資料３

次世代シークエンスについて

Sequencerの比較

	Provider	Sequencer	Maximum Read Length (bp)	Maximum sequence yield per run (Gb)
Short read sequencing	**Illumina**	HiSeq 4000	2 × 150	1,500
		HiSeq 3000	2 × 150	750
		HiSeq 2500	2 × 125	1,000
		NextSeq 500	2 × 150	120
		MiSeq	2 × 300	15
		MiniSeq	2 × 150	7.5
	ThermoFisher Scientific	Ion Proton	200	10
		Ion PGM	400	2
		Ion Torrent S5	600	10-15 (though there is a trade-off between read length and output)
Long read sequencing	Pacific Biosciences (PacBio)	PacBio RSII	Half of data in reads >20,000 Max length >60,000	1
		PacBio Sequel	Half of data in reads >20,000 Max length >60,000	7
	Oxford Nanopore (ONT)	MinION	Read length = DNA Fragment length Longest reported approaching 1 x 10^6	10–20
		GridION X5	Read length = DNA Fragment length Longest reported approaching 1 x 10^6	50–100
		PromethION	Read length = DNA Fragment length Longest reported approaching 1 x 10^6	11,000 (theoretical maximum)

Short read sequencing

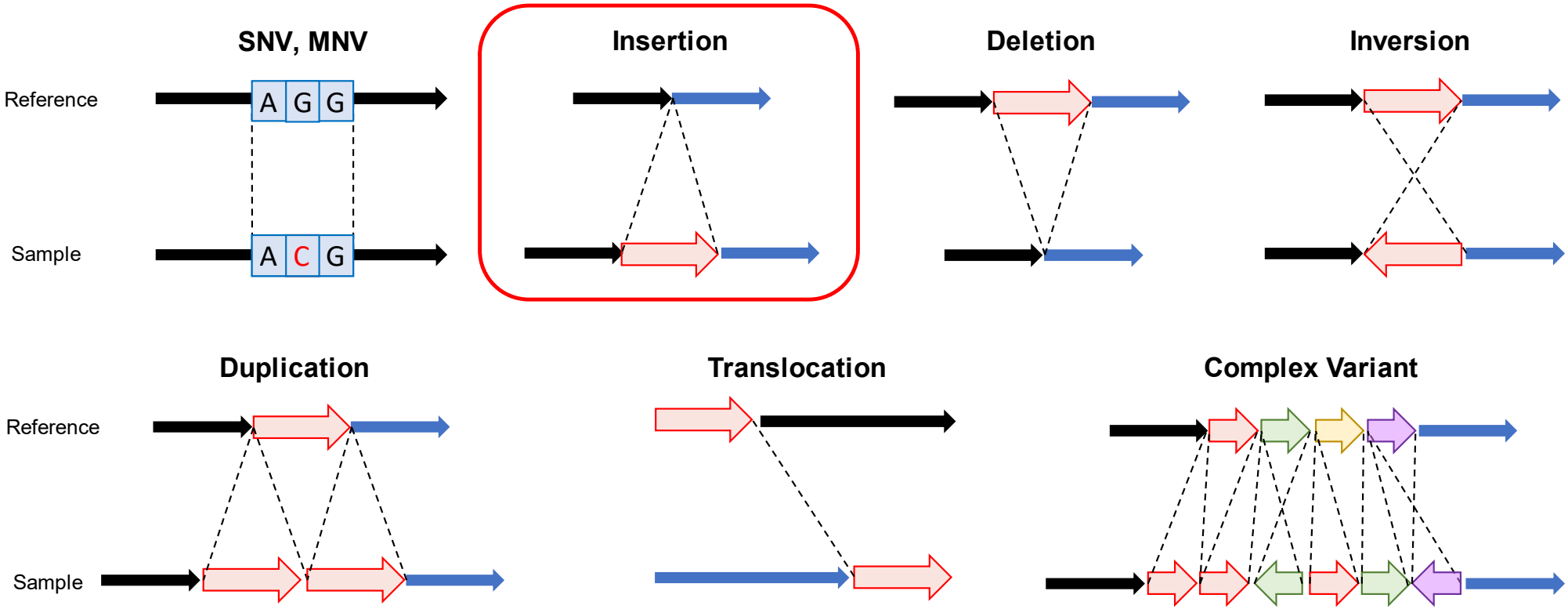
- 正確性が高い
- リピート配列が読めない

Long read sequencing

- 比較的リピート配列が読める
- 正確性がやや低い

Food Control (2019) 101, 134–143

起こりうる主な遺伝子変異



mapping統計情報の例

