

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クソン育成講座

⑥



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

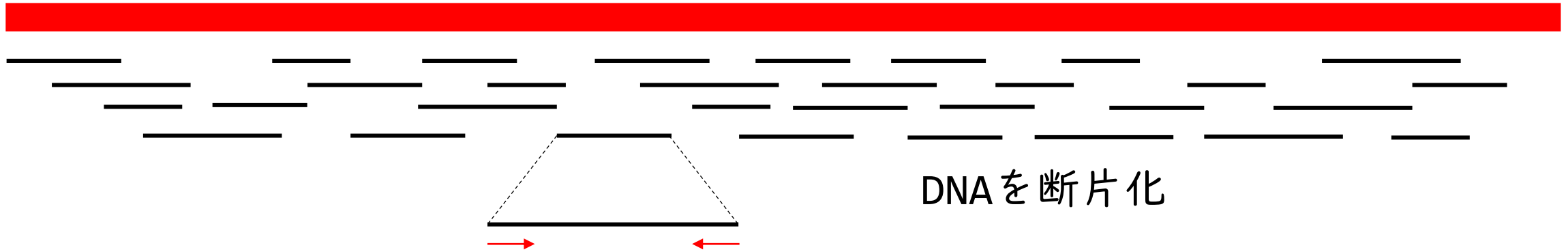
- 1000\$ゲノムプロジェクトの過程で開発されたシーケンサー
- 従来型と比較してパフォーマンスが劇的に増加している。

	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

reference genome



DNAを断片化

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

The screenshot displays the NCBI Sequence Read Archive (SRA) homepage. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a 'Sign in to NCBI' button. Below this is a search bar with 'SRA' selected in the dropdown and a 'Search' button. The main content area features a large banner with the SRA logo and a description: 'Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.' Below the banner are three columns of links: 'Getting Started' (How to Submit, Log in to SRA, Log in to Submission Portal, SRA Documentation, Download Guide, SRA Fact Sheet (.pdf)), 'Tools and Software' (Download SRA Toolkit, SRA Toolkit Documentation, SRA-BLAST, SRA Run Browser, SRA Run Selector), and 'Related Resources' (Submission Portal, Trace Archive, dbGaP Home, BioProject, BioSample). The footer contains a breadcrumb trail 'You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)' and a 'Support Center' link. The footer also has five columns of links: 'GETTING STARTED' (NCBI Education, NCBI Help Manual, NCBI Handbook), 'RESOURCES' (Chemicals & Bioassays, Data & Software, DNA & RNA), 'POPULAR' (PubMed, Bookshelf, PubMed Central), 'FEATURED' (Genetic Testing Registry, GenBank, Reference Sequences), and 'NCBI INFORMATION' (About NCBI, Research at NCBI, NCBI News & Blog). The Windows taskbar is visible at the bottom with the search bar and various application icons.

<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

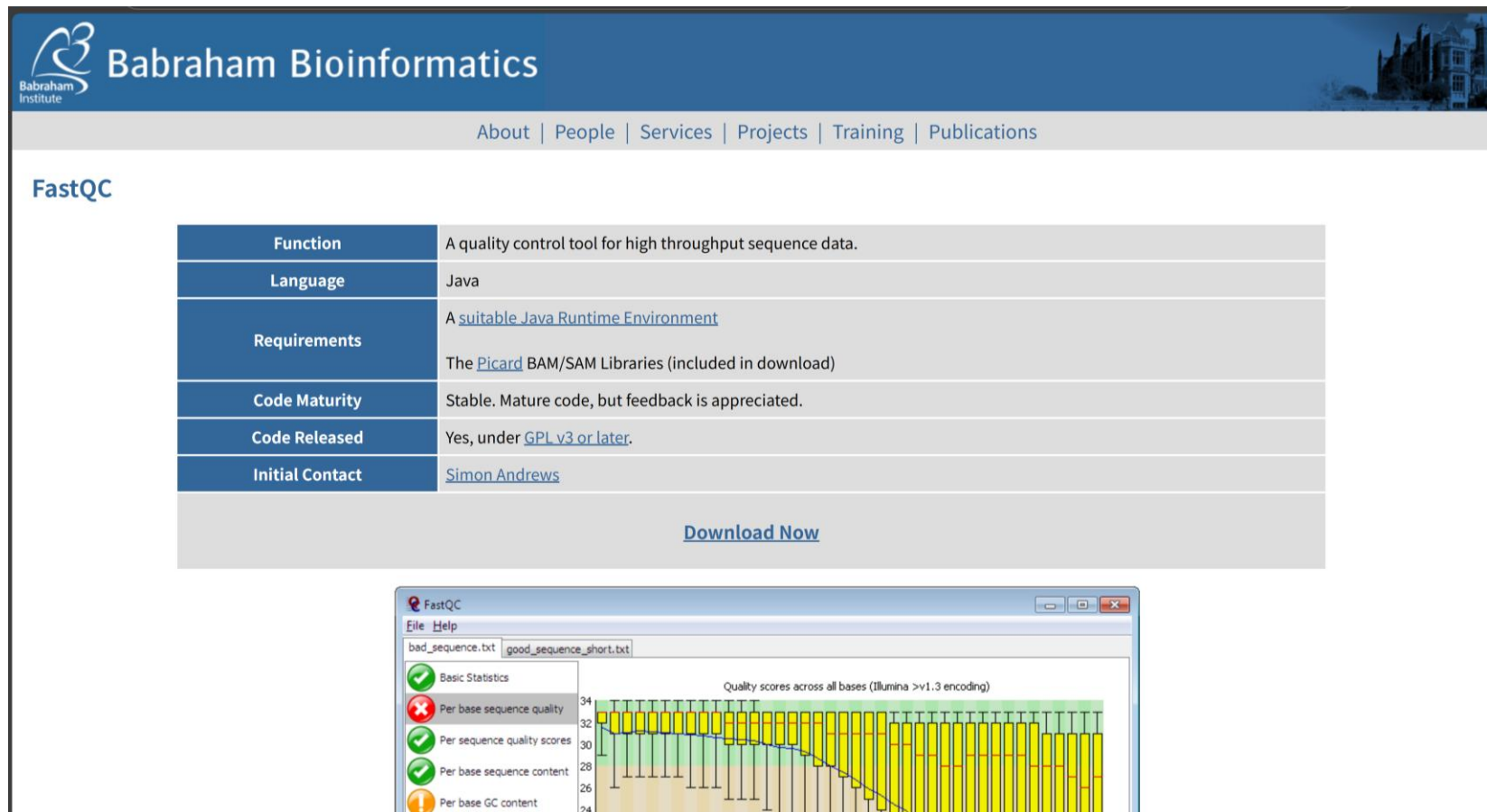
+

!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



The screenshot shows the Babraham Bioinformatics website. The header includes the Babraham Institute logo and the text "Babraham Bioinformatics". Below the header is a navigation menu with links for "About", "People", "Services", "Projects", "Training", and "Publications". The main content area is titled "FastQC" and contains a table with the following information:

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

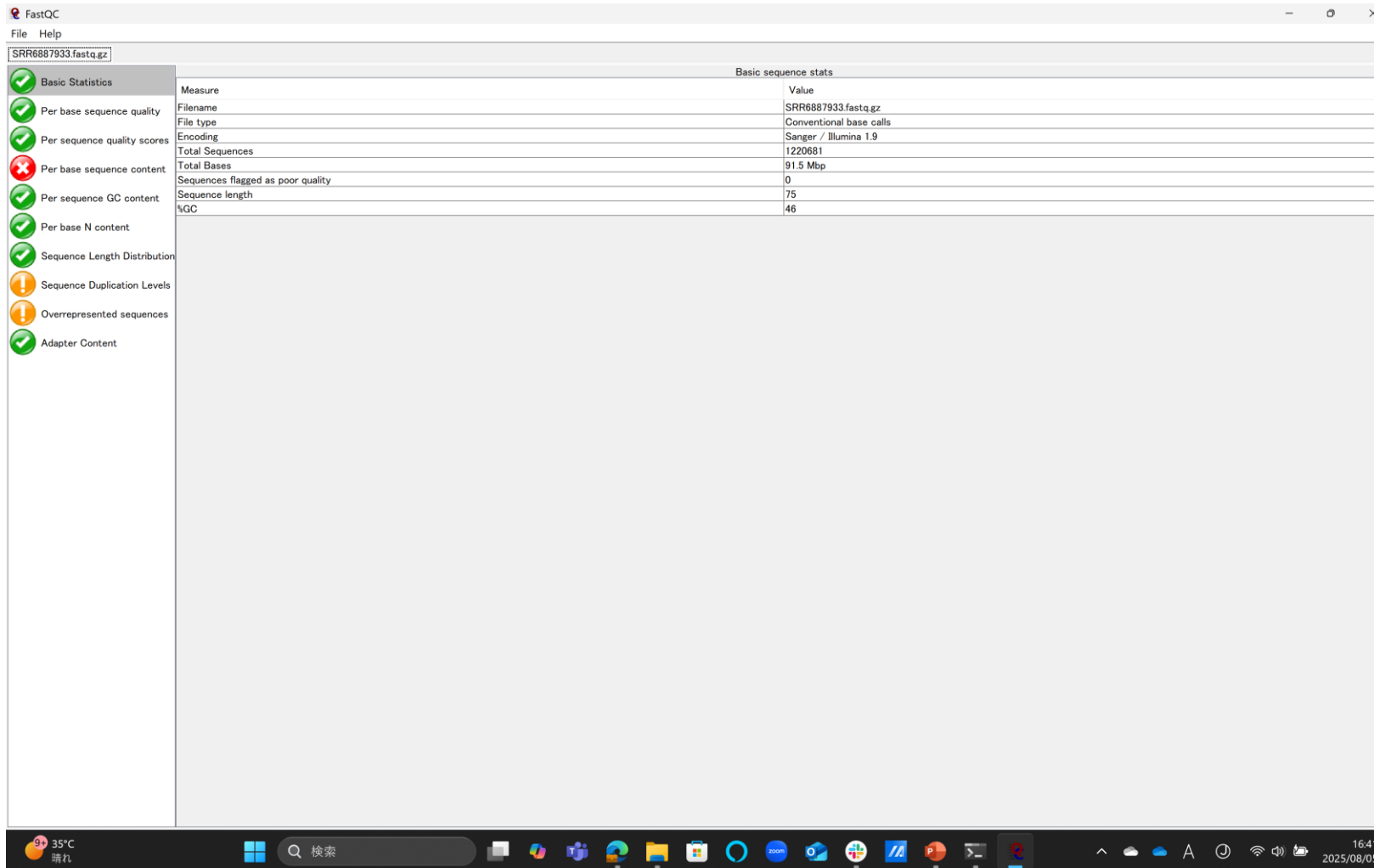
Below the table is a "Download Now" button. At the bottom of the screenshot, there is a preview of the FastQC software interface. The interface shows a window titled "FastQC" with a menu bar (File, Help) and two tabs: "bad_sequence.txt" and "good_sequence_short.txt". The main display area shows a box plot titled "Quality scores across all bases (Illumina >v1.3 encoding)". The y-axis represents quality scores from 24 to 34. The plot shows a distribution of quality scores for each base, with a blue line indicating the overall trend. A legend on the left side of the interface shows the following items:

- Basic Statistics (Green checkmark)
- Per base sequence quality (Red X)
- Per sequence quality scores (Green checkmark)
- Per base sequence content (Green checkmark)
- Per base GC content (Yellow warning icon)



