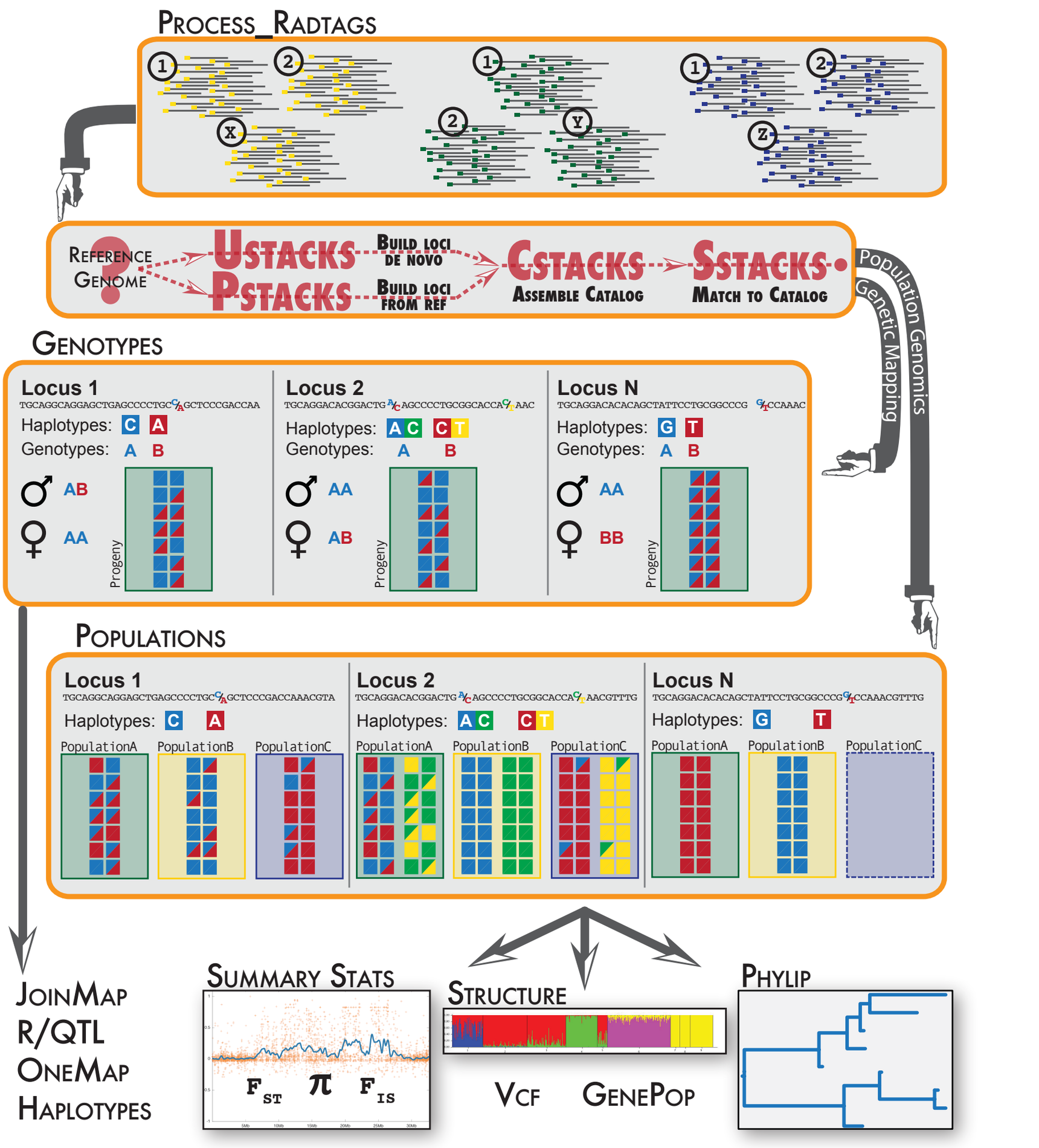
Cleaning, Demultiplexing, and Deduplicating with Stacks

Stacks



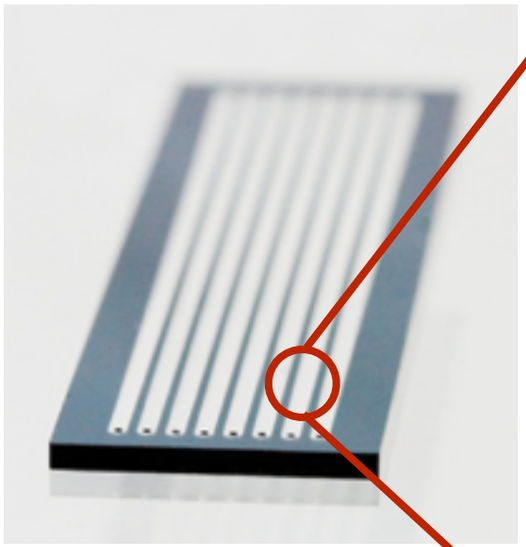
Stacks: Building and Genotyping Loci *De Novo* From Short-Read Sequences

