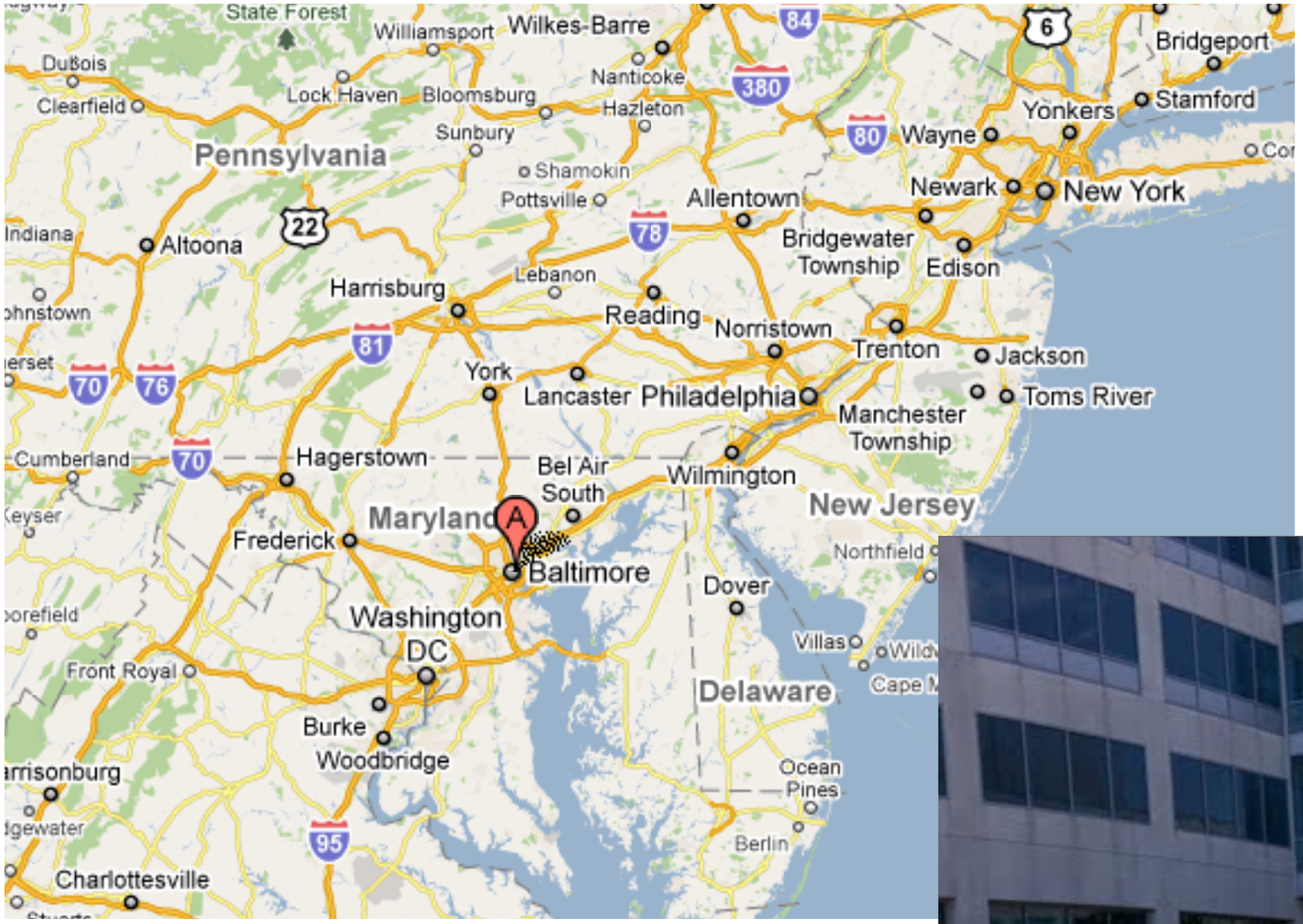
Basics of HT sequencing technologies

Kasper Daniel Hansen <khansen@jhsph.edu>
Postdoc w/Rafael Irizarry
Johns Hopkins Bloomberg School of Public Health

