ゲノム勉強会

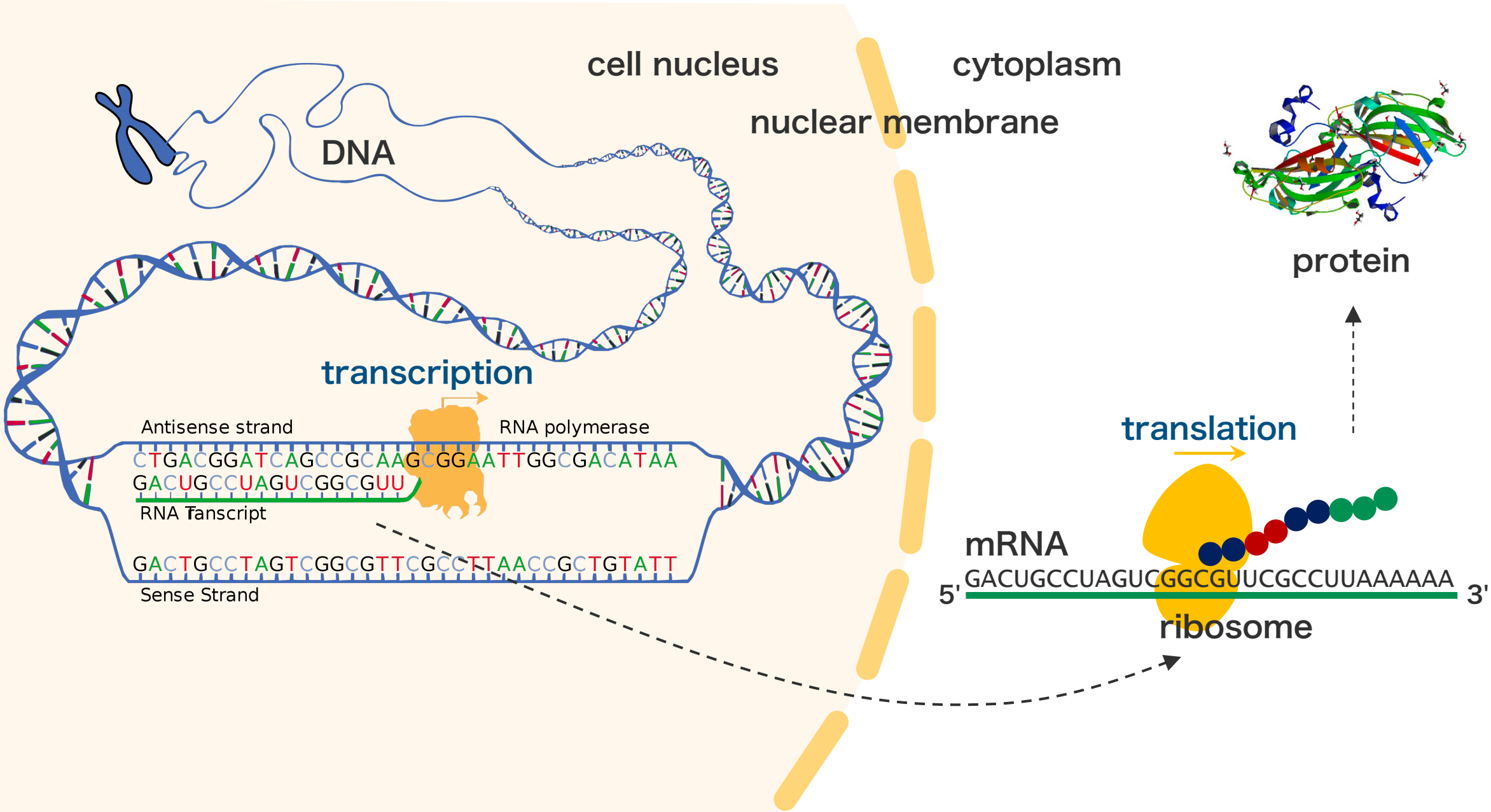
2020-08-04 – 2020-08-07

孫 建強

RNA-Seq

- ○ Introduction
- ○ Quality control
- ○ Mapping and quantification
- ○ Assembly

Central dogma

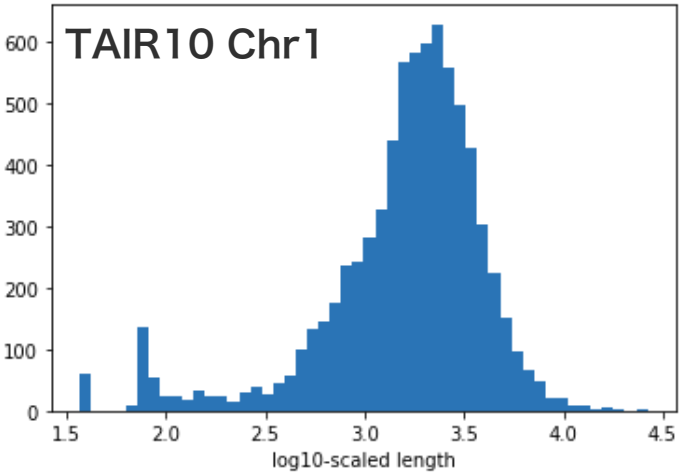


RNA-Seq

sequecing

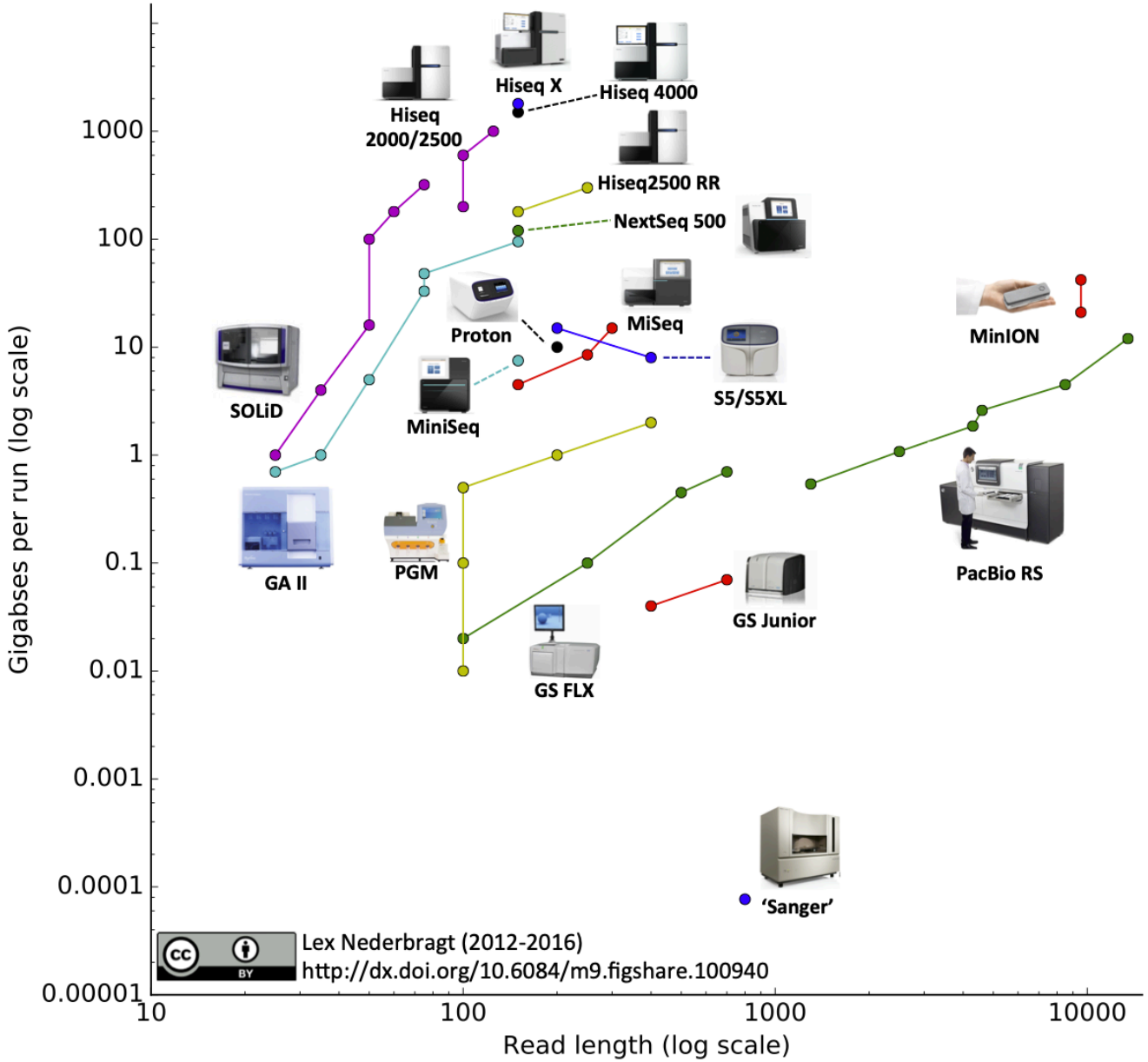
ATGATGCATTCAGCTACAGC
ATGCACATGA
ATGATGCATTCAGCTACAGC
ATGCAGTA
ATGCACATGA
ATGTACGTAGCTAGCTAGCTACGTAG
ATGTACGTAGCTAGCTAGCTACGTAG
ATGCACATGA
ATGCACATGA
ATGTACGTAGCTAGCTAGCTACGTAG

FT
LFY
FT1
STM
LFY
EIN3
EIN3
LFY
LFY
EIN3



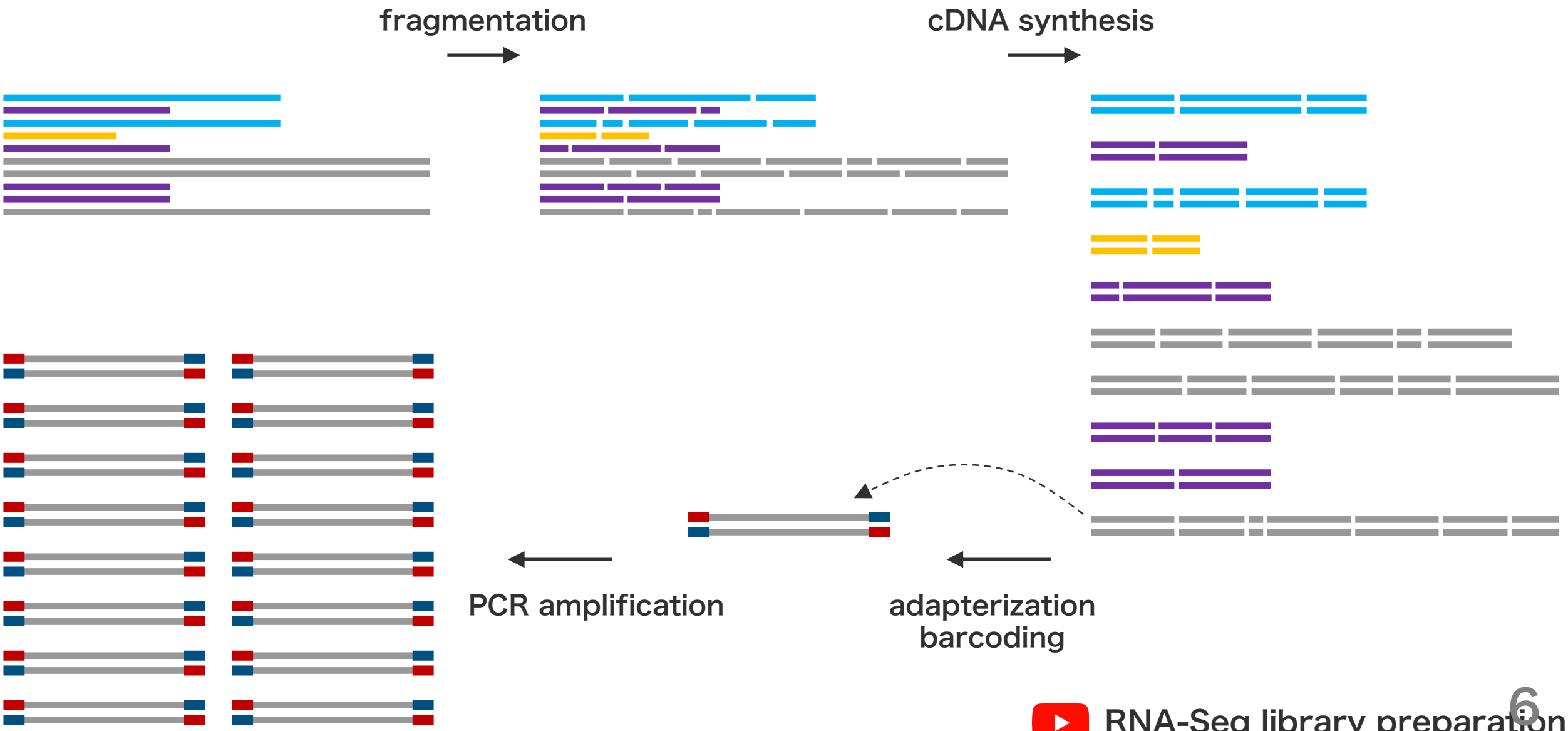
gene length (intron included)

RNA-Seq



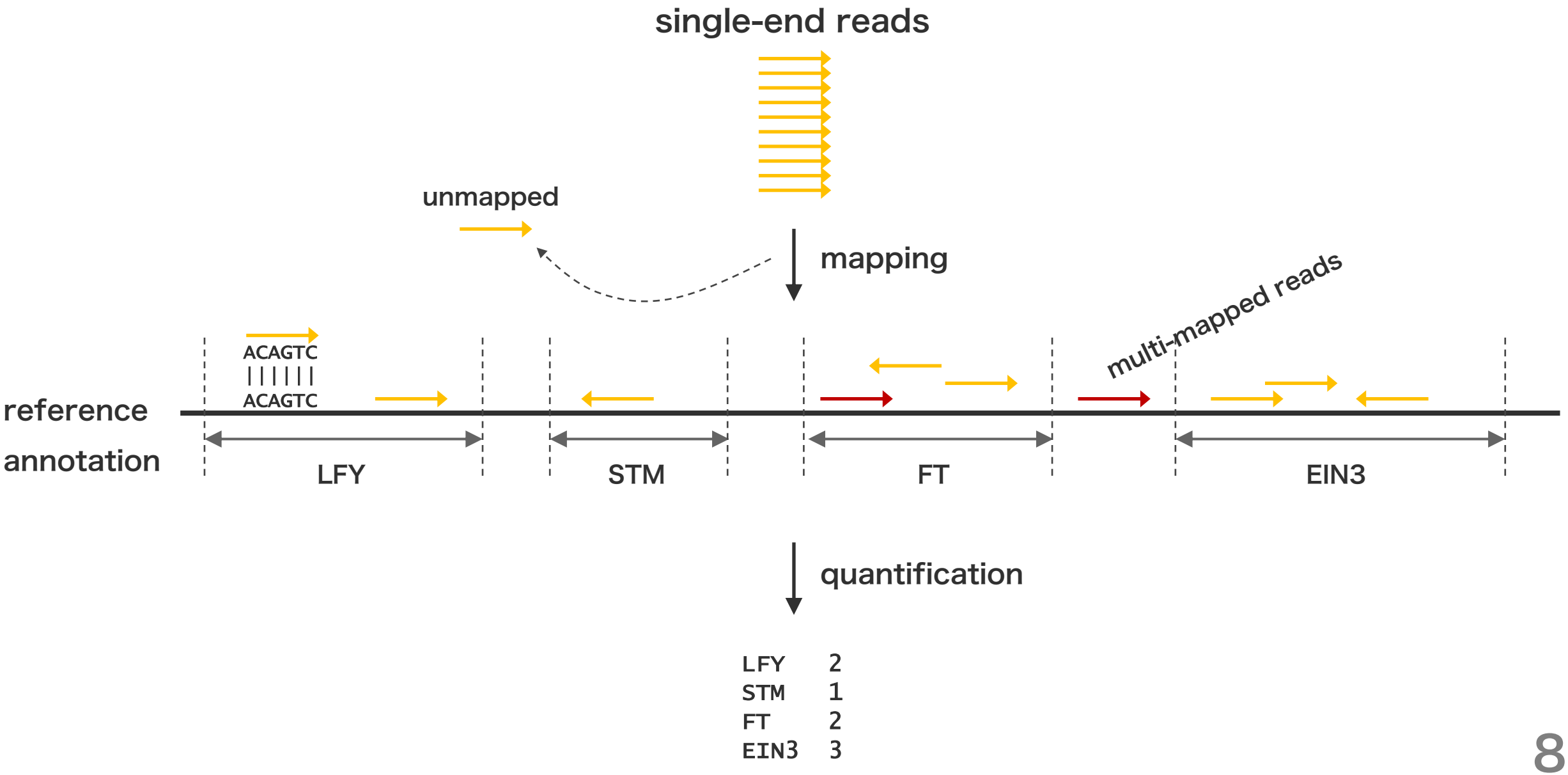
- Illumina HiSeq / NextSeq / MiSeq
- IonTorrent
- PacBio
- Oxford nanopore

RNA-Seq library preparation

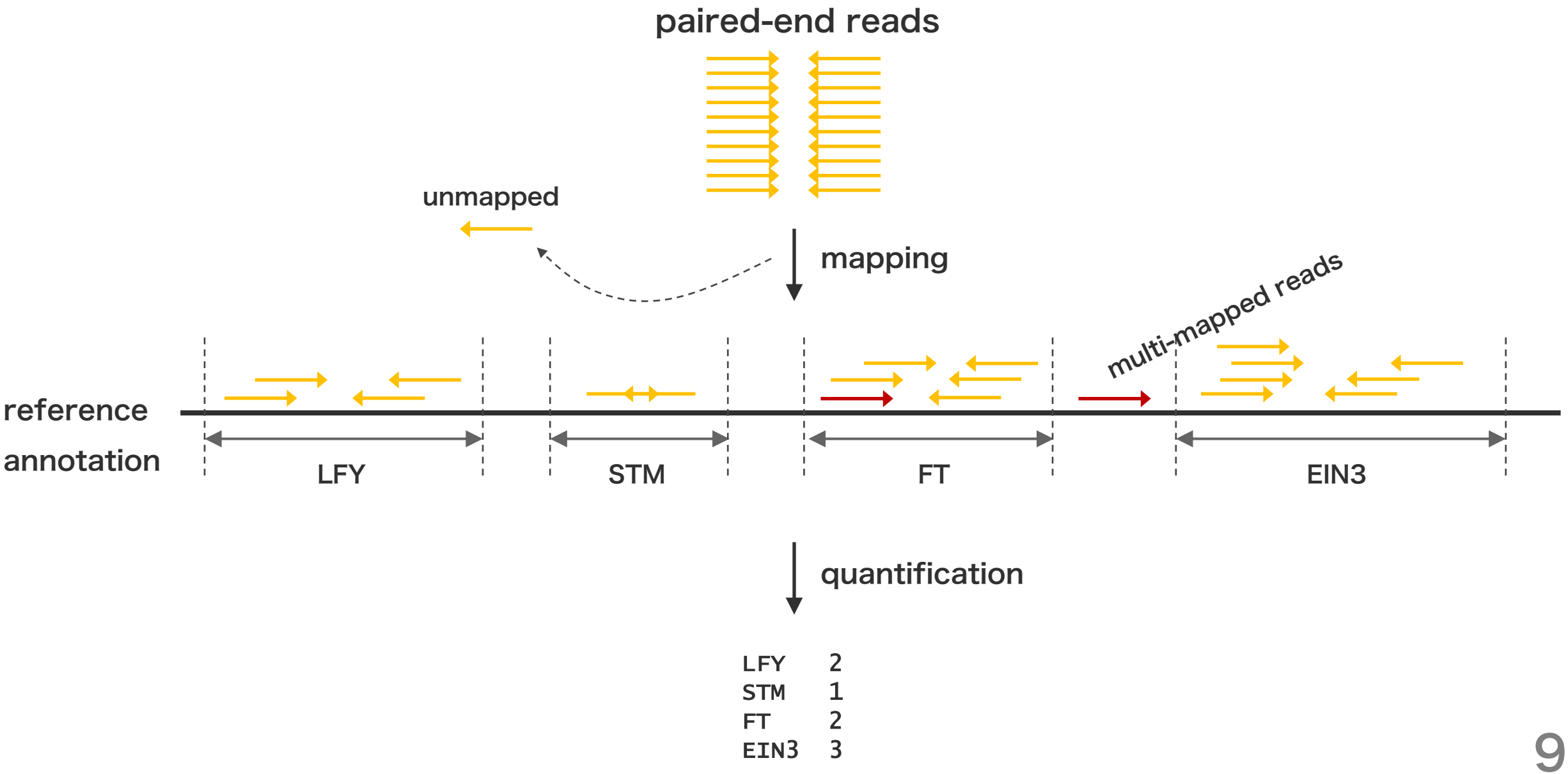


RNA-Seq library preparation

RNA-Seq / single-end approach



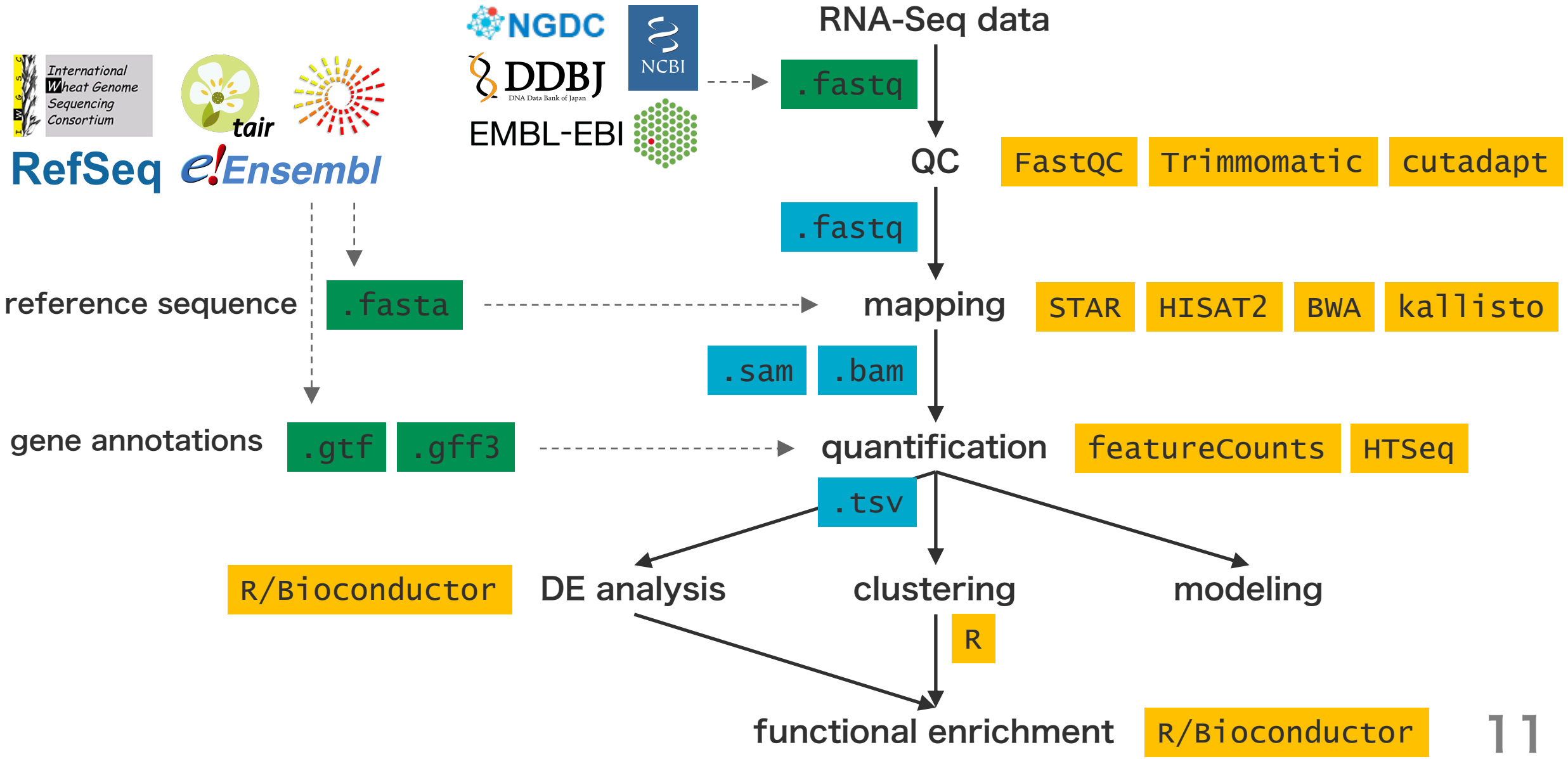
RNA-Seq / paired-end approach



RNA-Seq

- ○ Introduction
- ○ Quality control
- ○ Mapping and quantification
- ○ Assembly

RNA-Seq


International Wheat Genome Sequencing Consortium
tair
RefSeq
Ensembl
NGDC
DDBJ
DNA Data Bank of Japan
NCBI
EMBL-EBI
RNA-Seq data
.fastq
QC
FastQC
Trimmomatic
cutadapt
.fastq
reference sequence
.fasta
mapping
STAR
HISAT2
BWA
kallisto
.sam
.bam
gene annotations
.gtf
.gff3
quantification
featureCounts
HTSeq
.tsv
R/Bioconductor
DE analysis
clustering
modeling
R
functional enrichment
R/Bioconductor

FASTQ

read sequence →

Quality score →

```
@HWI:2:2102:31421:12730
TAAGGAGATCAAGAACGGGCGACTTGCGCTGTTGGCGTT
+
@@1BDDDDB??D:CGFIFEBGGABGFGFFFBABBBB?<>
@HWI:2:1102:21126:37630
GCCCCAGCCCCAGTCCTAGCCCCAGCTCCAGTTCCGGCT
+
CCCFFFFFHHGHHJJJJJJIJJIJJJJJJIJJHIDDD@CB
@HWI:2:1102:17396:37651
AAAAAAAGGAACAAAAGAGGACAAGACCCAATCCACAAT
+
@@CFDFDFFBHHAGGFHIIGCHJJIGHIEGIJJJEECD:
@HWI:2:1102:17975:37502
TTGATGTTAGGCAAAGTCAAGAAGTTCTTGGTGATGTGA
+
CCCFFFFFHHHHHJJJIJJJJJJJJIJIEEDDDDDDDDDB
```

FASTQ

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

```
@HWI:2:2102:31421:12730
TAAGGAGATCAAGAACGGGCGACTTGCGCTGTTGGCGTT
+
@@1BDDDDB??D:CGFIFEBGGABGFGFFFBABBBB?<>
```

Quality score = (ASCII code) - 33

= 31

error rate p

$$p = 10^{\frac{Q}{-10}} = 10^{\frac{64-33}{-10}} = 7.9\times10^{-4}$$

FASTQ

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

```
@HWI:2:2102:31421:12730
TAAGGAGATCAAGAACGGGCGACTTGCGCTGTTGGCGTT
+
@@1BDDDDB??D:CGFIFEBGGABGFGFFFBABBBB?<>
```

Quality score = (ASCII code) - 33

= 16

error rate p

$$p = 10^{\frac{Q}{-10}} = 10^{\frac{49-33}{-10}} = 2.5\times10^{-2}$$

FASTQ

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

```
@HWI:2:2102:31421:12730
TAAGGAGATCAAGAACGGGCGACTTGCGCTGTTGGCGTT
+
@@1BDDDDB??D:CGFIFEBGGABGFGFFFBABBBB?<>
```

Quality score = (ASCII code) - 33

= 30

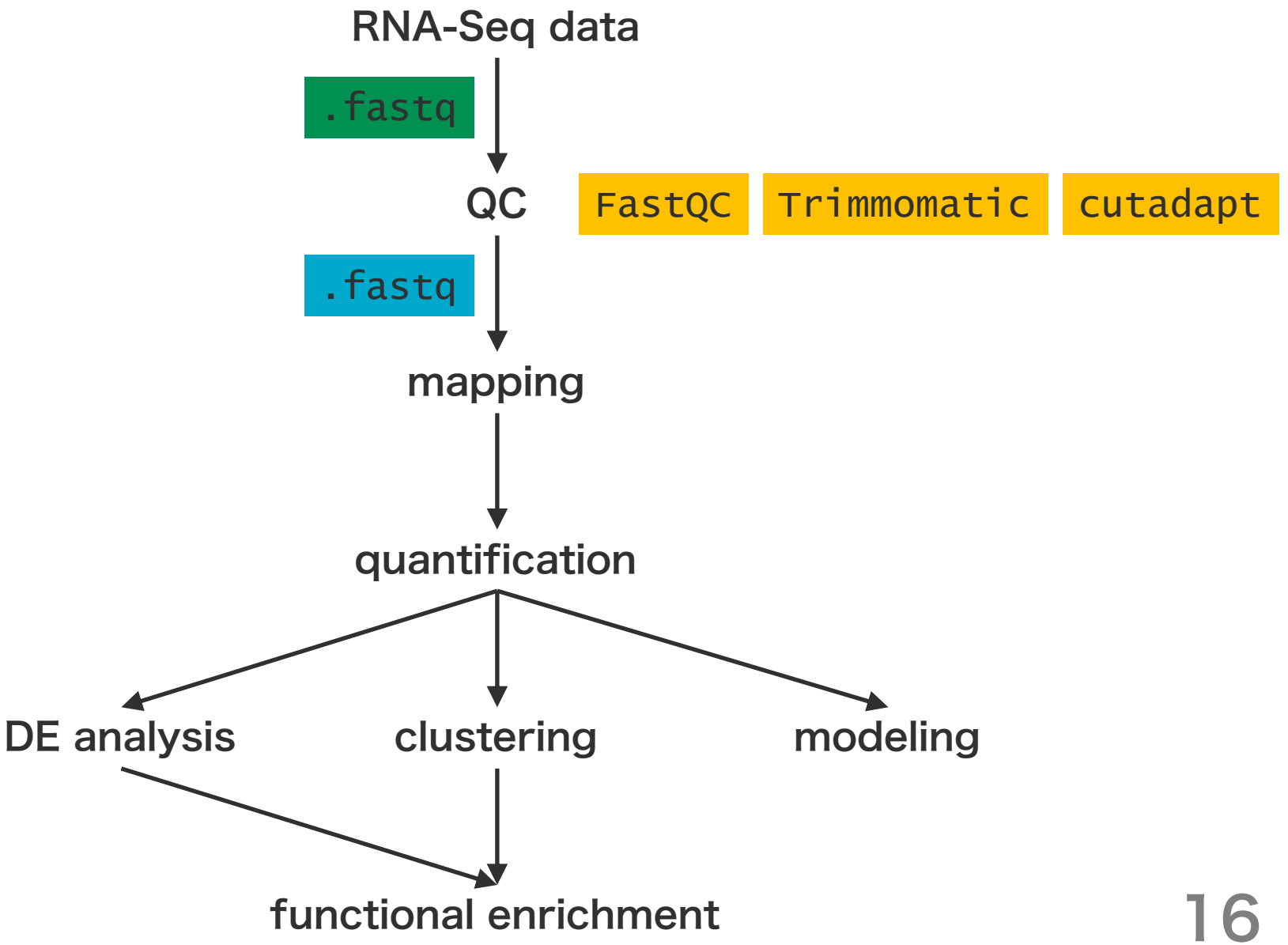
error rate p

$$p = 10^{\frac{Q}{-10}} = 10^{\frac{63-33}{-10}} = 1.0\times10^{-3}$$

Quality control (QC)

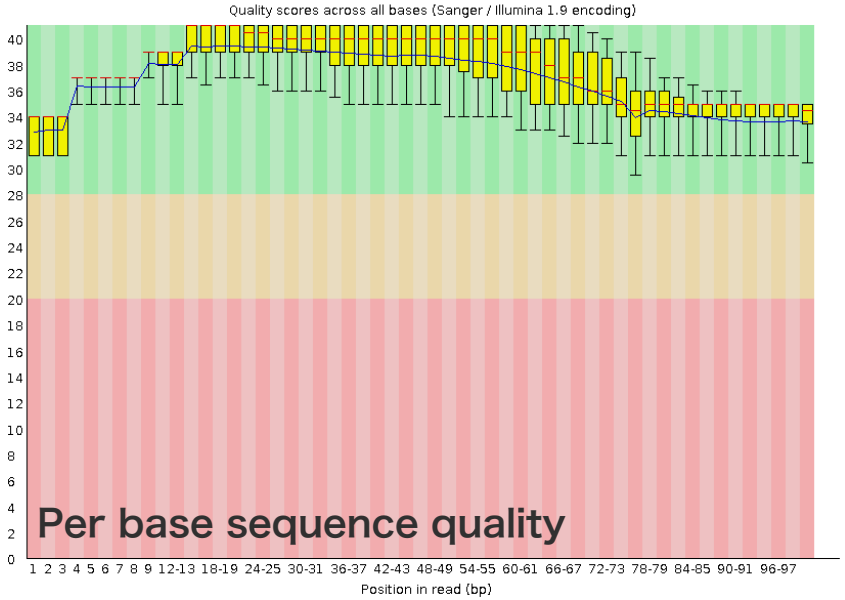
Qquality Control (QC)

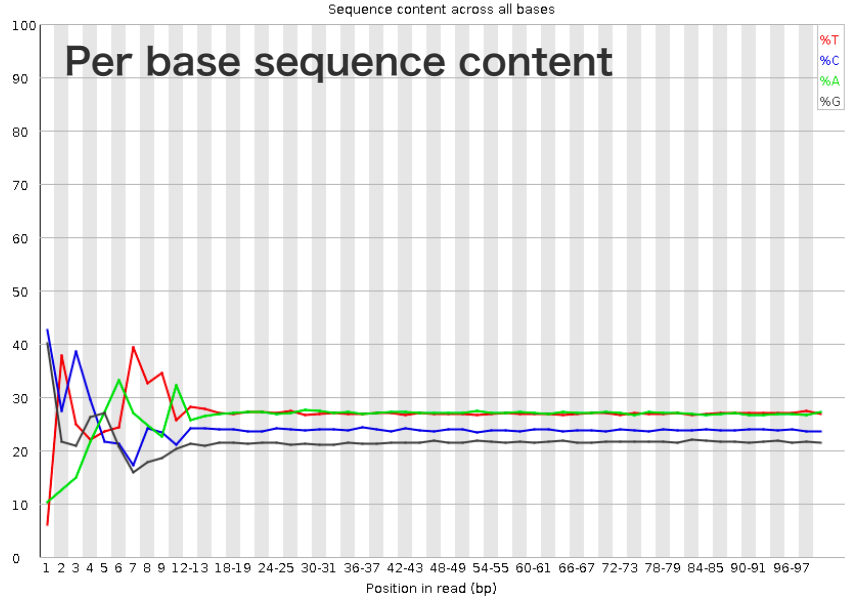
- trim low quality parts
- trim poly-A/T
- trim adapter sequences
- discard short reads

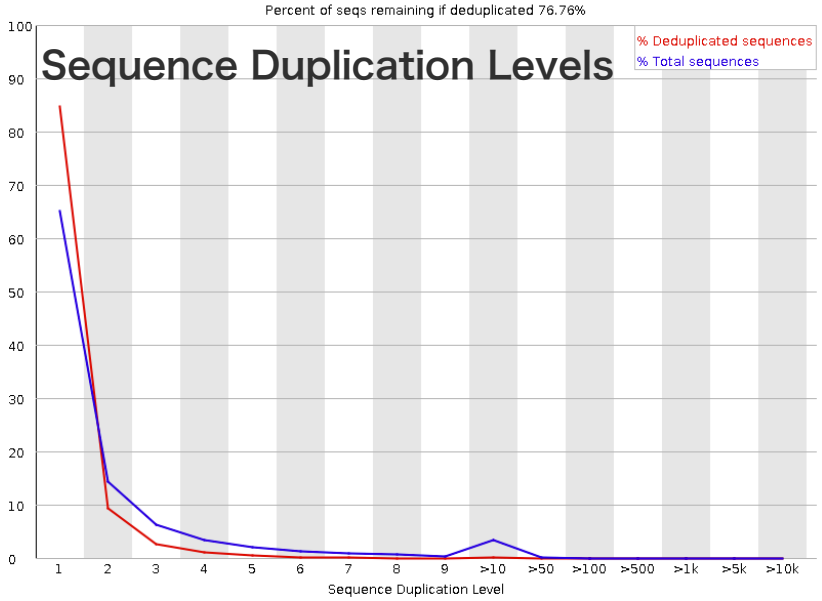


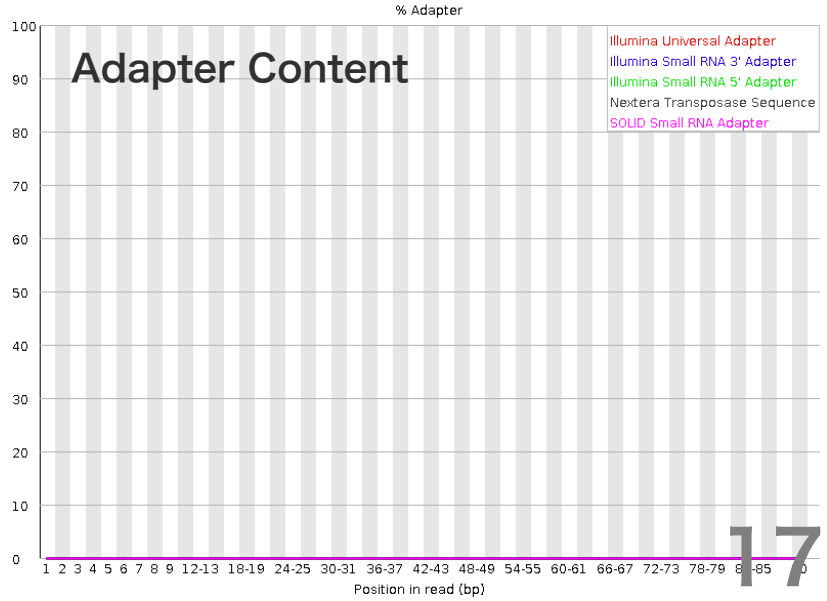
QC report

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content





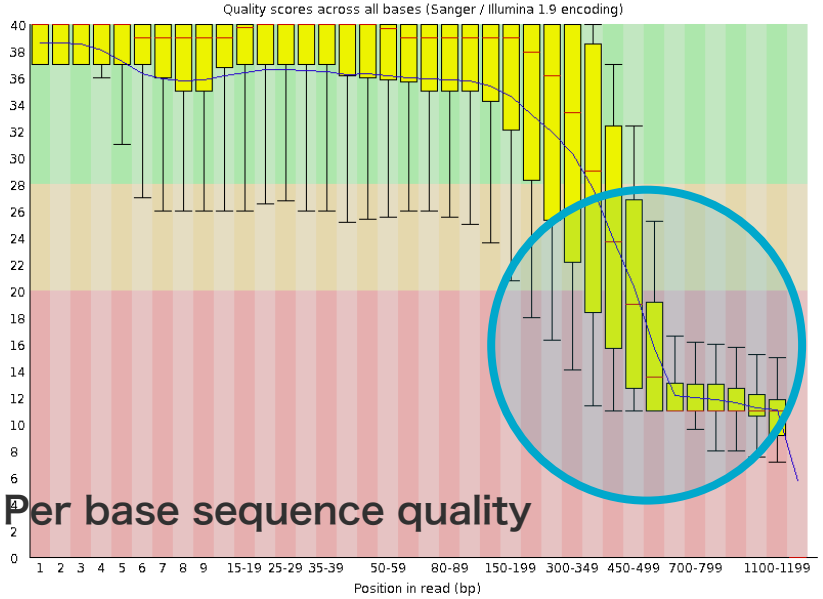


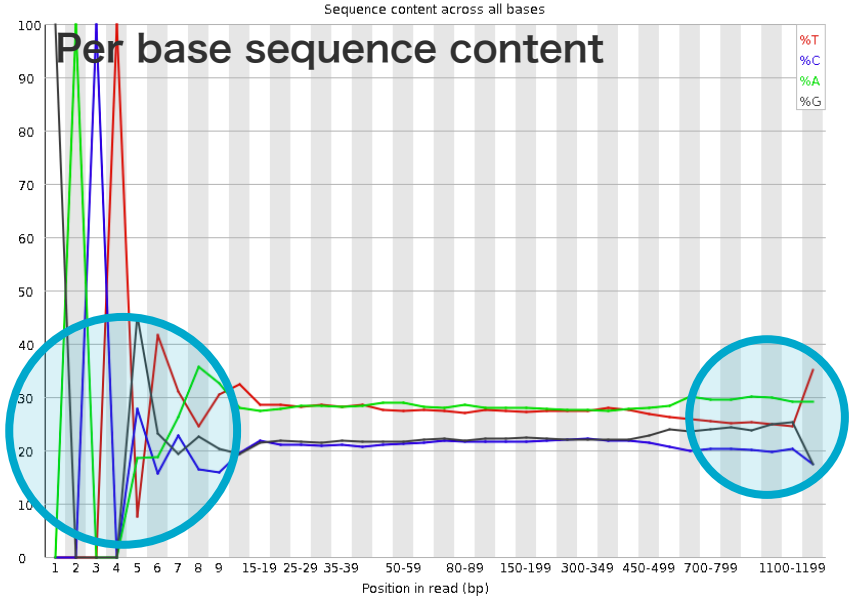


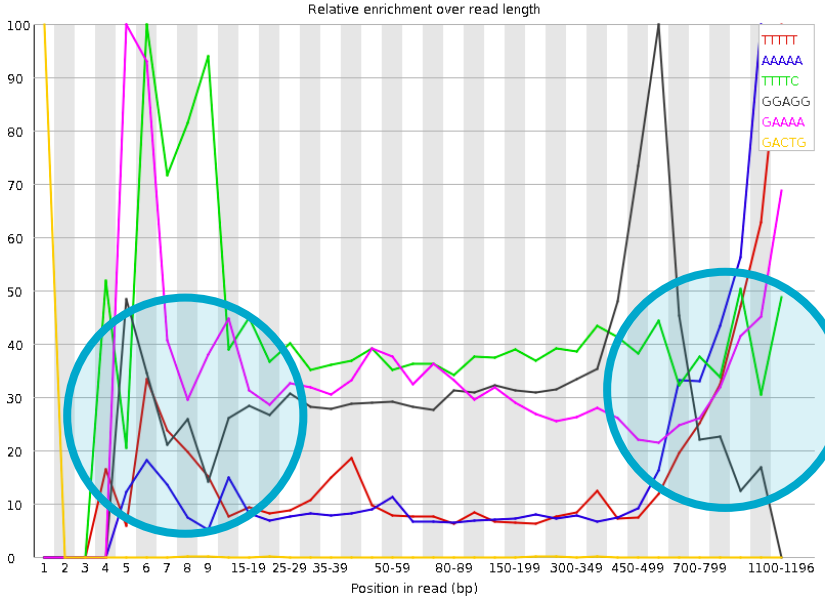
QC report was generated by FastQC

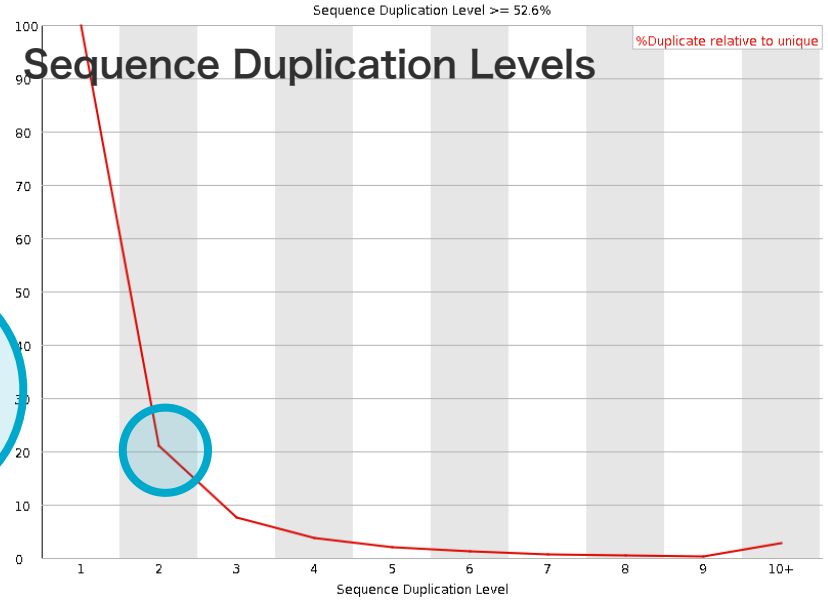
QC report / NG

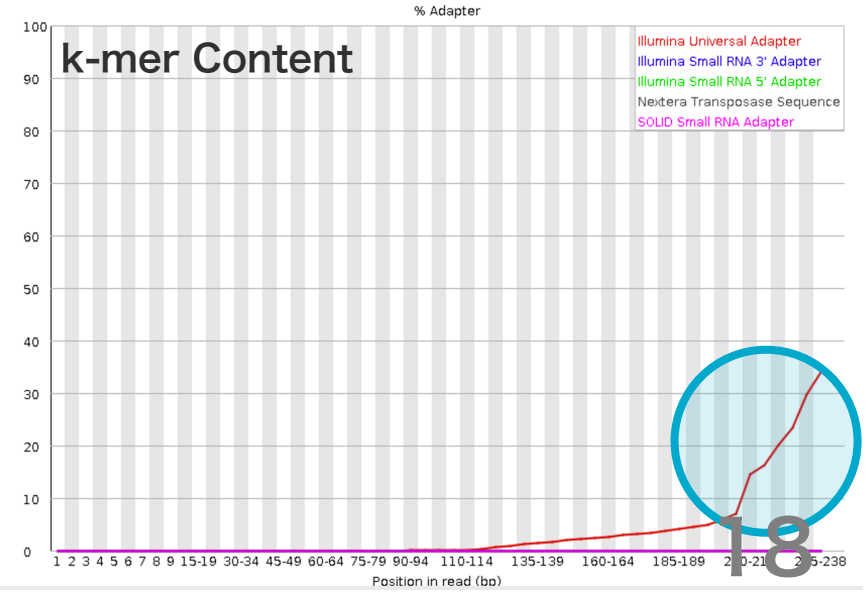
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content







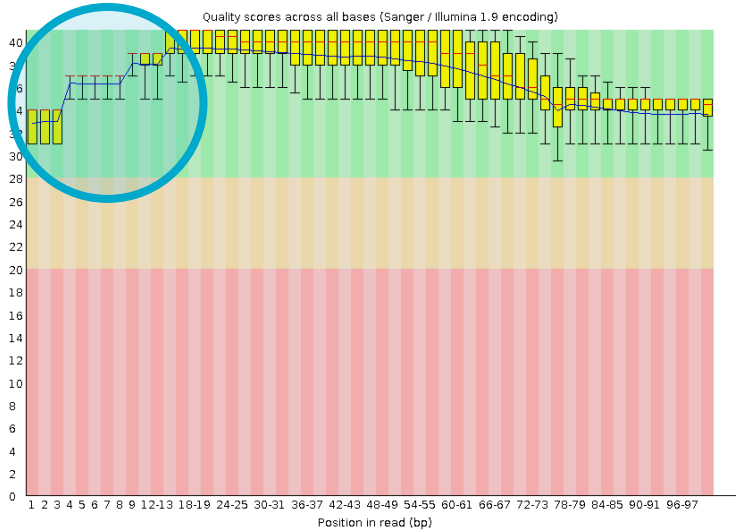




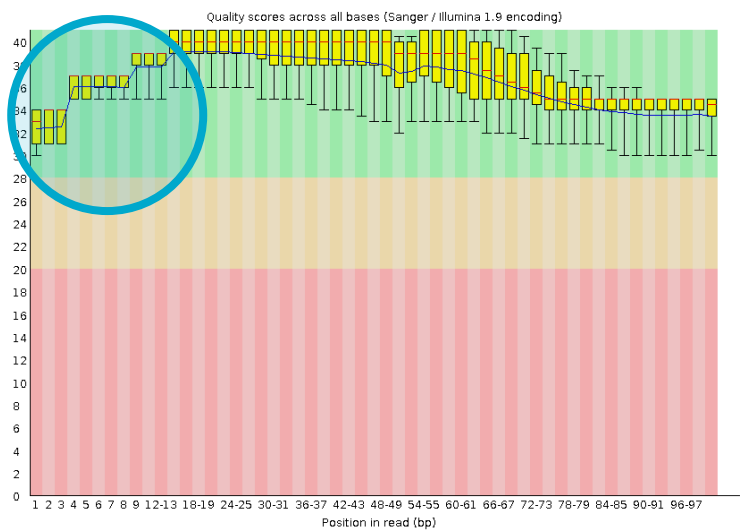
QC report / before QC

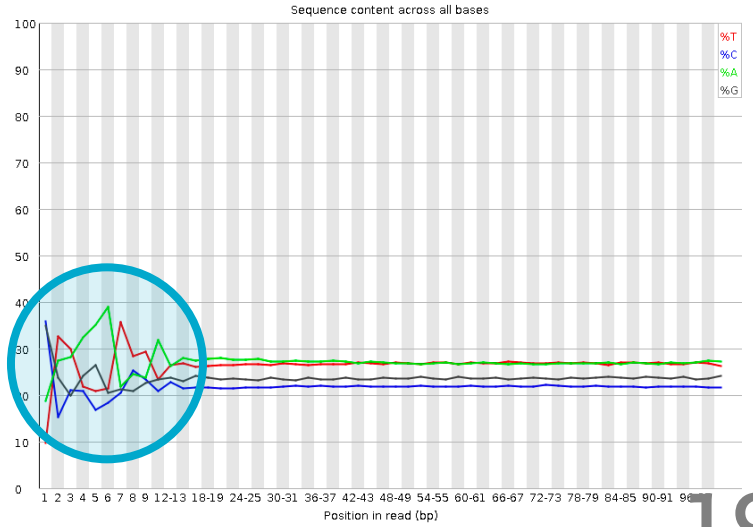
paired-end reads

forward reads



reversed reads

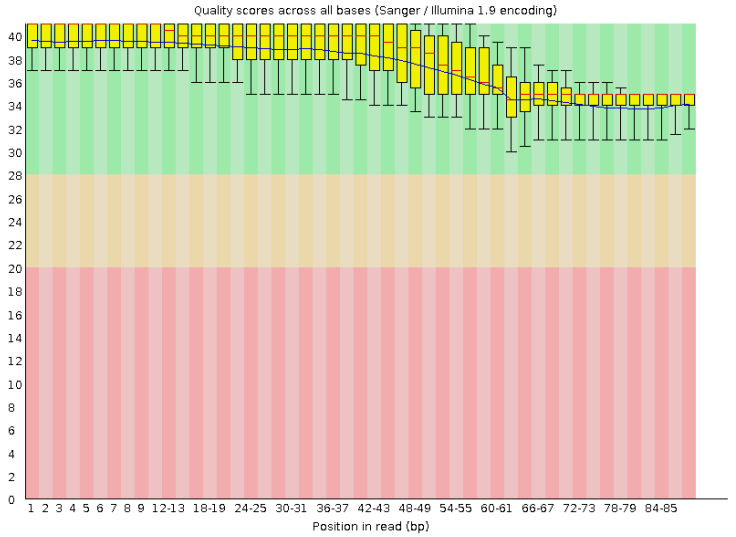


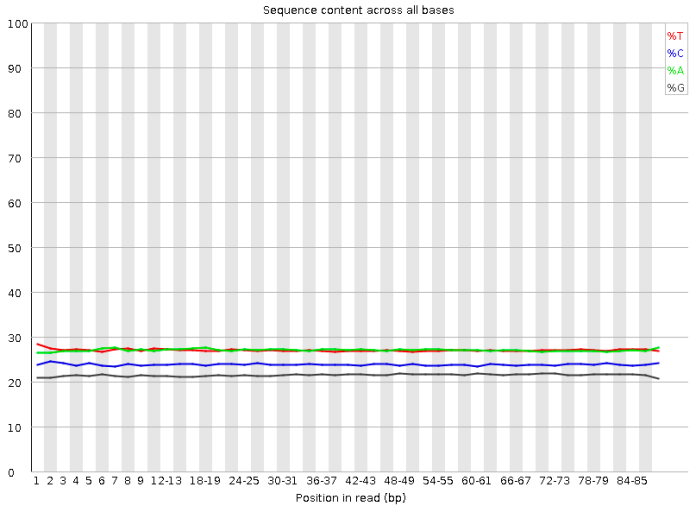


QC report / after QC

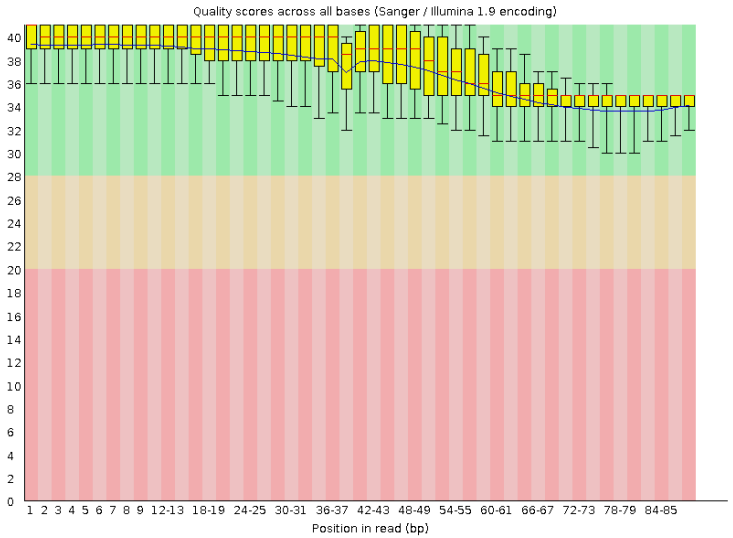
paired-end reads

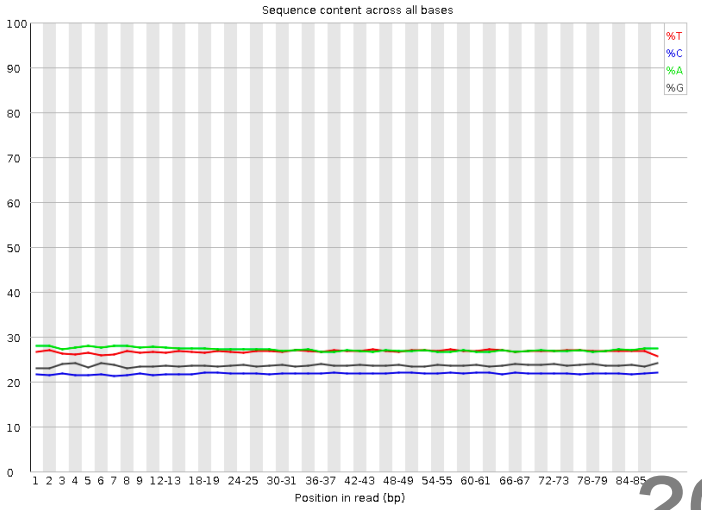
forward reads





reversed reads





QC / Trimmomatic

Cut the read to a specified length

`HEADCROP:20`

Cut the specified number of bases from the start of the read

`CROP:120`

`LEADING:30`

Cut bases off the start of a read, if below a threshold quality

`SLIDINGWINDOW:20:30`

Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.