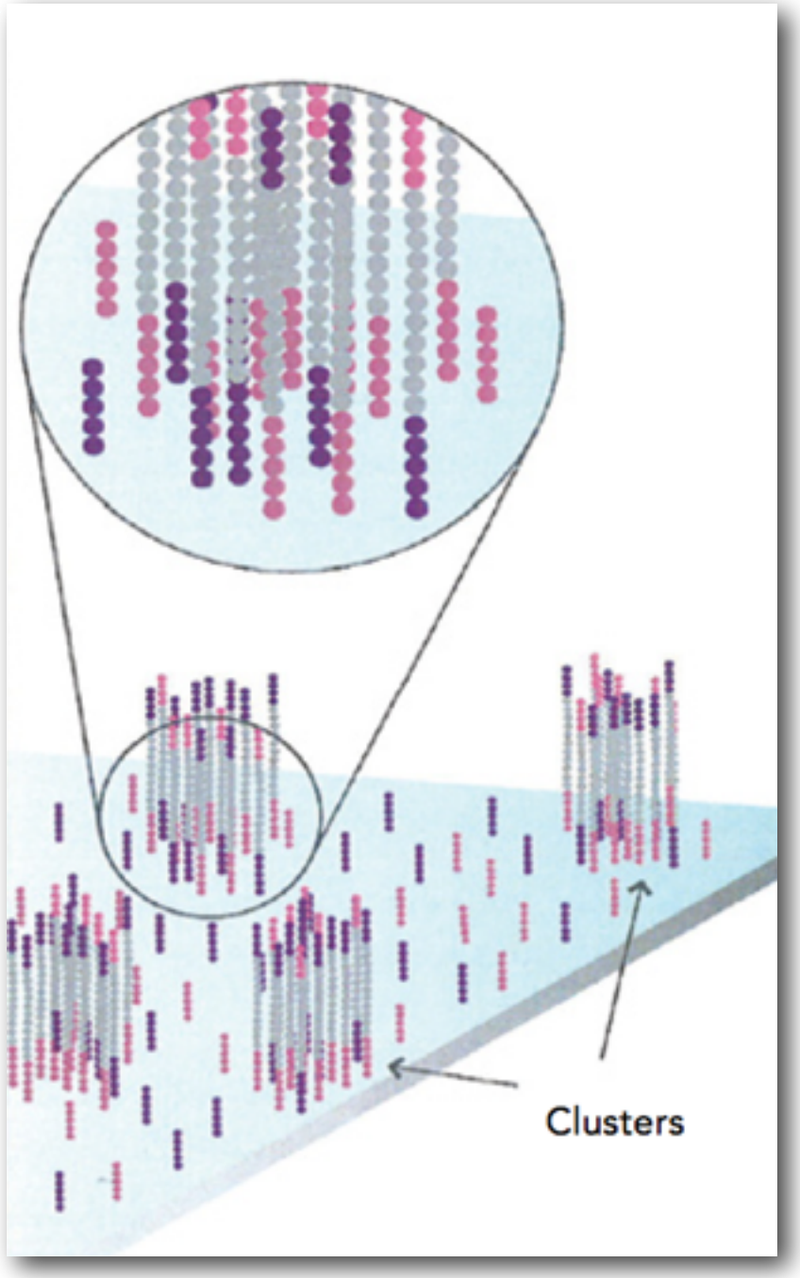
Julian M. Catchen,* Angel Amores,† Paul Hohenlohe,* William Cresko,* and John H. Postlethwait‡,1
*Center for Ecology and Evolutionary Biology and †Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403

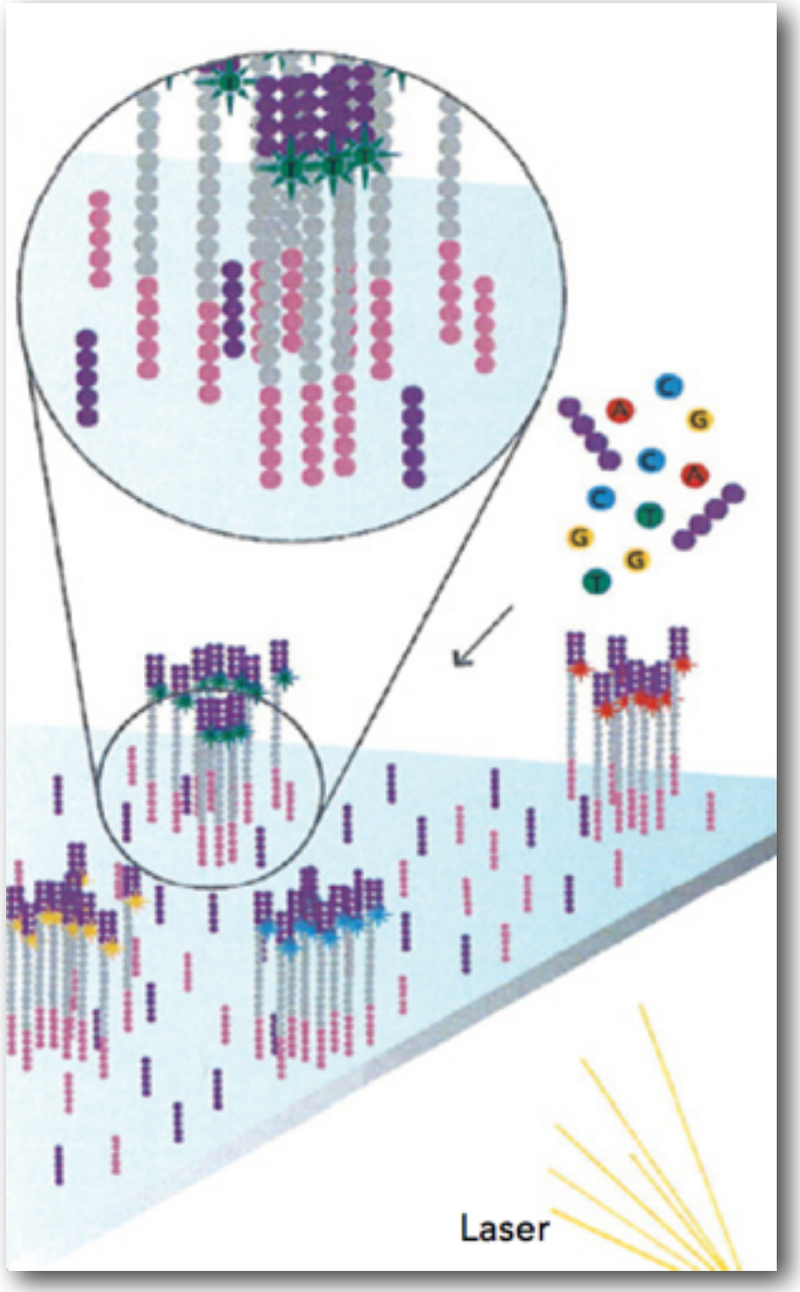
Stacks: an analysis tool set for population genomics

