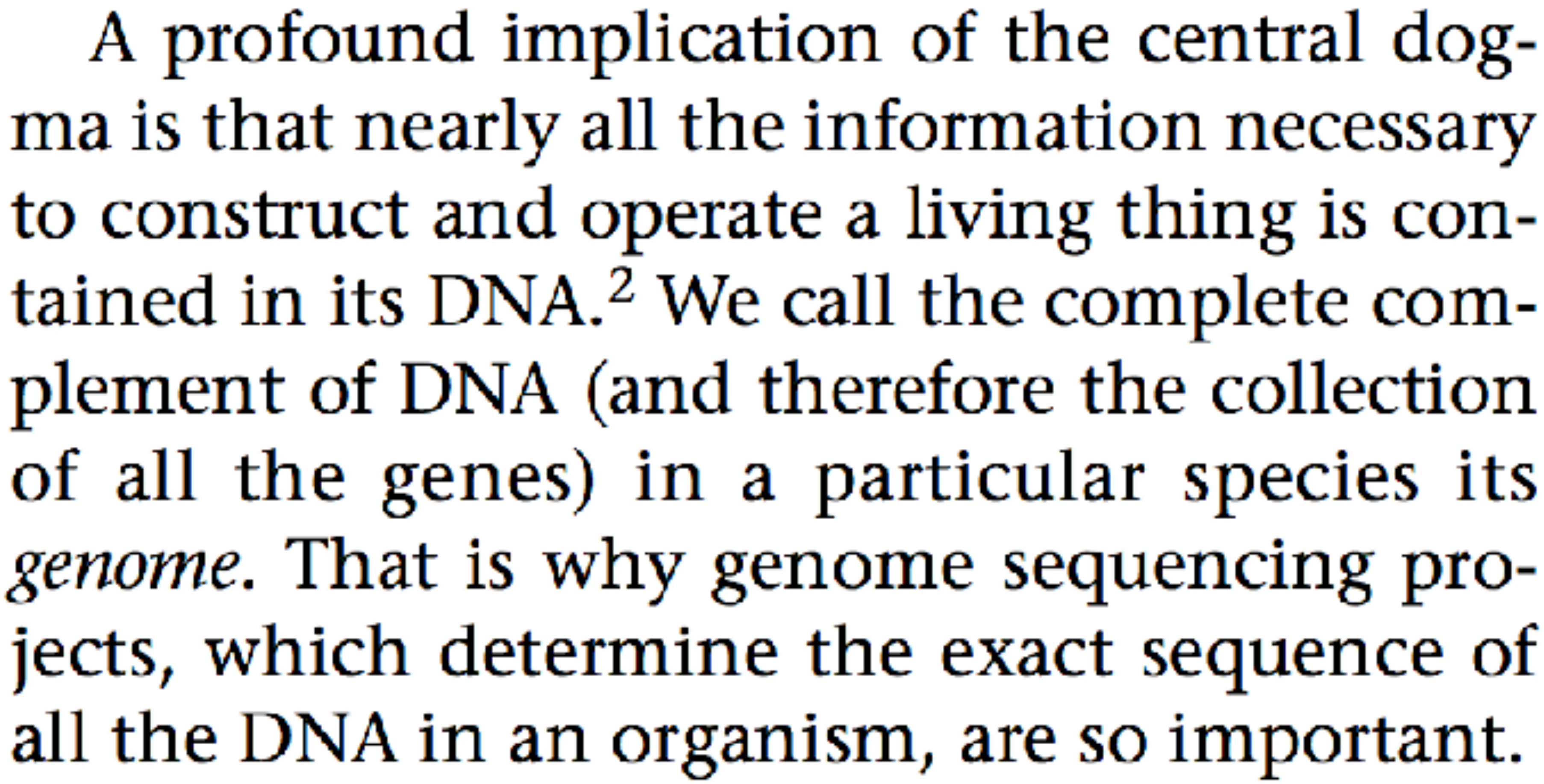
DNA Sequencing

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

DNA



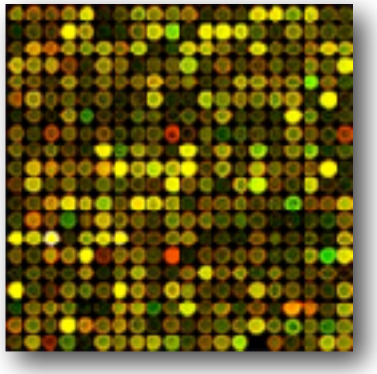
A profound implication of the central dogma is that nearly all the information necessary to construct and operate a living thing is contained in its DNA.[2] We call the complete complement of DNA (and therefore the collection of all the genes) in a particular species its *genome*. That is why genome sequencing projects, which determine the exact sequence of all the DNA in an organism, are so important.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

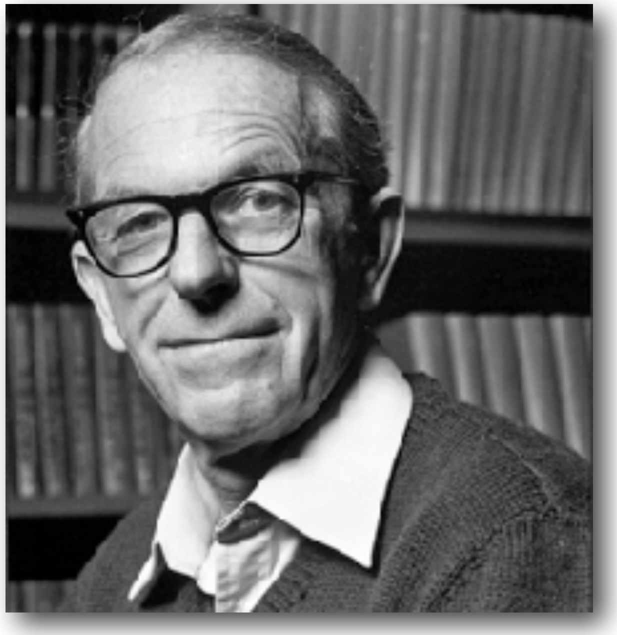
Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010



Fred Sanger
1918-2013

“Chain termination” sequencing

Sanger sequencing

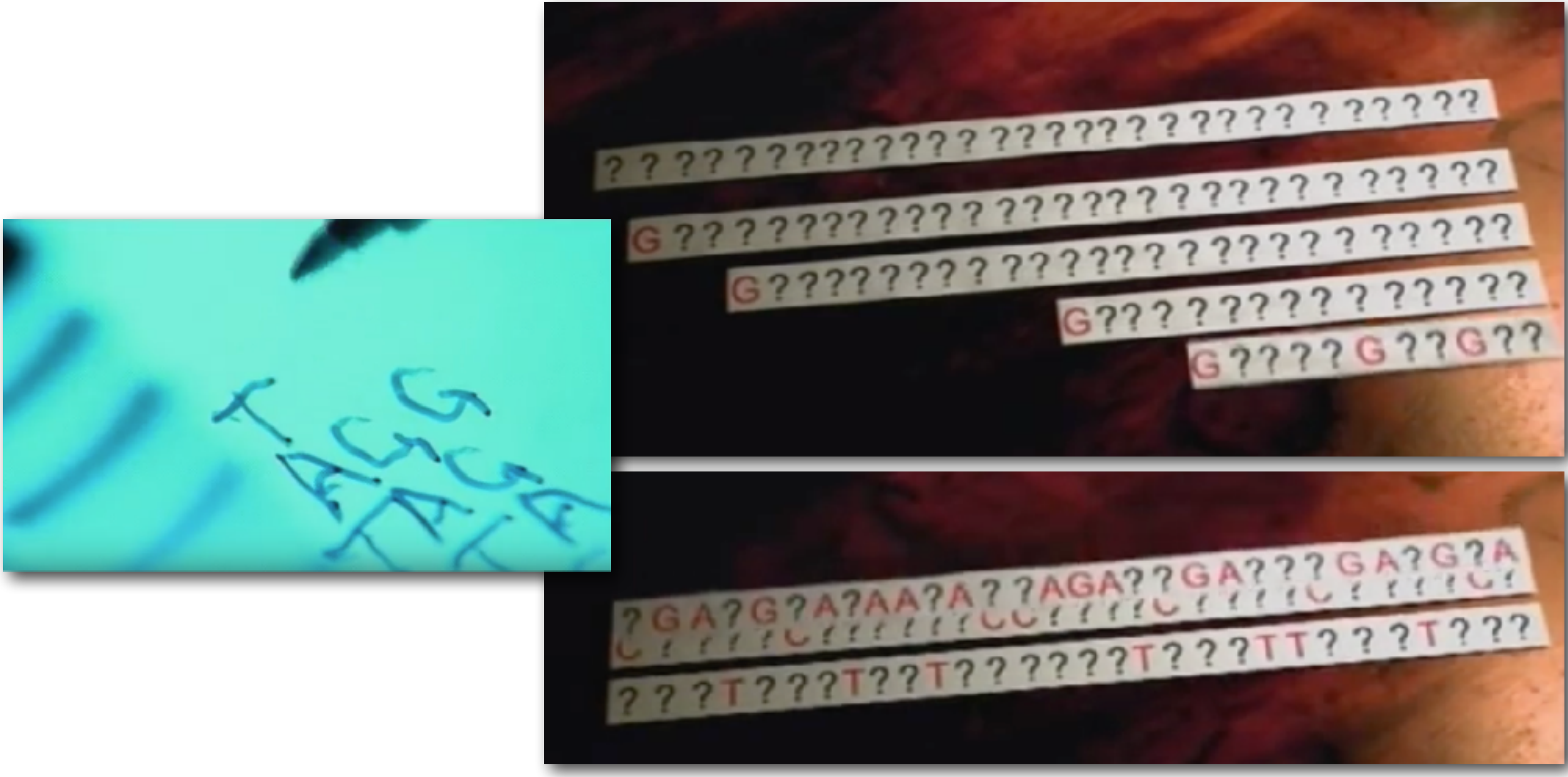
Sanger sequencing
1977-1990s

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.

Sanger sequencing

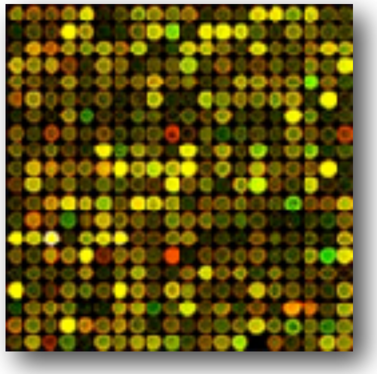

G???????????????????????????????
G??????????????????????????
G????????????
G????G??G??
?GA?G?A?AA?A??AGA??GA???GA?G?A
???T????T???T?????????T???TT???T???

From "DNA" documentary, episode 3

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010

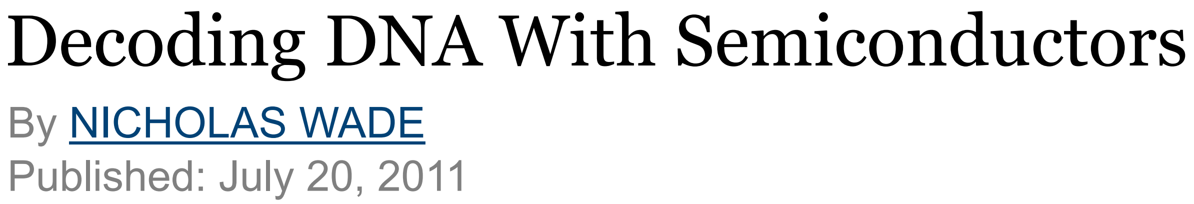
Solexa

illumina®

Sequencing

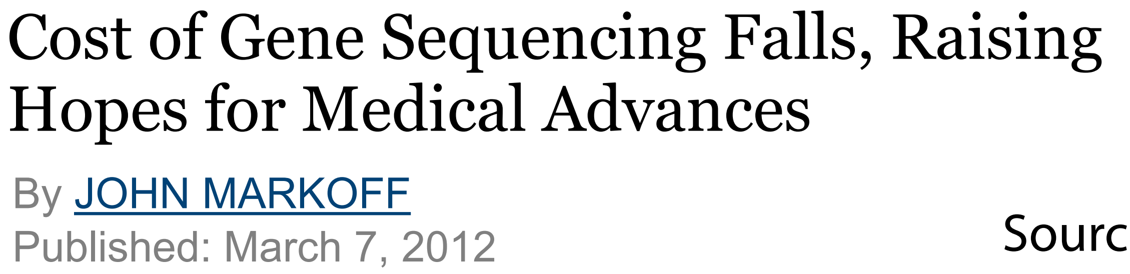
No sequencing technology yet invented can read much more than 10,000 nucleotides at a time with reasonable cost, throughput, accuracy

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



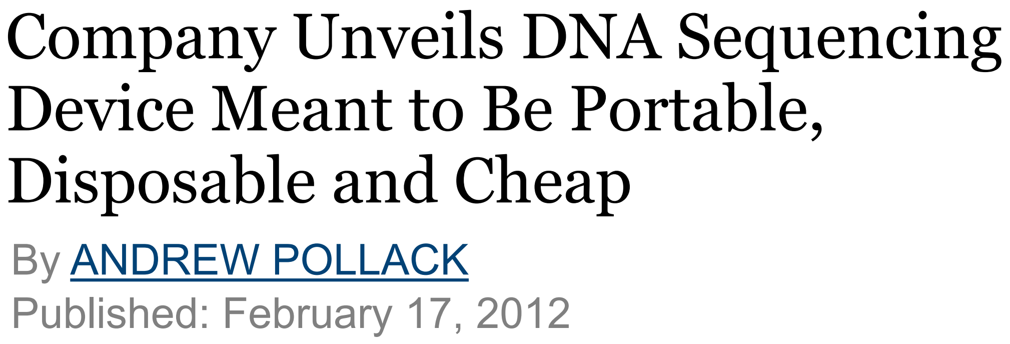
Decoding DNA With Semiconductors

By NICHOLAS WADE
Published: July 20, 2011



Cost of Gene Sequencing Falls, Raising Hopes for Medical Advances

By JOHN MARKOFF
Published: March 7, 2012



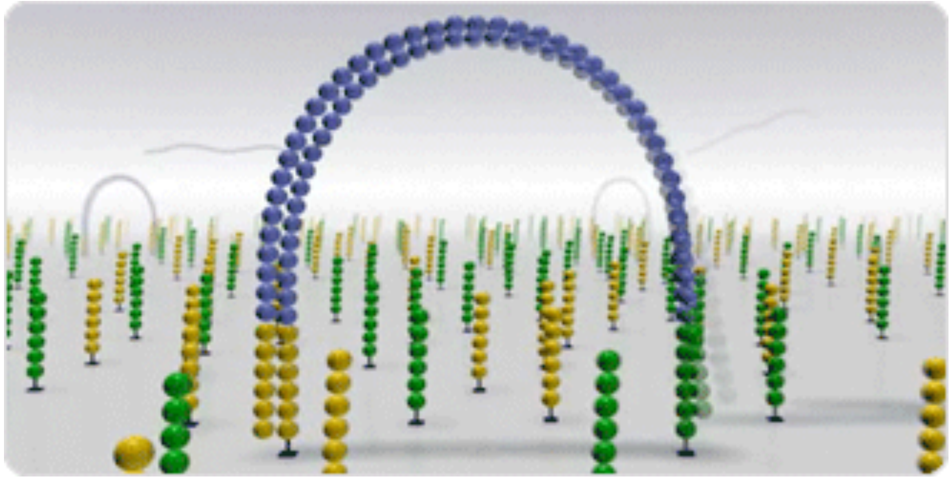
Company Unveils DNA Sequencing Device Meant to Be Portable, Disposable and Cheap

By ANDREW POLLACK
Published: February 17, 2012

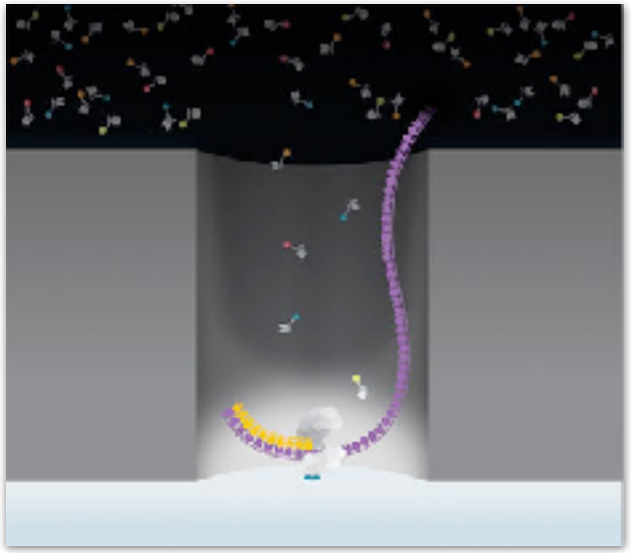
Source: nytimes.com

Sequencing

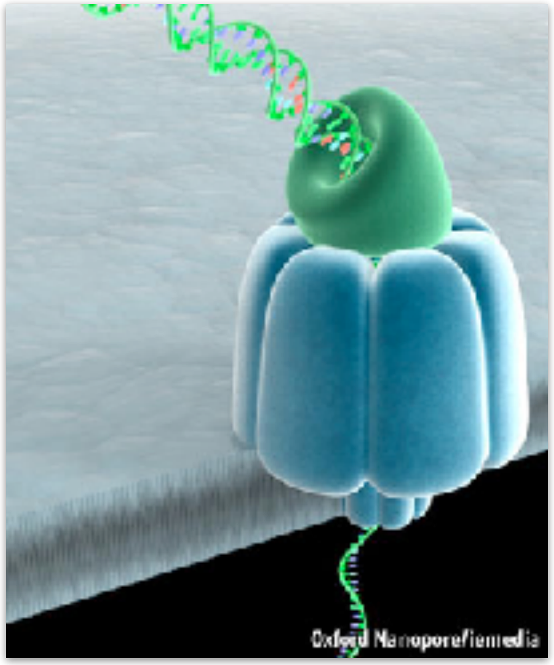
Since 2005, many DNA sequencing instruments have been described and released. They are based on a few different principles



