Bioinformatics: DNA Sequencing and Beyond

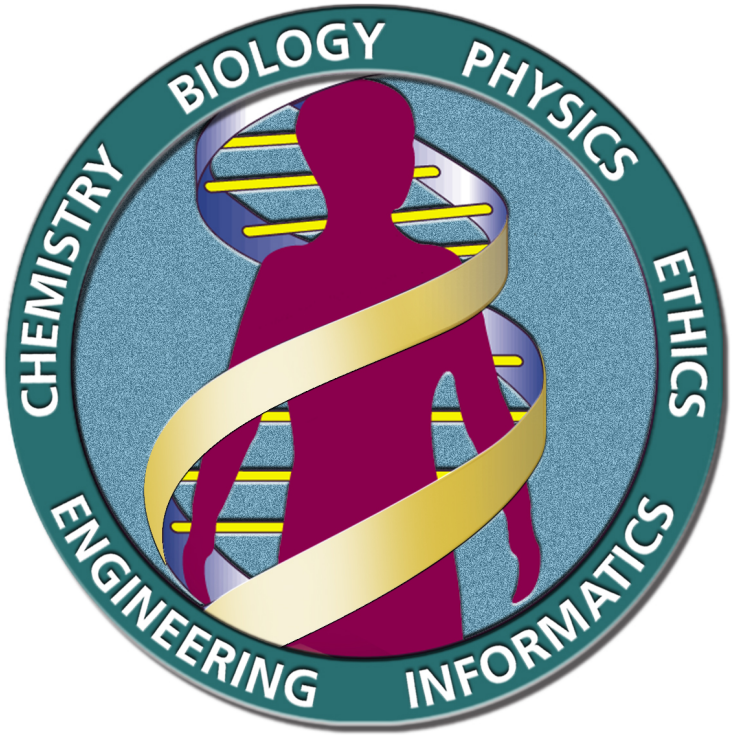
Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Book Reference

The Human Genome Project

- Proposed in 1987 by the U.S. Department of Energy (not NIH!)
- Biology's "Manhattan project"
- Officially started in 1989
- Joint effort of NIH and DOE in the United States, plus many other countries
- the Sanger Centre in England was the largest center outside the US



Goals

- sequence 3 billion basepairs for \$1/base by 2005

BRAC UNIVERSITY
Inspiring Excellence

The race to sequence the genome: early 1990s

- Scientists around the world were busy creating "maps"
- Maps take small or large pieces of DNA and place them somewhere on the genome
- Maps also take particular genes and identify their approximate location
- 1995: TIGR sequences first complete bacterial genome ever, *Haemophilus influenzae*
  - 1.8 million bases 1742 genes
  - Project led by Craig Venter (TIGR) and Hamilton Smith (Johns Hopkins)
- 1998: the race begins
  - new sequencing machine developed by Applied Biosystems
  - Craig Venter, Ham Smith & others leave TIGR to form Celera Genomics, a for-profit company

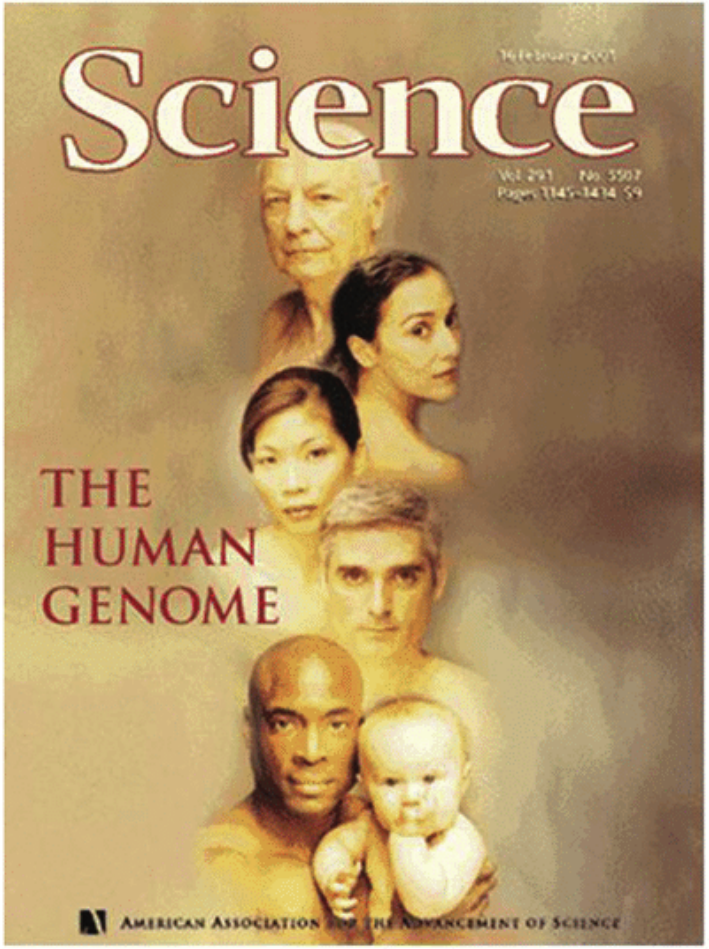
BRAC UNIVERSITY
Inspiring Excellence

The race to sequence the genome: 1999-2000

- 1999: NIH merges its efforts into 3 larger centers
- 2000: Celera sequences and publishes the complete genome of the fruit fly, Drosophila melanogaster
- Proves the whole-genome shotgun technique works on a 20X larger scale than previously
- 1999: Craig Venter announces that Celera will finish by 2001
- 1999: NIH and the Sanger Centre announce that the public HGP will finish a "draft" genome by 2001
- 2000: NIH, Sanger Centre, and Celera talk about publishing jointly
- Late 2000: talks fall apart; 2 papers planned