JULIAN CATCHEN,* PAUL A. HOHENLOHE,*† SUSAN BASSHAM,* ANGEL AMORES‡ and WILLIAM A. CRESKO*
*Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403-5289, USA, †Biological Sciences, University of Idaho, Moscow, ID 83844-3051, USA, ‡Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254, USA

Sequencing on Illumina's Flow cell

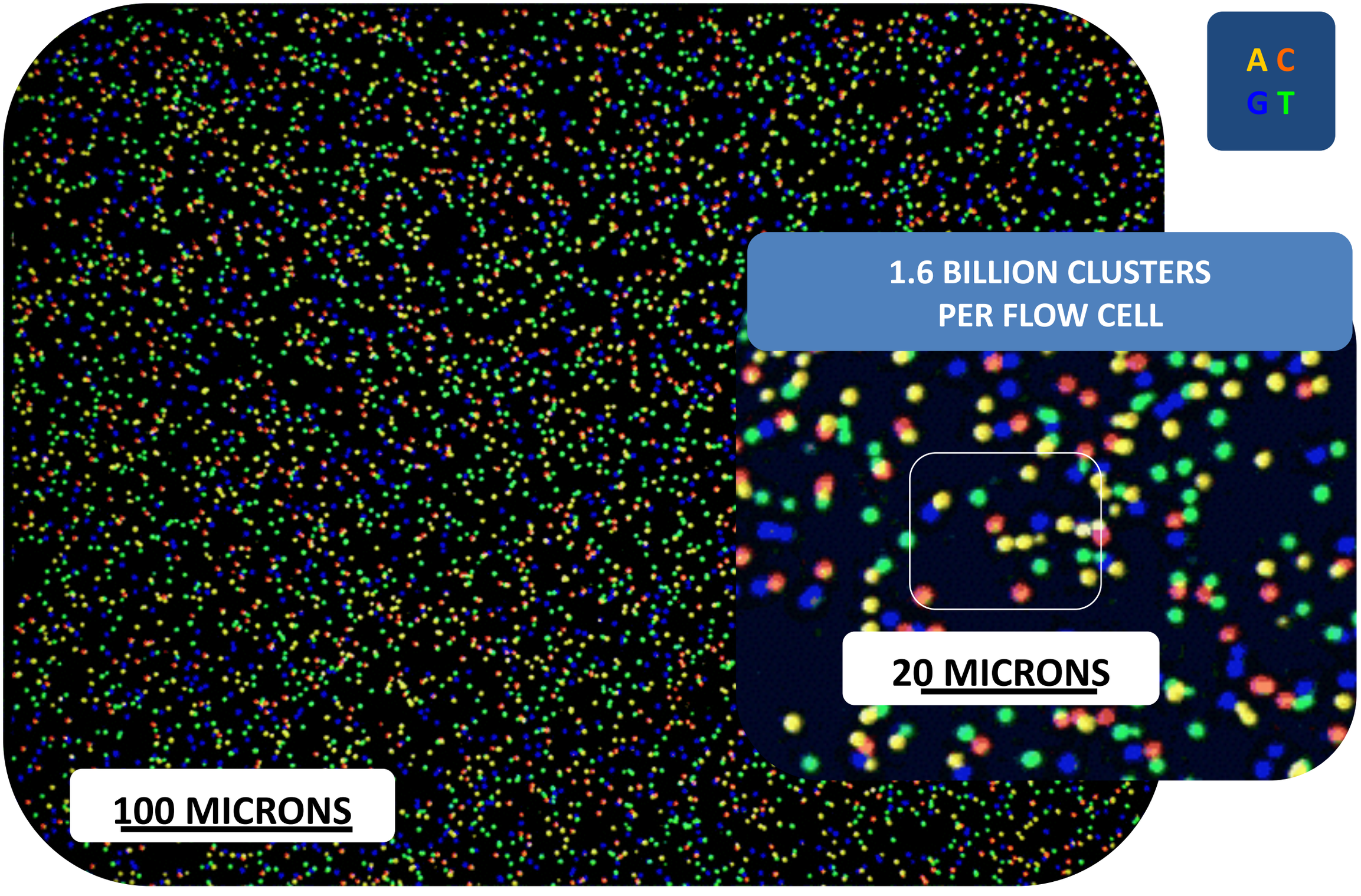




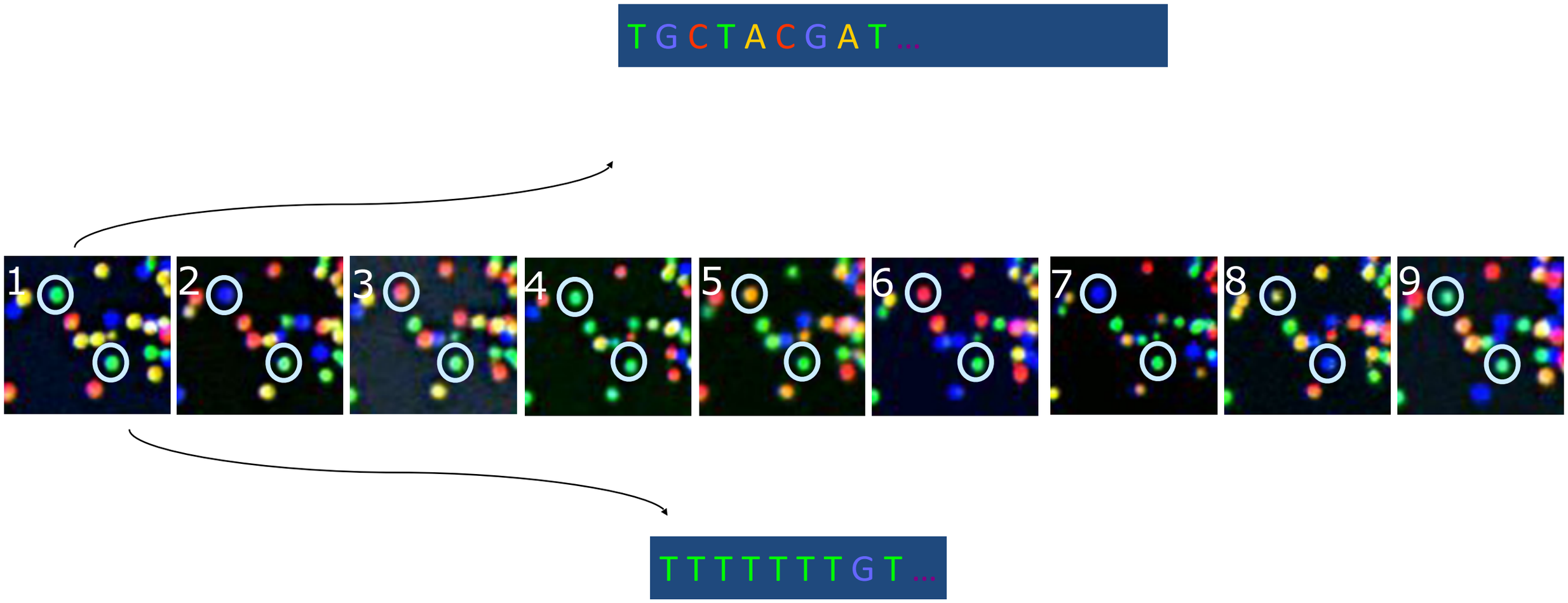


Source: Illumina Publication 770-2007-002 01May07

Illumina Sequencing : How it looks



Base calling from raw data

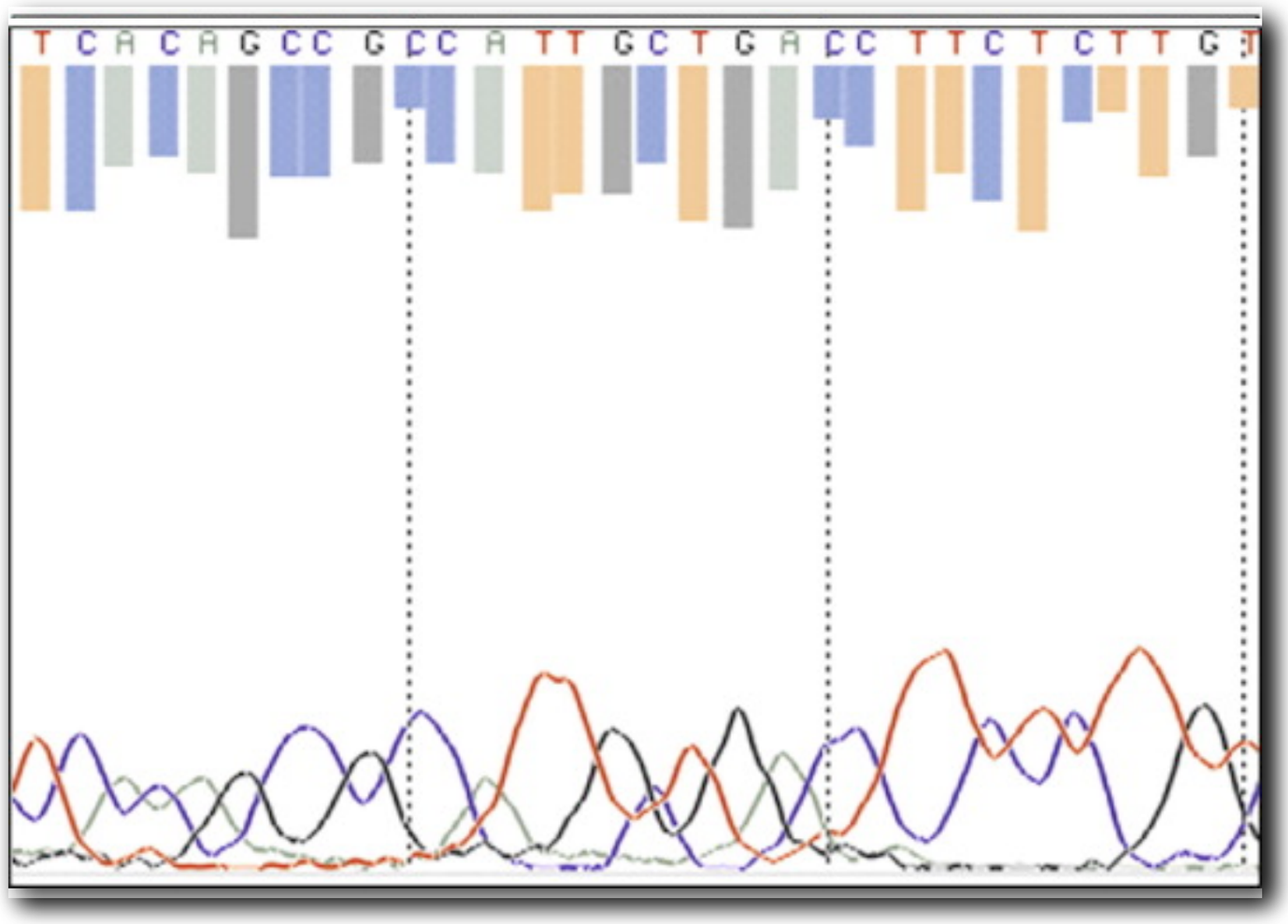


The identity of each base of a cluster is read off from sequential images.

