次世代シーケンサーとシーケンスデータ解析

東京科学大学　ILA国府台
中林　潤

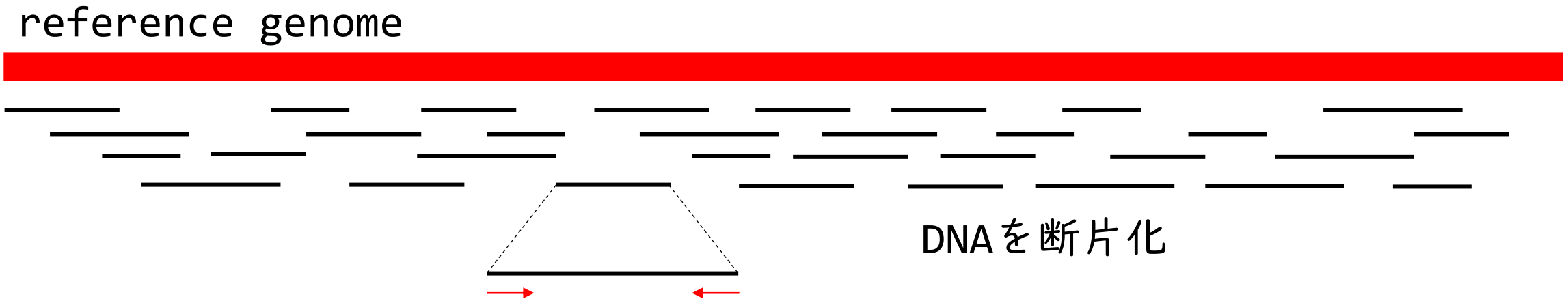
JH人材育成課　バイオインフォマティシャン育成講座　⑥

次世代シーケンサー（Next Generation Sequencer NGS）

- 1000$ゲノムプロジェクトの過程で開発されたシーケンサー
- 従来型と比較してパフォーマンスが劇的に増加している。

	従来型シーケンサー	次世代シーケンサー
ヒトゲノムを読むのにかかる日数	13年間	3.5日
ヒトゲノムを読むのにかかる費用	1400万ドル	1000ドル

NGSデータ



GCATCGATCGAGC
GCATGCCGCAT
AGGTGCATG
…AGGTGCATGCCGCATCGATCGAGC…

リード数：N
リード長：L
ゲノム長：G
coverage $C = (N \times L)/G$

大量のDNA断片の配列を参照ゲノムにマッピングして元の塩基配列を再構成する

reference genome

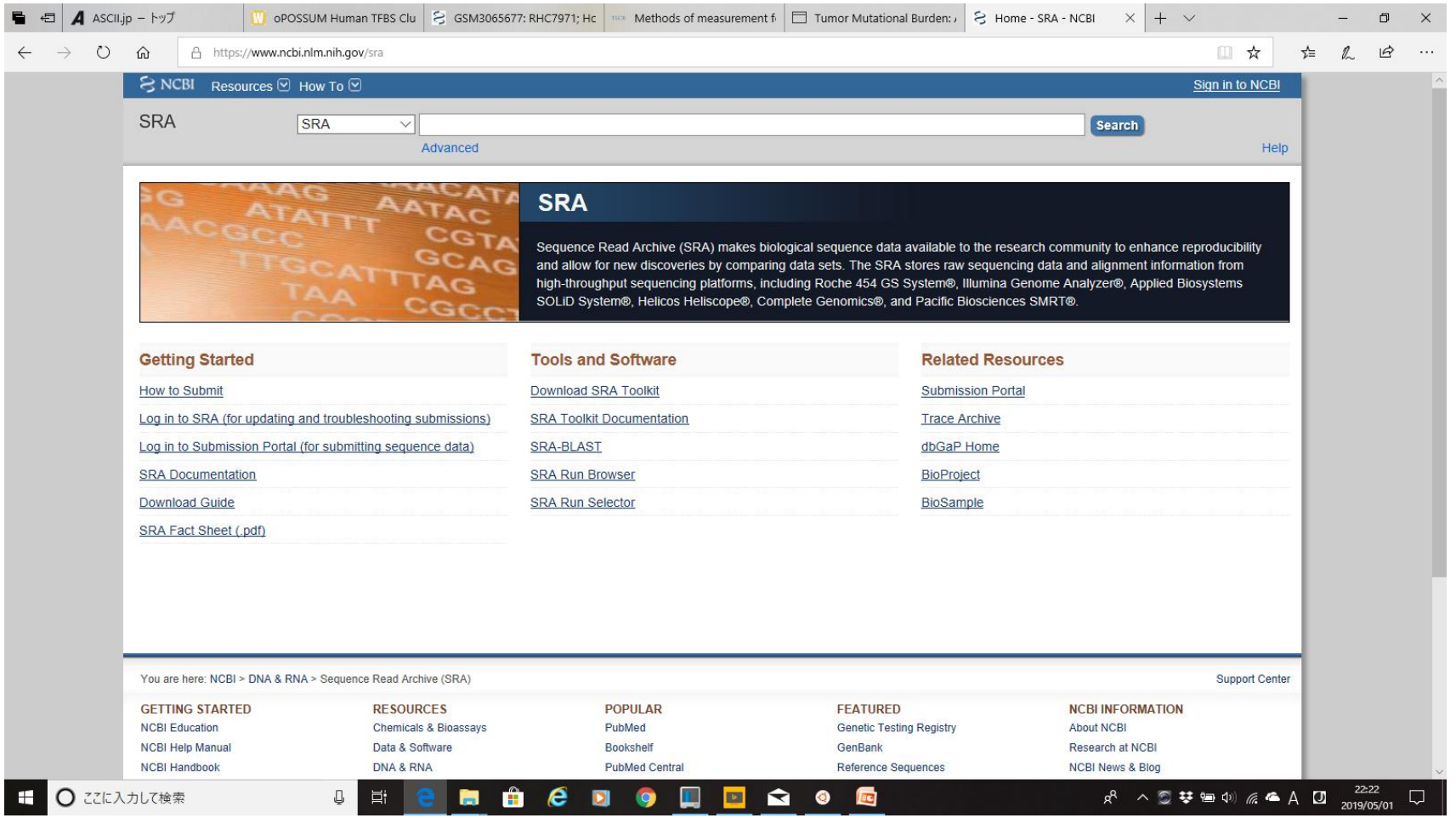
- University of California Santa Cruz（UCSC）版
  Human hg38 2013年12月
  Mouse mm39 2020年1月

