

次世代シークエンサ実習II

ITの子カラで研究を支援



アメリエフ株式会社

本講義にあたって

- **代表的な解析の流れを紹介します**
 - 論文でよく使用されているツールを使用します
- **コマンドを沢山実行します**
 - スペルミスが心配な方は、コマンド例がありますのでコピーして実行してください
 - `/home/admin1409/amelieff/ngs/Reseq_command.txt`
- **TRY!** マークのコマンドは実行してください。
 - 実行が遅れてもあせらずに、応用や課題の間に追いついてください

本講義の内容

• Reseq解析

公開データ取得



クオリティコントロール



マッピング



変異検出

SNVとIndel検出
を行います。

• RNA-seq解析

公開データ取得



クオリティコントロール



マッピング



発現定量

FPKMを算出します。

Reseq解析：検出可能な変異

- ・ ショートリードのシーケンスでも様々な変異を検出可能

SNV

Duplication

InDel

Translocation

Inversion

CNV

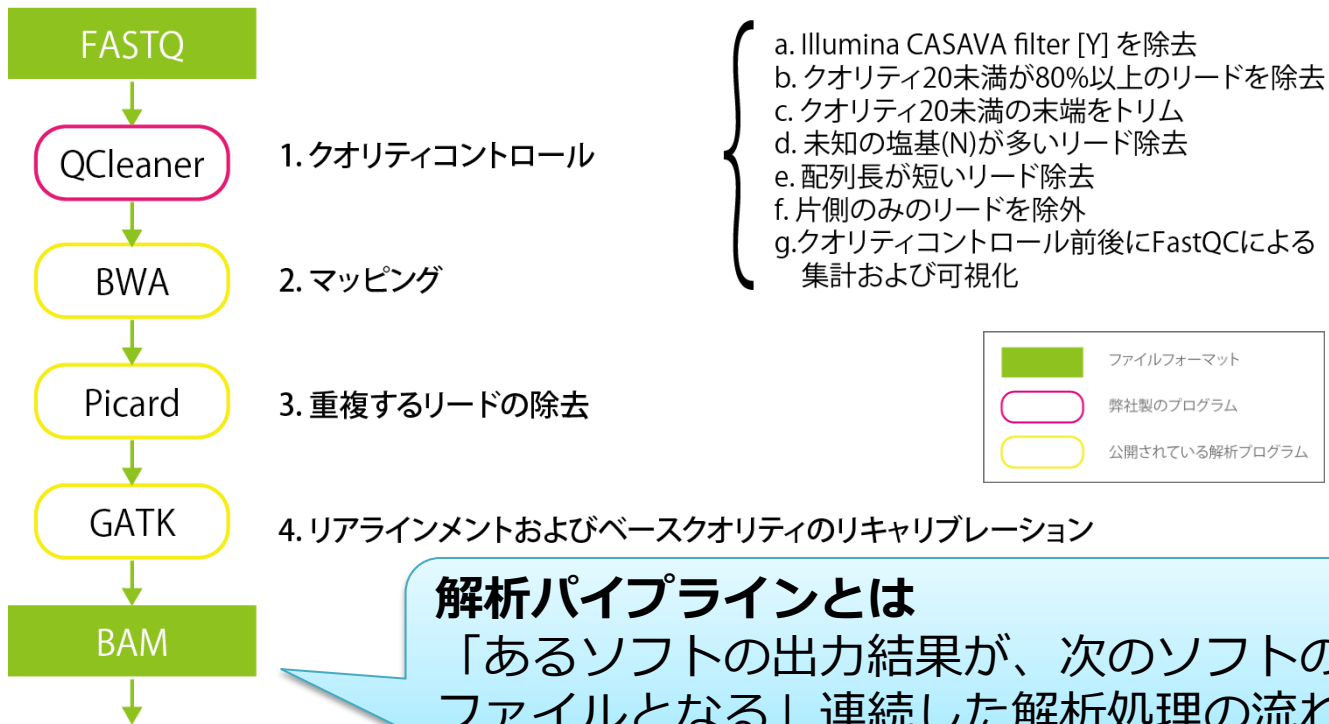
- ・ 検出アルゴリズムとソフトウェア

Paired-end mapping : BreakDancer、VariationHunter
Split-read mapping : Pindel
Others、Complex : CREST、DELLY