Brixen, 26 June 2011

Many slides are courtesy of
Hector Corrada Bravo and Ben Langmead

Baltimore / Johns Hopkins Bloomberg SPH



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

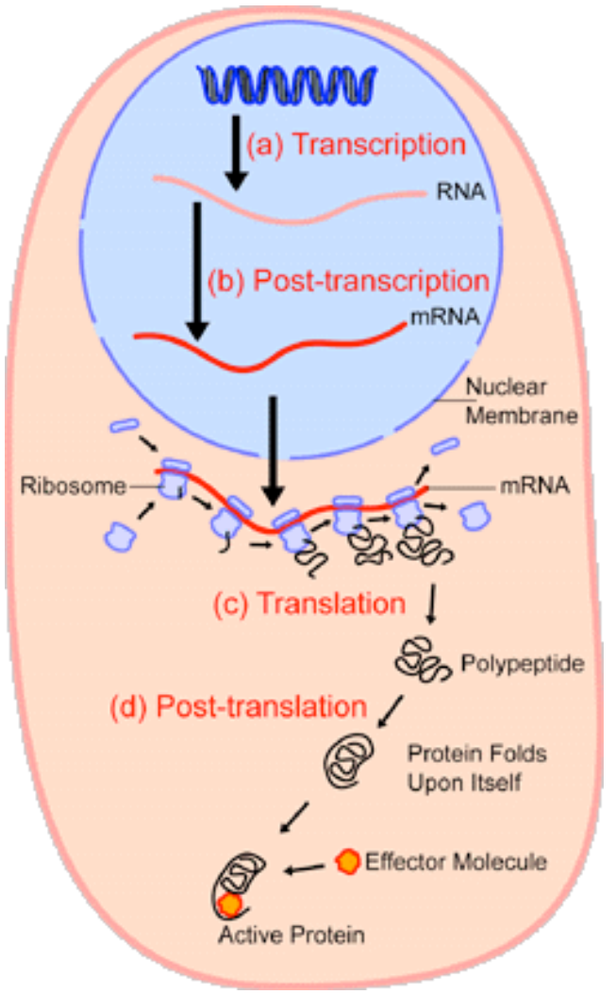
Microscope

Source:
http://blog.savcds.org/swanson/files/2009/10/Epidermal-Cell-onion.jpg

Sequencing

Illumina HiSeq 2000

Source: http://www.illumina.com/systems/hiseq_2000.ilmn

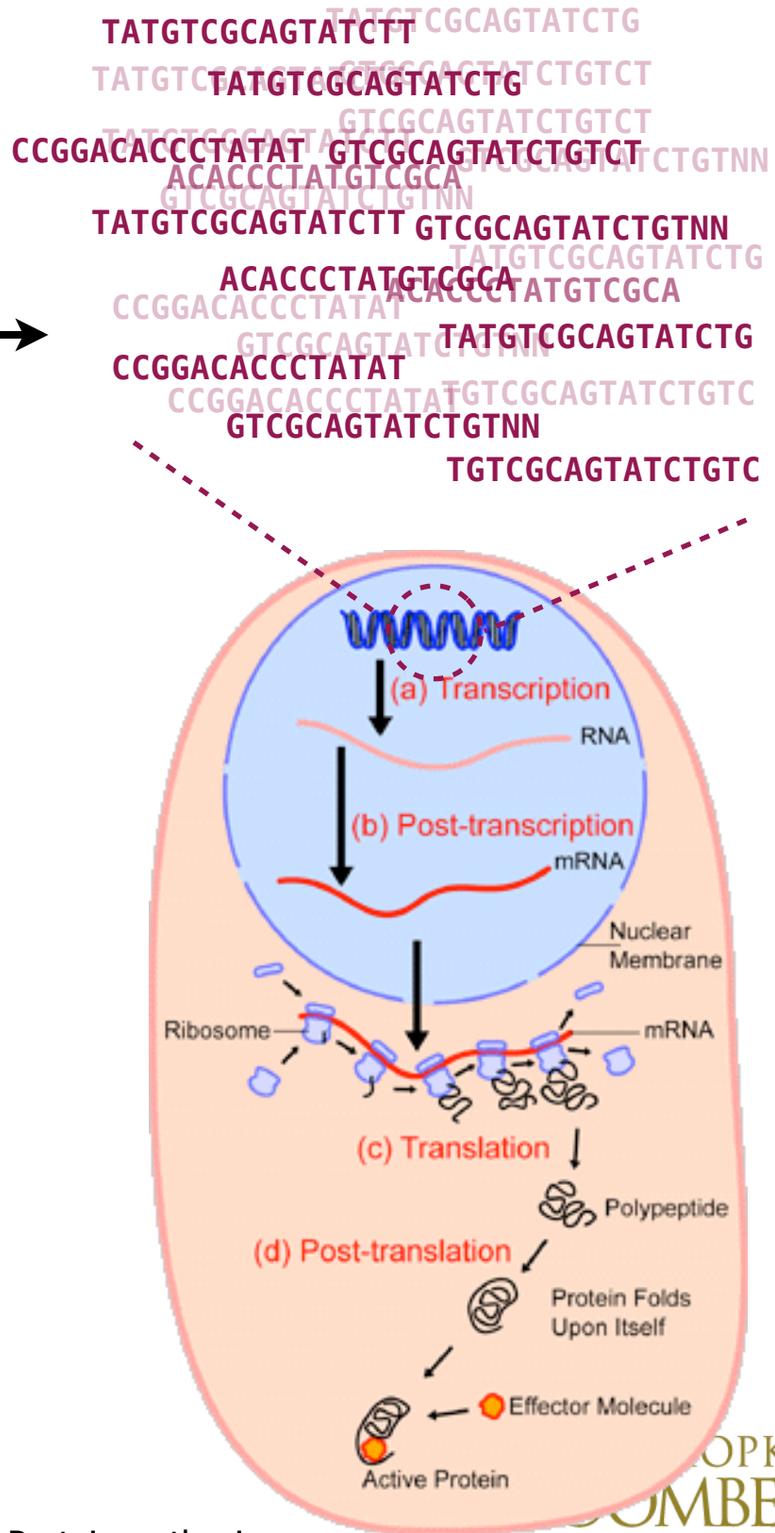


Source: http://en.wikipedia.org/wiki/File:Proteinsynthesis.png

Sequencing

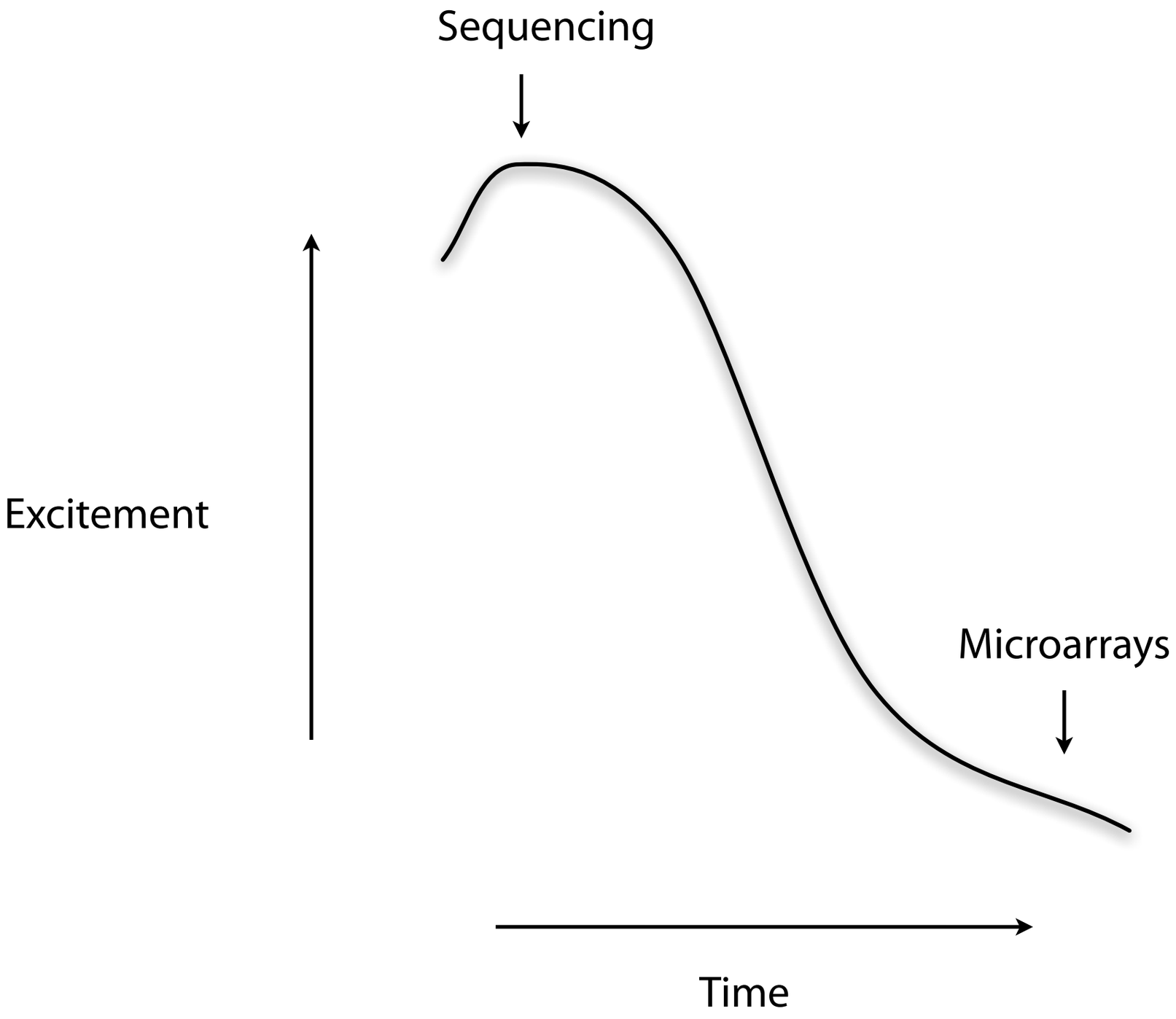
Illumina HiSeq 2000

Source: http://www.illumina.com/systems/hiseq_2000.ilmn



Source: http://en.wikipedia.org/wiki/File:Proteinsynthesis.png

Excitement



One instrument, many applications

Object of study → **DNA fragments** → **Text file**

Library preparation yields
DNA seq (genotyping)
ChIP seq
RNA seq
Small RNA seq
Ribosome profiling
CAGE seq
RACE seq
BS seq
...

Sequencing

Analysis, bottleneck →

Results

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

A steep curve

GA II
1.6 billion nt per day
(2008)

GA IIx
5 billion nt per day
(2009)

HiSeq 2000
25 billion nt per day
(2010)

HiSeq 2000
75 billion nt per day
(2011)

Images: www.illumina.com/systems

Numbers: www.politigenomics.com/next-generation-sequencing-informatics

Dates: Illumina press releases

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

The main players

- Illumina (GA, GA-II, GA-IIx, Hiseq-2000).
  Instrument market leader
- ABI SOLiD (1, 2, 3, 3+, 4, 5500)
- Roche 454
- Illumina Miseq
- Ion Torrent
- Complete Genomics (service, only human resequencing)
- PacBio

