Tin Sinh học
Bioinformatics

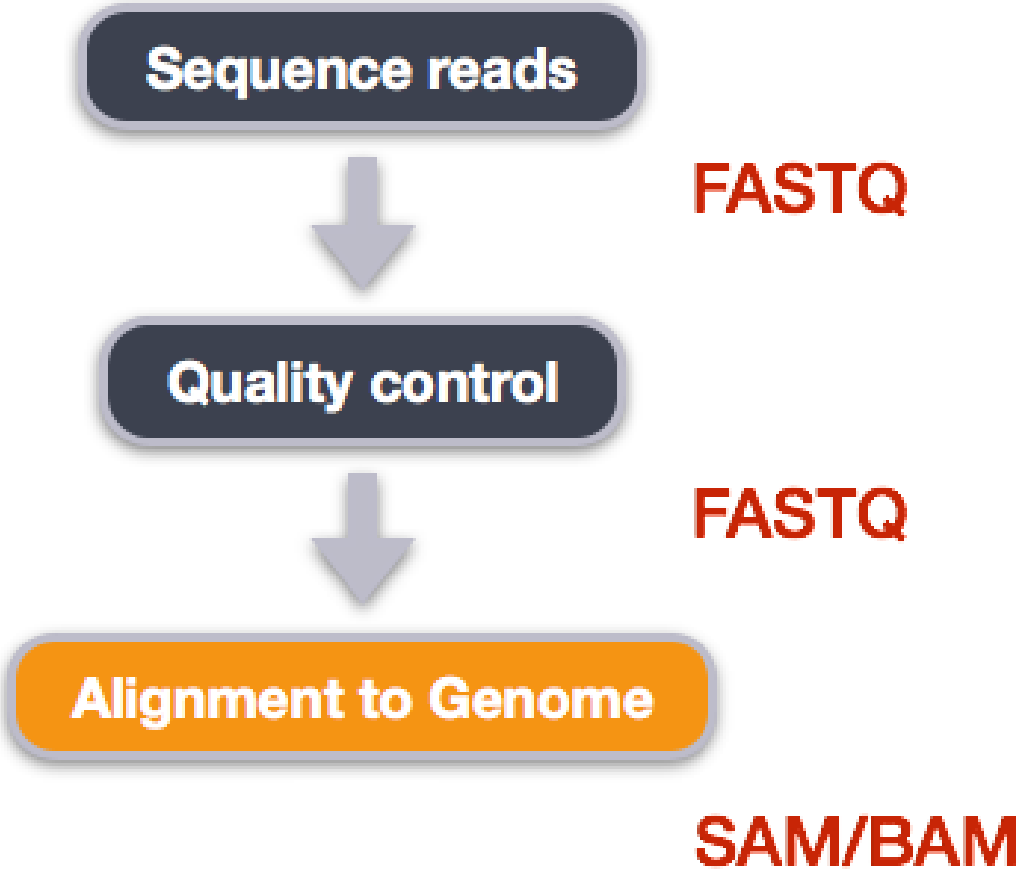
Data Wrangling and Processing for Genomics

https://datacarpentry.org/wrangling-genomics

1. Đặt vấn đề
2. Assessing Read Quality
3. Trimming and Filtering
4. Variant Calling Workflow

Mục tiêu

- Xác định các đột biến của các chủng E. coli ra đời sau so với chủng E. coli gốc REL606
- Bước 1: Alignment với hệ gen tham chiếu



Cài đặt các công cụ

- BWA: http://bio-bwa.sourceforge.net/

sudo apt update

sudo apt install bwa

- Samtools

sudo apt install samtools

- bcftools

sudo apt install bcftools

Download hệ gen tham chiếu

```
$ cd ~/dc_workshop
$ mkdir -p data/ref_genome
$ curl -L -o data/ref_genome/ecoli_rel606.fasta.gz
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/017/
985/GCA_000017985.1_ASM1798v1/GCA_000017985
.1_ASM1798v1_genomic.fna.gz
$ gunzip data/ref_genome/ecoli_rel606.fasta.gz
$ head data/ref_genome/ecoli_rel606.fasta
CP000819.1 Escherichia coli B str. REL606, complete
genome
```