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。



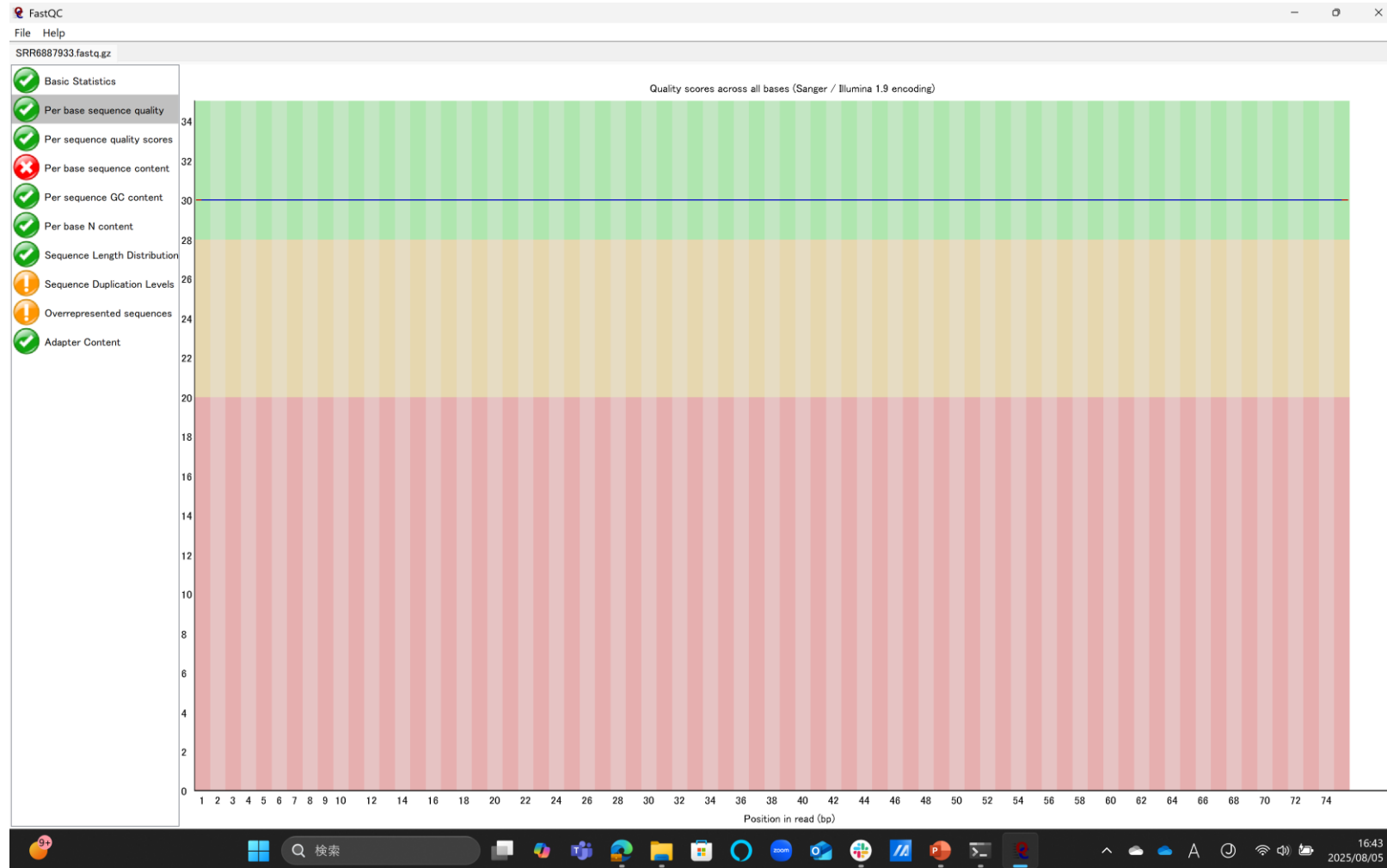
The screenshot shows the FastQC application window. The title bar reads 'FastQC'. The menu bar contains 'File' and 'Help'. The main window displays the file 'SRR6887933.fastq.gz'. On the left, there is a sidebar with various analysis modules, each with a status icon: 'Basic Statistics' (green checkmark), 'Per base sequence quality' (green checkmark), 'Per sequence quality scores' (green checkmark), 'Per base sequence content' (red X), 'Per sequence GC content' (green checkmark), 'Per base N content' (green checkmark), 'Sequence Length Distribution' (green checkmark), 'Sequence Duplication Levels' (orange exclamation mark), 'Overrepresented sequences' (orange exclamation mark), and 'Adapter Content' (green checkmark). The main area shows a table titled 'Basic sequence stats' with the following data:

Measure	Value
Filename	SRR6887933.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1220681
Total Bases	91.5 Mbp
Sequences flagged as poor quality	0
Sequence length	75
%GC	46