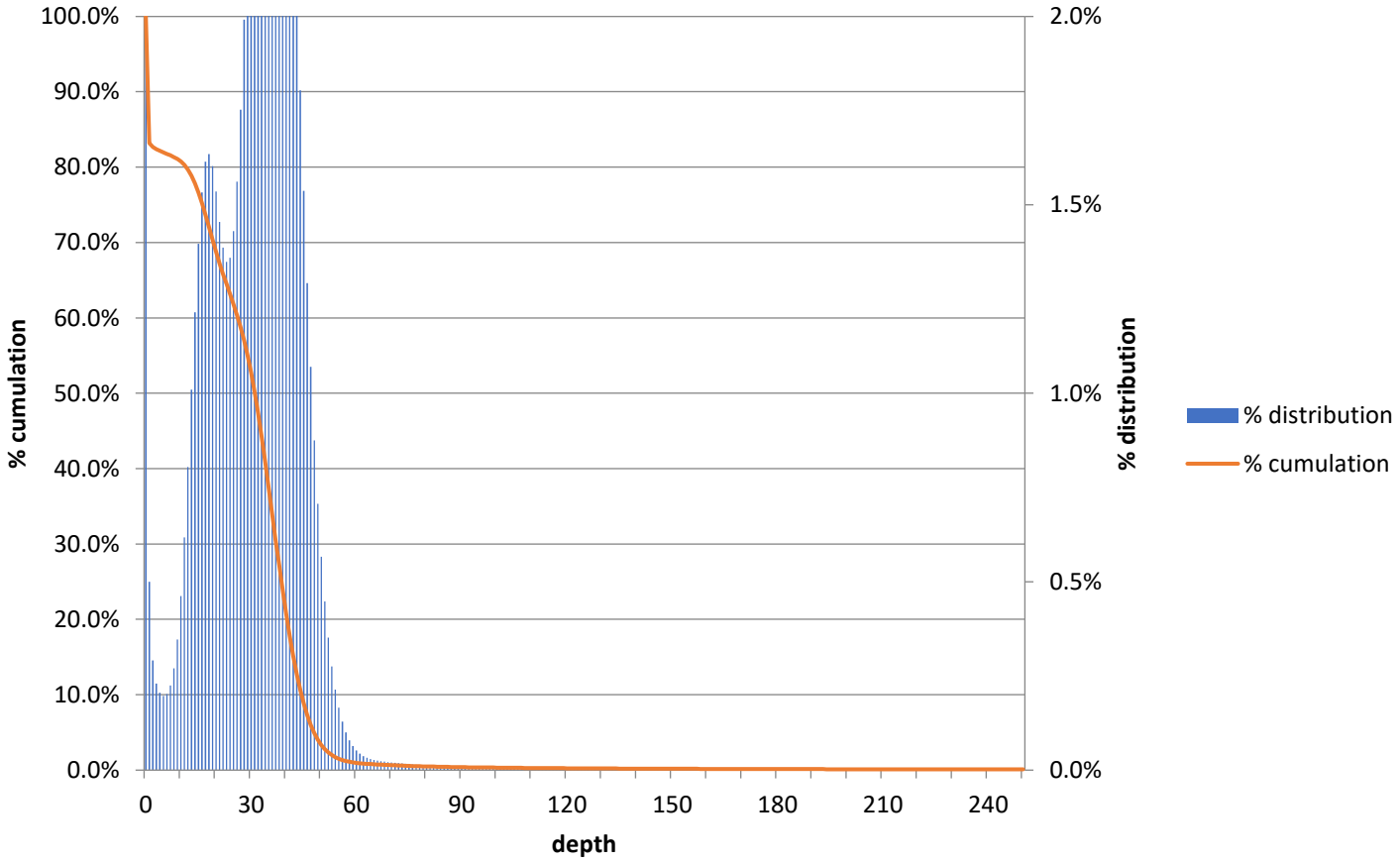
Map summary	PG3676_01_a	PG3676_03_a
Total input reads	795,743,304	714,037,020
Total input reads(%)	100	100
Number of duplicate marked reads	124,654,147	104,371,473
Number of duplicate marked reads(%)	16	15
Number of unique reads (excl. duplicate marked reads)	671,089,157	609,665,547
Number of unique reads (excl. duplicate marked reads)(%)	84	85
Reads with mate sequenced	795,743,304	714,037,020
Reads with mate sequenced(%)	100	100
Reads without mate sequenced	0	0
Reads without mate sequenced(%)	0	0
QC-failed reads	0	0
QC-failed reads(%)	0	0
Mapped reads	760,352,741	683,134,856
Mapped reads(%)	96	96
Number of unique & mapped reads (excl. duplicate marked reads)	635,698,594	578,763,383
Number of unique & mapped reads (excl. duplicate marked reads)(%)	80	81
Unmapped reads	35,390,563	30,902,164
Unmapped reads(%)	4	4

ave coverage=40x

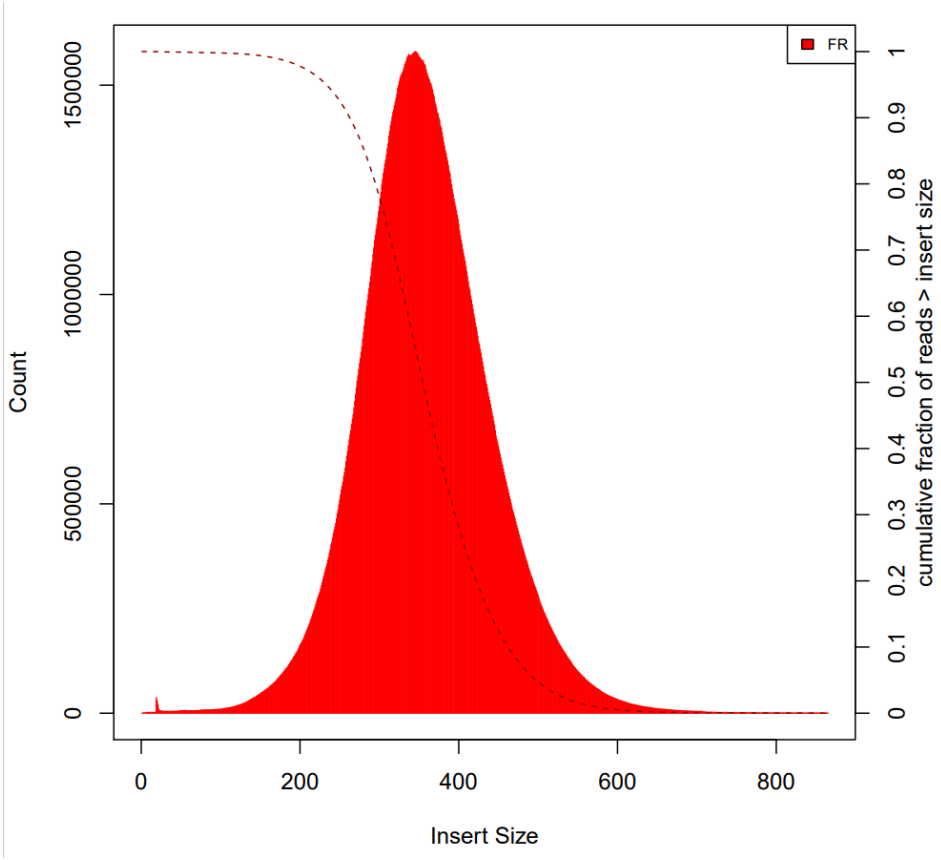
pair-end 150 bp

ave coverage=32x

Coverage分布


100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
% cumulation
2.0%
1.5%
1.0%
0.5%
0.0%
% distribution
0
30
60
90
120
150
180
210
240
depth
% distribution
% cumulation