Synthesis / ligation



SMRT cell

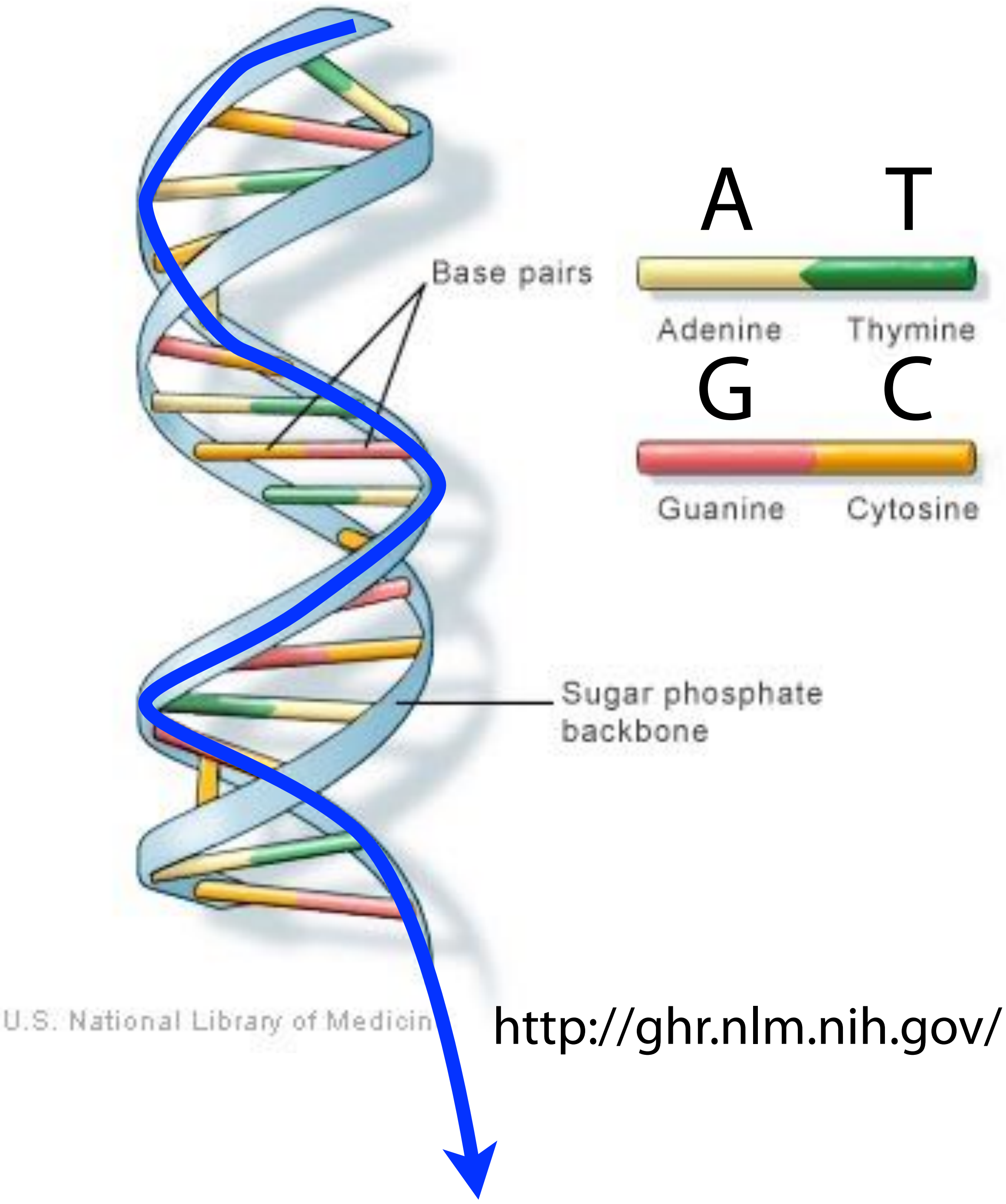


Nanopore

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/

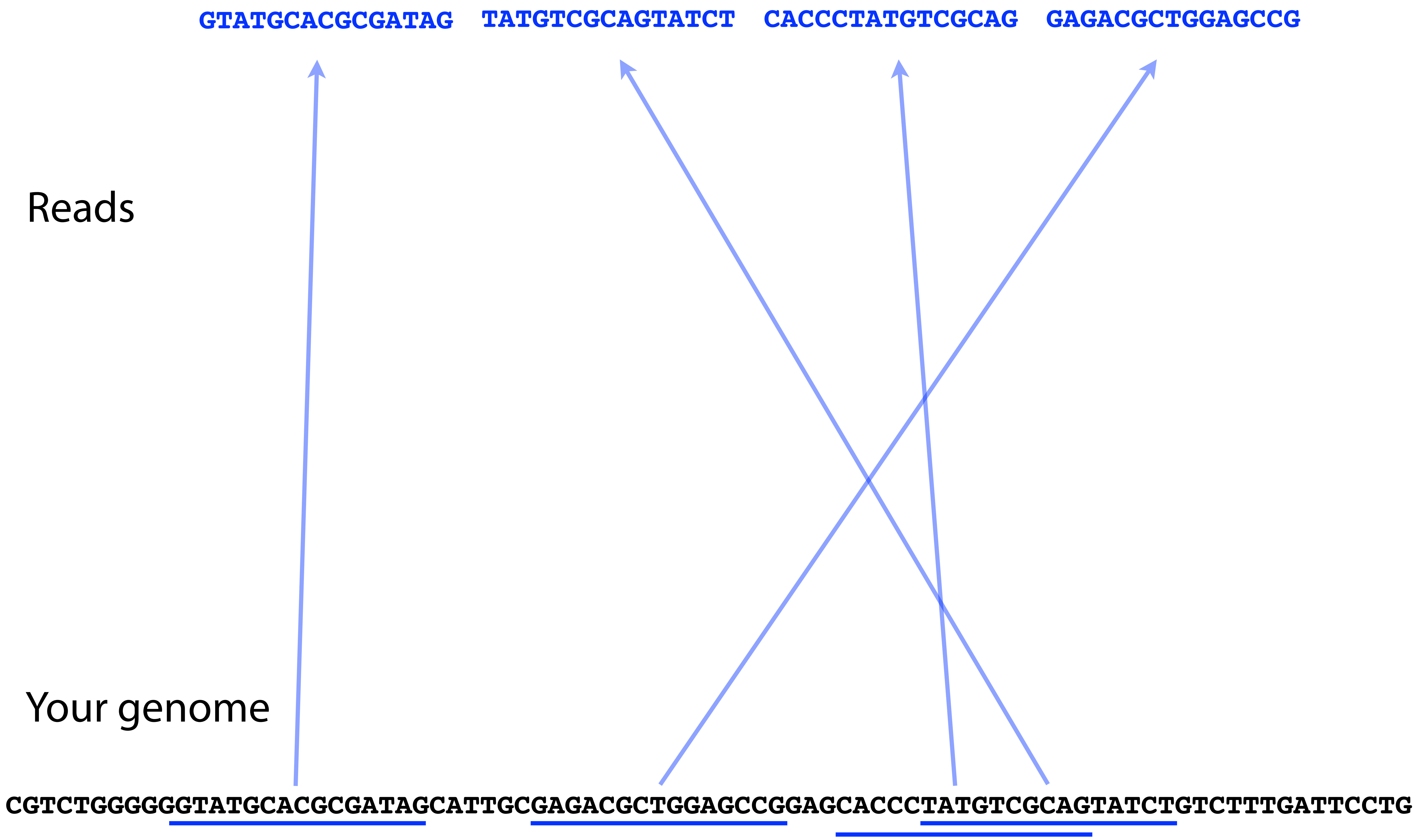
DNA: double helix



U.S. National Library of Medicin

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

TCACACTGAGCGTGCTG

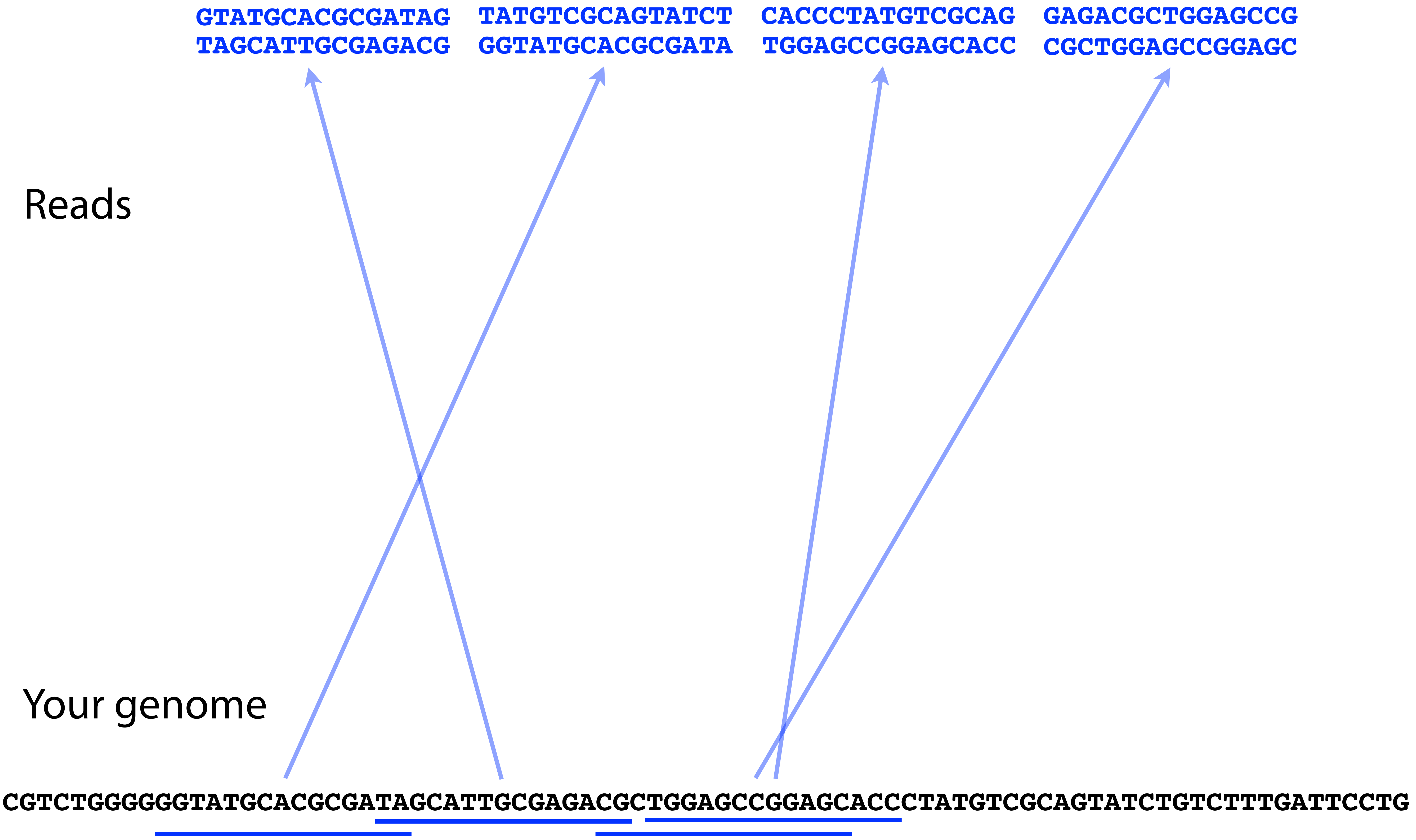


GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG



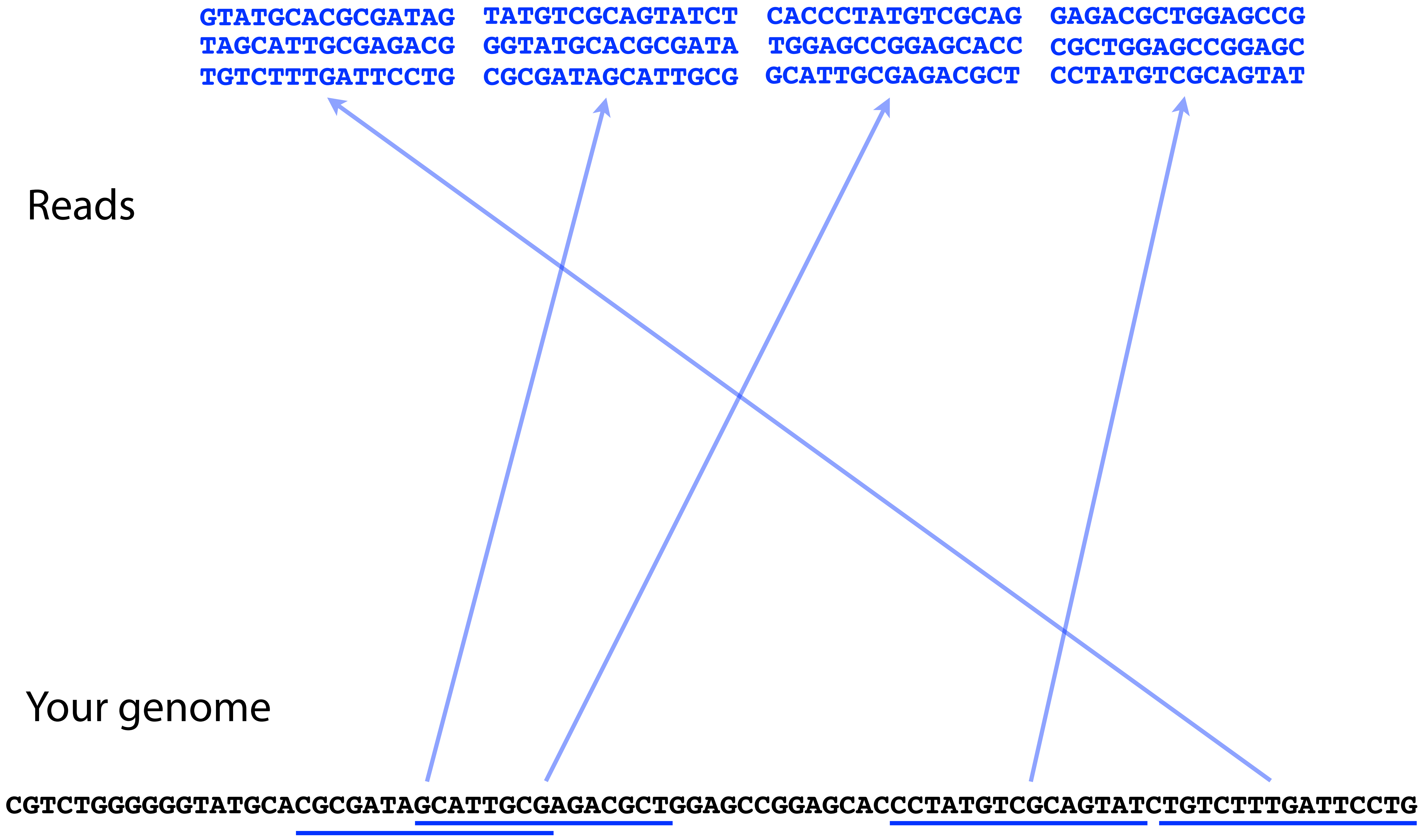
GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG

TAGCATTGCGAGACG GGTATGCACGCGATA TGGAGCCGGAGCACC CGCTGGAGCCGGAGC

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG



```
GTATGCACGCGATAG  TATGTCGCAGTATCT  CACCCTATGTCGCAG  GAGACGCTGGAGCCG
TAGCATTGCGAGACG  GGTATGCACGCGATA  TGGAGCCGGAGCACC  CGCTGGAGCCGGAGC
TGTCTTTGATTCCTG  CGCGATAGCATTGCG  GCATTGCGAGACGCT  CCTATGTCGCAGTAT
```

Reads

Your genome

```
CGTCTGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG
```

Reads

GTATGCACGCGATAG
TAGCATTGCGAGACG
TGTCTTTGATTCCTG
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA
GTATGCACGCGATAG
GCGAGACGCTGGAGC
CCTACGTTCAATATT
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA

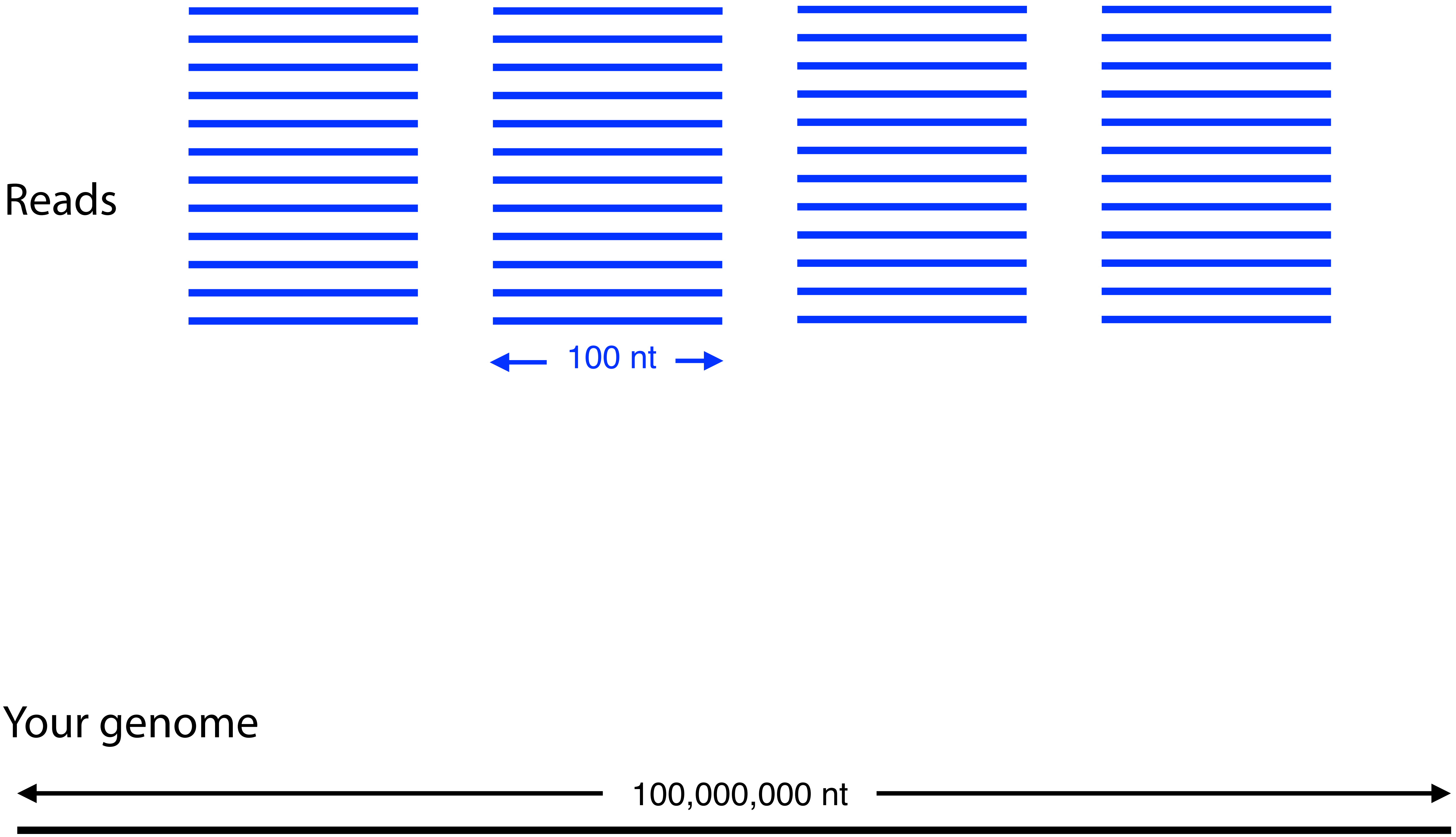
TATGTCGCAGTATCT
GGTATGCACGCGATA
CGCGATAGCATTGCG
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT
ACCTACGTTCAATAT
CTATCACCCTATTAA
GCACCTACGTTCAAT
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT

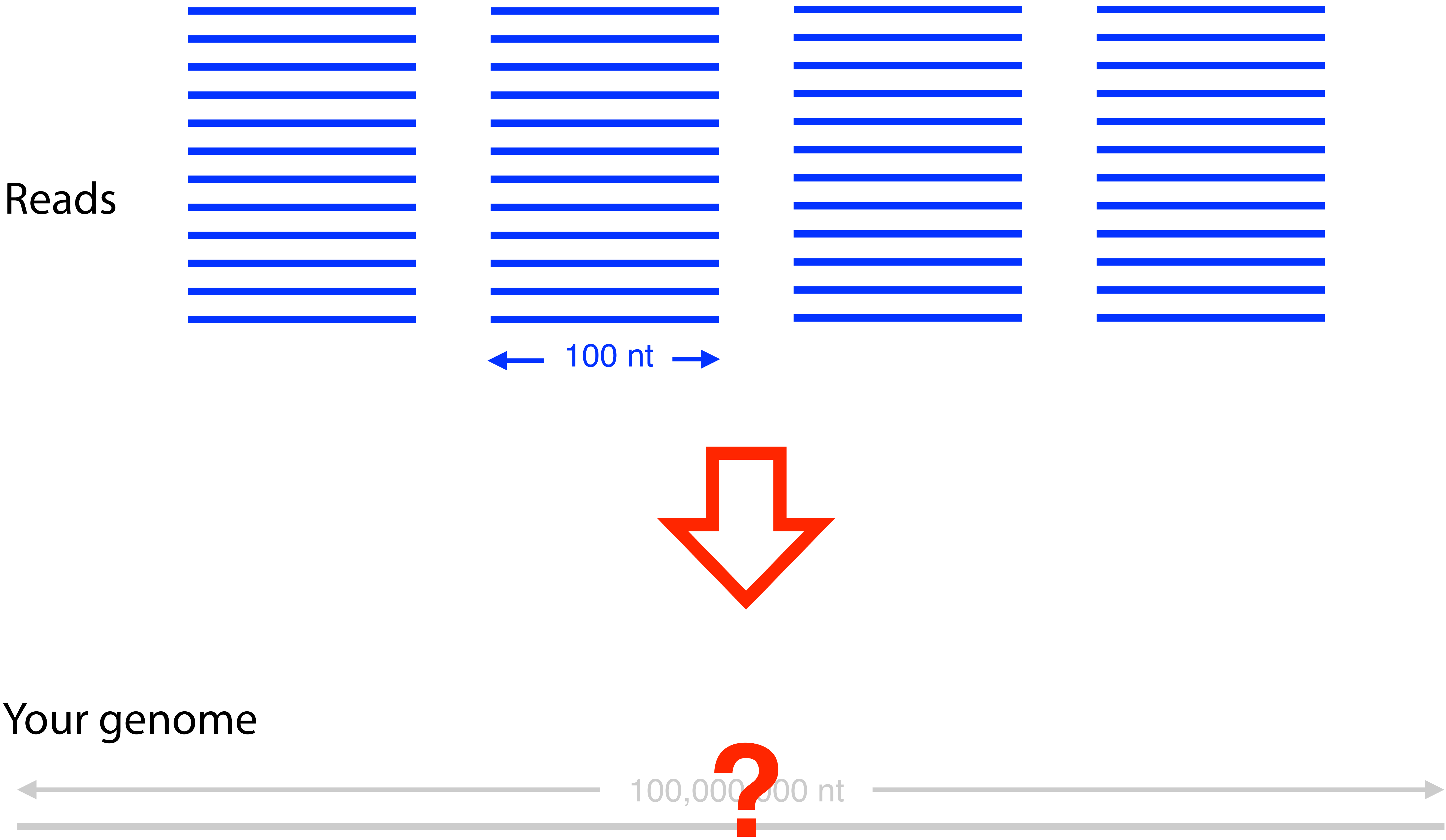
CACCCTATGTCGCAG
TGGAGCCGGAGCACC
GCATTGCGAGACGCT
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGTATG
TATTTATCGCACCTA
CTGTCTTTGATTCCT
GTCTGGGGGTATGC
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGTATG

GAGACGCTGGAGCCG
CGCTGGAGCCGGAGC
CCTATGTCGCAGTAT
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG
CCACTCACGGGAGCT
ACTCACGGGAGCTCT
AGCCGGAGCACCCTA
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG

Your genome

CGTCTGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG


Reads
100 nt
Your genome
100,000,000 nt



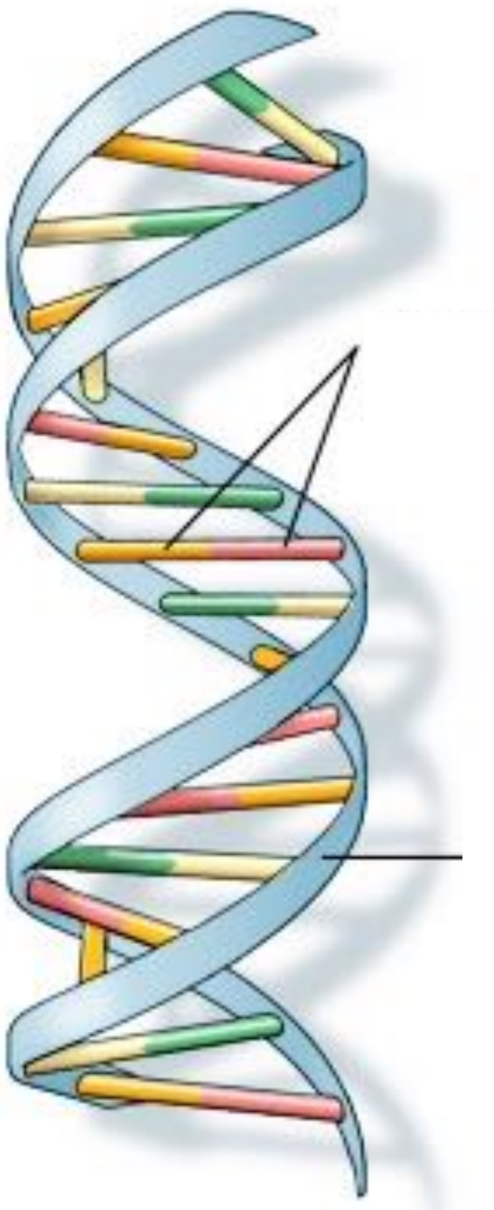
Reads

100 nt

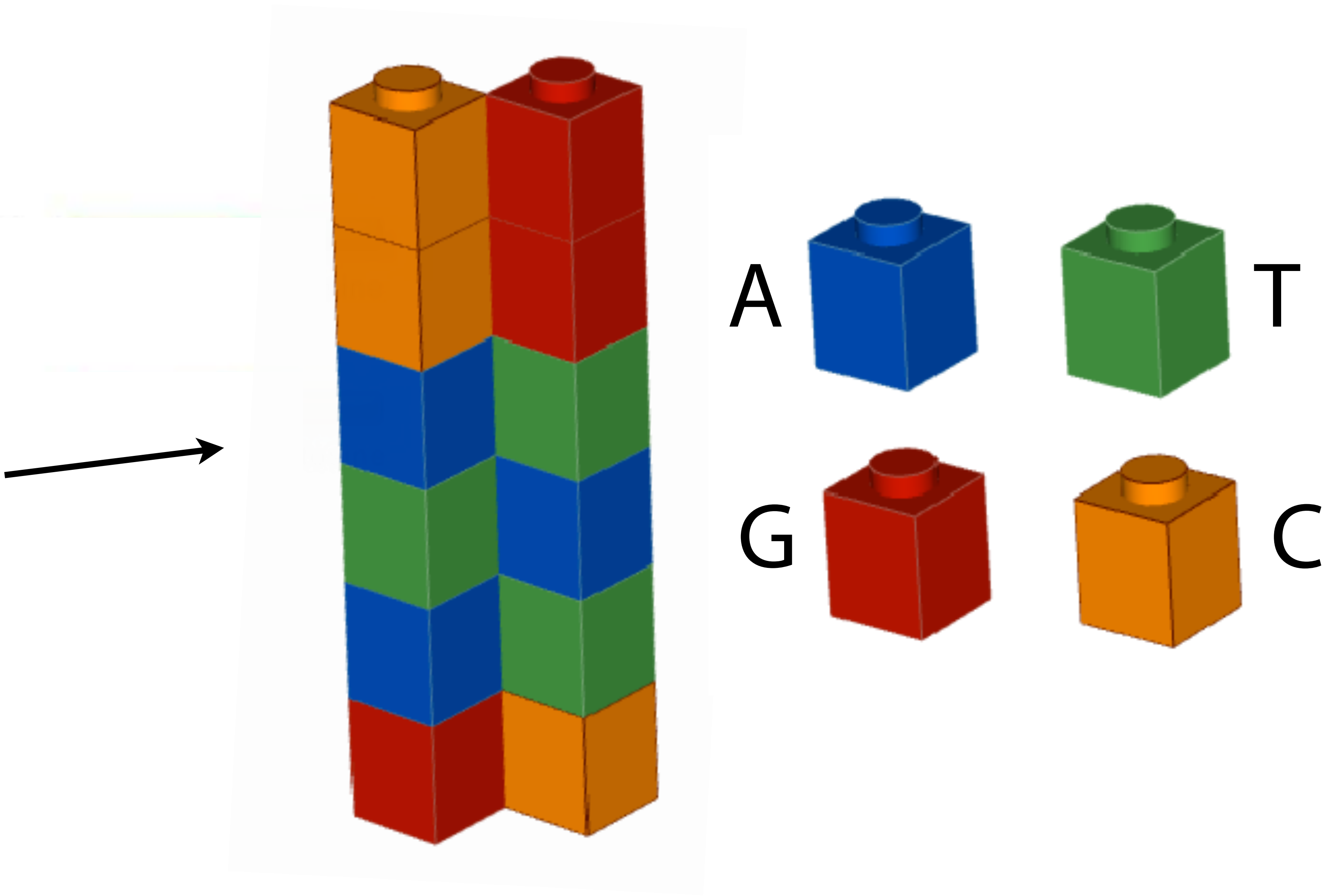
Your genome

100,000,000 nt

?