Current read lengths = 36-150 nt

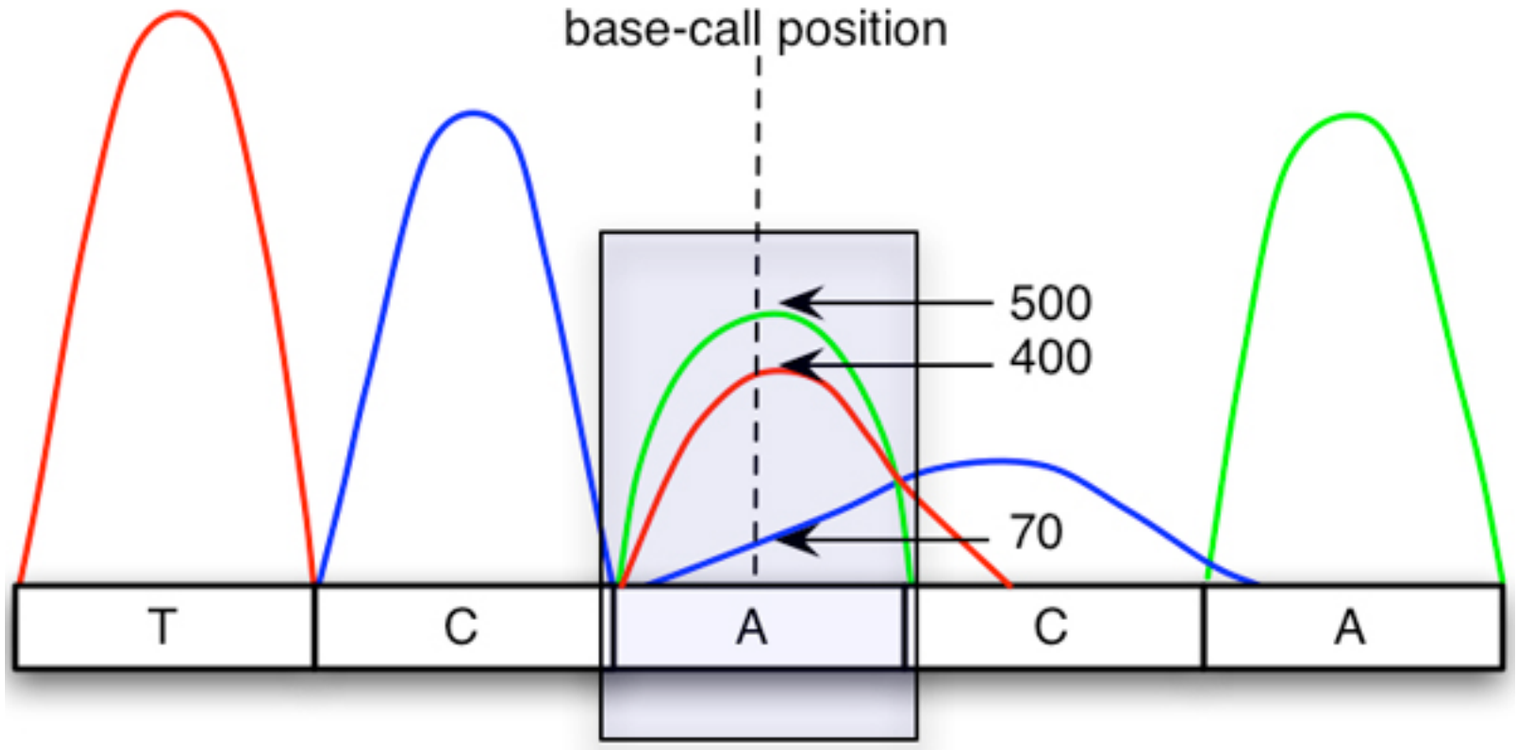
Total sequence data for 1 paired-end lane with 125bp = 500Gb!

Sequencing on Illumina's Flow cell, ctd.