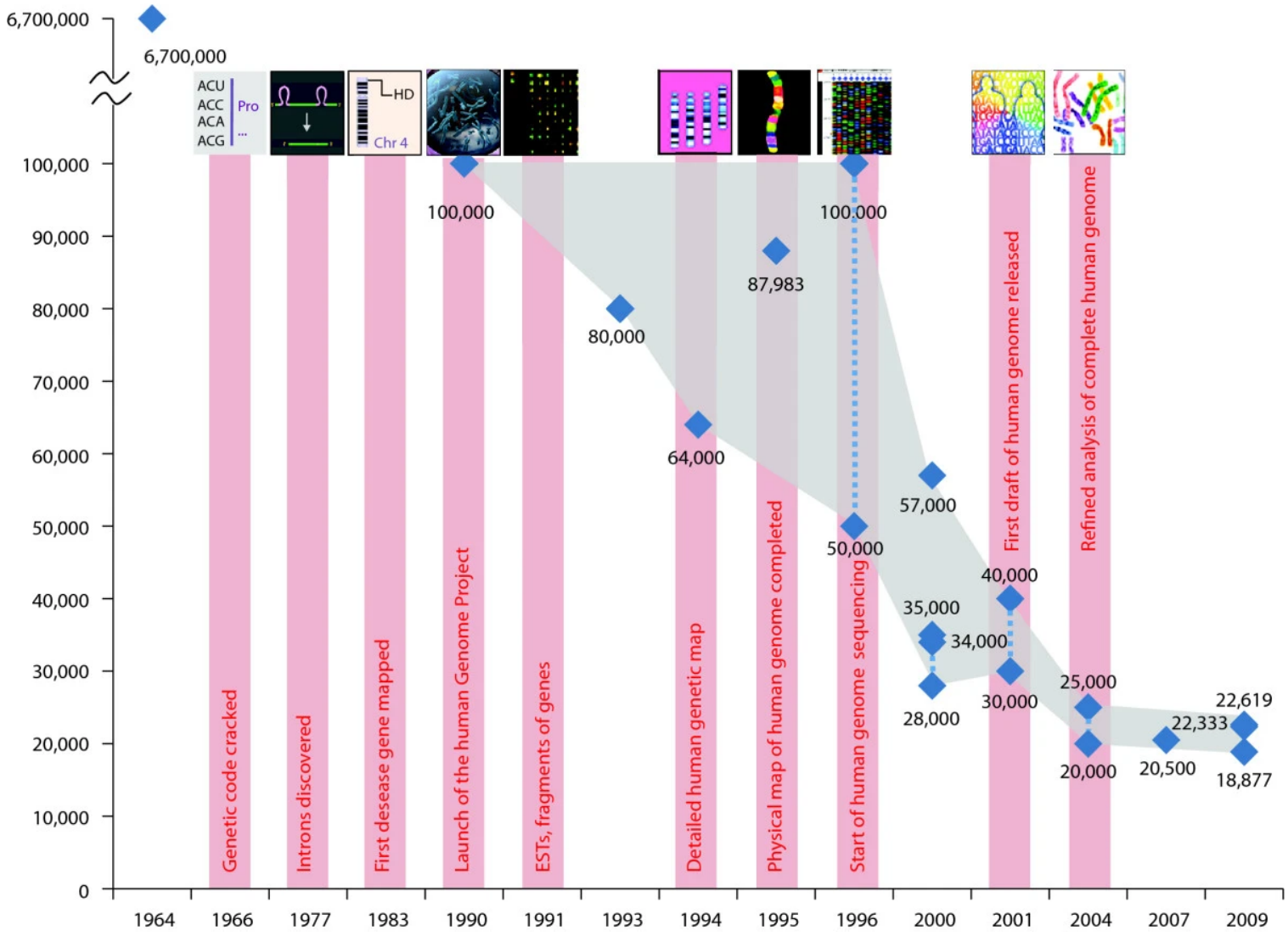
BRAC UNIVERSITY
Inspiring Excellence

Two Papers!



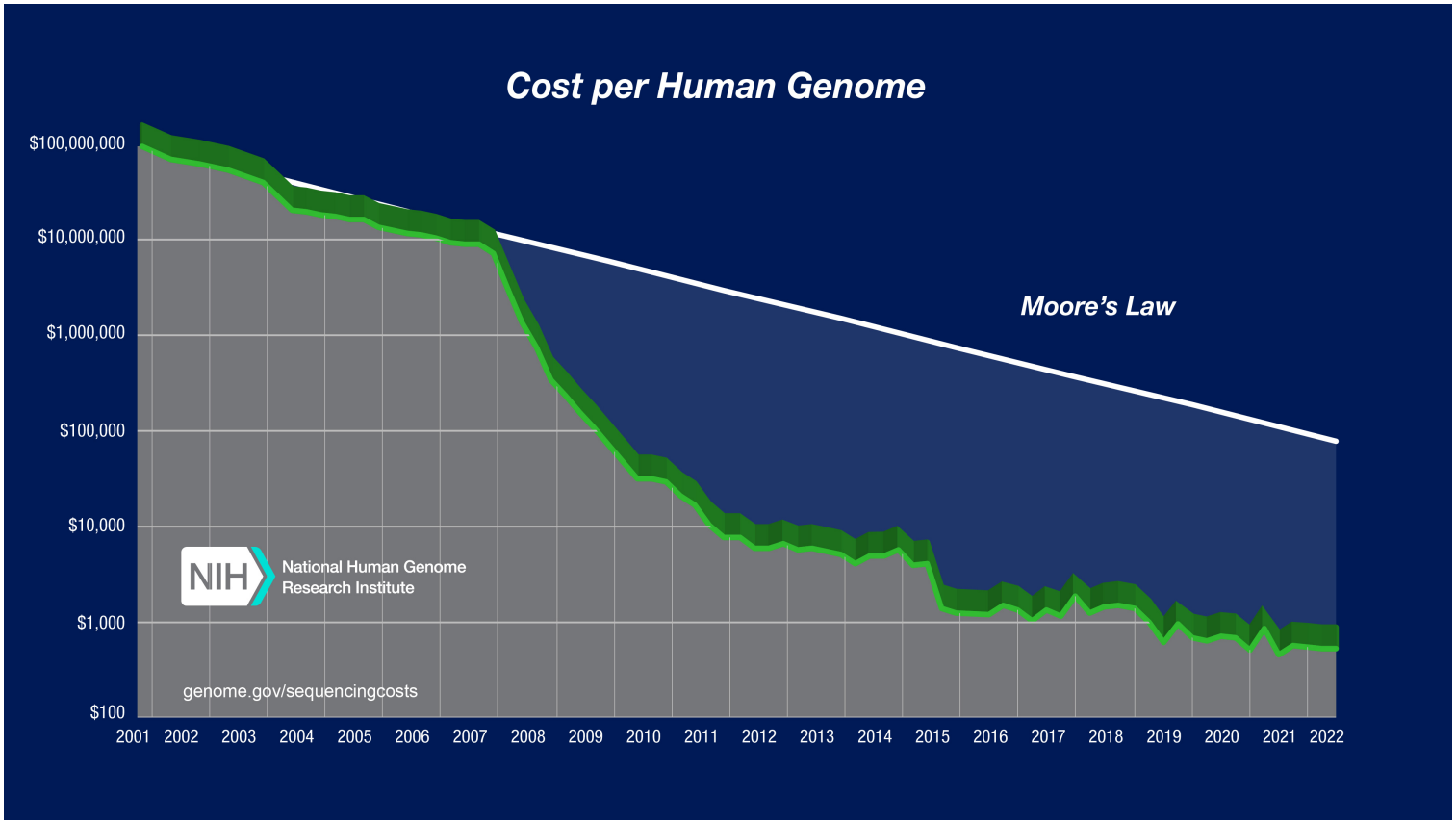
BRAC UNIVERSITY
Inspiring Excellence

Evolution of Gene Count!



BRAC UNIVERSITY
Inspiring Excellence

Evolution of Cost - Beyond Second Generation Sequencing!



BRAC UNIVERSITY
Inspiring Excellence

How Second Generation Sequencing Works!

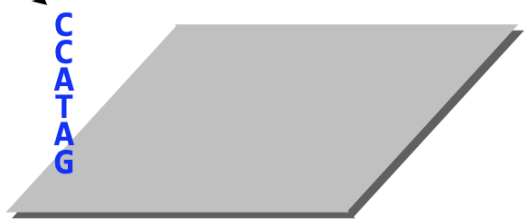
Input DNA

CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT

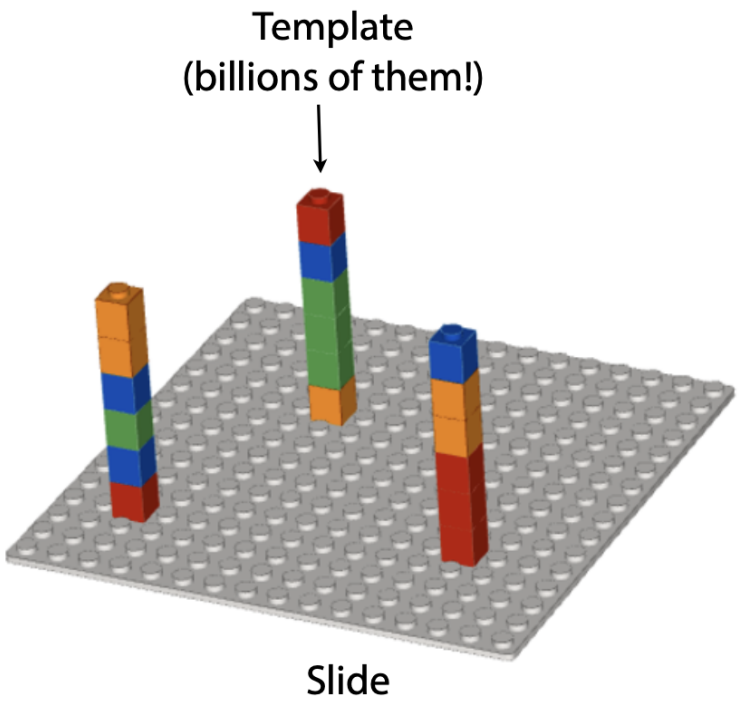
Cut into snippets

CCATAGTA TATCTCGG CTCTAGGCCCTC ATTTTTT
CCA TAGTATAT CTCGGCTCTAGGCCCTCA TTTTTT
CCATAGTAT ATCTCGGCTCTAG GCCCTCA TTTTTT
CCATAG TATATCT CGGCTCTAGGCCCT CATTTTTT

