

DTU





**DTU Health Technology  
Bioinformatics**

## **Data basics**

*Shyam Gopalakrishnan*  
*Associate Professor*  
*Section of Evolutionary Genomics, KU*  
*Section of Bioinformatics, DTU*  
*shyam.g@gmail.com*

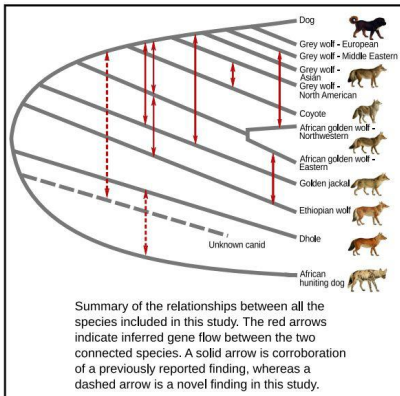
# About me

- + PhD in Biostatistics – Population genetics, University of Michigan
- + Postdoc at University of Chicago
- + Associate Professor at Globe Institute, KU
- + Associate Professor at Bioinformatics, Health Tech, DTU

## Current Biology

### Interspecific Gene Flow Shaped the Evolution of the Genus *Canis*

#### Graphical Abstract



#### Authors

Shyam Gopalakrishnan, Mikkel-Holger S. Sinding, Jazmin Ramos-Madriral, ..., Øystein Wiig, Anders J. Hansen, M. Thomas P. Gilbert

#### Correspondence

shyam@snm.ku.dk

#### In Brief

Gopalakrishnan et al. present evidence of pervasive gene flow among species of the genus *Canis*. In addition to previously known admixture events, they find evidence of gene flow from a “ghost” canid, related to the dhole, into the ancestor of the gray wolf and coyote. Further, they suggest that the African golden wolf is a species of hybrid origin.

## MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE | Open Access | CC BY

### Using DNA metabarcoding for simultaneous inference of common vampire bat diet and population structure

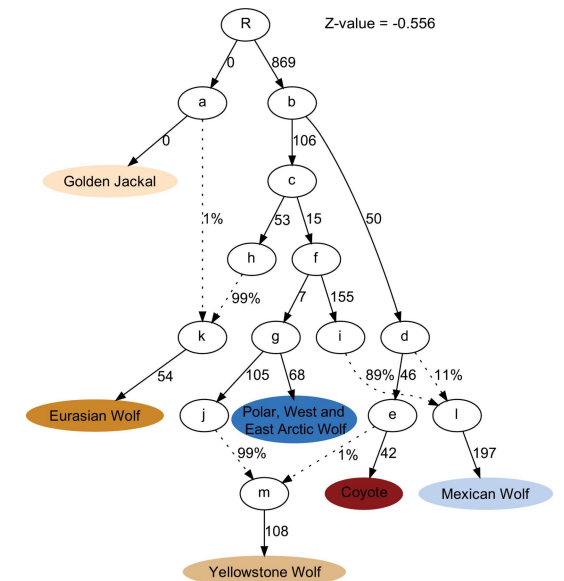
Kristine Bohmann, Shyam Gopalakrishnan, Martin Nielsen, Luisa dos Santos Bay Nielsen, Gareth Jones, Daniel G. Streicker, M. Thomas P. Gilbert

First published: 19 April 2018 | <https://doi.org/10.1111/1755-0998.12891> | Cited by: 2

### Ancient genomes from Iceland reveal the making of a human population

S. Sunna Ebenesersdóttir<sup>1,2,\*</sup>, Marcela Sandoval-Velasco<sup>3</sup>, Ellen D. Gunnarsdóttir<sup>1,2</sup>, Anuradha Jagadeesan<sup>1,2</sup>, Valdis B. G...  
\* See all authors and affiliations

Science 01 Jun 2018;  
Vol. 360, Issue 6392, pp. 1028-1032  
DOI: 10.1126/science.aar2625