DNA library prep

↓ Fragmentation

↓ End repair
↓ Adaptor ligation

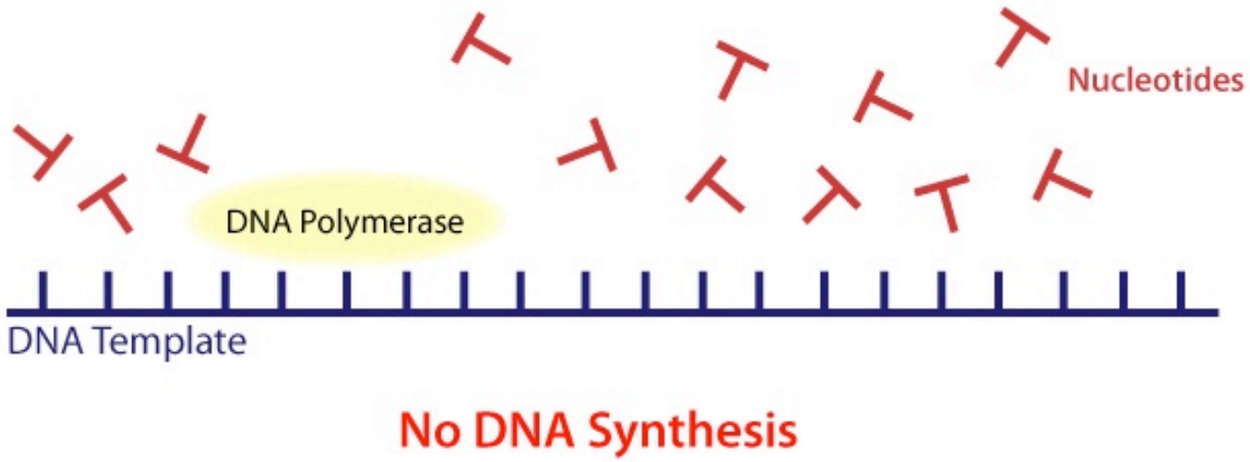
↓ Size selection

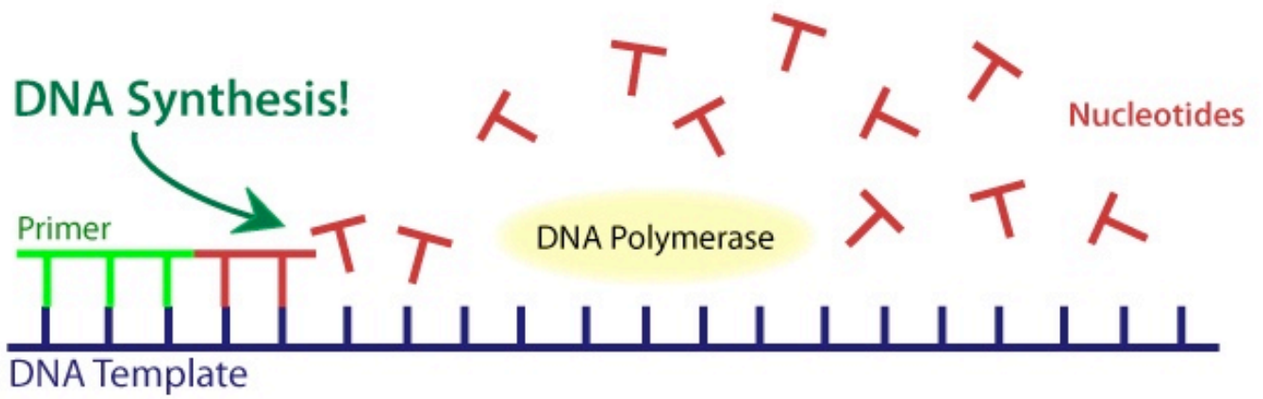
↓ PCR

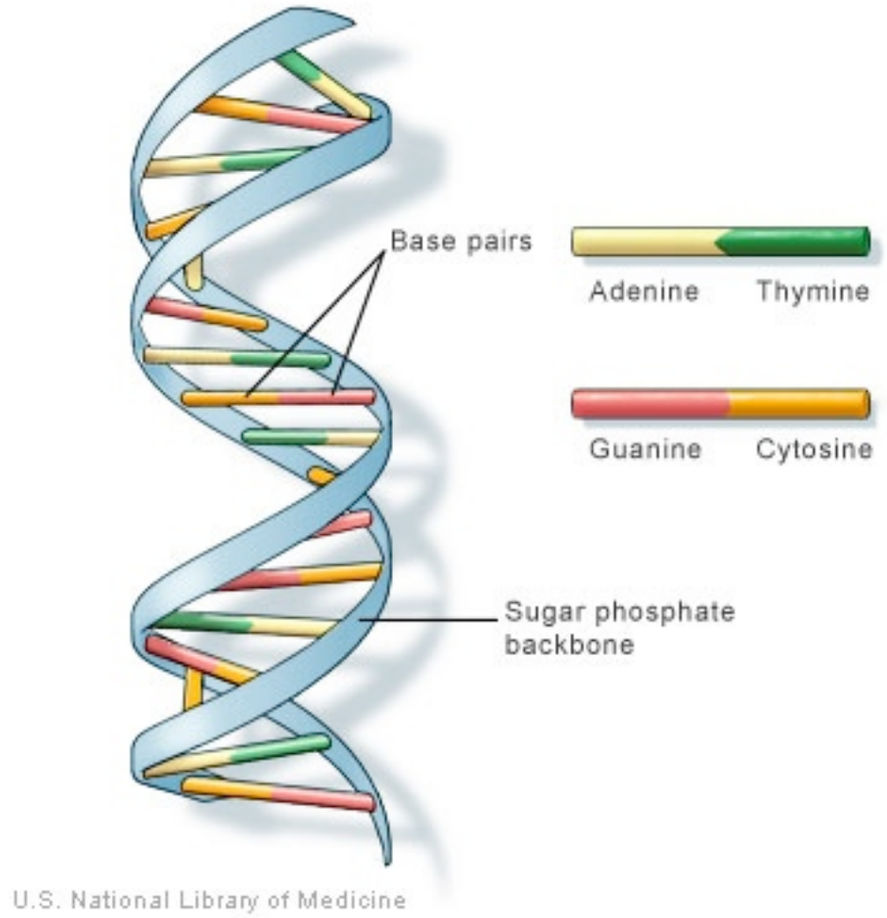
→ Flowcell

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

DNA synthesis



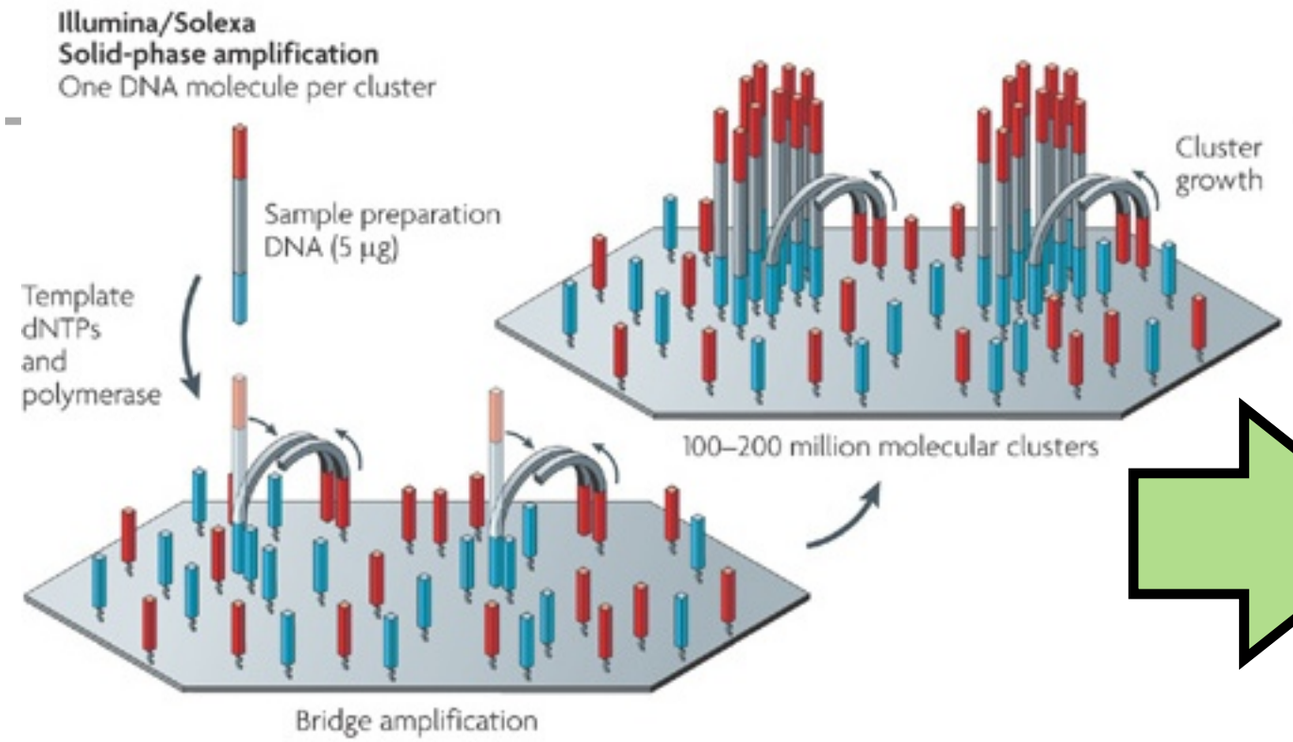




www.visionlearning.com/library/module_viewer.php?mid=180&l=

www.dna-sequencing-service.com/dna-sequencing/dna-hydrogen-bonds-2/

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

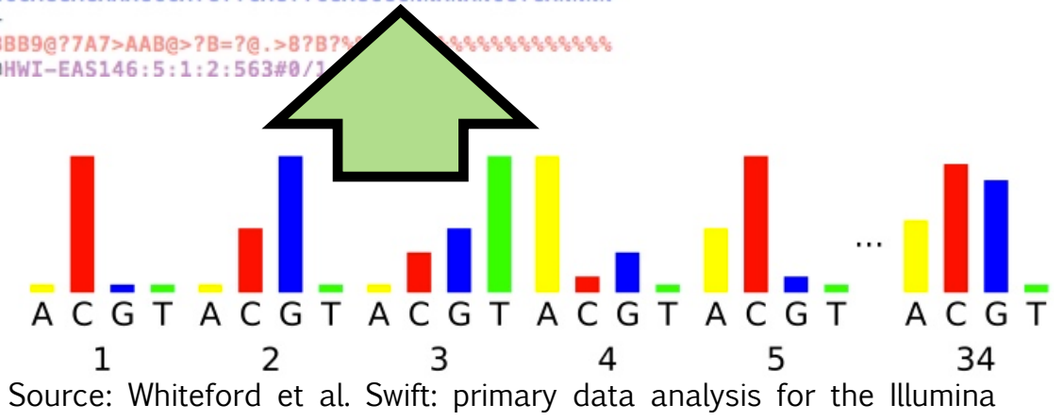


Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

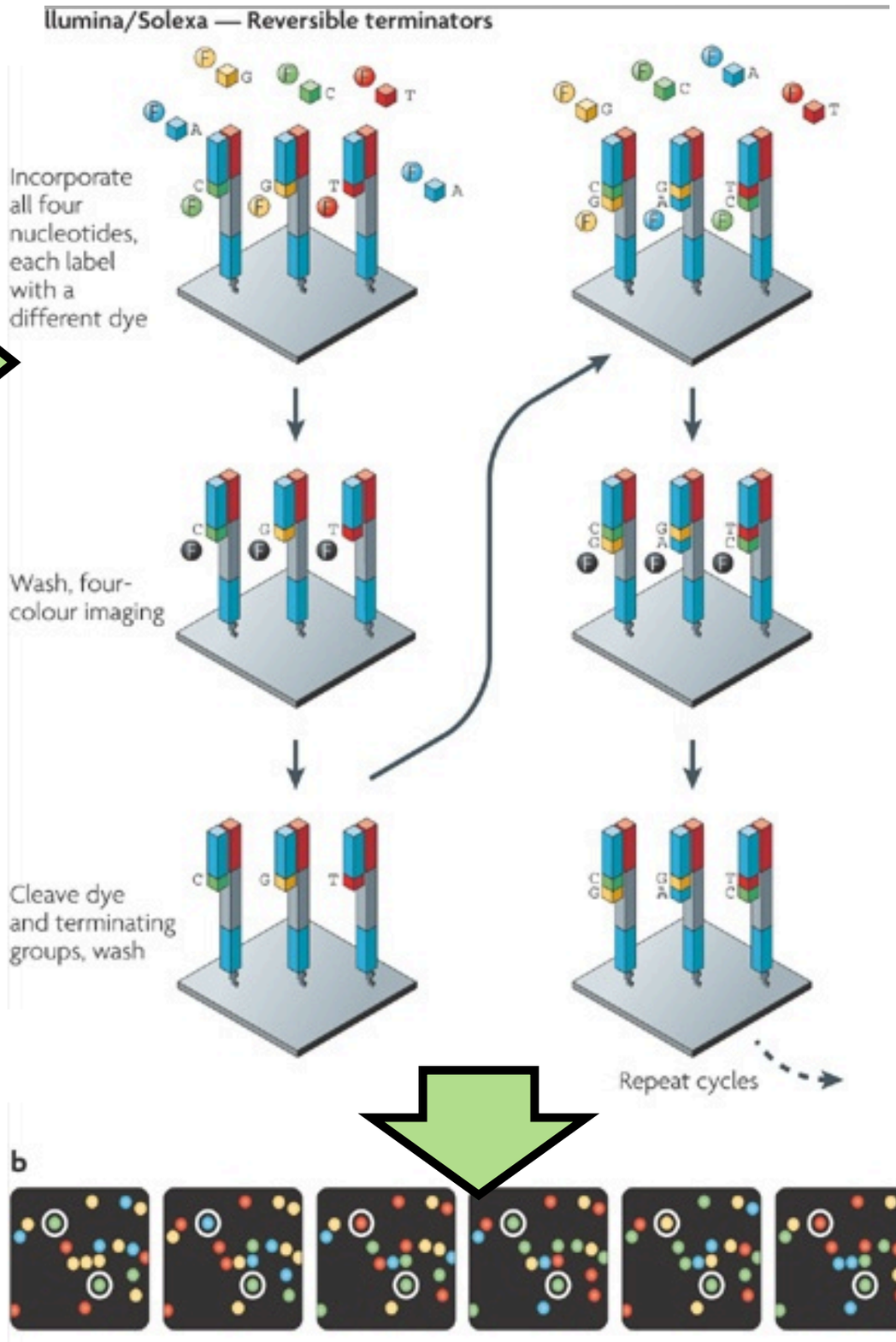
```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B7&7:>B@:A>?9:<;:>?4?%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAGTTCGAGCGCNNANANCGTGANNNN
+
BBB9@?7A7>AAB@>?B=?@.>B?B?%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
```

name
sequence
quality scores

x 100s of millions

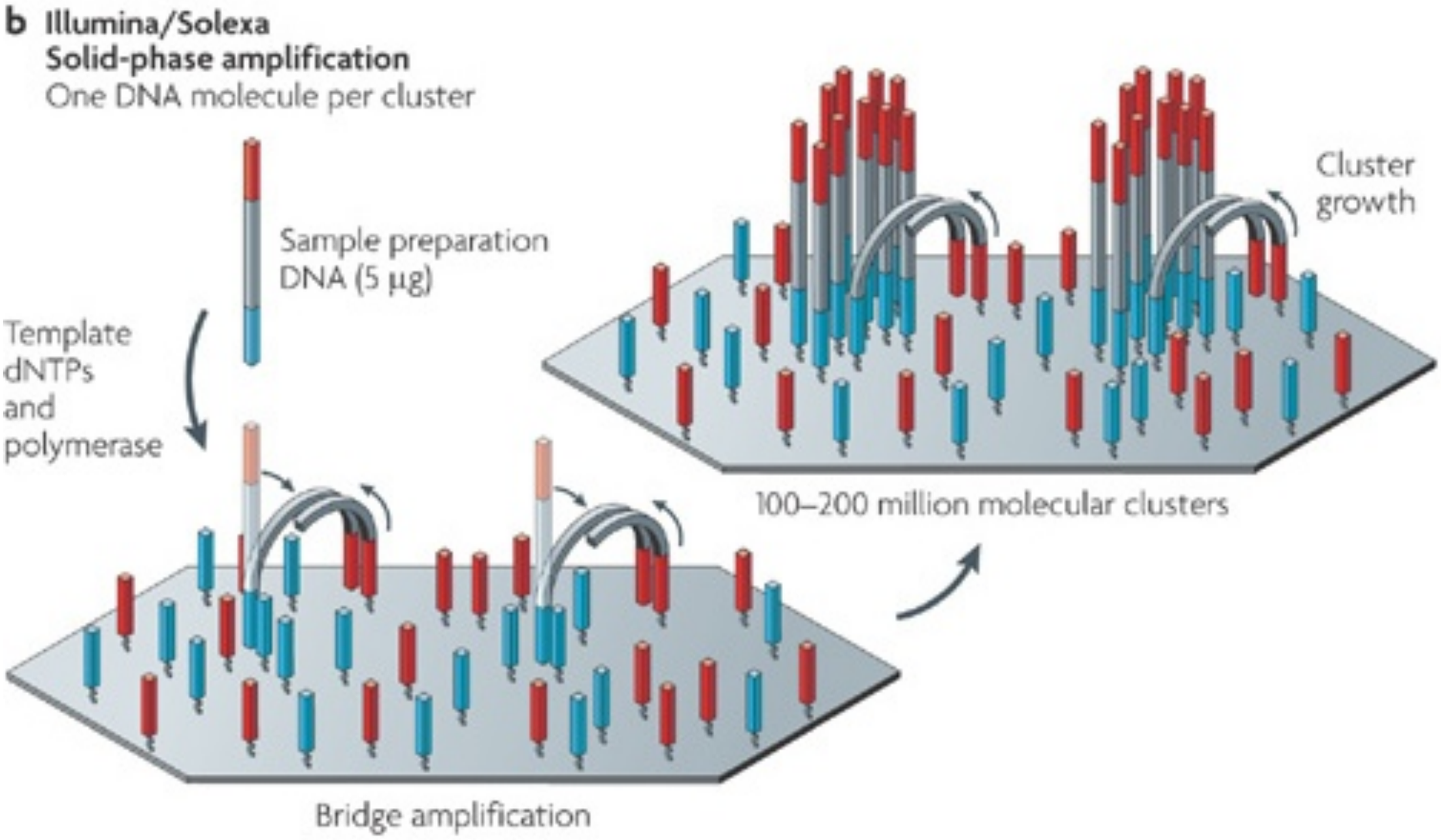


Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009



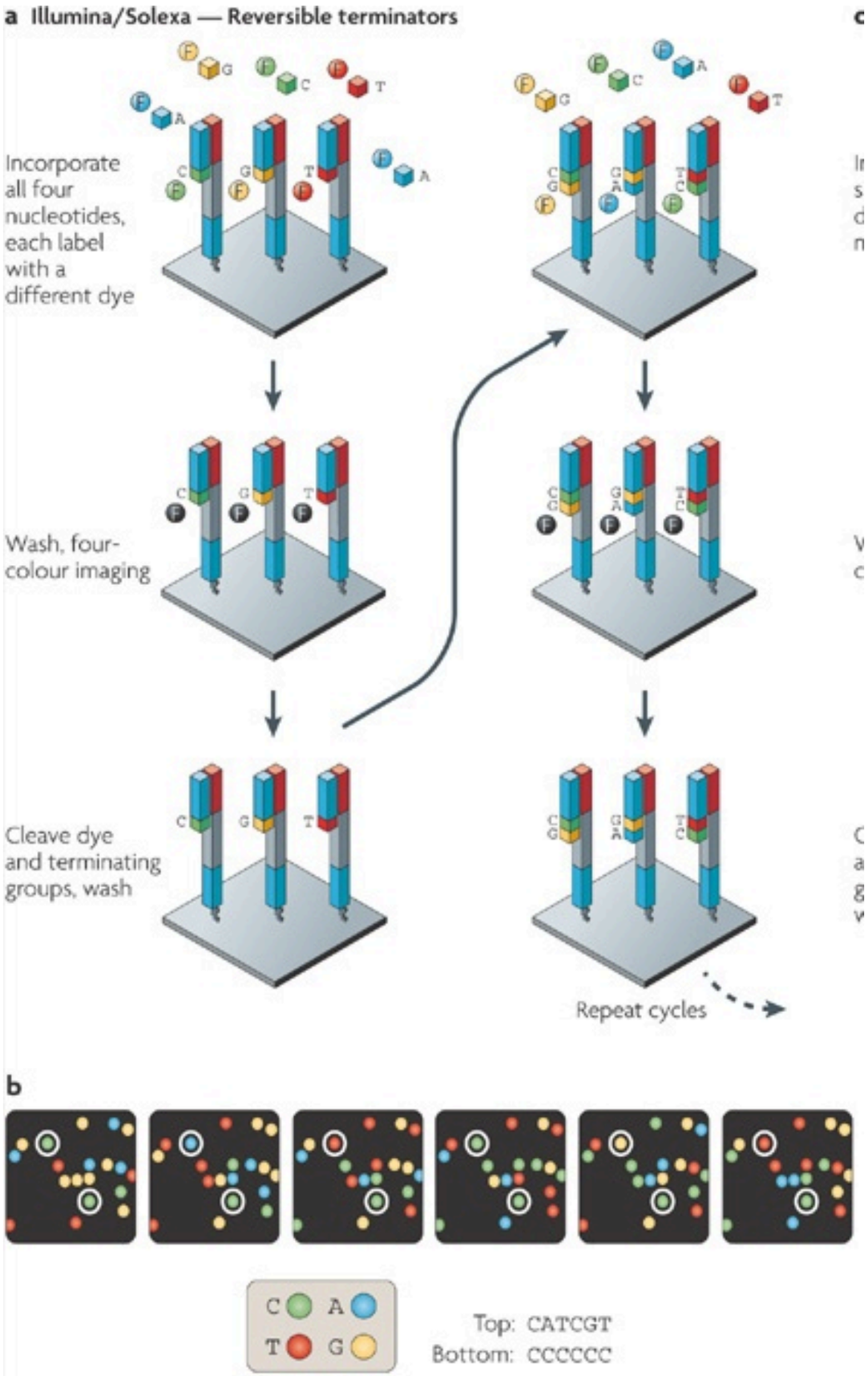
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