Deposit on slide

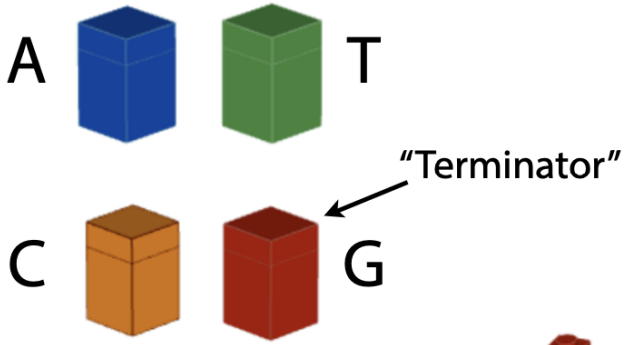


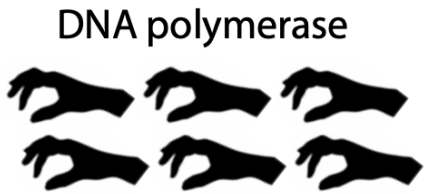
How Second Generation Sequencing Works!

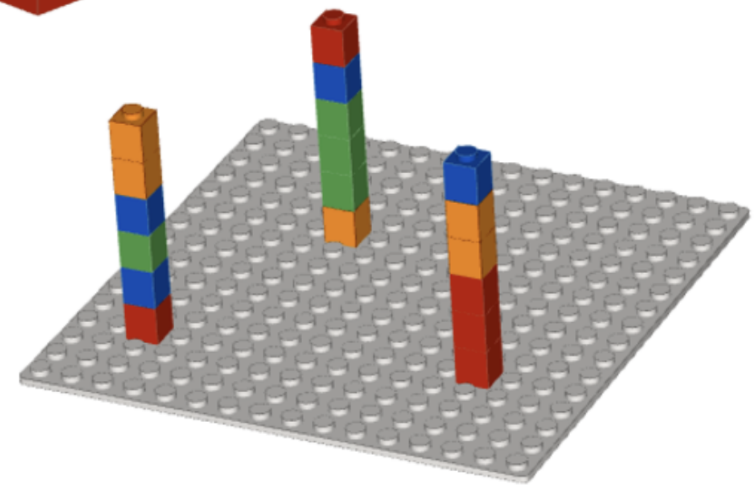


BRAC UNIVERSITY
Inspiring Excellence

How Second Generation Sequencing Works!

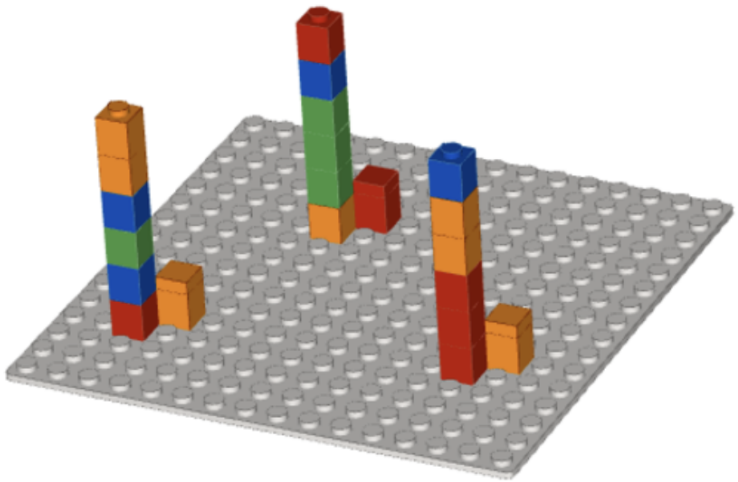




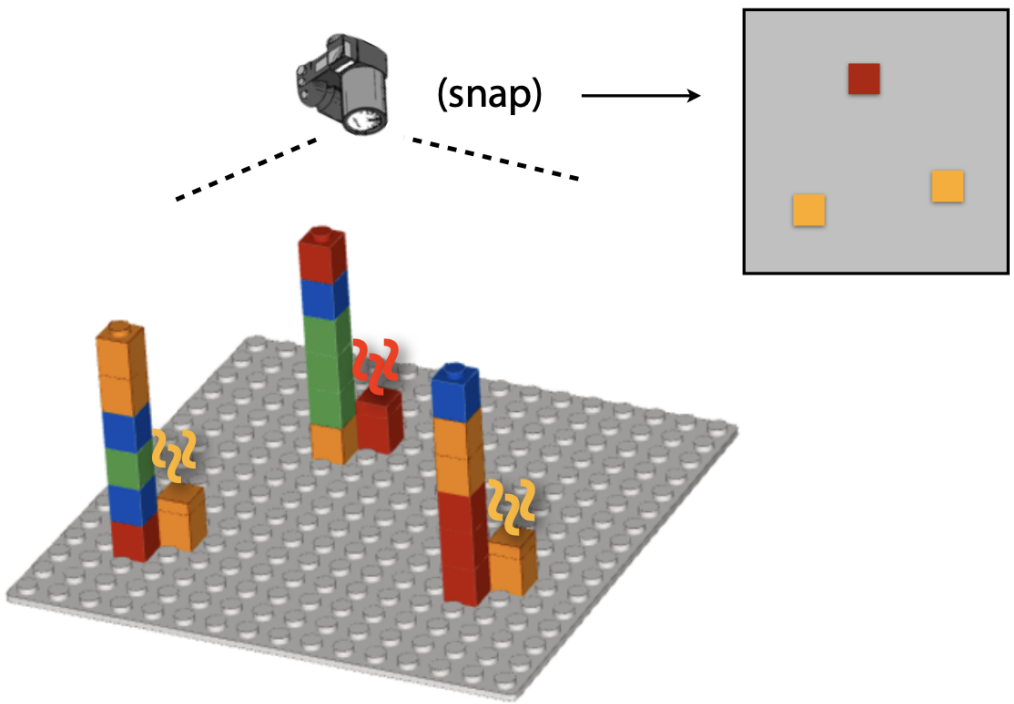


BRAC UNIVERSITY
Inspiring Excellence

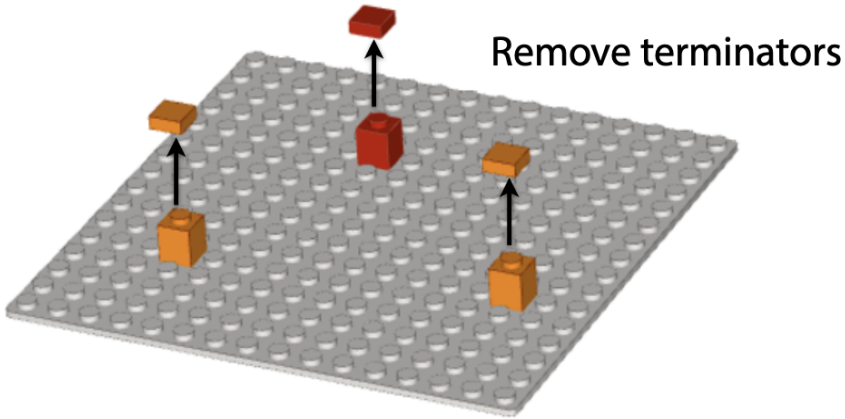
How Second Generation Sequencing Works!



