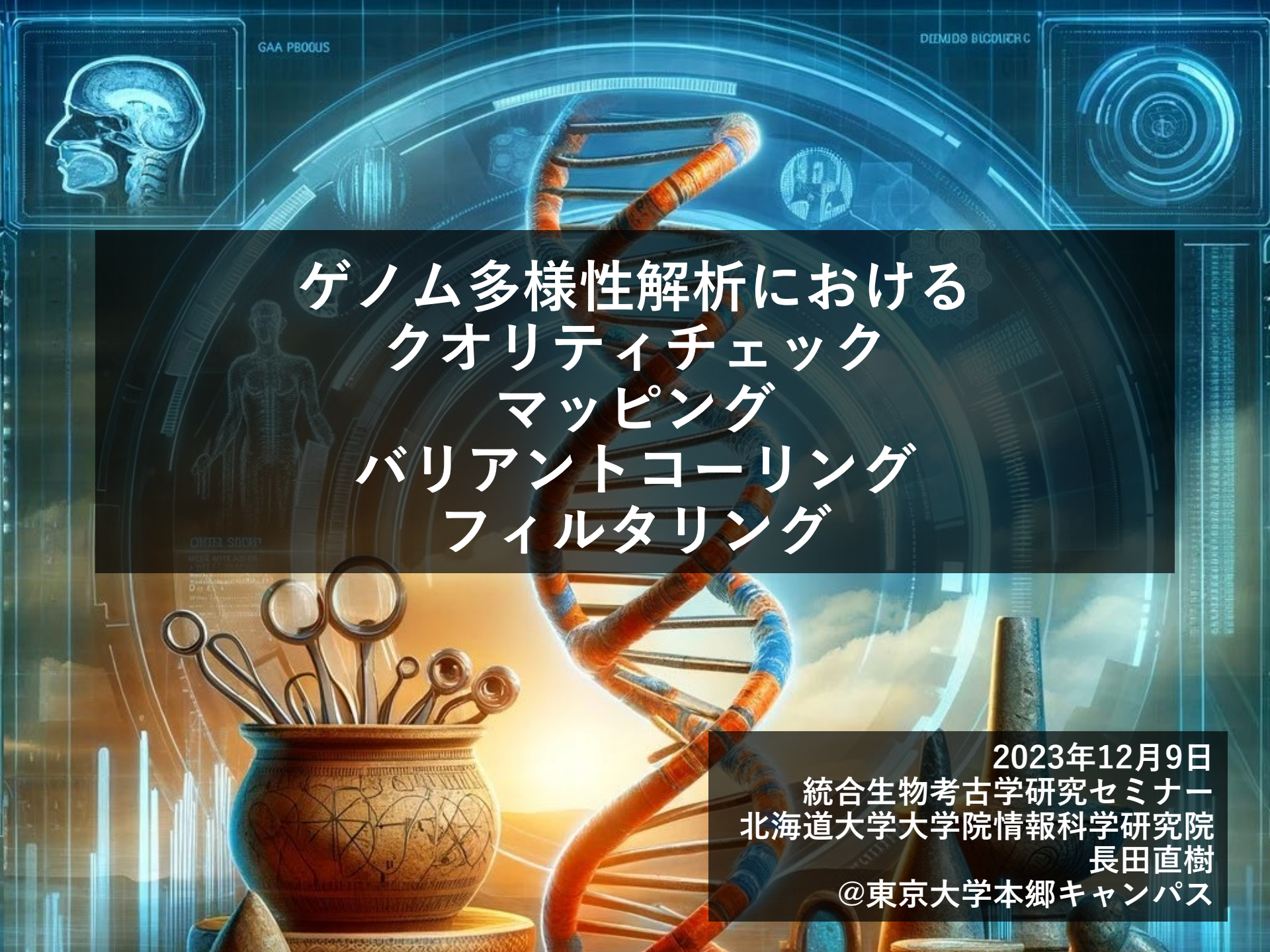


ゲノム多様性解析における
クオリティチェック
マッピング
バリアントコーリング
フィルタリング

2023年12月9日
統合生物考古学研究セミナー
北海道大学大学院情報科学研究院
長田直樹
@東京大学本郷キャンパス

講義内容

ゲノム多様性解析（またはpopulation genomics）解析に必要な基礎的な知識と技術を学ぶ

ゲノム多様性解析とは？

同種・近縁種に属する複数個体からゲノム情報を取得し，生物が過去に経た歴史を推定したり，どのような自然選択がゲノムにはたらいたのかを推定したりすることにより，生物進化のメカニズムや生物と環境とのかかわりを明らかにすることを目的とした研究

ゲノム配列解析のコストが下がったために，モデル生物だけではなく，幅広い生物に適用可能になった

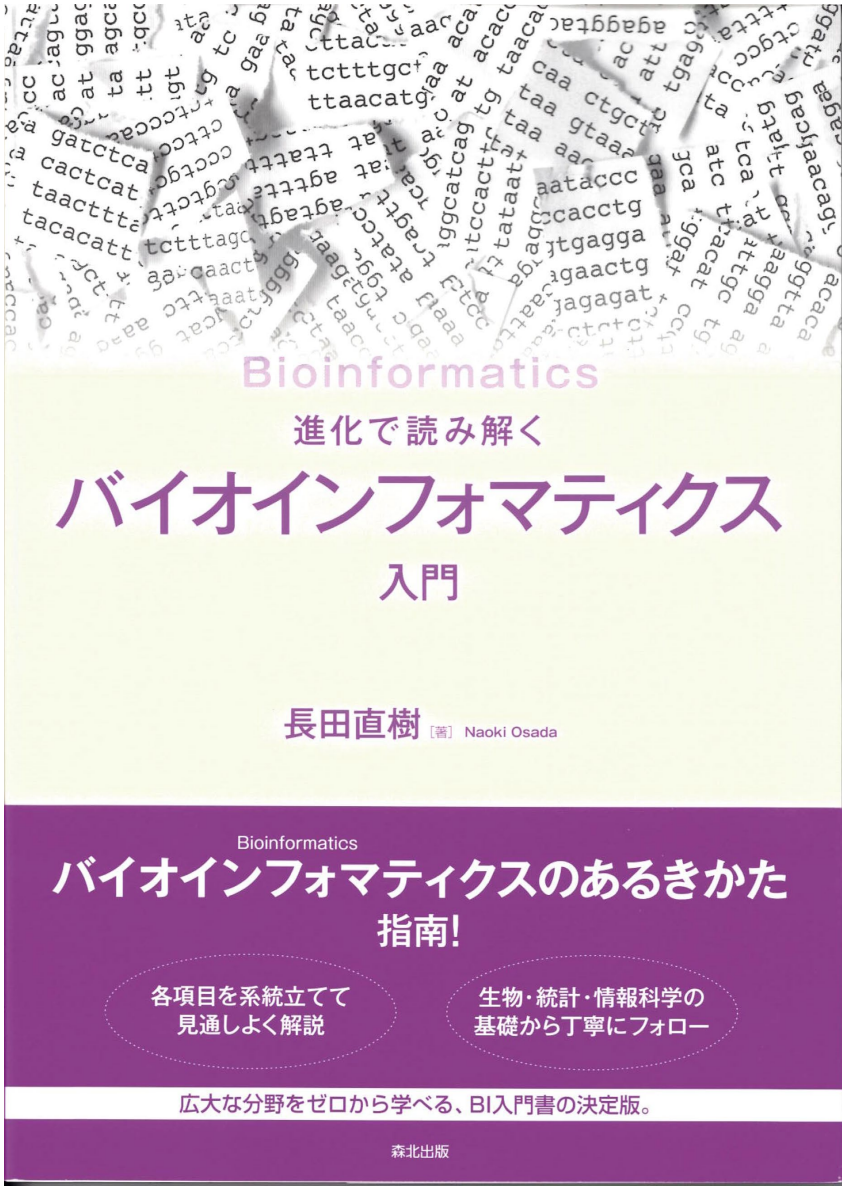
本講義で扱うもの

２倍体（以上）の生物の核ゲノム解析

本講義で扱わないもの

細菌，ウイルスのゲノム進化
ミトコンドリアゲノムの解析
遠く離れた生物同士の系統関係
体細胞ゲノム解析（がんゲノム解析）
遺伝子発現，機能解析

本日のトピック以外の参考書



森北出版
238ページ
¥3520

ゲノム多様性解析のワークフロー

大量配列解析実験（次世代シークエンサー）

↓

fastq ファイル（エラー率つきの塩基配列）

↓ 参照ゲノム配列へのマッピング（bwa）

sam/bam ファイル

↓ 変異検出（GATK，varScan）

vcf ファイル（変異データ）

↓ フォーマット変換ソフトウェア（PLINK）

カスタムフォーマットの変異ファイル

多検体変異ファイルへの統合

↓ eigenstrat, STRUCTURE, treemix など

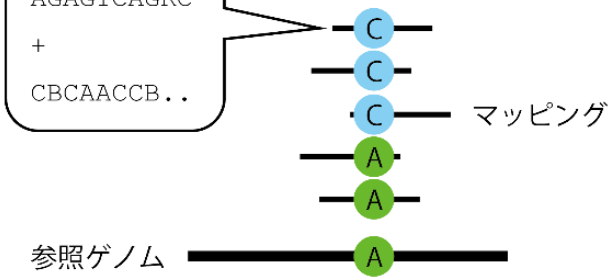
さまざまな解析プログラム

↓ R, python など

グラフの描画や結果の解釈

fastq ファイル

```
@read1
AGAGTCAGRC
+
CBCAACCB..
```



変異ファイル（vcf ファイル）

```
chr1 1010 A T 176.7 . AC=1;AF=0.500 ...
chr1 2014 G A 157.7 . AC=1;AF=1.000 ...
chr1 7888 C T 423.3 . AC=1;AF=0.500 ...
...
```

主成分分析など

遺伝学・集団遺伝学の基礎知識

集団遺伝学を苦手な人は多い

(3.31) $$u(p) = \lim_{t \to \infty} u(p,t).$$

For this probability,

$$\frac{\partial u}{\partial t} = 0$$

and $u(p)$ satisfies the ordinary differential equation

(3.32) $$\frac{V_{\delta p}}{2} \frac{d^2 u(p)}{dp^2} + M_{\delta p} \frac{du(p)}{dp} = 0$$

with boundary conditions

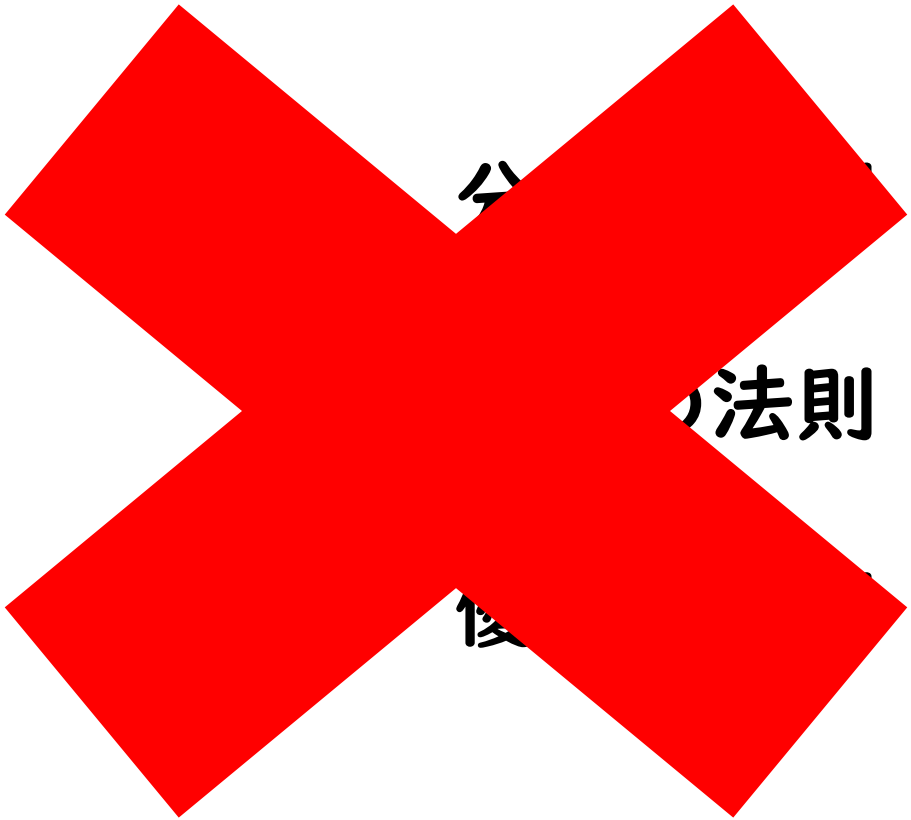
(3.33) $$u(0) = 0, \; u(1) = 1.$$

Equation (3.29) may readily be extended to multivariate cases: consider n independent loci each with a pair of alleles, a normal and a mutant allele. We will denote by $p^{(i)}$ the initial frequency of the mutant gene at the ith locus ($i = 1, 2, \cdots, n$). Let $u(p^{(1)}, p^{(2)}, \cdots, p^{(n)}; t)$ be the probability that all the n mutant genes become fixed in the population by the tth generation, given that their frequencies are $p^{(1)}, p^{(2)}, \cdots, p^{(n)}$ at $t = 0$. Then $u(p^{(1)}, \cdots, p^{(n)}; t)$ satisfies

(3.34) $$\frac{\partial u}{\partial t} = \frac{1}{2} \sum_{i=1}^{n} V_{\delta p^{(i)}} \frac{\partial^2 u}{\partial^2 p^{(i)}} + \sum_{i>j} W_{\delta p^{(i)} \delta p^{(j)}} \frac{\partial^2 u}{\partial p^{(i)} \partial p^{(j)}} + \sum_{i=1}^{n} M_{\delta p^{(i)}} \frac{\partial u}{\partial p^{(i)}}.$$