Thử nghiệm với tập dữ liệu con

$ curl -L -o sub.tar.gz https://ndownloader.figshare.com/files/14418248

$ tar xvf sub.tar.gz

$ mv sub/ ~/dc_workshop/data/trimmed_fastq_small

- Tạo thư mục chứa các file kết quả

$ mkdir -p results/sam results/bam results/bcf results/vcf

- Indexing reference genome

$ bwa index data/ref_genome/ecoli_rel606.fasta

Align reads to reference genome

```
$ bwa mem data/ref_genome/ecoli_rel606.fasta
data/trimmed_fastq_small/SRR2584866_1.trim.sub.fastq
data/trimmed_fastq_small/SRR2584866_2.trim.sub.fastq >
results/sam/SRR2584866.aligned.sam
```

- Convert SAM file to BAM format:

```
$ samtools view -S -b results/sam/SRR2584866.aligned.sam >
results/bam/SRR2584866.aligned.bam
```

- Sort BAM file by coordinates:

```
samtools sort -o results/bam/SRR2584866.aligned.sorted.bam
results/bam/SRR2584866.aligned.bam
```

SAM Format

```
HWI-ST330:304:H045HADXX:20934#0  16  chr1  60023  50  100M  *  0  0
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT  ?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR
HWI-ST330:304:H045HADXX:20934#0	16	chr1	60023	50	100M

```
HWI-ST330:304:H045HADXX:20934#0  16  chr1  60023  50  100M  *  0  0
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT  ?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

SEQ	QUAL	MRNM	MPOS	ISIZE
		*	0	0

SAM/BAM format

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

https://samtools.github.io/hts-specs/SAMv1.pdf

Xem thông tin alignment

samtools flagstat results/bam/SRR2584866.aligned.sorted.bam

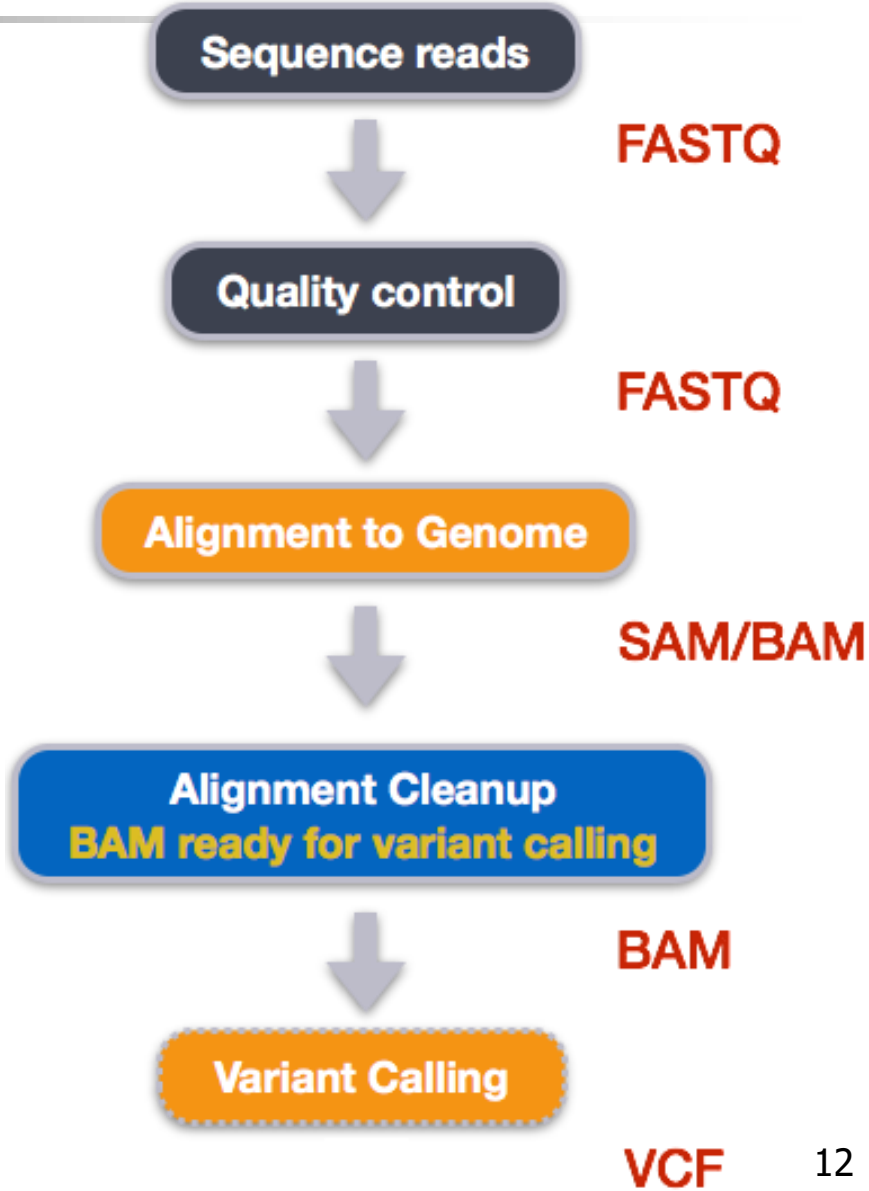
- 351169 + 0 in total (QC-passed reads + QC-failed reads)
- 0 + 0 secondary
- 1169 + 0 supplementary
- 0 + 0 duplicates
- 351103 + 0 mapped (99.98% : N/A)
- 350000 + 0 paired in sequencing

