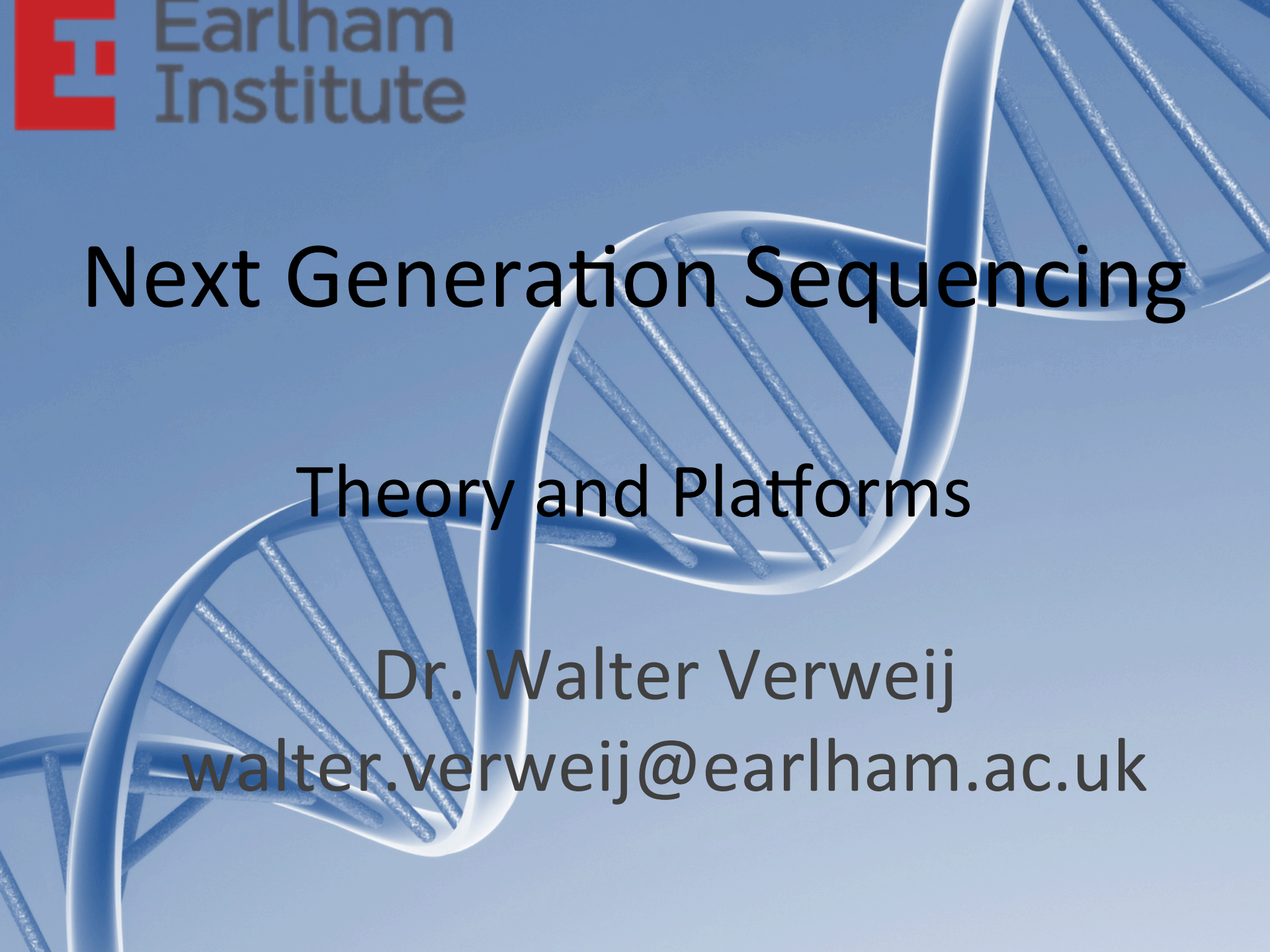


Next Generation Sequencing

Theory and Platforms

Dr. Walter Verweij
walter.verweij@earlham.ac.uk


Earlham Institute
Earlham Institute
At the Norwich Research Park (NRP)

Sequencing Platforms

Illumina HiSeq 2500 x 2
Illumina HiSeq 2000 x 1
Illumina HiSeq 4000 x 2

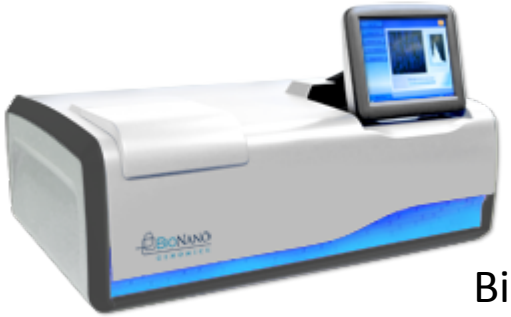
Illumina MiSeq x 3

PacBio RSII x 1
PacBio Sequel x 1

Sequencing Platforms

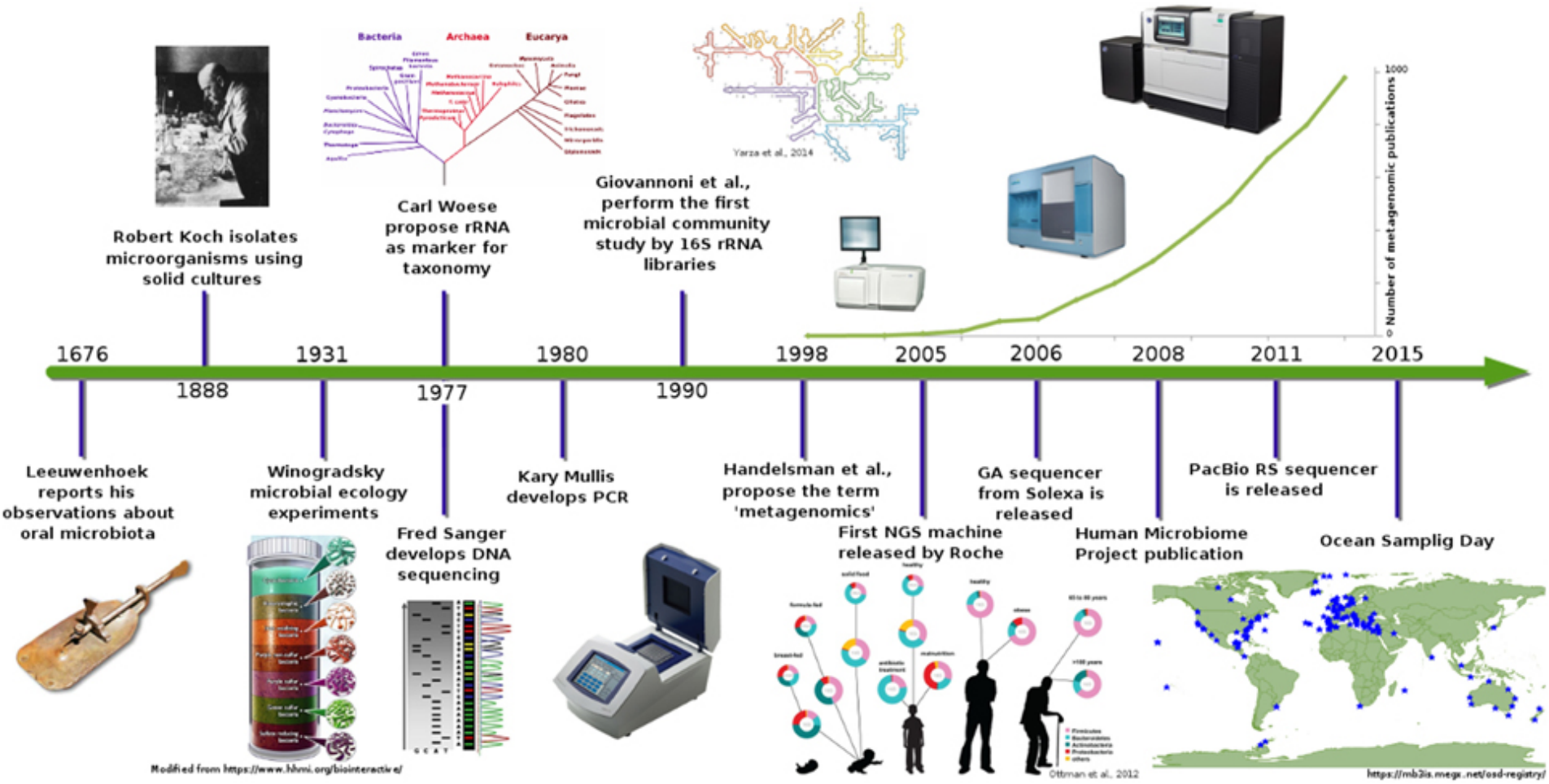
OpGen Argus x 1

Oxford Nanopore MinIon x …



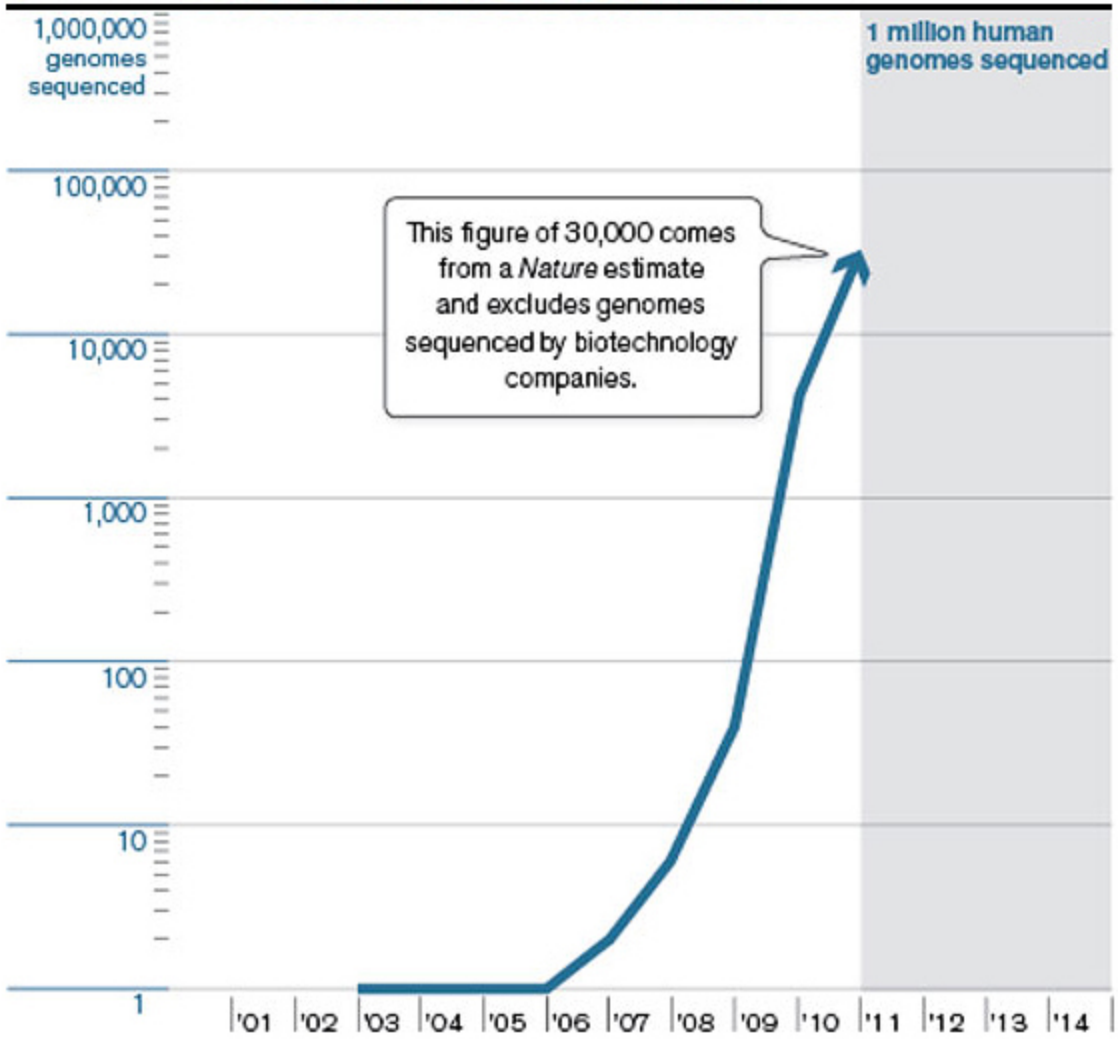
Bionano Optical mapping x 1

Sequence history

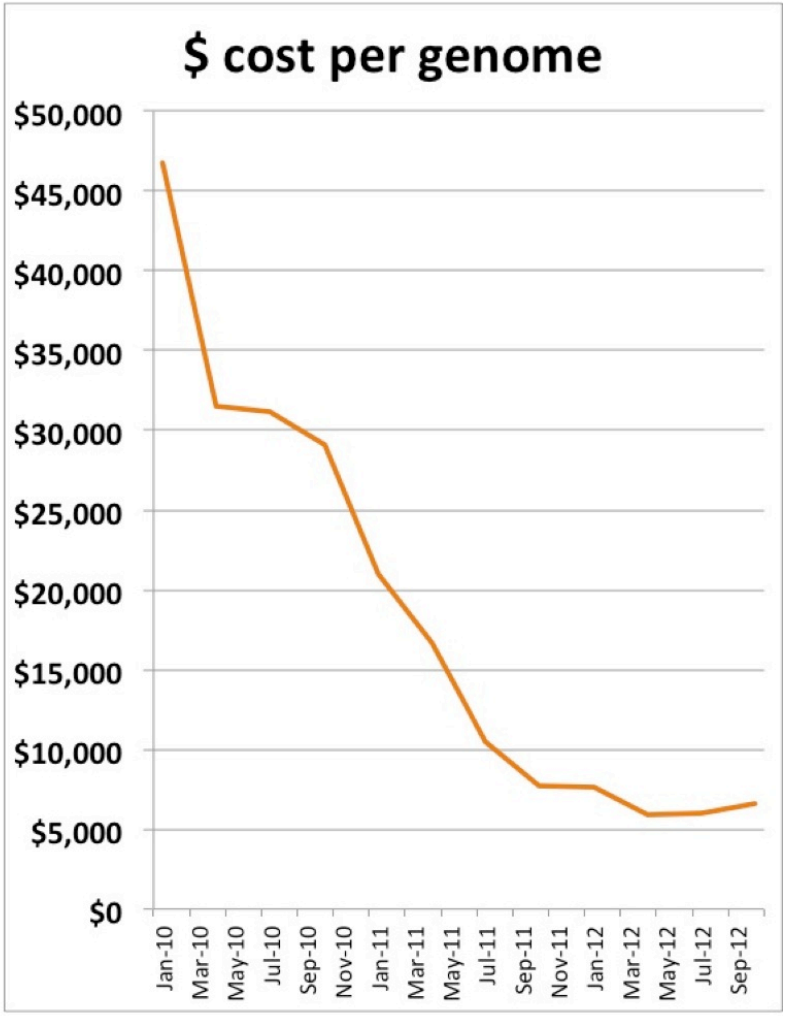


Output Skyrocketing

Number sequenced

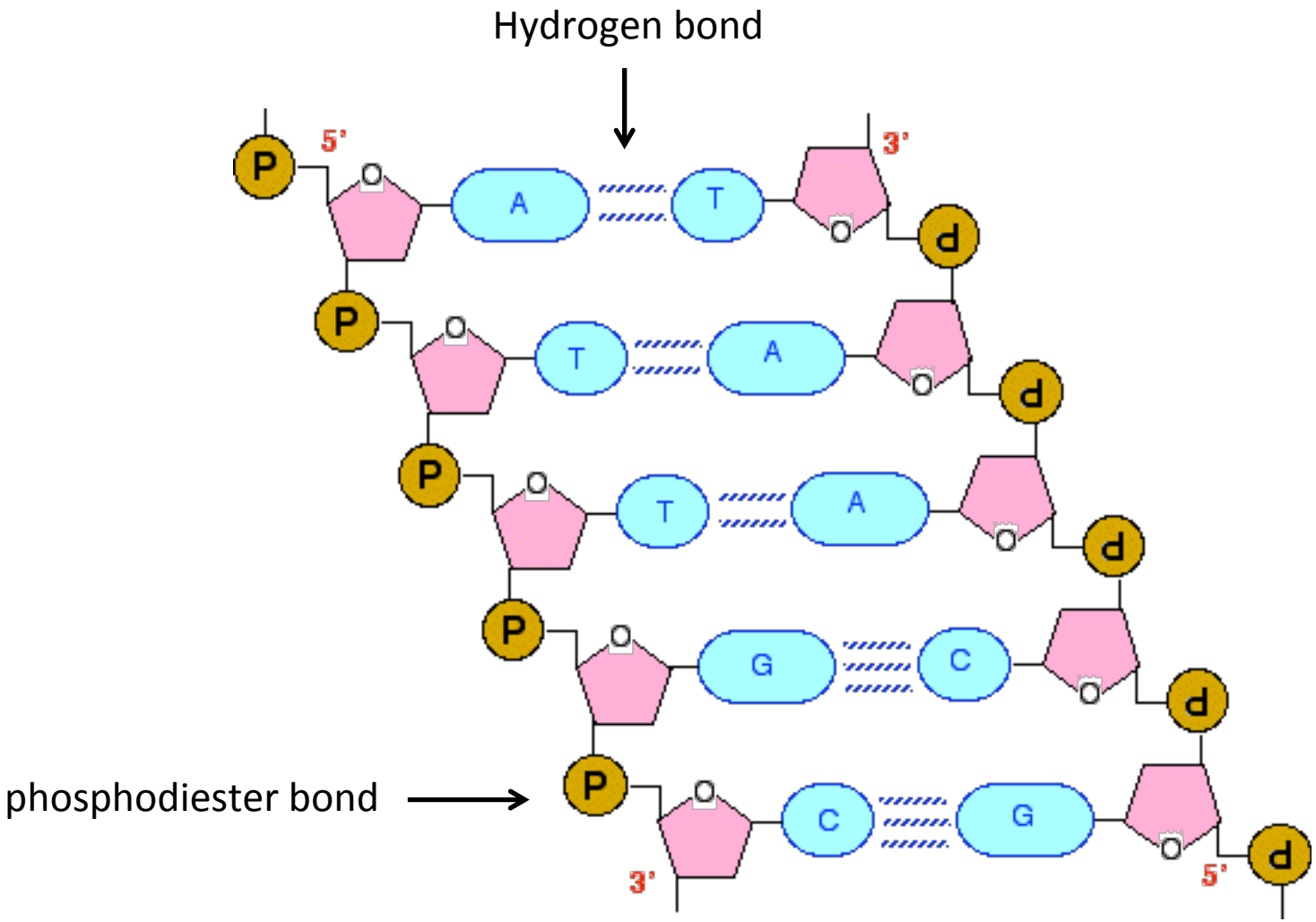



$ cost per Mb
$0.60
$0.50
$0.40
$0.30
$0.20
$0.10
$0.00
Jan-10
Mar-10
May-10
Jul-10
Sep-10
Nov-10
Jan-11
Mar-11
May-11
Jul-11
Sep-11
Nov-11
Jan-12
Mar-12
May-12
Jul-12
Sep-12


$ cost per genome
$50,000
$45,000
$40,000
$35,000
$30,000
$25,000
$20,000
$15,000
$10,000
$5,000
$0
Jan-10
Mar-10
May-10
Jul-10
Sep-10
Nov-10
Jan-11
Mar-11
May-11
Jul-11
Sep-11
Nov-11
Jan-12
Mar-12
May-12
Jul-12
Sep-12

DNA sequencing

- Capillary Sequencing (Sanger sequencing)
- Next Gen Sequencing Platforms
  - Illumina
  - Pacific Biosciences


Hydrogen bond
phosphodiester bond
5'
3'
P
O
A
T
T
A
T
A
G
C
C
G
d

Why sequence?

1. ***De novo***

   A species genomic sequence allows genetic analysis.

2. **Re-sequencing**

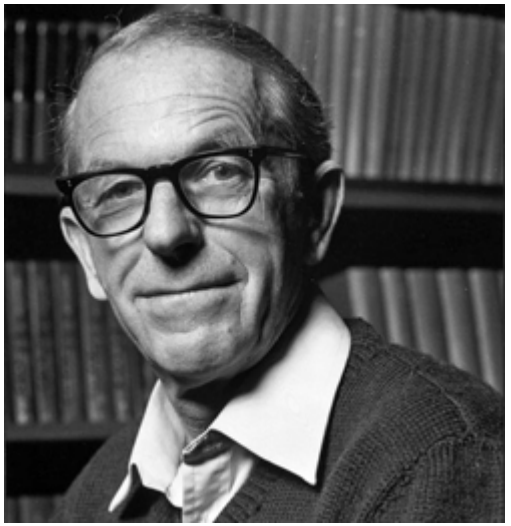
   Understand the nature of genetic variation – population genetics

3. **Counting**

   Functional genomics e.g. gene expression etc.

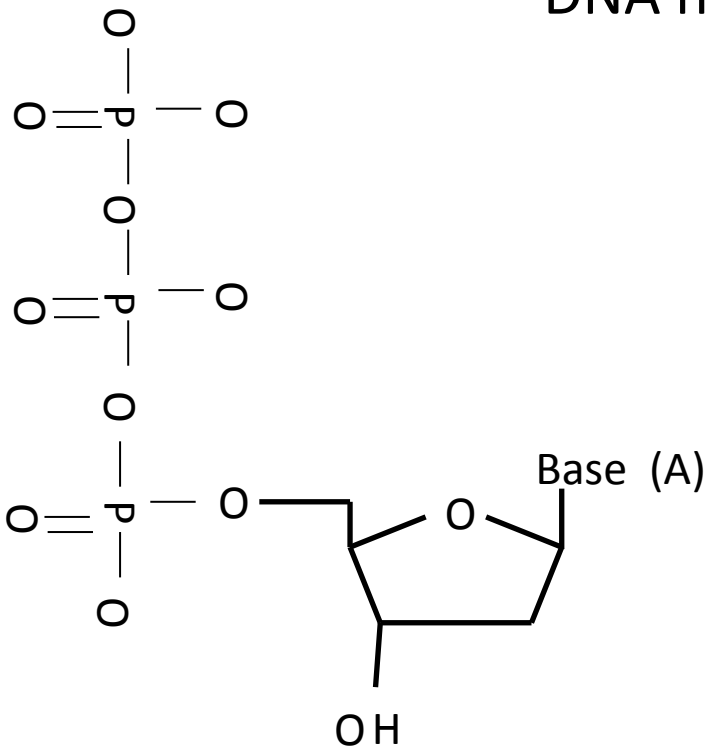
Capillary Sequencing

The DNA sequencing technology developed by Fred Sanger in 1977 is know as the dideoxy method, the chain termination method, or the Sanger sequencing method.



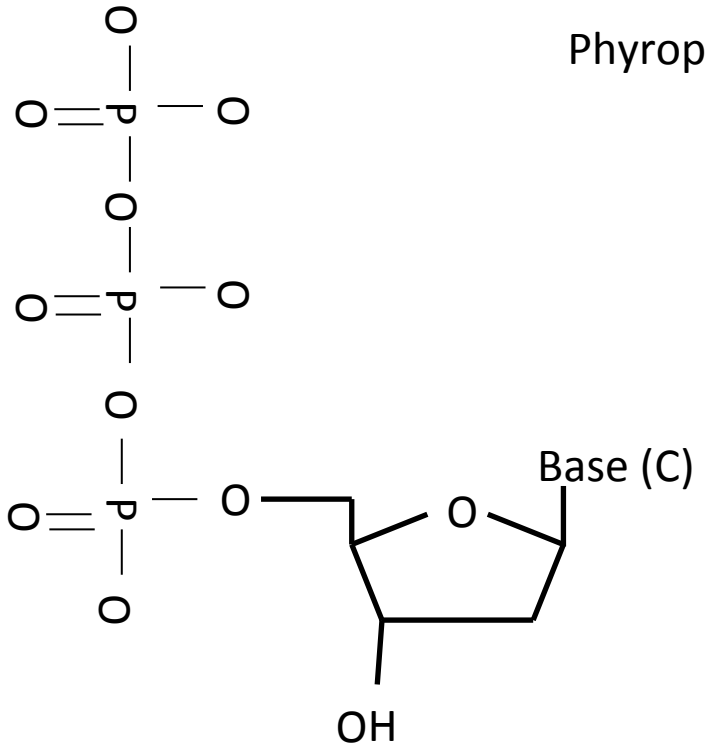
Based on the technique of separation by electrophoresis.