十分に精度が高いとは言えません。

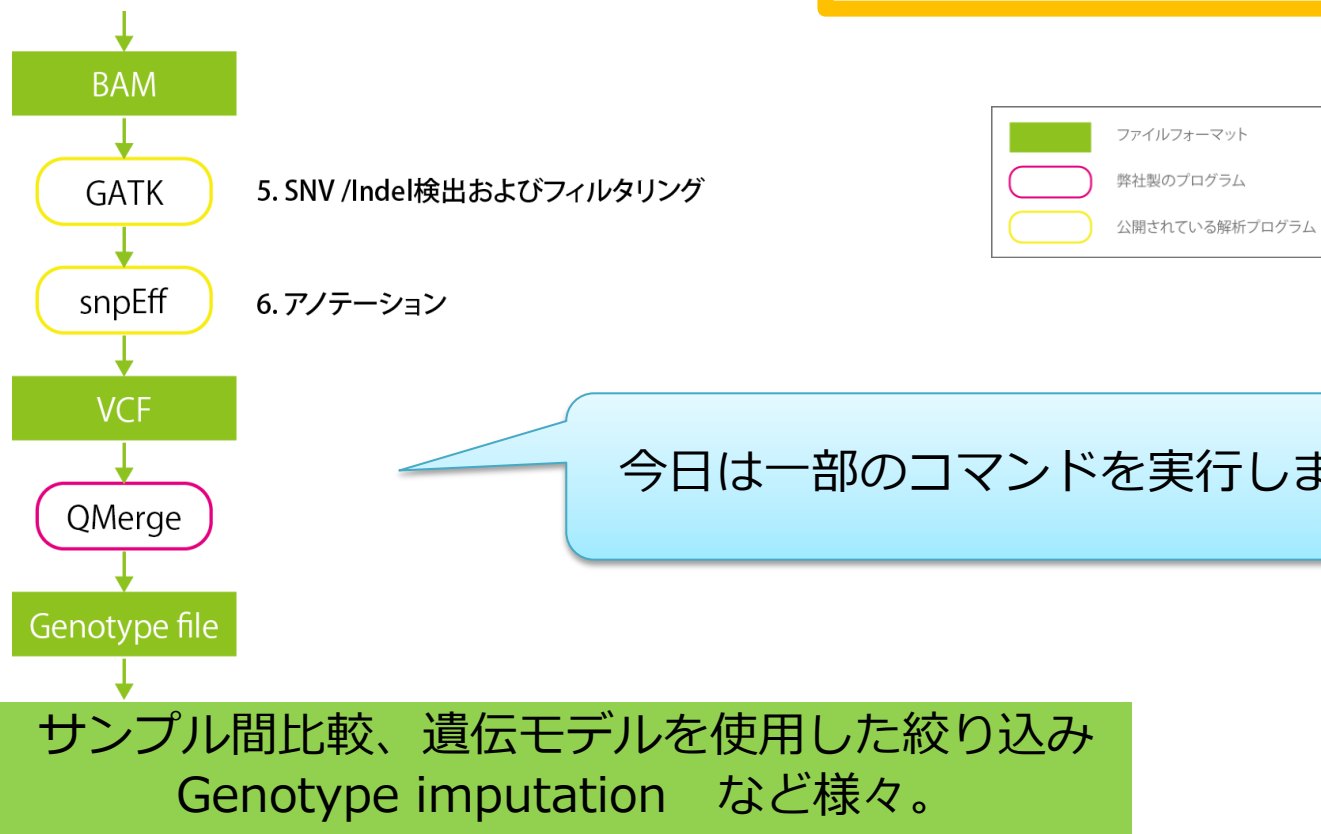
Reseq解析：パイプライン

データ取得 → **クオリティコントロール** → **マッピング** → 変異検出



Reseq解析：パイプライン

データ取得 → クオリティコントロール → **マッピング → 変異検出**



Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンス取得

illumina iGenomes



Saccharomyces cerevisiae

NCBI

build2.1

70 MB May 15 22:36

build3.1


70 MB May 15 22:36

Reseq解析：データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンス取得

Illumina iGenomes



Saccharomyces cerevisiae	NCBI	build2.1	70 MB	May 15 22:36
		build3.1	70 MB	May 15 22:36

リファレンスのfastaのみではなく、マッピングソフトのインデックスファイルや遺伝子情報ファイルも一緒に圧縮されて公開しています。

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンス取得 (実行済み)

ダウンロードして、解凍します。

```
$ wget ftp://igenome:G3nom3s4u@usd-  
ftp.illumina.com/Saccharomyces_cerevisiae/NCBI  
/build3.1/Saccharomyces_cerevisiae_NCBI_build3  
.1.tar.gz  
$ tar zxvf  
Saccharomyces_cerevisiae_NCBI_build3.1.tar.gz
```

※ファイル容量が大きいため、使用しないデータを一部削除しています

TRY!

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンスを確認

```
$ cd  
/home/admin1409/genome/Saccharomyces_cerevisiae/NCBI/build3.1/Sequence  
$ ll
```

```
drwxrwxr-x 2 admin1409 admin1409 4096 Jun  4 01:53 AbundantSequences  
drwxrwxr-x 2 admin1409 admin1409 4096 Apr 11 2012 Bowtie2Index  
drwxrwxr-x 4 admin1409 admin1409 4096 Mar 16 2012 BWAIndex  
drwxrwxr-x 2 admin1409 admin1409 4096 Mar 17 2012 Chromosomes  
drwxrwxr-x 2 admin1409 admin1409 4096 May  9 2013 WholeGenomeFasta
```

TRY!

Reseq 解析 : データ

データ取得

→ クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンスを確認

```
$ ll WholeGenomeFasta
```

```
-rwxrwxr-x 1 admin1409 admin1409      2310 Mar 16 2012 genome.dict
-rwxrwxr-x 1 admin1409 admin1409 12330859 Mar 16 2012 genome.fa
-rwxrwxr-x 1 admin1409 admin1409      412 Mar 16 2012 genome.fa.fai
-rwxrwxr-x 1 admin1409 admin1409      2318 May 9 2013 GenomeSize.xml
```

Reseq 解析 : データ

データ取得

→ クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンスを確認

```
$ less WholeGenomeFasta/genome.fa
```

ヘッダには、コンティグ名が記載されます。

```
>I  
CCACACCACACCCACACACCCACACACCACACACCACACACCACACCCACACACACACATCCTAAC  
CTACCCTAACACAGCCCTAATCTAACCCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTC  
CACTCGTTACCCTGTCCCATTCAACCATAACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTC  
ACCCACCGTTACCCTCCAATTACCCATATCCAACCCACTGCCACTTACCCTACCATTACCCTACCATCCA
```

「q」で閲覧を終了します。

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンスを確認

```
$ less WholeGenomeFasta/genome.fa.fai
```

I	230218	3	70	71
II	813184	233514	70	71
III	316620	1058320	70	71

インデックスファイルを開きます。
SamToolsで作成できます。

- 1列目 : コンティグ名 (fastaファイルのヘッダ)
- 2列目 : **コンティグの長さ**
- 3列目 : ファイルの先頭から見た、染色体の第一塩基目の位置
- 4列名 : fastaの1行の文字数
- 5列目 : 各行のバイト数

応用) ヒトリファレンスの話

GRCh Build37 + デコイ配列 Version 5

ヒトWhole Genome Sequencing Cloneを「ヒトゲノム+ヒトヘルペスウイルスHHV-4」にマッピングして、よくマップできなかったものを集めたもの。