`TRAILING:30`

Cut bases off the end of a read, if below a threshold quality

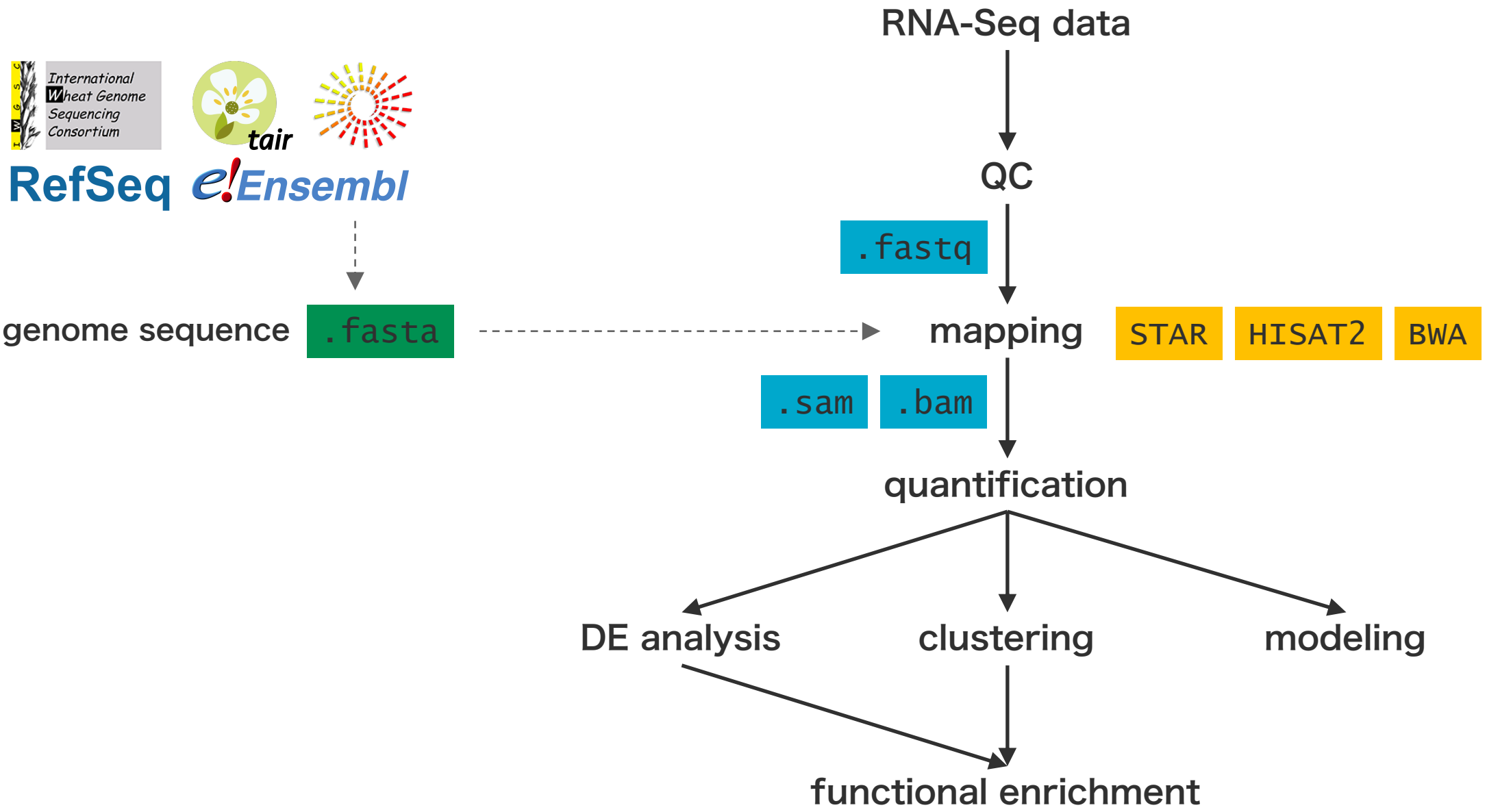
`MINLEN:20` Drop the read if it is below a specified length

http://www.usadellab.org/cms/?page=trimmomatic

RNA-Seq

- ○ Introduction
- ○ Quality control
- ○ Mapping and quantification
- ○ Assembly

Mapping


International Wheat Genome Sequencing Consortium
tair
RefSeq
Ensembl
genome sequence
.fasta
RNA-Seq data
QC
.fastq
mapping
STAR
HISAT2
BWA
.sam
.bam
quantification
DE analysis
clustering
modeling
functional enrichment

Genome sequence

Ensembl

http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index

Gene annotation

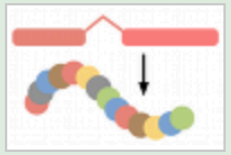
What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins - FASTA - GFF3

Update your old Ensembl IDs

Example gene



Example transcript

International Wheat Genome Sequencing Consortium

tair

RefSeq Ensembl

Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz

Arabidopsis_thaliana.TAIR10.47.gff3.gz

Mapping

reads

TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT

mapping

reference GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

Mapping

reads

TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
CCTTAGTAAGGGTCGCGC

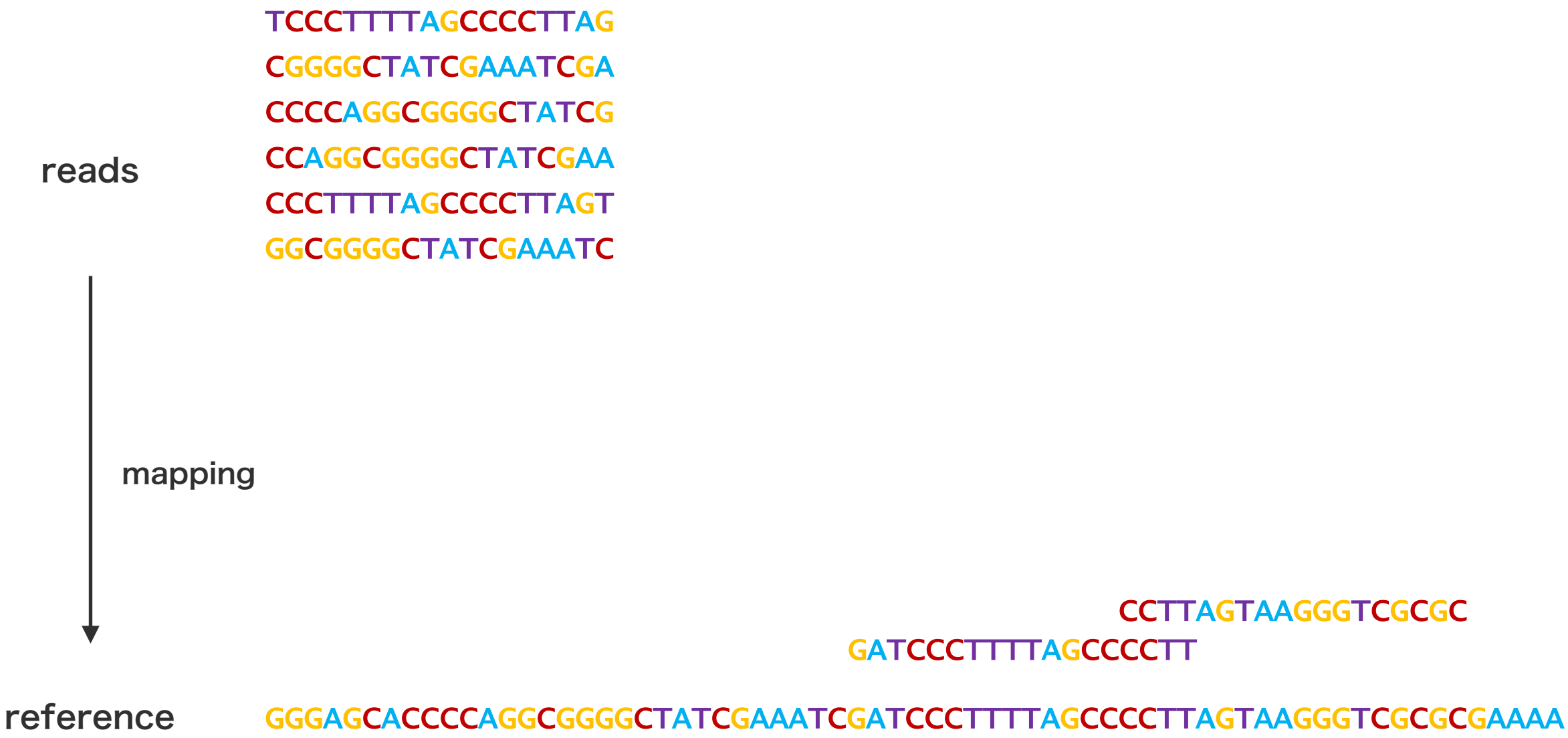
mapping

GATCCCTTTTAGCCCCTT

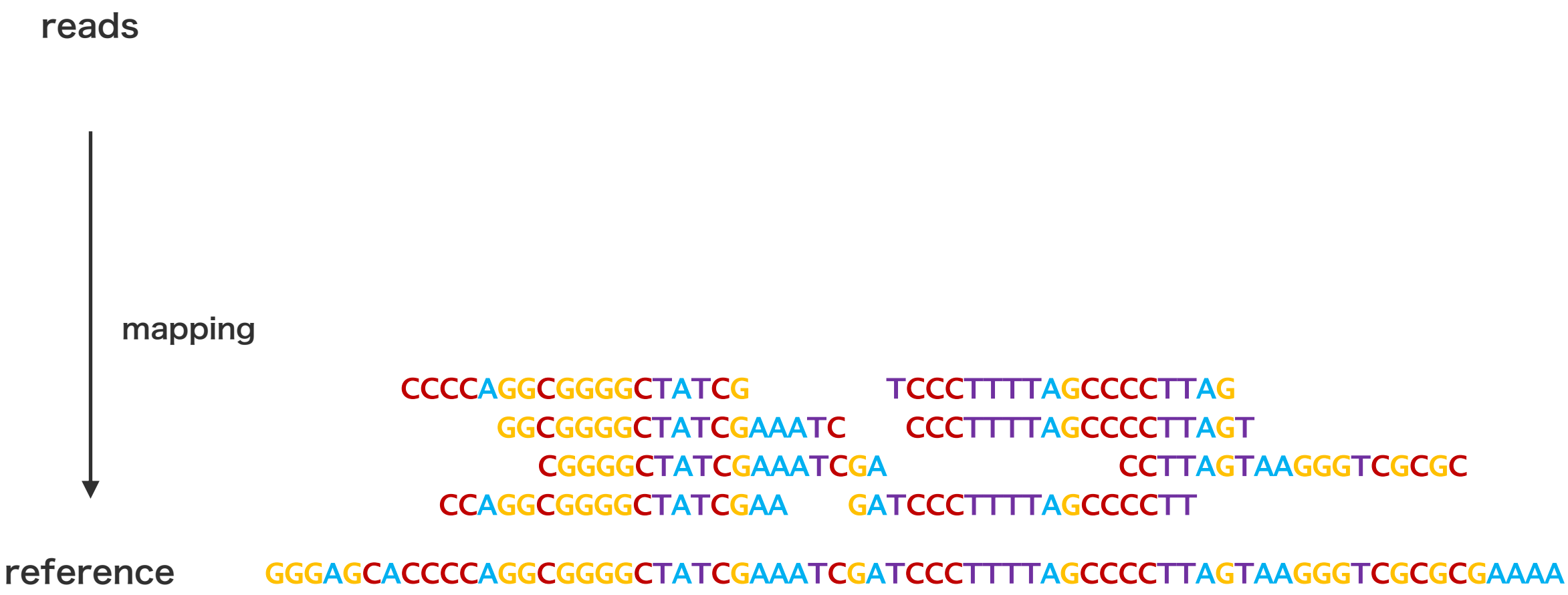
reference

GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

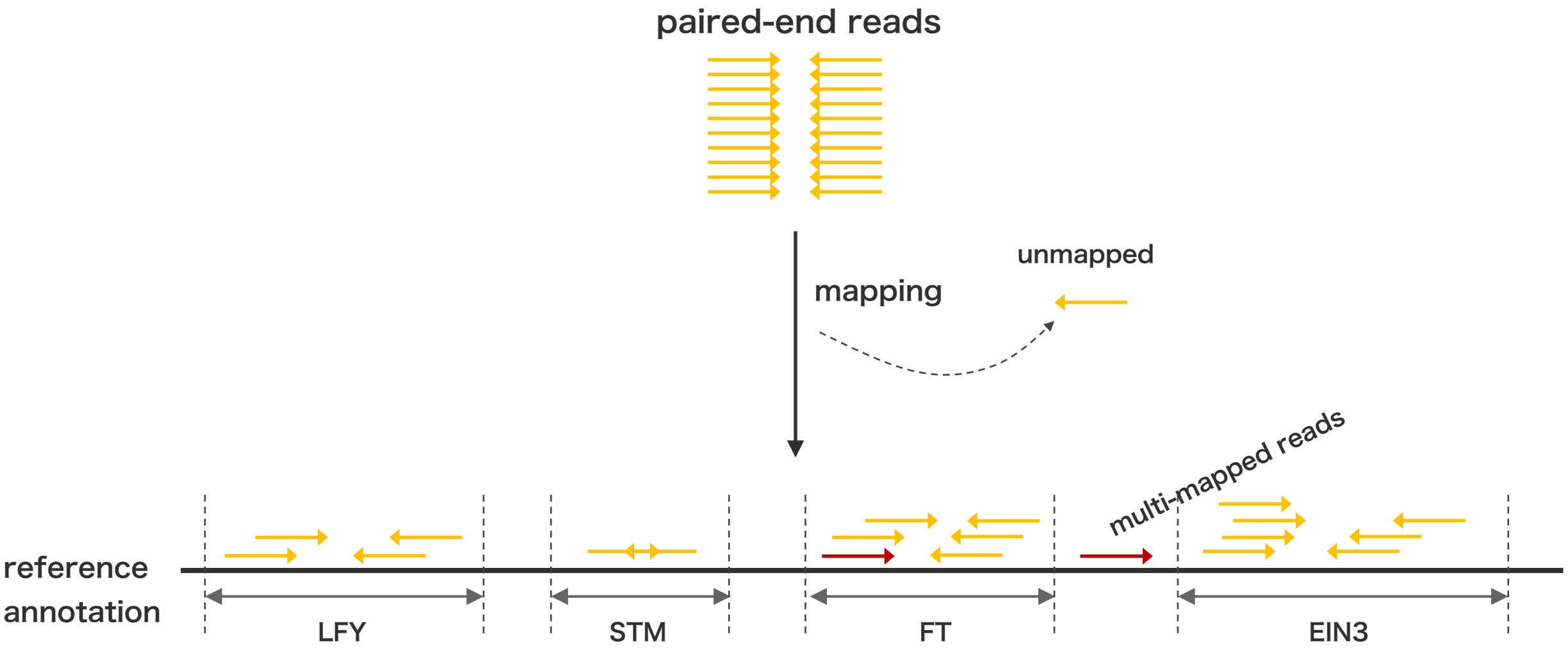
Mapping


TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
reads
mapping
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT
reference
GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

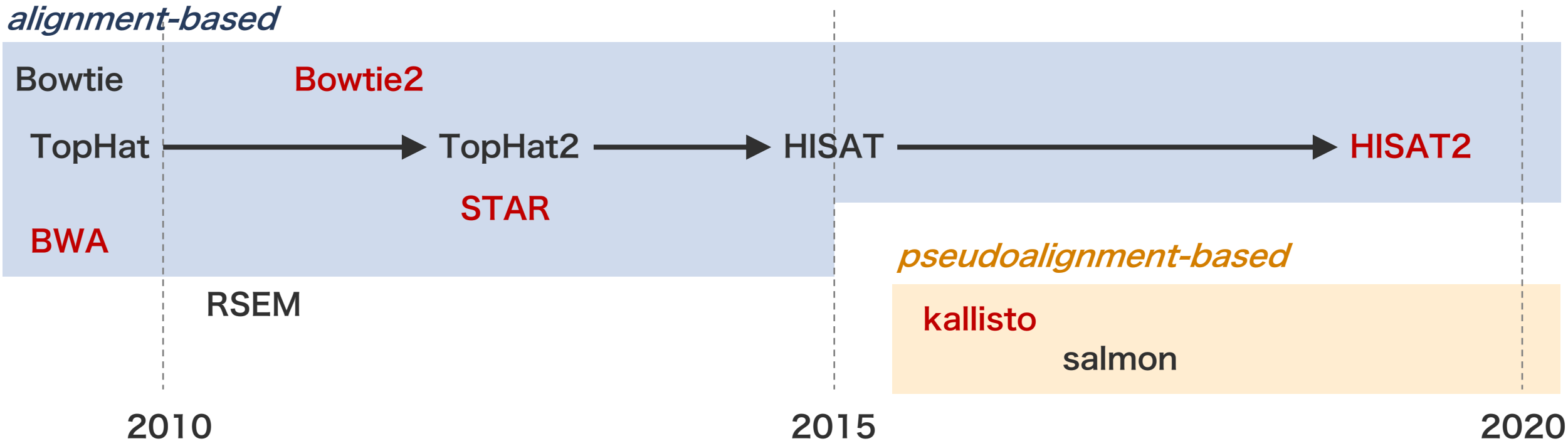
Mapping


reads
mapping
CCCCAGGCGGGGCTATCG
TCCCTTTTAGCCCCTTAG
GGCGGGGCTATCGAAATC
CCCTTTTAGCCCCTTAGT
CGGGGCTATCGAAATCGA
CCTTAGTAAGGGTCGCGC
CCAGGCGGGGCTATCGAA
GATCCCTTTTAGCCCCTT
reference
GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

Mapping


paired-end reads
mapping
unmapped
multi-mapped reads
reference
annotation
LFY
STM
FT
EIN3

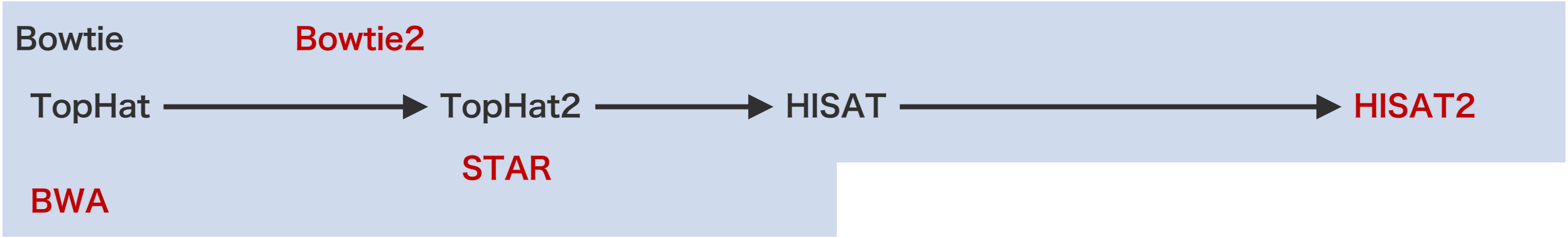
Mapping software



(Bowtie) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10:R25, 2009.
Fast gapped-read alignment with Bowtie 2. Nature Methods, 9:357–359, 2012.
TopHat: Discovering Splice Junctions With RNA-Seq. Bioinformatics, 25(9):1105-11, 2009.
TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 14: R36, 2013.
HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12:357–360, 2015.
Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37:907–915, 2019.
Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60, 2009.
Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, 26:589-595, 2010.
STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29:1, 2013.
RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12:323, 2011.
(kallisto) Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology, 34:525–527, 2016.
Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14,:417–419, 2017.

Mapping software

alignment-based

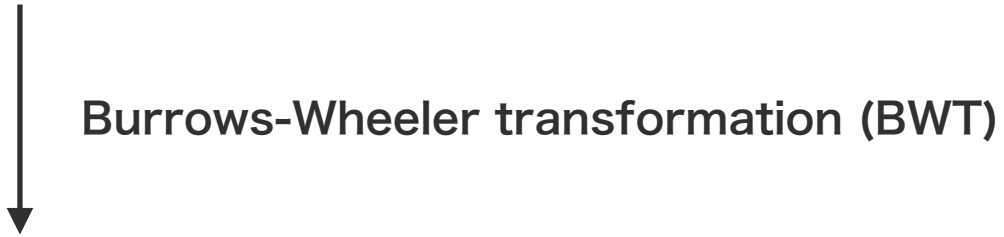


Alignment-based software align reads based on

- text search by FM-index using Burrows-Wheeler transformation text
- local alignment by Smith-Waterman algorithm

BWT

The ability to resolve the inherent complexity of gene families related to important agronomic traits demonstr ates the impact of IWGSC RefSeq v1.0 on dissecting quan titative traits genetically and implementing modern bre eding strategies for future wheat improvement.



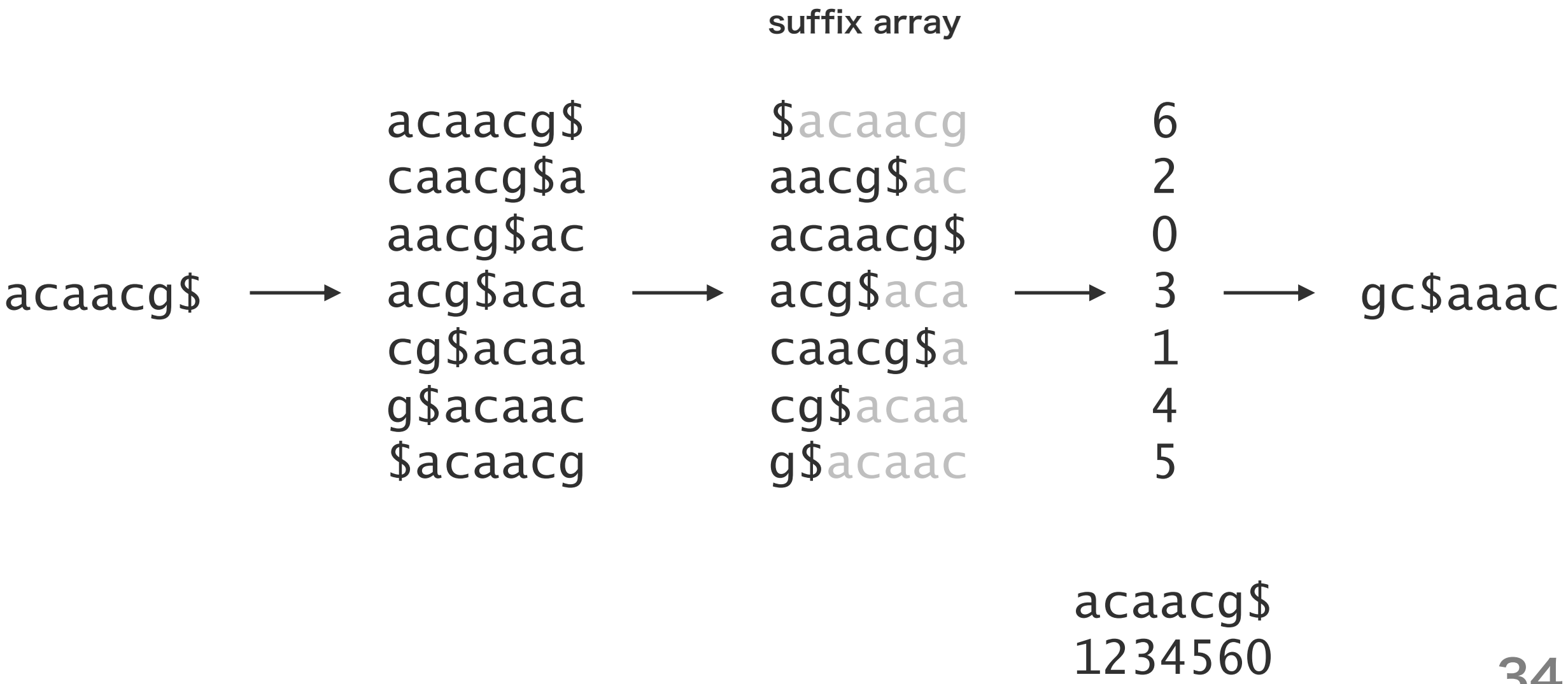
fCetyntsnesrfsedotegtyOgsogsedyceqe1t.vSW Gf.I p rrcf
tuelrrta ii aene oe hnhhvvrhsterRtrvldggrmmShdiitrnlooe
  nnn eaTttwnmtglmb ttd dtaaxlteppiiaoleeoa eiioii
roaeeiiiiooaeeeattm snc rmfprmmmmme ottttub e egpotete
esei nnnacnriaaa encna ssiiuiiqtf lio eltt

- high compression ratio
- high-performance text searching

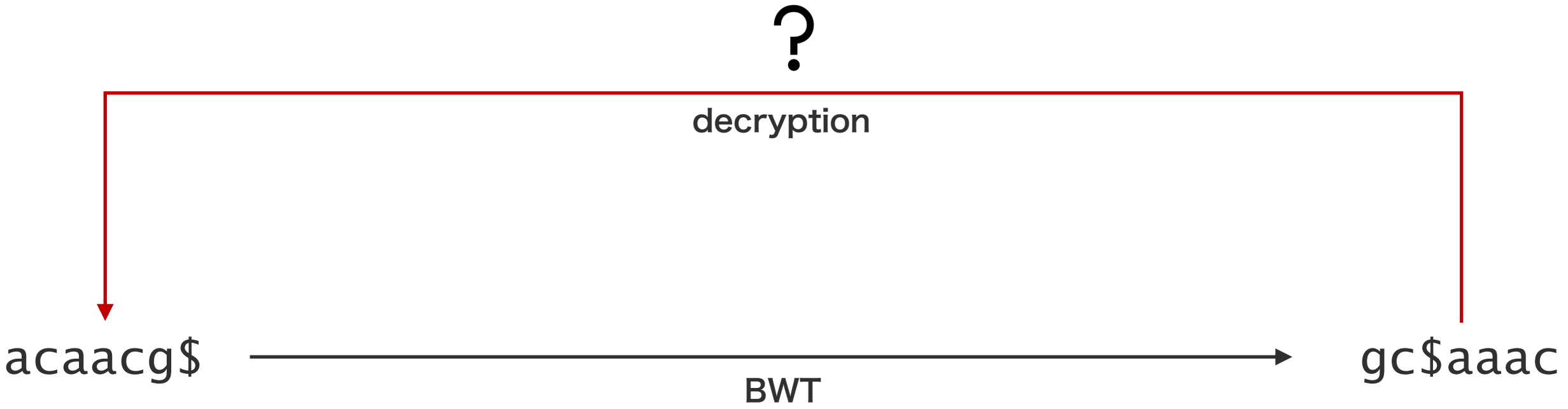
BWT


acaacg$
acaacg$
caacg$a
aacg$ac
acg$aca
cg$acaa
g$acaac
$acaacg
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
gc$aaac

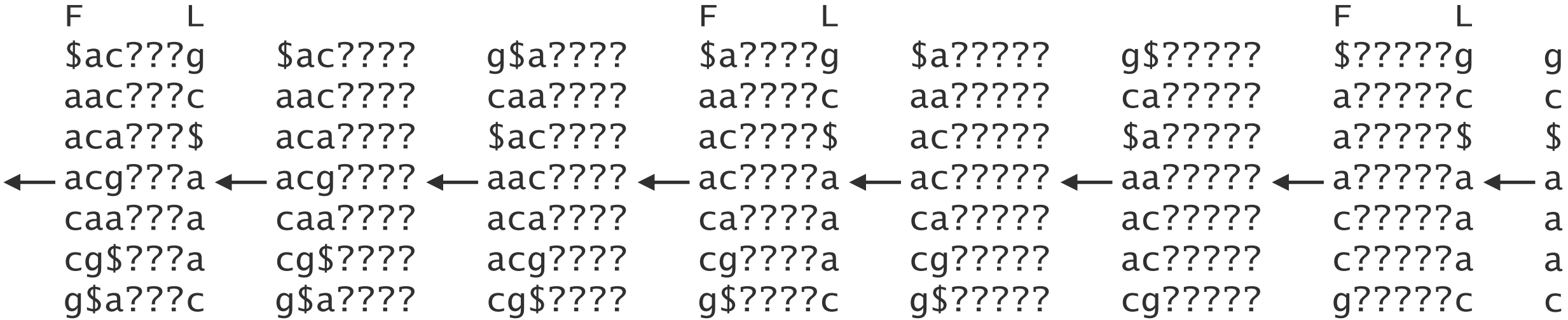
BWT


suffix array
acaacg$
acaacg$
caacg$a
aacg$ac
acg$aca
cg$acaa
g$acaac
$acaacg
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
6
2
0
3
1
4
5
gc$aaac
acaacg$
1234560

BWT



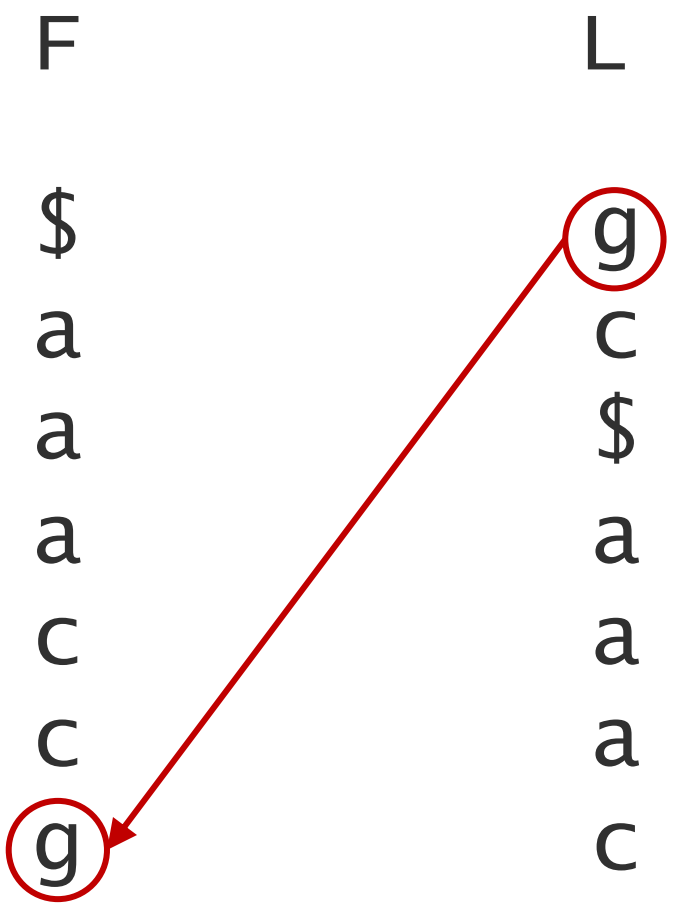
BWT / decryption



- F-column equals to L-column sorted in ascended rank
- L-column equals to BW translated string when F-column is sorted in ascending rank

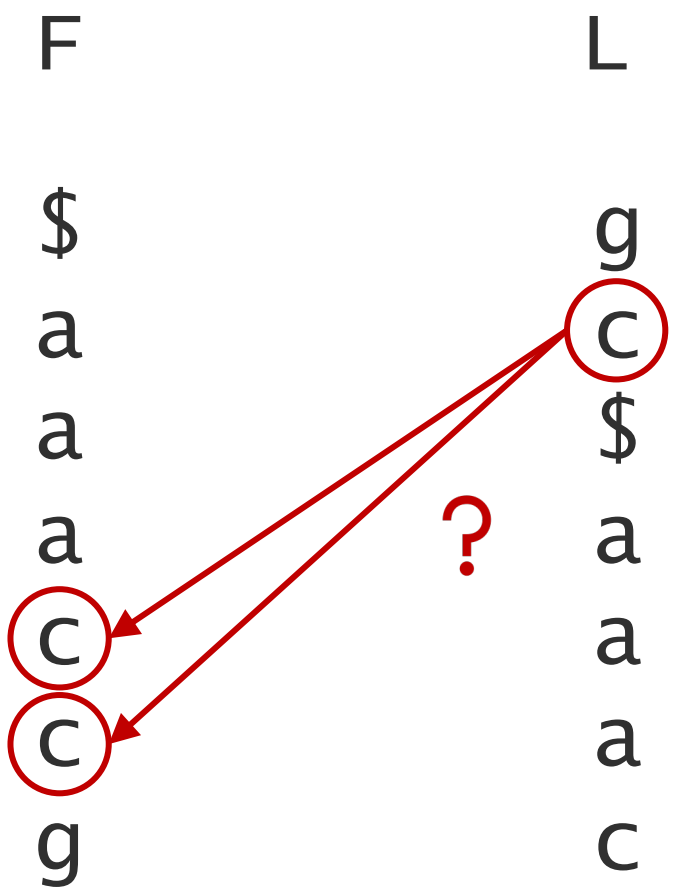
LF mapping

LF mapping is an algorithm mapping from the last column of the BWT to the first column.

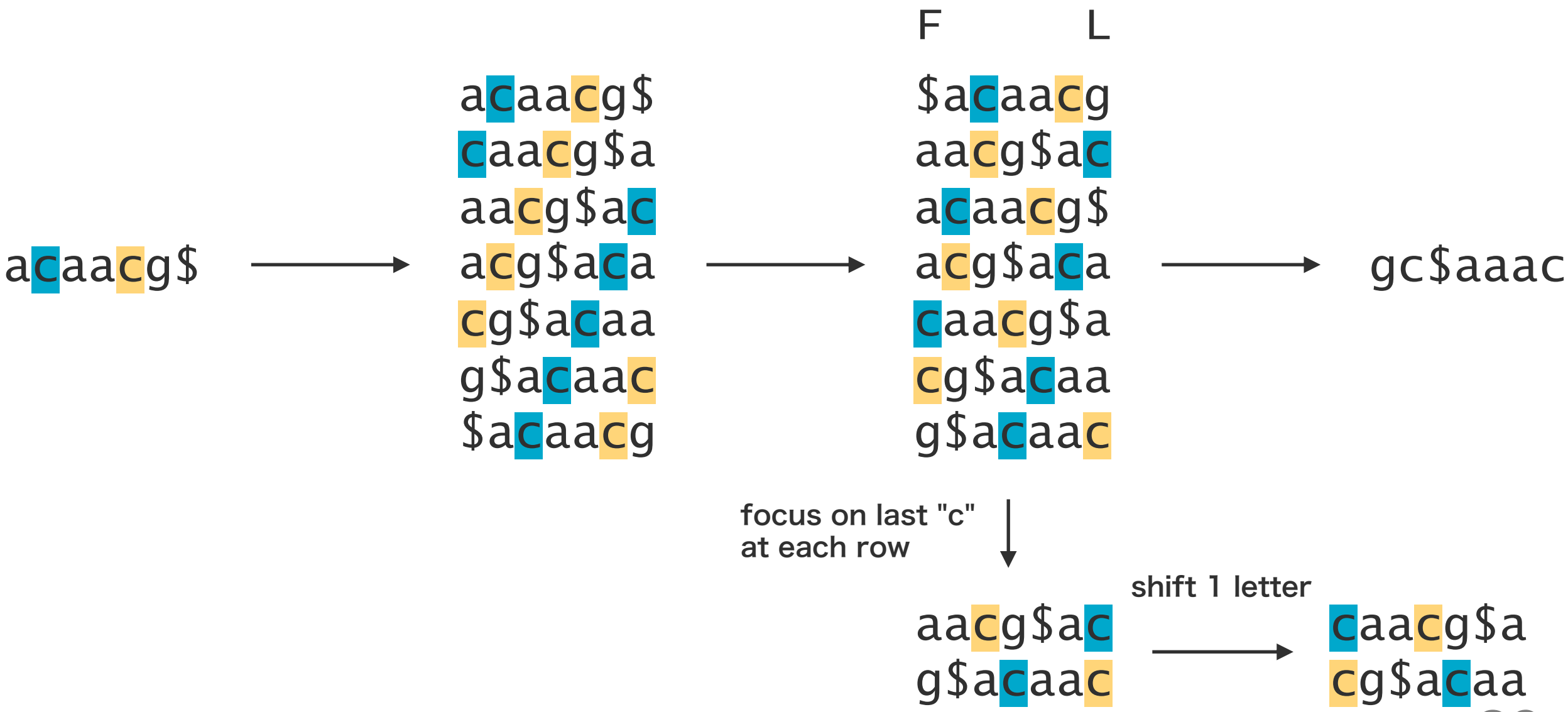


LF mapping

LF mapping is an algorithm mapping from the last column of the BWT to the first column.

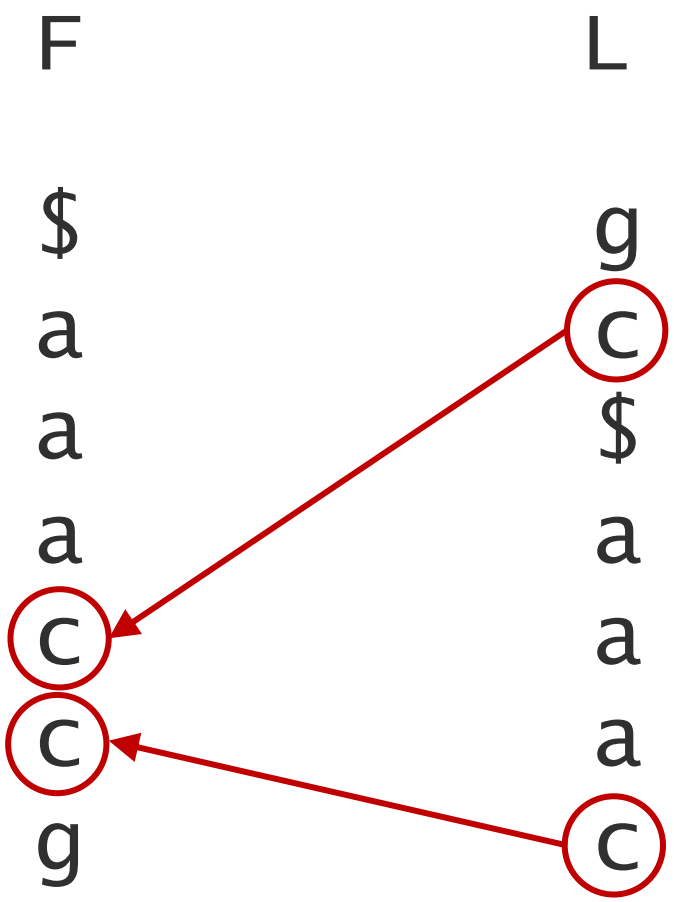


BWT


acaacg$
acaacg$
caacg$a
aacg$ac
acg$aca
cg$acaa
g$acaac
$acaacg
F
L
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
gc$aaac
focus on last "c"
at each row
aacg$ac
g$acaac
shift 1 letter
caacg$a
cg$acaa

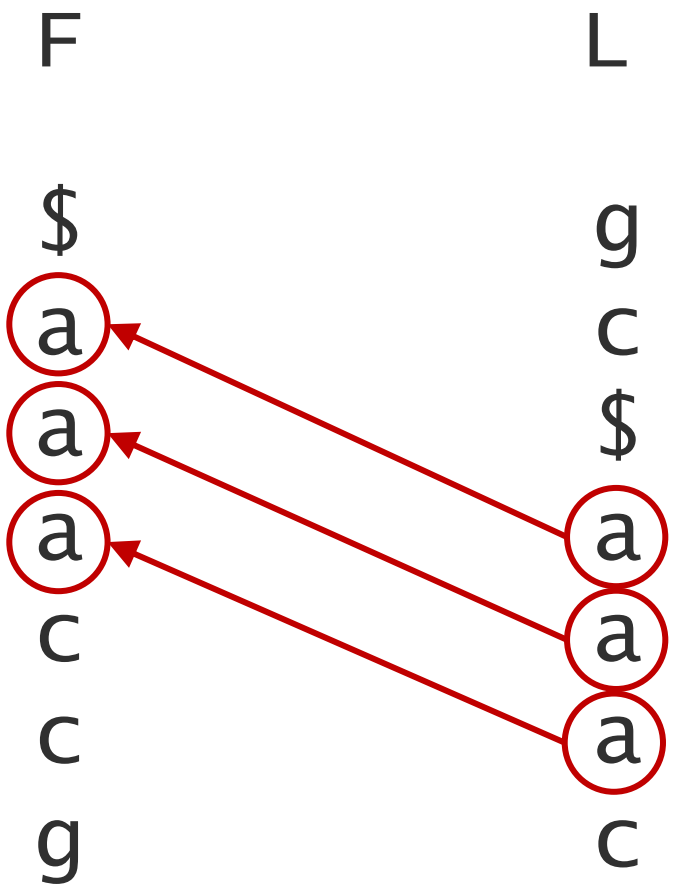
LF mapping

LF mapping is an algorithm mapping from the last column of the BWT to the first column.



LF mapping

LF mapping is an algorithm mapping from the last column of the BWT to the first column.



LF mapping

LF mapping is an algorithm mapping from the last column of the BWT to the first column.

$$LF(i) = C[L[i]] + ooc(L[i], 0, i)$$

C[*s*] index into F column where *c* begin

ooc(*s*, *from*, *to*) a function that calculates the number of occurrences of character *s* between *from* and *to*

```
F      L
$?????g
a?????c
a?????$
a?????a
c?????a
c?????a
g?????c
```

C[*s*] and occ(*s*, *from*, *to*) can be calculated before the analysis.

```
C[$] = 0
C[a] = 1
C[c] = 4
C[g] = 6
```

```
occ(g, 0, 0) = 1
occ(c, 0, 1) = 1
occ($, 0, 2) = 1
occ(a, 0, 3) = 1
occ(a, 0, 4) = 2
occ(a, 0, 5) = 3
occ(c, 0, 6) = 2
```

LF mapping

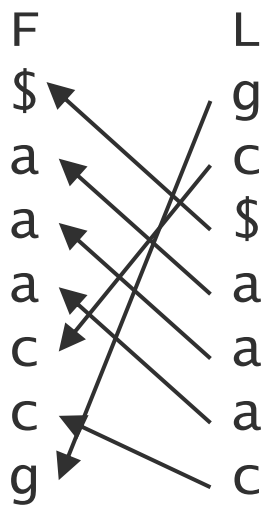
LF mapping is an algorithm mapping from the last column of the BWT to the first column.

$$LF(i) = C[L[i]] + ooc(L[i], 0, i)$$

Compute LF(i) when i = 0, 1, 2, ..., 6

```
LF(0) = C[g] + ooc(g, 0, 0) = 6 + 1 = 7
LF(1) = C[c] + ooc(c, 0, 1) = 4 + 1 = 5
LF(2) = C[$] + ooc($, 0, 2) = 0 + 1 = 1
LF(3) = C[a] + ooc(a, 0, 3) = 1 + 1 = 2
LF(4) = C[a] + ooc(a, 0, 4) = 1 + 2 = 3
LF(5) = C[a] + ooc(a, 0, 5) = 1 + 3 = 4
LF(6) = C[c] + ooc(c, 0, 6) = 4 + 2 = 6
```

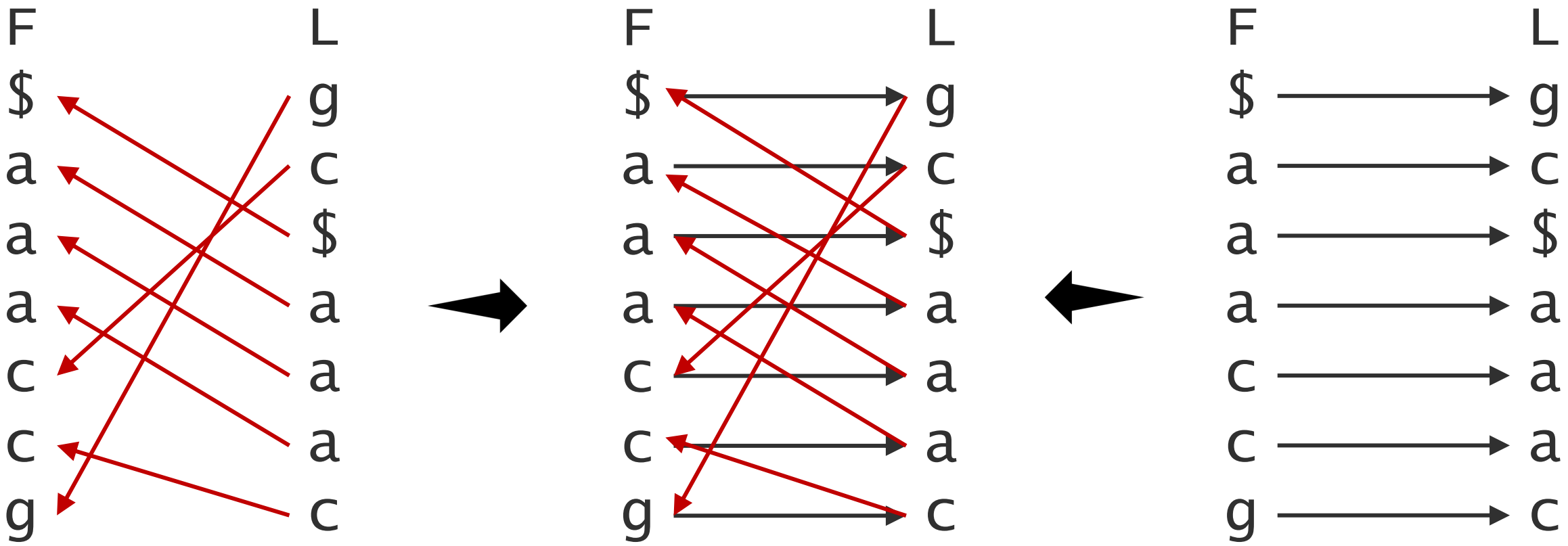
```
F      L
$?????g
a?????c
a?????$
a?????a
c?????a
c?????a
g?????c
```




```
C[$] = 0
C[a] = 1
C[c] = 4
C[g] = 6
```

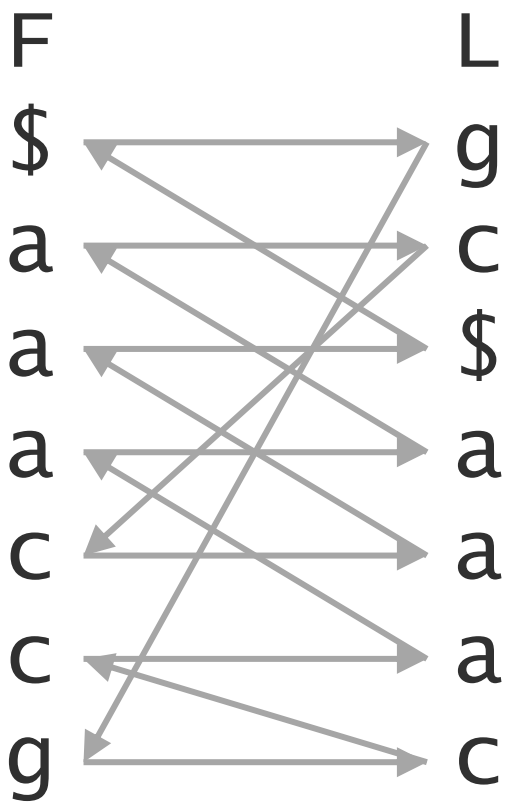
```
occ(g, 0, 0) = 1
occ(c, 0, 1) = 1
occ($, 0, 2) = 1
occ(a, 0, 3) = 1
occ(a, 0, 4) = 2
occ(a, 0, 5) = 3
occ(c, 0, 6) = 2
```