Kimura 1964

メンデルの法則



これはDNAとか染色体とかわからない時代に作られた法則

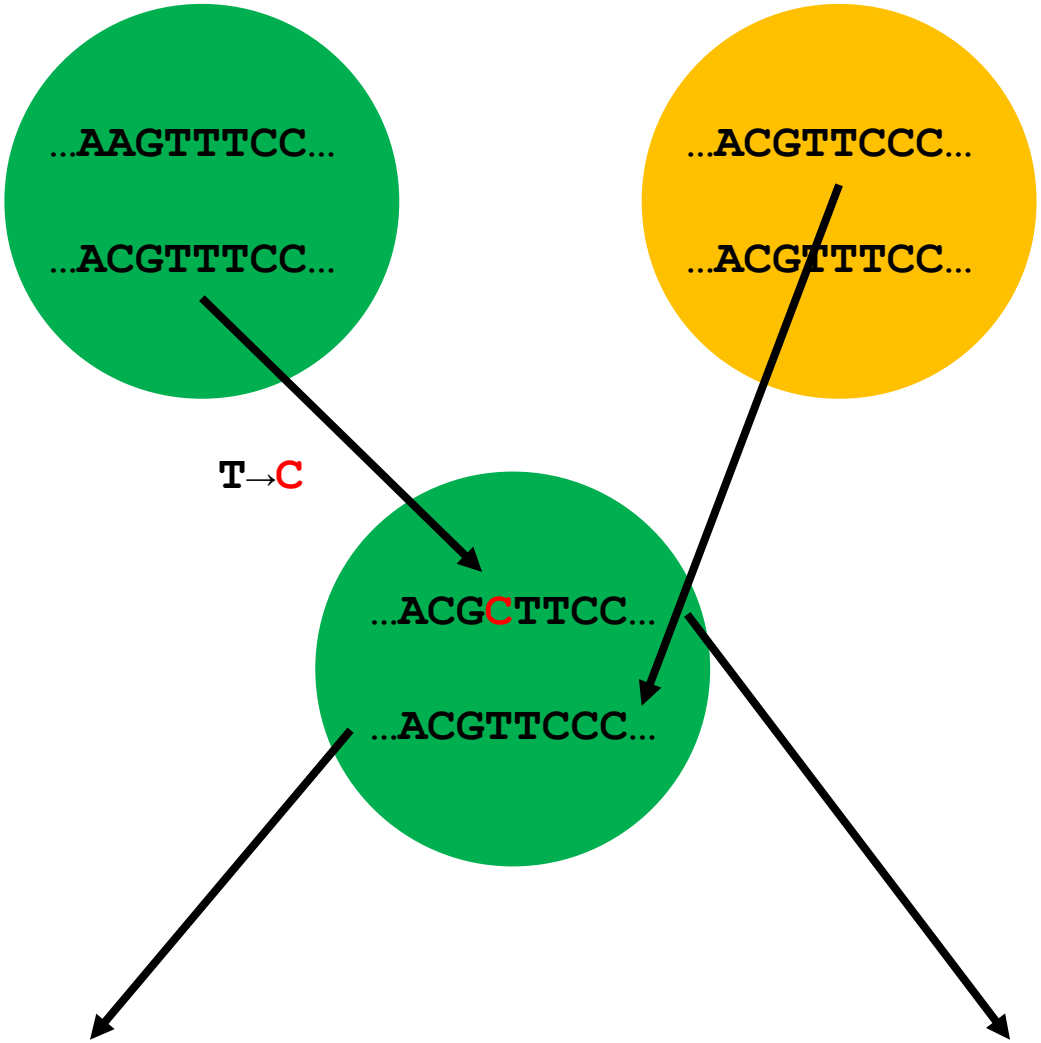
メンデルの法則

（2倍体生物の場合）
ゲノムの（塩基）変異は，その周りの領域と一緒に，
1/2の確率で子供に伝わる

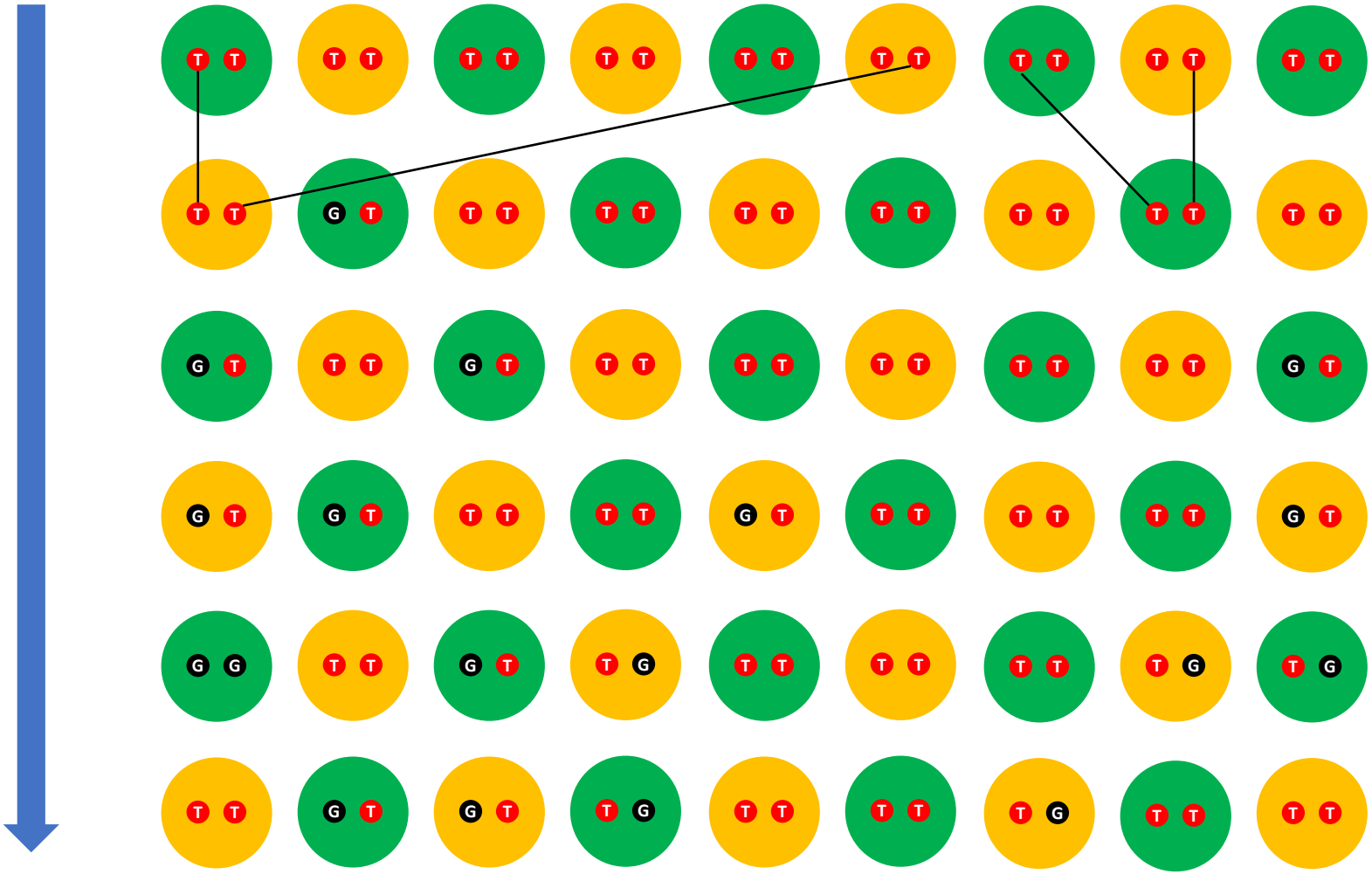
世代を経るごとに，ある確率で，塩基配列（アレル）に突然変異が起こる

変異が表現型に与える影響は，どのような場所にどのような変異が存在するかによって変化する

メンデル遺伝と突然変異

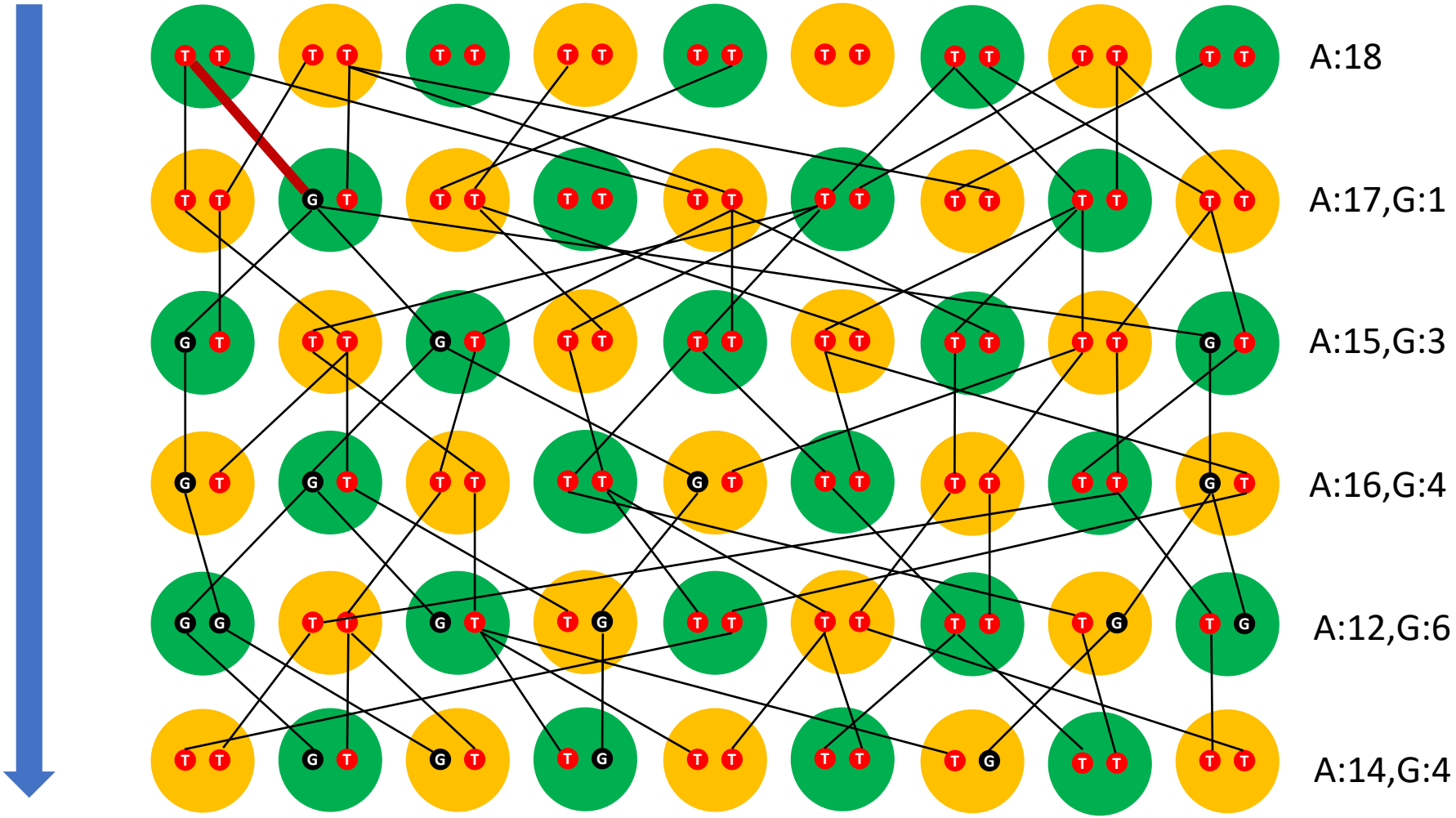


Wright-Fisher モデル



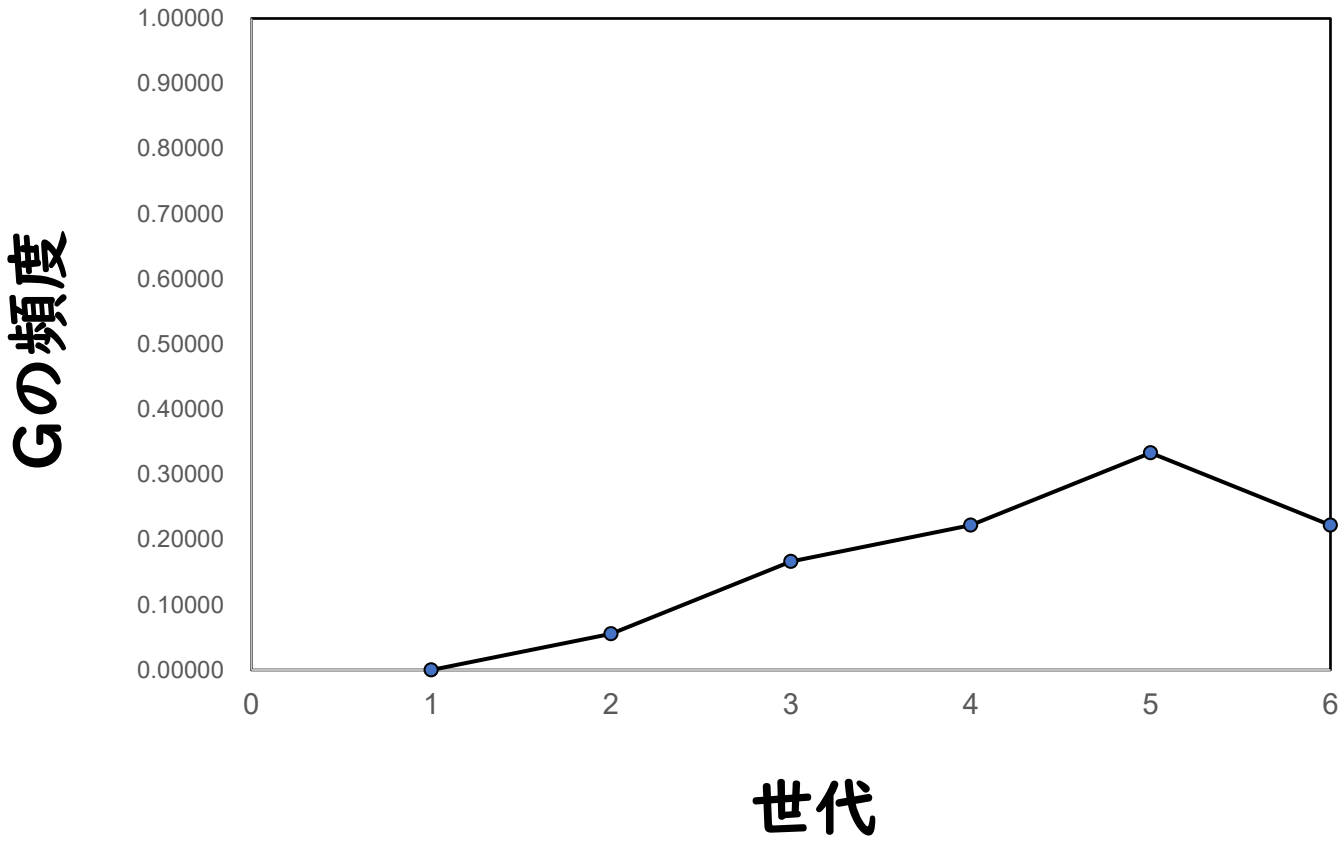
時間

集団遺伝学はメンデル遺伝の積み重ね

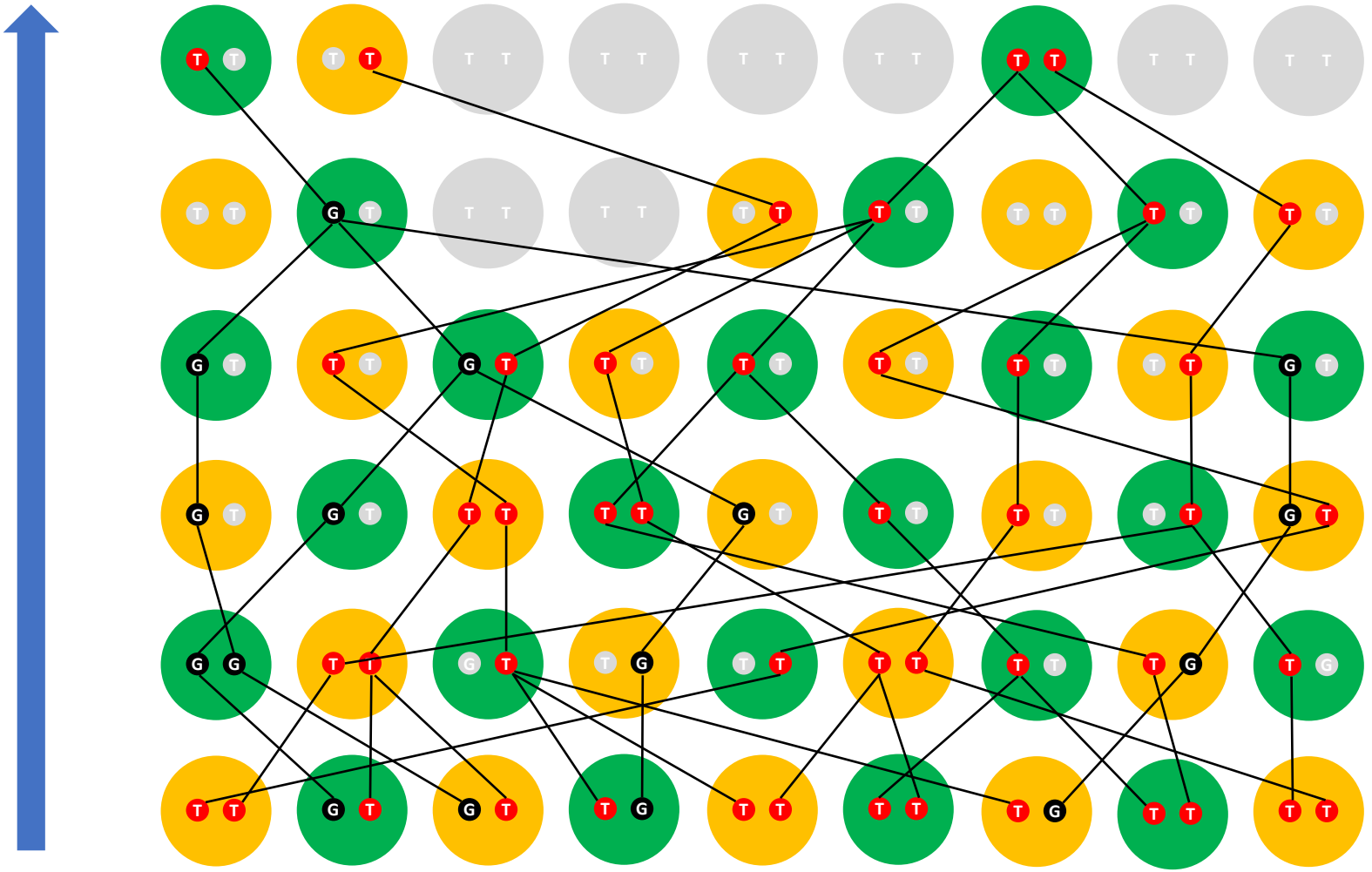


前向きな考え方

遺伝的浮動


1.00000
0.90000
0.80000
0.70000
0.60000
0.50000
0.40000
0.30000
0.20000
0.10000
0.00000
Gの頻度
0
1
2
3
4
5
6
世代