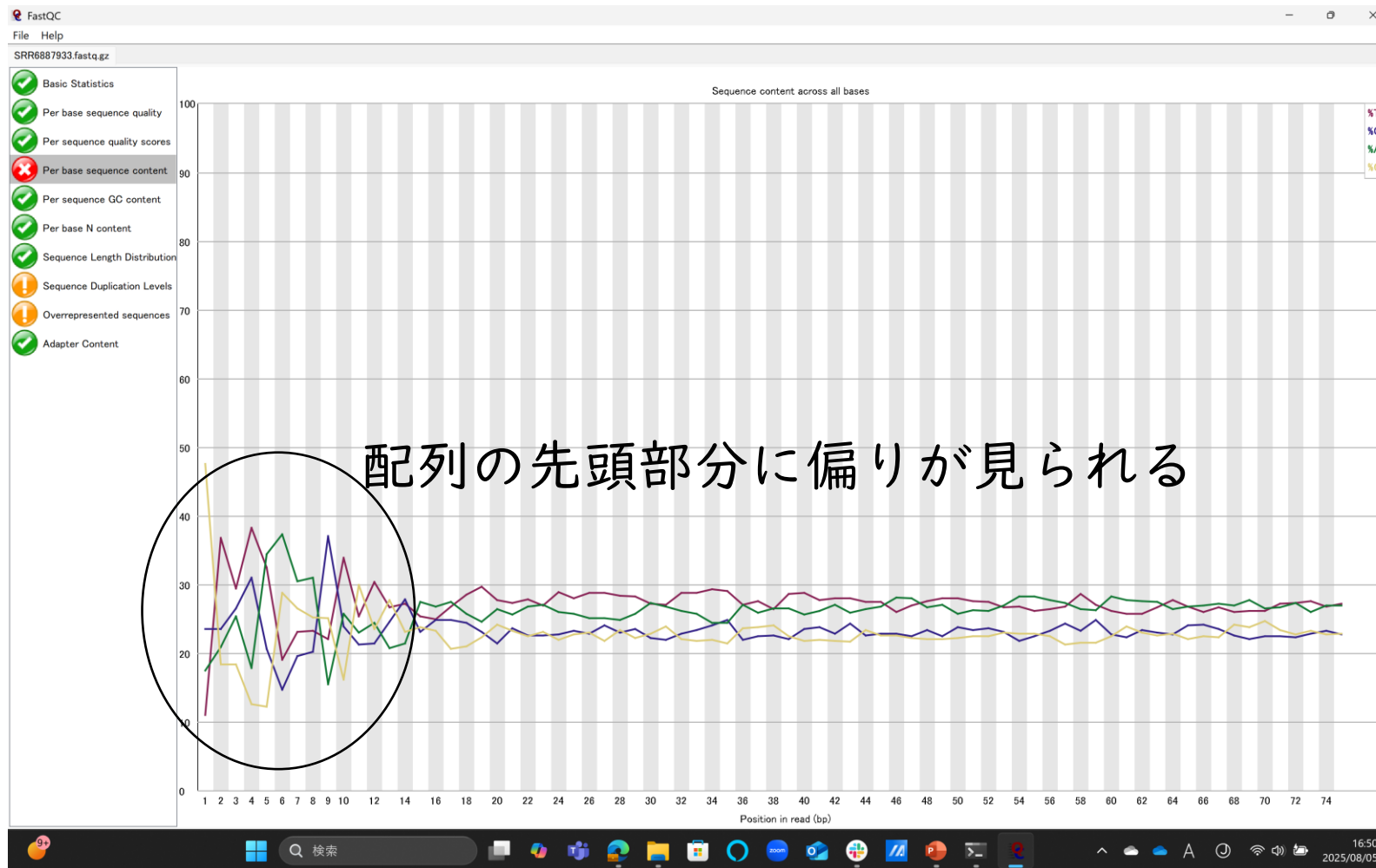
Per Base Sequence Quality

シーケンスクオリティを示す。
28以上が合格。



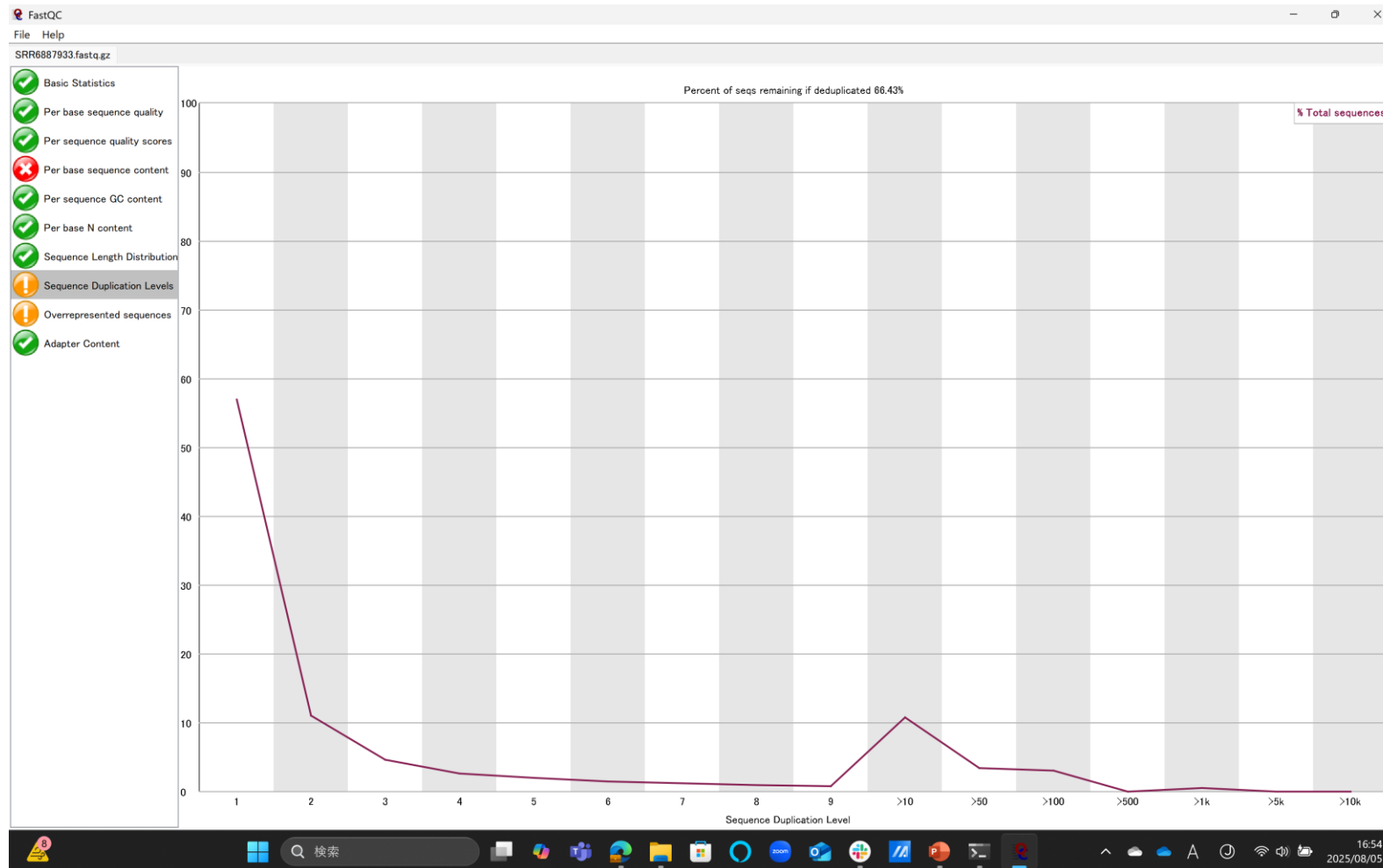
Per Base Sequence Content

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

$BWT(X) = \text{ard\$rcaaaaabb}$

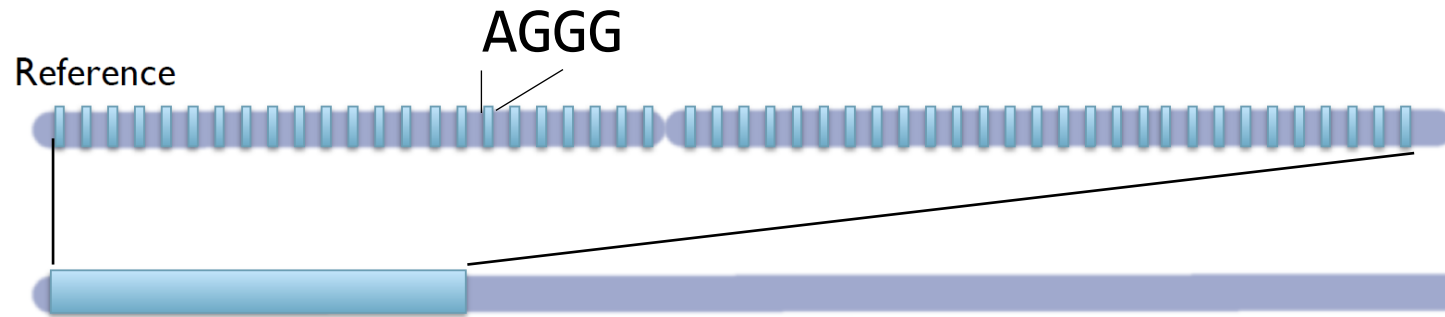


辞書式にソートしているので
同じアルファベットが集まる

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



BWTの利点



BWT(Reference)

GGGA

GGGC

GGGT

...

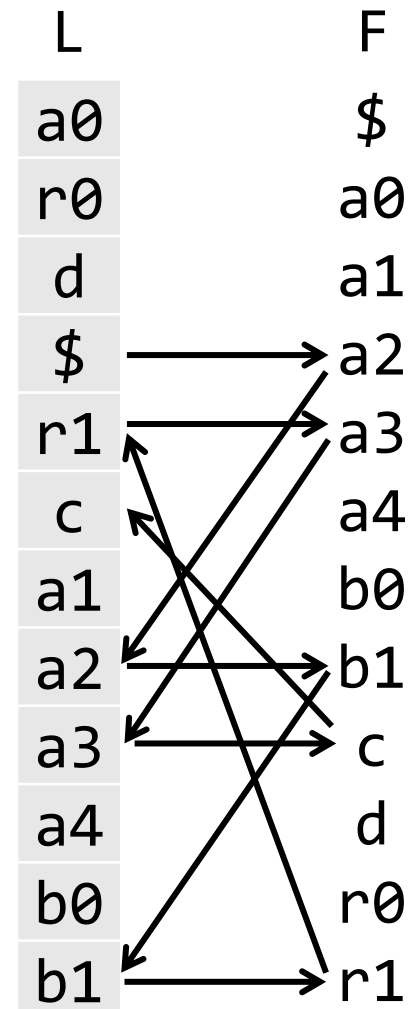
- 同じ文字が固まる傾向がある
- 検索しやすい



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とFを並べる
(複数個ある文字には番号を付けてある)



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...