Bridge amplification



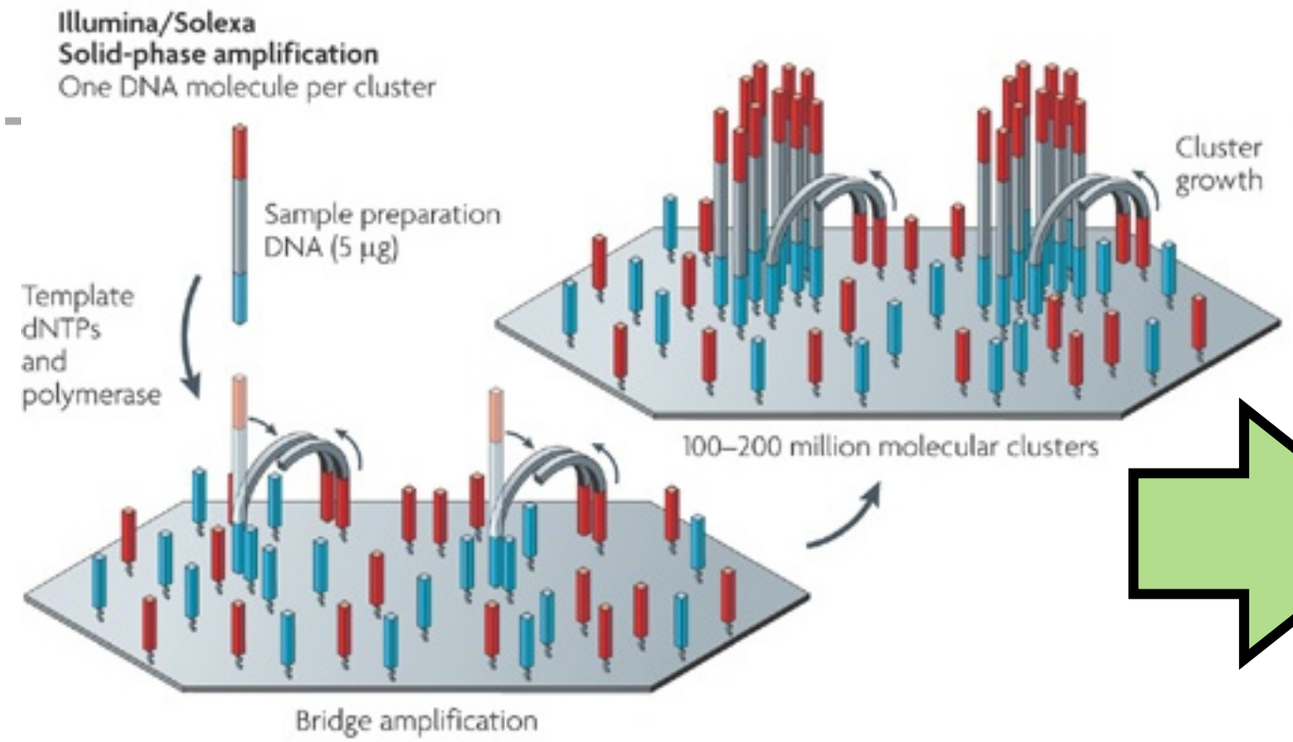
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

Sequencing a cluster



Consequences for error profiles

- Insertions and deletions are very rare (in the reads)
- Focus on substitutions (which are easier to deal with)
- Error rates increases with position in the read (cycle)

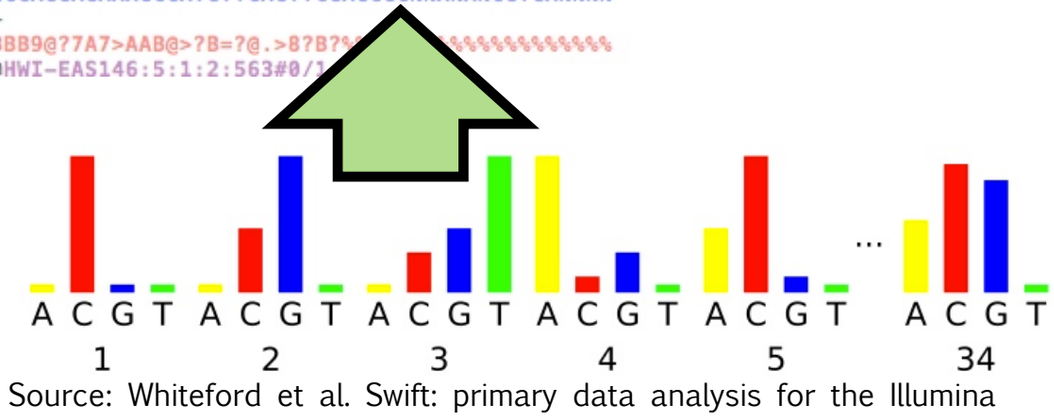


Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

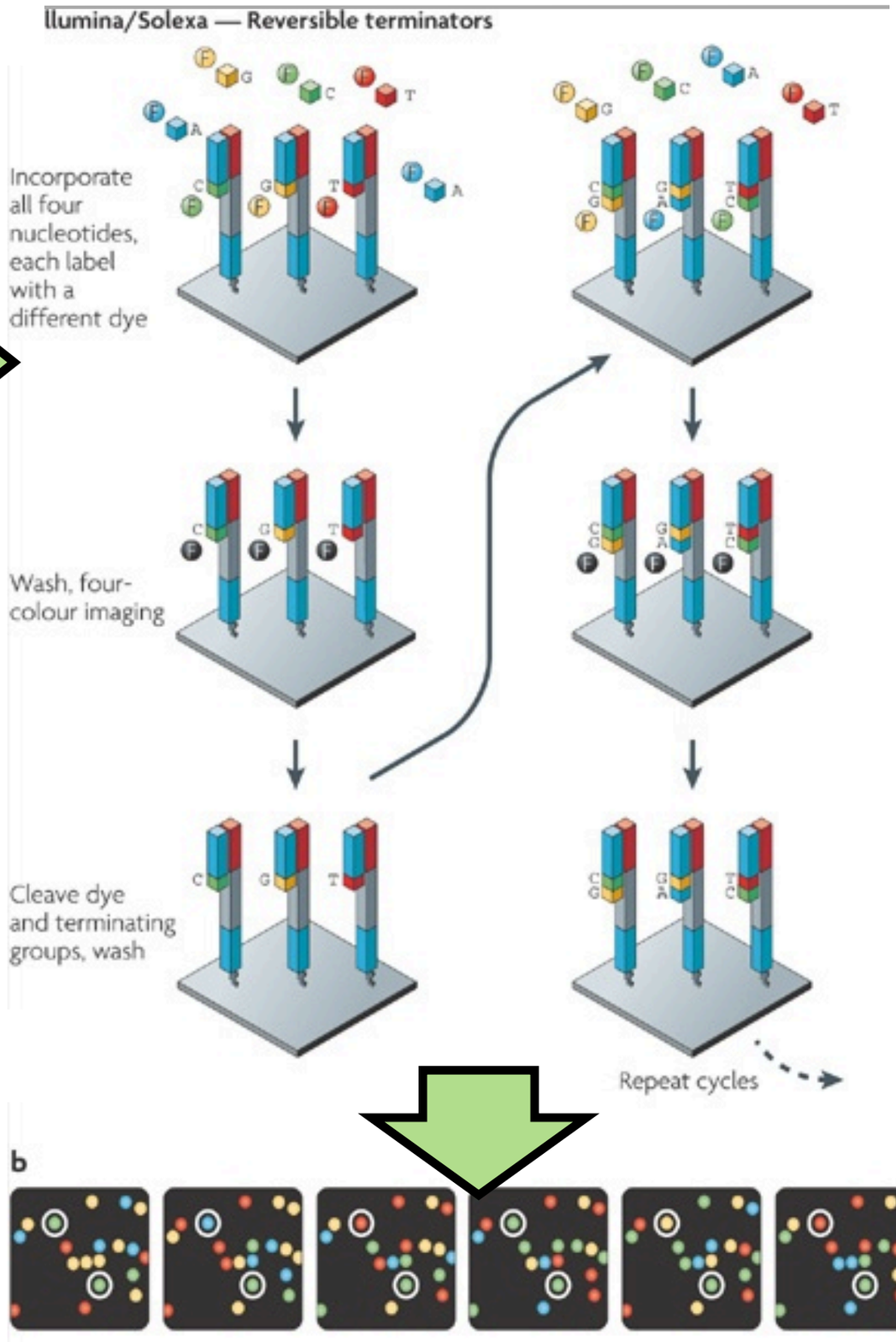
```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B7&7:>B@:A>?9:<;:>?4?%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAGTTCGAGCGCNNANANCGTGANNNN
+
BBB9@?7A7>AAB@>?B=?@.>B?B?%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
```