How Second Generation Sequencing Works!

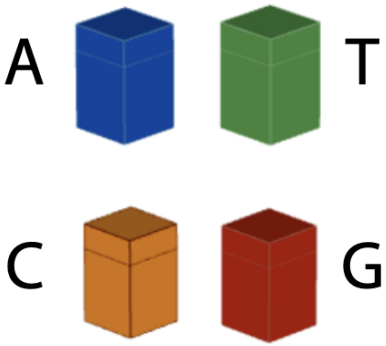


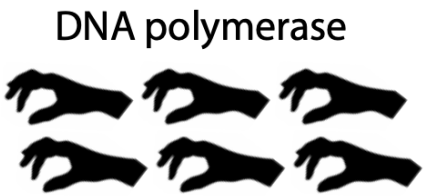
How Second Generation Sequencing Works!

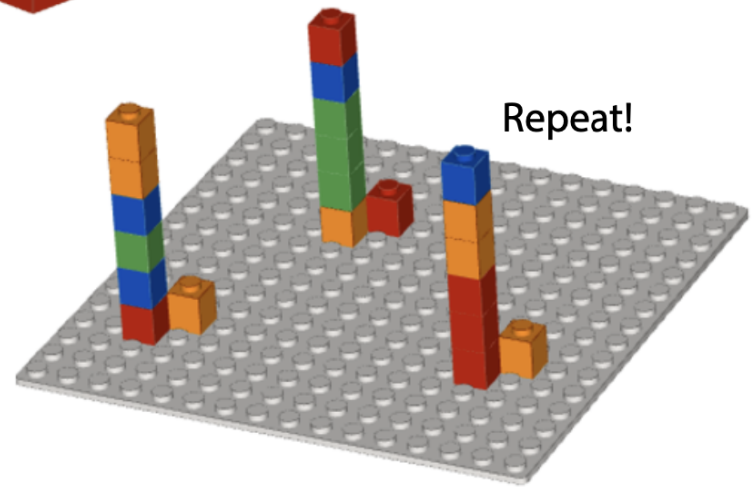


BRAC UNIVERSITY
Inspiring Excellence

How Second Generation Sequencing Works!

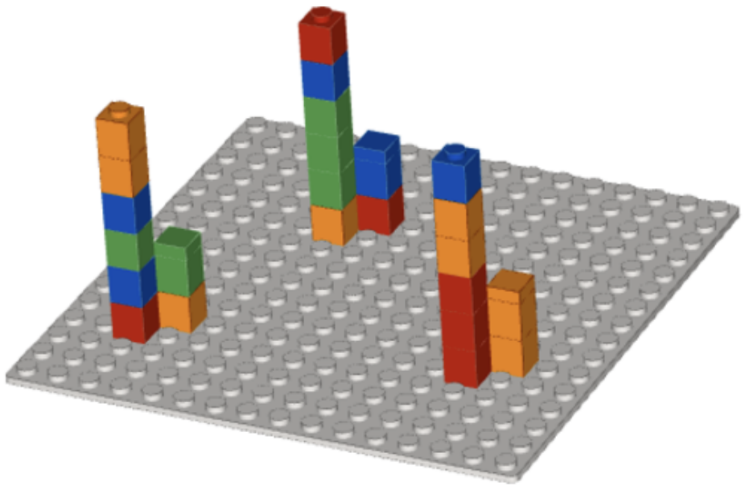






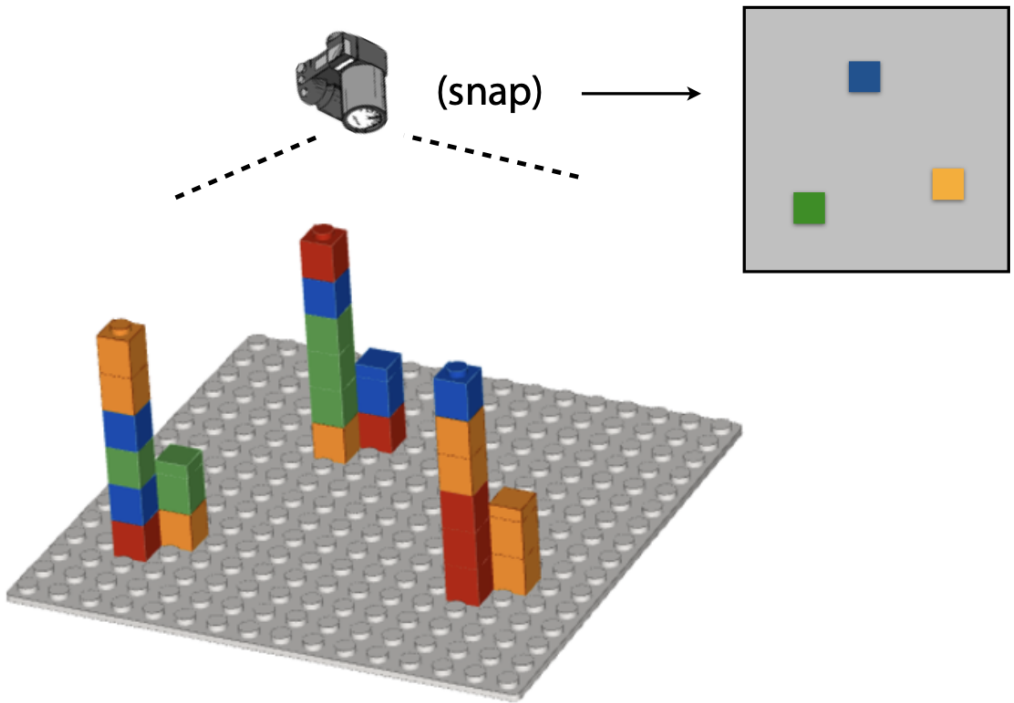
BRAC UNIVERSITY
Inspiring Excellence

How Second Generation Sequencing Works!