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

The screenshot shows a web browser displaying the hisat2 website. The browser's address bar shows the URL <https://ccb.jhu.edu/software/hisat2/index.shtml>. The website header includes the hisat2 logo and the text "graph-based alignment of next generation sequencing reads to a population of genomes". Below the header, there is a paragraph describing hisat2 as a fast and sensitive alignment program. The main content area is organized into sections for different releases, each with a list of changes. On the right side, there is a sidebar with navigation links (Site Map, Home, Manual, FAQ), a "News and Updates" section, a "Getting Help" section, and a "Releases" section with a table of download links for various operating systems and genome versions. At the bottom of the browser window, the Windows taskbar is visible, showing the time as 17:00 on 2019/08/09.

HISAT2
graph-based alignment of next generation sequencing reads to a population of genomes

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome). Based on an extension of BWT for graphs [Sirián et al. 2014], we designed and implemented a graph FM index (GFM), an original approach and its first implementation to the best of our knowledge. In addition to using one global GFM index that represents a population of human genomes, HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover the human population). These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM Index (HGFM).

HISAT2 2.1.0 release 6/8/2017

- This major version includes the first release of HISAT-genotype, which currently performs HLA typing, DNA fingerprinting analysis, and CYP typing on whole genome sequencing (WGS) reads. We plan to extend the system so that it can analyze not just a few genes, but a whole human genome. Please refer to the HISAT-genotype website for more details.
- HISAT2 can be directly compiled and executed on Windows system using Visual Studio, thanks to Hilgel Dyer.
- Implemented --new-summary option to output a new style of alignment summary, which is easier to parse for programming purposes.
- Implemented --summary-file option to output alignment summary to a file in addition to the terminal (e.g. stderr).
- Fixed discrepancy in HISAT2's alignment summary.
- Implemented --no-templaten-adjustment option to disable automatic template length adjustment for RNA-seq reads.

HISAT2 2.0.5 release 11/4/2016

Version 2.0.5 is a minor release with the following changes.

- Due to a policy change (HTTP to HTTPS) in using SRA data ("--sra-option"), users are strongly encouraged to use this version. As of 11/9/2016, NCBI will begin a permanent redirect to HTTPS, which means the previous versions of HISAT2 no longer works with "--sra-acc" option soon.
- Implemented -l and -X options for specifying minimum and maximum fragment lengths. The options are valid only when used with --no-spliced-alignment, which is used for the alignment of DNA-seq reads.
- Fixed some cases where reads with SNPs on their 5' ends were not properly aligned.
- Implemented --no-softclip option to disable soft-clipping.
- Implemented --max-seeds to specify the maximum number of seeds that HISAT2 will try to extend to full-length alignments (see the manual for details).

HISAT, StringTie and Ballgown protocol published at Nature Protocols 8/11/2016

HISAT2 2.0.4 Windows binary available here, thanks to Andre Osorio Falcao 5/24/2016

HISAT2 2.0.4 release 5/18/2016

Version 2.0.4 is a minor release with the following changes.

- Improved template length estimation (the 9th column of the SAM format) of RNA-seq reads by taking introns into account.
- Introduced two options, --remove-chrname and --add-chrname, to remove "chr" from reference names or add "chr" to reference names in the alignment output, respectively (the 3rd column of the SAM format).
- Changed the maximum of mapping quality (the 5th column of the SAM format) from 255 to 60. Note that 255 is an undefined value according to the SAM manual and some programs would not work with this value (255) properly.
- Fixed NH (number of hits) in the alignment output.
- HISAT2 allows indels of any length pertaining to minimum alignment score (previously, the maximum length of indels was 3 bp).
- Fixed several cases that alignment goes beyond reference sequences.
- Fixed reporting duplicate alignments.

HISAT2 2.0.3-beta release 3/28/2016

Version 2.0.3-beta is a minor release with the following changes.

- Fixed graph index building when using both SNPs and transcripts. As a result, genome_snp_tran indexes here on the HISAT2 website have been rebuilt.
- Included some missing files needed to follow the small test example (see the manual for details).

HISAT2 2.0.2-beta release 3/17/2016

Note (3/19/2016): this version is slightly updated to handle reporting splice sites with the correct chromosome names.

Version 2.0.2-beta is a major release with the following changes.

- Memory mapped IO (--mm option) works now.
- Building linear index can be now done using multi-threads.
- Changed the minimum score for alignment in keeping with read lengths, so it's now --score-min L,0.0,-0.2, meaning a minimum score of -20 for 100-bp reads and -30 for 150-bp reads.
- Fixed a bug that the same read was written into a file multiple times when --un-conc was used.
- Fixed another bug that caused reads to map beyond reference sequences.
- Introduced --haplotype option in the hisat2-build (index building), which is used with --snp option together to incorporate those SNP combinations present in the human population. This option also prevents graph construction from exploding due to exponential combinations of SNPs in small genomic regions.
- Provided a new python script to extract SNPs and haplotypes from VCF files, hisat2_extract_snps_haplotypes_VCF.py
- Changed several python script names as follows
 1. extract_splice_sites.py to hisat2_extract_splice_sites.py
 2. extract_exons.py to hisat2_extract_exons.py
 3. extract_snps.py to hisat2_extract_snps_haplotypes_UCSC.py

HISAT2 2.0.1-beta release 11/19/2015

Version 2.0.1-beta is a maintenance release with the following changes.

- Fixed a bug that caused reads to map beyond reference sequences.

