DNA Sequencing

Ben Langmead

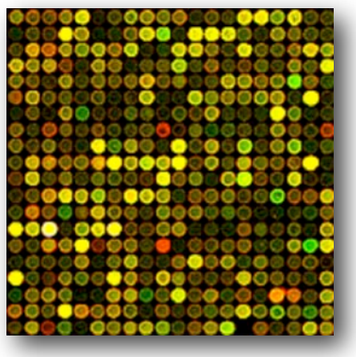
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

Technology for studying nucleic acids

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

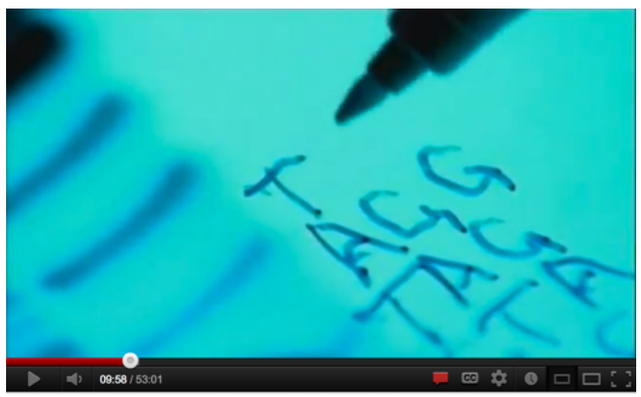
Since ~2010

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sanger (“first generation”) DNA sequencing

Sanger sequencing
1977-1990s

Fred Sanger in episode 3 of PBS documentary “DNA”

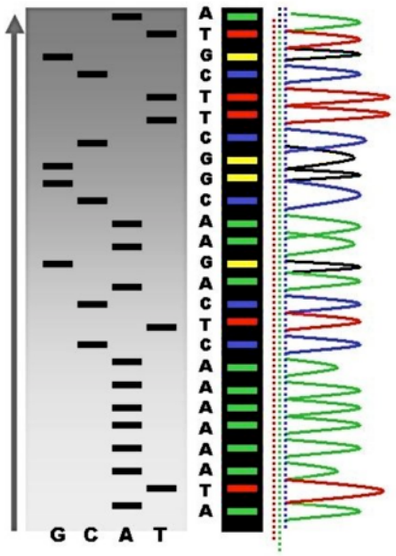


Not-so-high-throughput Sanger sequencing

First practical method invented by Fred Sanger in 1977. Initially used to sequence shorter genomes, e.g. viral genomes 10,000s of bases long.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNA sequencing: Human Genome Project



Improvements both in technology and in algorithms allowed it to eventually be used for human genome project.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing

No sequencing technology yet invented can read much more than 10,000 nucleotides at a time with reasonable cost, throughput, accuracy

Instead, there's a vigorous race to see whose sequencer can read "short" fragments of DNA (around 100s of nucleotides) with best cost, throughput, accuracy

Decoding DNA With Semiconductors

By NICHOLAS WADE
Published: July 20, 2011

Company Unveils DNA Sequencing Device Meant to Be Portable, Disposable and Cheap

By ANDREW POLLACK
Published: February 17, 2012

Cost of Gene Sequencing Falls, Raising Hopes for Medical Advances

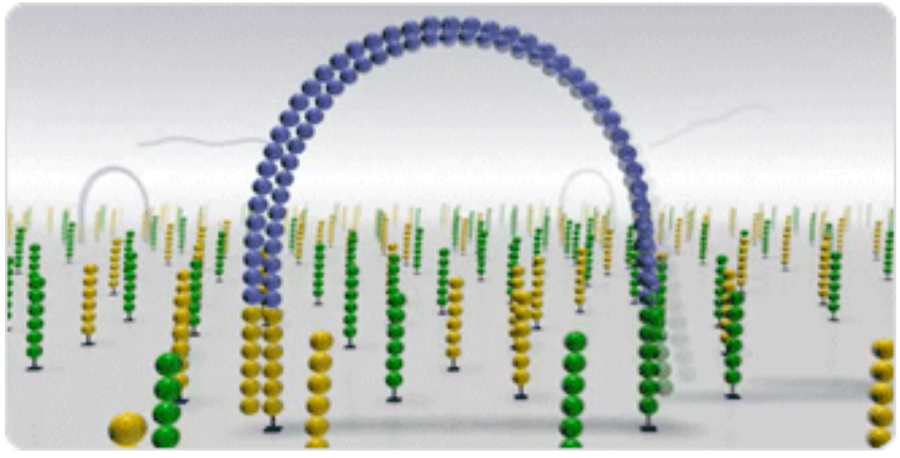
By JOHN MARKOFF
Published: March 7, 2012