Xem thông tin alignment

- 175000 + 0 read1
- 175000 + 0 read2
- 346688 + 0 properly paired (99.05% : N/A)
- 349876 + 0 with itself and mate mapped
- 58 + 0 singletons (0.02% : N/A)
- 0 + 0 with mate mapped to a different chr
- 0 + 0 with mate mapped to a different chr (mapQ>=5)

Bước 2. Variant Calling

- Single Nucleotide Variant (SNV), Single Nucleotide Polymorphism (SNP)



Step 1: Calculate the read coverage of positions in the genome

```
$ bcftools mpileup -O b
-o results/bcf/SRR2584866_raw.bcf
-f data/ref_genome/ecoli_rel606.fasta
results/bam/SRR2584866.aligned.sorted.bam
```

- Step 2: Detect the single nucleotide variants (SNVs)

```
$ bcftools call --ploidy 1 -m -v -o
results/vcf/SRR2584866_variants.vcf
results/bcf/SRR2584866_raw.bcf
```

Step 3: Filter and report the SNV variants in variant calling format (VCF)

- Filter the SNVs for the final output in VCF format, using vcfutils.pl:

- $ vcfutils.pl varFilter results/vcf/SRR2584866_variants.vcf > results/vcf/SRR2584866_final_variants.vcf
- $ less -S results/vcf/SRR2584866_final_variants.vcf

