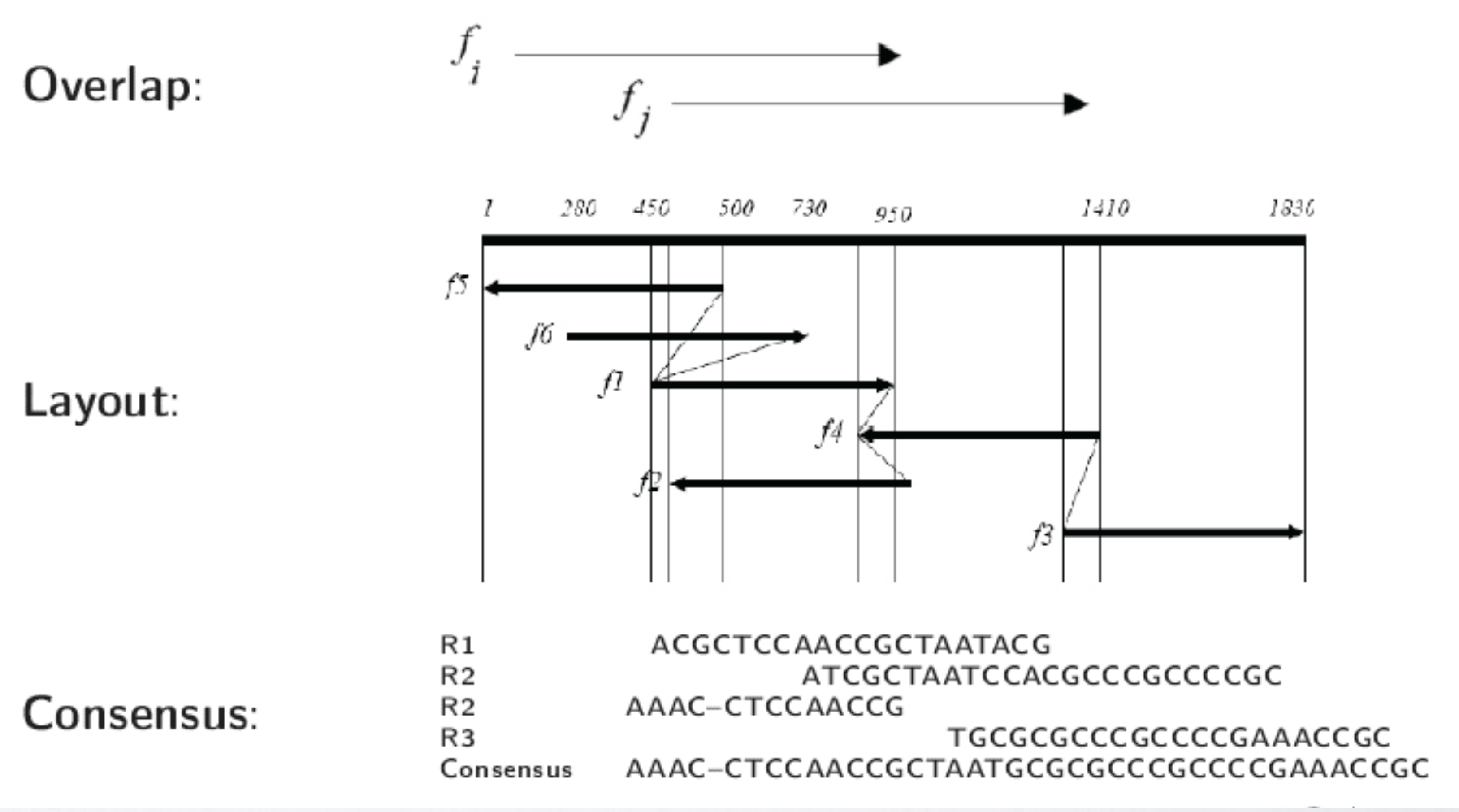
The greedy algorithms apply one basic operation: given any read or contig, add one more contig. The basic operation is repeated until no more operations are possible. Each operation uses the next highest-scoring overlap to make the next join.

Overlap Layout C

step 1 overlap discovery
step 2 build and use the over
step 3 multiple sequence alig

de Bruijn graph b

The de Bruijn graph approach
Rather than using the reads 'a
length k within the reads) are
k-mers. Such a paradigm hand
and makes the computation o



```
R1           ACGCTCCAACCGCTAATACG
R2                     ATCGCTAATCCACGCCCGCCCCGC
R2          AAAC-CTCCAACCG
R3                                TGCGCGCCCGCCCCGAAACCGC
Consensus   AAAC-CTCCAACCGCTAATGCGCGCCCGCCCCGAAACCGC
```

TOOLS

Assembler

Greedy

The greedy algorithms apply one basic operation: given any read or contig, add one more contig. The basic operation is repeated until no more operations are possible. Each operation uses the next highest-scoring overlap to make the next join.

Overlap Layout Consensus (OLC)

step 1 overlap discovery
step 2 build and use the overlap graph
step 3 multiple sequence alignment

de Bruijn graph bases

The de Bruijn graph approach circumvents the problems of overlap consensus assembly. Rather than using the reads 'as is' and trying to link them, the k-mers (all subsequences of length k within the reads) are computed and the reads are represented as a path through the k-mers. Such a paradigm handles redundancy better than the overlap consensus approach and makes the computation of paths more tractable.

TOOLS

Assembler

Greedy

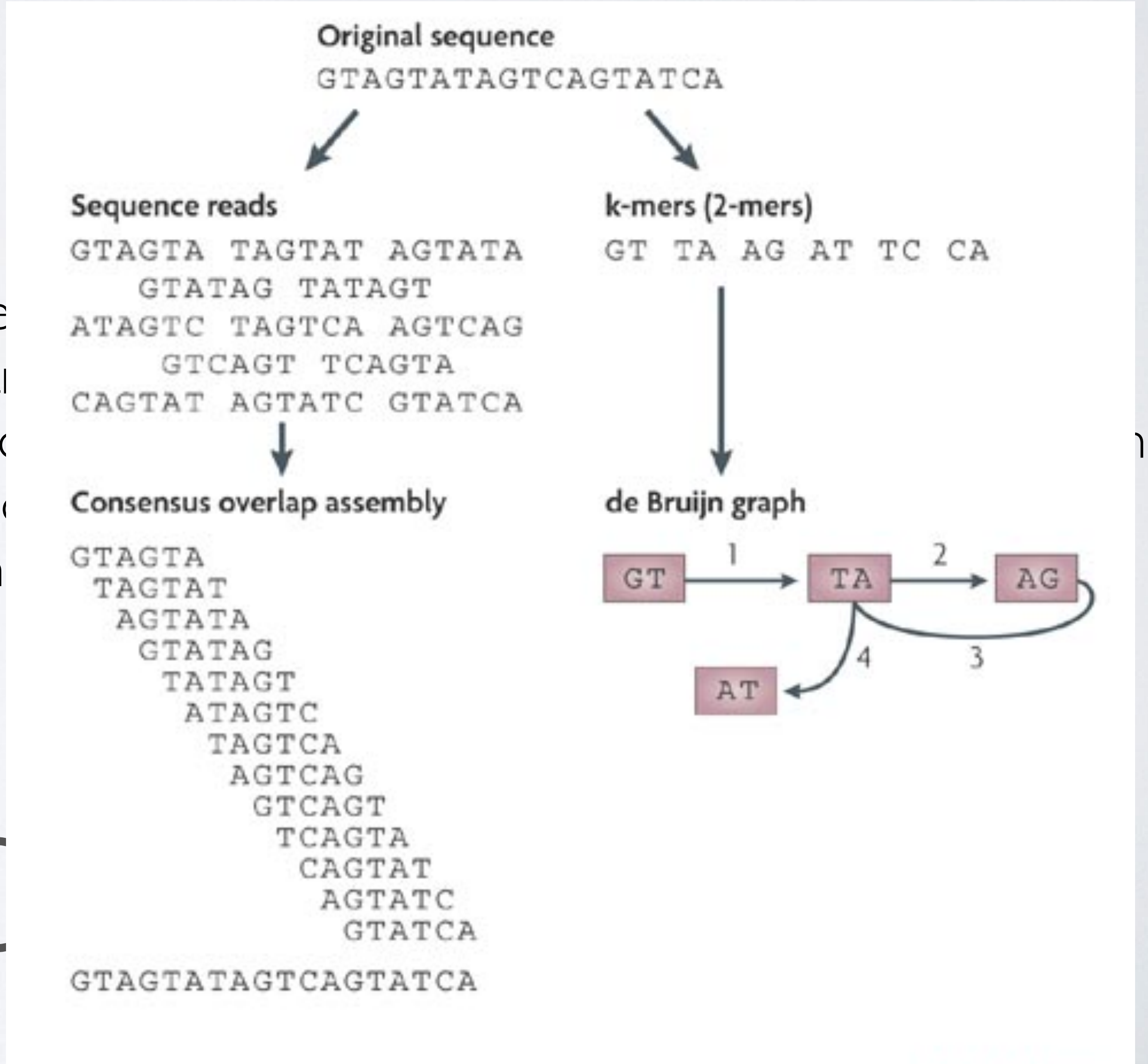
The greedy algorithms apply one basic operation: given any read or contig, add one more contig. The basic operation is repeated until no more operations are possible. Each operation uses the next highest-scoring overlap to make the next join.