Source: nytimes.com

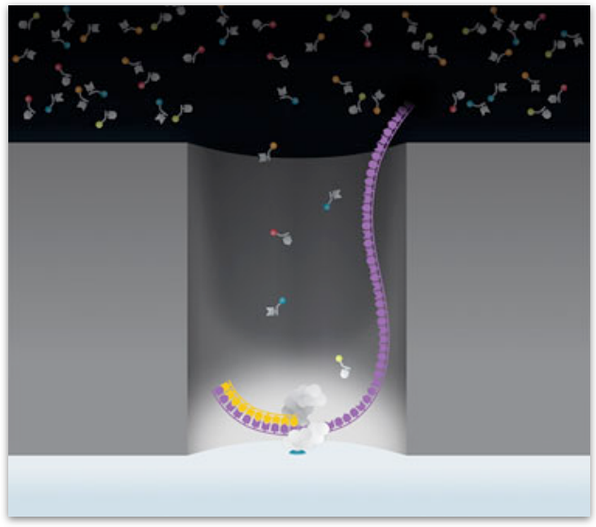
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing

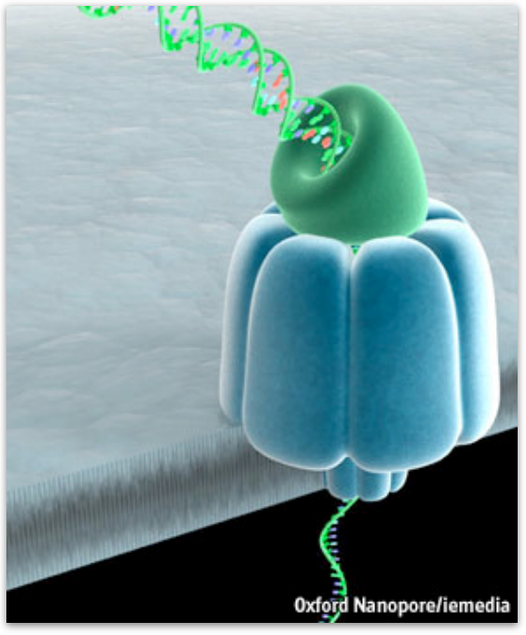
Since 2005, many DNA sequencing instruments have been described and released. They are based on a few different principles