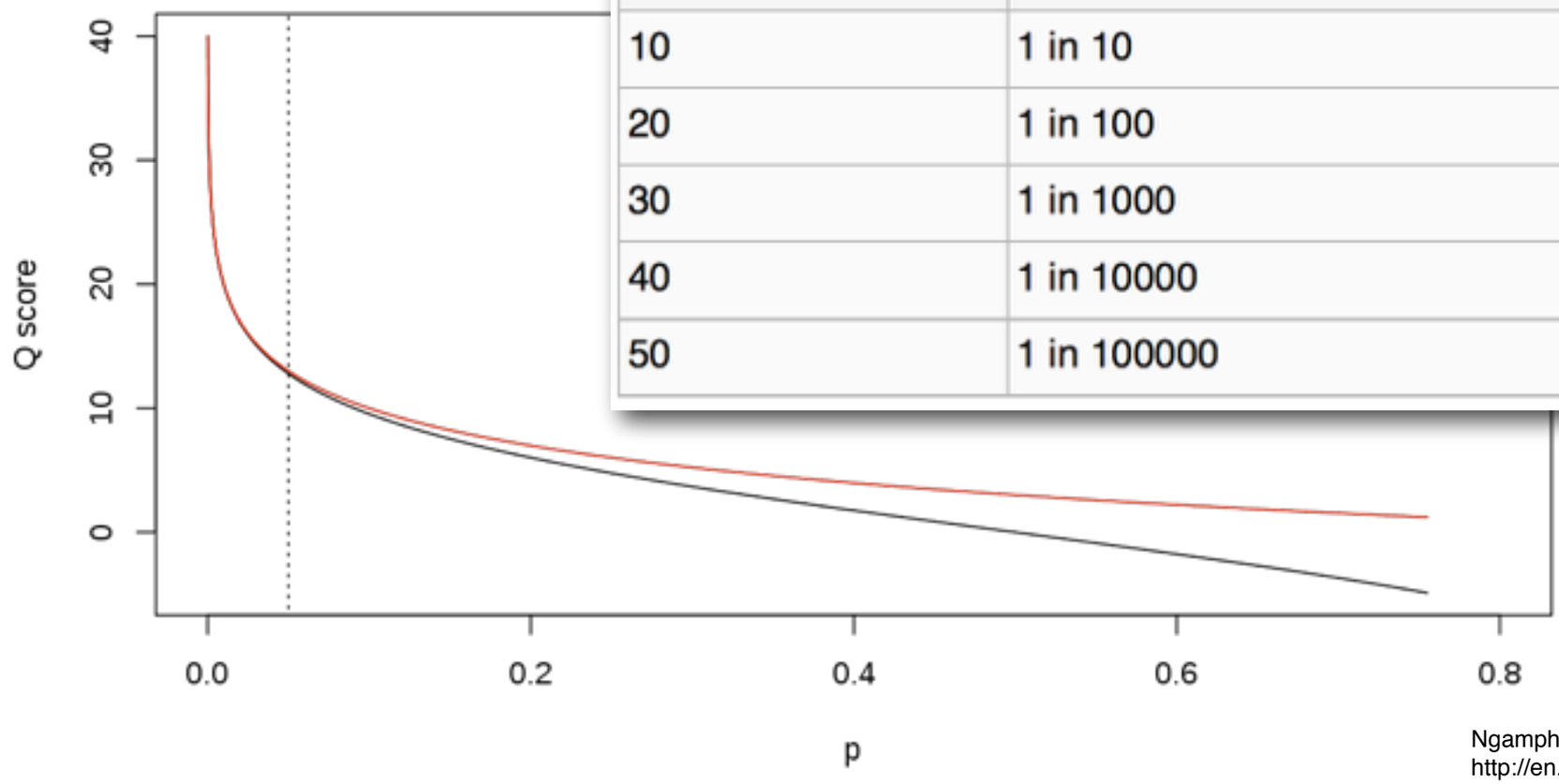


Source: Kim, et al. doi:10.1016/j.pnsc.2007.12.011

Phred Quality Score



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%



Ngamphiw, et al. doi:10.1186/1471-2105-9-S12-S9
http://en.wikipedia.org/wiki/FASTQ_format

The FASTQ File Format

FASTQ

```
@Sequence_137
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+Sequence_137
<?@DDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

```
@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+
<?@DDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

http://4toc.com/fb/FBHelp/gfx/AppF_ASCIITable.png

ASCII Code

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	Δ	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	∑	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

The FASTQ File Format, ctd

```
@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+
<?@DDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

Quality Scores

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
  ........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                          |    |        |                                  |                     |
 33                         59   64       73                                104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

ASCII values $33 - 73 = 0 - 40$

‘F’ = 70 **70 - 33 = 37**

http://en.wikipedia.org/wiki/FASTQ_format

Single-end, Inline Barcodes

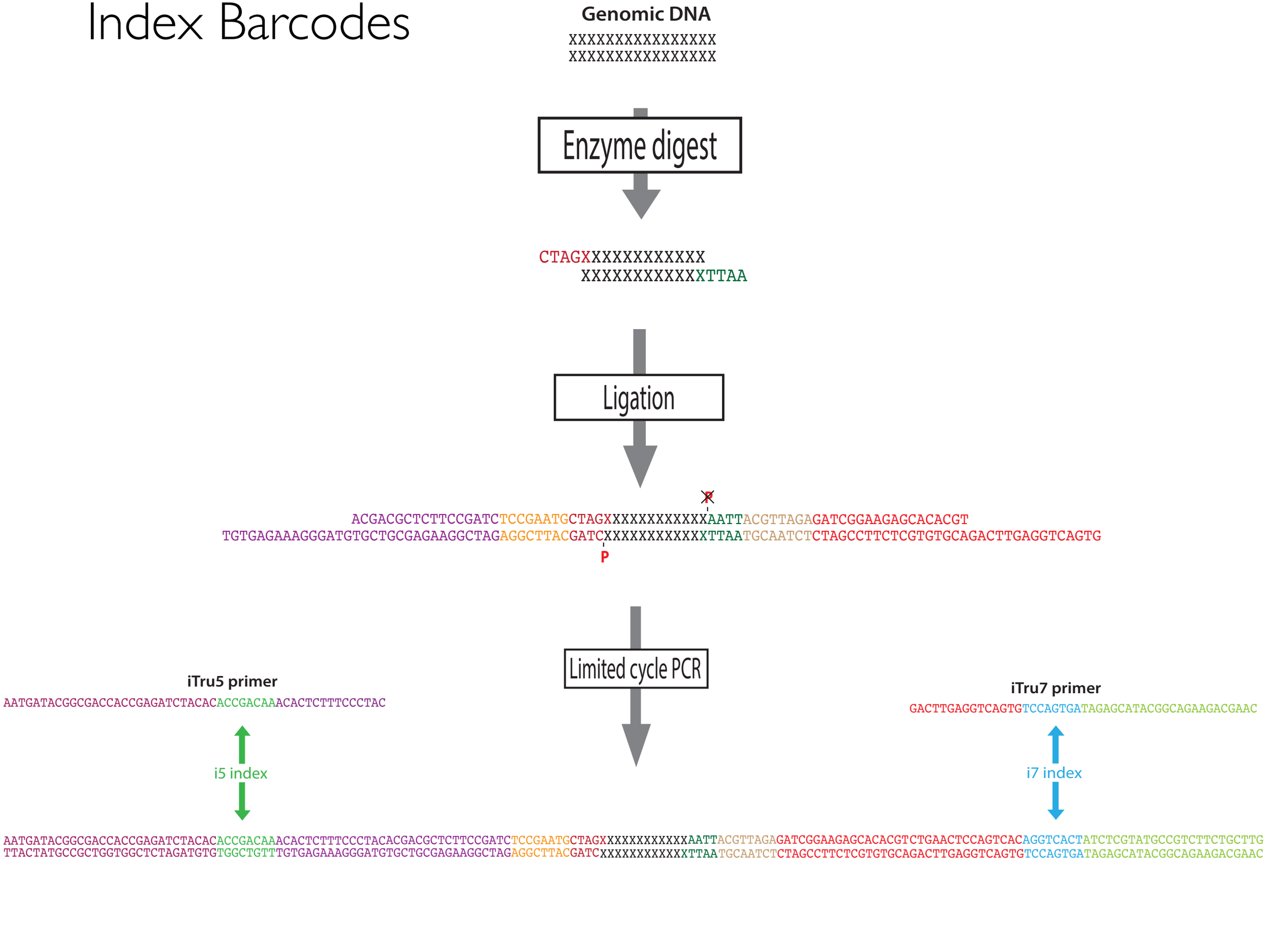
```
@HWI-ST0747:188:C09HWACXX:1:1101:2968:2083 1:N:0:
TTATGATGCAGGACCAGGATGACGTCAGCACAGTGCGGGTCCTCCATGGATGCTCCTCGGTCGTGGTTGGGGGAGGAGGCA
+
@@@DDDDDBHHFBF@CCAGEHHHBFGIIFGIIGIEDBBGFHCGIIGAEEEDCC;A?;;5,:@A?=B5559999B@BBBBBA
@HWI-ST0747:188:C09HWACXX:1:1101:2863:2096 1:N:0:
TTATGATGCAGGCAAATAGAGTTGGATTTTGTGTCAGTAGGCGGTTAATCCCATACAATTTTACACTTTATTCAAGGTGGA
+
CCCFFFFFHHHHHJJGHIGGAHHIIGGIIJDHIGCEGHIFIJIH7DGIIIAHIJGEDHIDEHJJHFEEECEFEFFDECDDD
@HWI-ST0747:188:C09HWACXX:1:1101:2837:2098 1:N:0:
GTGCCTTGCAGGCAATTAAGTTAGCCGAGATTAAGCGAAGGTTGAAAATGTCGGATGGAGTCCGGCAGCAGCGAATGTAAA
```

Paired-end, Inline Barcodes