サイズ： 合計35.4Mb、N50=22.9kb

特徴： 50%はサテライト配列またはシンプルリピート、
20%はレトロトランスポゾン

※現在は、2013/12/24にメジャーアップしたGRCh38が公開されています。

応用) ヒトリファレンスの話

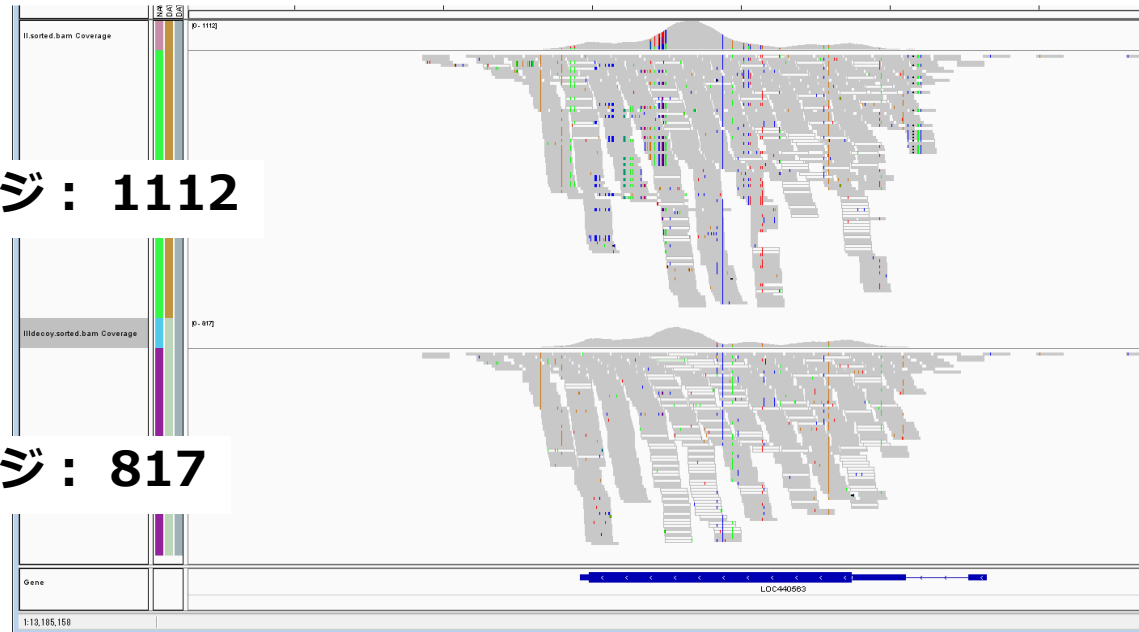
GRCh Build37 + デコイ配列 Version 5

Without Decoy

最大カバレッジ: 1112

With Decoy

最大カバレッジ: 817



Reseq解析は、リファレンスに対して変異検出するので、リファレンス自体がどの程度確かなのかが非常に大切

TRY!

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- 酵母のゲノムのリファレンスを確認

```
$ ll BWAIndex/version0.6.0
```

BWA バージョン0.6の
インデックスファイルを開きます。

リンクの「|」 シンボリックリンク名 -> 実体のファイル

```
lrwxrwxrwx  
-rwxrwxr-x  
-rwxrwxr-x  
-rwxrwxr-x  
-rwxrwxr-x  
-rwxrwxr-x
```

...

```
genome.fa -> ../../WholeGenomeFasta/genome.fa  
genome.fa.amb  
genome.fa.ann  
genome.fa.bwt  
genome.fa.pac  
genome.fa.sa
```


Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- リファレンスのインデックスを作成

BWA バージョン0.7のインデックスファイルを作成します。
BWAの使い方を確認します。

※既にbwa0.7にPATHが通っている場合はbwaだけで実行できます。

```
$ /home/admin1409/amelieff/ngs/bwa-0.7.10/bwa index
```

```
Usage: bwa index [-a bwtsv|is] [-c] <in.fasta>
```

```
$ mkdir BWAIndex/version0.7.10  
$ cd BWAIndex/version0.7.10
```

TRY!

Reseq 解析 : データ

データ取得

→ クオリティコントロール → マッピング → 変異検出

- リファレンスのインデックスを作成

シンボリックリンクを作成します。

```
$ ln -s 実体のファイル
```

```
$ ln -s ../../WholeGenomeFasta/genome.fa  
$ ll
```

```
lrwxrwxrwx ... genome.fa -> ../../WholeGenomeFasta/genome.fa
```

Reseq 解析 : データ

データ取得

→ クオリティコントロール → マッピング → 変異検出

- リファレンスのインデックスを作成

インデックスを作成します。

```
$ /home/admin1409/amelieff/ngs/bwa-0.7.10/bwa index genome.fa  
$ ll
```

```
lrwxrwxrwx genome.fa -> ../../WholeGenomeFasta/genome.fa  
-rw-rw-r-- genome.fa.amb  
-rw-rw-r-- genome.fa.ann  
-rw-rw-r-- ... genome.fa.bwt  
-rw-rw-r-- genome.fa.pac  
-rw-rw-r-- genome.fa.sa
```

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータ取得 <http://trace.ddbj.nig.ac.jp/dra/index.html>

DDBJ
DNA Data Bank of Japan

Login & Submit | Databases▼ | Japanese | Contact

Google™ Custom Search

Home | Handbook | FAQ | **Search** | Download▼ | Pipeline | About DRA

News

2014-05-13: [New DRA submission system is released.](#) less...

We have released the new DRA submission system. For major changes, please see the [slides](#) and [new handbook](#).

(6th, June, 2014)

For submissions with status "new" which had been created before 12th, May, 2014, addition or deletion of metadata objects could cause errors. It is recommended that download metadata as a tab-delimited text file and upload it into a newly created submission.