name
sequence
quality scores

x 100s of millions

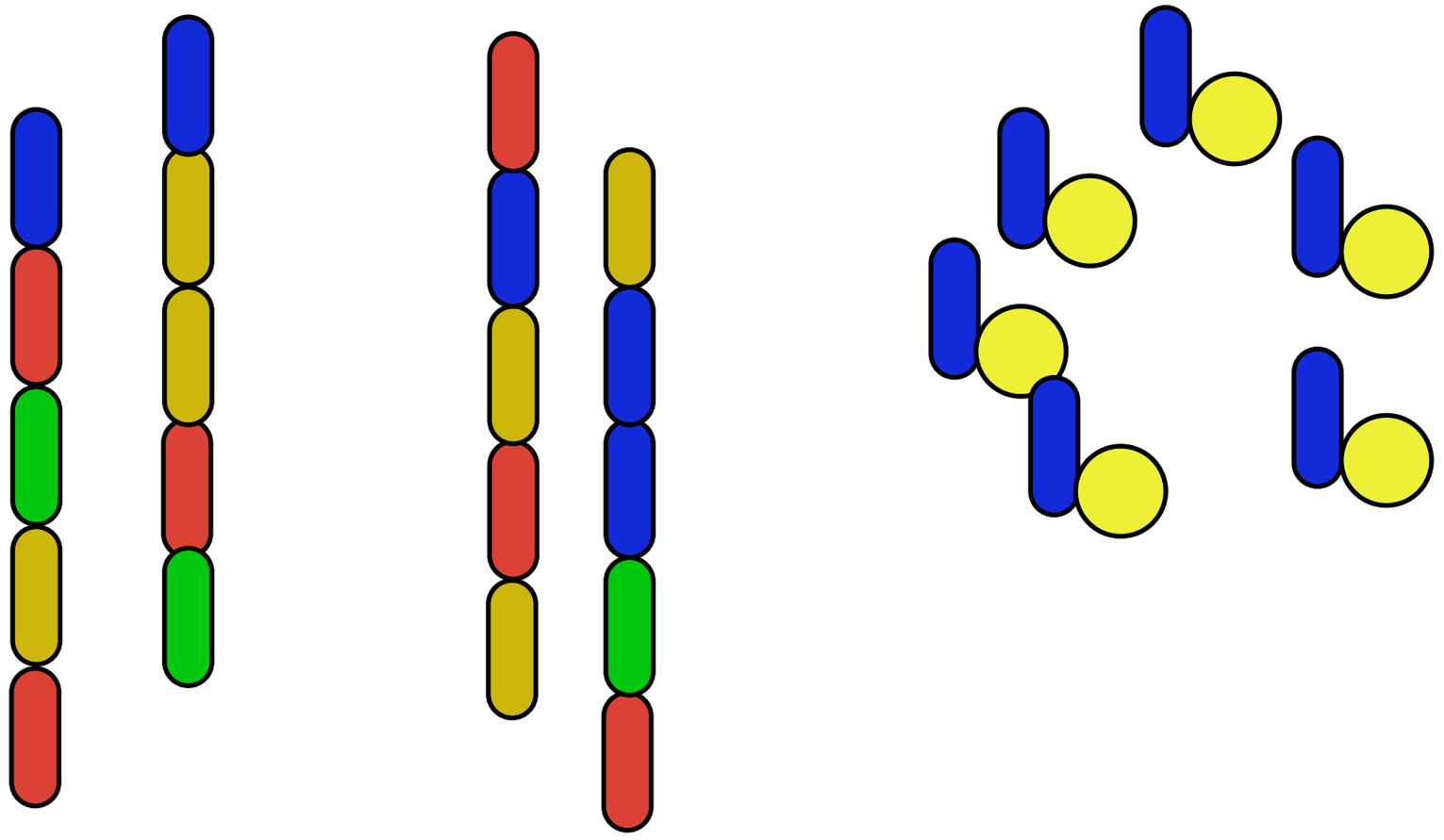


Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009

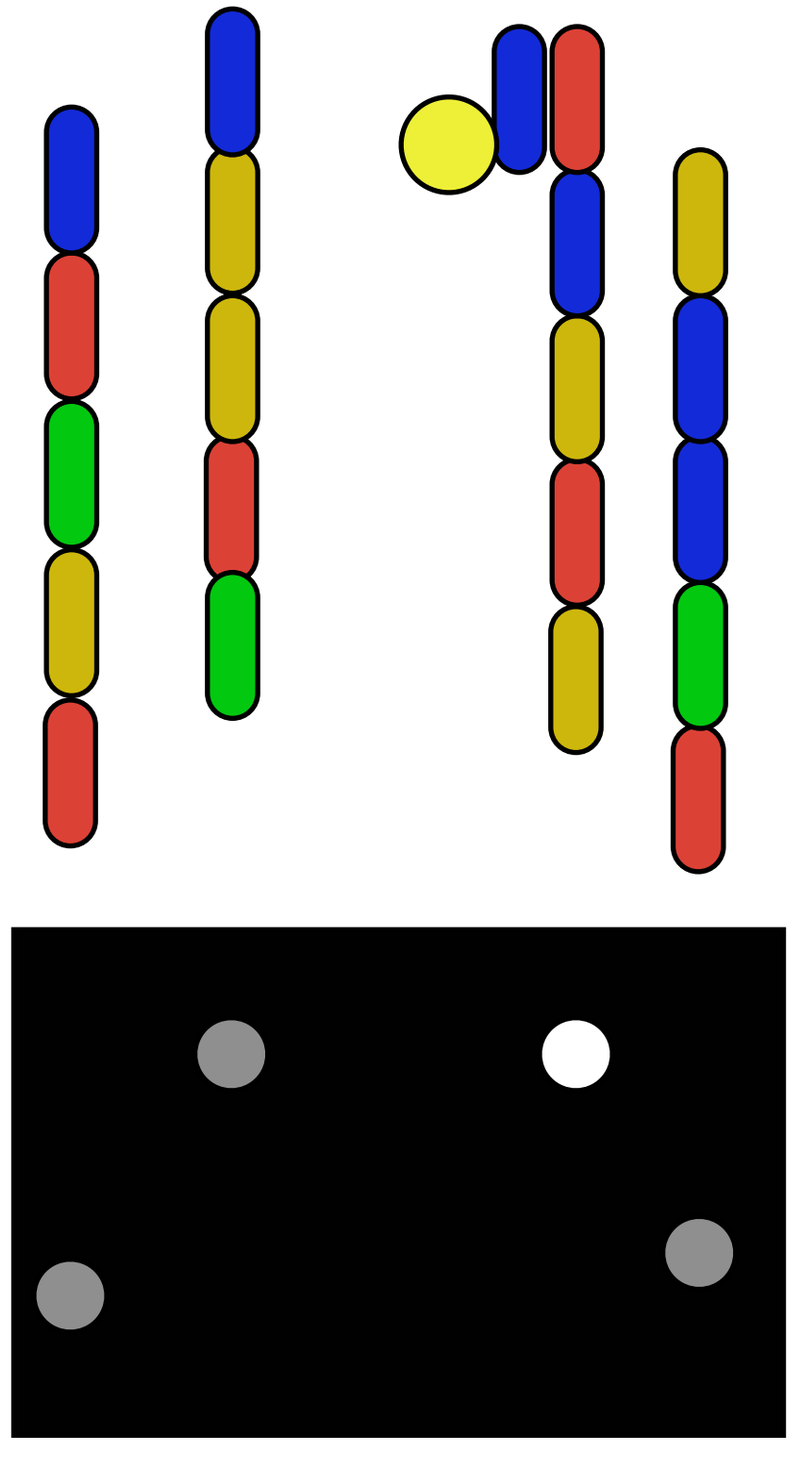


Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

Sec-gen Sequencing

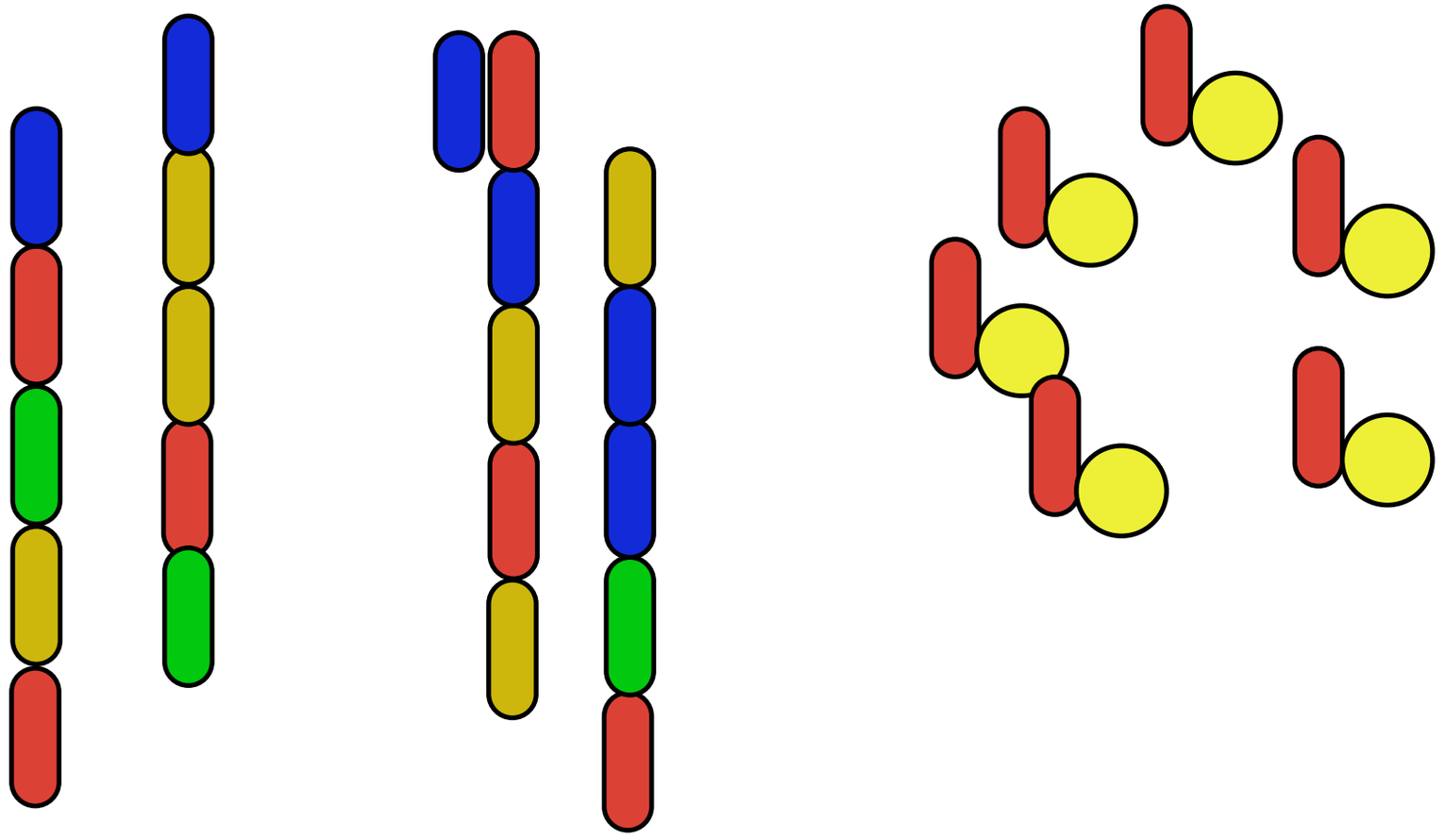


Sec-gen Sequencing

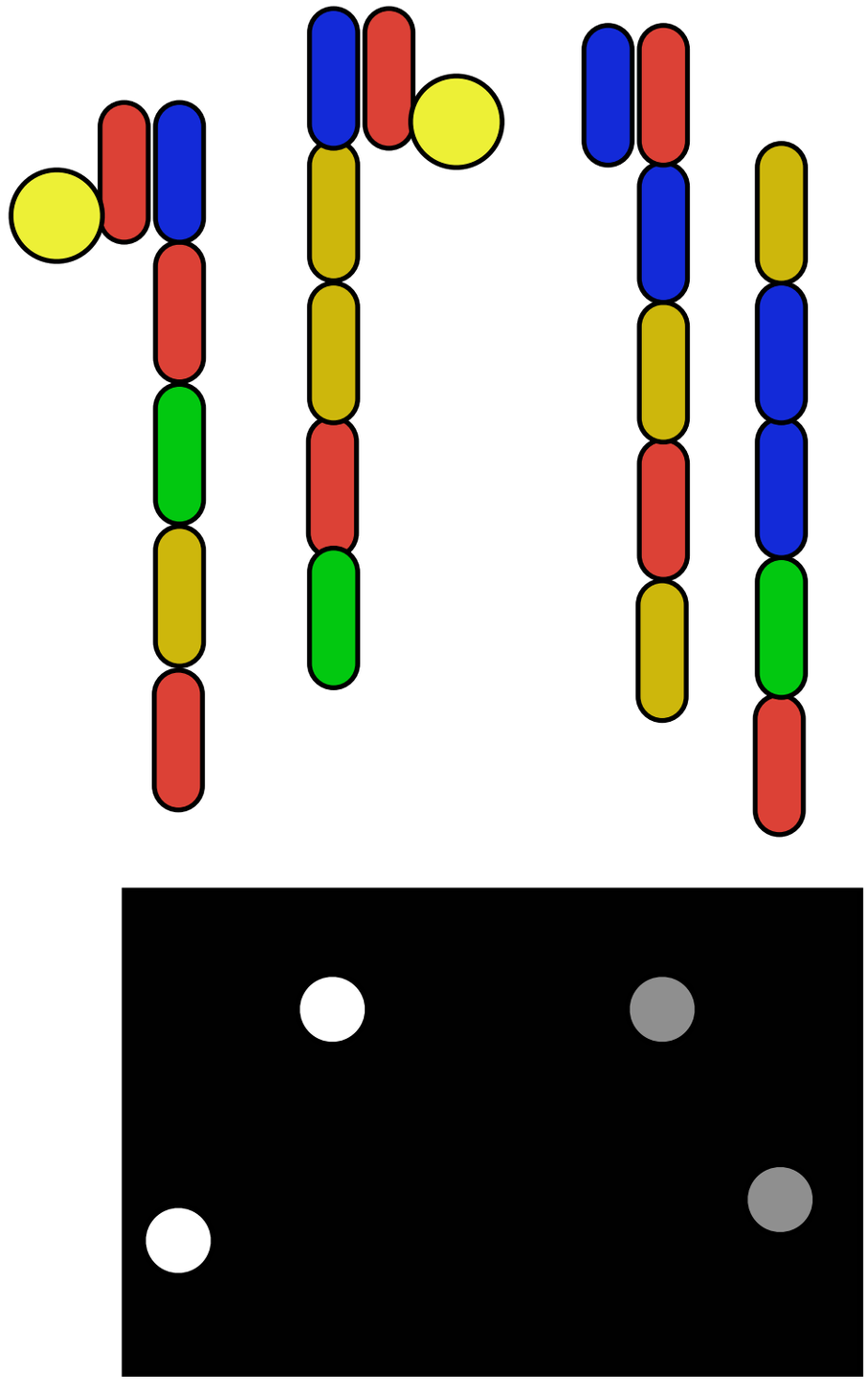


JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

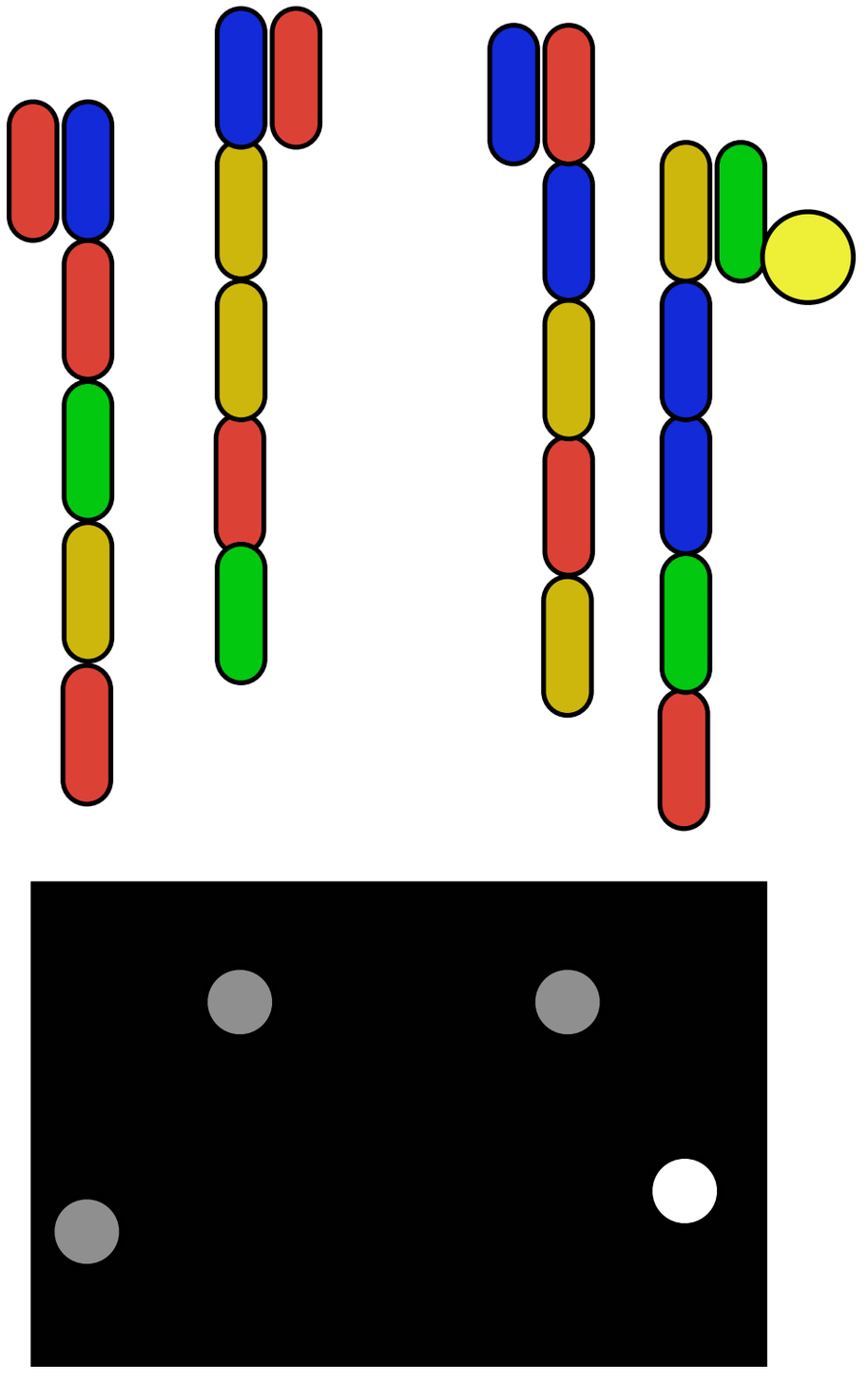
