CMSC 858D :

Details of Sequencing Technologies: short & long-read sequencing

NOTE: Illumina sequence slides are taken from
http://www.slideshare.net/USDBioinformatics/illumina-sequencing

CMSC 858D :

Details of Sequencing Technologies: short & long-read sequencing

NOTE: Illumina sequence slides are taken from
http://www.slideshare.net/USDBioinformatics/illumina-sequencing

Illumina Diagram

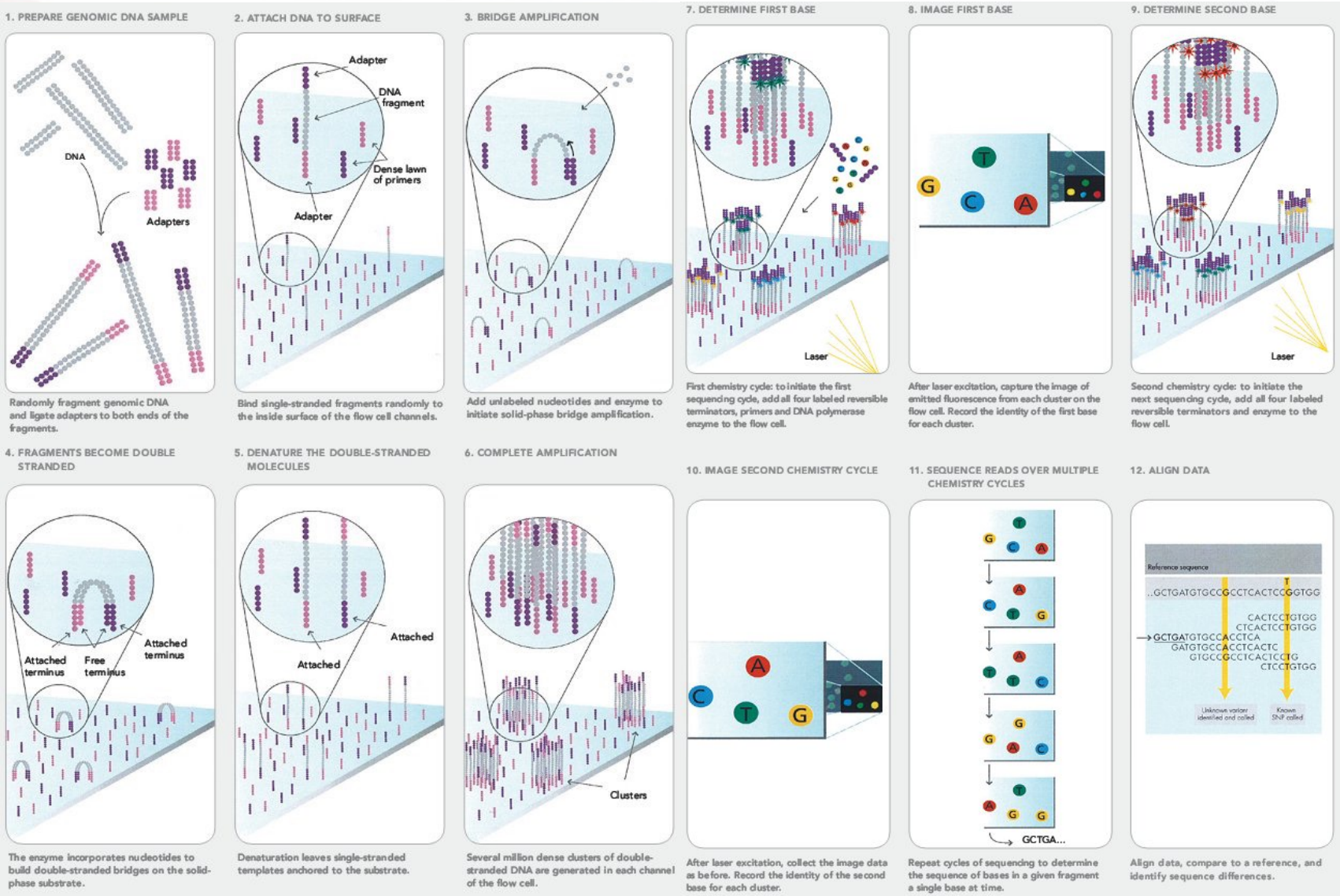


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina Diagram

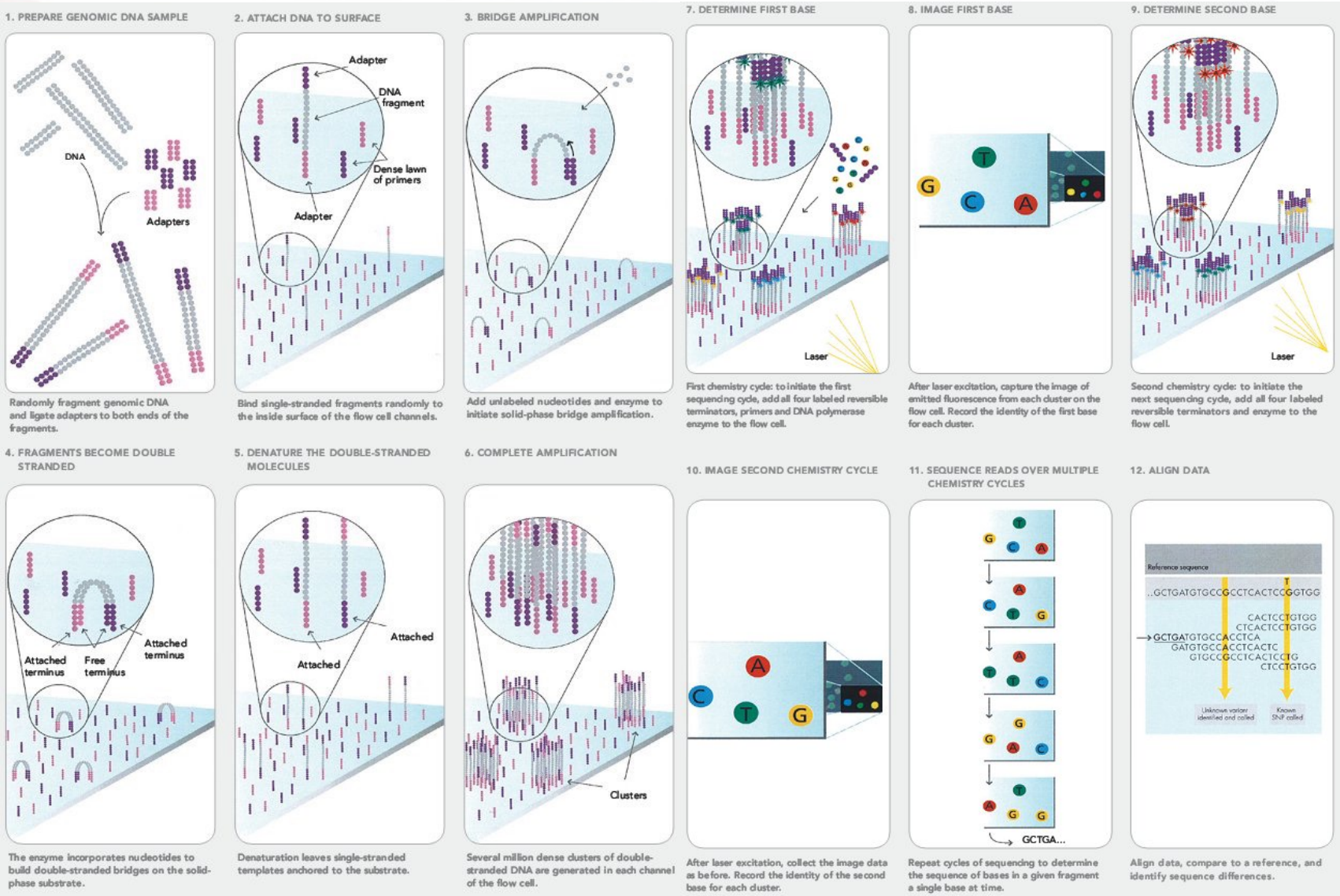


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Prepare Genomic DNA Sample

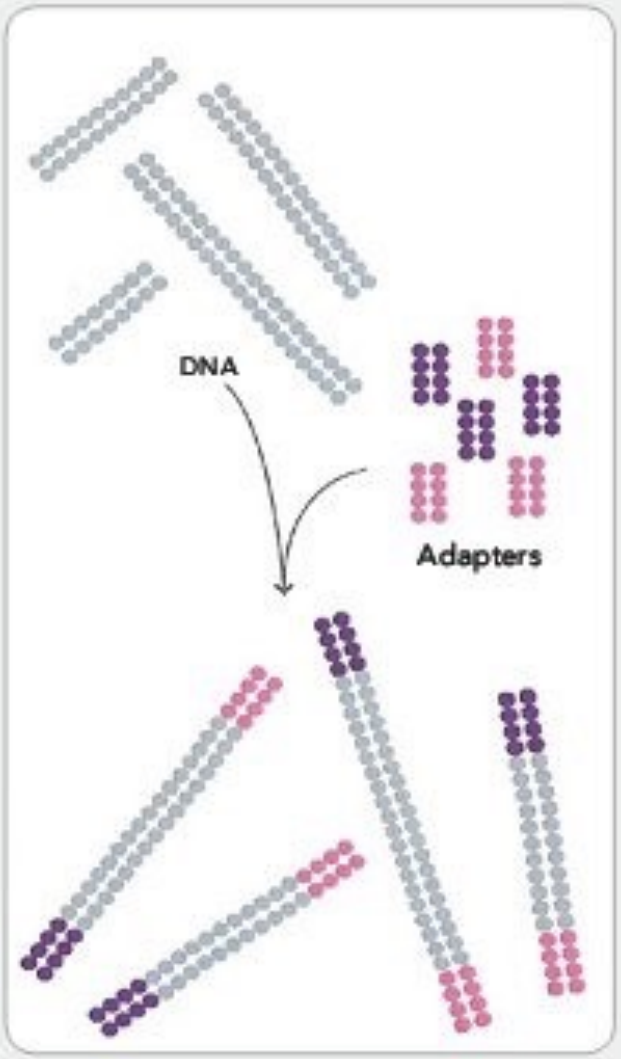


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- Fragment DNA of interest into smaller strands that are able to be sequenced
  - Sonication
  - Nebulization
  - Enzyme digestion
- Ligate Adapters
- Denature dsDNA into ssDNA by heating to 95° C

Prepare Genomic DNA Sample

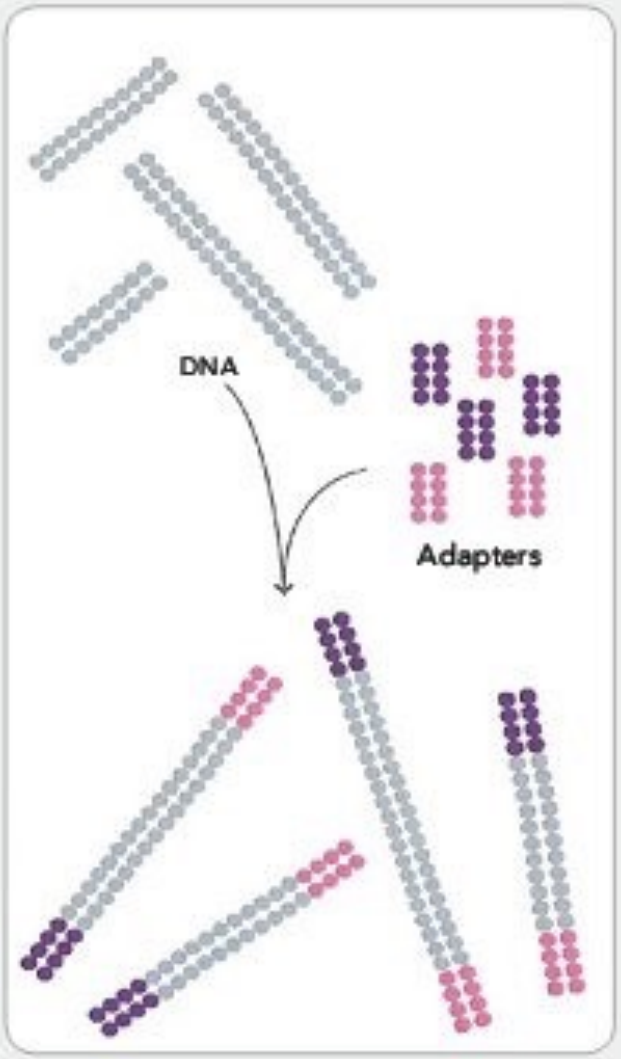
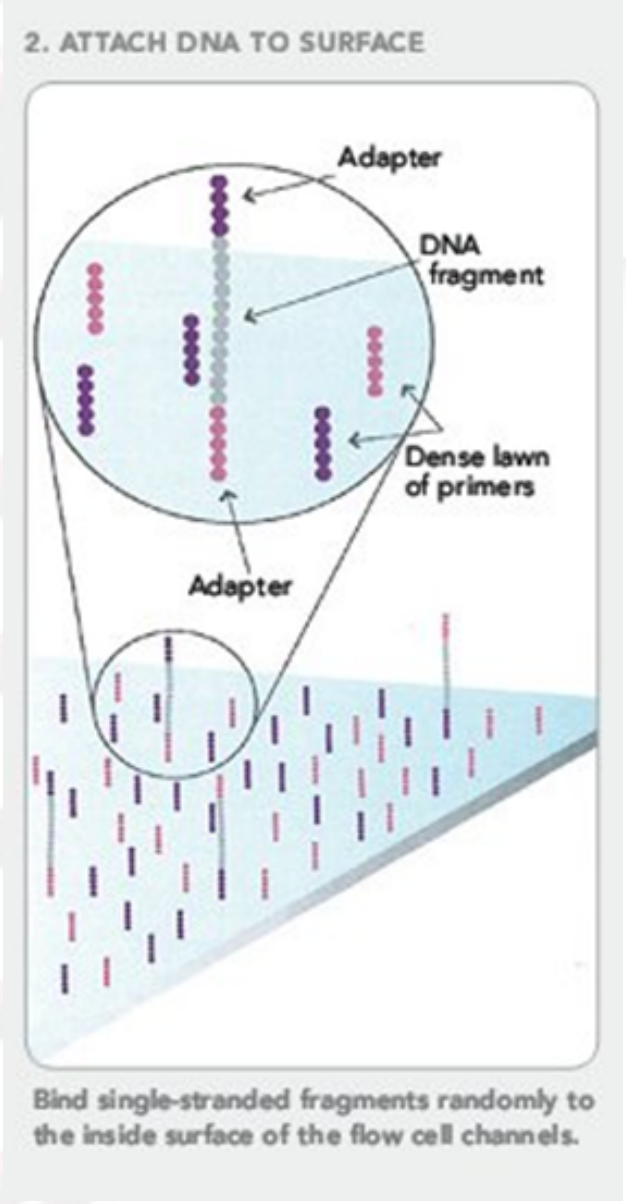


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

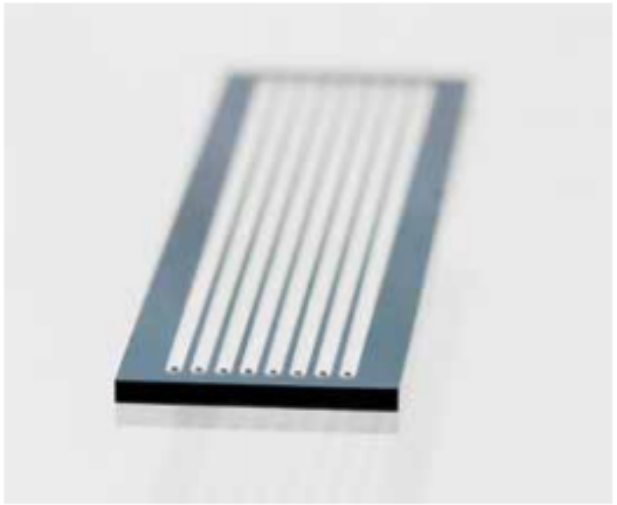
- Fragment DNA of interest into smaller strands that are able to be sequenced
  - Sonication
  - Nebulization
  - Enzyme digestion
- Ligate Adapters
- Denature dsDNA into ssDNA by heating to 95° C

Attach DNA to Surface



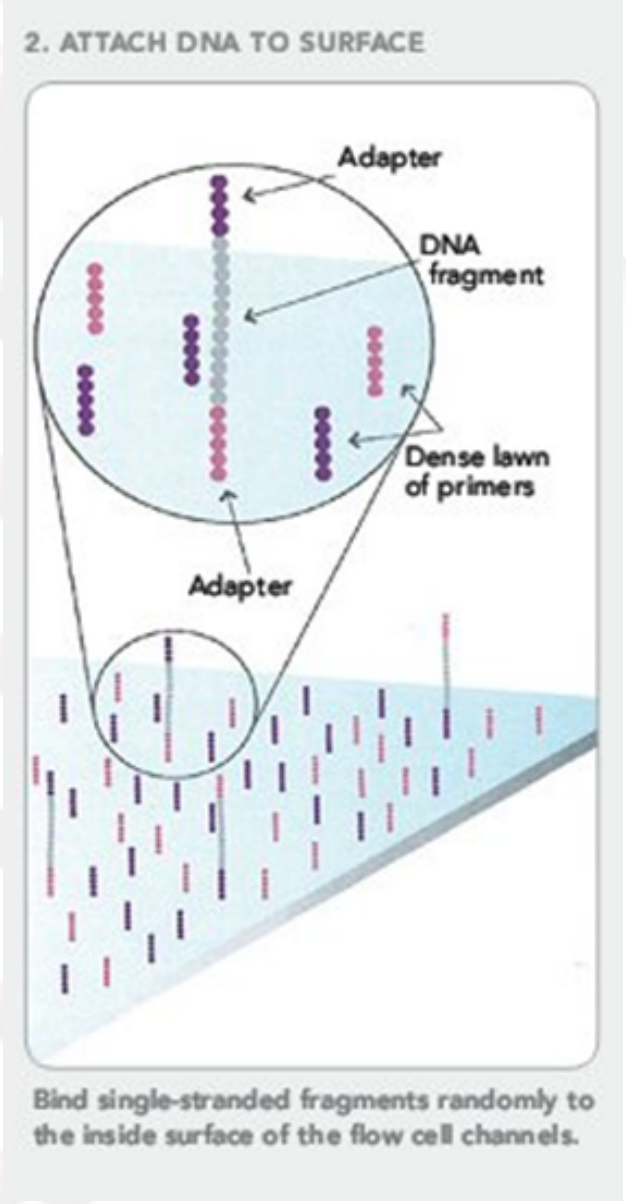
- ssDNA is then bound to inside surface of flow cell channels
- Dense lawn of primer on the surface of the flow cell

Flow Cell



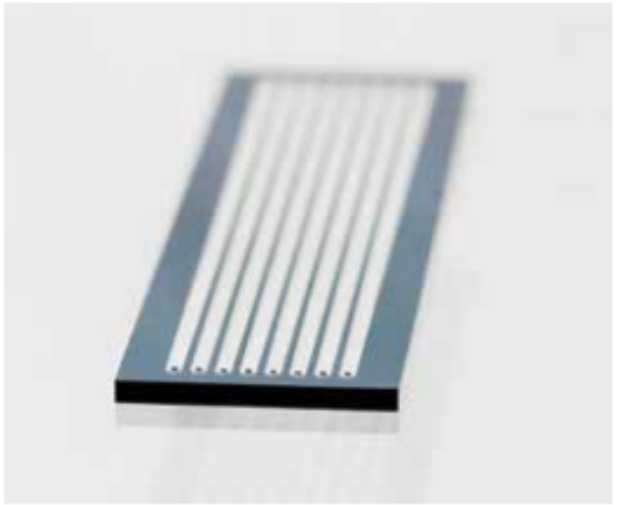
Images retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Attach DNA to Surface



- ssDNA is then bound to inside surface of flow cell channels
- Dense lawn of primer on the surface of the flow cell

Flow Cell



Images retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

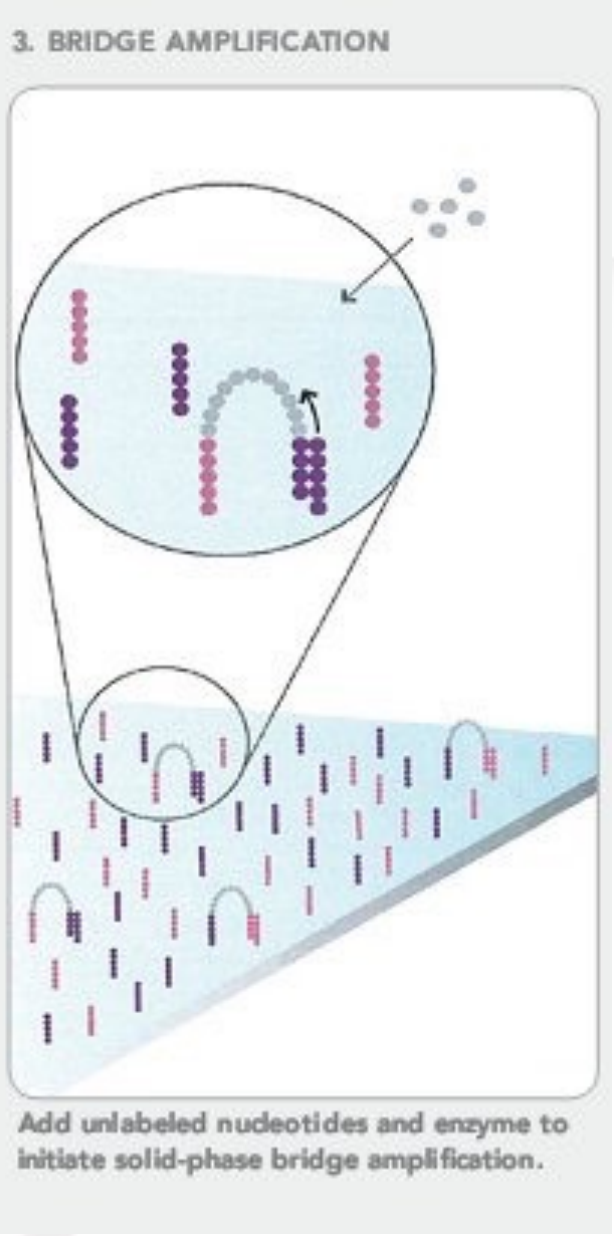


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Bridge Amplification

- Unlabeled nucleotides and polymerase enzyme are added to initiate the solid phase bridge amplification