LF mapping



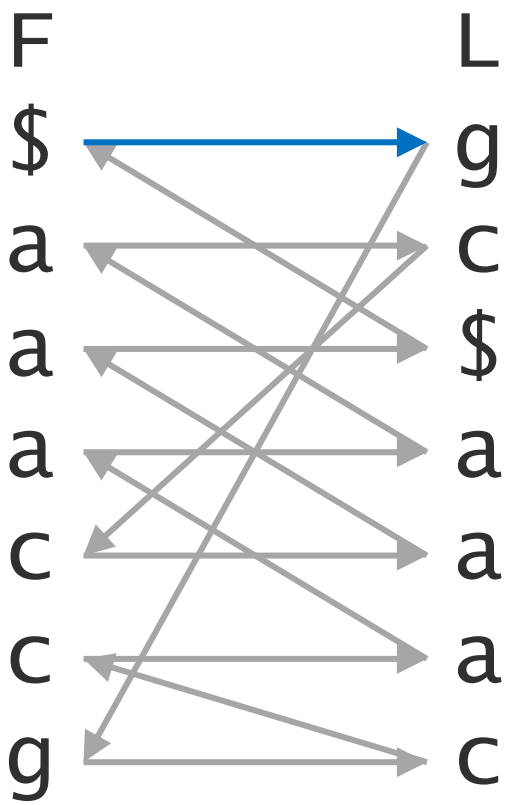
LF mapping

LF mapping



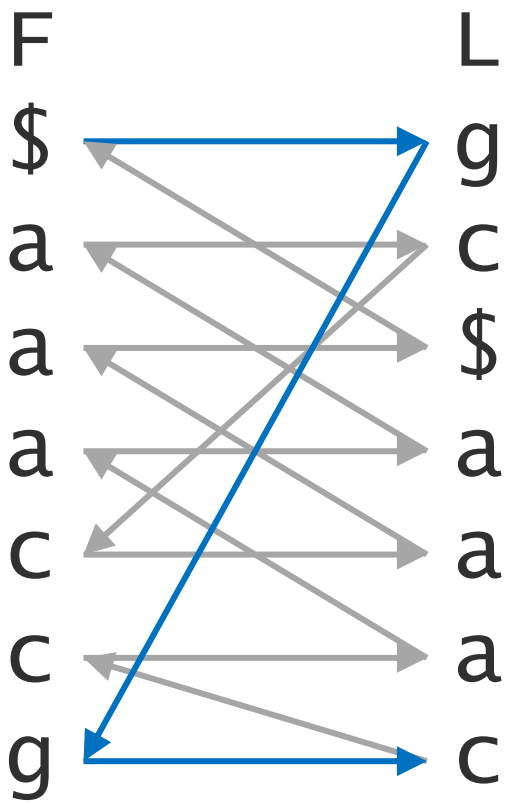
$

LF mapping



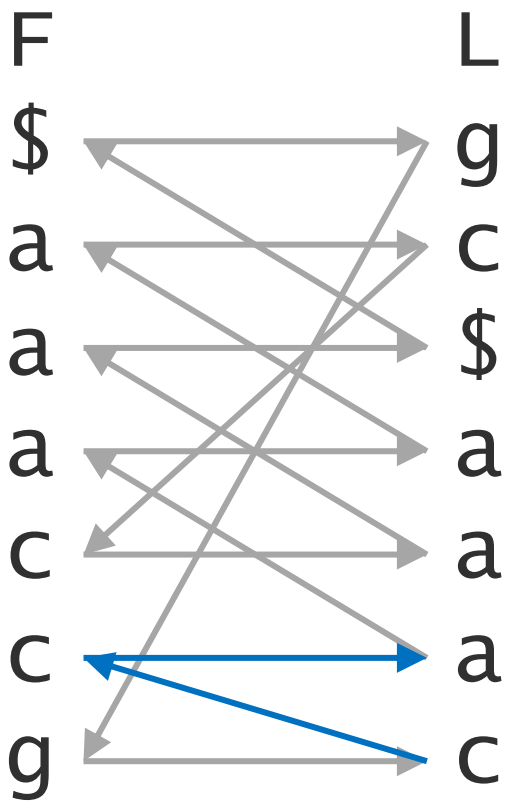
$g

LF mapping



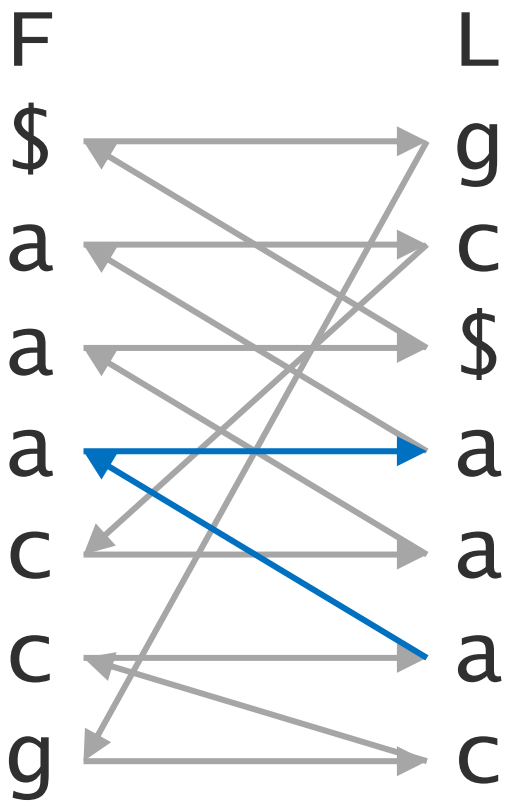
$gc

LF mapping



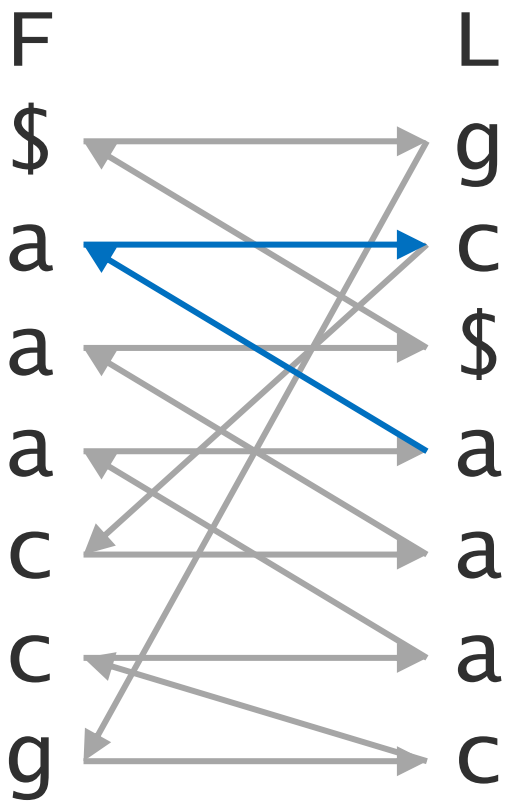
$gca

LF mapping



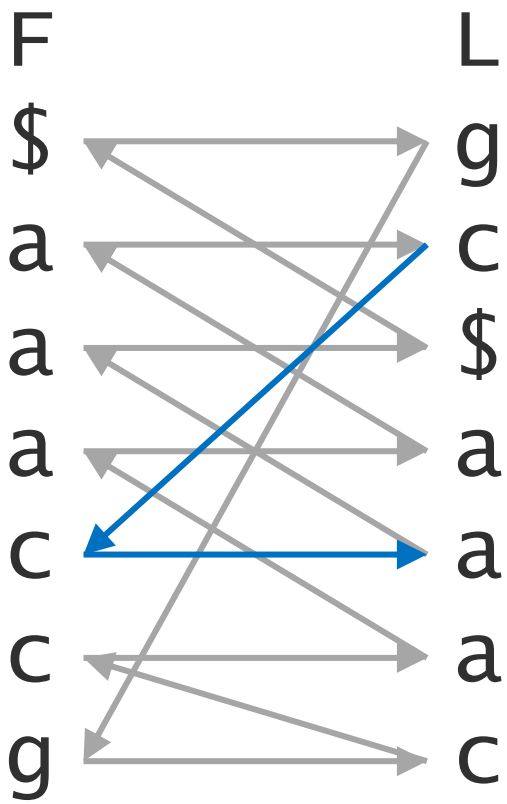
$gcaa

LF mapping



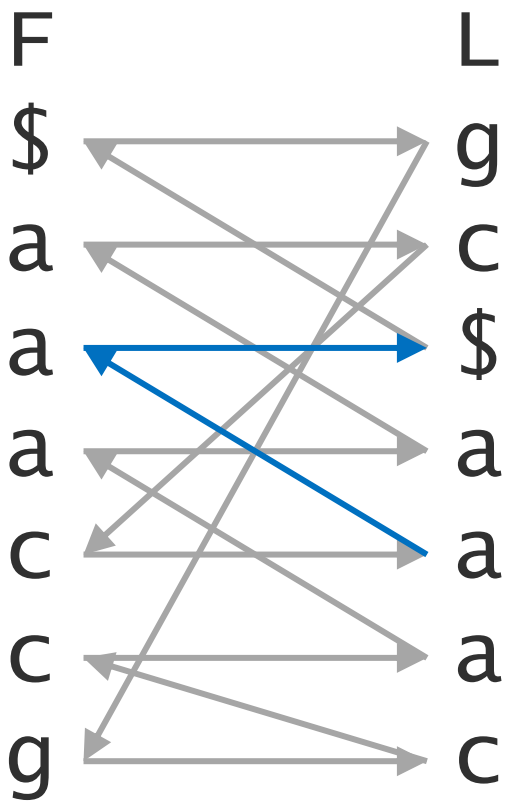
$gcaac

LF mapping



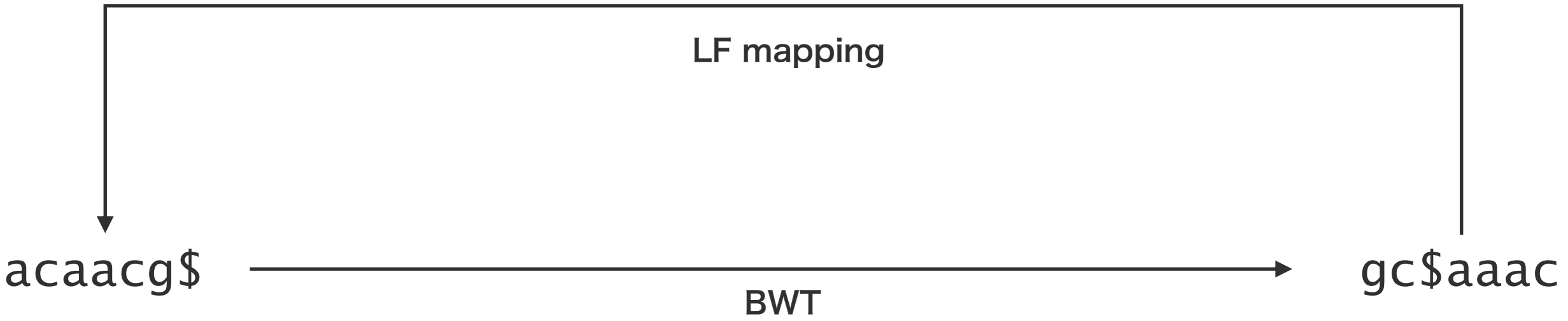
$gcaaca

LF mapping



$gcaaca → (reverse) → acaacg$

BWT



suffix array

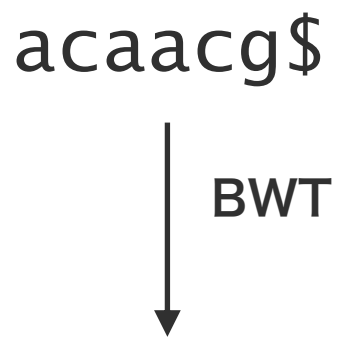
acaacg$

BWT / suffix array

Query = aac

```
$acaacg  6
aacg$ac  2
acaacg$  0
acg$aca  3
caacg$a  1
cg$acaa  4
g$acaac  5
```

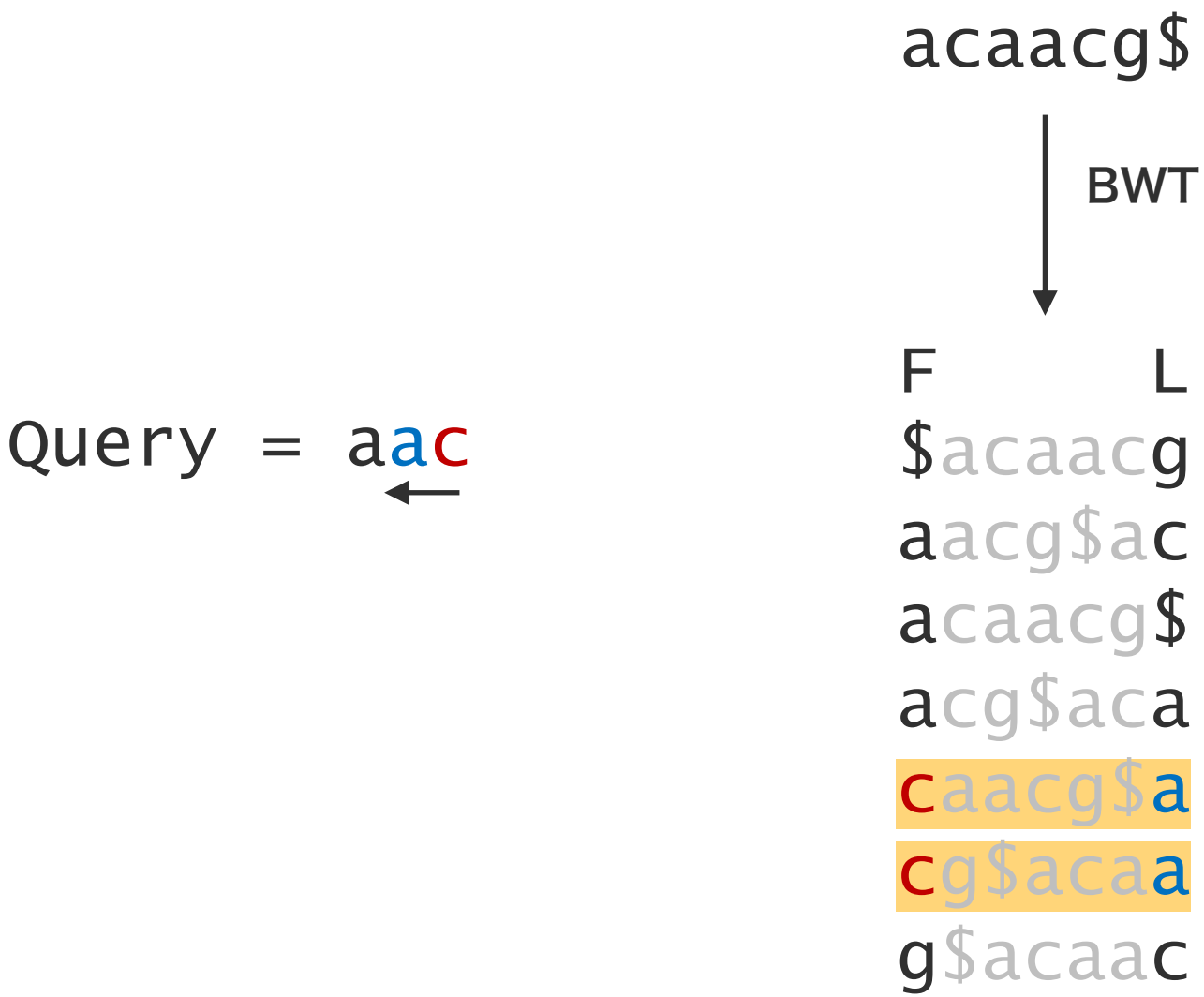
FM index



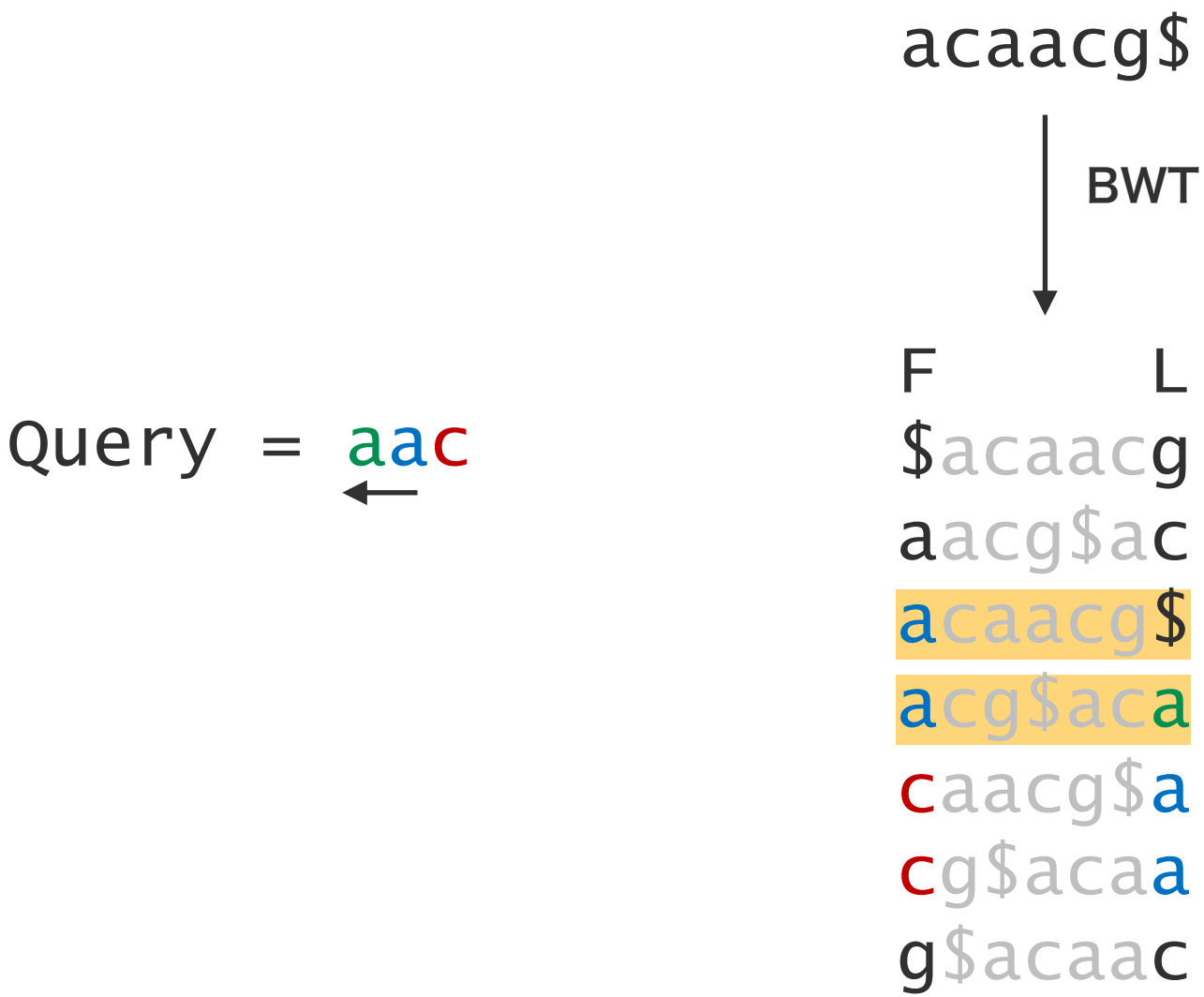
Query = aac

F	L
$	acaacg
a	acg$ac
a	caacg$
a	cg$aca
c	aacg$a
c	g$acaa
g	$acaac

FM index

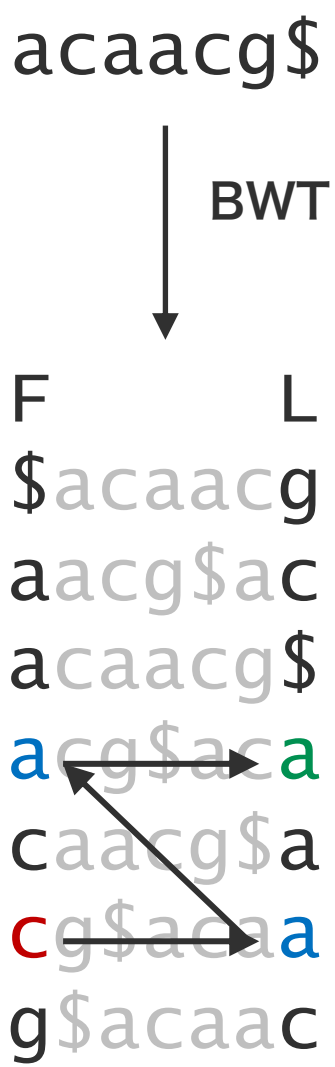

acaacg$
BWT
F L
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
Query = aac

FM index


acaacg$
BWT
F L
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
Query = aac

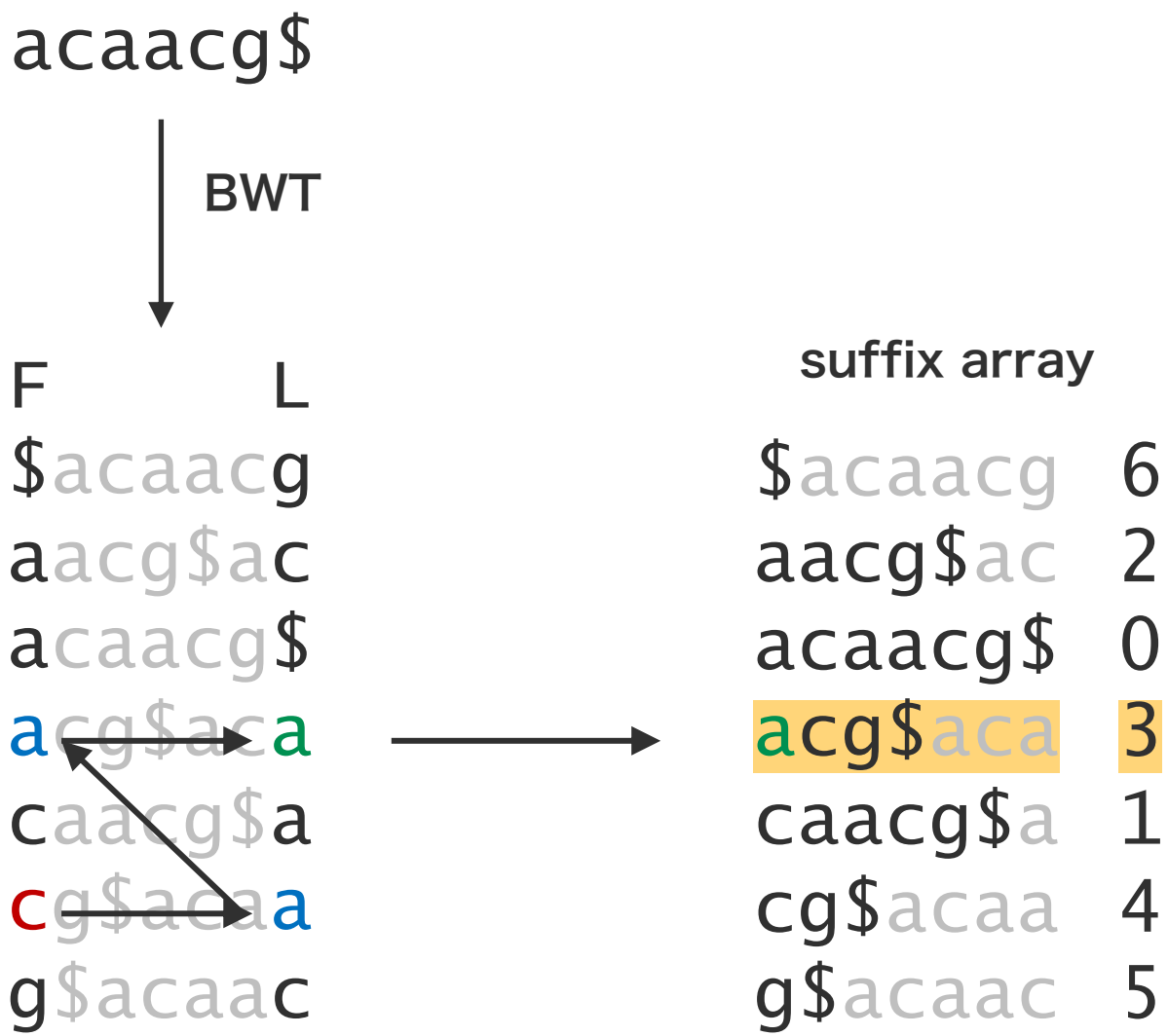
FM index

Query = aac



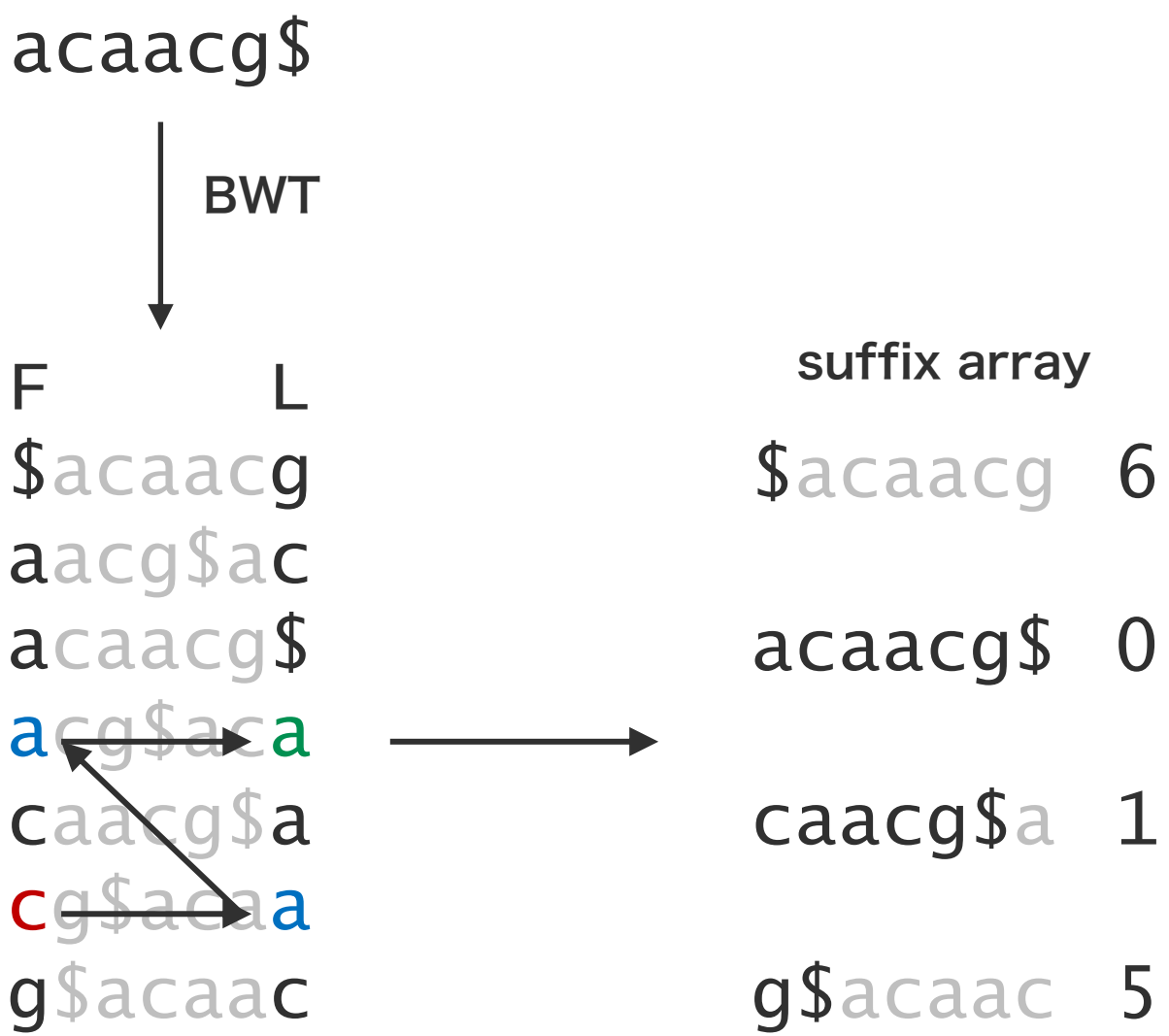
FM index

Query = aac



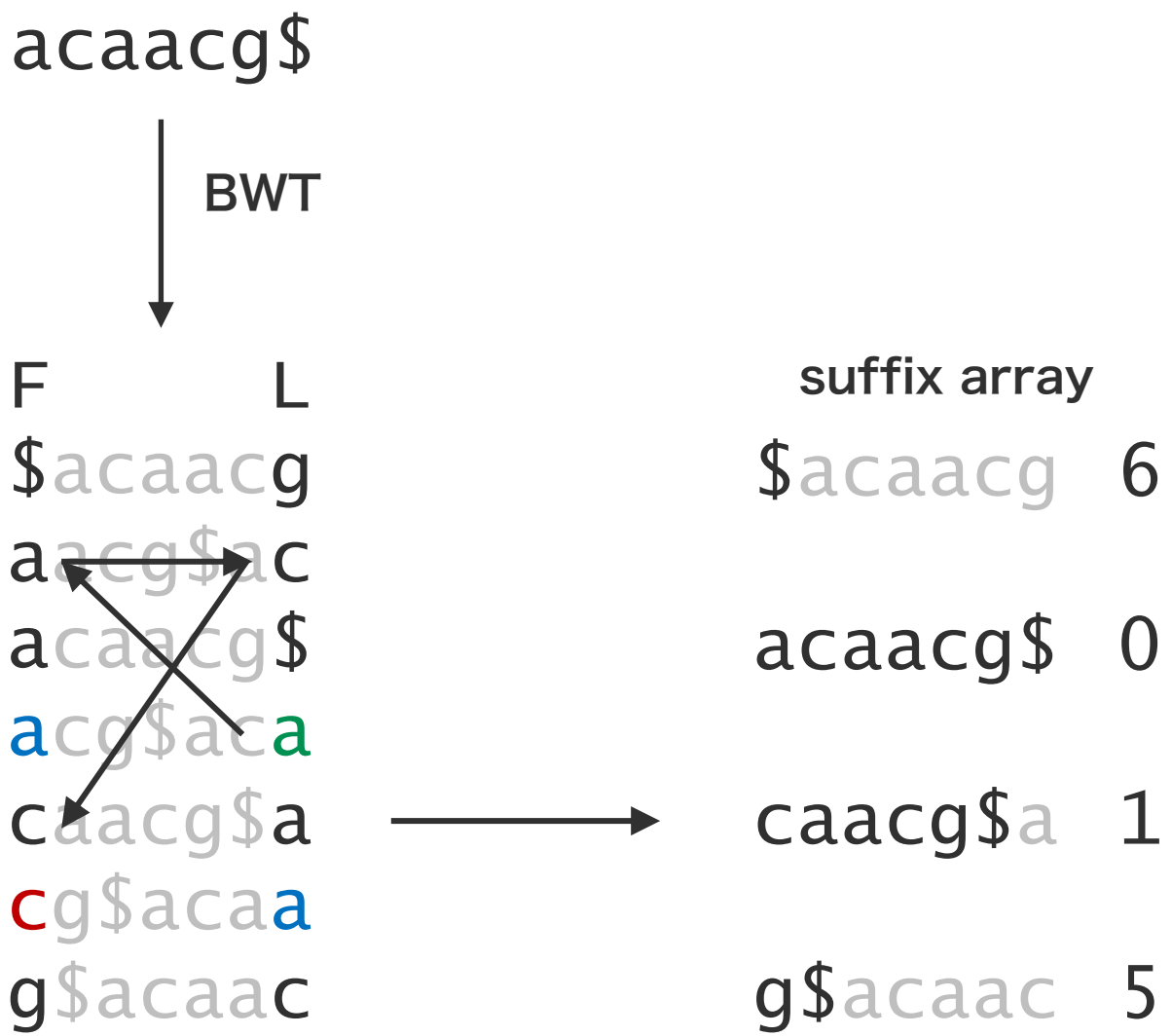
FM index

Query = aac



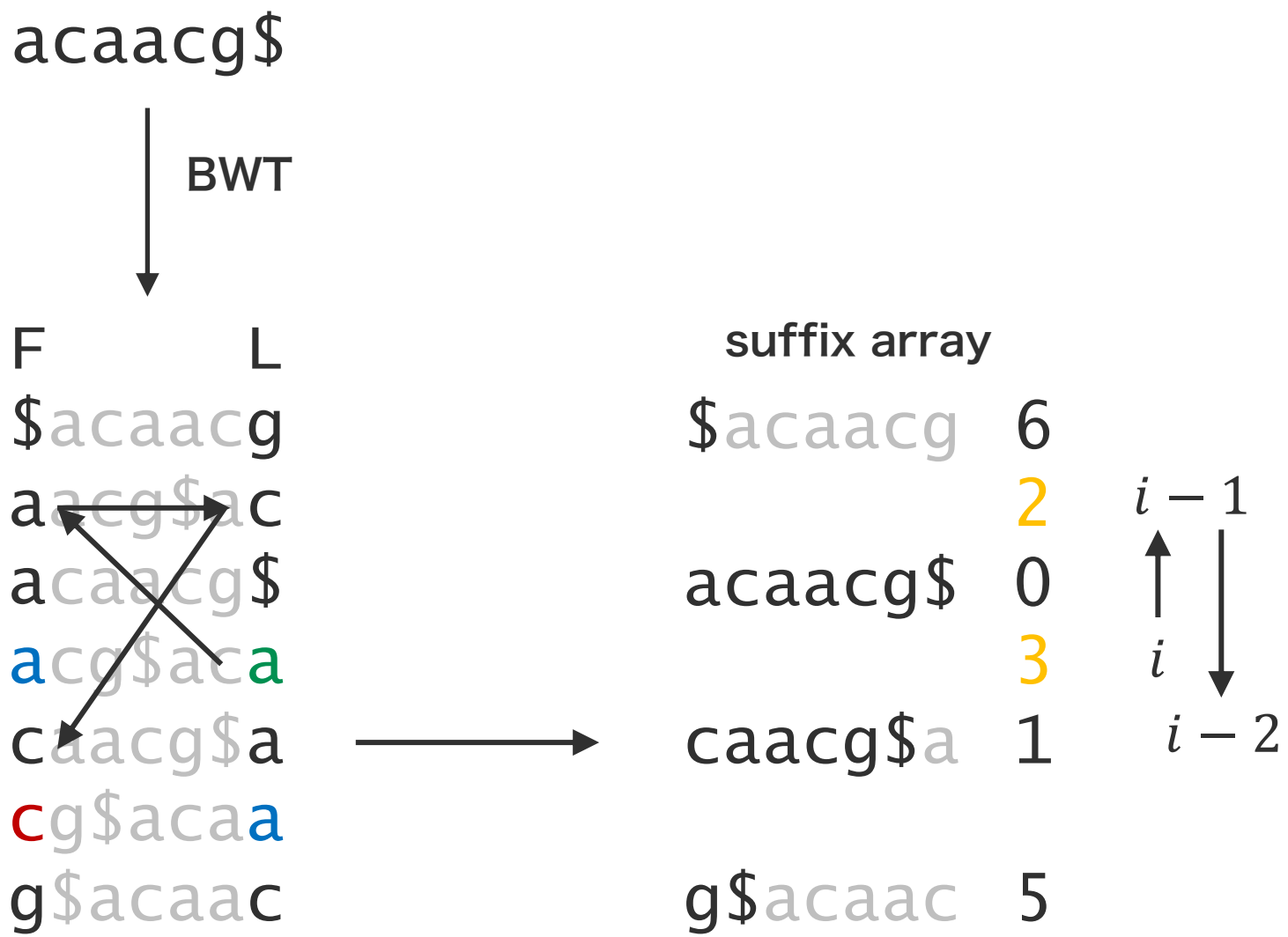
FM index

Query = aac



FM index

Query = aac



Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0							
a	0							
c	0							

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

$s(C,C) = 2$ $s(C,\text{^}C) = -2$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	+2 -1 -1 2						
a	0							
c	0							

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

$s(C,C) = 2$ $\quad$ $s(C,\text{^}C) = -2$ $\quad$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	2	1					
a	0	2						
c	0							

-2 -1
-1
+2 -1
-1

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

$s(C,C) = 2$ $s(C,\text{^}C) = -2$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	2	1					
a	0	2						
c	0							

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

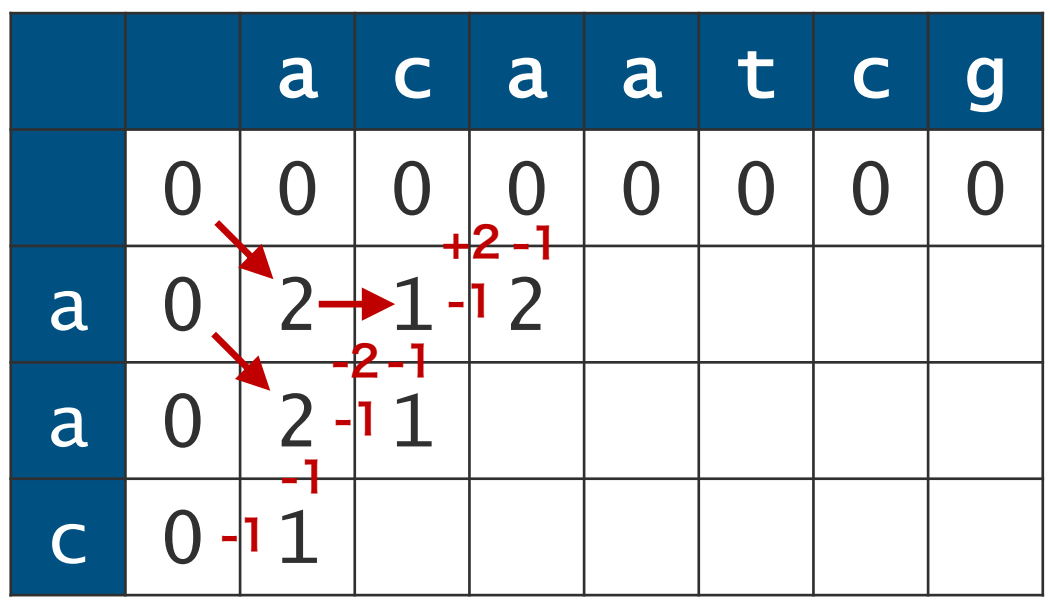
$s(C,C) = 2$ $\quad$ $s(C,\text{^}C) = -2$ $\quad$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	2	1	2				
a	0	2	1					
c	0	1						



$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

$s(C,C) = 2$ $\quad$ $s(C,\hat{}C) = -2$ $\quad$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	2	1	2				
a	0	2	1					
c	0	1						

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

$s(C,C) = 2$ $\quad$ $s(C,\text{^}C) = -2$ $\quad$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	2	1	2	2	1	0	0
a	0	2	1	3	4	3	2	1
c	0	1	4	3	3	2	5	4

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

$s(C,C) = 2$ $s(C,\text{^}C) = -2$ $d = 1$

Smith-Waterman algorithm

Reference

Query

		a	c	a	a	t	c	g
	0	0	0	0	0	0	0	0
a	0	2	1	2	2	1	0	0
a	0	2	1	3	4	3	2	1
c	0	1	4	3	3	2	5	4

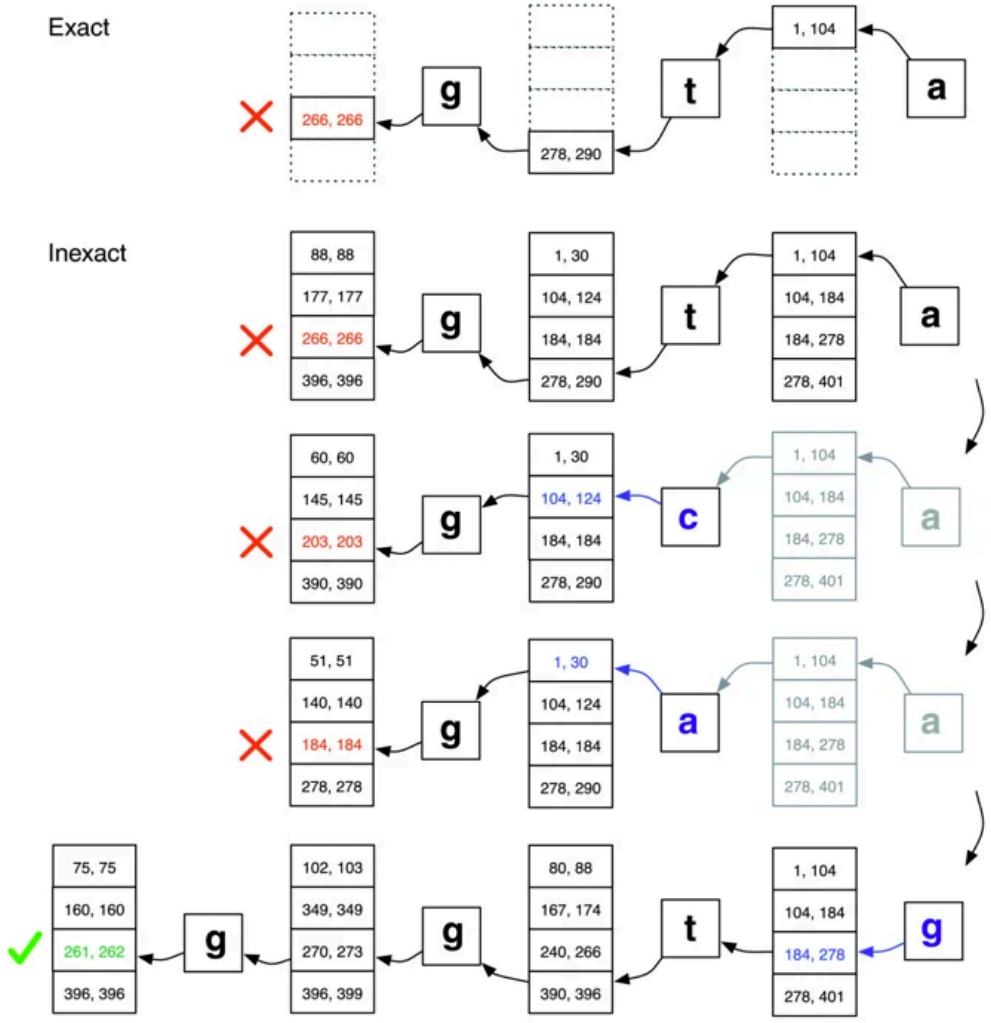
```
acaatcg
 || |
aa-c
```

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

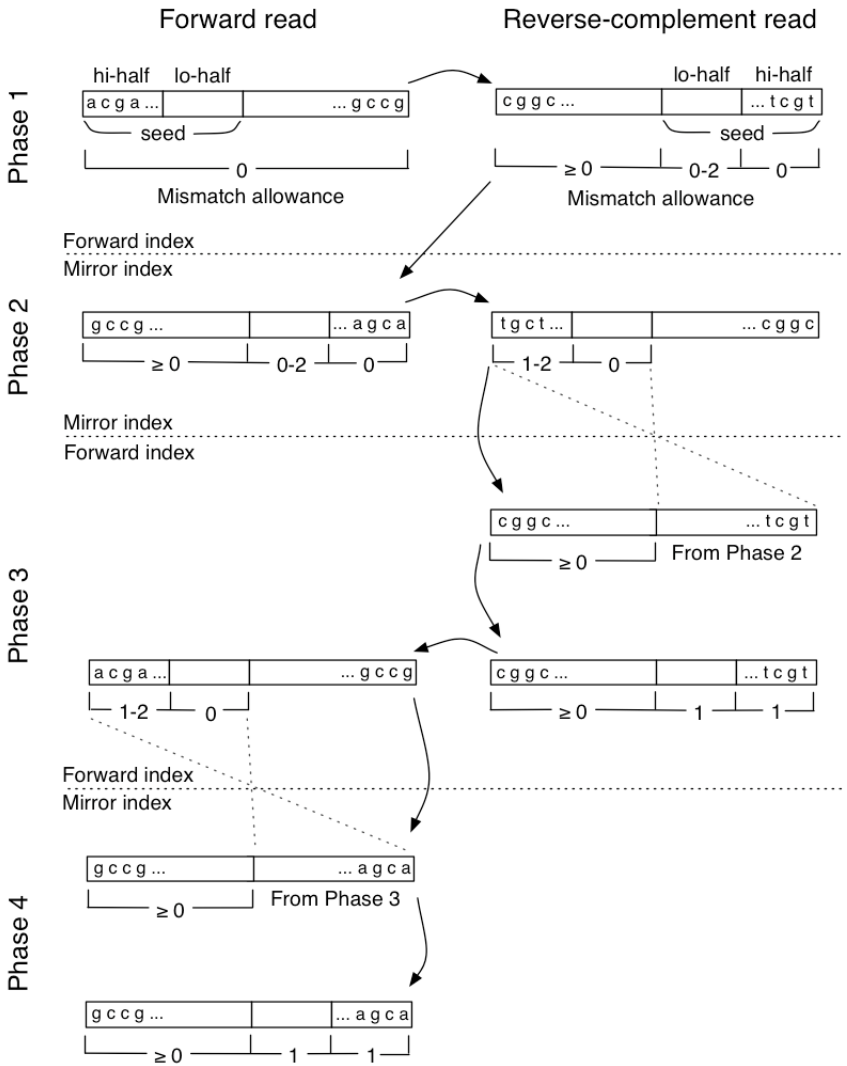
$s(C,C) = 2$ $\quad$ $s(C,\hat{}C) = -2$ $\quad$ $d = 1$

Bowtie

backtracking search



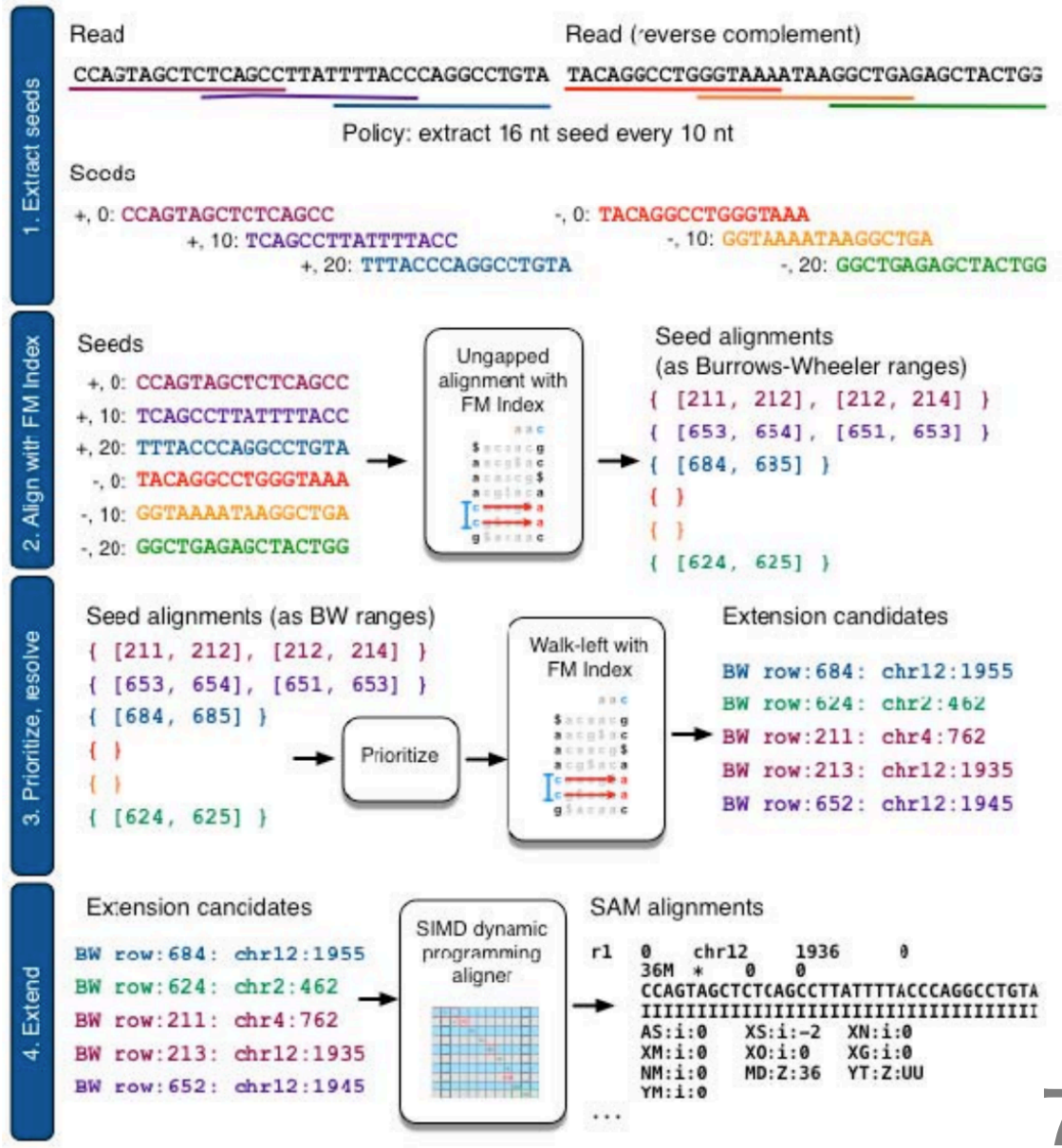
double index



Bowtie2

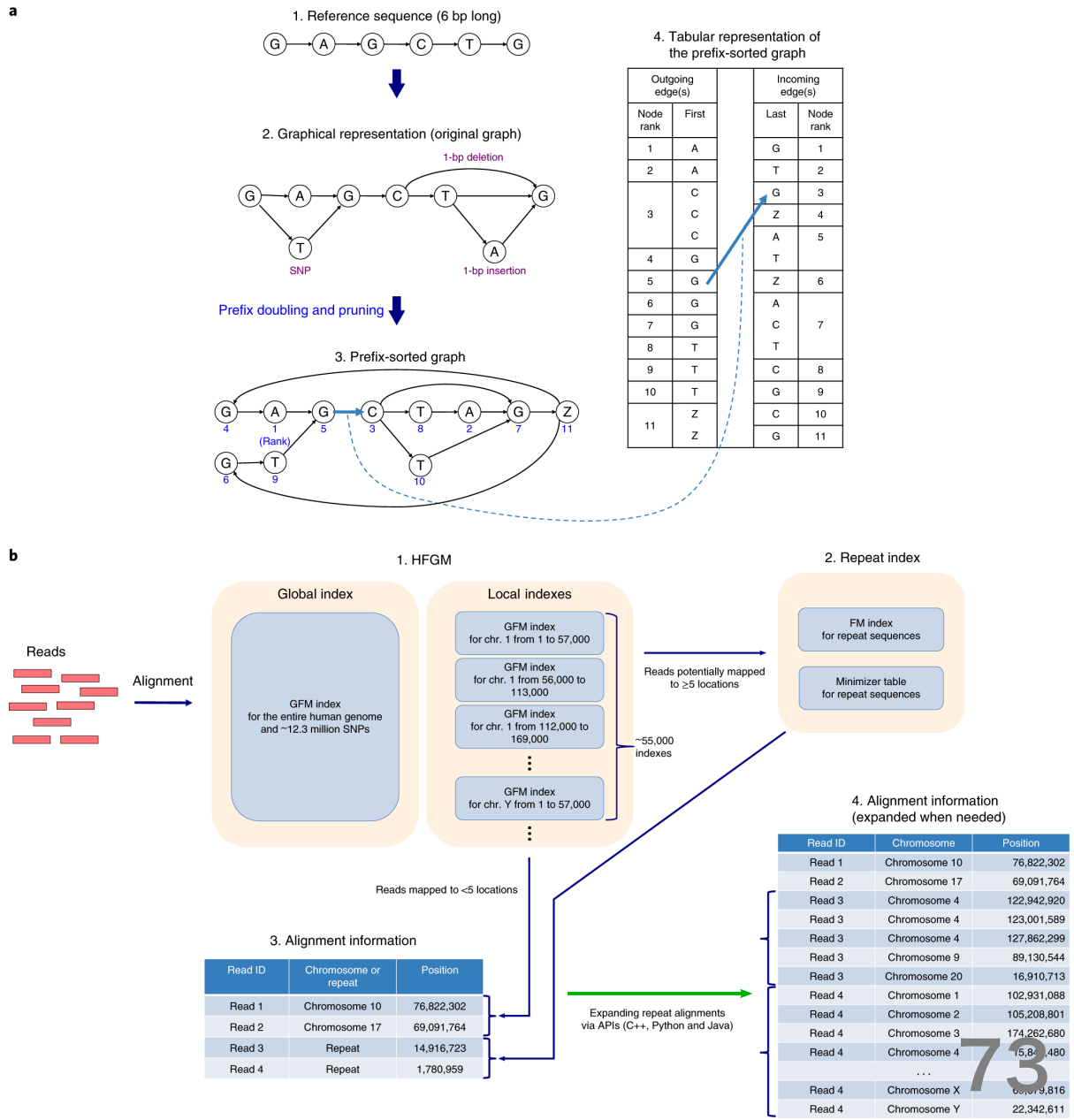
- Similar to Bowtie, Bowtie2 uses BWT index and use 16 bases as a seed

- Extend seed to both ends with Smith-Waterman algorithm
  - allow gaps



HISAT2

- Alignment with FM index algorithm
  - global FM index
  - local FM index
    - generated from 64,000 bp subsequences of genome
    - overlap is 1024 bp between two subseqeuences
    - covered whole genome
- Find alignemnt candidates with global FM index and build a correct alignment with local FM index

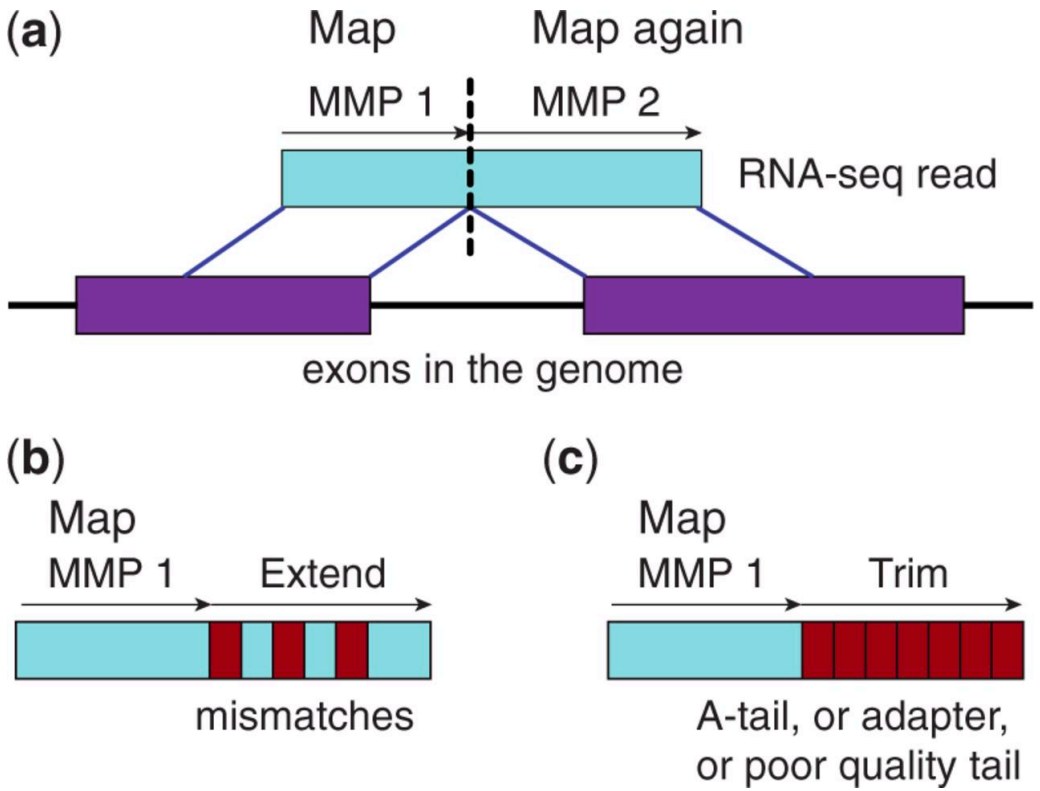


STAR

- STAR consists of two major steps: seed searching and clustering/stitching/scoring

- seed searching
  - search MMP (Maximal Mappable Prefix) from 5'-end
  - then, search MMP from the unmapped position of the given read
  - trim poly-A tails and poor quality tails

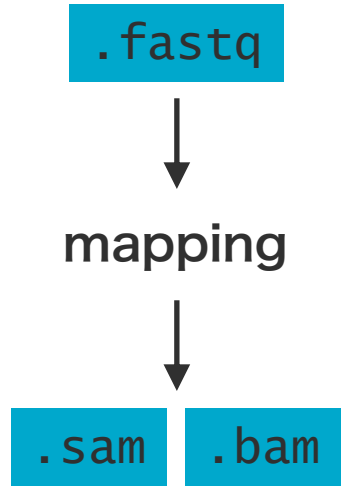
- clustering/stitching/scoring
  - intron size
  - mismatches
  - gaps



SAM format

HEADER

```
@HD VN:1.0
@SD SN:1        LN:30427671
@SD SN:2        LN:19698289
@SD SN:3        LN:23459830
@RG ID:MAK1     PL:ILLUMINA   LB:IL200     SM:Col-0
@PG ID:bwa      PN:bwa        VN:0.6.2
```



RECORDS

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	TAG
SRR7508939.23	163	1	3848	60	66M82N22M	=	3890	128	GCC...GAG	HIJ...DD@	AS:i:0 XN:i:0 XM:i:0 ... NH:i:1

SAM format

FLAG

```
   1 read paired
   2 read mapped in proper pair
   4 read unmapped
   8 mate unmapped
  16 read being reverse complemented
  32 mate being reverse complemented
  64 first in the template
 128 last in the template
 256 not primary alignment
 512 not passing platform/vendor quality checks
1024 read is PCR or optical duplicate
2048 supplementary alignment
```

CIGAR

```
M alignment match (can be a sequence match or mismatch)
I insertion to the reference
D deletion from the reference
N skipped region from the reference
S soft clipping (clipped sequences present in SEQ)
H hard clipping (clipped sequences NOT present in SEQ)
```

TAG

```
AS alignment score
XO number of best hits
XM number of mismatches in the alignment
NH number of reported alignments that contain the query in the current record
```