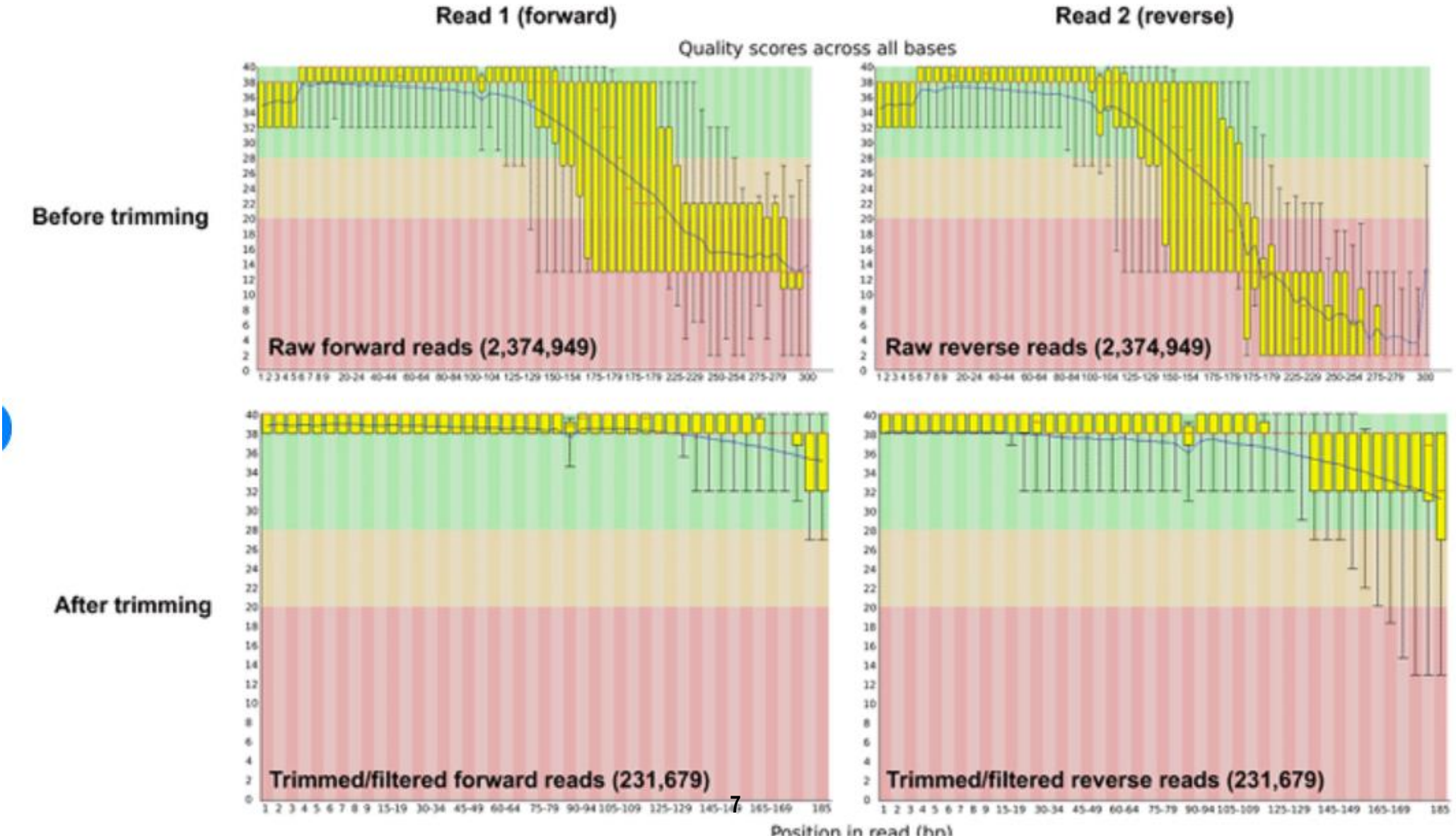
library size分布



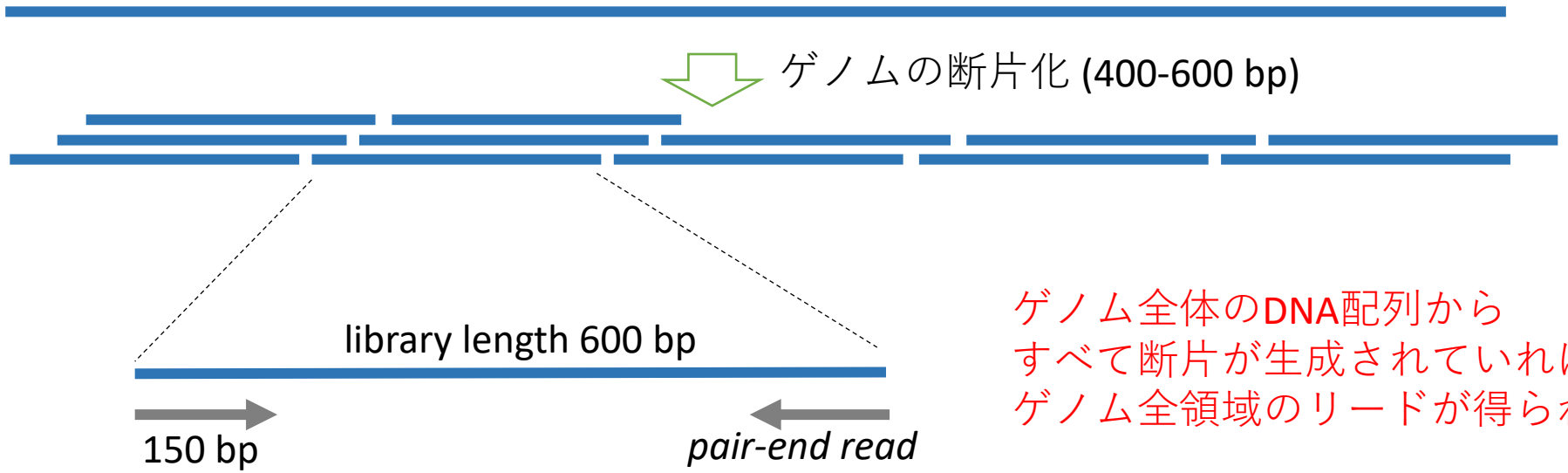
400 bp

150 bp

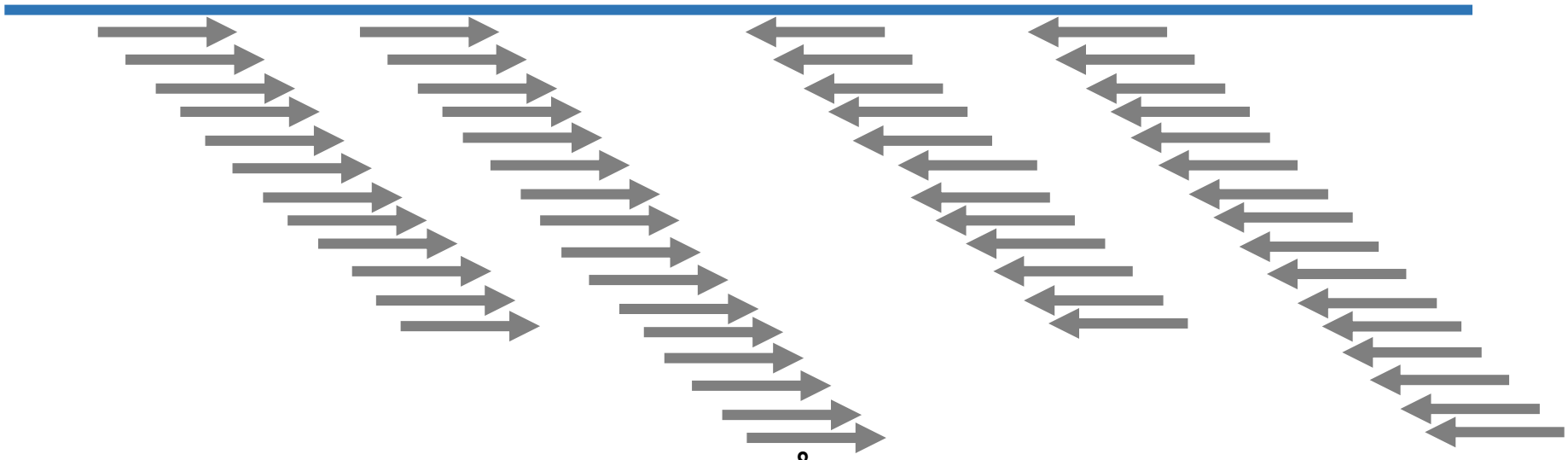
read quality分布



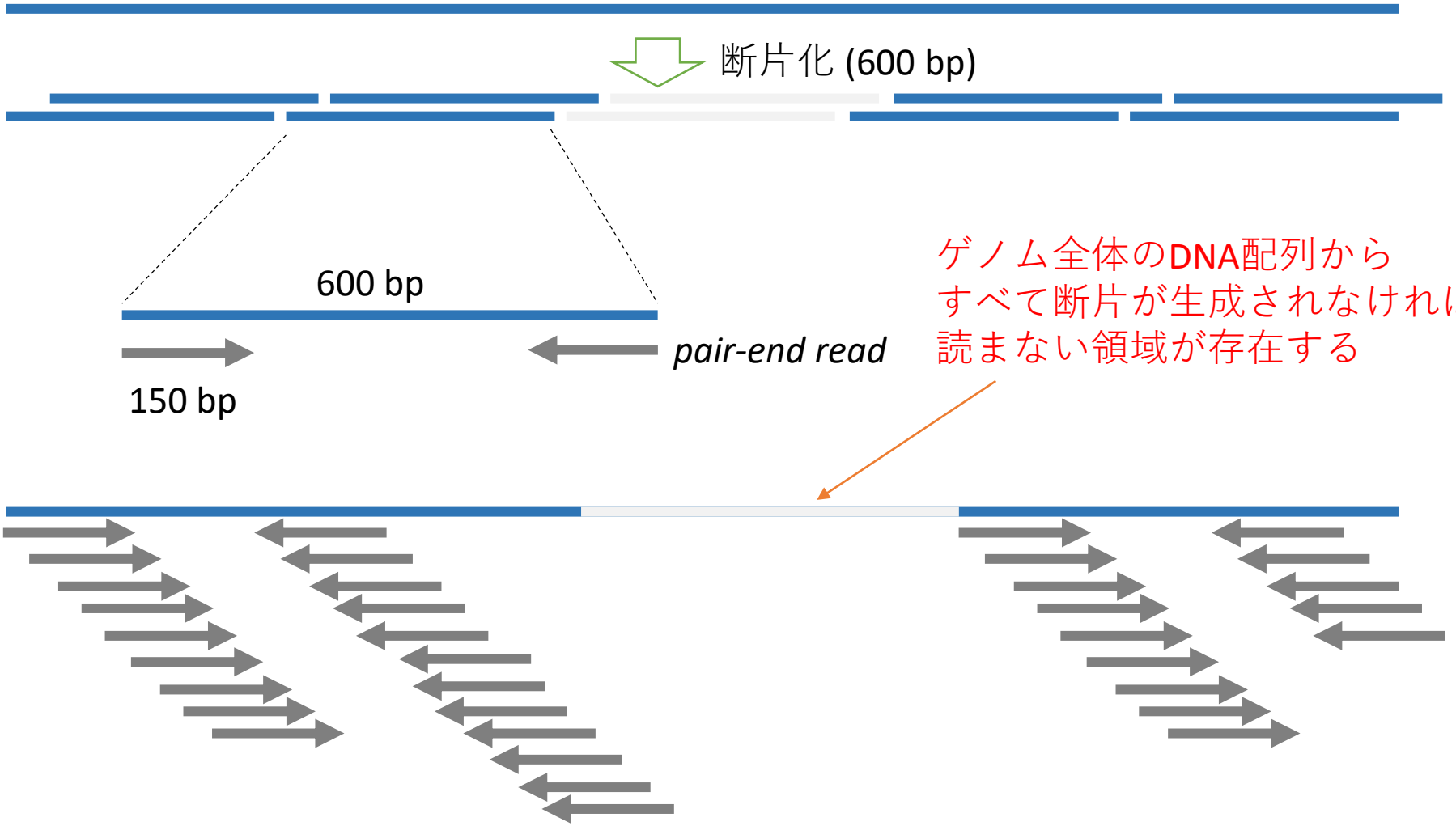
Illumina short readはどのように読まれるか

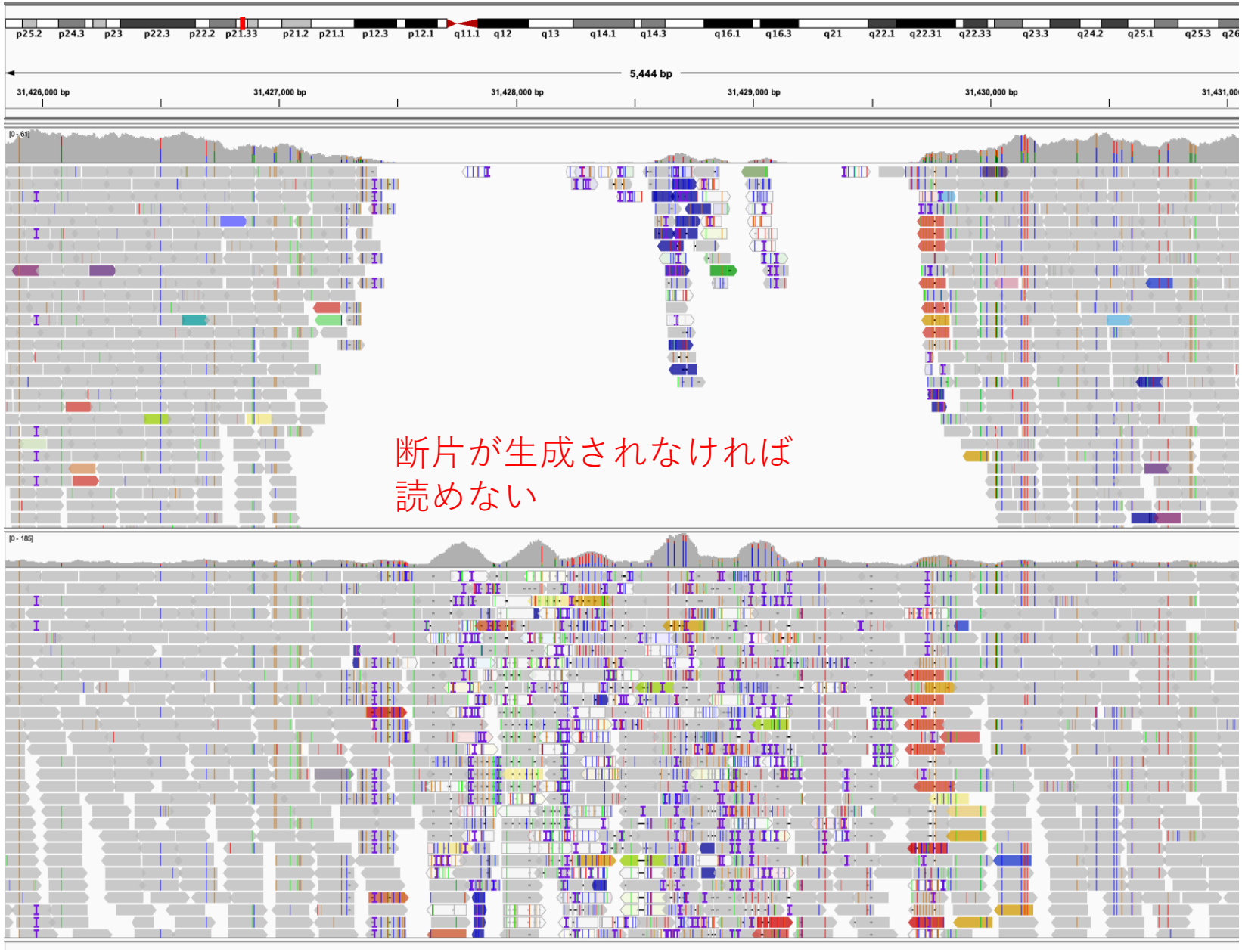


ゲノム全体のDNA配列から
すべて断片が生成されていれば
ゲノム全領域のリードが得られる



libraryができなかった領域は読めない

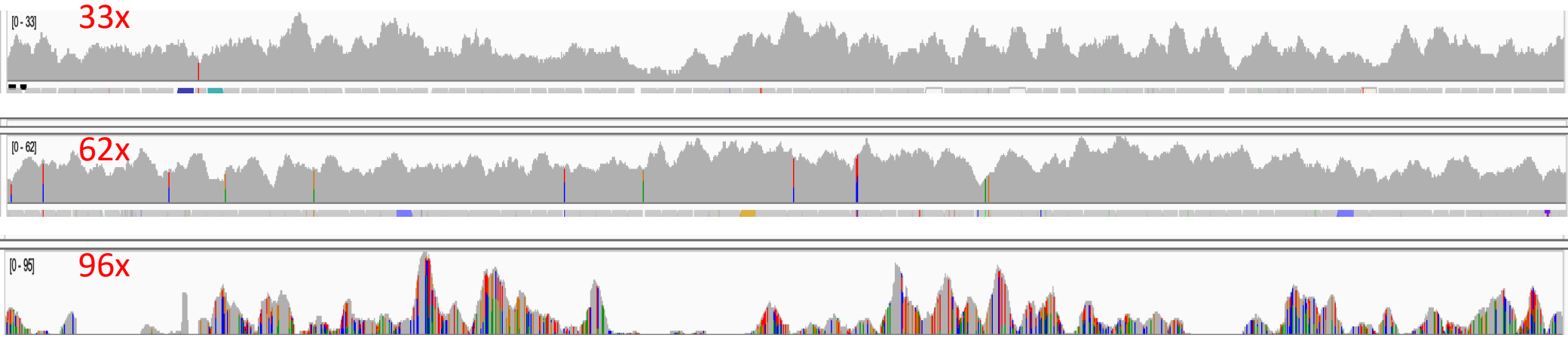