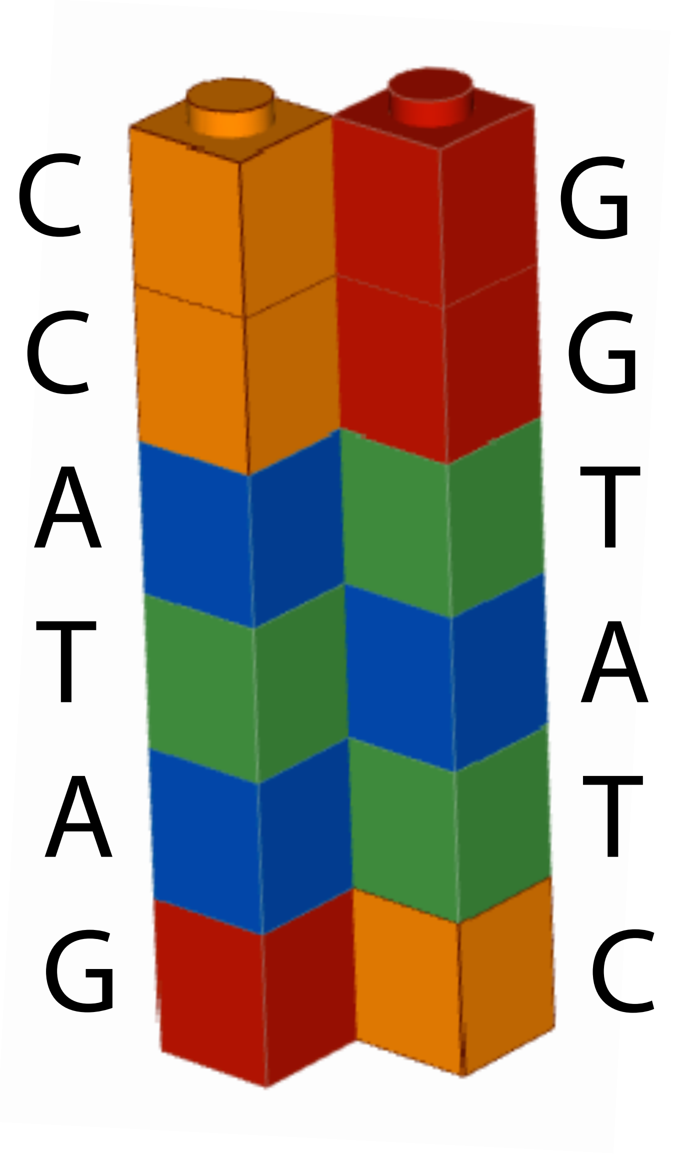


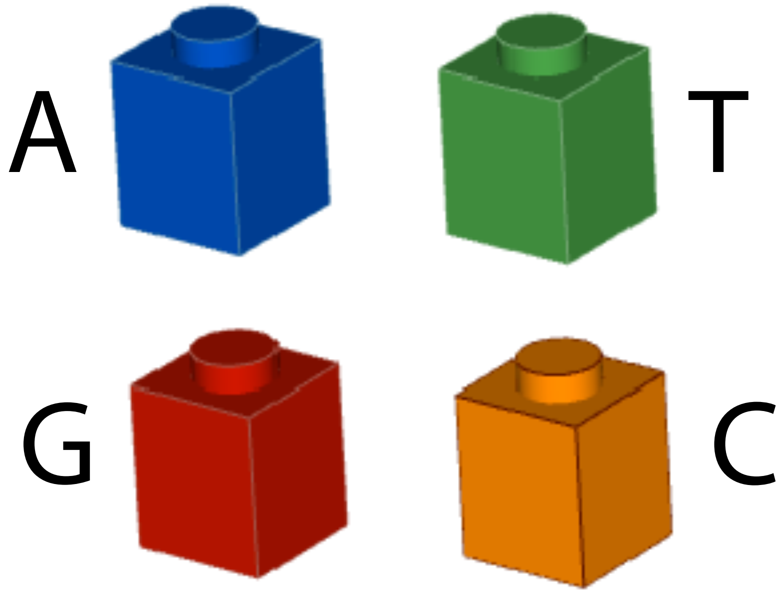
U.S. National Library of Medicine

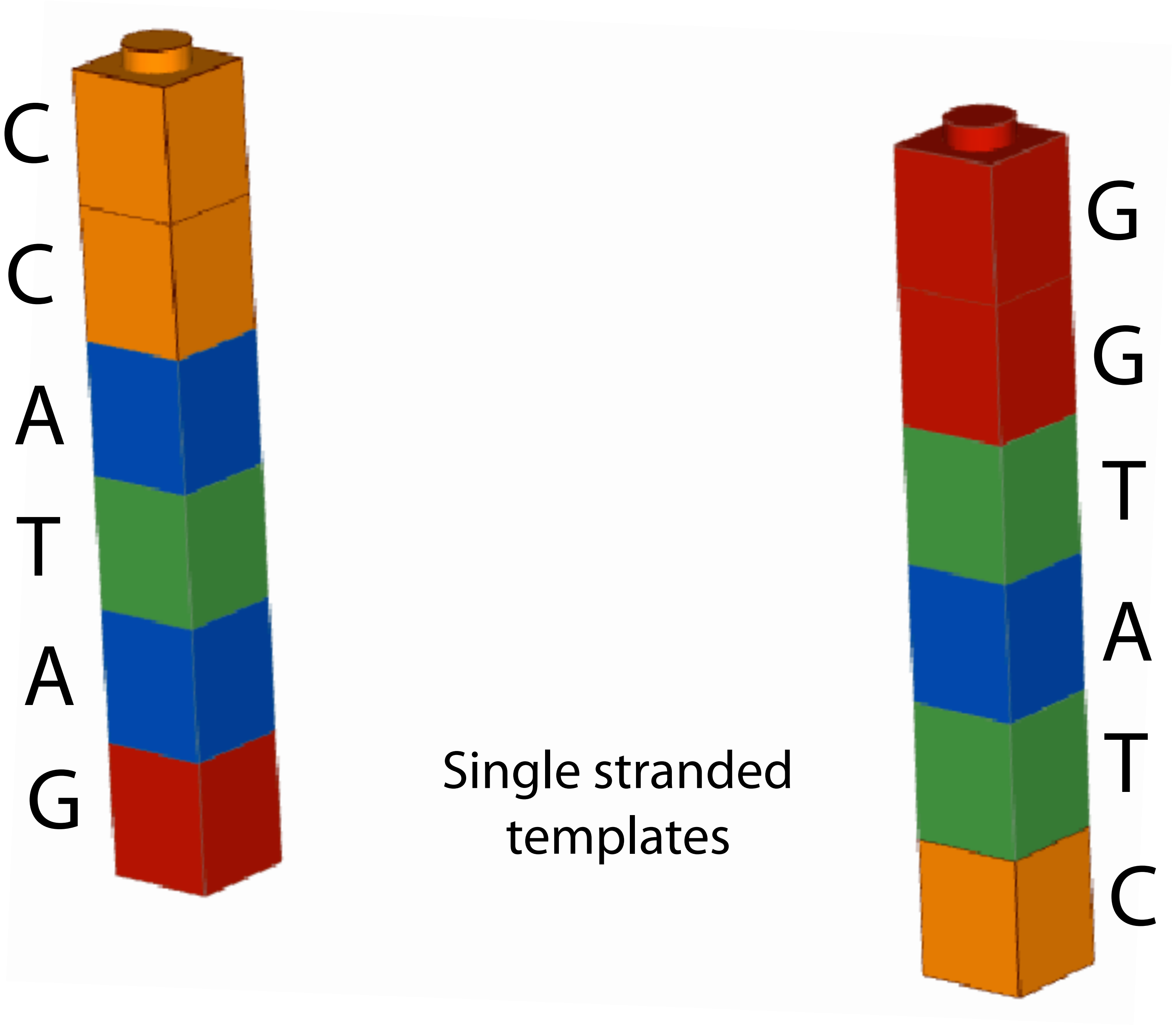


Double stranded DNA (double helix)

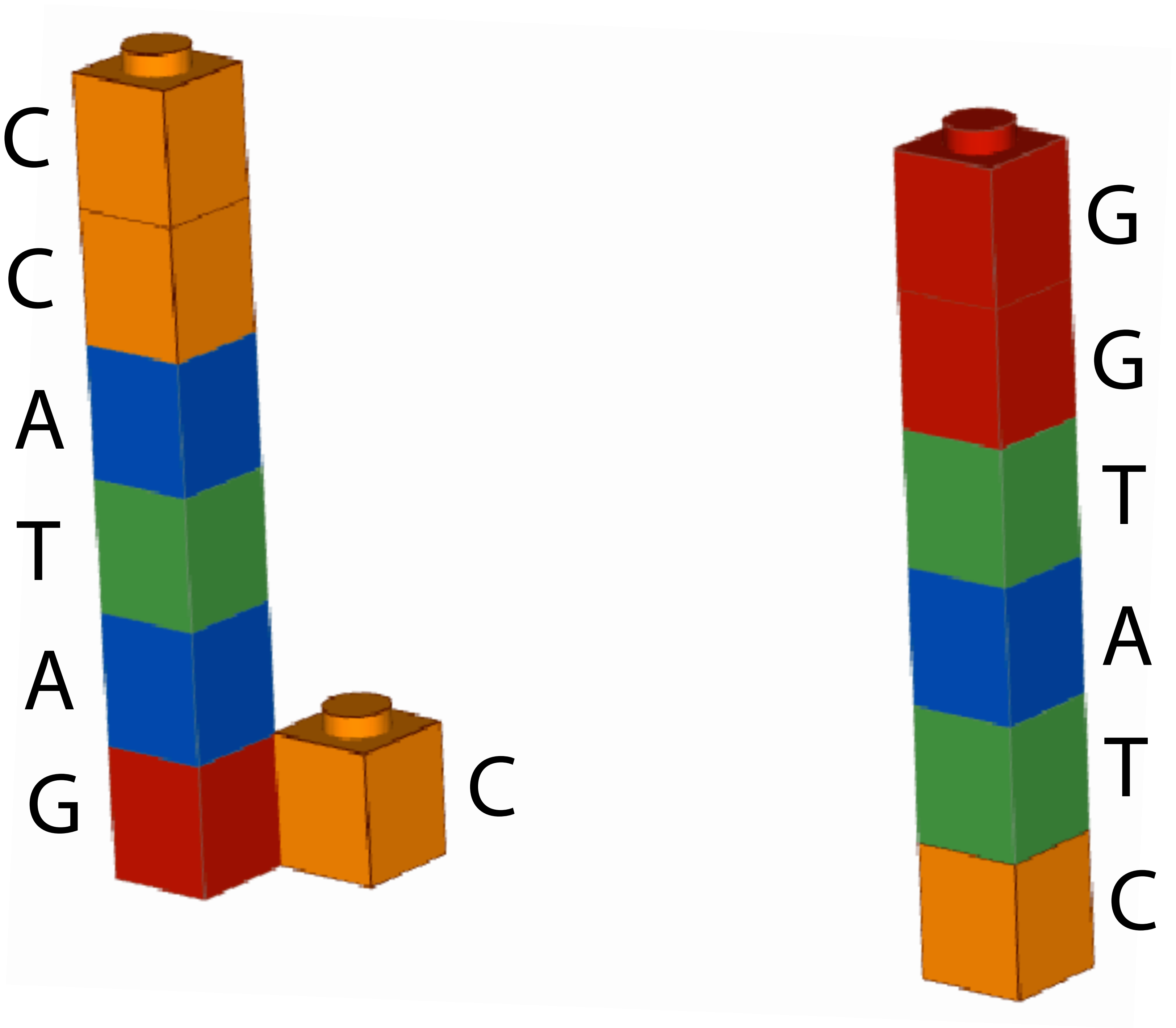
Double stranded DNA (lego version)