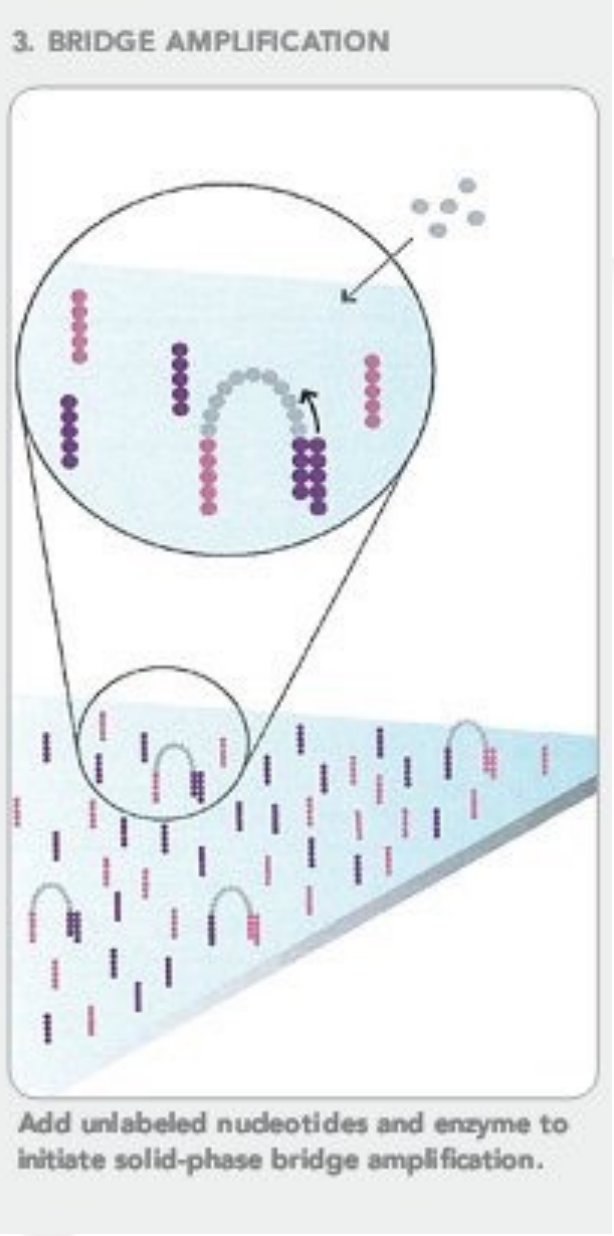


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Bridge Amplification

- Unlabeled nucleotides and polymerase enzyme are added to initiate the solid phase bridge amplification

Fragments Become Double Stranded

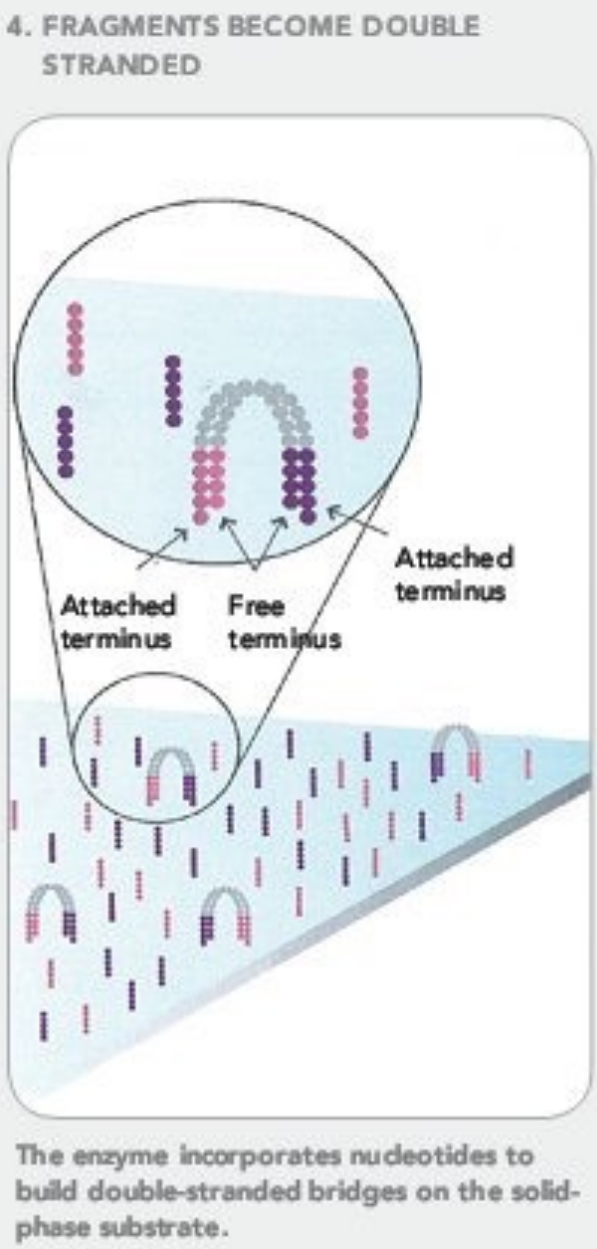


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- In this step it demonstrates the work done by the sequencing reagents
  - Primers
  - Nucleotides
  - Polymerase enzymes
  - Buffer

Fragments Become Double Stranded

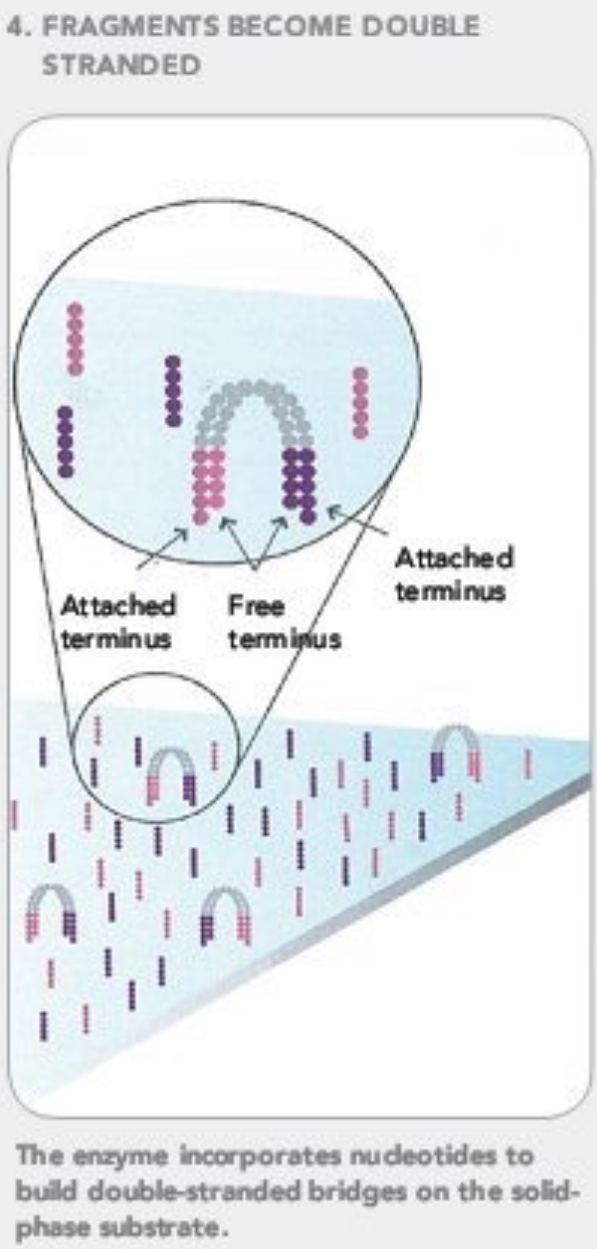
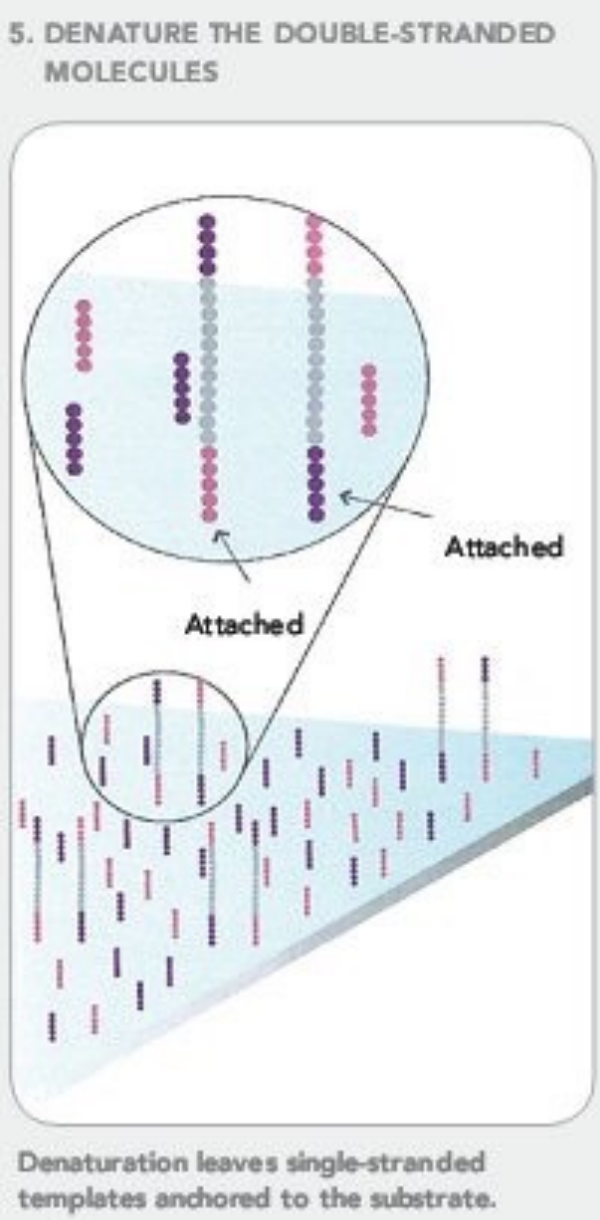


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- In this step it demonstrates the work done by the sequencing reagents
  - Primers
  - Nucleotides
  - Polymerase enzymes
  - Buffer

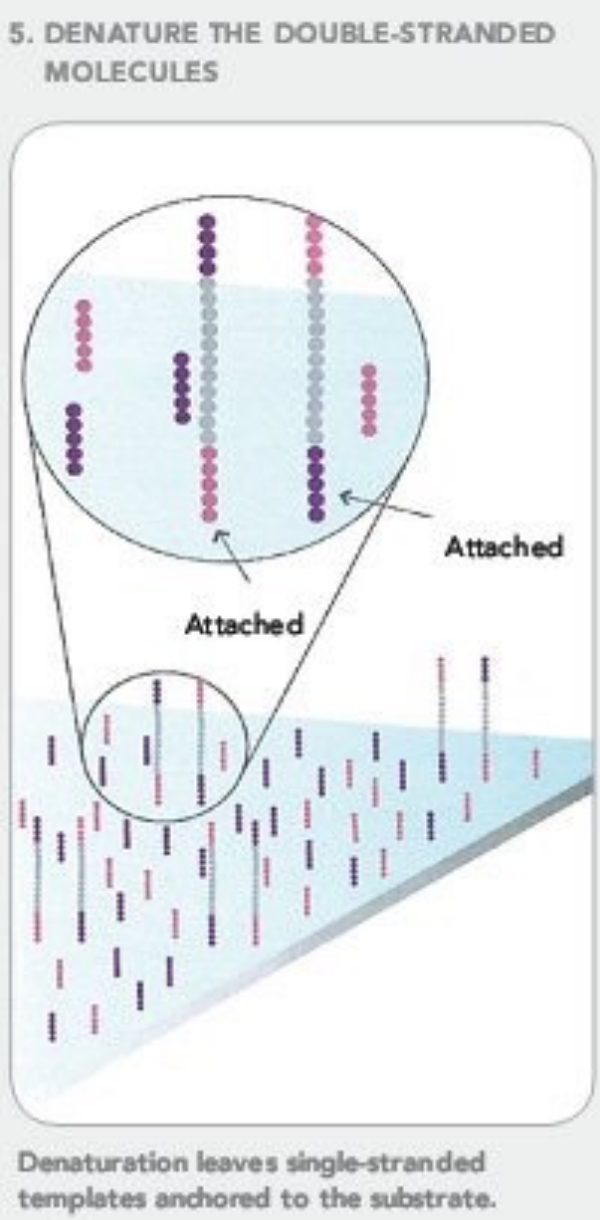
Denature the Double Stranded Molecules



- The original strand is then washed away, leaving only the strands that had been synthesized to the oligos attached to the flow cell

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Denature the Double Stranded Molecules



- The original strand is then washed away, leaving only the strands that had been synthesized to the oligos attached to the flow cell

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Steps 5-7 Repeats

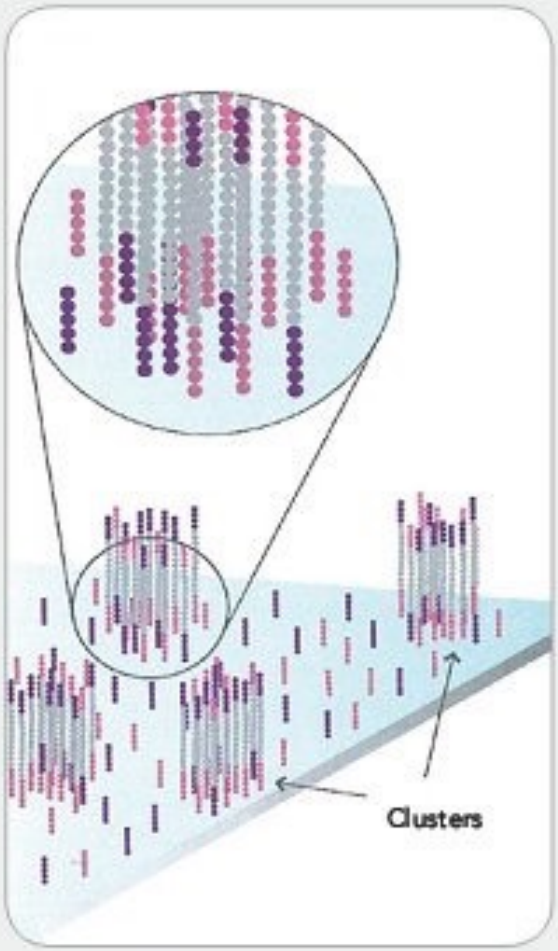
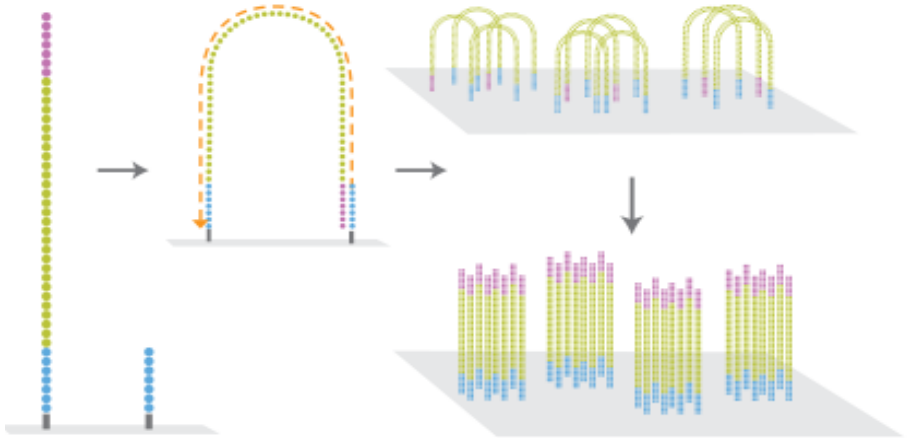


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- Cycle of new strand synthesis and Denaturation to make multiple copies of the same sequence (amplification)
  - Fragments Become Double Stranded
  - Denature the Double Strand Molecules



Steps 5-7 Repeats

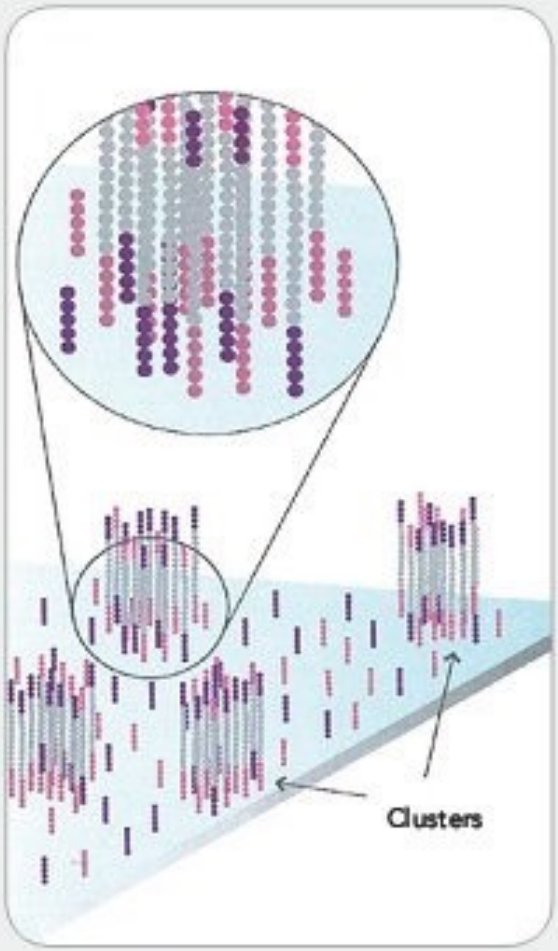
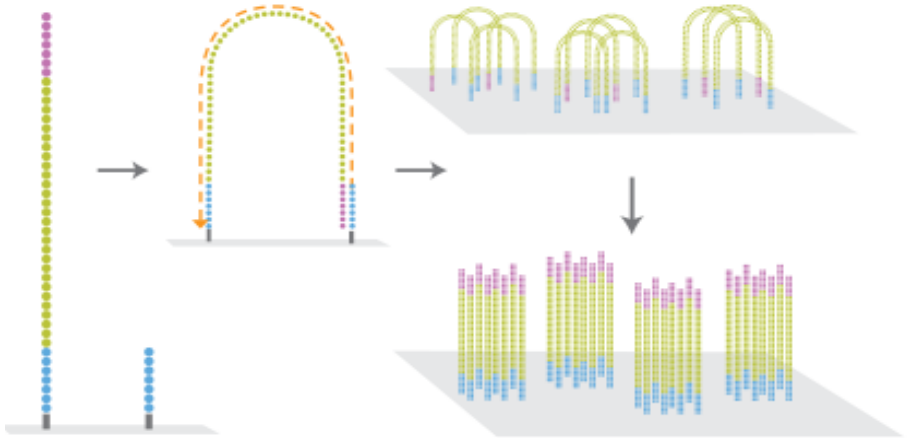
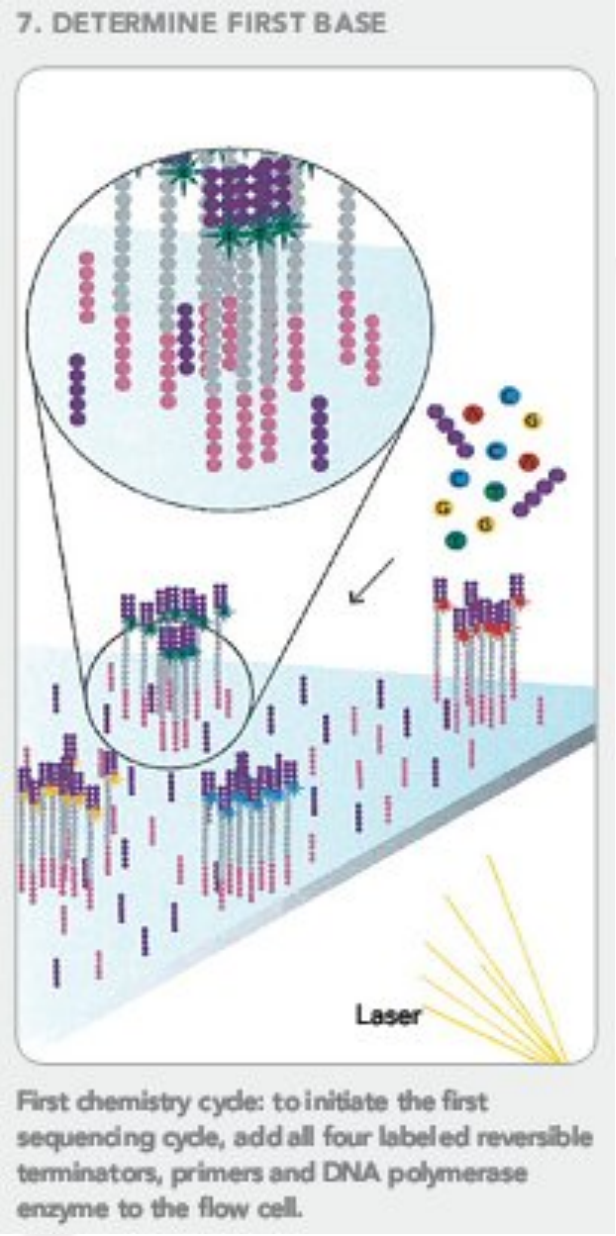