Overlap Layout Consensus (OLC)

step 1 overlap discovery
step 2 build and use the overlap graph
step 3 multiple sequence alignment

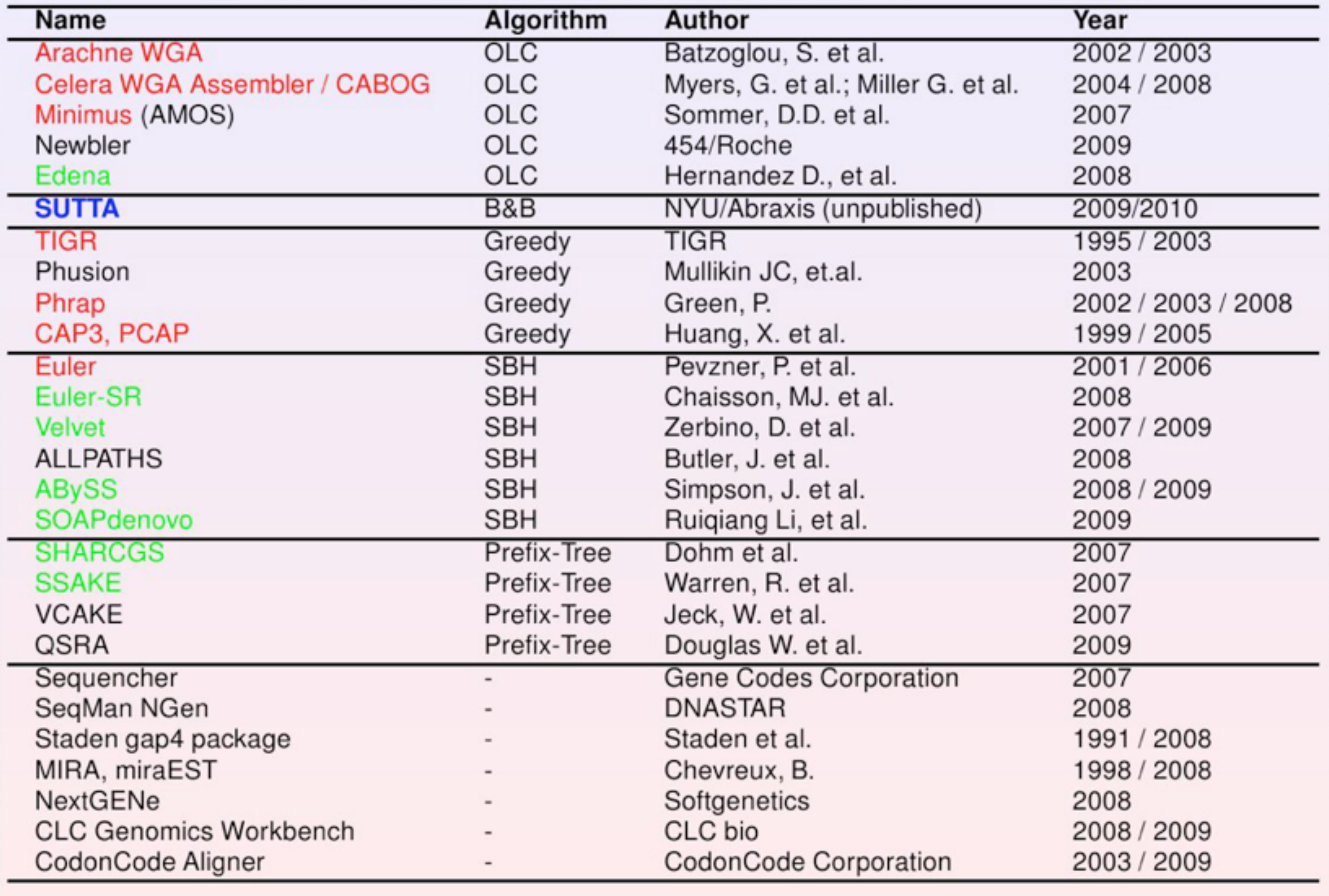
de Bruijn graph bases

The de Bruijn graph approach circumve
Rather than using the reads 'as is' and t
length k within the reads) are compute   ne
k-mers. Such a paradigm handles redun
and makes the computation of paths m



TC

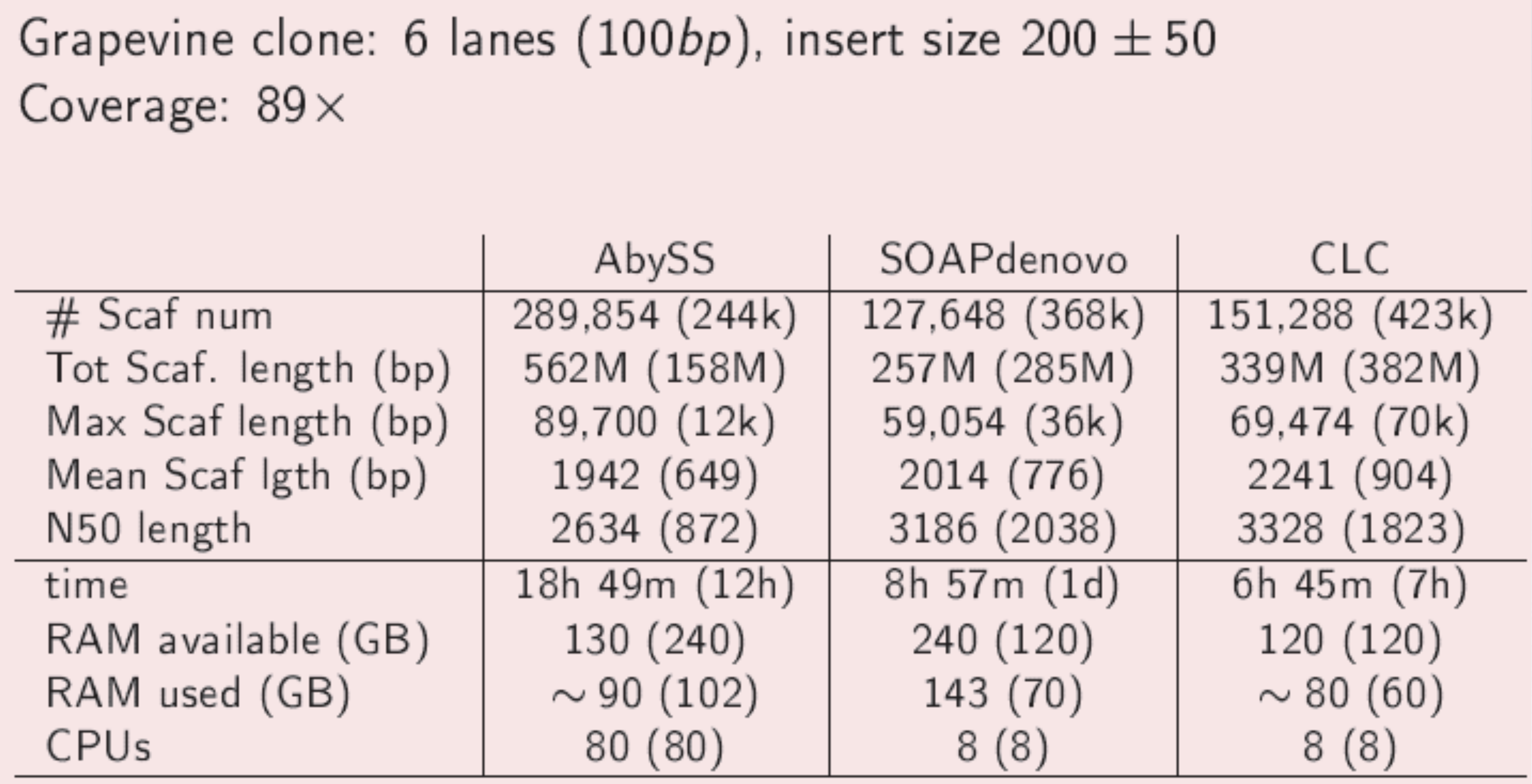
Assemblers



Name	Algorithm	Author	Year
Arachne WGA	OLC	Batzoglou, S. et al.	2002 / 2003
Celera WGA Assembler / CABOG	OLC	Myers, G. et al.; Miller G. et al.	2004 / 2008
Minimus (AMOS)	OLC	Sommer, D.D. et al.	2007
Newbler	OLC	454/Roche	2009
Edena	OLC	Hernandez D., et al.	2008
SUTTA	B&B	NYU/Abraxis (unpublished)	2009/2010
TIGR	Greedy	TIGR	1995 / 2003
Phusion	Greedy	Mullikin JC, et.al.	2003
Phrap	Greedy	Green, P.	2002 / 2003 / 2008
CAP3, PCAP	Greedy	Huang, X. et al.	1999 / 2005
Euler	SBH	Pevzner, P. et al.	2001 / 2006
Euler-SR	SBH	Chaisson, MJ. et al.	2008
Velvet	SBH	Zerbino, D. et al.	2007 / 2009
ALLPATHS	SBH	Butler, J. et al.	2008
ABySS	SBH	Simpson, J. et al.	2008 / 2009
SOAPdenovo	SBH	Ruiqiang Li, et al.	2009
SHARCGS	Prefix-Tree	Dohm et al.	2007
SSAKE	Prefix-Tree	Warren, R. et al.	2007
VCAKE	Prefix-Tree	Jeck, W. et al.	2007
QSRA	Prefix-Tree	Douglas W. et al.	2009
Sequencher	-	Gene Codes Corporation	2007
SeqMan NGen	-	DNASTAR	2008
Staden gap4 package	-	Staden et al.	1991 / 2008
MIRA, miraEST	-	Chevreux, B.	1998 / 2008
NextGENe	-	Softgenetics	2008
CLC Genomics Workbench	-	CLC bio	2008 / 2009
CodonCode Aligner	-	CodonCode Corporation	2003 / 2009

TOOLS

Assemblers



Grapevine clone: 6 lanes (100*bp*), insert size 200 ± 50
Coverage: 89×

	AbySS	SOAPdenovo	CLC
# Scaf num	289,854 (244k)	127,648 (368k)	151,288 (423k)
Tot Scaf. length (bp)	562M (158M)	257M (285M)	339M (382M)
Max Scaf length (bp)	89,700 (12k)	59,054 (36k)	69,474 (70k)
Mean Scaf lgth (bp)	1942 (649)	2014 (776)	2241 (904)
N50 length	2634 (872)	3186 (2038)	3328 (1823)
time	18h 49m (12h)	8h 57m (1d)	6h 45m (7h)
RAM available (GB)	130 (240)	240 (120)	120 (120)
RAM used (GB)	∼ 90 (102)	143 (70)	∼ 80 (60)
CPUs	80 (80)	8 (8)	8 (8)

Grapevine genome size: 475Mb

Policriti et al per. com.

TOOLS

☐ Quality control

☐ Mapping

☐ Assembly

☐ Digital gene expression

- DESeq, BaySeq, edgeR are R package to analyse count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression.