# Outline

- Starting point – what we learned yesterday
- Sequence data storage

Fasta format

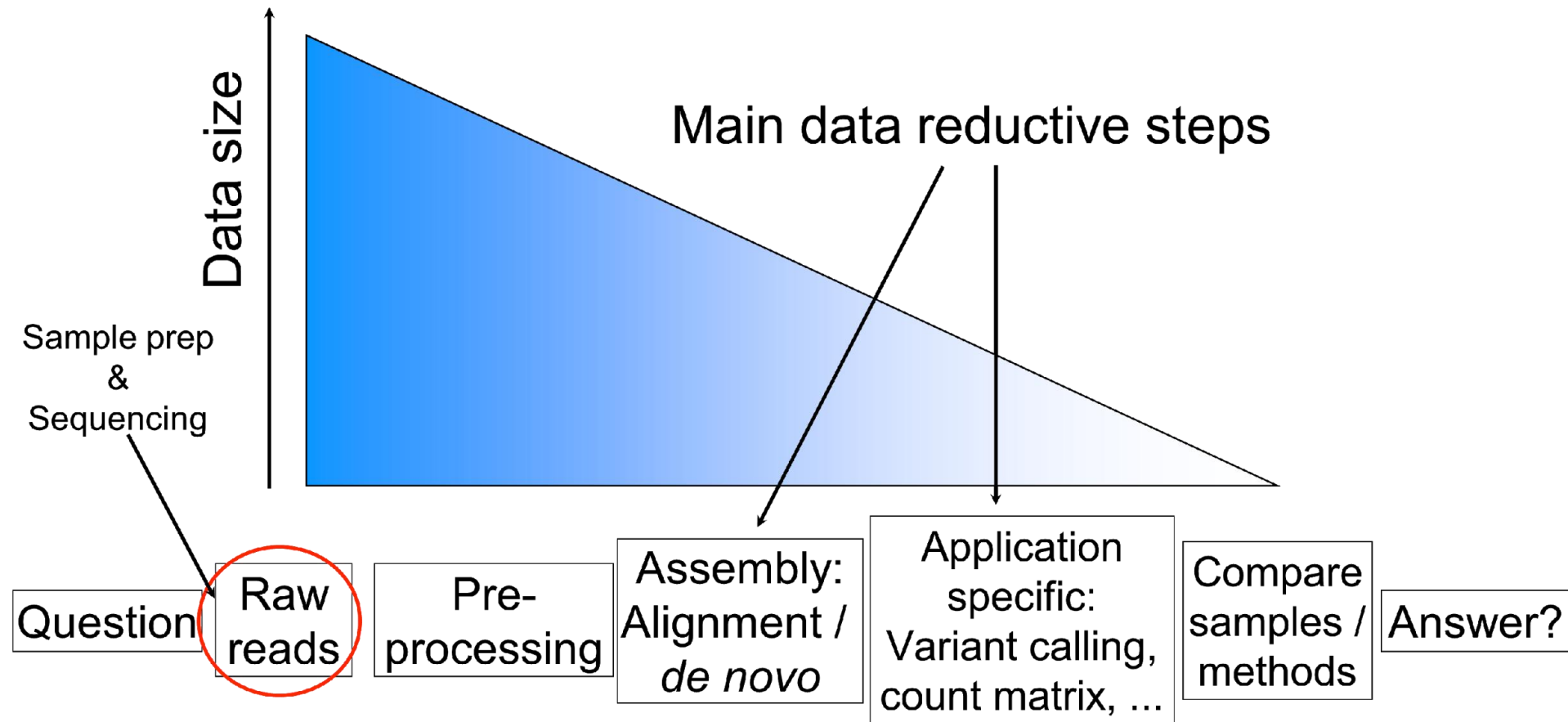
Fastq format

- Quality scores
- Multiplexing/Demultiplexing
- Sequencing read types

Single end

Paired end

# Generalized NGS analysis



# How would you store sequencing data?

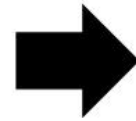
- + Sequencing technologies



# How would you store sequencing data?

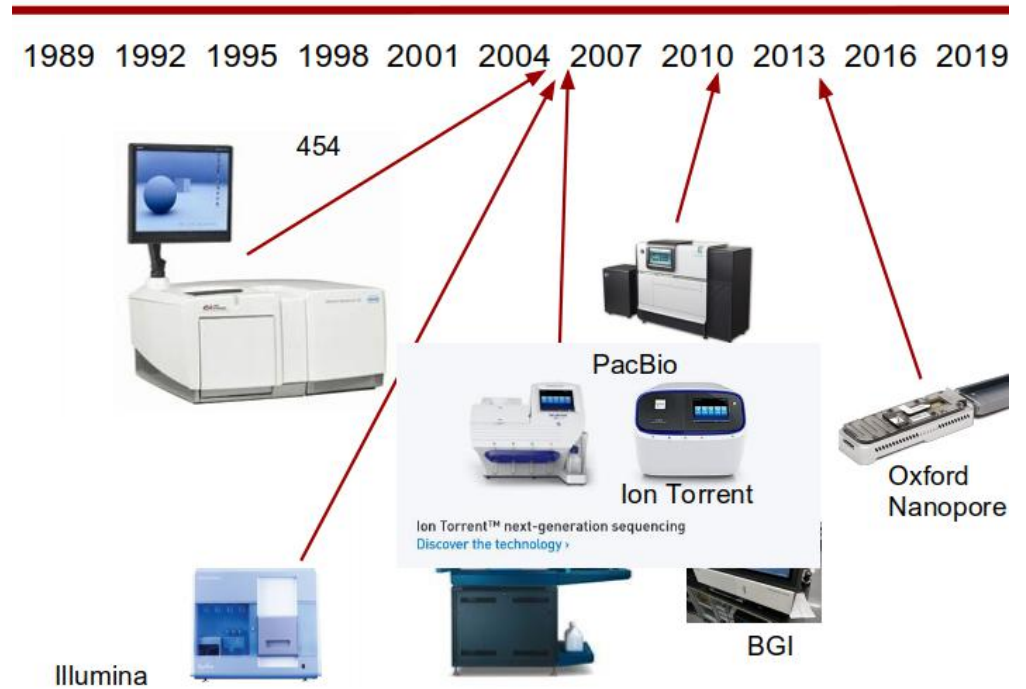
- + Sequencing technologies
- + Key concepts
  - Read length
  - Error types
  - Throughput

template



# How would you store sequencing data?

- + Sequencing technologies
- + Key concepts
  - Read length
  - Error types
  - Throughput
- + 3<sup>rd</sup> generation sequencing





# Storing sequence data – Fasta format

## Sequences are stored in fasta-files

Header

```
>gi|218693476|ref|NC_011748.1| Escherichia coli 55989 chromosome, complete genome
```

Sequence →

```
GTAAGTATTTTTCAGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAGAGT
GTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTTAAATTTTATTGACTTAGGTCACTAA
ATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACG
CATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAA
ACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCAT
GCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTG
GAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGG
TGGCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTT
TGCCGAACTTTTGACGGGACTCGCCGCCGCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTC
GATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCA
```

**E.coli ~ 4.5 - 6 Mbases**

**Human ~ 3.2 Gbases**

# Storing output from sequencing machines

## Fastq

Header  
@ILLUMINA-C90280\_0030\_FC:5:1:2675:1090#NNNNNN/1

Sequence → ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC  
+

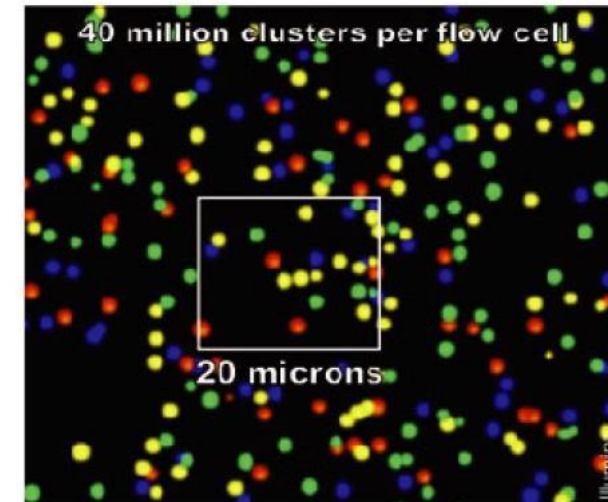
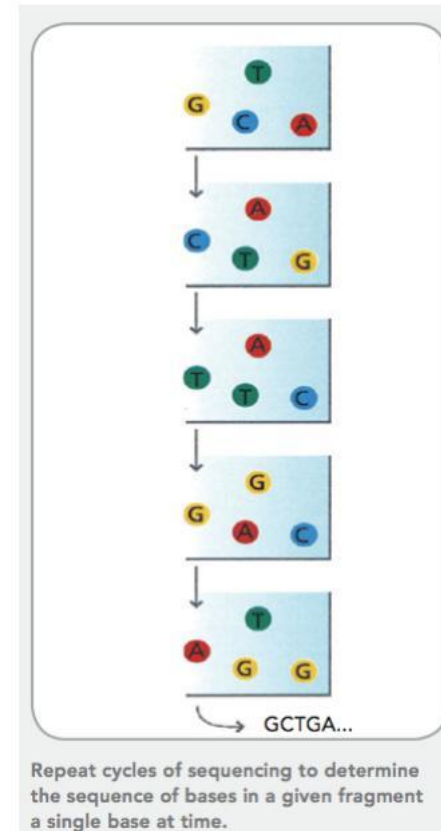
Qualities ↗ BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?  
(probability that base call is wrong)