- Genome Reference Consortium（GRC）版
  Human Grch38
  Mouse GRCm39

- Refseq,Ensembleから入手することも出来る

Sequence Read Archive（SRA）

raw sequence dataが登録されているデータベース



https://www.ncbi.nlm.nih.gov/sra

NGSデータ処理の流れ

1．シーケンスデータのクオリティチェック

2．参照ゲノム配列にマッピング

3．解析

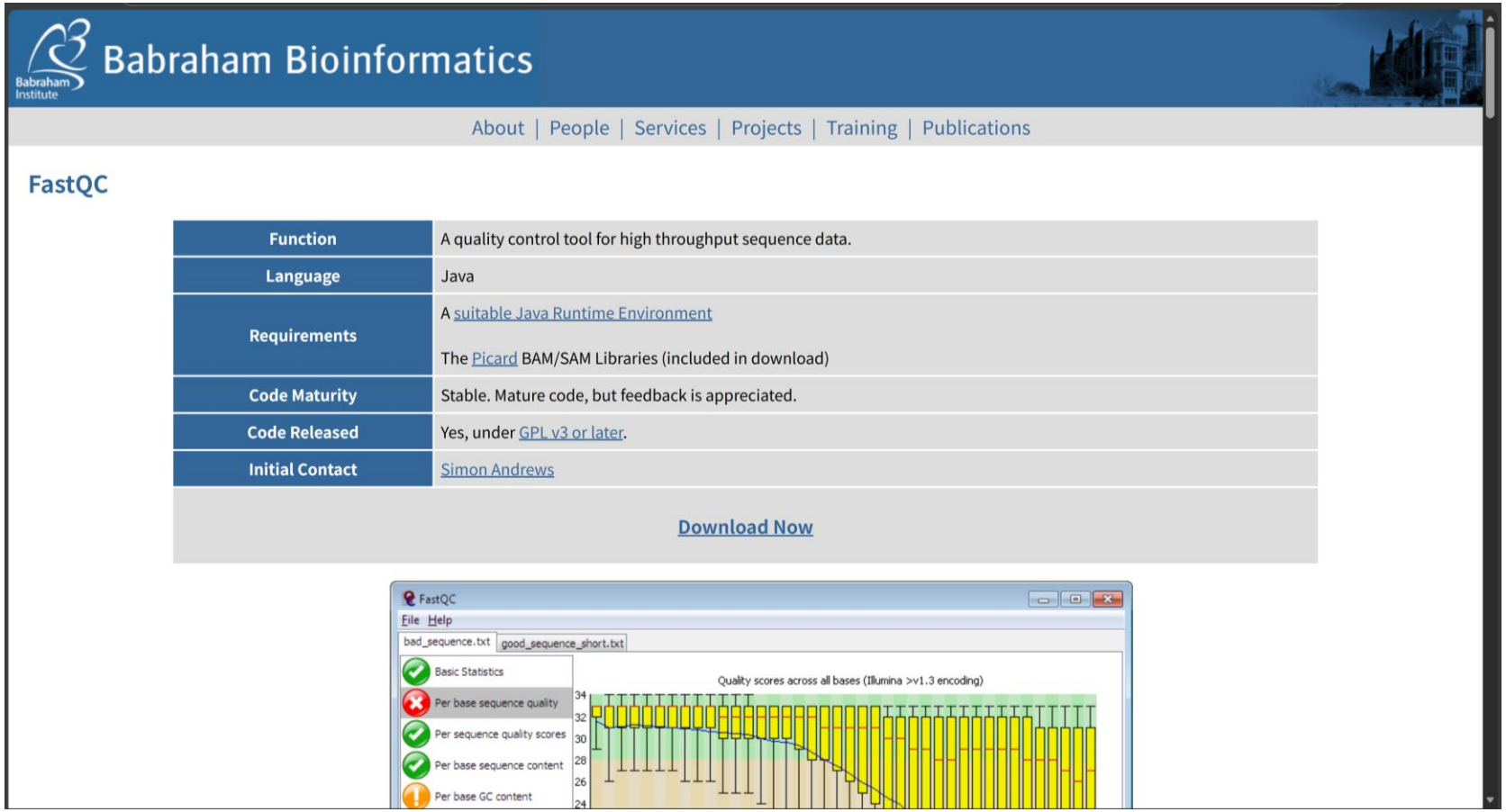
FASTQフォーマット

- 1行目：@配列ID
- 2行目：塩基配列
- 3行目：+配列ID 説明
- 4行目：クオリティー値（シーケンスエラーの生じる確率）

```
@Seq-ID
AGGTGCATCGATGCGCGAATAAT
+
!1’’*))++//?”AAA{{
```

FASTQファイルのクオリティチェック

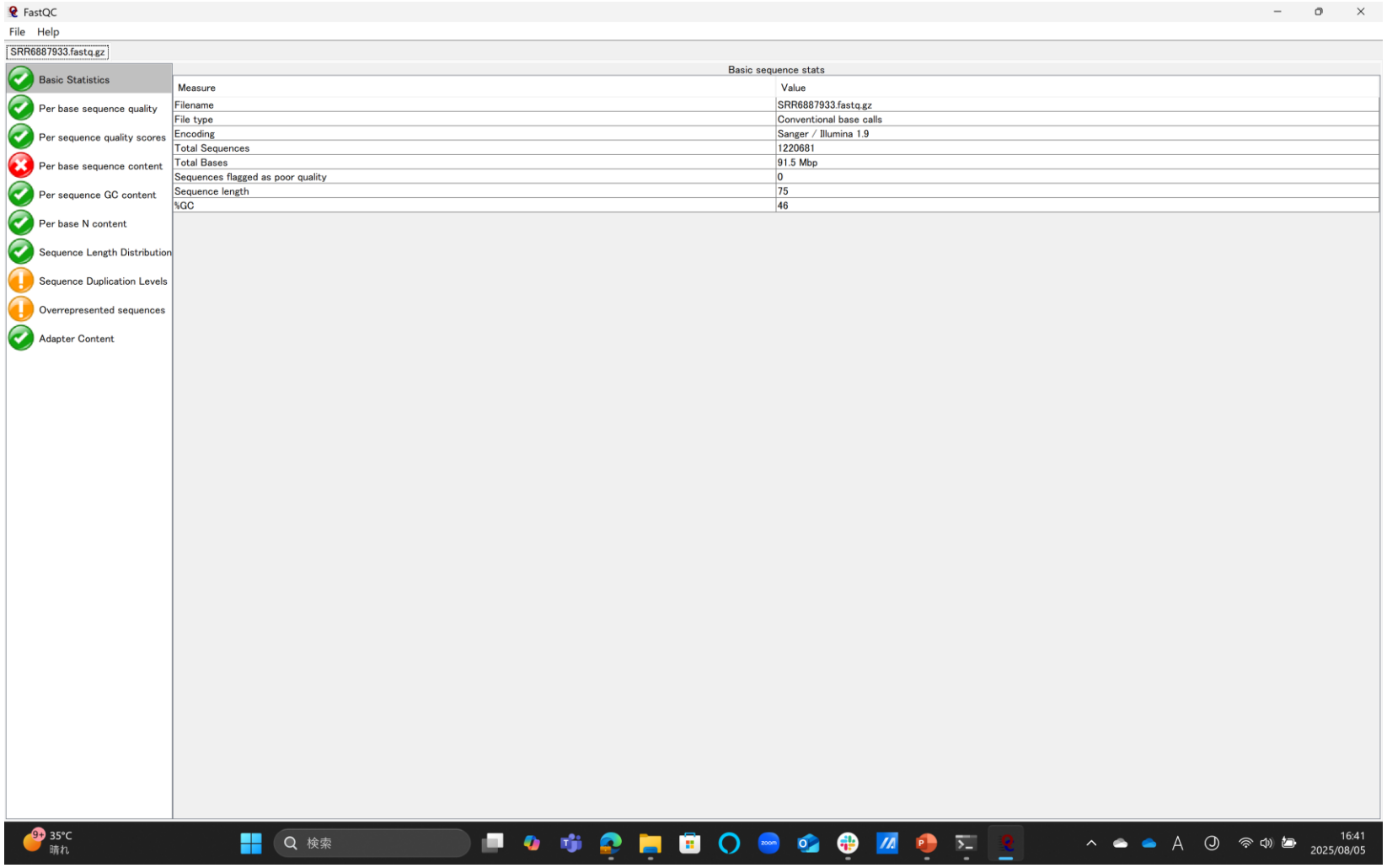
FastQC：シーケンスデータのクオリティチェック用プログラム
Javaで書かれていて、MacやWindowsで実行可能