<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



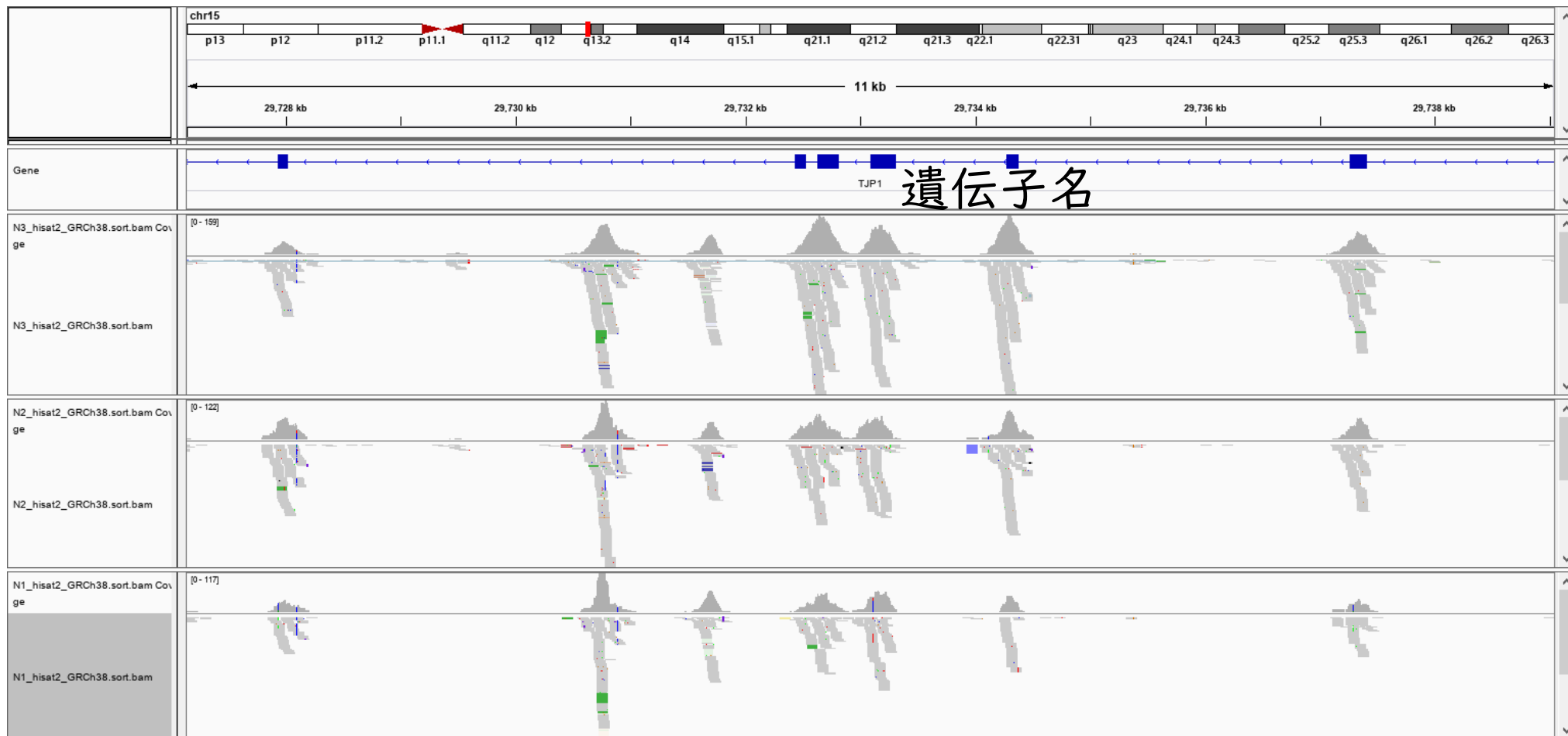
<http://www.broadinstitute.org/igv/>



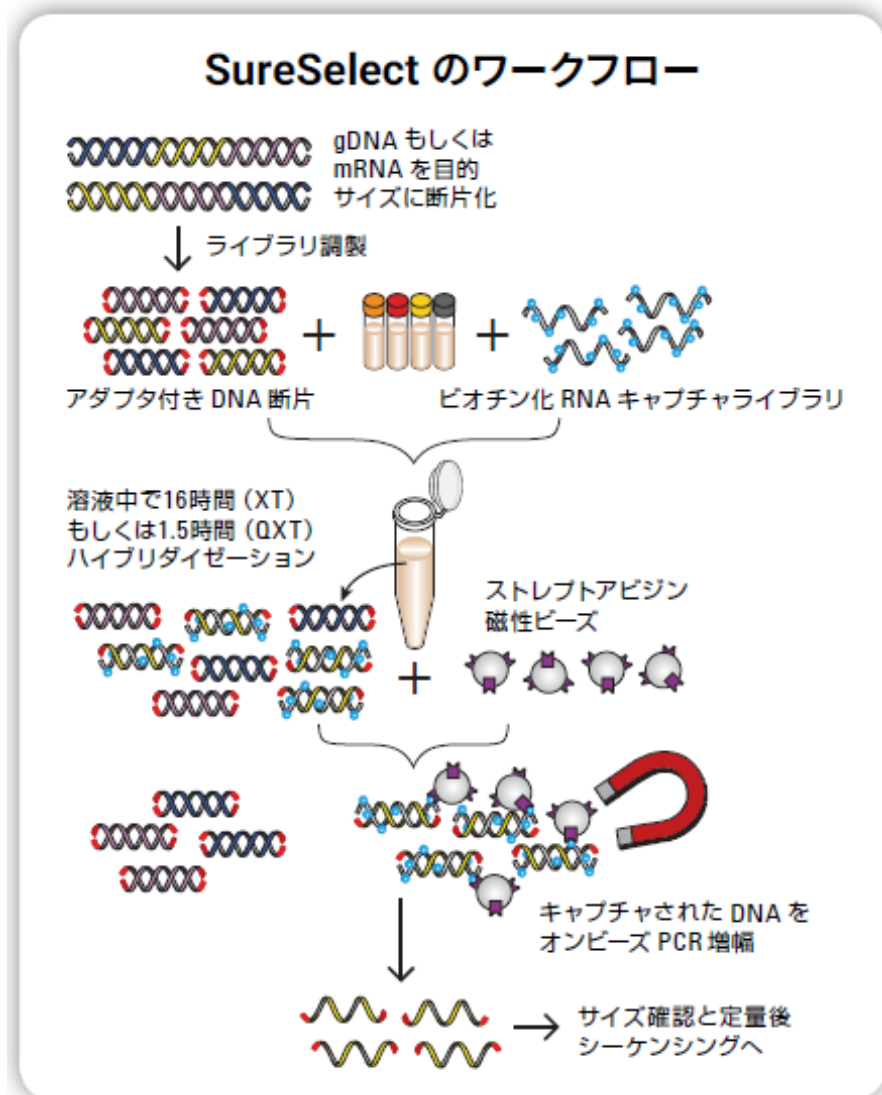
配列断片の可視化

染色体番号

参照ゲノム



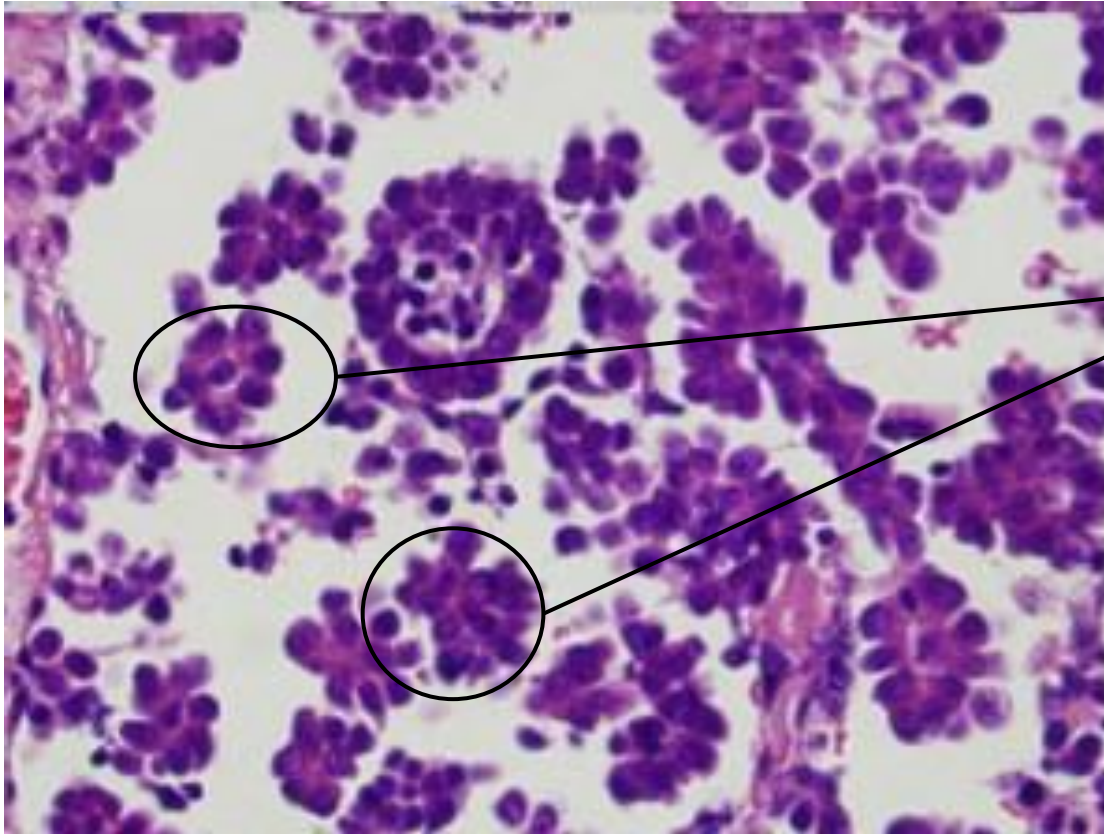
Whole Exome Sequencing (WES)