BAM

mapping

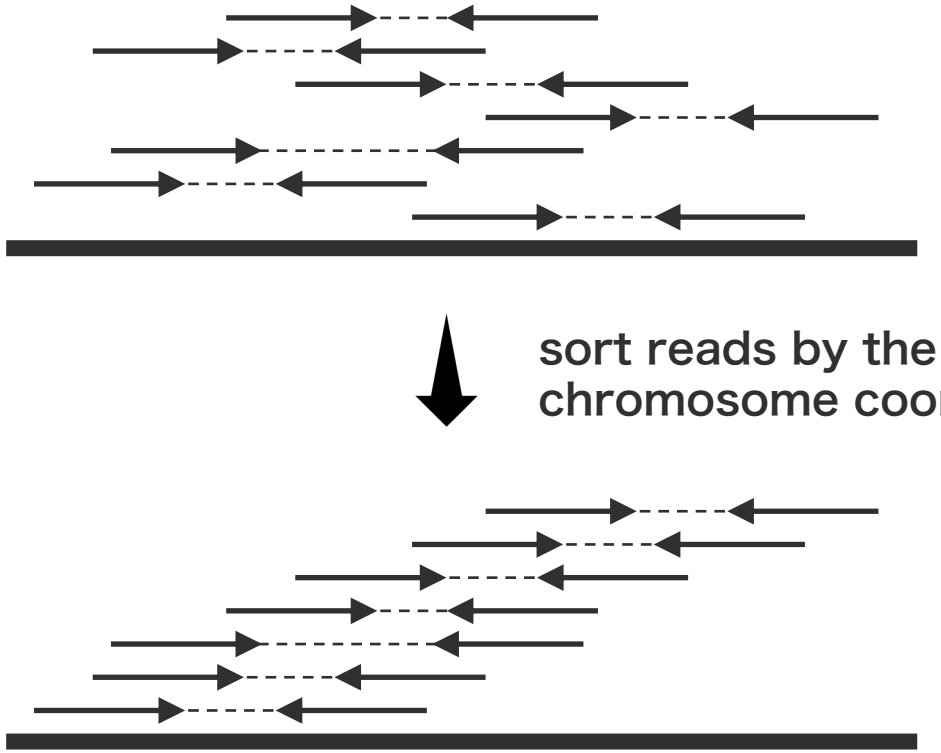
.sam .bam

sorting samtools

.bam

indexing samtools

.bai .bam

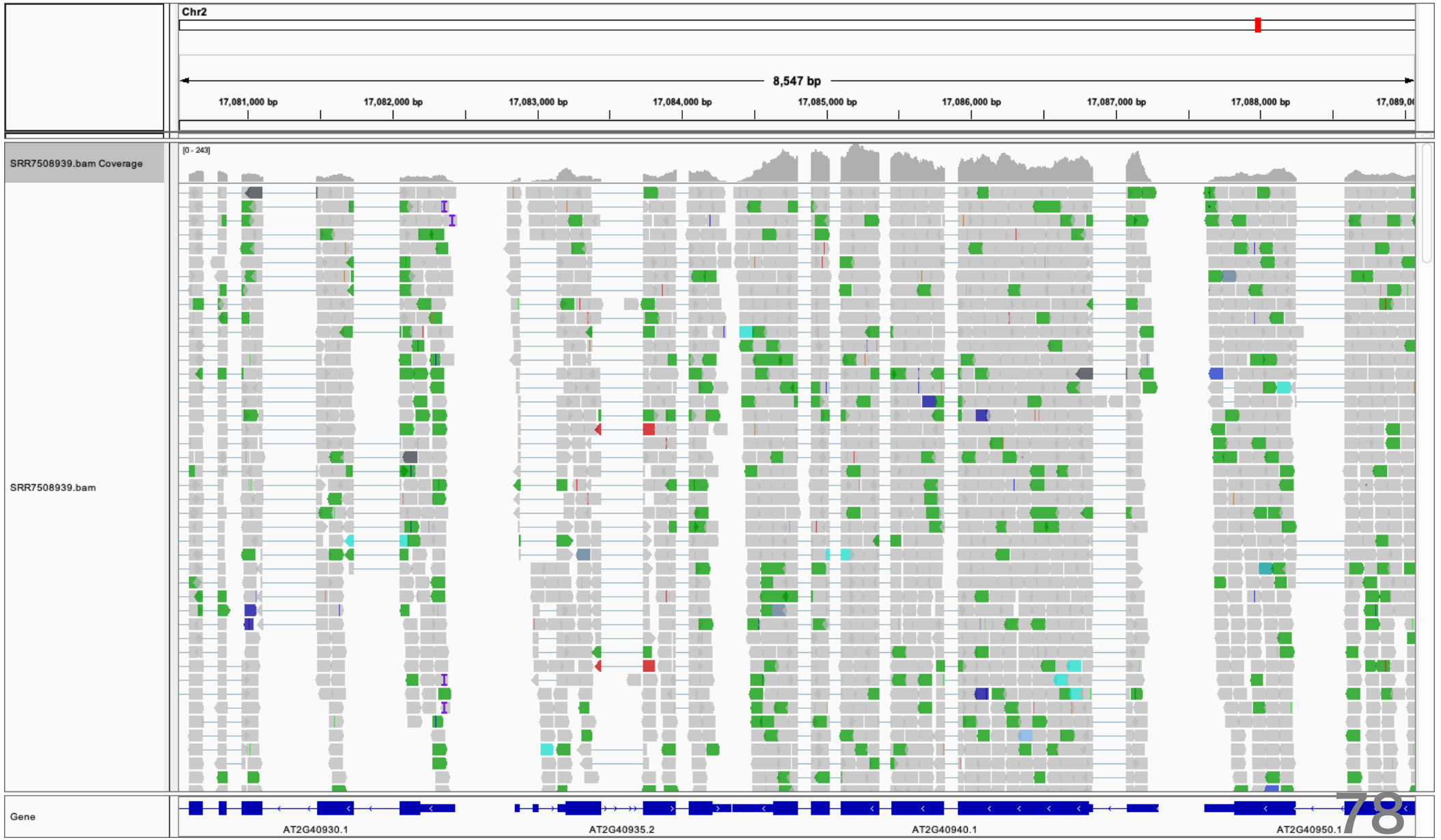


IGV

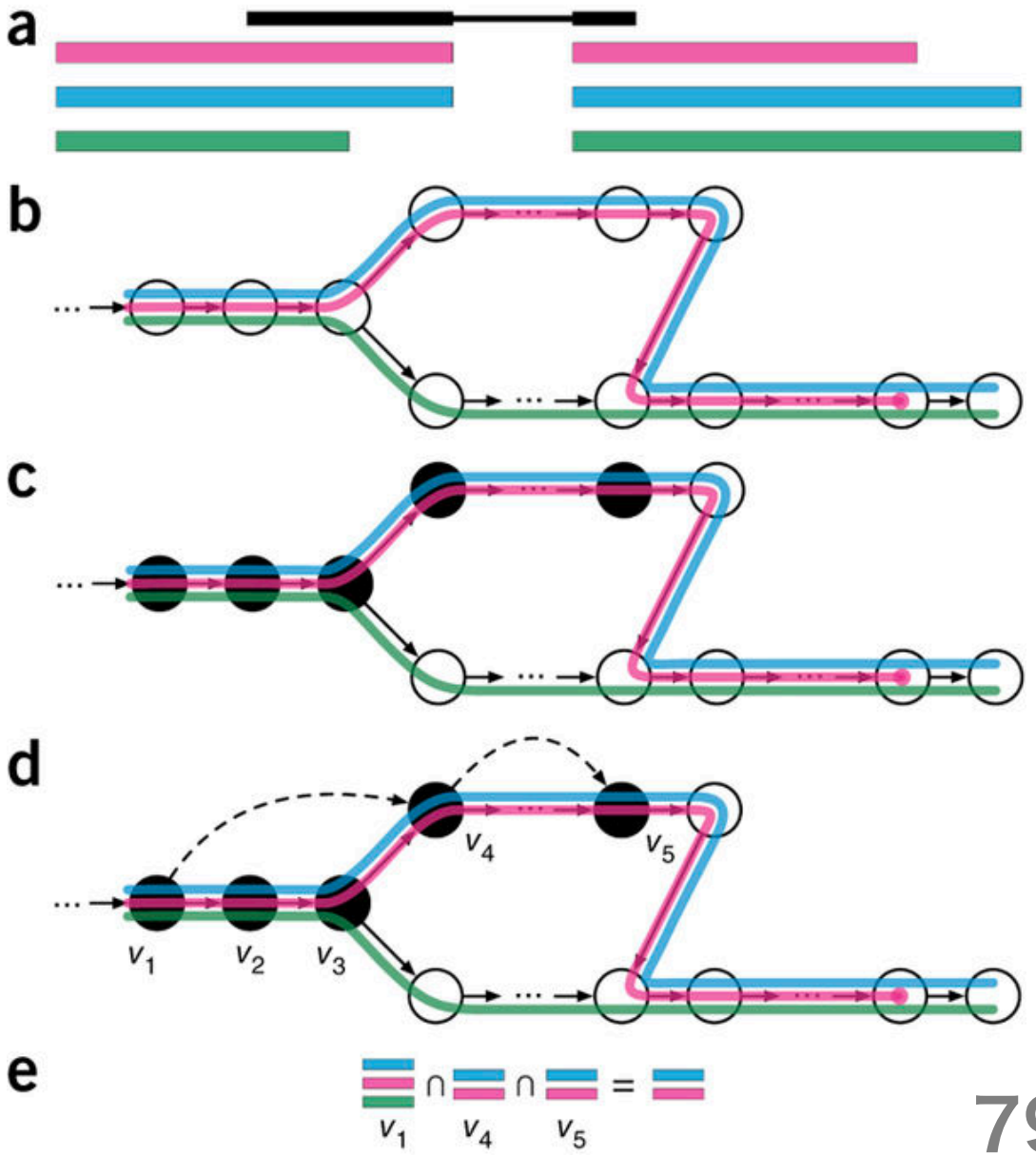
.bam

.bai

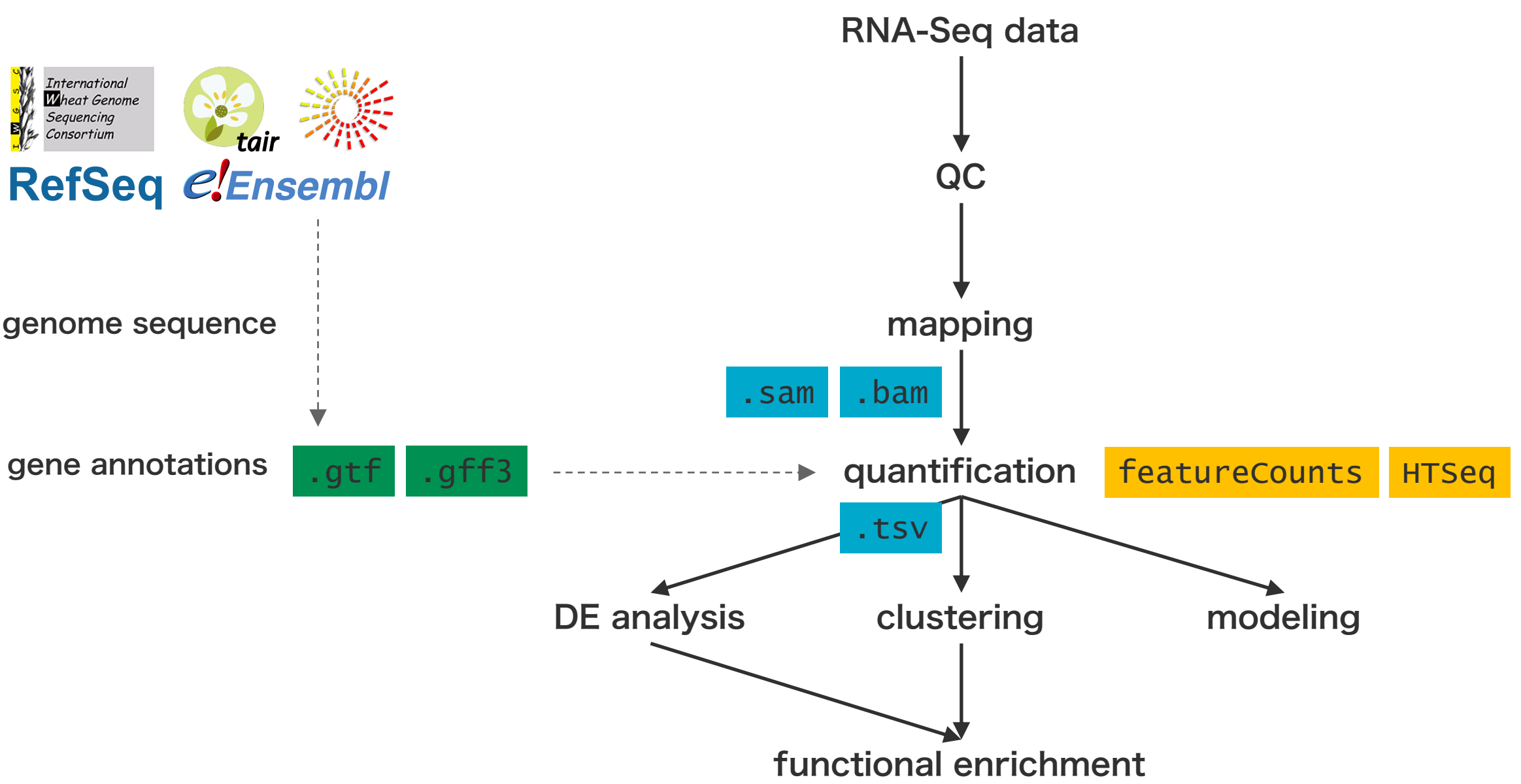
.gtf



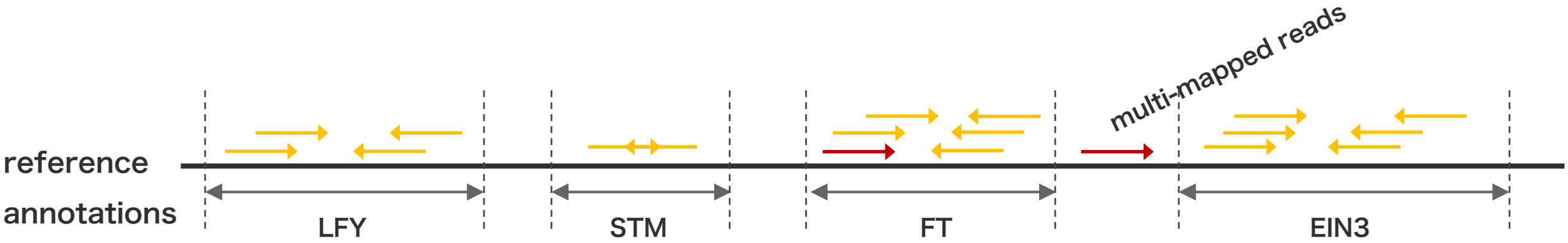
kallisto



Quantification



Quantification



quantification

`featureCounts` `HTSeq`

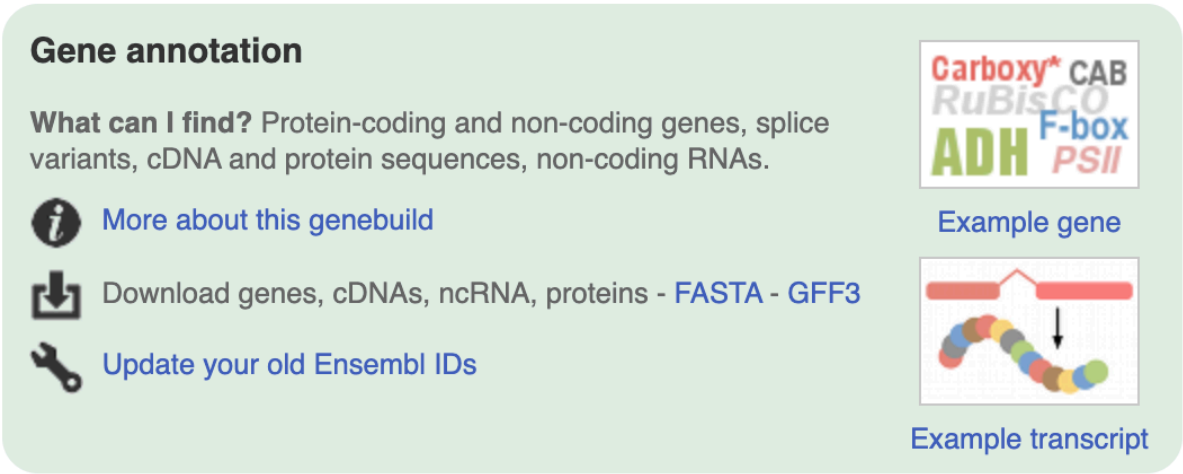
gene	count
LFY	2
STM	1
FT	2
EIN3	3

Gene annotation

- Gene annotations are recorded by GTF or GFF3 format
  - tab delimited text file
  - nine columns records chromosome name, feature, position, strand, gene ID, exon ID, and so on

Ensembl

http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index



Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz

Arabidopsis_thaliana.TAIR10.47.gff3.gz

GTF/GFF format

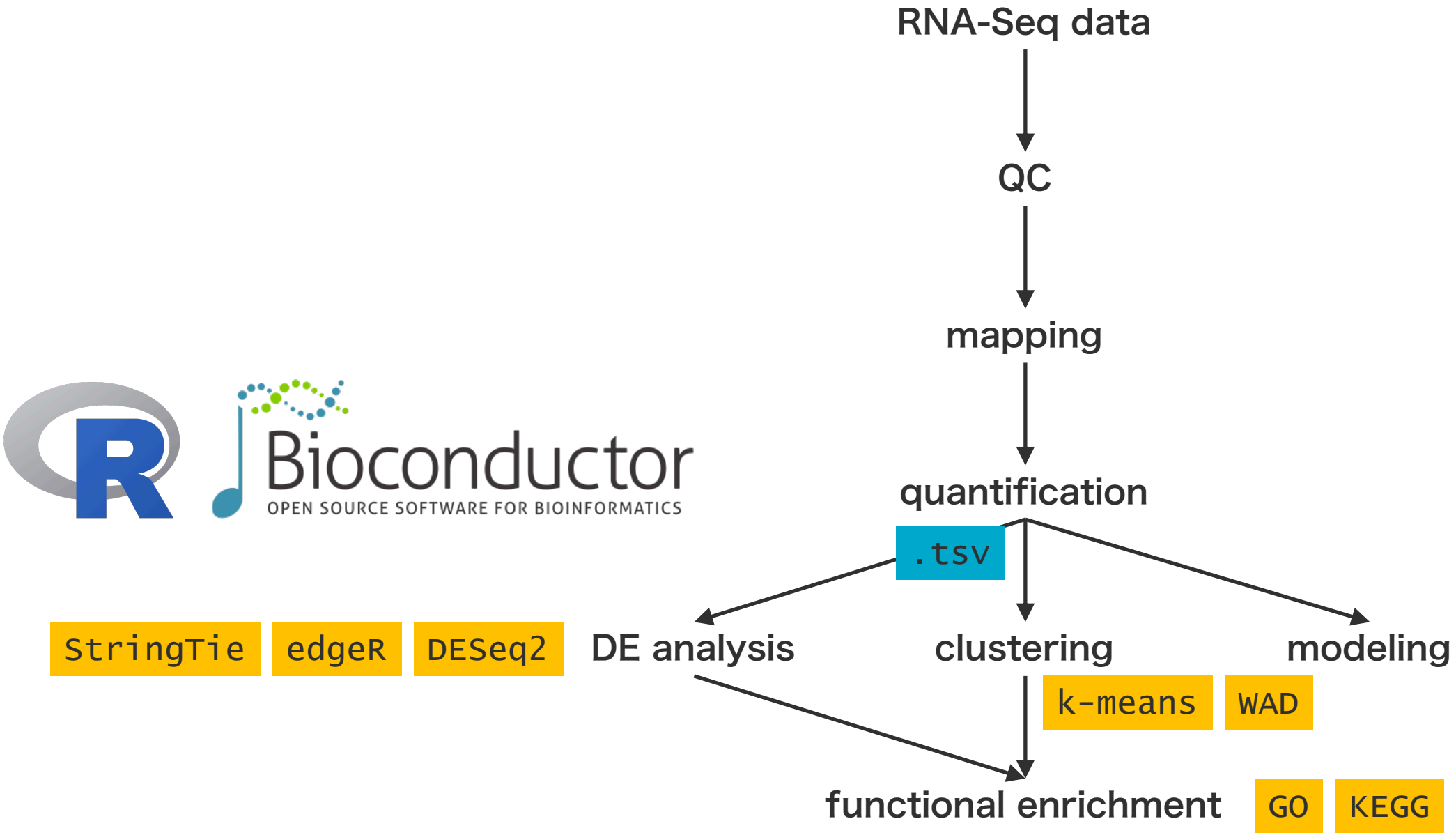
GTF (Gene Transfer Format)

```
chr1 at exon 9873504 9874841 . + . gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at CDS  9873504 9874841 . + 0 gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at exon 9877488 9877679 . + . gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at CDS  9877488 9877679 . + 0 gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at exon 9888412 9888586 . + . gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at CDS  9888412 9888586 . + 0 gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at exon 9891475 9891998 . + . gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
chr1 at CDS  9891475 9891995 . + 2 gene_id "G01"; transcript_id "T01"; gene_name "ZNF366";
```

GTF (General Feature Format)

```
ctg123 at mRNA 1300 9950 . + . ID=t_012143;gene_name=EDEN
ctg123 at exon 1300 1500 . + . Parent=t_012143
ctg123 at exon 3000 3902 . + . Parent=t_012143
ctg123 at CDS  3301 3902 . + 0 Parent=t_012143
ctg123 at exon 5000 5500 . + . Parent=t_012143
ctg123 at CDS  5000 5500 . + 1 Parent=t_012143
ctg123 at exon 7000 9000 . + . Parent=t_012143
ctg123 at CDS  7000 7600 . + 1 Parent=t_012143
ctg123 at exon 9400 9950 . + . Parent=t_012143
```

Data analysis



RNA-Seq

- ○ Introduction
- ○ Quality control
- ○ Mapping and quantification
- ○ Assembly

Assembly

de novo assembly

GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

sequencing

CCCCAGGCGGGGCTATCG
GGCGGGGCTATCGAAATC
CGGGGCTATCGAAATCGA
CCAGGCGGGGCTATCGAA

TCCCTTTTAGCCCCTTAG
CCCTTTTAGCCCCTTAGT
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT

de novo assembly

GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

sequencing

CCCCAGGCGGGGCTATCG
GGCGGGGCTATCGAAATC
CGGGGCTATCGAAATCGA
CCAGGCGGGGCTATCGAA

TCCCTTTTAGCCCCTTAG
CCCTTTTAGCCCCTTAGT
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT

assembly

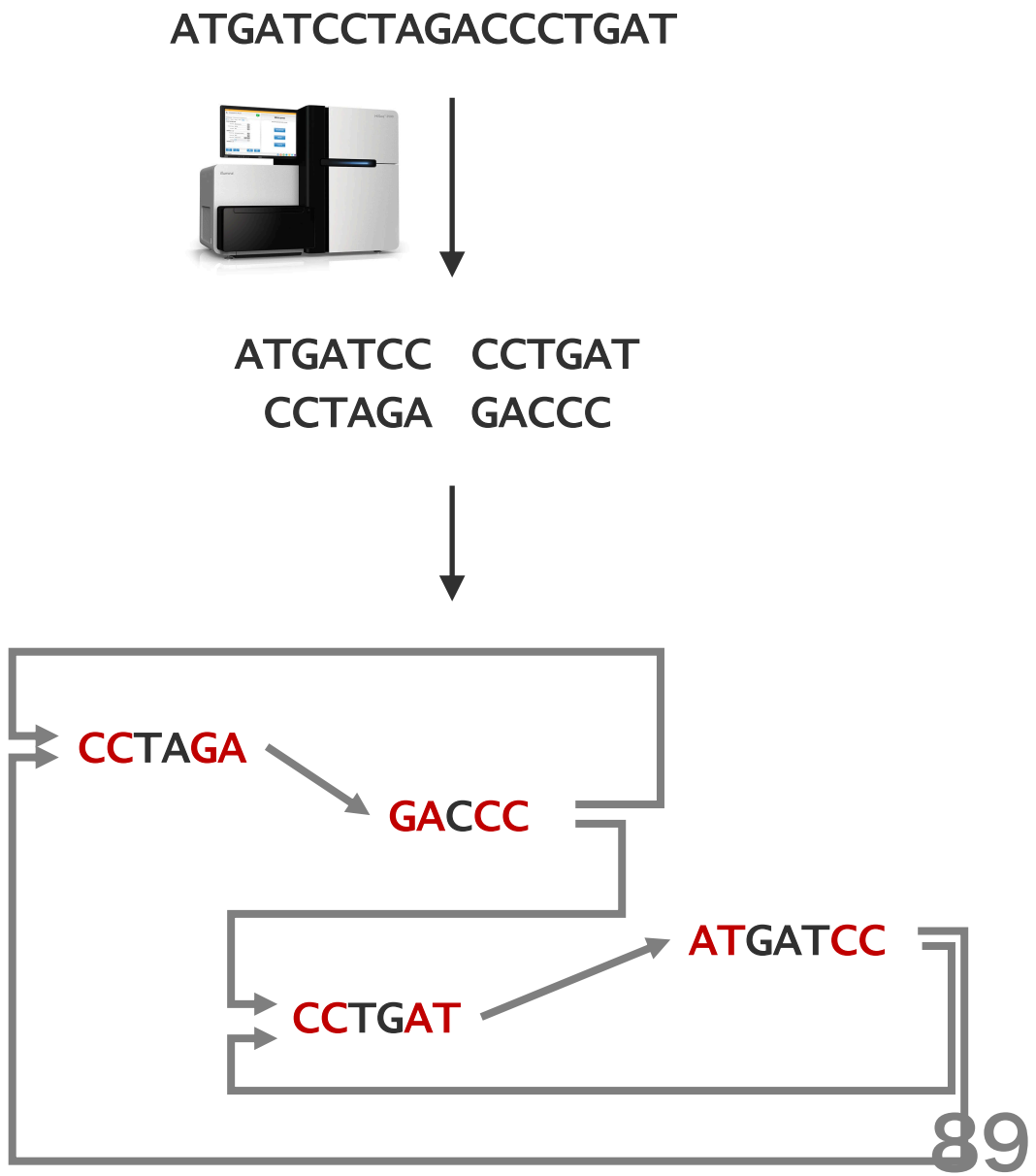
CCCCAGGCGGGGCTATCG
GGCGGGGCTATCGAAATC
CGGGGCTATCGAAATCGA
CCAGGCGGGGCTATCGAA

TCCCTTTTAGCCCCTTAG
CCCTTTTAGCCCCTTAGT
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT

GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

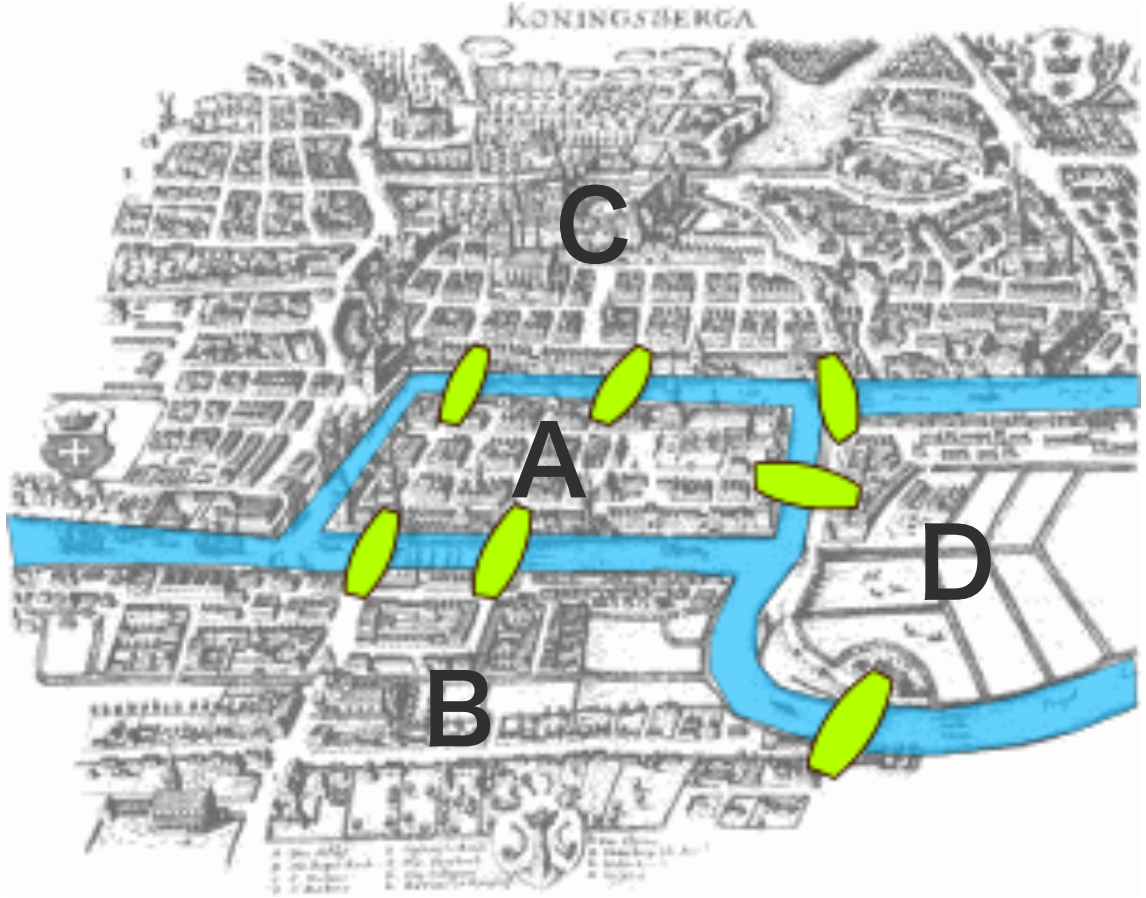
de novo assembly

- Hamiltonian path problem

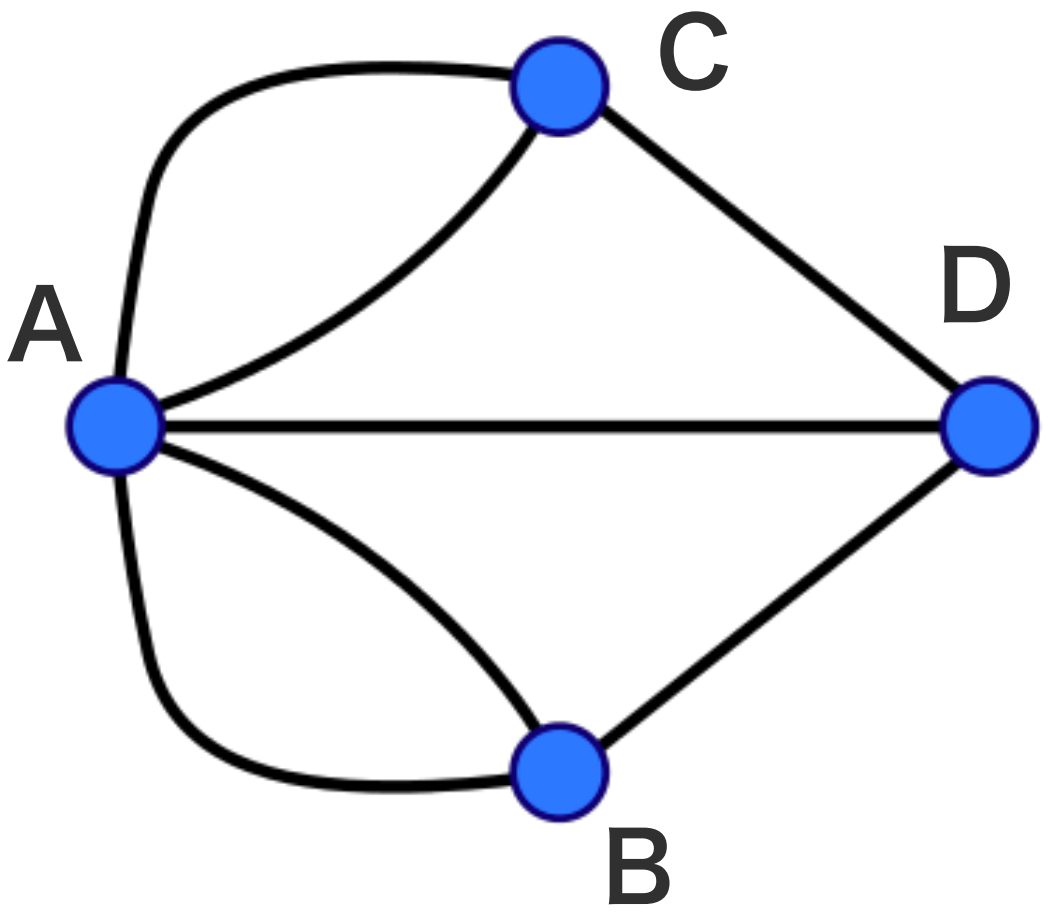


de novo assembly

Hamiltonian path problem



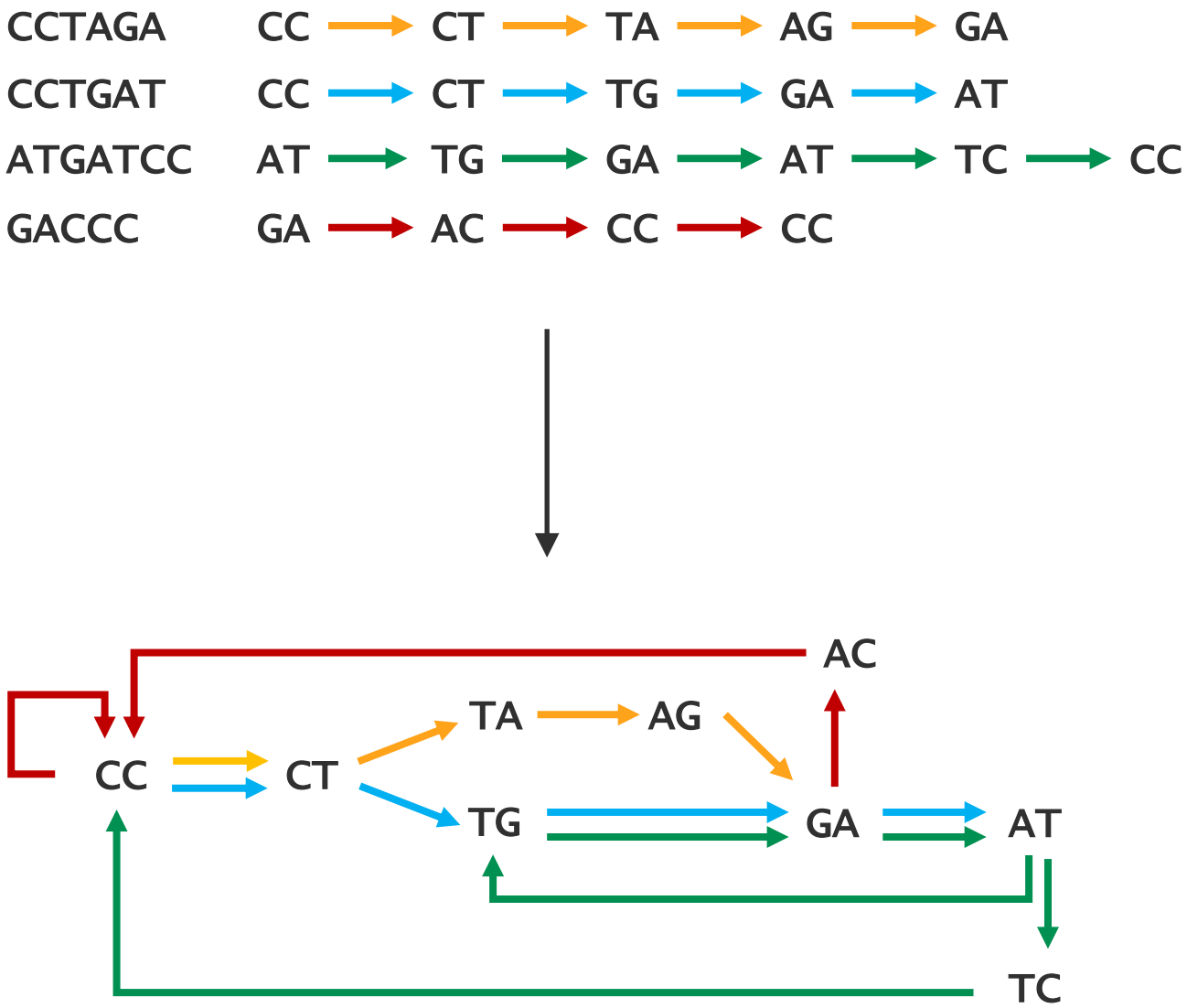
Eulerian path



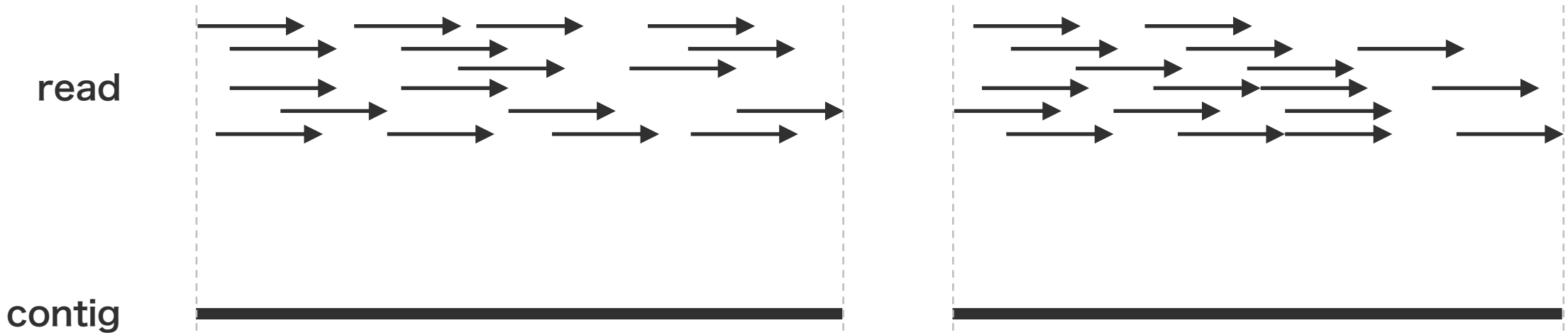
(c) wikipedia

de novo genome assembly

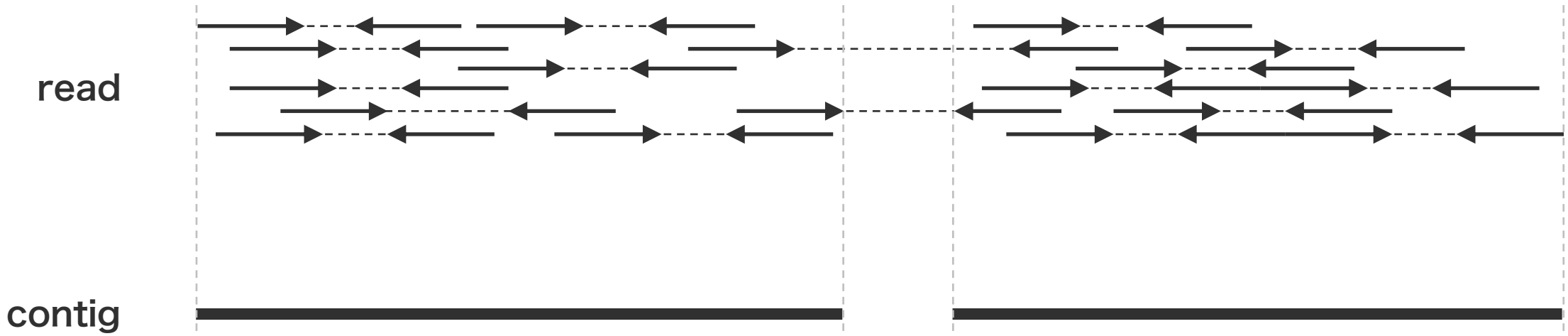
- Eulerian path



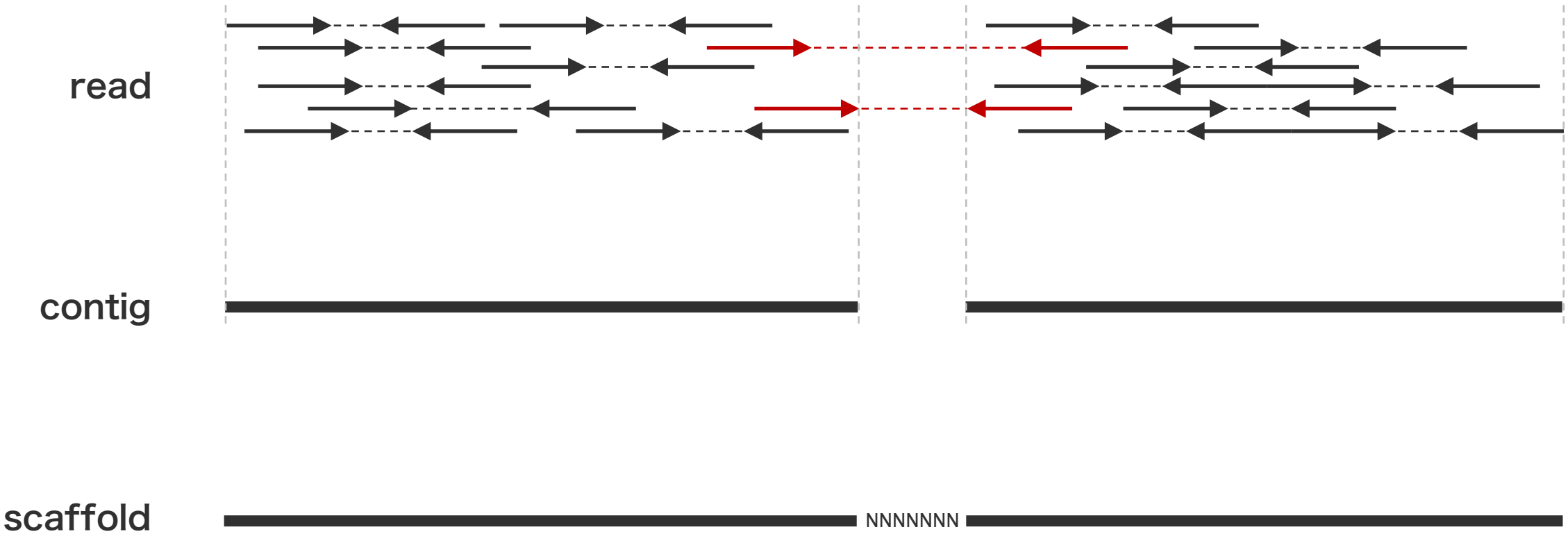
de novo assembly



de novo assembly



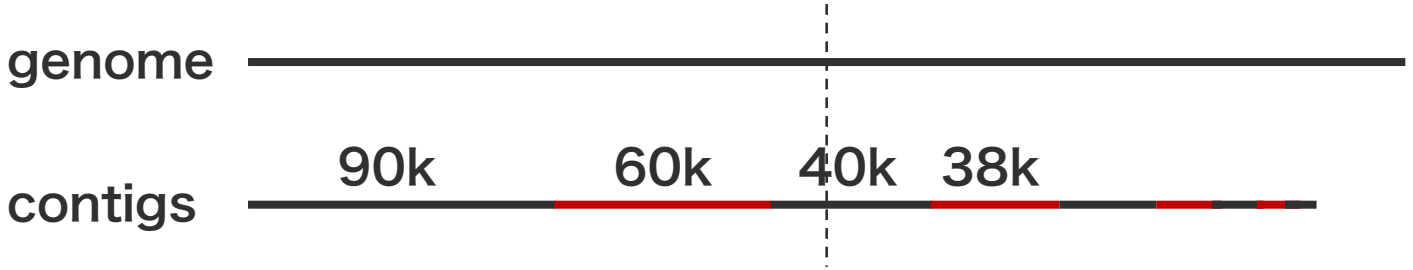
de novo assembly



assembly quality

Table 1. Assembly statistics of IWGSC RefSeq v1.0.

Assembly characteristics	Values
Assembly size	14.5 Gb
Number of scaffolds	138,665
Size of assembly in scaffolds ≥ 100 kb	14.2 Gb
Number of scaffolds ≥ 100 kb	4,443
N50 contig length	51.8 kb
Contig L50 number	81,427
N90 contig length	11.7 kb
Contig L90 number	294,934
Largest contig	580.5 kb
Ns in contigs	0
N50 scaffold length	7.0 Mb
Scaffold L50 number	571
N90 scaffold length	1.2 Mb
Scaffold L90 number	2,390
Largest scaffold	45.8 Mb
Ns in scaffolds	261.9 Mb
Gaps filled with BAC sequences	183 (1.7 Mb)
Average size of inserted BAC sequence	9.5 kb
N50 superscaffold length	22.8 Mb
Superscaffold L50 number	166
N90 superscaffold length	4.1 Mb
Superscaffold L90 number	718
Largest superscaffold	165.9 Mb
Sequence assigned to chromosomes	14.1 Gb (96.8%)
Sequence ≥ 100 kb assigned to chromosomes	14.1 Gb (99.1%)
Number of superscaffolds on chromosomes	1,601
Number of oriented superscaffolds	1,243
Length of oriented sequence	13.8 Gb (95%)
Length of oriented sequence ≥ 100 kb	13.8 Gb (97.3%)
Smallest number of superscaffolds per subgenome chromosome	35 (7A), 68 (2B), 36 (1D)
Largest number of superscaffolds per subgenome chromosome	111 (4A), 176 (3B), 90 (3D)
Average number of superscaffolds per chromosome	76

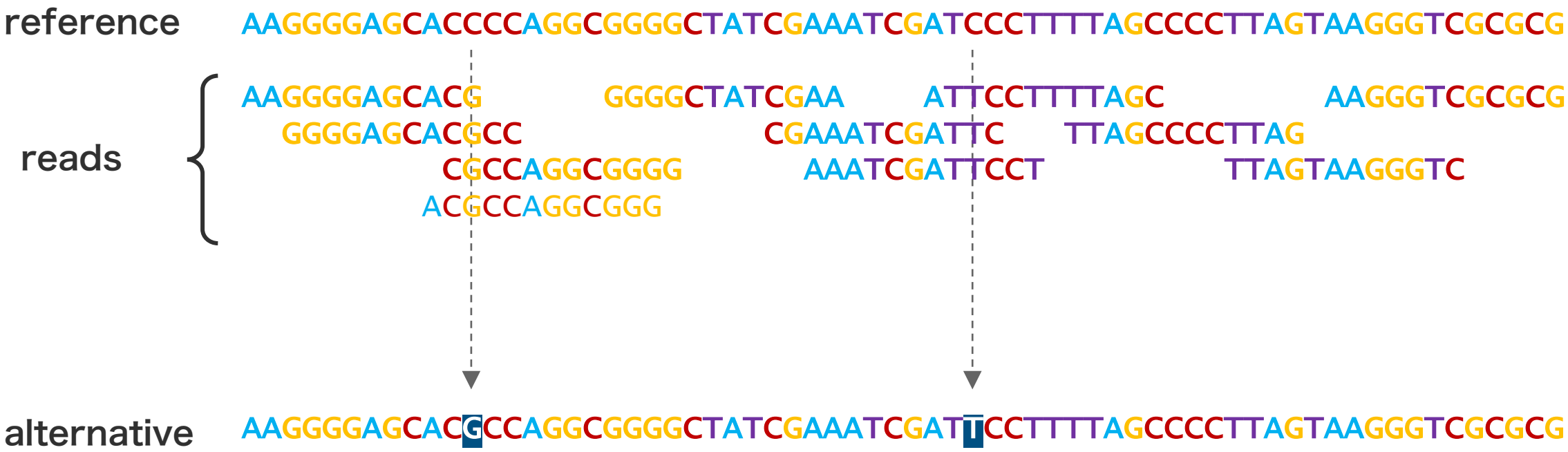


N50 = 40k

L50 = 3

IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191, 2018.

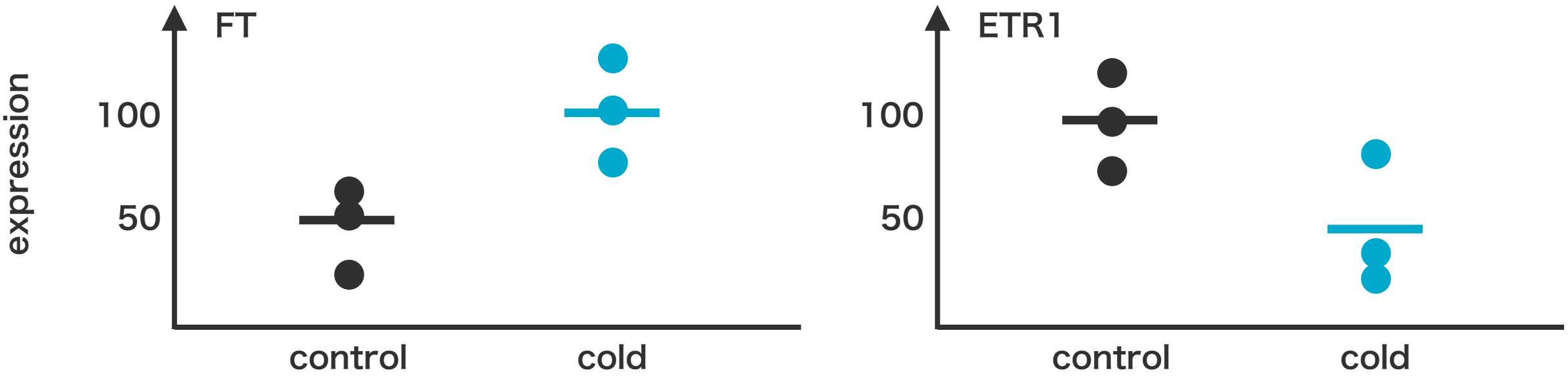
reference-guided assembly



Expression analysis

- DE analysis
- Functional analysis

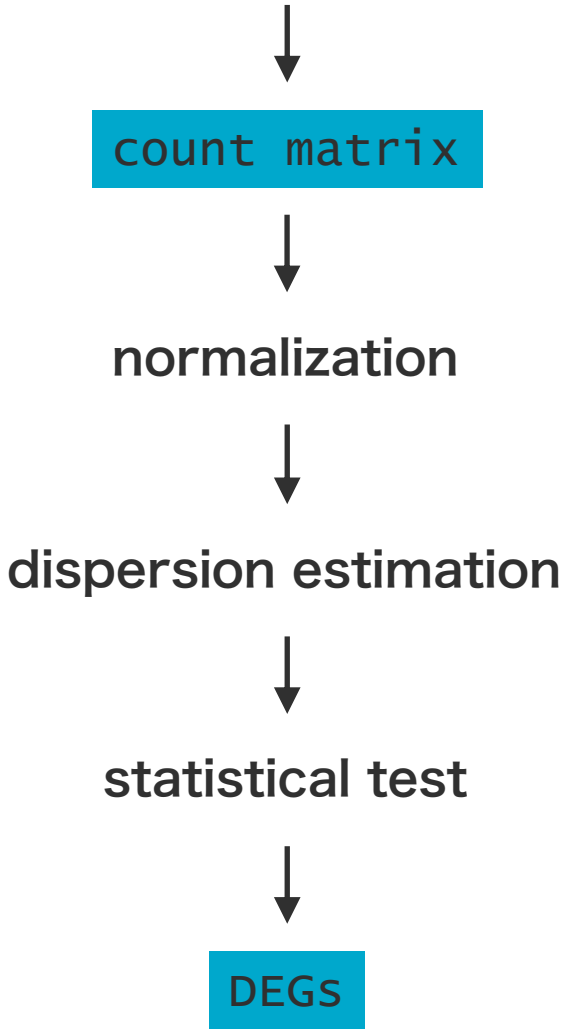
DE analysis



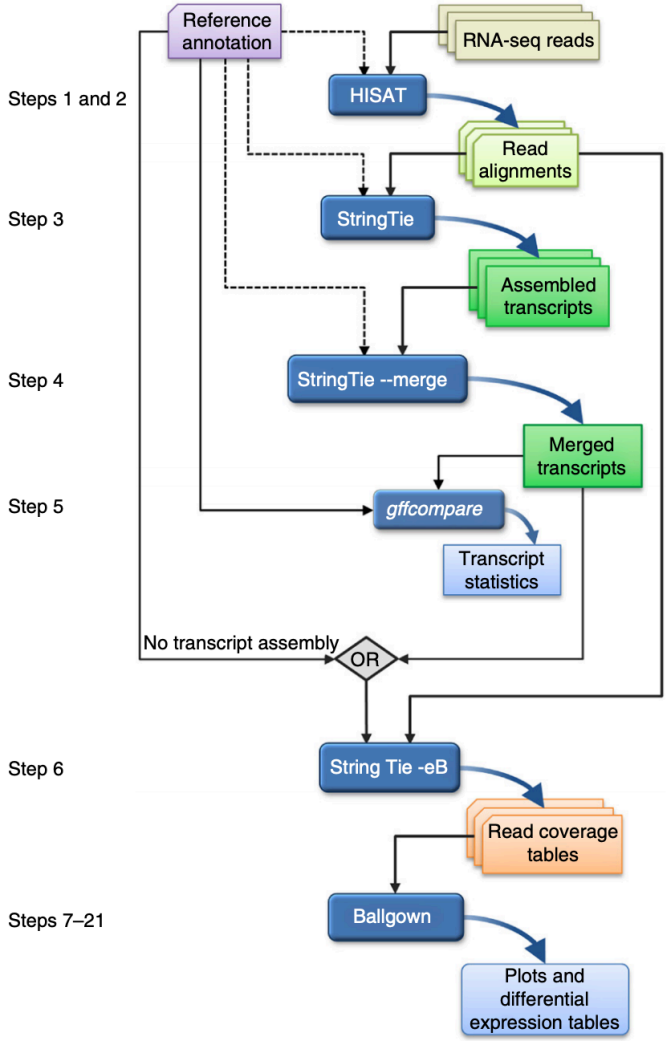
→ Cold treatment enhanced FT, whereas suppressed ETR1.

