Quality control and trimming

Igor S. Pessi & Antti Karkman, University of Helsinki

Quality control

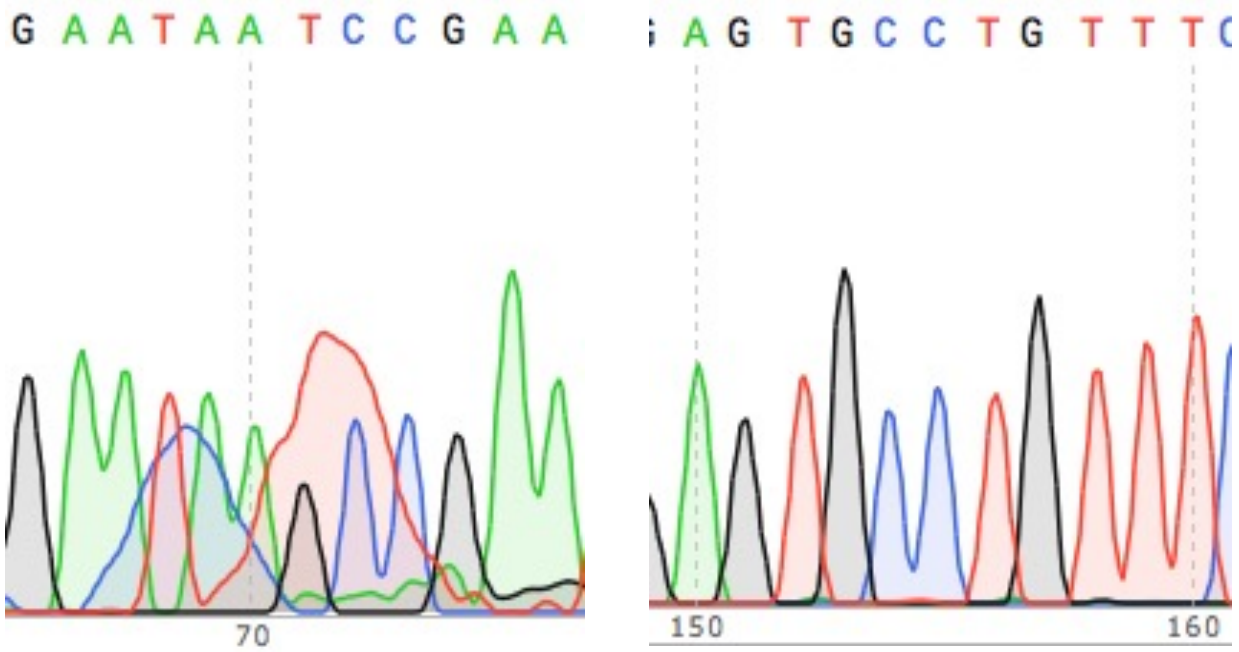
- Things to consider
  - Sequence quality
  - Adapter contamination