DNA nucleotide binding

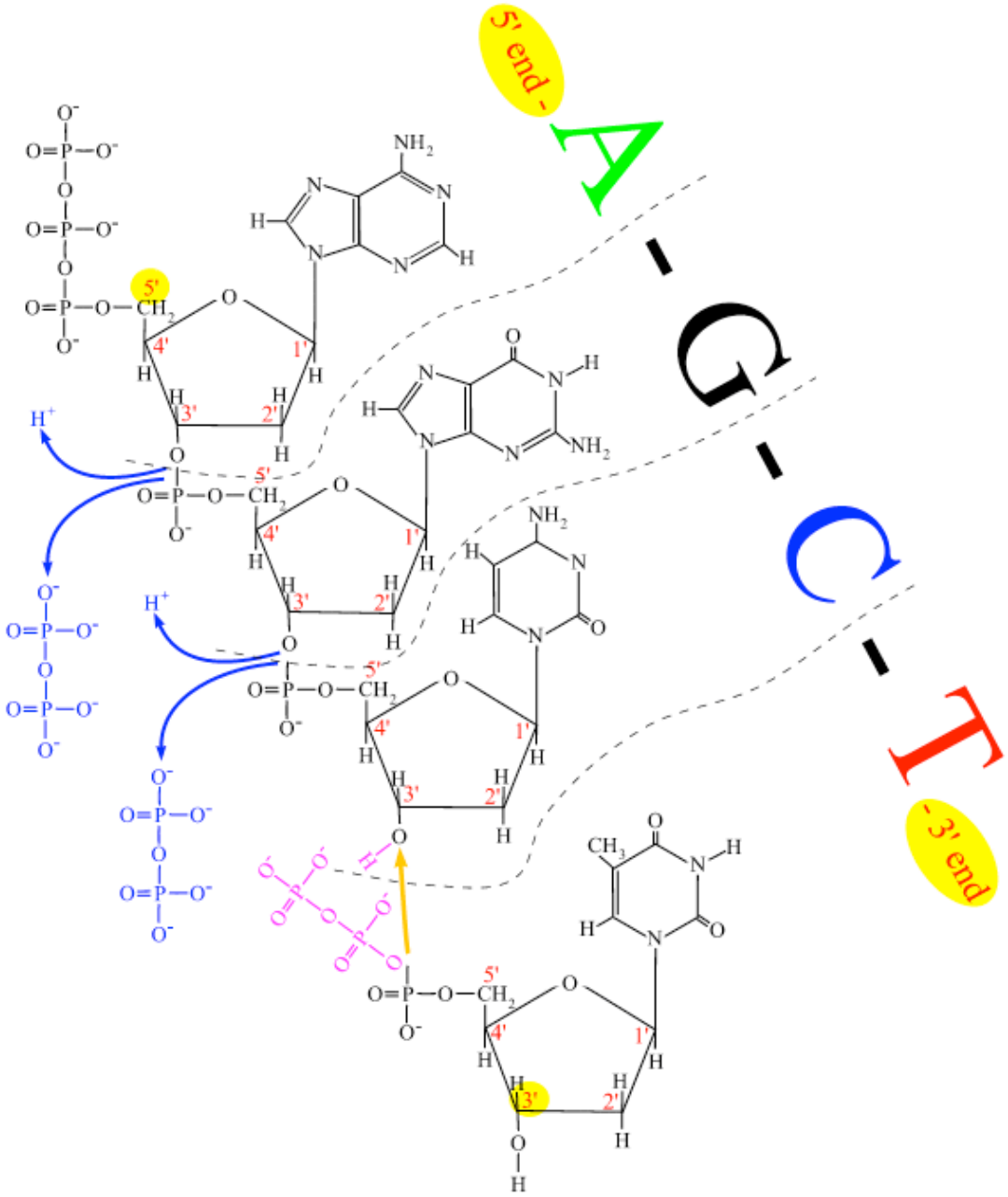


hydrogen

Phyrophosphate



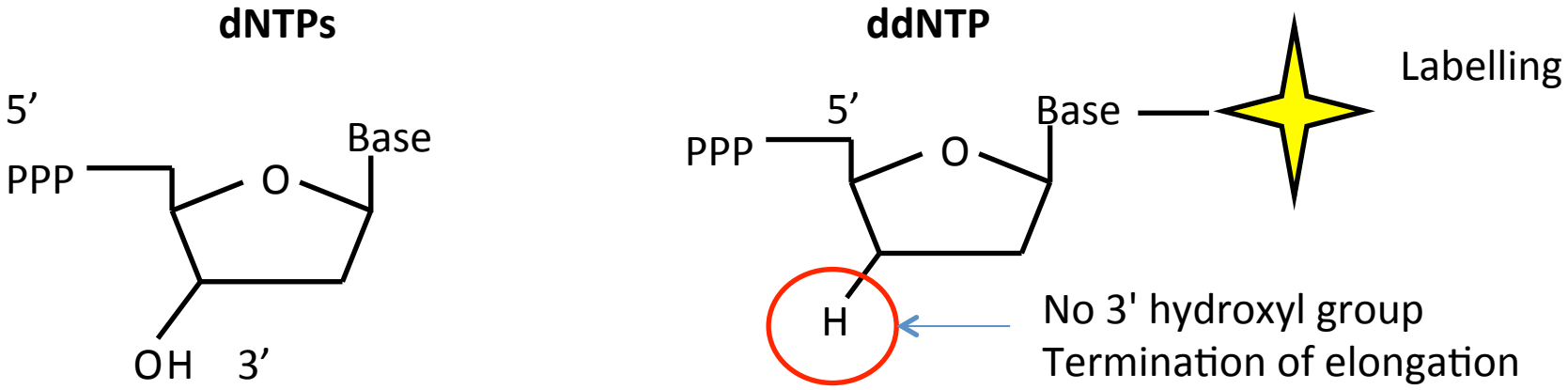
From nucleotide to DNA



(Andy Vierstraete 1999)

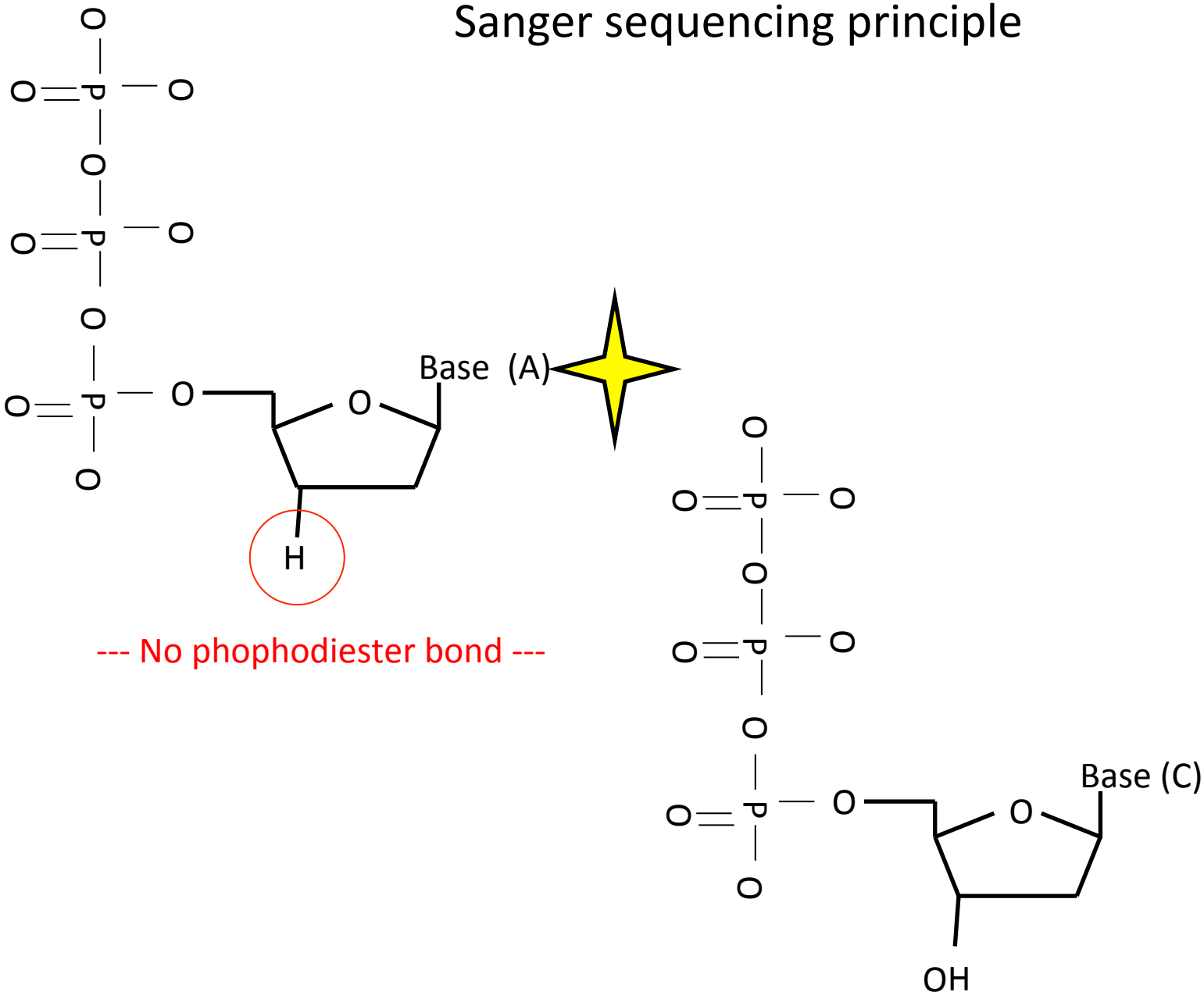
Capillary Sequencing – Chain Termination

Key principle of Sanger sequencing is the use of (radioactive or fluorescence) labelled dideoxy NTPs (ddNTPs) which terminate elongation of a DNA fragment

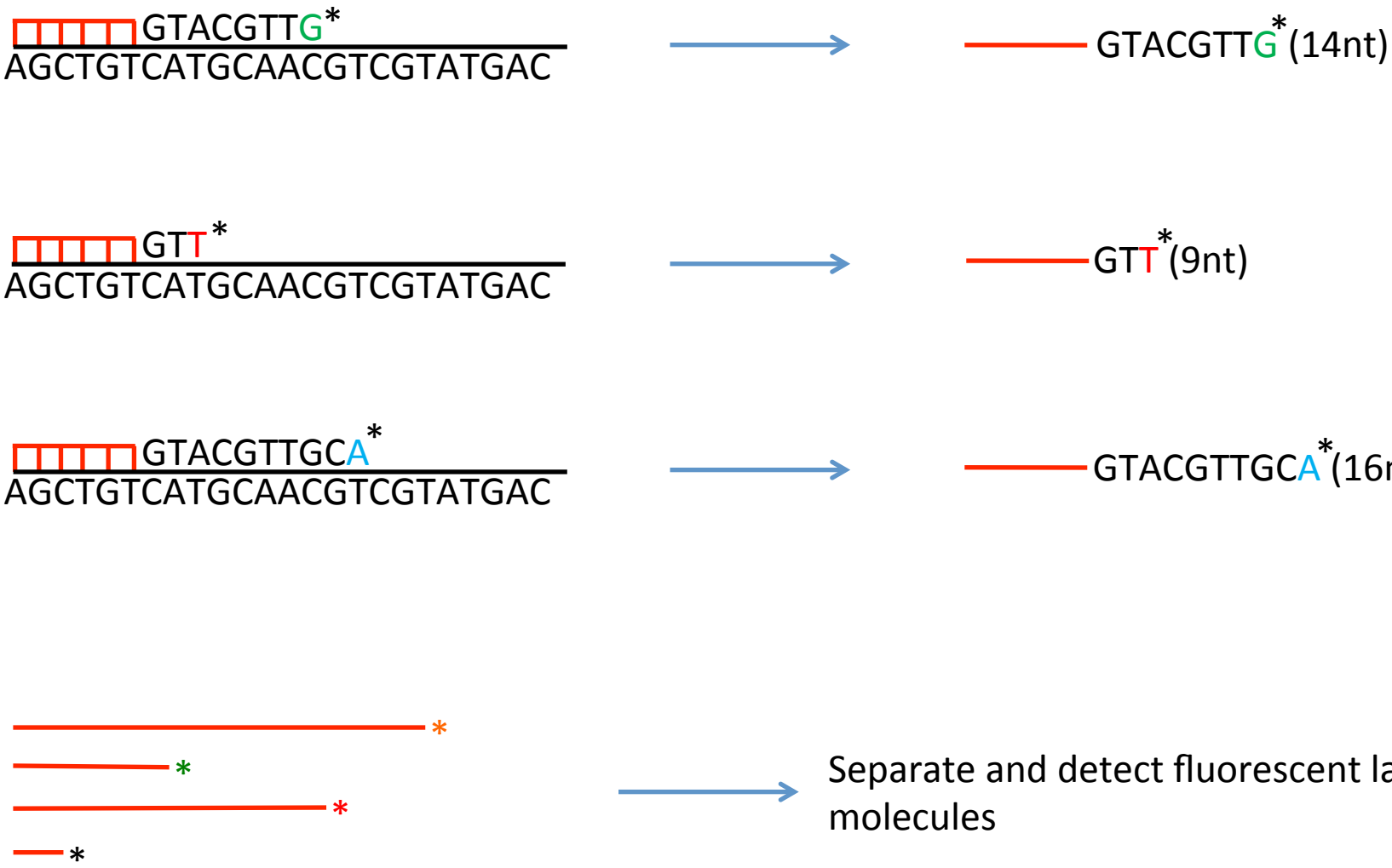


Normal and labelled ddNTPs in reaction

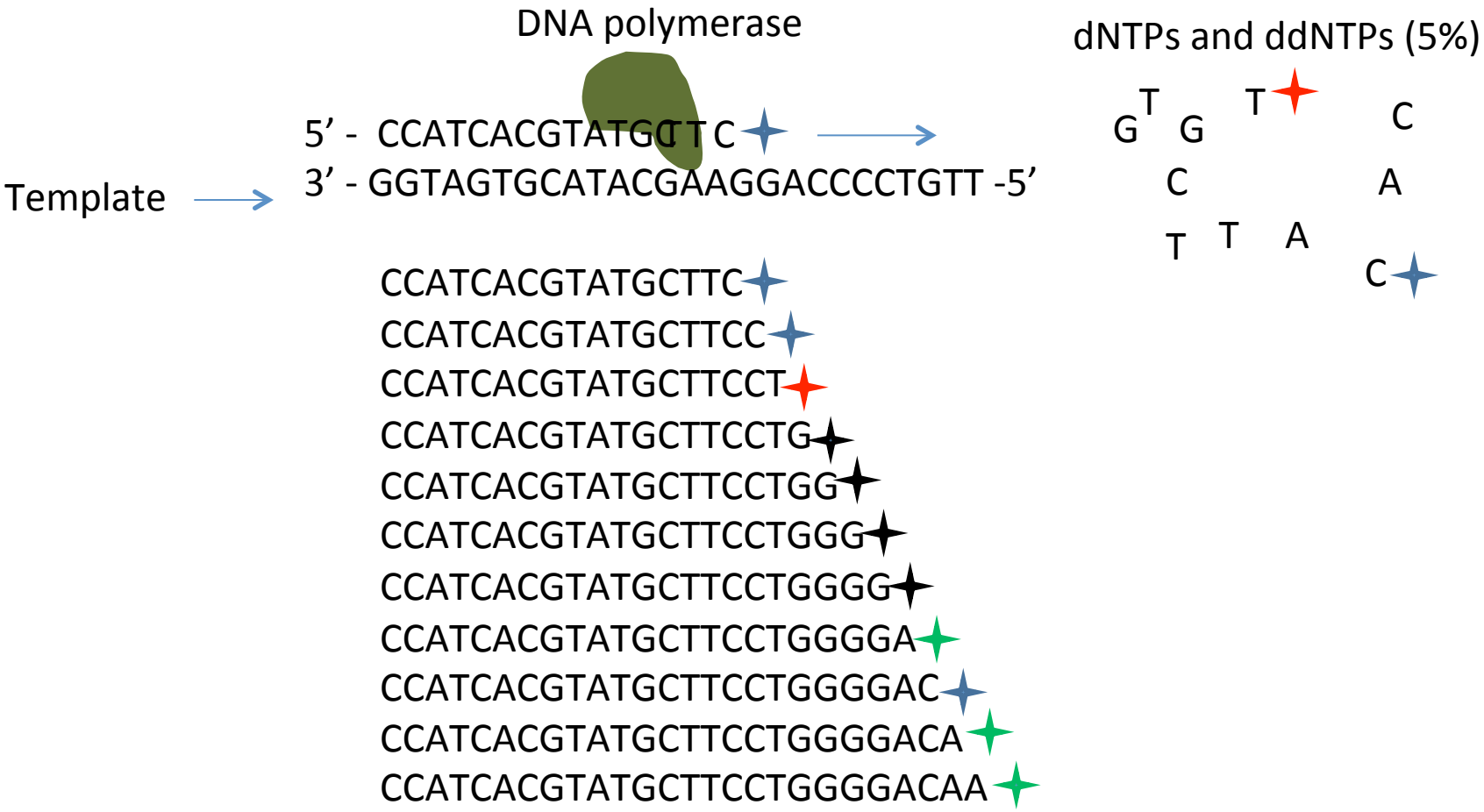
Sanger sequencing principle



Amplification is stopped for THIS molecule

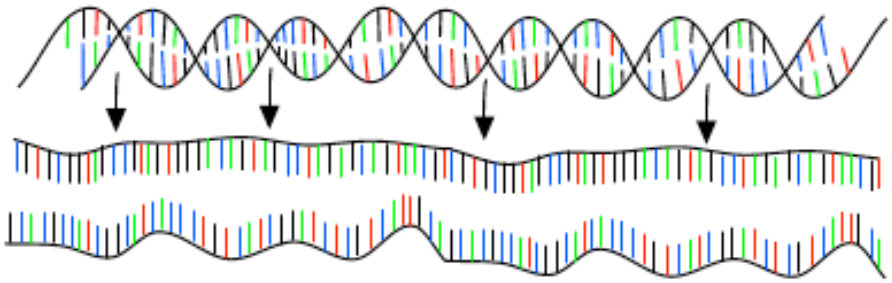

GTACGTTG*
AGCTGTCATGCAACGTCGTATGAC
GTACGTTG*(14nt)
GTT*
AGCTGTCATGCAACGTCGTATGAC
GTT*(9nt)
GTACGTTGCA*
AGCTGTCATGCAACGTCGTATGAC
GTACGTTGCA*(16nt)
Separate and detect fluorescent labeled molecules
Add:
- DNA polymerase
- dNTPs
- primer
- ddNTPs (5%)
- Buffer

Chain Termination



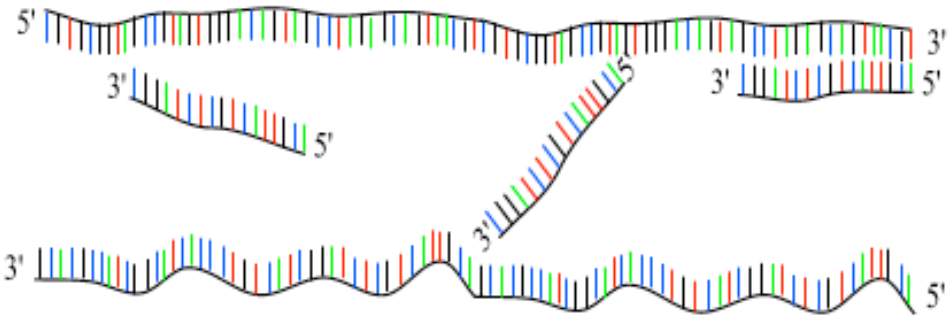
Sequencing

30 cycles of 3 steps :



Step 1 : denaturation

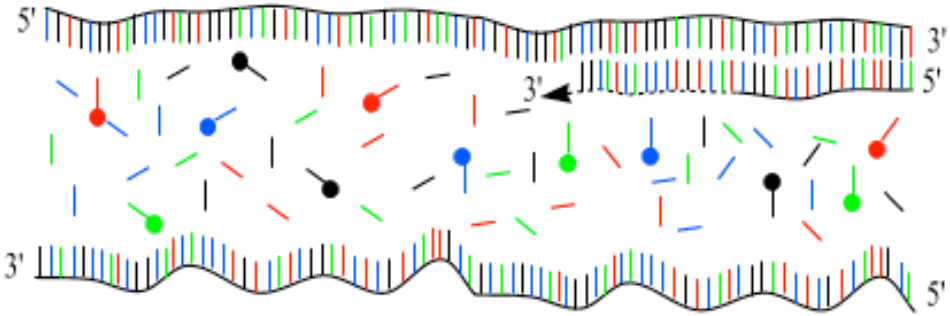
1 minut 94 °C



Step 2 : annealing

15 seconds 50 °C

1 primer !!!!

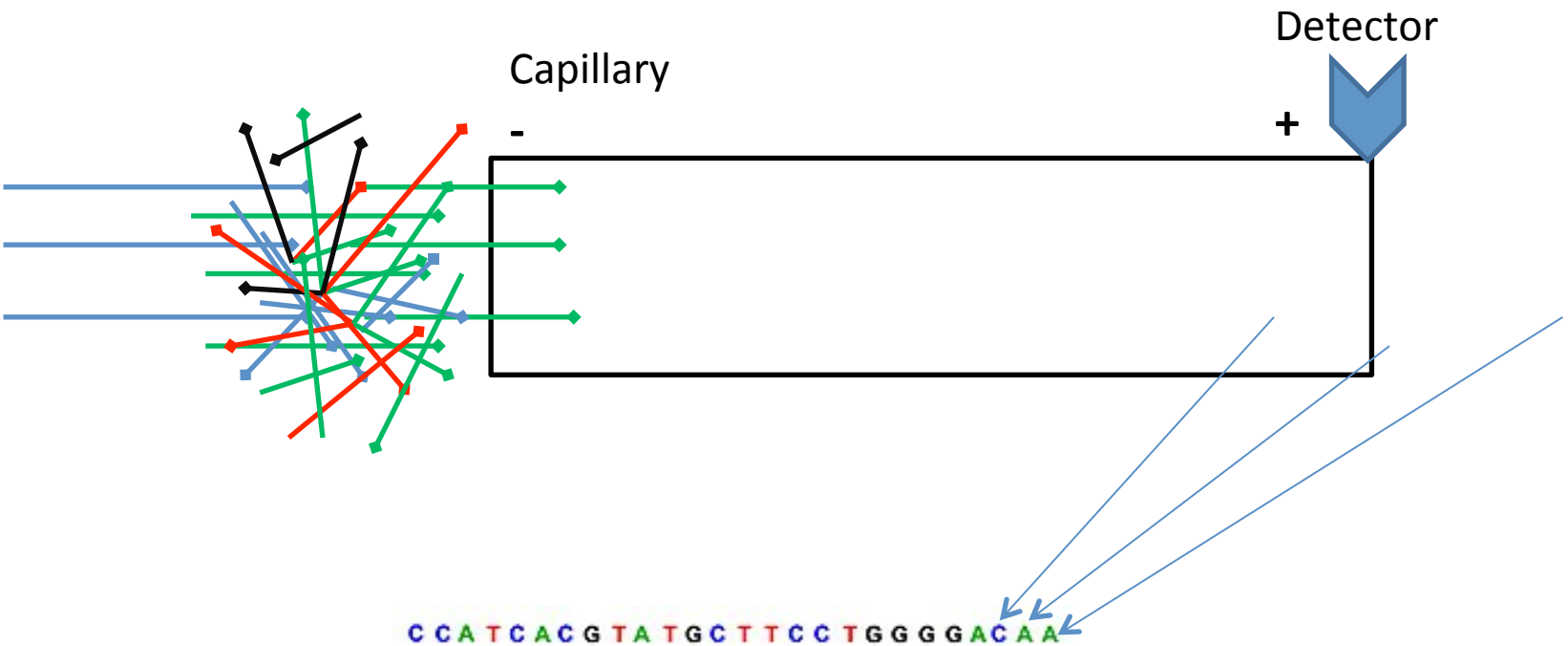


Step 3 : extension

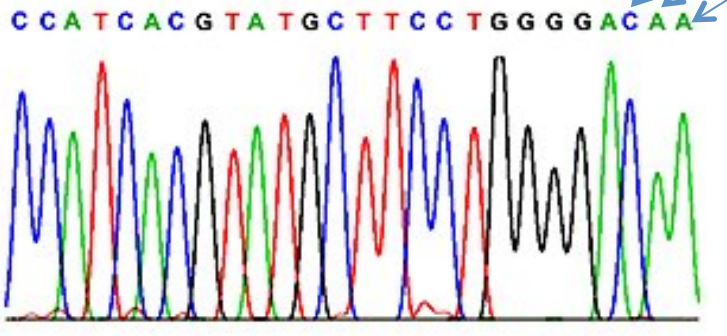
4 minutes 60 °C
mixture of dNTP's
and ddNTP's

(Andy Vierstraete 1999)