DE analysis

count-based

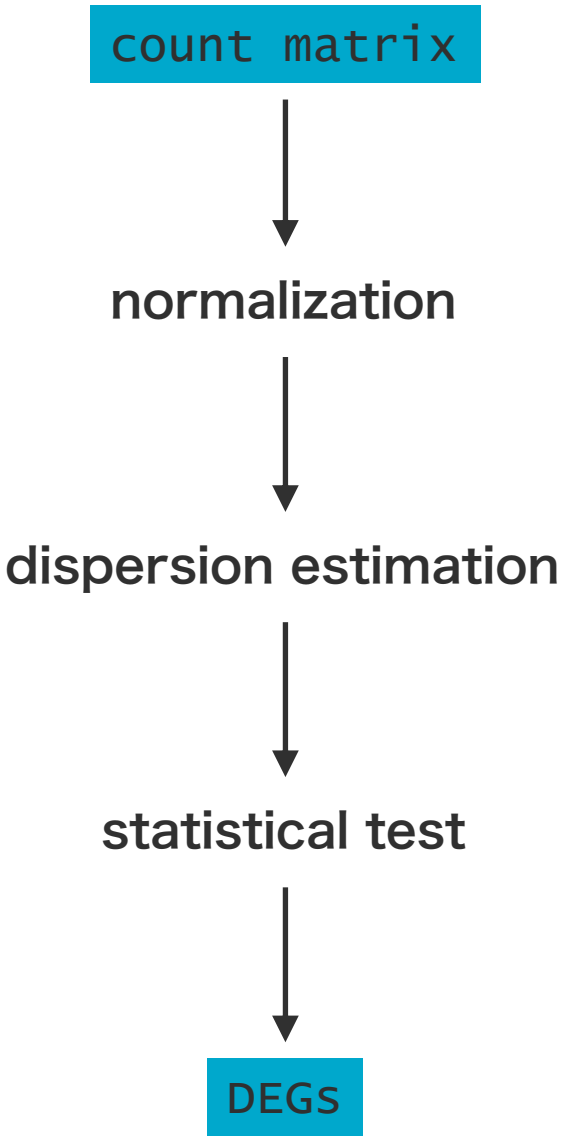


StringTie-Ballgown

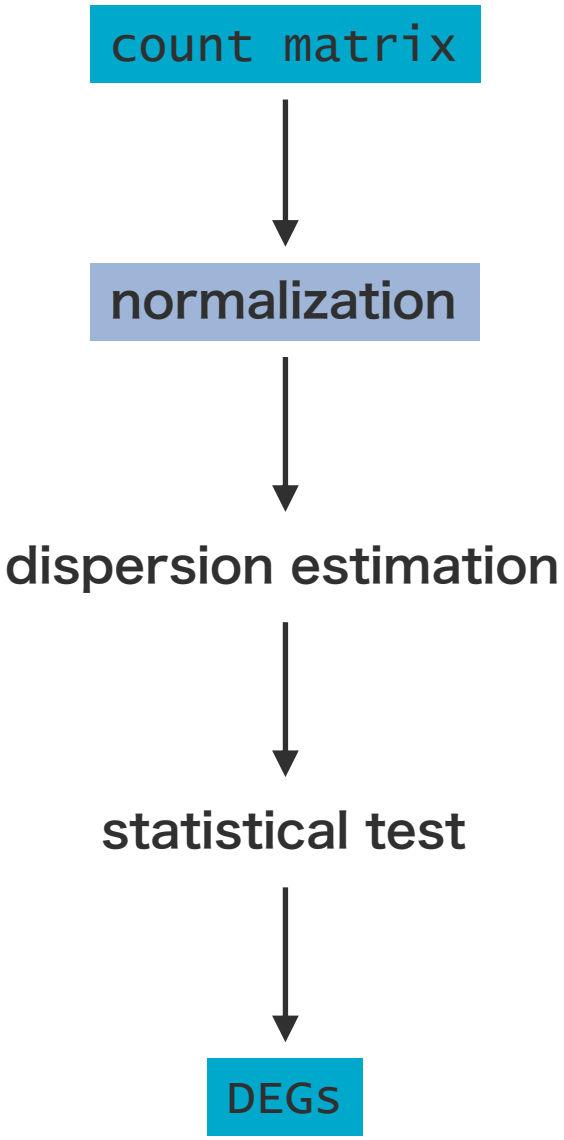


https://doi.org/10.1038/nprot.2016.095

DE analysis


count matrix
normalization
dispersion estimation
statistical test
DEGs

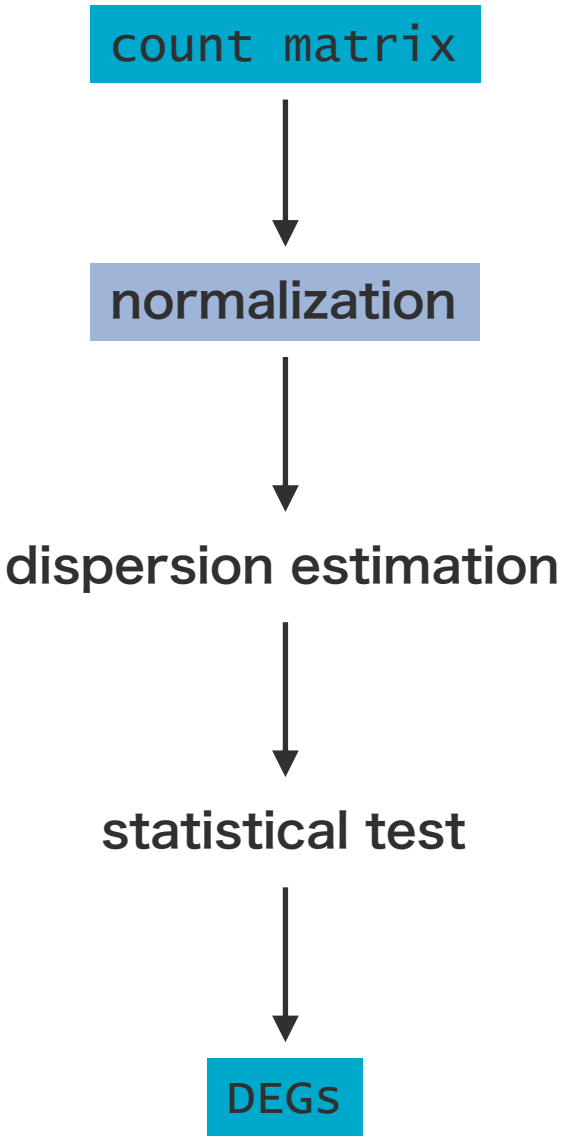
DE analysis / normalization



gene	ctrl 1	ctrl 2	ctrl 3	cold 1	cold 2	cold 3
FT	12	11	6	29	26	19
LFY	42	36	18	84	58	52
STM	3	4	1	17	33	23
PIN	32	23	14	42	38	25
EIN3	231	364	107	553	371	429
ERS1	96	103	45	182	91	85

❓ Most genes are up-regulated by cold treatment?

DE analysis / normalization



gene	ctrl 1	ctrl 2	ctrl 3	cold 1	cold 2	cold 3
FT	12	11	6	29	26	19
LFY	42	36	18	84	58	52
STM	3	4	1	17	33	23
PIN	32	23	14	42	38	25
EIN3	231	364	107	553	371	429
ERS1	96	103	45	182	91	85
lib size	416	541	191	907	617	633

normalization

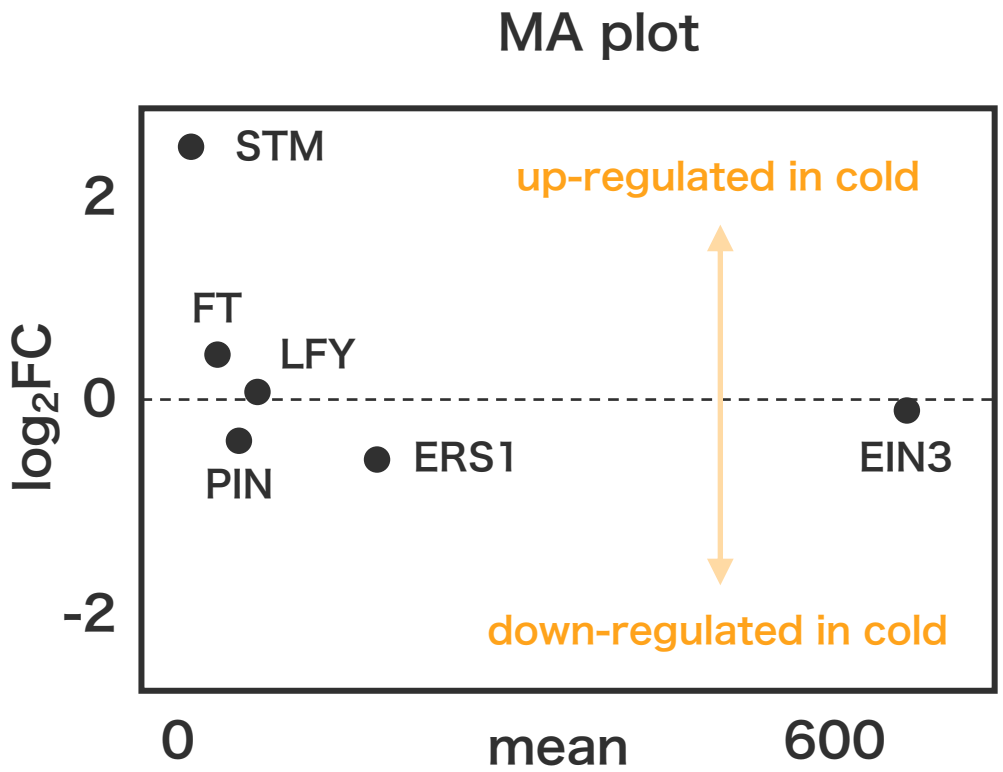
gene	ctrl 1	ctrl 2	ctrl 3	cold 1	cold 2	cold 3
FT	28.8	20.4	31.4	32.0	42.1	30.0
LFY	101.0	66.5	94.2	93.0	94.0	82.1
STM	7.2	7.4	5.3	18.7	53.5	36.3
PIN	76.9	42.5	73.3	45.3	61.6	39.5
EIN3	555.3	672.8	560.2	609.3	601.3	677.8
ERS1	230.8	190.4	235.6	201.7	147.5	134.3
lib size	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0

MA plot

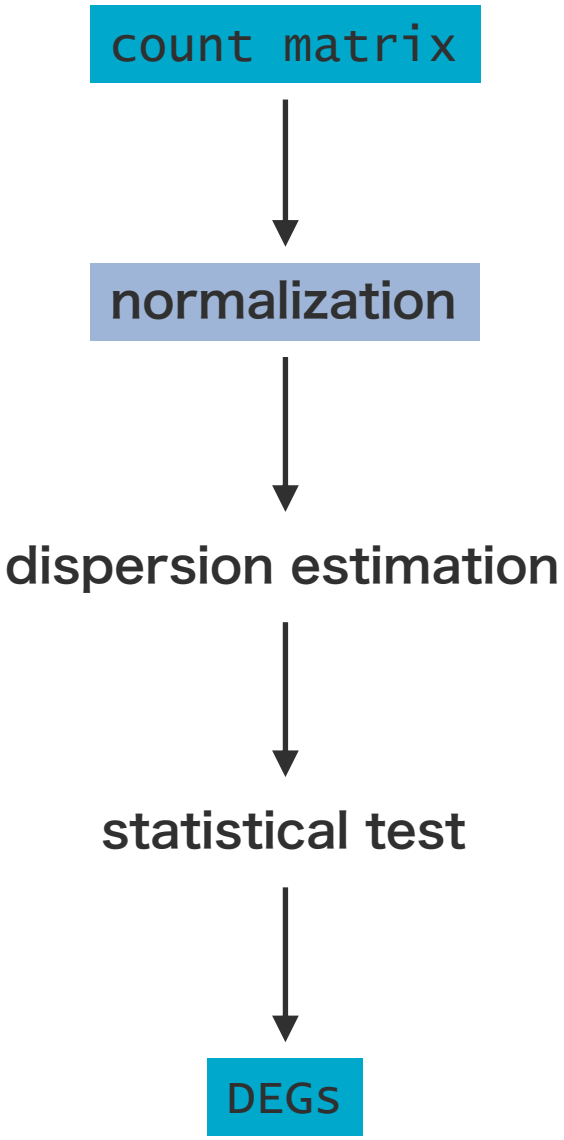
gene	ctrl 1	ctrl 2	ctrl 3	cold 1	cold 2	cold 3
FT	28.8	20.4	31.4	32.0	42.1	30.0
LFY	101.0	66.5	94.2	93.0	94.0	82.1
STM	7.2	7.4	5.3	18.7	53.5	36.3
PIN	76.9	42.5	73.3	45.3	61.6	39.5
EIN3	555.3	672.8	560.2	609.3	601.3	677.8
ERS1	230.8	190.4	235.6	201.7	147.5	134.3

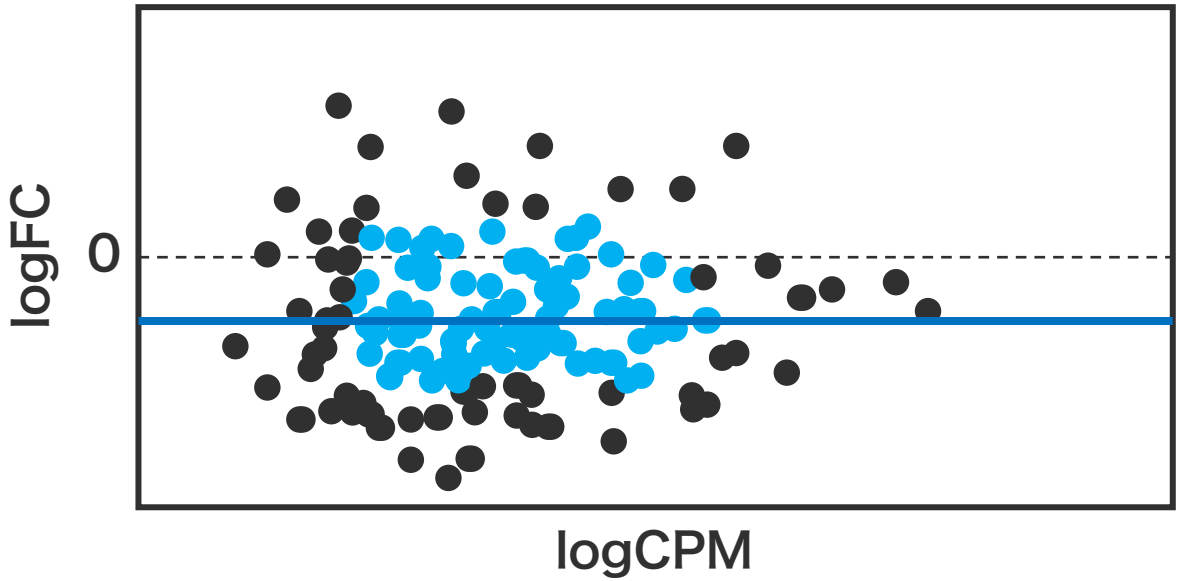
gene	ctrl	cold
FT	26.9	34.0
LFY	87.2	89.7
STM	6.6	36.2
PIN	64.2	48.8
EIN3	642.3	629.5
ERS1	218.9	161.2

gene	mean	$\log_2$FC
FT	30.5	0.34
LFY	88.5	0.04
STM	21.4	2.46
PIN	56.5	-0.396
EIN3	635.9	-0.029
ERS1	190.1	-0.441



DE analysis / count-based normalization





DE analysis / count-based normalization

count matrix

↓

normalization

↓

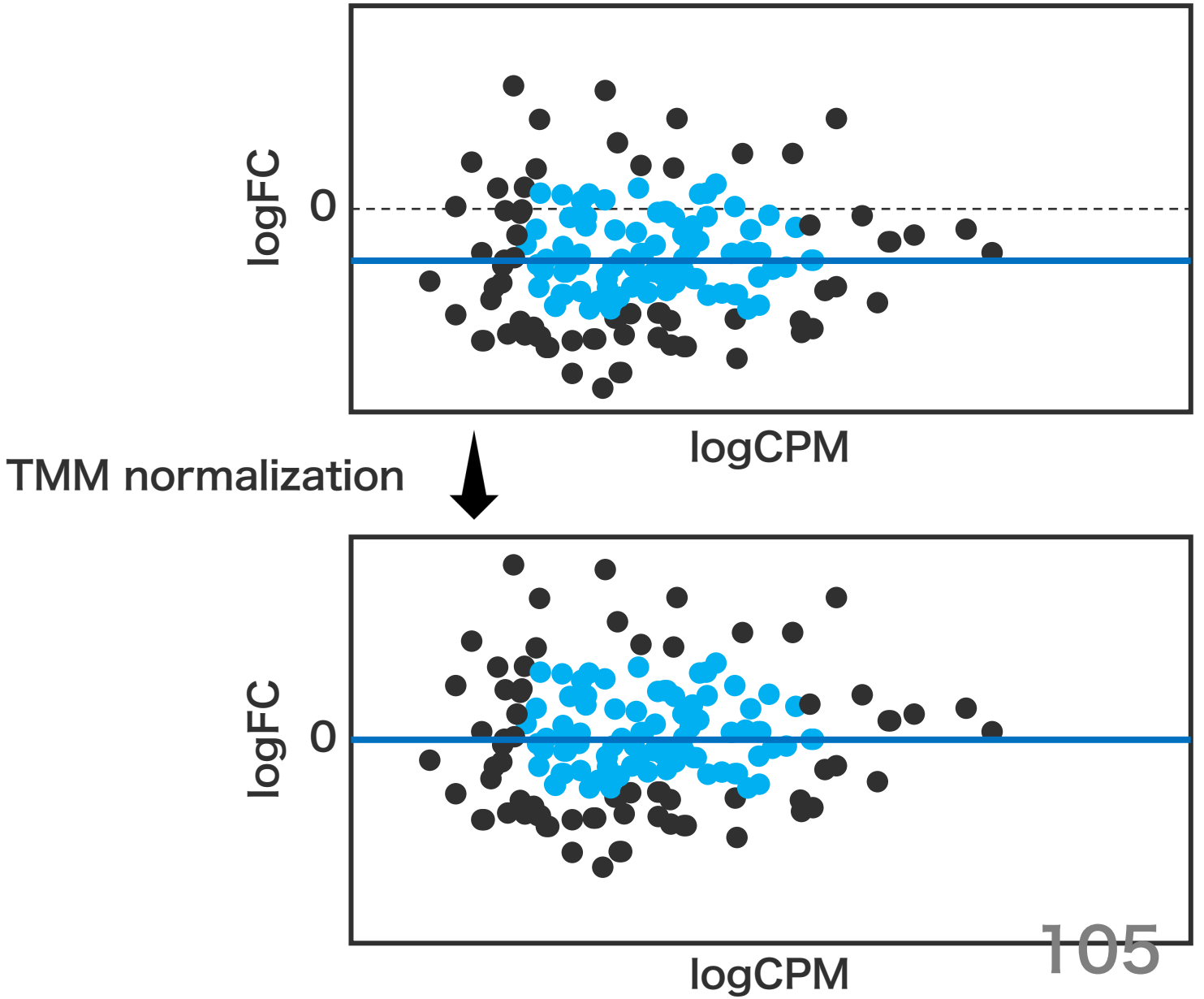
dispersion estimation

↓

statistical test

↓

DEGs



DE analysis / variance

count matrix

↓

normalization

↓

dispersion estimation

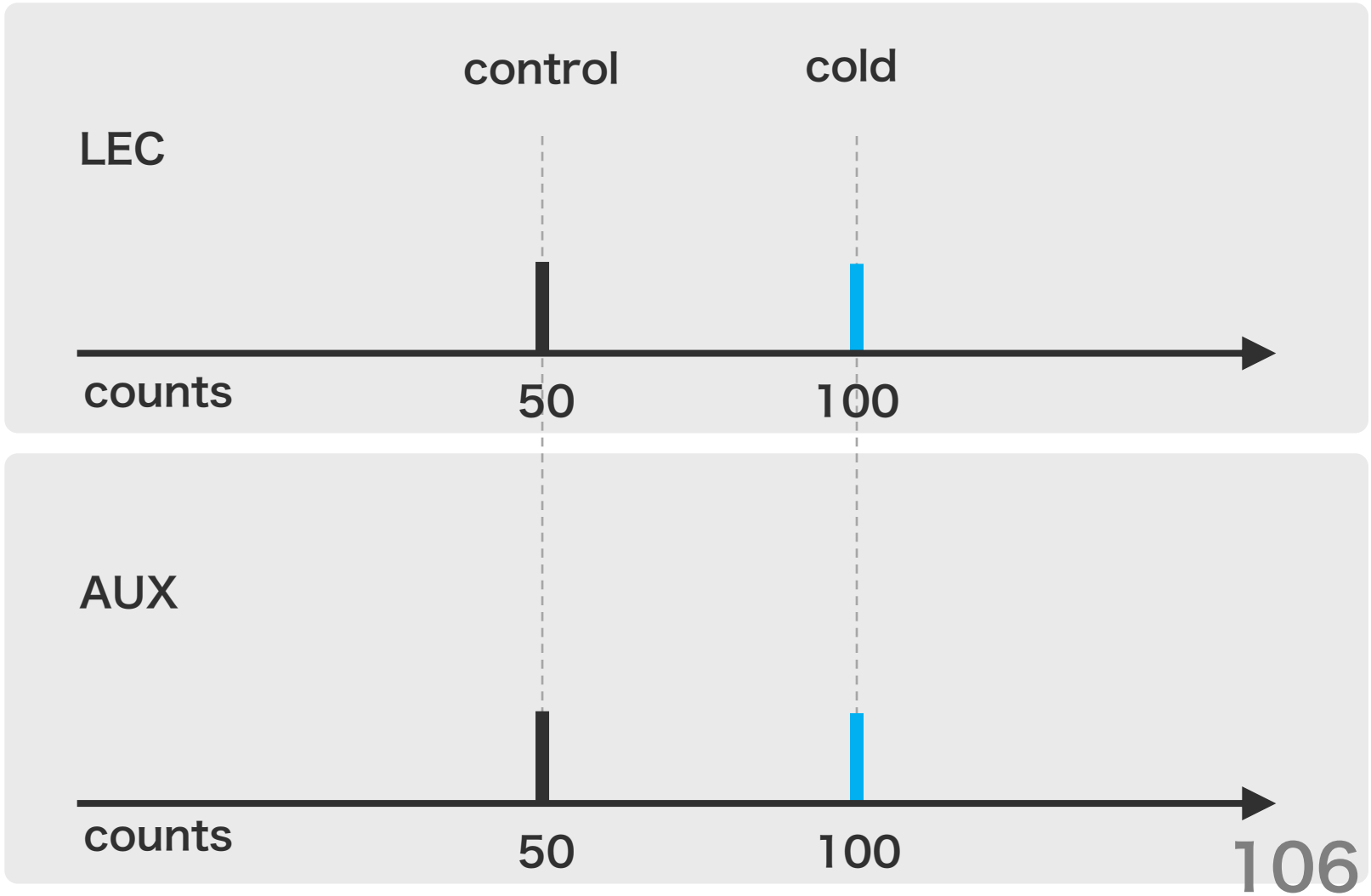
↓

statistical test

↓

DEGs

Are LEC and AUX up-regulated in cold condition?



DE analysis / variance

count matrix

↓

normalization

↓

dispersion estimation

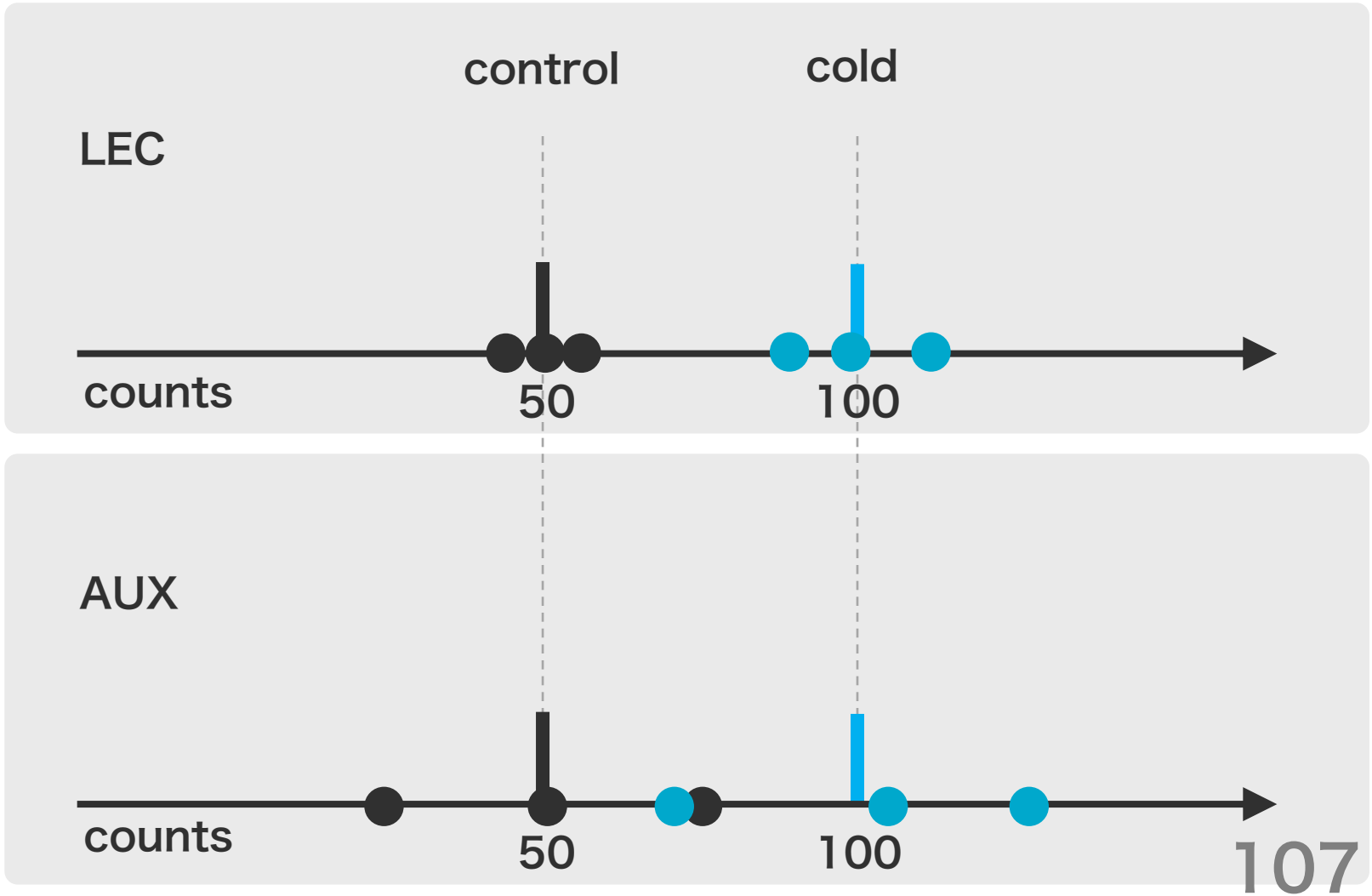
↓

statistical test

↓

DEGs

Estimating the variance with biological replicates is required!



DE analysis / variance

count matrix

↓

normalization

↓

dispersion estimation

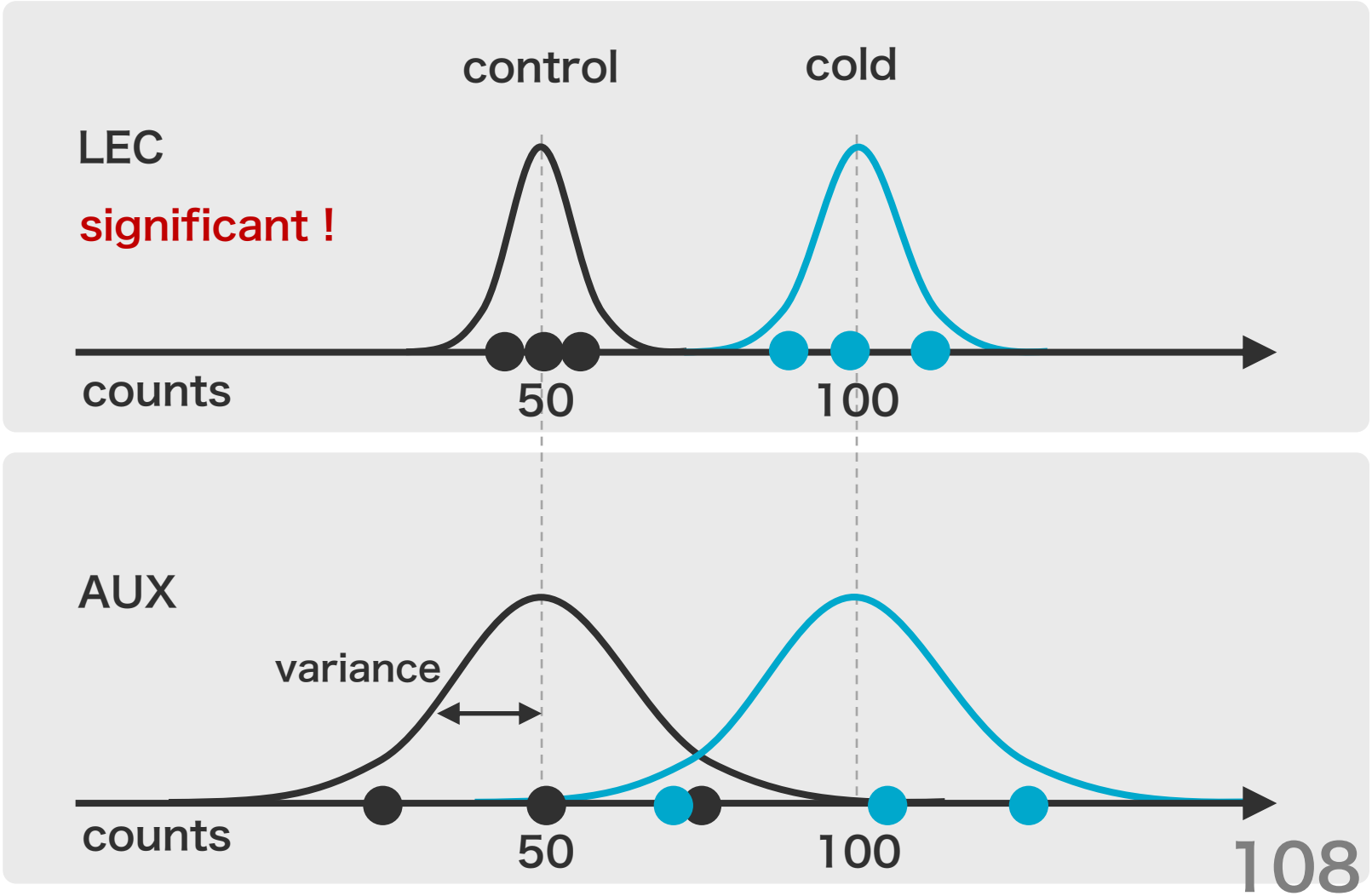
↓

statistical test

↓

DEGs

dispersion will be an important factor for DEGs identificcation



DE analysis / variance

count matrix

↓

normalization

↓

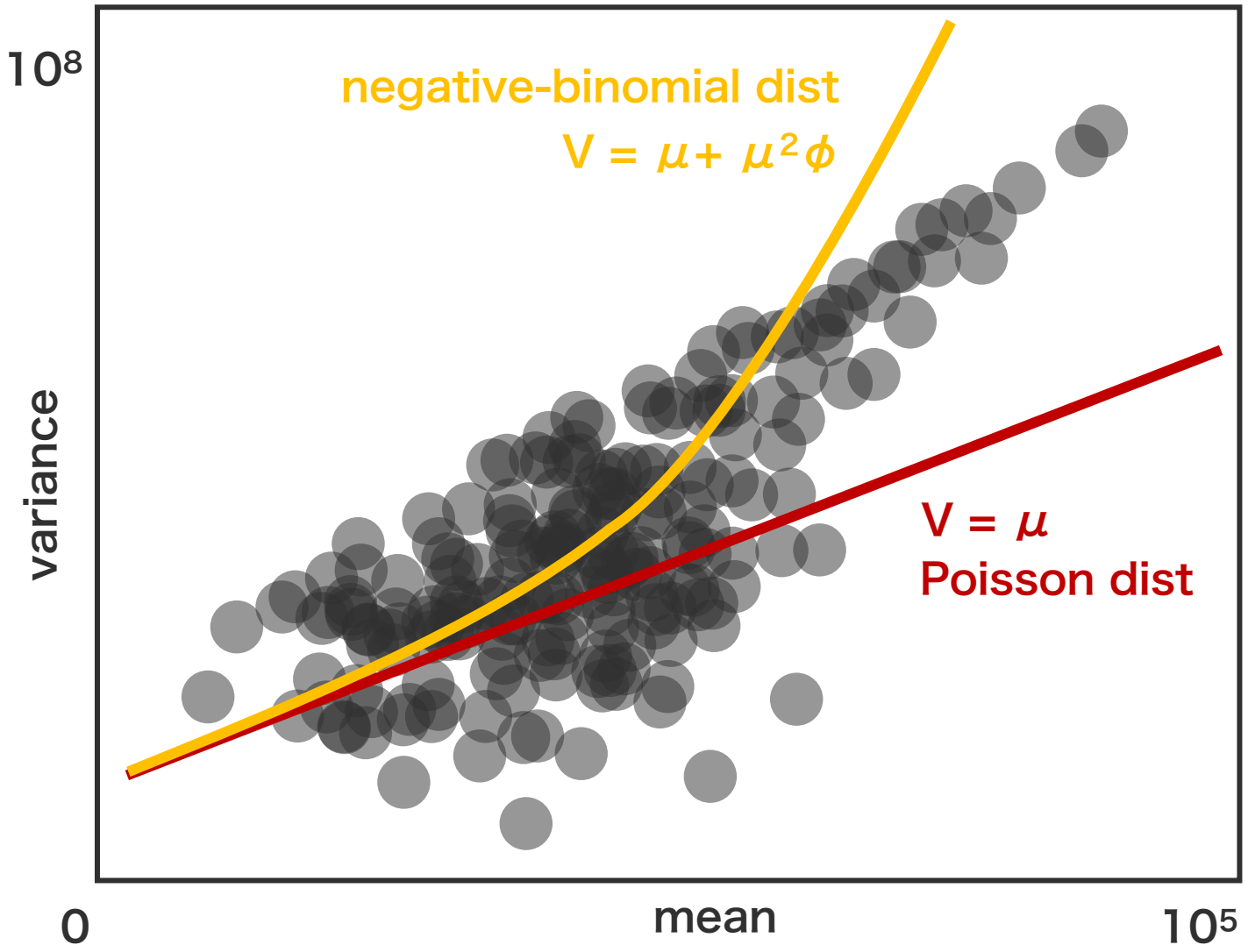
dispersion estimation

↓

statistical test

↓

DEGs



DE analysis / variance

count matrix

↓

normalization

↓

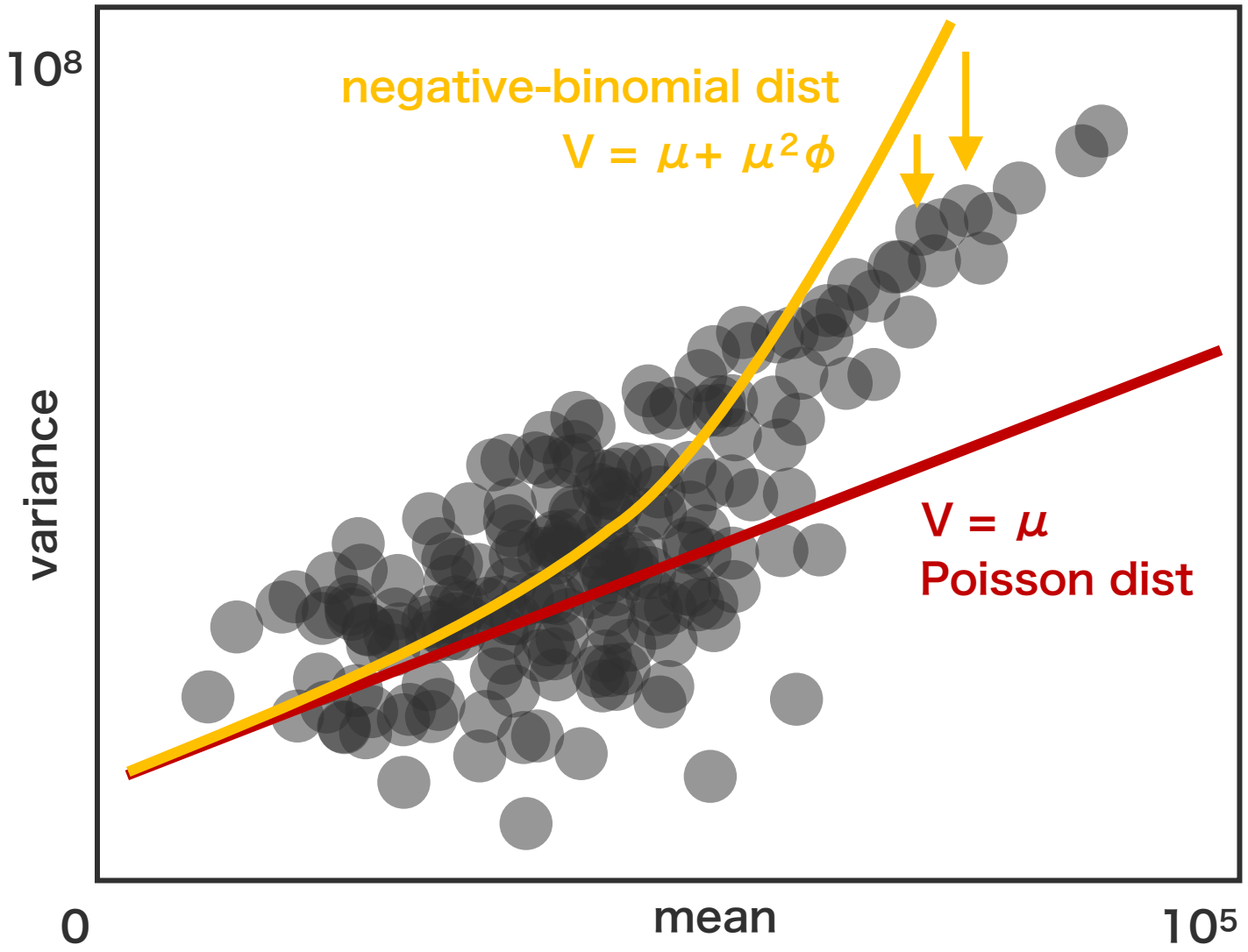
dispersion estimation

↓

statistical test

↓

DEGs



DE analysis / variance

count matrix

↓

normalization

↓

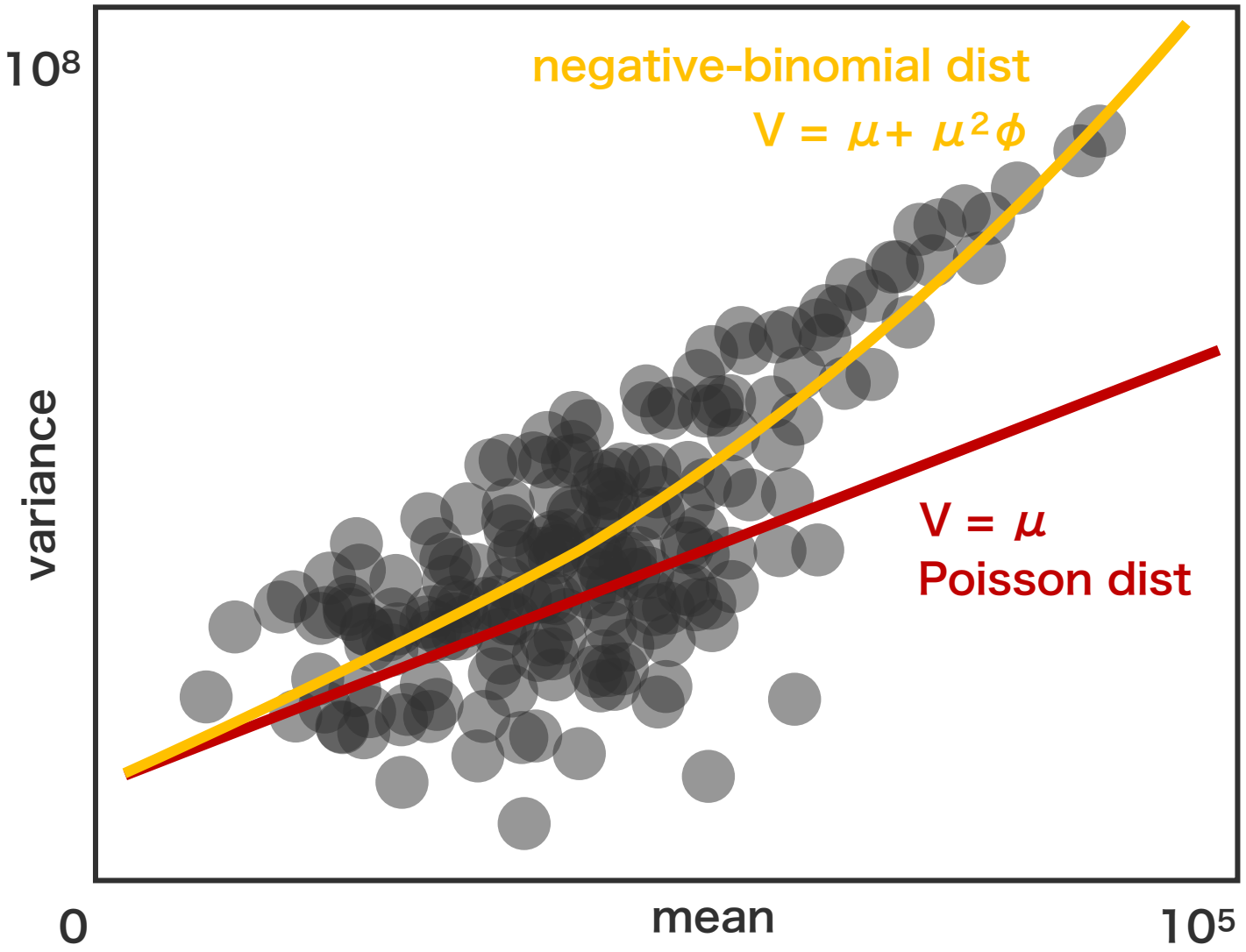
dispersion estimation

↓

statistical test

↓

DEGs



DE analysis / dispersion

count matrix

↓

normalization

↓

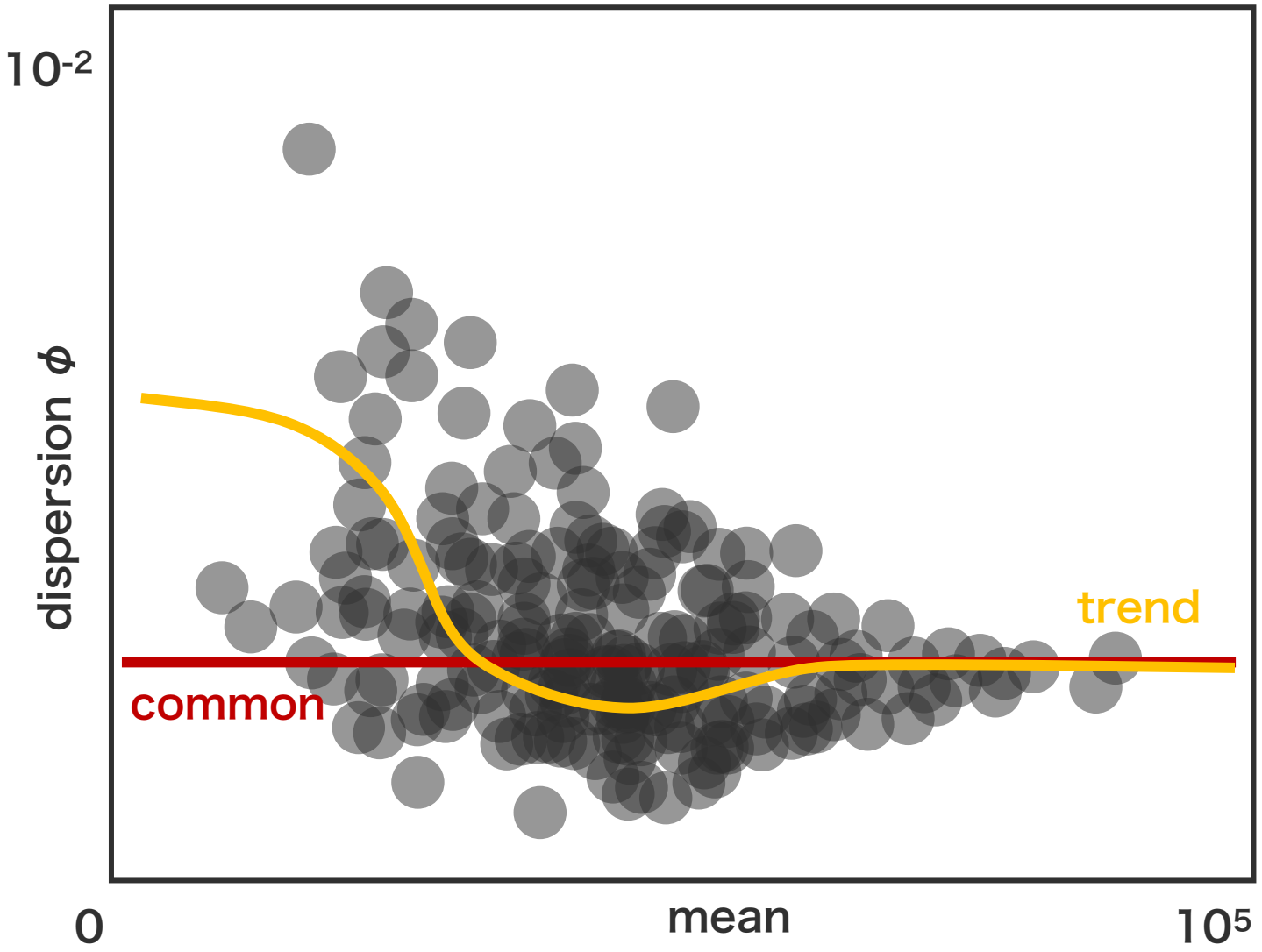
dispersion estimation

↓

statistical test

↓

DEGs



DE analysis / dispersion

count matrix

↓

normalization

↓

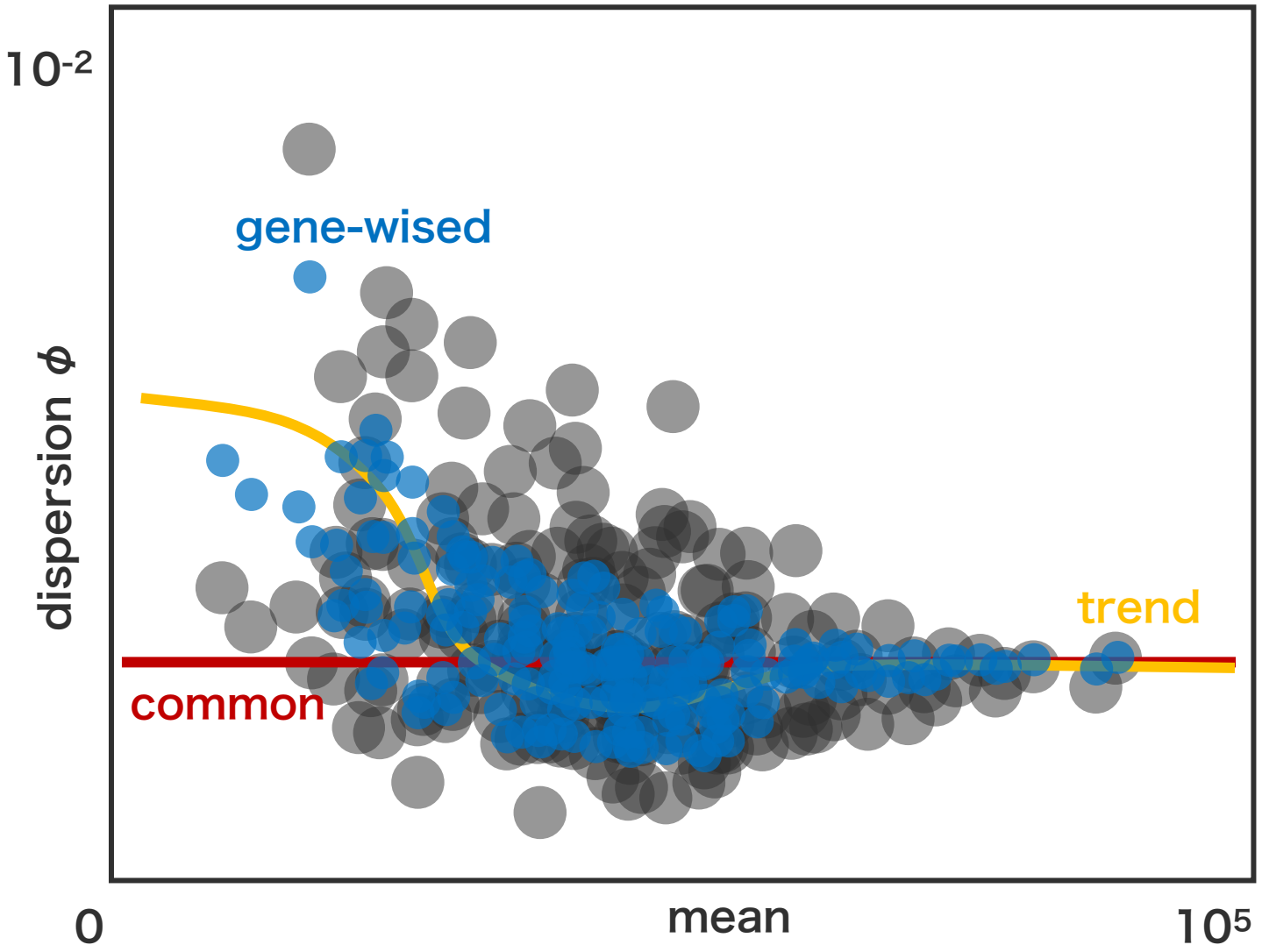
dispersion estimation

↓

statistical test

↓

DEGs



DE analysis / statistical test

count matrix

↓

normalization

↓

dispersion estimation

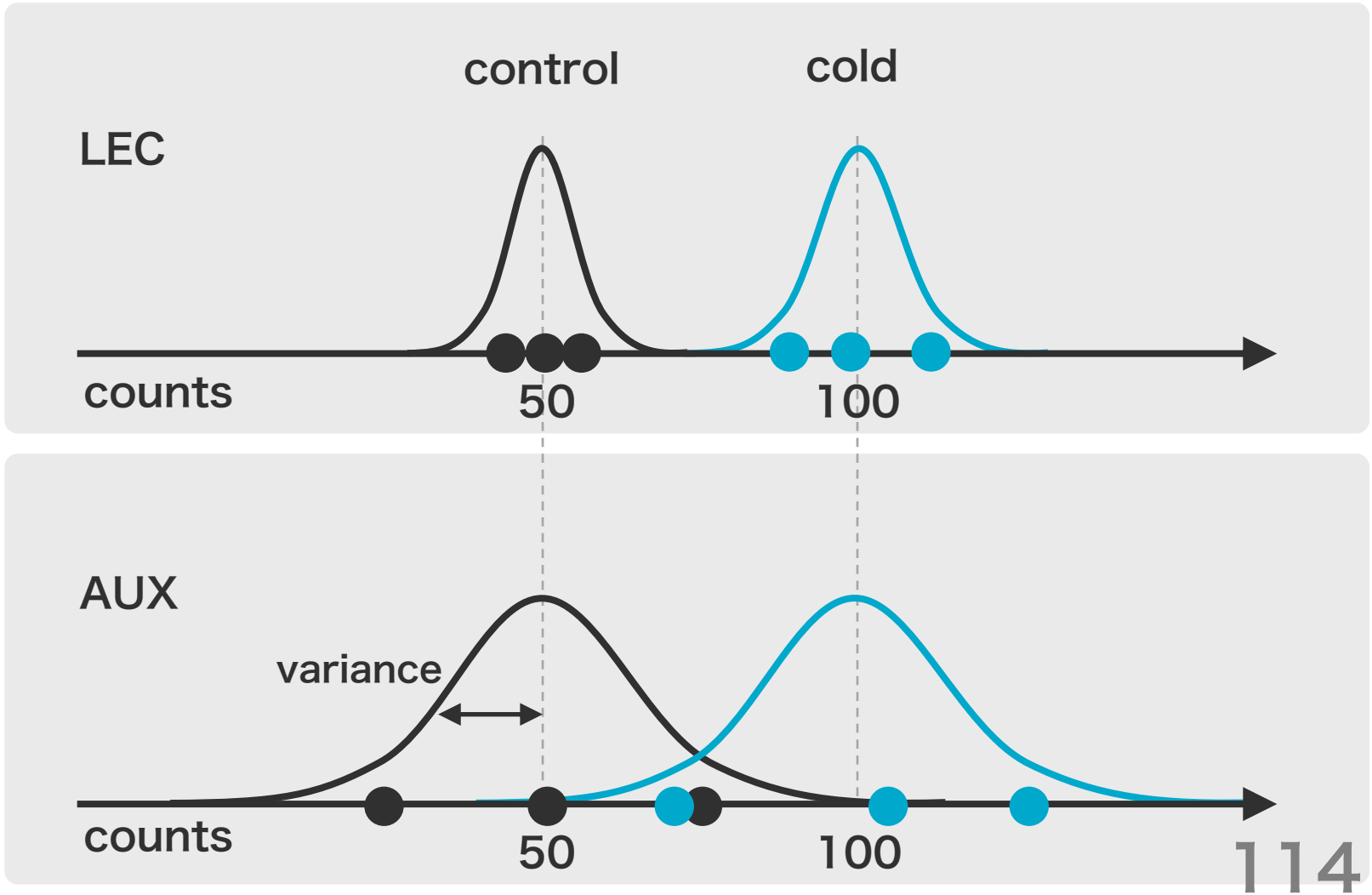
↓

statistical test

↓

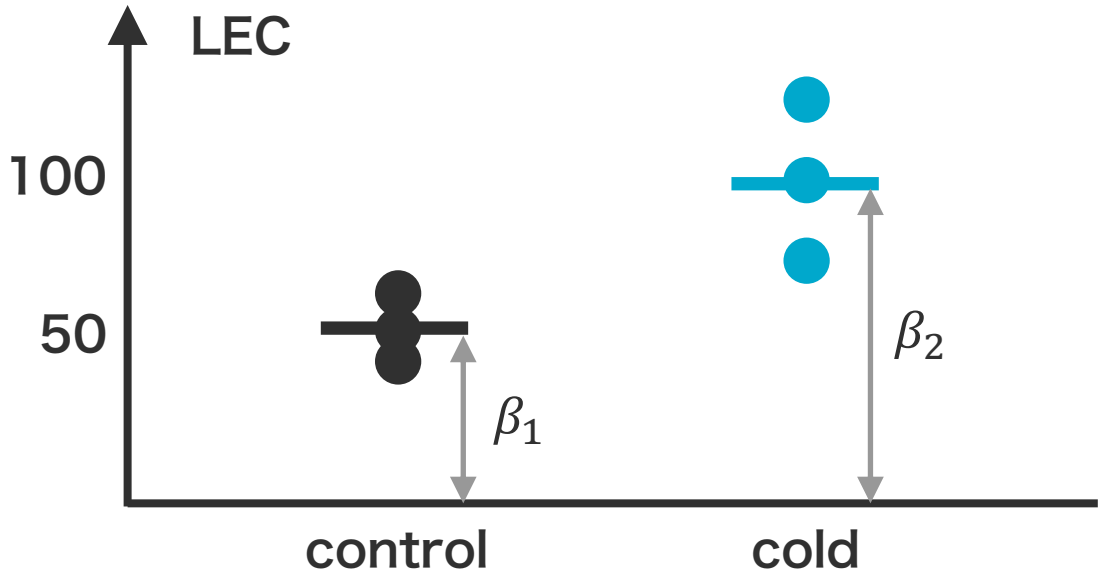
DEGs

statistical test will be performed against all genes independently



statistical test / GLM

How do we compare the expression between groups?