Synthesis / ligation



SMRT cell



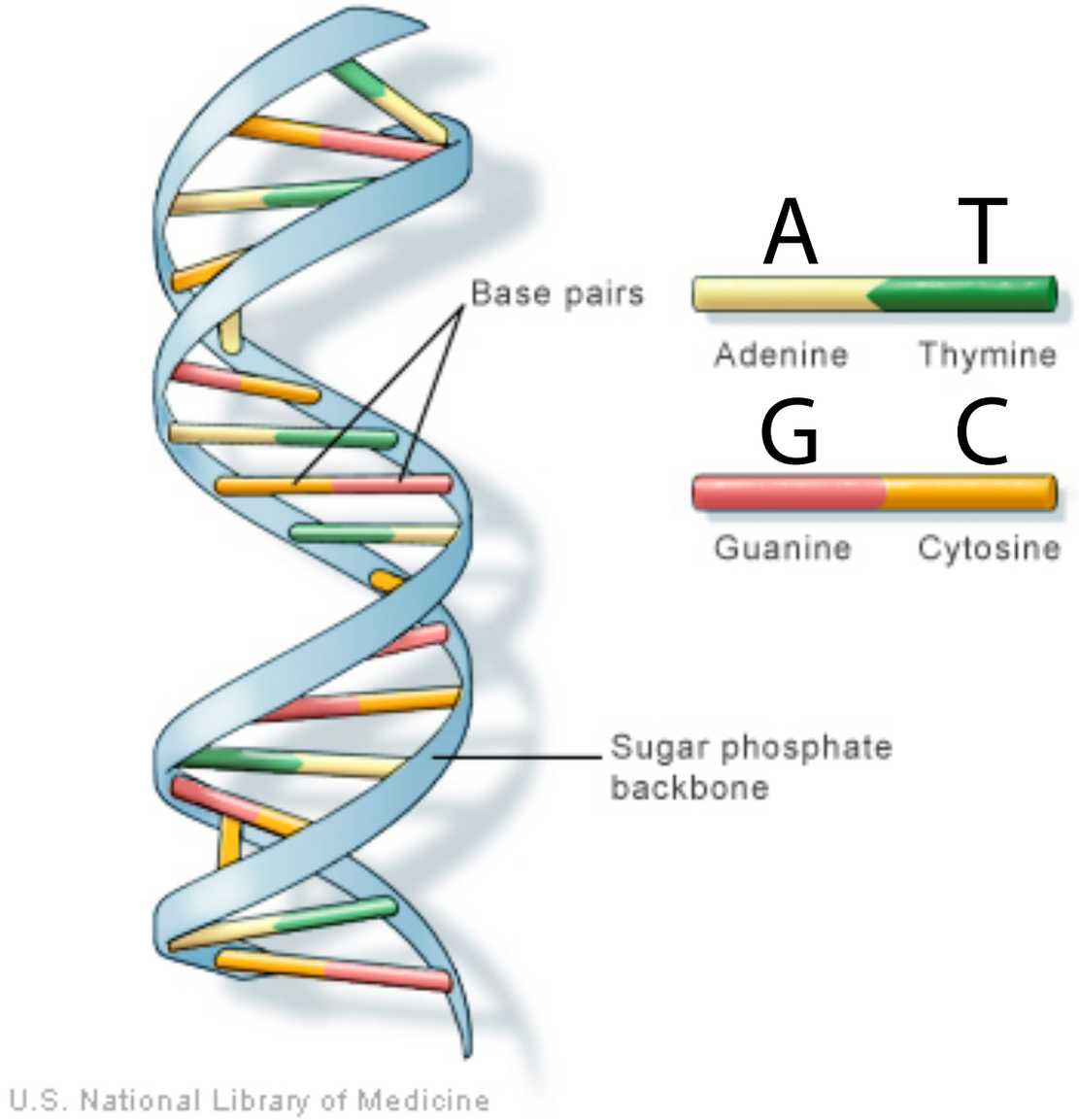
Nanopore

Sequencing by synthesis (“massively parallel sequencing”) provides greatest throughput, and is the most prevalent today

Pictures: http://www.illumina.com/systems/miseq/technology.ilmn, http://www.genengnews.com/gen-articles/third-generation-sequencing-debuts/3257/