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- Cycle of new strand synthesis and Denaturation to make multiple copies of the same sequence (amplification)
  - Fragments Become Double Stranded
  - Denature the Double Strand Molecules



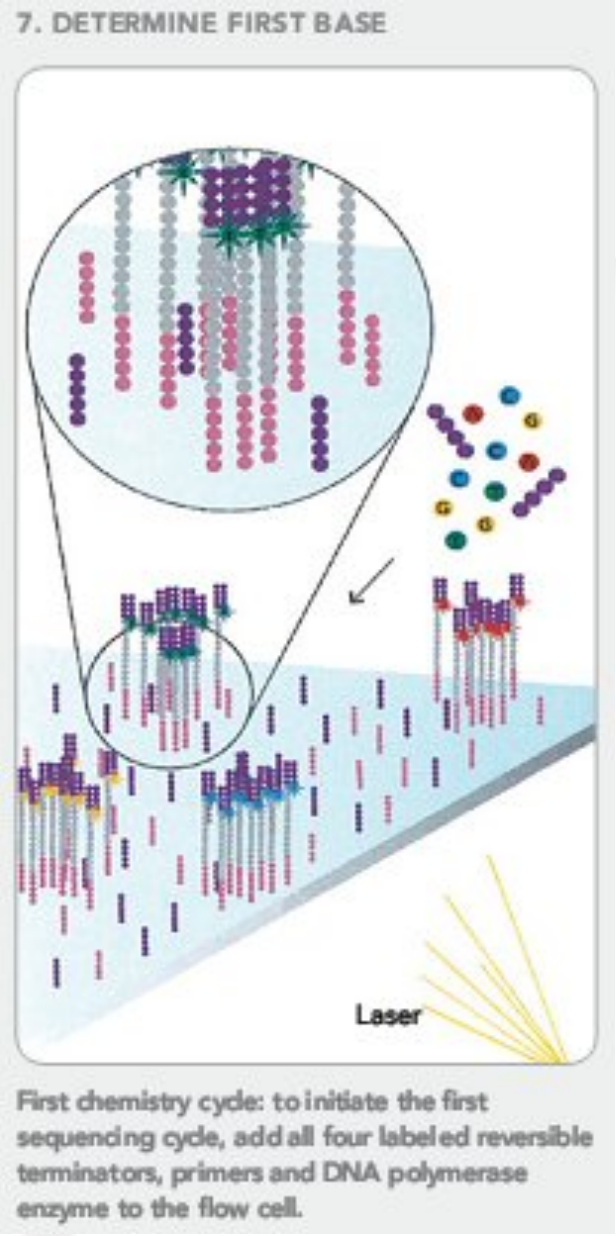
Determine First Base



- The P5 region is cleaved
- Add sequencing reagents
  - Primers
  - Polymerase
  - Fluorescently labeled nucleotides
  - Buffer
- First base incorporated

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Determine First Base



- The P5 region is cleaved
- Add sequencing reagents
  - Primers
  - Polymerase
  - Fluorescently labeled nucleotides
  - Buffer
- First base incorporated

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

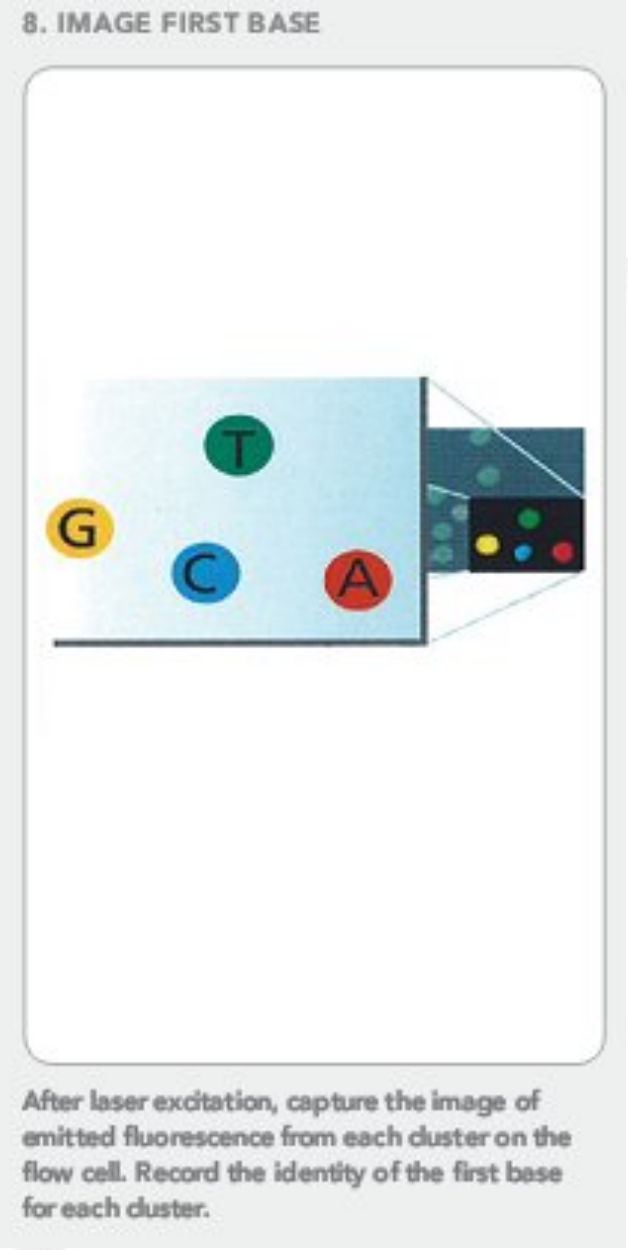


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Image First Base

- Remove unincorporated bases
- Detect Signal
- Deblock and remove the fluorescent signal →new cycle

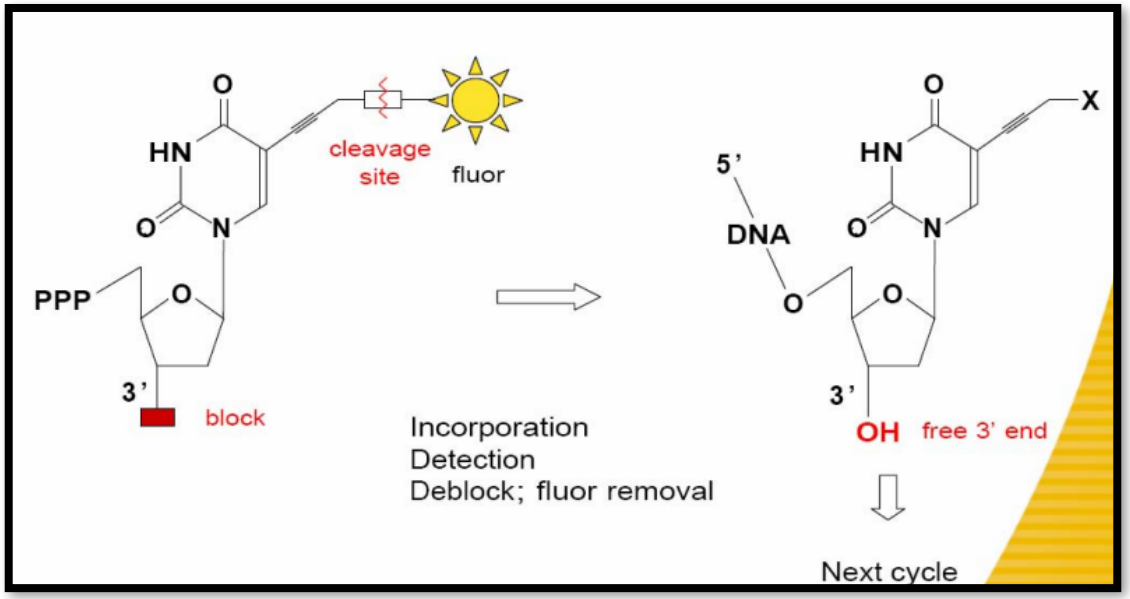


Image retrieved from http://research.stowers-institute.org/microscopy/external/PowerpointPresentations/ppt/Methods_Technology/KSH_Tech&Methods_012808Final.pdf

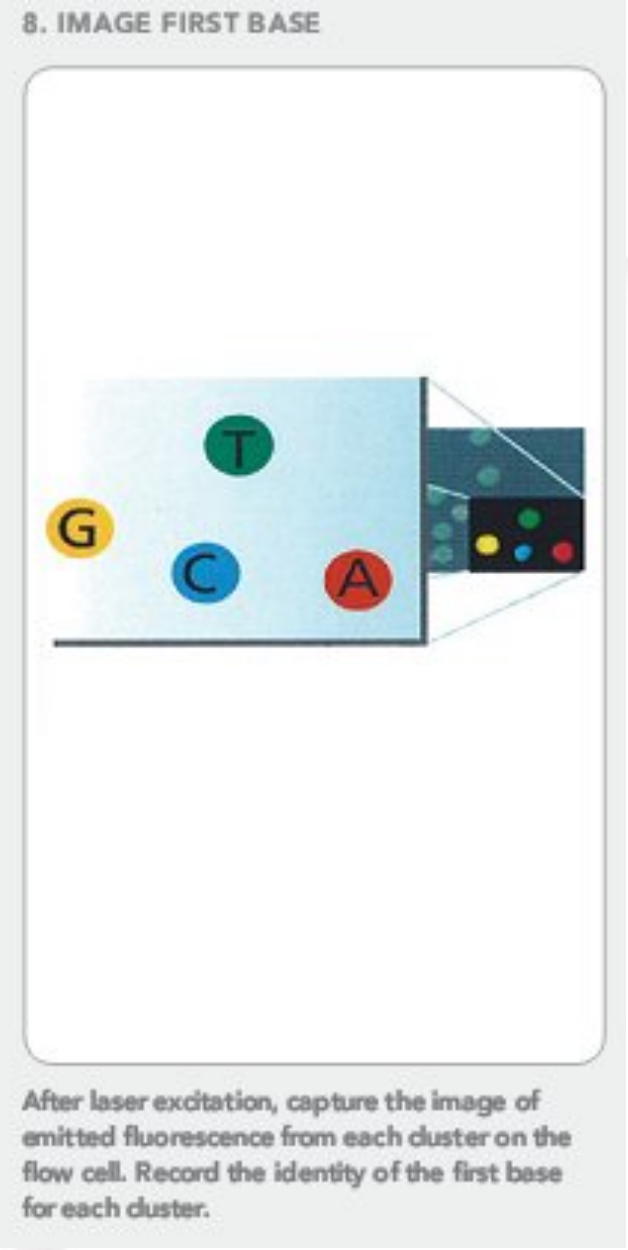


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Image First Base

- Remove unincorporated bases
- Detect Signal
- Deblock and remove the fluorescent signal →new cycle

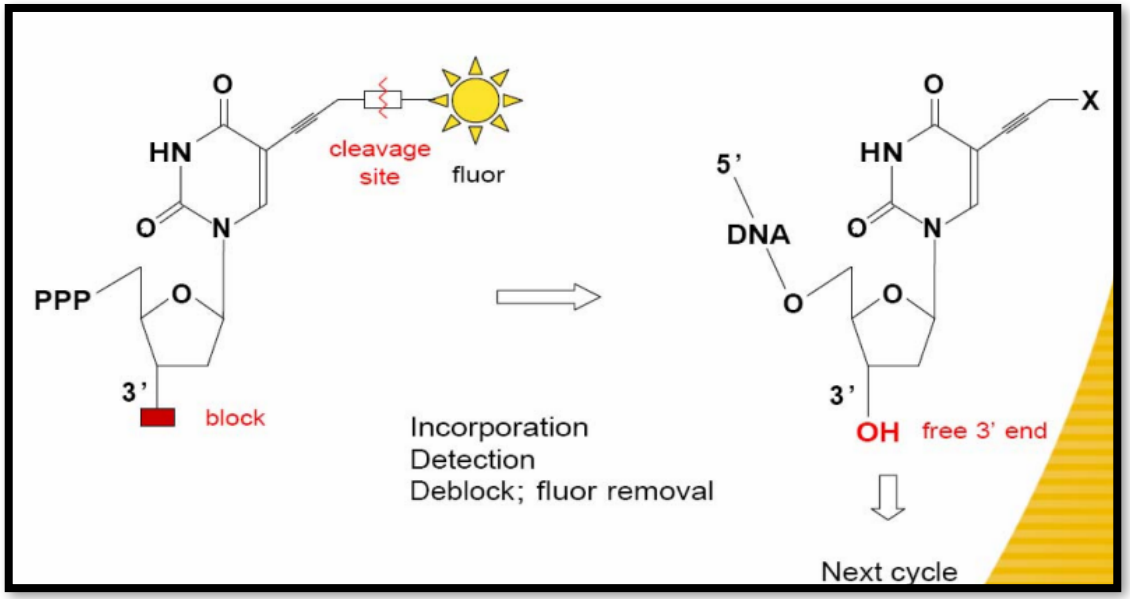
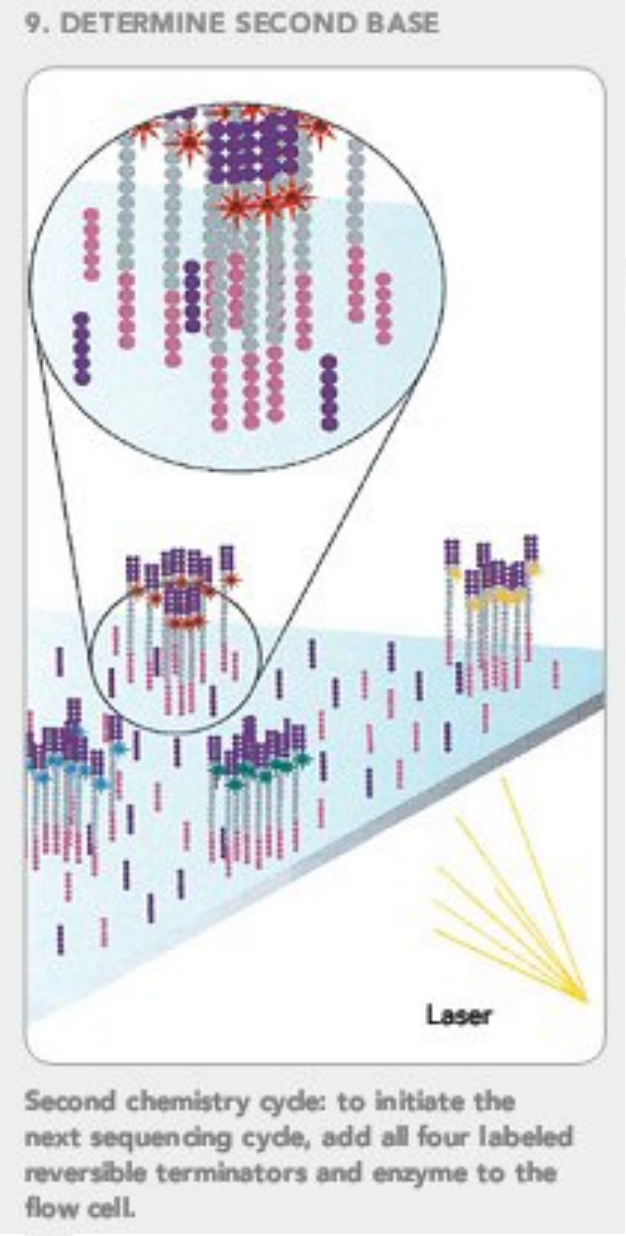


Image retrieved from http://research.stowers-institute.org/microscopy/external/PowerpointPresentations/ppt/Methods_Technology/KSH_Tech&Methods_012808Final.pdf

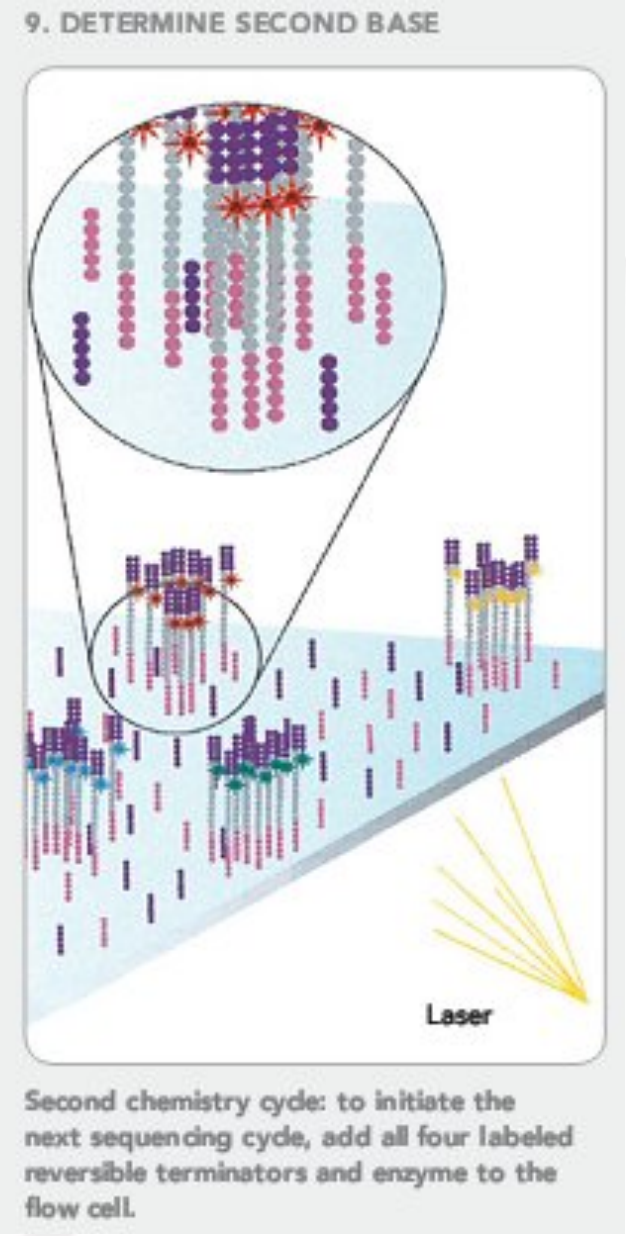
Determine Second Base



- Add sequencing reagents
  - Primers
  - Polymerase
  - Fluorescently labeled nucleotides
  - Buffer
- Second base incorporated

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

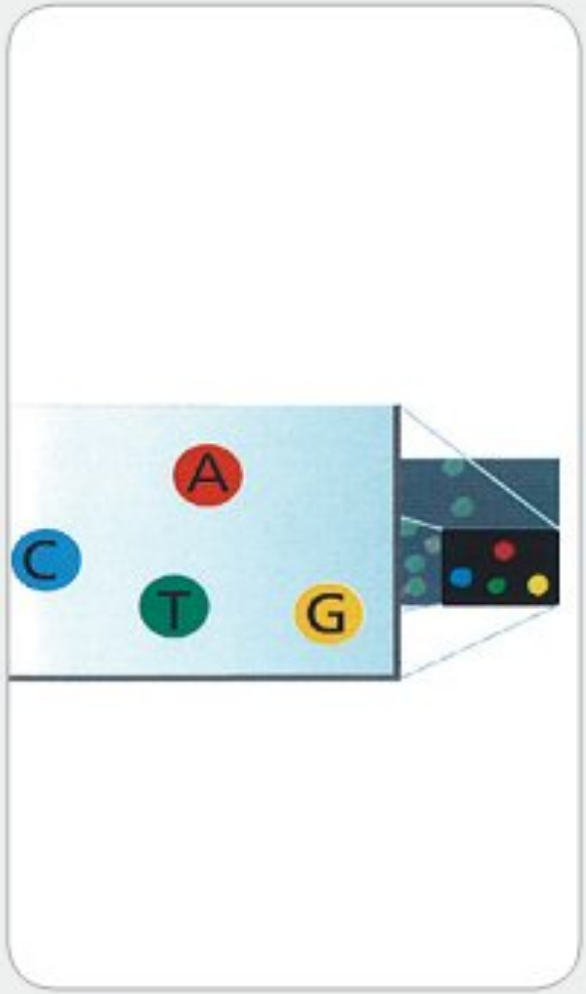
Determine Second Base



- Add sequencing reagents
  - Primers
  - Polymerase
  - Fluorescently labeled nucleotides
  - Buffer
- Second base incorporated