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

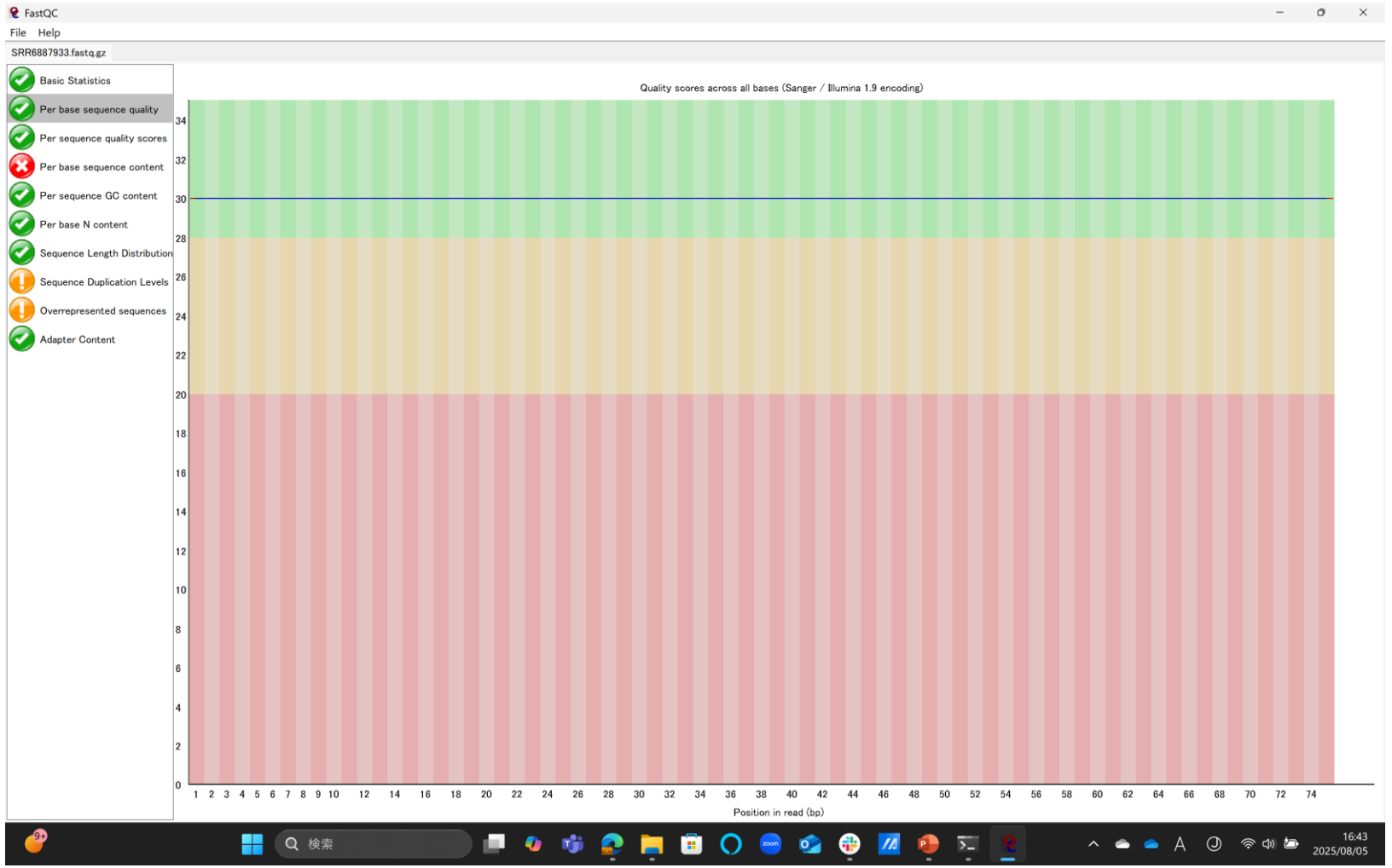
FastQC実行画面

ダウンロードしたプログラムを実行できる。
GUIが整備されている。



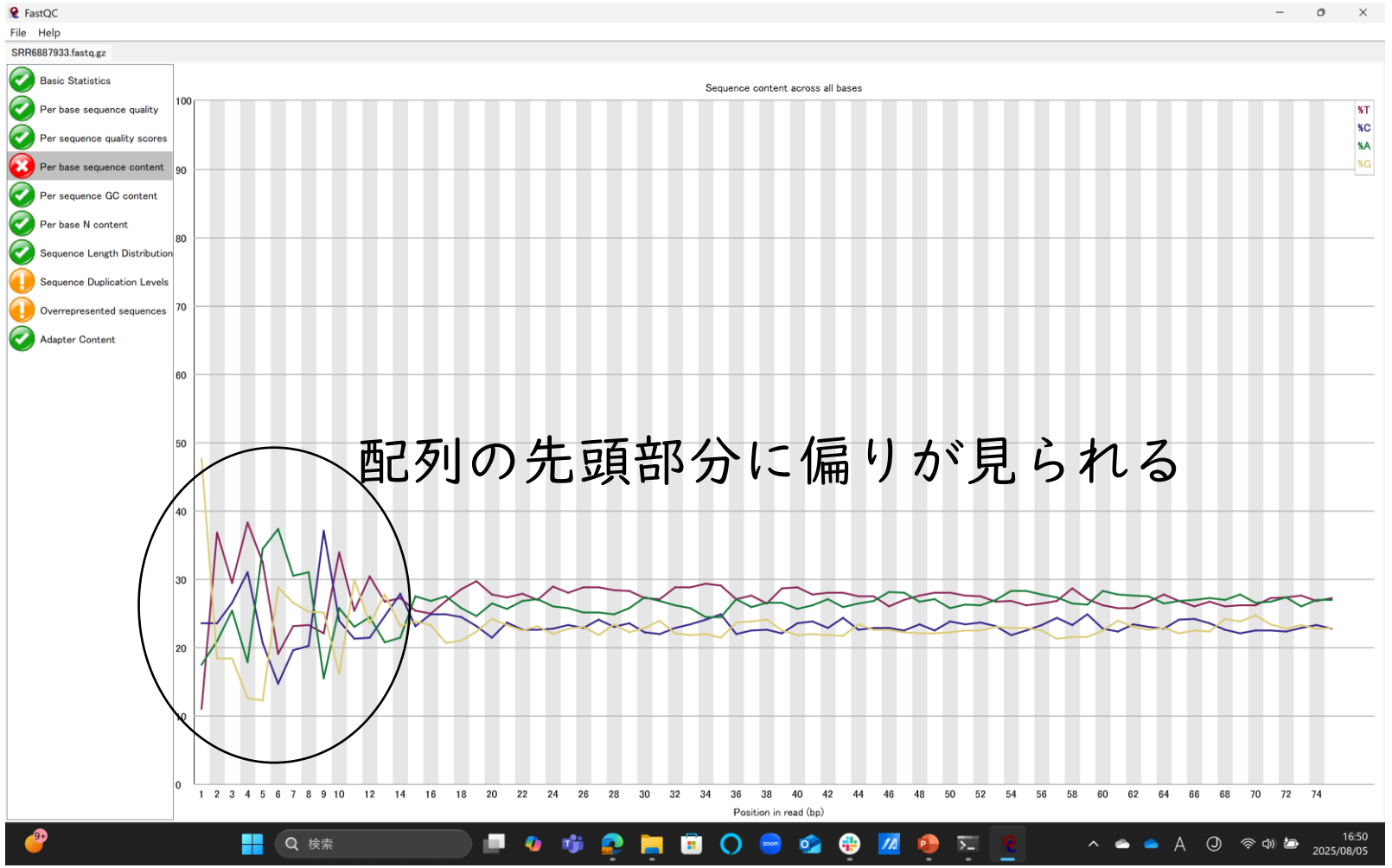
Per Base Sequence Quality

シーケンスクオリティを示す。
28以上が合格。



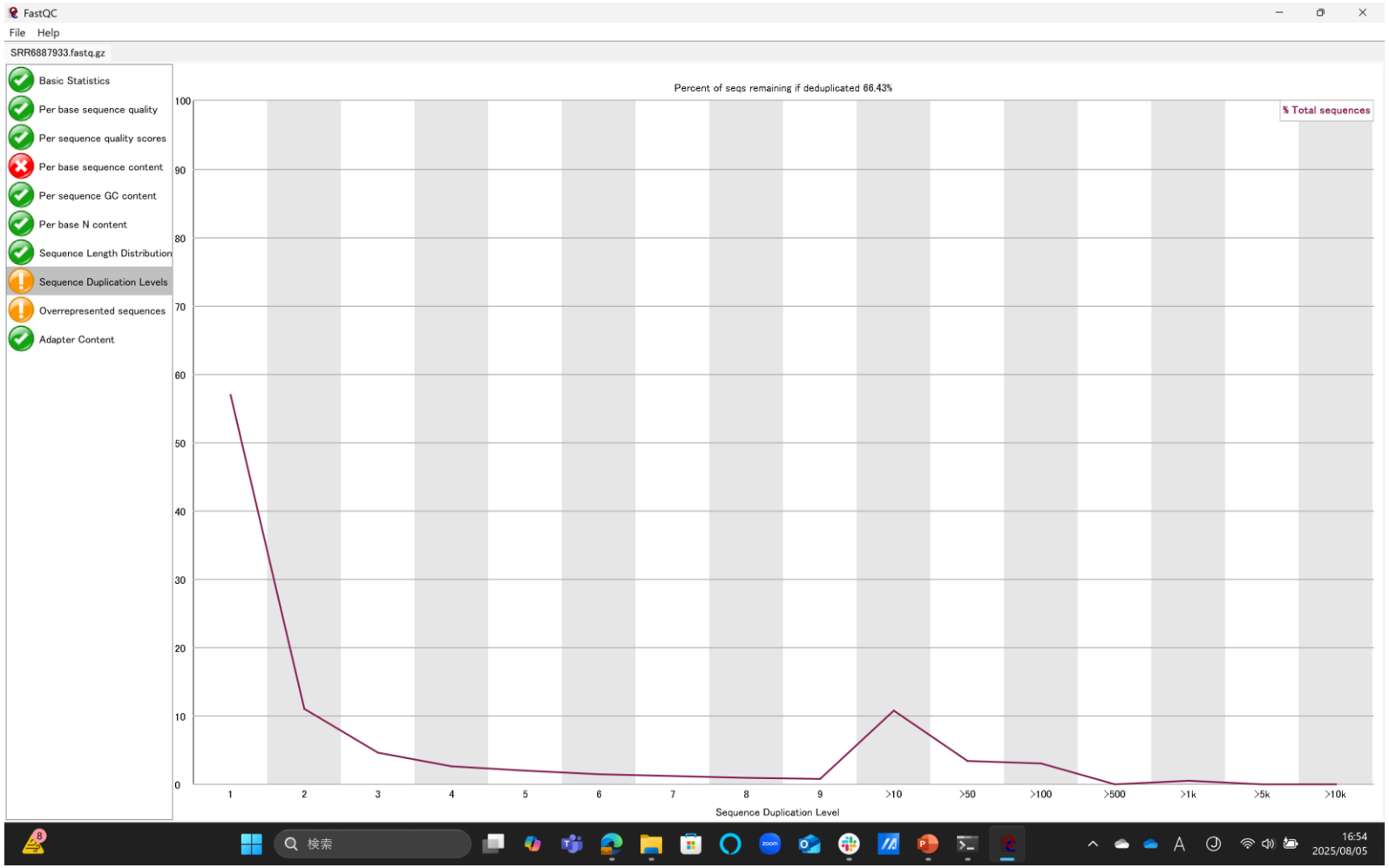
Per Base Sequence Content

各塩基の割合を示す。
配列全体で均等に分布していることが望ましい。



Sequence Duplication Level

重複している配列の割合を示す。
non-unique sequenceが20%以上で警告が出る。



マッピング

- シーケンス断片を参照ゲノムの配列に割り当てる。

BWA
   https://bio-bwa.sourceforge.net/
Bowtie2
   https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
HISAT2
   https://daehwankimlab.github.io/hisat2/manual/
STAR
   https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

Burrows Wheeler transformation

- Michael BurrowsとDavid Wheelerによって提唱されたデータ圧縮アルゴリズム

- ファイル圧縮に応用されている（bzip2圧縮）