DDBJのSequence Read Archive → Search

DDBJ Seq... ng machines including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, and others. DRA is a member of the International Nucleotide Sequence Database Collaboration (INSDC) and archiving the data in a close collaboration with NCBI Sequence Read Archive (SRA) and EBI Sequence Read Archive (ERA). Please submit the trace data from conventional capillary sequencers to DDBJ Trace Archive.

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータ取得

 **DRASearch**

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword :

Show records Sort by

Accessionに「ERR038793」と入力 → Search

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータ取得

DRASearch Send Feedback

ERR038793 [FASTQ](#) [SRA](#) ここからダウンロード

Run Detail	
Alias	SC_RUN_6178_5#1
Instrument model	
Date of run	
Run center	
Number of spots	739,873
Number of bases	147,974,600

Navigation	
Submission	ERA038218
Study	ERP000277
Experiment	ERX015989
Sample	ERS021158

READS (joined) quality show 10 rows << < 1 / 73988 Page > >>

```
>ERR038793.1
GGACAAGGTTACTTCCTAGATGCTATATGTCCTACGGCCCTTGCTAACACCATCCAGCATGCAATAAGGTGACATAGAT
ATACCCACACACCCACCCCTGTGGAGTTGGATATGGGTAATTGGAAGGTAAACGGTGGGTGAGTGGTAGTAAGTAGAGGGA
TGGATGGTGGTTCGGAGGGGTATGGTTGGATGGGACAGGG
>ERR038793.2
TGGTGGTAT
GCAGGATAG
ACCCACACCCACACCCACACCCACACCCACTAACCCCTAA
```

NavigationエリアのExperiment → 「ERX015989」をクリック

Reseq解析：データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータ取得

Experiment Detail	
Title	
Design Description	Illumina sequencing of library 2414804, ERP000547. This is part of an Illumina run tagged with the sequence ATCACGTT.
Organism	Saccharomyces cerevisiae
Library Description	
Name	2414804
Strategy	WGS
Source	GENOMIC
Selection	RANDOM
Layout	PAIRED

他にも、シーケンサのプラットフォームやリード長などの情報も記載されています。

Whole Genome Sequencing

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータ取得 (実行済み)

ダウンロードします。

```
$ wget  
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/ERA038/ERA038218/ERX015989/ERR038793_1.fastq.bz2
```

```
$ wget  
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/ERA038/ERA038218/ERX015989/ERR038793_2.fastq.bz2
```


Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータ取得 (実行済み)

解凍して、先頭1000リードを抽出します。

```
$ bunzip2 ERR038793_1.fastq.bz2  
$ bunzip2 ERR038793_2.fastq.bz2
```

```
$ head -4000 ERR038793_1.fastq > 1K_ERR038793_1.fastq  
$ head -4000 ERR038793_2.fastq > 1K_ERR038793_2.fastq
```

Reseq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 変異検出

- シーケンスデータを確認

```
$ cd /home/admin1409/amelieff/ngs  
$ ll
```

```
-rw-rw-r-- 1 admin1409 admin1409 315892 Jul 16 18:45 1K_ERR038793_1.fastq  
-rw-rw-r-- 1 admin1409 admin1409 315892 Jul 16 18:45 1K_ERR038793_2.fastq  
-rw-rw-r-- 1 admin1409 admin1409 346770 Dec  3  2013 1K_SRR518891_1.fastq
```

:

行数を数えます。1リードは4行で表記されます。

```
$ wc -l 1K_ERR038793_1.fastq
```

```
4000 1K_ERR038793_1.fastq
```

TRY!

Reseq 解析 : クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 変異検出

- シーケンスデータのクオリティを確認

インストールされているFastQCの、バージョンと使い方を確認します。

```
$ fastqc -version
```

```
FastQC v0.10.1
```

```
$ fastqc -help
```

Fastqのみではなく、bamとsamも入力可能

```
fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam] [-c contaminant file] seqfile1 .. seqfileN
```

複数のファイルも指定可能

TRY!

Reseq 解析 : クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 変異検出

- シーケンスデータのクオリティを確認

FastQCを実行します。

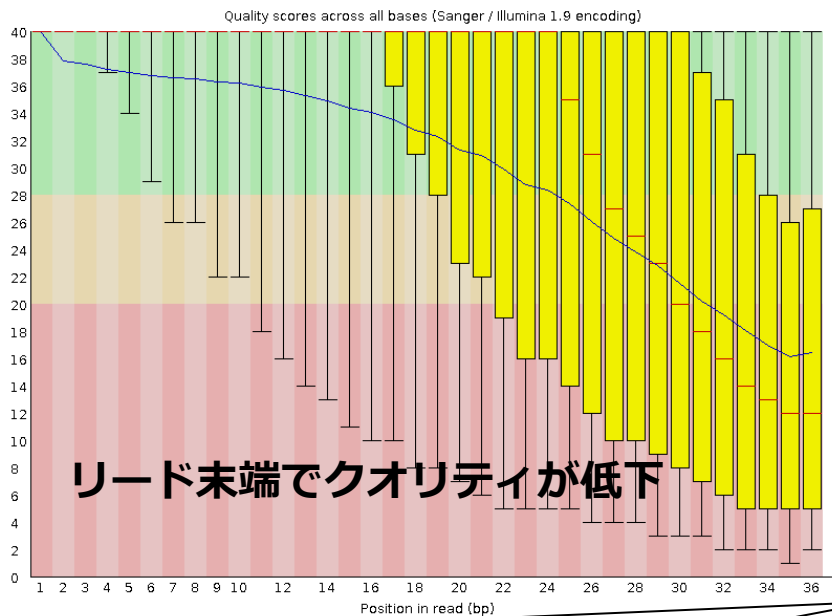
```
$ mkdir reseq
$ fastqc -o reseq -f fastq 1K_ERR038793_1.fastq
1K_ERR038793_2.fastq
```