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

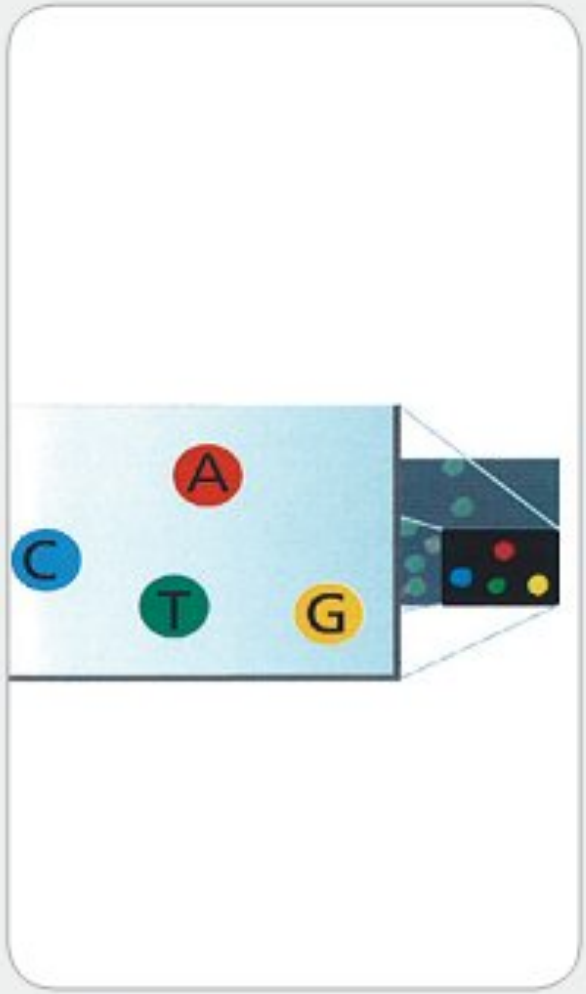
Image Second Chemistry Cycle



- Remove unincorporated bases
- Detect Signal
- Deblock and remove the fluorescent signal → new cycle

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Image Second Chemistry Cycle



- Remove unincorporated bases
- Detect Signal
- Deblock and remove the fluorescent signal → new cycle

Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Sequence Reads Over Multiple Chemistry Cycles

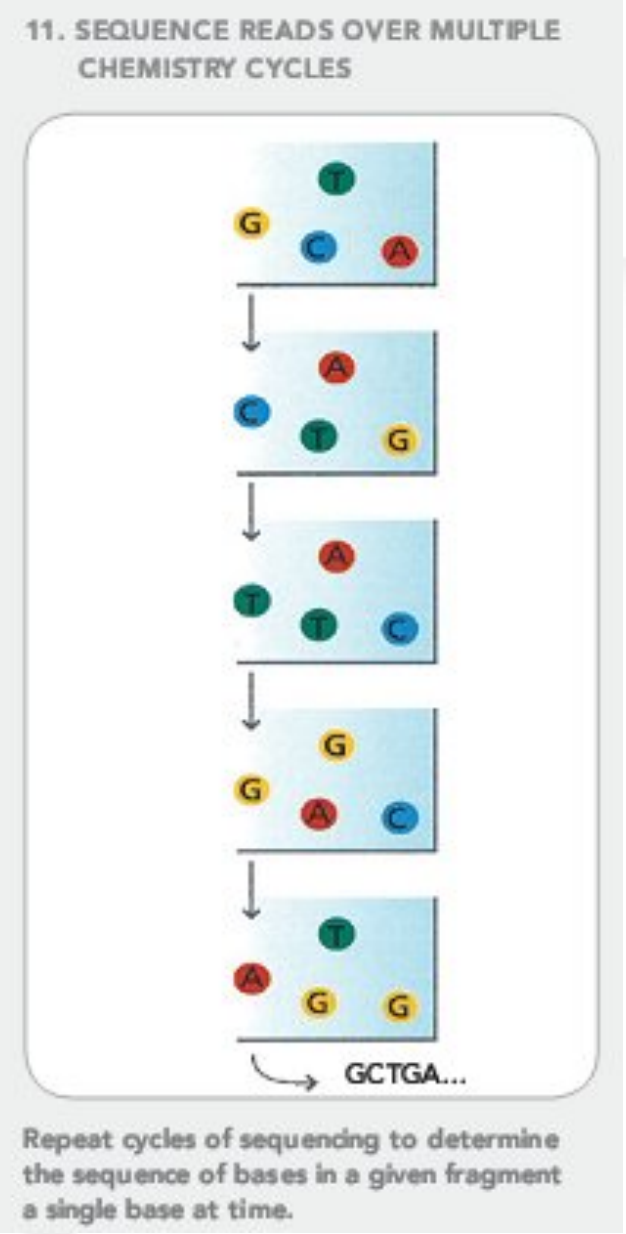


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- The identity of each base of a cluster is read off from sequential images

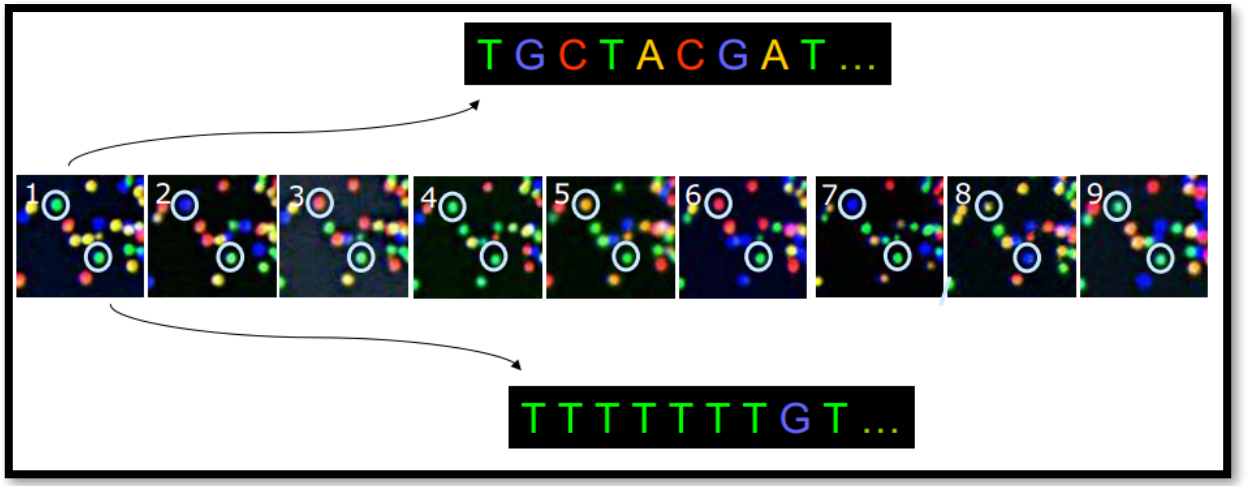


Image retrieved from http://research.stowers-institute.org/microscopy/external/PowerpointPresentations/ppt/Methods_Technology/KSH_Tech&Methods_012808Final.pdf

Sequence Reads Over Multiple Chemistry Cycles

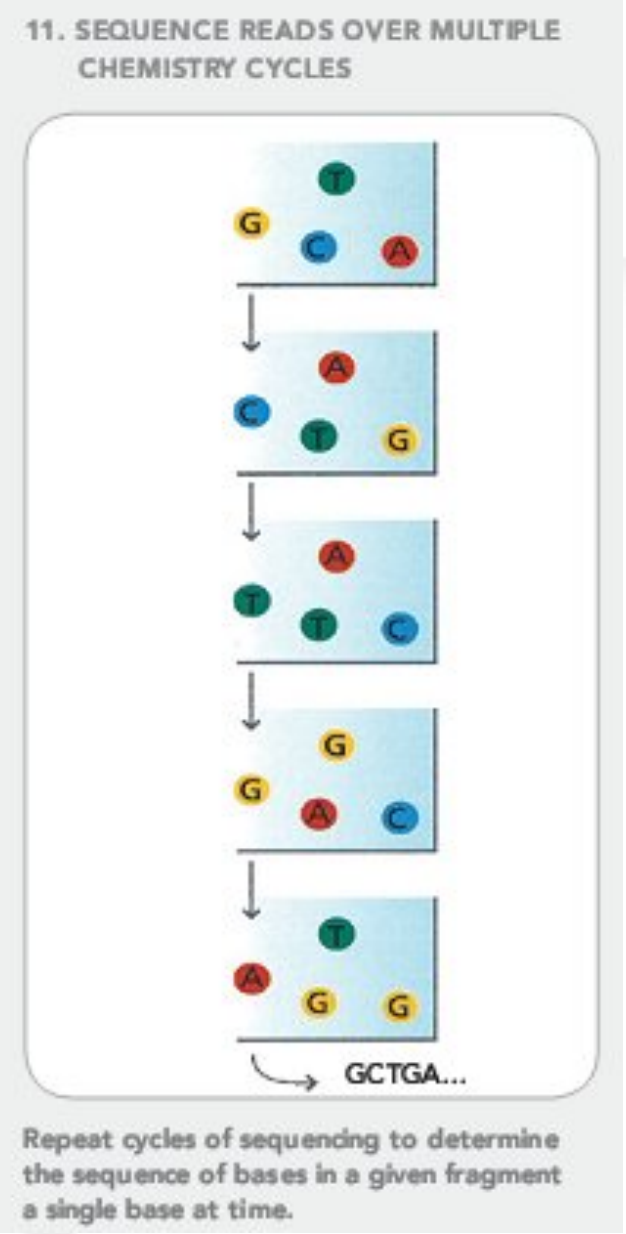


Image retrieved from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

- The identity of each base of a cluster is read off from sequential images

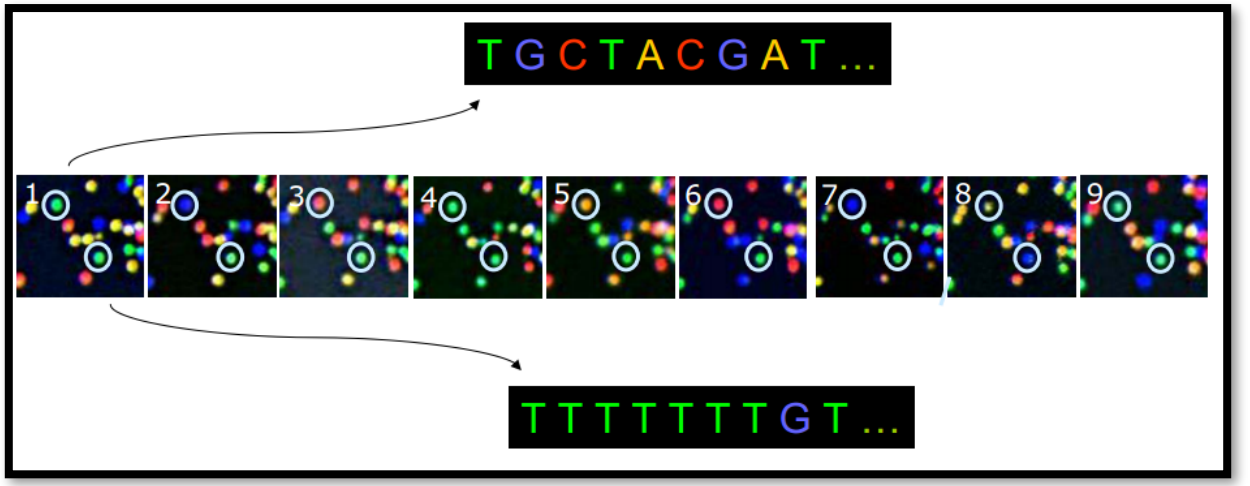
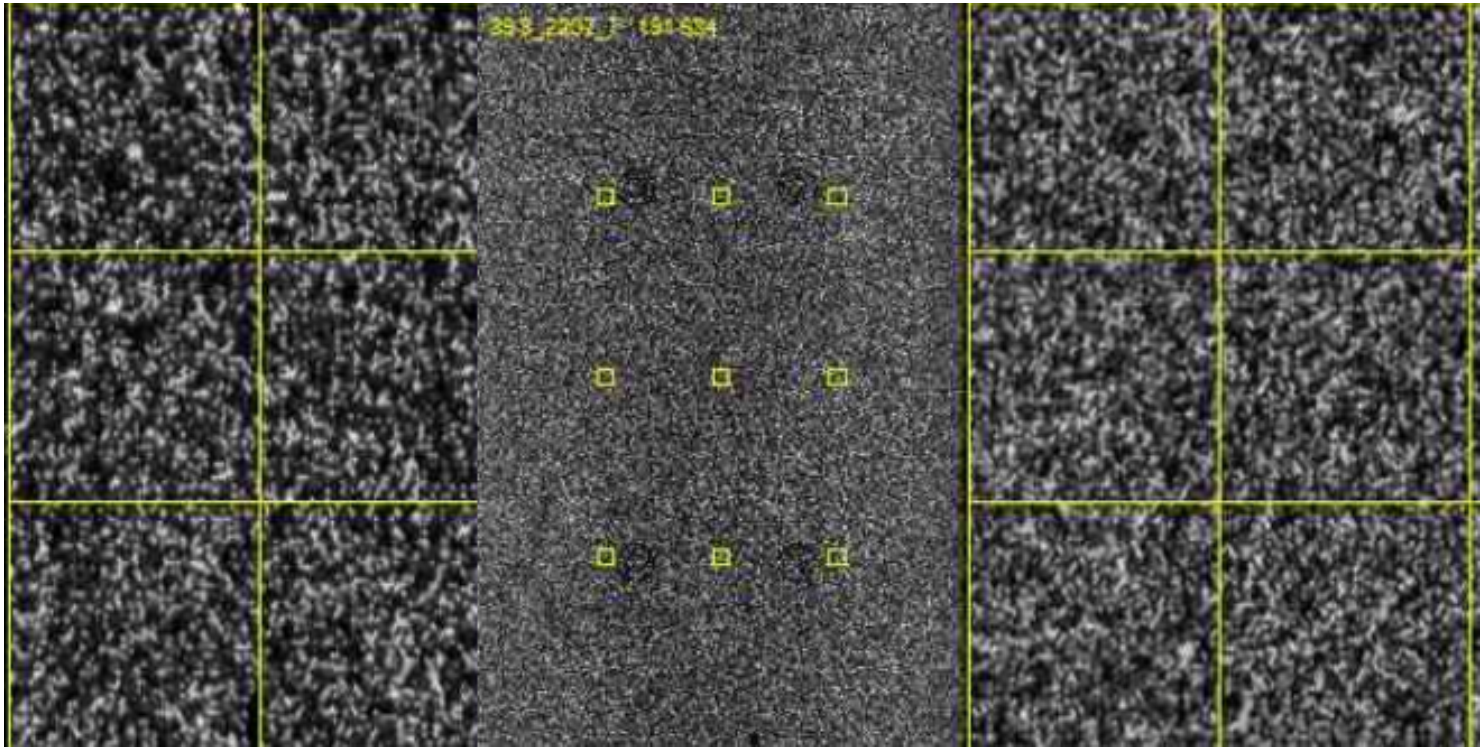


Image retrieved from http://research.stowers-institute.org/microscopy/external/PowerpointPresentations/ppt/Methods_Technology/KSH_Tech&Methods_012808Final.pdf

Sequencing by synthesis



Actual Illumina HiSeq 3000 image

http://dnatech.genomecenter.ucdavis.edu/2015/05/07/first-hiseq-3000-data-download/

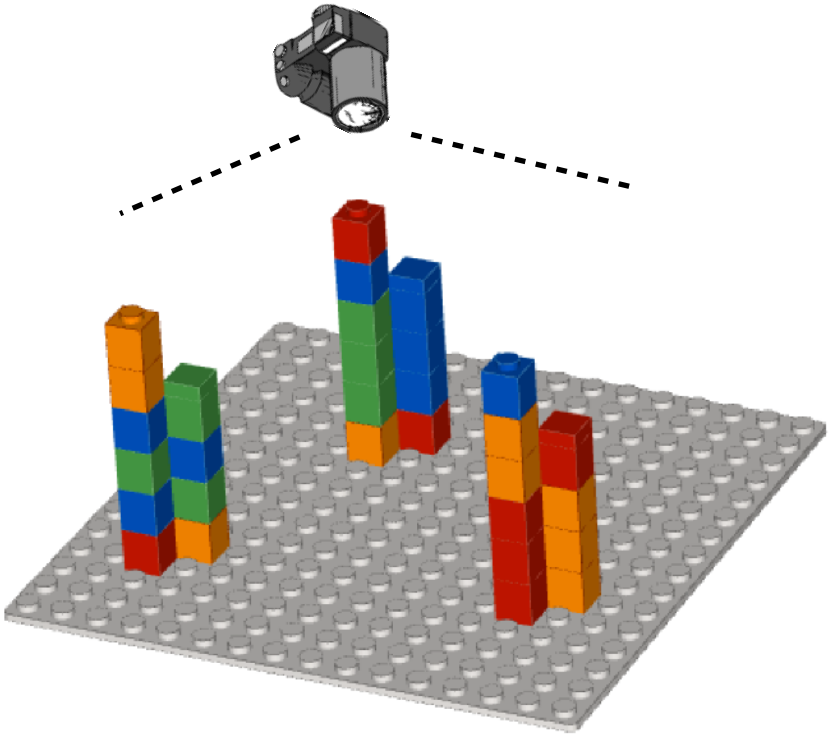
This and following slides marked with * courtesy of Ben Langmead