**Millions to billions of these  
reads per experiment**

# Quality score encoding

- Quality score is the combination of these two (Illumina):
- Quality predictor values of clusters:  
Intensity profiles, phasing and SNR
- Quality model/table:
  - Pre-calculated combinations of the above
  - Depend on machine, chemistry, software

Illumina



# A closer look at the qualities

Header

```
@ILLUMINA-C90280_0030_FC:5:1:2675:1090#NNNNNN/1
```

Sequence

```
→ ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC
```

+

Qualities

```
↗ BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?
```

(prob. that base call is wrong)

One character encodes a number using  
ascii table (0-255)

This number ( $Q$ ) can be converted to  $P$

Phred-scale

$$Q = -10 * \log_{10} P$$

$$P = 10^{(-Q/10)}$$

# Phred scale

@ILLUMINA-C90280\_0030\_FC:5:1:2675:1090#NNNNNN/1

ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC

+

BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?

Q ~ Prob

10 ~ 0.1

20 ~ 0.01

30 ~ 0.001

40 ~ 0.0001

ASCII\_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41 )	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII\_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93 ]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

# Phred-scaled probabilities

- Base qualities, read mapping qualities, variant qualities, ...
- Straight-forward, except for when they are used in reads!
  - Offset: Sanger = 33 (“Phred+33”), Illumina = 64 (“Phred+64”)

```
@ILLUMINA-C90280_0030_FC:5:1:2675:1090#NNNNNN/1
ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC
+
BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?
```

Q ~ Prob

10 ~ 0.1

20 ~ 0.01

30 ~ 0.001

40 ~ 0.0001

Sanger: 33 32 32 ~0.001

Illumina: 2 1 1 ~1

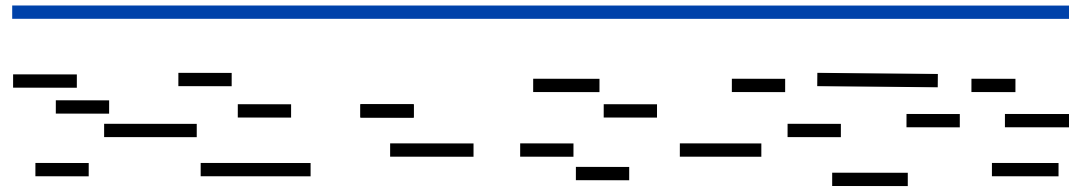
**HUGE difference!**

# Sanger vs. Illumina vs. Solexa

- 454, Ion Torrent, Pac Bio, Nanopore, Sanger: “Sanger” encoding
- Illumina reads: “Illumina” or “Sanger” encoding. New reads are all “Sanger”
- Solexa data: Solexa encoding (bought by Illumina)
- All data from SRA/ENA: “Sanger”

# Read types

Fragment DNA:

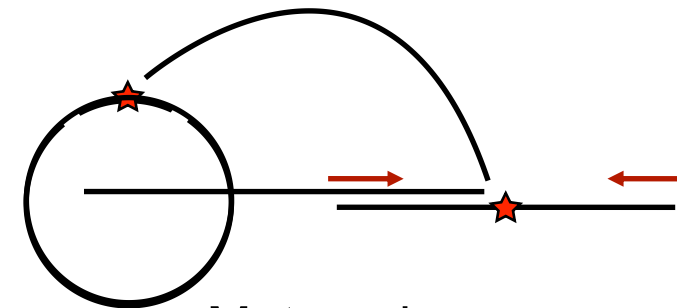


Single end



Paired end

Ins: 200-800 bp



Mate pair

Ins: 2kb - 40kb (~5kb)


Protocol/technology dependent



# Example sequencing read

DNA Fragment 32 bp: AGGTGTGAGCTCTACCCTAGCCCAGTTGGACC

Single end (R1) 10 bp:  AGGTGTGACC

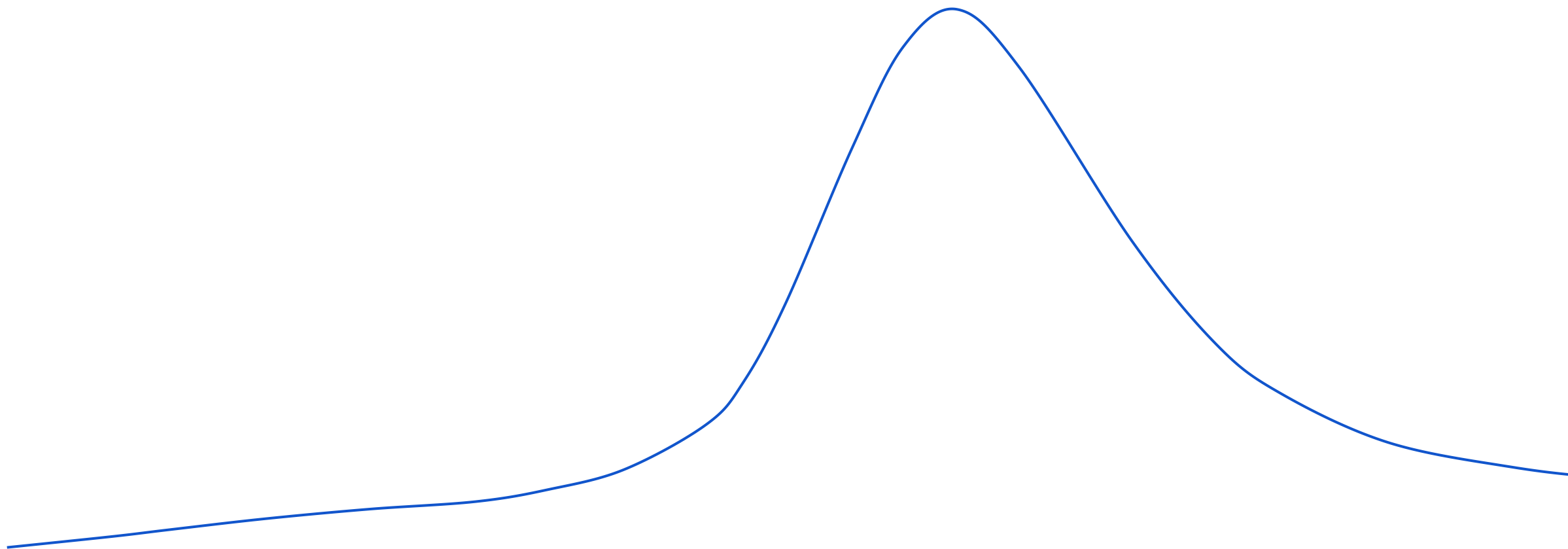
Paired end R1 10 bp:  AGGTGTGAGC  
R2 10 bp:

  
GTCAACCTGG

What would happen if the reads were 30 bp long?  
What are the effects of variation in the insert sizes?

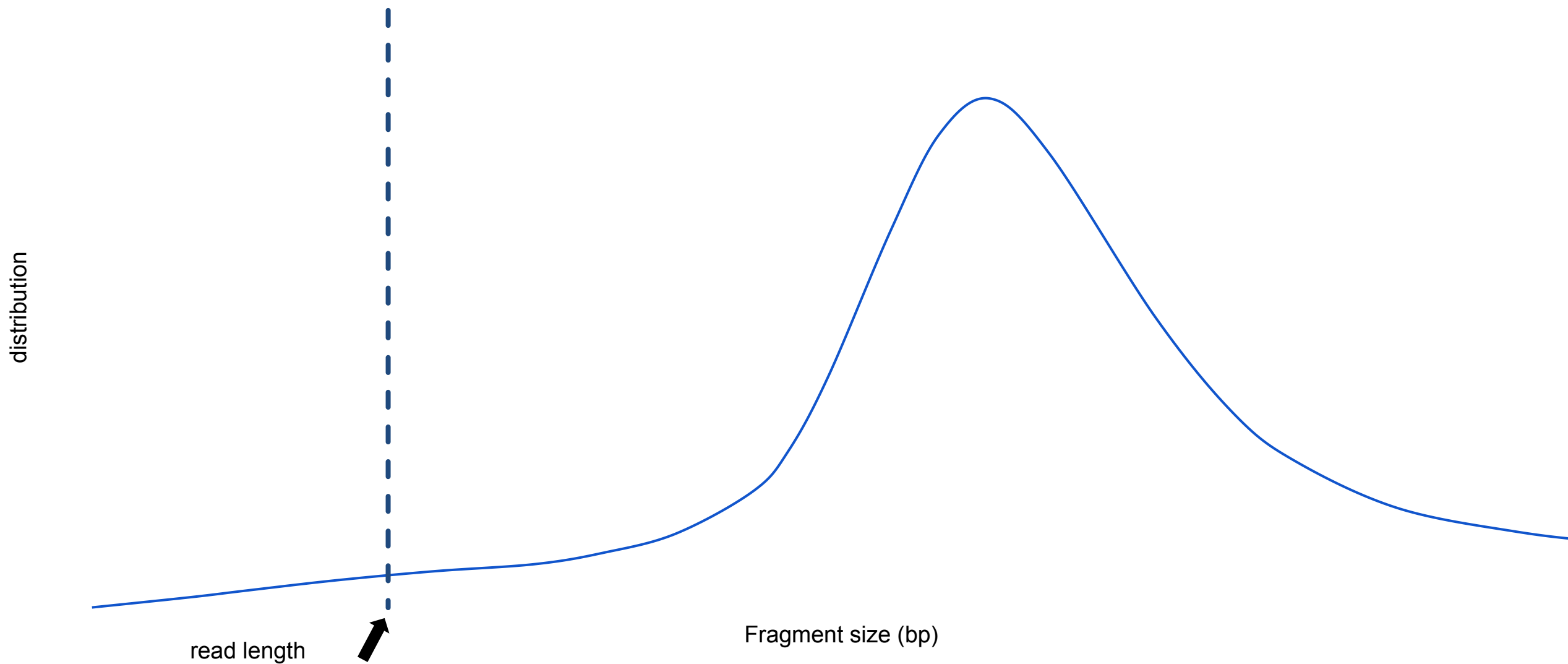
# Insert size distribution

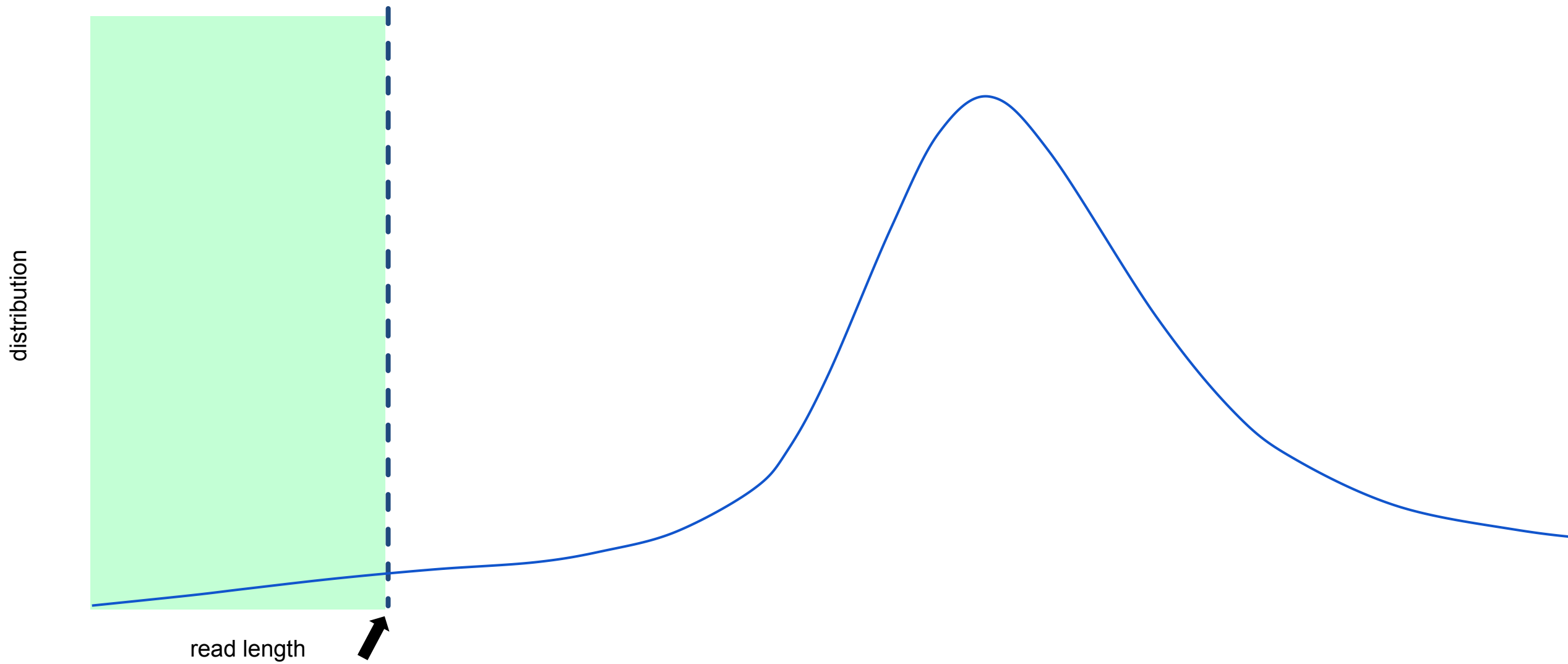
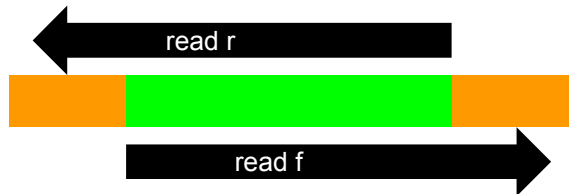
distribution

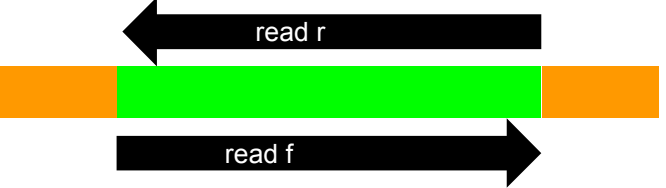


Fragment size (bp)