fastqc_report.htmlを、ウェブブラウザで開きます。

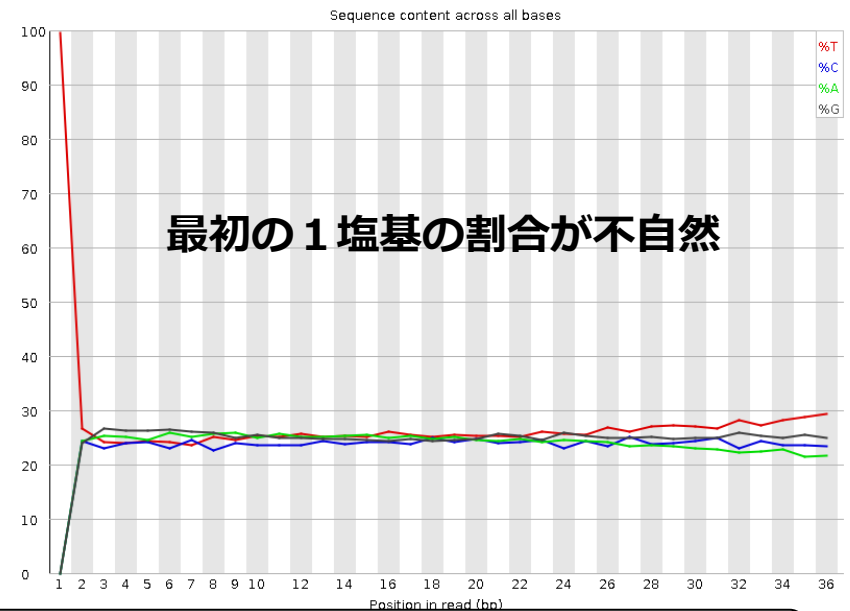
```
$ firefox reseq/1K_ERR038793_1_fastqc/fastqc_report.html
$ firefox reseq/1K_ERR038793_2_fastqc/fastqc_report.html
```

応用) とあるシーケンスデータの実例

✖ Per base sequence quality



✖ Per base sequence content



マッピング率が低下や、変異の偽陽性が増加するなどの問題を引き起こす。

シーケンス技術が向上しクオリティの高いデータを目にする機会が増えましたが、試料・シーケンス・トリミングなどに、問題がないか確認することをおすすめします。

TRY!

Reseq 解析 : クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 変異検出

- クオリティ30以上の塩基が90%未満のリードを削除

インストールされているfastq_quality_filterの使い方を確認します。

```
$ fastq_quality_filter -h
```

```
usage: fastq_quality_filter [-h] [-v] [-q N] [-p N] [-z] [-i INFILE] [-o OUTFILE]
Part of FASTX Toolkit 0.0.13.1 by A. Gordon (gordon@cshl.edu)
```

```
[-h]           = This helpful help screen.
[-q N]         = Minimum quality score to keep.
[-p N]         = Minimum percent of bases that must have [-q] quality.
[-z]           = Compress output with GZIP.
[-i INFILE]    = FASTA/Q input file. default is STDIN.
[-o OUTFILE]   = FASTA/Q output file. default is STDOUT.
[-v]           = Verbose - report number of sequences.
```

TRY!

Reseq 解析：クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 変異検出

- クオリティ30以上の塩基が90%未満のリードを削除

```
$ fastq_quality_filter -i 1K_ERR038793_1.fastq -o  
reseq/1K_ERR038793_1_qual.fastq -q 30 -p 90 -Q 33 -v
```

```
Quality cut-off: 30  
Minimum percentage: 90  
Input: 1000 reads.  
Output: 802 reads.  
discarded 198 (19%) low-quality reads.
```

ターミナルに直接解析のサマリー
を出力するソフトもあります。

以降の解析は、片側のリードのみ使用します。

応用) クオリティコントロールの順番も大切

FASTQ形式にマッチするかチェック

データクオリティチェック (FastQC)

Illumina CASAVA filter [Y] を除去

クオリティ20未満が80%以上のリードを除去

クオリティ20未満の末端をトリム

未知の塩基(N)が多いリード除去

配列長が短いリード除去

片側のみのリードを除外

データクオリティチェック (FastQC)

ロングリードの場合、リードの大半が除外されてしまう可能性。

ペアエンドリードの場合、ペアが揃っていないとマッピングソフトが停止する可能性。

TRY!

Reseq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 変異検出

- Bwa memコマンドの使い方を確認

```
$ bwa-0.7.10/bwa mem
```

```
Usage: bwa mem [options] <idxbase> <in1.fq> [in2.fq]
```

```
-R STR read group header line such as '@RG\tID:foo\tSM:bar' [null]
```

※RG (read groups)
platform (PL) および sample (SM)が必要
PLの例 : 454, LS454, Illumina, Solid, ABI_Solid

TRY!

Reseq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 変異検出

- マッピング

```
$ cd reseq
$ ../bwa-0.7.10/bwa mem -R
"@RG\tID:1K_ERR038793_1\tSM:ERR038793\tPL:Illumina"
/home/admin1409/genome/Saccharomyces_cerevisiae/NCBI
/build3.1/Sequence/BWAIndex/version0.7.10/genome.fa
1K_ERR038793_1_qual.fastq > 1K_ERR038793_1_qual.sam
$ ll
```

1K_ERR038793_1_qual.sam

TRY!

Reseq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 変異検出

- SAMをBAMに変換

```
$ samtools view -Sb 1K_ERR038793_1_qual.sam  
> 1K_ERR038793_1_qual.bam  
$ ll -h
```

```
56K Jul 24 20:46 1K_ERR038793_1_qual.bam  
248K Jul 24 19:12 1K_ERR038793_1_qual.fastq  
239K Jul 24 20:43 1K_ERR038793_1_qual.sam
```

1/4程度にファイルサイズが小さくなりました。

TRY!