Sequencing by synthesis

Billions of templates on a slide

Massively parallel: photograph captures all templates simultaneously

Terminators are “speed bumps,” keeping reactions in sync



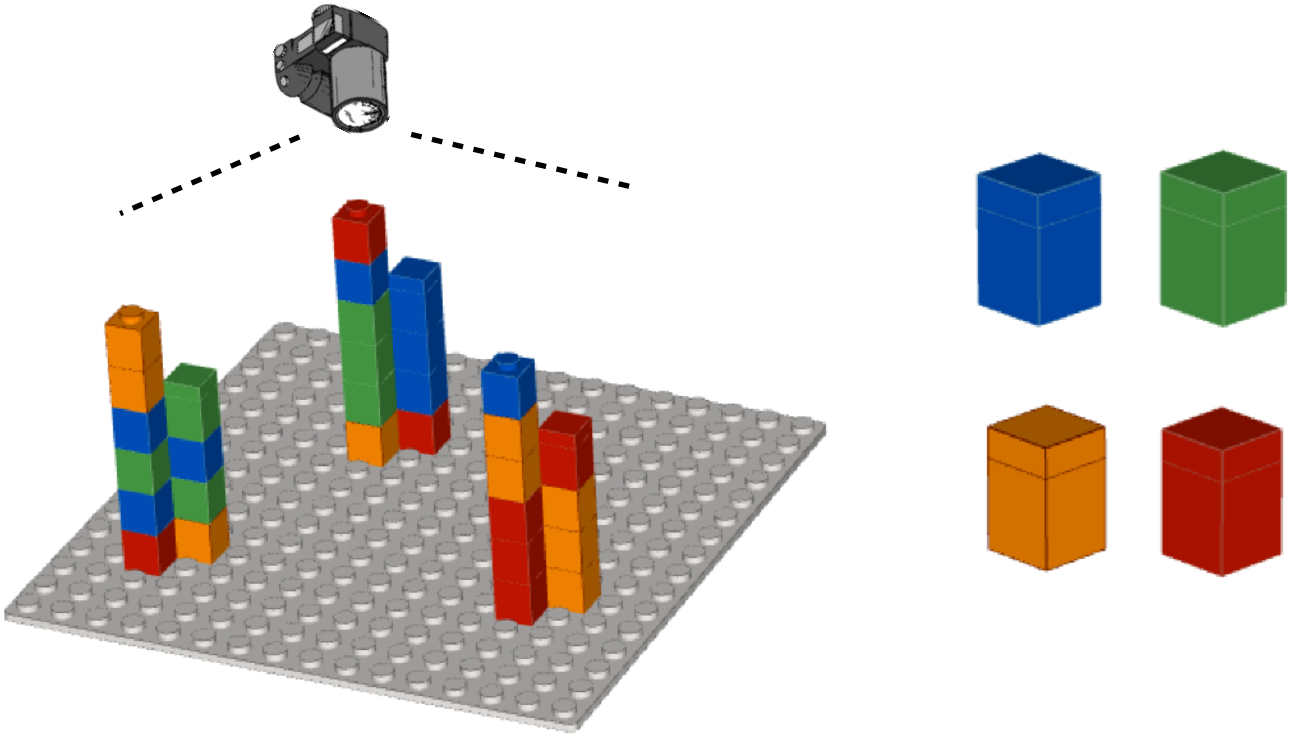
*

Sequencing by synthesis

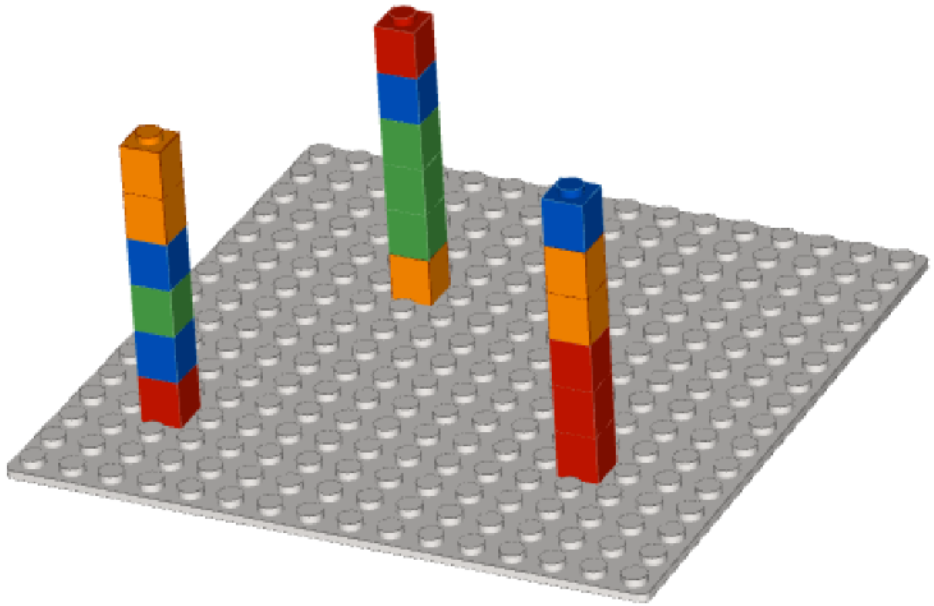
Billions of templates on a slide

Massively parallel: photograph captures all templates simultaneously

Terminators are “speed bumps,” keeping reactions in sync

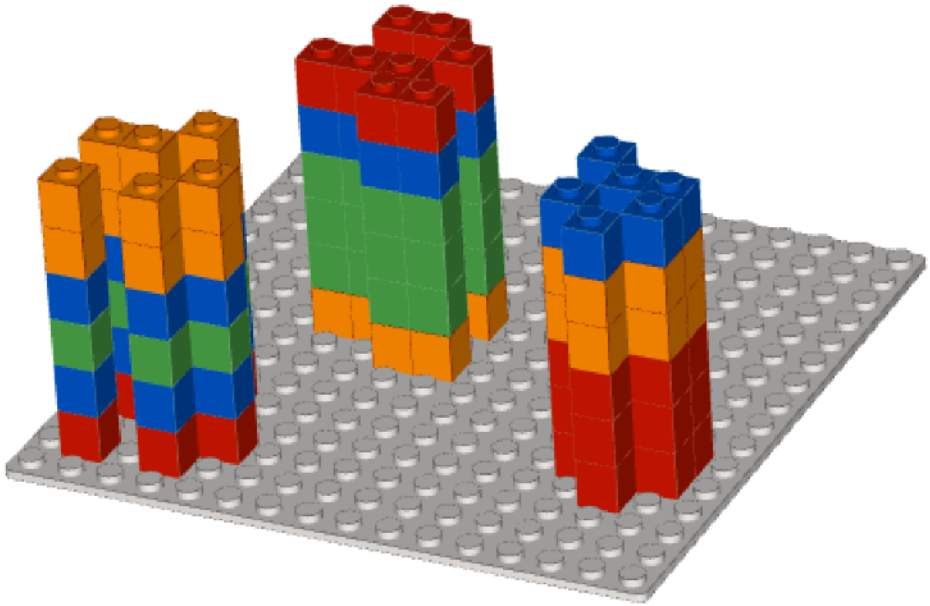


*

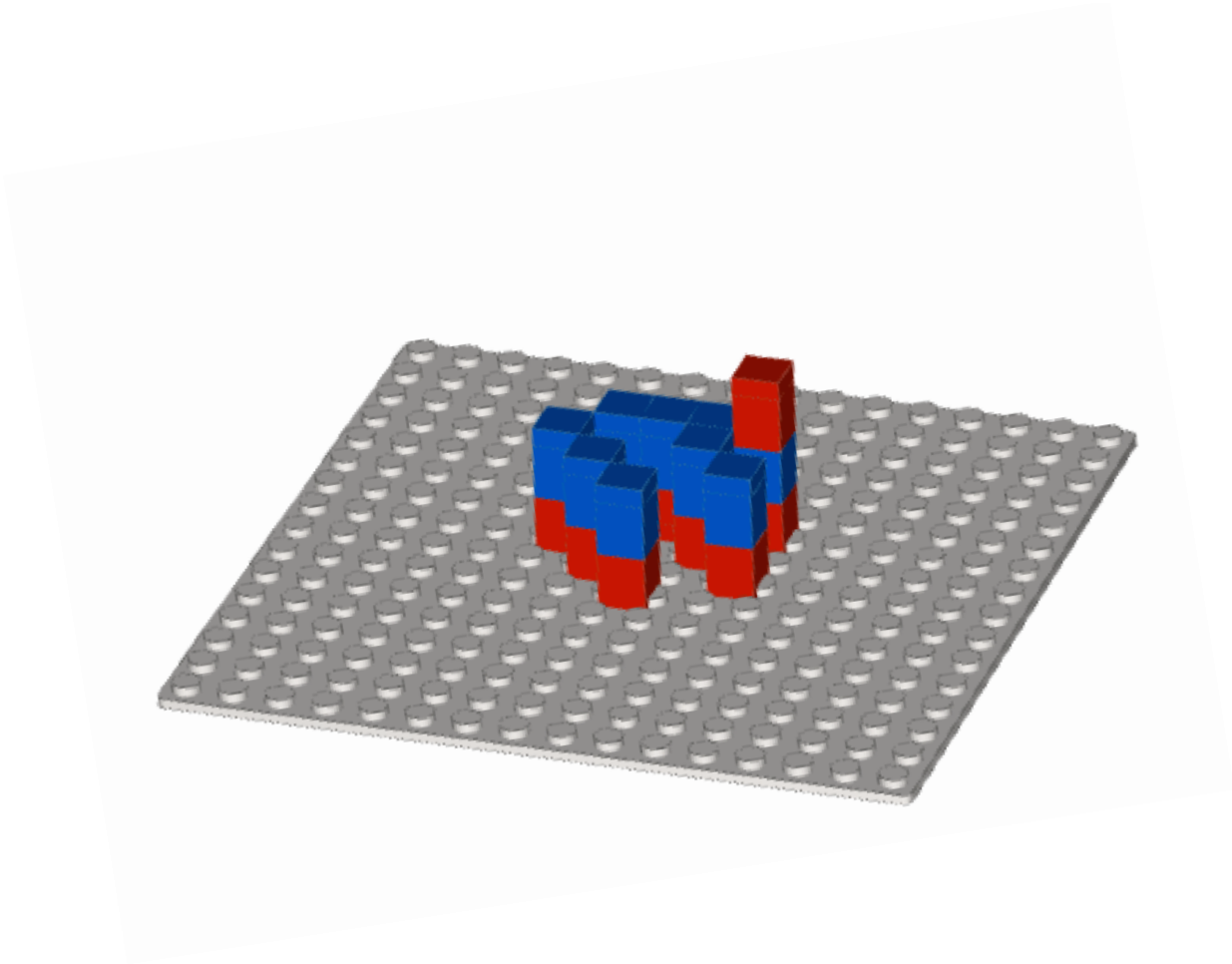


*

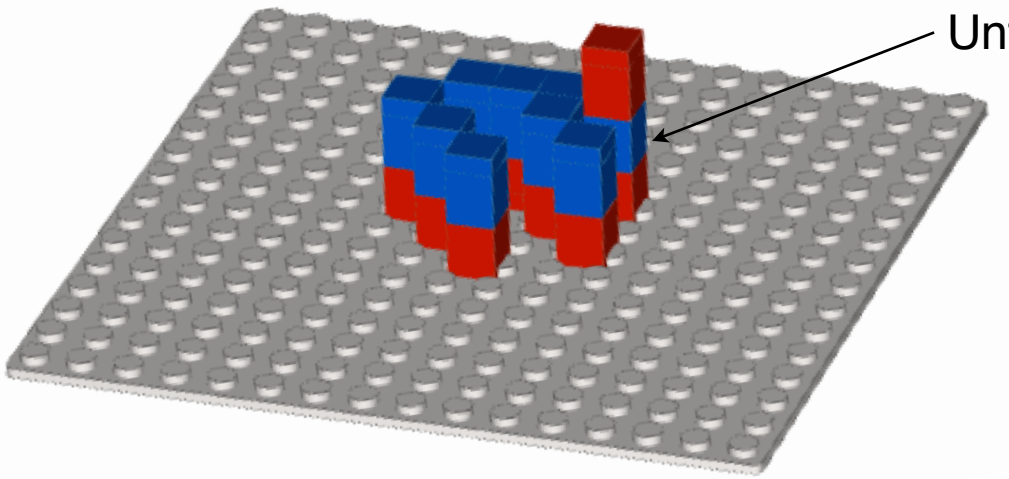
Cluster of clones



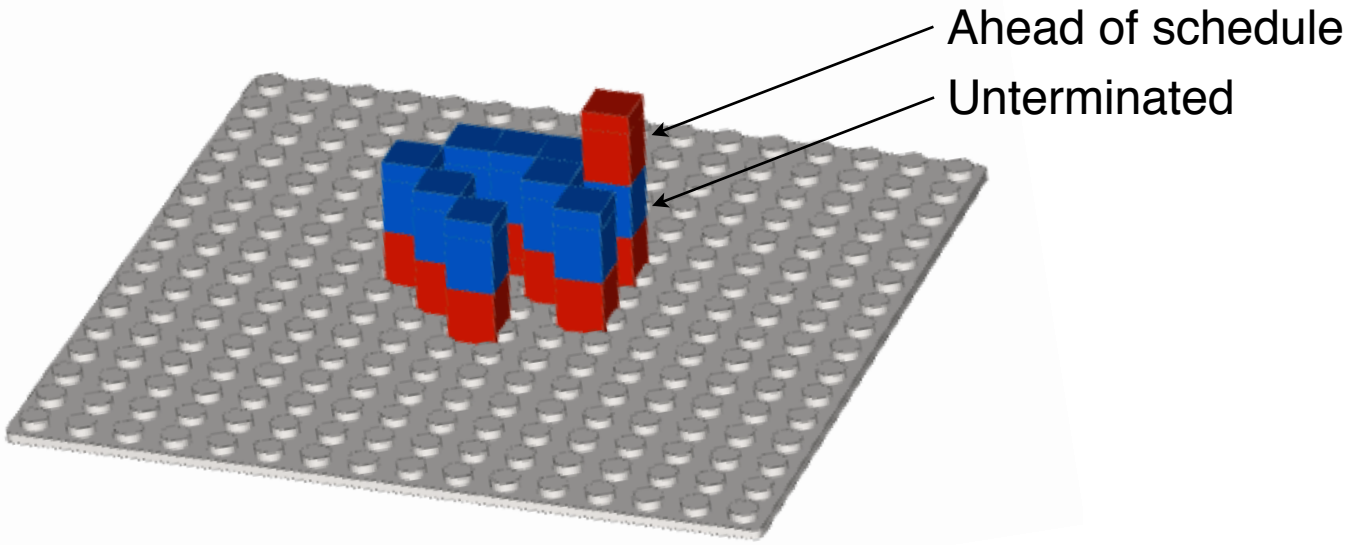
*



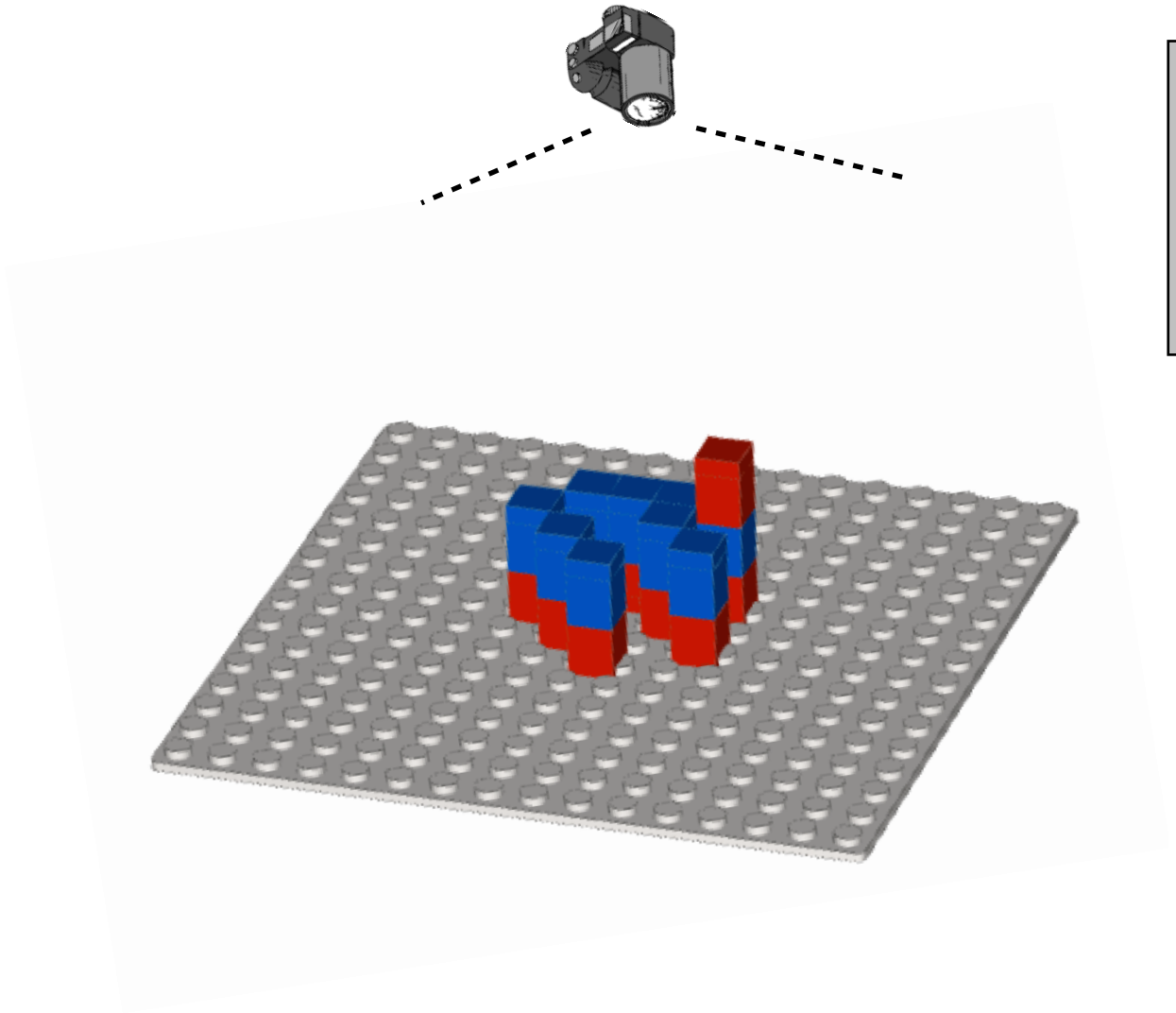
*

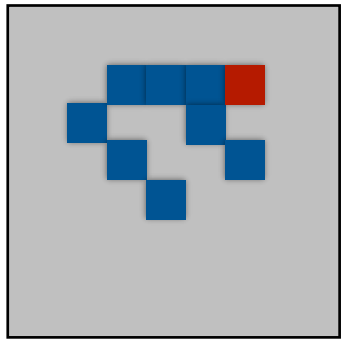

Unterminated

*

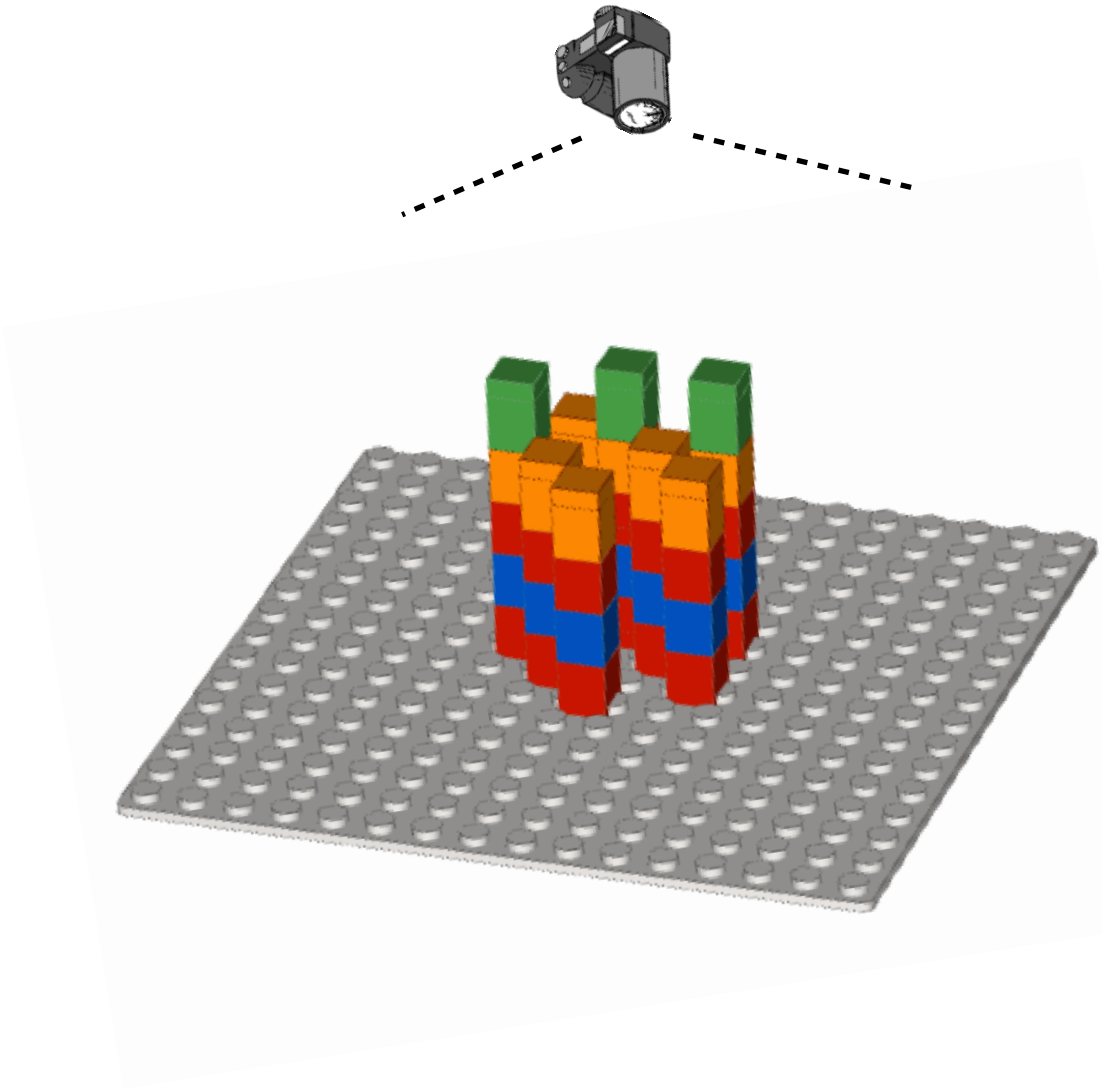

Ahead of schedule
Unterminated

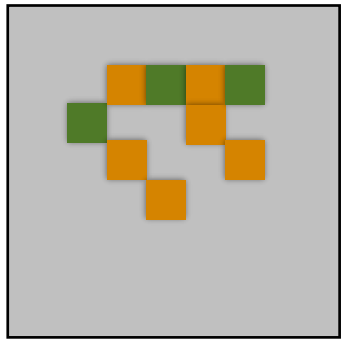
*





*





*

$$Q = -10 \cdot \log_{10} p$$

*

$$Q = -10 \cdot \log_{10} p$$

Base quality

*

$$Q = -10 \cdot \log_{10} p$$

Base quality

Probability that base call is incorrect

*

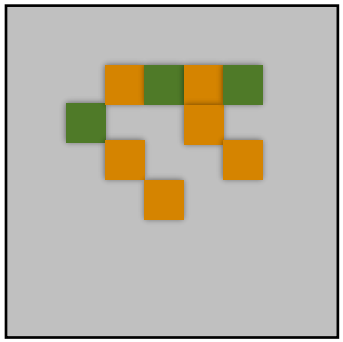
$$Q = -10 \cdot \log_{10} p$$

Base quality

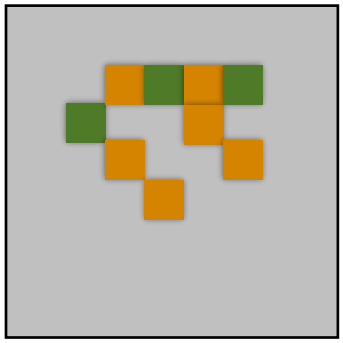
Probability that base call is incorrect

Q = 10 → 1 in 10 chance call is incorrect
Q = 20 → 1 in 100
Q = 30 → 1 in 1,000

*

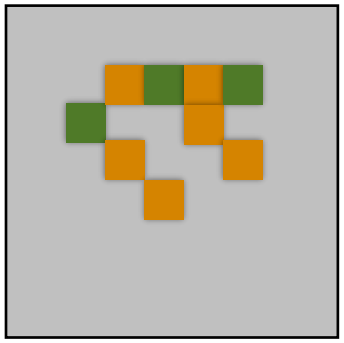


*



Call: orange (C)

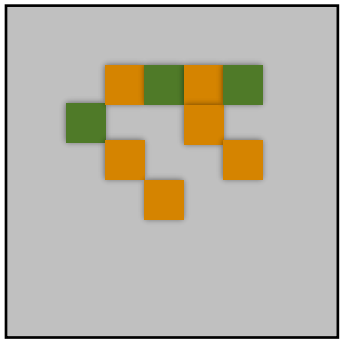
*



Call: orange (C)

Estimate p, probability incorrect:

*

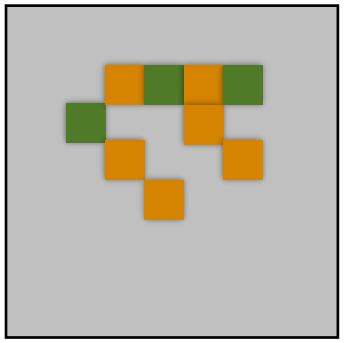


Call: orange (C)

Estimate p, probability incorrect:

non-orange light / total light

*



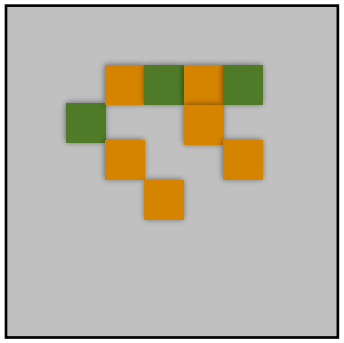
Call: orange (C)