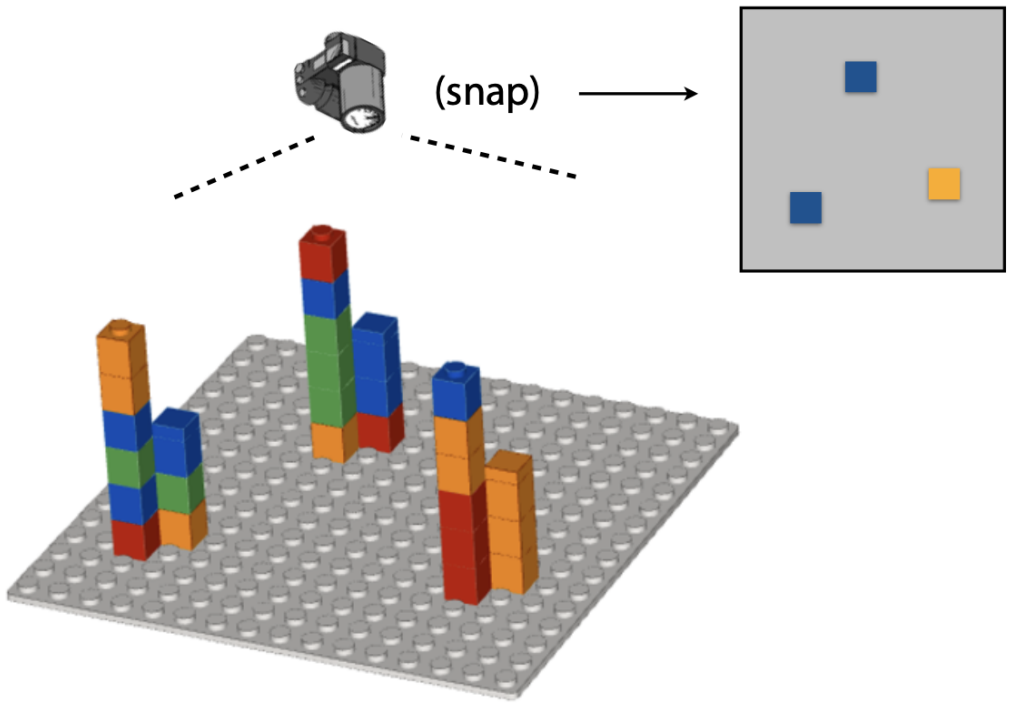


BRAC UNIVERSITY
Inspiring Excellence

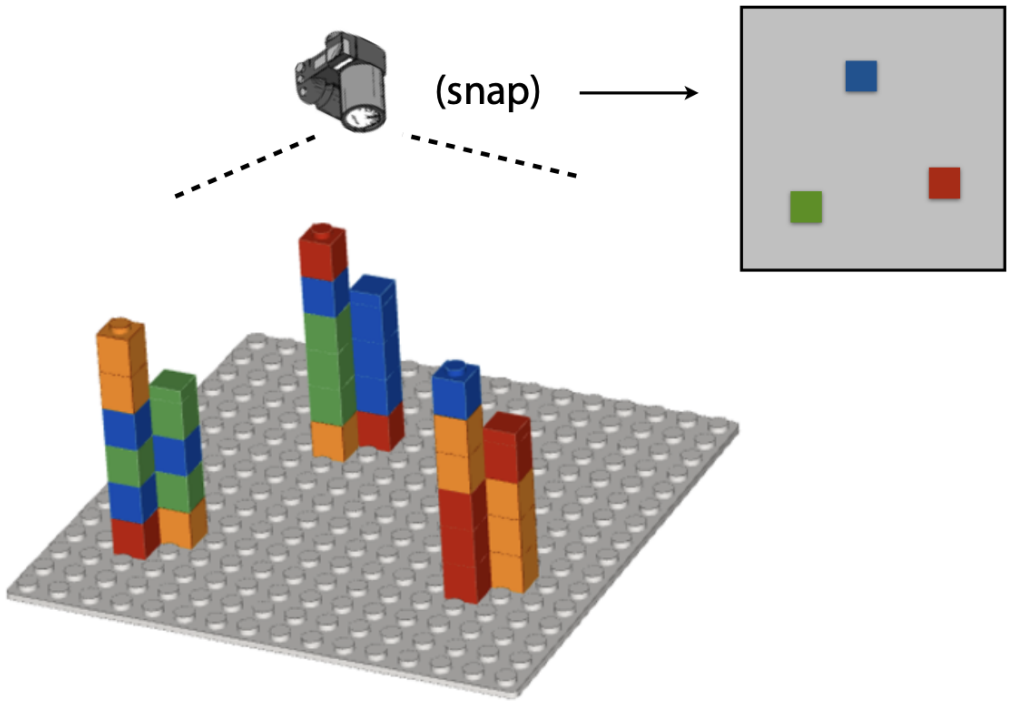
How Second Generation Sequencing Works!



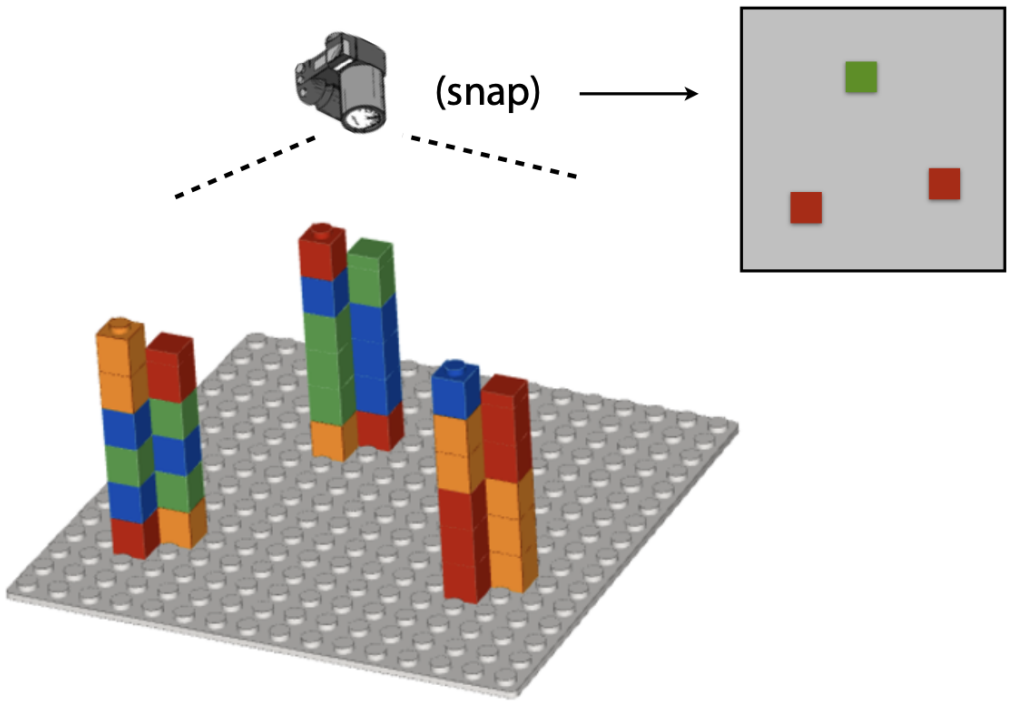
How Second Generation Sequencing Works!



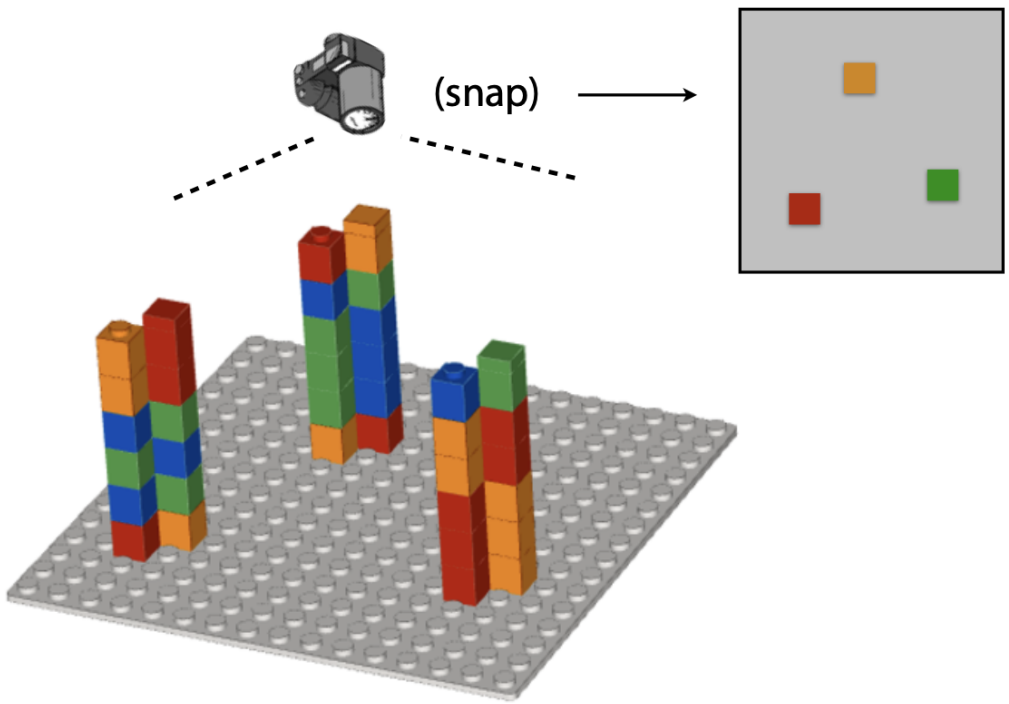
How Second Generation Sequencing Works!