BWTのアルゴリズム　圧縮①

original
X = abracadabra
末尾に$を付ける
左に一文字ずつシフトして
ローテーション

a	b	r	a	c	a	d	a	b	r	a	$
b	r	a	c	a	d	a	b	r	a	$	a
r	a	c	a	d	a	b	r	a	$	a	b
a	c	a	d	a	b	r	a	$	a	b	r
c	a	d	a	b	r	a	$	a	b	r	a
a	d	a	b	r	a	$	a	b	r	a	c
d	a	b	r	a	$	a	b	r	a	c	a
a	b	r	a	$	a	b	r	a	c	a	d
b	r	a	$	a	b	r	a	c	a	d	a
r	a	$	a	b	r	a	c	a	d	a	b
a	$	a	b	r	a	c	a	d	a	b	r
$	a	b	r	a	c	a	d	a	b	r	a

BWTのアルゴリズム　圧縮②

アルファベット順にソート
$はaよりも先

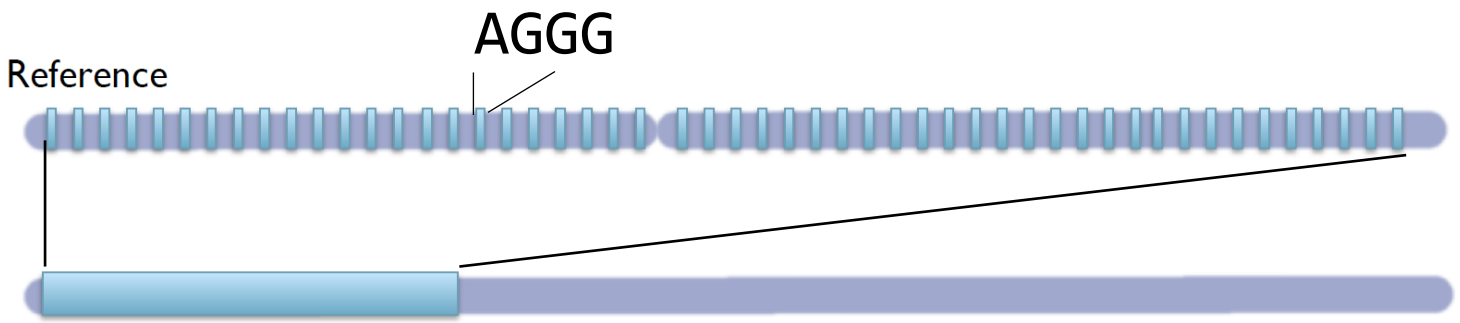
Lが返還後の文字列

BWT(X) = ard$rcaaaabb

辞書式にソートしているので
同じアルファベットが集まる

F										L
$	a	b	r	a	c	a	d	a	b	r	a
a	$	a	b	r	a	c	a	d	a	b	r
a	b	r	a	$	a	b	r	a	c	a	d
a	b	r	a	c	a	d	a	b	r	a	$
a	c	a	d	a	b	r	a	$	a	b	r
a	d	a	b	r	a	$	a	b	r	a	c
b	r	a	$	a	b	r	a	c	a	d	a
b	r	a	c	a	d	a	b	r	a	$	a
c	a	d	a	b	r	a	$	a	b	r	a
d	a	b	r	a	$	a	b	r	a	c	a
r	a	$	a	b	r	a	c	a	d	a	b
r	a	c	a	d	a	b	r	a	$	a	b