Estimate p, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

*



Call: orange (C)

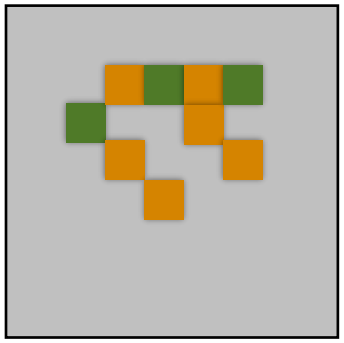
Estimate p, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

$Q = -10 \log_{10} 1/3$

*



Call: orange (C)

Estimate p, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

$Q = -10 \log_{10} 1/3 = 4.77$

*

A read in FASTQ format

```
@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1
ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
+
?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G
```

*

A read in FASTQ format

Name `@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1`

```
ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
+
?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G
```

*

A read in FASTQ format

Name `@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1`

Sequence `ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT`

```
+
?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G
```

*

A read in FASTQ format

Name `@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1`

Sequence `ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT`

(ignore) `+`

`?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G`

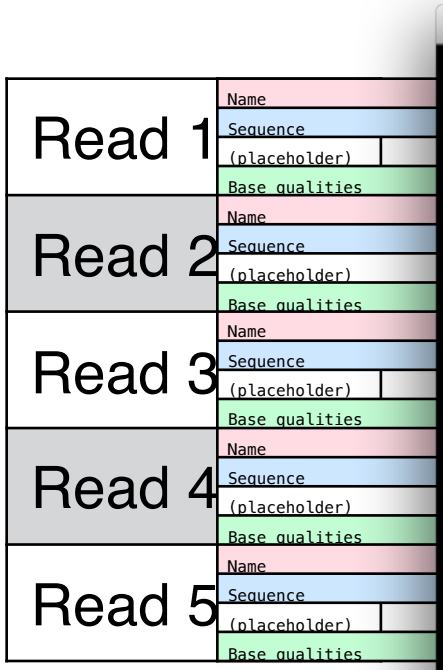
*

A read in FASTQ format

Name	`@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1`
Sequence	`ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT`
(ignore)	`+`
Base qualities	`?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G`

*

FASTQ


Read 1
Name
Sequence
(placeholder)
Base qualities
Read 2
Name
Sequence
(placeholder)
Base qualities
Read 3
Name
Sequence
(placeholder)
Base qualities
Read 4
Name
Sequence
(placeholder)
Base qualities
Read 5
Name
Sequence
(placeholder)
Base qualities


reads — Example — bash — 104×25
$ head -20 SRA_HISEQ2000_FC1.shuffle.2M.1.fastq
@509.6.64.20524.149722
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGAGCCTTCTCTCCACCCTGAAAATAGCTTCTGGCTGNTGGGTGAACTATGGAGAGAAAGCGTTTTATTAT
+
HHHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHIHFHHFHHHIHGEHHIIFIHBC#@:@9,--541436D9?;E###############
@509.4.62.19231.2763
GTTGATAAGCAAGCATCTCATTTTGTGCATATACCTGGTCTTTCGTATTCTGGCGTGAAGTCGCCGNCTGAATGCCAGCAATCTCTTTTTGAGTCTCATT
+
HHHHHHHHHHHHHHHEHHHHHHHHHHHHHHHHHHHHHHHHHDHHHHHHGHGHHHHHHHHHH=EF?DHE4#555=;===GGHEGGEGHG@C@<7<3@?F<A9@<
@509.6.47.3027.76579
CCTTTTCGACTAGAGACTGCCAAGTGCCAAAATATCCACTTGCAGATACTACAACAAGAGTGTTTCNAAACTGCTCAATCAAAAGAAATGTTCAACTCTT
+
HHHHHHHHHHHHHHHHHHHIHHHHHHHHHHHHHHHHHHHHGHHHHHHHHGHHHHHHHHHHEH?HH4#554DDADDHHHHHH@GHHFGBBFFHFHFHEHHH
@509.2.7.2951.186312
AAAGATACAACATACCACAATCTTTGAGACACACCTAAGACAATAAGGCAGTGTTAAGAGGAAAATTAATAGCACTAAATGCCCACATCAAAAAGTTAGA
+
HHHHHHHHHHHHHHHHHHHGHDHHHHHFHEHHHGHHGHHHHHHHHGHHHHHHEHEF<?<@=BBFFFGCFFE?<;@AFG=GA;@D@D?FDFFB=B;F=>AA@
@509.6.25.8102.140546
GGACACATTCAAACCATTGCATCCATCCTCTGCATTCAGAAAGATAGTCCAACAGAAAGATCTGGANTCAAGAGACCCAGCTGATTACCAATTCCAGTTT
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHIHHHIHHHIHG#FFDCDD@@GGGHHFIHEGIFIEIIIIGIIGFGF
$

*

Base qualities

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
||||||||||||||||||||||||||||||||
HHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

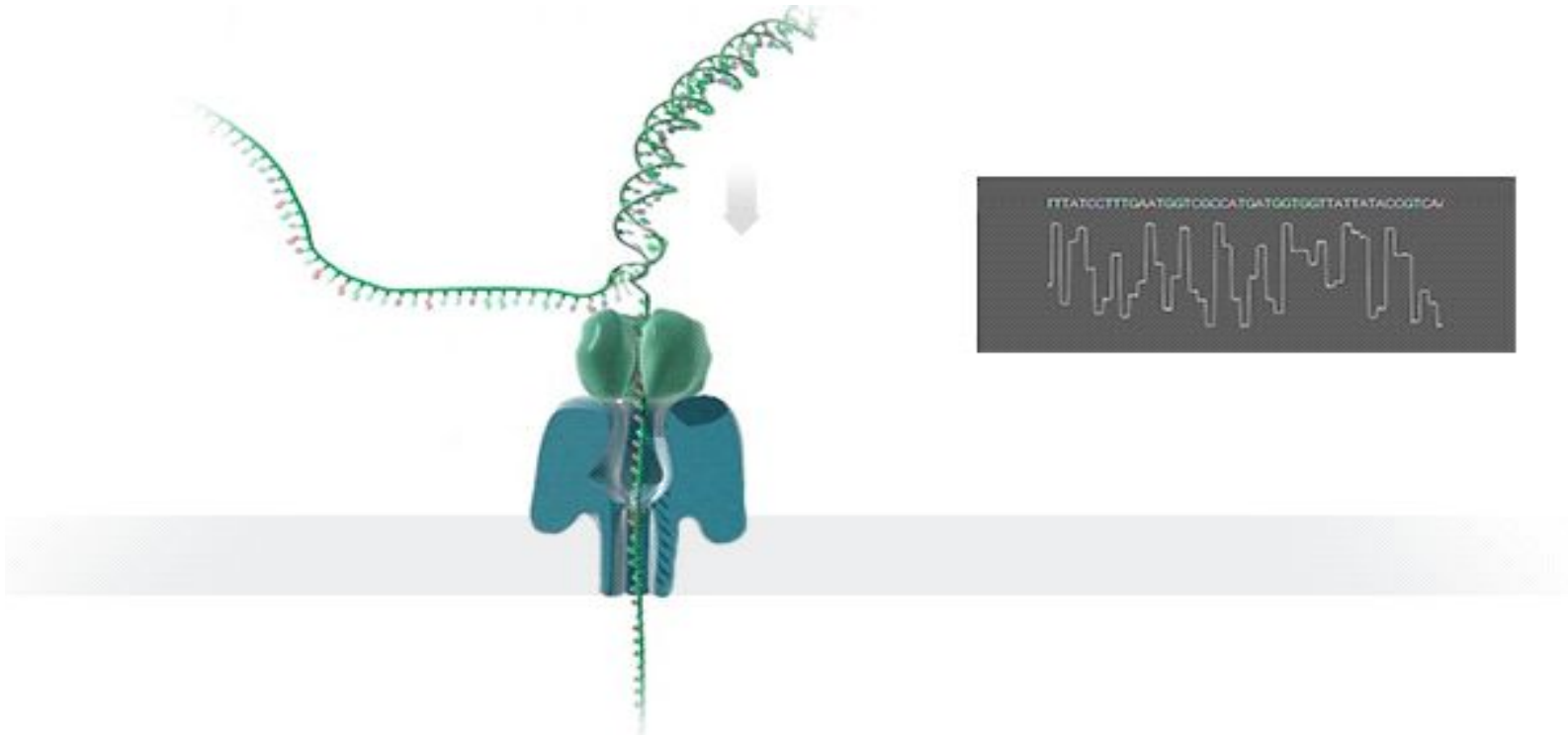
Base quality is ASCII-encoded version of $Q = -10 \log_{10} p$

Long-read sequencing via nanopores

Following slides taken or adapted from Mike Schatz:
http://schatz-lab.org/appliedgenomics2019/lectures05.LinkedAndLongReads.pdf

Nanopore Sequencing

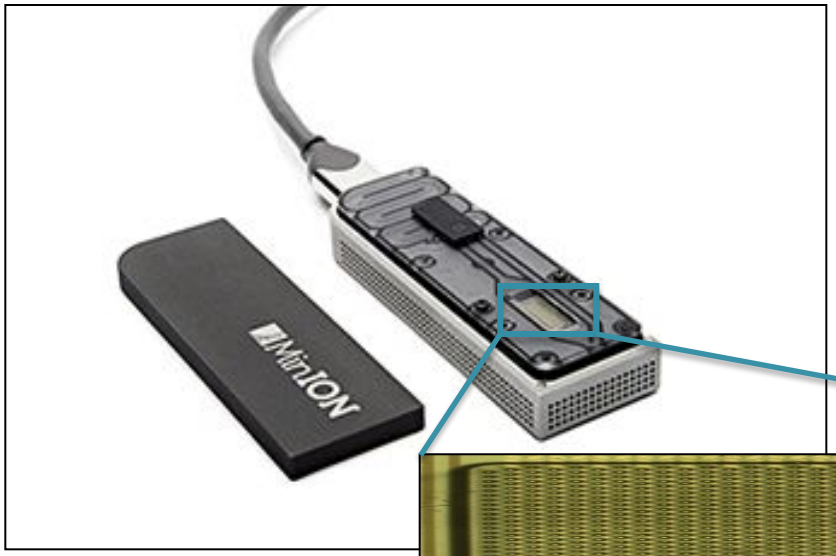
Sequences DNA/RNA by measuring changes in ionic current as nucleotide strand passes through a pore

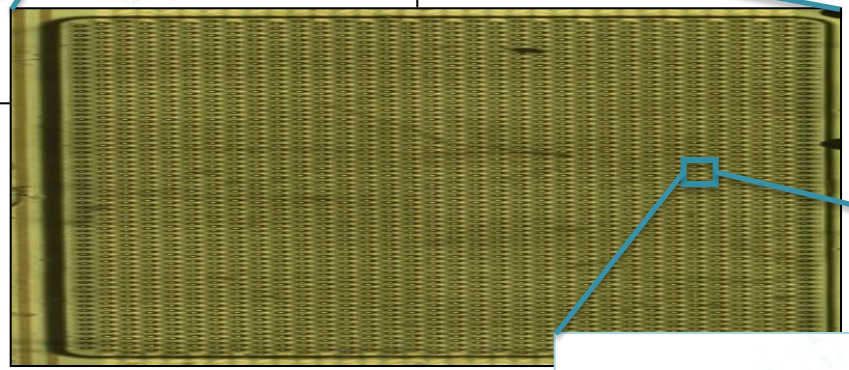


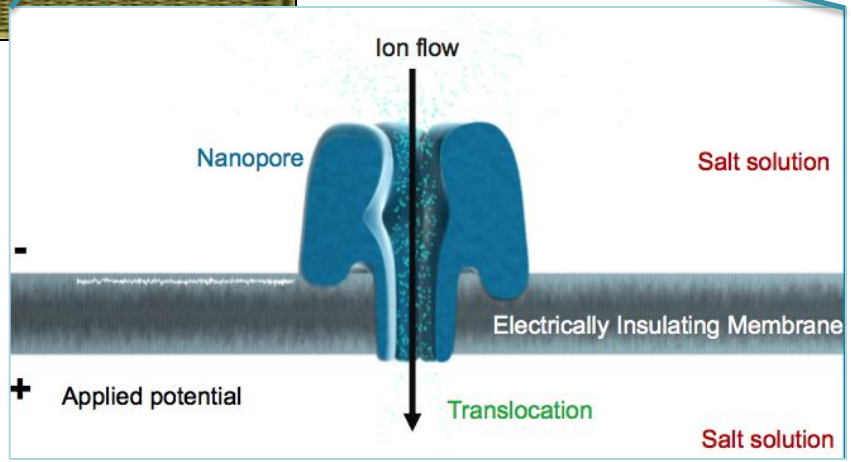
https://www.youtube.com/watch?v=CE4dW64x3Ts

nanoporetech.com/applications/dna-nanopore-sequencing

Oxford Nanopore MinION







- Thumb drive sized sequencer powered over USB
- Contains 512 channels
- Four pores per channel, only one pore active at a time

- Early access began in 2014
- Officially released in 2015

“Ultra-Long Read” Assembly

nature biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. *De novo* assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). We developed a protocol to generate ultra-long reads (N50 > 100 kb, read lengths up to 882 kb). Incorporating an additional 5× coverage of these ultra-long reads more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4-Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38.

The human genome is used as a yardstick to assess performance of DNA sequencing instruments[1–5]. Despite improvements in sequencing technology, assembling human genomes with high accuracy and completeness remains challenging. This is due to size (~3.1 Gb), heterozygosity, regions of GC% bias, diverse repeat families, and segmental duplications (up to 1.7 Mbp in size) that make up at least 50% of the genome[6]. Even more challenging are the pericentromeric, centromeric, and acrocentric short arms of chromosomes, which contain satellite DNA and tandem repeats of 3–10 Mb in length[7,8]. Repetitive structures pose challenges for *de novo* assembly using “short read” sequencing technologies, such as Illumina’s. Such data, while enabling highly accurate genotyping in non-repetitive regions, do not provide contiguous *de novo* assemblies. This limits the ability to reconstruct repetitive sequences, detect complex structural variation, and fully characterize the human genome.

Single-molecule sequencers, such as Pacific Biosciences’ (PacBio), can produce read lengths of 10 kb or more, which makes *de novo* human genome assembly more tractable[9]. However, single-molecule sequencing reads have significantly higher error rates compared with Illumina sequencing. This has necessitated development of *de novo* assembly algorithms and the use of long noisy data in conjunction with accurate short reads to produce high-quality reference genomes[10]. In May 2014, the MinION nanopore sequencer was made available to early-access users[11]. Initially, the MinION nanopore sequencer was used to sequence and assemble microbial genomes or PCR products[12–14] because the output was limited to 500 Mb to 2 Gb of sequenced bases. More recently, assemblies of eukaryotic genomes including yeasts, fungi, and *Caenorhabditis elegans* have been reported[15–17].

Recent improvements to the protein pore (a laboratory-evolved *Escherichia coli* CsgG mutant named R9.4), library preparation techniques (1D ligation and 1D rapid), sequencing speed (450 bases/s), and control software have increased throughput, so we hypothesized that whole-genome sequencing (WGS) of a human genome might be feasible using only a MinION nanopore sequencer[17–19].

We report sequencing and assembly of a reference human genome for GM12878 from the Utah/CEPH pedigree, using MinION R9.4 1D chemistry, including ultra-long reads up to 882 kb in length. GM12878 has been sequenced on a wide variety of platforms, and has well-validated variation call sets, which enabled us to benchmark our results[20].

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA. [2]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA. [3]Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA. [5]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, Utah, USA. [6]Michael Smith Laboratories and Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada. [7]Surgical Research Laboratory, Institute of Cancer & Genomic Science, University of Birmingham, UK. [8]DeepSeq, School of Life Sciences, University of Nottingham, UK. [9]Norwich Medical School, University of East Anglia, Norwich, UK. [10]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA. [11]Ontario Institute for Cancer Research, Toronto, Canada. [12]Department of Computer Science, University of Toronto, Toronto, Canada. [13]These authors contributed equally to this work. Correspondence should be addressed to N.J.L. (n.j.loman@bham.ac.uk) or M.L. (matt.loose@nottingham.ac.uk).

Received 20 April 2017; accepted 11 December 2017; published online 29 January 2018; doi:10.1038/nbt.4060

Current Nanopore Assembly

nature biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

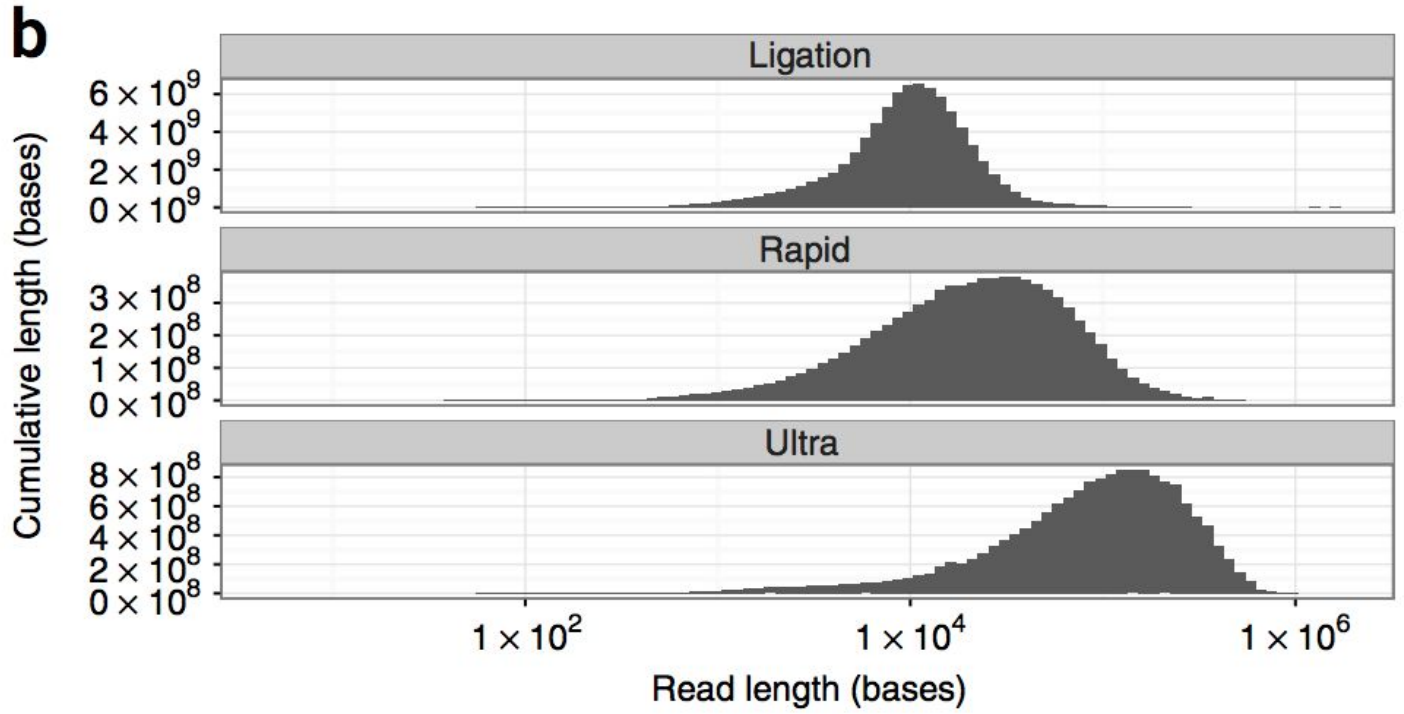
We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30x theoretical coverage, were produced. Referenc
modifications. *De novo* assembly o
protocol to generate ultra-long reac
of these ultra-long reads more thar
2,867 million bases in size, coveri
short-read sequencing data, exceei
histocompatibility complex (MHC)
reference human genome assembly

The human genome is used as a yards
DNA sequencing instruments[1–5]. Desp
ing technology, assembling human gen
completeness remains challenging. Thi
erozygosity, regions of GC% bias, dive
mental duplications (up to 1.7 Mbp in s
of the genome[6]. Even more challenging
tromeric, and acrocentric short arms of
satellite DNA and tandem repeats of 3–
structures pose challenges for *de novo*
sequencing technologies, such as Illumi
highly accurate genotyping in non-repe
contiguous *de novo* assemblies. This lir
repetitive sequences, detect complex s
characterize the human genome.

Single-molecule sequencers, such as
can produce read lengths of 10 kb or mor
genome assembly more tractable[9]. Howe
ing reads have significantly higher error
sequencing. This has necessitated deve

[1]UC Santa Cruz Genomics Institute, University
Branch, National Human Genome Research In
UK. [4]Department of Human Genetics, Universi
Utah, USA. [6]Michael Smith Laboratories and D
Laboratory, Institute of Cancer & Genomic Scie
Medical School, University of East Anglia, Nor
for Cancer Research, Toronto, Canada. [12]Depar
Correspondence should be addressed to N.J.L.

Received 20 April 2017; accepted 11 Decemb



Current Nanopore Assembly

nature biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Referenc
modifications. *De novo* assembly o
protocol to generate ultra-long reac
of these ultra-long reads more thar
2,867 million bases in size, coveri
short-read sequencing data, excee
histocompatibility complex (MHC)
reference human genome assembly

The human genome is used as a yards
DNA sequencing instruments[1–5]. Desp
ing technology, assembling human gen
completeness remains challenging. Thi
erozygosity, regions of GC% bias, dive
mental duplications (up to 1.7 Mbp in s
of the genome[6]. Even more challenging
tromeric, and acrocentric short arms of
satellite DNA and tandem repeats of 3–
structures pose challenges for *de novo*
sequencing technologies, such as Illumi
highly accurate genotyping in non-repe
contiguous *de novo* assemblies. This lir
repetitive sequences, detect complex s
characterize the human genome.