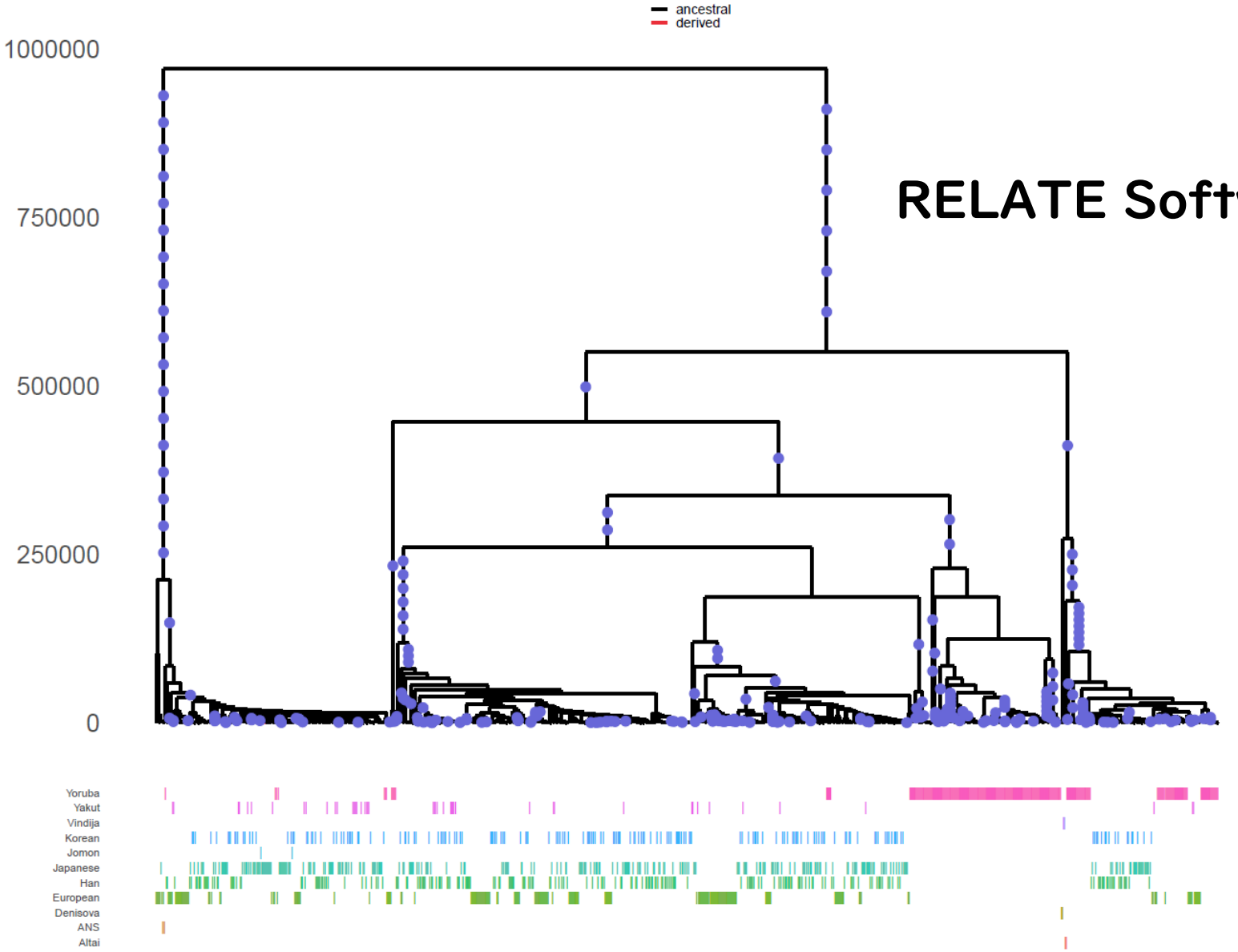
集団遺伝学はメンデル遺伝の積み重ね



時間

後ろ向きな考え方

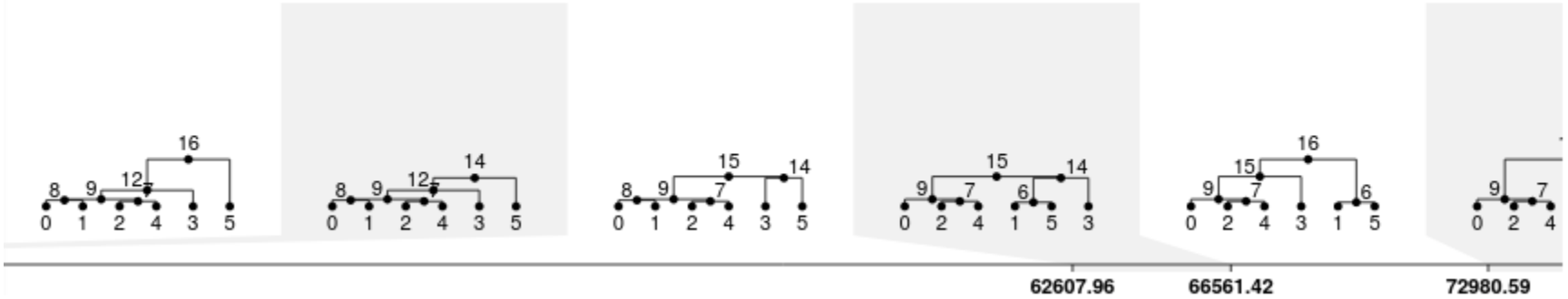
遺伝子系図（genealogy）



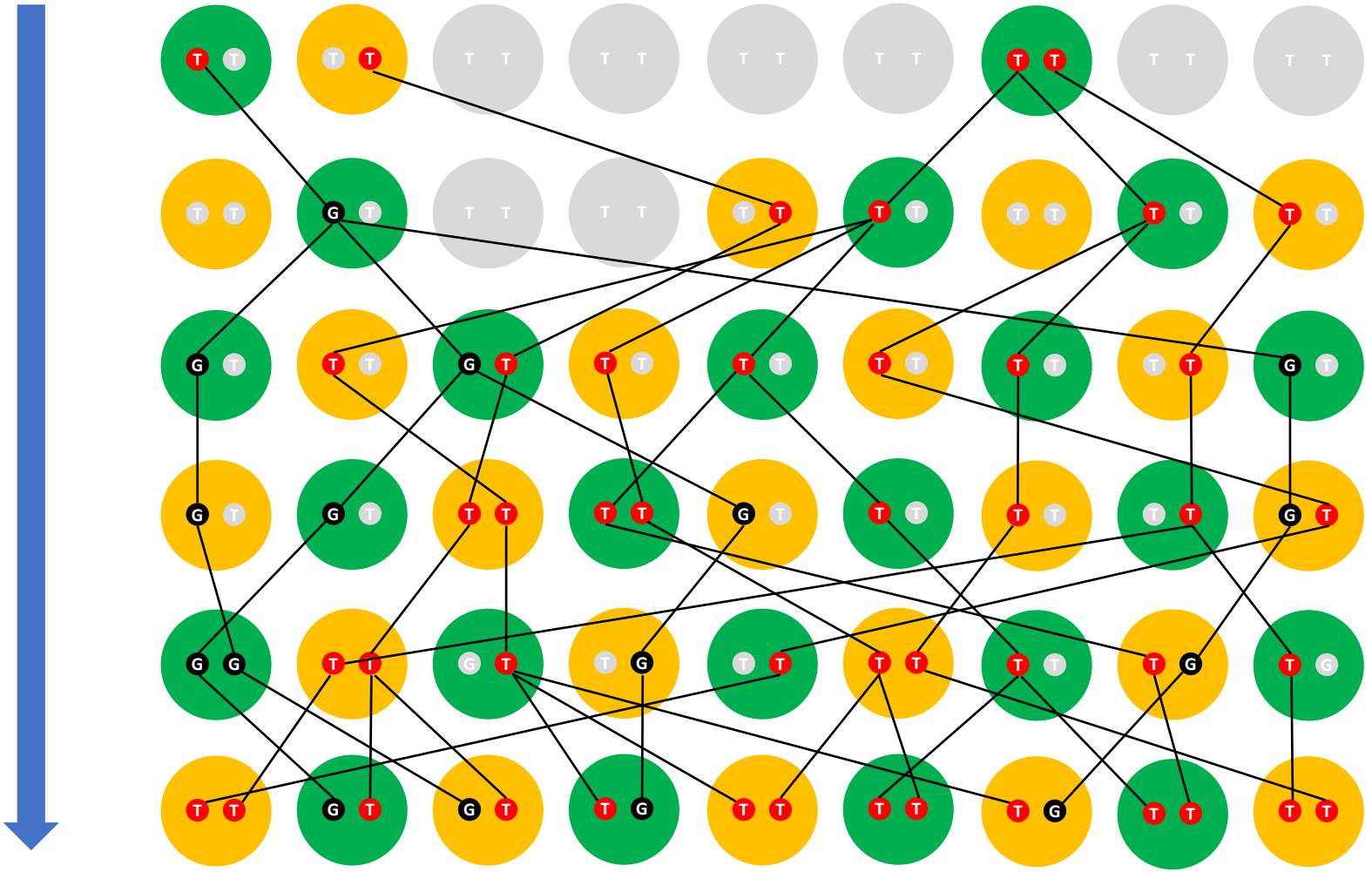
RELATE Software

一般的に観察は「難しい」

遺伝子系図はゲノムの場所によって変わる

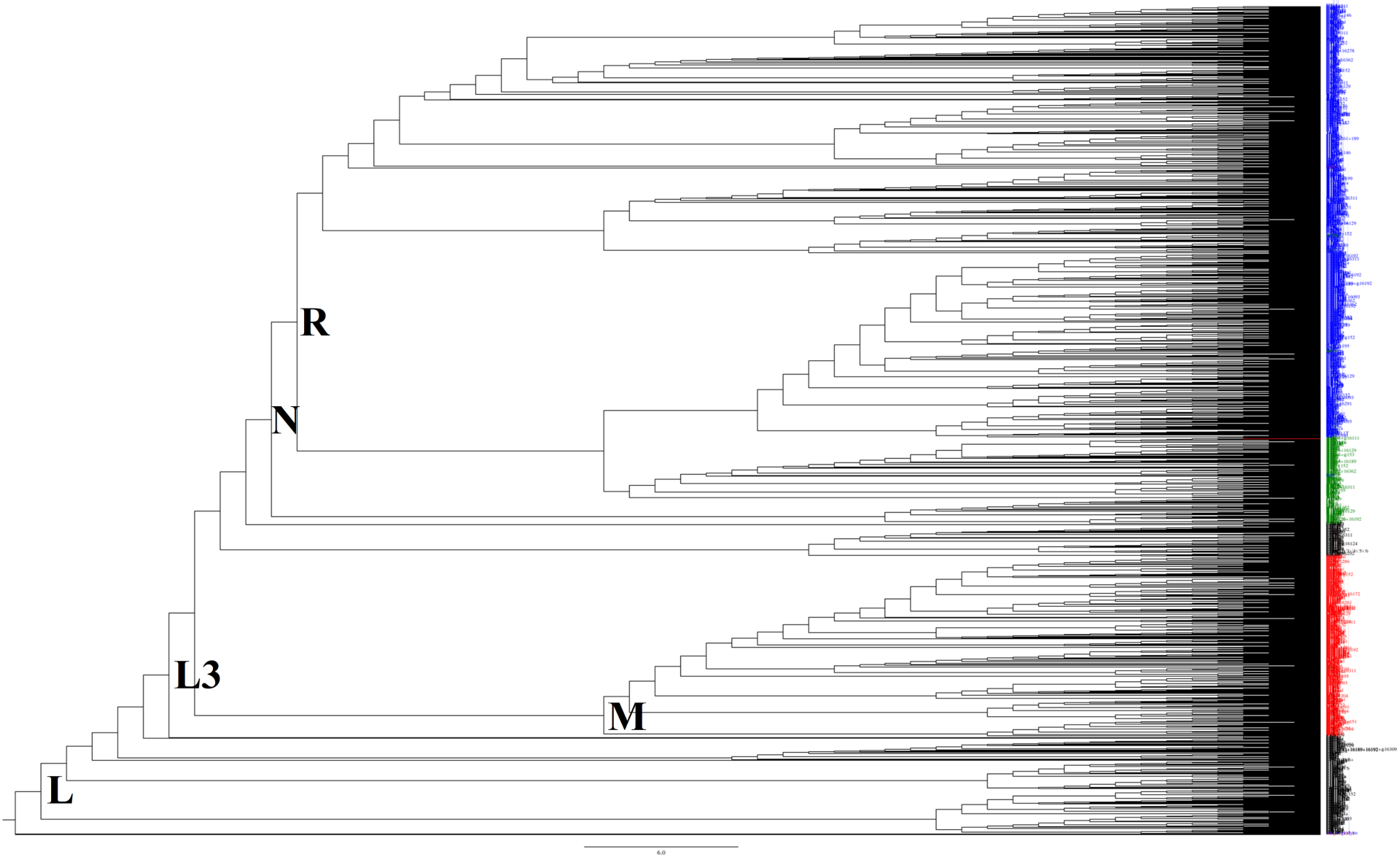


色々な場所を見ればほぼすべての祖先が見える



時間

ミトコンドリアハプロタイプ



遺伝子系図

藤原一道修士論文

核ゲノム解析の難しさ

ミトコンドリア：遺伝子系図が見える！

核ゲノム：「直接」観察できない！
　→（抽象的なモデルを立てる必要がある）

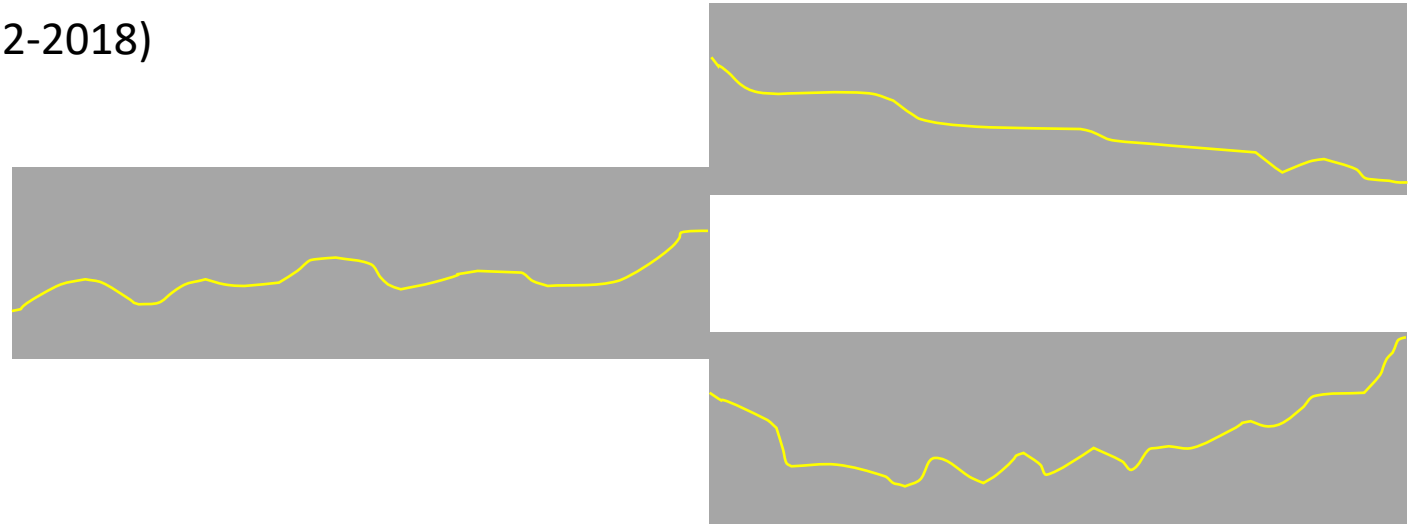
F_{st}（集団分化の指標）

集団A	集団B	Fst
A:100, G:0	A:0, G:100	1
A: 20, G:80	A: 20, G:80	0
A: 20, G:80	A: 80, G:20	0.72

１－（アレル頻度の相関係数）と考えることができる

Cavalli-Sforza

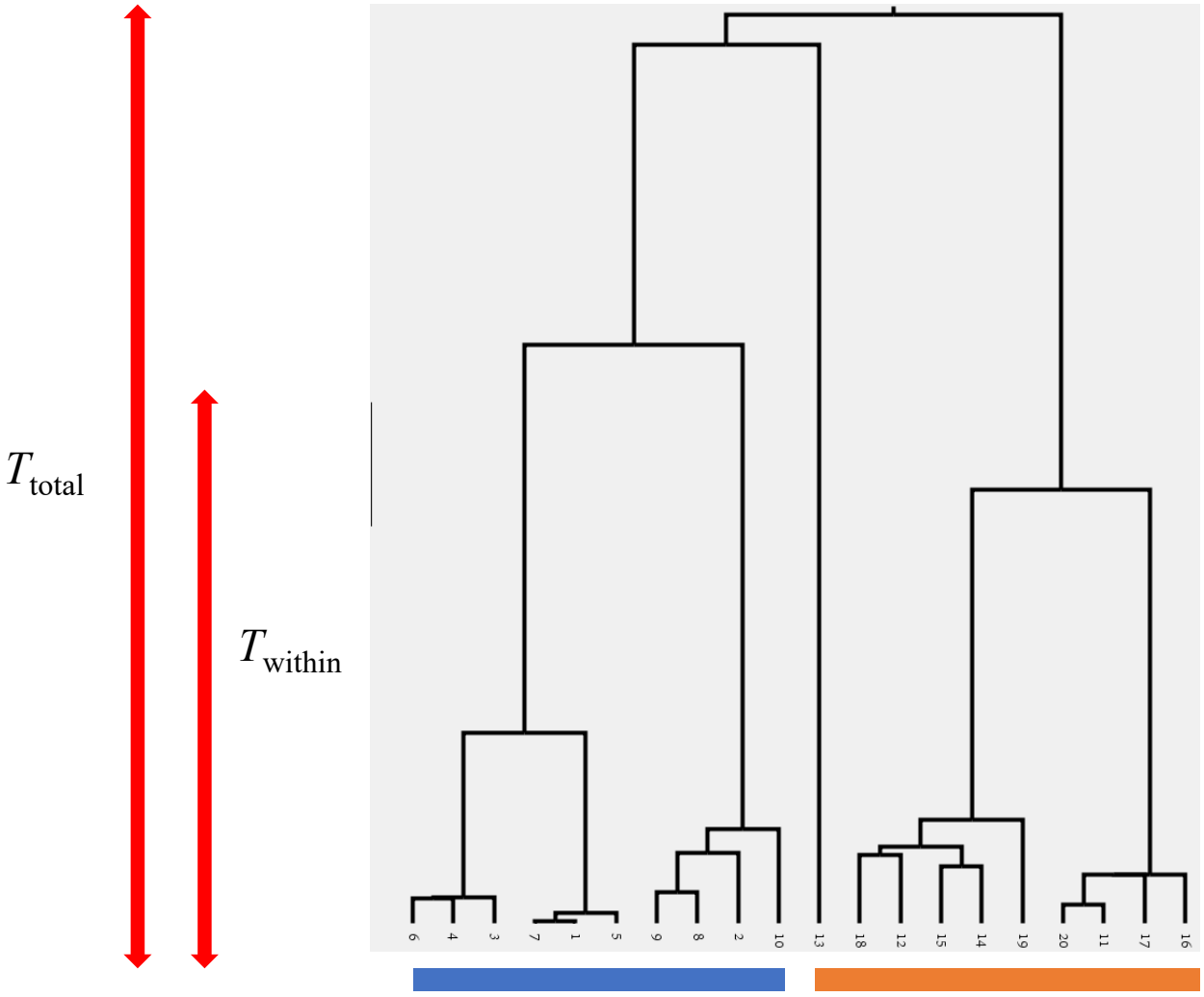
(1922-2018)