BWTの利点

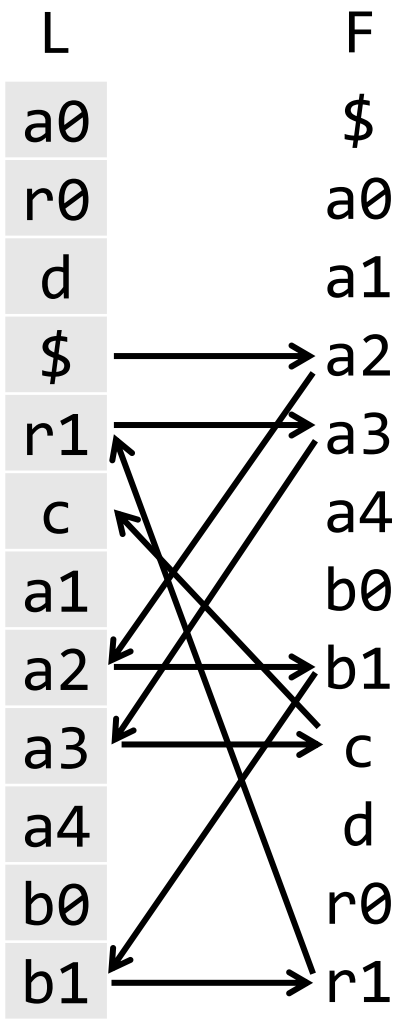


GGGA
GGGC
GGGT
…

- 同じ文字が固まる傾向がある
- 検索しやすい

BWTのアルゴリズム　複合

L	F
a0	$
r0	a0
d	a1
$	a2
r1	a3
c	a4
a1	b0
a2	b1
a3	c
a4	d
b0	r0
b1	r1



Lの$から出発して
L→F→L→F…と
たどっていくと
オリジナルの文字列を
復元できる

a2b1r1a3ca4…

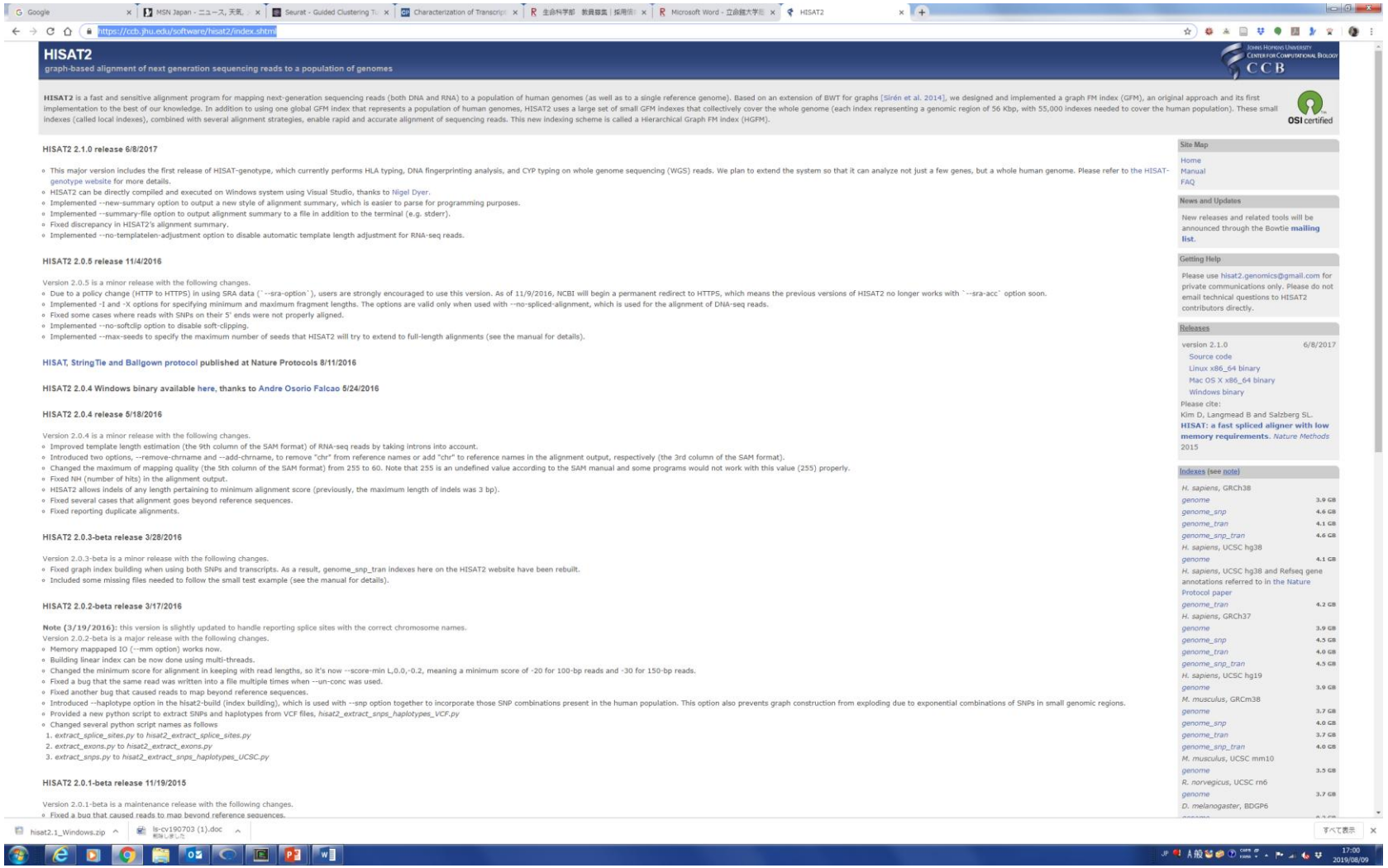
LとFを並べる
（複数個ある文字には番号を付けてある）

簡単に複合できる

hisat2

- bowtie2の後継マッピングツール。
- Burrows Wheeler transform（データ圧縮技術）を利用している。
- bowtie2より高速であるが、メモリの消費は少ない。