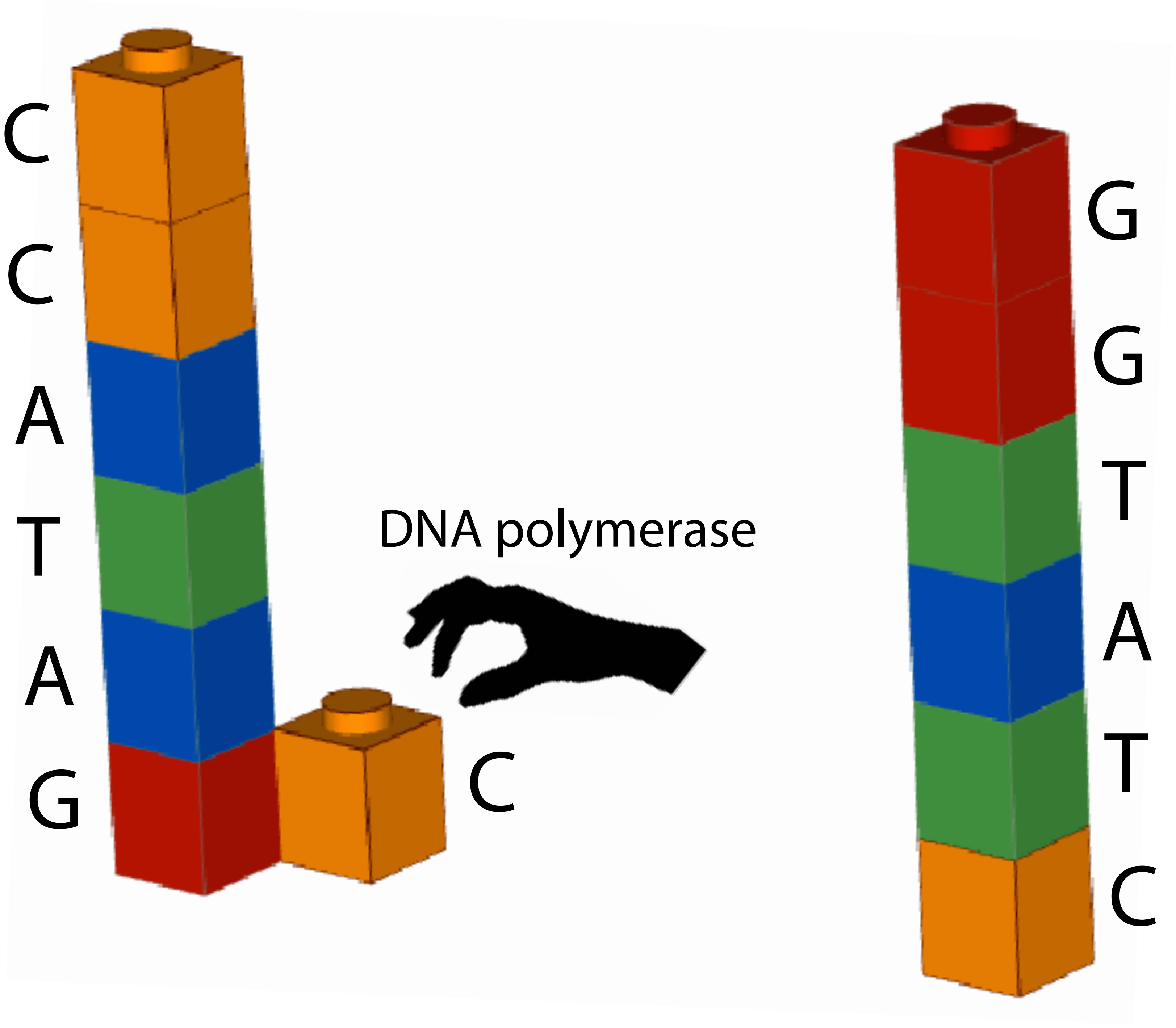

C
G
C
G
A
T
T
A
A
T
G
C

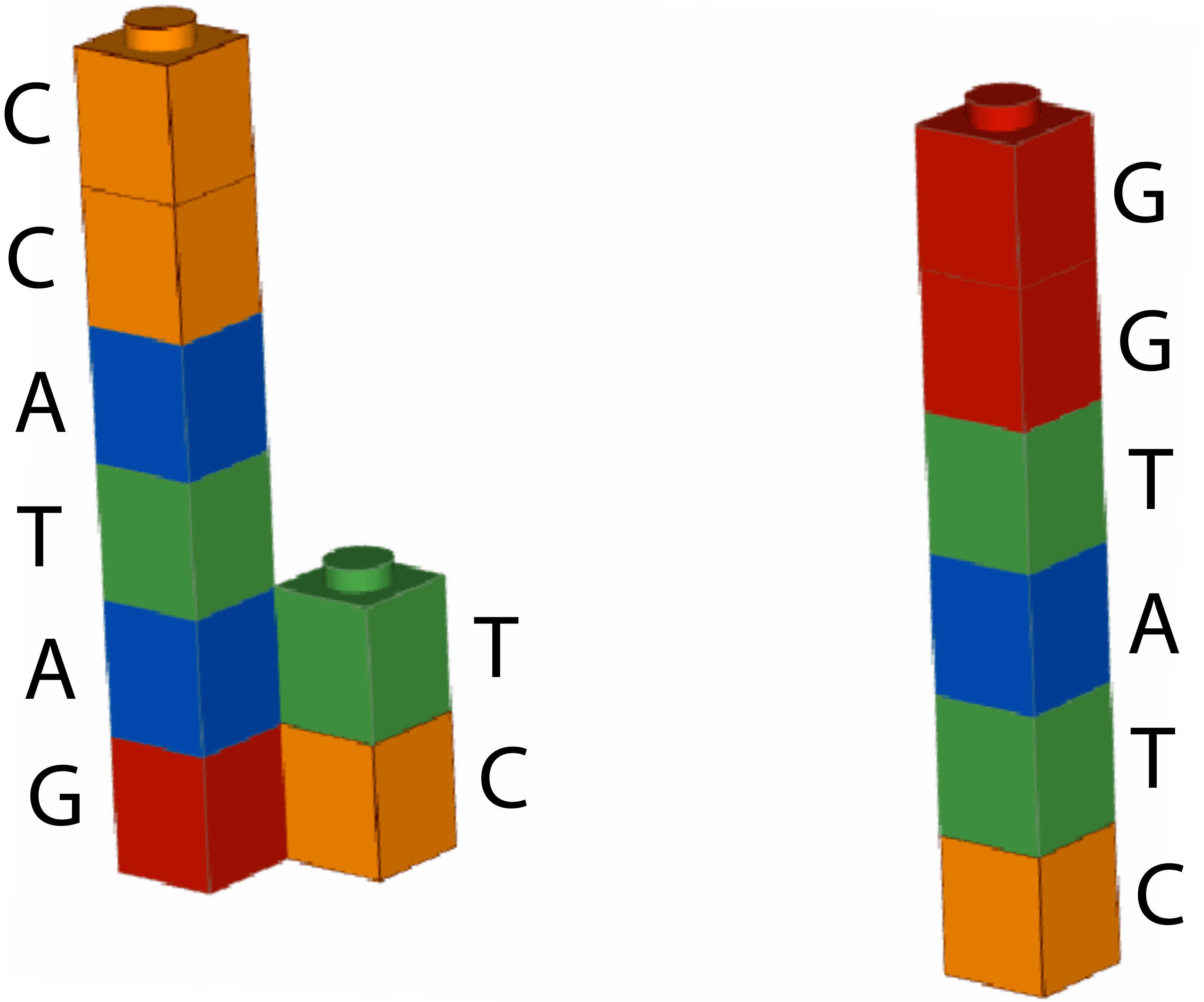

A
T
G
C

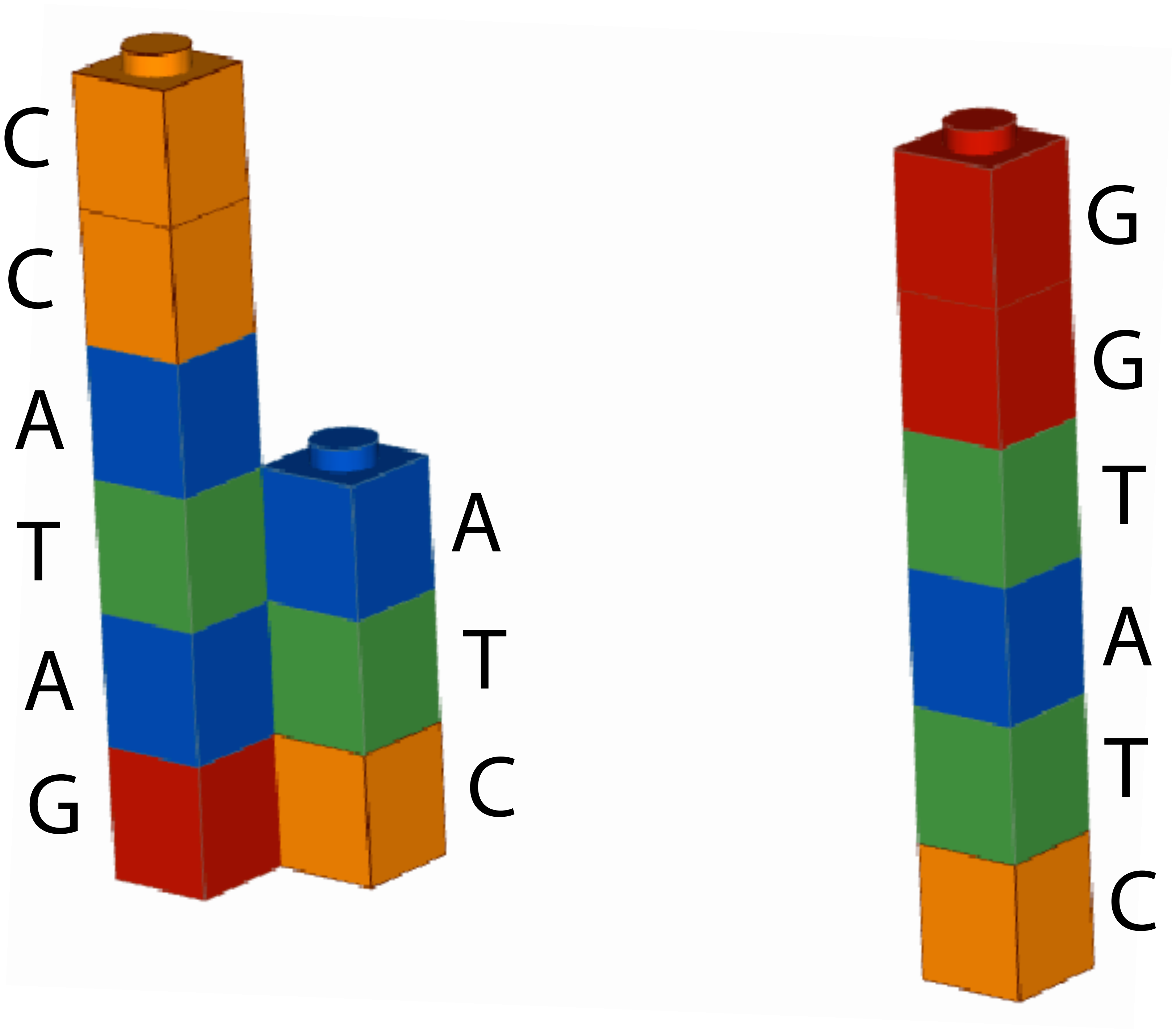


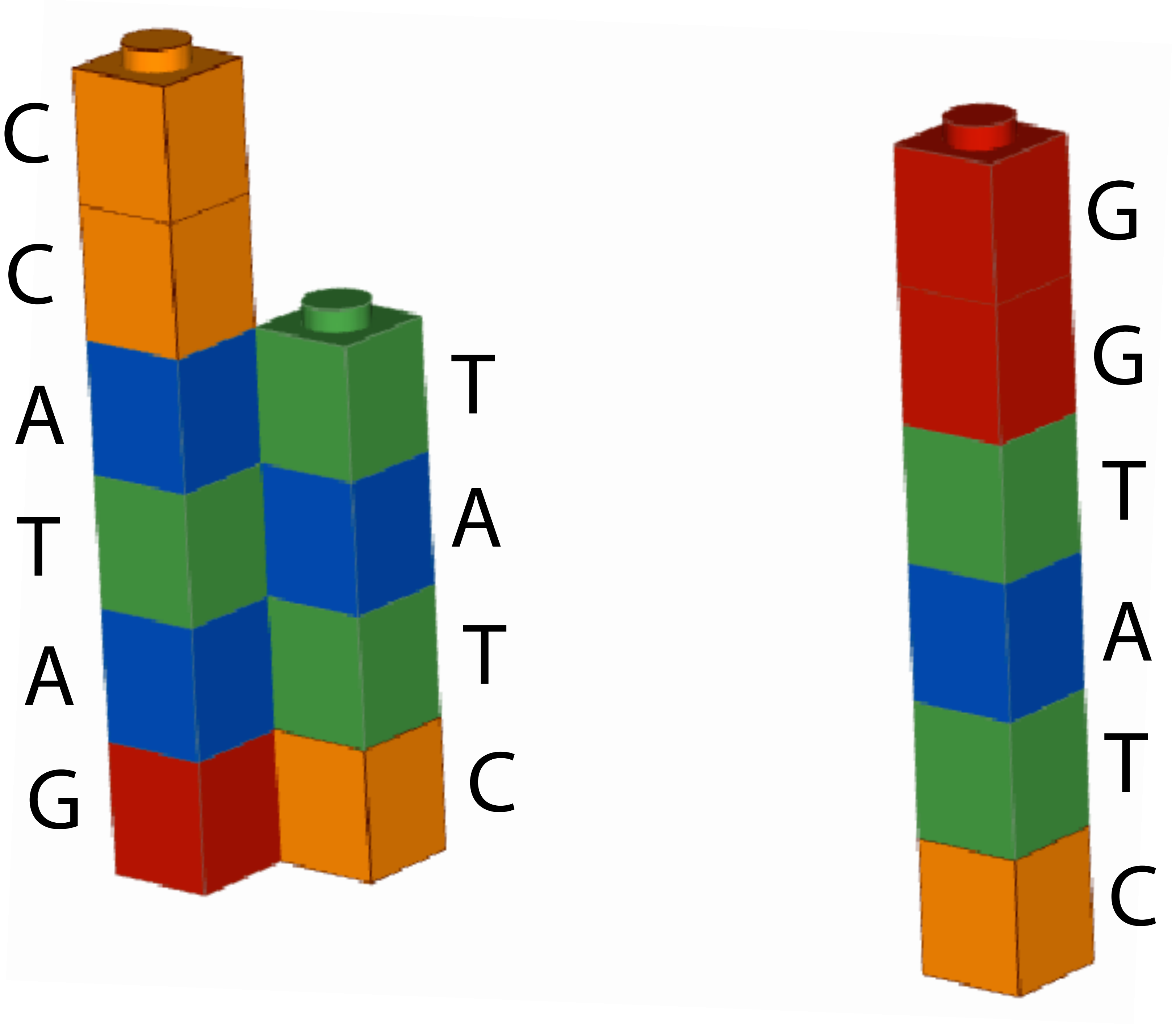
Single stranded templates

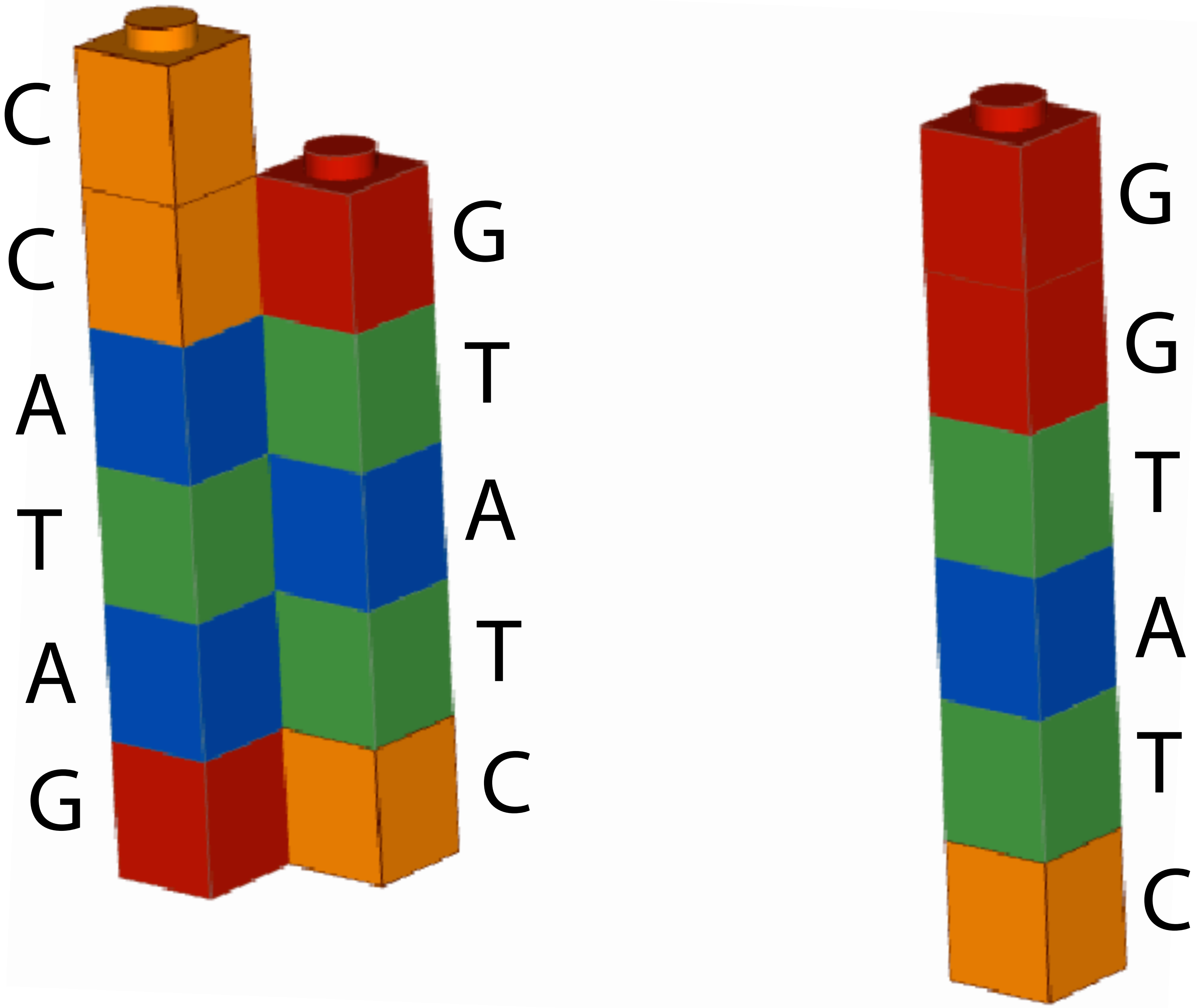

C
C
A
T
A
G
C
G
G
T
A
T
C

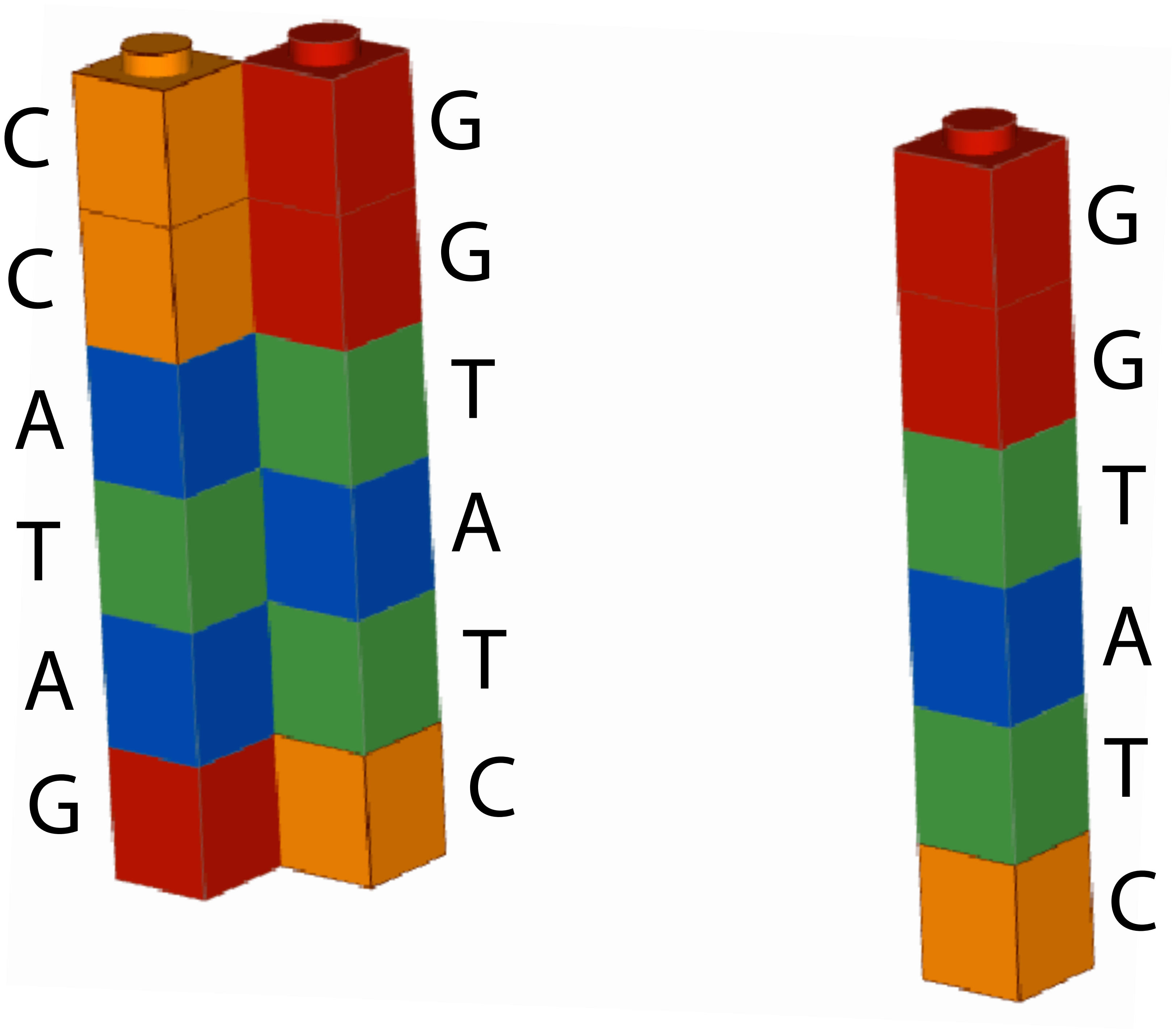

C
C
A
T
A
G
C
DNA polymerase
G
G
T
A
T
C

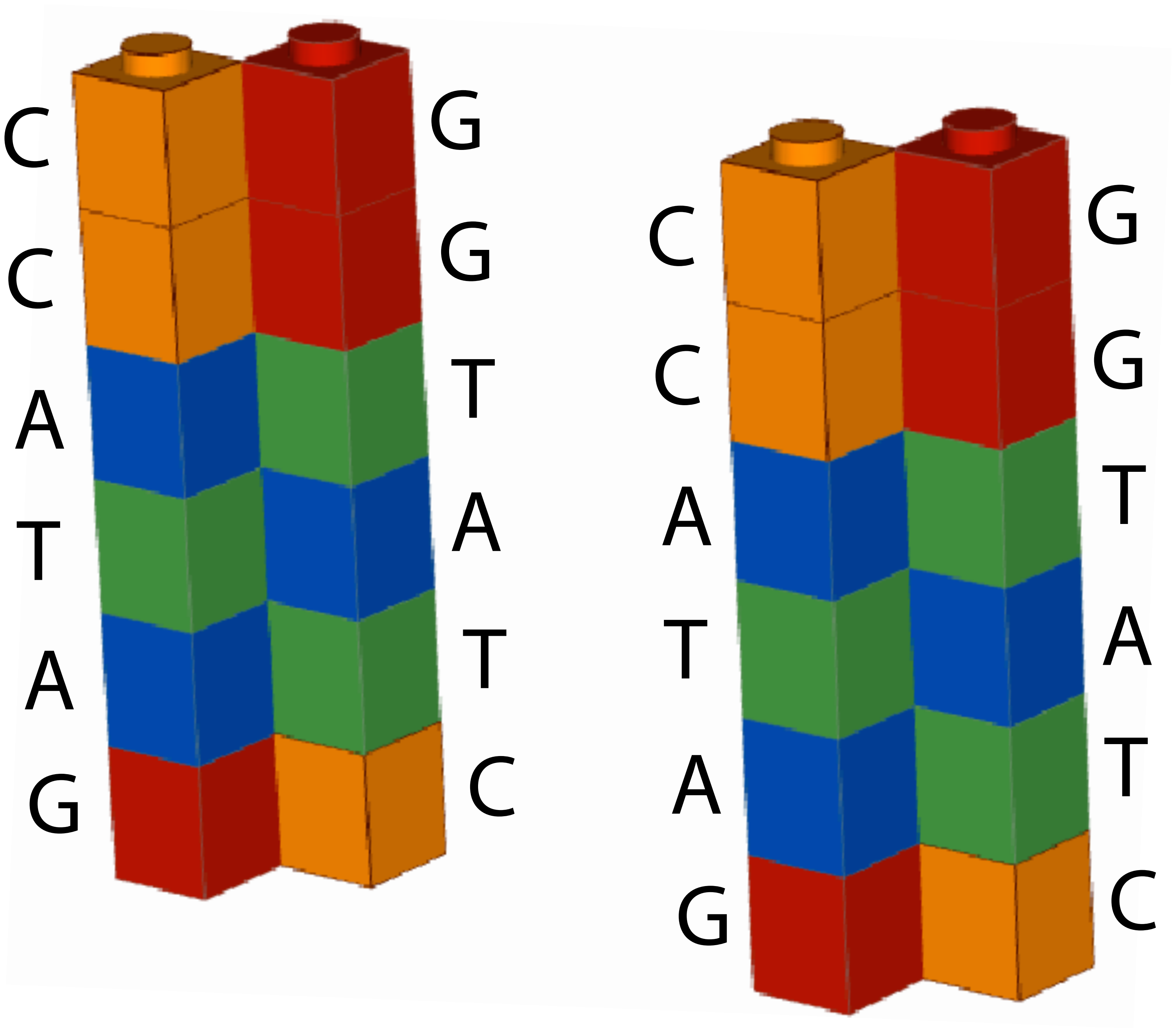

C
C
A
T
A
G
T
C
G
G
T
A
T
C


C
C
A
T
A
G
A
T
C
G
G
T
A
T
C


C
C
A
T
A
G
T
A
T
C
G
G
T
A
T
C


C
C
A
T
A
G
G
T
A
T
C
G
G
T
A
T
C


C
C
A
T
A
G
G
G
T
A
T
C
G
G
T
A
T
C


C
C
A
T
A
G
G
G
T
A
T
C
C
C
A
T
A
G
G
G
T
A
T
C

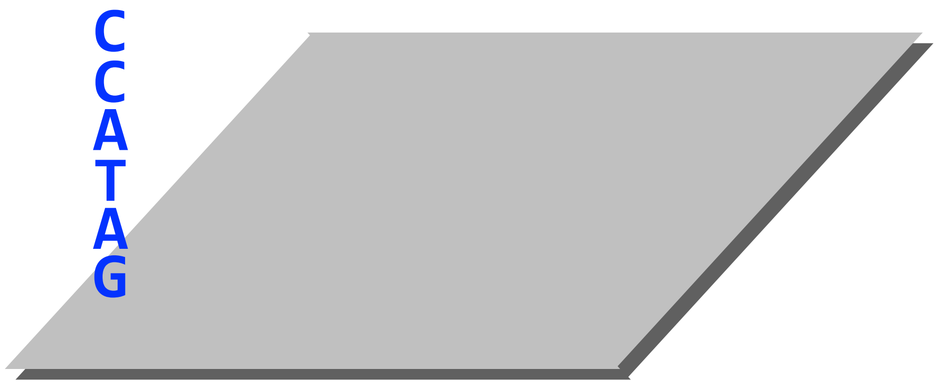
Input DNA

CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT

Cut into snippets

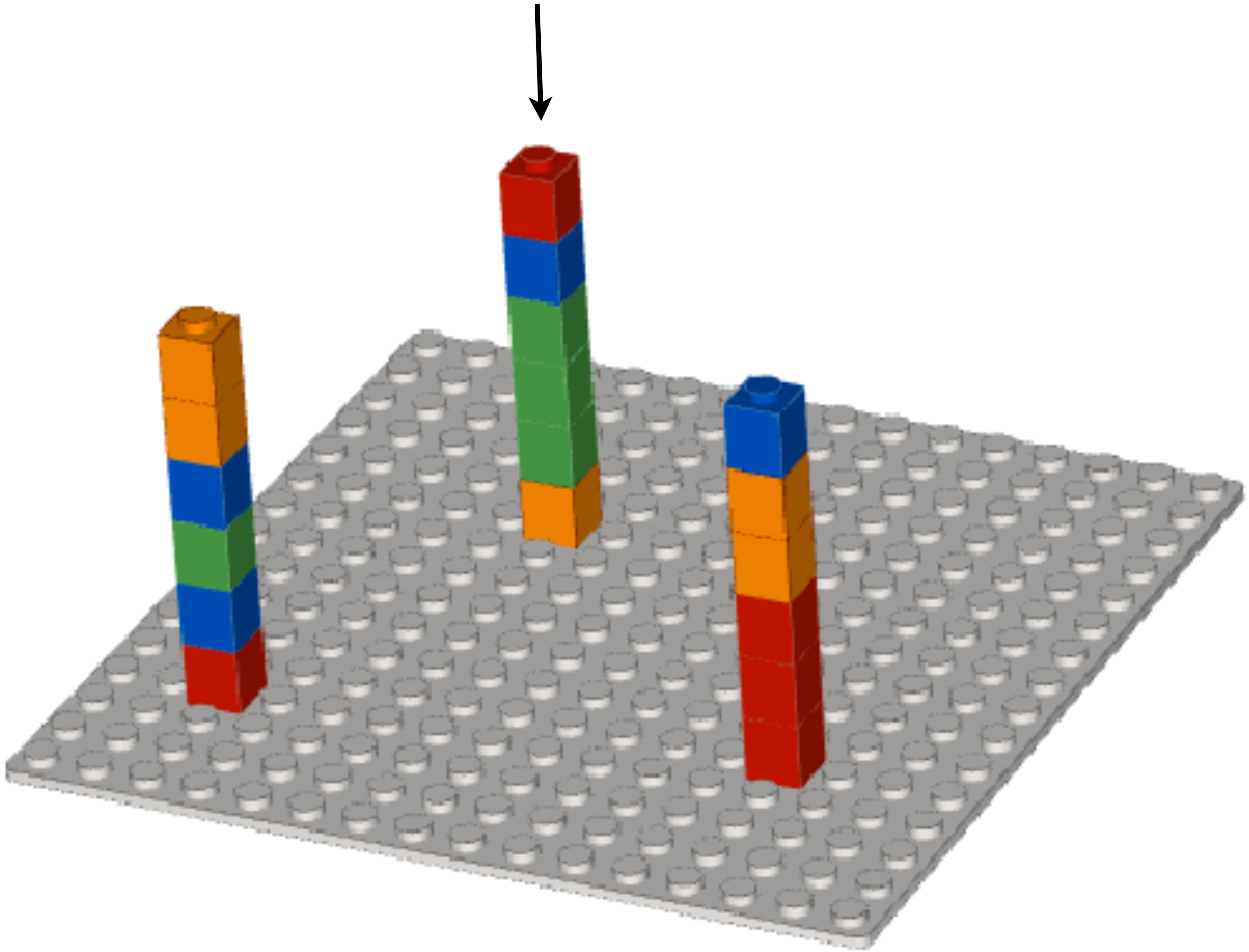
CCATAGTA TATCTCGG CTCTAGGCCCTC ATTTTTT
CCA TAGTATAT CTCGGCTCTAGGCCCTCA TTTTTT
CCATAGTAT ATCTCGGCTCTAG GCCCTCA TTTTTT
CCATAG TATATCT CGGCTCTAGGCCCT CATTTTTT