- 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)



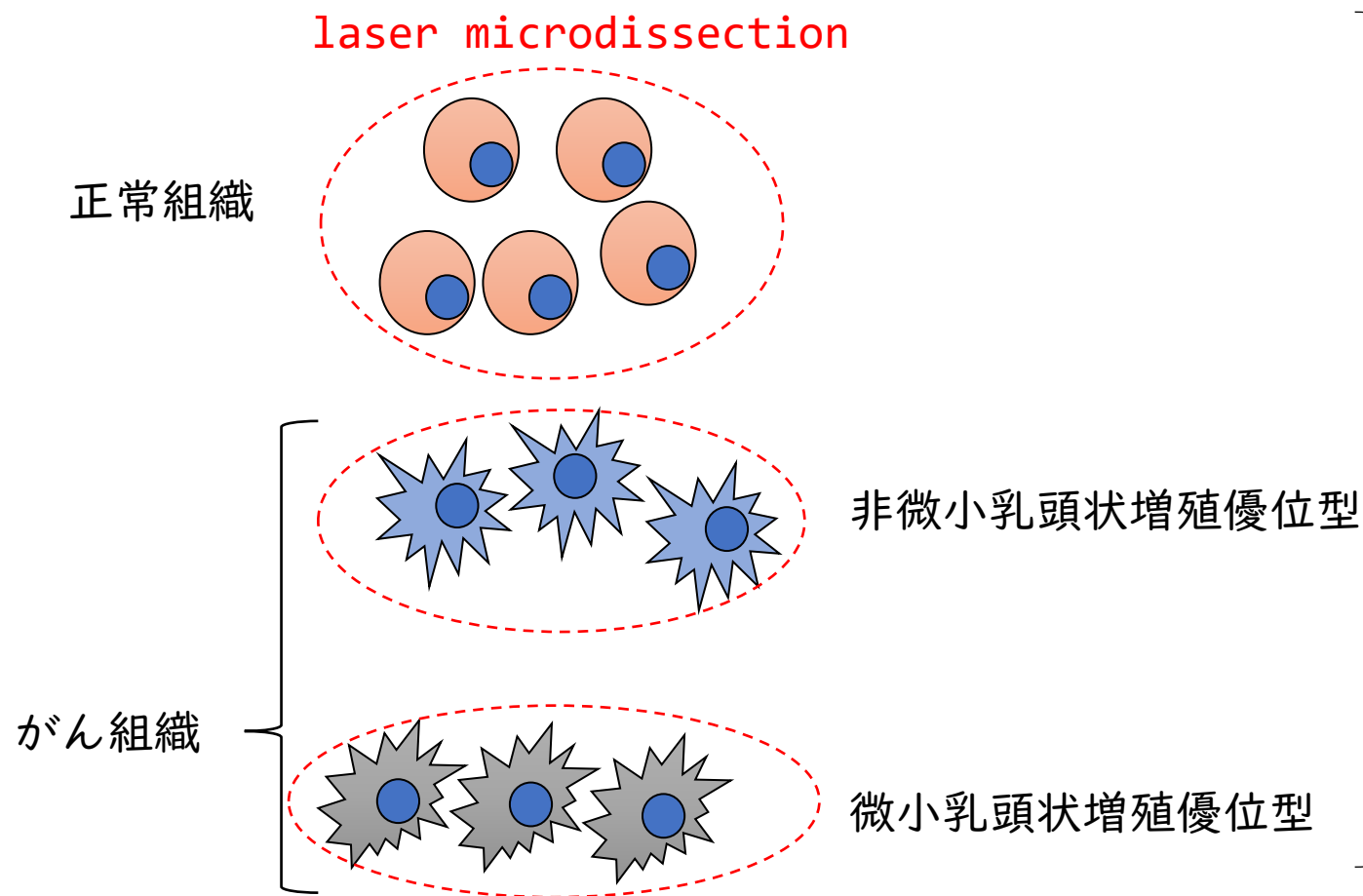
微小乳頭状増殖優位型

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



Whole Exome Sequencing (WES)

82歳 女性 非喫煙者



	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)