p25.2 p24.3 p23 p22.3 p22.2 p21.33 p21.2 p21.1 p12.3 p12.1 q11.1 q12 q13 q14.1 q14.3 q16.1 q16.3 q21 q22.1 q22.31 q22.33 q23.3 q24.2 q25.1 q25.3
5,444 bp
31,426,000 bp
31,427,000 bp
31,428,000 bp
31,429,000 bp
31,430,000 bp
[0 - 61]
断片が生成されなければ
読めない
[0 - 185]

read coverageは全ゲノム領域で均一ではない

average read coverage （平均冗長度）とは、
全リード塩基数をゲノムサイズ（塩基数）で割っただけの値

average read coverage = 50 としても、各領域でcoverageは大きく違う

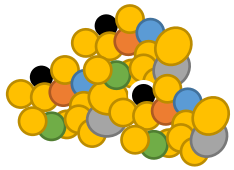


大事なのは解析したい領域で十分なcoverageを確保できているか

Q:　coverageはどれくらい必要か？

A: どういうサンプルで何を見たいかで変わる！

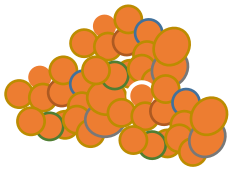
例えば、がんゲノム解析では100 x または300 x で全ゲノム解析
組織から取ったがん細胞は、それぞれ違う変異を持っている



100個細胞があり、その中の1つの細胞がある遺伝子に変異があったときに
その変異の検出確率は100分の1のため最低100リード必要.
通常、その変異をサポートするリードが複数必要なので、3個とすると
300 x で解析していないと検出できない

Q:　coverageはどれくらい必要か？

A: どういうサンプルで何を見たいかで変わる！