Reseq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 変異検出

- ソートとインデキシング

```
$ samtools sort 1K_ERR038793_1_qual.bam  
1K_ERR038793_1_qual_sorted  
  
$ samtools index 1K_ERR038793_1_qual_sorted.bam  
  
$ ll
```

```
1K_ERR038793_1_qual_sorted.bam  
1K_ERR038793_1_qual_sorted.bam.bai
```

Reseq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 変異検出

- マッピングされたリード数

```
$ samtools idxstats 1K_ERR038793_1_qual_sorted.bam
```

I	230218	713	0
II	813184	8	0

コンティグ名、コンティグの長さ、マッピングされたリード、マッピングされなかったリードの順に表示されます。

3列目を足し合わせると、マッピングされたリード数がわかります。

応用) 列の合計を計算するコマンド

```
$ samtools idxstats 1K_ERR038793_1_qual_sorted.bam > tmp  
$ awk '{a += $3} END {print a}' tmp
```

1行読み込むたびに、3列目を「a」に足す。

803 マッピングされたリード

```
$ awk '{a += $4} END {print a}' tmp
```

0 マッピングされなかったリード

**802リードのfastqをマッピングしたはずが、1本増えています。
マルチヒットしたリードがあると考えられます。**

応用) マルチヒットしたリードを探索

```
$ samtools view 1K_ERR038793_1_qual_sorted.bam | awk  
'{print $1}' | sort | uniq -c
```

```
1 ERR038793.100  
1 ERR038793.1000  
1 ERR038793.101
```

登場した回数、配列ID
1回以上登場した配列を探します。

```
$ samtools view 1K_ERR038793_1_qual_sorted.bam | awk  
'{print $1}' | sort | uniq -c | awk '$1>1 {print}'
```

```
2 ERR038793.217
```

本当に2回マッピングされているか
確認します。

応用) マルチヒットしたリードを探索

```
$ samtools view 1K_ERR038793_1_qual_sorted.bam | grep  
ERR038793.217
```

ERR038793.217	2048	II	8052	0	70H30M	*	0	0
@GFFF7F@EE@BGE;F;		NM:i:0	MD:Z:30	AS:i:30	XS:i:30	RG:Z:1K_ERR038793_1		
ERR038793.217	16	XV	1082774	0	47S53M	*	0	0

異なるコンティグにマッピングされています。

TRY!

Reseq 解析 : 変異検出

データ取得 → クオリティコントロール → マッピング → 変異検出

- GATK UnifiedGenotyperコマンドの使い方を確認

```
$ java -jar /usr/local/src/GenomeAnalysisTK-1.6-13-g91f02df/GenomeAnalysisTK.jar -T UnifiedGenotyper -h
```

```
-R, --reference_sequence <reference_sequence>
```

```
-glm, --genotype_likelihoods_model <genotype_likelihoods_model>
```

SNP, INDEL, BOTH から選べます。デフォルトはSNP

TRY!

Reseq解析：変異検出

データ取得 → クオリティコントロール → マッピング → 変異検出

- SNV/Indel検出

```
$ java -jar /usr/local/src/GenomeAnalysisTK-1.6-13-g91f02df/GenomeAnalysisTK.jar -T UnifiedGenotyper -glm BOTH -R /home/admin1409/genome/Saccharomyces_cerevisiae/NCBI/build3.1/Sequence/BWAIndex/version0.7.10/genome.fa -I 1K_ERR038793_1_qual_sorted.bam -o 1K_ERR038793_1_qual_sorted.vcf $ 11
```

```
1K_ERR038793_1_qual_sorted.vcf  
1K_ERR038793_1_qual_sorted.vcf.idx
```

TRY!

Reseq解析：変異検出

データ取得 → クオリティコントロール → マッピング → 変異検出

- 検出したSNV/Indelの数を確認

```
$ less 1K_ERR038793_1_qual_sorted.vcf  
$ awk '!/^#/ ' 1K_ERR038793_1_qual_sorted.vcf | wc -l
```

```
100
```