☐ Visualization

TOOLS

□ Quality control

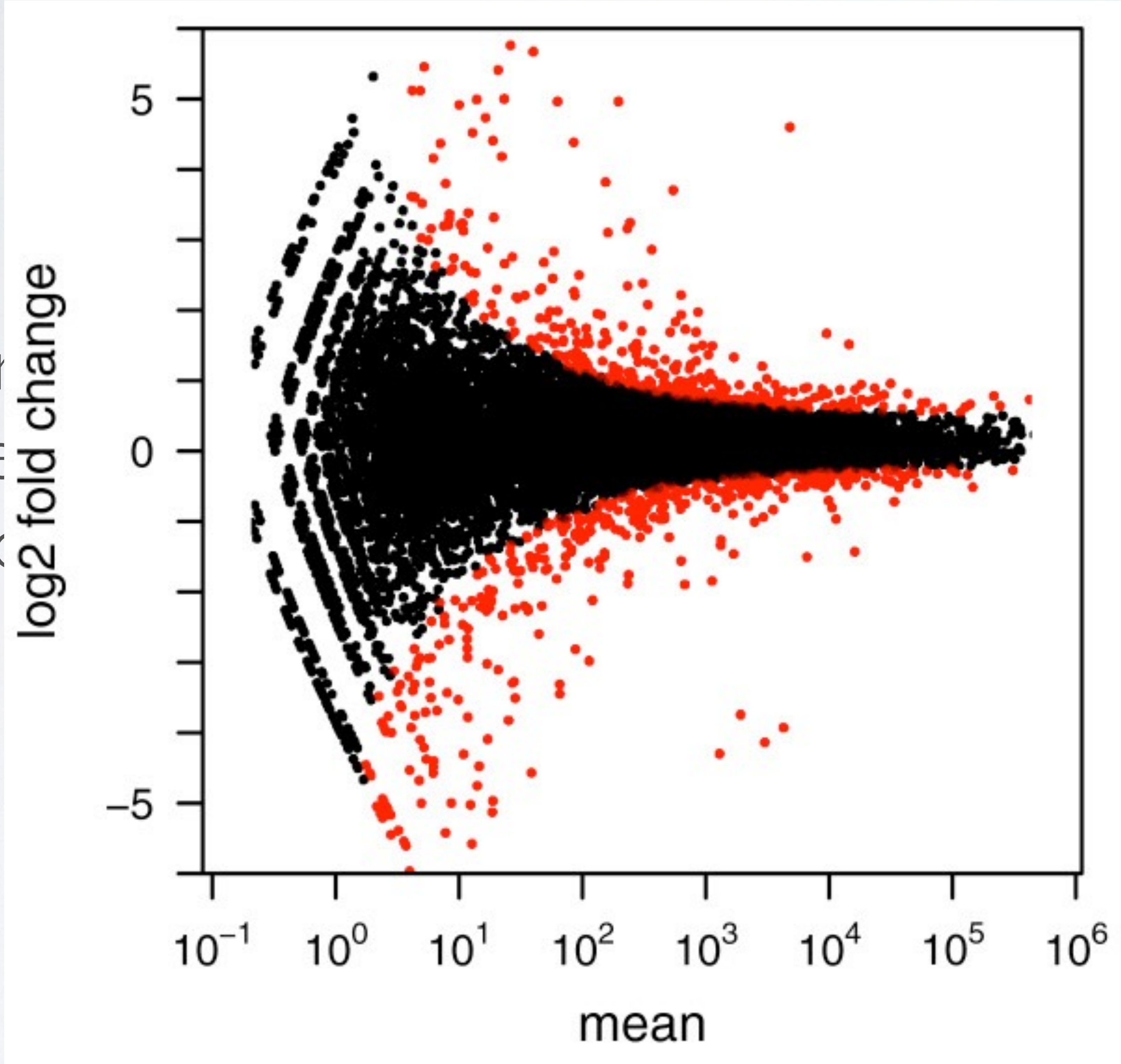
□ Mapping

□ Assembly

□ Digital gene expr

- DESeq, BaySeq, e t data from high-thro RNA-Seq and test

□ Visualization



TOOLS

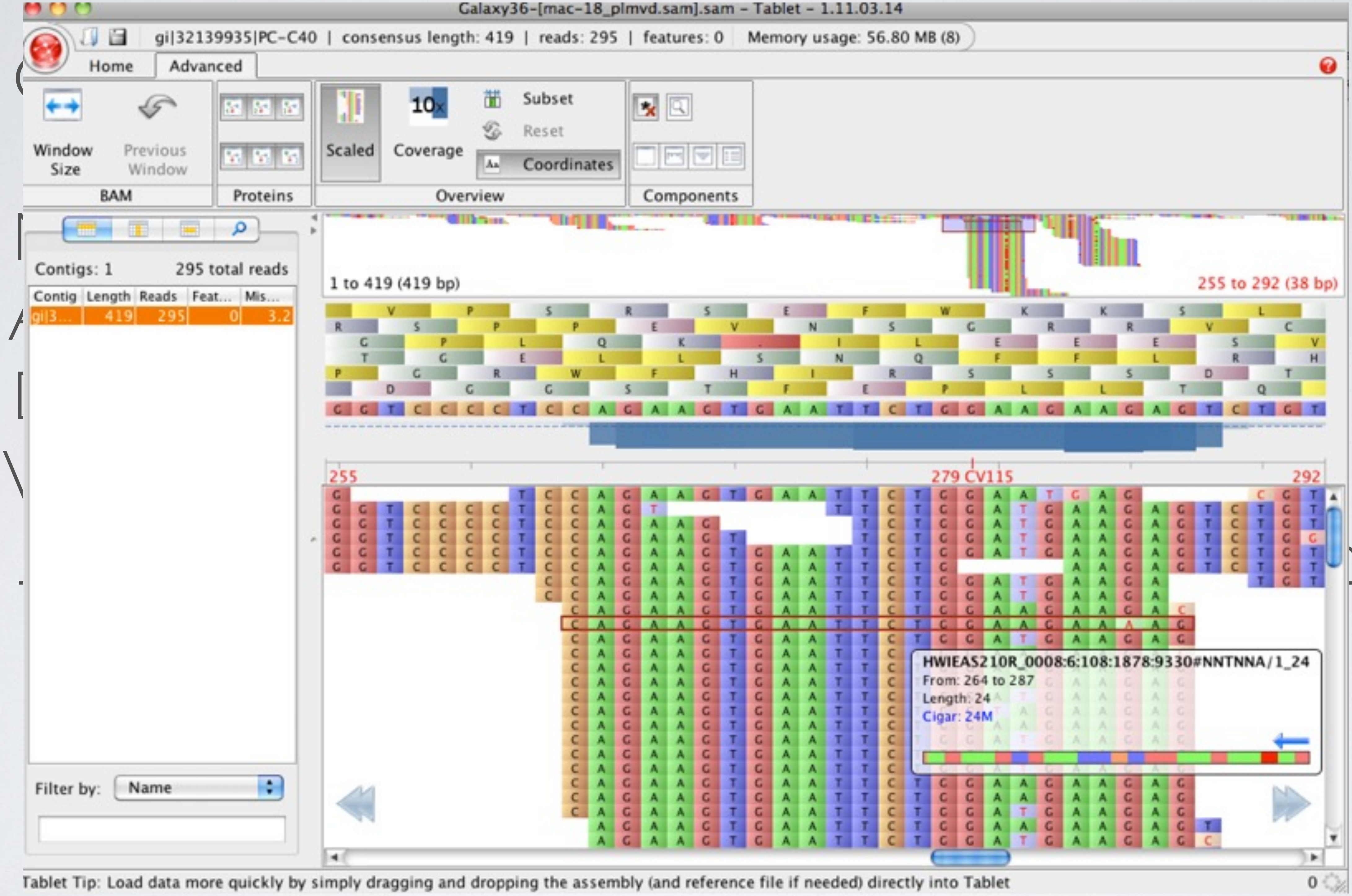
TOOLS

□ Quality control

□ Mapping
□ Assembly
□ Digital gene expression
□ Visualization

- http://lh3lh3.users.sourceforge.net/NGSalnview.shtml

Tablet (http://bioinf.scri.ac.uk/tablet/)



Galaxy36-[mac-18_plmvd.sam].sam – Tablet – 1.11.03.14

gi|32139935|PC-C40 | consensus length: 419 | reads: 295 | features: 0 | Memory usage: 56.80 MB (8)

Contigs: 1 295 total reads

Contig	Length	Reads	Feat...	Mis...
gi|3...	419	295	0	3.2

Filter by: Name

1 to 419 (419 bp)

255 to 292 (38 bp)

HWIEAS210R_0008:6:108:1878:9330#NNTNNA/1_24
From: 264 to 287
Length: 24
Cigar: 24M

Tablet Tip: Load data more quickly by simply dragging and dropping the assembly (and reference file if needed) directly into Tablet