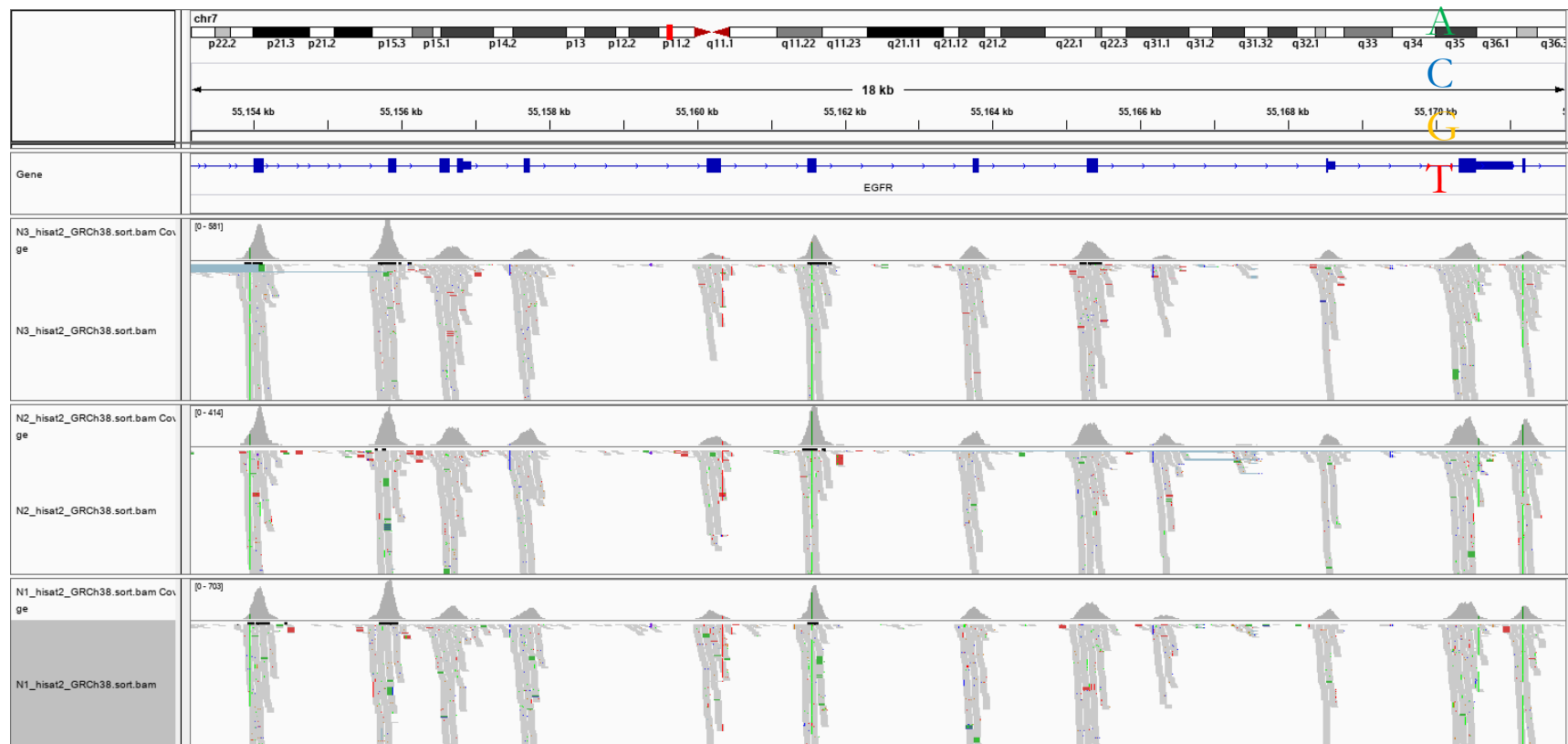
	Total Reads (100bp paired end)	Overall Mapping Rate	Coverage
正常	58,037,525	98.13%	193×
非微小乳頭状増殖優位型	62,830,751	97.69%	207×
微小乳頭状増殖優位型	64,318,300	97.47%	210×

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



Whole Exome Sequencing (WES)

EGFR



正常

非微小乳頭状増殖優位型

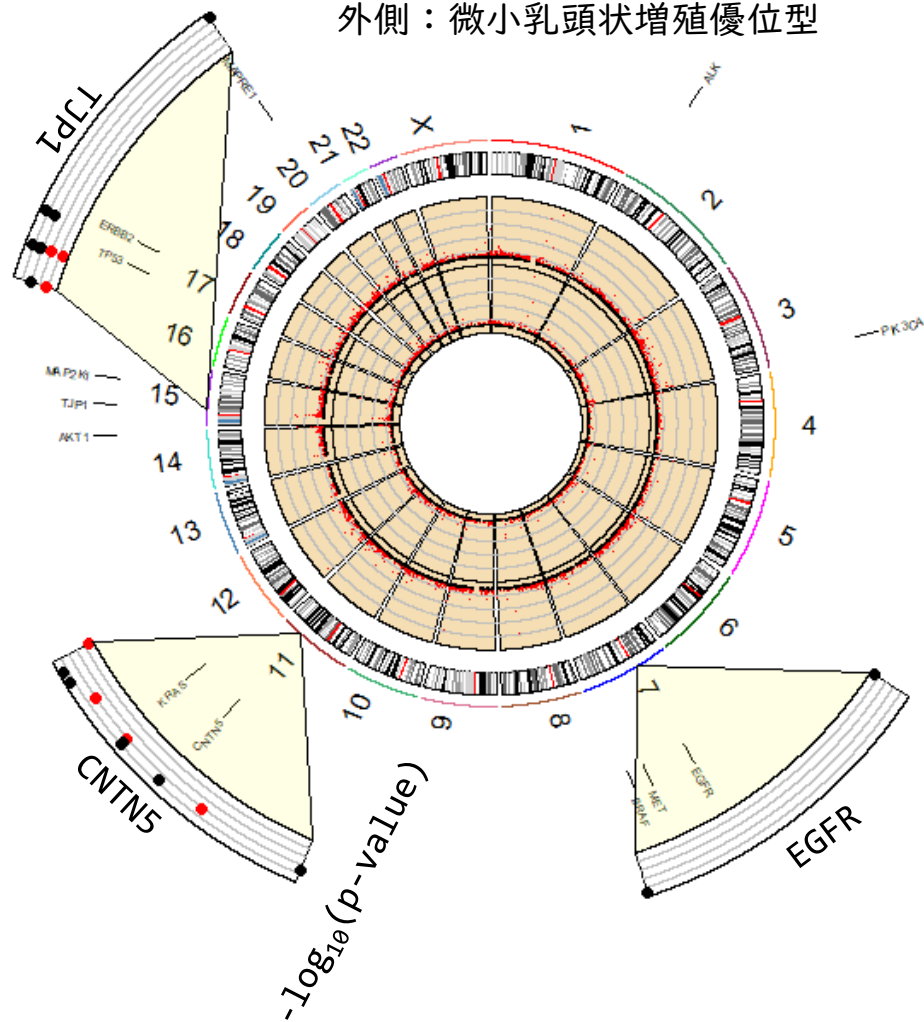
微小乳頭状増殖優位型



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