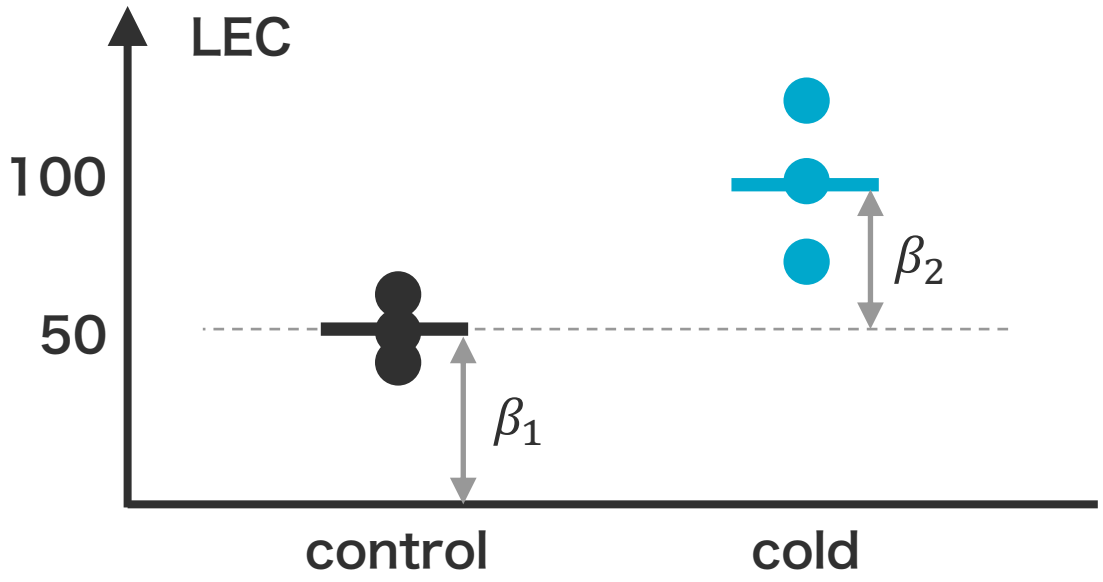
Assume that

$$E[ctrl] = \beta_1$$
$$E[cold] = \beta_2$$

then, test the null hypothesis $\beta_1 = \beta_2$.

statistical test / GLM

How do we compare the expression between groups?



Assume that

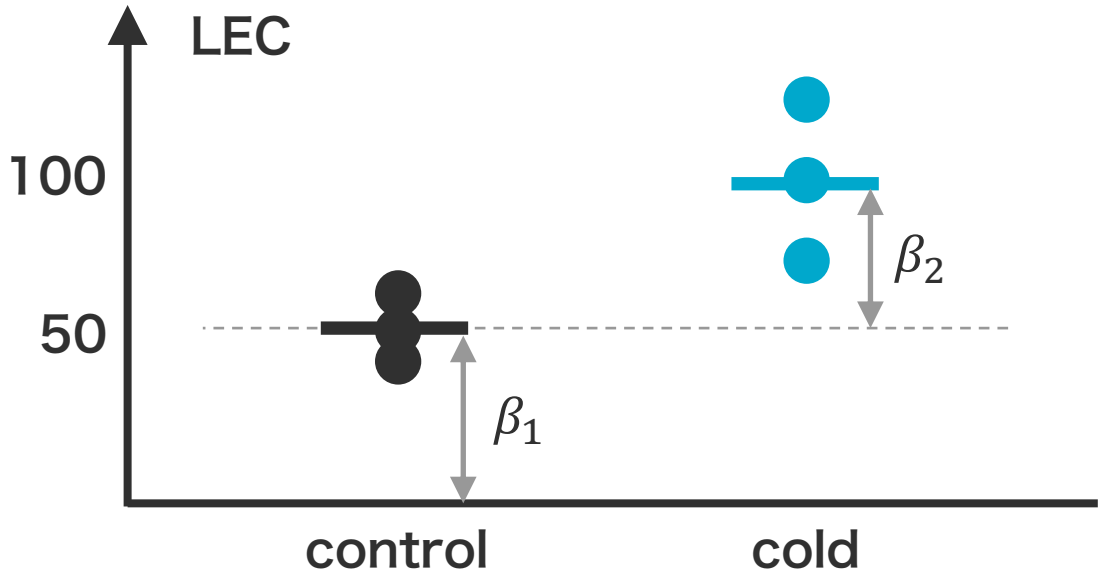
$$E[ctrl] = \beta_1$$

$$E[cold] = \beta_1 + \beta_2$$

then, test the null hypothesis $\beta_2 = 0$.

statistical test / GLM

How do we compare the expression between groups?



Assume that

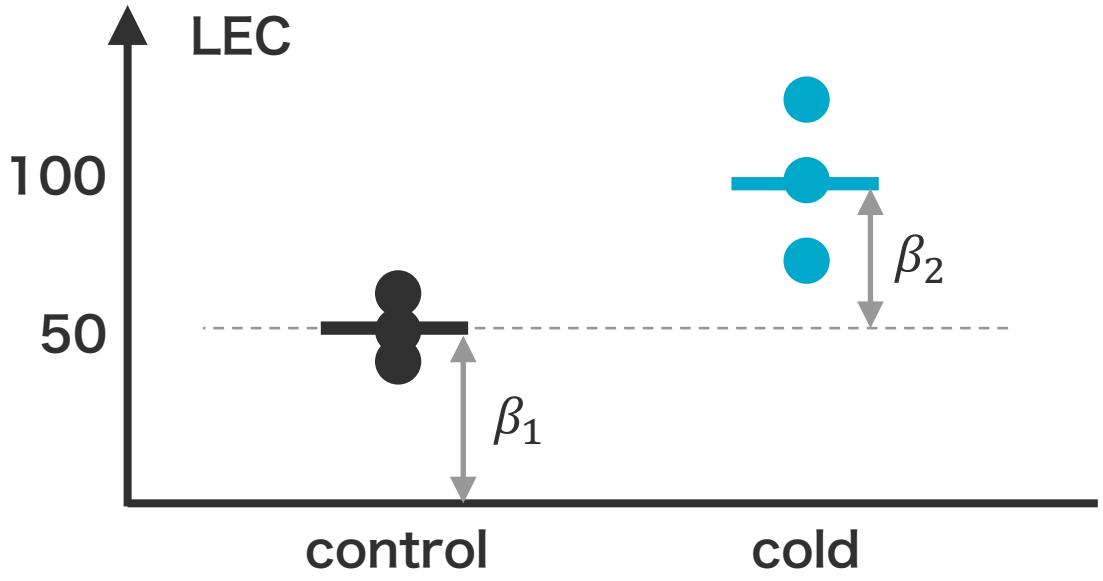
$$E[ctrl] = \beta_1$$
$$E[cold] = \beta_1 + \beta_2$$

then, test the null hypothesis $\beta_2 = 0$.
In this experiment, we have 6 replicates,

$$E[r1] = \beta_1$$
$$E[r2] = \beta_1$$
$$E[r3] = \beta_1$$
$$E[r4] = \beta_1 + \beta_2$$
$$E[r5] = \beta_1 + \beta_2$$
$$E[r6] = \beta_1 + \beta_2$$

statistical test / GLM

How do we compare the expression between groups?



Assume that

$$E[ctrl] = \beta_1$$
$$E[cold] = \beta_1 + \beta_2$$

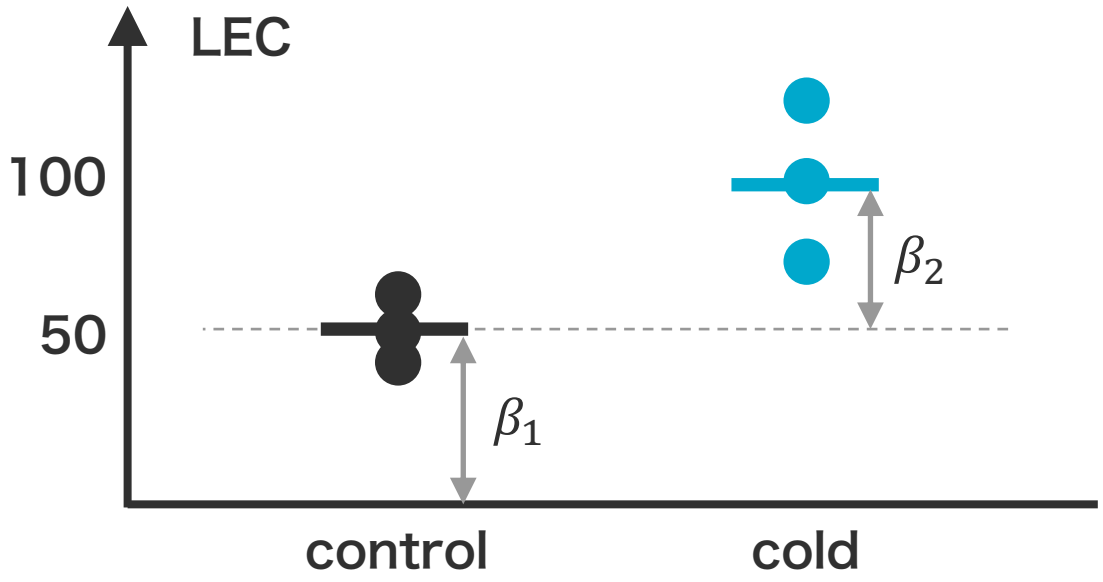
then, test the null hypothesis $\beta_2 = 0$.
In this experiment, we have 6 replicates,

$$(E[r1] \quad E[r2] \quad E[r3] \quad E[r4] \quad E[r5] \quad E[r6]) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

design matrix

statistical test / GLM

How do we compare the expression between groups?



To test null hypothesis $\beta_2 = 0$, we create two models and decide which one is better.

full model

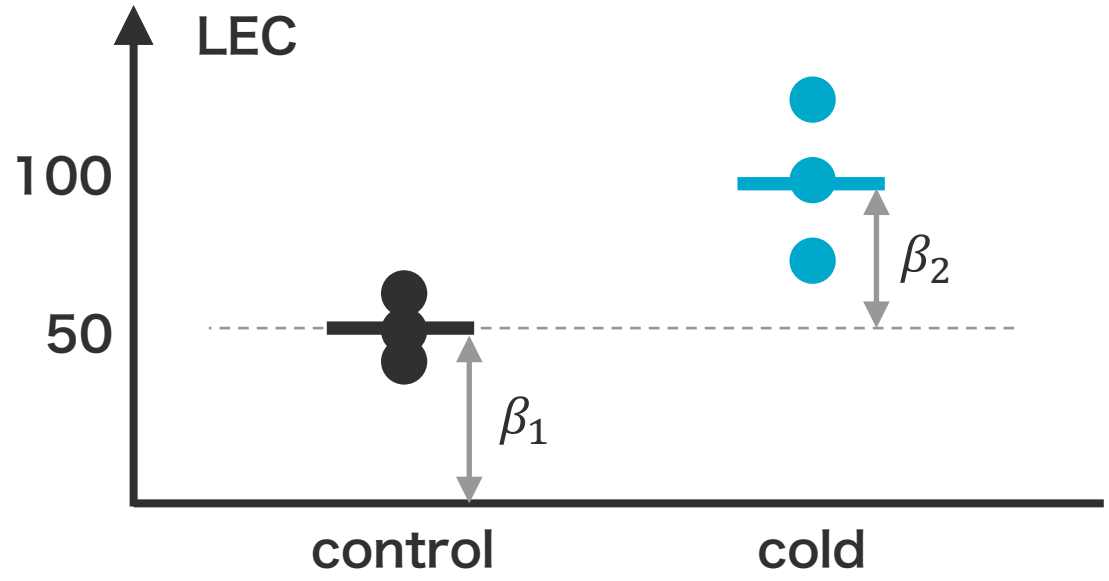
$$(E[r1] \quad E[r2] \quad E[r3] \quad E[r4] \quad E[r5] \quad E[r6]) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

reduced model

$$(E[r1] \quad E[r2] \quad E[r3] \quad E[r4] \quad E[r5] \quad E[r6]) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

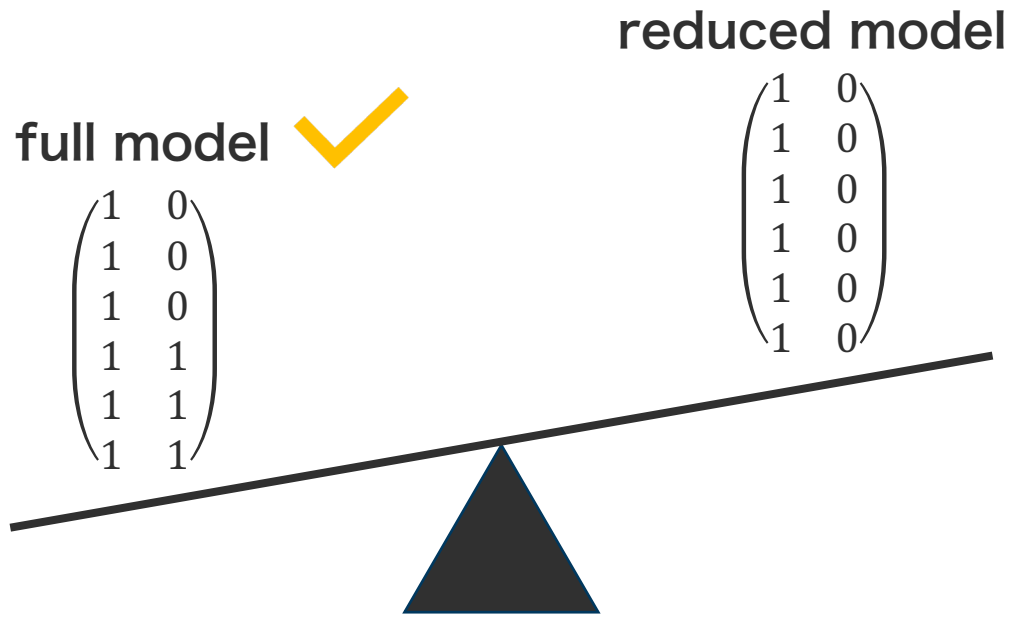
statistical test / GLM

How do we compare the expression between groups?



the null hypothesis $\beta_2 = 0$ can be rejected

To test null hypothesis $\beta_2 = 0$, we create two models and decide which one is better.



Likelihood ratio test (LRT) and Wald test can be used for comparing the two models.

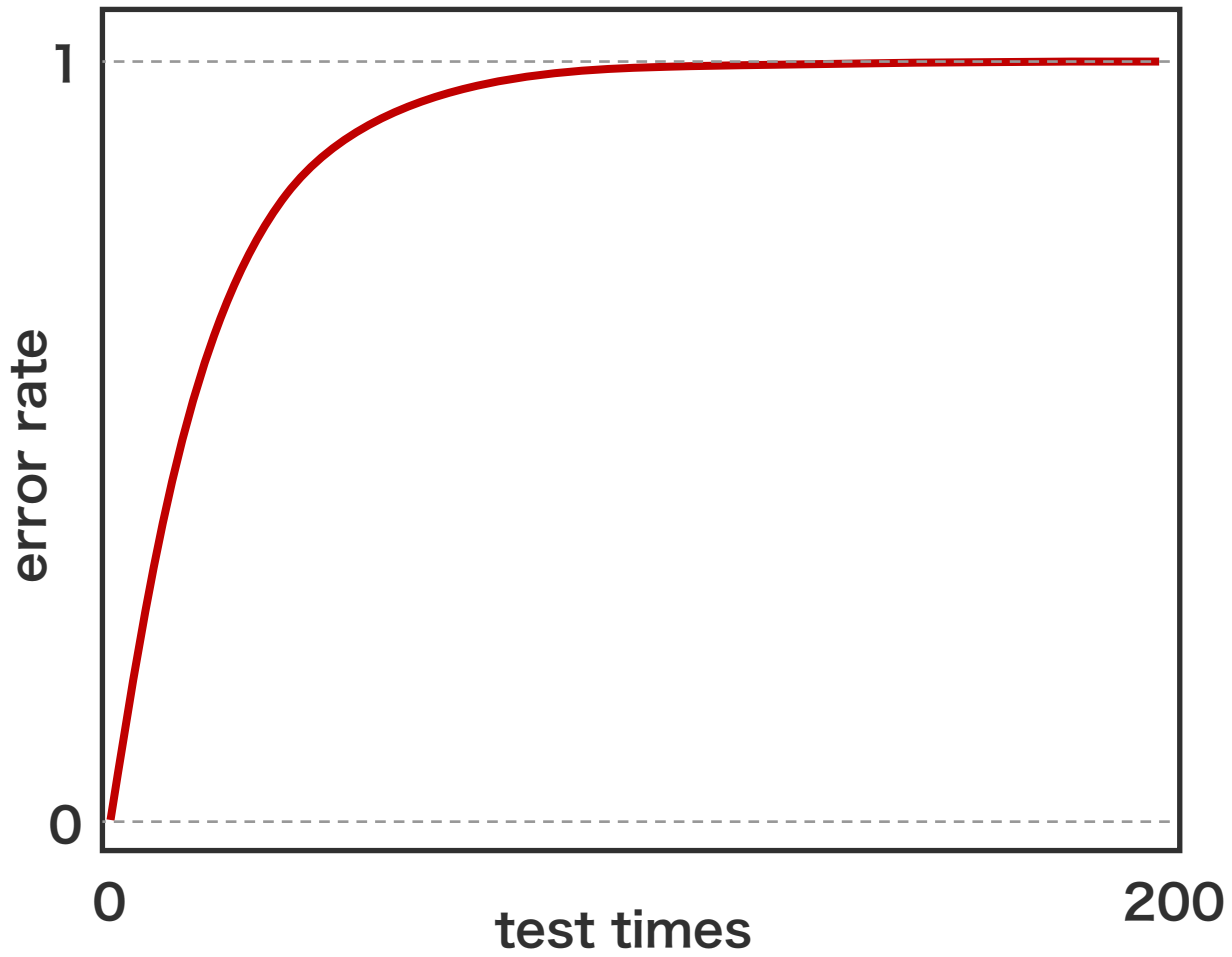
Multiple testing corrections

If we test a null hypothesis with cutoff 0.05, the type I error rate should be less than

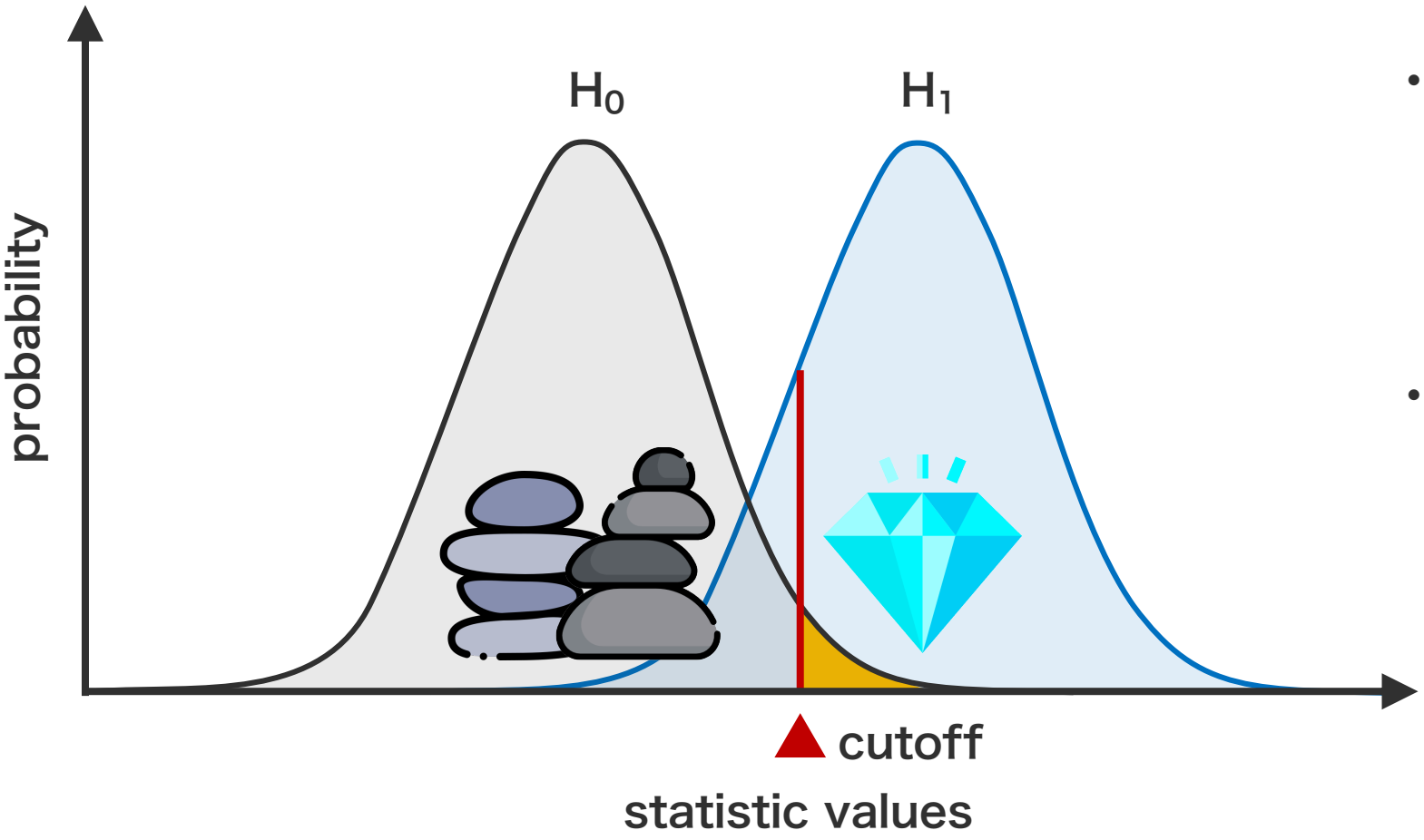
$$0.05$$

If we test two null hypotheses with cutoff 0.05 indenpendently, the type I error rate should be less than

$$1-(1-0.05)^2=0.0975$$

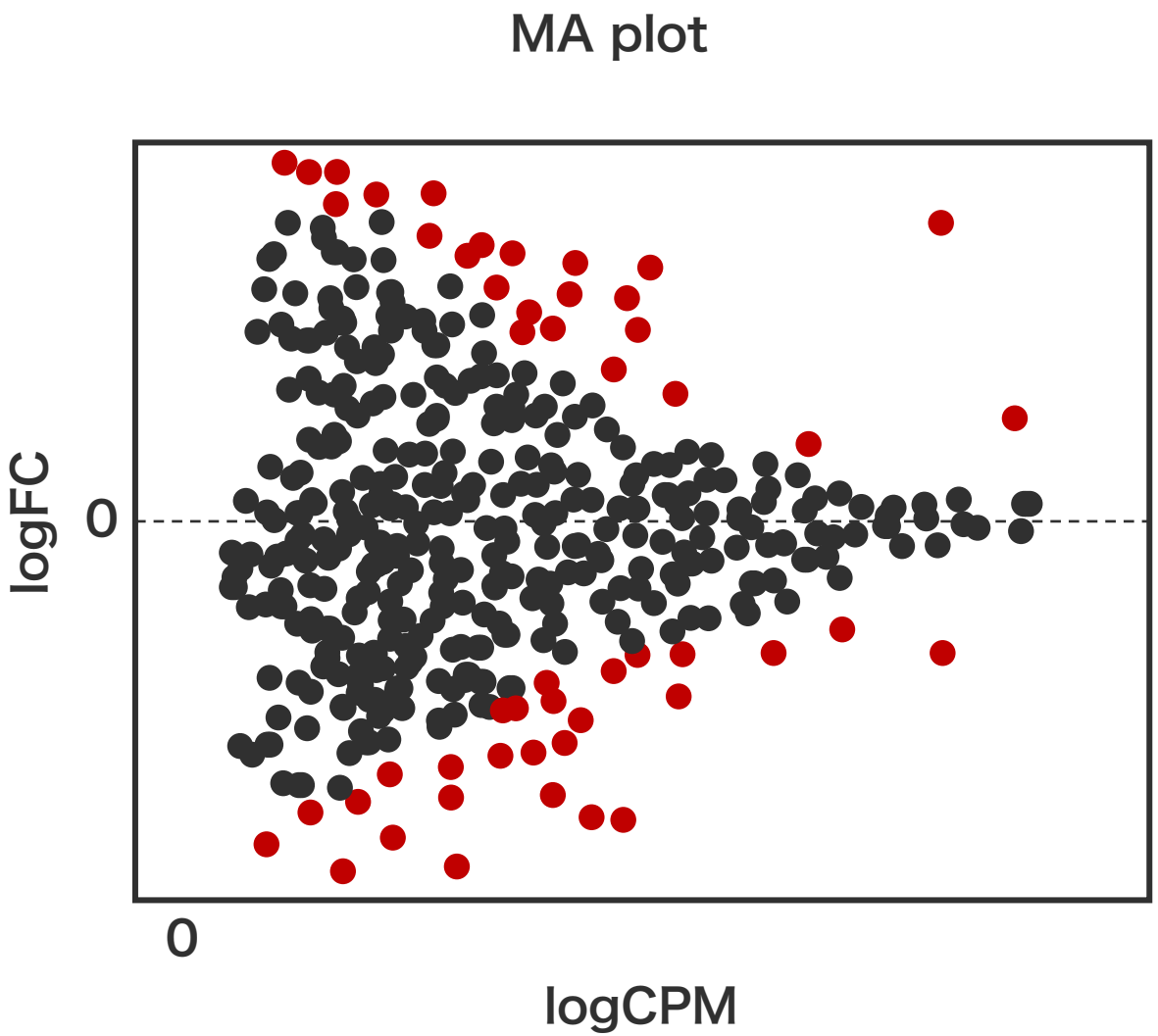


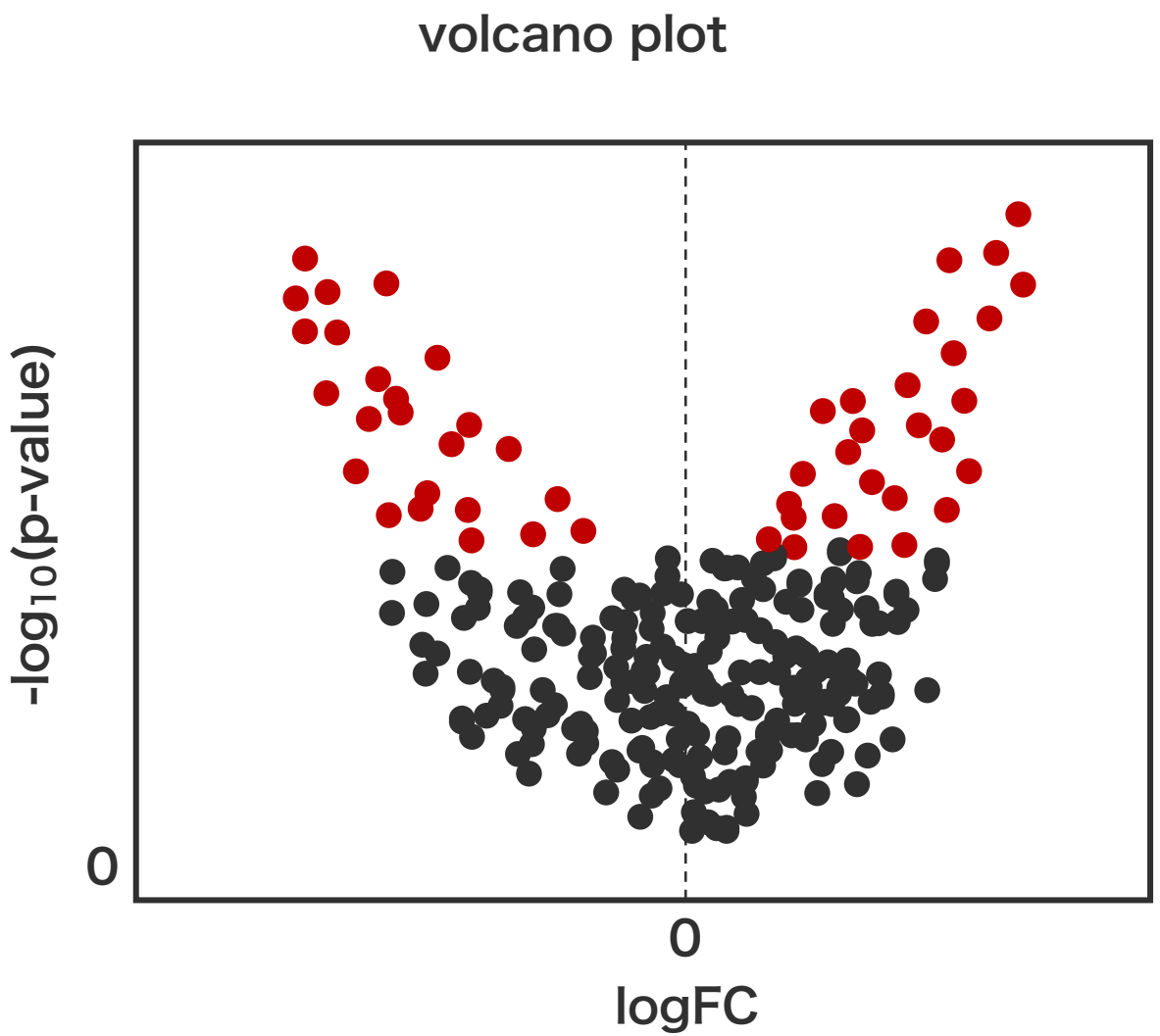
FDR



- Type I error
  - the ratio of the orange area diveided by the grey area
- FDR (flase discovery rate)
  - the ratio of the orange area diveded by the blue area on the right side of red line

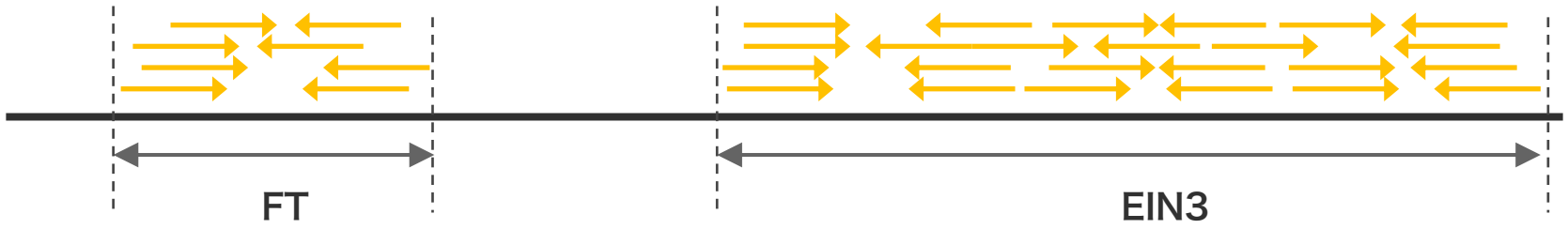
DE analysis





Normalization / TPM

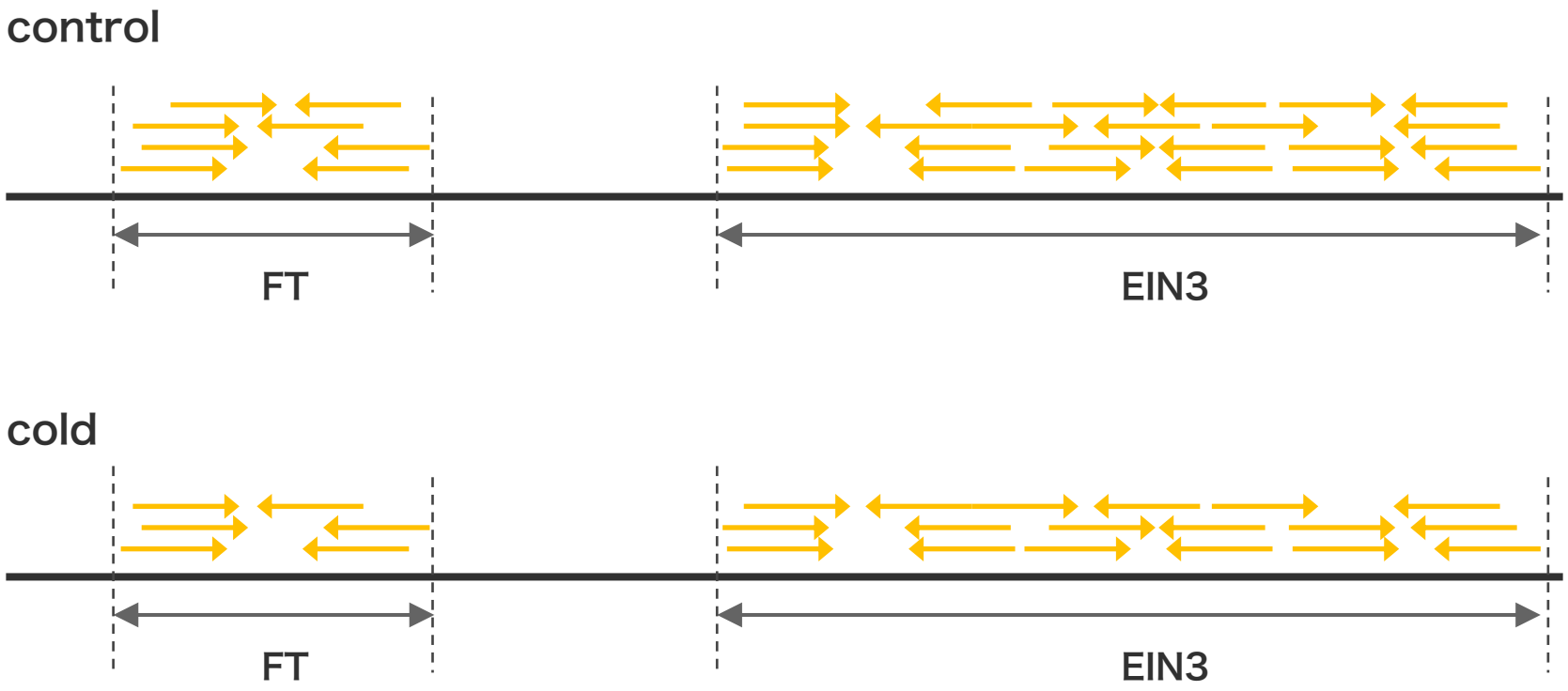
note



- long genes are expected to have a large number of reads
- normalize read counts by both gene length and library size
  1. normalize gene length to 1 kbp
  2. normalize library size to 1 million

Normalization / count-based normalization

note



- gene length between control and cold treatment are same
- to identify differentially expressed genes between groups
  - normalize by library size is enough

Expression analysis

- DE analysis
- Functional analysis

Gene ontology

AT1G62360 (STM)

BP carpel development, cytokinin biosynthetic process, floral meristem determinacy, plasmodesmata-mediated intercellular transport, regulation of meristem structural organization, regulation of transcription by RNA polymerase II, stem cell population maintenance

CC cytoplasm, cytoplasm, endosome, microtubule cytoskeleton, nucleus, plasma membrane, plasmodesma

MF DNA-binding transcription factor activity, DNA-binding transcription factor activity, RNA polymerase II-specific, RNA binding, RNA polymerase II cis-regulatory region sequence-specific DNA binding, protein binding, protein homodimerization activity

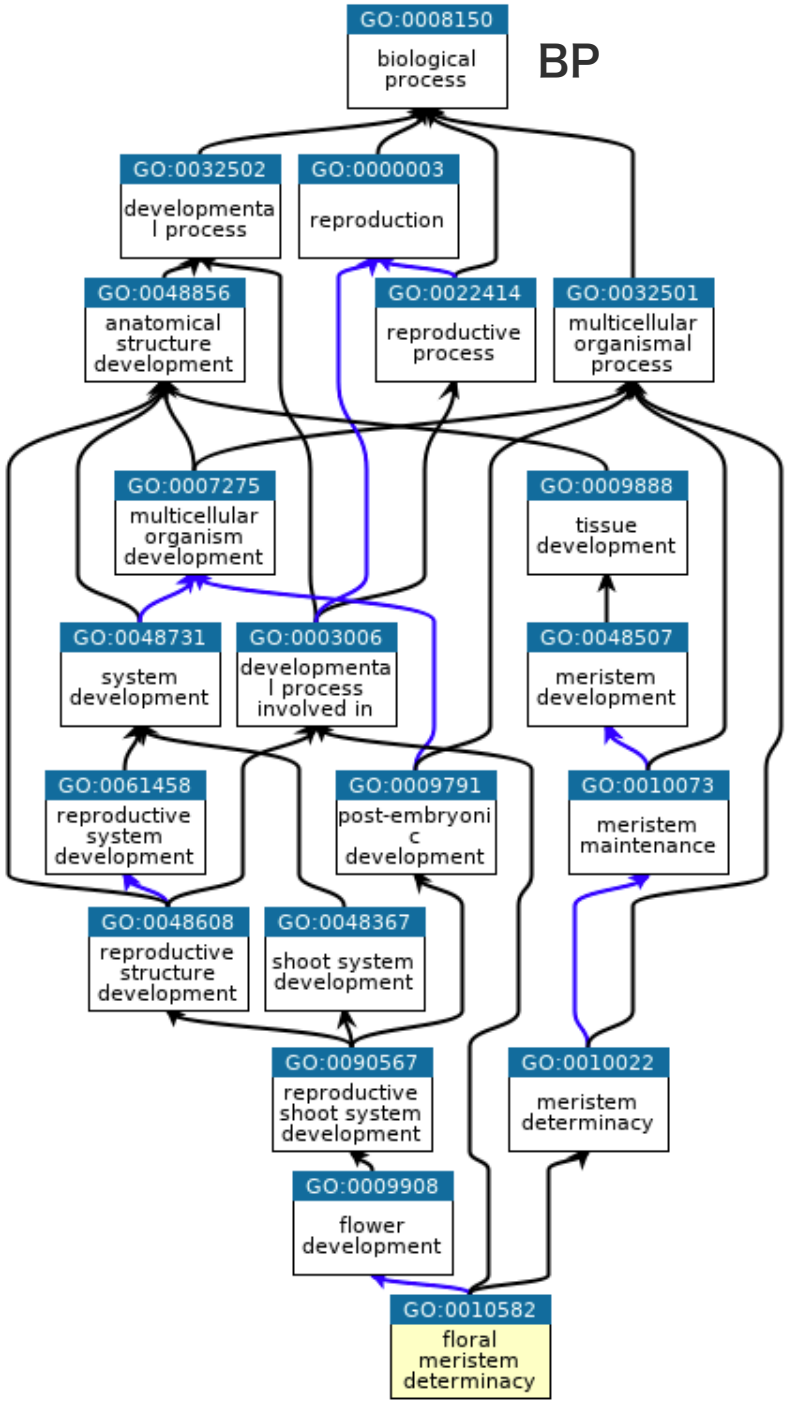
Gene ontology

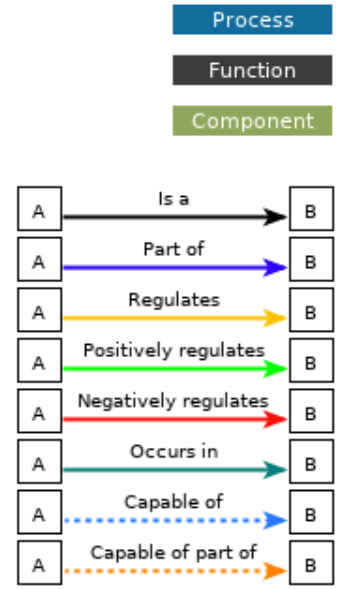
GO ID

GO term

GO:0010582

floral meristem determinacy

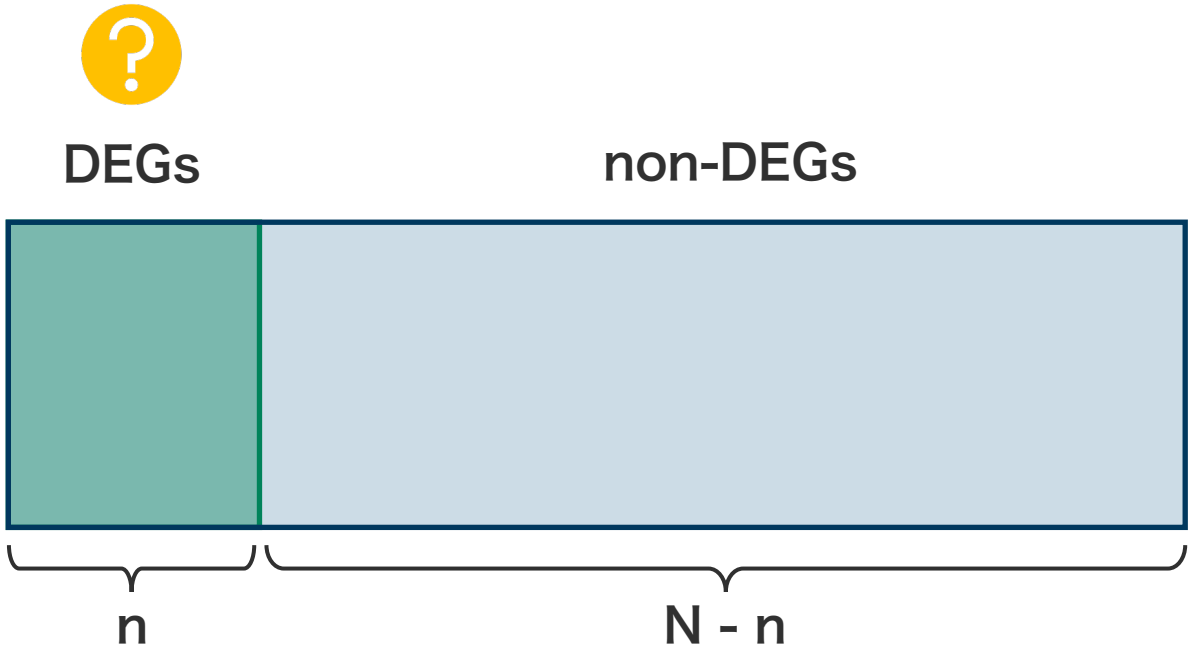




QuickGO - https://www.ebi.ac.uk/QuickGO

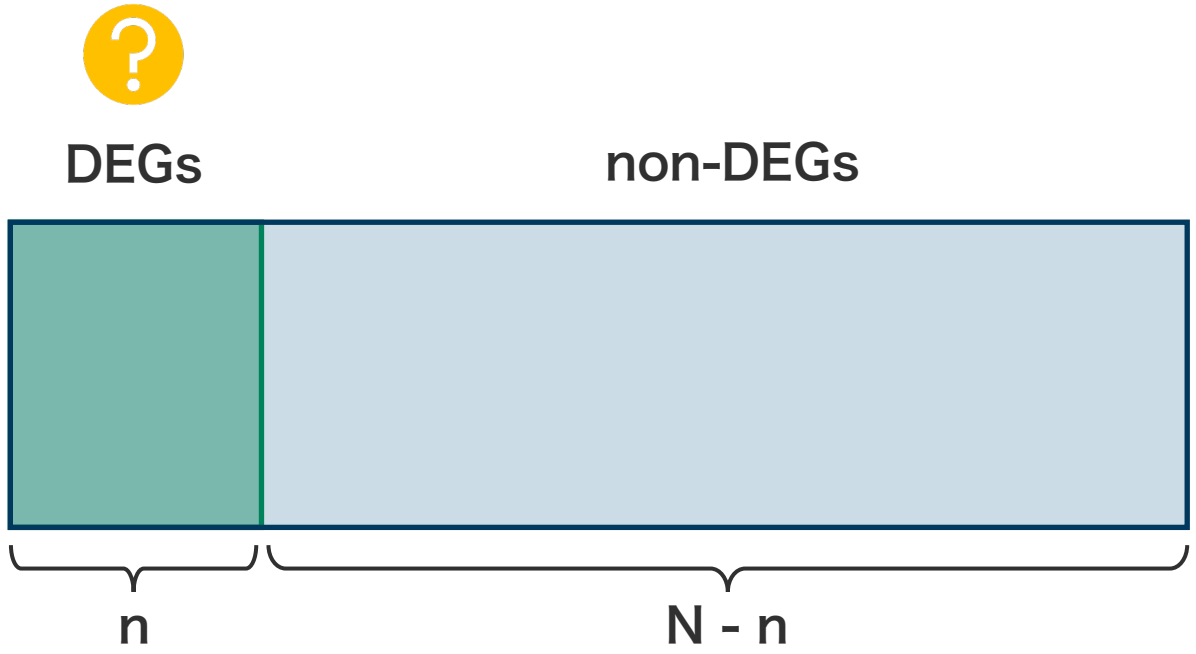
Enrichment analysis

What functions are these DEGs related to?



Enrichment analysis

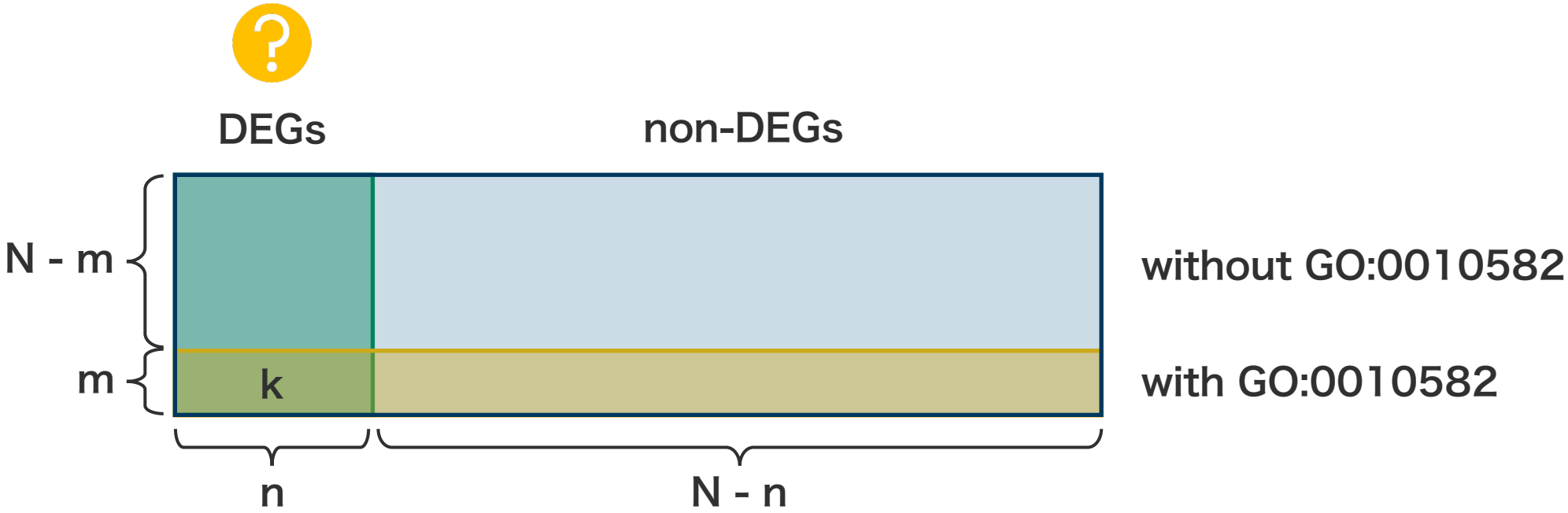
What functions are these DEGs related to?



GO:0010582
GO:0009723
GO:0009415
GO:0008150

Enrichment analysis

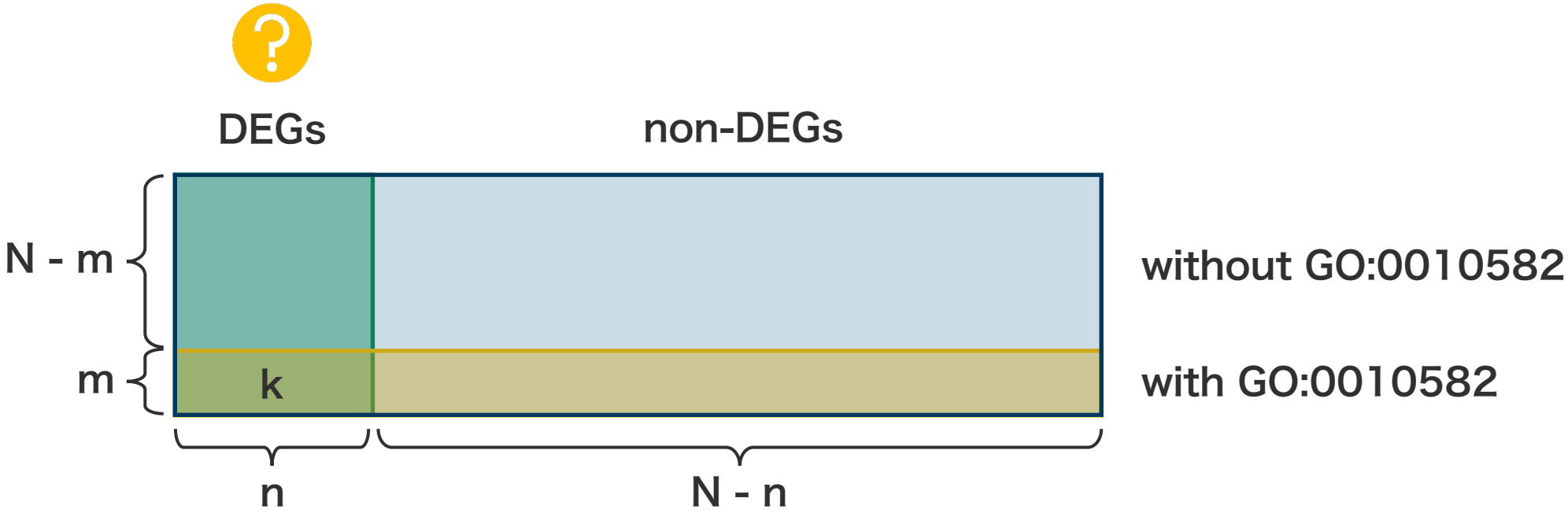
What functions are these DEGs related to?



GO:0010582
GO:0009723
GO:0009415
GO:0008150

Enrichment analysis

What functions are these DEGs related to?

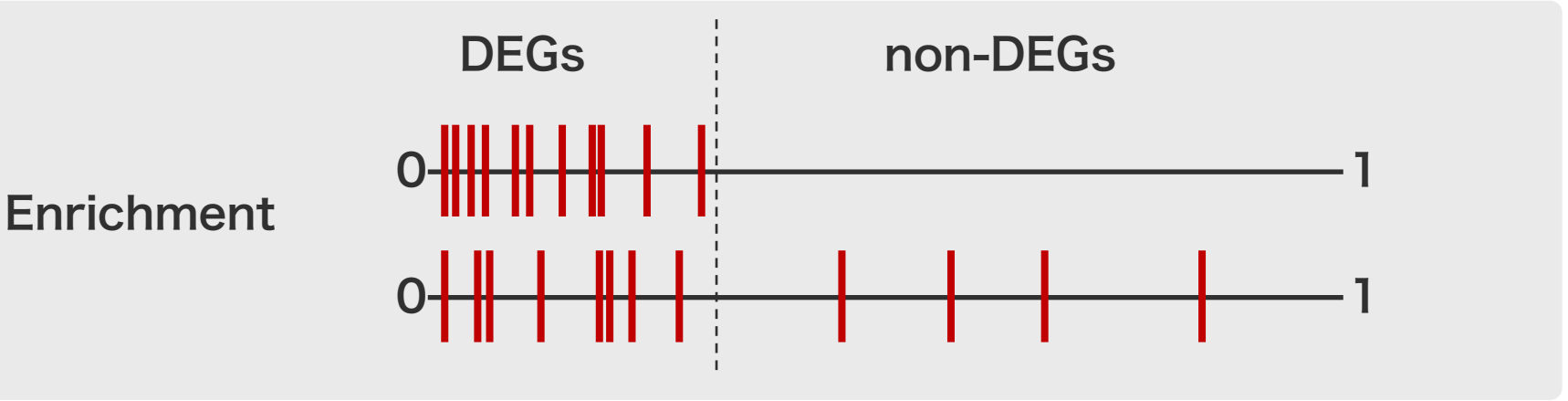


	DEGs	nonDEGs	total
with GO:0010582	k	m - k	m
without GO:0010582	n - k	N - m - n + k	N - m
total	n	N - n	N

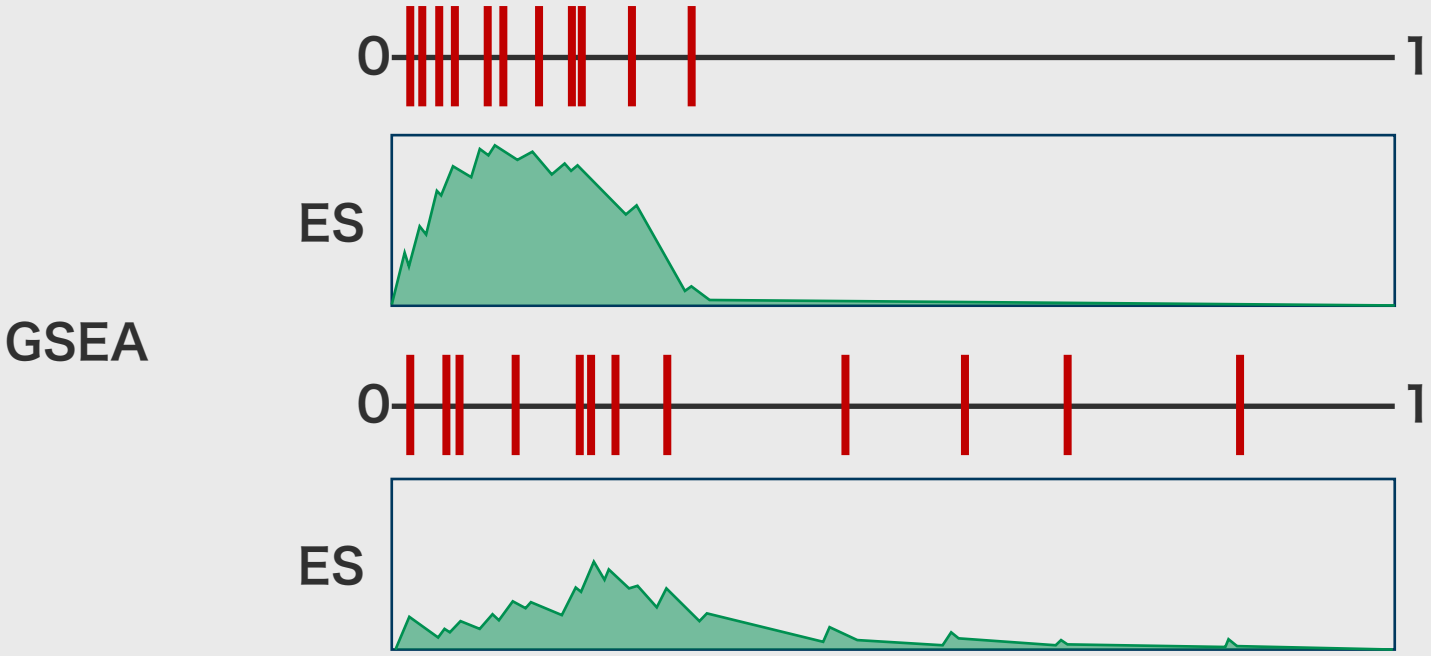
Fisher's exact test

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

GSEA (gene set enrichment analysis)



Focus on DEGs annotated with specific GO terms.