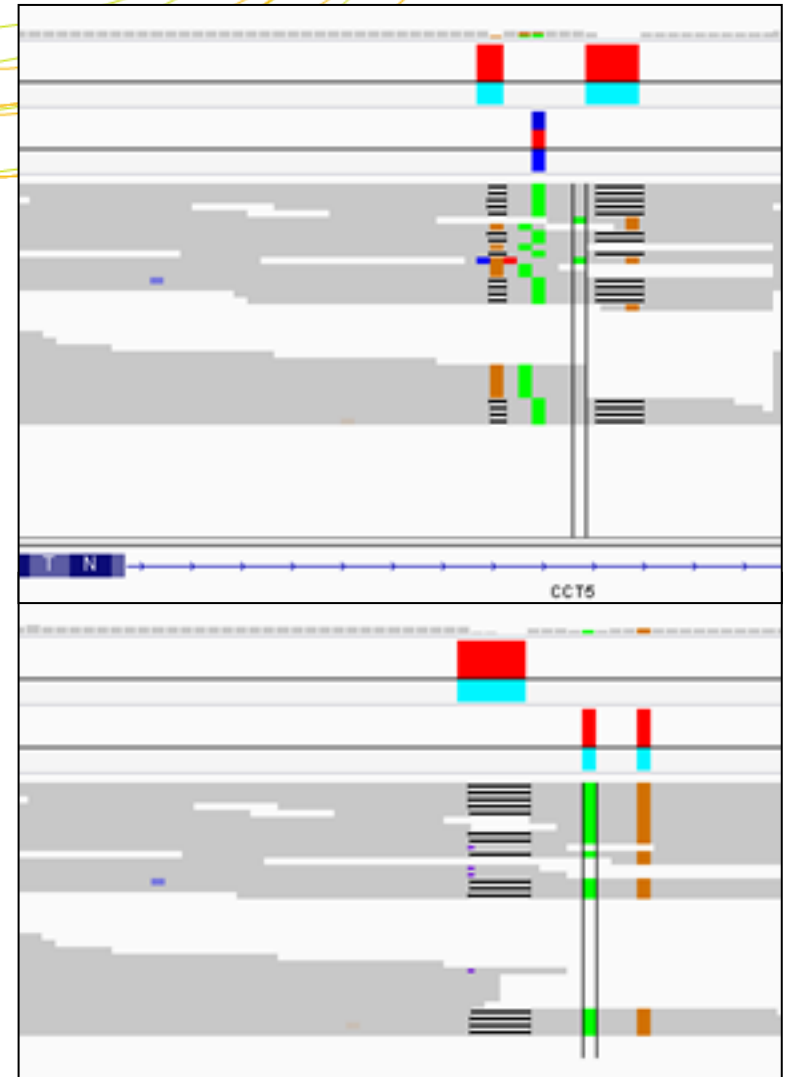
100個の変異が検出されました

応用) リアライメント

リアライメントは必要?

BWAでは、1本のリードに複数の変異が含まれる場合に、アライメントスコアの計算上、SNVやIndelの正確な位置を決めることが出来ません。

このような領域を対象領域として抜き出して、改めて丁寧にアライメントを行う。



TRY!

Reseq解析：変異検出

データ取得 → クオリティコントロール → マッピング → **変異検出**

- 検出したSNV/Indelを可視化

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
I	111	.	C	T	114.96	.

...
ERR038793
GT:AD:DP:GQ:PL 0/1:2,4:6:64.40:145,0,64

ジェノタイプがC/Tのヘテロ

カバレッジが6

```
$ igv.sh
```

TRY!

Genomes → Load Genome from File...

/home/admin1409/genome/
Saccharomyces_cerevisiae/
NCBI/build3.1/Sequence/
WholeGenomeFastaまで移動

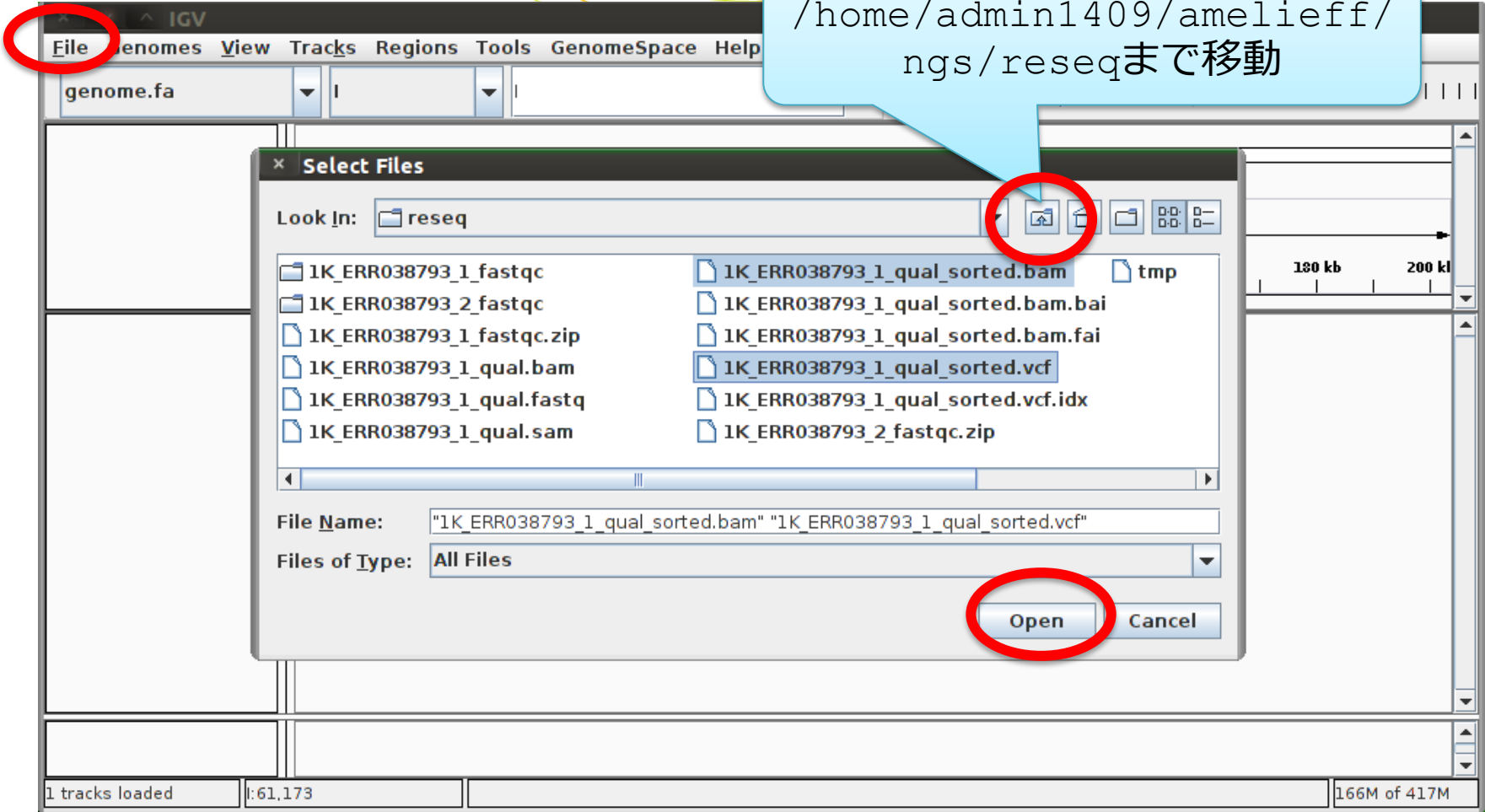
The screenshot shows the IGV (Integrative Genomics Viewer) interface. The 'Genomes' menu is circled in red. A 'Load Genome' dialog box is open, showing the 'Look In' path as 'WholeGenomeFasta'. The file list includes 'genome.dict', 'genome.fa', 'genome.fa.fai', and 'GenomeSize.xml'. The 'genome.fa' file is circled in red. The 'Open' button is also circled in red.