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

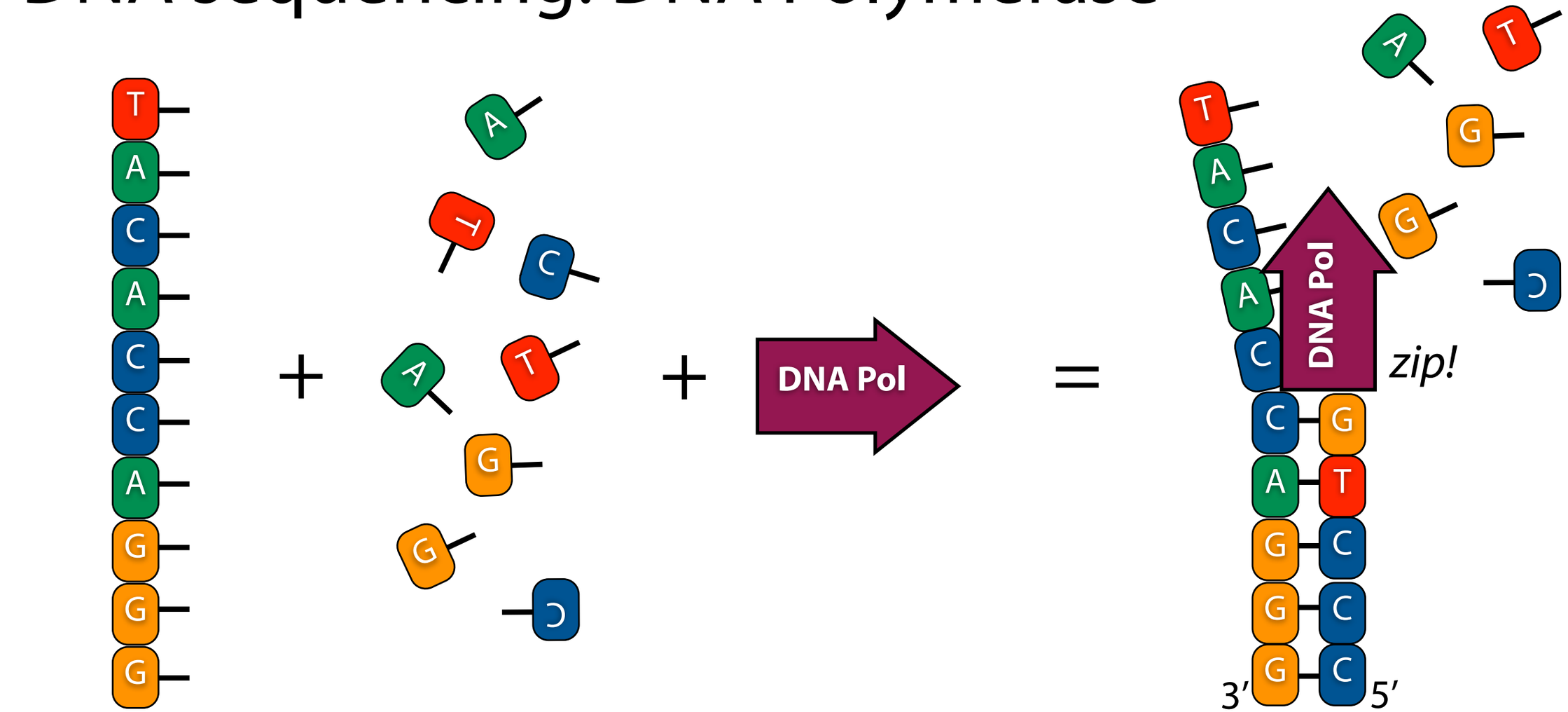
DNA sequencing: double helix



Picture: http://ghr.nlm.nih.gov/handbook/basics/dna

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNA sequencing: DNA Polymerase



Single-stranded DNA *template* | Free nucleotides | DNA polymerase | Strand synthesis

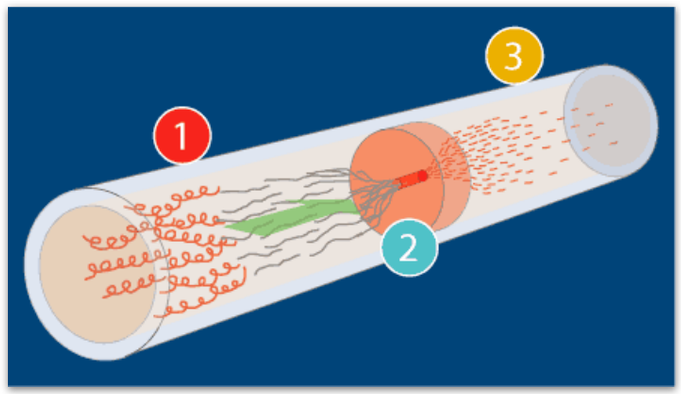
DNA polymerase moves along the template in one direction, integrating complementary nucleotides as it goes

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing by synthesis

1. Take DNA sample, which includes many copies of the genome, and chop it into single-stranded fragments (“templates”)

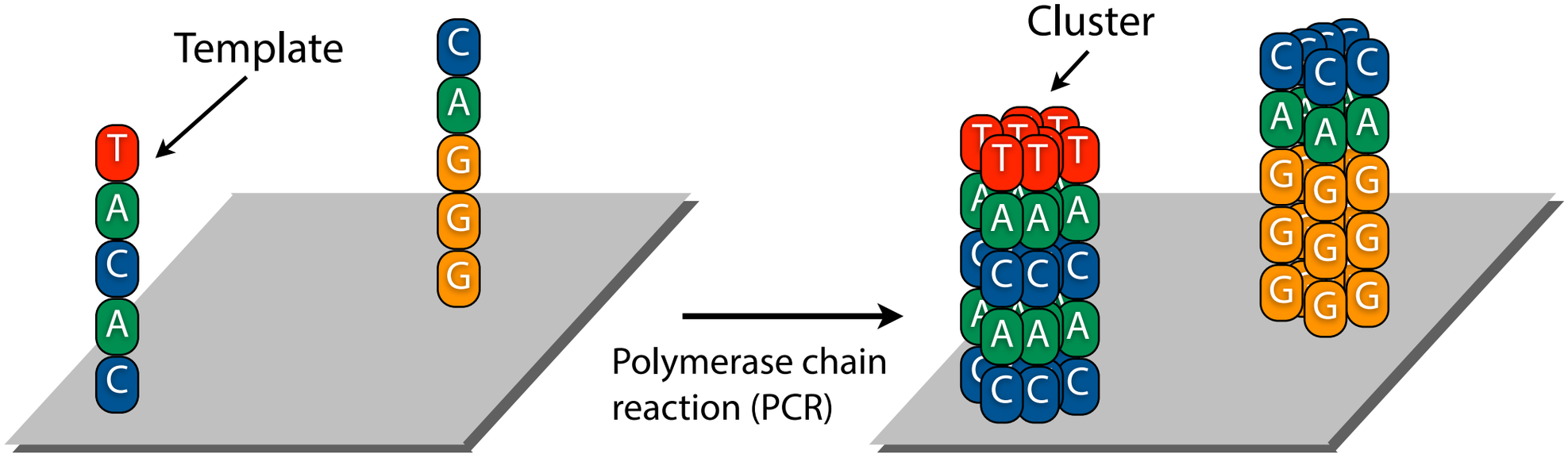
   E.g. with ultrasound waves, water-jet shearing (pictured), divalent cations