集団が分かれて時間がたつと，遺伝的浮動によってアレル頻度の違いが大きくなる（ブラウン運動を仮定すると期待値は0）

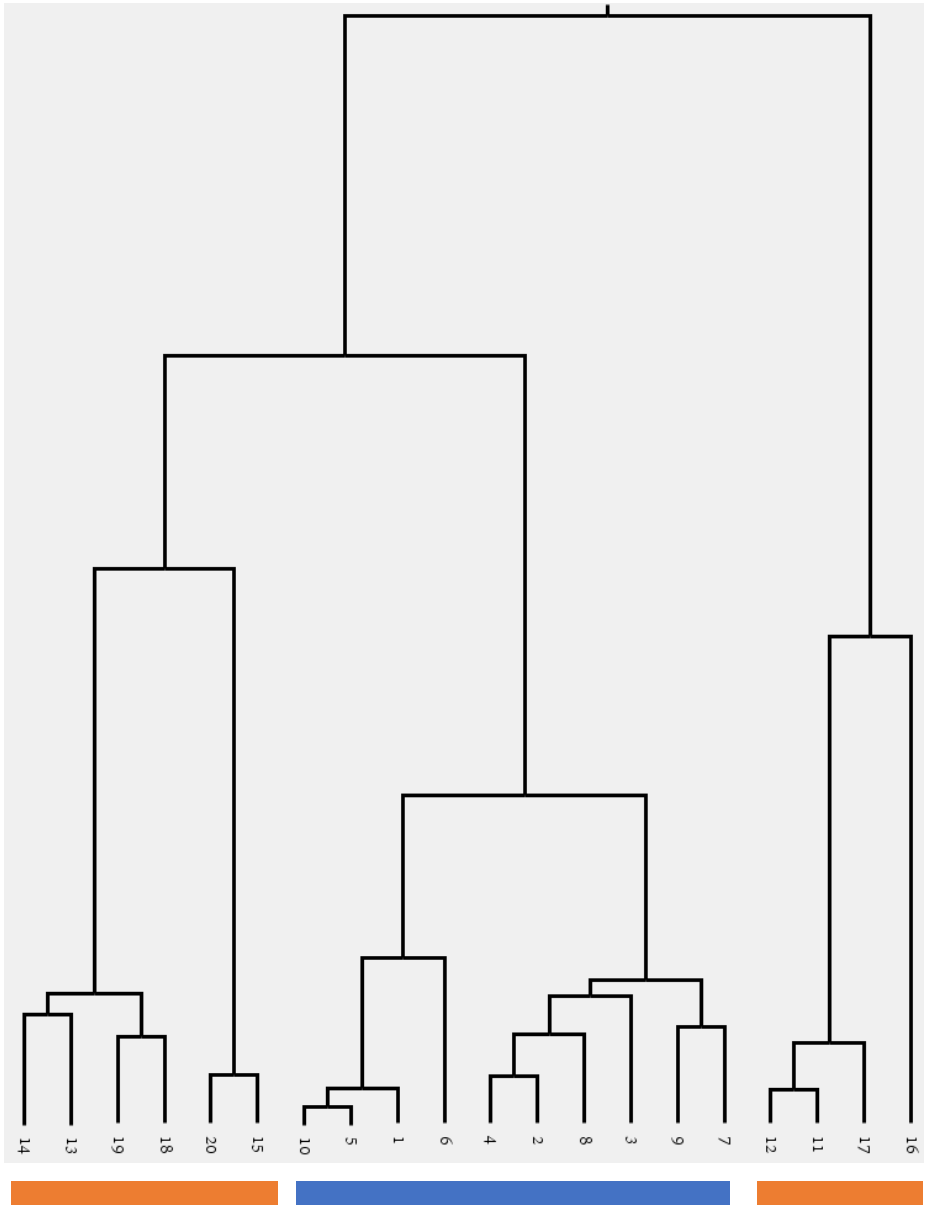
SNPの頻度を使って集団構造などの推定を行う手法はほぼこの系譜

F_{st}の後ろ向きな解釈



$$F_{ST} = \frac{T_{total} - T_{within}}{T_{total}}$$

F_{st}の後ろ向きな解釈



クオリティチェック

クオリティチェック

クオリティチェック（QC）を行うことにより，うまく読めなかったサンプルや，サンプルの取り違えを行う

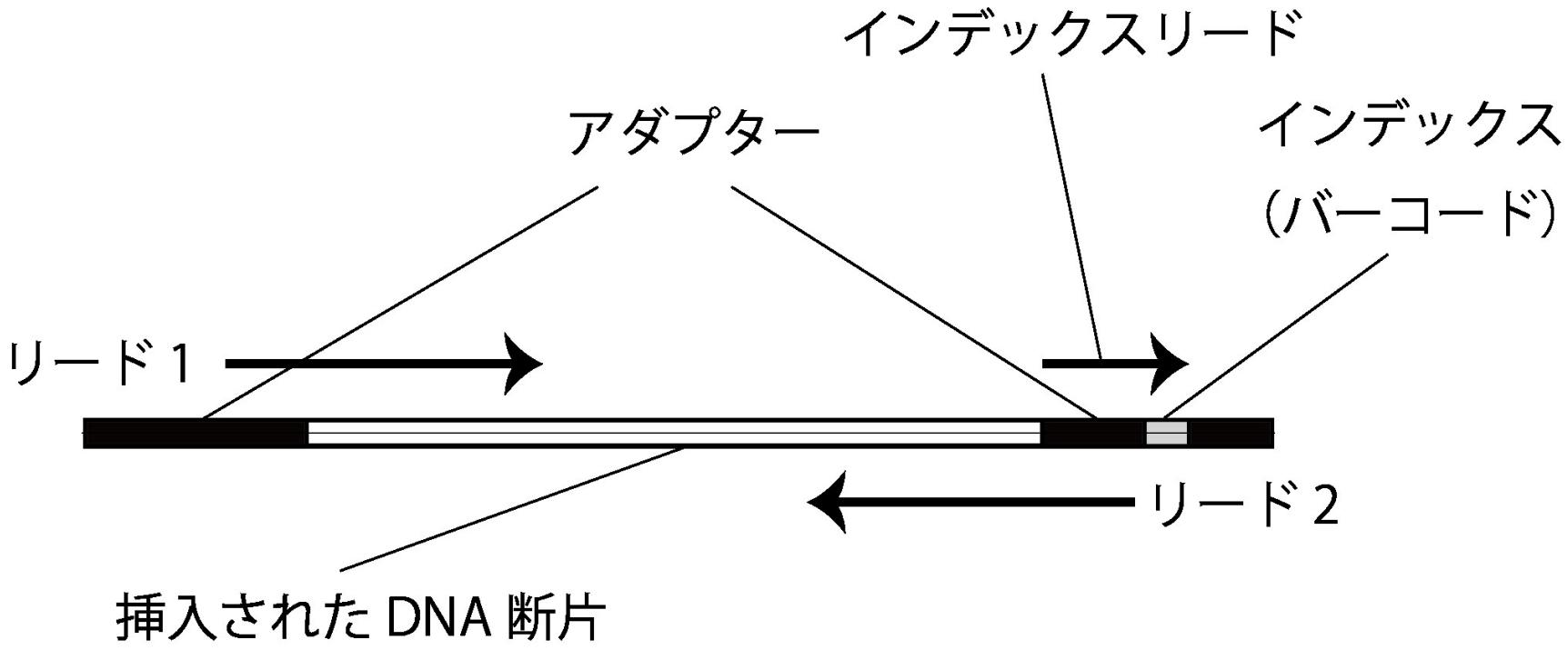
配列解読のクオリティチェック

- リードのクオリティチェック
- マッピングによるクオリティチェック
- 集団解析によるクオリティチェック

すべての段階でクオリティチェックが可能であり，解析を正しく進めるうえで重要なステップとなる

サンプルの取り違え（コンタミ）は最も重要で厄介な問題

一般的なショートリードライブラリ



挿入された断片が短いと，アダプターが読まれる
アダプターがたくさん入っている場合には除いた方
が良いことがある

Phredスコア

$$P = 10^{-\frac{Q}{10}}$$

$Q = 30$: 【　1,000　】個に一つのエラー
$Q = 40$: 【　10,000　】個に一つのエラー

FASTQフォーマット

```
>sequence_id
ATGTAGTTACTAGCT
```

【　FASTAフォーマット　】

```
@sequence_id
ATGTAGTTACTAGCT
+
CAT’’+(>>>>AGBG
```

一つの文字で【　エラー率　】を示す

アスキーコード

ASCII	文字	ASCII	文字	ASCII	文字	ASCII	文字	ASCII	文字
33	!	40	(	47	/	54	6	61	=
34	“	41	)	48	0	55	7	62	>
35	#	42	*	49	1	56	8	63	?
36	$	43	+	50	2	57	9	64	@
37	%	44	,	51	3	58	:	65	A
38	&	45	-	52	4	59	;	66	B
39	‘	46	.	53	5	60	<	67	C

FASTQファイル

通常，ForwardリードとReverseリードが別のファイルに分かれて格納される（または同じファイルに交互）

SP01_R1.fastq.gz

FASTQファイルのQC

何をすべきか

極端に低いクオリティの配列ばかりか？
アダプター配列が多く混入していないか？

悪い結果であってもとりあえず先に進み，マッピングやバリアントコーリングの結果を見てから決めてもよい

配列のQCはあくまでも目安

FastQC

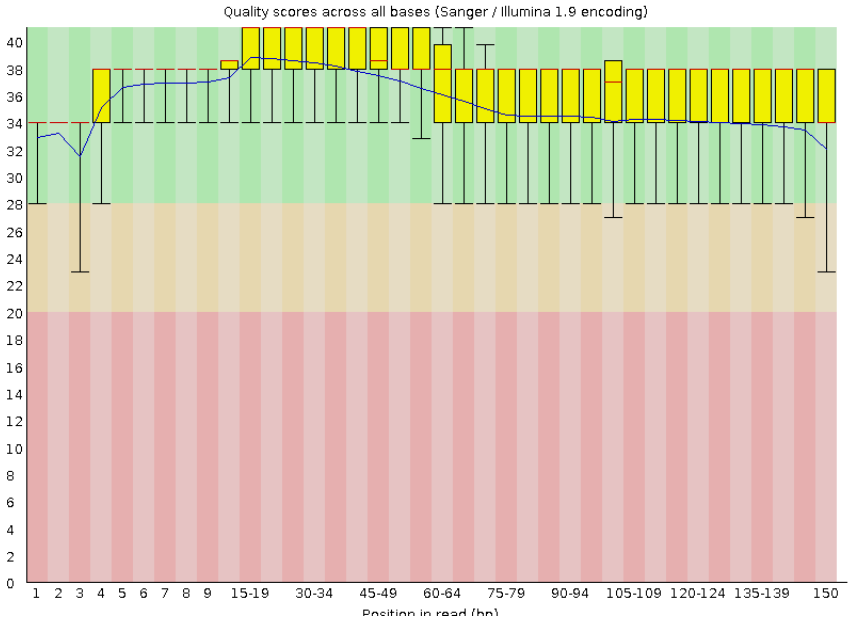
FastQC Report

Summary

Basic Statistics

Measure	Value
Filename	SP01_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2586391
Sequences flagged as poor quality	0
Sequence length	150
%GC	41

Per base sequence quality



MultiQC

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2022-12-07, 11:22 JST based on data in: `/mnt/e/Google/working/ゲノム多様性テキスト/ゲノム多様性解析入門_共有/data/2/fastqc_results`

Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

Copy table | Configure Columns | Plot | Showing 4/4 rows and 3/5 columns.

Sample Name	% Dups	% GC	M Seqs
SP01_R1	2.0%	41%	2.6
SP01_R2	0.9%	42%	2.6
TK01_R1	2.1%	41%	2.7
TK01_R2	0.9%	42%	2.7

FastQC

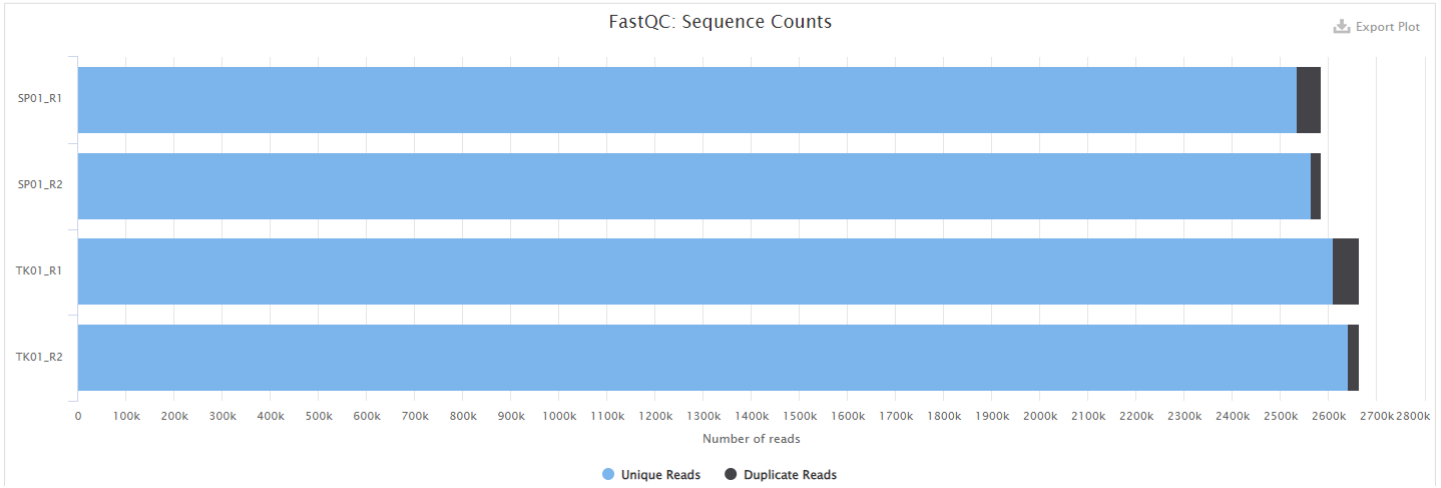
FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Counts

Help

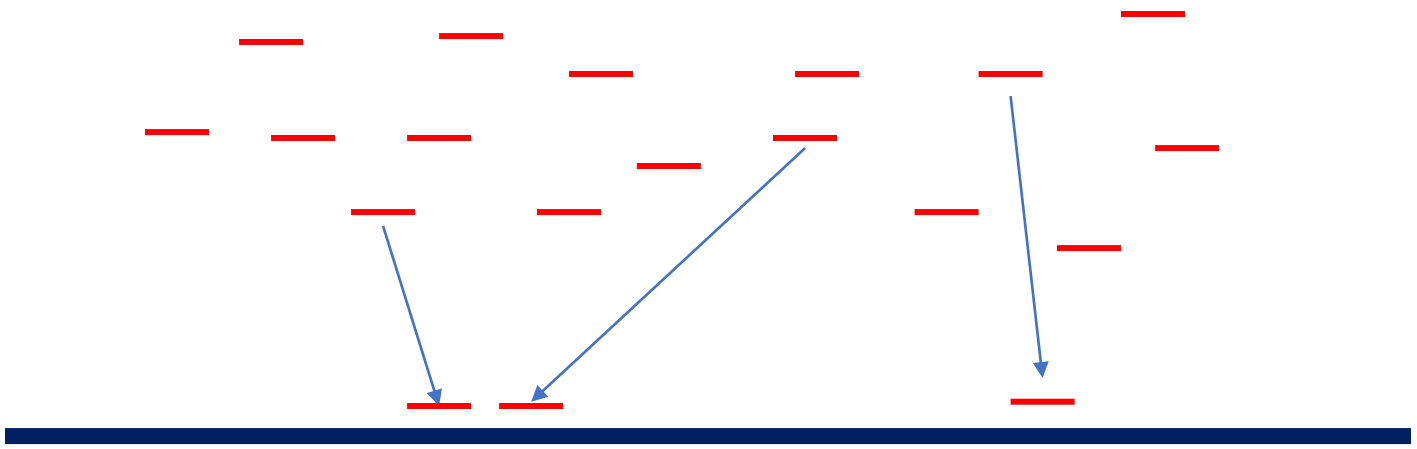
Sequence counts for each sample. Duplicate read counts are an estimate only.