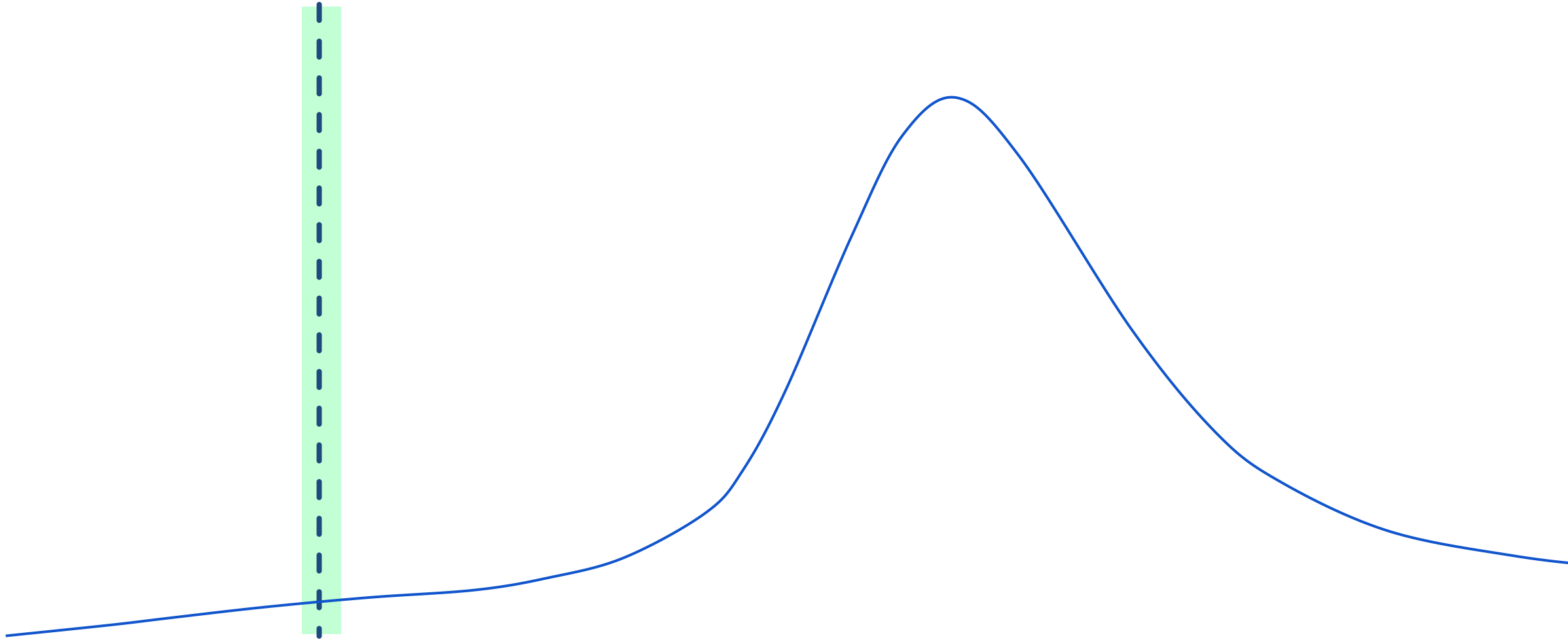
# Insert size distribution





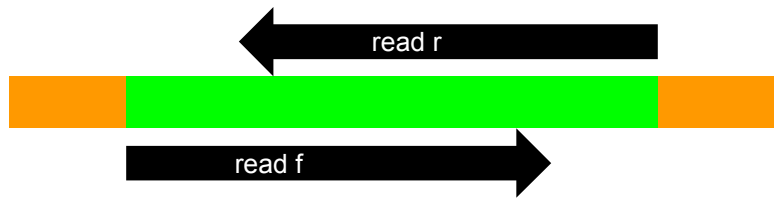


distribution

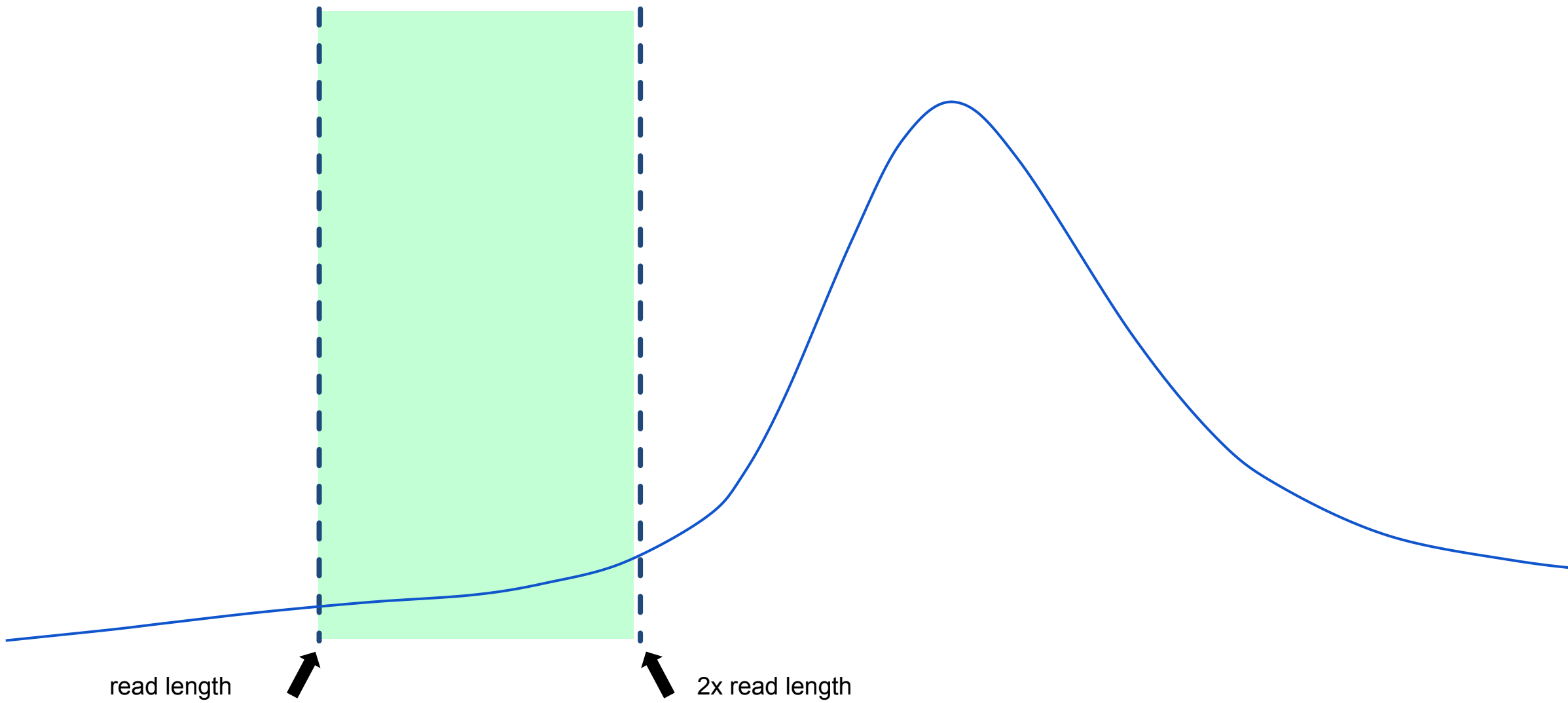


read length

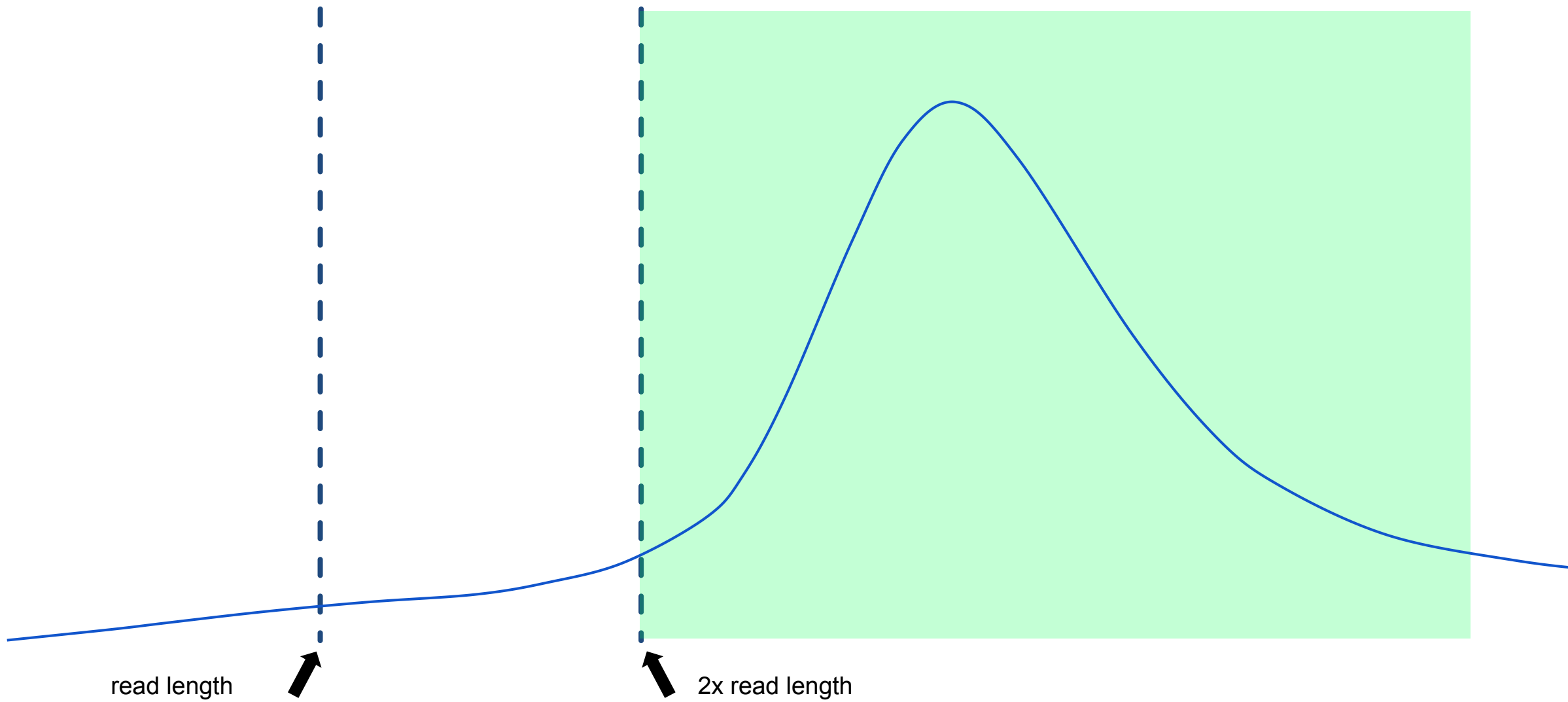




distribution