Capillary Sequencing

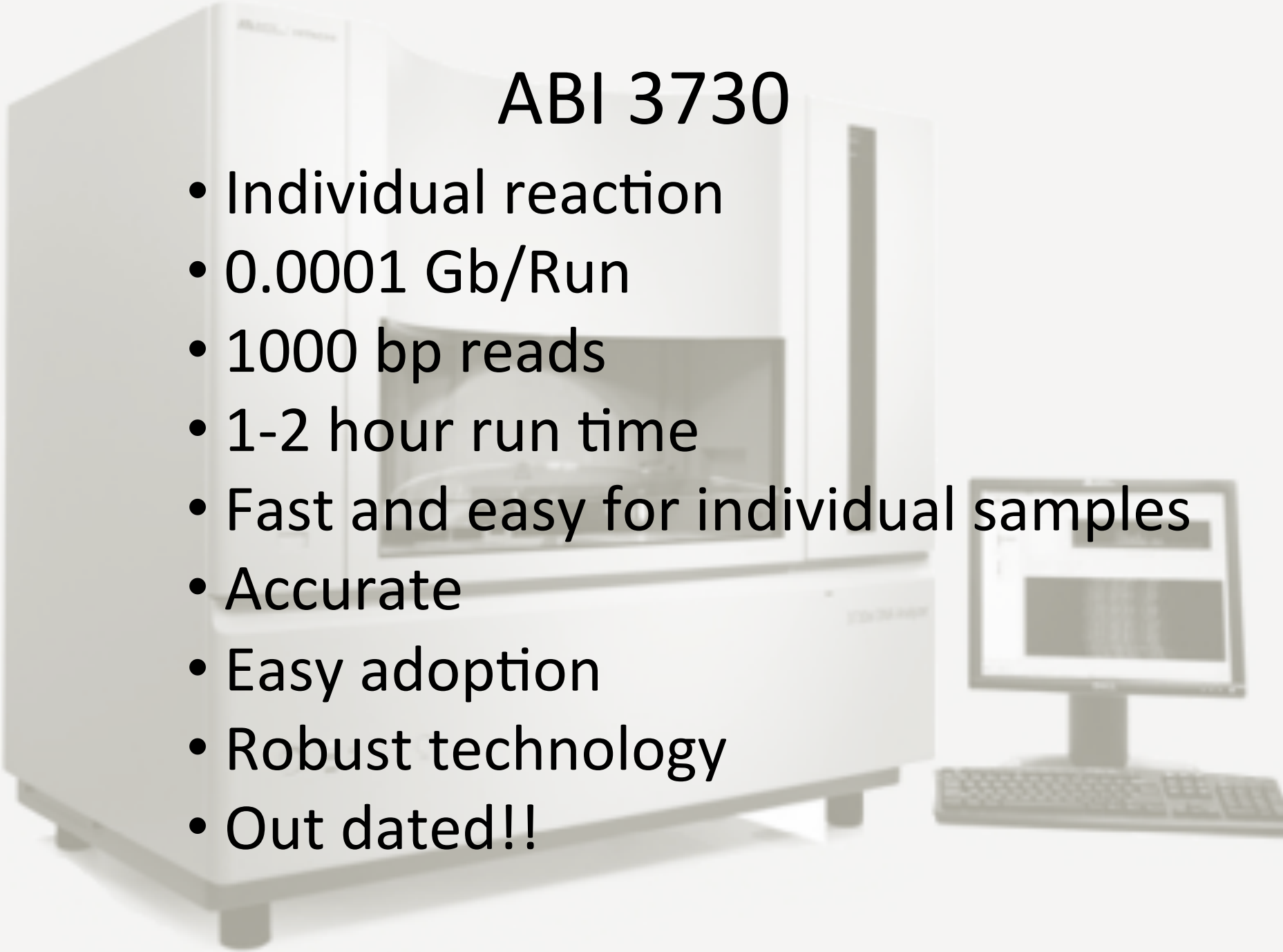


Chromatogram



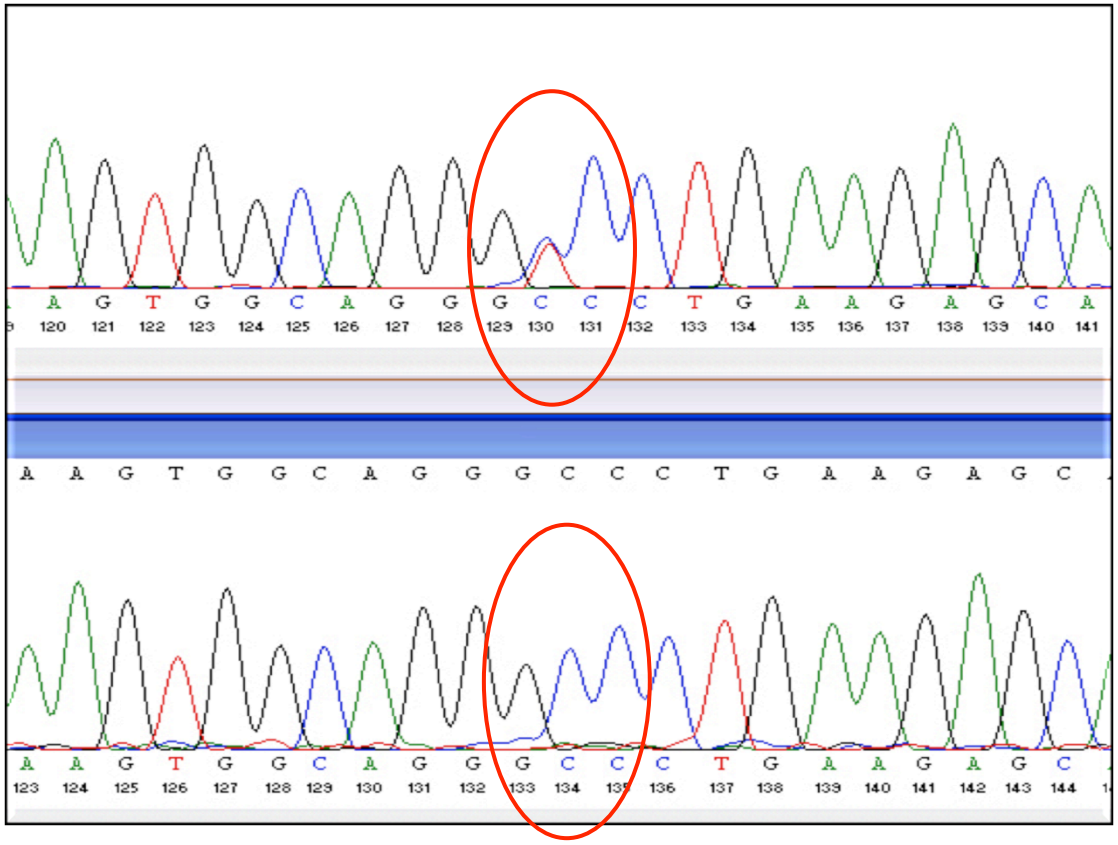
ABI 3730

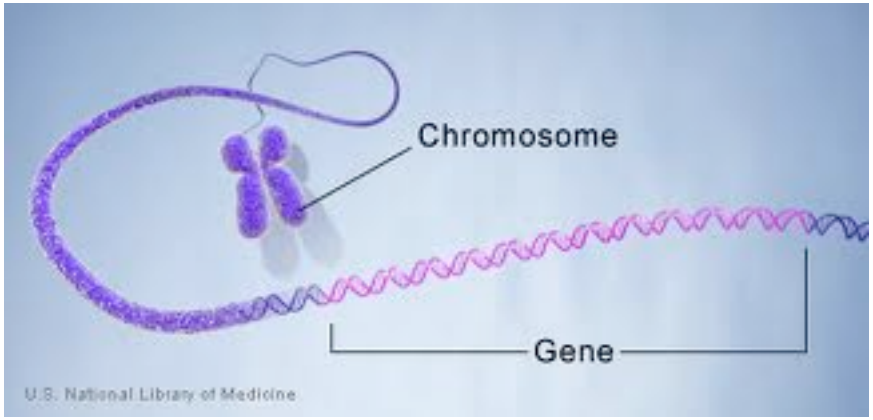
- Individual reaction
- 0.0001 Gb/Run
- 1000 bp reads
- 1-2 hour run time
- Fast and easy for individual samples
- Accurate
- Easy adoption
- Robust technology
- Out dated!!



Question:

1) How can you explain the following sanger sequence chromatogram?

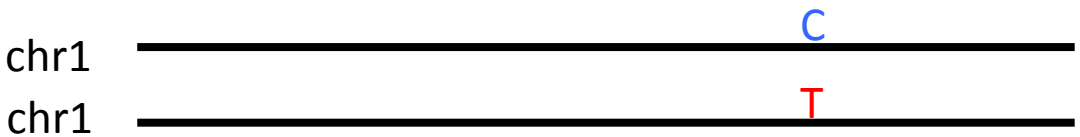




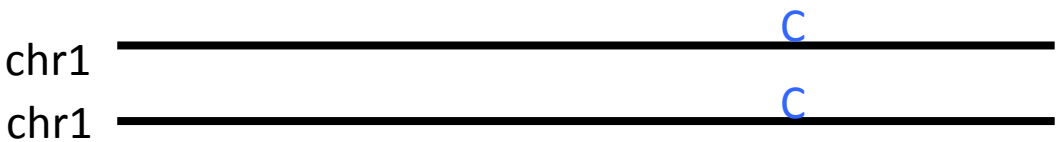
2 copies of the chromosome in one cell
Amplify the gene by PCR

→ homozygous: these tow copies are IDENTICAL

→ Heterozygous: these two copies are NOT identical (SNPs)



Heterozygous



Homozygous

Next Gen Sequencing (NGS) Platforms

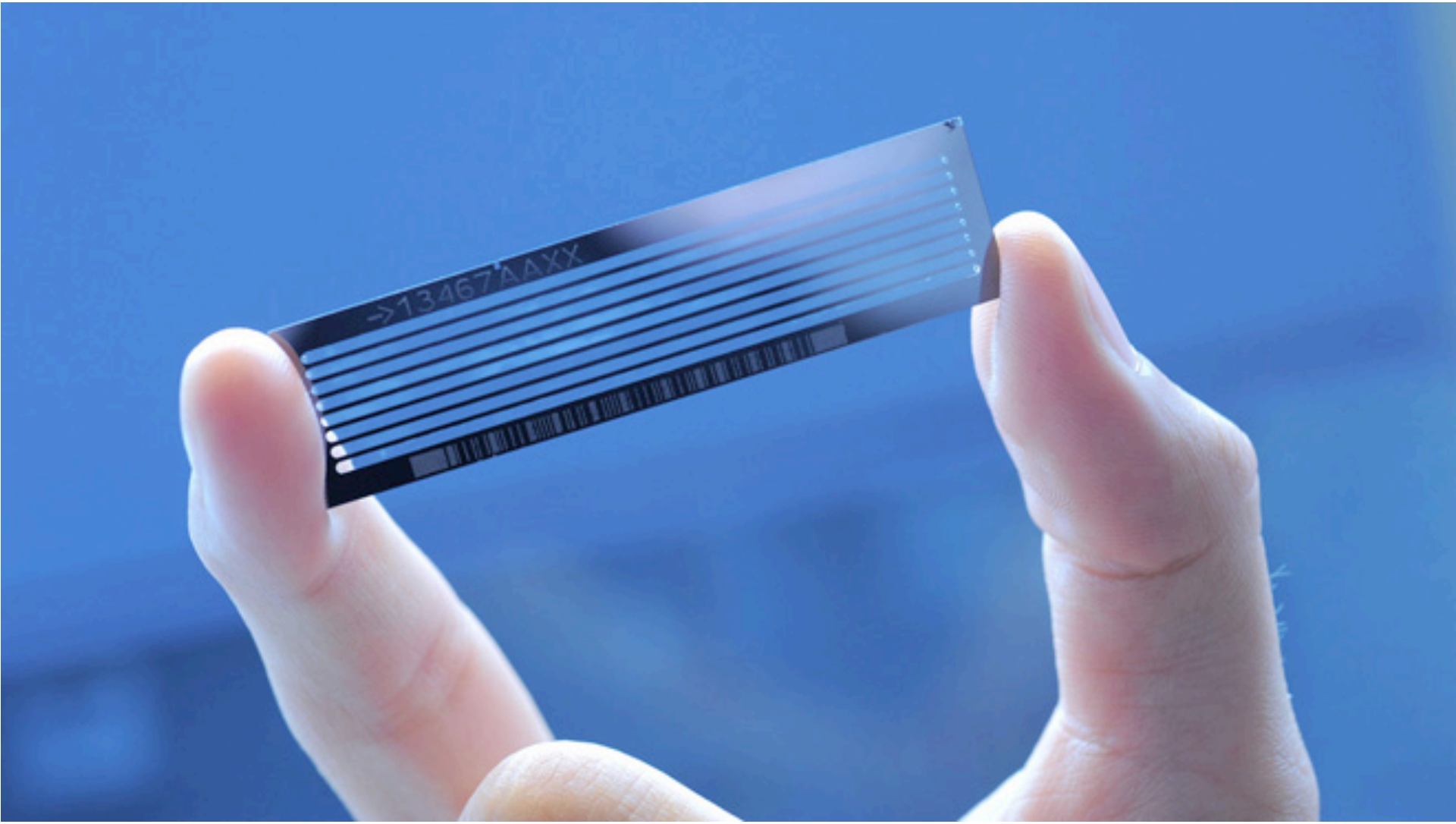
Illumina

Pacific Biosciences (PacBio)

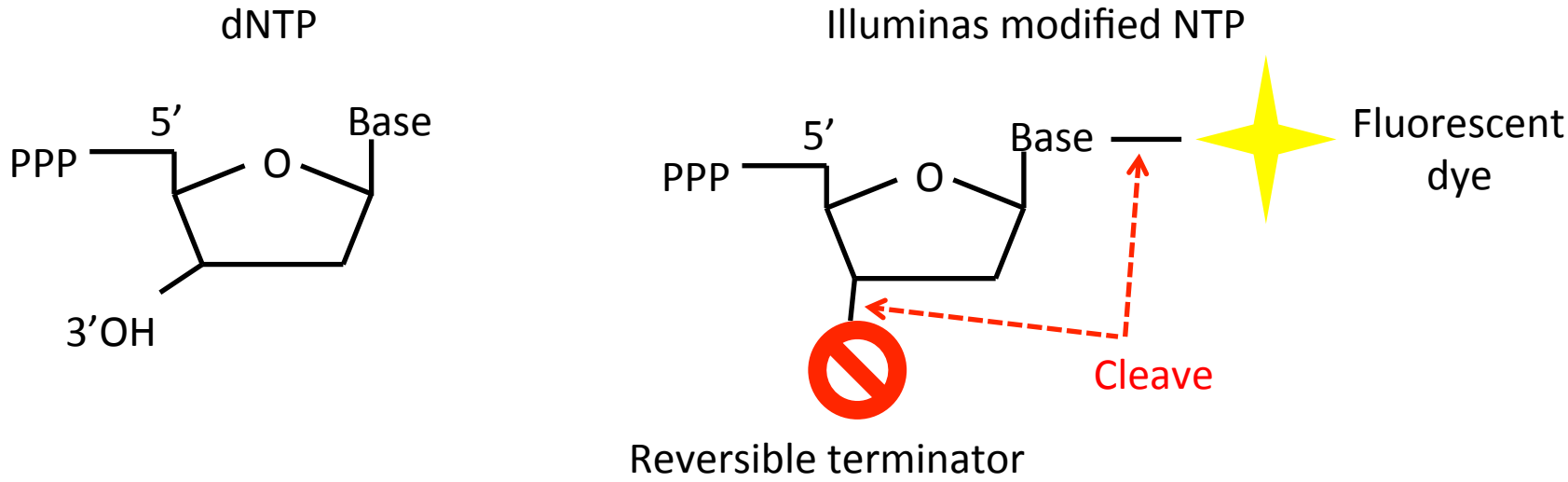
Illumina Sequencing

Has become the most popular next generation sequencing platform.

Synthetic sequencing using fluorescent reversible terminator technology.

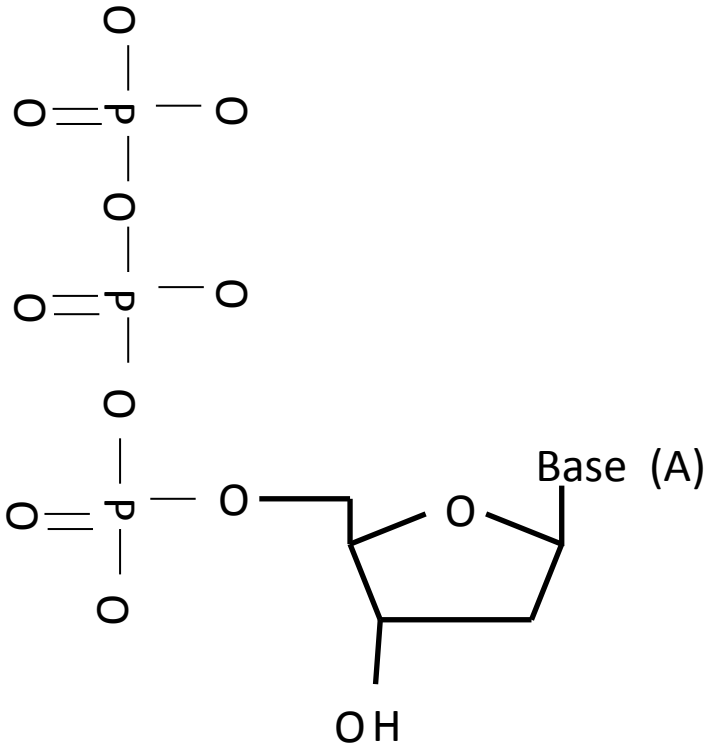

->13467AAXX

Illumina Sequencing

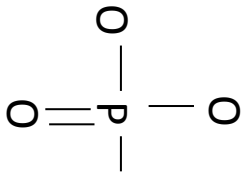


Every single base labelled (Each of the 4 bases have a different fluorescent dye)

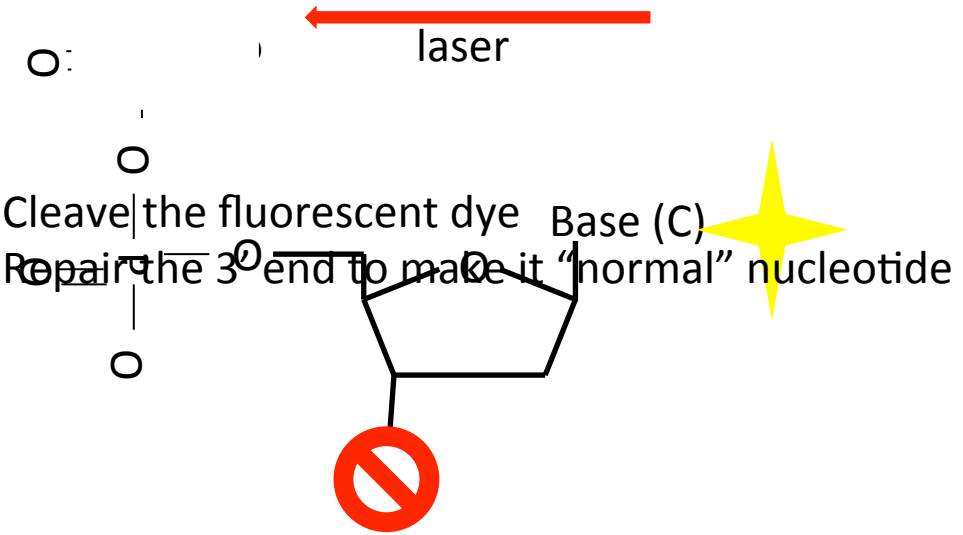
Every base has a terminator group -**cleaved off** after each round of sequencing



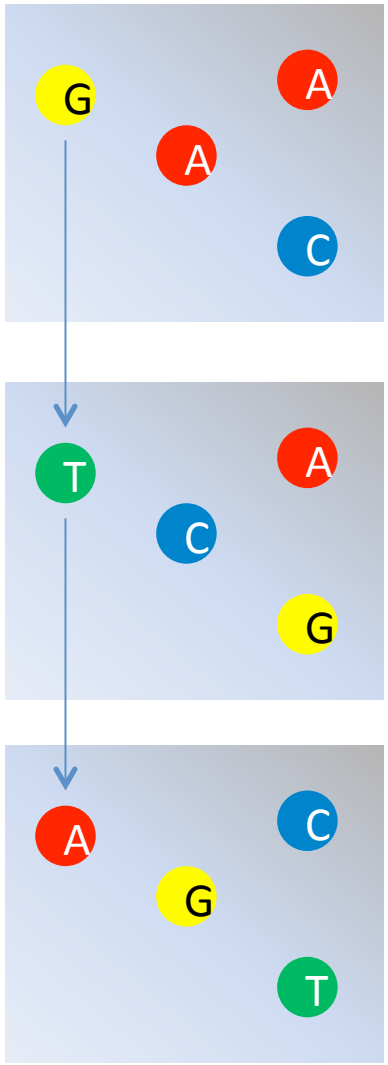
hydrogen



Phyrophosphate



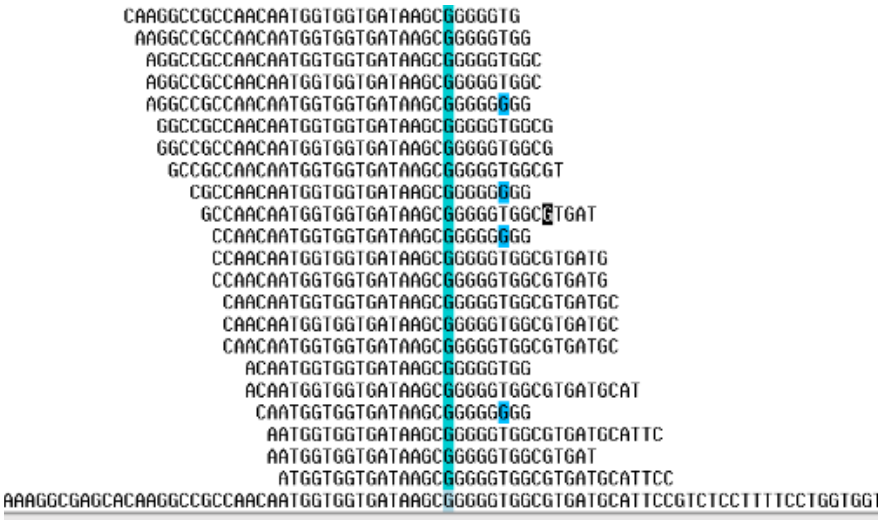
Illumina Base Calling



Cycle 1

Cycle 2

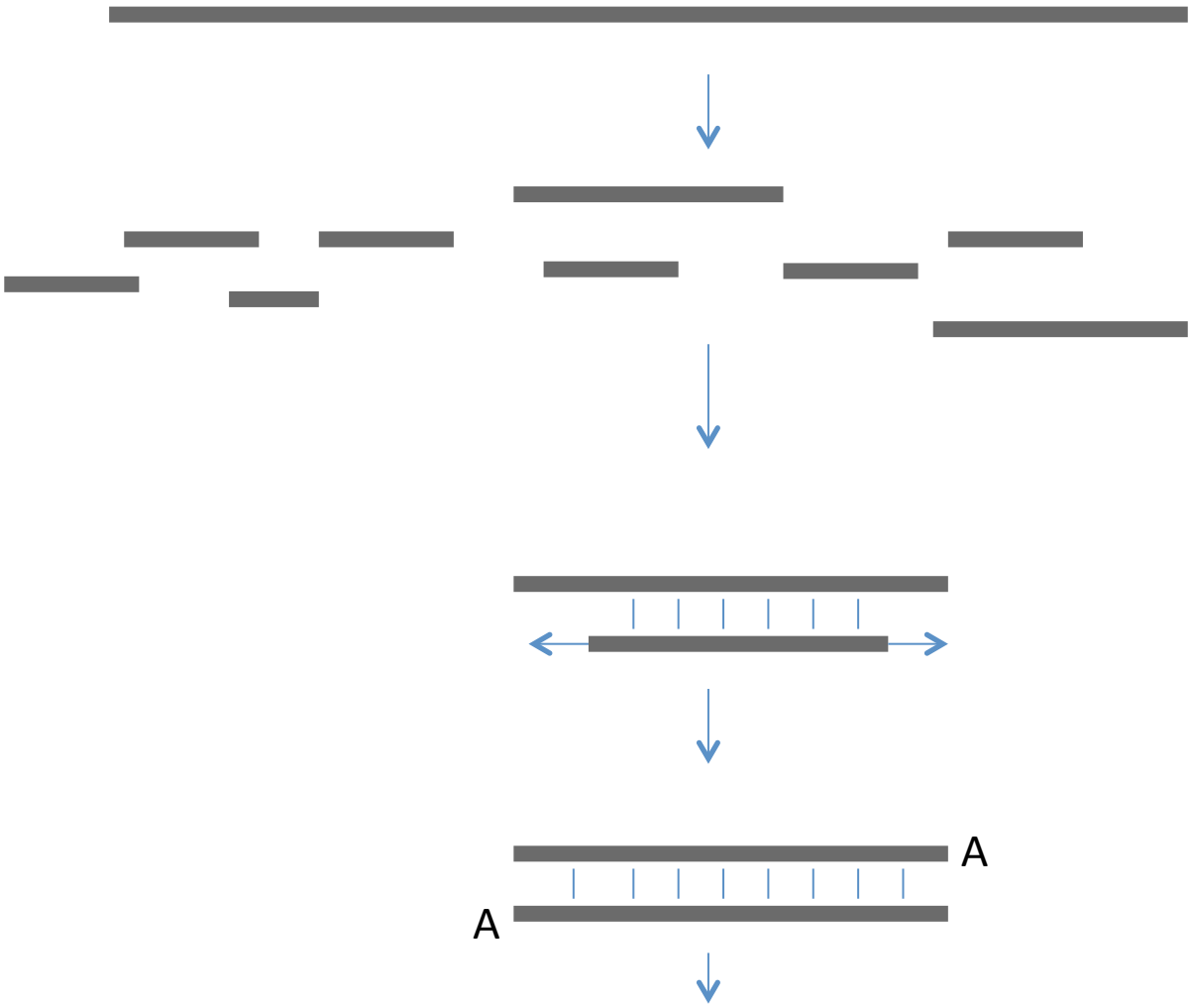
Cycle 3



Aligned to consensus

G T A....