C

I

A

I

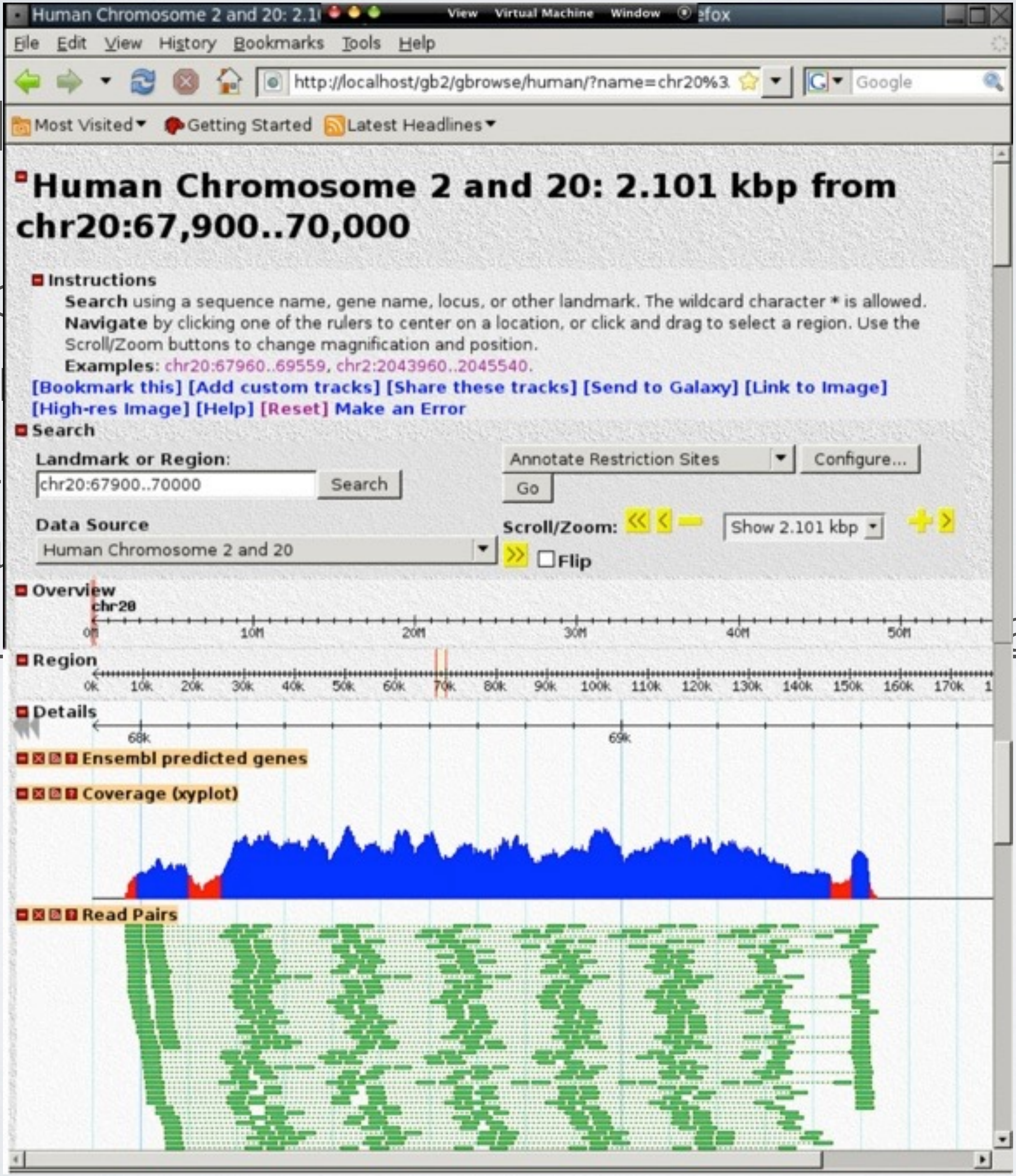
V

html

TOOLS

GBrowse (http://gmod.org/wiki/GBrowse/)

- □ Qual
- □ Mapp
- □ Asse
- □ Digit
- □ Visua
  - http: alnview.shtml



Human Chromosome 2 and 20: 2.1

File Edit View History Bookmarks Tools Help

http://localhost/gb2/gbrowse/human/?name=chr20%3

Most Visited Getting Started Latest Headlines

Human Chromosome 2 and 20: 2.101 kbp from chr20:67,900..70,000

Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed.
Navigate by clicking one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.
Examples: chr20:67960..69559, chr2:2043960..2045540.

[Bookmark this] [Add custom tracks] [Share these tracks] [Send to Galaxy] [Link to Image] [High-res Image] [Help] [Reset] Make an Error

Search

Landmark or Region: chr20:67900..70000 Search

Annotate Restriction Sites Configure... Go

Data Source: Human Chromosome 2 and 20

Scroll/Zoom: Show 2.101 kbp □ Flip

Overview

Region

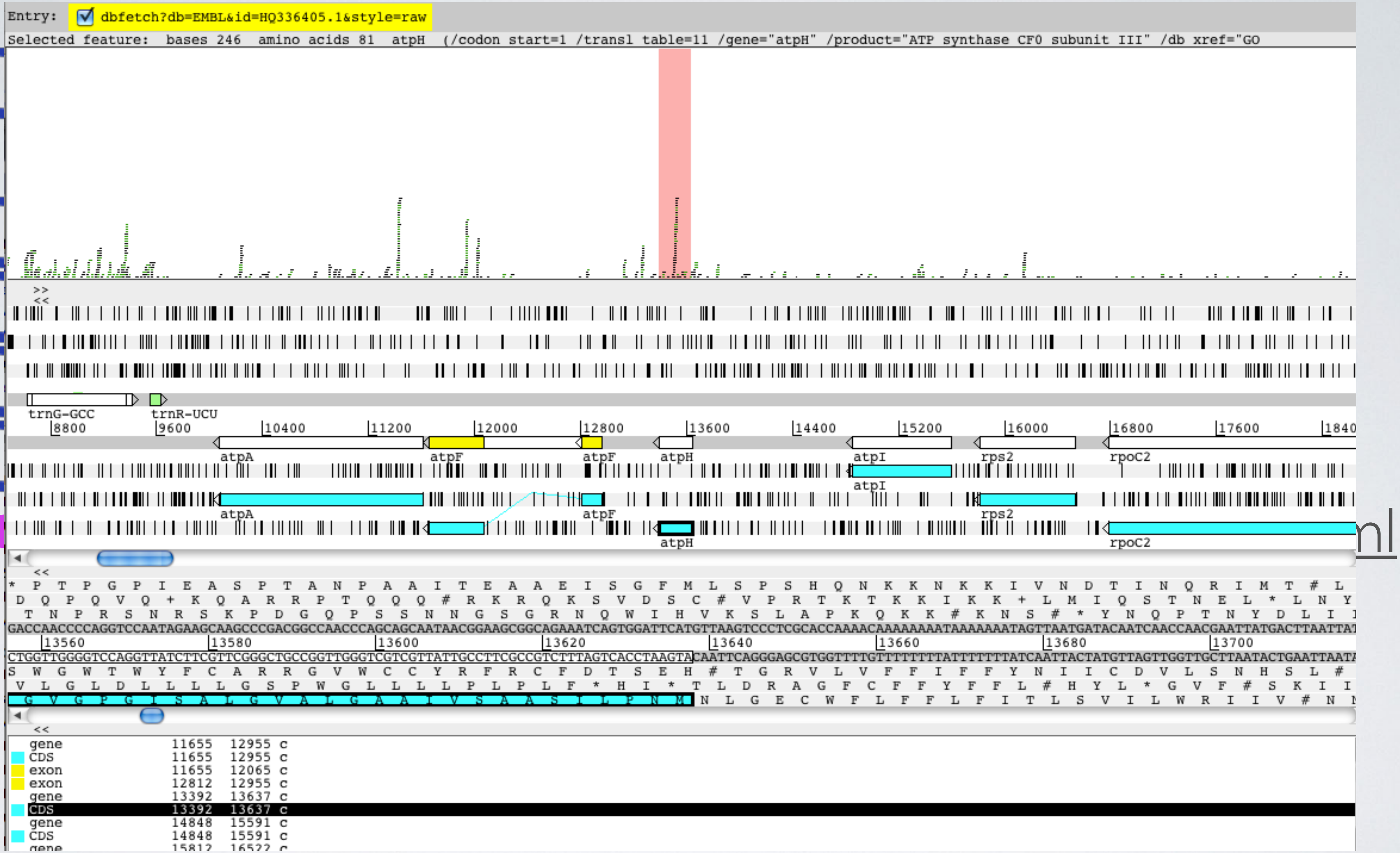
Details

Ensembl predicted genes

Coverage (xyplot)

Read Pairs

Artemis (http://www.sanger.ac.uk/resources/software/artemis/)



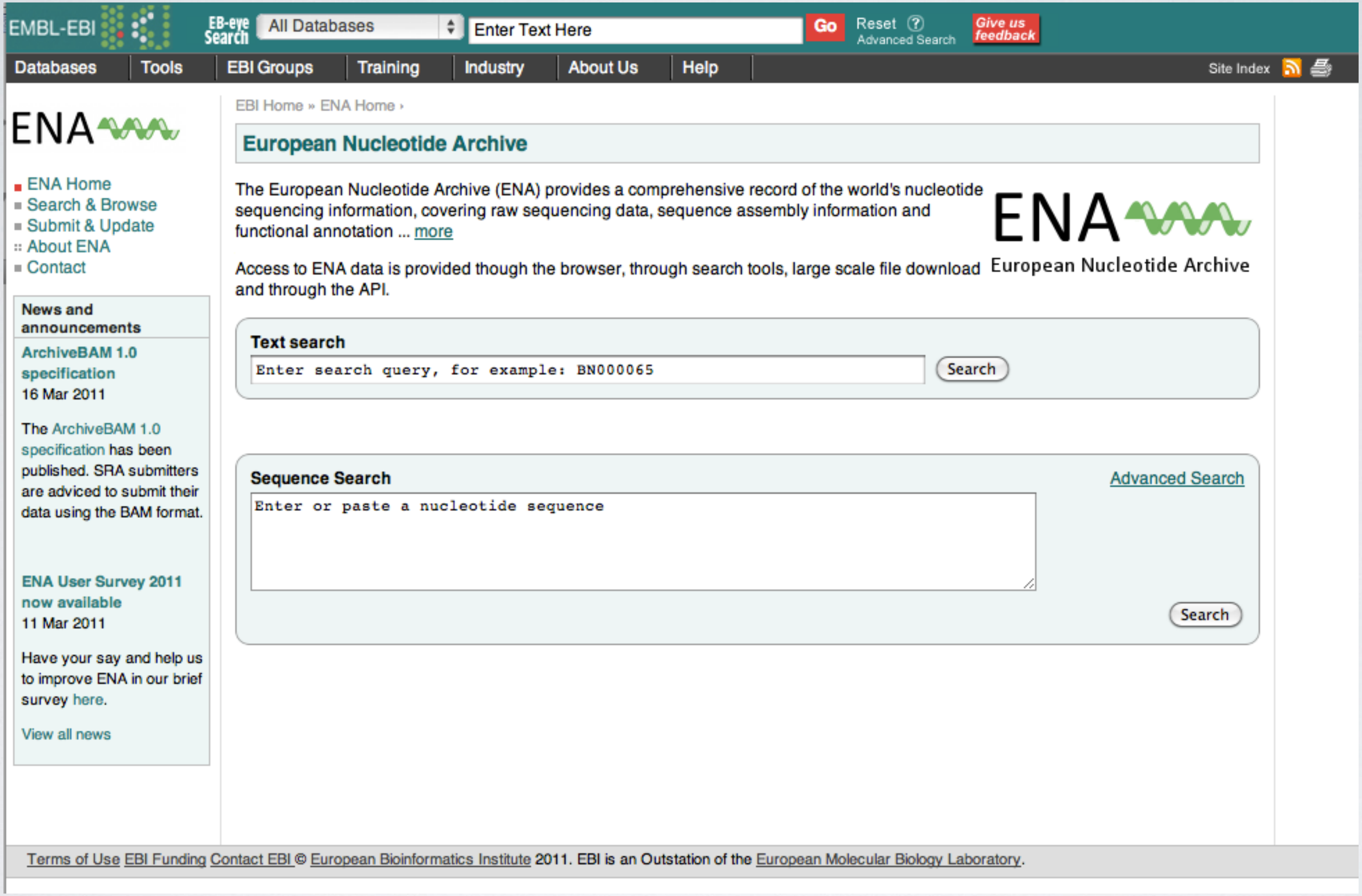
```
Entry: ☑ dbfetch?db=EMBL&id=HQ336405.1&style=raw
Selected feature:  bases 246  amino acids 81  atpH  (/codon start=1 /transl table=11 /gene="atpH" /product="ATP synthase CF0 subunit III" /db xref="GO
```

```
* P T P G P I E A S P T A N P A A I T E A A E I S G F M L S P S H Q N K K N K K I V N D T I N Q R I M T # L
 D Q P Q V Q + K Q A R R P T Q Q Q # R K R Q K S V D S C # V P R T K T K K I K K + L M I Q S T N E L * L N Y
  T N P R S N R S K P D G Q P S S N N G S G R N Q W I H V K S L A P K Q K K # K N S # * Y N Q P T N Y D L I I
GACCAACCCCAGGTCCAATAGAAGCAAGCCCGACGGCCAACCCAGCAGCAATAACGGAAGCGGCAGAAATCAGTGGATTCATGTTAAGTCCCTCGCACCAAAACAAAAAAAATAAAAAAAATAGTTAATGATACAATCAACCAACGAATTATGACTTAATTAT
     13560        13580        13600        13620        13640        13660        13680        13700
CTGGTTGGGGTCCAGGTTATCTTCGTTCGGGCTGCCGGTTGGGTCGTCGTTATTGCCTTCGCCGTCTTTAGTCACCTAAGTACAATTCAGGGAGCGTGGTTTTGTTTTTTTTATTTTTTTTATCAATTACTATGTTAGTTGGTTGCTTAATACTGAATTAATA
S W G W T W Y F C A R R G V W C C Y R F R C F D T S E H # T G R V L V F F I F F Y N I I C D V L S N H S L #
 V L G L D L L L L G S P W G L L L L P L P L F * H I * T L D R A G F C F F Y F F L # H Y L * G V F # S K I I
  G V G P G I S A L G V A L G A A I V S A A S I L P N M N L G E C W F L F F L F I T L S V I L W R I I V # N N
```

```
   gene        11655   12955 c
🟦 CDS         11655   12955 c
🟨 exon        11655   12065 c
🟨 exon        12812   12955 c
   gene        13392   13637 c
🟦 CDS         13392   13637 c
   gene        14848   15591 c
🟦 CDS         14848   15591 c
   gene        15812   16522 c
```