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.8+htslib-1.8
##bcftoolsCommand=mpileup -O b -o results/bcf/SRR2584866_raw.bcf -f data/ref_genome/ecoli_rel606.fasta results/bam/SRR2584866.aligned.sorted.bam
```

##reference=file://data/ref_genome/ecoli_rel606.fasta

##contig=<ID=CP000819.1,length=4629812>

##ALT=<ID=*,Description="Represents allele(s) other than observed.">

##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">

##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">

##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">

##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">

##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version=

##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">

##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">

##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">

##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">

##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">

##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">

##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">

##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">

##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">

##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">

##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">

- ##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
- ##bcftools_callVersion=1.8+htslib-1.8
- ##bcftools_callCommand=call --ploidy 1 -m -v -o results/bcf/SRR2584866_variants.vcf results/bcf/SRR2584866_raw.bcf; Date=Tue Oct 9 18:48:10 2018

- Followed by information on each of the variations observed:

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT results/bam/SRR2584866.aligned.sorted.bam

CP000819.1 1521 . C T 207 . DP=9;VDB=0.993024;SGB=-0.662043;MQSB=0.974597;MQ0F=0;AC=1;AN=1;DP4=0,0,4,5;MQ=60

CP000819.1 1612 . A G 225 . DP=13;VDB=0.52194;SGB=-0.676189;MQSB=0.950952;MQ0F=0;AC=1;AN=1;DP4=0,0,6,5;MQ=60

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

CP000819.1 9092 . A G 225 . DP=14;VDB=0.717543;SGB=-0.670168;MQSB=0.916482;MQ0F=0;AC=1;AN=1;DP4=0,0,7,3;MQ=60

CP000819.1 9972 . T G 214 . DP=10;VDB=0.022095;SGB=-0.670168;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,2,8;MQ=60 GT:PL

CP000819.1 10563 . G A 225 . DP=11;VDB=0.958658;SGB=-0.670168;MQSB=0.952347;MQ0F=0;AC=1;AN=1;DP4=0,0,5,5;MQ=60

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

CP000819.1 22257 . C T 127 . DP=5;VDB=0.0765947;SGB=-0.590765;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,2,3;MQ=60 GT:PL

CP000819.1 38971 . A G 225 . DP=14;VDB=0.872139;SGB=-0.680642;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,4,8;MQ=60 GT:PL

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

CP000819.1 42306 . A G 225 . DP=15;VDB=0.969686;SGB=-0.686358;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,5,9;MQ=60 GT:PL

CP000819.1 45277 . A G 225 . DP=15;VDB=0.470998;SGB=-0.680642;MQSB=0.95494;MQ0F=0;AC=1;AN=1;DP4=0,0,7,5;MQ=60

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

CP000819.1 56613 . C G 183 . DP=12;VDB=0.879703;SGB=-0.676189;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,8,3;MQ=60 GT:PL

CP000819.1 62118 . A G 225 . DP=19;VDB=0.414981;SGB=-0.691153;MQSB=0.906029;MQ0F=0;AC=1;AN=1;DP4=0,0,8,10;MQ=59

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

CP000819.1 64042 . G A 225 . DP=18;VDB=0.451328;SGB=-0.689466;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,7,9;MQ=60 GT:PL

Exercise

- Use the grep and wc commands you have learned to assess how many variants are in the vcf file.
- $ grep -v "#" results/vcf/SRR2584866_final_variants.vcf | wc -l

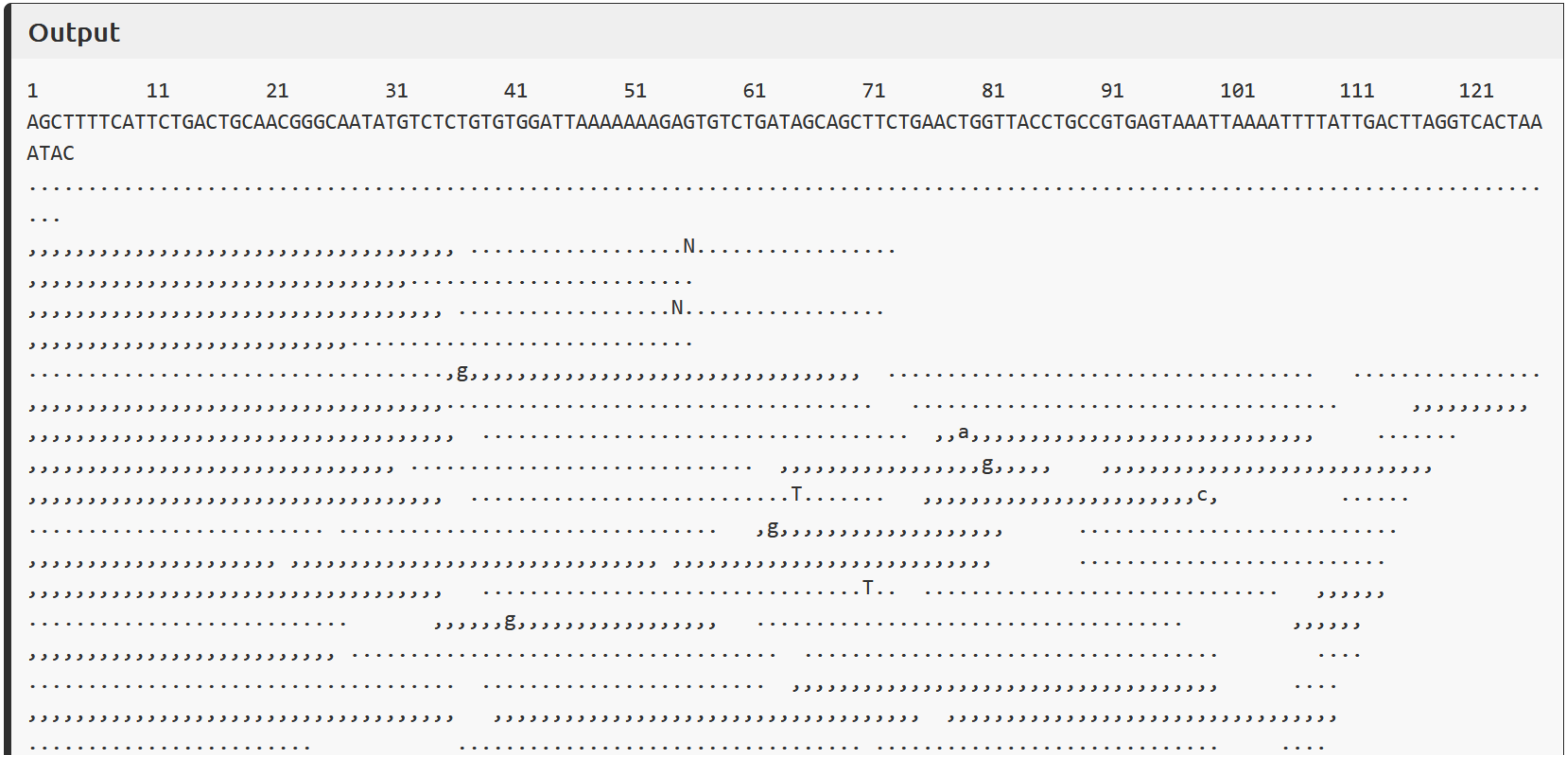
Assess the alignment (visualization) - optional step

- index the BAM file:

$ samtools index results/bam/SRR2584866.aligned.sorted.bam

- visualize our mapped reads:

$ samtools tview results/bam/SRR2584866.aligned.sorted.bam data/ref_genome/ecoli_rel606.fasta



Output