Sequence quality

- PCR – what happens in PCR, stays in PCR

Sequence quality

- PCR – what happens in PCR, stays in PCR
- Sequencing

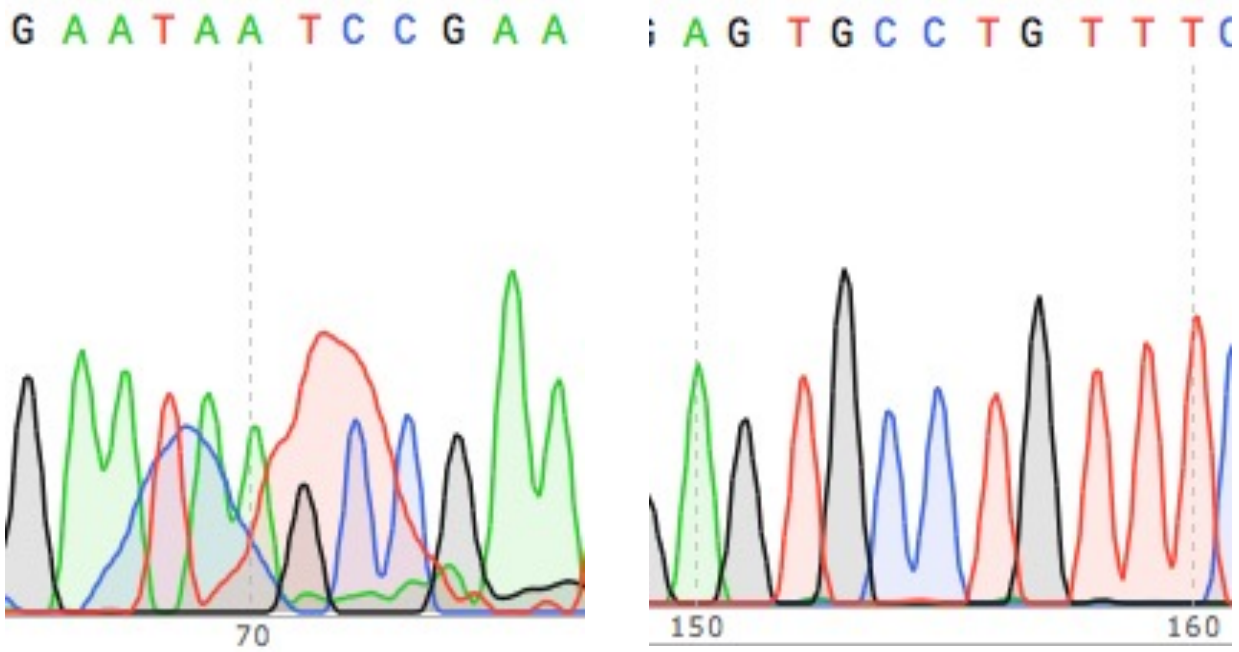


https://bitesizebio.com/10709/how-to-analyze-dna-sequencing-results-properly/

Sequence quality

- PCR – what happens in PCR, stays in PCR
- Sequencing

```
@A00464:69:H7L3KDSXX:1:1101:7726:1000 1:N:0:GTCTAATG+NGTGGTCA
GNATTGCCATCGGGTAAAGCGTCAGGAAGCCGAGCAGCAGGATCAGCAGGGCGATGCCGGCAG
GCCGGCCGAAGCGGCGGAGGCCGCGCGGCAGCGCAGCCGTCGGAAGAGTGACGGCAACCGTCA
TCGGCAGCCGCCCATCGTCAATCGT
+
F#FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFF
```



https://bitesizebio.com/10709/how-to-analyze-dna-sequencing-results-properly/

Sequence quality

- PCR – what happens in PCR, stays in PCR
- Sequencing

```
@A00464:69:H7L3KDSXX:1:1101:7726:1000 1:N:0:GTCTAATG+NGTGGTCA
GNATTGCCATCGGGTAAAGCGTCAGGAAGCCGAGCAGCAGGATCAGCAGGGCGATGCCGGCAG
GCCGGCCGAAGCGGCGGAGGCCGCGCGGCAGCGCAGCCGTCGGAAGAGTGACGGCAACCGTCA
TCGGCAGCCGCCCATCGTCAATCGT
+
F#FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFF
```

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Sequence quality

- PCR – what happens in PCR, stays in PCR
- Sequencing

```
@A00464:69:H7L3KDSXX:1:1101:7726:1000 1:N:0:GTCTAATG+NGTGGTCA
GNATTGCCATCGGGTAAAGCGTCAGGAAGCCGAGCAGCAGGATCAGCAGGGCGATGCCGGCAG
GCCGGCCGAAGCGGCGGAGGCCGCGCGGCAGCGCAGCCGTCGGAAGAGTGACGGCAACCGTCA
TCGGCAGCCGCCCATCGTCAATCGT
+
F#FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFF
```

Q = 0
E = 1

ASCII 33

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Sequence quality

- PCR – what happens in PCR, stays in PCR
- Sequencing

```
@61fab8a7-c3c9-4850-a217-f44f926c9fb9 runid=5f43d83b33652c5908fece0571c30130da773d95
sampleid=m12208 read=28498 ch=109 start_time=2020-06-30T23:21:26Z
ATCGGTATGCTTCGTTCAATTCATATTGCCTCGTAGCTGGACCACTATTTTCGTCTGAATCAGGATGTAAGGACATGGGTAATTC
AATTCCATAGCTGAATGGTGTGGAGCACAGATTATATCTTTCTCTCTCTTCTATTAATACAGAATGGAAGCAGCT
TTGATAGATAAAATTTTTTCCATATGTTGAATTGAATCATTTCTTTCATGATCTCTTCTTTCAAATATTTTAAACAATATTCGTC
TAAAGTGCAGAAATGTTGTCAGTTAGTACTGATTATTTGAGTGAAAAGCATGTGTTTGGTGTCAGCCAAACTCTT
GATTTAATGCATTAGCTATCAGATAATTCATCCTGAACCGACCAGTTAGCTTGCGTTTTTT
+
$$%#%%$&$)*)&+06+$$.)%&%%&)00/5'$$&'(,%'$$&*026@EE<<7;A8>B?@/.<<ED400267<-
)%$&%%%$%,$&%$(((%&%18:78/>:6>>9<92;<4/)++02',*.+''(%$$#&''*-
%***.53+%$&$1&+%%/)('&$#$
$&')634=<<73479??A-00.5.64,/),+2523<CA?BKB;EGCD63102.*-
0+$)+**34<.*;==399@;BD83==81>;499.&,%&''*.34(?(+9,*&88&4:9;2.0;=245;6(0..0--
,'()5303CD>0/55:/2'..6846C820
;<@D5--.%&&+/+&&(&('%&%$#%$$%$12,($(&'(&;0+')0.&+432-,3>BCB9.
```