Deposit on slide

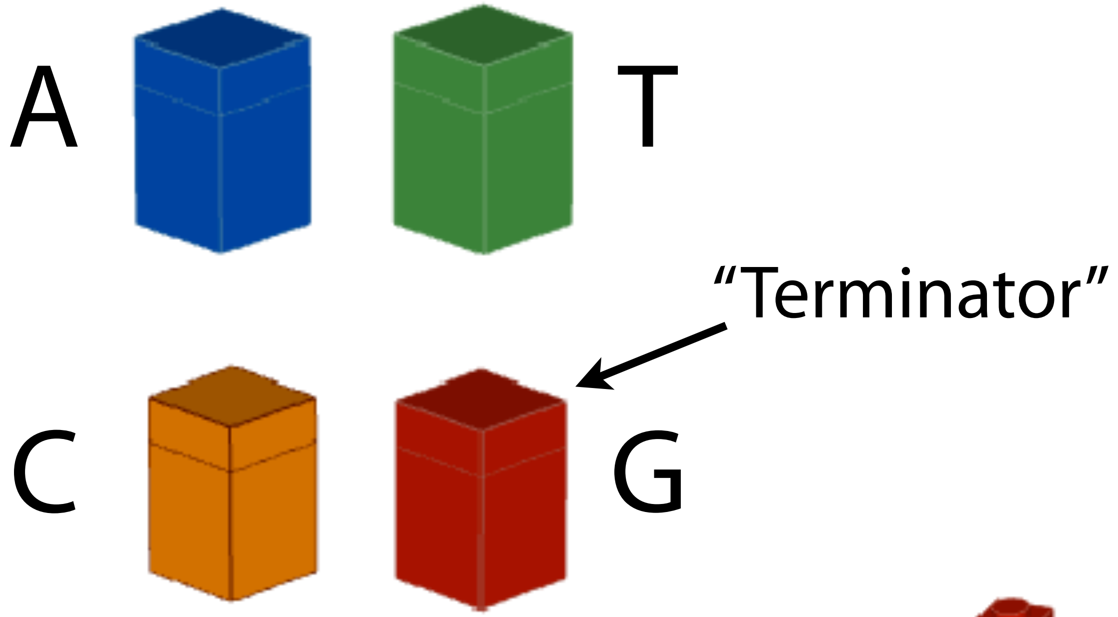


More details: Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6;456(7218):53-9

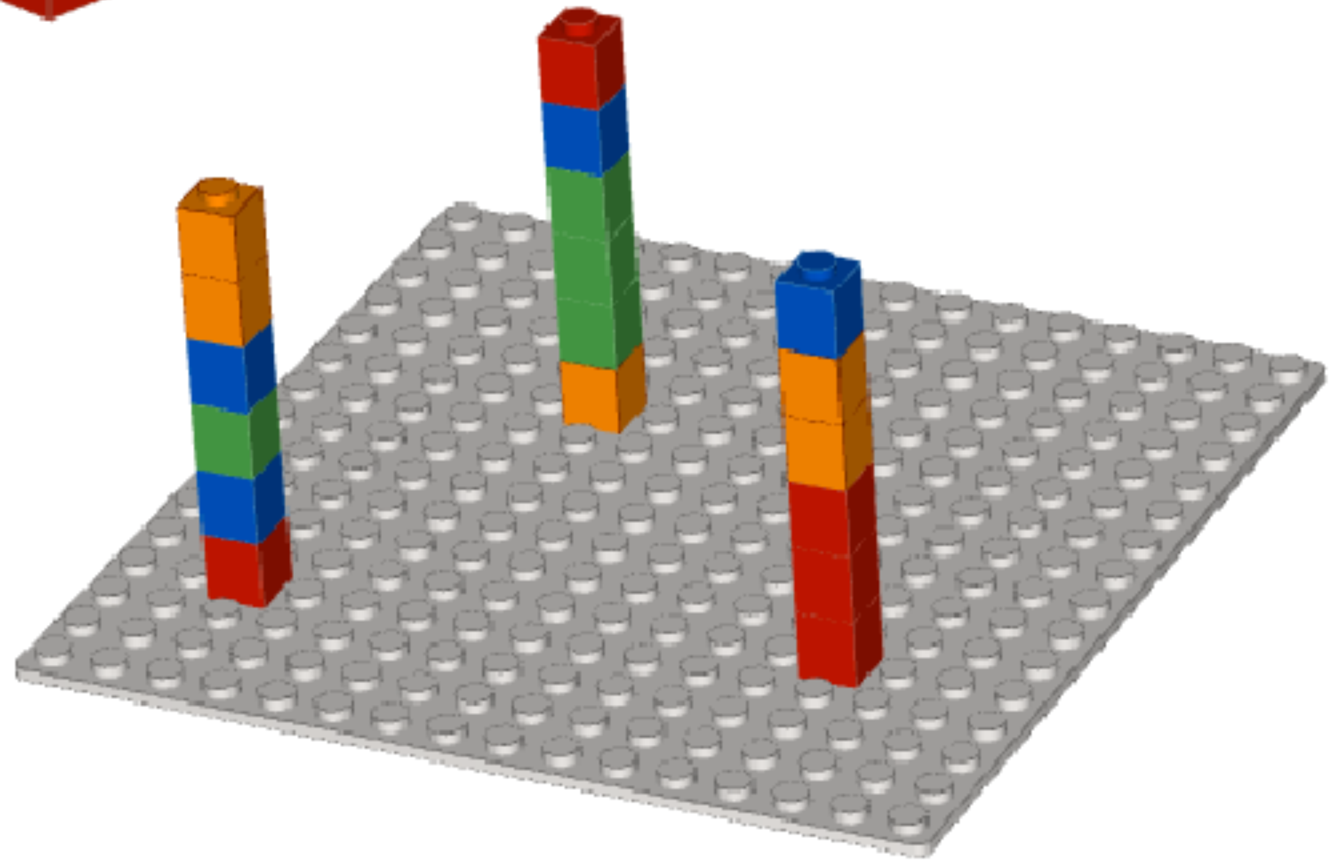
Template
(billions of them!)

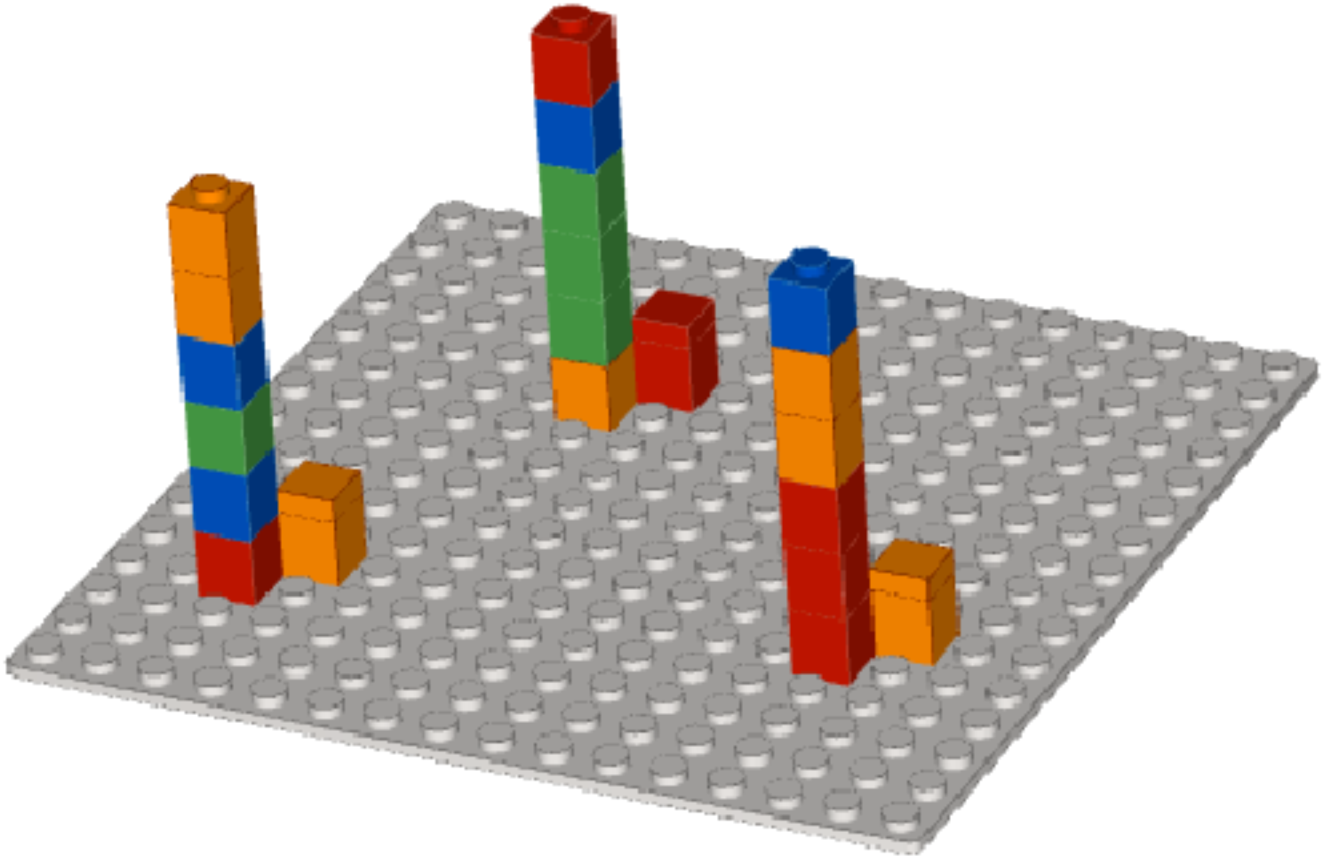


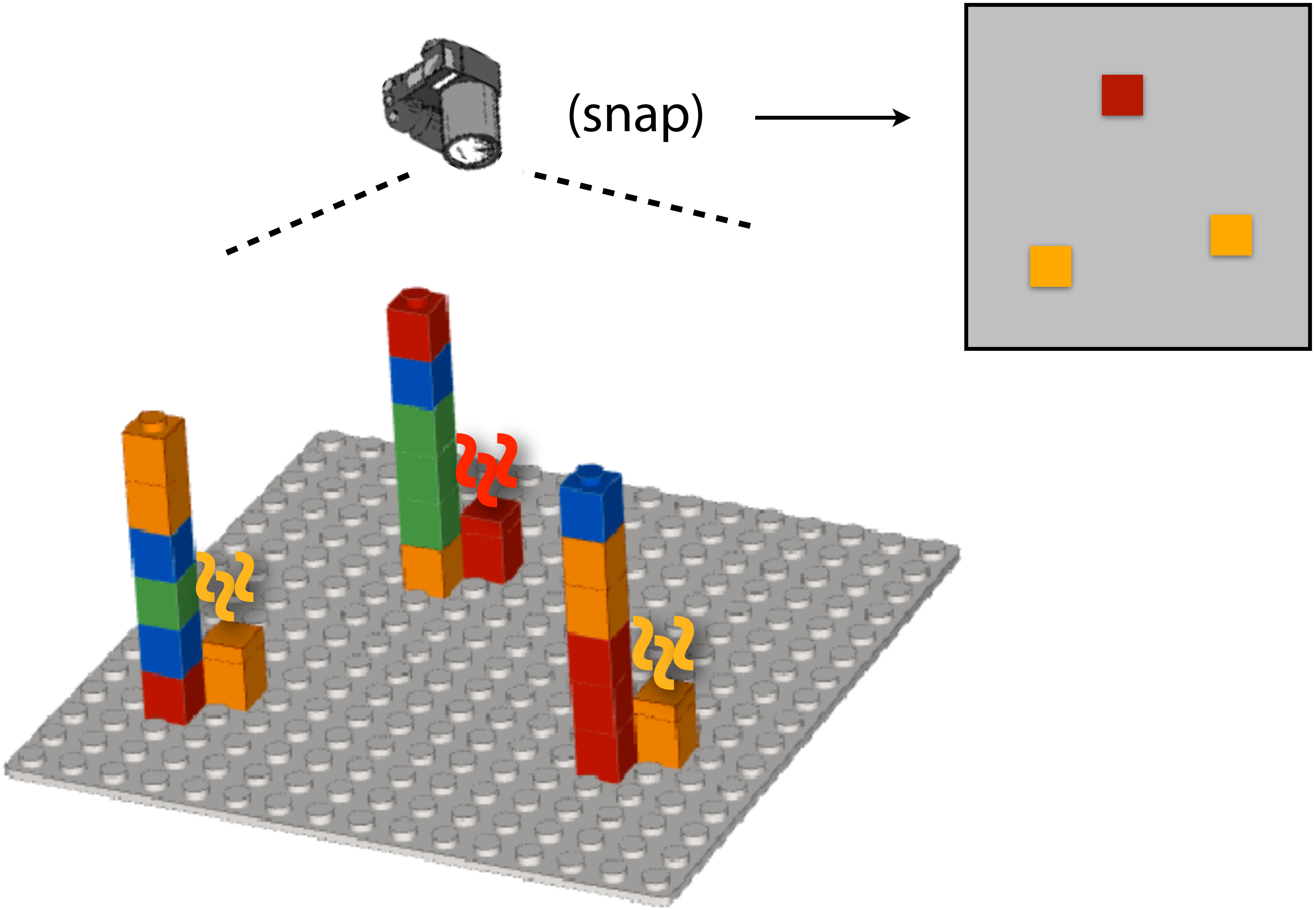
Slide



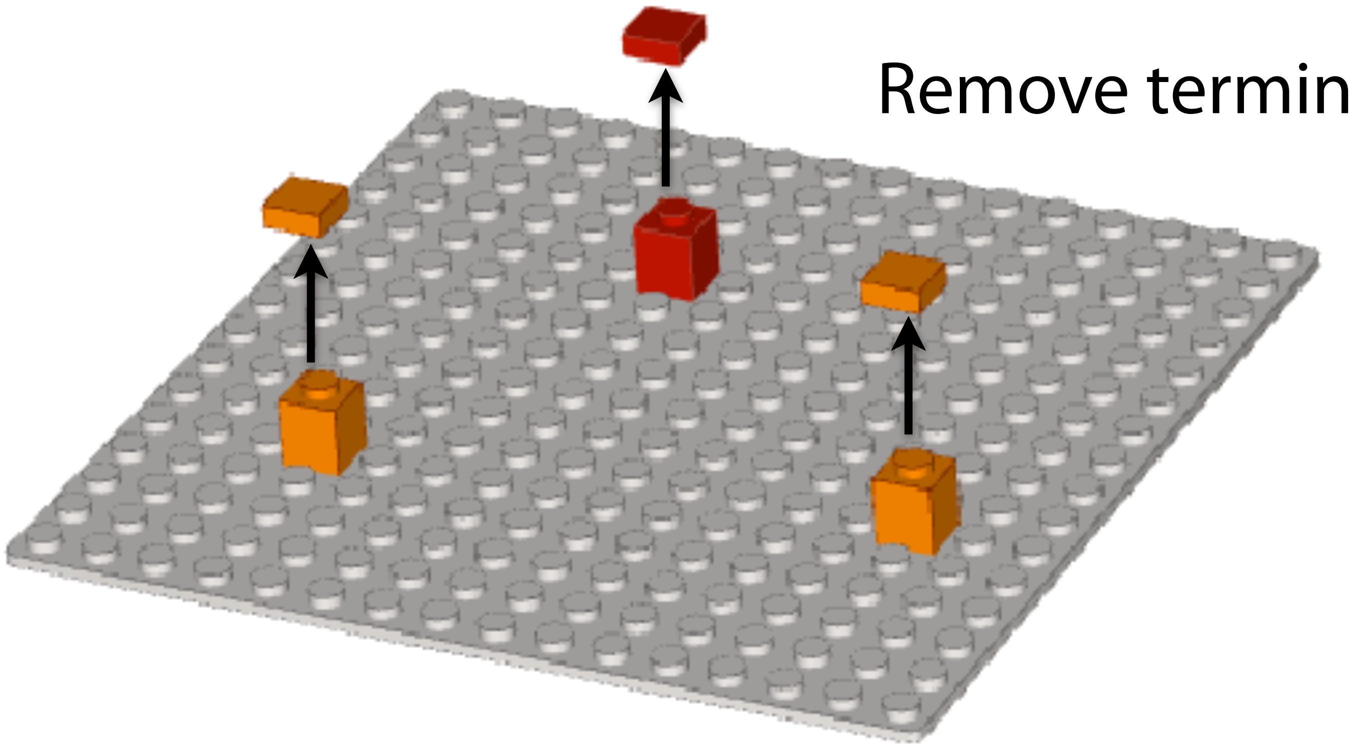
DNA polymerase



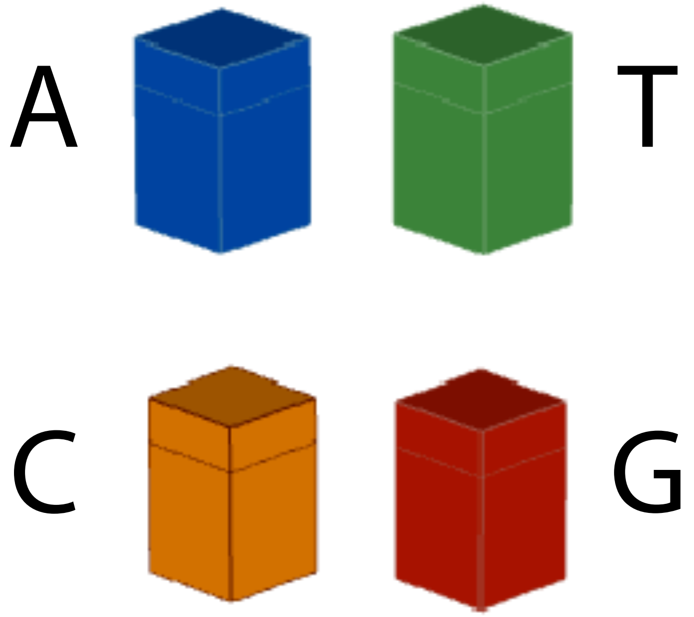




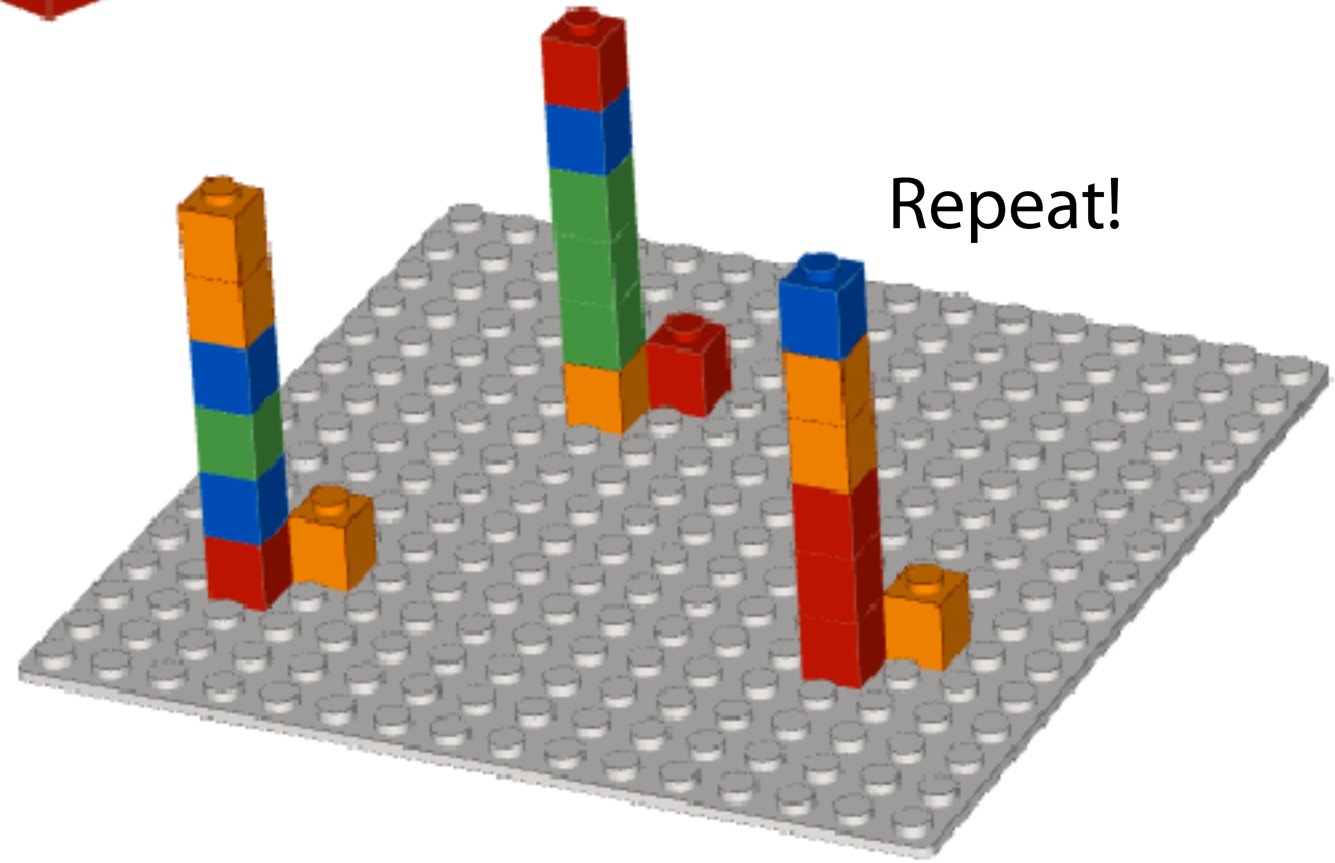
(snap)