```
1             11            21            31            41            51            61            71            81            91            101           111           121
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAA
ATAC
...............................................................................................................................
...
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ..................N.................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,........................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ..................N.................
,,,,,,,,,,,,,,,,,,,,,,,,,,,.............................
...................................,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ....................................   ................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,....................................   ....................................      ,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ....................................  ,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,     .......
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, .............................  ,,,,,,,,,,,,,,,,,g,,,,,    ,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ...........................T.......   ,,,,,,,,,,,,,,,,,,,,,,,c,          ......
......................... ................................   ,g,,,,,,,,,,,,,,,,,,,      ...........................
,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,,,,       ..........................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ................................T..  ..............................   ,,,,,,
...........................       ,,,,,,g,,,,,,,,,,,,,,,,,   ....................................         ,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,, ....................................  ...................................        ....
....................................  ........................  ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,      ....
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,   ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
........................            .................................. .............................     ....
```

- To navigate to a specific position, type g. A dialogue box will appear.
- In this box, type the name of the “chromosome” followed by a colon and the position of the variant you would like to view (e.g. for this sample, type CP000819.1:50 to view the 50th base.
- Type Ctrl^C or q to exit tview.

Exercise

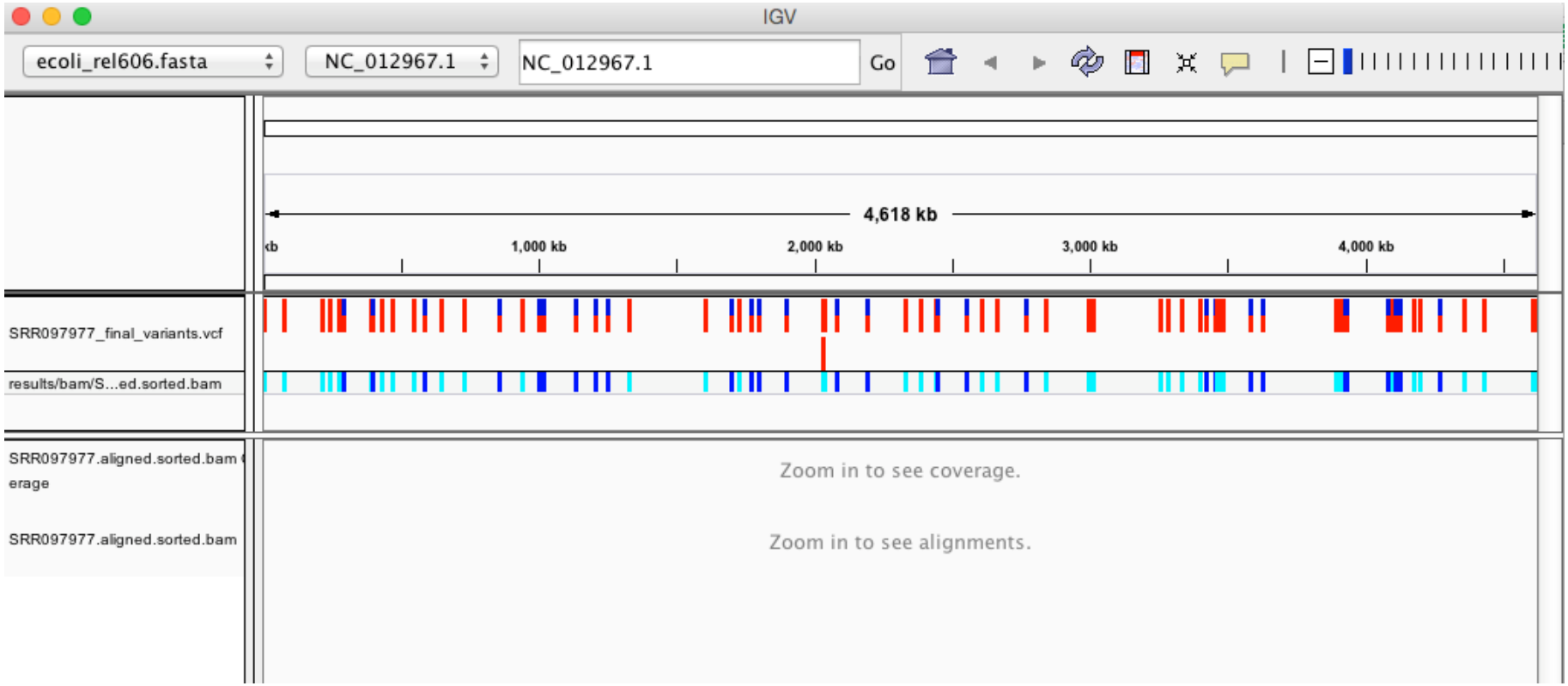
- Visualize the alignment of the reads for our SRR2584866 sample.
- What variant is present at position 4377265?
- What is the canonical nucleotide in that position?
- $ samtools tview ~/dc_workshop/results/bam/SRR2584866.aligned.sorted.bam ~/dc_workshop/data/ref_genome/ecoli_rel606.fasta

- Then type g. In the dialogue box, type CP000819.1:4377265. G is the variant. A is canonical.
- This variant possibly changes the phenotype of this sample to hypermutable.
- It occurs in the gene mutL, which controls DNA mismatch repair.

Viewing with IGV