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
32	20	[SPACE]	64	40	@	96	60	`
33	21	!	65	41	A	97	61	a
34	22	"	66	42	B	98	62	b
35	23	#	67	43	C	99	63	c
36	24	$	68	44	D	100	64	d
37	25	%	69	45	E	101	65	e
38	26	&	70	46	F	102	66	f
39	27	'	71	47	G	103	67	g
40	28	(	72	48	H	104	68	h
41	29	)	73	49	I	105	69	i
42	2A	*	74	4A	J	106	6A	j
43	2B	+	75	4B	K	107	6B	k
44	2C	,	76	4C	L	108	6C	l
45	2D	-	77	4D	M	109	6D	m
46	2E	.	78	4E	N	110	6E	n
47	2F	/	79	4F	O	111	6F	o
48	30	0	80	50	P	112	70	p
49	31	1	81	51	Q	113	71	q
50	32	2	82	52	R	114	72	r
51	33	3	83	53	S	115	73	s
52	34	4	84	54	T	116	74	t
53	35	5	85	55	U	117	75	u
54	36	6	86	56	V	118	76	v
55	37	7	87	57	W	119	77	w
56	38	8	88	58	X	120	78	x
57	39	9	89	59	Y	121	79	y
58	3A	:	90	5A	Z	122	7A	z
59	3B	;	91	5B	[	123	7B	{
60	3C	<	92	5C	\	124	7C	|
61	3D	=	93	5D	]	125	7D	}
62	3E	>	94	5E	^	126	7E	~
63	3F	?	95	5F	_	127	7F	[DEL]

Sequence quality

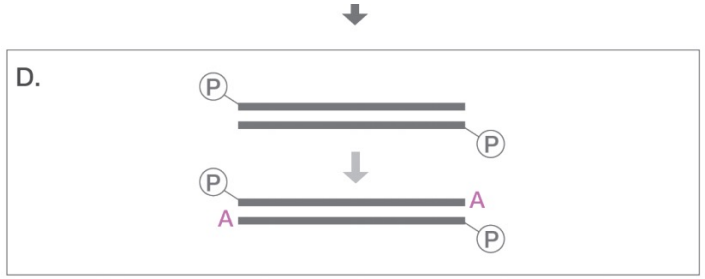
- PCR – what happens in PCR, stays in PCR
- Sequencing
- Illumina:
  - Occurrence: substitutions > indels
  - Quality scores: substitutions < indels
  - Overall quality: R1 > R2; beginning > end

> ”G seems to be preferentially incorporated if an A, C or T is sequenced and if G is sequenced a T is falsely incorporated for the majority of substitutions.”
>
> Schirmer et al. 2016. BMC Bioinformatics

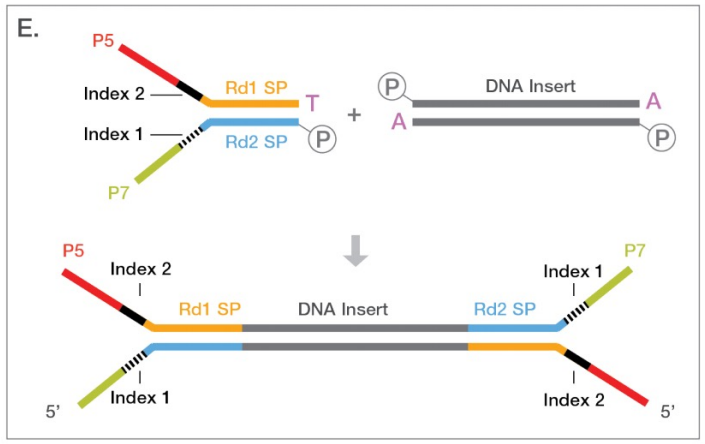
Sequence quality

- PCR – what happens in PCR, stays in PCR
- Sequencing
- Illumina:
  - Occurrence: substitutions > indels
  - Quality scores: substitutions < indels
  - Overall quality: R1 > R2; beginning > end
- Nanopore: substitutions