Number of reads | Percentages



マッピング

マッピングとは



リファレンスゲノム配列

一度の実験で数100万～数1000万リードを張り付けなければいけない

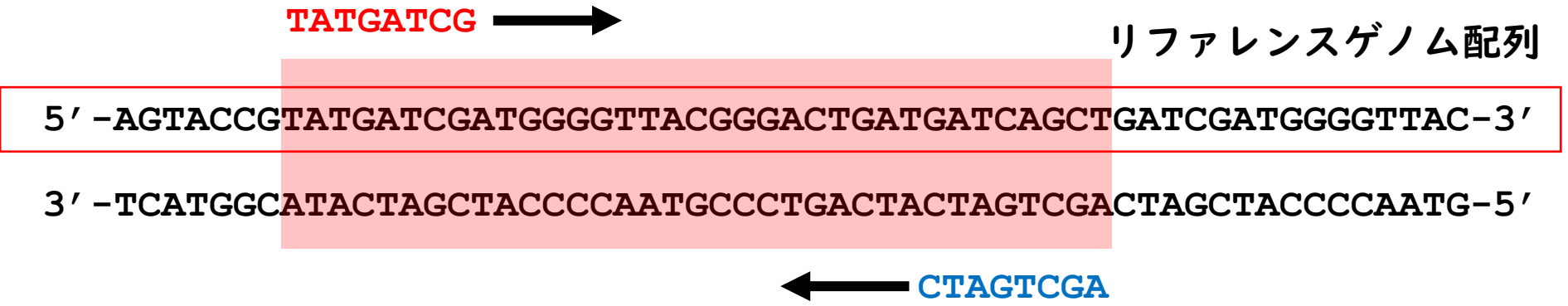
さまざまなソフトウェアが存在
e.g. BWA, Bowtie, …

どれがベストなのかは種やゲノム構造にも依る

メイトペア配列①

TATGATCGATGGGGTTACGGGACTGATGATCAGCT
ATACTAGCTACCCCAATGCCCTGACTA**CTAGTCGA**

R1.fastq **TATGATCG**
R2.fastq **ACGACTAG**

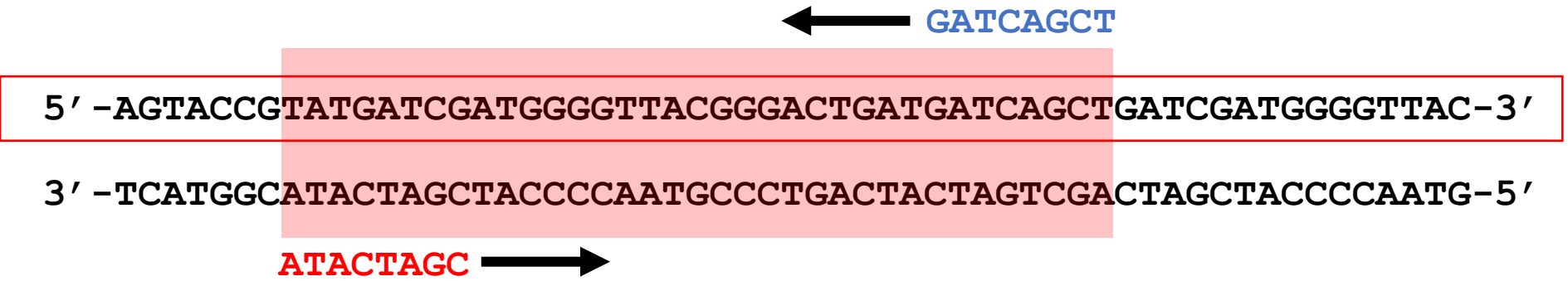


Read１はForward
Read2はReverse
の向きにヒット

メイトペア配列②

TATGATCGATGGGGTTACGGGACTGATGATCAGCT
ATACTAGCTACCCCAATGCCCTGACTACTAGTCGA

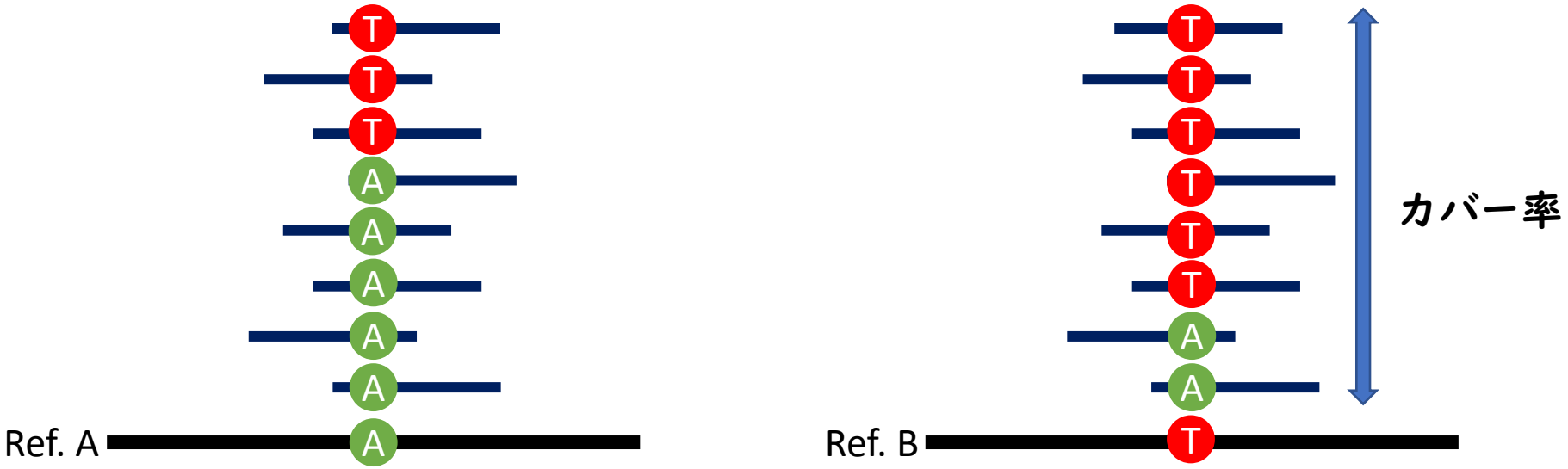
R1.fastq ATACTAGC
R2.fastq AGCTGATC



Read1はReverse
Read2はForward
の向きにヒット

リファレンス（参照）ゲノム配列

リファレンス配列の由来は重要

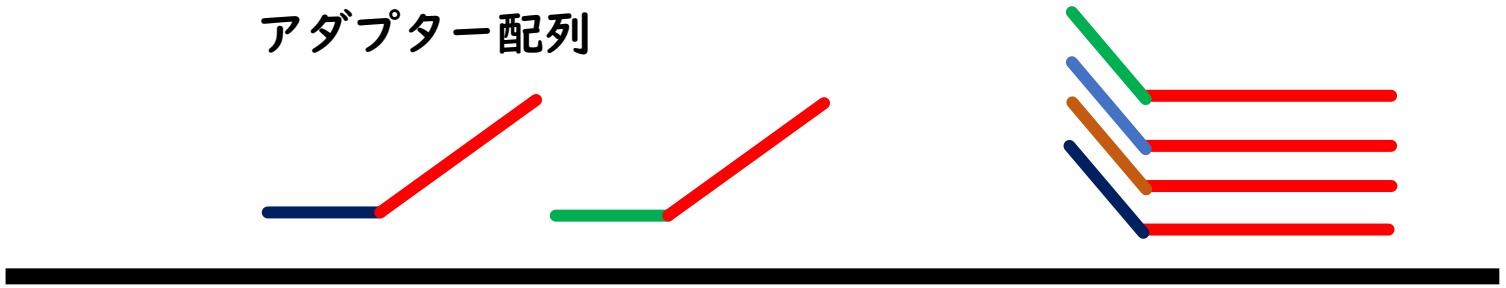


基本的にリファレンス配列に近いリードはマッピングされやすいというバイアスがあるので注意

とくに，カバー率が低いものは注意

割り切ってアウトグループにマッピングするという手もある

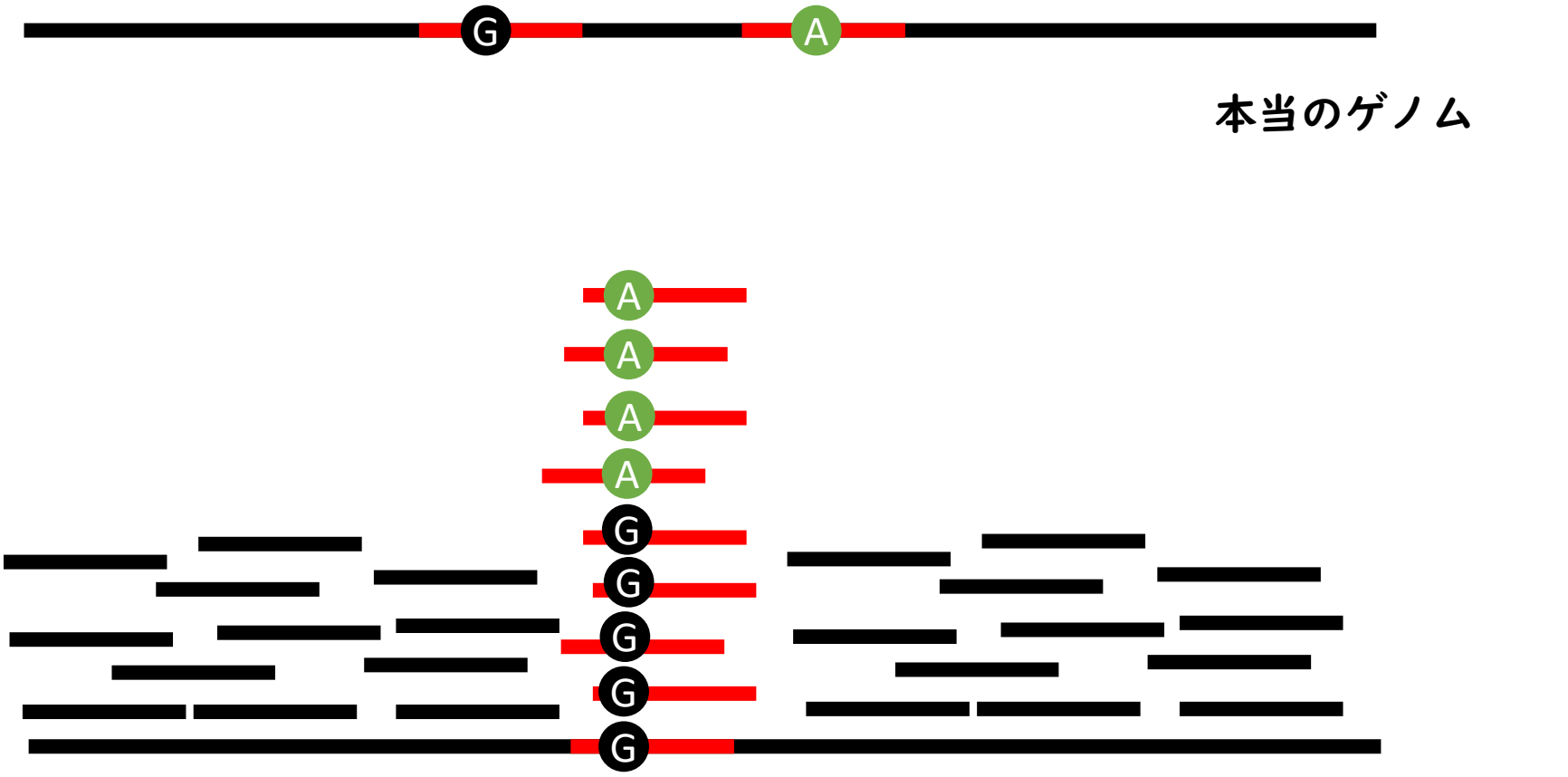
アダプター配列のマッピングへの影響



ヒトの場合はたぶん気にしなくてよい（アダプターはヒトゲノムにマッチしないようにデザインされているはず）

それ以外の生物ゲノムの場合は注意する必要がある

反復配列の影響

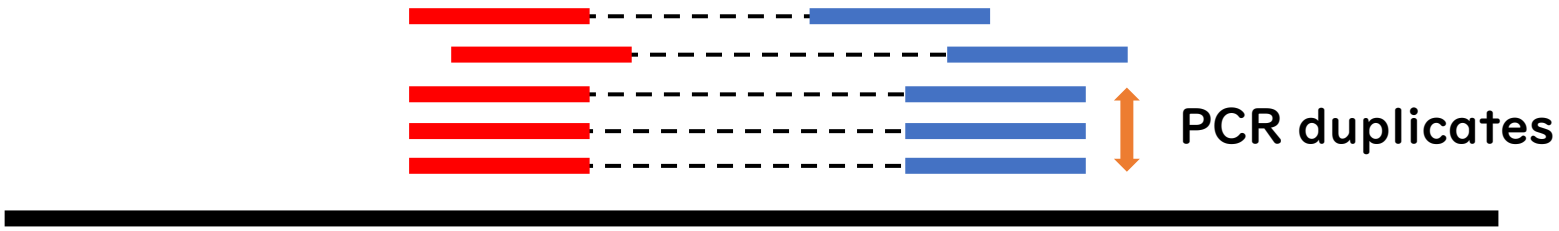


PCR duplicates

次世代シークエンスのPCR増幅
　　ライブラリ調整
　　フローセル上で1つのスポットが2つ以上に分かれて認識される

データ解析上は，両端の位置が同じフラグメントがPCR duplicateと認識され，ラベルされる（picard, samblasterなど）

一般的なバリアントコールプロトコルでは除かれる



マッピング結果の保存（アラインメント）

SAM，BAM，CRAMファイル

SAMファイル
　テキストファイル

BAMファイル
　圧縮されたバイナリファイル

CRAMファイル
　マッピングされなかったリードを効率よく保存
　ファイルサイズ小
　ソフトウェアによっては取り扱えない

これらのファイルの操作にはsamtoolsが広く使われている

Samtools

```
HK01_1-2508128  83      chr1    5445993 60      150M    =       5445687 -456    ATTCATATTATAAAAATGAAAAAATTAAGGATAATATAAAAGATTTAAAATTTAAATGTATTTG
TTGCATAGAGTATATATATCTTAAGAGTATTTTCTTACAGAGAACCATTTTTATTCAAAAGTACTTACATTTATTTTTAGTGCATC  CCCCGGCGGGCGCCGCGGGGGGCG$CCGGGGGGGGCCC1GG$GCGGGCCJCGGGGG
GGG8CCGGGGCGG$CCGGGCCGCGGCC8CJJJCGGJJJCJJCGGJGJGJJ1GGJJGCGGGJJCGJJJC1JGGJCJGJJGGJGGGGGGGGGCCCC  NM:i:3  MD:Z:41T27A58A21        MC:Z:150M
AS:i:135        XS:i:21 RG:Z:SP01
HK01_1-2508128  163     chr1    5445687 60      150M    =       5445993 456     TAAATACTAACCGCATAAAGCGGTGGCCTAATCCTCTAAGTGGGGGTTGGGTGCTCTGTGTGGG
AGGGGGGTATCTGCAAAGTCACCCCCTGGGGCAATTATACACCCCTGAAGATTCAGCAACCGCAGCAGGCCCAATTACGCAACATG  CC1GGGGGGGGGGJJ=CGJJJGGJJ8JJG=GJJGGGJCJ=JJ=GGJGG8JGGCJGG
JCCCGGJGGCJCGGCJCGCGCGCGGG(GGGGGGGGCCG$$G$C=GJ1$=JJGCGGCG$=CGGGGGGGG$GGGGC$GG$G=GCGGCG$$GCCC$G  NM:i:7  MD:Z:94T0T1T15A10A8A8T7 MC:Z:150M
AS:i:115        XS:i:0  RG:Z:SP01
HK01_1-2508126  83      chr1    19909190        60      150M    =       19908895        -445    AGCTGAAACCACCCAGTATTAGCGTGAGTTCCCAGAAGCCCCAGCCCA
AGACGAAGCAATCCCAGCTGCTGGCTGCCCAGCAAAAGAAGGTACAGGCCAACAAGGCGACCAAGCTGGACAAGAAGGAGAAGAGTCCGGCGAGGTCCTTGA  GCGGG=CGCCGGGGCGGCCGGGGGGGGCCCGCCGC=GGCG