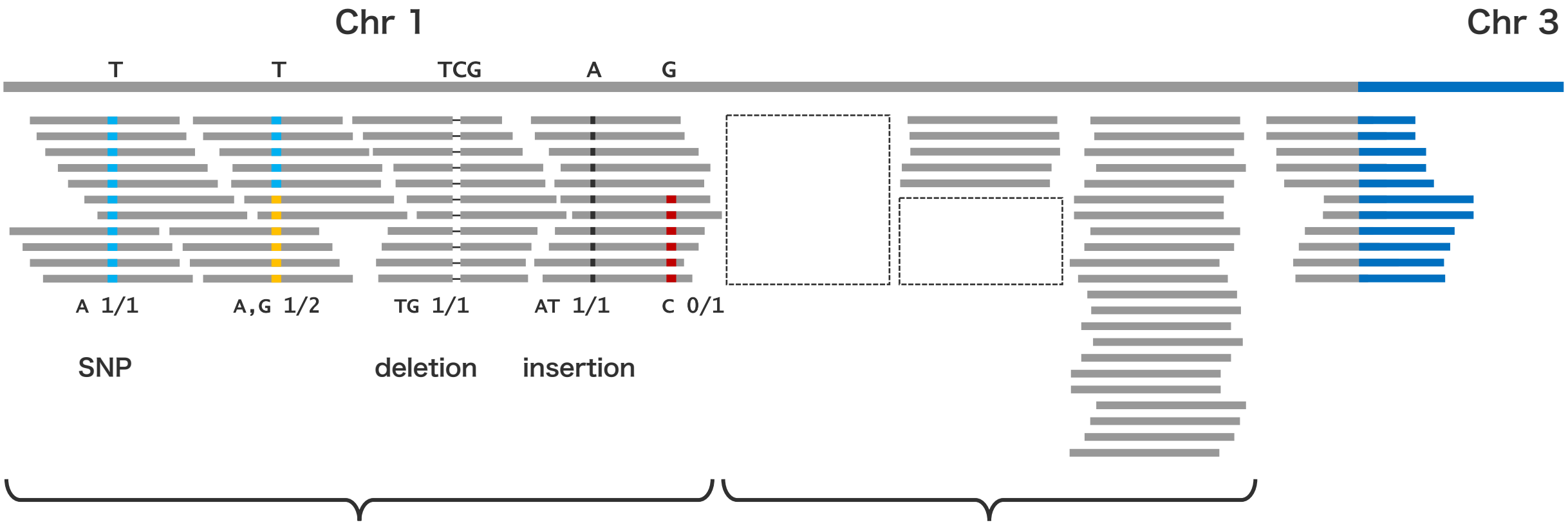


Focus on all genes ranked by the statistics, and check whether the gene is annotated with specific GO terms or not from the top untill the bottom of the rank.

GWAS

- ◯ Introduction
- ◯ Mapping
- ◯ Variant calling

Genomic variants



GWAS

Cardamine amara
inhabits in water

Cardamine rivularis
inhabits in glassland

```
Ancestor  REF  GGGAACCCCAGGGCTATCGAAATCGATCCCAA
          A1   GGGAACCGCAGGGCAATCGAAATCGACCCCAA
          A2   GGGAACCGCAGGGCAATCGAAATCGACCCCAA
          A3   GGGAACCCCAGGGCAATCGAAATCGACCCCAA
          A4   GGGAACCCCAGGGCAATCGAAATCGACCCCAA
          A5   GGGAACCCCAGGGCAATCGAAATCGACCCCAA
          R1   GGGAACCCCAAGGCTATCGTAATCGACCCCAA
          R2   GGGAACCCCAAGGCTATCGTAATCGACCCCAA
          R3   GGGAACCCCAAGGCTATCGTAATCGACCCCAA
          R4   GGGAACCCCAGGGCTATCGTAATCGACCCCAA
          A5   GGGAACCCCAGGGCTATCGTAATCGACCCCAA
```

▲ ▲

contribute water-stress tolerance?

GWAS

```
REF        GGGAACCCCAGGGCTATCGAAATCGATCCCAA

XI-1A.1    GGGAACCGCAGGGCAATCGAAATCGACCCCAA
XI-1A.2    GGGAACCGCAGGGCAATCGAAATCGACCCCAA
XI-1B.1    GGCAACCCCAGGGCAATCGAAATCGACCCCAA
XI-1B.2    GGCAACCCCAGGGCAATCGAAATCGACCCCAA
XI-2.1     GGGAACCCCAAGGGTATCGAAA-CGATCGCAA
XI-2.2     GGGAACCCCAAGGCTATCGTAA-CGATCGCAA
XI-2.3     GGGAACCCCAAGGCTATCGTAA-CGATCGCAA
:          GGGAACCCCAAGGCTATCGTAATCGATCGCAA
:          GGGAACCCCAGGGCTATCGTAATCGATCCCAA
Admix.3    GGGAACCCCAGGGCTATCGTAATCGATCCCAA
```

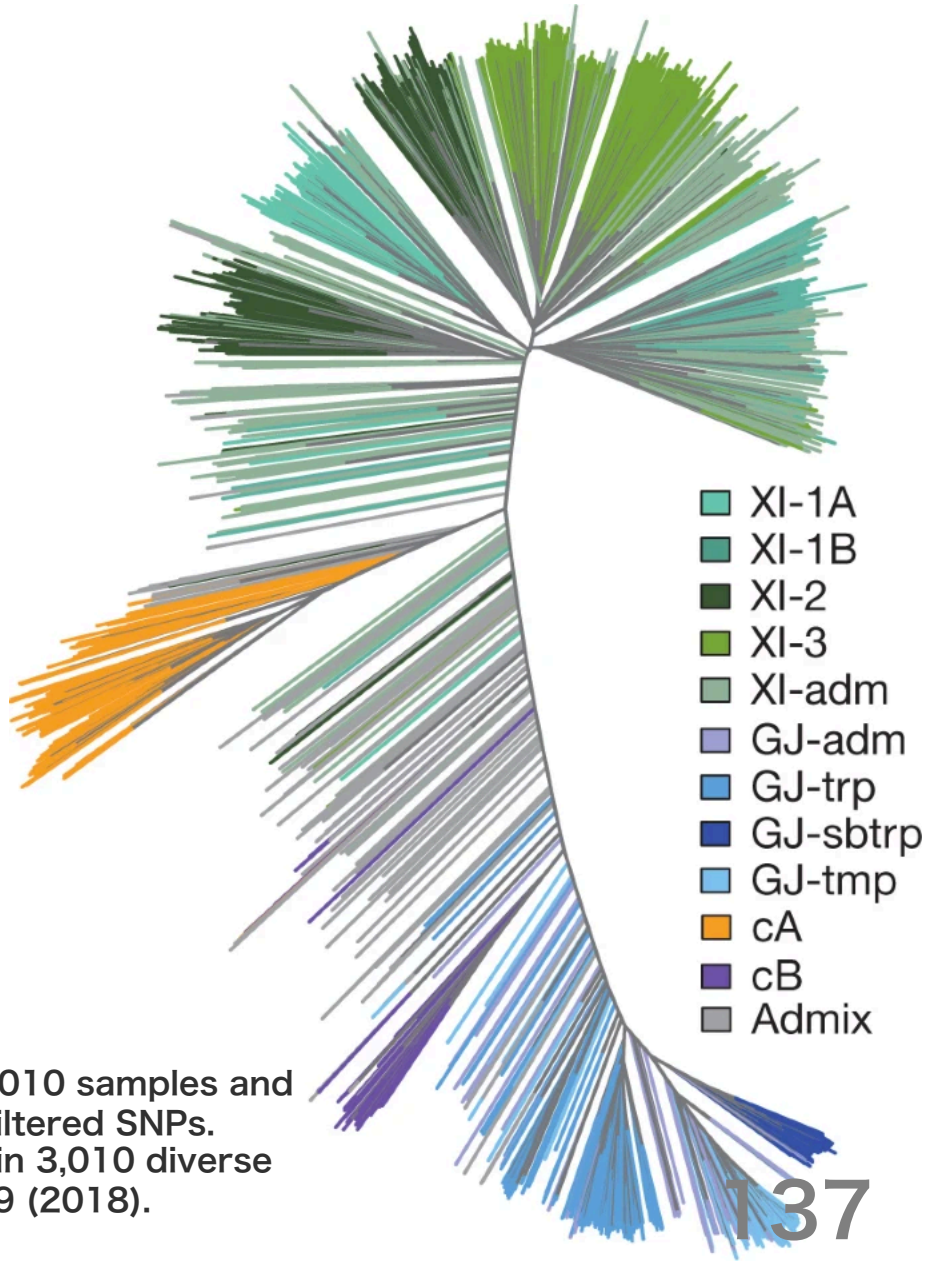
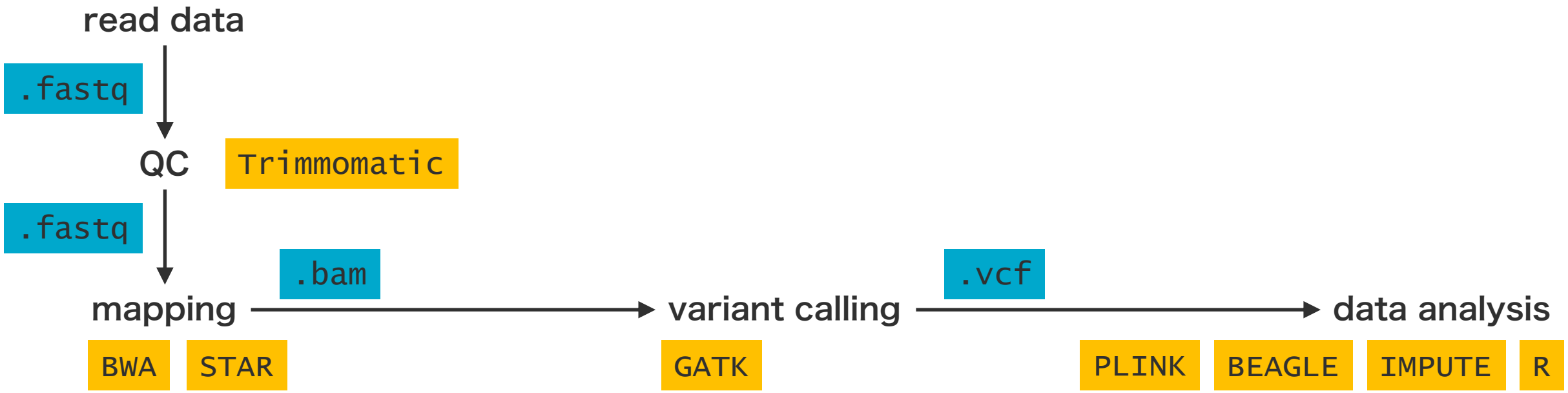


Fig. 1: Unweighted neighbour-joining tree based on 3,010 samples and computed on a simple matching distance matrix for filtered SNPs. Wang, W., Mauleon, R., Hu, Z. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557, 43–49 (2018). https://doi.org/10.1038/s41586-018-0063-9

GWAS

- ◯ Introduction
- ◯ Mapping
- ◯ Variant calling

Variant calling



pre-processing	variant calling
mark duplicates	variant calling
local re-alignment	variant filtering
base quality recalibration (BQSR)	VQSR

Mapping

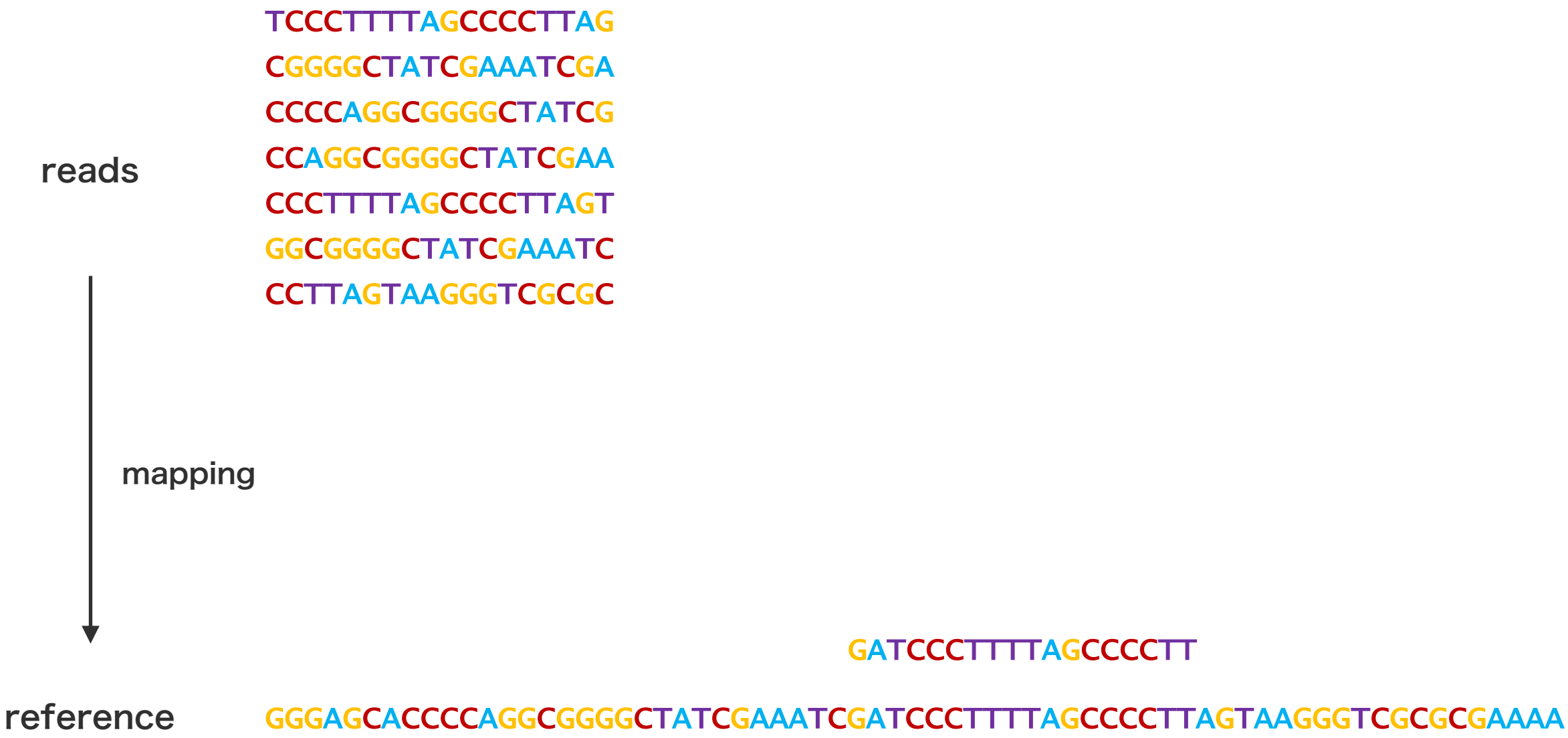
reads

TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT

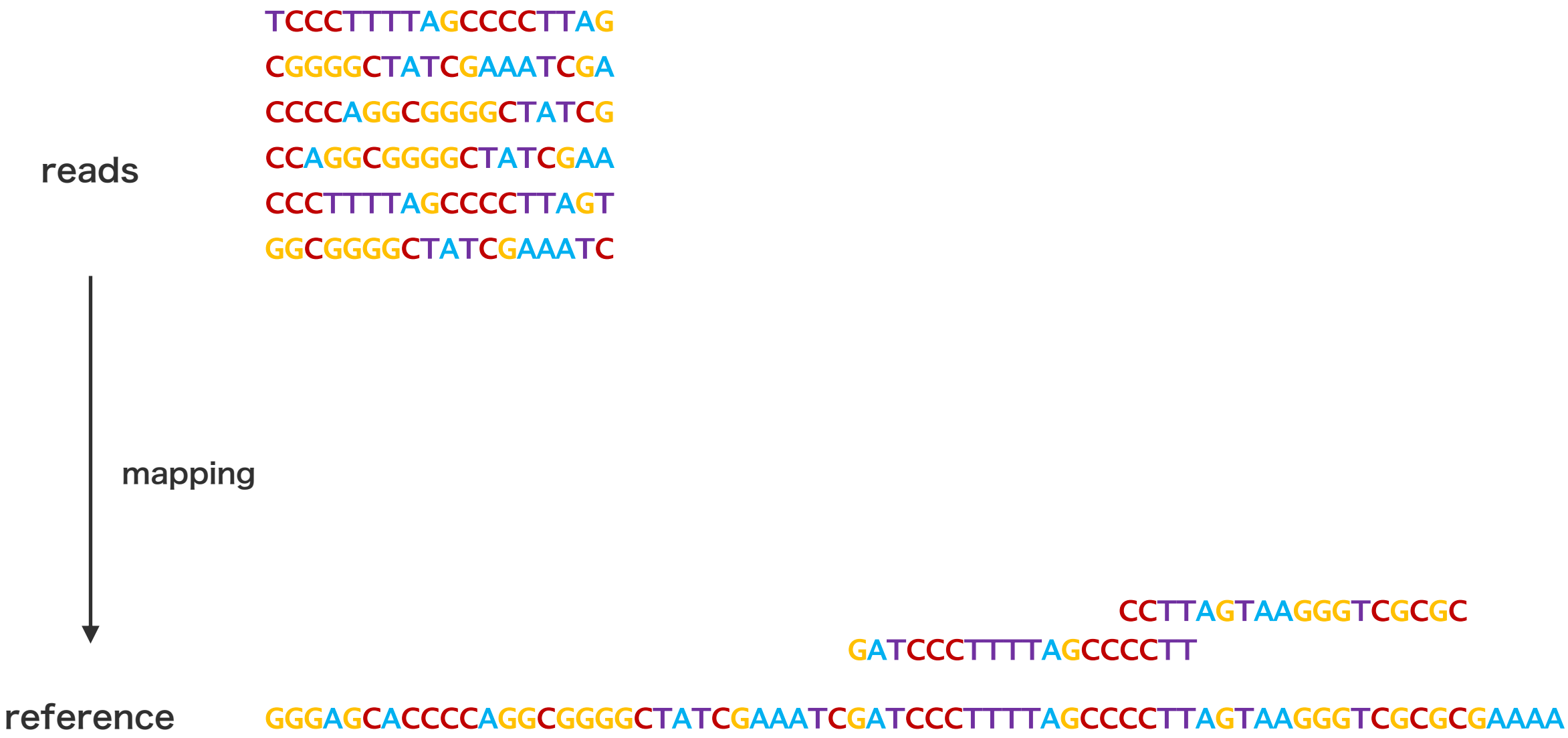
mapping

reference GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

Mapping


TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
CCTTAGTAAGGGTCGCGC
reads
mapping
GATCCCTTTTAGCCCCTT
reference
GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

Mapping


TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
reads
mapping
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT
reference
GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

Mapping

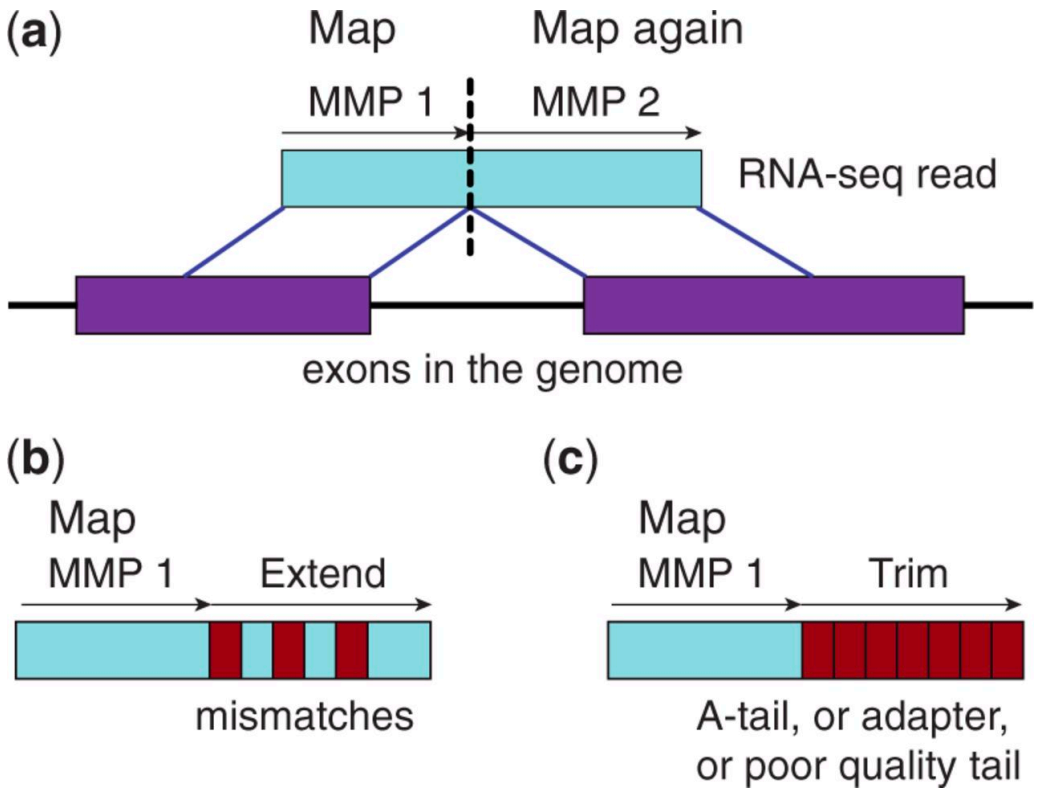

reads
TCCCTTTTAGCCCCTTAG
CGGGGCTATCGAAATCGA
CCCCAGGCGGGGCTATCG
CCAGGCGGGGCTATCGAA
CCCTTTTAGCCCCTTAGT
GGCGGGGCTATCGAAATC
CCTTAGTAAGGGTCGCGC
GATCCCTTTTAGCCCCTT
mapping
CCCCAGGCGGGGCTATCG
TCCCTTTTAGCCCCTTAG
GGCGGGGCTATCGAAATC
CCCTTTTAGCCCCTTAGT
CGGGGCTATCGAAATCGA
CCTTAGTAAGGGTCGCGC
CCAGGCGGGGCTATCGAA
GATCCCTTTTAGCCCCTT
reference
GGGAGCACCCCAGGCGGGGCTATCGAAATCGATCCCTTTTAGCCCCTTAGTAAGGGTCGCGCGAAAA

STAR

- STAR consists of two major steps: seed searching and clustering/stitching/scoring

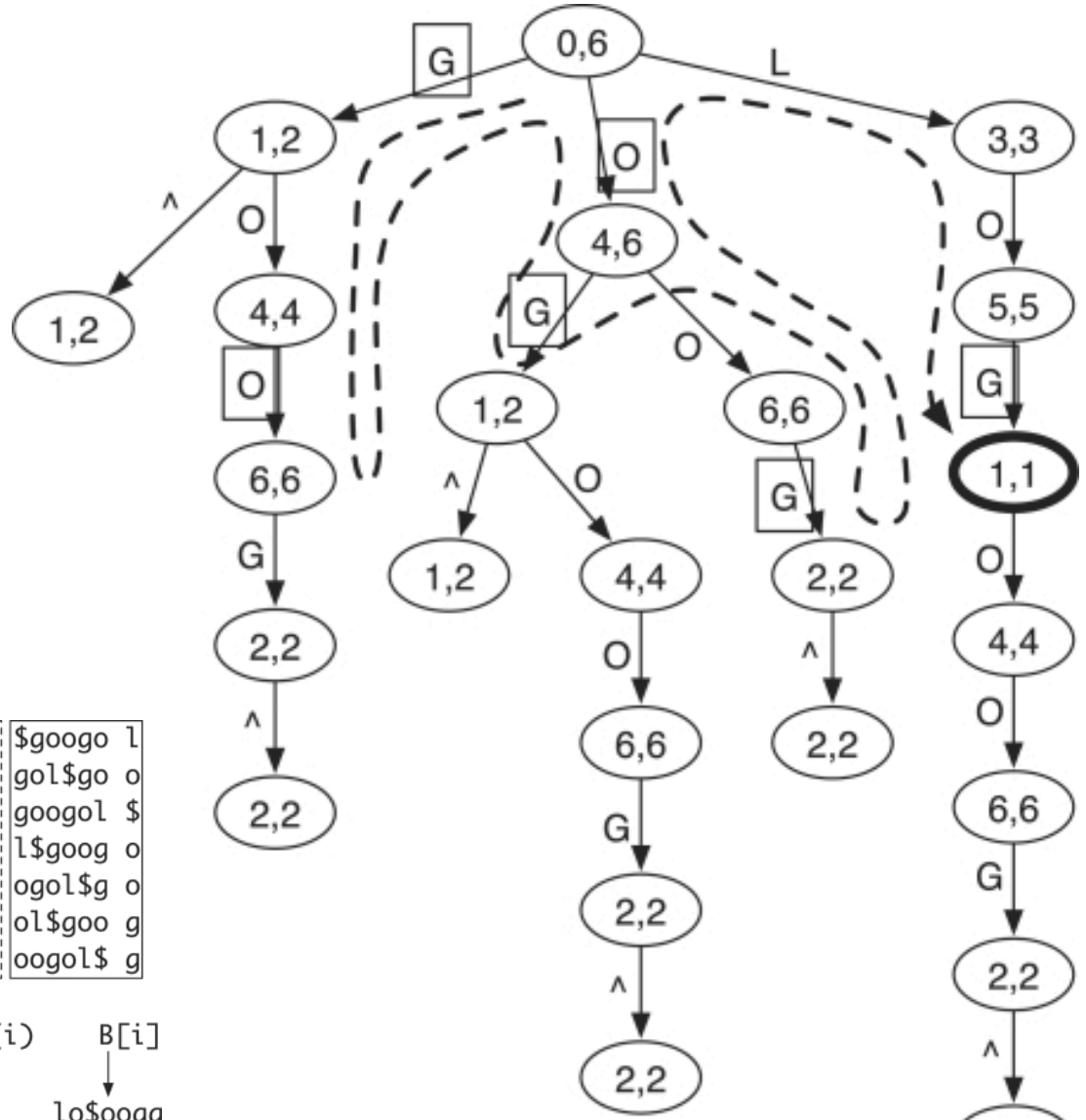
- seed searching
  - search MMP (Maximal Mappable Prefix) from 5'-end
  - then, search MMP from the unmapped position of the given read
  - trim poly-A tails and poor-quality tails

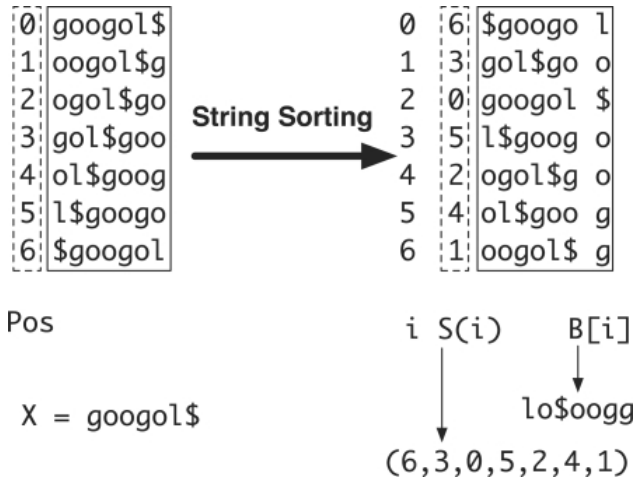
- clustering/stitching/scoring
  - intron size
  - mismatches
  - gaps



BWA

- There are three alignment modes
  - short reads (< 100 bp)
    - BWA-backtrack
  - long reads (70 < 1M bp)
    - BWA-SW
    - BWA-MEM





SAM format

HEADER

```
@HD VN:1.0
@SD SN:1        LN:30427671
@SD SN:2        LN:19698289
@SD SN:3        LN:23459830
@RG ID:MAK1     PL:ILLUMINA   LB:IL200     SM:Col-0
@PG ID:bwa      PN:bwa        VN:0.6.2
```

RECORDS

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	TAG
SRR7508939.23	163	1	3848	60	66M82N22M	=	3890	128	GCC...GAG	HIJ...DD@	AS:i:0 XN:i:0 XM:i:0 ... NH:i:1

SAM format

FLAG

```
   1 read paired
   2 read mapped in proper pair
   4 read unmapped
   8 mate unmapped
  16 read being reverse complemented
  32 mate being reverse complemented
  64 first in the template
 128 last in the template
 256 not primary alignment
 512 not passing platform/vendor quality checks
1024 read is PCR or optical duplicate
2048 supplementary alignment
```

CIGAR

```
M alignment match (can be a sequence match or mismatch)
I insertion to the reference
D deletion from the reference
N skipped region from the reference
S soft clipping (clipped sequences present in SEQ)
H hard clipping (clipped sequences NOT present in SEQ)
```

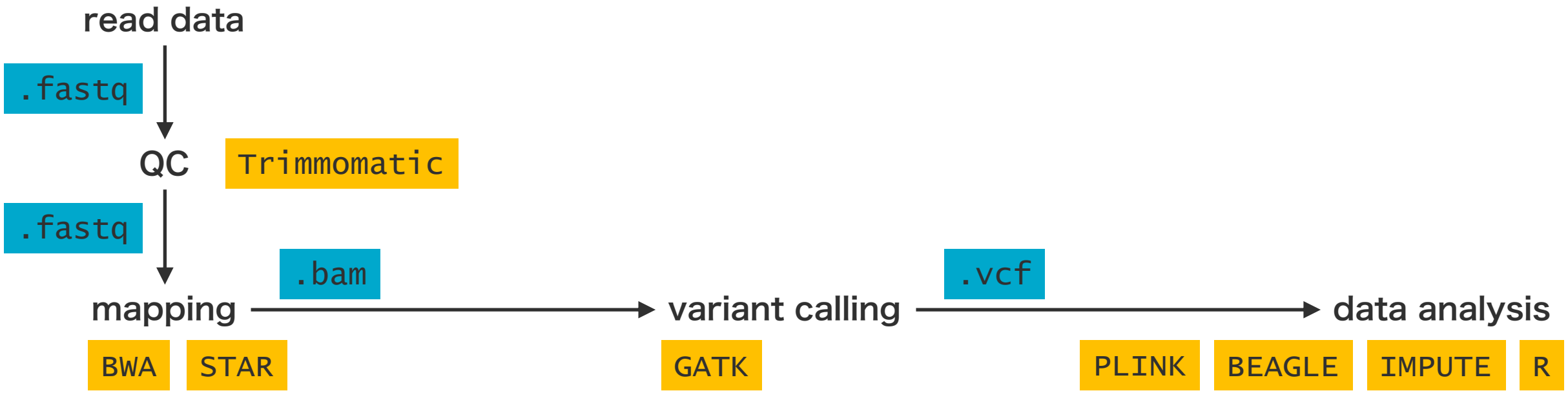
TAG

```
AS alignment score
XO number of best hits
XM number of mismatches in the alignment
NH number of reported alignments that contain the query in the current record
```

GWAS

- Introduction
- Mapping
- Variant calling
-

Variant calling



pre-processing	variant calling
mark duplicates	variant calling
local re-alignment	variant filtering
base quality recalibration (BQSR)	VQSR

pre-processing / mark duplicates

.bam

↓

mark duplicates

↓

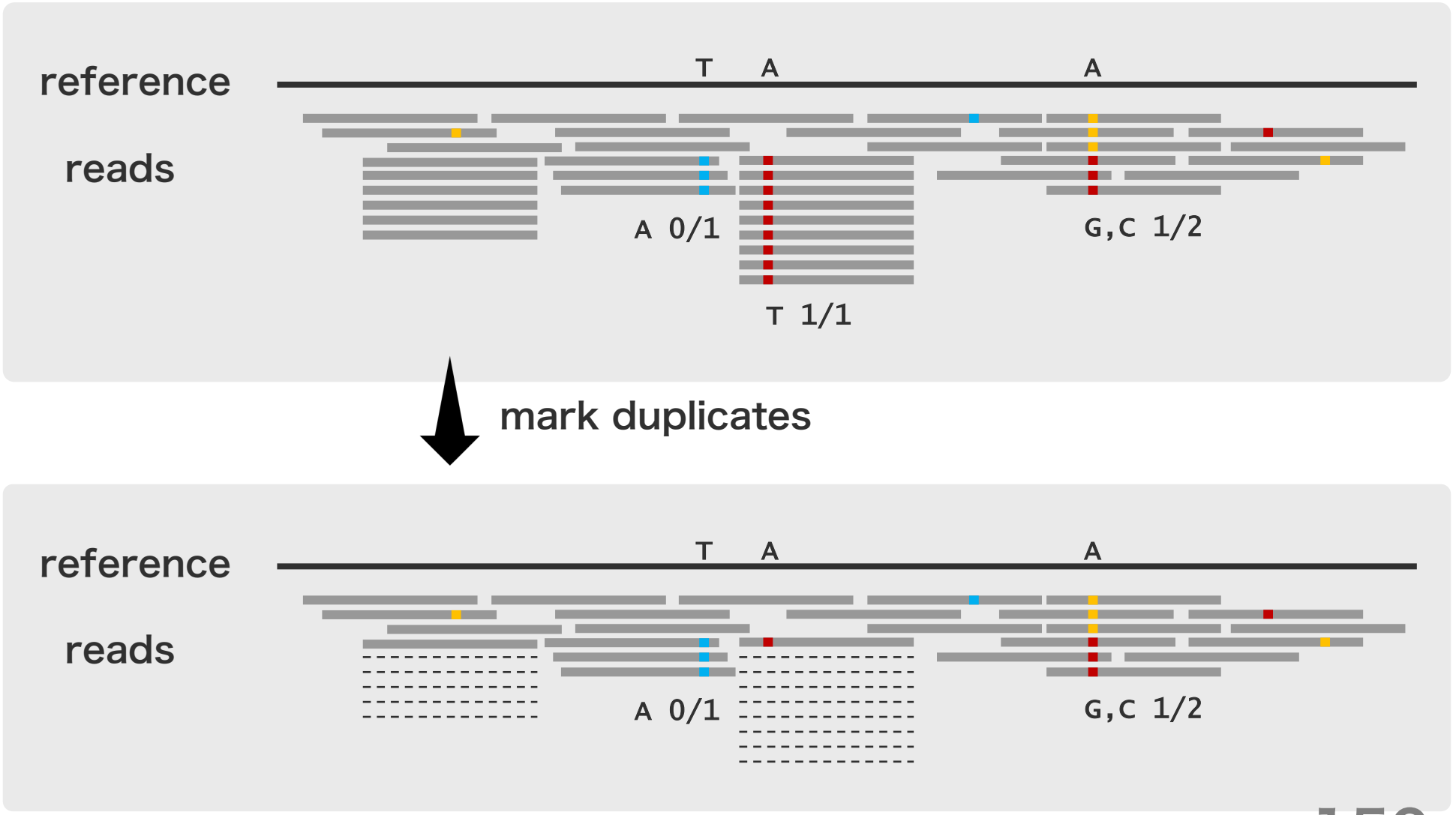
local realignment

↓

BQSR

↓

.bam



pre-processing / local realignment

.bam

↓

mark duplicates

↓

local realignment

↓

BQSR

↓

.bam

pre-processing / BQSR

.bam

↓

mark duplicates

↓

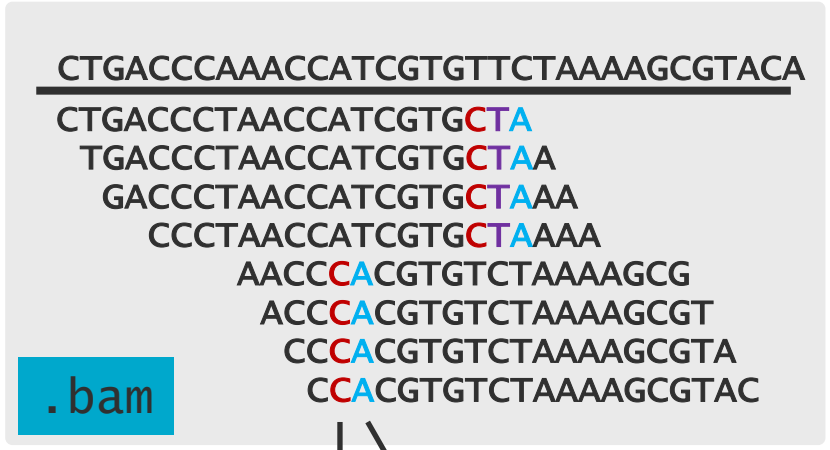
local realignment

↓

BQSR

↓

.bam



quality score 10 40

error rate 10% 0.01%

↓

Should we trust these quality score?

- Sequencers make systematic errors in base calling
- PCR-based errors are not considerated

.fastq

```
@HWI:2:2102
AACCCACGTGTCTAAAAGCG
+
@@1B+IDDB??D:CGFI?<>
@HWI:2:1102
ACCCACGTGTCTAAAAGCGT
+
CCC+IIFFHHGHHJJJJJJB
@HWI:2:1324
CCCACGTGTCTAAAAGCGTA
+
DD+I??D:CGFFFFHJJJFF
@HWI:2:6229
CCACGTGTCTAAAAGCGTAC
+
F+IFHJBA@@@1BDDDDB??
```

pre-processing / BQSR

.bam

↓

mark duplicates

↓

local realignment

↓

BQSR

↓

.bam

↓

quality score 30

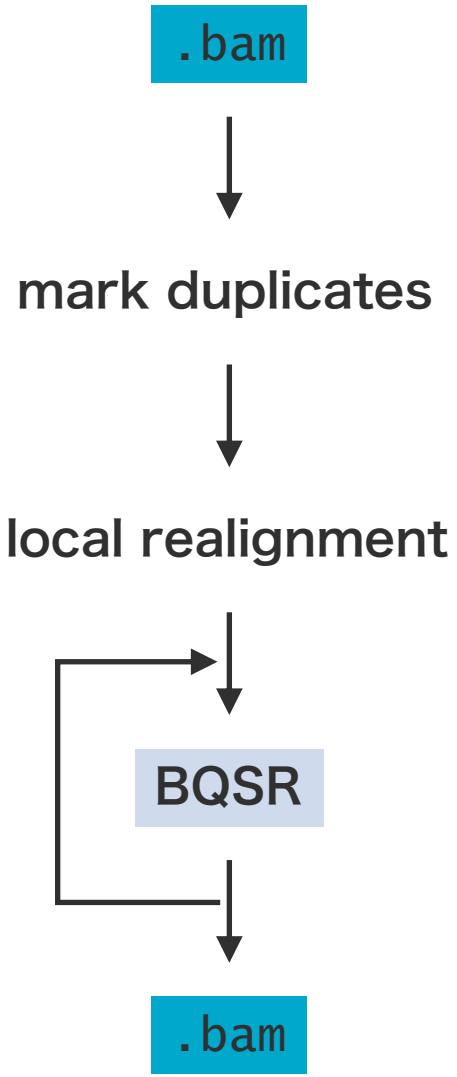
error rate 0.1%

.fastq

```
@HWI:2:2102
AACCCACGTGTCTAAAAGCG
+
????????????????????
@HWI:2:1102
ACCCACGTGTCTAAAAGCGT
+
????????????????????
@HWI:2:1324
CCCACGTGTCTAAAAGCGTA
+
????????????????????
@HWI:2:6229
CCACGTGTCTAAAAGCGTAC
+
????????????????????
```

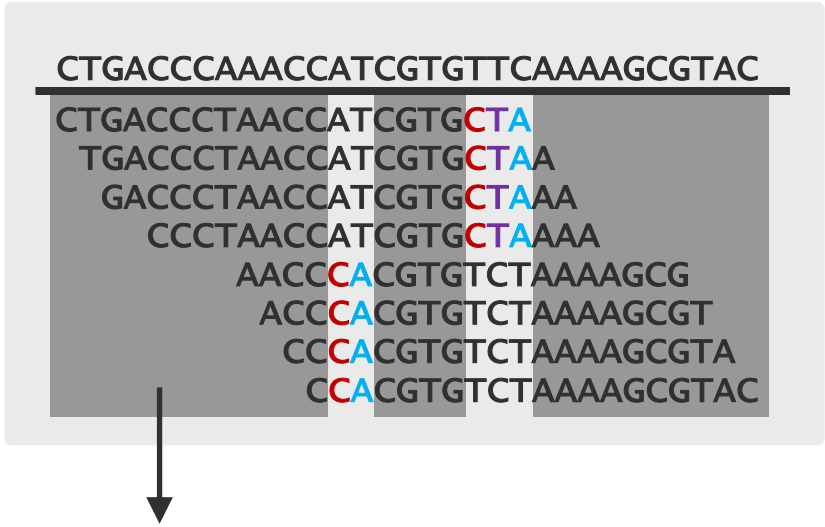
- After we performed QC, the bases are expected to containe 0.1% error due to quality score
- If we sequenced 4.1 Gb (135Mb x 30), the result is expected to contain 4 Mb errors!

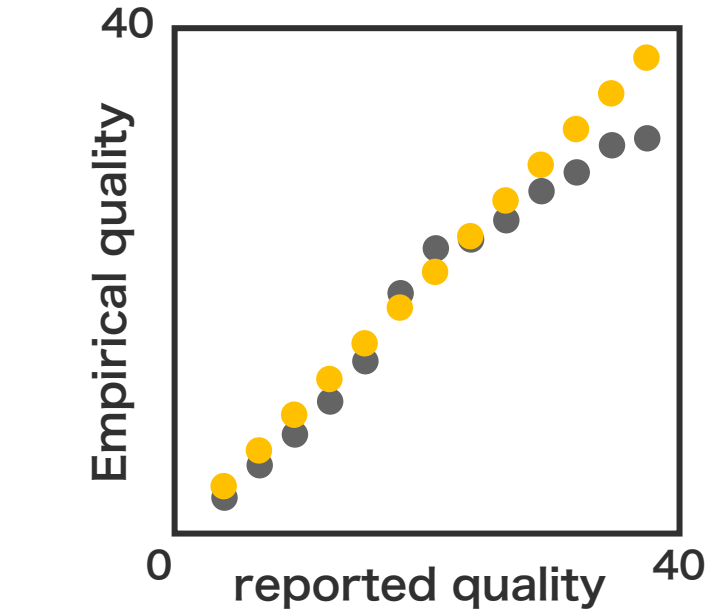
pre-processing / BQSR



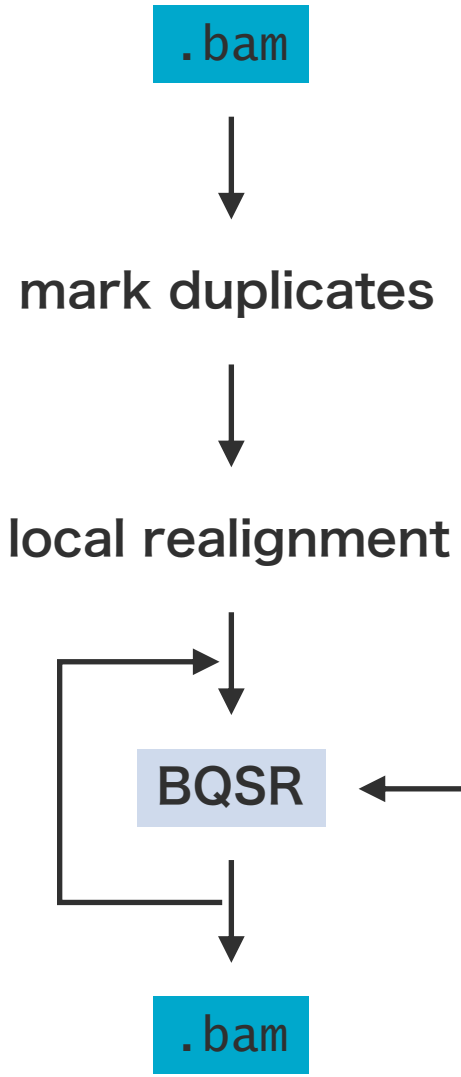
Base Quality Score Recalibration

- Build a correction model with non-variants loci
- Use @RG metadata to correct the reported quality score
- BQSR can performed iteratively until covergence



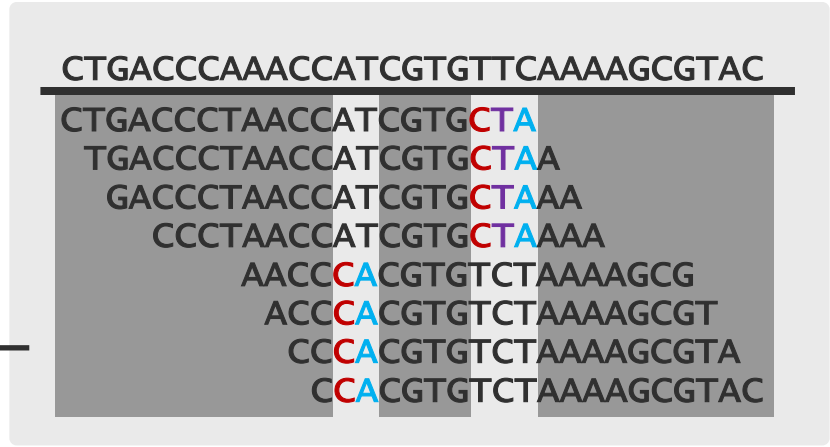


pre-processing / BQSR

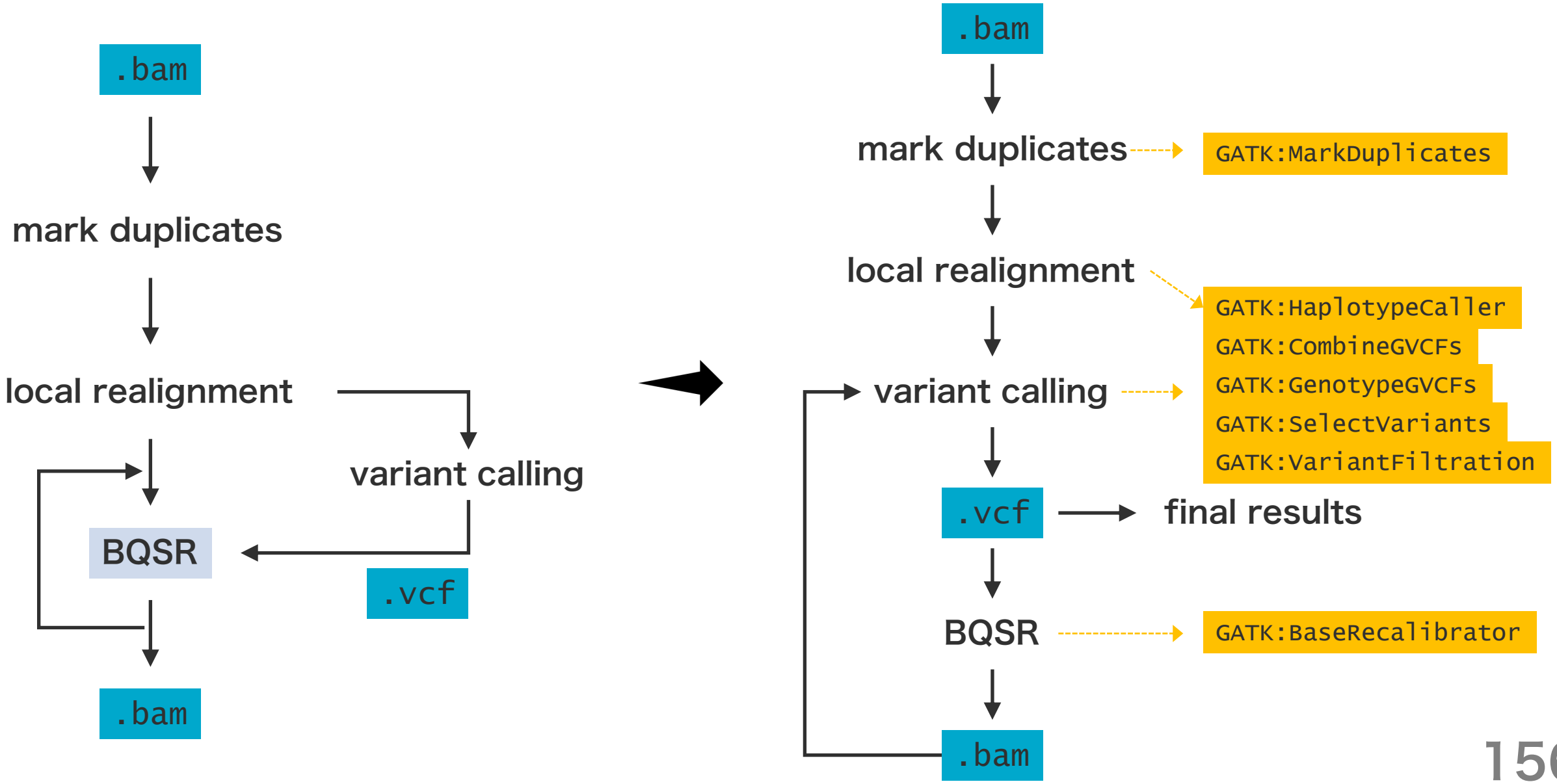


We need variants information before BQSR!