Genome.faファイルを選択

TRY!

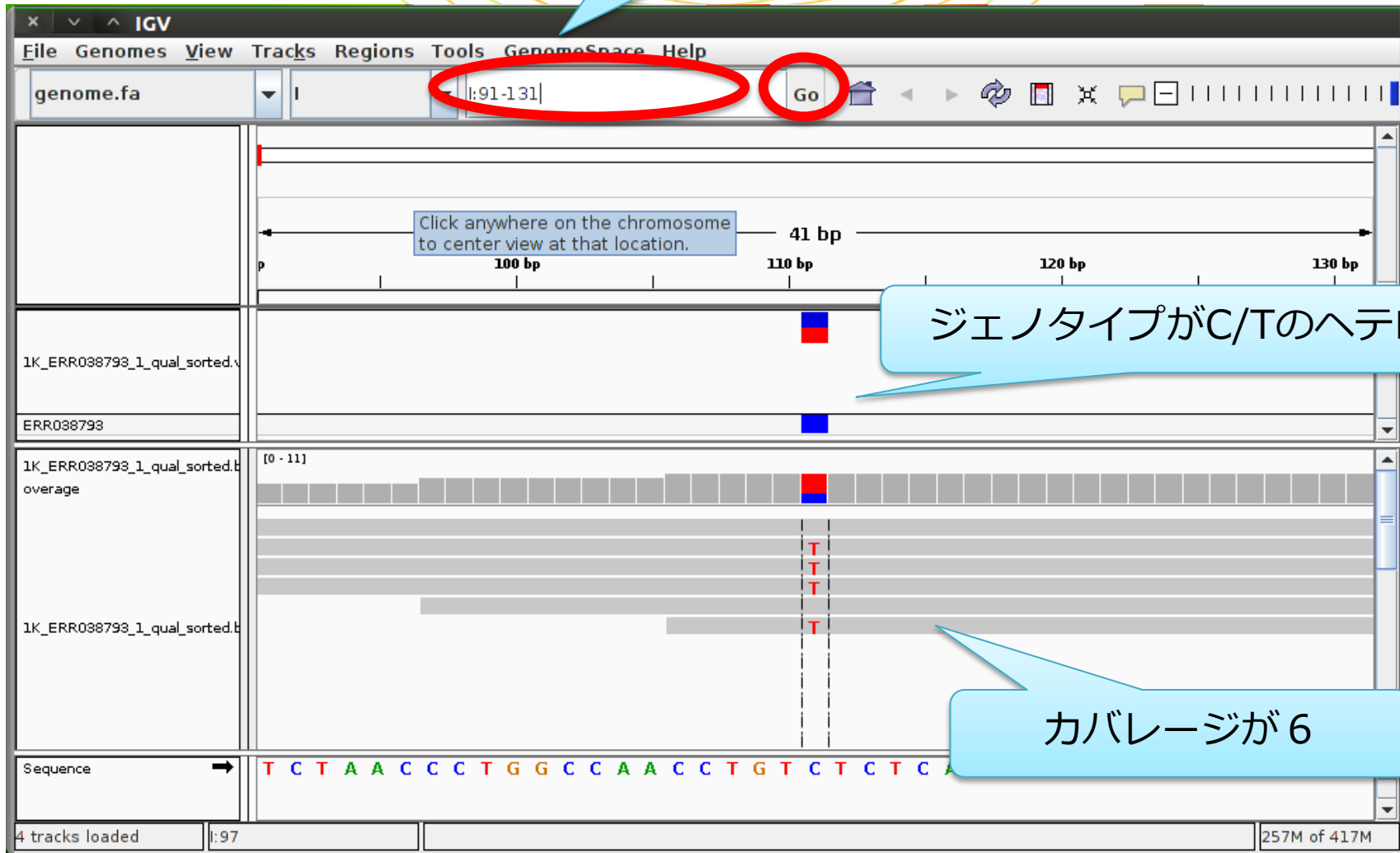
File → Load from File...

/home/admin1409/amelieff/
ngs/reseqまで移動



TRY!

「I:111」と入力



応用) Indelの見方

TTの欠失

TTTの欠失

I 5068 . ATTT AT,A 525.47 .

... GT:AD:DP:GQ:PL 1/2:1,0,7:16:99:604,205,160,221,0,179

ジェノタイプは、AT/A

ホモポリマーではシーケンスエラーによって、偽陽性のIndelが検出されやすい。

A T A T T A A A T A T A C A T T T T G C A T T T T T T T T T T T T T T T C T G T A T T

応用) 変異のフィルタリング

- GATKのVariantFiltrationコマンドでフィルタリングをします

```
$ java -jar /usr/local/src/GenomeAnalysisTK-1.6-13-g91f02df/GenomeAnalysisTK.jar -T VariantFiltration -R /home/admin1409/genome/Saccharomyces_cerevisiae/NCBI/build3.1/Sequence/BWAIndex/version0.7.10/genome.fa -V 1K_ERR038793_1_qual_sorted.vcf -o 1K_ERR038793_1_qual_sorted_fil.vcf --clusterWindowSize 10 --filterExpression "DP < 10" --filterName "LowCoverage"
```

VCFファイルのFILTER列に、条件を通過した場合“PASS”、そうでない場合は“filterName”が記入されます。

本講義の内容

• Reseq解析

公開データ取得



クオリティコントロール



マッピング



変異検出

• RNA-seq解析

公開データ取得



クオリティコントロール



マッピング



発現定量

FPKMを算出します。