Picture: http://www.jgi.doe.gov/sequencing/education/how/how_1.html

2. Attach templates to a surface

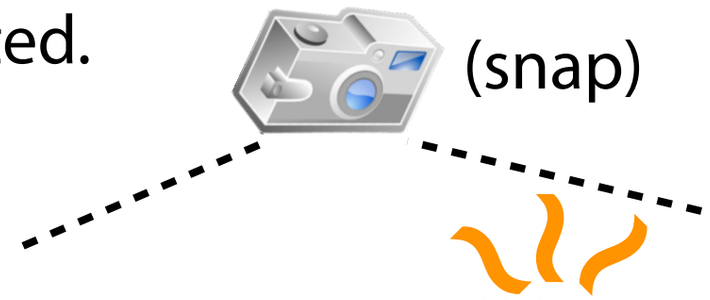
3. Make copies so that each template becomes a “cluster” of clones

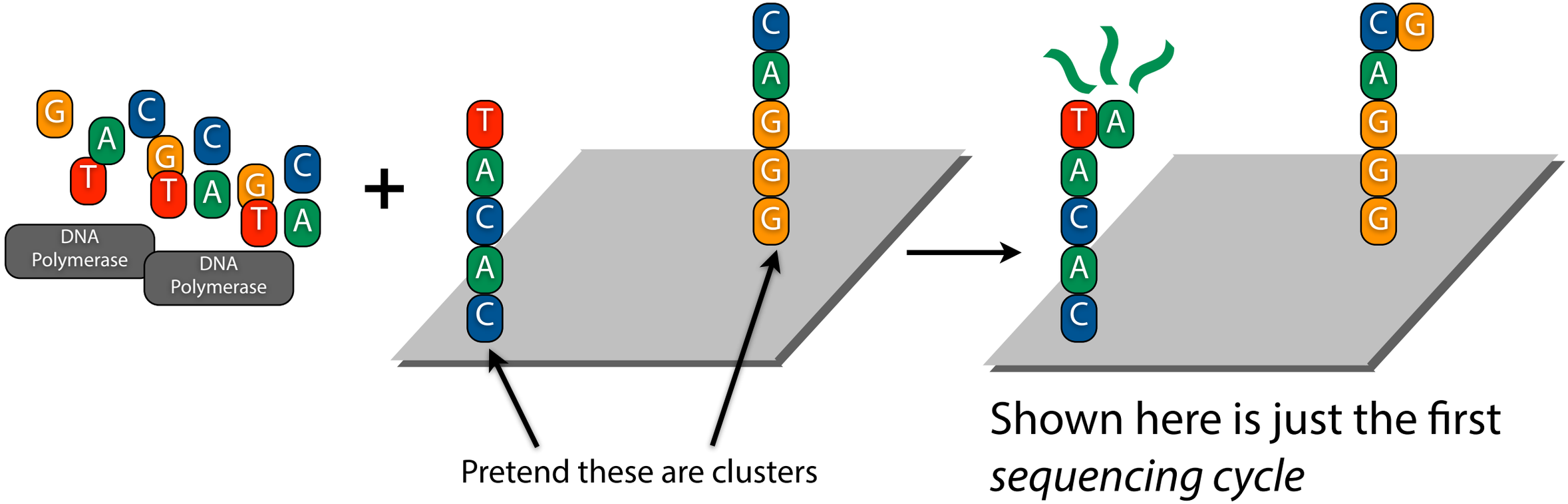


JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Sequencing by synthesis

4. Repeatedly inject mixture of *color-labeled* nucleotides (A, C, G and T) and DNA polymerase. When a complementary nucleotide is added to a cluster, the corresponding color of light is emitted. Capture images of this as it happens.