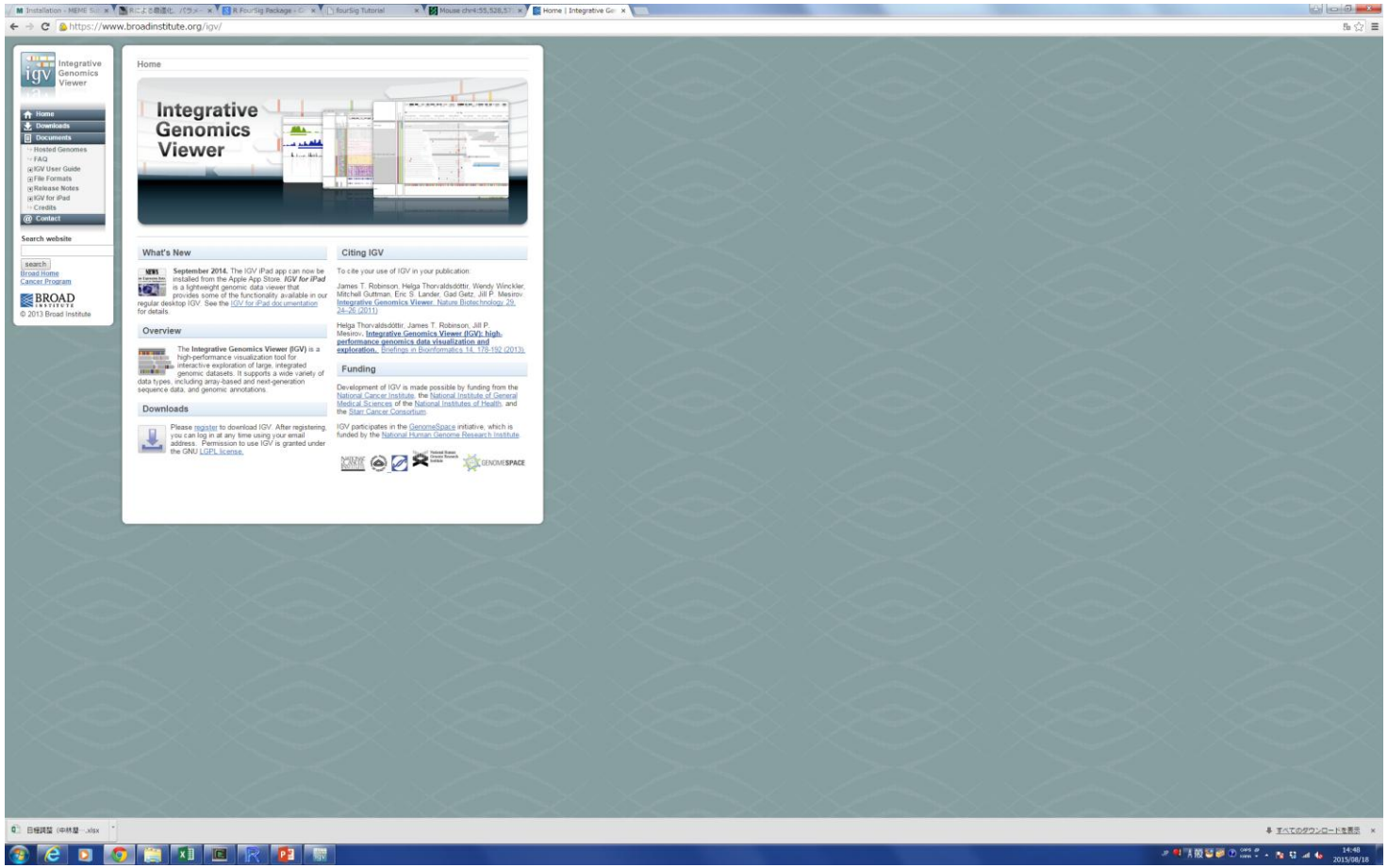
hisat2



https://ccb.jhu.edu/software/hisat2/index.shtml

Integrative Genomics Viewer（IGV）

マッピングされた配列断片の可視化ツール

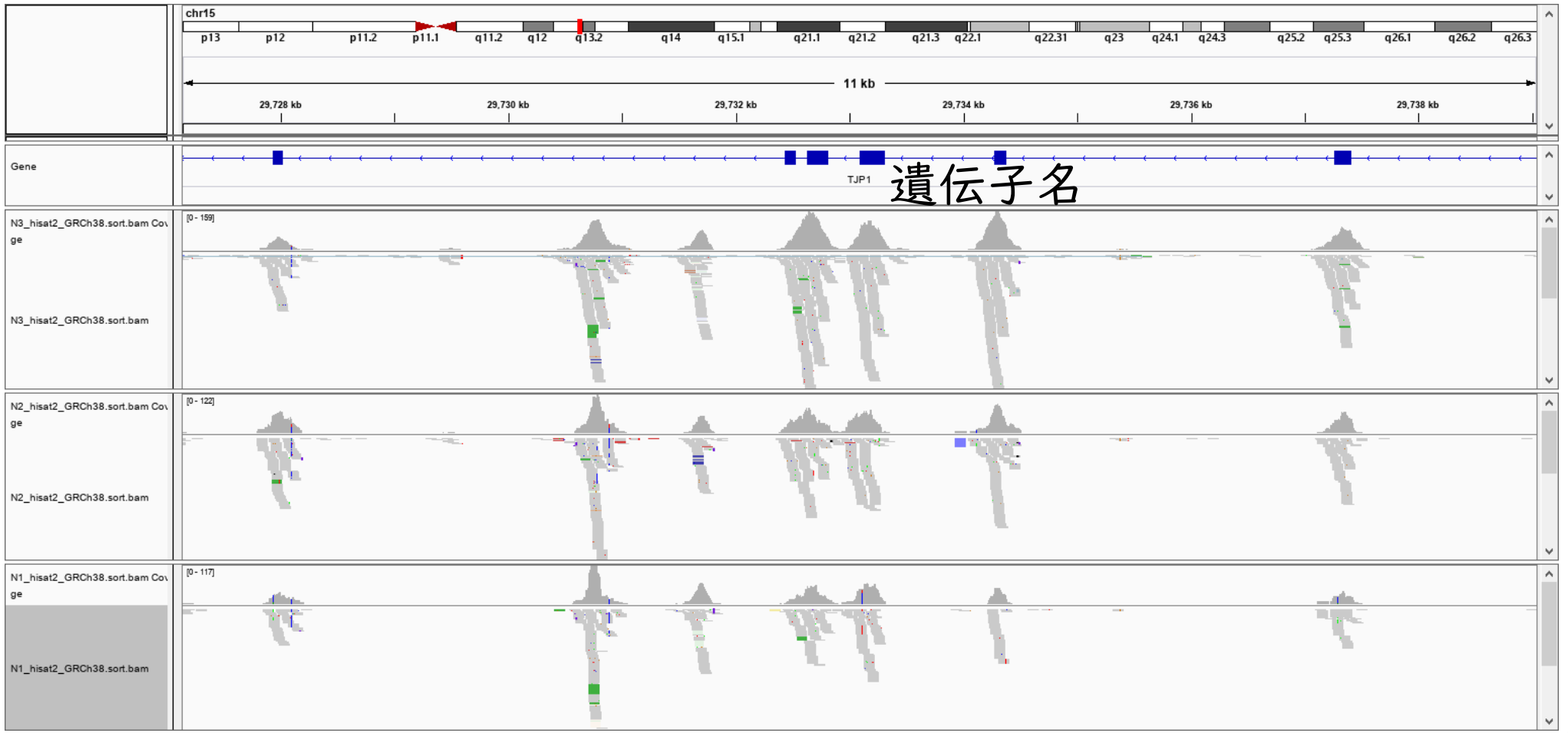


http://www.broadinstitute.org/igv/

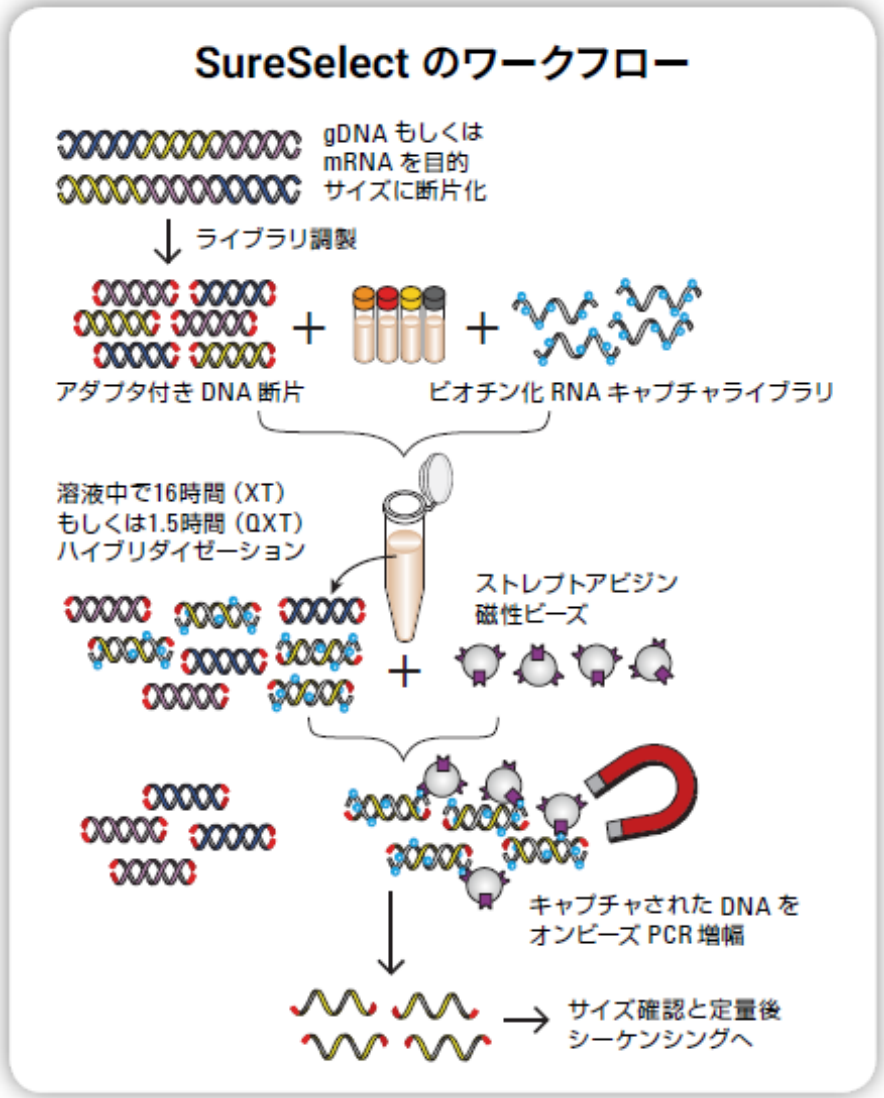
配列断片の可視化

染色体番号

参照ゲノム

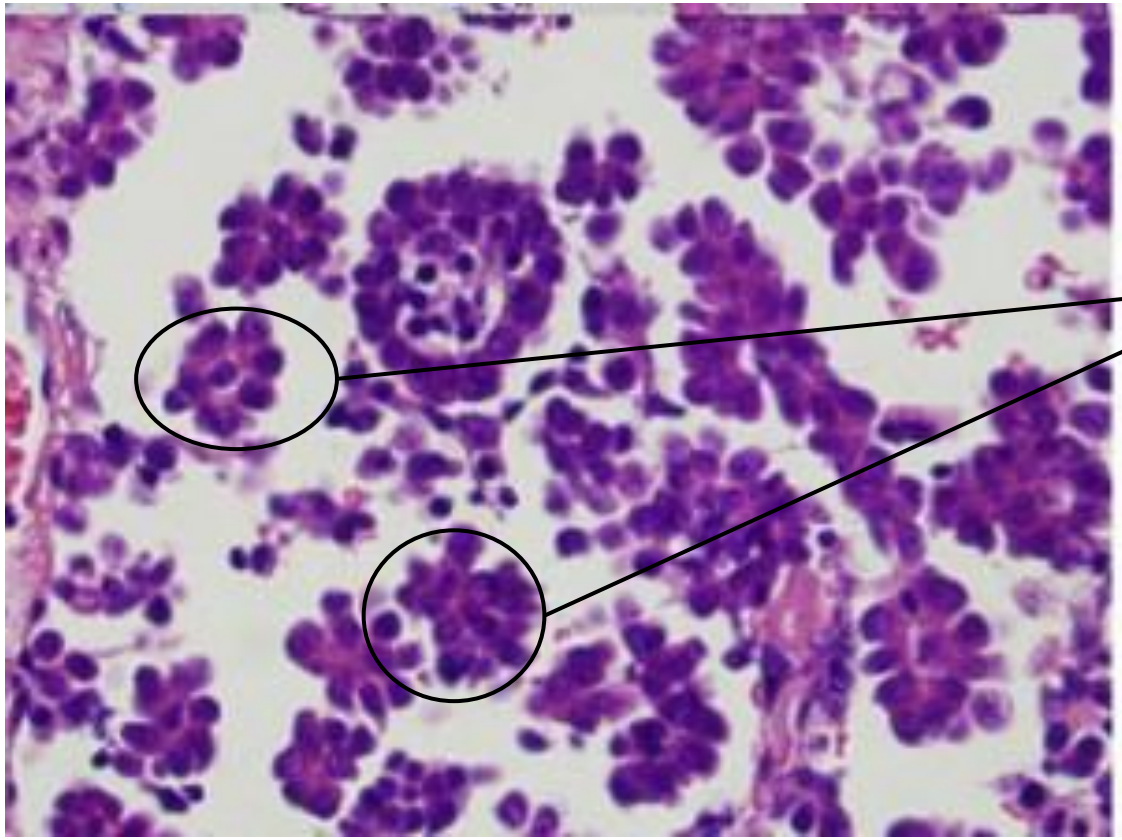


Whole Exome Sequencing（WES）



- exon（coding sequence）だけを集める
- ゲノムの1～2%がエクソン
- 機能領域にフォーカスして解析
- シーケンスコストを節約できる

Whole Exome Sequencing（WES）



微小乳頭状増殖優位型

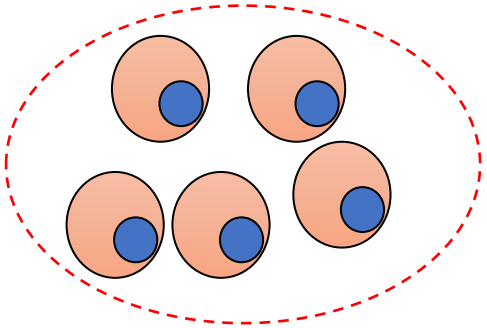
https://www.pathologyoutlines.com/topic/lungtumormicropapillary.html