Adapter contamination

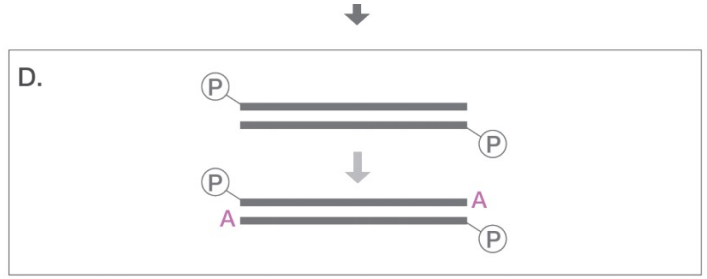


A-base is added.

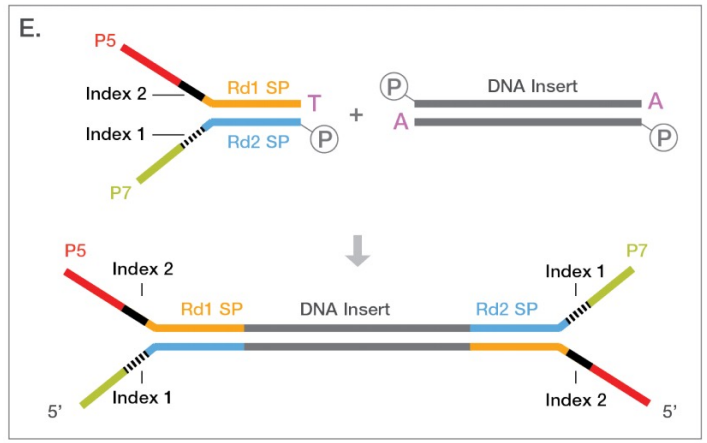


Dual-index adapters are ligated to the fragments* and final product is ready for cluster generation.

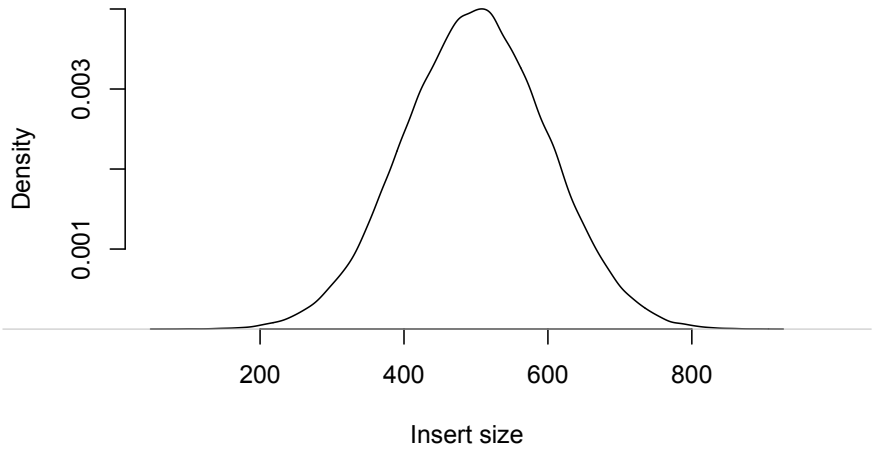
Adapter contamination



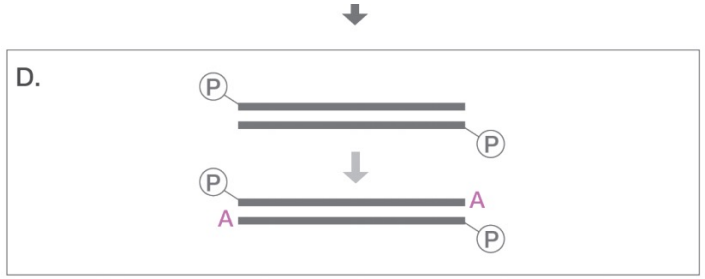
A-base is added.



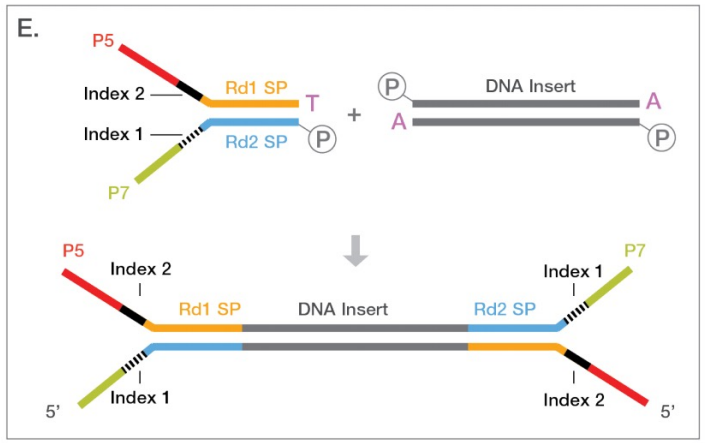
Dual-index adapters are ligated to the fragments* and final product is ready for cluster generation.