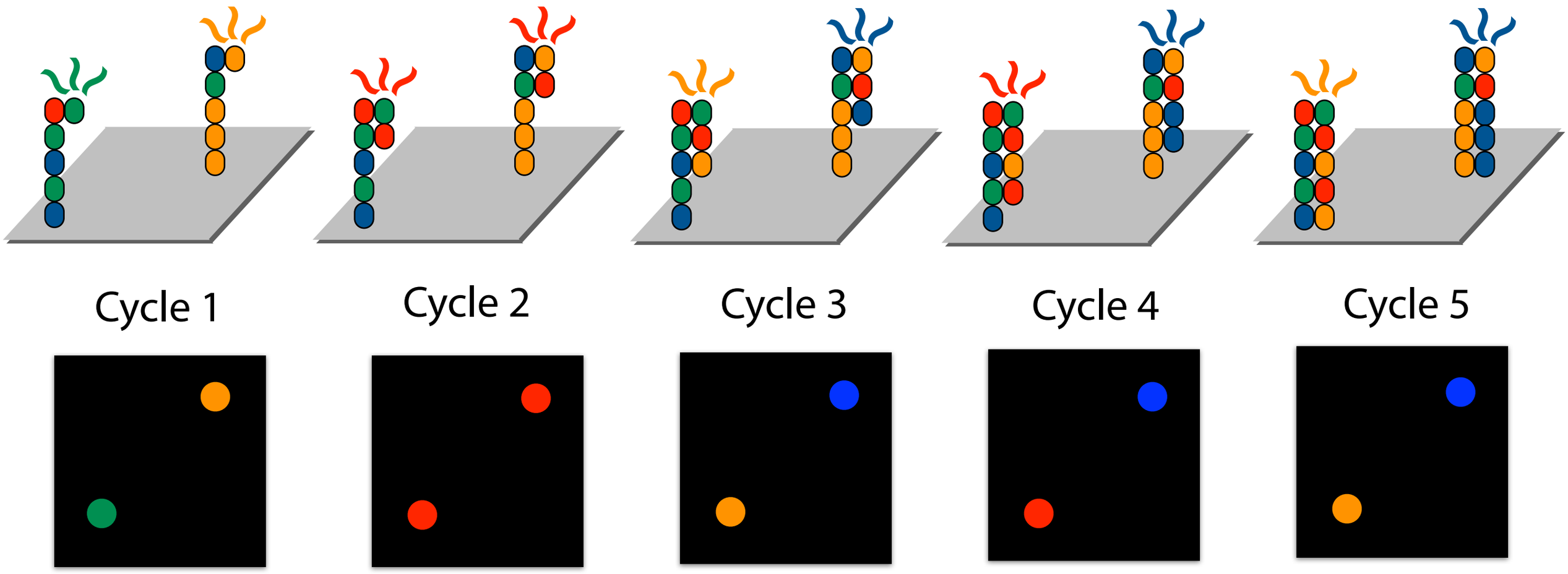




JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

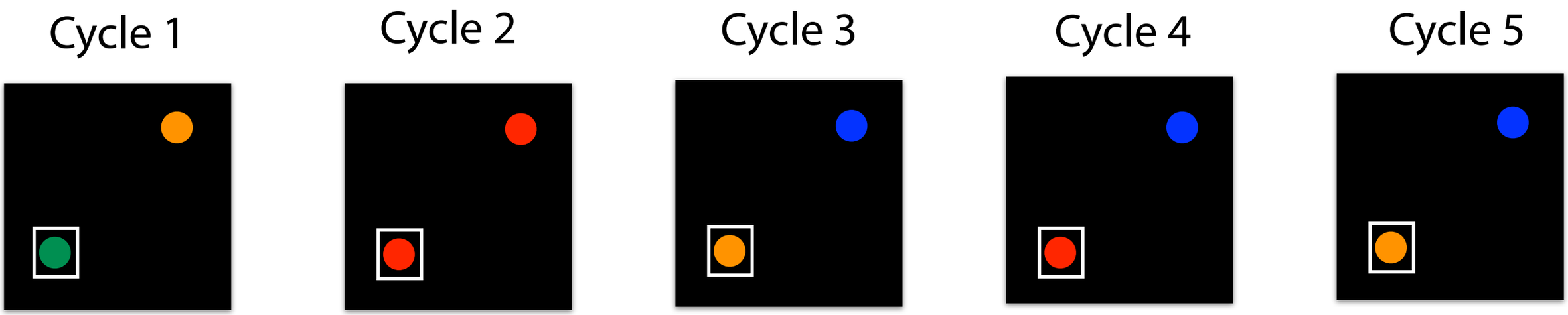
Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides

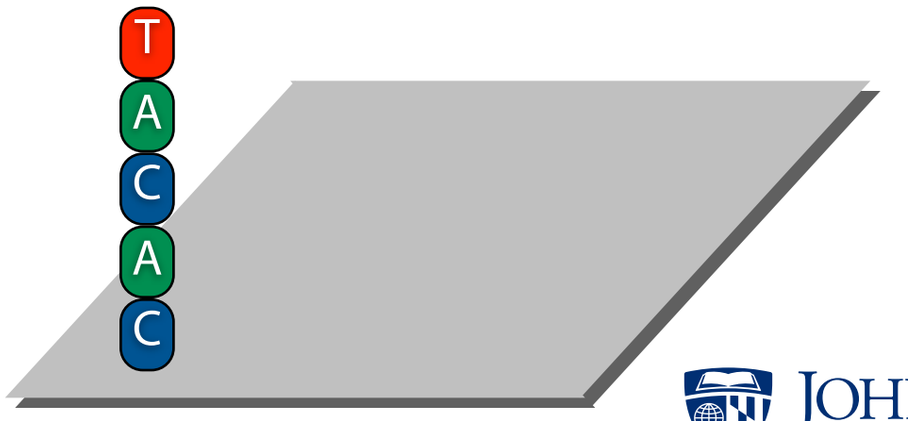


Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



"Base caller" software looks at this cluster across all images and "calls" the complementary nucleotides: TACAC, corresponding to the template sequence



TACAC is a "sequence read," or "read." Actual reads are usually 100 or more nucleotides long.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing by synthesis

A modern sequencing-by-synthesis instrument such as the HiSeq sequences *billions* of clusters simultanously

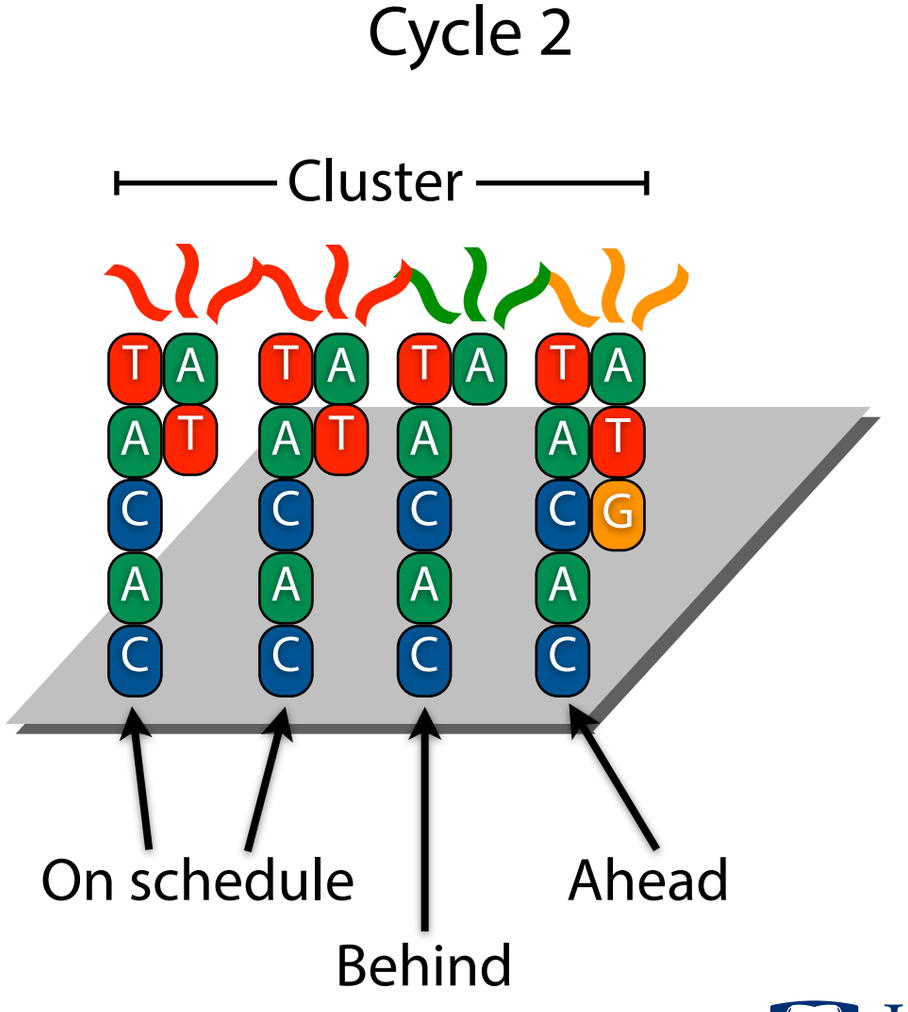
A single "run" takes about 10 days to generate about 600 billion nucleotides of data

Cost of the reagents is $5-10K per run; multiplexing (sequencing many samples per run) further reduces cost per genome

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing by synthesis: errors

Errors creep in when some templates get "out of sync," by missing an incorporation or by incorporating 2 or more nucleotides at once



Base caller must deal with this uncertainty. Actual base callers report a *quality score* (confidence level) along with each nucleotide.