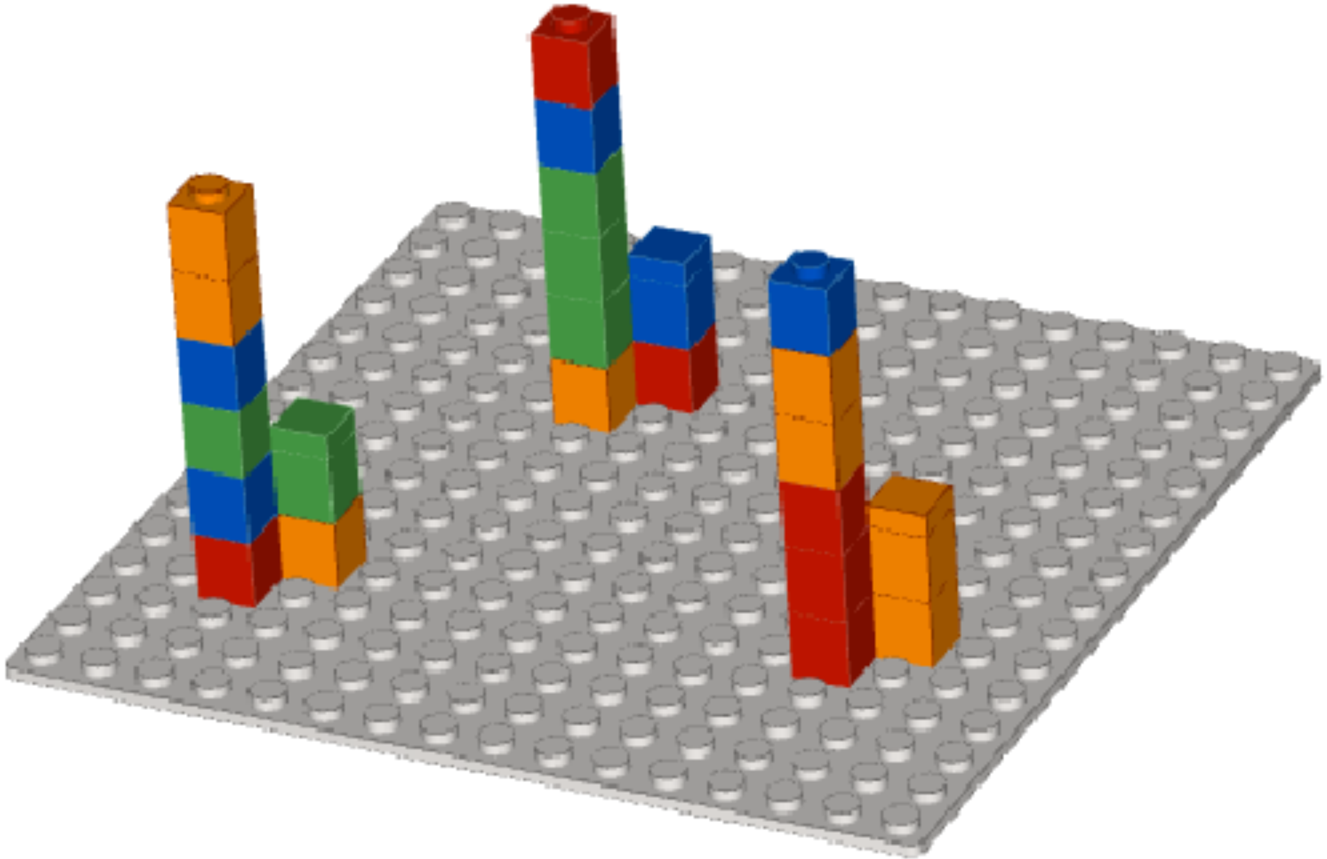
Remove terminators

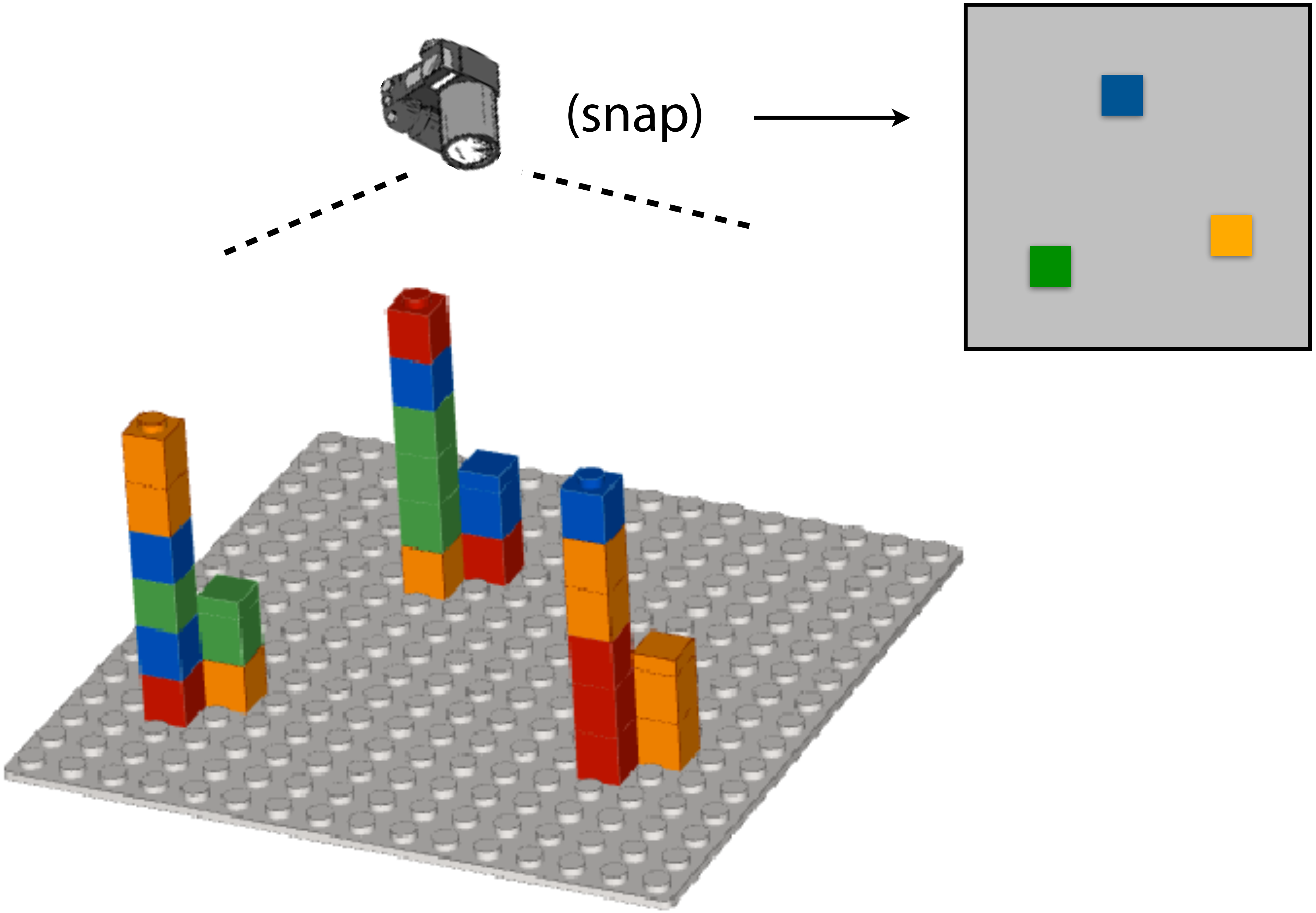


DNA polymerase

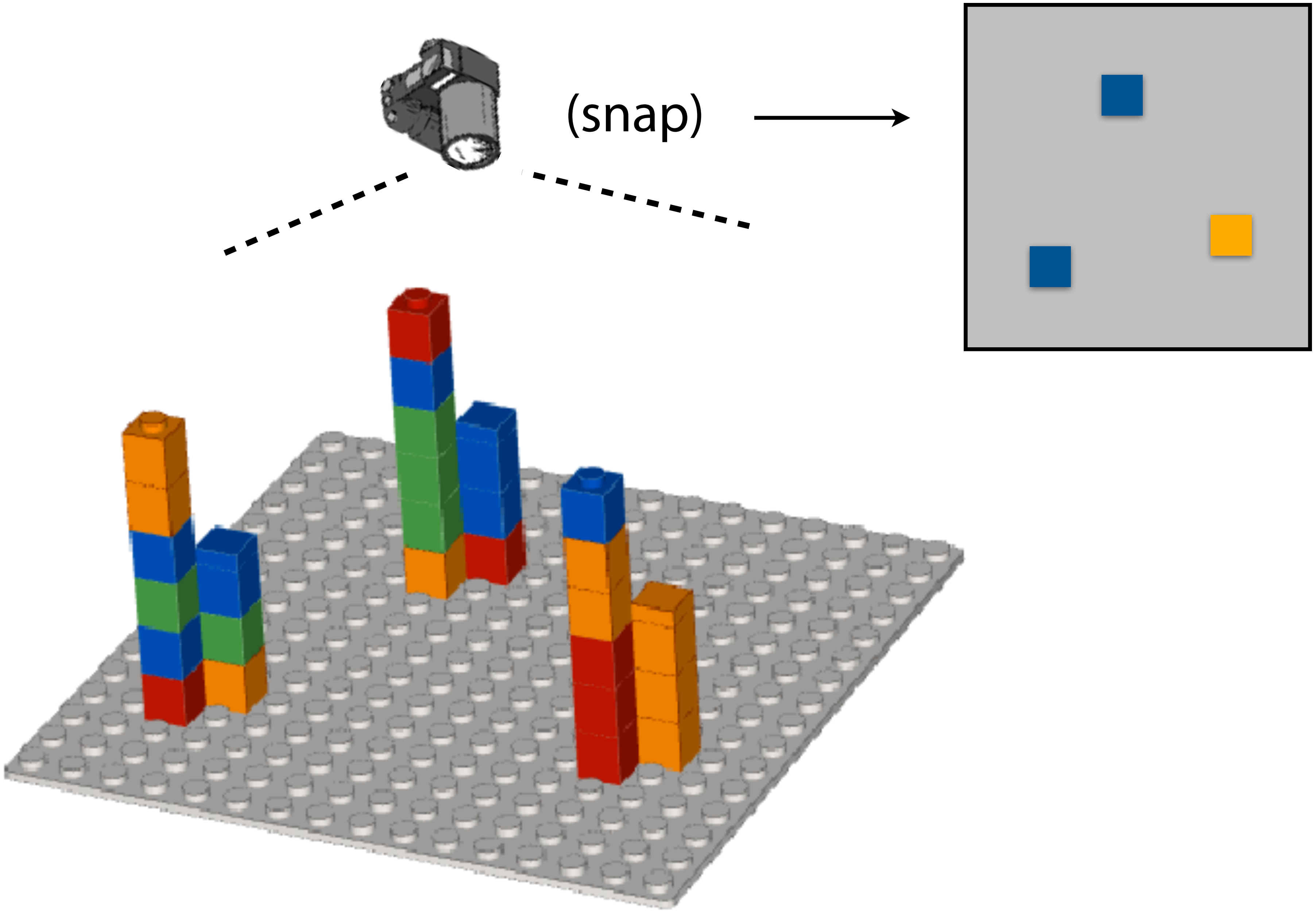


Repeat!

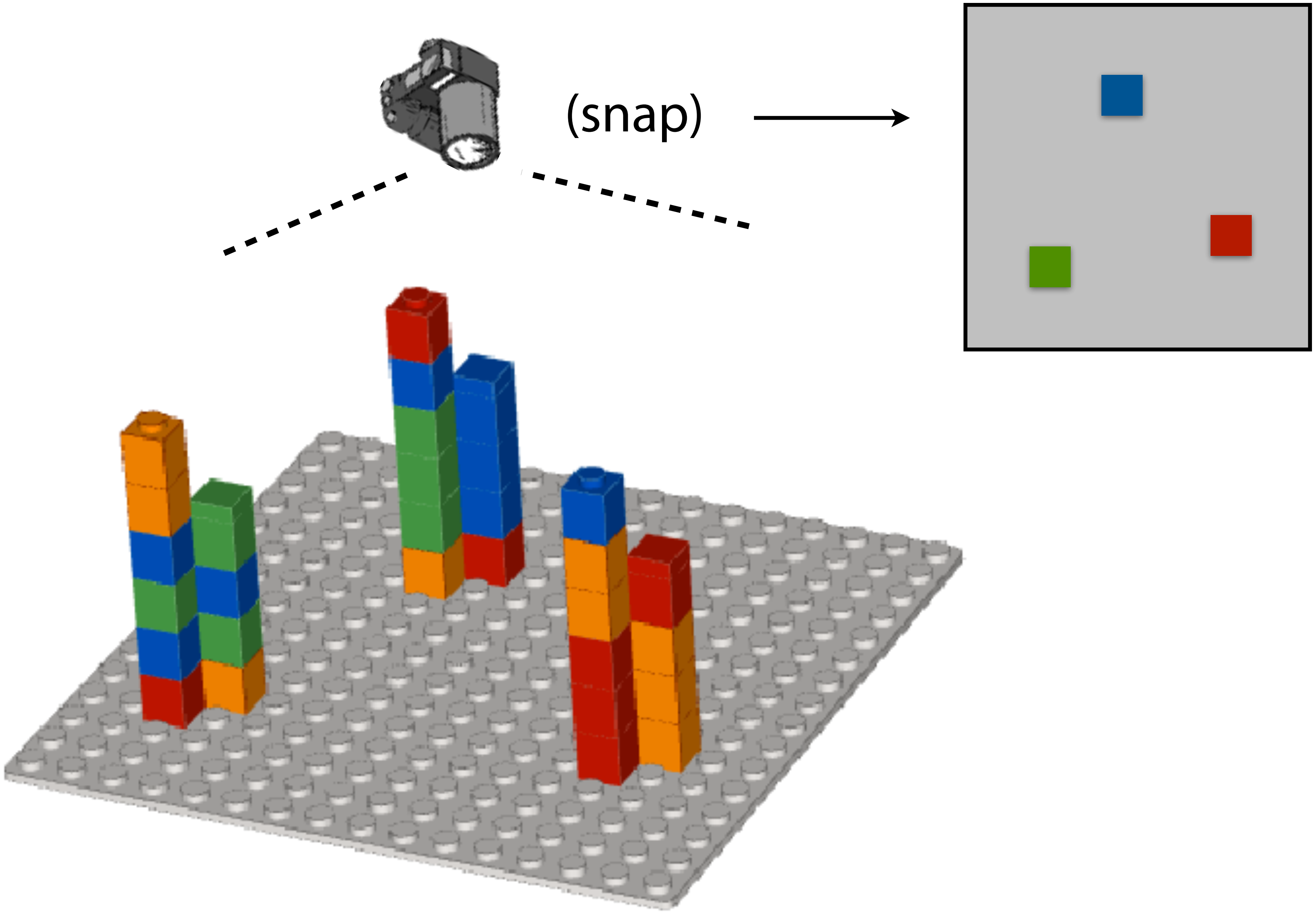




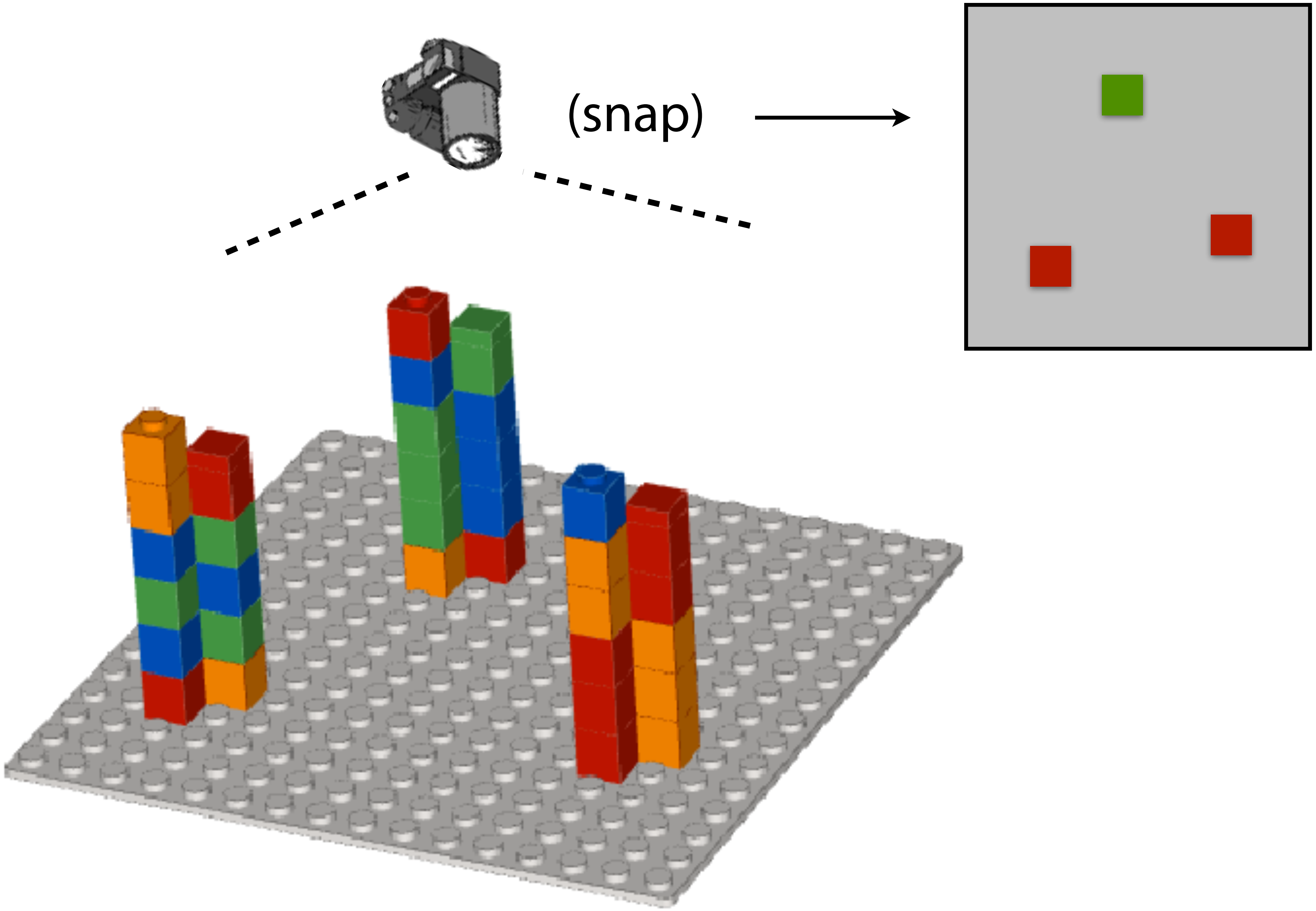
(snap)



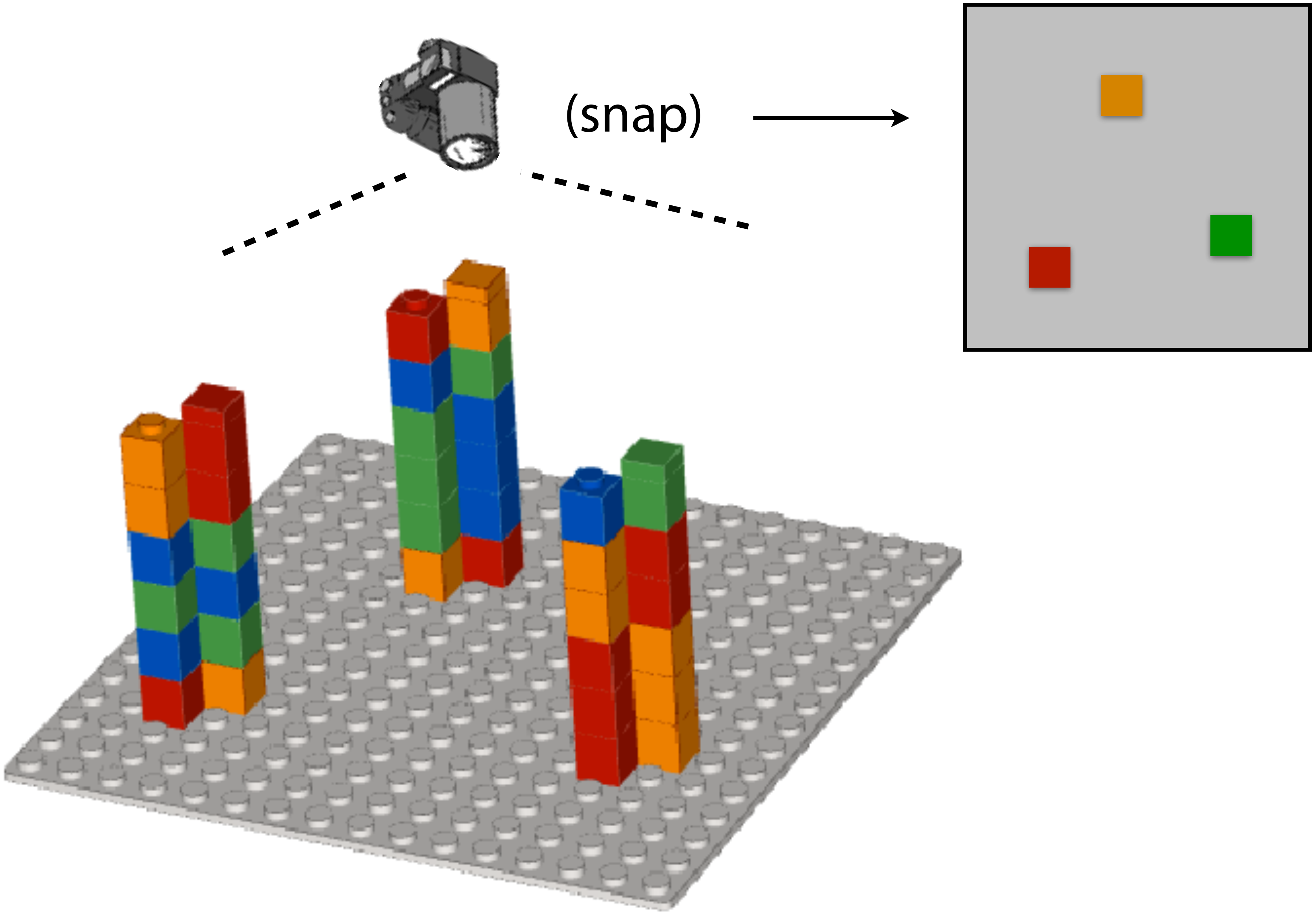
(snap)



(snap)

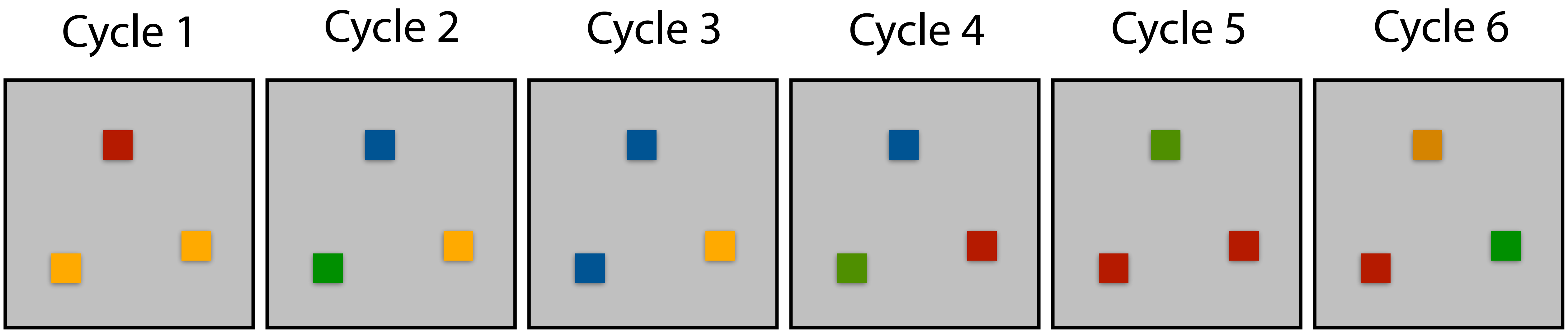


(snap)