Sec-gen Sequencing



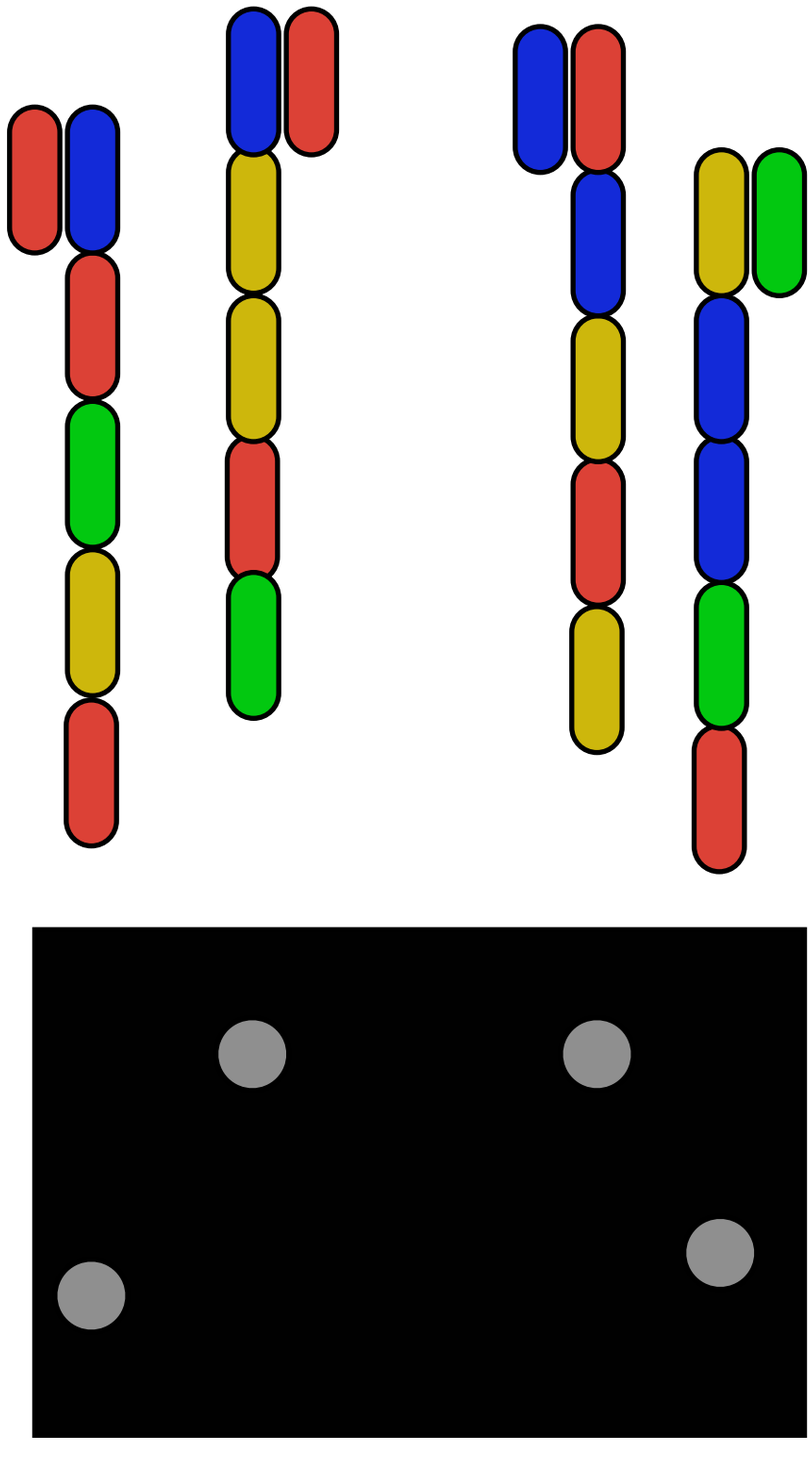
Sec-gen Sequencing



Sec-gen Sequencing

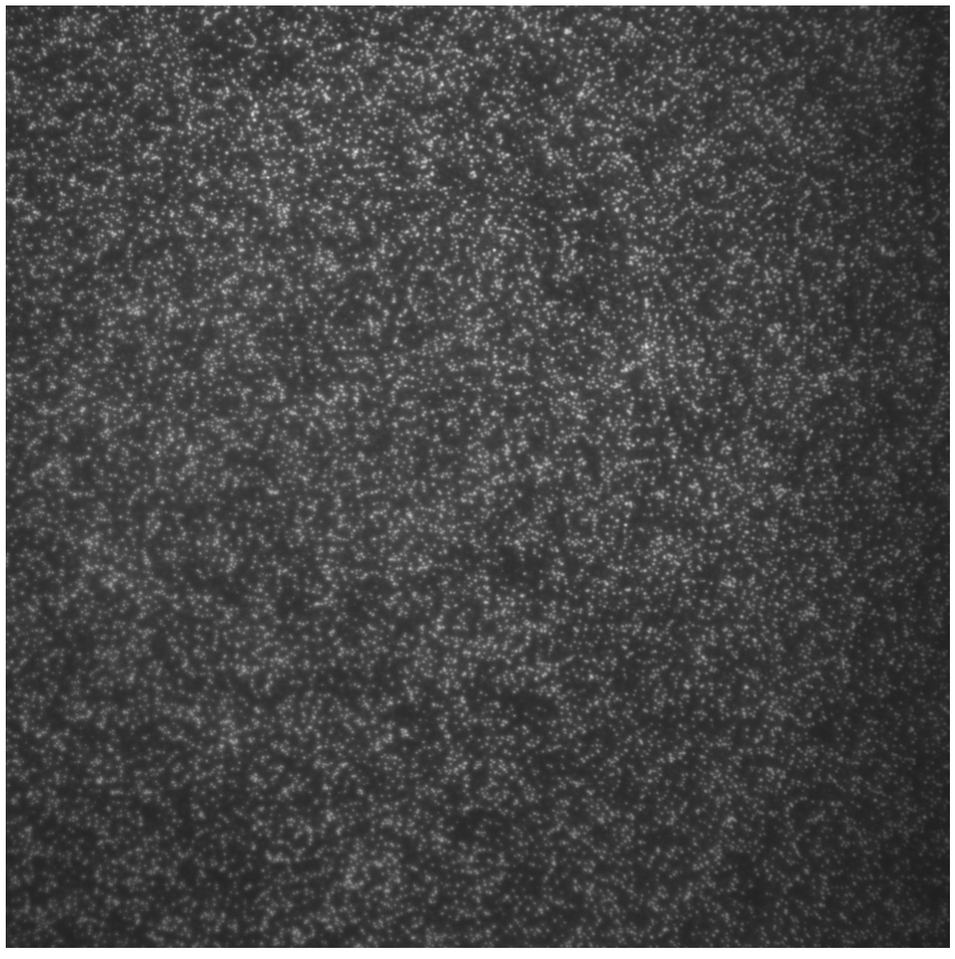


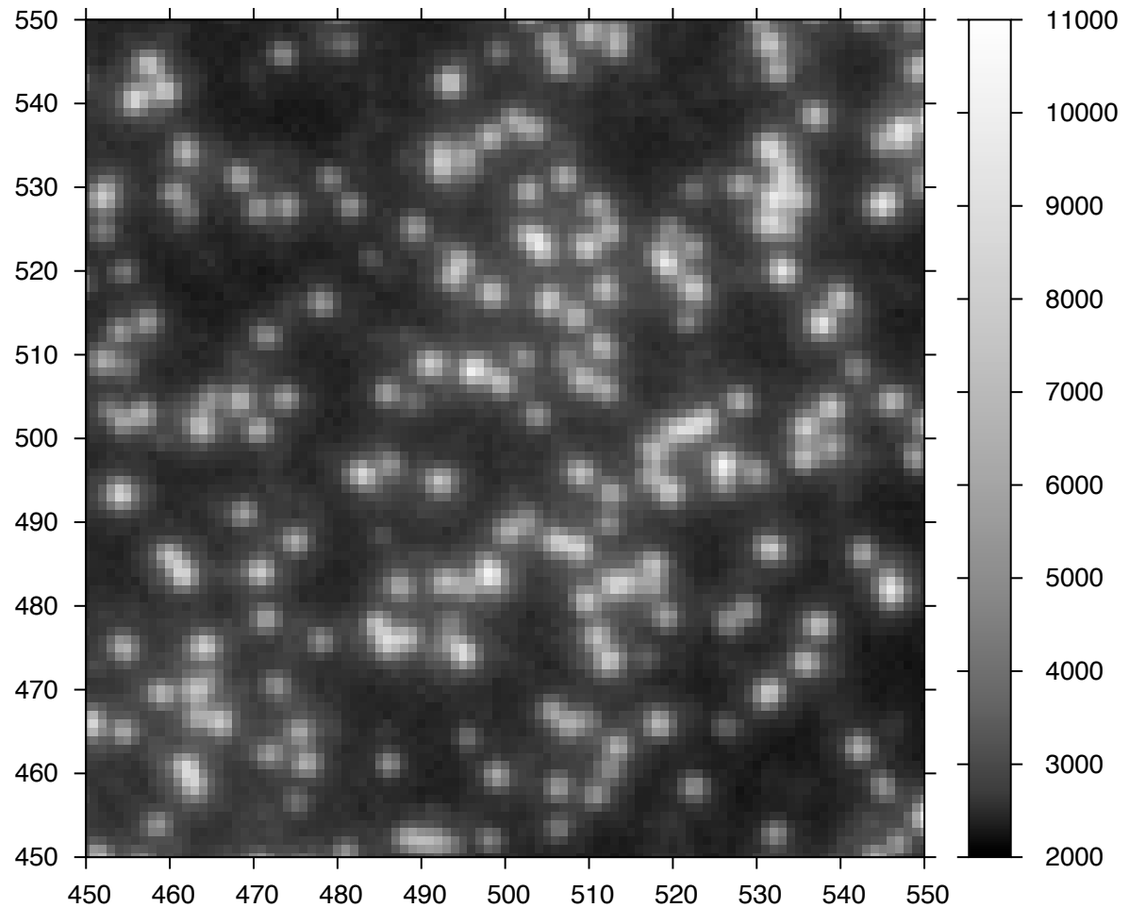
Sec-gen Sequencing



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Image Analysis





An input image and zoomed in section

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Image Analysis

4 images per cycle

~100 tiles

Analysis:

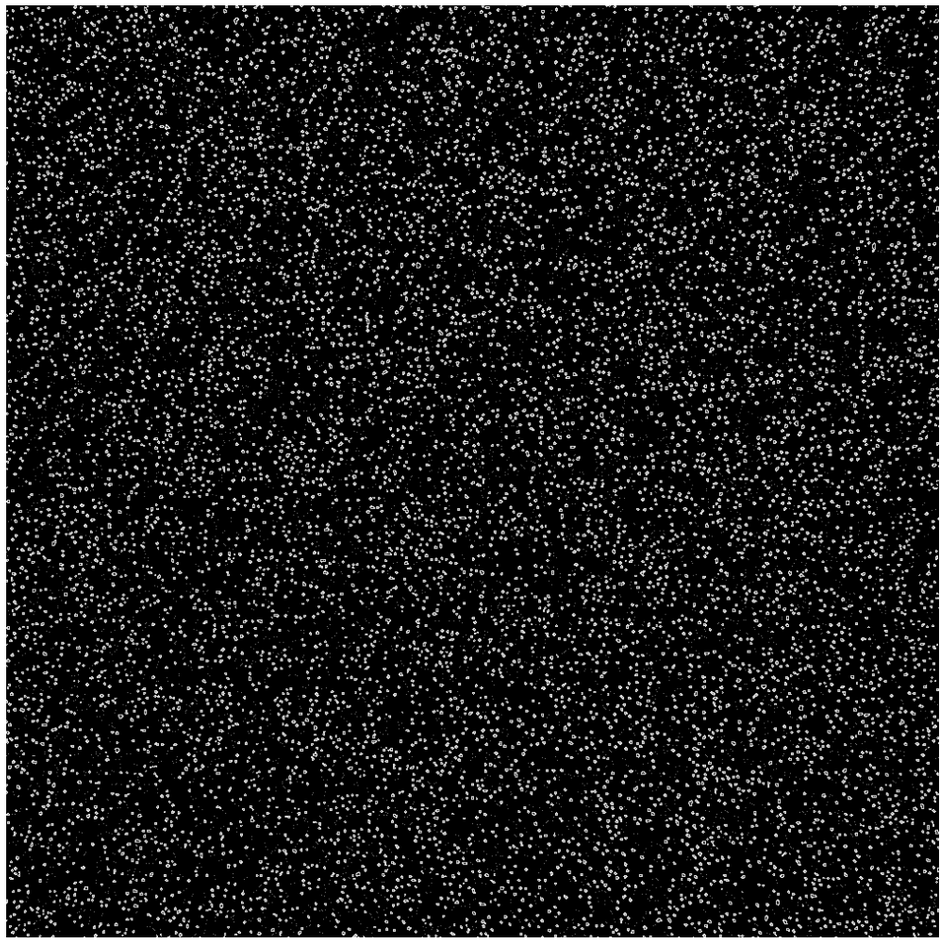
Filtering

Background subtraction

Thresholding

Each image analysis independent (so can parallelize)

Image Analysis



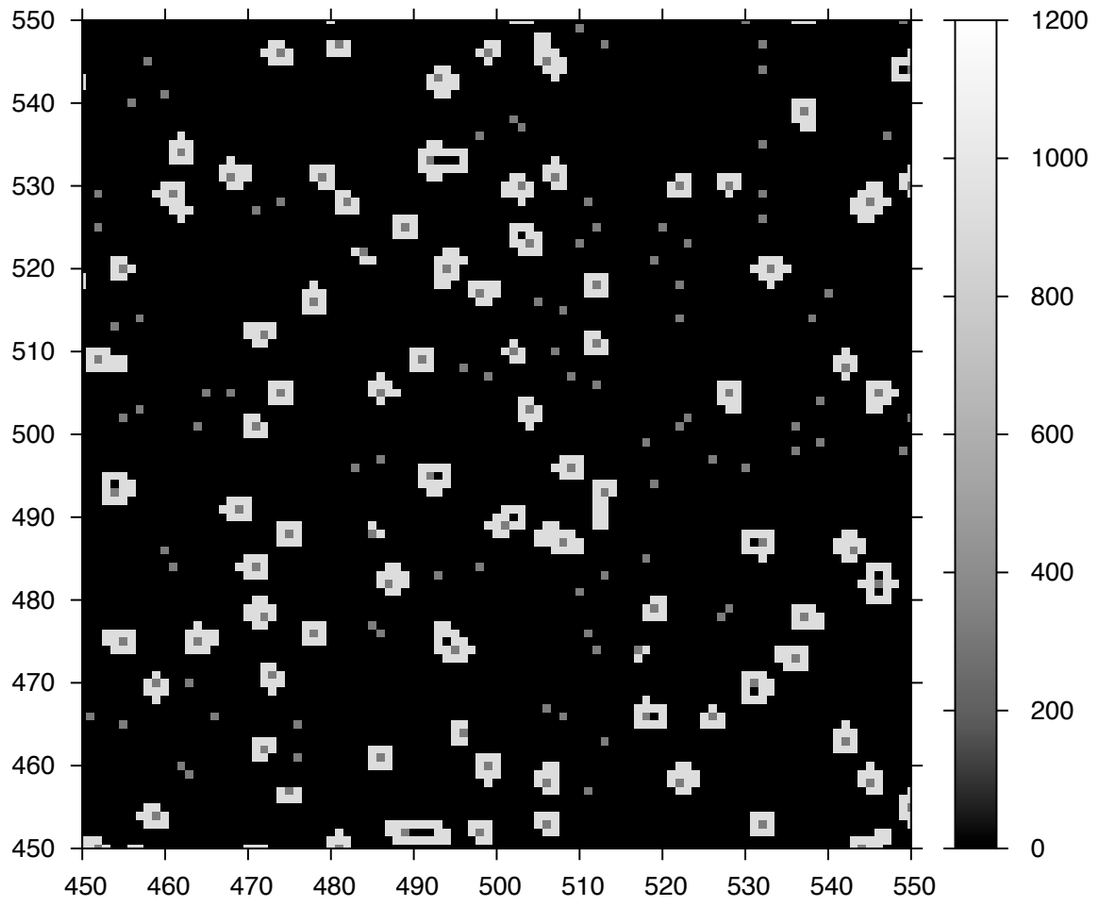
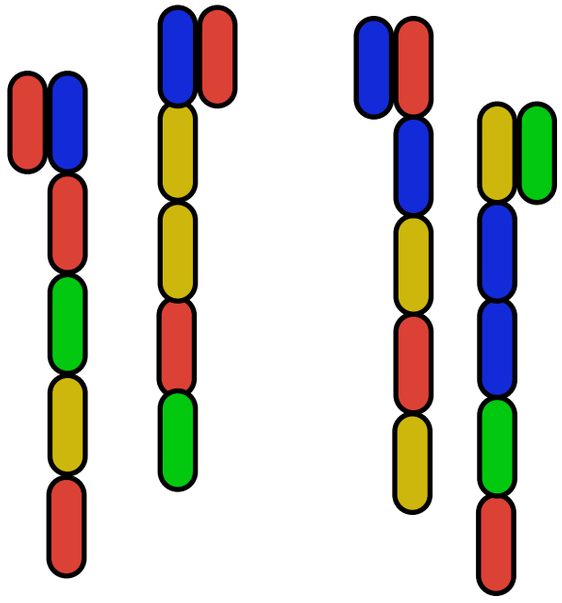


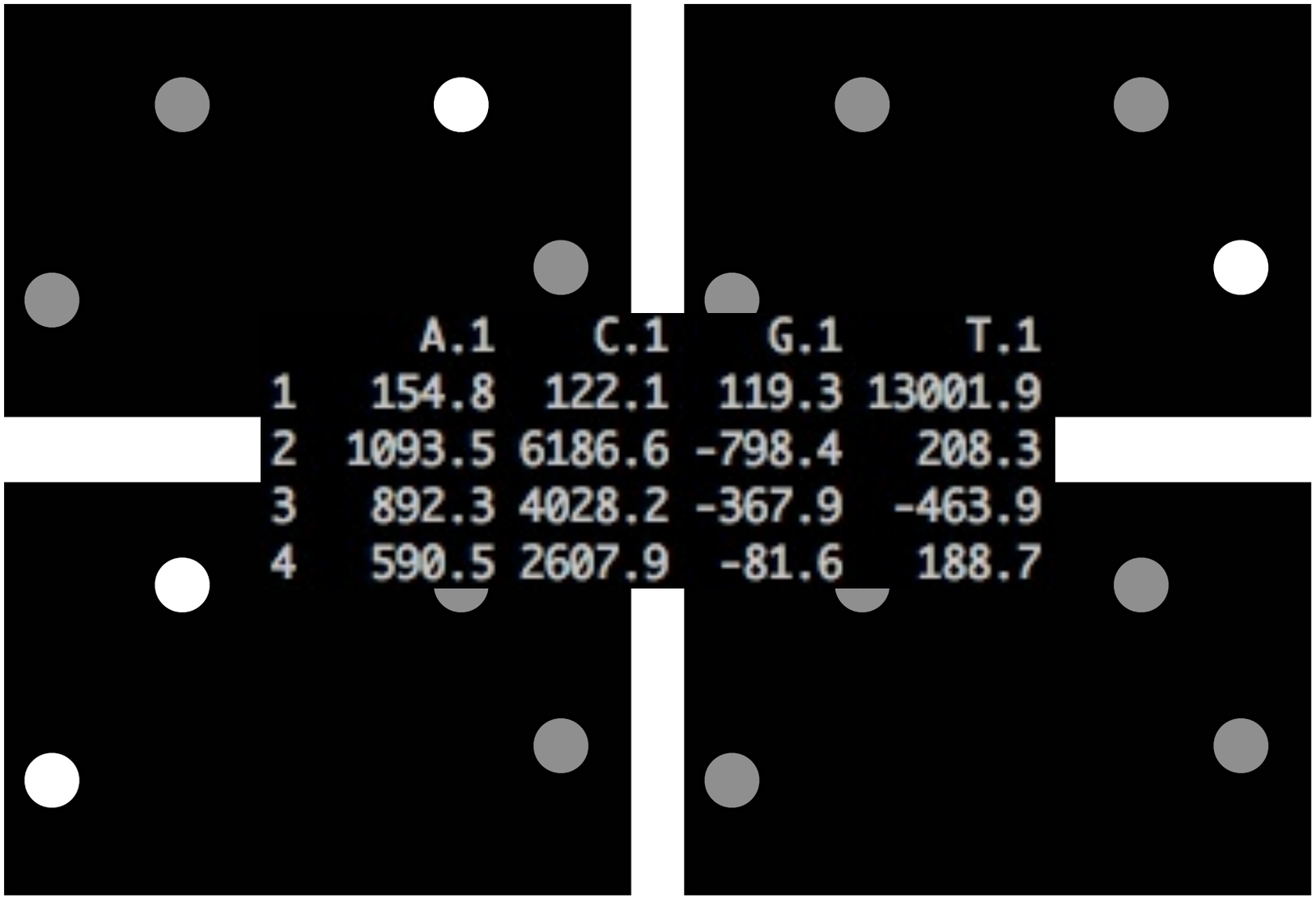
Image after processing. *This is old, cluster density is much higher now*