```
@9432NS1:54:C1K8JACXX:7:1101:5584:1725 1:N:0:
ACTGGCATGATGATCATAGTATAACGTGGGATACATATGCCTAAGGCTAAAGATGCCTTGAAGCTTGGCTTATGTT
+
DDDBHHFBF@CCAGEHHHBFGIIFGIIGIEDBBGFHCGIIGAEEEDCC;A?;;5,:@A?=B5559999B@BBBBA
@9432NS1:54:C1K8JACXX:7:1101:5708:1737 1:N:0:
TTCGACATGTGTTTACAACGCGAACGGACAAAGCATTGAAAATCCTTGTTTTGGTTTCGTTACTCTCTCCTAGCAT
+
CCCFFFFFHHHHHJJGHIGGAHHIIGGIIJDHIGCEGHIFIJIH7DGIIIAHIJGEDHIDEHJJHFEEECEFEFFD
```

```
@9432NS1:54:C1K8JACXX:7:1101:5584:1725 2:N:0:
AATTTACTTTGATAGAAGAACAACATAAGCCAAGCTTCAAGGCATCTTTAGCCTTAGGCATATGTATCCCACGTTA
+
@@@DDDDDBHHFBF@CCAGEHHHBFGIIFGIIGIEDBBGFHCGIIGAEEEDCC;A?;;5,:@A?=B5559999B@B
@9432NS1:54:C1K8JACXX:7:1101:5708:1737 2:N:0:
AGTCTTGTGAAAAACGAAATCTTCCAAAATGCTAGGAGAGAGTAACGAAACCAAAACAAGGATTTTCAATGCTTTG
+
CCCFFFFFHHHHHJJGHIGGAHHIIGGIIJDHIGCEGHIFIJIH7DGIIIAHIJGEDHIDEHJJHFEEECEFEFFD
```

Index Barcodes



Index Barcodes

```
@9432NS1:54:C1K8JACXX:8:1101:6912:1869 1:N:0:ATGACT
TCAGGCATGCTTTCGACTATTATTGCATCAATGTTCTTTGCGTAATCAGCTACAATATCAGGTAATATCAGGCGCA
+
CCCFFFFFHHHHHJJJJJJJJIJJJJJJJJJJJJHIIJJJJJJIJJJJJJJJJJJJJJJJJJJJGIJJJJJJJHHHFF
@9432NS1:54:C1K8JACXX:8:1101:6822:1873 1:N:0:ATGACT
CAGCGCATGAGCTAATGTATGTTTTACATTCCAGAAAGAGAGCTACTGCTGCAGGTTGTGATAAAATAAAGTAAGA
+
B@@FFFFFHFFHHJJJJFHIJHGGGHIJIIJIJCHJIIGGIIIGGIJEHIJJHII?FFHICHFFGGHIIGG@DEHH
@9432NS1:54:C1K8JACXX:8:1101:6793:1916 1:N:0:ATGACT
TTTCGCATGCCCTATCCTTTTATCACTCTGTCATTCAGTGTGGCAGCGGCCATAGTGTATGGCGTACTAAGCGAAA
+
@C@DFFFFHGHHHGIGHHJJJJJJJGIJIJJIGIJJJJHIGGGHGII@GEHIGGHDHEHIHD6?493;AAA?;=;=
```

Paired-end, Index Barcodes