逆に、homogenousな細胞集団ですべての細胞が同じ変異を持っているのであれば10 x もあれば十分とも言える
ゲノムのある領域ではその半分以下のcoverageであるため、実際にはその数倍のcoverageが必要（30-40xあれば）

2倍体でヘテロ体であれば、その倍のcoverageで十分

そこに、各ゲノム領域でcoverageは変化するので、＋アルファする

ある特定の領域だけを解析する、または
全ゲノムは30xで解析して、特定領域（挿入領域など）をdeep解析したい場合はAmplicon seqで5000x-30,000xで解析できる

全ゲノム領域を厚く読むのは費用が急に上がる
最初から高深度でNGS解析する例は少ない

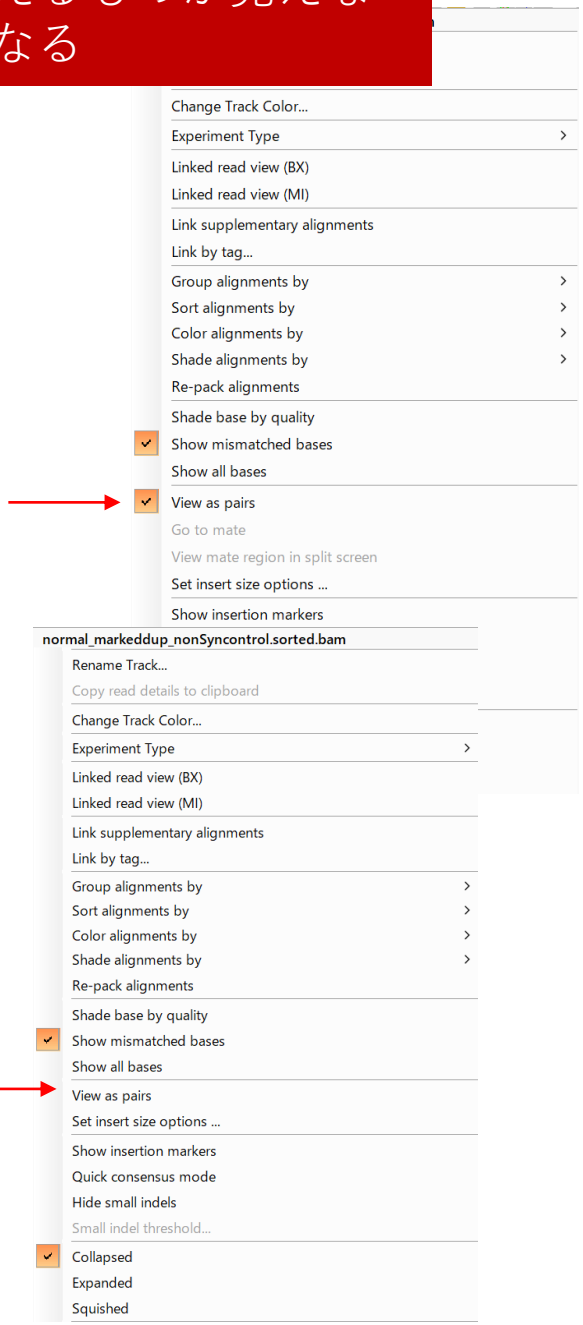
*実行可能性の点から：
現在の多くの受託NGS全ゲノム解析の価格はかなり安くなっているが、このときのデータ取得量はヒトゲノム換算で35-40xである