Single-molecule sequencers, such as
can produce read lengths of 10 kb or mor
genome assembly more tractable[9]. Howe
ing reads have significantly higher error
sequencing. This has necessitated deve

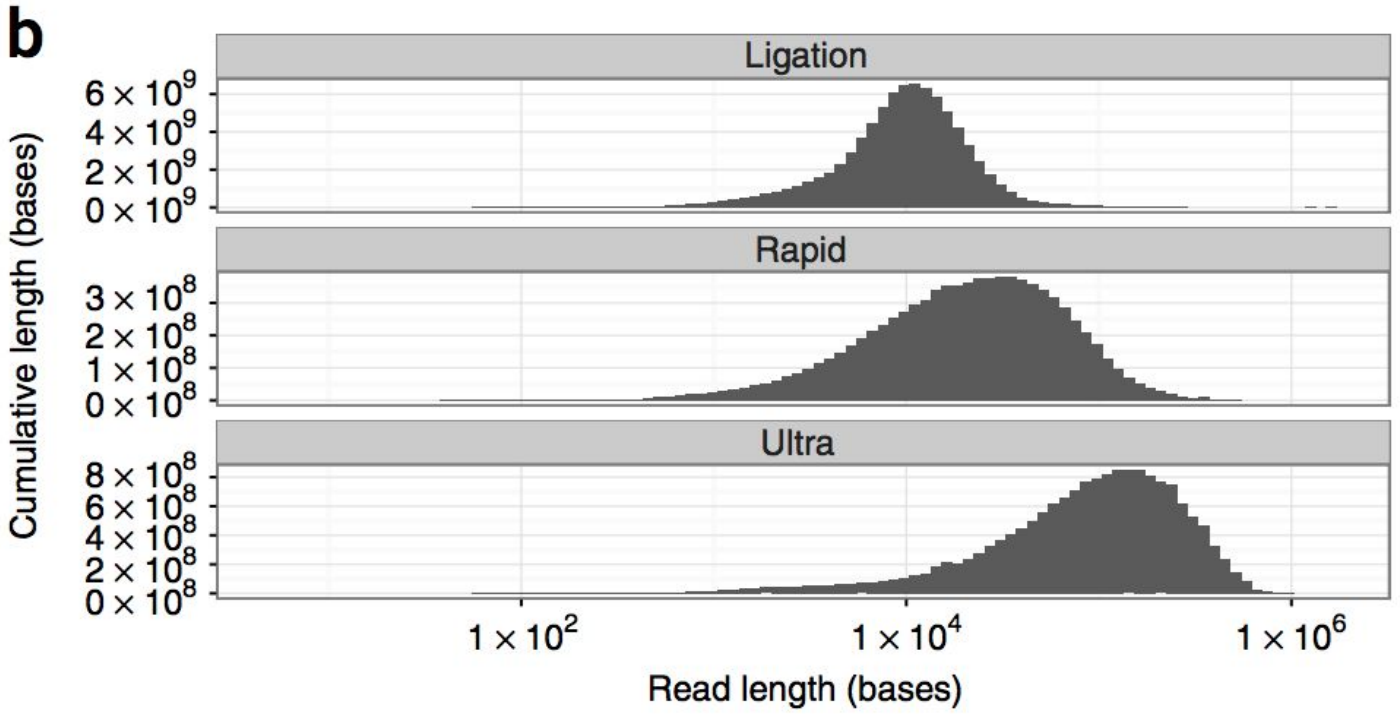
[1]UC Santa Cruz Genomics Institute, University
Branch, National Human Genome Research In
UK. [4]Department of Human Genetics, Universi
Utah, USA. [6]Michael Smith Laboratories and D
Laboratory, Institute of Cancer & Genomic Scie
Medical School, University of East Anglia, Nor
for Cancer Research, Toronto, Canada. [12]Depar
Correspondence should be addressed to N.J.L.

Received 20 April 2017; accepted 11 Decemb

Same group recently reported a read 2.3 million bases long!

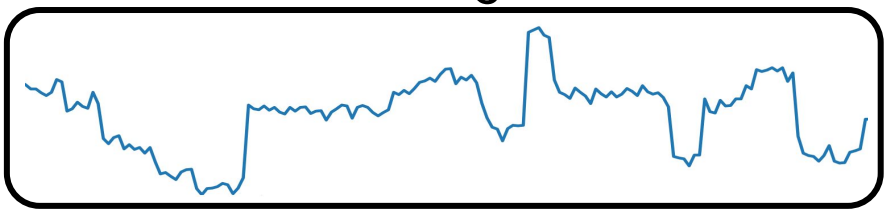
2,272,580 nt, to be exact.

No theoretical upper limit



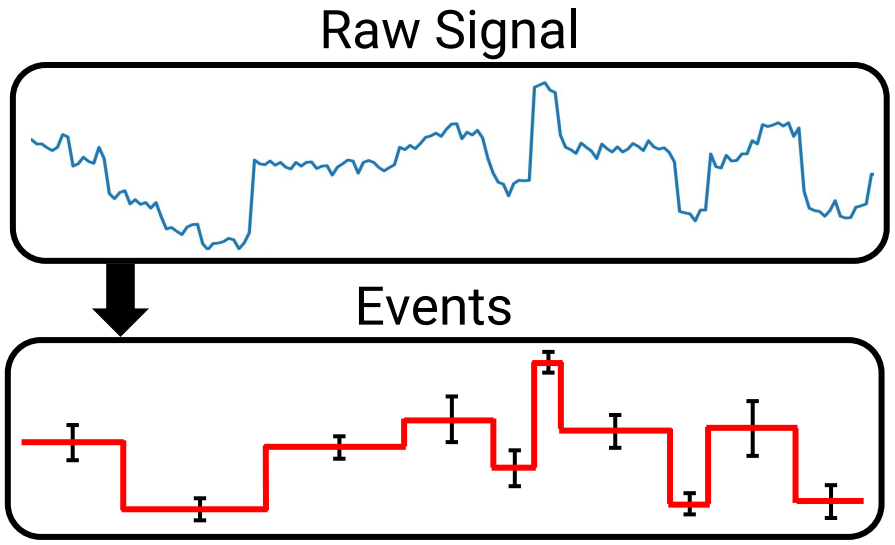
Nanopore Basecalling

Raw Signal



Translation of raw signal into basepairs

Nanopore Basecalling

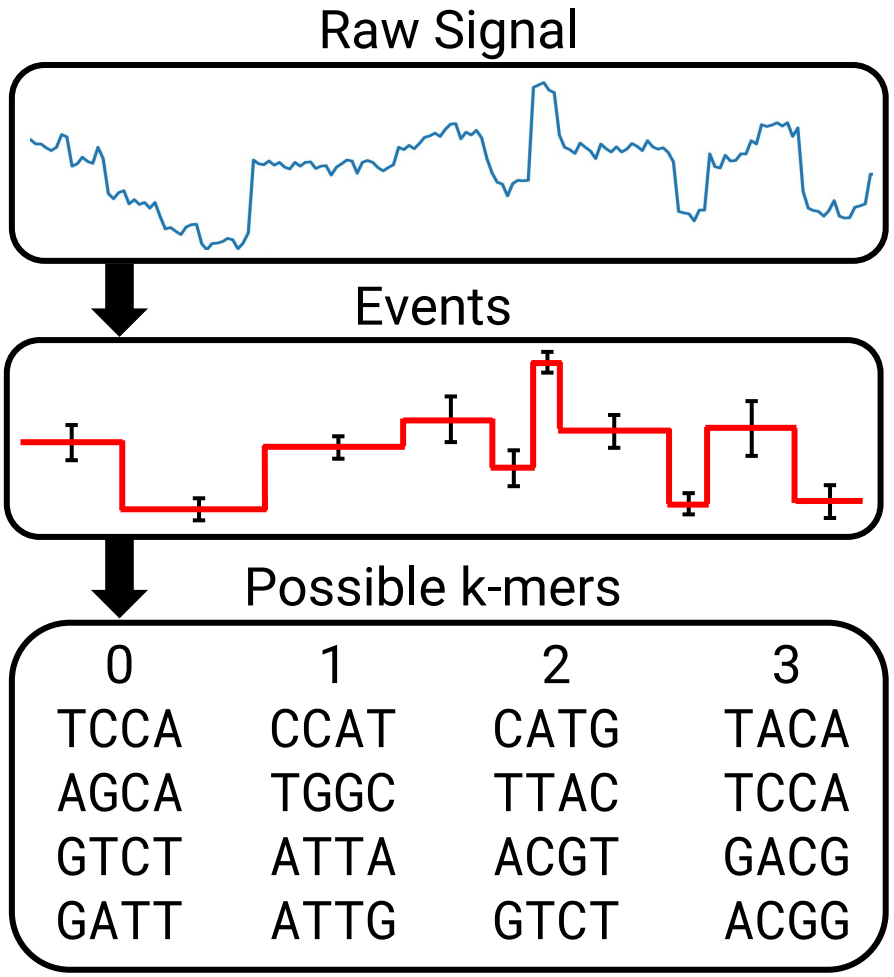


Translation of raw signal into basepairs

Early basecallers began by estimating k-mer boundaries using “events”, which were then input to an HMM

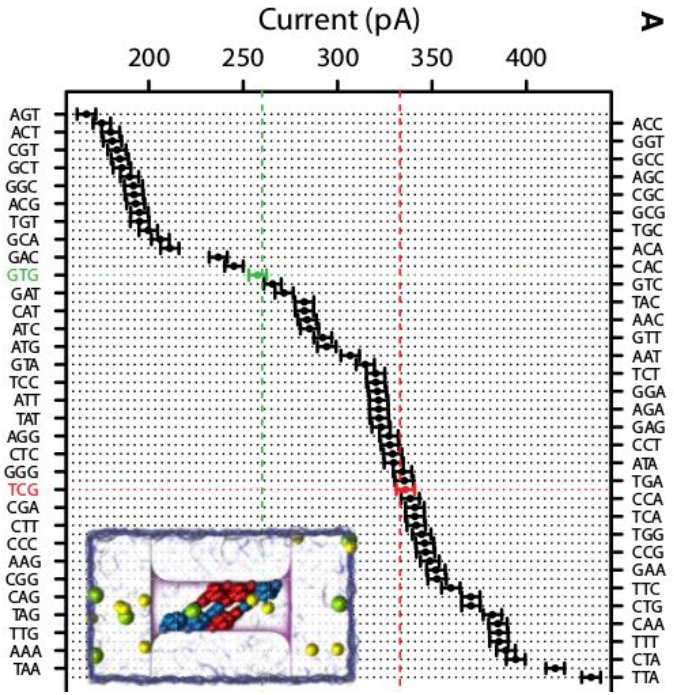
Modern basecalers use neural networks directly on raw signal

Nanopore Basecalling



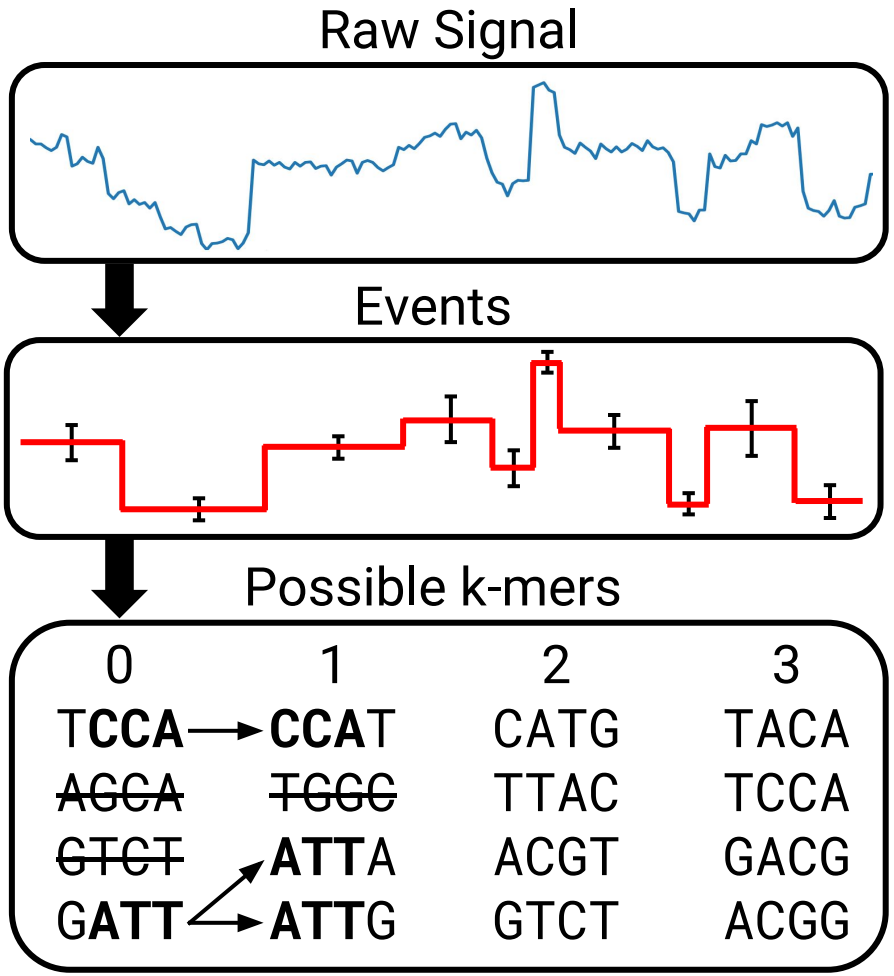
(Based on probability of event matches)

ONT releases k-mer models with expected current distribution of every k-mer

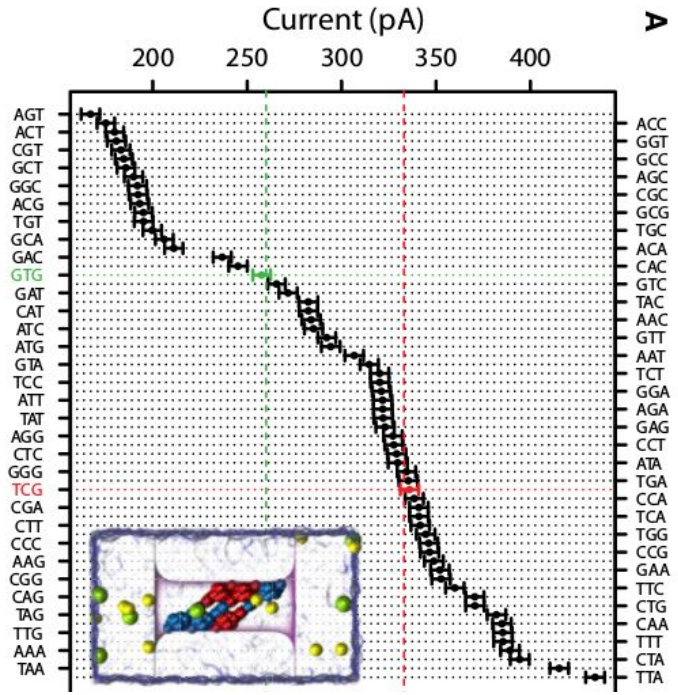


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling

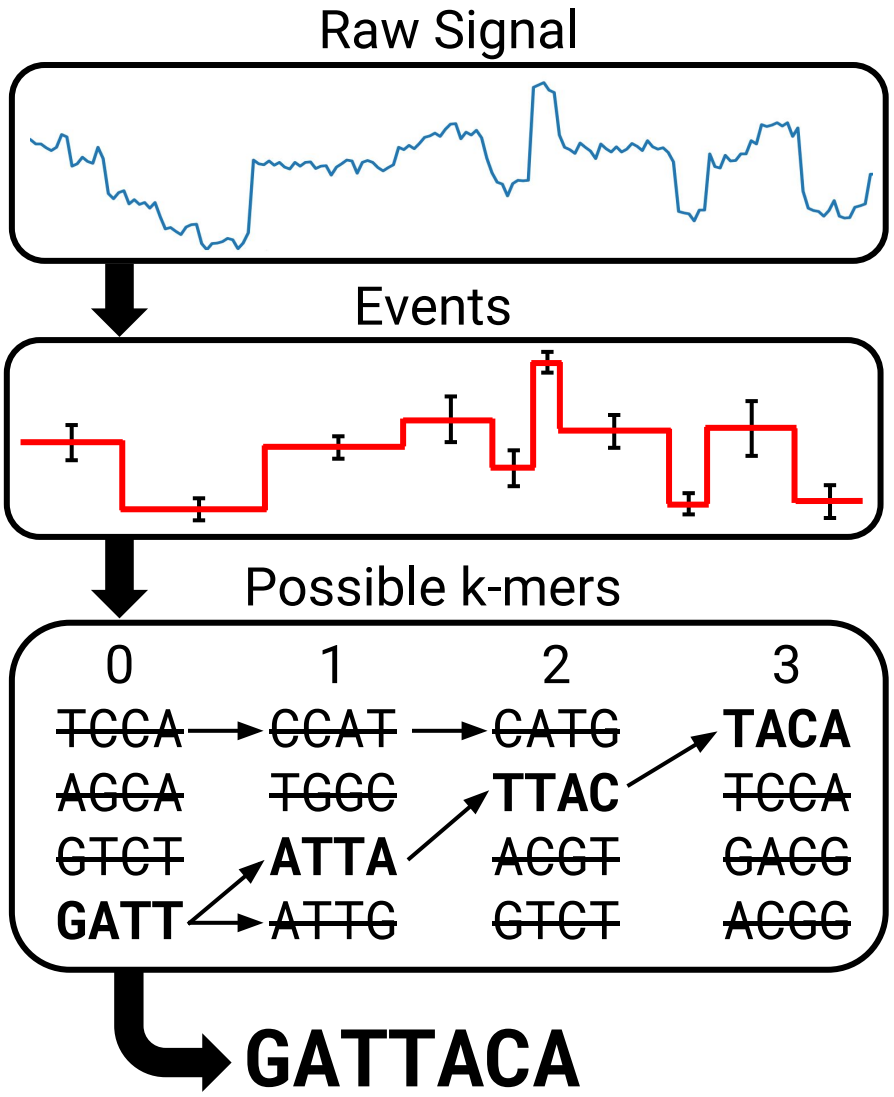


Certain k-mers can be eliminated based on possible transitions

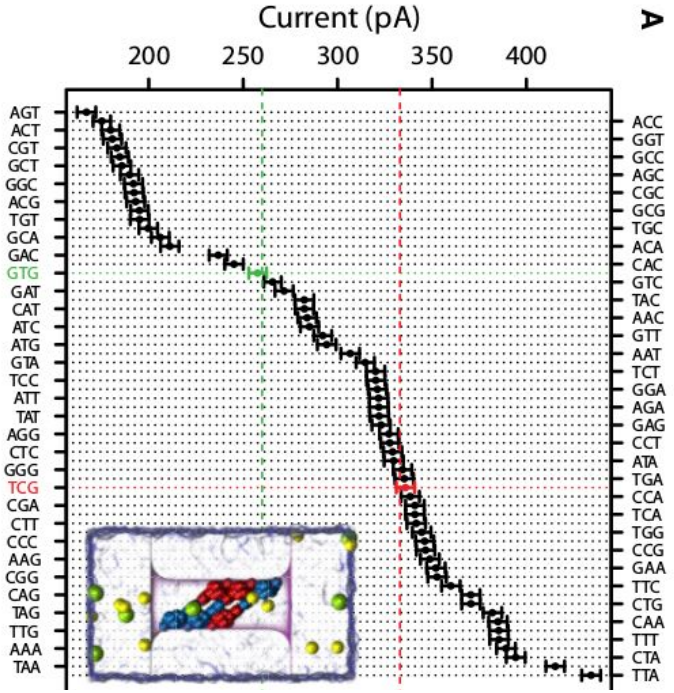


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling



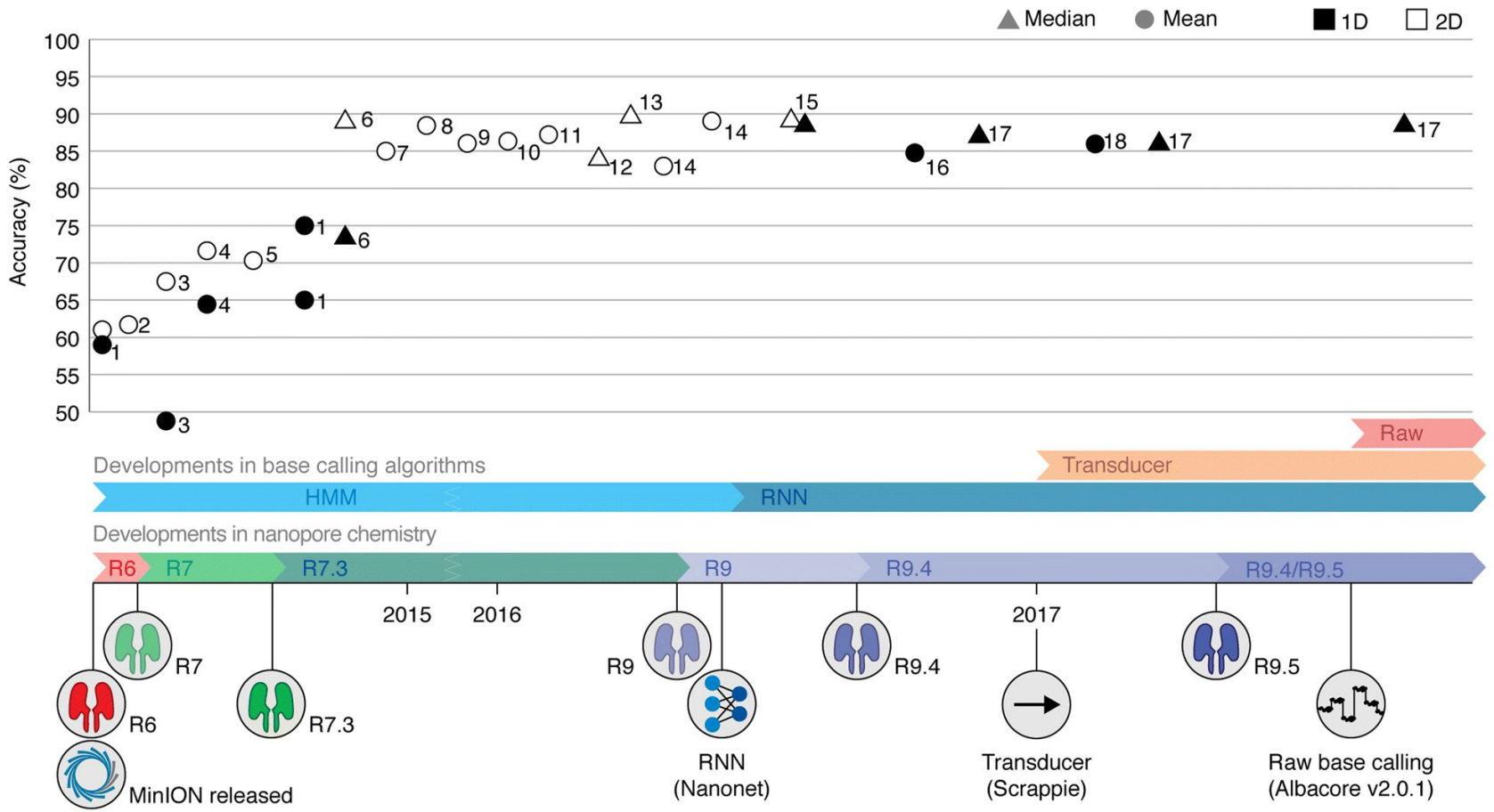
Final sequence determined by most probable k-mers