TOOLS

http://www.ebi.ac.uk/ena/



European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation ... more

Access to ENA data is provided though the browser, through search tools, large scale file download and through the API.

Text search

Enter search query, for example: BN000065 [Search]

Sequence Search Advanced Search

Enter or paste a nucleotide sequence

[Search]

News and announcements

ArchiveBAM 1.0 specification
16 Mar 2011

The ArchiveBAM 1.0 specification has been published. SRA submitters are adviced to submit their data using the BAM format.

ENA User Survey 2011 now available
11 Mar 2011

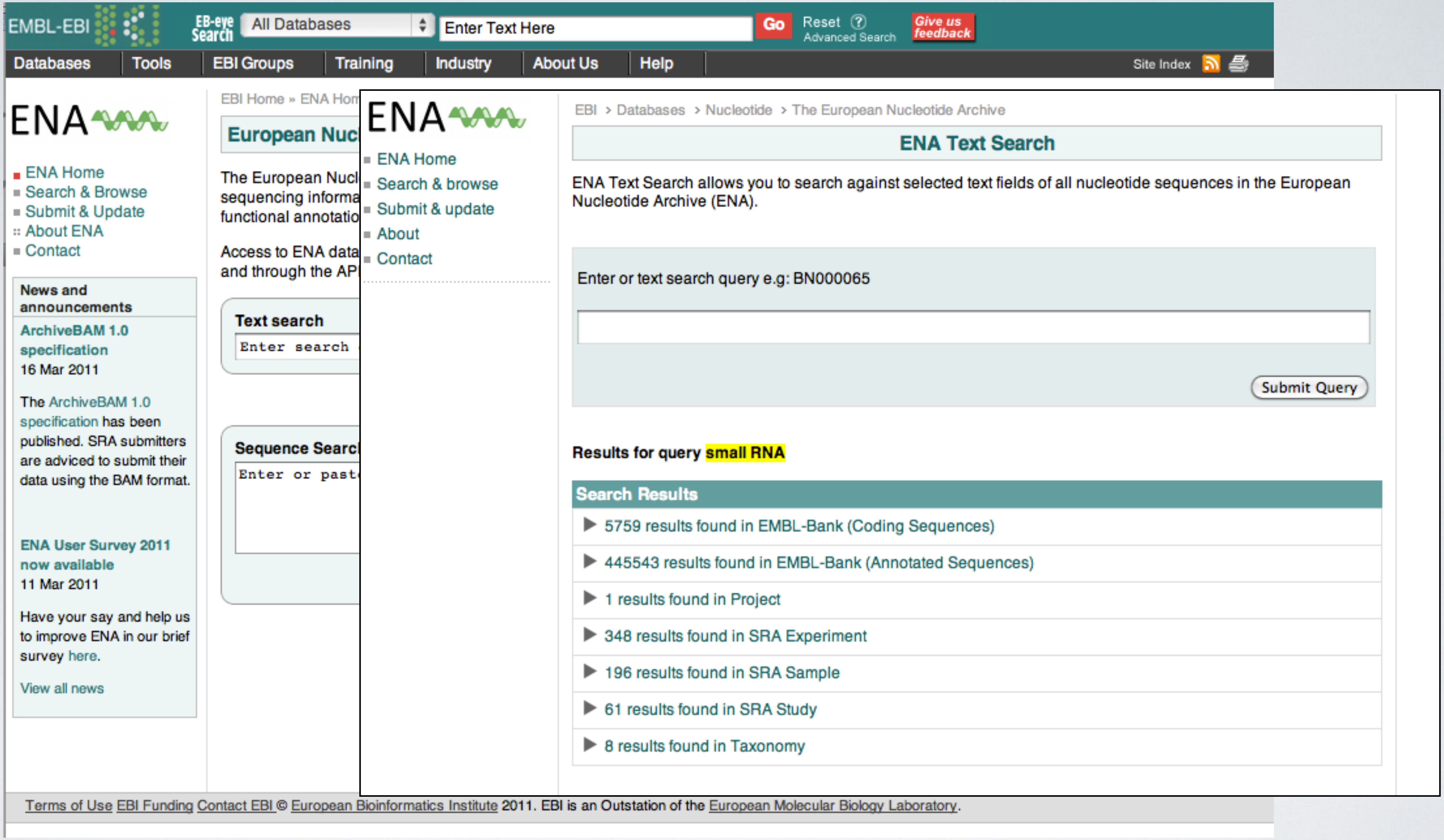
Have your say and help us to improve ENA in our brief survey here.

View all news

Terms of Use EBI Funding Contact EBI © European Bioinformatics Institute 2011. EBI is an Outstation of the European Molecular Biology Laboratory.

DATA REPOSITORY

http://www.ebi.ac.uk/ena/



News and announcements

ArchiveBAM 1.0 specification
16 Mar 2011

The ArchiveBAM 1.0 specification has been published. SRA submitters are adviced to submit their data using the BAM format.

ENA User Survey 2011 now available
11 Mar 2011

Have your say and help us to improve ENA in our brief survey here.

View all news

The European Nucl
sequencing informa
functional annotatio

Access to ENA data
and through the API

ENA Text Search

ENA Text Search allows you to search against selected text fields of all nucleotide sequences in the European Nucleotide Archive (ENA).

Enter or text search query e.g: BN000065

Results for query small RNA

Search Results

- ▶ 5759 results found in EMBL-Bank (Coding Sequences)
- ▶ 445543 results found in EMBL-Bank (Annotated Sequences)
- ▶ 1 results found in Project
- ▶ 348 results found in SRA Experiment
- ▶ 196 results found in SRA Sample
- ▶ 61 results found in SRA Study
- ▶ 8 results found in Taxonomy

Terms of Use EBI Funding Contact EBI © European Bioinformatics Institute 2011. EBI is an Outstation of the European Molecular Biology Laboratory.

DATA REPOSITORY

THANKS!!!

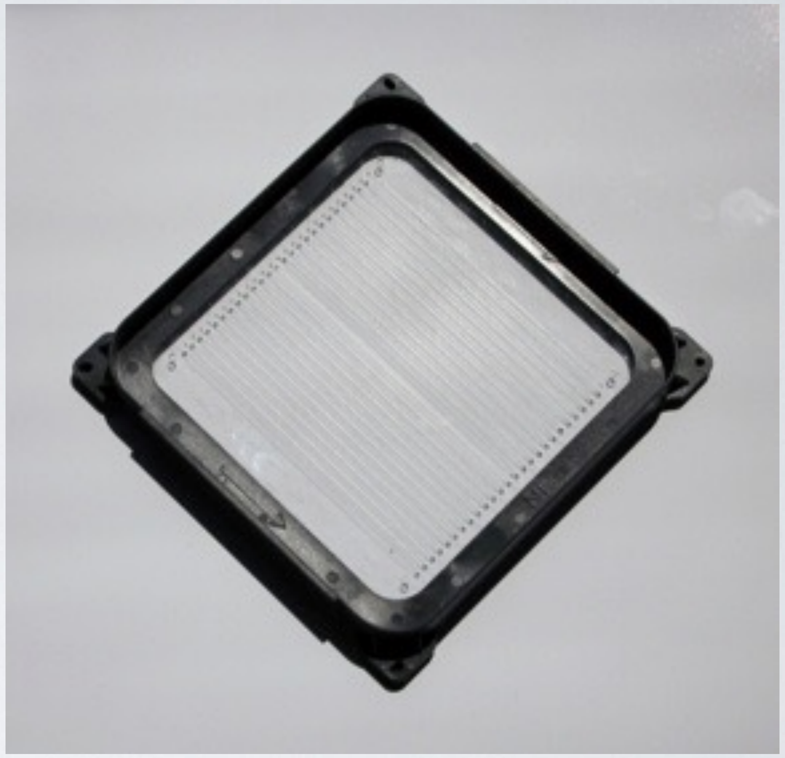
THANKS!!!