Illumina library



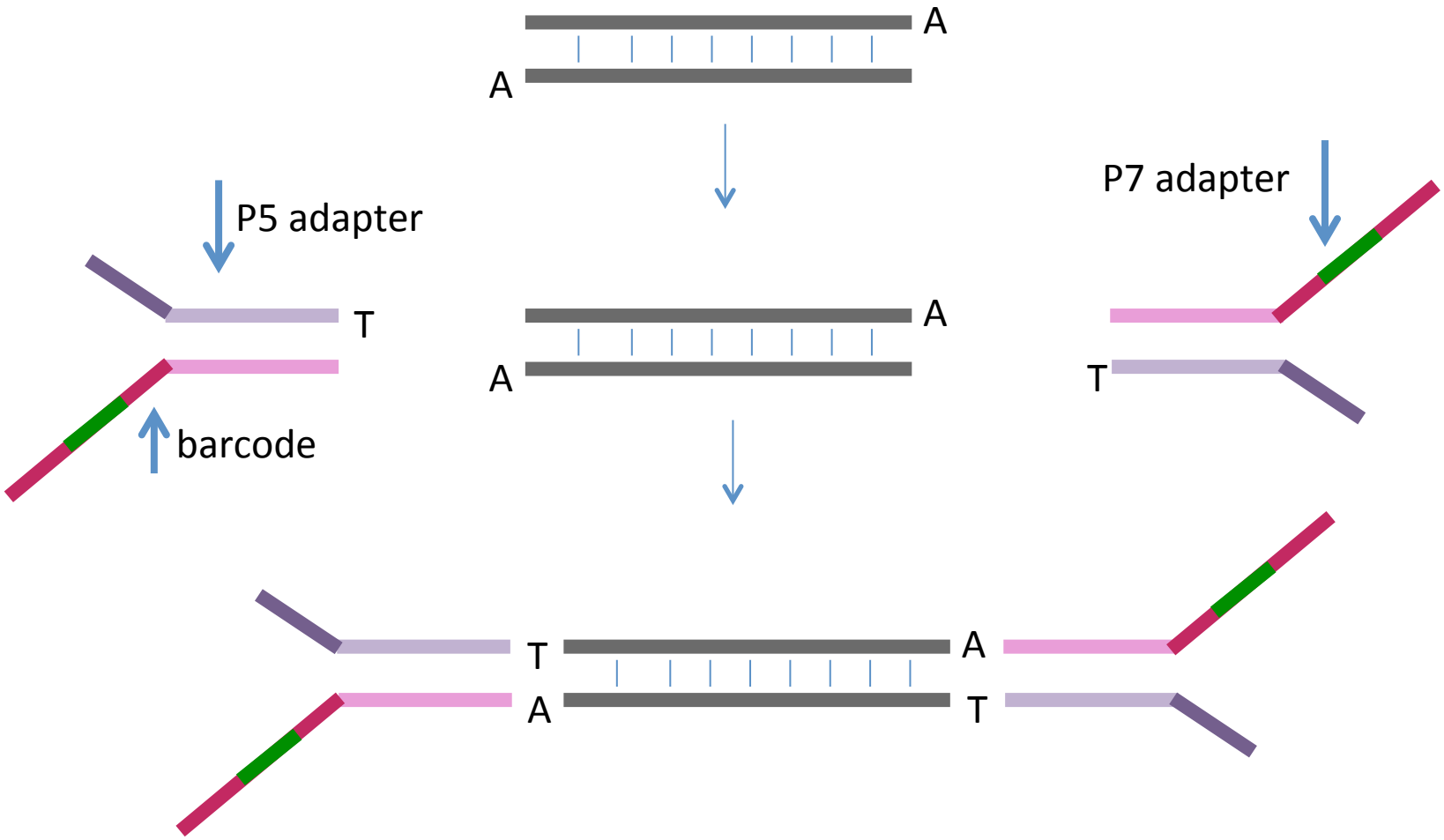
Genome

Fragment Genome
50 to 500bp

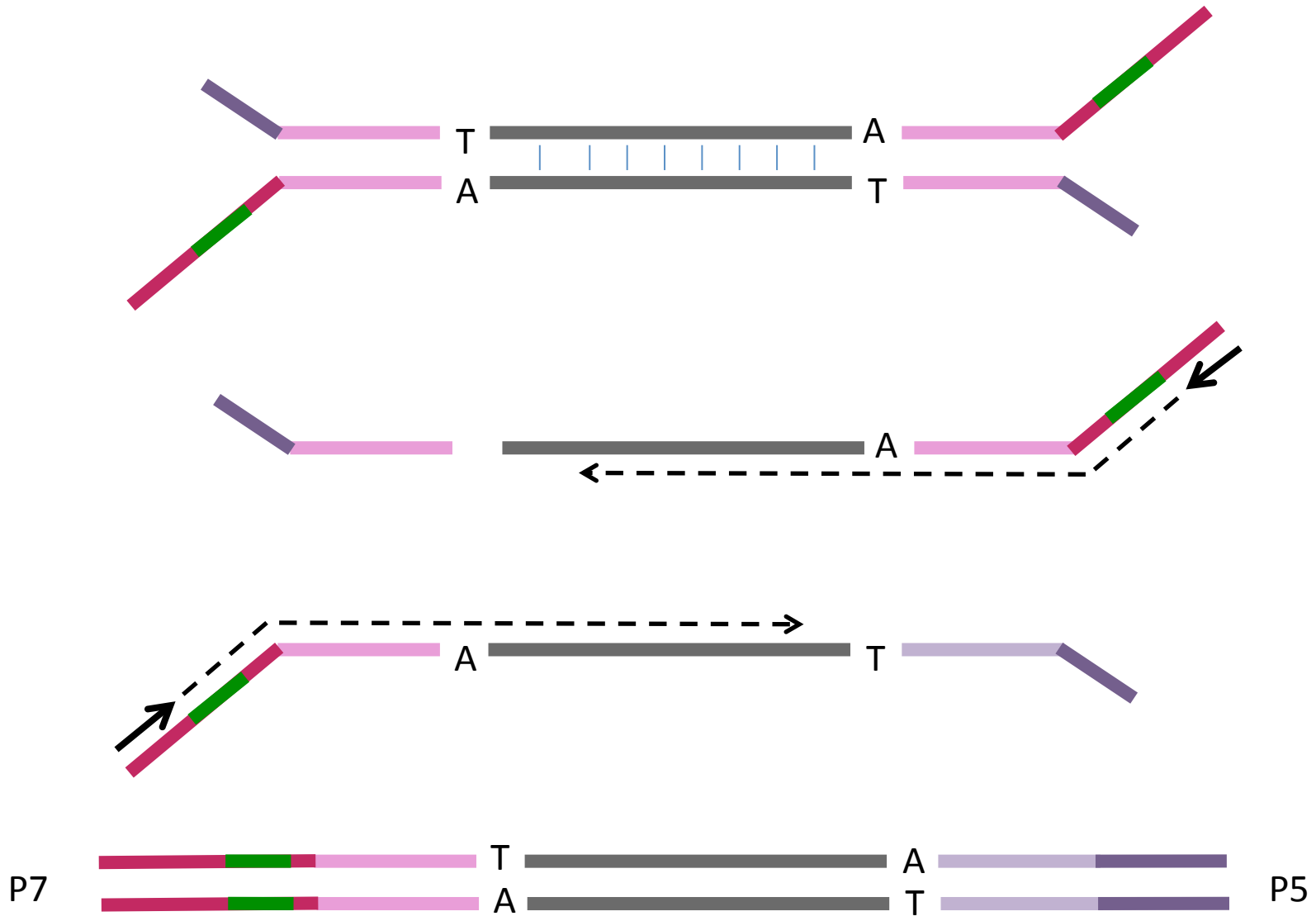
End Repair :DNA ends are blunted and phosphorylated

A tailing to the 3’-end of each DNA fragment

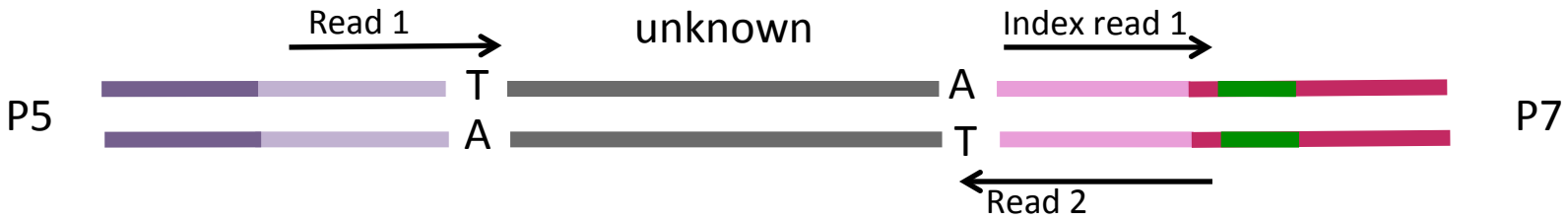
Illumina library



Illumina library



A billion times, coverage the entire genome



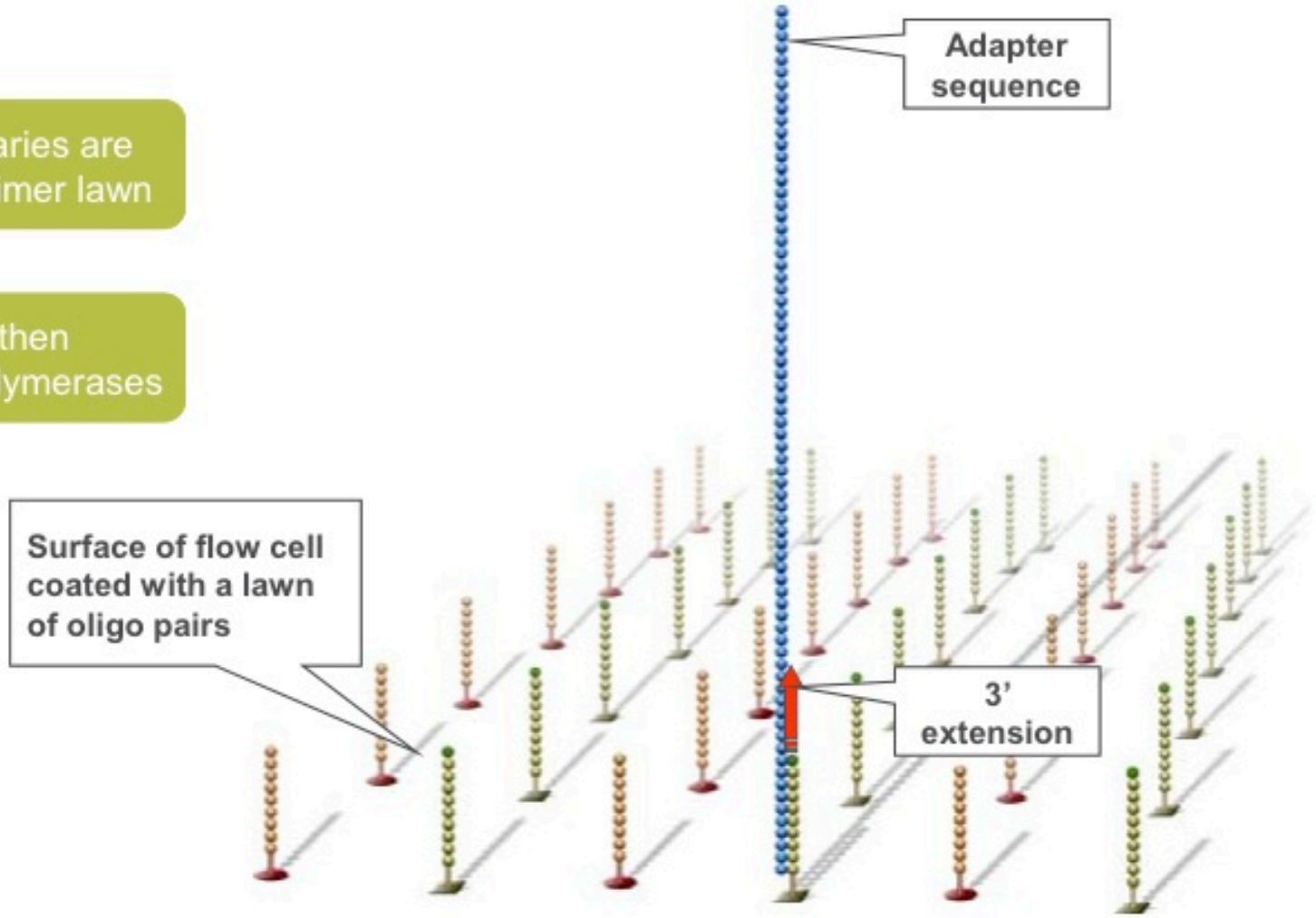
5’- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC* T 3’

5’- /5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC BC ATCTCGTATGCCGTCTTCTGCTT*G 3

Hybridize Fragment & Extend

Single DNA libraries are hybridized to primer lawn

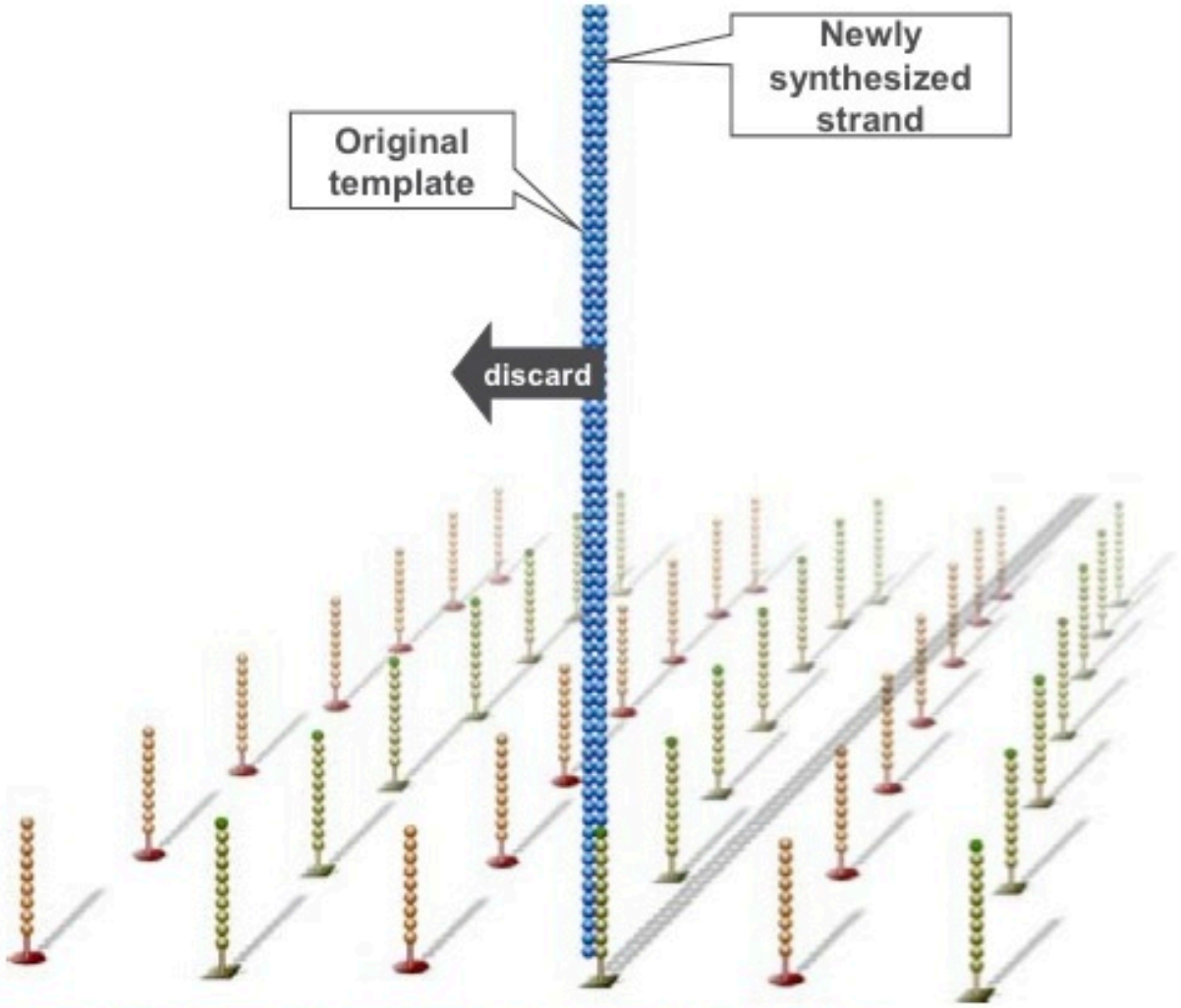
Bound libraries then extended by polymerases



illumina

Denature Double-Stranded DNA

- Double-stranded molecule is denatured
- Original template washed away
- Newly synthesized strand is covalently attached to flow cell surface

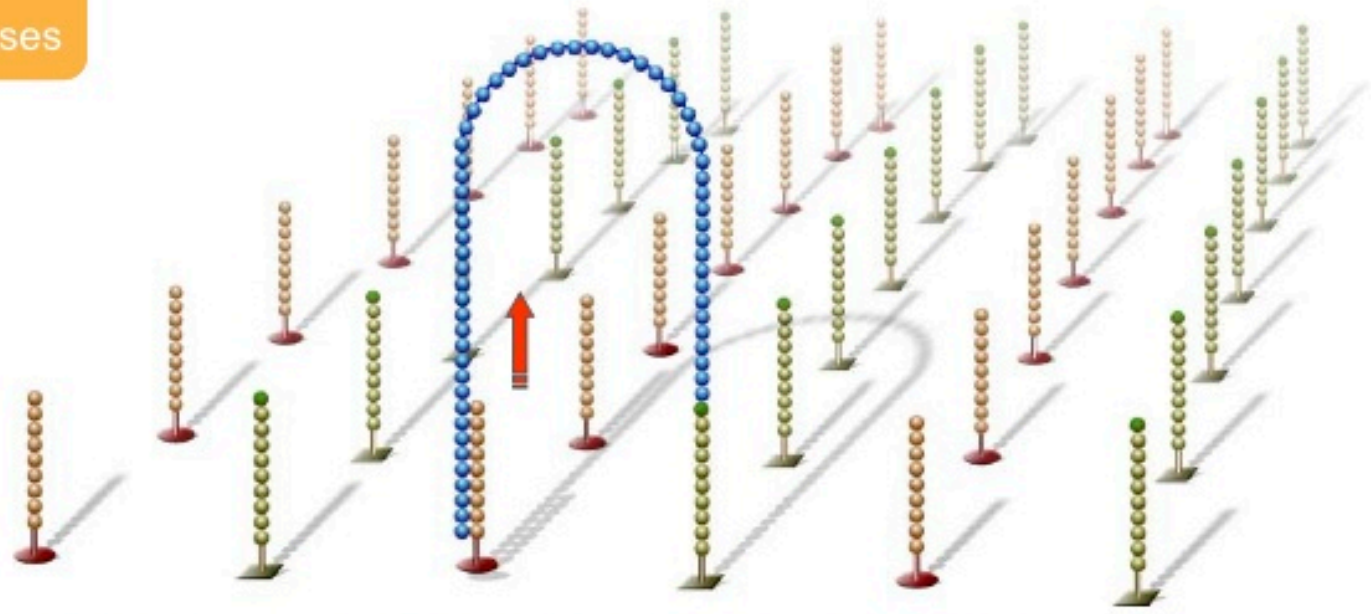


illumina

Bridge Amplification

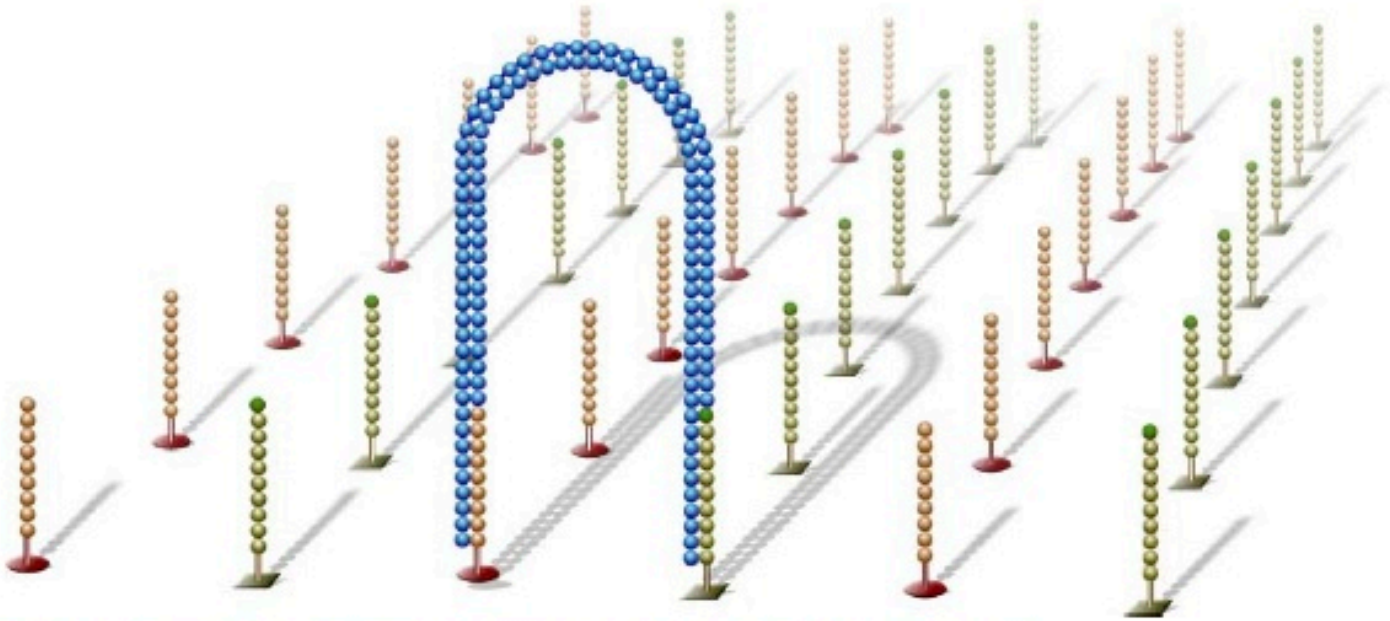
Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases



Bridge Amplification

Double-stranded bridge is formed

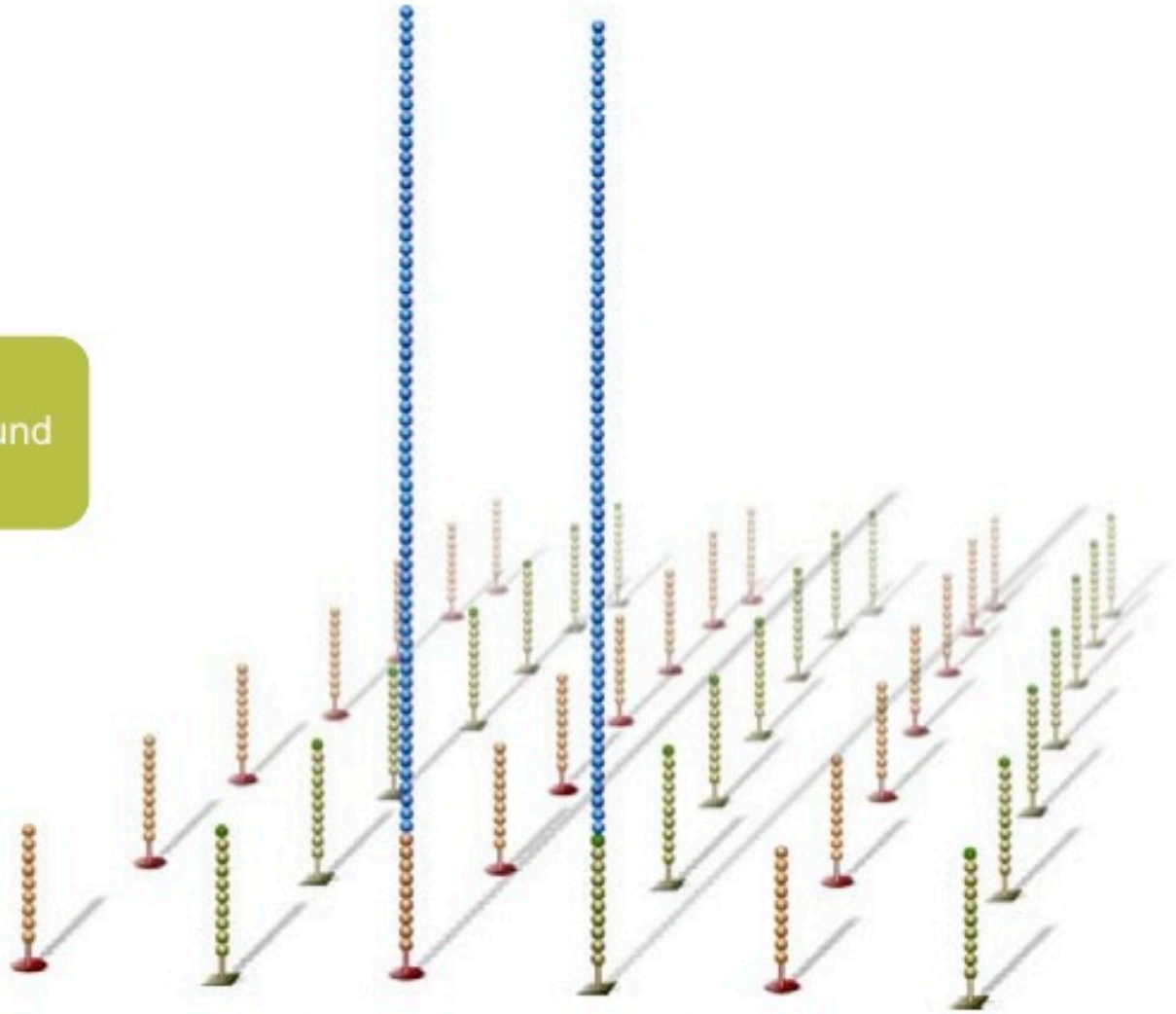


illumina®

Denature Double-Stranded Bridge

Double-stranded bridge is denatured

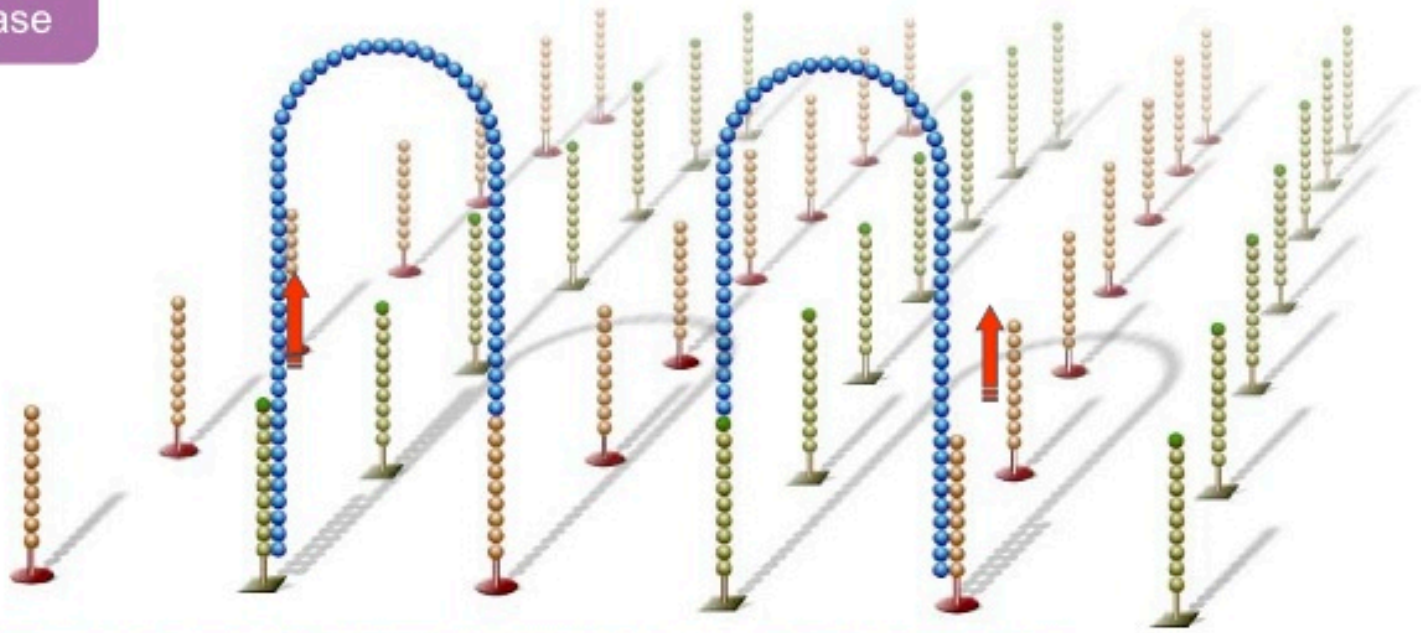
Result:
Two copies of covalently bound single-stranded templates