仮に75-80xでデータを取得するには数倍の費用と、解析リソースが必要になる

マッピング結果の表示はIGV (Integrative Genome Viewer) が標準

IGV 表示方法の違いで見えるものが見えなくなる

一つのread

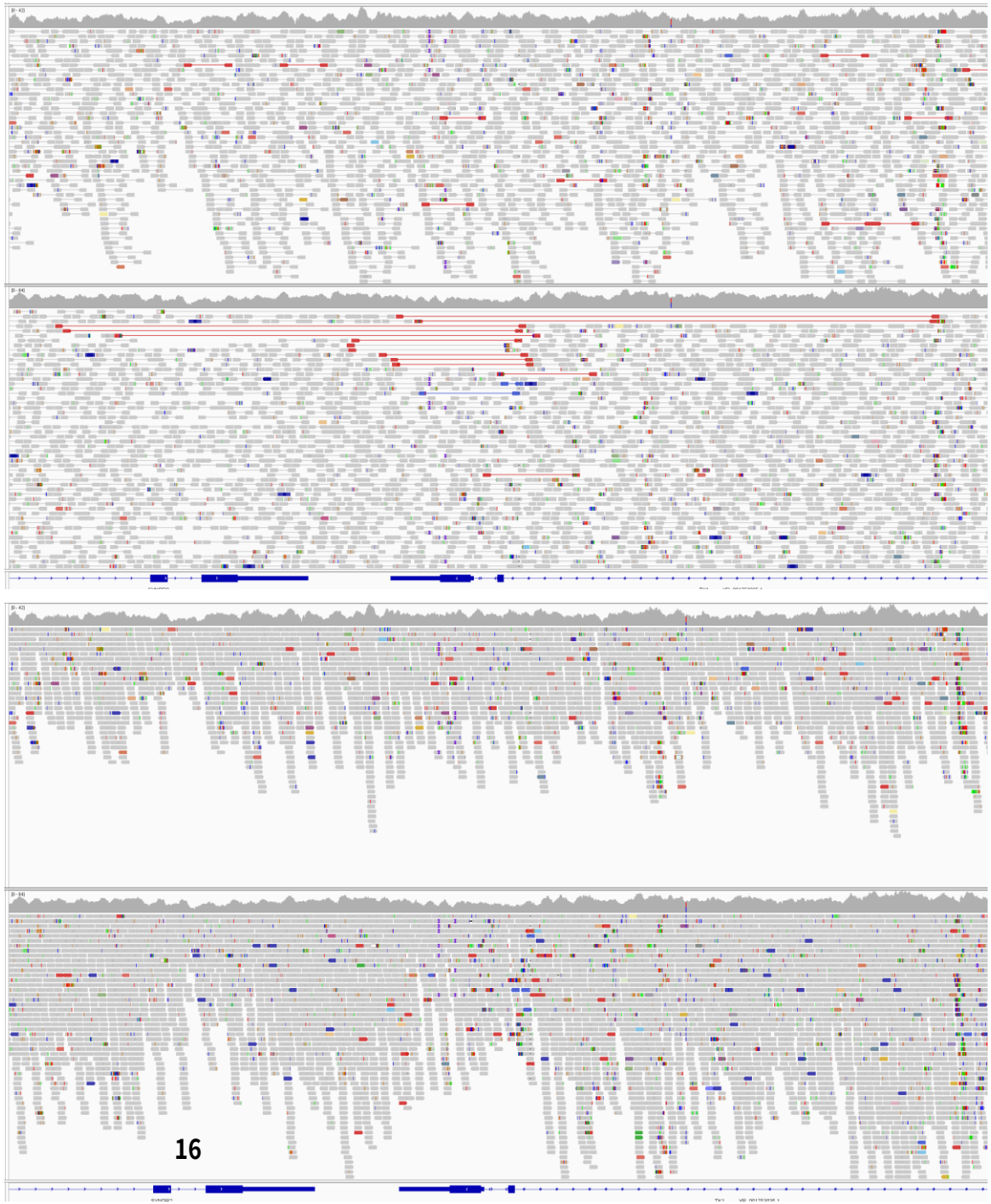


control

CRISPR

control

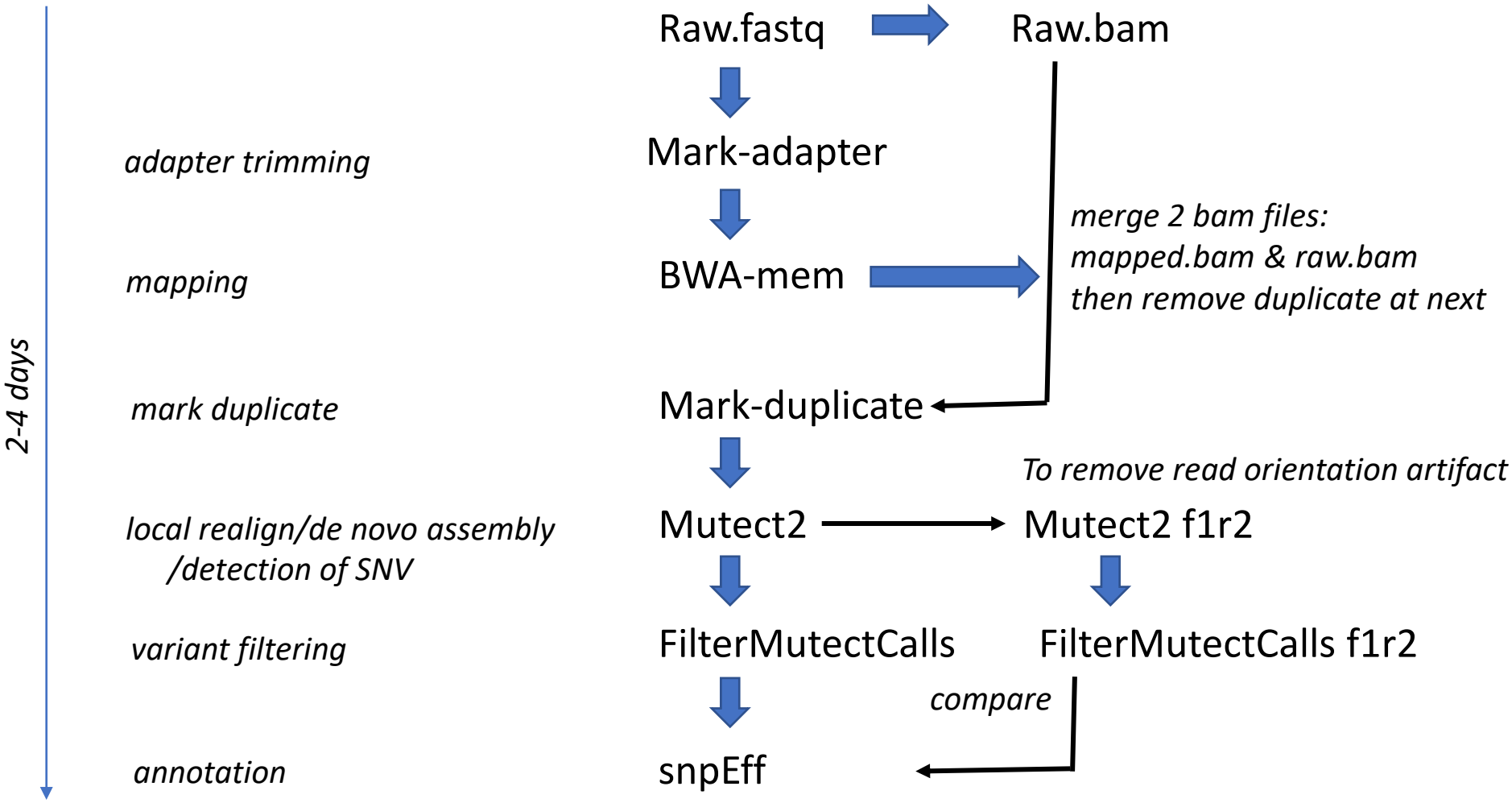
CRISPR



Nanopore sequencing

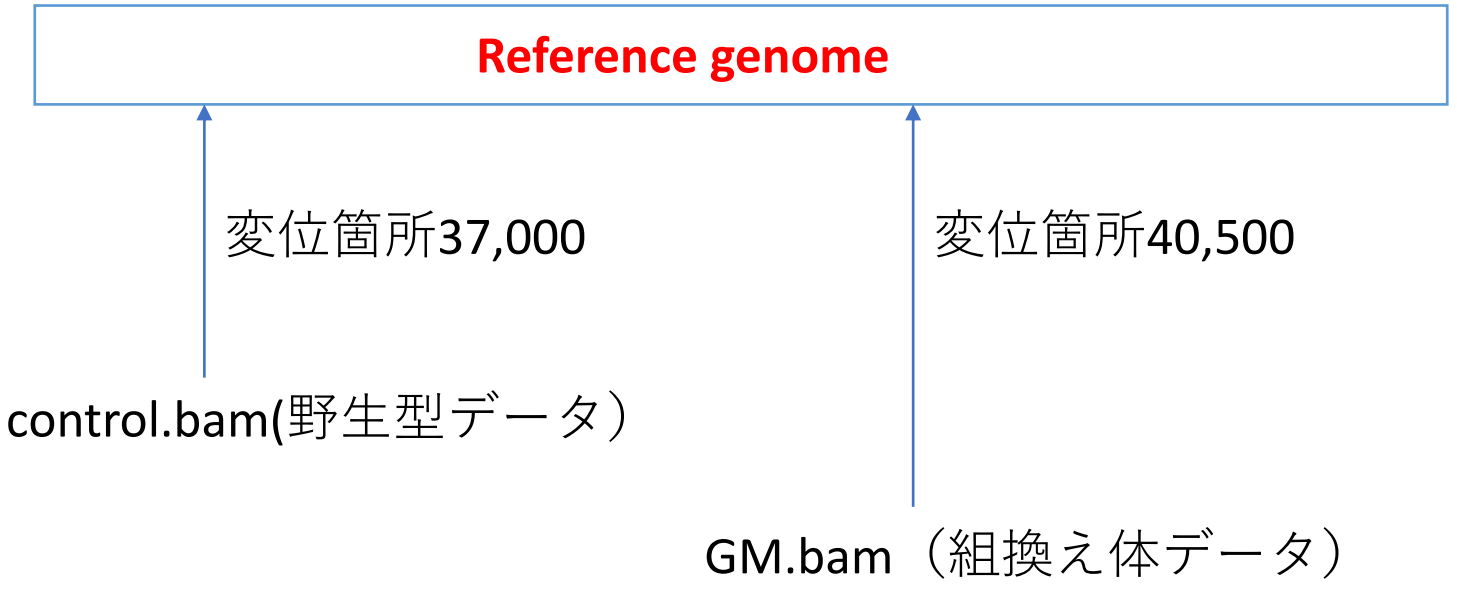
Long-read では数10kbを一気に読むことが可能

Workflow_1 (picard/gatk)



hardware: MacBook-pro 16inch 8-core 64 GB-RAM

Workflow_1 (picard/gatk)



controlになかった変異だけを検出すればいい
（精度の高いreference genomeでなくても問題ない
ある程度完成したdraft genomeがあれば精度よく
GMのみにある変異を高精度に検出できる）

Q: **NGS解析に求める要件は？**

１．read qualityの分布図
２．library調整法とlibrary (insert size)の分布とread length
　insert size=600 bp, pair-end 150bp
３．coverageの分布図
４．統計情報一覧
５．解析ソフトと使用したparameter
６．マッピングIGV図と表示設定
７．用いた参照ゲノム（versionなど）

－－－－－－－－－－－－－－－－－－－－－－－－

１０．1-6を求めても本当にその通り解析しているかは確認できない
１１．blindのポジコンデータセット（いろんな変位が入った）を
　作製して、これを同時に解析してもらうことでデータの解析
　精度を保証できるが。。。。