Whole Exome Sequencing（WES）

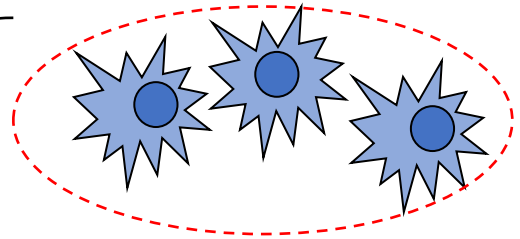
82歳　女性　非喫煙者

laser microdissection

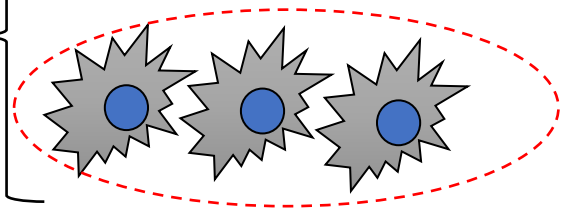
正常組織



がん組織



非微小乳頭状増殖優位型



微小乳頭状増殖優位型

	total DNA (ng)
正常	153.7
非微小乳頭状増殖優位型	10.3
微小乳頭状増殖優位型	195.7

Whole
Exome
Sequencing

Whole Exome Sequencing (WES)

	Total Reads (100bp paired end)	Overall Mapping Rate	Coverage
正常	58,037,525	98.13%	193×
非微小乳頭状増殖優位型	62,830,751	97.69%	207×
微小乳頭状増殖優位型	64,318,300	97.47%	210×

$$C = \text{mapped read counts} \times 2 \times 100 / (3{,}000{,}000{,}000 \times 0.02)$$