GGGGG1GGGJGGG=GGGGC1GC$CGCGCCCCG=CCJGG1GJGGCGJCGJC$GGJGJJJJGCJJCJJJJJGGJCJCGJJJGJGGJJJJJJCJJJGGGJGGGGGGGGCGC==  NM:i:2  MD:Z:62A27A59   MC:Z:150
M       AS:i:140        XS:i:20 RG:Z:SP01
HK01_1-2508126  163     chr1    19908895        60      150M    =       19909190        445     TATATCATTTCTATATGATTATTAAAAAATGTATCTAATTCCAGCCTC
CGGCCCCCGAGCATTAAGAGCGGCATTCCCTCGCCGAGCAGTCCGCAGACCGCTCCCCAGACGCACTCGATGCTGGACAAGCTGAAGCTCTTCAGCAAGGAG  CCC1G1GGGGGGGJJJJ8GJJJJGJJGCJGJJJCJ8CGJC
JGJGJGJJGC$=GJGCGGGC(GGJG8JGCGGJ1GCGCCGGGGG==CGG8GG=CGGGGGGGGCJJ=CJGG$C=GCCGGGCGGCGGGGG=G=(CGGGGCG=GGG$8GCGC==  NM:i:3  MD:Z:50A58A32A7 MC:Z:150
M       AS:i:135        XS:i:0  RG:Z:SP01
HK01_1-2508124  83      chr1    1861867 60      150M    =       1861484 -533    CCAGGGGTTCGAAGGGTGATTGTGGGTTATGCACATAAAAACAAGTGATGCCCAAAAGTCAAAT
GGCGCAAGTGCGGGCTTTAATTCAACGAAATAACTTAATGCCCGGATGGCTTGGCAAACTGCGACGCCAGAGCCACCGCGTCCTCC  8GGCGGCGGGC=GGGG$GCG=GG$GGG$CGGG8GGGCGGG8CGCGGCCGJC=GGCG
GCGCGC=CCG8GCG8GGGJGGGGGCGJJJJGGJCCCJGCGCGJGCGJJJJJGG8JGJ=JG8CGJGJCJJJJJJJJJGGJJGGGGCGGG1G8CCC  NM:i:1  MD:Z:16G133     MC:Z:150M       AS:i:145
        XS:i:0  RG:Z:SP01
HK01_1-2508124  163     chr1    1861484 60      150M    =       1861867 533     CTTTTGGCTTGTTTGTTTATTCTATGGCAATGTTGTTCTTTGCACTCCCCCACCCCCATCCCCT
GCCACCCTTCCCCTTAAATTGGGGAAAATAAACATTGTTGGTTTTGTACAGCAGCTGCCATTGTCATTCCTTTTAGTCAAGCTGTA  =CCCGGGGGGGGGGGJCJJ(JJJJJGJJJ$GJCJJGJGCCJGGJJJGCJJCGJCCGG
JJ$JCGJJG=GGCJGGGGGCGGGGGGG1GGGGGGCGCCGGCGGGGC=J$J=GG$GGG$GCGCGGG$G(CGGG$G=GCCGC=G$GGGCCG$GGCC  NM:i:7  MD:Z:58A1A43C4T18T9T6T4 MC:Z:150M
AS:i:115        XS:i:0  RG:Z:SP01
HK01_1-2508122  99      chr1    23342272        40      150M    =       23342608        486     GAAACTGATTTTATTTGGAAATATCTTCGGTTTAAATAGGTGACATGA
GAATCGCATCTTACAGTAAATGGCCTACGCAGAGGCCTAAGTACAGTGTCCCCGCCTTATCGAGGCCCCACGCTCGGGCACATCTGCCTATCTAGAGCAGCG  CCCGGGGGGGGGGGGJJGJJCGJJJJJ=JG8JJJGJGJJJJ
GGJJJGJJCCJG=GGGGGCGG$GJGCCJJJGCGCCGGGGGGG=GGCGCGGG$G$$GCCGCCCCGGGGCGCGCG$CGGGG1CC=GC=GGCGGG=CCGGG=C$$GG1G$GCG  NM:i:7  MD:Z:61A29A1T0A18T27T4G3
        MC:Z:150M       AS:i:116        XS:i:116        RG:Z:SP01
HK01_1-2508122  147     chr1    23342608        60      150M    =       23342272        -486    GCCCAAAATTGTAAGTCTTTAAATATATTCGTGTTCATGTGTGAACAC
GAAGTTTGATGCAACAACGGGGCTTCTGCAAAGGCCAAATTACTCGTGCGCATAATAATGCCTTAAAATTTGTTGAAGACATTCATTCAGTGCAAACAATAG  CCCGCGG=C$GCC1CCGCGGGGGCGGG81G8G=GG1GGGG
GC11CJCCC$GGG(GGGG$8CGC1GGCGCCGG=CGGCGG=$GGCCJGGGGJCGGGGG=JJCJJCJCCCGGJJGJJJGGJGJGJGJJJJJ1JJJJCJJGGGCGCGGGG1CC  NM:i:5  MD:Z:49G3G4T21A43T25
MC:Z:150M       AS:i:125        XS:i:49 RG:Z:SP01
```