How Second Generation Sequencing Works!

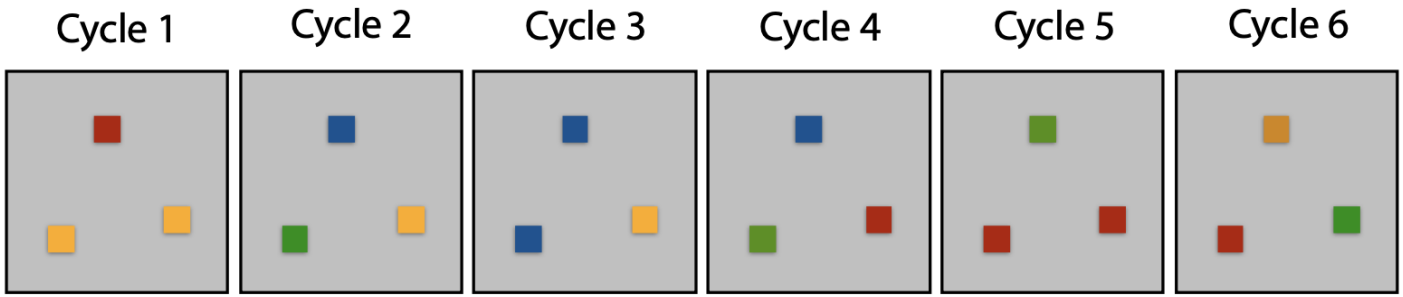


How Second Generation Sequencing Works!



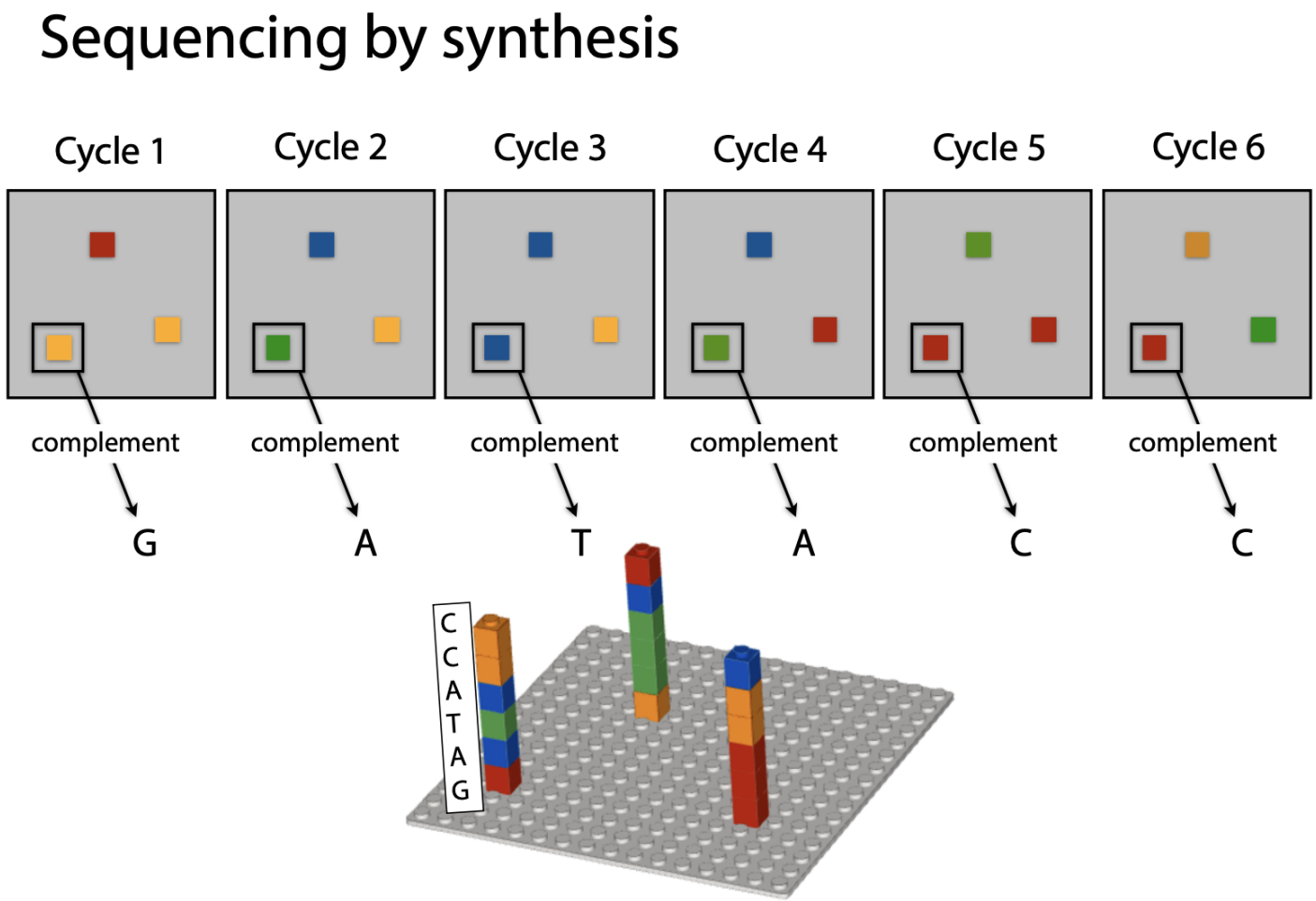
How Second Generation Sequencing Works!

Sequencing by synthesis



BRAC UNIVERSITY
Inspiring Excellence

How Second Generation Sequencing Works!



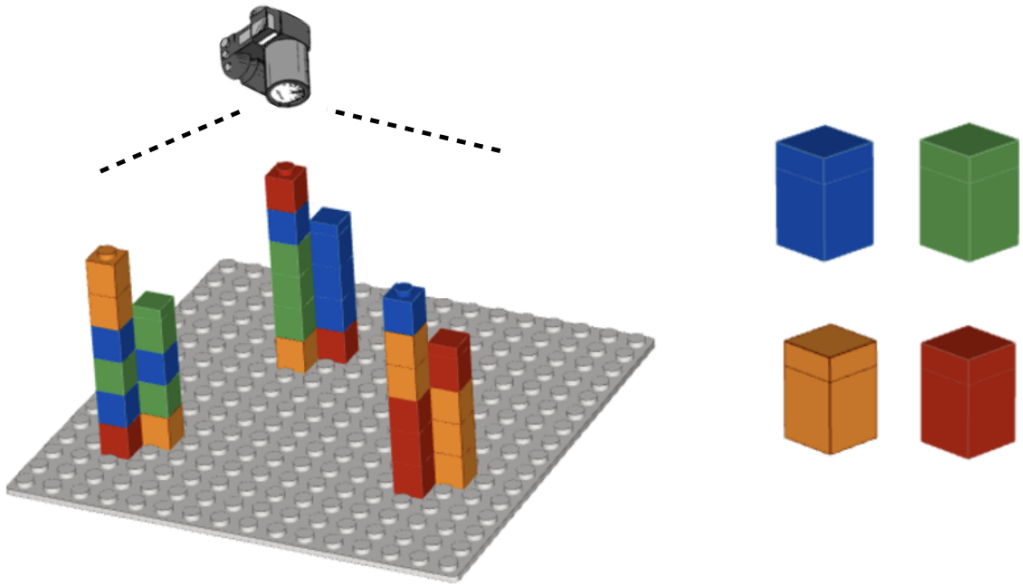
How Second Generation Sequencing Works!