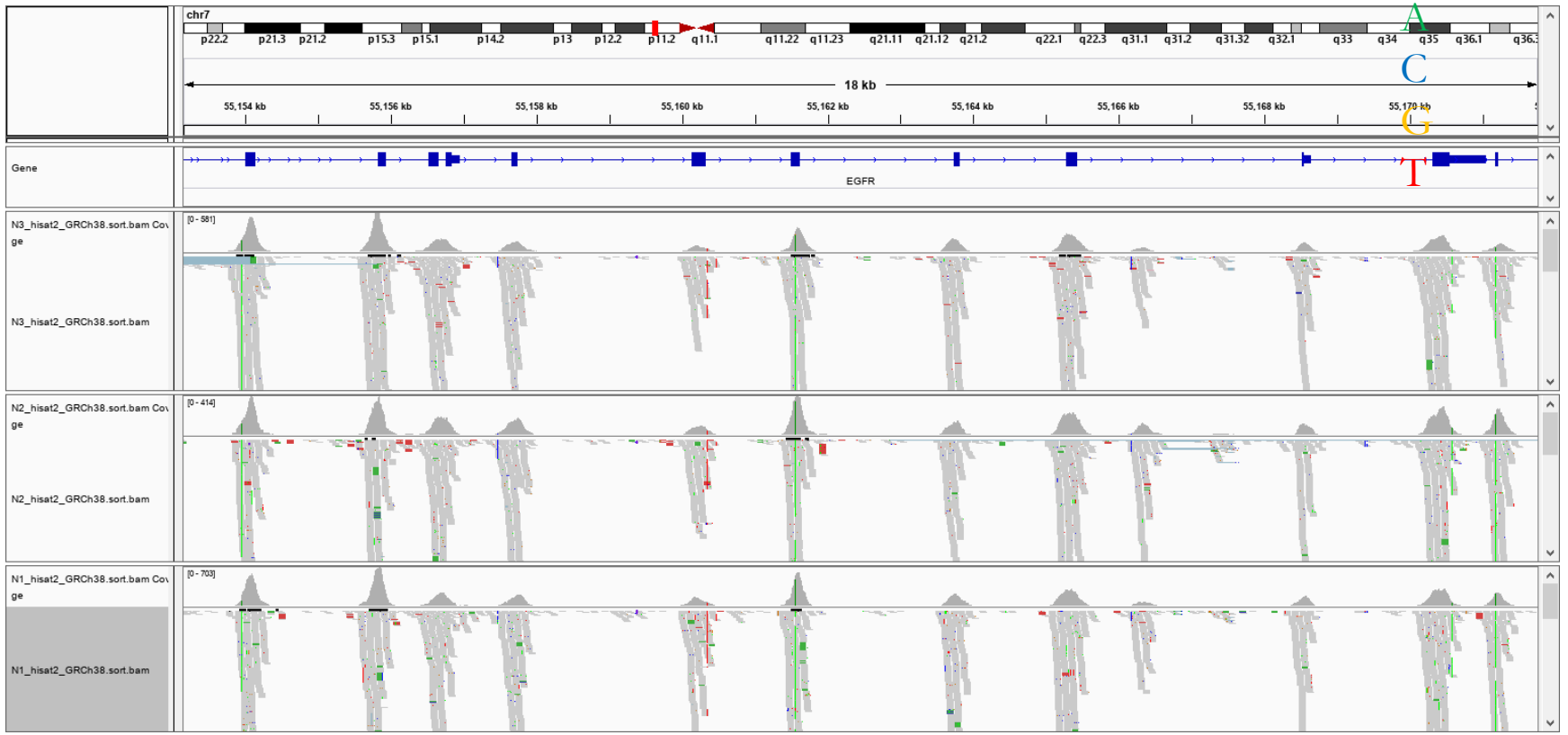
Whole Exome Sequencing（WES）

EGFR

正常

非微小乳頭状増殖優位型

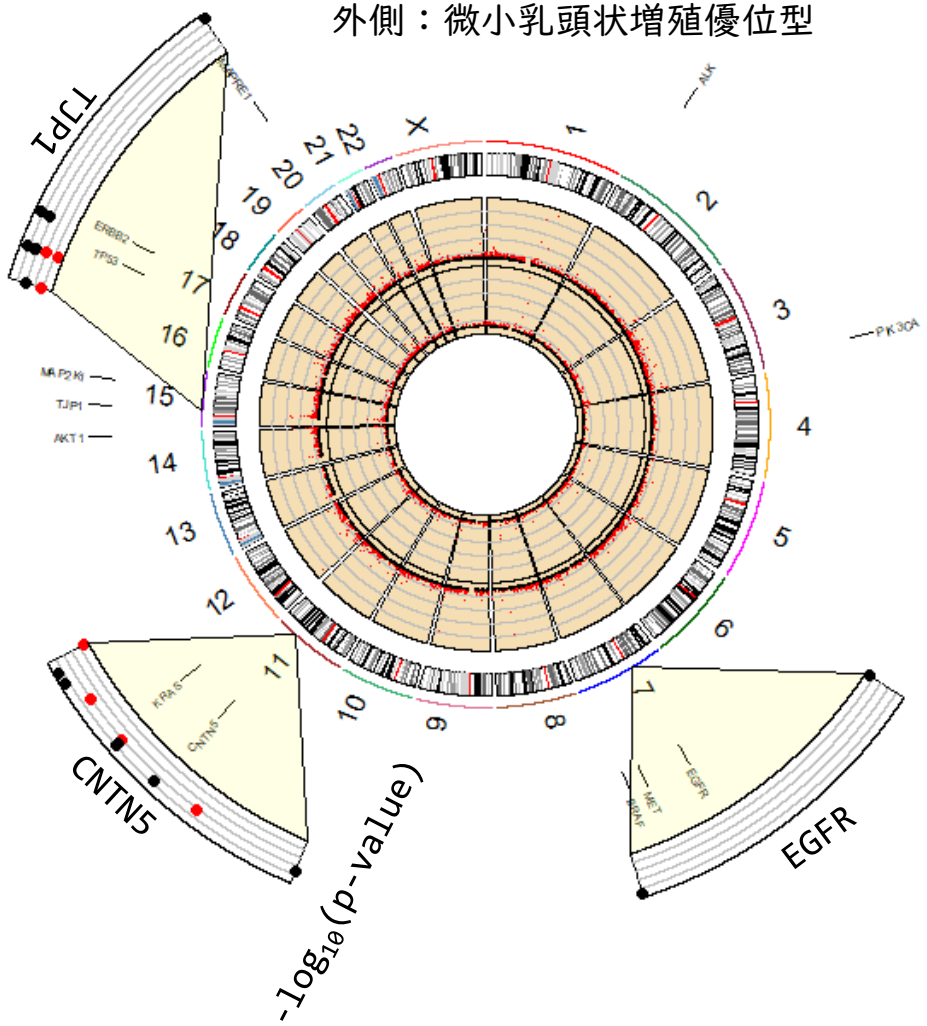
微小乳頭状増殖優位型



Whole Exome Sequencing (WES)

hi-confident somatic mutations

内側：非微小乳頭状増殖優位型
外側：微小乳頭状増殖優位型



	Somatic Mutations p<0.05	p<0.0001
非微小乳頭状増殖優位型	13,652	1671
微小乳頭状増殖優位型	21,925	3632

まとめ

- NGSデータはシーケンス配列の断片
- クオリティチェック→マッピング→解析
- カバー率$C = (N \times L)/G$
  N：リード数
  L：リード長
  G：ゲノム長