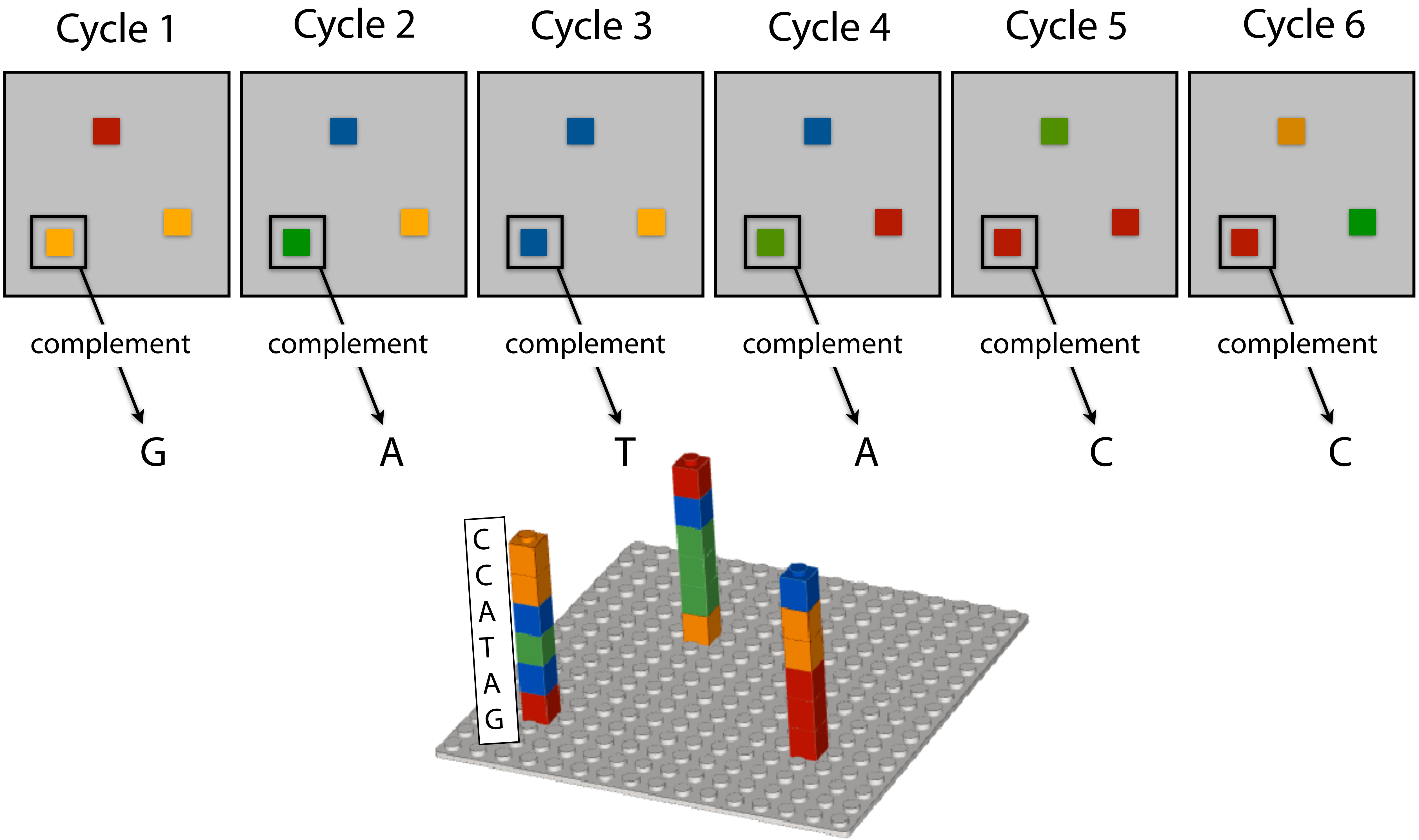


(snap)

Sequencing by synthesis



Sequencing by synthesis



Cycle 1 complement G

Cycle 2 complement A

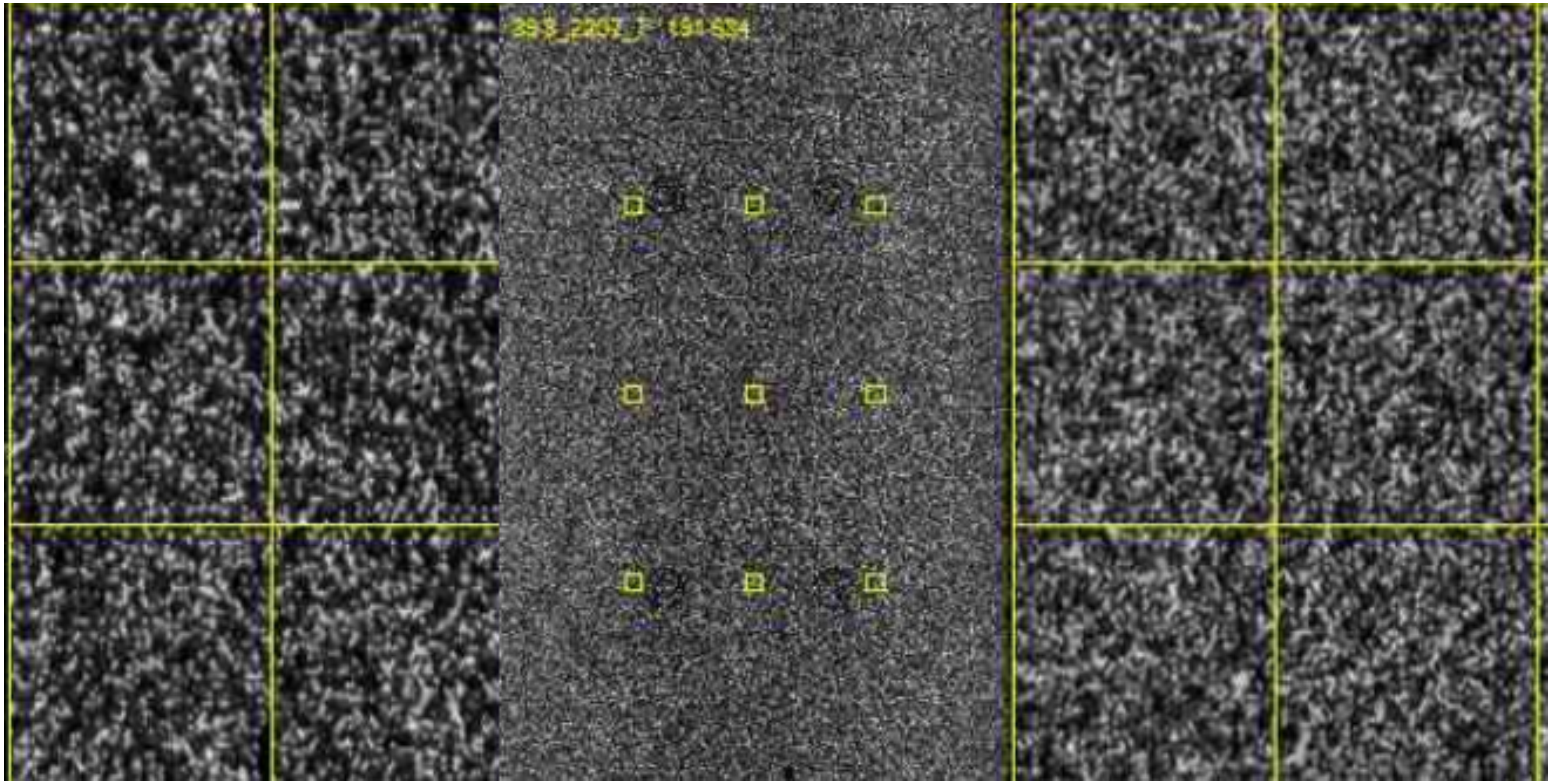
Cycle 3 complement T

Cycle 4 complement A

Cycle 5 complement C

Cycle 6 complement C

Sequencing by synthesis



Actual Illumina HiSeq 3000 image

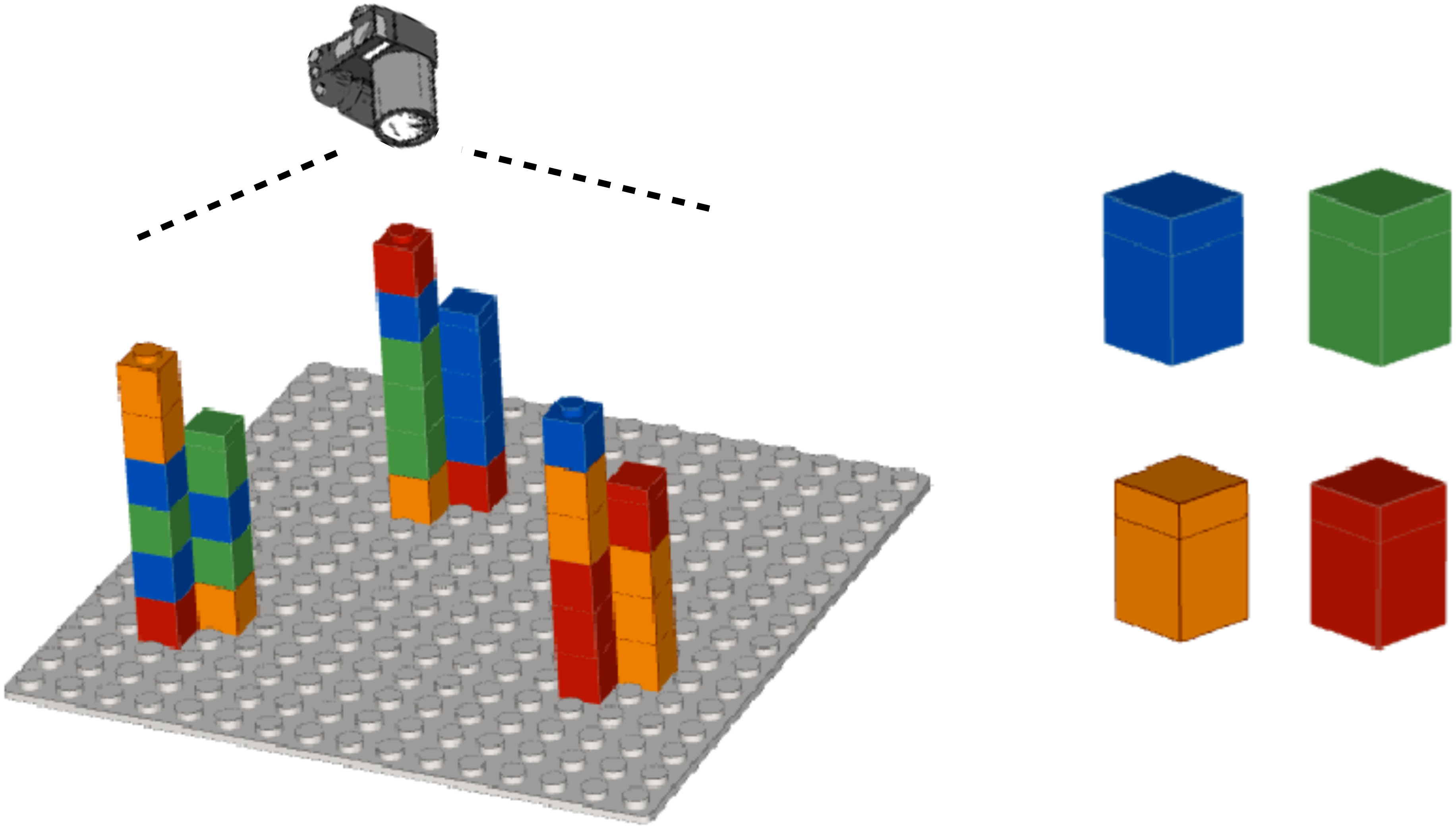
http://dnatech.genomecenter.ucdavis.edu/2015/05/07/first-hiseq-3000-data-download/

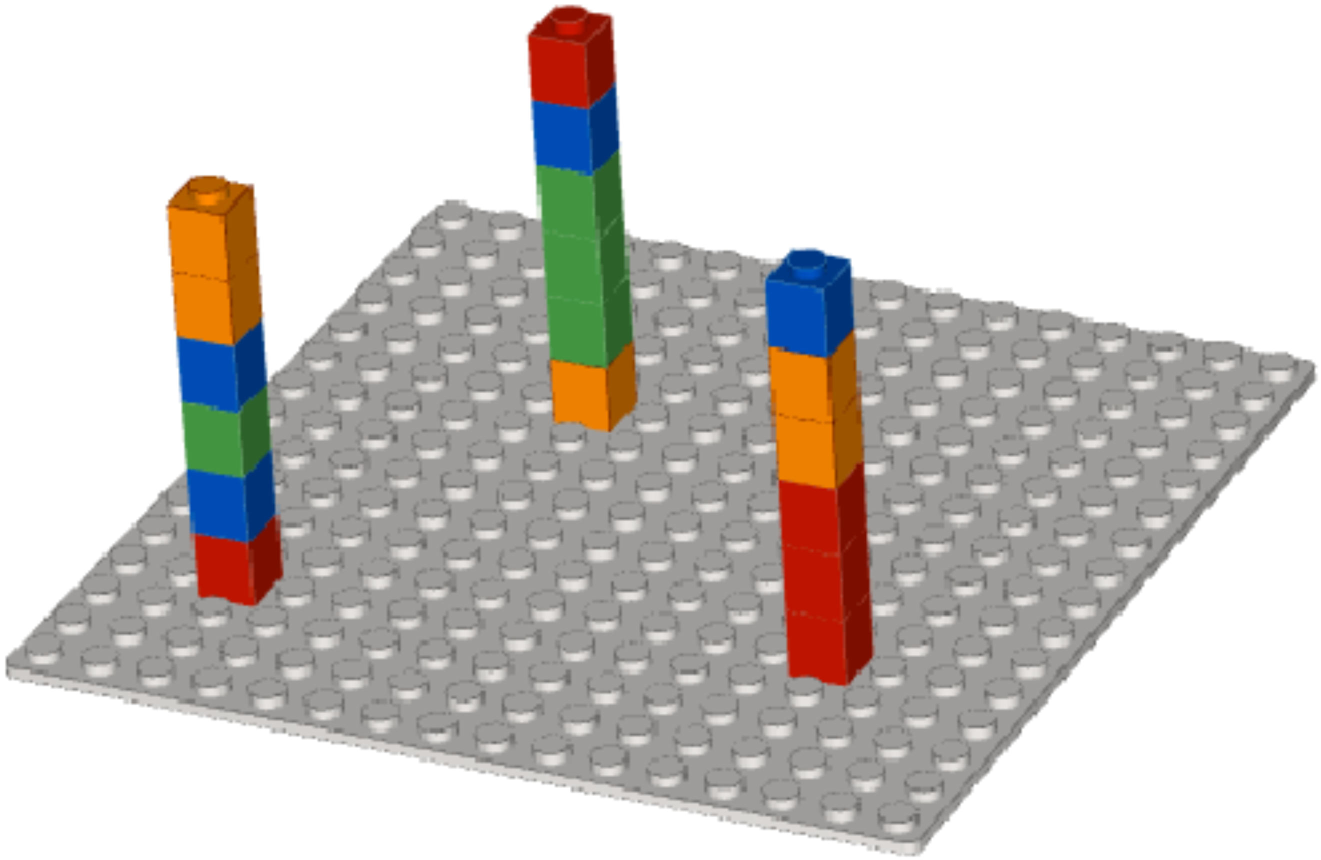
Sequencing by synthesis

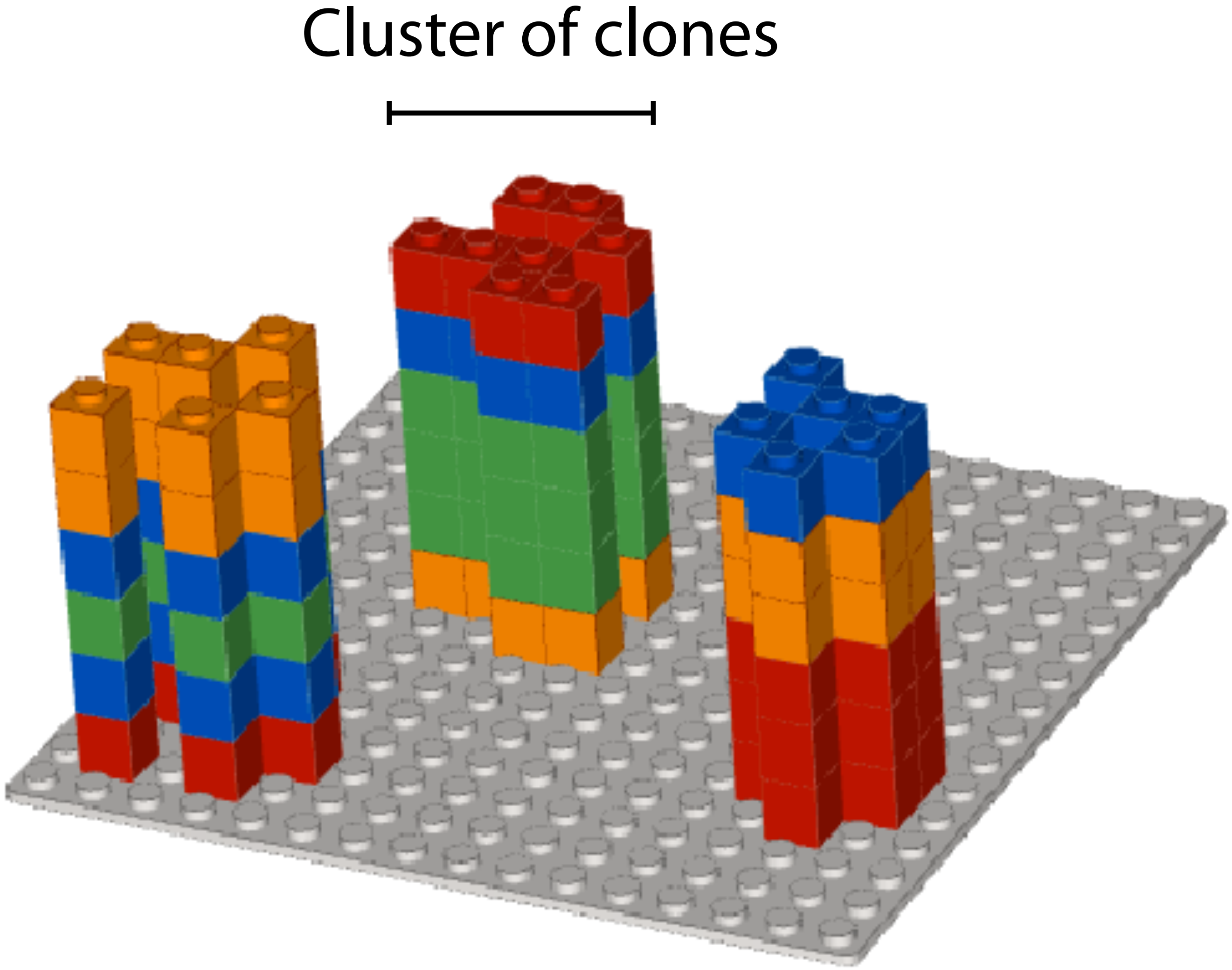
Billions of templates on a slide

Massively parallel: photograph captures all templates simultaneously

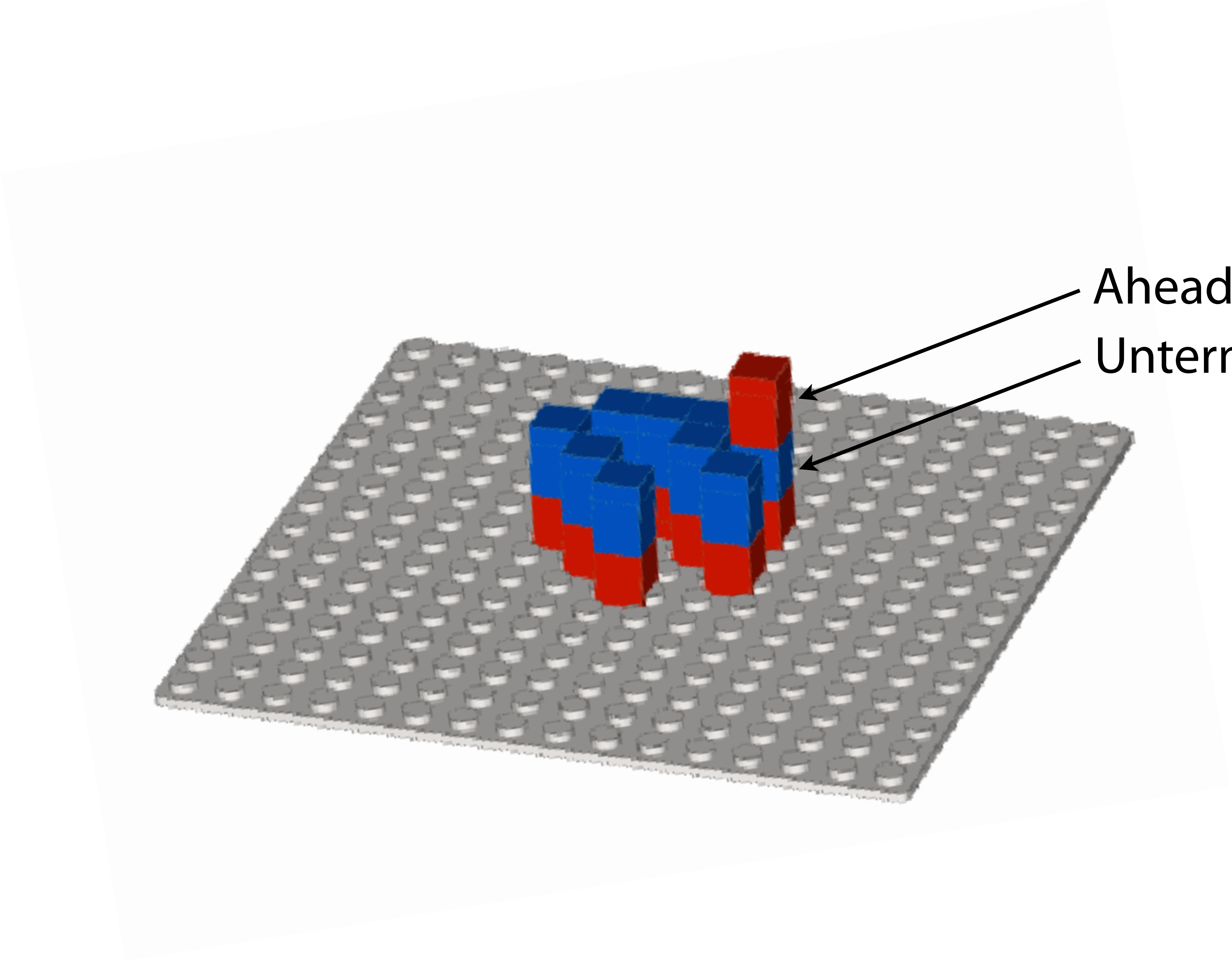
Terminators are "speed bumps," keeping reactions in sync