Sequencing by synthesis

Billions of templates on a slide

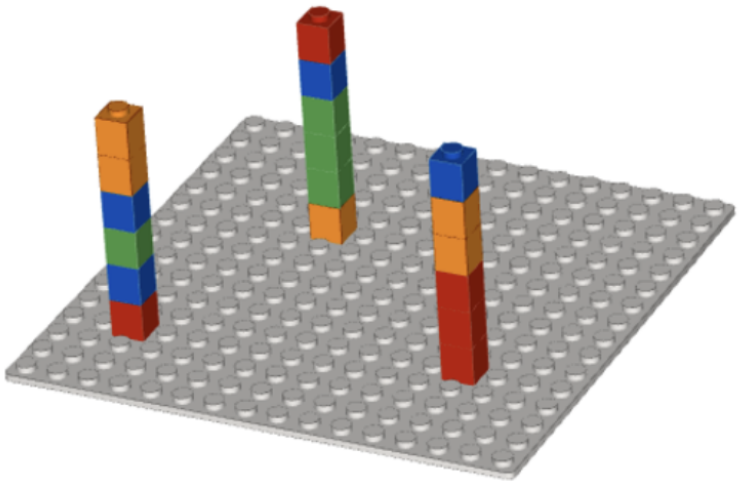
Massively parallel: photograph captures all templates simultaneously

Terminators are "speed bumps," keeping reactions in sync



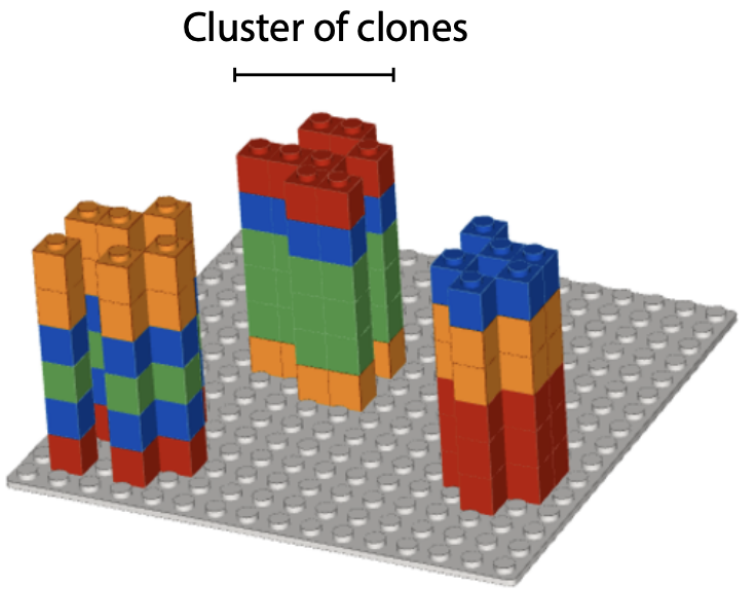
BRAC UNIVERSITY
Inspiring Excellence

Understanding Sequencing Error!

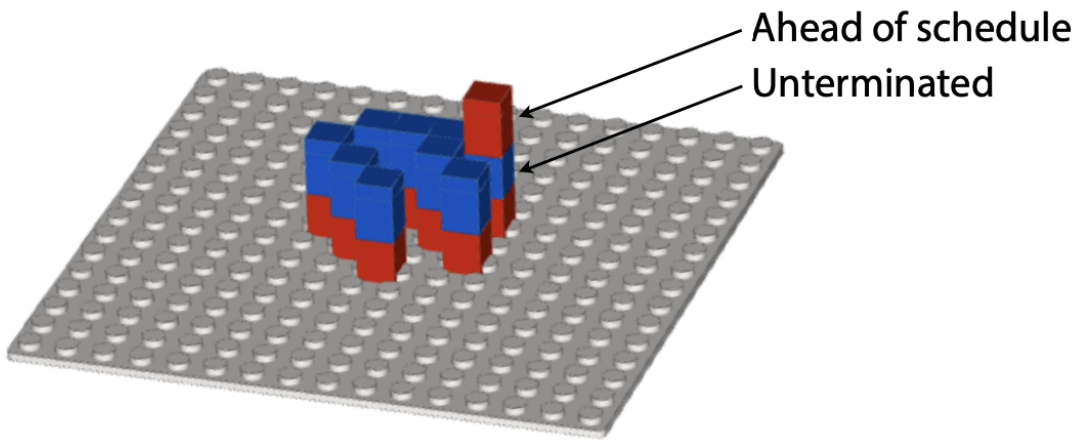


BRAC UNIVERSITY
Inspiring Excellence

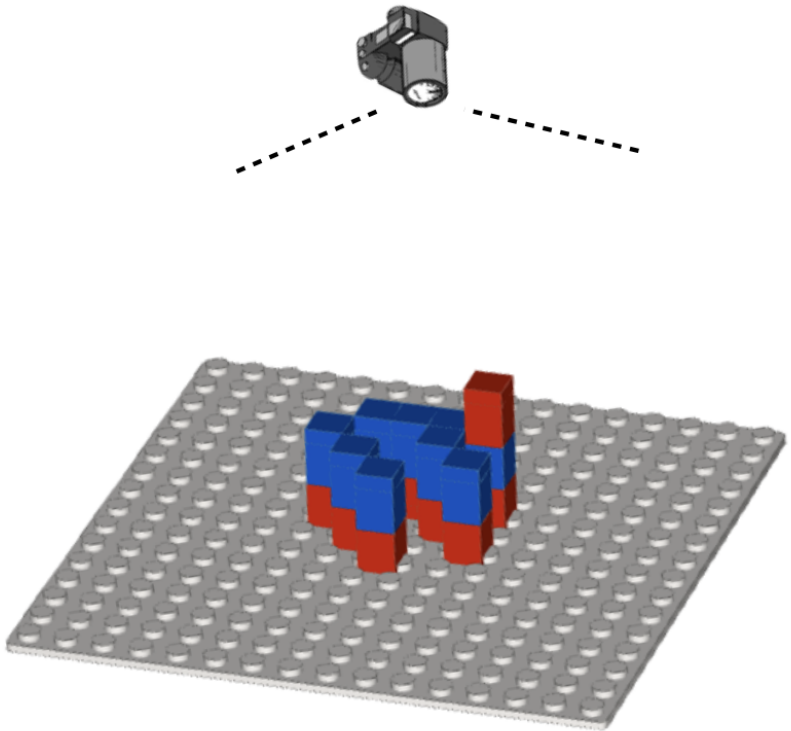
Understanding Sequencing Error!