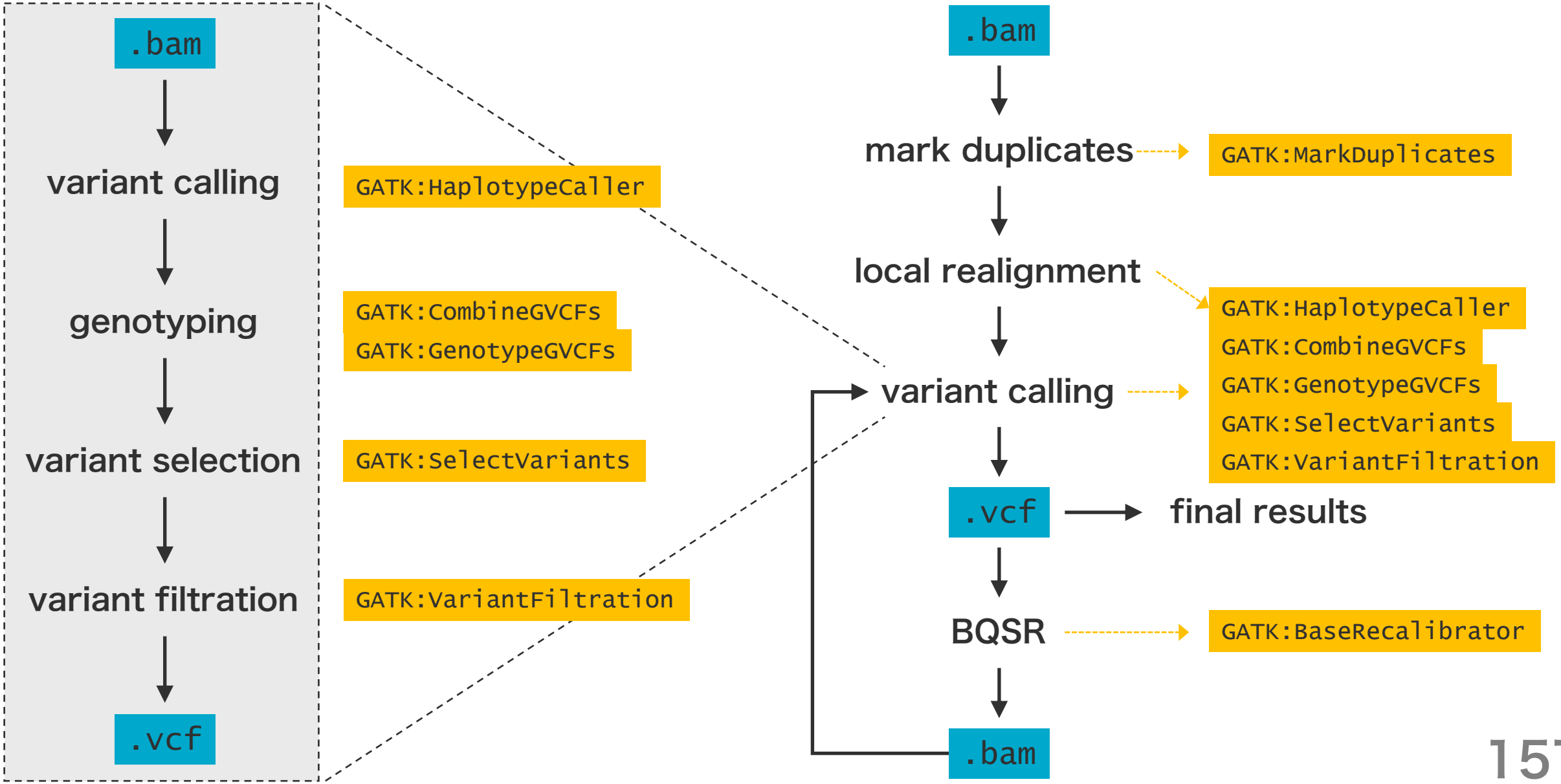
- We can download these information from database
- We can provisionally perform variant calling



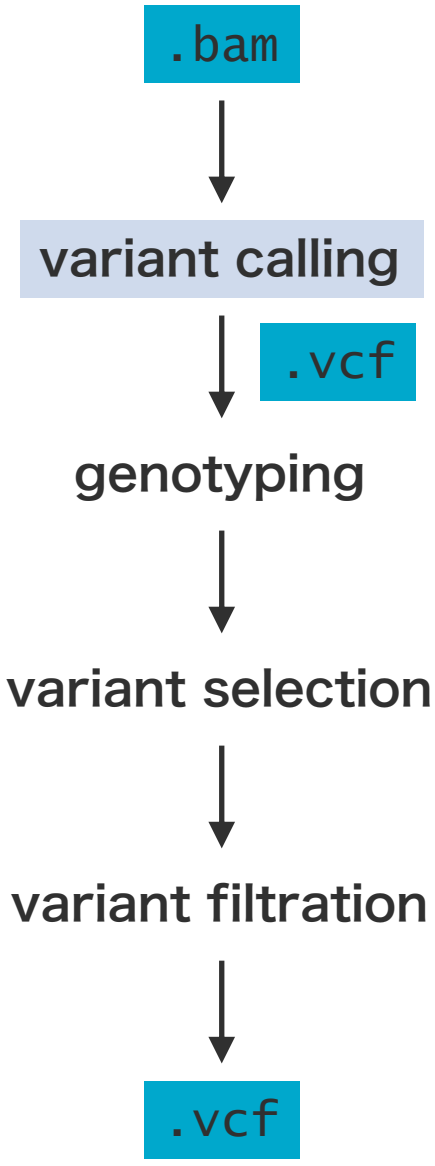
pre-processing / BQSR



Variant calling

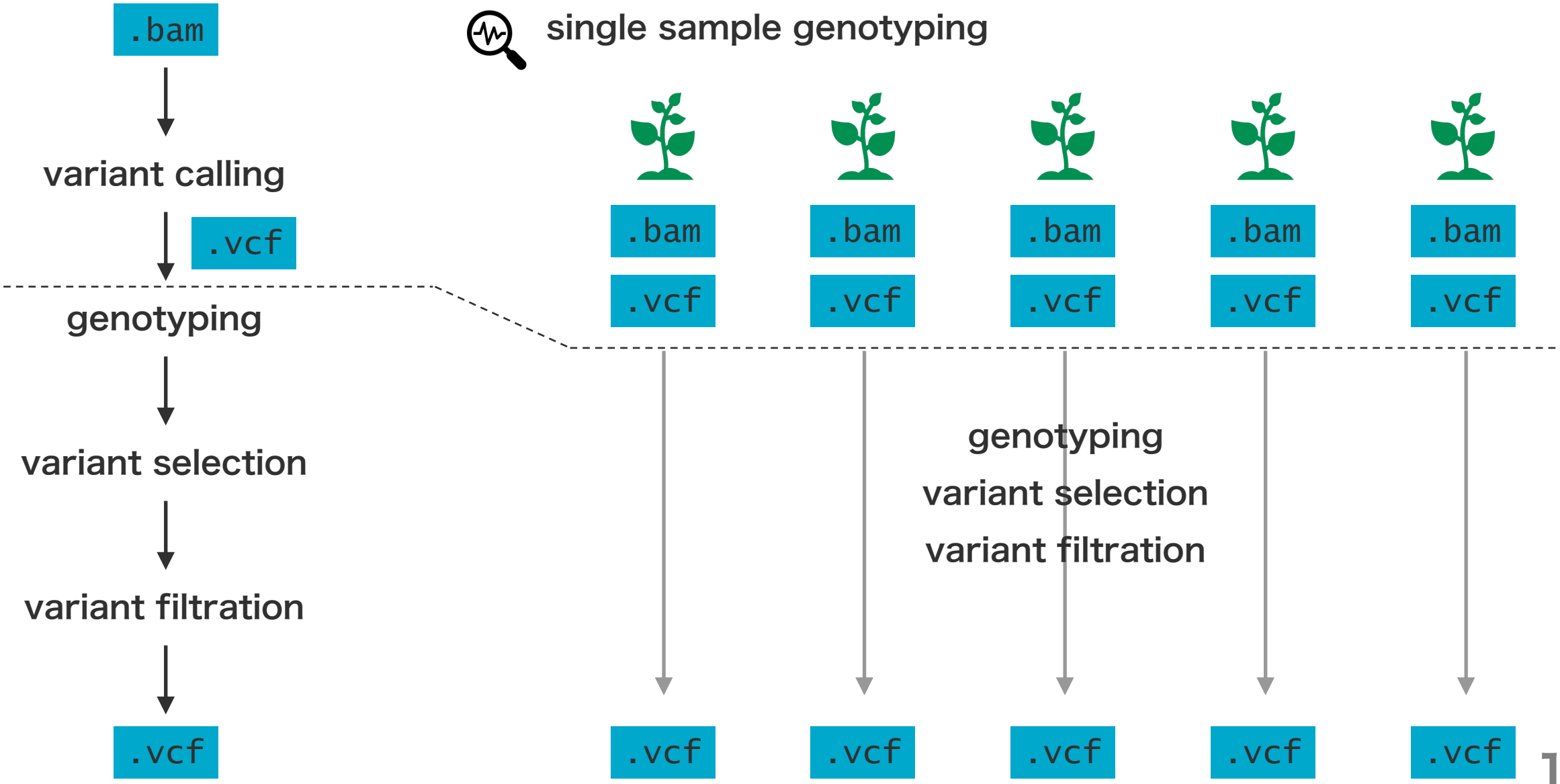


Variant calling

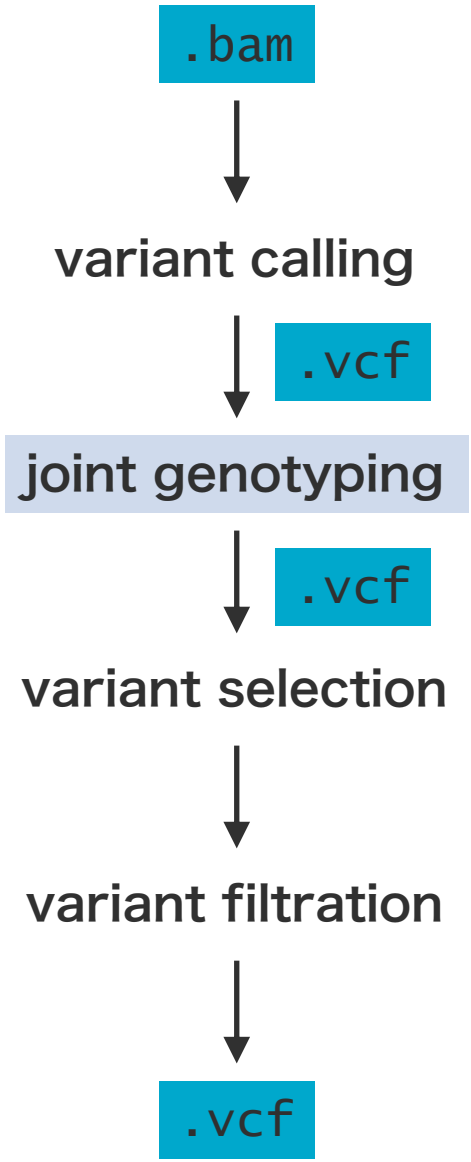


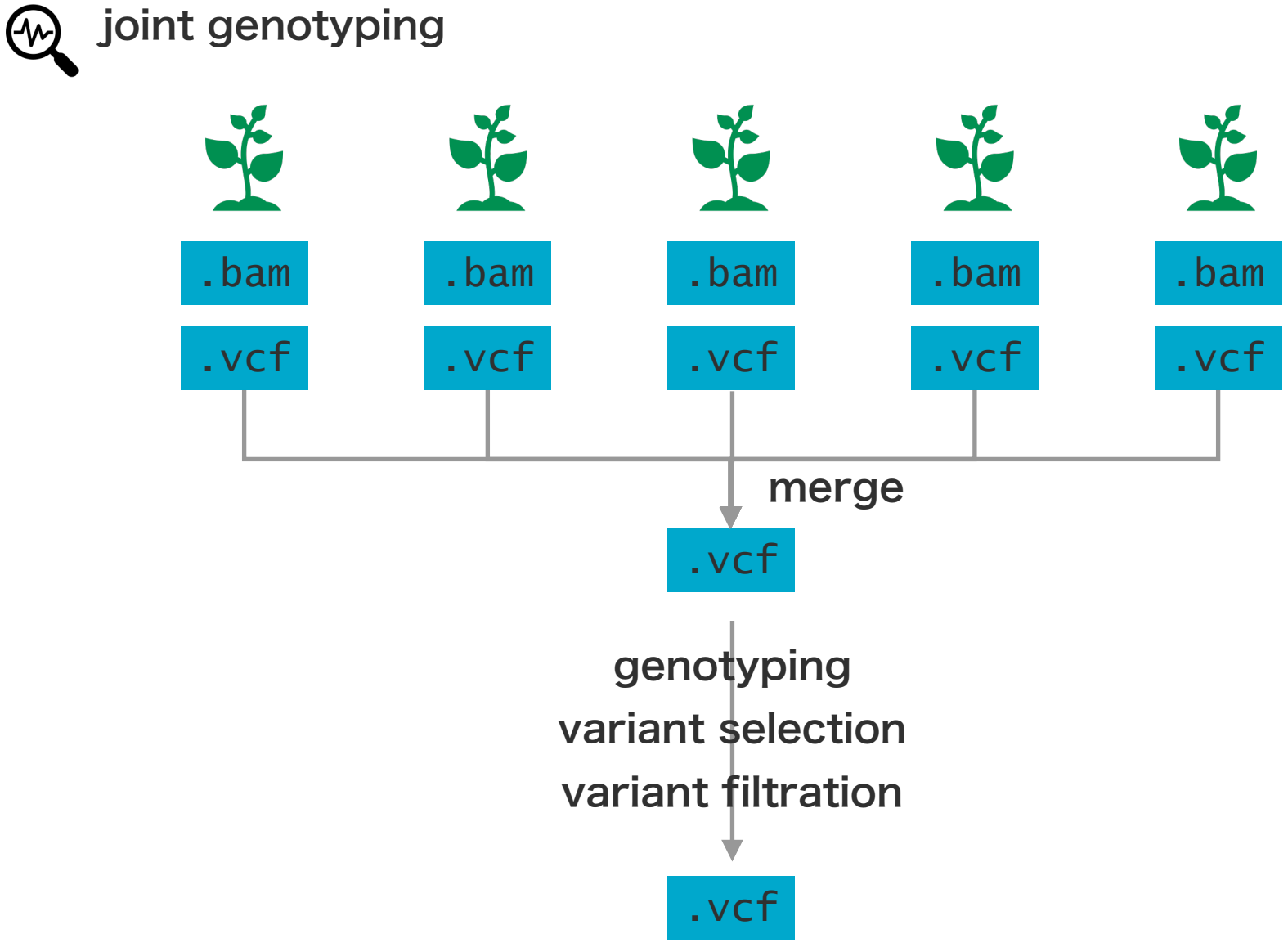
- Define active region
- Determine haplotypes by assembly of the active region
  - local realignment of atcive regions with Smith-Waterman algorithm
  - identify possible haplotypes
- Determine likelihoods of the haplotypes given the read data
  - pairwise alignment of each read against each ahplotype with PairHMM algorithm
  - ouput likelihoods of alleles for each potential variants
- Assign sample genotypes
  - calculate likehoods of each genotype per sample with Bayes' rule

Variant calling



Variant calling





Variant calling

.bam

↓

variant calling

↓ .vcf

joint genotyping

↓ .vcf

variant selection

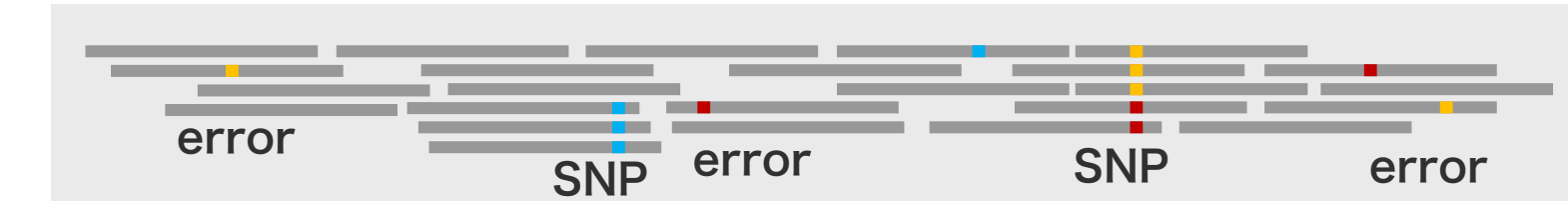
↓

variant filtration

↓

.vcf

joint genotyping



Variant calling

.bam

↓

variant calling

↓ .vcf

joint genotyping

↓ .vcf

variant selection

↓

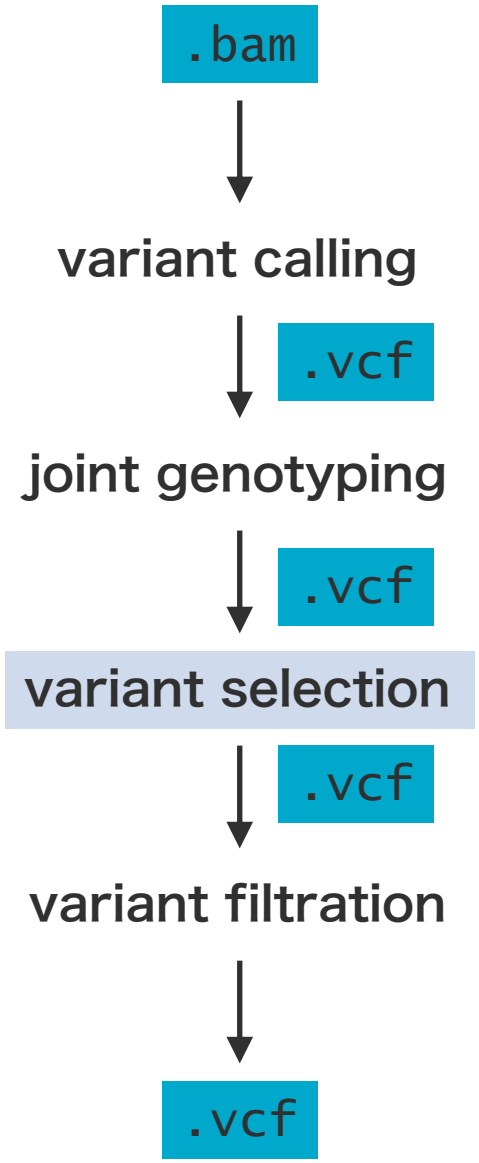
variant filtration

↓

.vcf

joint genotyping

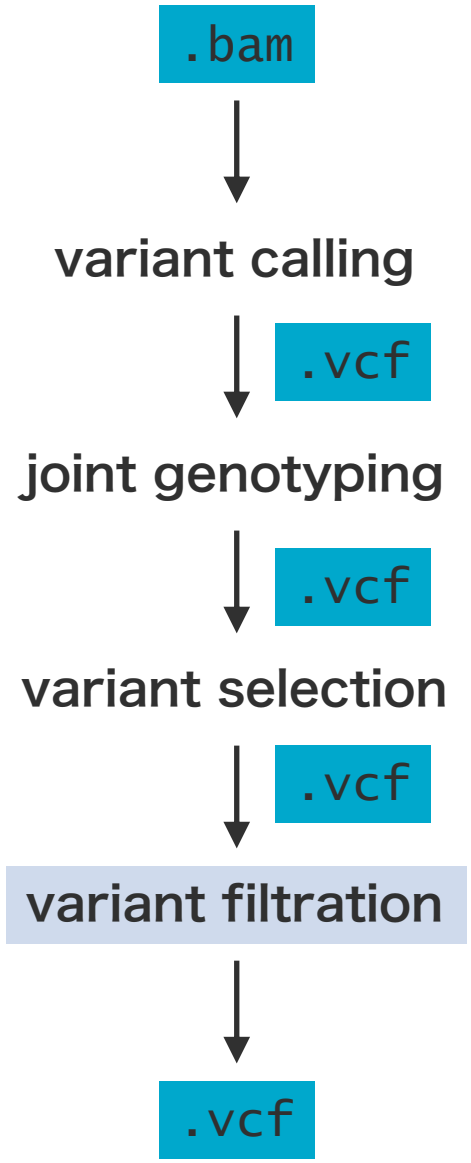
Variant calling

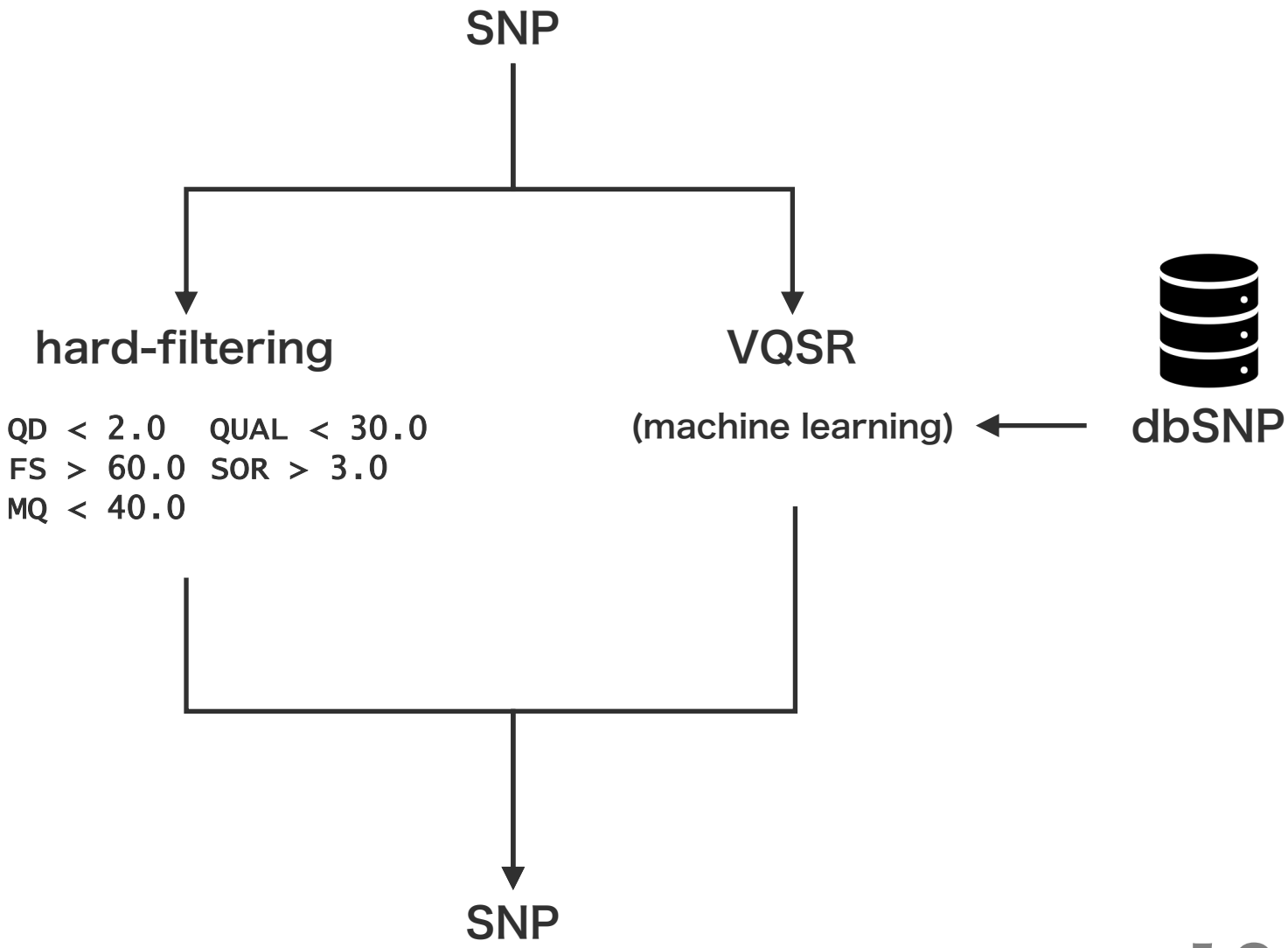


GATK outputs five major types of variant

- SNP: single nucleotide polymorphism
- INDEL: insetion and deletion
- MIXED: combination of SNPs and indels at a single position
- MNP: multi-nucleotidde polymorphism
- SYMBOLIC

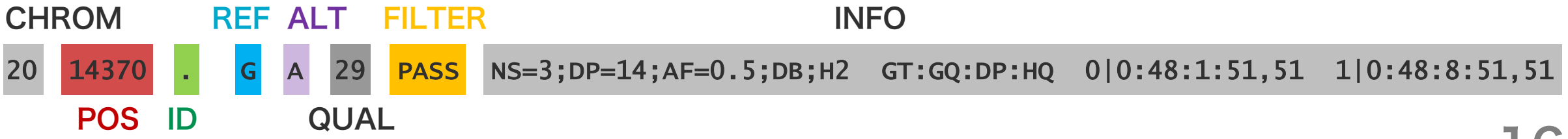
Variant calling





VCF format

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID REF  ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   .  G    A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .  T    A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3:.,.
20     1110696 .  A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4:.,.
20     1230237 .  T    .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2:.,.
20     1234567 .  GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
:
:
:
```

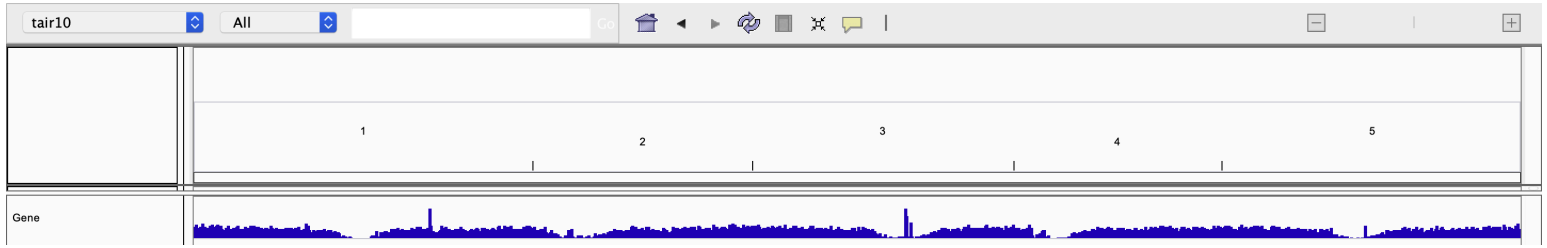


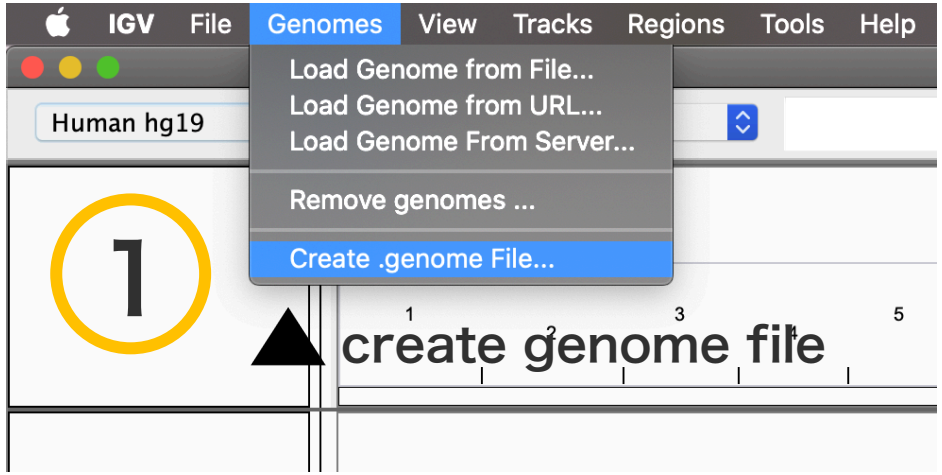
IGV

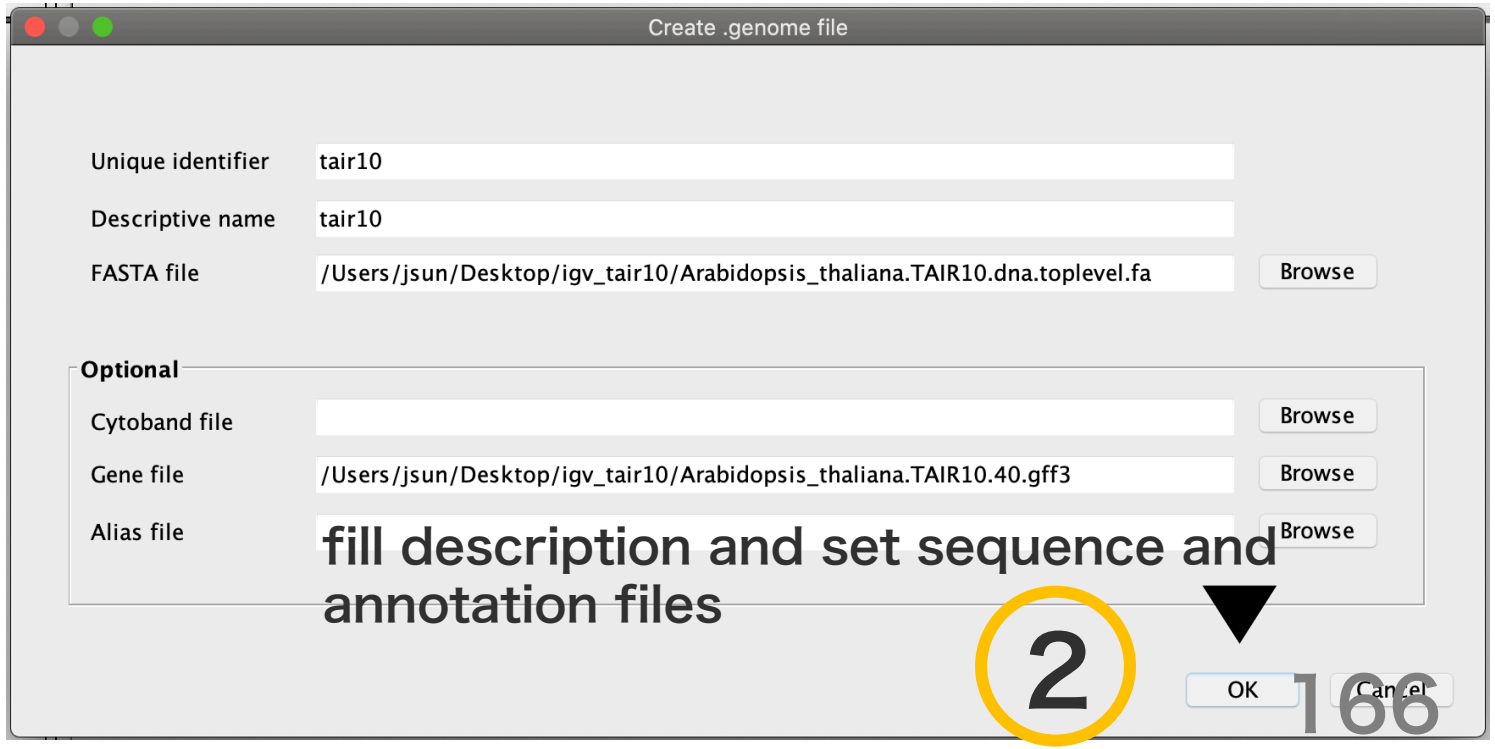
.fasta

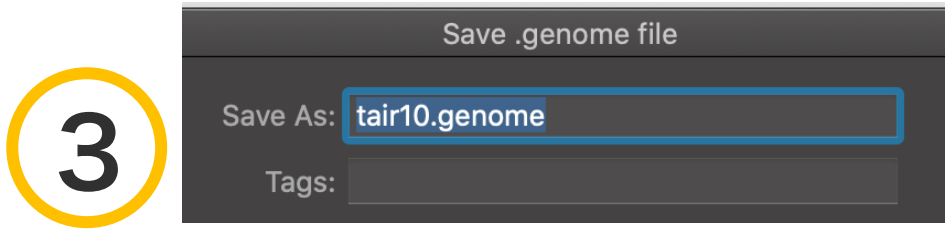
.gff

.genome





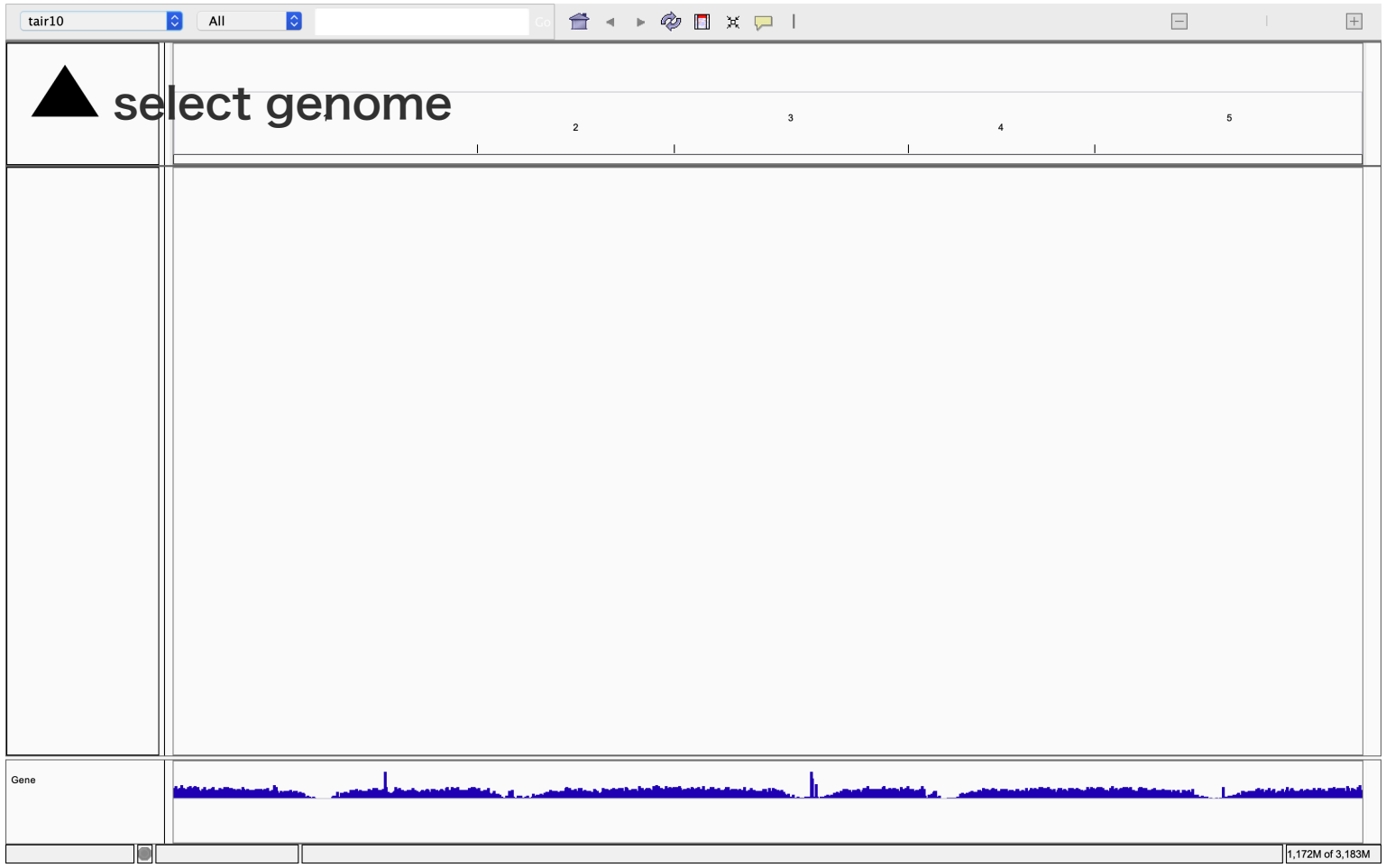




IGV

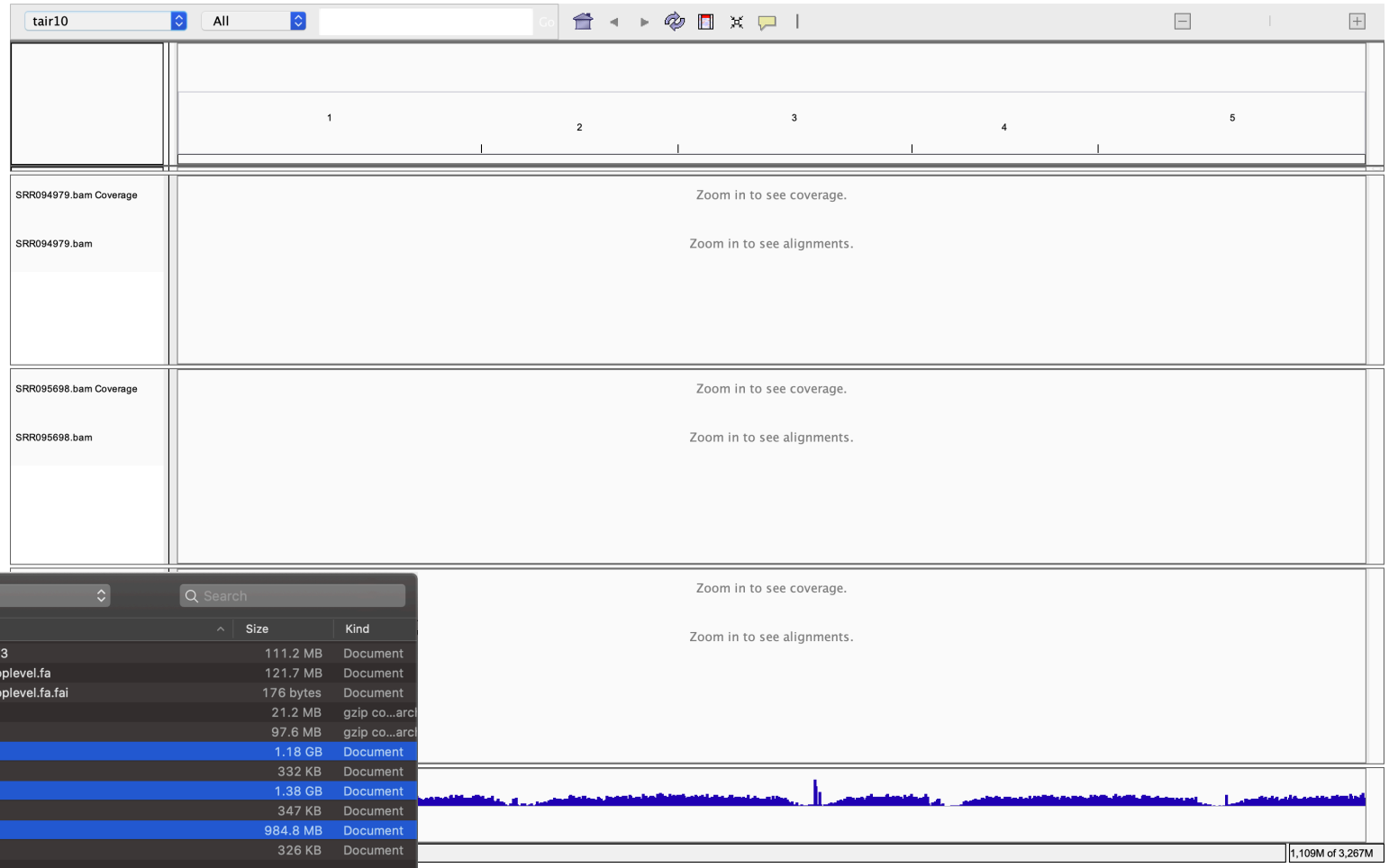
.bam

.bai

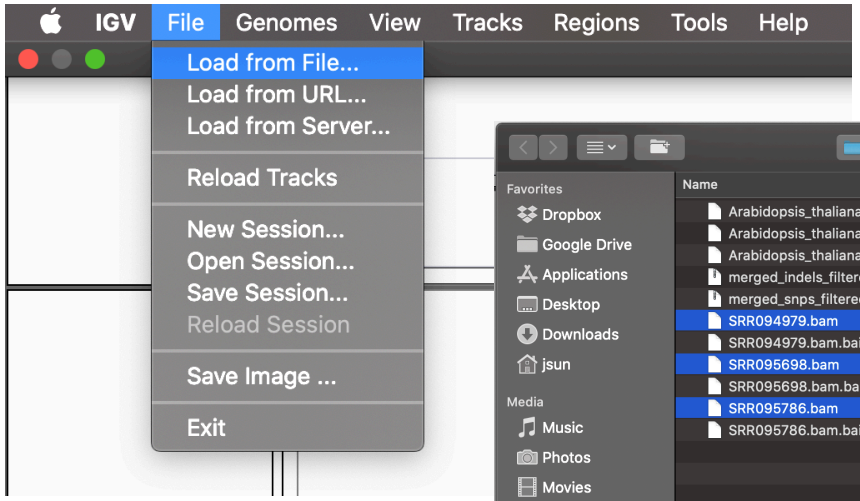


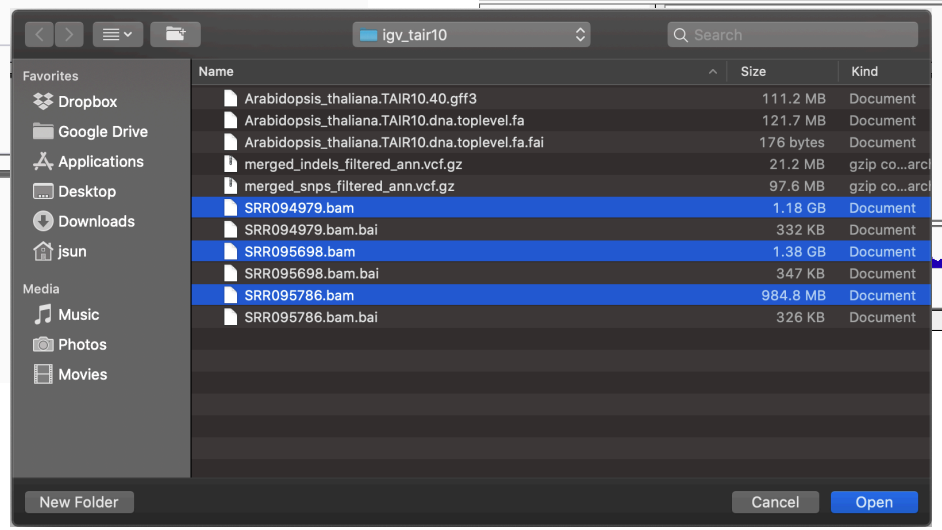
IGV

.bam
.bai



▼ select BAM files.



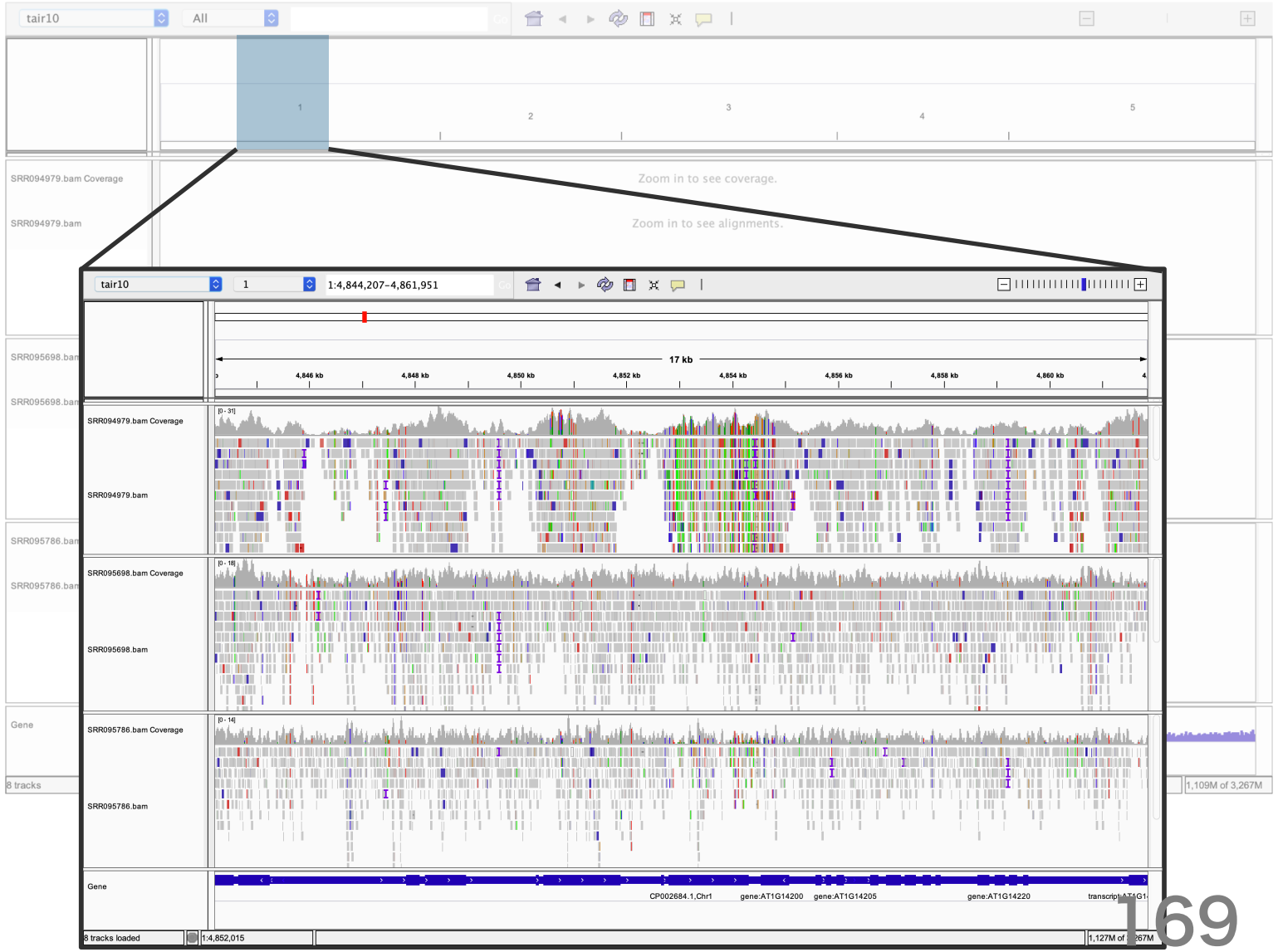


IGV

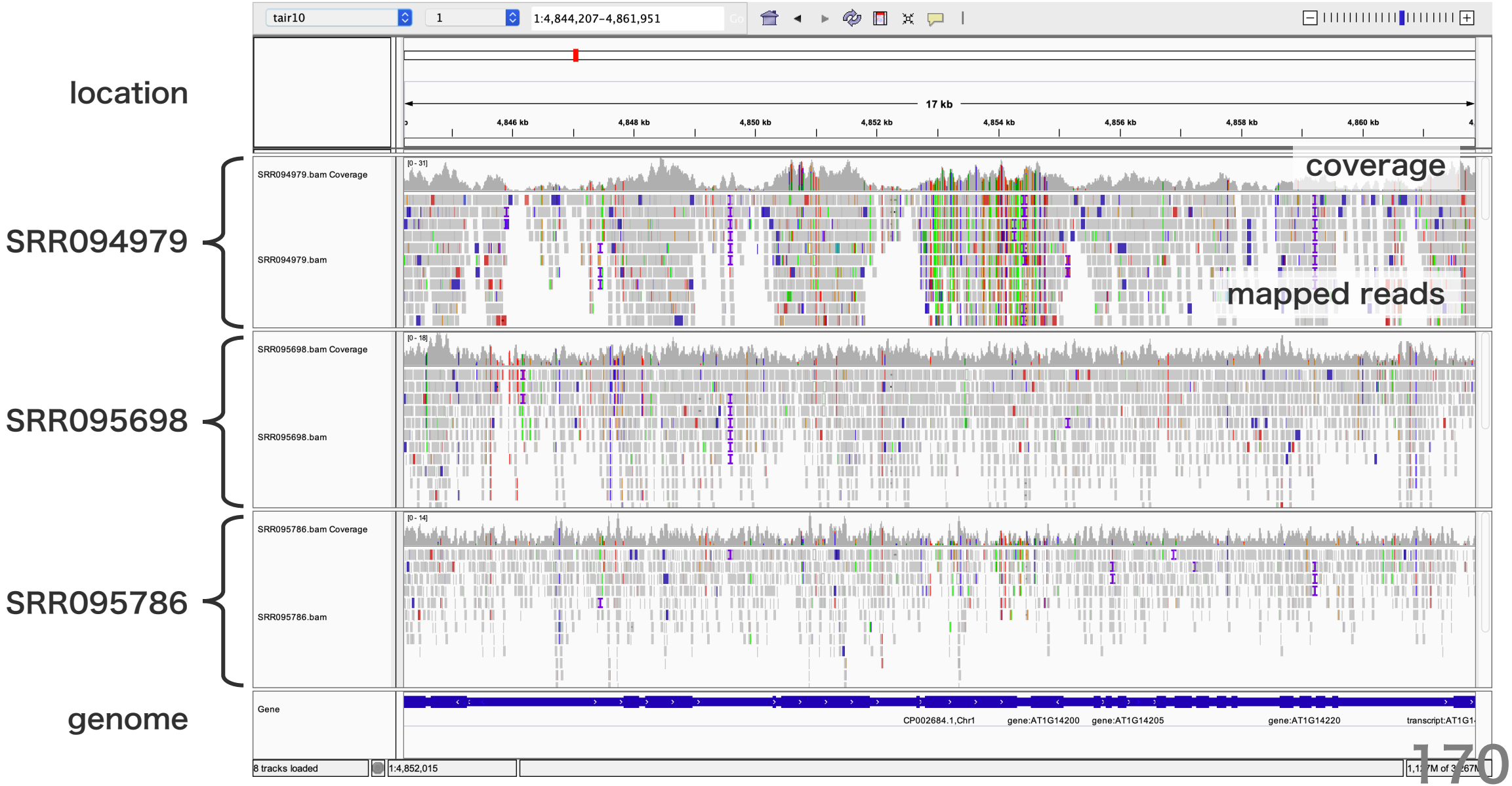
.bam

.bai

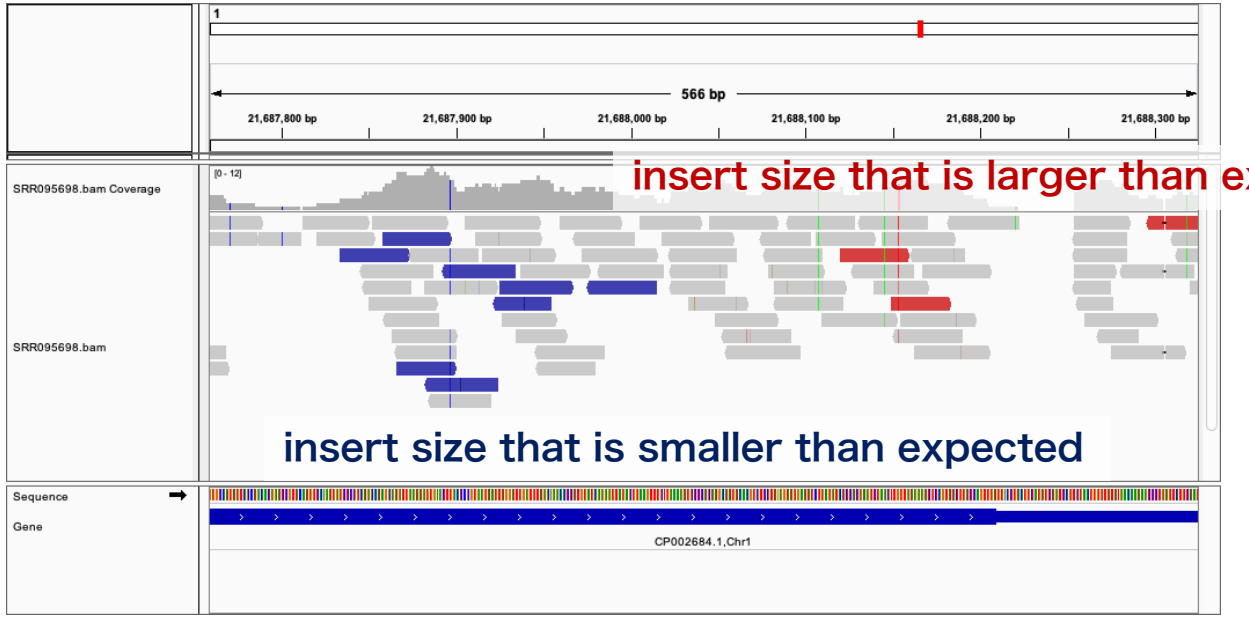

igv


tair10
All
SRR094979.bam Coverage
Zoom in to see coverage.
SRR094979.bam
Zoom in to see alignments.
tair10
1
1:4,844,207-4,861,951
17 kb
4,846 kb
4,848 kb
4,850 kb
4,852 kb
4,854 kb
4,856 kb
4,858 kb
4,860 kb
SRR094979.bam Coverage
SRR094979.bam
SRR095698.bam Coverage
SRR095698.bam
SRR095786.bam Coverage
SRR095786.bam
Gene
CP002684.1,Chr1
gene:AT1G14200
gene:AT1G14205
gene:AT1G14220
8 tracks loaded
1:4,852,015
1,109M of 3,267M

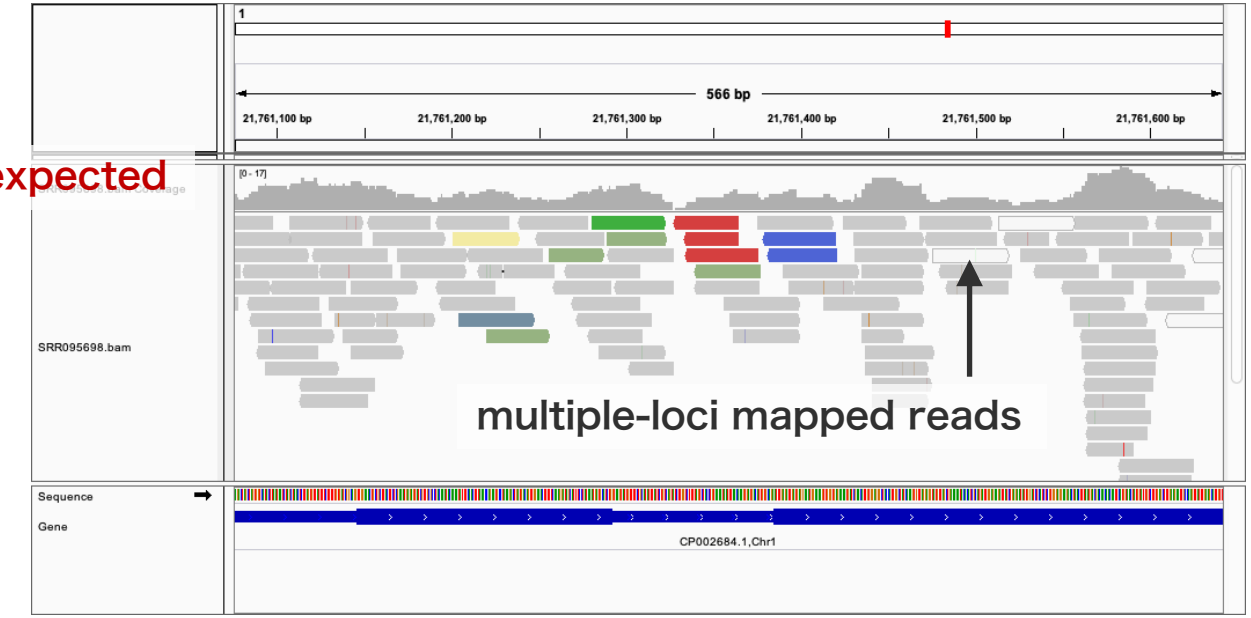
IGV


location
SRR094979
SRR095698
SRR095786
genome
tair10
1
1:4,844,207-4,861,951
17 kb
4,846 kb
4,848 kb
4,850 kb
4,852 kb
4,854 kb
4,856 kb
4,858 kb
4,860 kb
SRR094979.bam Coverage
SRR094979.bam
[0 - 31]
coverage
mapped reads
SRR095698.bam Coverage
SRR095698.bam
[0 - 18]
SRR095786.bam Coverage
SRR095786.bam
[0 - 14]
Gene
CP002684.1,Chr1
gene:AT1G14200
gene:AT1G14205
gene:AT1G14220
transcript:AT1G1
8 tracks loaded
1:4,852,015

IGV



insert size that is larger than expected



The pair reads that are coded by the chromosome on which their mates can be found