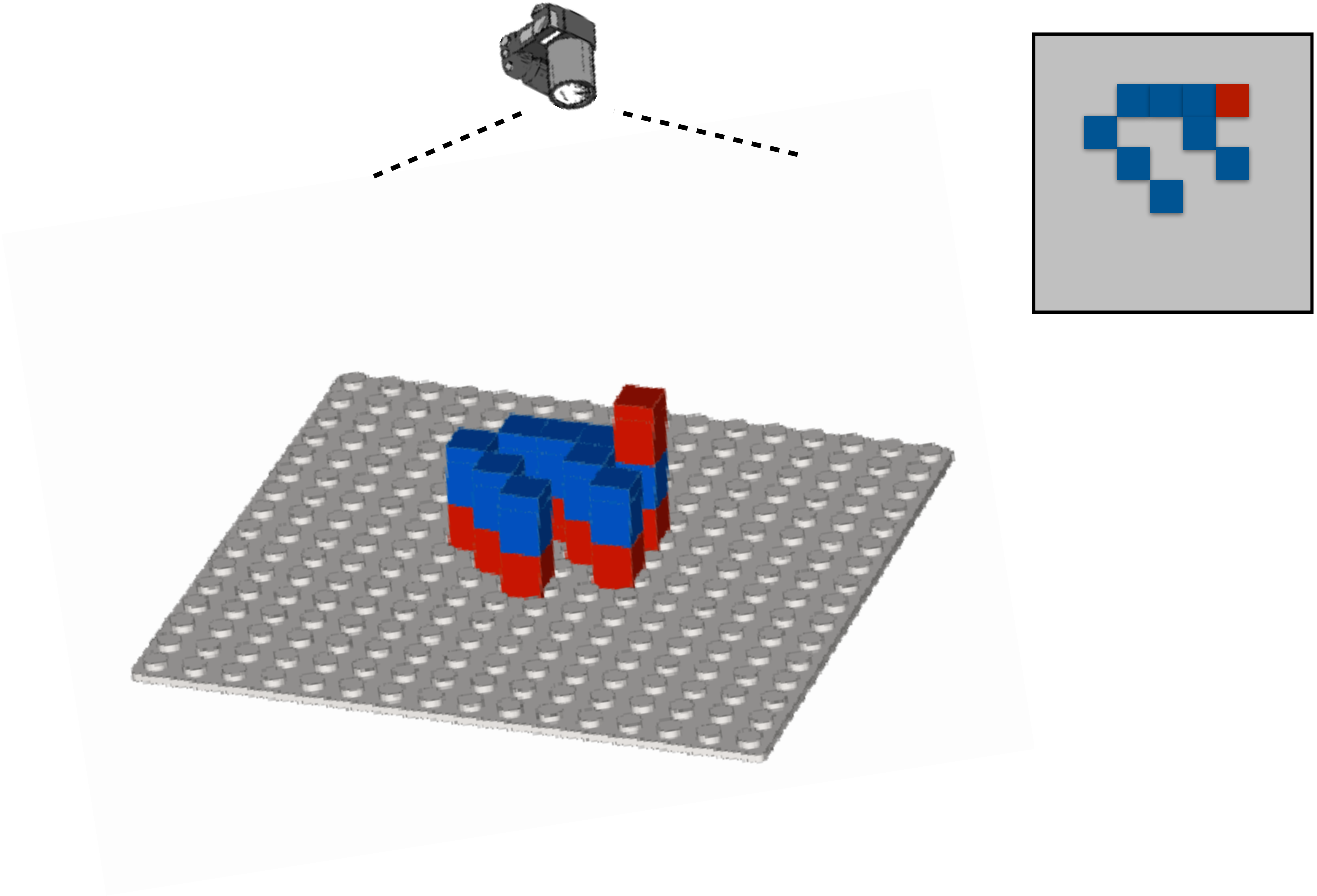


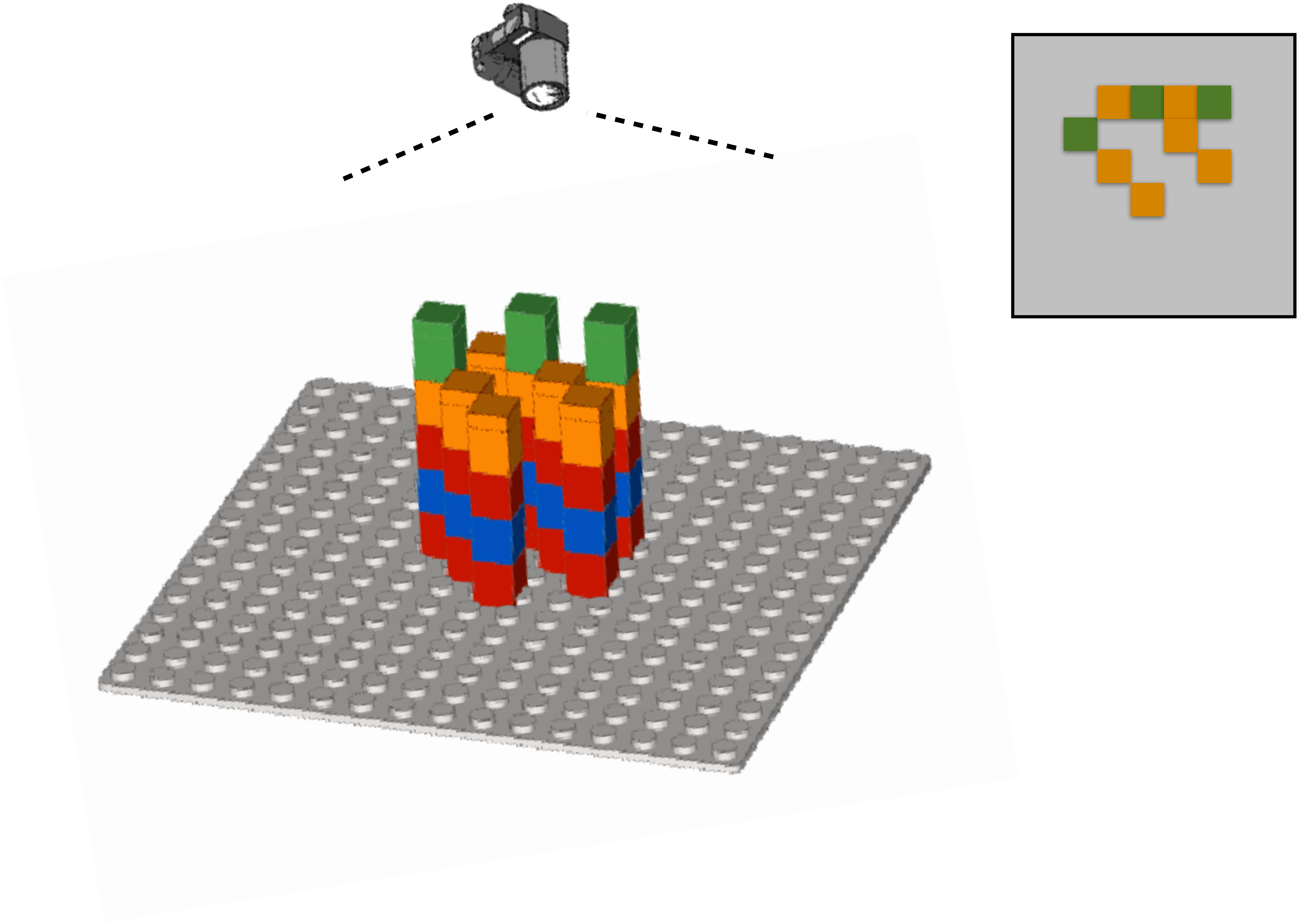
Cluster of clones



Ahead of schedule

Unterminated





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

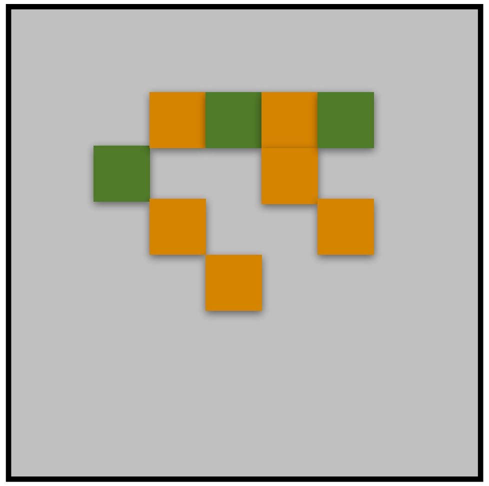
Base quality

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)

Estimate *p*, probability incorrect:

non-orange light / total light

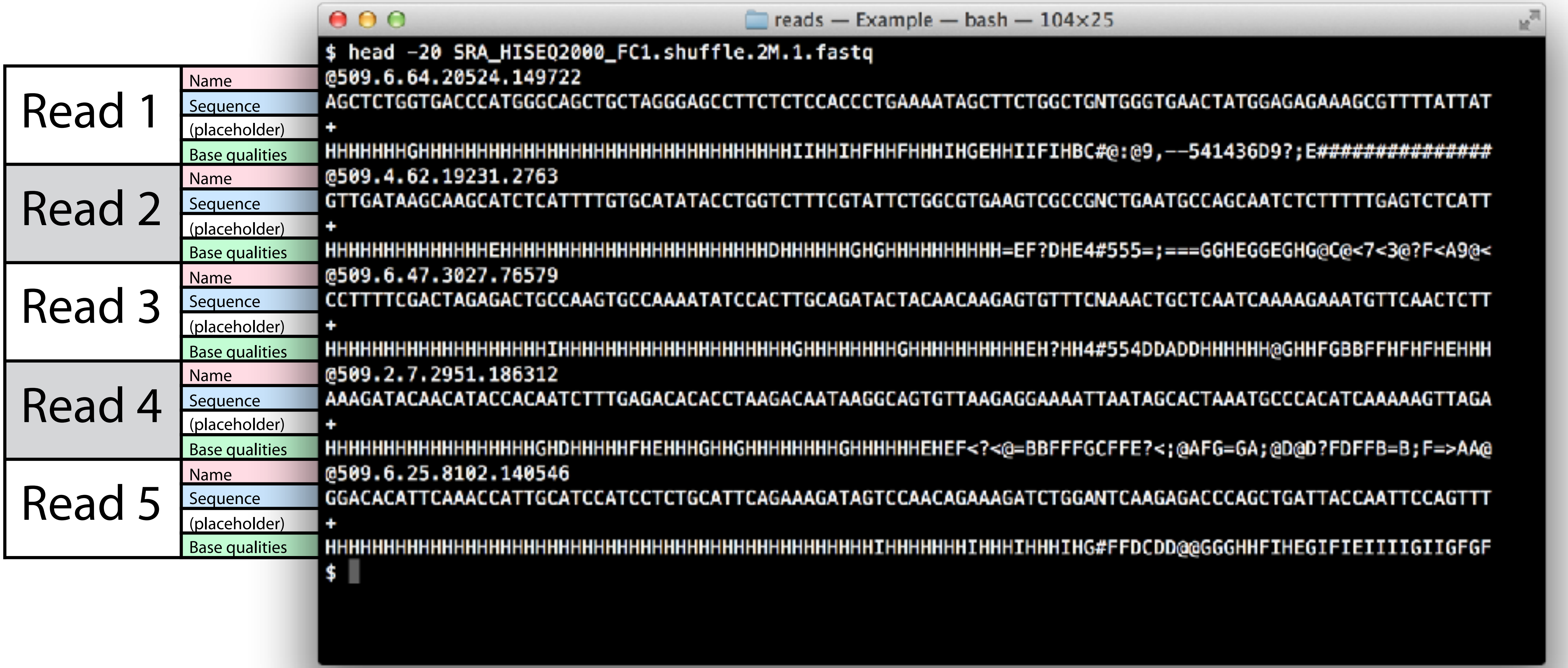
p = 3 green / 9 total = 1/3

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

A read in FASTQ format

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

FASTQ



```
$ head -20 SRA_HISEQ2000_FC1.shuffle.2M.1.fastq
@509.6.64.20524.149722
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGAGCCTTCTCTCCACCCTGAAAATAGCTTCTGGCTGNTGGGTGAACTATGGAGAGAAAGCGTTTTATTAT
+
HHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHIHFHHFHHHIHGEHHIIFIHBC#@:@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
+
HHHHHHHHHHHHHHHHHHGHDHHHHHFHEHHHGHHGHHHHHHHHGHHHHHHEHEF<?<@=BBFFFGCFFE?<;@AFG=GA;@D@D?FDFFB=B;F=>AA@
@509.6.25.8102.140546
GGACACATTCAAACCATTGCATCCATCCTCTGCATTCAGAAAGATAGTCCAACAGAAAGATCTGGANTCAAGAGACCCAGCTGATTACCAATTCCAGTTT
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHIHHHIHHHIHG#FFDCDD@@GGGHHFIHEGIFIEIIIIGIIGFGF
$ 
```

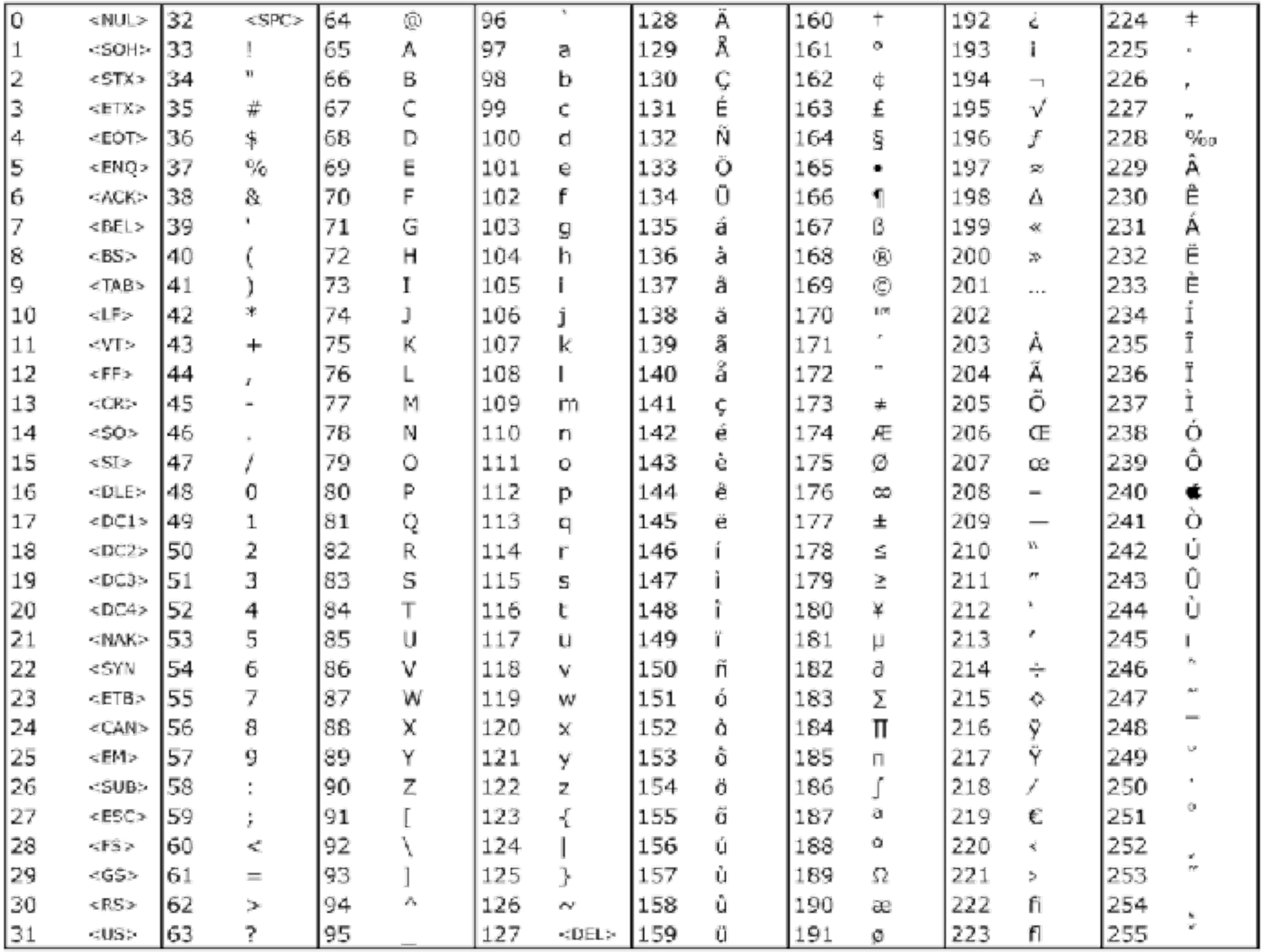
Base qualities

Bases and qualities line up:

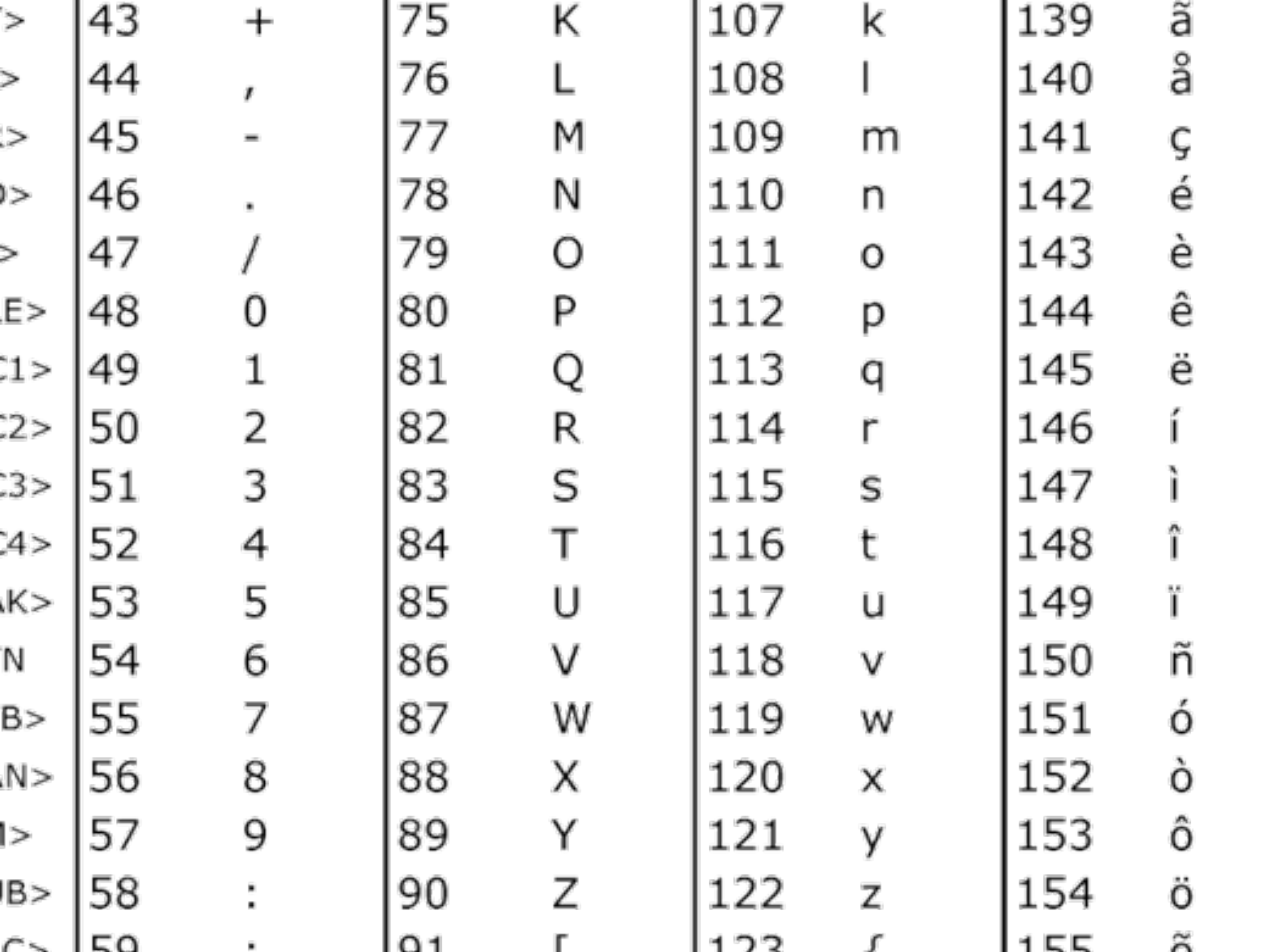
```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
||||||||||||||||||||||||||||||||
HHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

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

ASCII



0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	¡	225	·
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	‚
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
4	<EOT>	36	$	68	D	100	d	132	Ñ	164	§	196	ƒ	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Ö	165	•	197	≈	229	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	∆	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	…	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	´	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	¨	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Õ	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	–	240	
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	“	242	Ú
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	”	243	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	‘	244	Ù
21	<NAK>	53	5	85	U	117	u	149	ï	181	µ	213	’	245	ı
22	<SYN	54	6	86	V	118	v	150	ñ	182	∂	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	Σ	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	∏	216	ÿ	248	¯
25	<EM>	57	9	89	Y	121	y	153	ô	185	π	217	Ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	∫	218	⁄	250	˙
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	˚
28	<FS>	60	<	92	\	124	|	156	ú	188	º	220	‹	252	¸
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	›	253	˝
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	ﬁ	254	˛
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	ﬂ	255	ˇ



T>	43	+	75	K	107	k	139	ã
>	44	,	76	L	108	l	140	å
R>	45	-	77	M	109	m	141	ç
O>	46	.	78	N	110	n	142	é
>	47	/	79	O	111	o	143	è
LE>	48	0	80	P	112	p	144	ê
C1>	49	1	81	Q	113	q	145	ë
C2>	50	2	82	R	114	r	146	í
C3>	51	3	83	S	115	s	147	ì
C4>	52	4	84	T	116	t	148	î
AK>	53	5	85	U	117	u	149	ï
YN	54	6	86	V	118	v	150	ñ
B>	55	7	87	W	119	w	151	ó
AN>	56	8	88	X	120	x	152	ò
M>	57	9	89	Y	121	y	153	ô
UB>	58	:	90	Z	122	z	154	ö
C>	59	;	91	[	123	{	155	õ

Base qualities

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(int(round(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)