"DNA Base-Calling from a Nanopore Using a Viterbi Algorithm"
Timp et al. (2012) *Biophysical Journal*

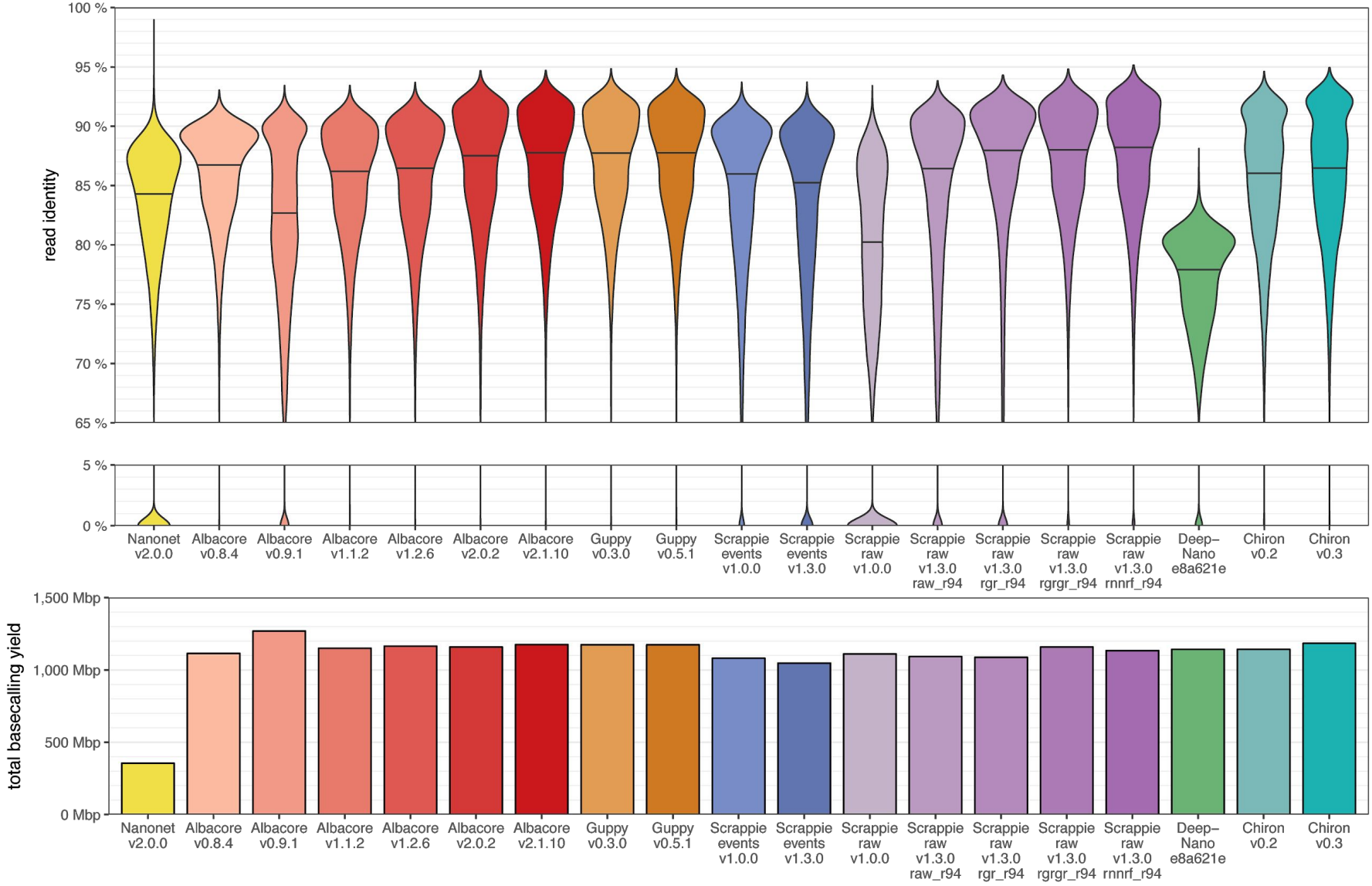
Basecaller/Pore Timeline

Development of both pore chemistry and basecalling algorithms is responsible for improvement in accuracy



From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology.* https://doi.org/10.1186/s13059-018-1462-9

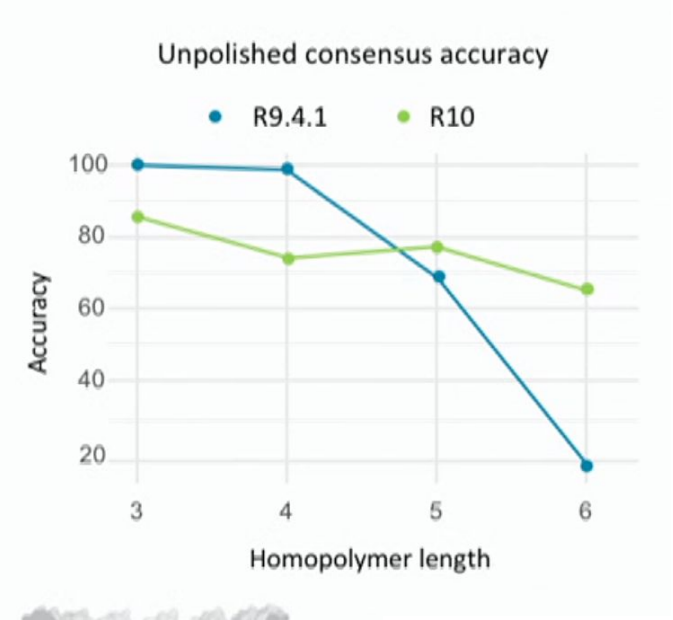
Basecaller Comparison

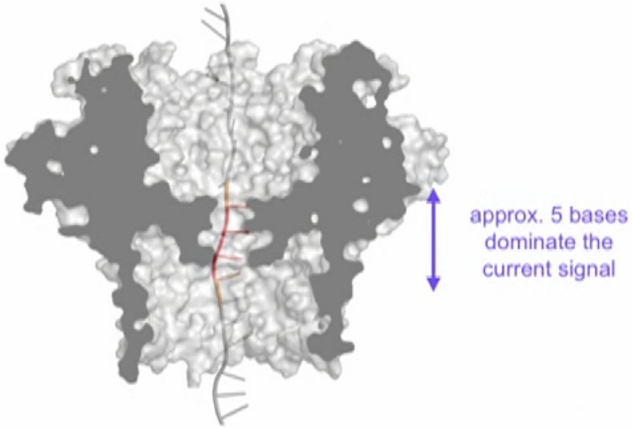


https://github.com/rrwick/Basecalling-comparison

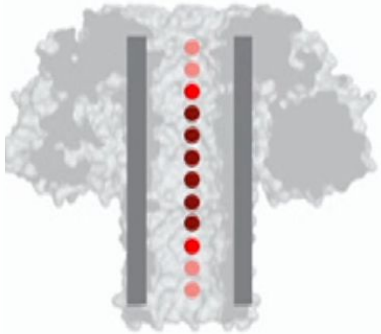
New Pore Chemistries

ONT is developing alternate pore chemistries to improve accuracy, particularly for homopolymers

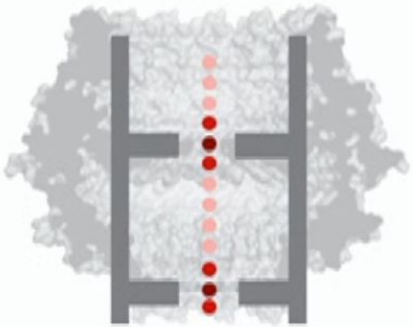




Standard pore chemistry
"R9"



Pore with long reader head Lysenin –
"R8"



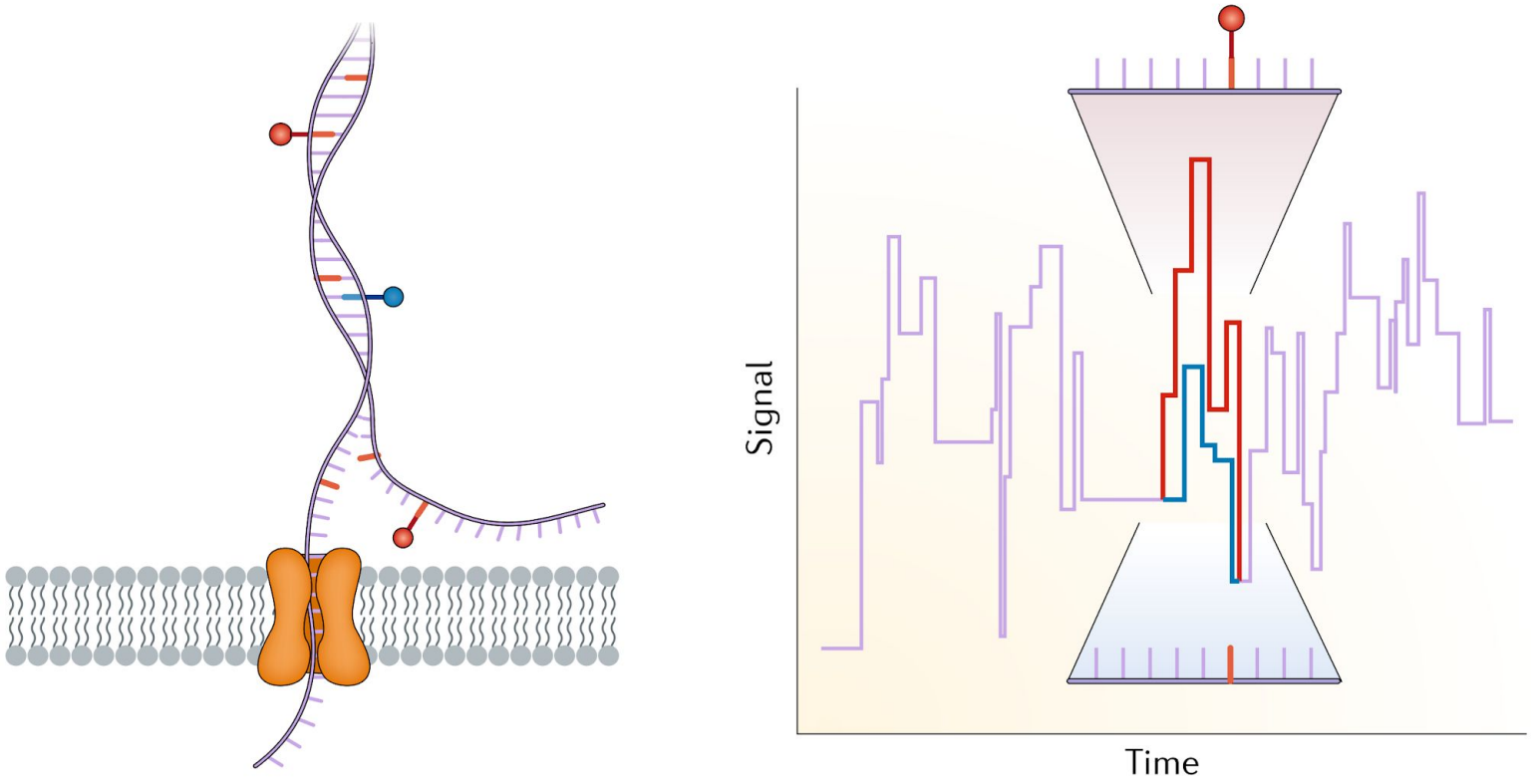
Multiple points of contribution
"R10"

From 2018 London Calling Keynote
https://vimeo.com/272526835

DNA Modification Detection

Like PacBio, ONT can detect methylation from raw signal

- Or any other modification that changes ionic current



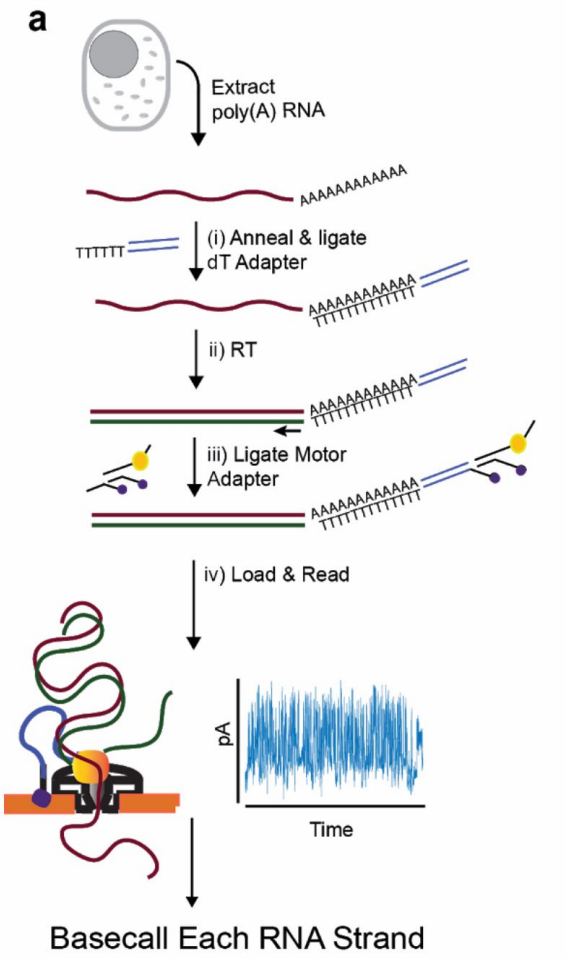
Piercing the dark matter: bioinformatics of long-range sequencing and mapping
Sedlazeck et al. (2018) *Nature Reviews Genetics.* 19:329

Direct RNA-seq

Standard RNA sequencing (RNA-seq) requires creation of complementary DNA (cDNA)

ONT recently introduced direct RNA sequencing

Allows detection of RNA modifications, and potentially secondary structure



Nanopore native RNA sequencing of a human poly(A) transcriptome

Workman et al. *BioRxiv* (https://www.biorxiv.org/content/10.1101/459529v1)

ReadUntil Sequencing

ONT machines can stop sequencing a read and immediately start on another in real-time

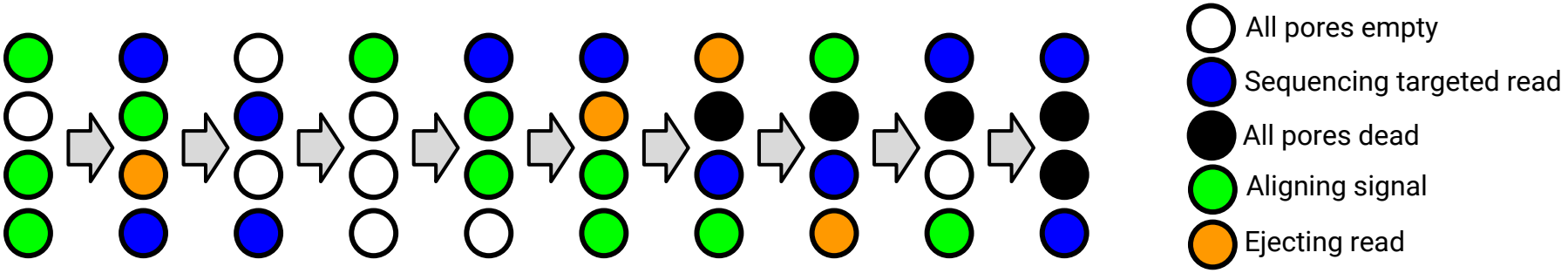
- Each channel has four pores, non-active pores have reads docked

Can potentially avoid sequencing unwanted reads

- For example: reads that align to the human genome, reads that *do not* align to a database of pathogens, reads that align to a region already sequenced to a desired depth

MinION has up to 512 active channels, each reading 450 bp/sec

- Actual number of active channels is variable



TRADEOFFS OF DIFFERENT TECHS.

- Illumina sequencing is cheap & ubiquitous
  - Can sequence *very* deeply, good for measuring abundance
  - Error rate is *very* low, can be good for detection of small variants
  - Can be combined with other technologies (e.g. linked-reads) to provide many different types of information
  - Reads are short (<=350bp each, fragments <= 1000bp), making assembly difficult
  - Library prep (prior to sequencing) can introduce many biases
- ONT sequencing is getting cheaper
  - Can sequence *very* long reads, transformative for assembly
  - Can be good for detection of large (e.g. structural) variants
  - Error rate is *much* higher than with short reads (getting better)
  - Fewer individual reads raise challenges in quantification-related tasks

Work Cited

- Illumina Sequencing Technology. (2010). In Techonlogy Spotlight: Illumina Sequencing. Retreived July 30, 2014, from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
- Illumina Solexa Sequencing. (Jan 22, 2010). On YouTube uploaded by Draven1983101. Retrieved July 30, 2014, from https://www.youtube.com/watch?v=77r5p8IBwJk
- Overview of Illumina Chemistry. In Massachusetts General Hospital. Retreived July 30, 2014, from http://nextgen.mgh.harvard.edu/IlluminaChemistry.html
- Introduction to Next Generation Sequencing Using the Illumina 1G Genome Analyzer (Solexa). (Jan 31, 2008). Retrieved July 30, 2014, from http://research.stowers-institute.org/microscopy/external/PowerpointPresentations/ppt/Methods_Technology/KSH_Tech&Methods_012808Final.pdf
- Sequencing technology – Past, Present and Future. (2013). Wei Chen. Berlin Institute for Medical Systems Biology. Max-Delbrueck-Center for Molecular Medicine Retrieved July 30, 2014, from http://www.molgen.mpg.de/899148/OWS2013_NGS.pdf
- DNA barcoding. (2014, July 30). In *Wikipedia, The Free Encyclopedia*. Retrieved July 30, 2014, from http://en.wikipedia.org/w/index.php?title=DNA_barcoding&oldid=619163634
- Reference genome. (2014, June 22). In Wikipedia, The Free Encyclopedia. Retrieved 19:54, August 4, 2014, from http://en.wikipedia.org/w/index.php?title=Reference_genome&oldid=613984719

Work Cited

- Illumina Sequencing Technology. (2010). In Techonlogy Spotlight: Illumina Sequencing. Retreived July 30, 2014, from http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
- Illumina Solexa Sequencing. (Jan 22, 2010). On YouTube uploaded by Draven1983101. Retrieved July 30, 2014, from https://www.youtube.com/watch?v=77r5p8IBwJk
- Overview of Illumina Chemistry. In Massachusetts General Hospital. Retreived July 30, 2014, from http://nextgen.mgh.harvard.edu/IlluminaChemistry.html
- Introduction to Next Generation Sequencing Using the Illumina 1G Genome Analyzer (Solexa). (Jan 31, 2008). Retrieved July 30, 2014, from http://research.stowers-institute.org/microscopy/external/PowerpointPresentations/ppt/Methods_Technology/KSH_Tech&Methods_012808Final.pdf
- Sequencing technology – Past, Present and Future. (2013). Wei Chen. Berlin Institute for Medical Systems Biology. Max-Delbrueck-Center for Molecular Medicine Retrieved July 30, 2014, from http://www.molgen.mpg.de/899148/OWS2013_NGS.pdf
- DNA barcoding. (2014, July 30). In *Wikipedia, The Free Encyclopedia*. Retrieved July 30, 2014, from http://en.wikipedia.org/w/index.php?title=DNA_barcoding&oldid=619163634
- Reference genome. (2014, June 22). In Wikipedia, The Free Encyclopedia. Retrieved 19:54, August 4, 2014, from http://en.wikipedia.org/w/index.php?title=Reference_genome&oldid=613984719