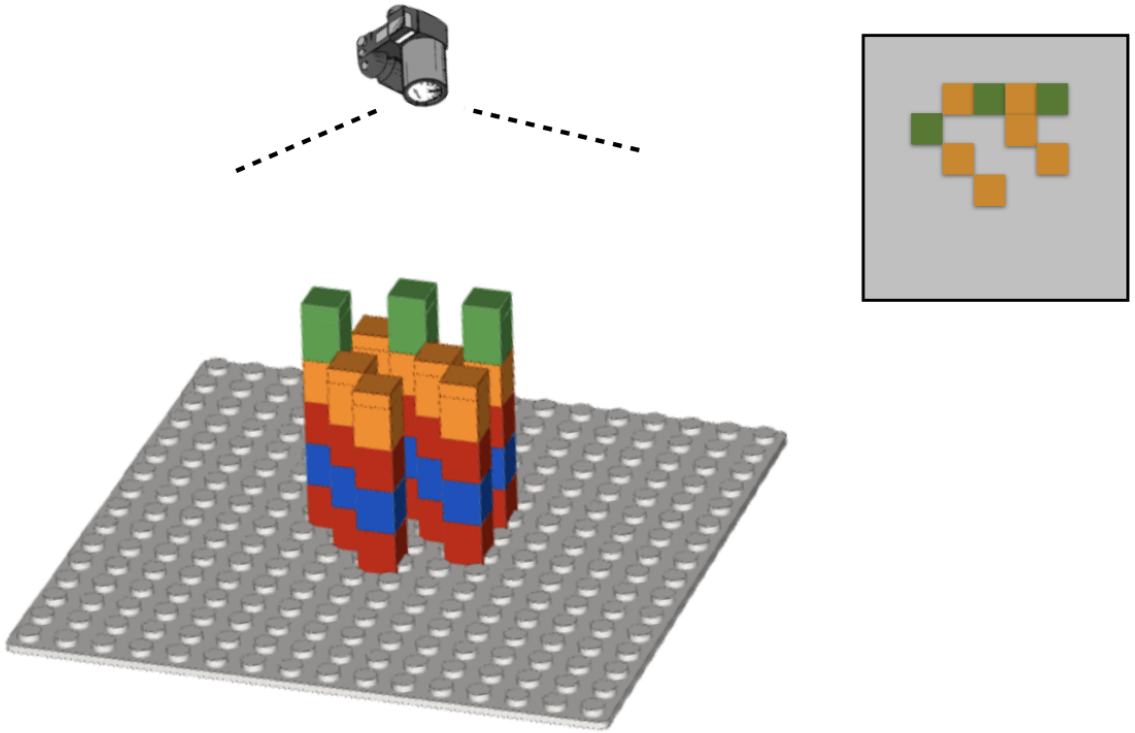
Understanding Sequencing Error!



Understanding Sequencing Error!



Understanding Sequencing Error!



Understanding Sequencing Error!

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

Base quality

Probability that base call is incorrect

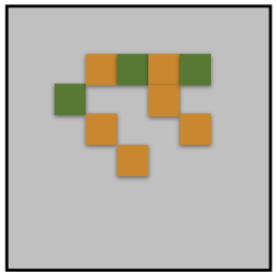
$Q = 10 \rightarrow$ 1 in 10 chance call is incorrect

$Q = 20 \rightarrow$ 1 in 100

$Q = 30 \rightarrow$ 1 in 1,000

BRAC UNIVERSITY
Inspiring Excellence

Understanding Sequencing Error!



Call: orange (C)

Estimate p, probability incorrect:

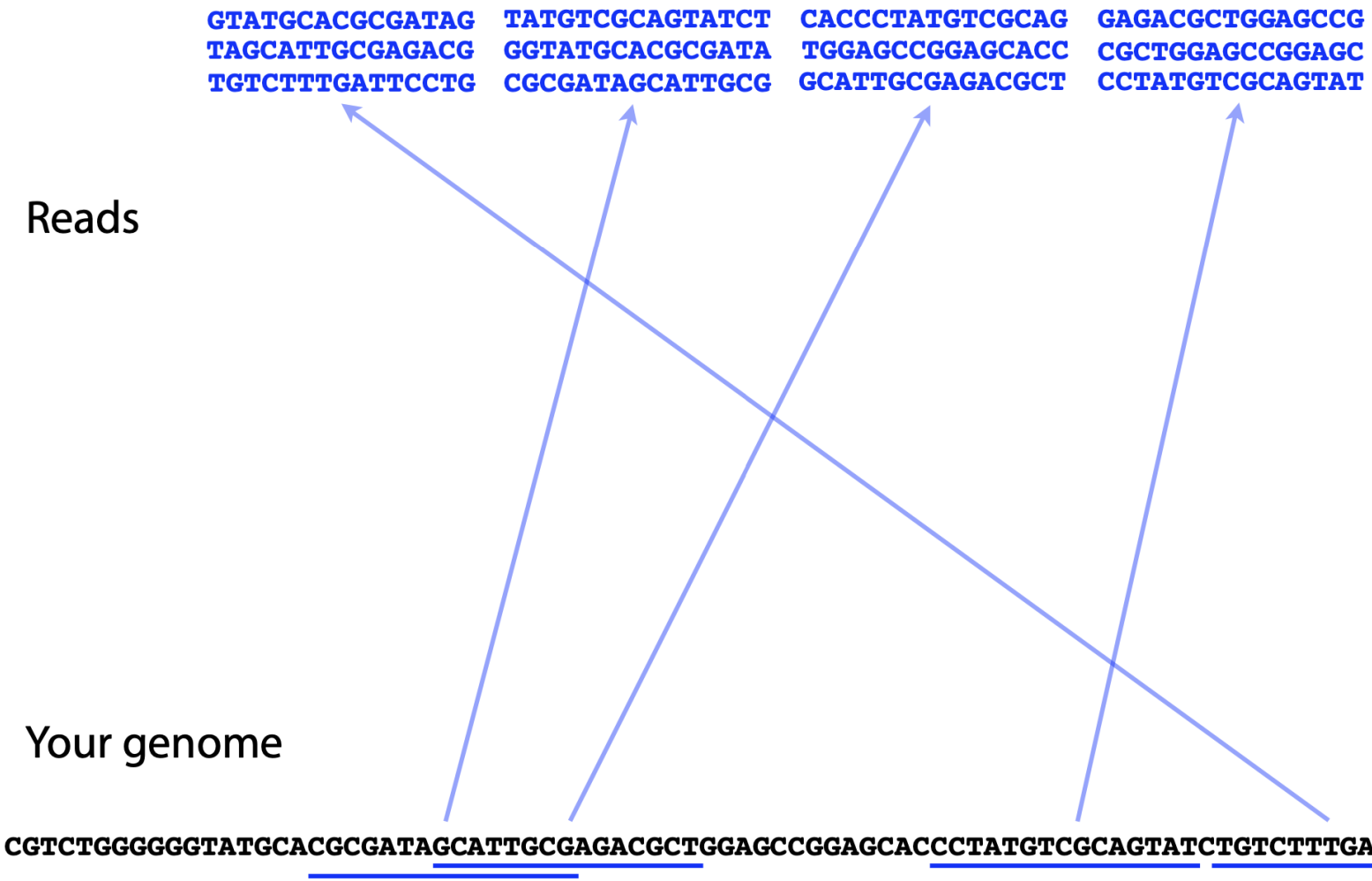
non-orange light / total light

p = 3 green / 9 total = 1/3

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

BRAC UNIVERSITY
Inspiring Excellence

Understanding Reads



Understanding Reads

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

Bases and qualities line up:

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

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

Understanding Reads

- Usual ASCII encoding is "Phred+33":
- take Q, rounded to integer, add 33, convert to character

```
def QtoPhred33(Q):
  """ Turn Q into Phred+33 ASCII-encoded quality """
  return chr(Q + 33)
```

(converts character to integer according to ASCII table)

```
def phred33ToQ(qual):
  """ Turn Phred+33 ASCII-encoded quality into Q """
  return ord(qual)-33
```

(converts integer to character according to ASCII table)

BRAC UNIVERSITY
Inspiring Excellence