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

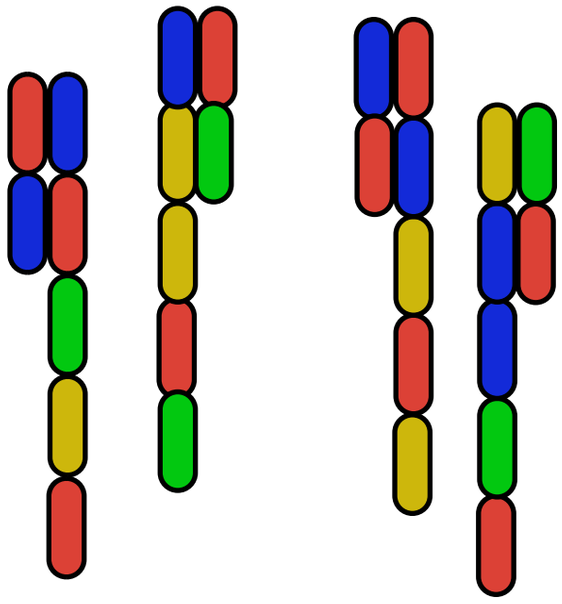
Sec-gen Sequencing



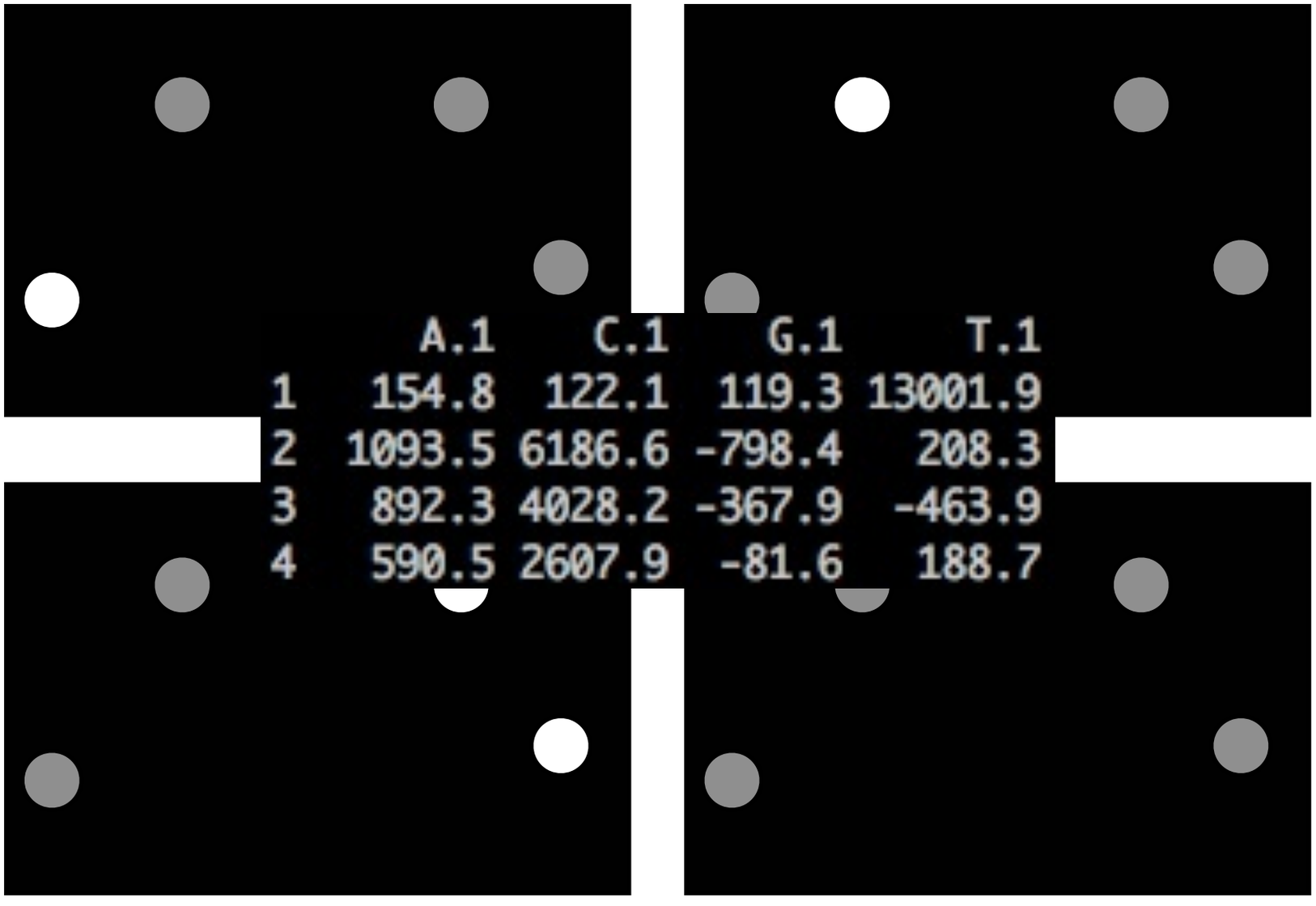
First Cycle



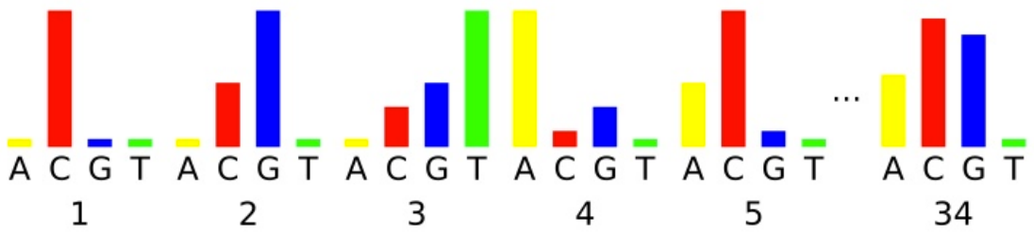
Sec-gen Sequencing



Second Cycle



Basecalling



Basecalling

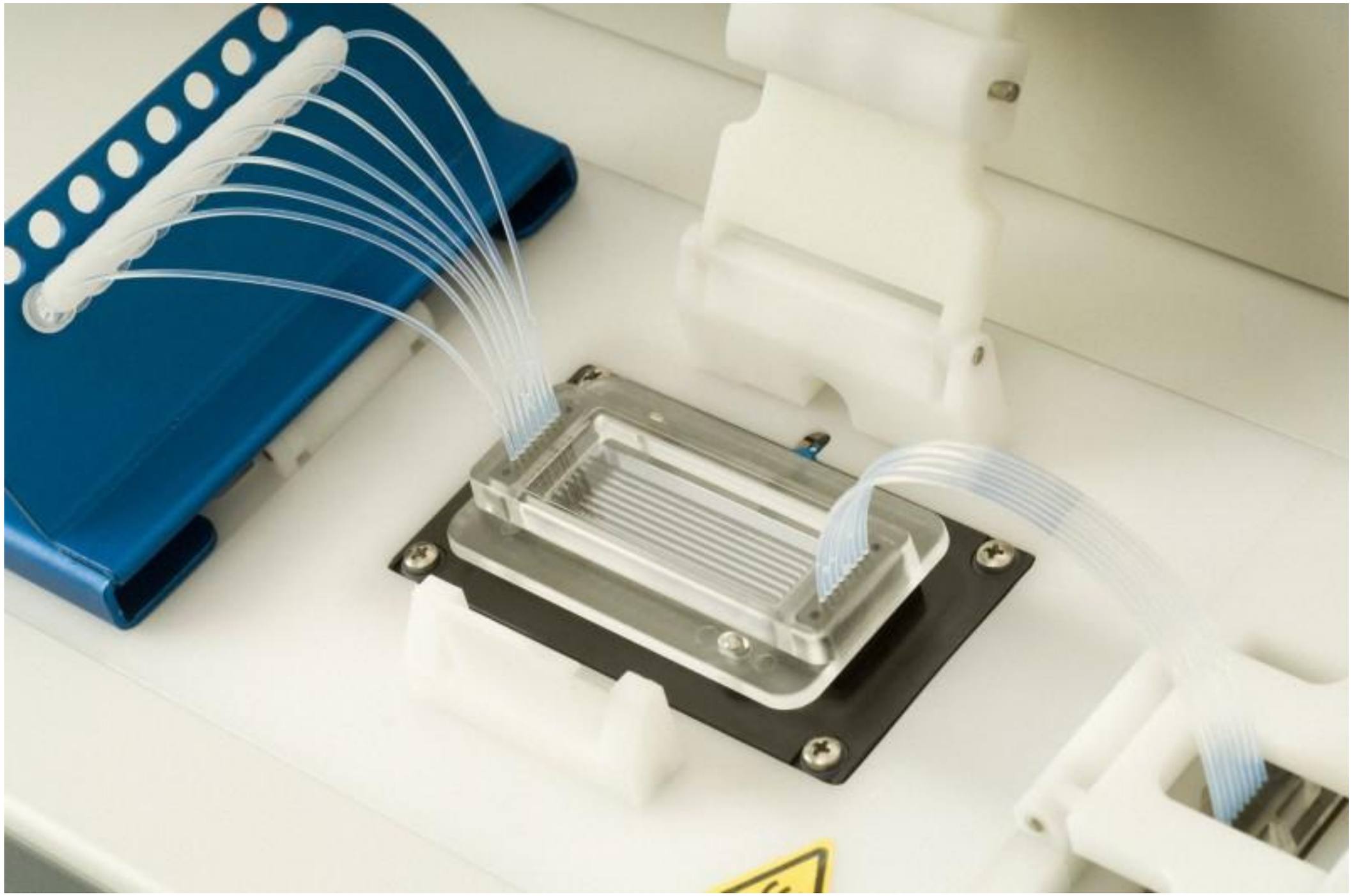
```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B7&7:>B@:A>?9:<;:>?4?%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAGTTCGAGCGCNNANANCGTGANNNN
+
BBB9@?7A7>AAB@>?B=?@.>8?B?%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
```

name
sequence
quality scores

x 100s of millions

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Flowcell / cluster station



8 lanes (often 1 is reserved for Phi X control)
Impacts experimental design

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

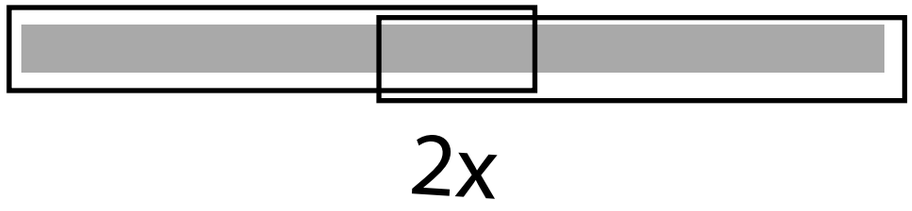
Tricks

200-300 bp fragment

100 bp read

Paired end sequencing

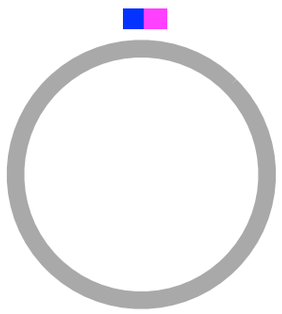
Insert



(Un)intentional

2x

long fragment



circularize

Long insert

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Major players

- **Illumina**
  2x100 bp, fragment~300bp
  No indels

- **ABI SOLiD**
  75+35 bp
  good barcoding
  No indels
  Every base gets read ‘twice’, outputs colorspace (pain)

- **Roche 454**
  500+ reads, variable length
  homopolymers are a problem
  Far fewer reads than Illumina/SOLiD

Flowgrams / 454 sequencing

ACCCTGGA

↓

A ~ 0.9
C ~ 3.4
T ~ 1.2
G ~ 1.9
A ~ 1.1

↓

Sequence

Homopolymer:
stretch of the ‘same’ nt
like ‘CCCCCC’

This explains why homopolymers are a problem for 454.