distribution



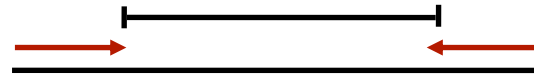
# Read orientation

Single end



Forward

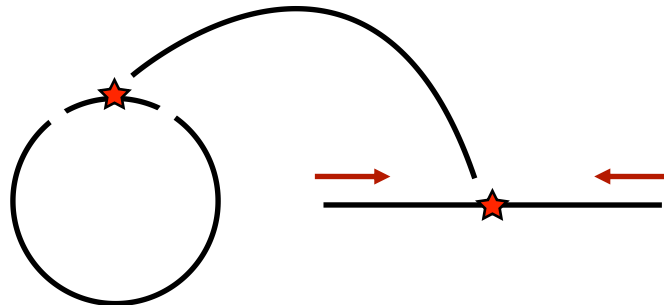
Paired end



Illumina: Forward - Reverse



Mate pair

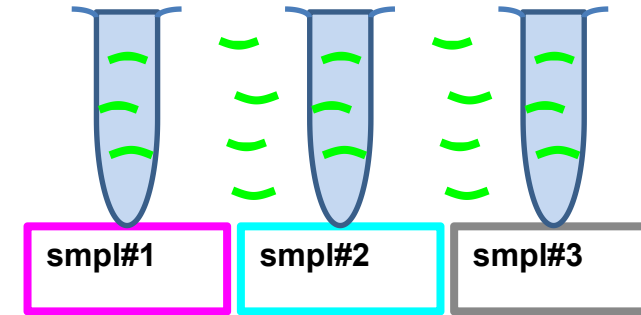
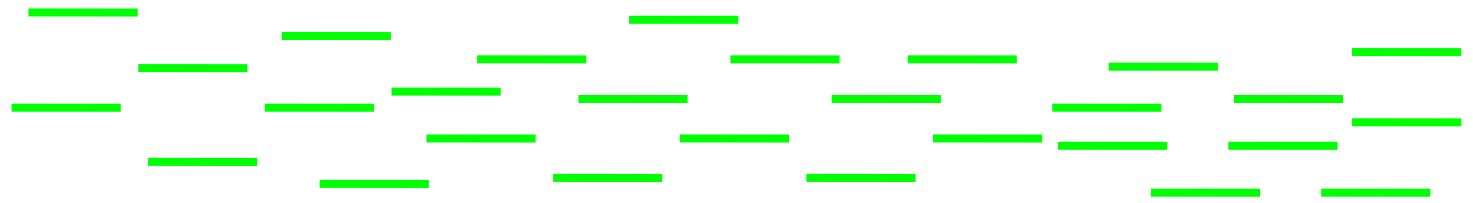