Bridge Amplification

Single-stranded molecules flip over to hybridize to adjacent primers

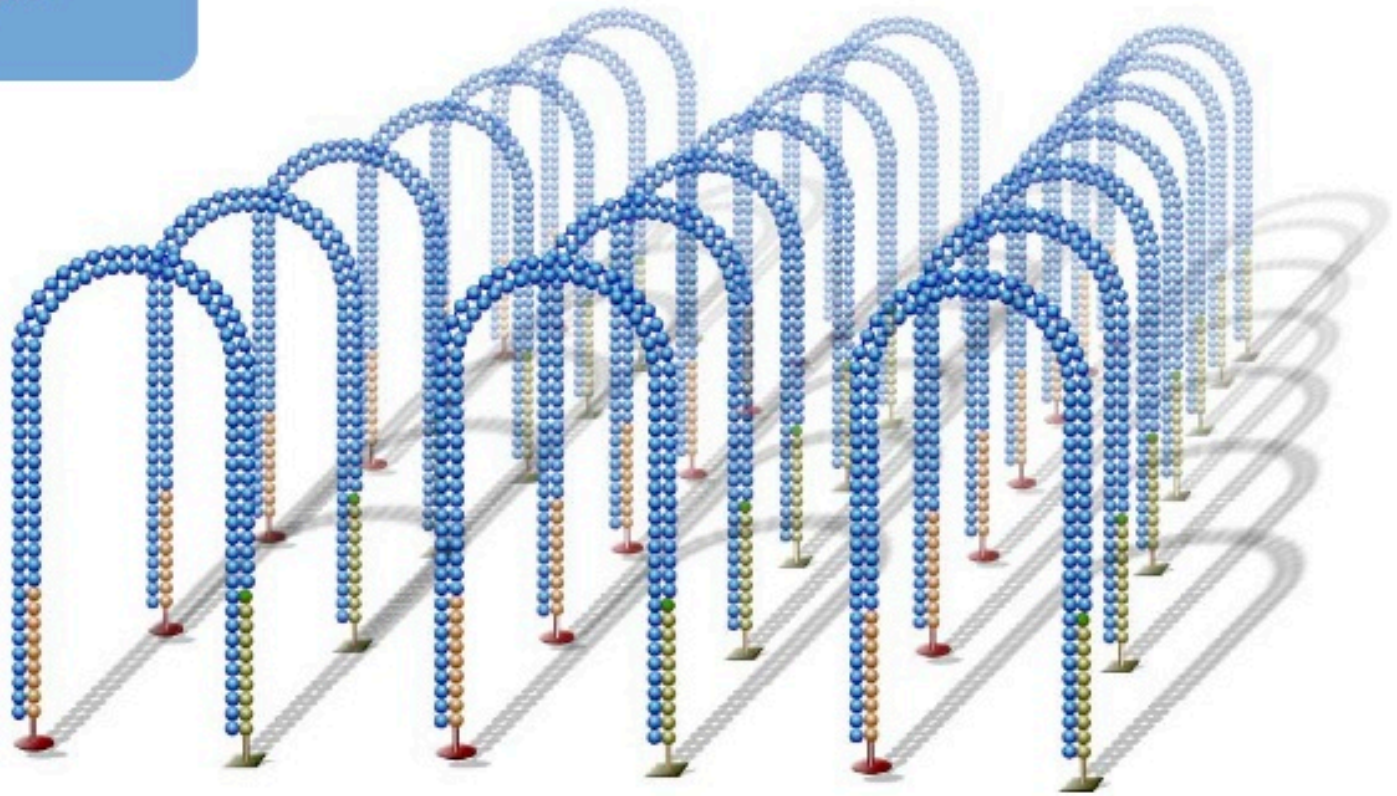
Hybridized primer is extended by polymerase



illumina

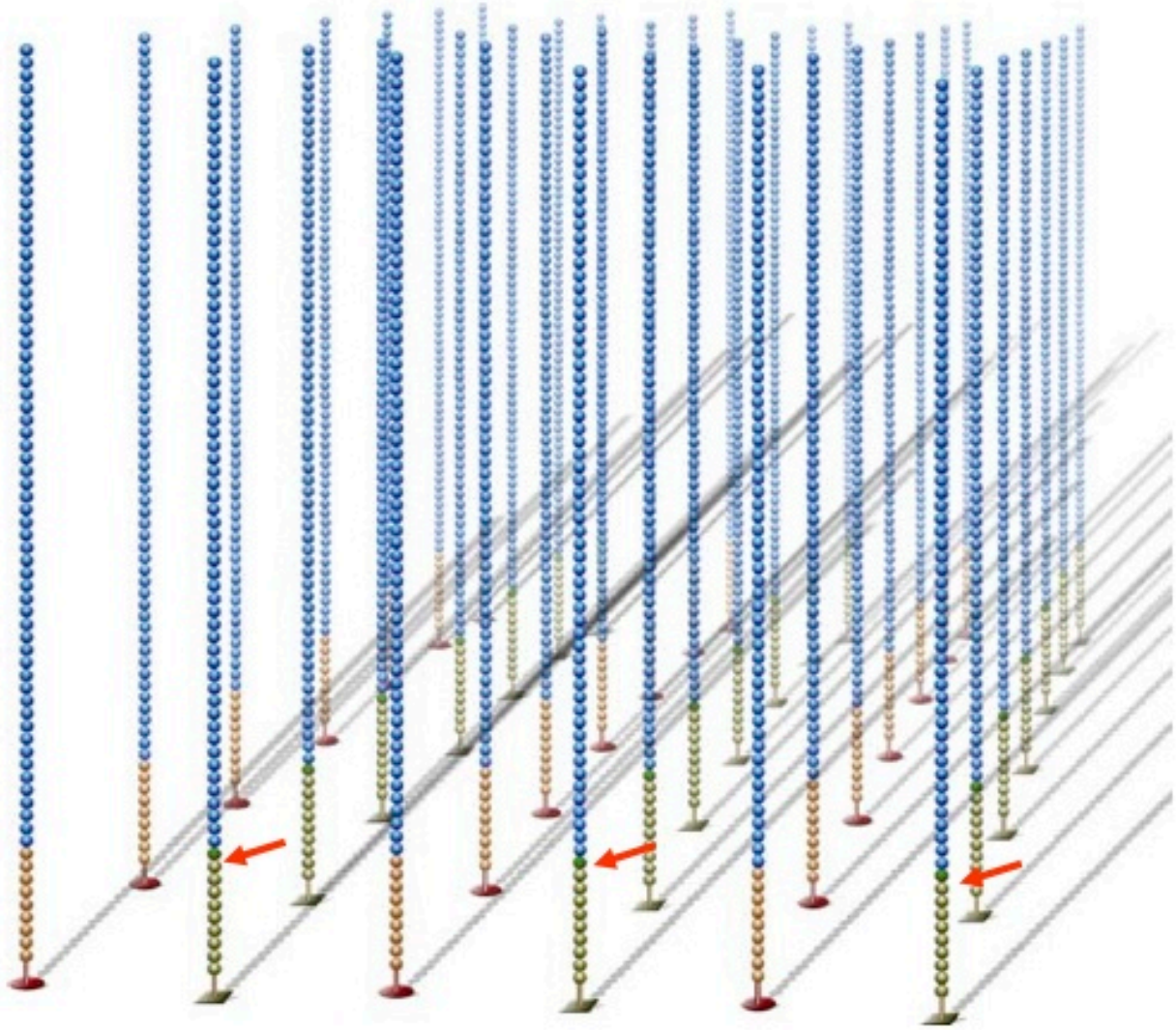
Bridge Amplification

Bridge amplification cycle repeated until multiple bridges are formed



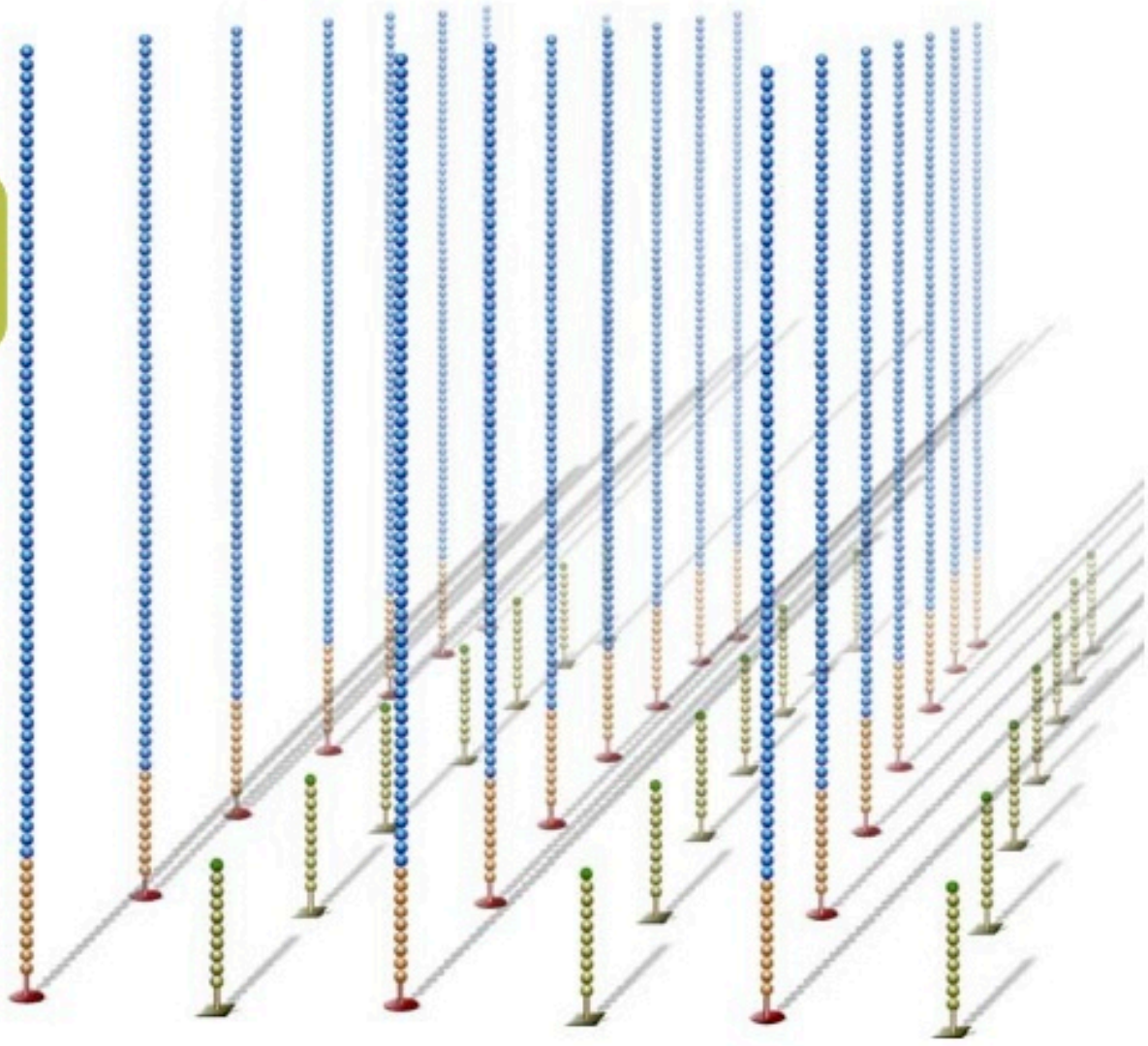
Linearization

dsDNA bridges are denatured



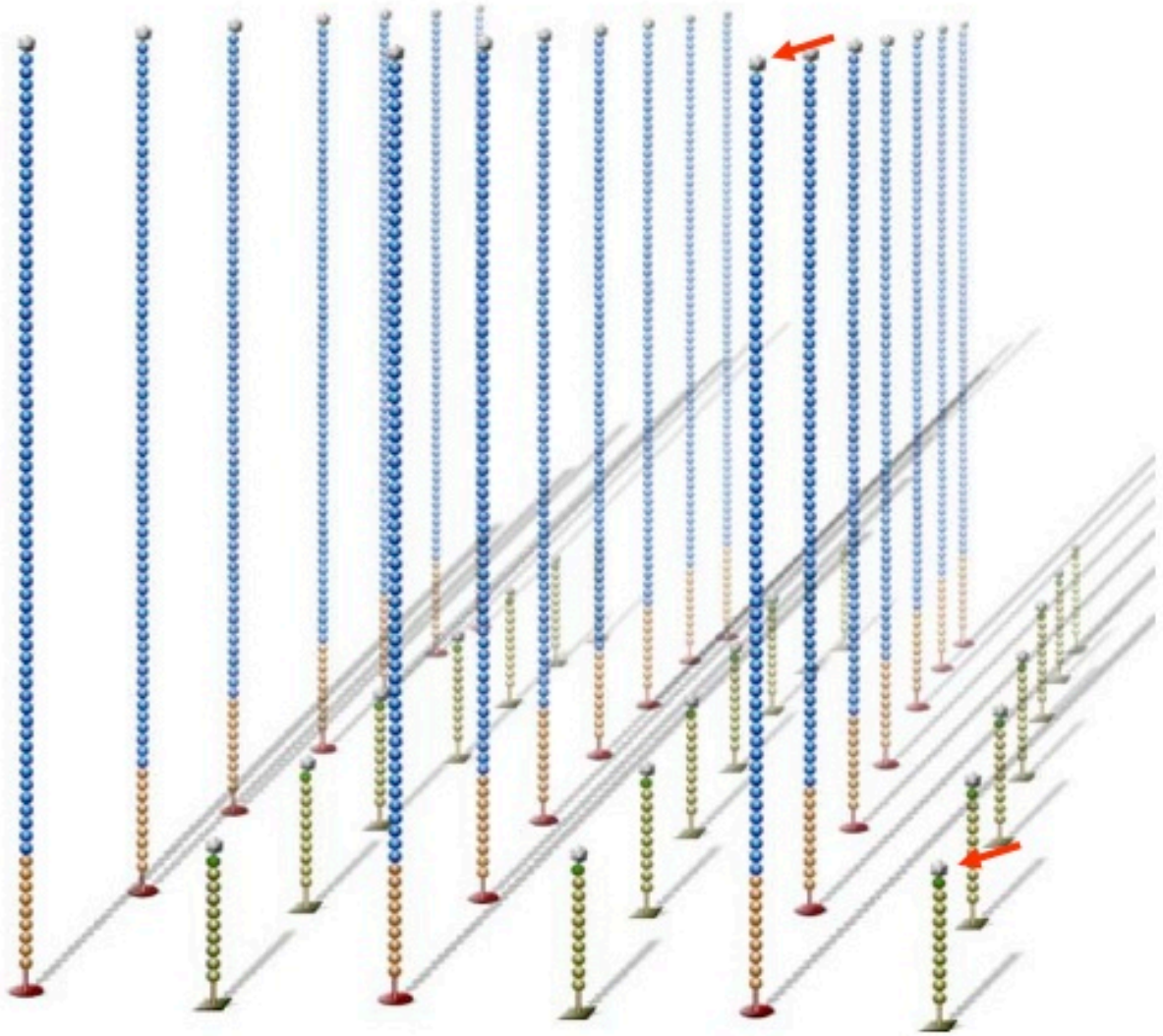
Reverse Strand Cleavage

Reverse strands cleaved and washed away, leaving a cluster with forward strands only



Blocking

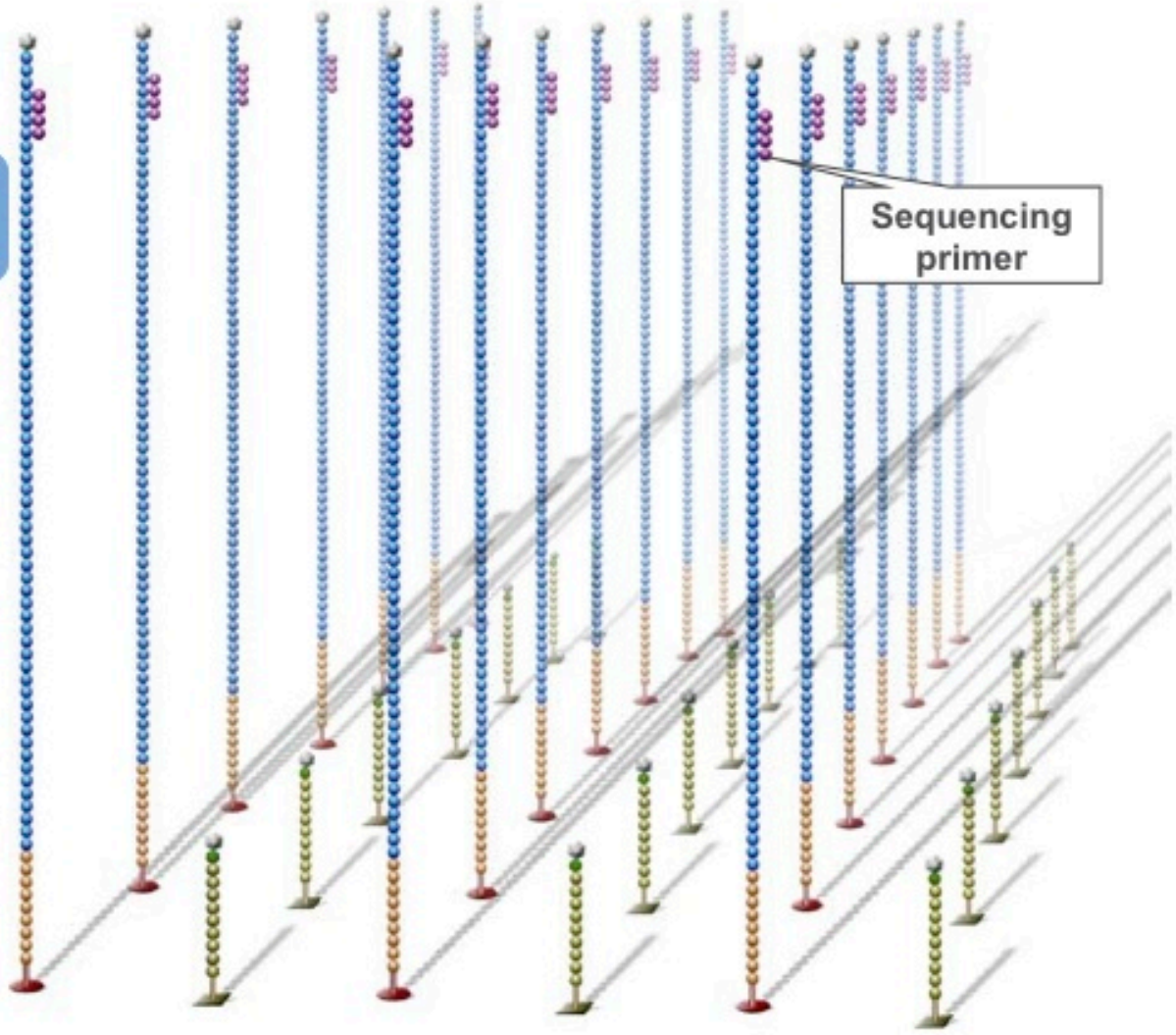
Free 3' ends are blocked to prevent unwanted DNA priming