HELICOS

□ DNA template immobilized to a flow cell
□ NO amplification (true single molecule sequencing)
□ Sequencing on flow cell (1000M reads)
□ Sequencing by synthesis (fluorescence)
□ Read length up to 50nt average 32
□ High error rate

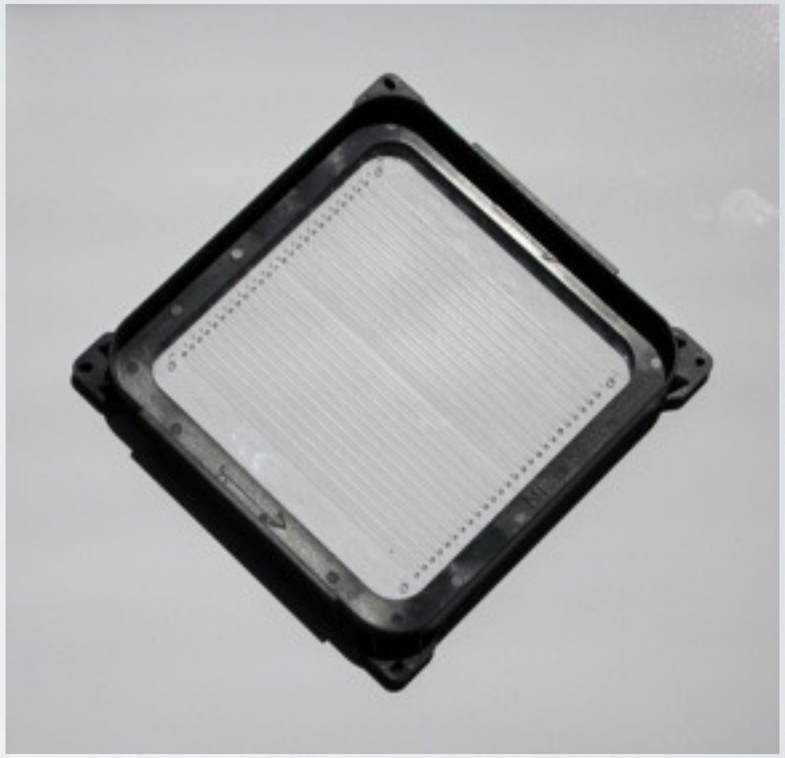
HELICOS

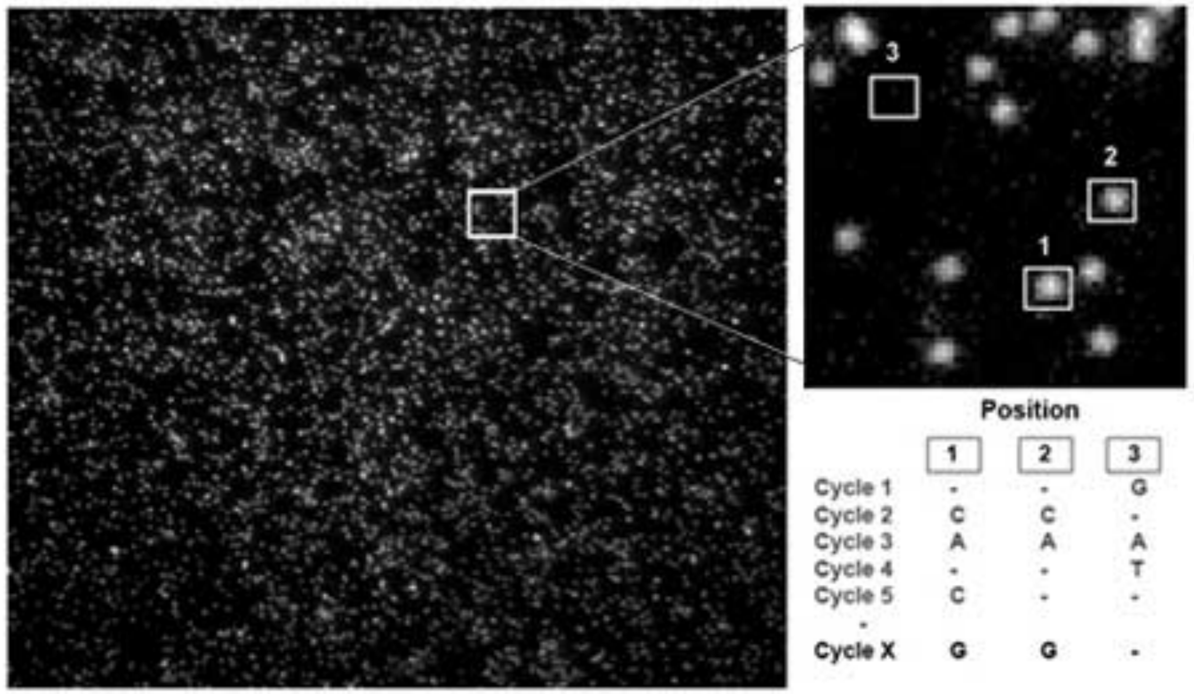
- ☐ DNA template immobilized to a flow cell
- ☐ NO amplification (true single molecule sequencing)
- ☐ Sequencing on flow cell (1000M reads)
- ☐ Sequencing by synthesis (fluorescence)
- ☐ Read length up to 50nt average 32
- ☐ High error rate



HELICOS

- □ DNA template immobilized to a flow cell
- □ NO amplification (true single molecule sequencing)
- □ Sequencing on flow cell (1000M reads)
- □ Sequencing by synthesis (fluorescence)
- □ Read length up to 50nt average 32
- □ High error rate





Position	1	2	3
Cycle 1	-	-	G
Cycle 2	C	C	-
Cycle 3	A	A	A
Cycle 4	-	-	T
Cycle 5	C	-	-
·			
Cycle X	G	G	-

An image taken by the HeliScope Single Molecule Sequencer.
Inset shows a close-up view of individual single molecules.

HELICOS

- □ Quality control

- □ Mapping
- □ Alignment
  - BLAST - Basic Local Alignment Search Tool
  - GAST - Global Alignment for Sequence Taxonomy

- □ Assembly
- □ Digital gene expression
- □ Visualisation

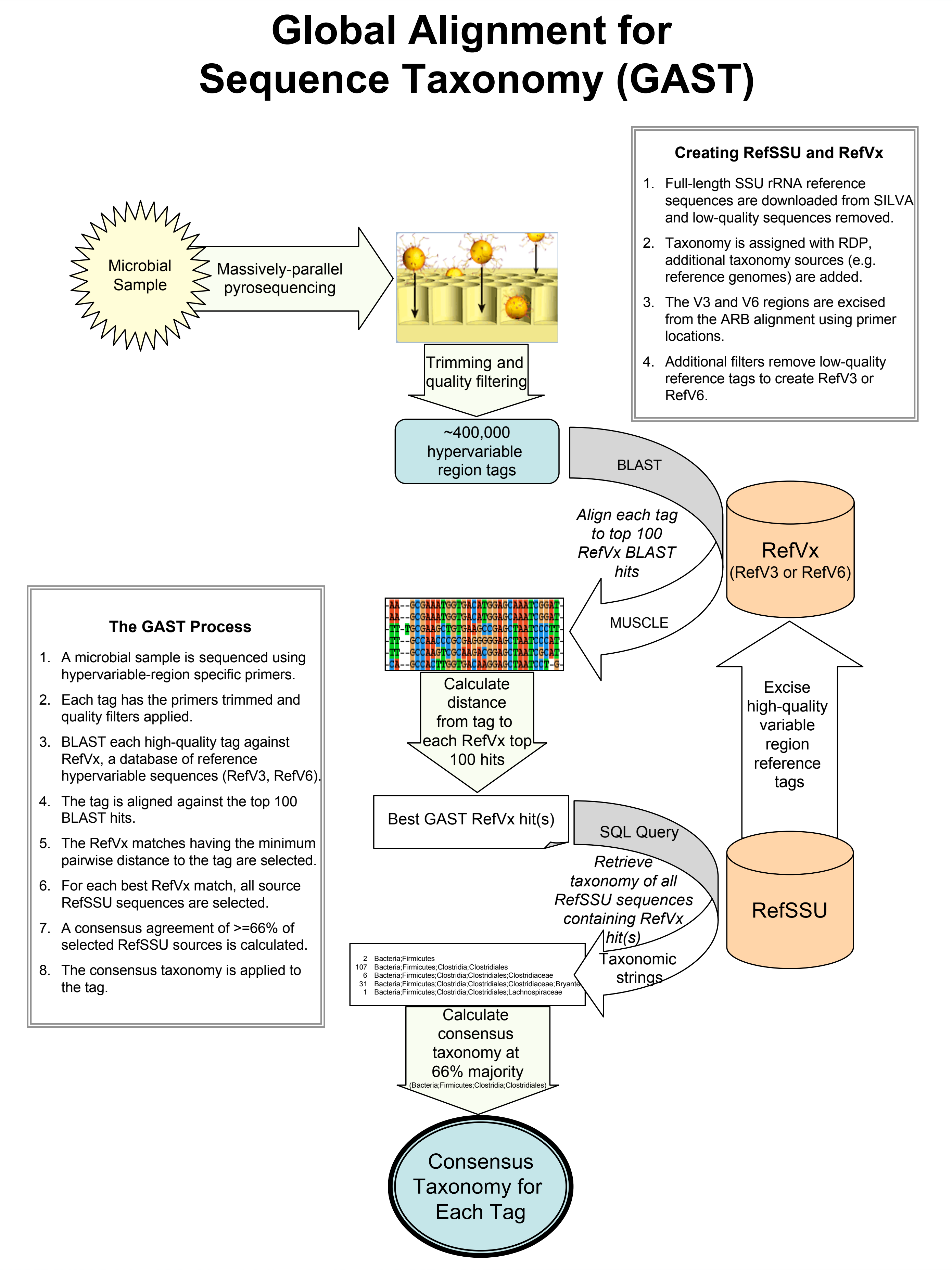
TOOLS

- □ Quality control

- □ Mapping
- □ Alignment
  - BLAST - Basic Lo
  - GAST - Global A

my

- □ Assembly
- □ Digital gene expr
- □ Visualisation



Global Alignment for Sequence Taxonomy (GAST)