1リード1行
2列目のフラグはリードを抜き出したりするときに便利

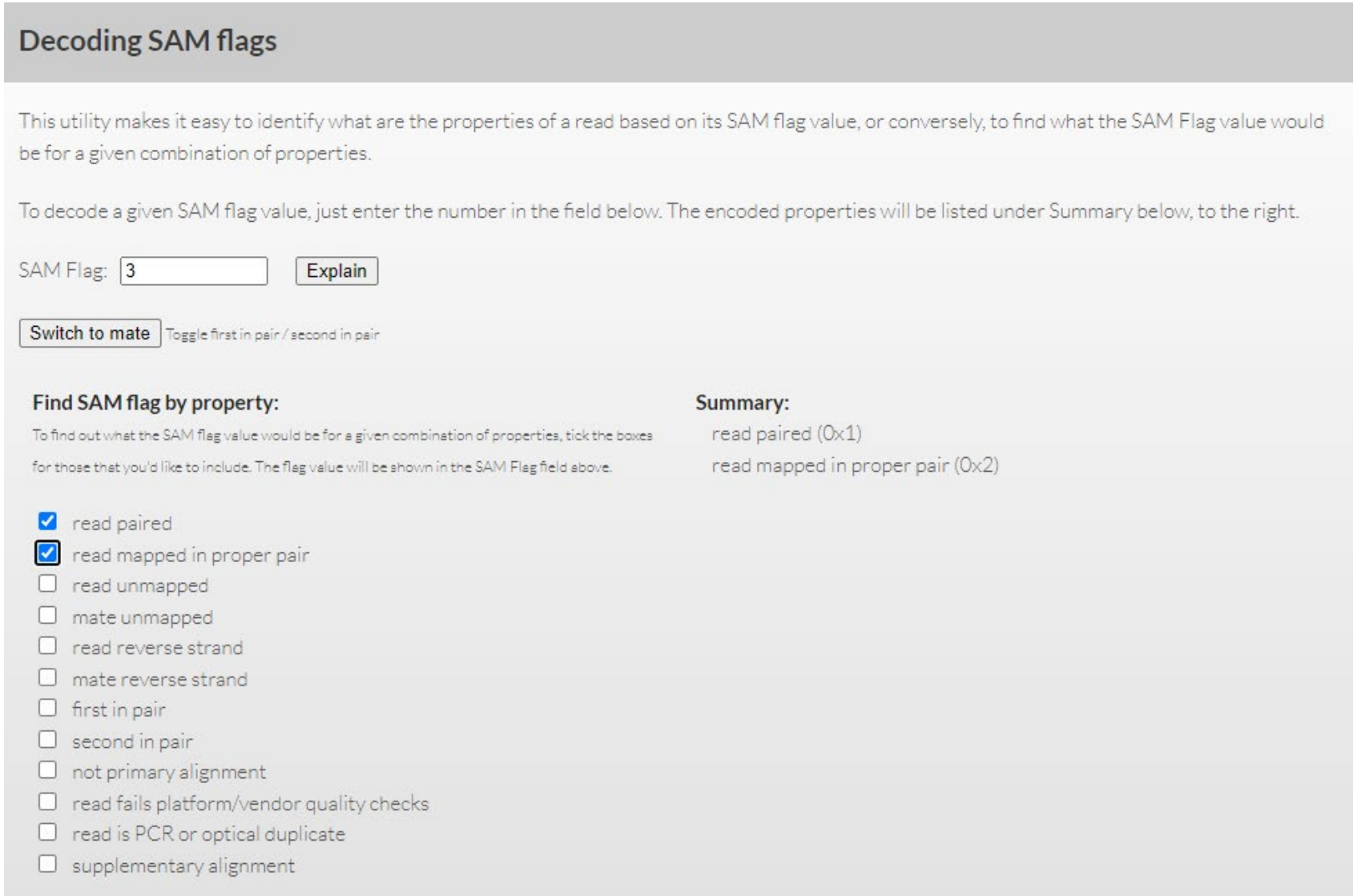
Samフォーマット

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

[11] Reference sequence names may contain any printable ASCII characters with the exception of certain punctuation characters, and may not start with '*' or '='. See Section 1.2.1 for details and an explanation of the [:rname:] notation.

5列目のマッピングクオリティ（MAPQ，MQ）は重要な指標
高いほどその場所にユニークに張り付いている
MAPQ=0はそのリードが他の場所にも同じ感じでマップされているということ

Flagの解説



https://broadinstitute.github.io/picard/explain-flags.html

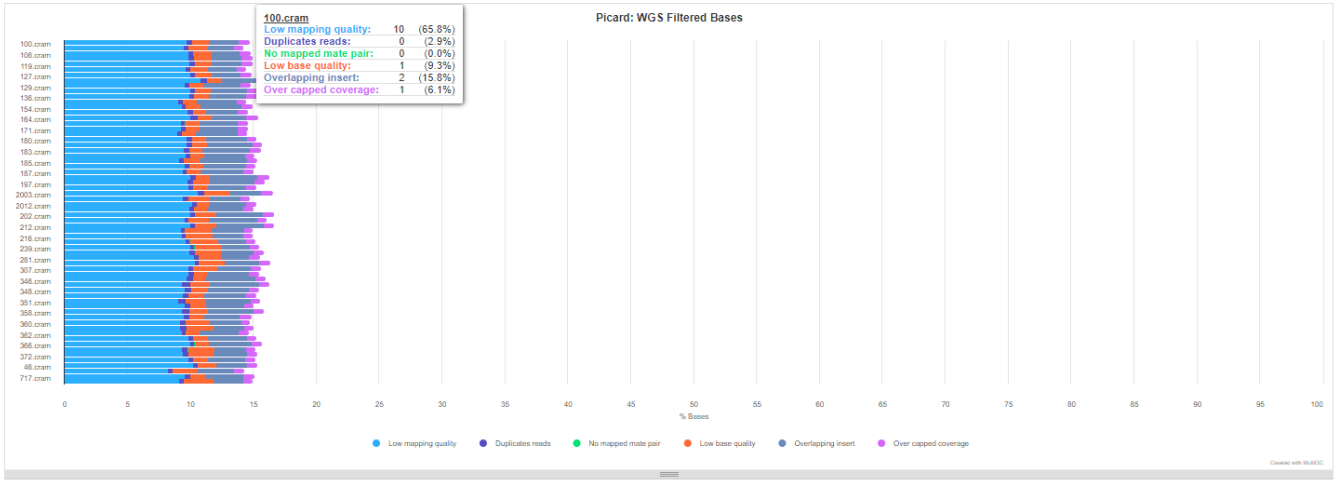
アラインメントのQC

例：picardのCollectWgsmetrics → multiQC

Sample Name	% Aligned	Median Coverage	Bases ≥ 30X
100	100%		
100.cram		41.0X	78%
102	100%		
102.cram		31.0X	55%
106	100%		
106.cram		32.0X	63%
112	100%		
112.cram		35.0X	73%
119	100%		
119.cram		36.0X	73%
125	100%		
125.cram		36.0X	74%
127	100%		
127.cram		31.0X	58%

WGS Filtered Bases

For more information about the filtered categories, see the Picard documentation.

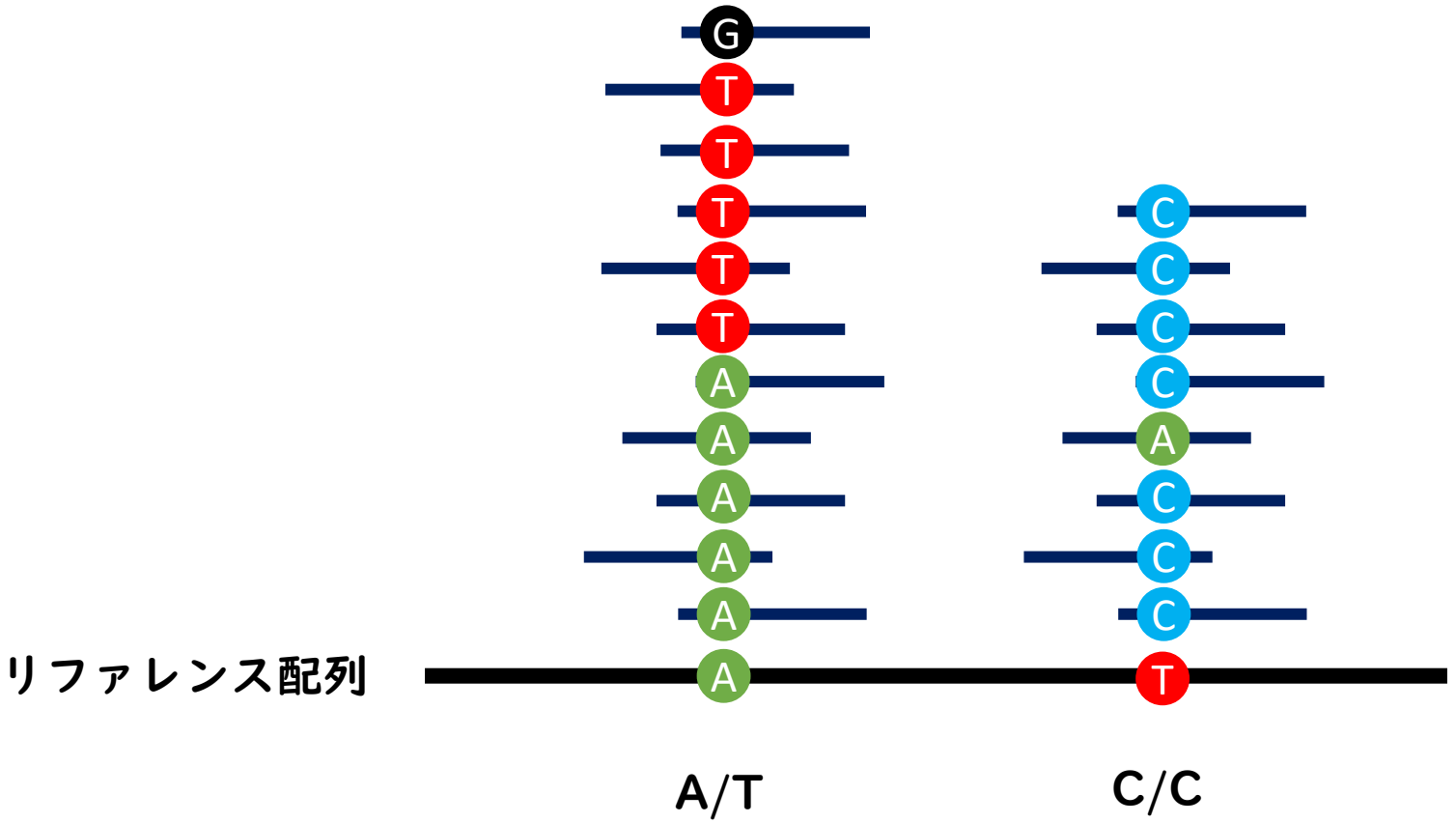


どれくらいのカバー率が必要か

- 理想は30x以上（ハイカバレッジゲノム）
- 多すぎても意味がない（100x読んでもお金の無駄）
- ２～３xのうす読みであっても，解析に役立てることは可能（古代ゲノムなどは1x程度のものも多い）

バリアントコール
（変異検出）

バリアントコール



遺伝子型（ジェノタイプ）を推定するのがバリアントコール

バリアントコールソフトウェア

GATK
- ヒトゲノム解析では一番基本的
- 少し面倒な仕様が多い
- ヒト以外への適用については未知数

Samtools (mpileup)
- 使うのが簡単
- GATKとそれほど変わらないという論文も

Angsd
- 集団内の頻度から尤度を計算することが可能

Deepvariant
- 機械学習
- 学習データがあればたぶん最強

どれを使うか？

Barbitoff *et al. BMC Genomics* (2022) 23:155
https://doi.org/10.1186/s12864-022-08365-3

BMC Genomics

RESEARCH ARTICLE Open Access

Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery

Yury A. Barbitoff[1,2,3*], Ruslan Abasov[1,4], Varvara E. Tvorogova[1,3], Andrey S. Glotov[2] and Alexander V. Predeus[1*]

Abstract

Background: Accurate variant detection in the coding regions of the human genome is a key requirement for molecular diagnostics of Mendelian disorders. Efficiency of variant discovery from next-generation sequencing (NGS) data depends on multiple factors, including reproducible coverage biases of NGS methods and the performance of read alignment and variant calling software. Although variant caller benchmarks are published constantly, no previous publications have leveraged the full extent of available gold standard whole-genome (WGS) and whole-exome (WES) sequencing datasets.