illumina

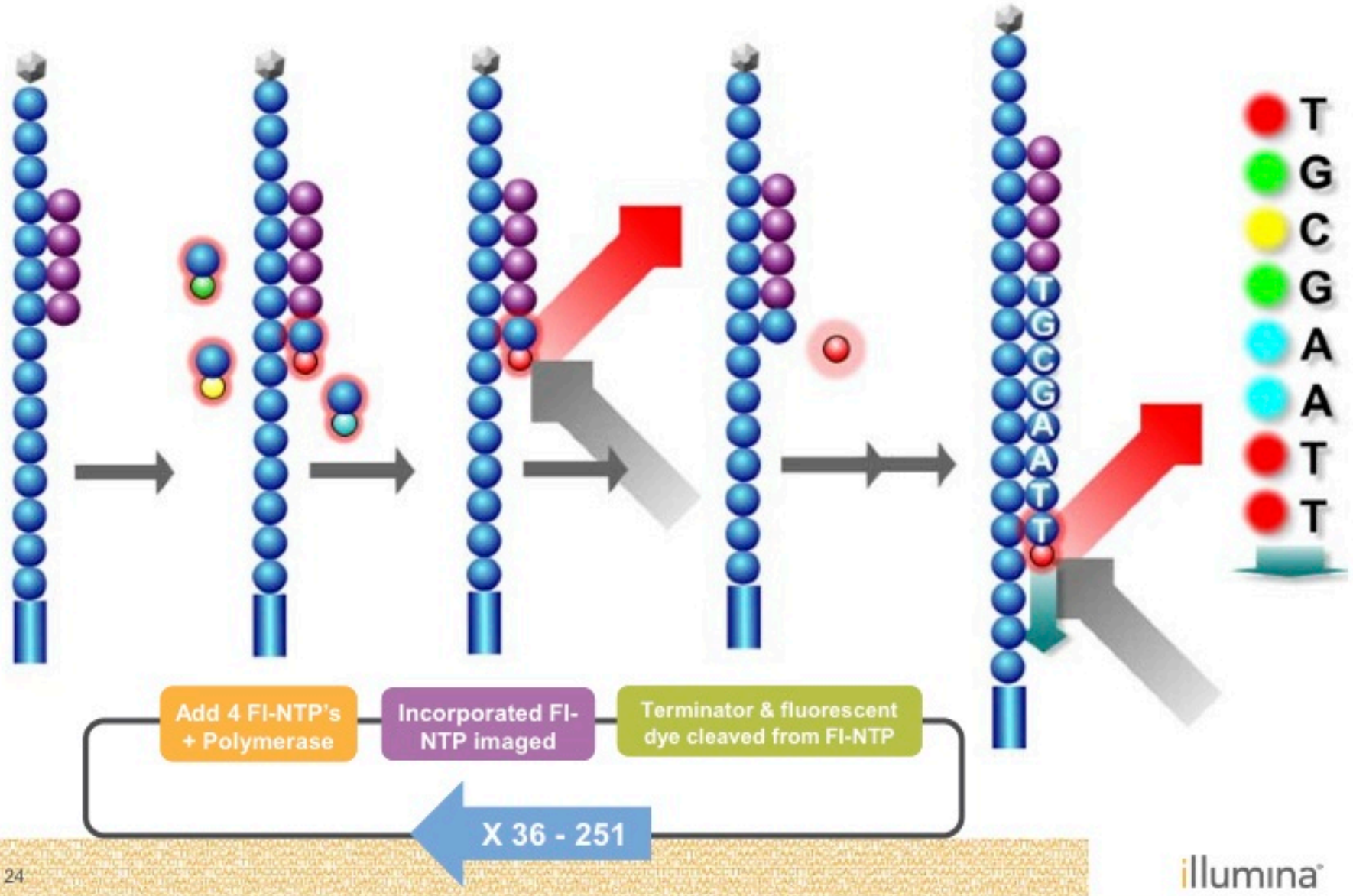
Read 1 Primer Hybridization

Sequencing primer is hybridized to adapter sequence

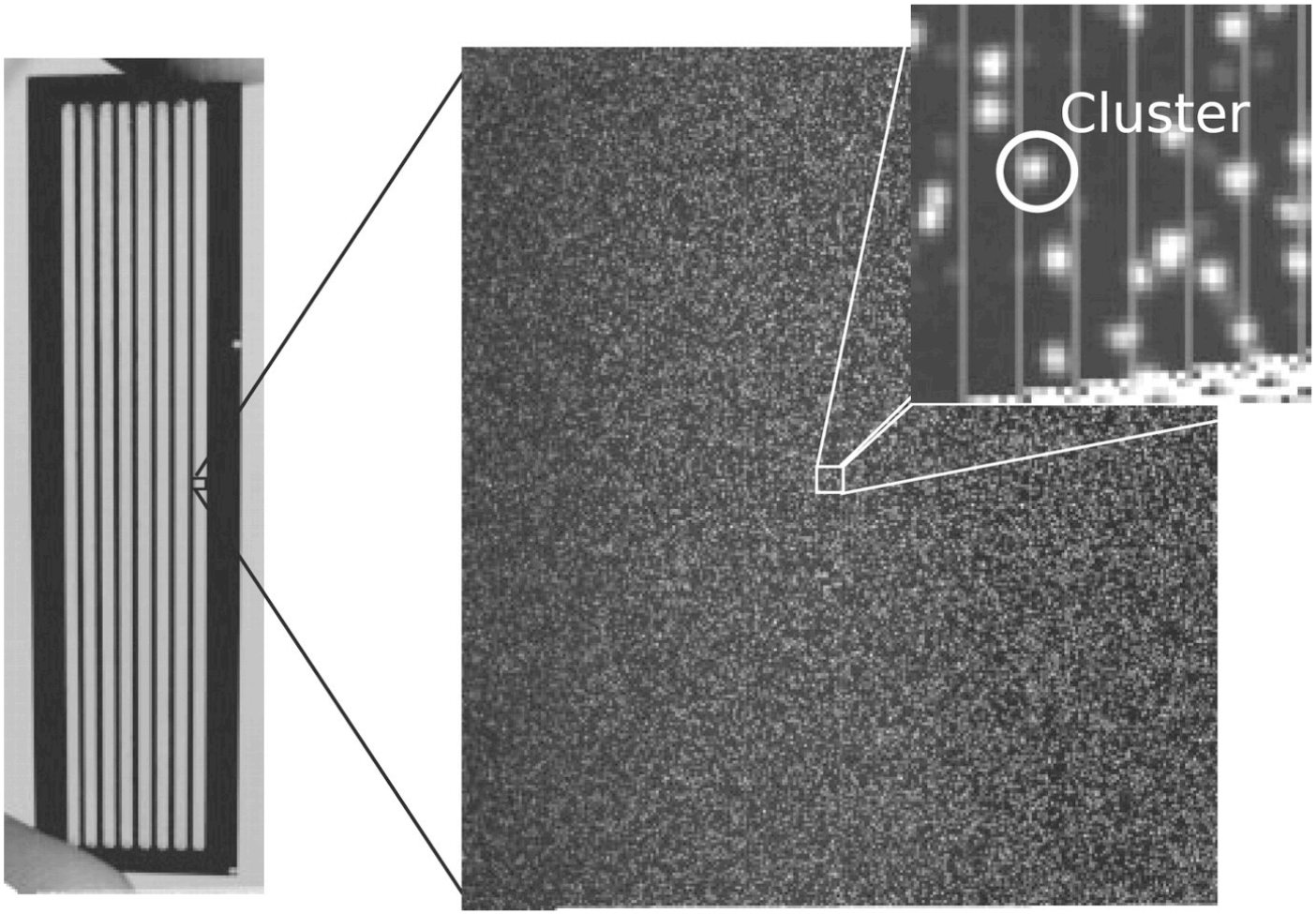


illumina

Sequencing by Synthesis

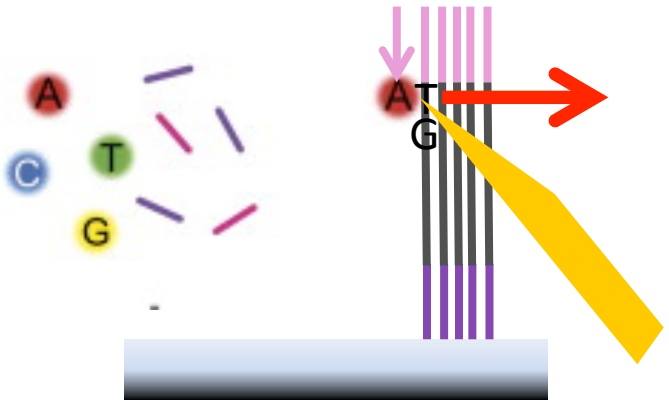


illumina



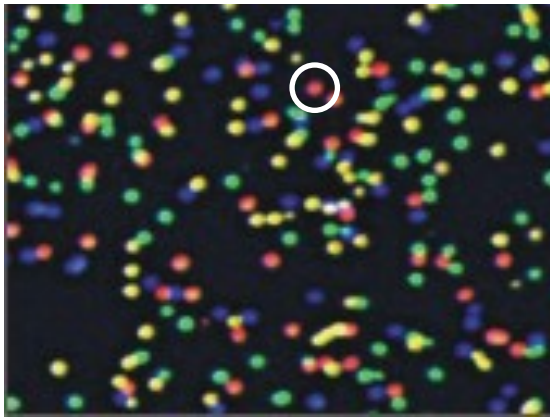
Tens of millions clusters per lane

Illumina Sequencing

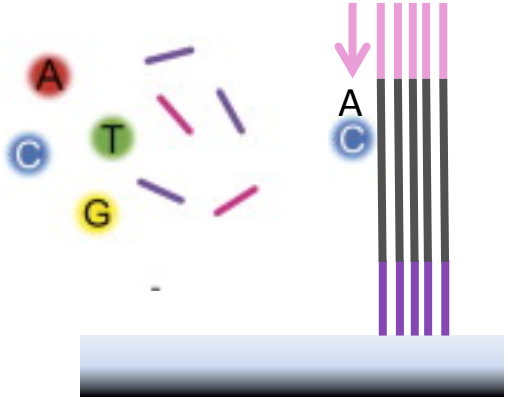


Primers and modified labelled nucleotides added

Capture

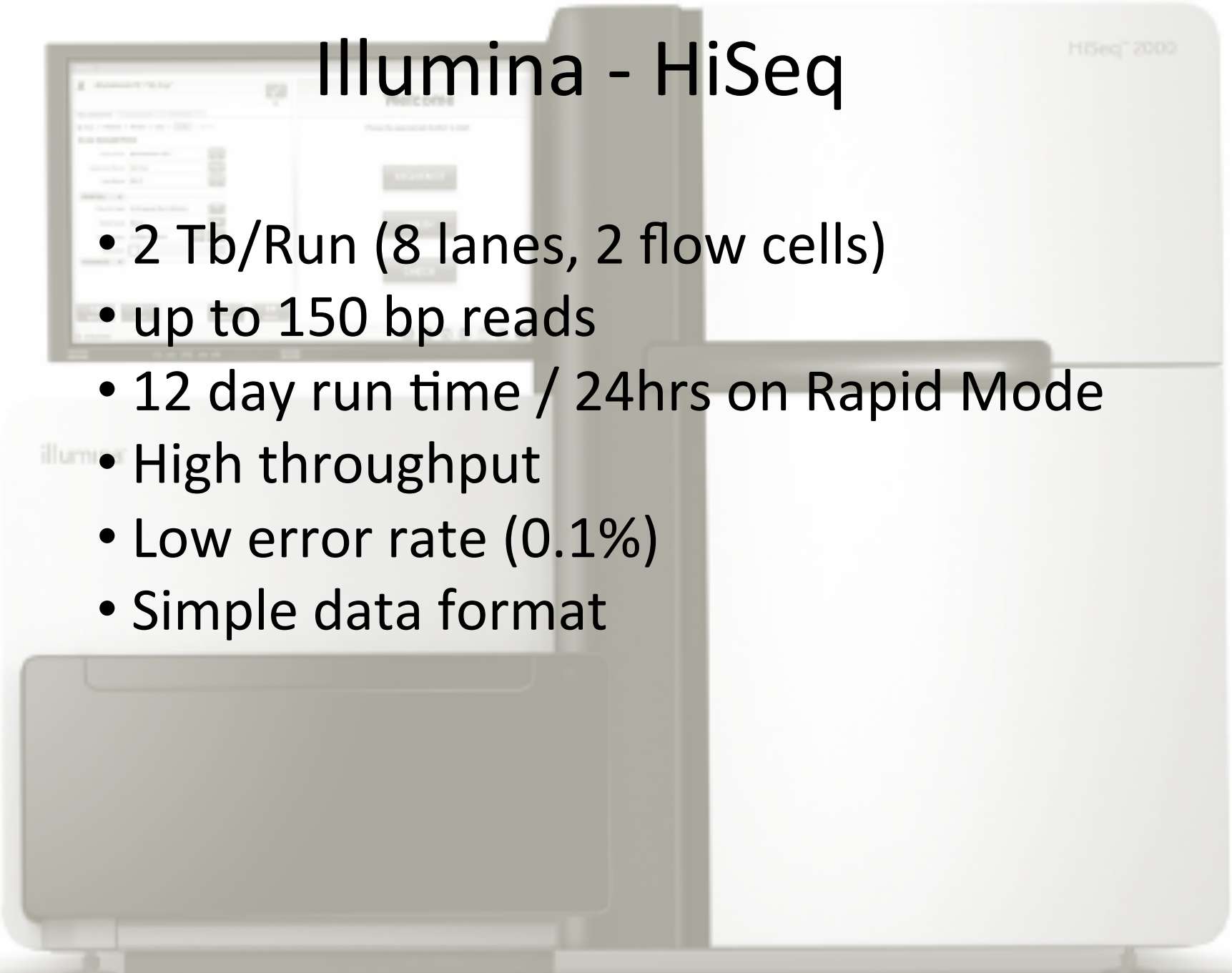


Cleave



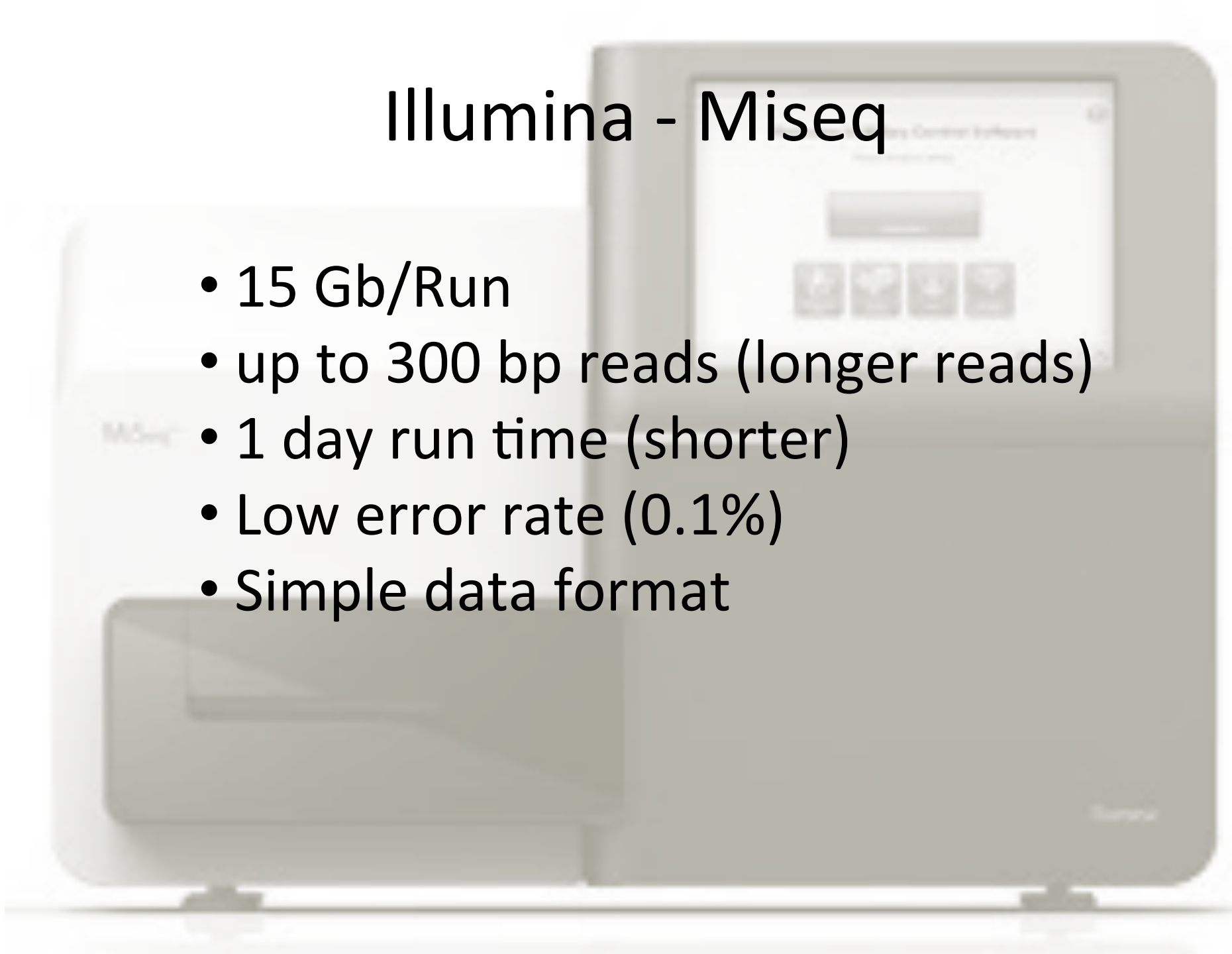
Repeat

Illumina - HiSeq



- 2 Tb/Run (8 lanes, 2 flow cells)
- up to 150 bp reads
- 12 day run time / 24hrs on Rapid Mode
- High throughput
- Low error rate (0.1%)
- Simple data format

Illumina - Miseq



- 15 Gb/Run
- up to 300 bp reads (longer reads)
- 1 day run time (shorter)
- Low error rate (0.1%)
- Simple data format

???

My supervisor gave me the task to find out how many samples (*Arabidopsis*, ~135Mb genome) he can run on one HiSeq lane in order to get 40x coverage for each genotype. He wants to use it for de novo assembly.

- 1 Hiseq lane produces 50Gbase / lane
- How many *Arabidopsis* samples can he pool??

135Mb x 40 = 5.4 Gb 50/5.4 = <u>9.25 samples/lane</u>

Pacific Biosciences (PacBio)

* Pyrosequencing in a zero mode waveguide (ZMW)

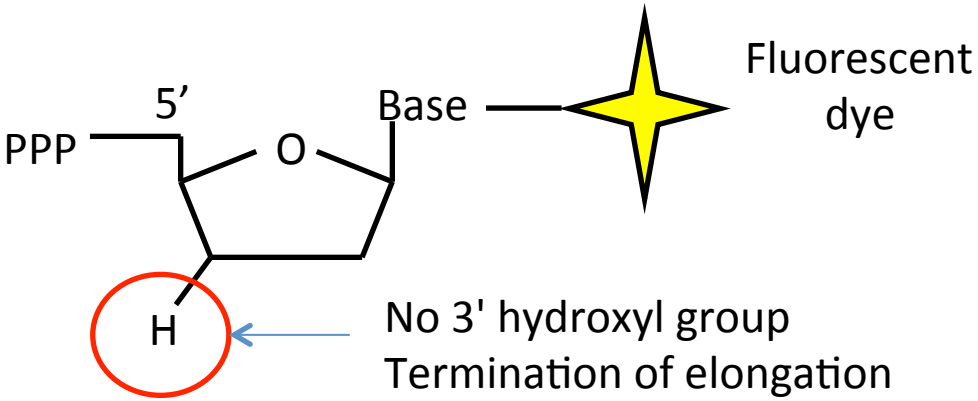
* Single molecule real time sequencing (SMRT)

* Accuracy ~85%

* Sequencing of long DNA molecules (20kb)

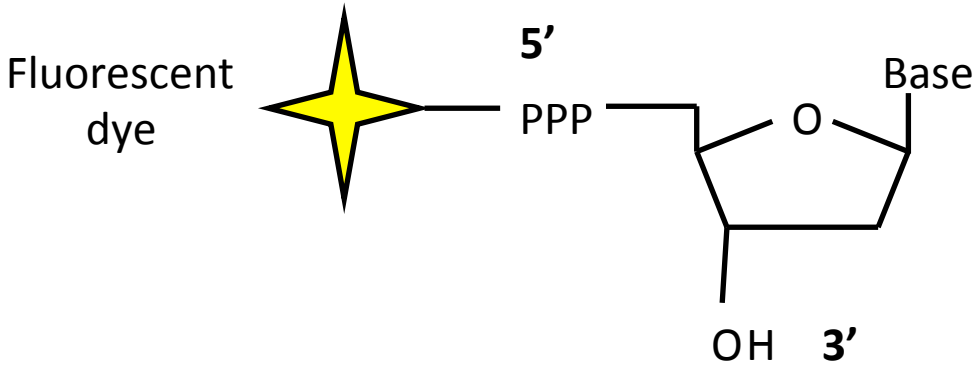
* Very suitable for sequencing repeats, cap filling and GC rich genomes

PacBio vs Sanger

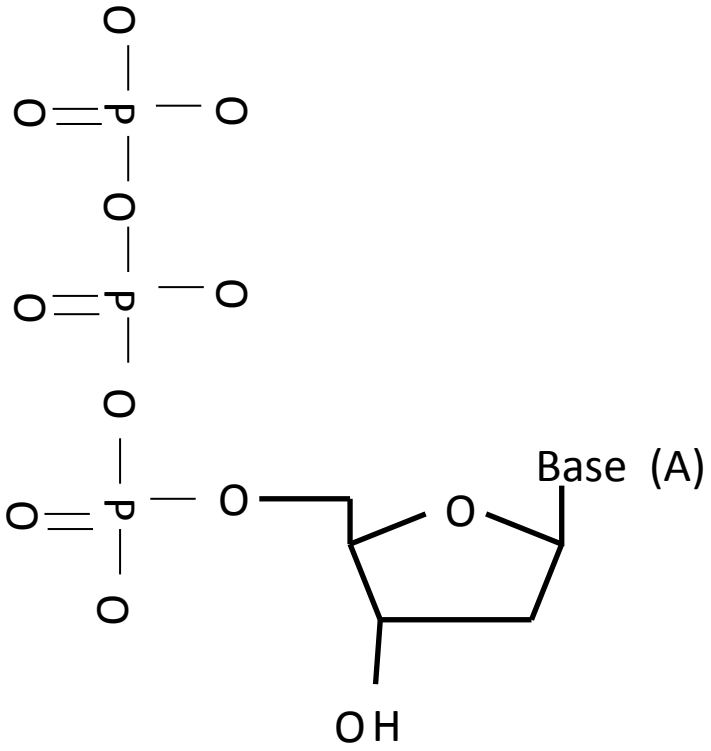


Sanger ddNTPs

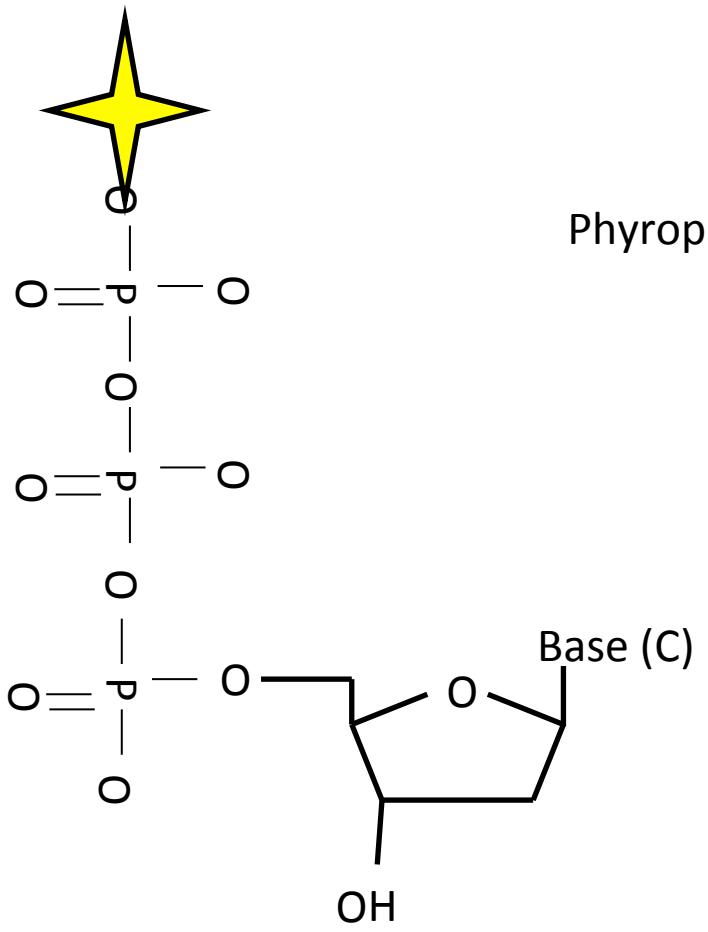
PacBio Phospho-Linked Nucleotides



During polymerase cleaves away fluorescent label, leaving natural DNA



hydrogen

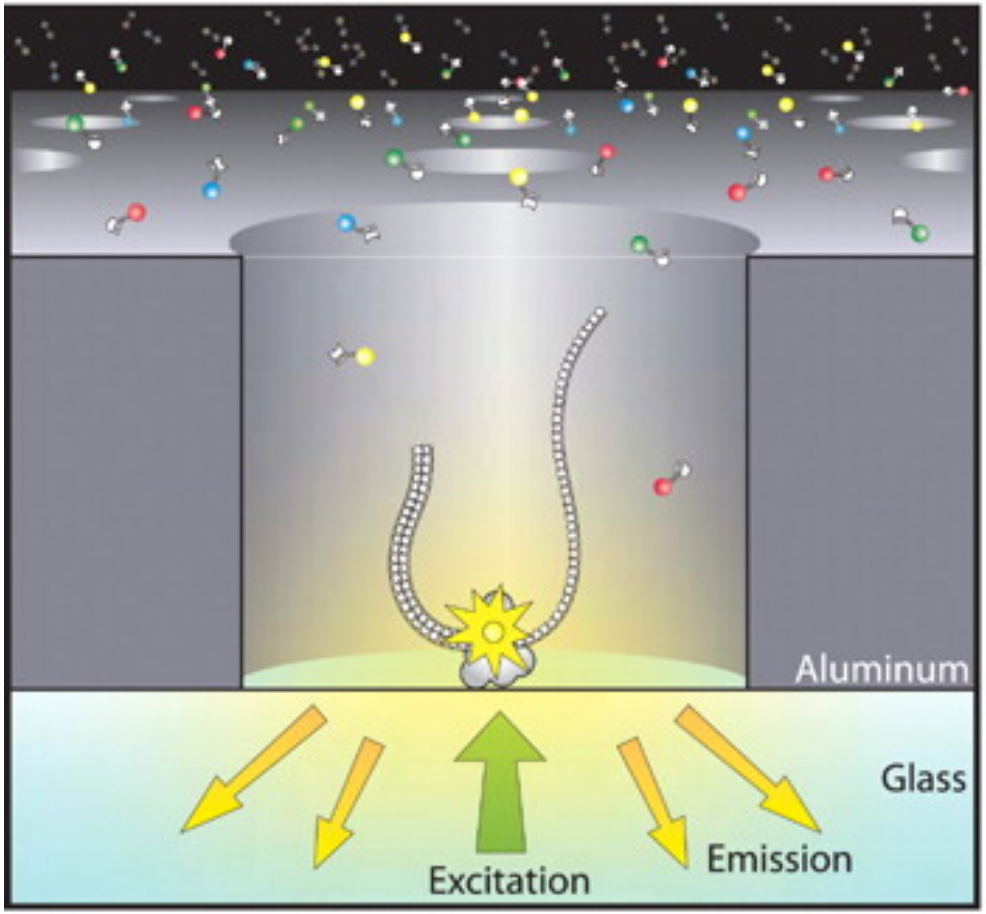


Phyrophosphate

Pacific Biosciences

Uses a zero mode waveguide (ZMW) to guide light energy to a single DNA polymerase with a single DNA template