Creating RefSSU and RefVx

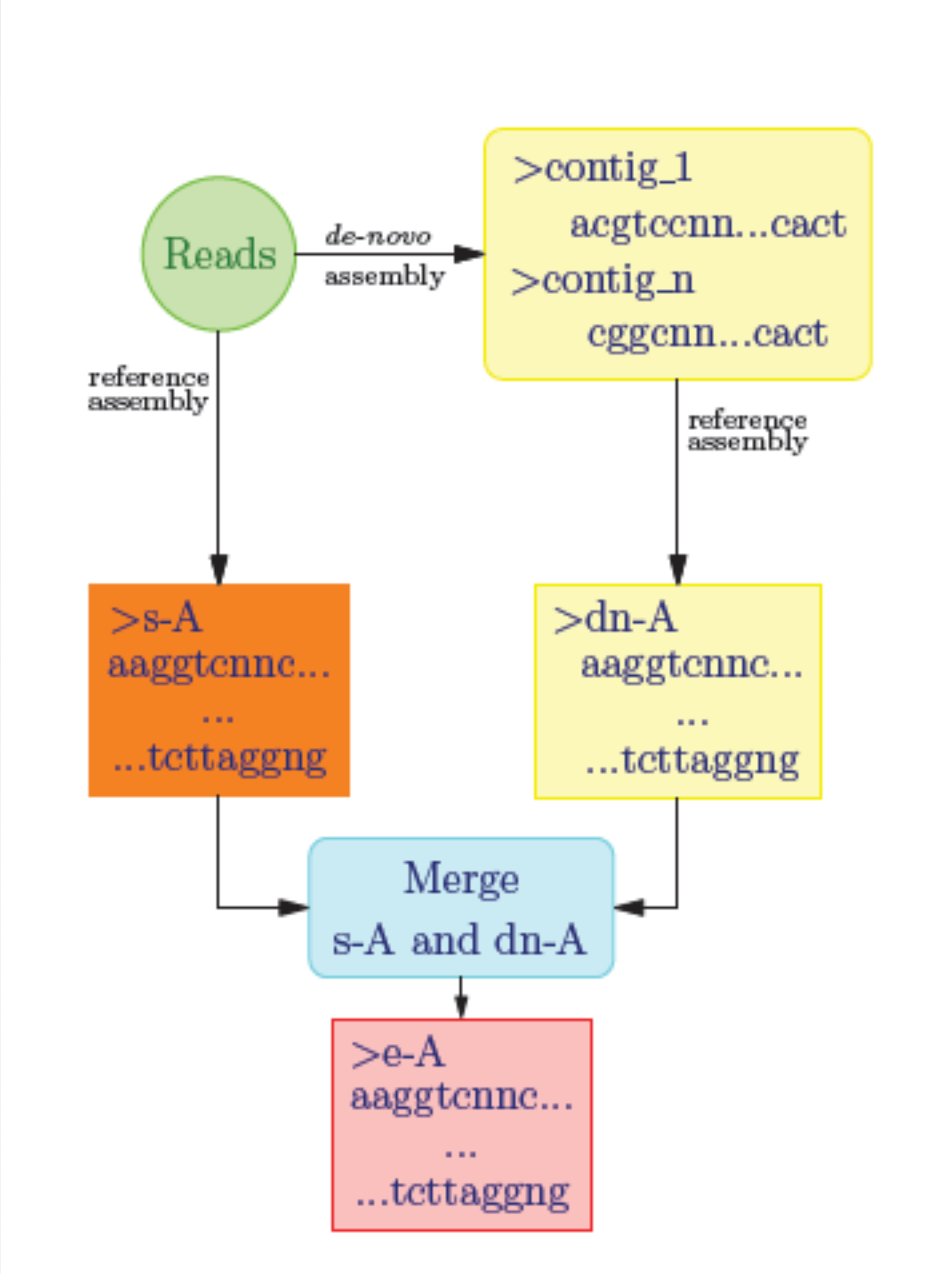
1. Full-length SSU rRNA reference sequences are downloaded from SILVA and low-quality sequences removed.
2. Taxonomy is assigned with RDP, additional taxonomy sources (e.g. reference genomes) are added.
3. The V3 and V6 regions are excised from the ARB alignment using primer locations.
4. Additional filters remove low-quality reference tags to create RefV3 or RefV6.

The GAST Process

1. A microbial sample is sequenced using hypervariable-region specific primers.
2. Each tag has the primers trimmed and quality filters applied.
3. BLAST each high-quality tag against RefVx, a database of reference hypervariable sequences (RefV3, RefV6).
4. The tag is aligned against the top 100 BLAST hits.
5. The RefVx matches having the minimum pairwise distance to the tag are selected.
6. For each best RefVx match, all source RefSSU sequences are selected.
7. A consensus agreement of >=66% of selected RefSSU sources is calculated.
8. The consensus taxonomy is applied to the tag.

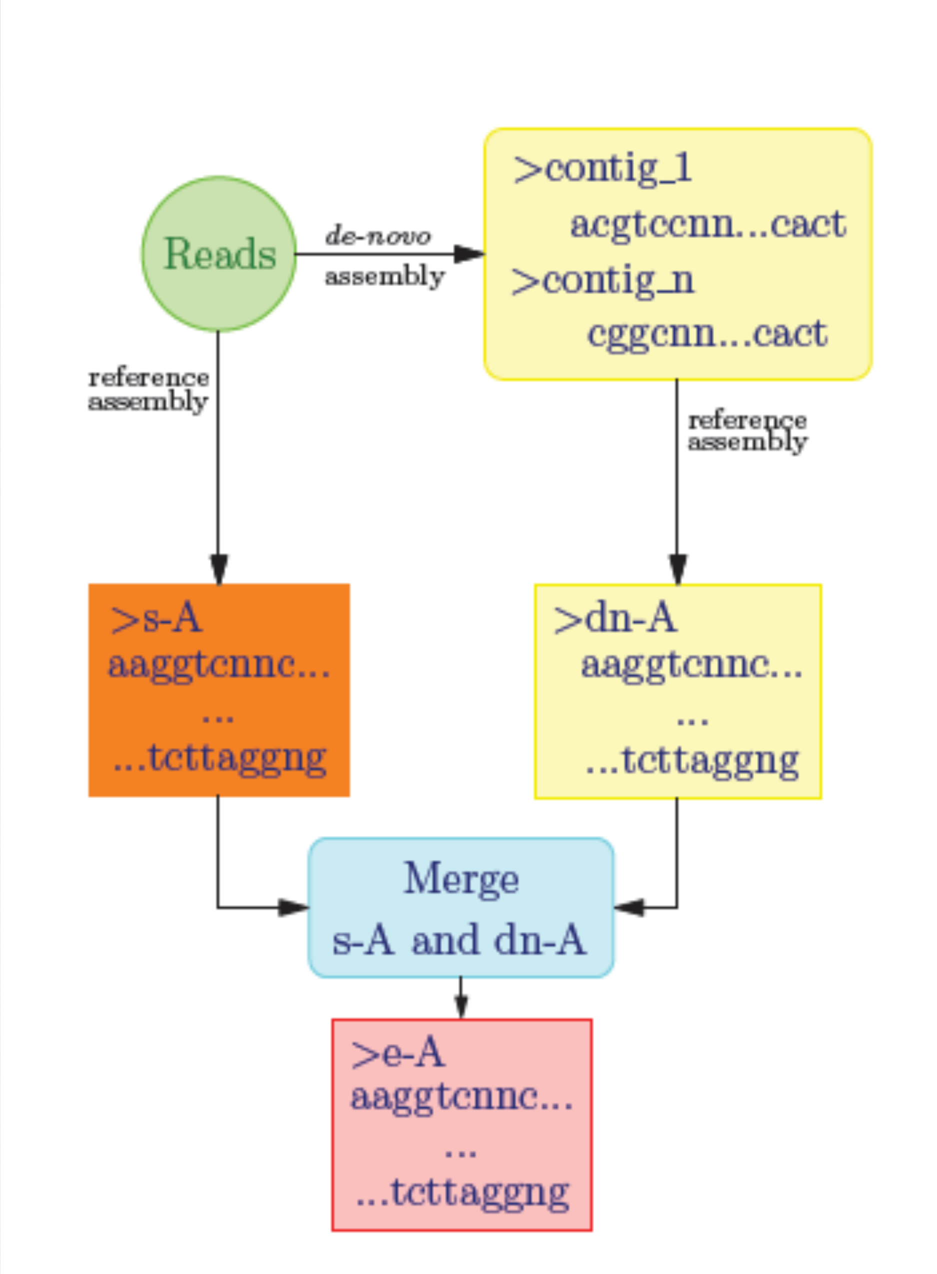
Microbial Sample
Massively-parallel pyrosequencing
Trimming and quality filtering
~400,000 hypervariable region tags
BLAST
Align each tag to top 100 RefVx BLAST hits
MUSCLE
RefVx
(RefV3 or RefV6)
Excise high-quality variable region reference tags
Calculate distance from tag to each RefVx top 100 hits
Best GAST RefVx hit(s)
SQL Query
Retrieve taxonomy of all RefSSU sequences containing RefVx hit(s)
RefSSU
Taxonomic strings
2 Bacteria;Firmicutes
107 Bacteria;Firmicutes;Clostridia;Clostridiales
6 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae
31 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Bryante
1 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae
Calculate consensus taxonomy at 66% majority
(Bacteria;Firmicutes;Clostridia;Clostridiales)
Consensus Taxonomy for Each Tag