Illumina: Reverse - Forward



Different for other technologies!



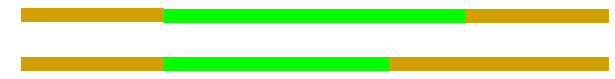
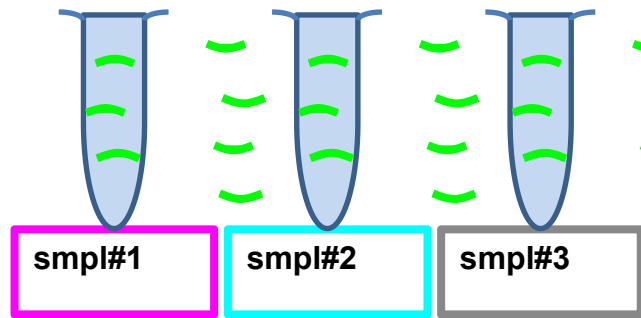
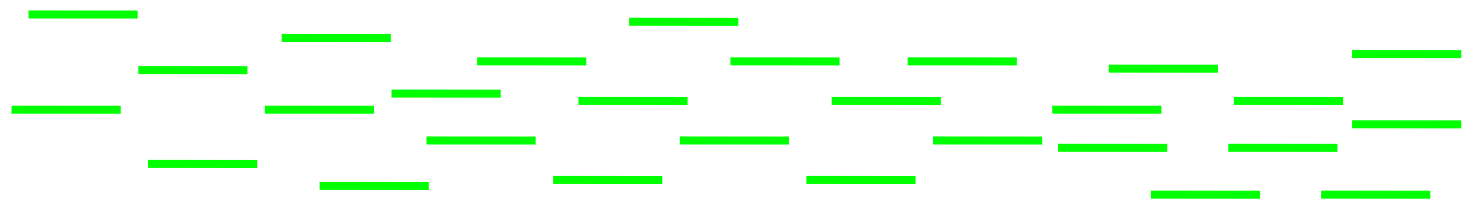
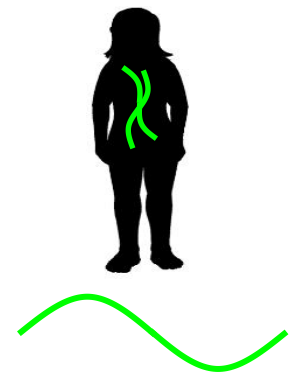