100nm

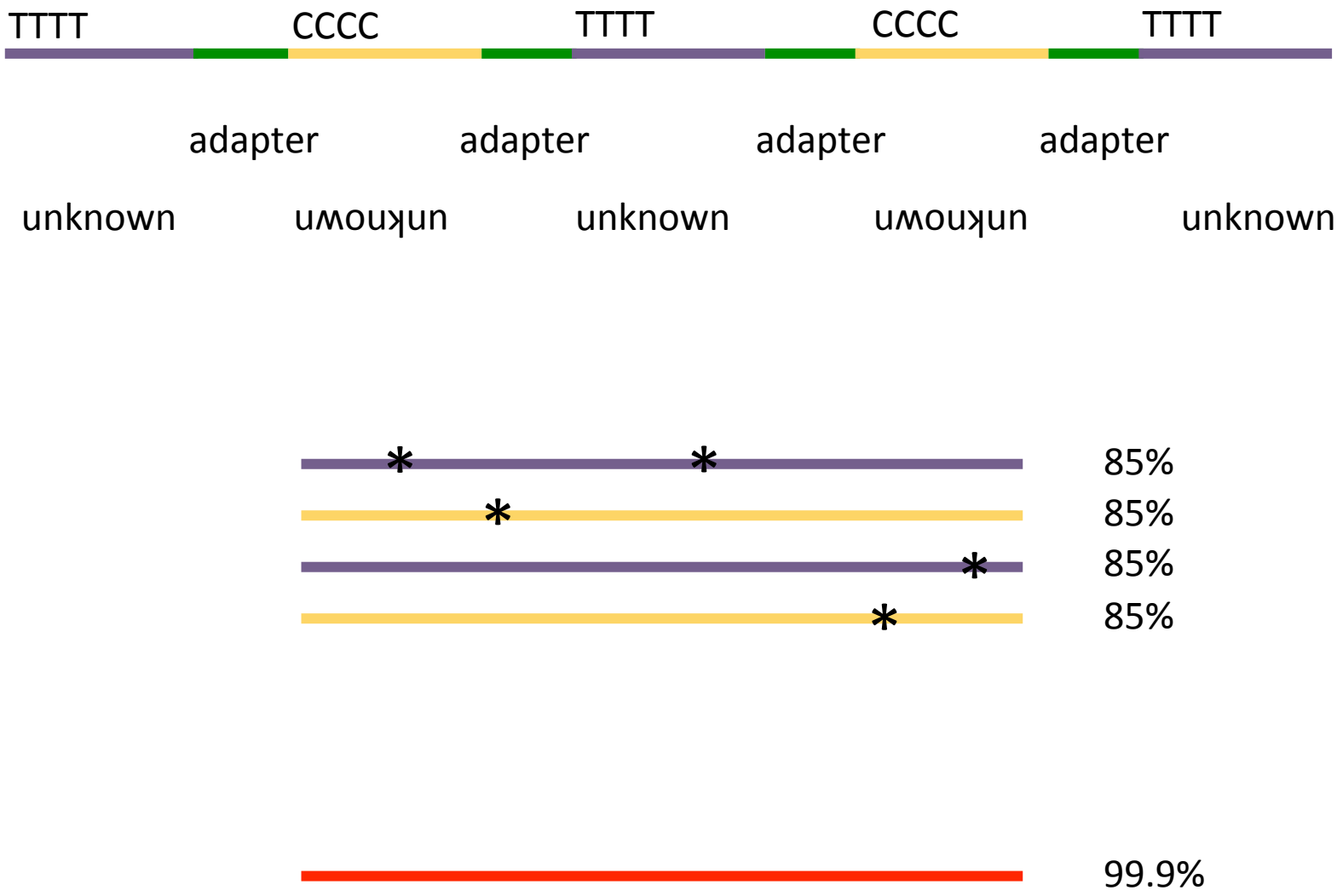


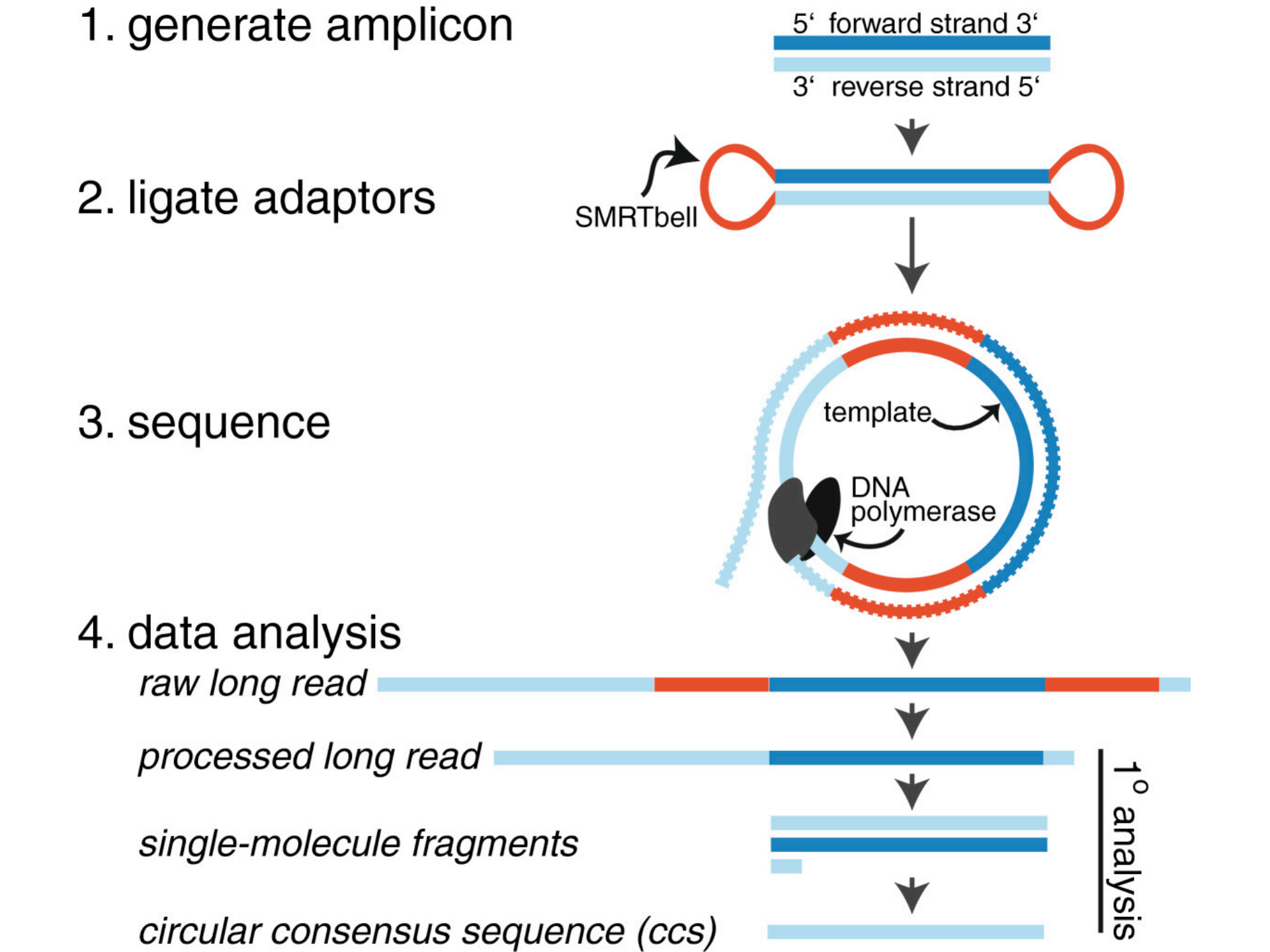
70nm hole in Aluminium


SMRT adapter
SMRT adapter

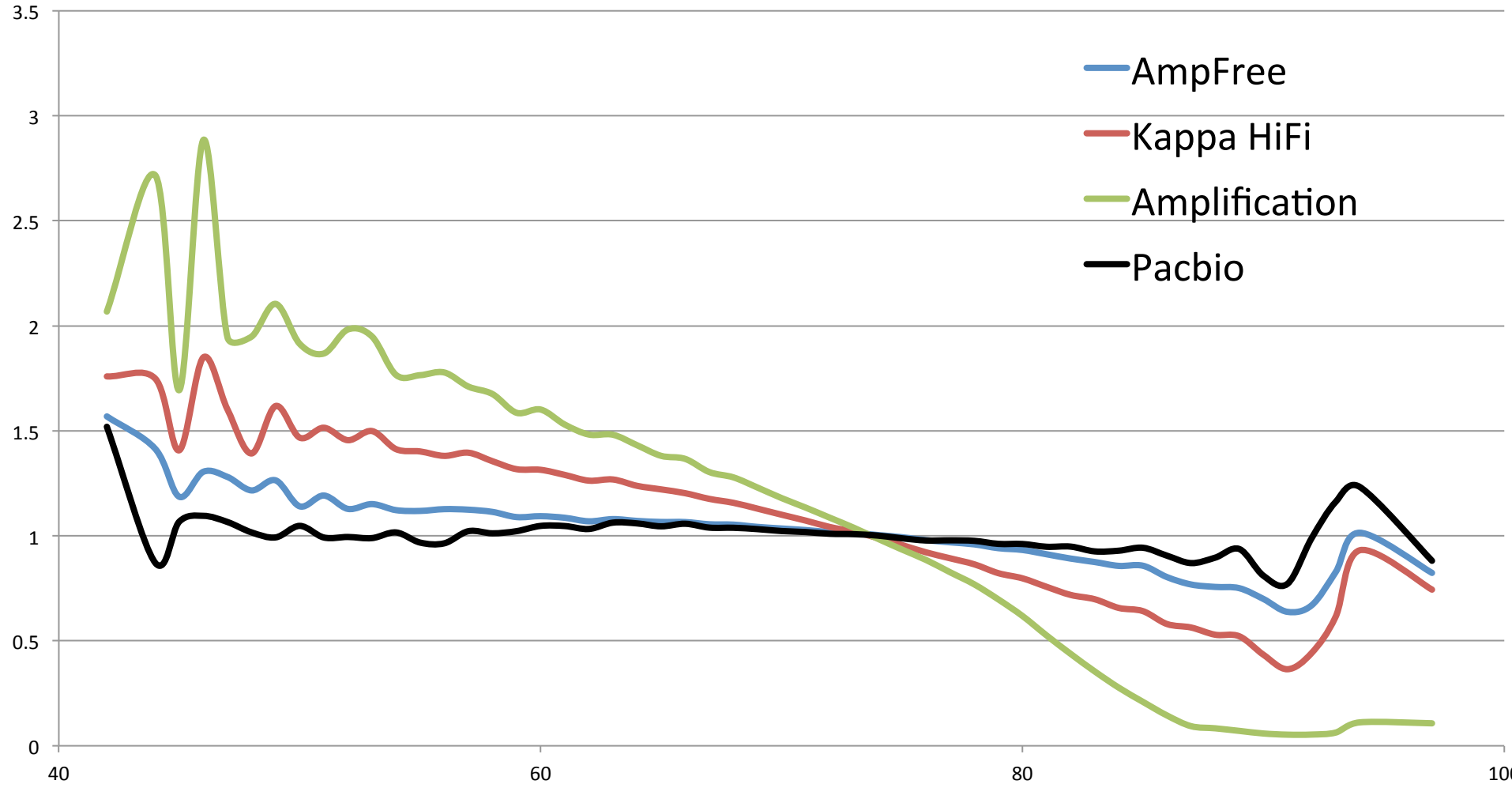

TTTT
GGGG
AAAA
CCCC


TTTT
CCCC
TTTT
CCCC
TTTT

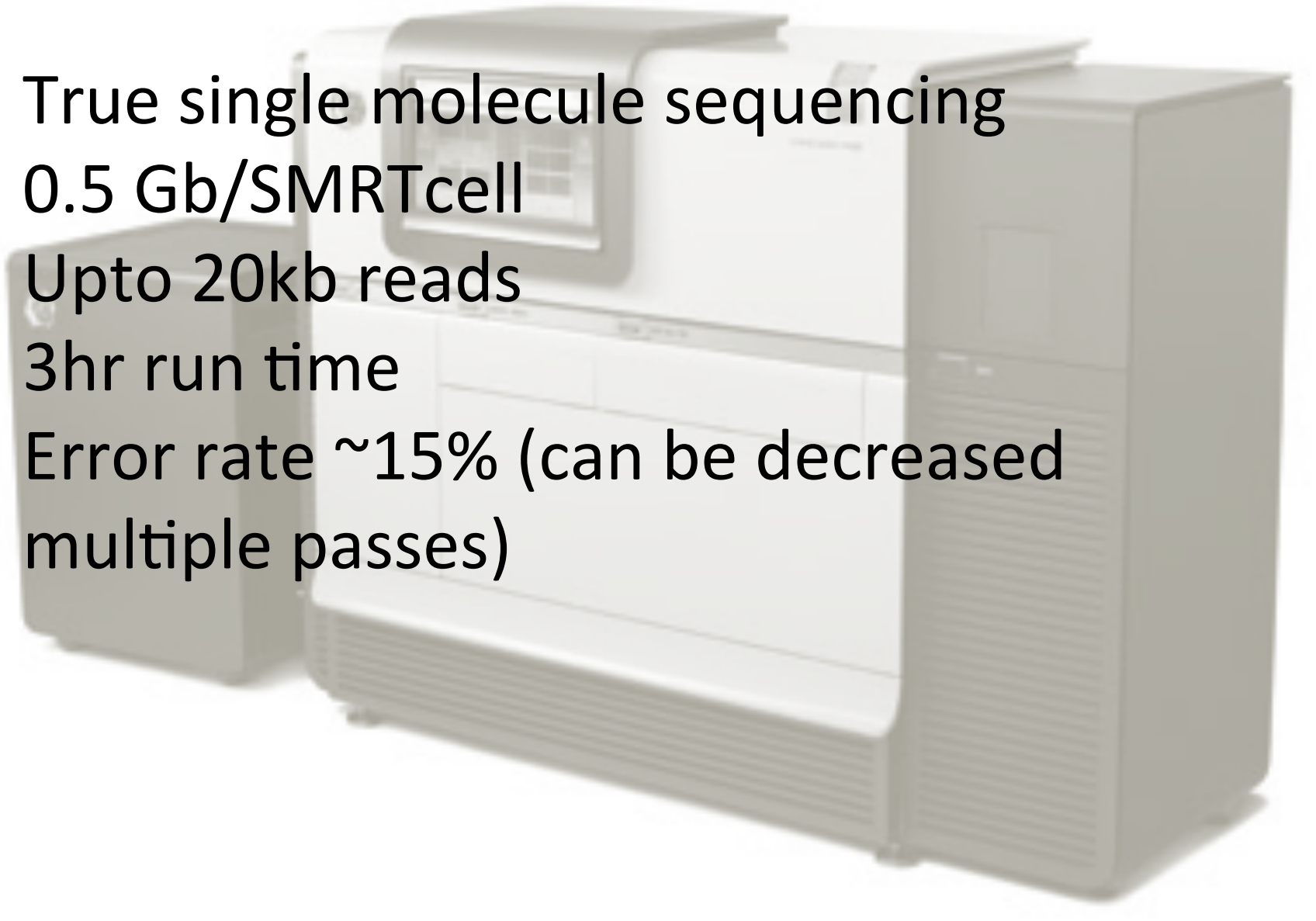

TTTT
CCCC
TTTT
CCCC
TTTT
adapter
adapter
adapter
adapter
unknown
unknown
unknown
unknown
unknown
85%
85%
85%
85%
99.9%


1. generate amplicon
5‘ forward strand 3‘
3‘ reverse strand 5‘
2. ligate adaptors
SMRTbell
3. sequence
template
DNA polymerase
4. data analysis
raw long read
processed long read
single-molecule fragments
circular consensus sequence (ccs)
1° analysis

GC bias – relative coverage



Pacific Biosciences



True single molecule sequencing
0.5 Gb/SMRTcell
Upto 20kb reads
3hr run time
Error rate ~15% (can be decreased multiple passes)

http://www.youtube.com/watch?v=_B_cUZ8hSYU

http://www.youtube.com/watch?v=v8p4ph2MAvI

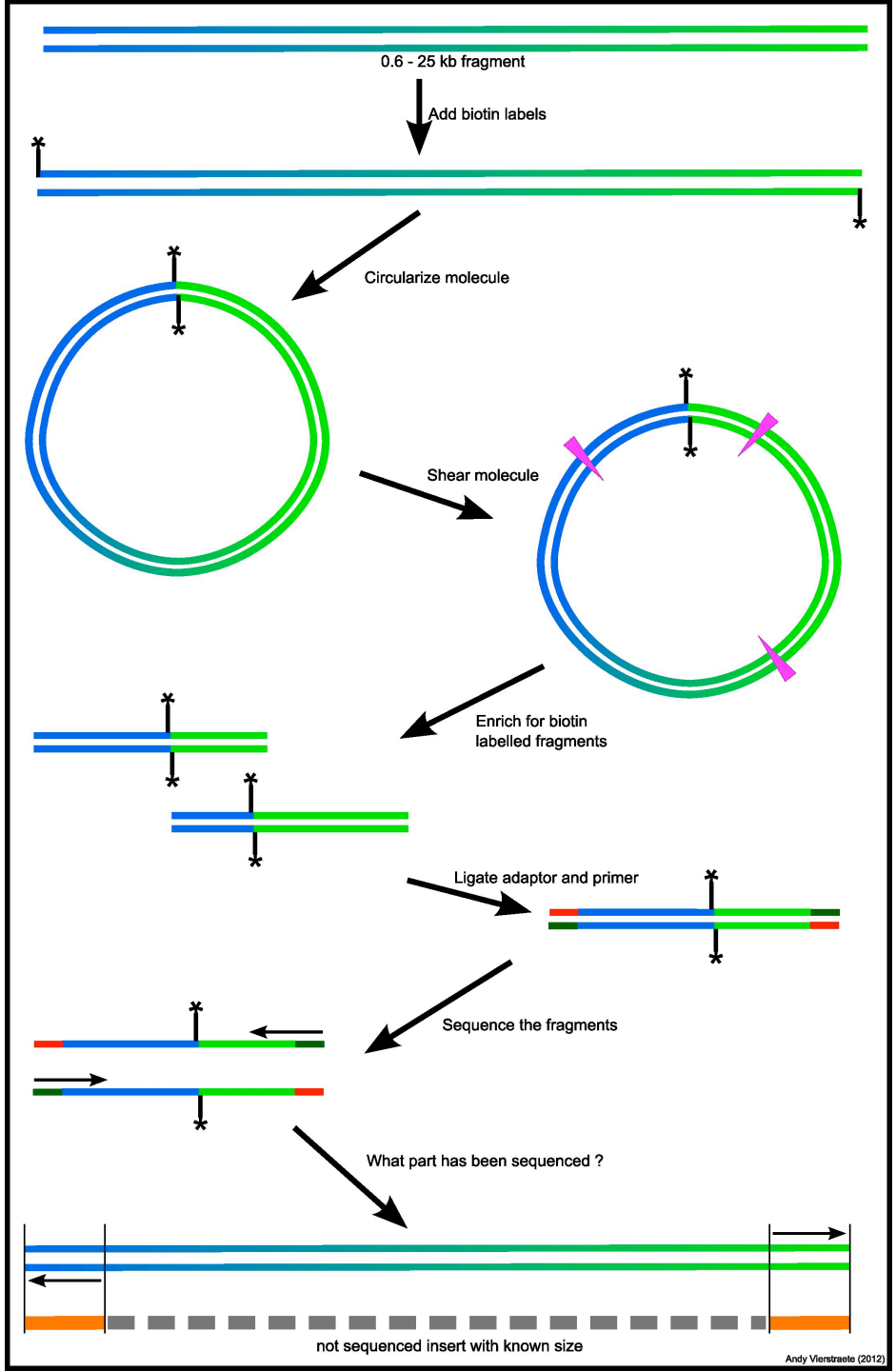
Summary

- Capillary sequencing
- NGS platforms
  - Illumina
  - Pacific Biosciences

Different types of libraries

Mate pair library

Mate pair library is very helpful in de novo sequencing; assembly genomes to sequence repeats.



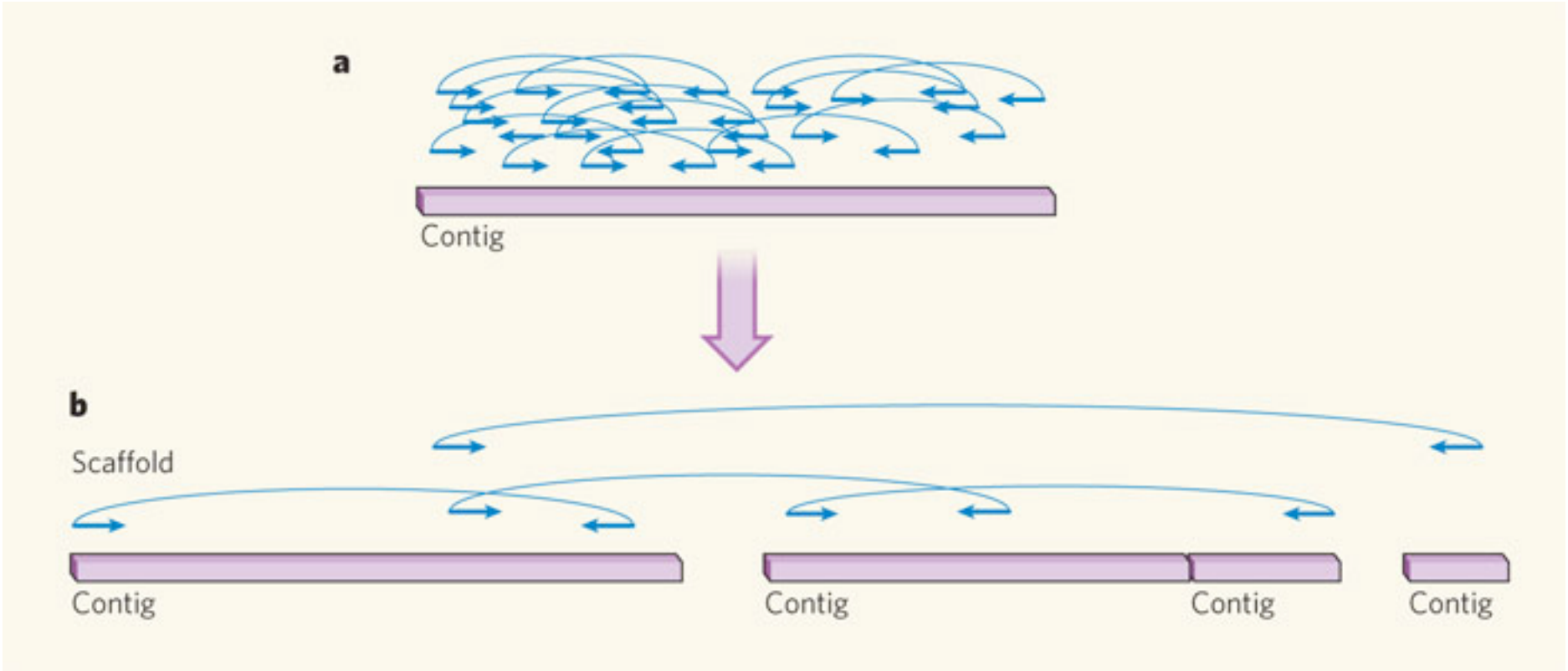
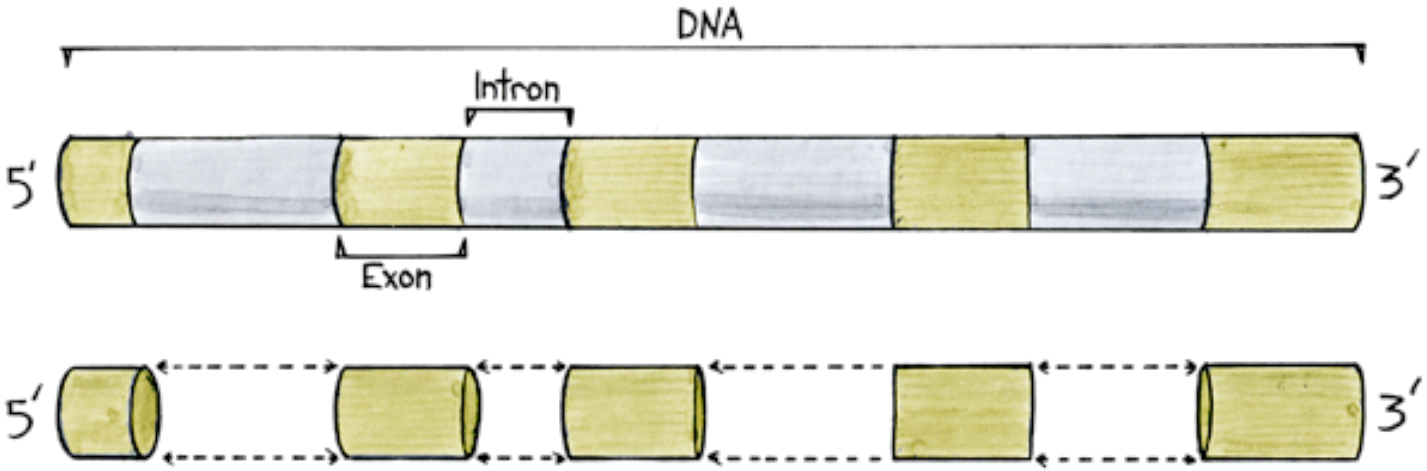
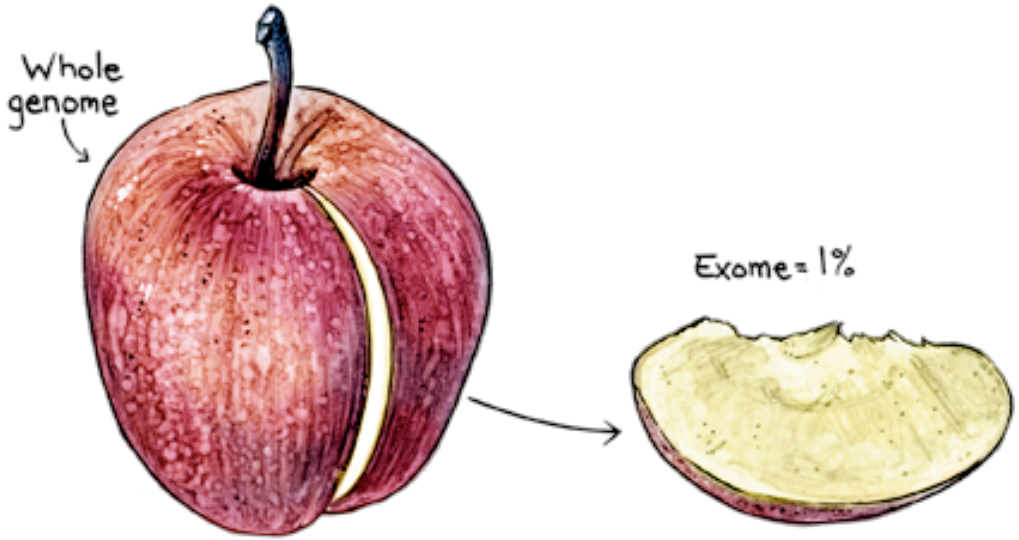

a
Contig
b
Scaffold
Contig
Contig
Contig
Contig

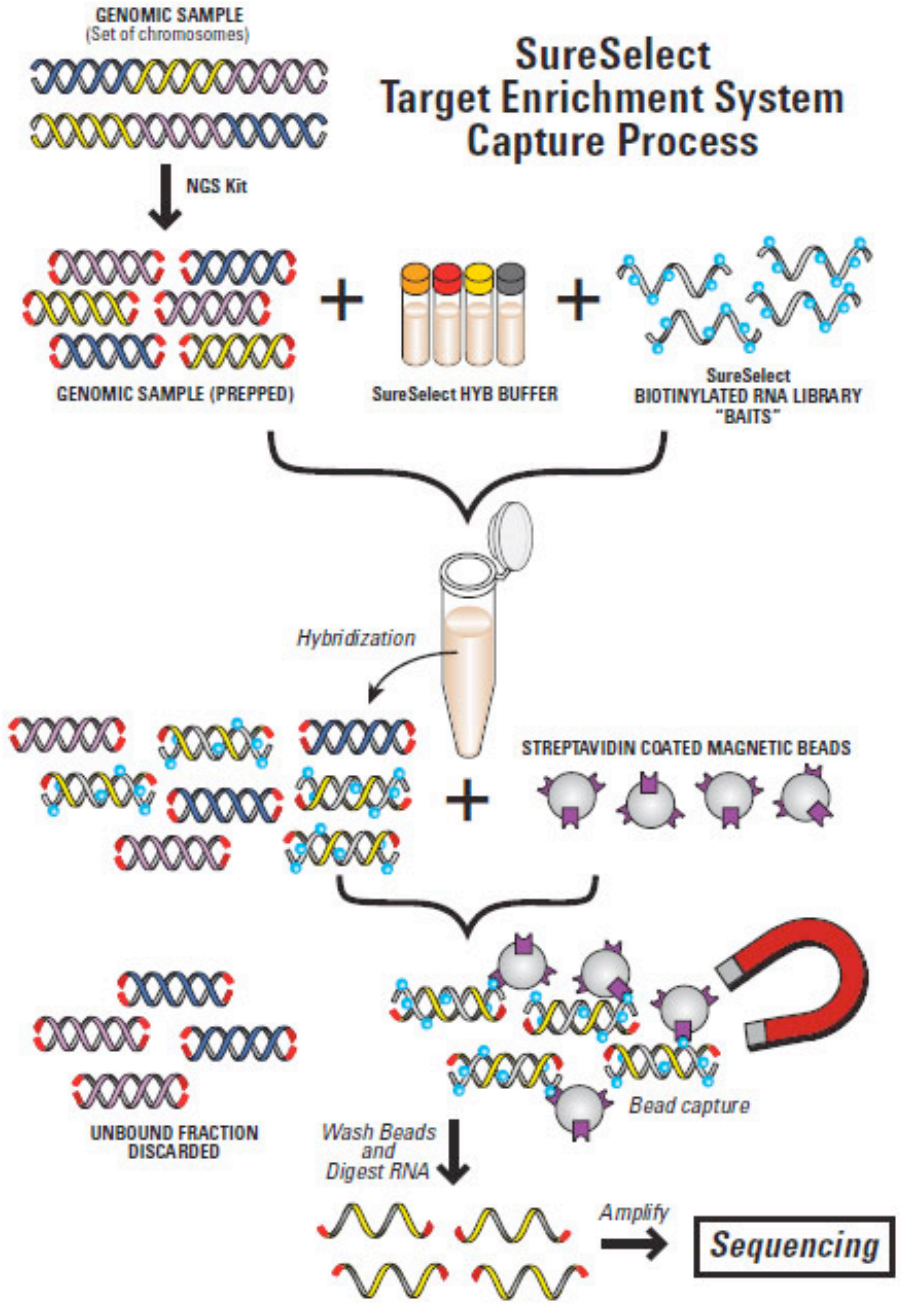
Table 2 Scaffolding value of different paired-read library combinations

No. of libraries*	Most efficient	Scaffold N50	Least efficient	Scaffold N50
1	15 kb	163,475	PE	37,694
2	5 kb + 25 kb	522,027	PE + 3 kb	46,699
2	5 kb + 20 kb	474,308	PE + 5 kb	141,403
2	8 kb + 25 kb	470,890	PE + 25 kb	142,007
3	5 kb + 20 kb + 25 kb	834,964	PE + 3 kb + 5 kb	158,525
3	5 kb + 15 kb + 25 kb	789,954	PE + 3 kb + 8 kb	171,253
3	8 kb + 20 kb + 25 kb	726,289	PE + 3 kb + 25 kb	198,696
7	ALL	1,287,609	N/A	N/A