```
@NS500216:197:HF7C5BGXX:1:11101:1693:1047 1:N:0:CTCCATGT+GGGGGGGG
TTAGGNAATCGGGCCTTCAGGAATGCTCCACGGCAAGGGGGCCACTAGGAACTCCAGTTCTTCACCATGTGGTGAC
+
AAAAA#E/EEEEEEEAEEEEEEEEEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEE
@NS500216:197:HF7C5BGXX:1:11101:4119:1051 1:N:0:TCGCTGTT+CTTCGTTC
GATACNGCTAGAAGTGTTATAAAATGCTAACTTGTAATGAATGTAAGAAAGGAGGATGAAGATGTGGTGTAGTGGT
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEAEEEEEEAEEEEEEEEEEAAEEEAEEEEEE
```

PCR Duplicates

Round 1

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

Round 2

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

Round 3

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

Round 4

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

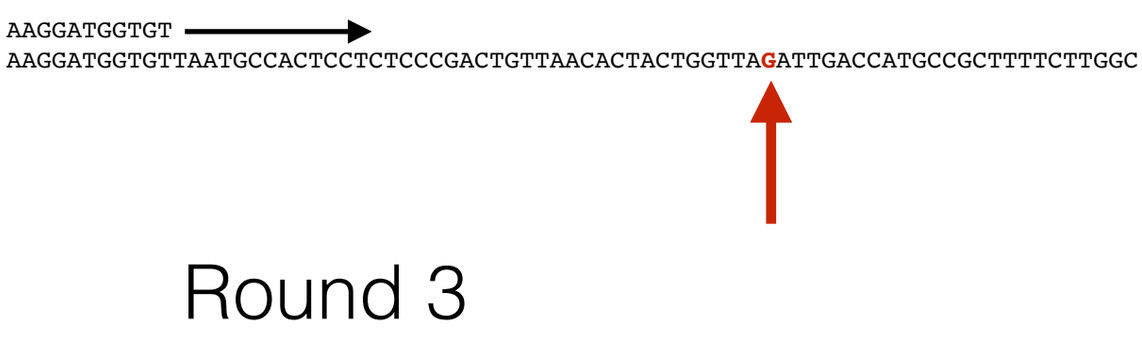
AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

AAGGATGGTGT →
AAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGC

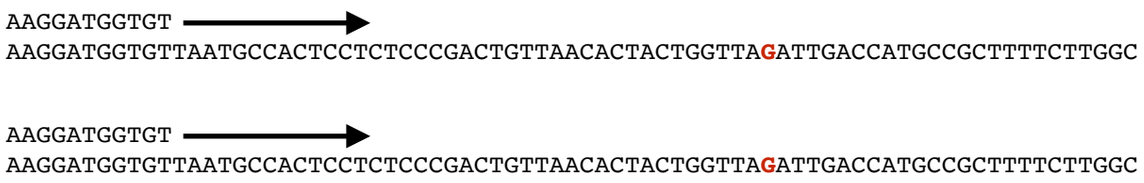
PCR Duplicates

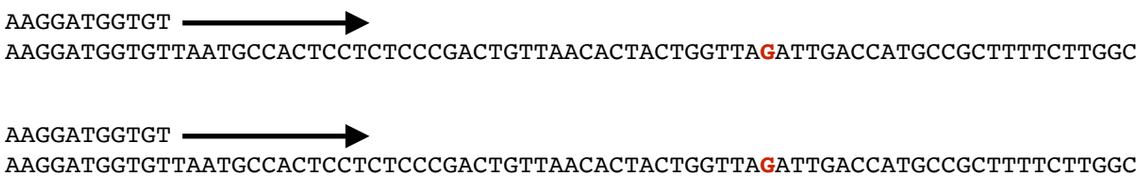
Round 1

Round 2

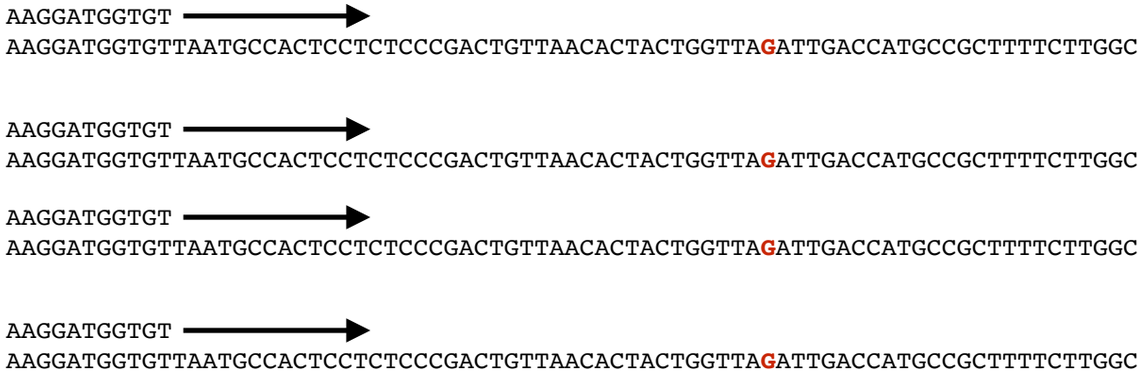


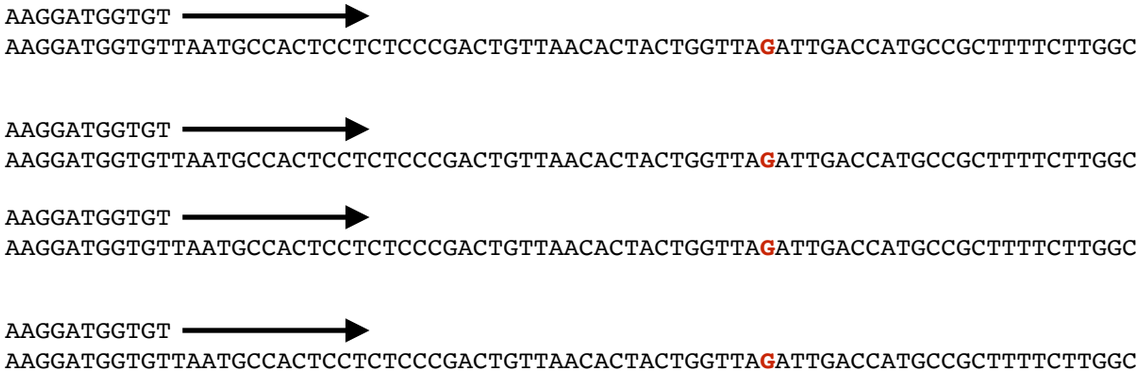
Round 3





Round 4





process_radtags

```
@Sequence_137
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+Sequence_137
<?@DDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

```
<?@DDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```