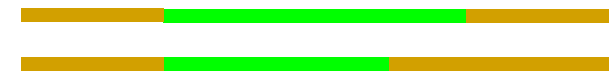
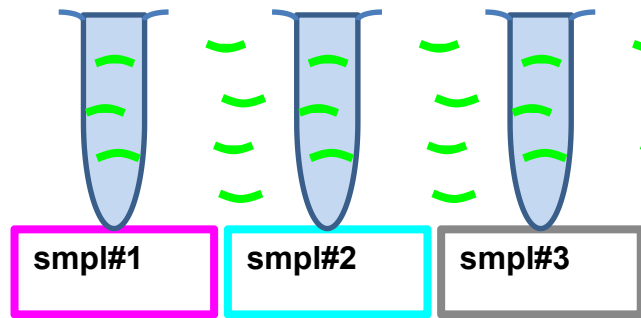
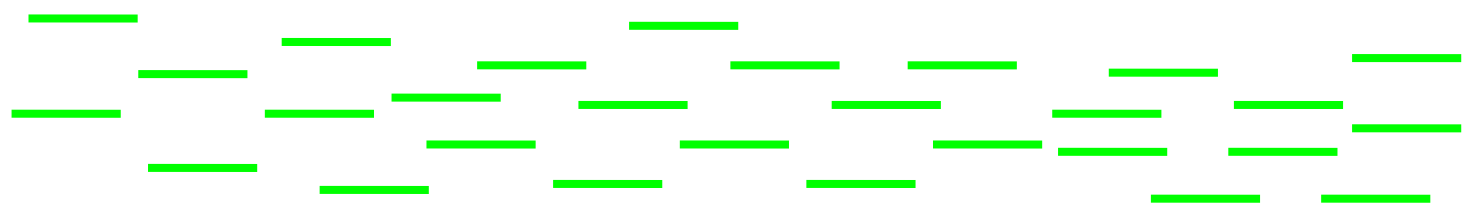
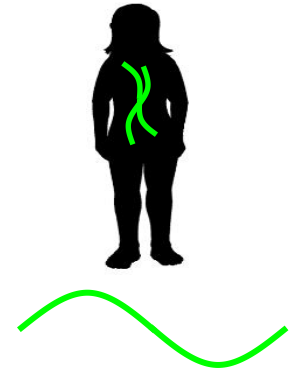
## Multiplexing

What is multiplexing?

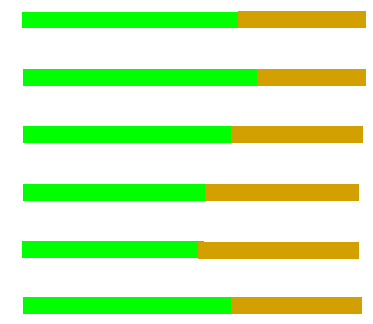
Why would you multiplex?

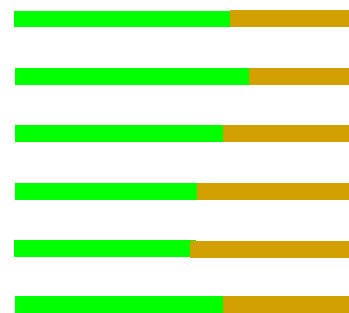
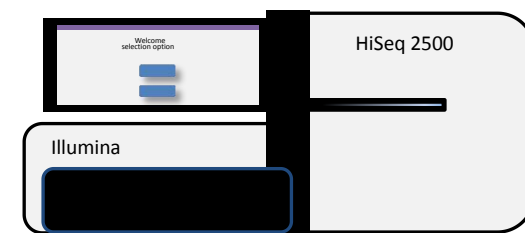
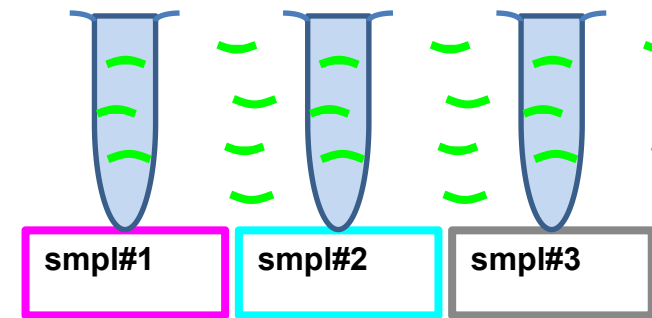
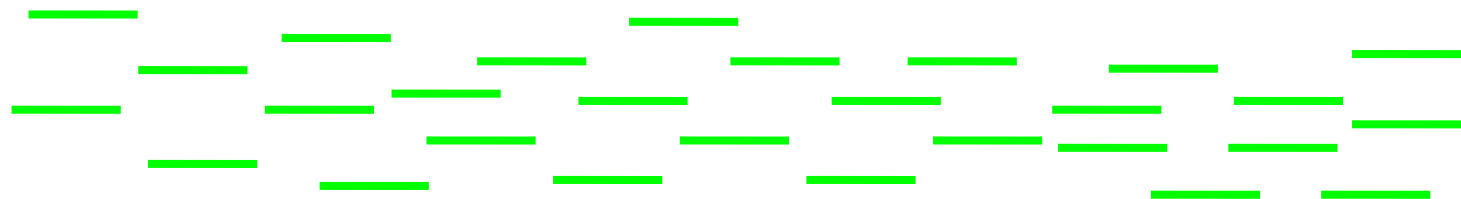
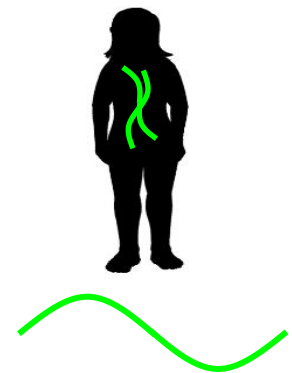
Advantages and challenges?



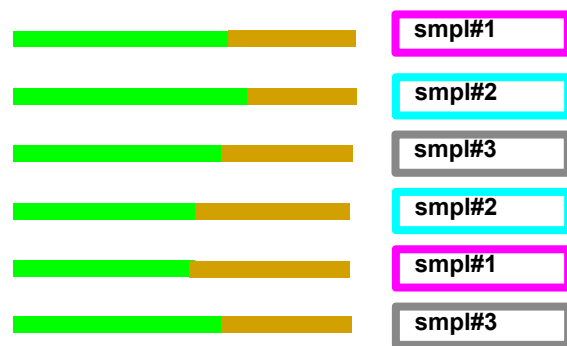


Basecalling





Demultiplexing



# Uses of multiplexing

- Environmental DNA studies
  - Lots of samples and replicates
  - Short target size
- Targeted sequencing
- Ancient DNA
  - Screening

# Exercise time!

[http://teaching.healthtech.dtu.dk/22126/index.php/Data\\_basics\\_exercise](http://teaching.healthtech.dtu.dk/22126/index.php/Data_basics_exercise)