Errors are more common in later sequencing cycles, as proportionally more templates fall out of sync

Sequencing: read format

Below is a FASTQ file. A chunk of 4 lines describes a read. For each read, the 4 lines are (1) read name, (2) nucleotide sequence, (3) (placeholder), (4) quality value sequence.

Read 1	Name
	Nucleotides
	(placeholder)
	Quality values
Read 2	Name
	Nucleotides
	(placeholder)
	Quality values
Read 3	Name
	Nucleotides
	(placeholder)
	Quality values
Read 4	Name
	Nucleotides
	(placeholder)
	Quality values
Read 5	Name
	Nucleotides
	(placeholder)
	Quality values

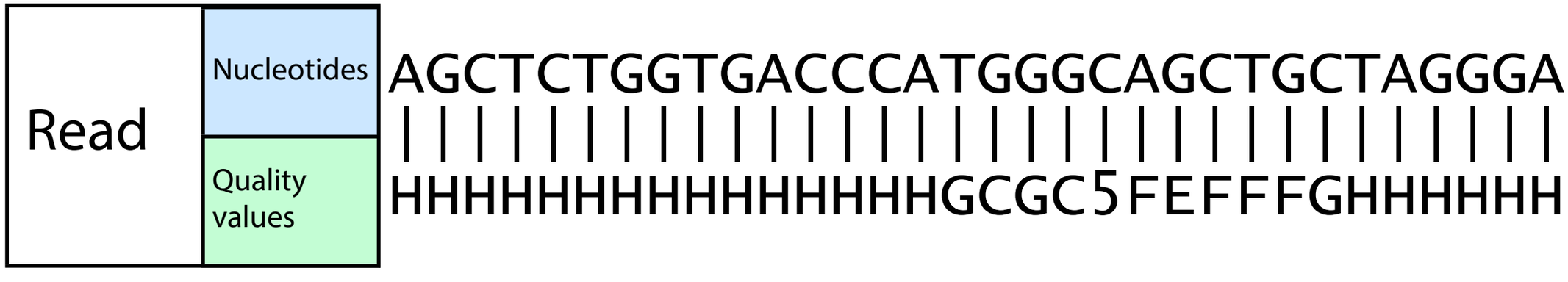
reads — Example — bash — 104×25

```
$ head -20 SRA_HISEQ2000_FC1.shuffle.2M.1.fastq
@509.6.64.20524.149722
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGAGCCTTCTCTCCACCCTGAAAATAGCTTCTGGCTGNTGGGTGAACTATGGAGAGAAAGCGTTTTATTAT
+
HHHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHIHFHHFHHHIHGEHHIIFIHBC#@:@9,--541436D9?;E###############
@509.4.62.19231.2763
GTTGATAAGCAAGCATCTCATTTTGTGCATATACCTGGTCTTTCGTATTCTGGCGTGAAGTCGCCGNCTGAATGCCAGCAATCTCTTTTTGAGTCTCATT
+
HHHHHHHHHHHHHHEHHHHHHHHHHHHHHHHHHHHHHHDHHHHHHGHGHHHHHHHHHH=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
+
HHHHHHHHHHHHHHHHHHHGHDHHHHHFHEHHHGHHGHHHHHHHHHGHHHHHHEHEF<?<@=BBFFFGCFFE?<;@AFG=GA;@D@D?FDFFB=B;F=>AA@
@509.6.25.8102.140546
GGACACATTCAAACCATTGCATCCATCCTCTGCATTCAGAAAGATAGTCCAACAGAAAGATCTGGANTCAAGAGACCCAGCTGATTACCAATTCCAGTTT
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHIHHHIHHHIHG#FFDCDD@@GGGHHFIHEGIFIEIIIIGIIGFGF
$
```

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing: qualities

Nucleotides and quality values line up 1-to-1:



A quality value is an ASCII encoding of a number, Q

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

where p is sequencer's estimate of the probability that the nucleotide at that position was called *incorrectly*

$Q = 10$: error probability is 1 in 10
$Q = 20$: error probability is 1 in 100
$Q = 30$: error probability is 1 in 1,000

Sometimes called "Phred scale"

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

See also: http://en.wikipedia.org/wiki/Fastq

Sequencing: qualities

Typical ASCII encoding is “Phred+33”:

take integer Q, add 33, convert to character

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

(converts character to integer)

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

(converts integer to character)

```
def QtoP(Q):
  """ Turn Phred-scaled integer into error probability """
  return 10.0 ** (-0.1 * Q)
```

(exponentiation)

```
def PtoQ(p):
  """ Turn error probability into Phred-scaled integer """
  import math
  return -10.0 * math.log10(p)
```