Results: In this work, we systematically evaluated the performance of 4 popular short read aligners (Bowtie2, BWA, Isaac, and Novoalign) and 9 novel and well-established variant calling and filtering methods (Clair3, DeepVariant, Octopus, GATK, FreeBayes, and Strelka2) using a set of 14 "gold standard" WES and WGS datasets available from Genome In A Bottle (GIAB) consortium. Additionally, we have indirectly evaluated each pipeline's performance using a set of 6 non-GIAB samples of African and Russian ethnicity. In our benchmark, Bowtie2 performed significantly worse than other aligners, suggesting it should not be used for medical variant calling. When other aligners were considered, the accuracy of variant discovery mostly depended on the variant caller and not the read aligner. Among the tested variant callers, DeepVariant consistently showed the best performance and the highest robustness. Other actively developed tools, such as Clair3, Octopus, and Strelka2, also performed well, although their efficiency had greater dependence on the quality and type of the input data. We have also compared the consistency of variant calls in GIAB and non-GIAB samples. With few important caveats, best-performing tools have shown little evidence of overfitting.

Conclusions: The results show surprisingly large differences in the performance of cutting-edge tools even in high confidence regions of the coding genome. This highlights the importance of regular benchmarking of quickly evolving tools and pipelines. We also discuss the need for a more diverse set of gold standard genomes that would include samples of African, Hispanic, or mixed ancestry. Additionally, there is also a need for better variant caller assessment in the repetitive regions of the coding genome.

どれでも良いが，バイアスの原因など，原理について理解する必要がある

パラメータはどうするのか？

さまざまなパラメータを試して結果を比較する余裕はわれわれには無い

むしろ，リファレンス配列の選定や，バリアントコール後のフィルタリングが重要

いくつかの基本パラメータは重要

例えば，GATKのHeterozygosity（SNVであることの事前確率，ディフォルトはヒトに合わせて0.001）

VCFフォーマット

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HK01
chr1 5666 . T G 1310.06 .
AC=2;AF=1.00;AN=2;DP=32;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=25.36;SOR=0.818 GT:AD:DP:GQ:PL 1/1:0,32:32:96:1324,96,0

0/0，0/1，1/1の3つの遺伝子型，複数のサンプルも扱える

変異がフェージングされている場合は，
0|0，0|1，1|0，1|1 の4種類

テキストファイル
gzip圧縮（.gz）しても大体のソフトウェアは読み込みOK

単独コールかジョイントコールか

一般的に，バリアントコールソフトウェアは変異のあった場所だけを出力する
（オプションで全サイトの遺伝子型を出力することもできることがある）

あるサイトが，読めなかったのか，変異が無かったのかの情報が得られない

そのサイトに変異が無いという情報は，集団解析にとってクリティカルな問題である

ジョイントコールは，複数のサンプルを同時にバリアントコールするので，どのサンプルで変異が無かったのかの情報を知ることが可能

ジョイントコールの例

実際の結果
（ジョイントコール）

サイト	サンプル1	サンプル2
1	0/0	1/1
2	0/1	./.
3	./.	0/1

サンプルごとに
コール

サイト	サンプル1
2	0/1

サイト	サンプル2
1	1/1
3	0/1

マージした結果

サイト	サンプル1	サンプル2
1	0/0	1/1
2	0/1	0/0
3	0/0	0/1

GVCFフォーマット

NC_041754.1 971 . A G,<NON_REF> 445.06 .
DP=11;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_MQandDP=37602,11 GT:AD:DP:GQ:PL:SB
1/1:0,11,0:11:33:459,33,0,459,33,459:0,0,11,0
NC_041754.1 972 . G <NON_REF> . . END=989 GT:DP:GQ:MIN_DP:PL 0/0:11:33:11:0,33,415
NC_041754.1 990 . T <NON_REF> . . END=1040 GT:DP:GQ:MIN_DP:PL 0/0:12:36:12:0,36,442
NC_041754.1 1041 . C <NON_REF> . . END=1046 GT:DP:GQ:MIN_DP:PL 0/0:12:30:12:0,30,450
NC_041754.1 1047 . T <NON_REF> . . END=1056 GT:DP:GQ:MIN_DP:PL 0/0:10:27:10:0,27,405
NC_041754.1 1057 . T <NON_REF> . . END=1072 GT:DP:GQ:MIN_DP:PL 0/0:9:24:8:0,24,313
NC_041754.1 1073 . A <NON_REF> . . END=1078 GT:DP:GQ:MIN_DP:PL 0/0:8:21:8:0,21,315
NC_041754.1 1079 . A <NON_REF> . . END=1092 GT:DP:GQ:MIN_DP:PL 0/0:7:18:6:0,18,222

サンプルごとに変異が無いところもまとめて出力

GATKパイプライン

GVCF形式でサンプルごとにバリアントコーリング

DBImport

データベースにインポート

GenotypeGVCFs

ジョイントコール

1000以上のサンプルに対応可能かつサンプルの追加が容易
小規模なプロジェクトには面倒

フィルタリング
(filtering out)

フィルタリング

次世代シークエンスはエラーが多いので（特にヘテロ接合サイト，繰り返し領域，インデル），コールされた変異から事後確率が高いものを選ぶ必要がある．

ハードフィルタリング

尤度，カバー率，変異を持ったリードのストランドの偏りなどをもとに，怪しいものを除いていく

機械学習などを用いたフィルタリング

正解のデータと比べながら，擬陽性をコントロール
正しいSNVのセットが必要
GATKのVQSR（Variant Quality Score Recalibration）など

ヘテロ接合フィルター

重複などの影響で，すべてのサンプルがヘテロ接合になるサイトが存在する

NC_041773.1 109 . T C 16607.21 PASS AC=65;AF=0.428;AN=152;BaseQRankSum=2.31;DP=1068;ExcessHet=96.0984;FS=1.400;InbreedingCoeff=-0.6910;MLEAC=66;MLEAF=0.434;MQRankSum=1.03;QD=18.94;ReadPosRankSum=-9.360e-01;SOR=0.946 GT:AD:DP:GQ:PGT:PID:PL:PS 0/0:3,0:3:9:.:.:0,9,128 0/1:12,10:22:99:.:.:326,0,343 0/0:7,0:7:0:.:.:0,0,184 0/1:7,5:12:99:.:.:171,0,220 0/1:5,14:19:99:.:.:482,0,117 0/1:7,5:12:99:.:.:147,0,185 0/1:4,7:11:99:.:.:246,0,109 0/1:11,11:22:99:.:.:373,0,338 0/1:3,10:13:56:.:.:353,0,56 0/1:5,7:12:99:.:.:227,0,136 0/1:10,6:16:99:.:.:165,0,283 0/1:6,3:10:66:.:.:66,0,188 0/1:8,14:22:99:.:.:472,0,175 0/1:11,8:19:99:.:.:230,0,294 0/1:5,4:9:99:.:.:130,0,109 0/1:4,7:11:99:.:.:261,0,101

ExcessHetタグを使って，bcftoolsなどでフィルター可能
VcftoolsでもHardy-Weinberg平衡からの逸脱（ヘテロ過剰）を基準にフィルター可能

マッパビリティ（mappability）フィルター

変異の機能などを推定する場合には，過度なフィルタリングは大事なものを見逃す可能性を上げてしまうが，集団解析などの定量的な評価においては余計なエラーを減らすことが重要である

例えば，ヘテロ接合度の評価

マッパビリティフィルターは繰り返し配列によるミスマッピングの結果を除くためのフィルタリング

MAPQによる評価より更に保守的にフィルタリング

Genmap mappability

GenMap computes the uniqueness of k-mers for each position in the genome while allowing for up to e mismatches. More formally, the uniqueness or (k,e)-mappability can be described for every position as the reciprocal value of how often each k-mer occurs approximately in the genome, i.e., with up to e mismatches. Hence, a mappability value of 1 at position i indicates that the k-mer in the sequence at position i occurs only once in the sequence with up to e errors. A low mappability value indicates that this k-mer belongs to a repetitive region. GenMap can be applied to single or multiple genomes and helps finding regions that are unique or shared by many or all genomes.

Below you can see the (4,1)-mappability and frequency `M` and `F` of the nucleotide sequence `T = ATCTAGGCTAATCTA`. The mappability value `M[1] = 0.33` means that the 4-mer starting at position 1 `T[1..3] = TCTA` occurs three times in the sequence with up to one mismatch: at positions `1 (TCTA)`, `6 (GCTA)` and `11 (TCTA)`.

i	0	**1**	2	3	4	5	**6**	7	8	9	10	**11**	12	13	14
$T[i]$	A	**T**	**C**	**T**	**A**	G	G	C	T	A	A	T	C	T	A
$M[i]$	.5	**.33**	.5	1.0	1.0	1.0	.33	.5	1.0	1.0	.5	.33	.0	.0	.0
$F[i]$	2	**3**	2	1	1	1	3	2	1	1	2	3	0	0	0

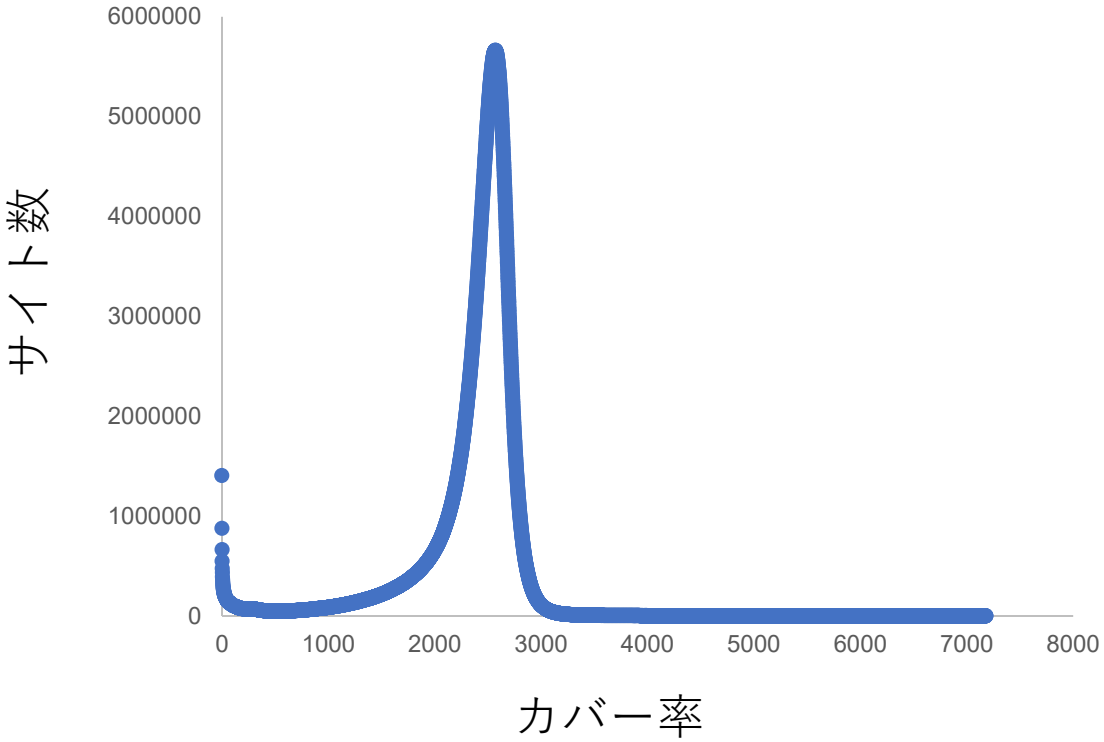
哺乳類のゲノムだと3割程度がマスク（フィルターによって除かれる領域）される

*maskファイルが「解析に含まれる領域」を示している場合もあるので注意

アクセシビリティ（accessibility）フィルター

きちんと読めている領域だけを抜き出す

さまざまな方法が考えられるが，例えば全サンプルを合わせてカバー率が一定範囲内のサイト



アクセシビリティ

Samtools coverage　などでカバー率の平均値を算出可能

Accessibilityがあるところは「読めている」という前提なので，vcfファイルのマージの問題も解決

一般的な哺乳類ゲノムでは，90%くらいのゲノム領域はaccessibleである

アクセシビリティの重要性

一般的なNGS解析は変異を同定することだが，進化解析では「変化がなかったサイト」の情報が重要

突然変異率を考えて年代推定等をすることが多いため，分母の情報が必要になってくる

アクセシビリティがあるサイトは読めているという前提に立てば，解析したサイト数，つまり分母は容易に推定が可能

全サイトをコールすることでも対応可能だが，必要な労力が多い（ファイルサイズも大きくなる）

その他のQC法

主成分分析などのpost-processing

ミトコンドリア・クロロプラスト配列のアセンブル

など，異なった視点からデータをチェックすることも重要

まとめ

大量配列解析実験（次世代シークエンサー）

↓

fastq ファイル（エラー率つきの塩基配列）

↓ 参照ゲノム配列へのマッピング（bwa）

sam/bam ファイル

↓ 変異検出（GATK，varScan）

vcf ファイル（変異データ）

↓ フォーマット変換ソフトウェア（PLINK）

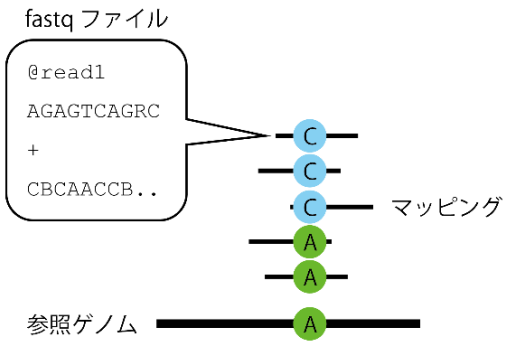
カスタムフォーマットの変異ファイル
多検体変異ファイルへの統合

↓ eigenstrat, STRUCTURE, treemix など

さまざまな解析プログラム

↓ R, python など

グラフの描画や結果の解釈



変異ファイル（vcf ファイル）

```
chr1 1010 A T 176.7 . AC=1;AF=0.500 ...
chr1 2014 G A 157.7 . AC=1;AF=1.000 ...
chr1 7888 C T 423.3 . AC=1;AF=0.500 ...
...
```

主成分分析など

すべてのプロセスでクオリティチェックは重要
進化解析には普通の変異解析では使わない情報も必要になってくる
ソフトウェアの使い方よりも「原理・原則」を理解することが大切