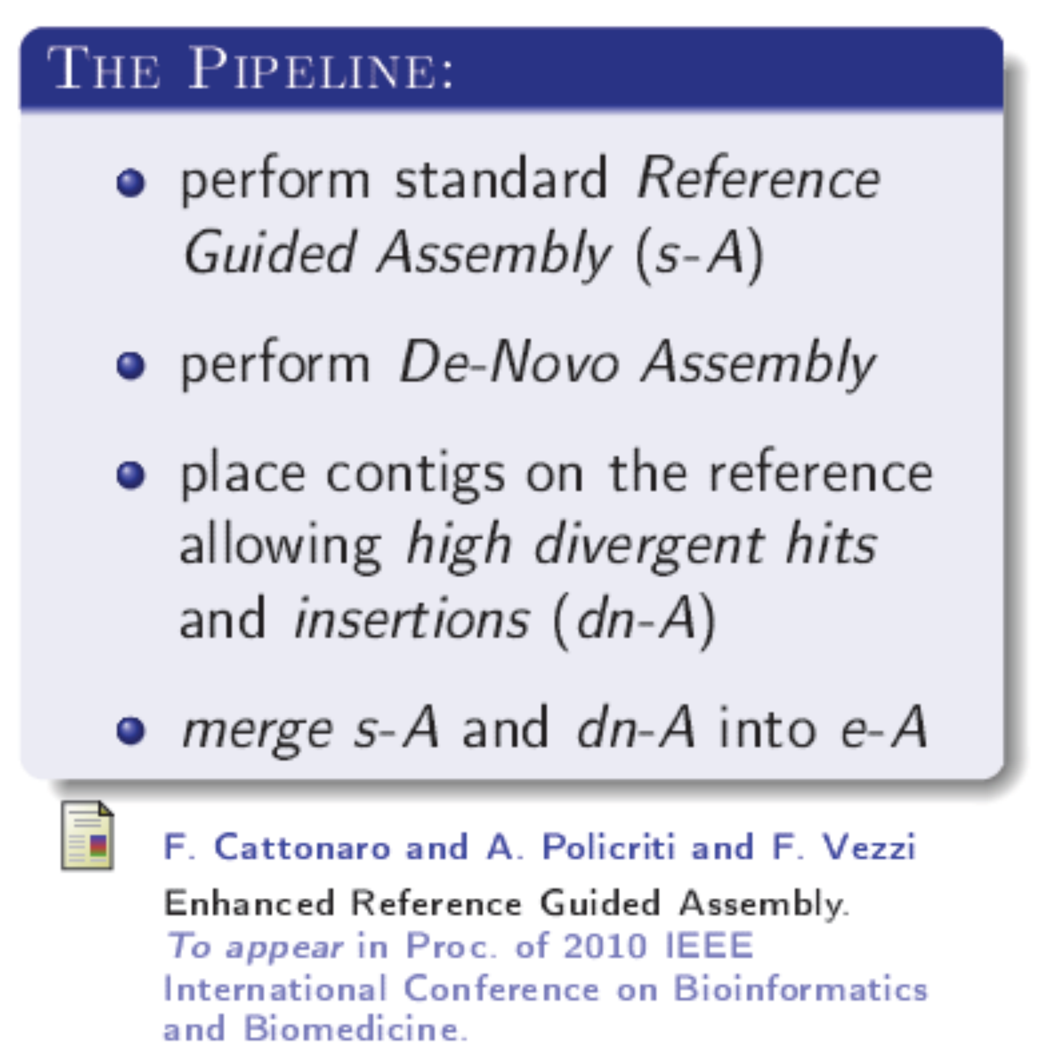
Assemblers - Enhanced Reference Guided Assembly



TOOLS

Assemblers - Enhanced Reference Guided Assembly



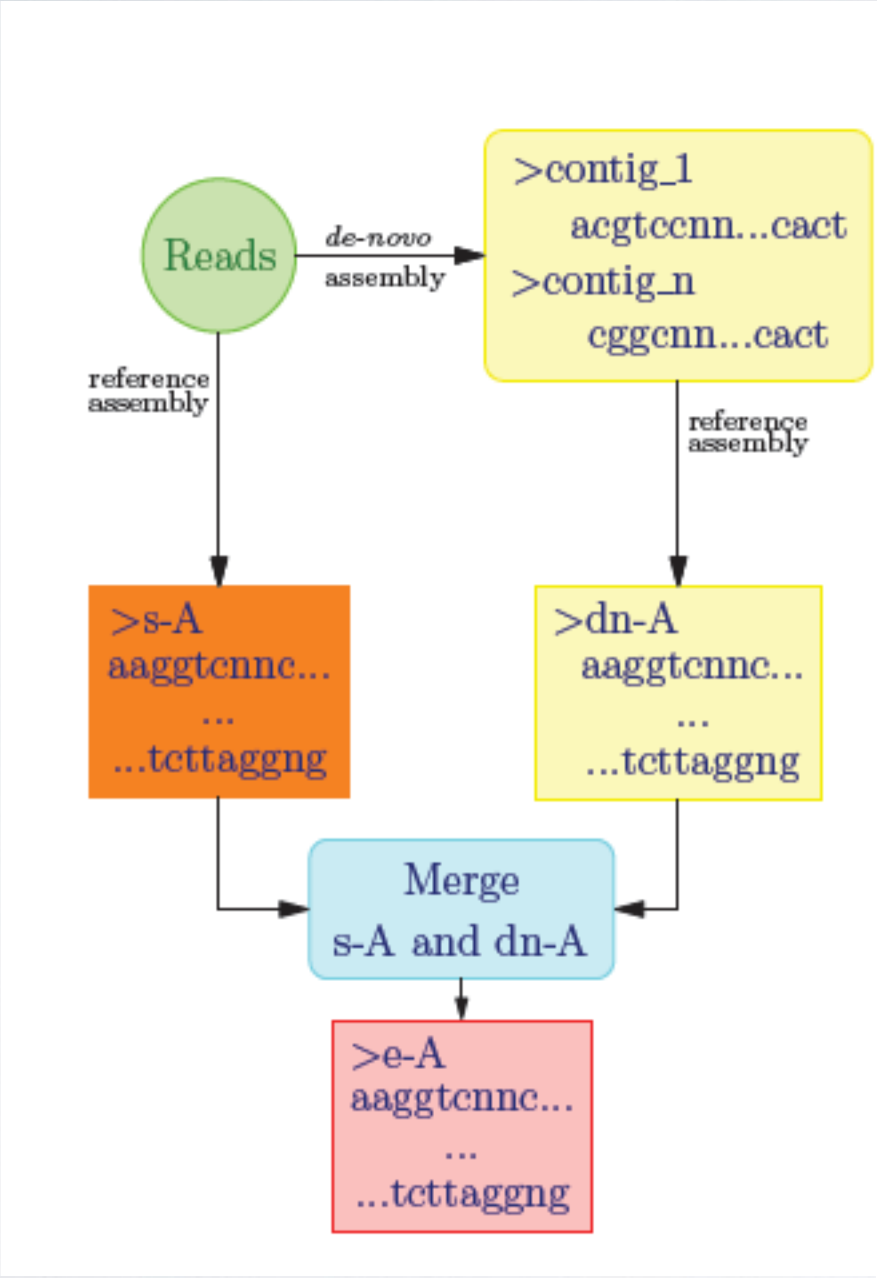


The Pipeline:

- perform standard *Reference Guided Assembly* (*s-A*)
- perform *De-Novo Assembly*
- place contigs on the reference allowing *high divergent hits* and *insertions* (*dn-A*)
- *merge s-A* and *dn-A* into *e-A*

F. Cattonaro and A. Policriti and F. Vezzi
Enhanced Reference Guided Assembly.
To appear in Proc. of 2010 IEEE International Conference on Bioinformatics and Biomedicine.

Assemblers - Enhanced Reference Guided Assembly



Reads

de-novo assembly

>contig_1
acgtccnn...cact
>contig_n
cggcnn...cact

reference assembly

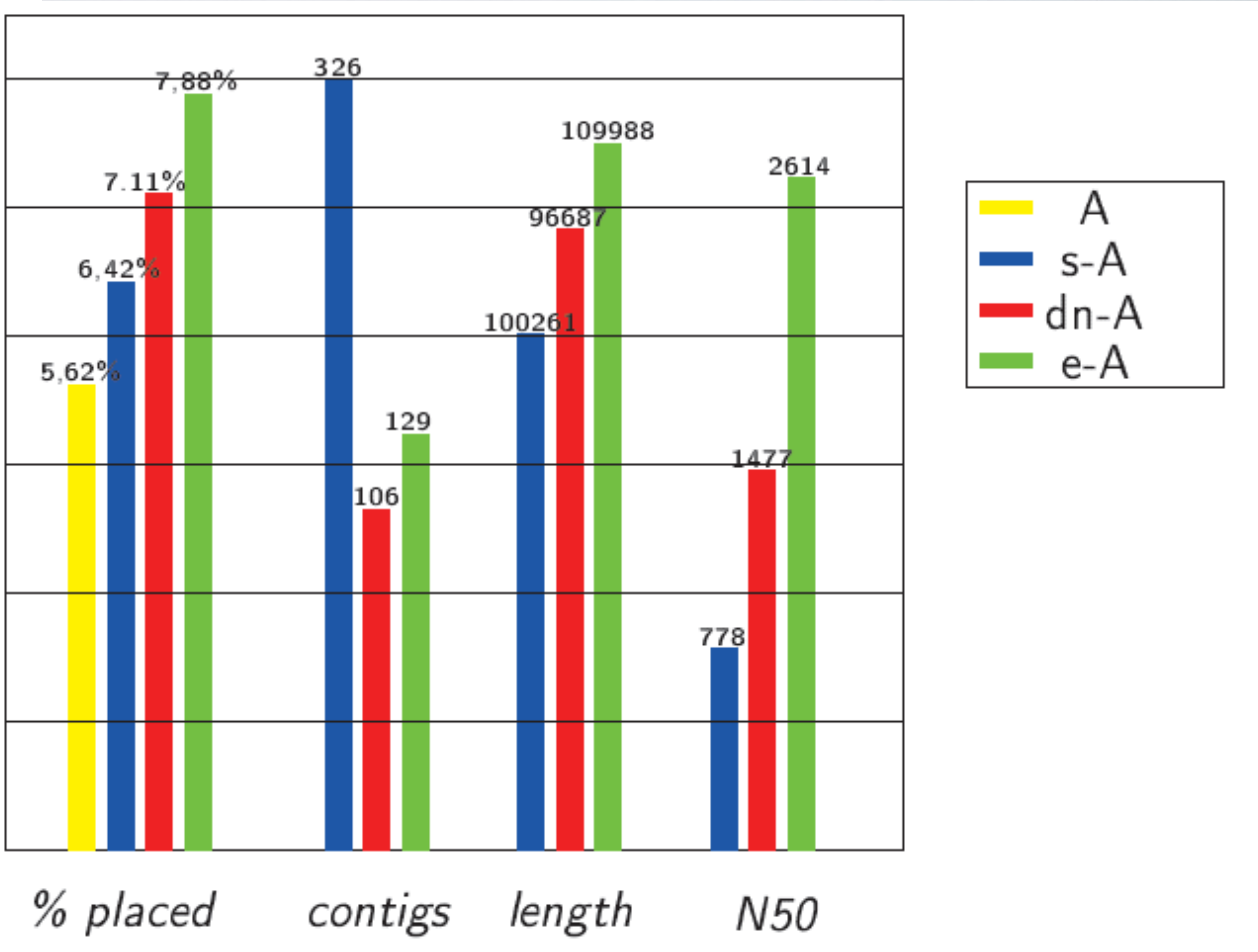
reference assembly

>s-A
aaggtcnnc...
...
...tcttaggng

>dn-A
aaggtcnnc...
...
...tcttaggng

Merge
s-A and dn-A

>e-A
aaggtcnnc...
...
...tcttaggng



TOOLS