Exome sequencing





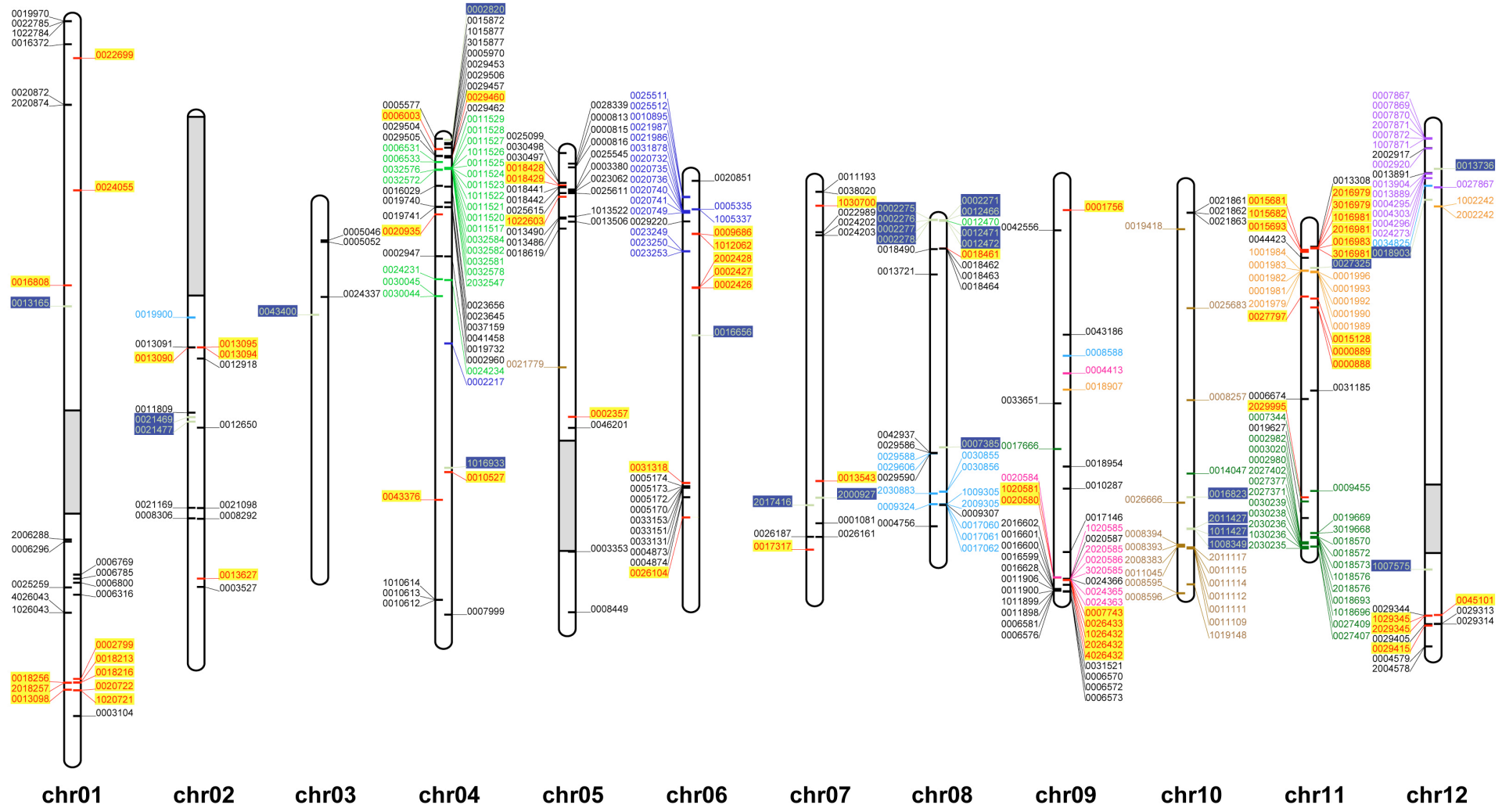


Only sequence data from regions of interest:

- Certain chromosomes
- Gene families
- Genes involved in cancer
- Etc…

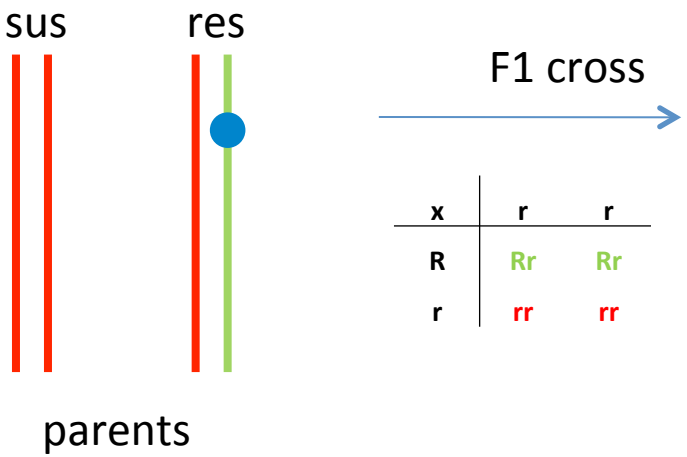
RenSeq (Resistance Enrichment Sequencing)

- In potato ~ 475 resistant gene present
- One or few R genes give resistance to Pi
- RenSeq = a method to quickly identify the R genes that gives resistance to *phytophthora infestans* (or any other pathogen)


chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12

resistant

susceptible


sus
res
parents
F1 cross
x r r
R Rr Rr
r rr rr
sus
res


PL521
PL535
PL544
PL501
PL542
PL491
PL488
PL566


PL959
PL906
PL941
PL955
PL930
PL936
PL942
PL925

BR

ccacgtgtagtcgtagcctacgt
tagtcgtagcctacgtgatgataa
tcgtagcctacgtgatgataaata
agtctgaccacgtgtagtcgtagcgtacgtgatgataaata
cgtgtagtcgtagcgtacg
acgtgtagtcgtagcgtac

50%

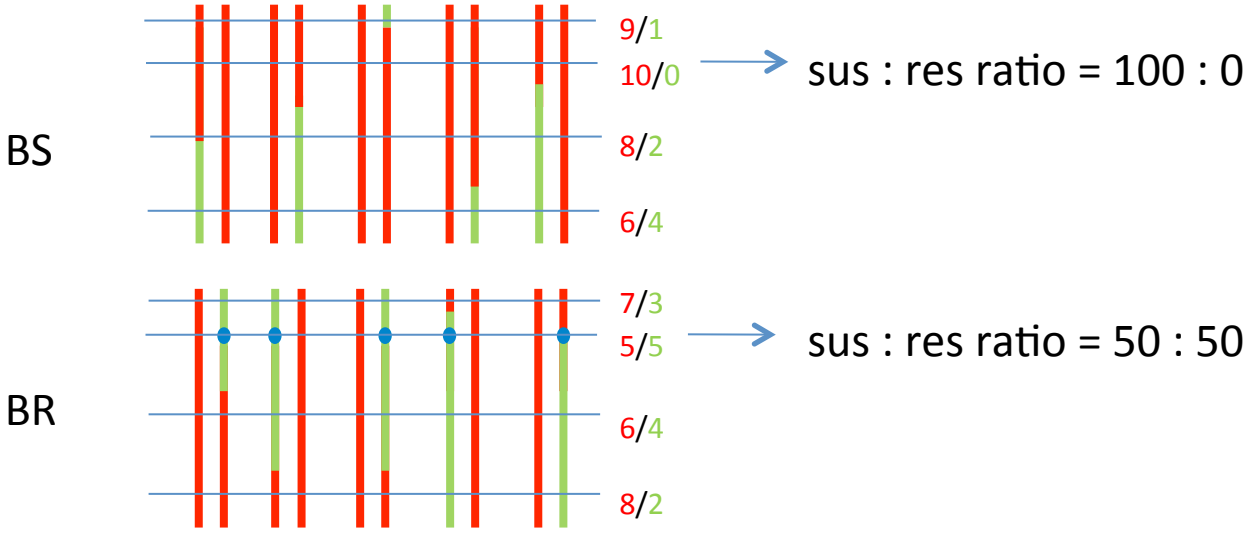
50%

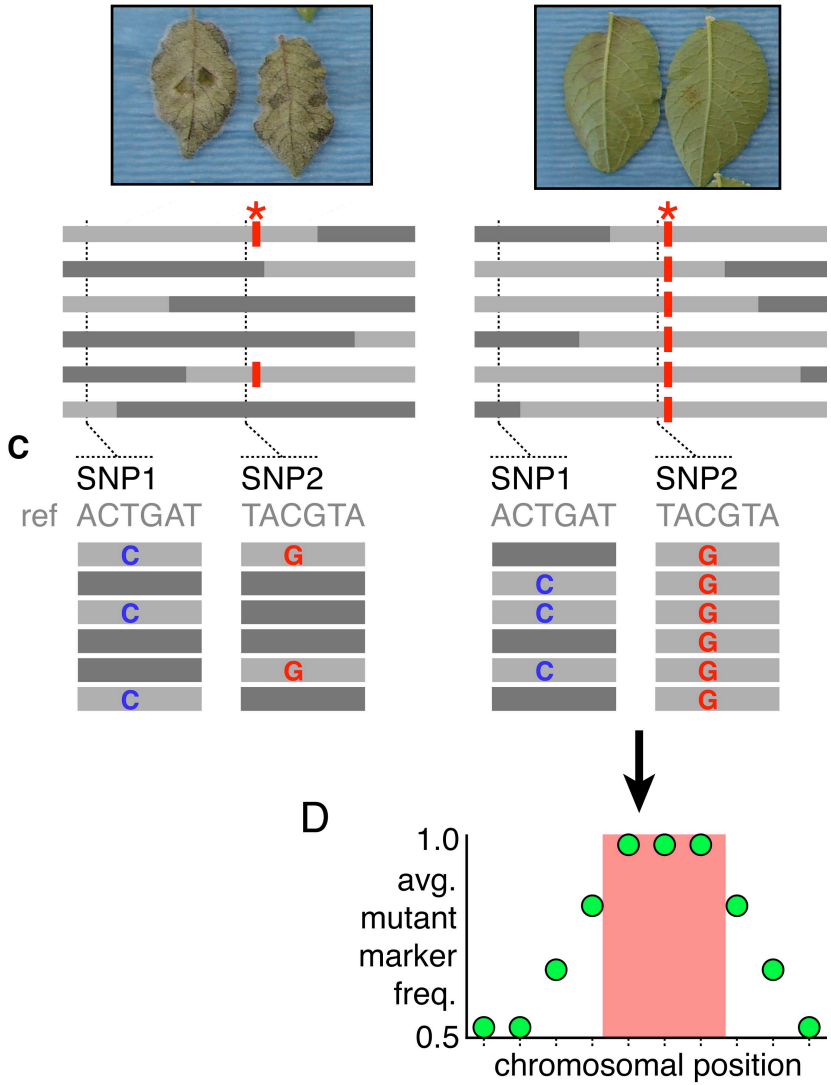
BS

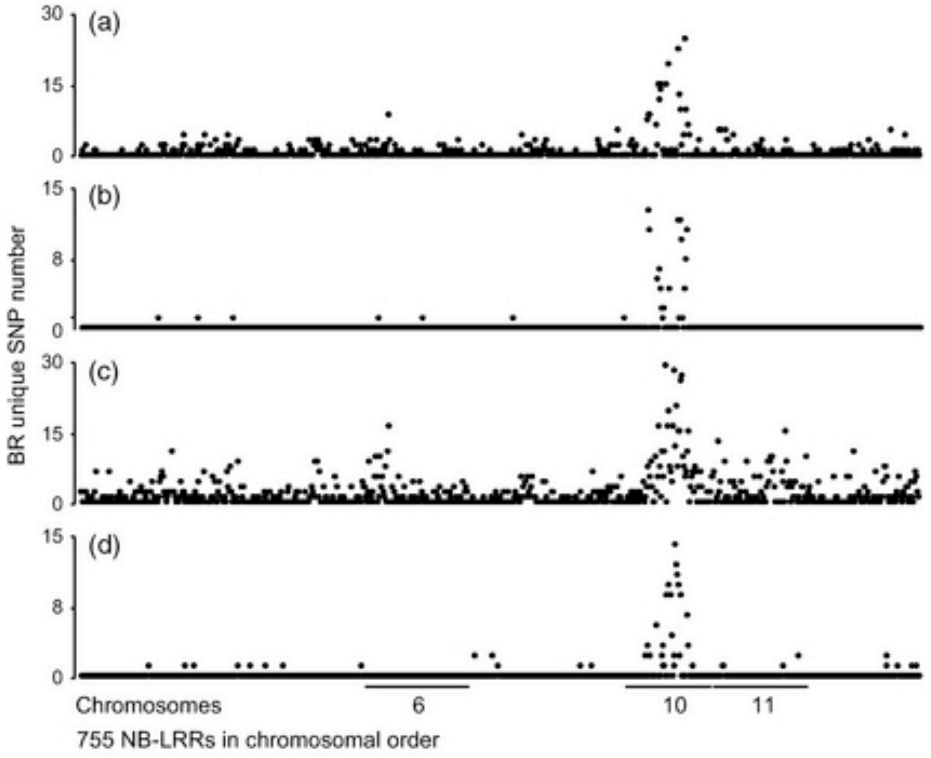
ccacgtgtagtcgtagcgtacgt
tagtcgtagcgtacgtgatgataa
tcgtagcgtacgtgatgataaata
agtctgaccacgtgtagtcg agcgtacgtgatgataaata
cgtgtagtcgtagcgtacg
acgtgtagtcgtagcgtac

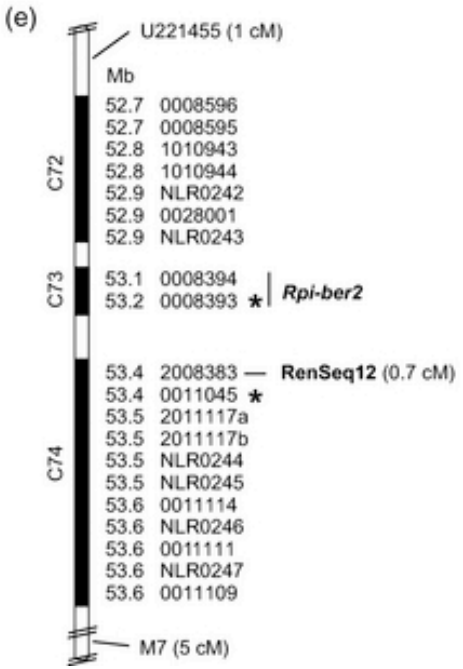
100%

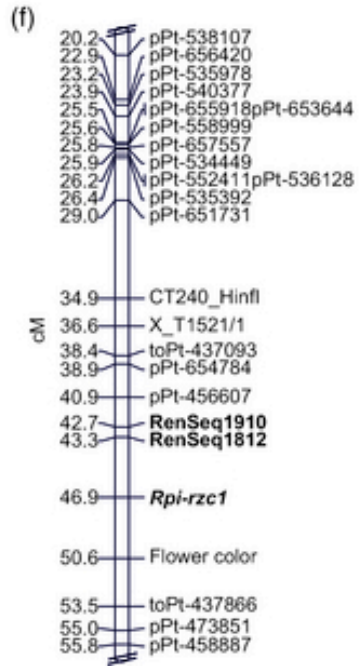
recombination




C
SNP1
SNP2
ref ACTGAT
TACGTA
SNP1
ACTGAT
SNP2
TACGTA
D
1.0
avg.
mutant
marker
freq.
0.5
chromosomal position





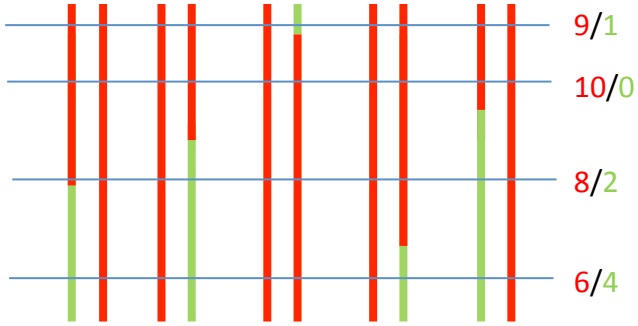


(a + c) = quick mapping; map reads to DM R genes

(b + d) = gene specific mapping; reads from BS were used for de novo assembly

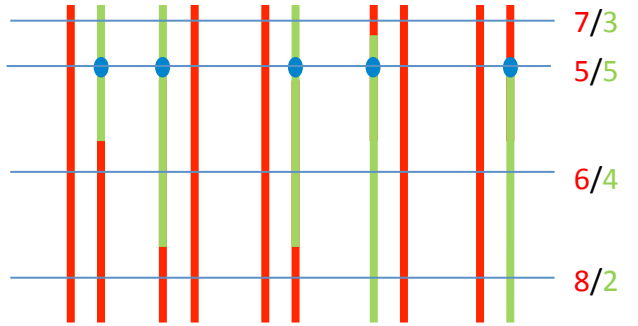
recombination

sus

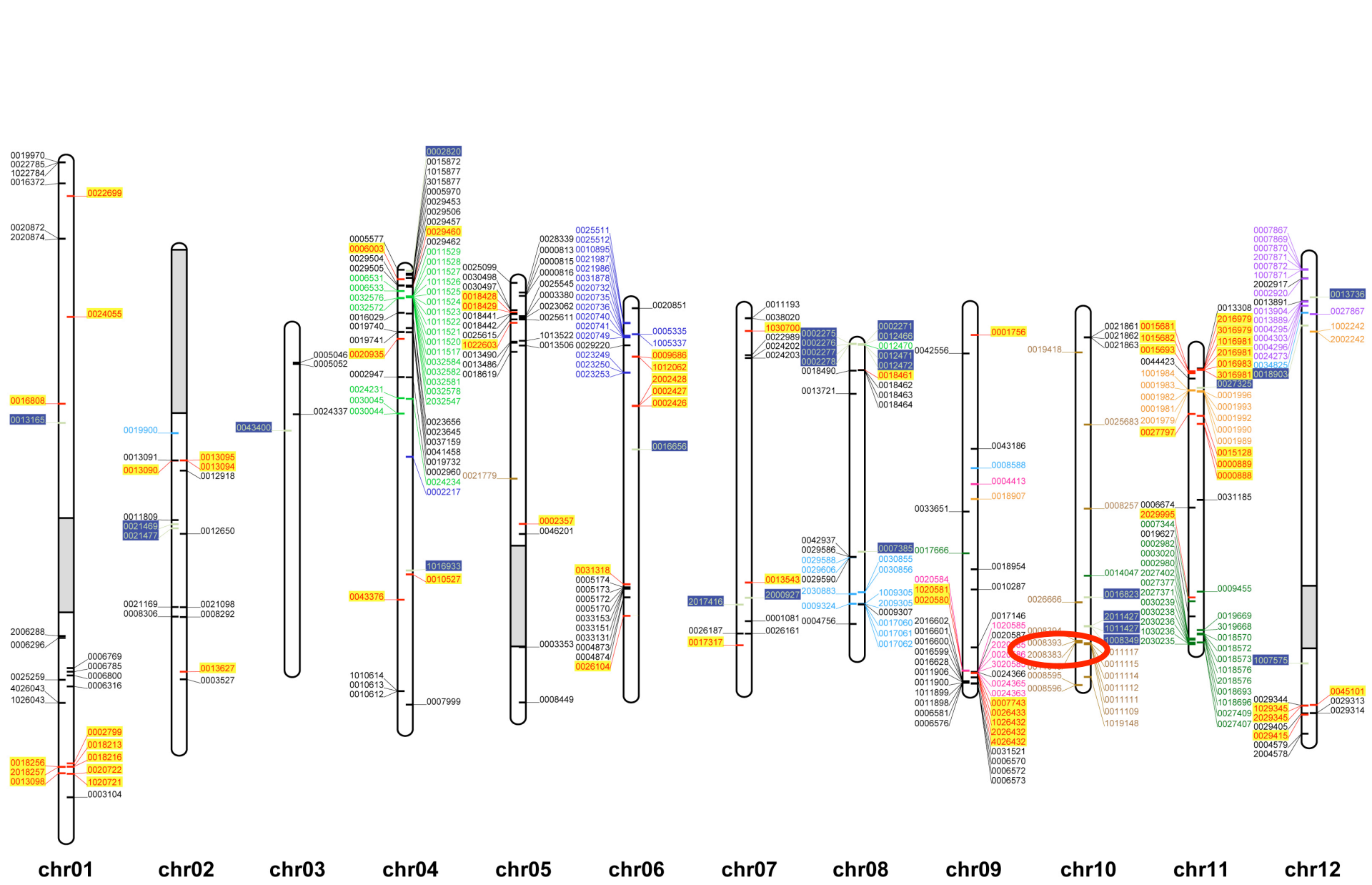


sus : res ratio = 100 : 0

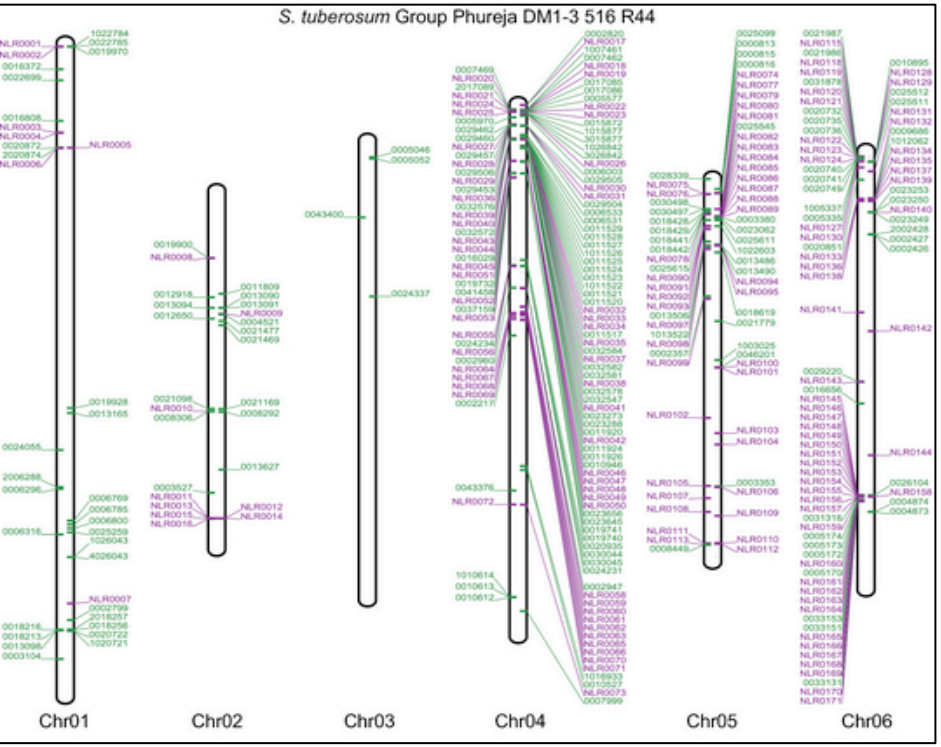
res

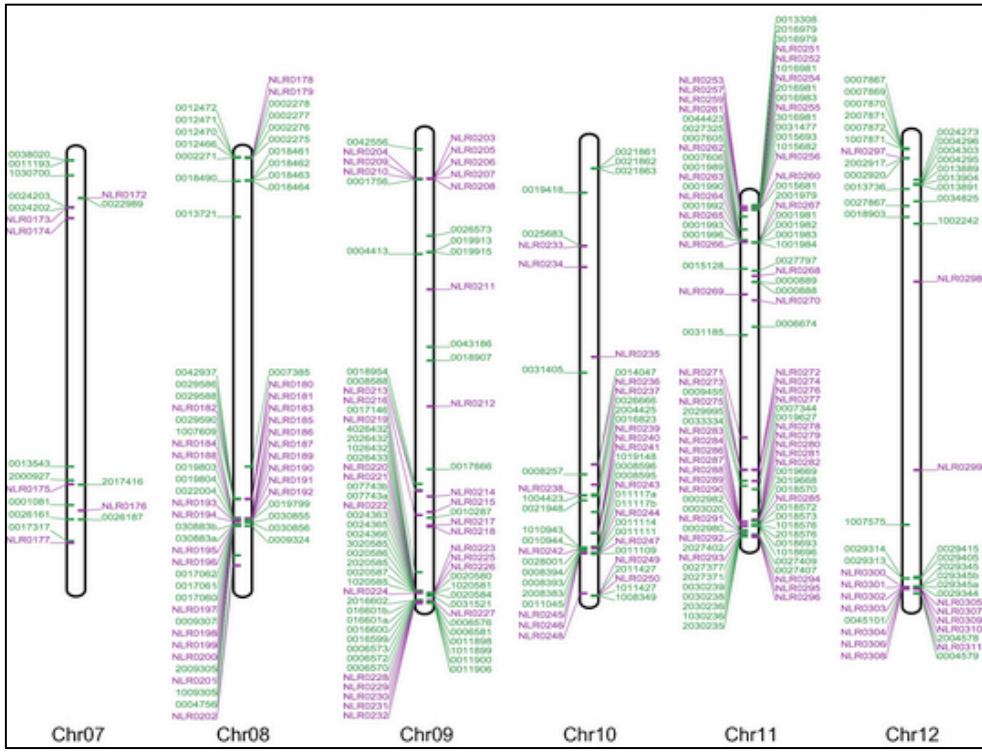


sus : res ratio = 50 : 50


chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12

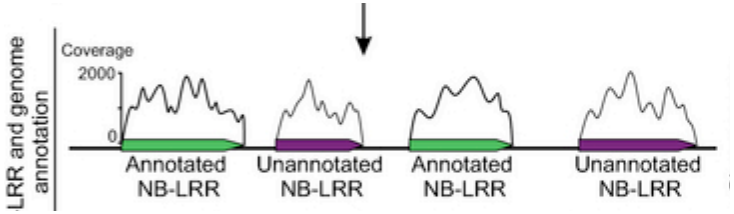
Also for discovery of un-known R genes





Green = known R genes

Purple = newly identified R genes



Questions?