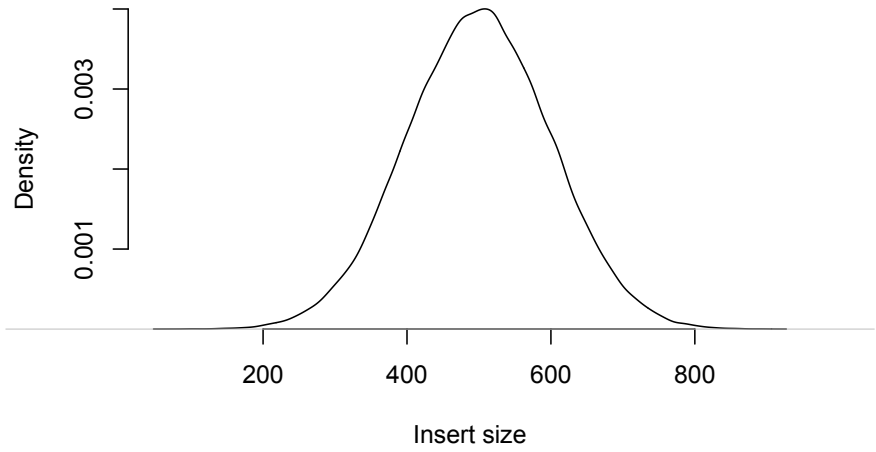
Adapter contamination

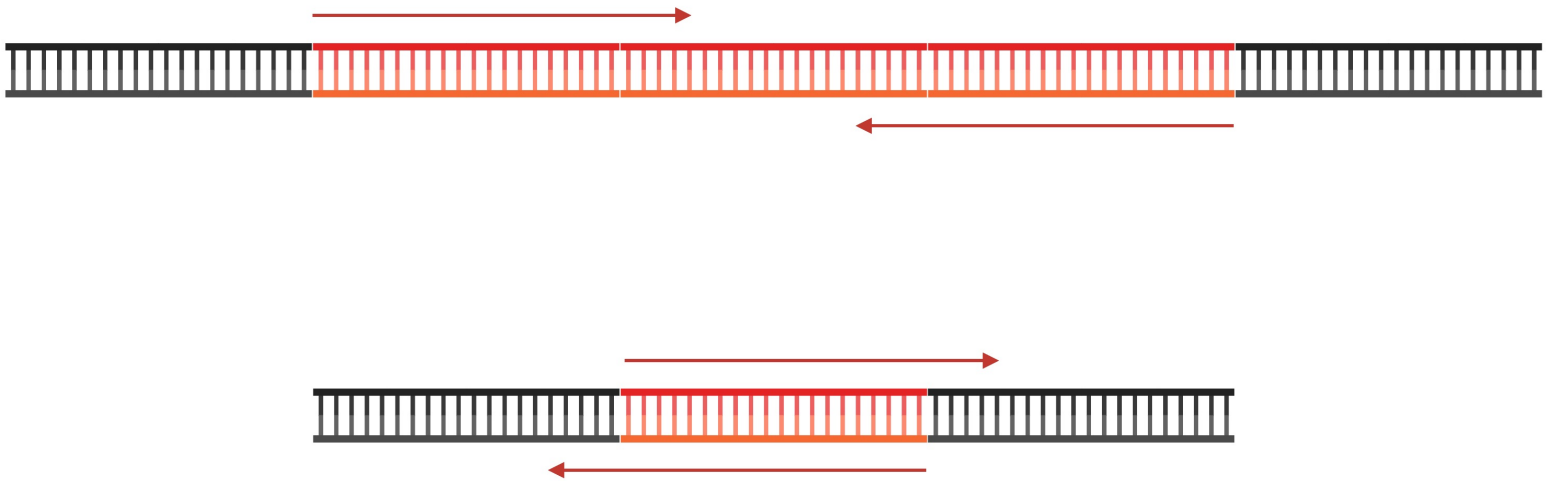


A-base is added.



Dual-index adapters are ligated to the fragments* and final product is ready for cluster generation.





Trimming

- What to do then?