```
$ mkdir ~/Desktop/files_for_igv
$ cd ~/Desktop/files_for_igv
$ cp results/bam/SRR2584866.aligned.sorted.bam ~/Desktop/files_for_igv
$ cp results/bam/SRR2584866.aligned.sorted.bam.bai ~/Desktop/files_for_igv
$ cp data/ref_genome/ecoli_rel606.fasta ~/Desktop/files_for_igv
$ cp results/vcf/SRR2584866_final_variants.vcf ~/Desktop/files_for_igv
```

- Open IGV.
- Load our reference genome file (ecoli_rel606.fasta) into IGV using the “Load Genomes from File…“ option under the “Genomes” pull-down menu.
- Load our BAM file (SRR2584866.aligned.sorted.bam) using the “Load from File…“ option under the “File” pull-down menu.
- Do the same with our VCF file (SRR2584866_final_variants.vcf).


IGV
ecoli_rel606.fasta
NC_012967.1
NC_012967.1
Go
4,618 kb
1,000 kb
2,000 kb
3,000 kb
4,000 kb
SRR097977_final_variants.vcf
results/bam/S...ed.sorted.bam
SRR097977.aligned.sorted.bam
erage
Zoom in to see coverage.
SRR097977.aligned.sorted.bam
Zoom in to see alignments.