Bversity

1st Ever Virtual College for Biotech

Name : Rajashree.U

Contact:rajashree5820@gmail.com

Mobile:+916383331696

Title: Genomics and Variant Calling analysis

Answer the following questions below and upload them in the google form.

1. Analyze how the E. coli Ara-3 population evolved over time by identifying genetic variants compared to the reference strain REL606.

a. Mention evolutionary trends observed in the Ara-3 population.

b. Discuss the implications of your findings of E. coli evolution in about 350 words.

Genomics and Variant Calling analysis

BRIEF:

E. coli, a common bacterium, offers a fascinating case study of evolution in action. We're exploring the Ara-3 population, which evolved a remarkable ability to utilize a previously inaccessible food source. To understand how this adaptation came about, we'll be comparing its DNA to its ancestor, REL606. By decoding the differences between their genetic blueprints, we can pinpoint the key mutations that fueled Ara-3's success. This will involve aligning the DNA of Ara-3 samples to the original strain's genome and identifying any differences, then analyzing these changes to understand their potential impact on Ara-3's adaptation.

steps involved in this variant calling analysis:

- Alignment to reference genome.The alignment process consists of two steps:
    - Indexing the reference genome
    - Aligning the reads to the reference genome
- the reference genome **for E. coli REL606** The name is **CP000819.1** Escherichia coli B str. REL606, complete genome. chromosome name (CP000819.1)

- The first step is **to index the reference genome** for use by **BWA.**

- The alignment process consists of choosing an appropriate reference genome to map our reads against and then deciding on an aligner. We will use the **BWA-MEM algorithm**, which is the latest and is generally recommended for high-quality queries as it is faster and more accurate.

- The **SAM file**, is a tab-delimited text file that contains information for each individual read and its alignment to the genome.

- The compressed binary version of SAM is called a **BAM file**. We use this version to reduce size and to allow for indexing, which enables efficient random access of the data contained within the file.

- convert the SAM file to BAM format using the samtools program with the view command and tell this command that the input is in SAM format (-S) and to output BAM format (-b)

- Next we sort the BAM file using the sort command from samtools. -o tells the command where to write the output

- SAM/BAM files can be sorted in multiple ways, e.g. by location of alignment on the chromosome, by read name, etc. It is important to be aware that different alignment tools will output differently sorted SAM/BAM, and different downstream tools require differently sorted alignment files as input. You can use samtools to learn more about this bam file as well.

- first pass on variant calling by counting read coverage with **bcftools**. We will use the command mpileup. The flag -O b tells bcftools to generate a bcf format output file, -o specifies where to write the output file, and -f flags the path to the reference genome:

Detect the single nucleotide variants (SNVs)

- Identify SNVs **using bcftools call**. We have to specify ploidy with the flag --ploidy, which is one for the haploid E. coli. -m allows for multiallelic and rare-variant calling, -v tells the program to output variant sites only (not every site in the genome), and -o specifies where to write the output file:

- You will see the header (which describes the format), the time and date the file was created, the version of bcftools that was used, the command line parameters used, and some additional information

Output:

```
##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


#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
```

The first few columns represent the information we have about a predicted variation.In an ideal world, the information in the QUAL column would be all we needed to filter out bad variant calls. the grep and wc commands you have learned to assess how many variants are in the vcf file.

- **visualization tools are useful for exploratory analysis .**

For us to visualize the alignment files, we will need to index the BAM file using 'samtools'

- **Viewing with tview**

Samtools implements a very simple text alignment viewer based on the GNU ncurses library, called tview. This alignment viewer works with short indels and shows MAQ consensus.

- **Viewing with IGV(Integrated Graphics Viewer)**

IGV is a stand-alone browser, which has the advantage of being installed locally and providing fast access. Web-based genome browsers, like Ensembl or the UCSC browser, are slower, but provide more functionality. They not only allow for more polished and flexible visualization, but also provide easy access to a wealth of annotations and external data sources.

STEPS:

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

The commands used to run variant calling analysis for three samples against reference genome taken are as follows:

Download genome:

- Input: Bash

```
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
```

- **Input: Getting easy-to-work FASTQ files**

  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

- **Input: Create result directories**

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

- **Index the reference genome**

- **Input: Indexing with BWA**

  bwa index data/ref_genome/ecoli_rel606.fasta

- **Input: Aligning with BWA(We have 3 different samples highlightened)**

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 > data/results/sam/SRR2584866.aligned.sam

bwa mem data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/SRR2584863_1.trim.sub.fastq data/trimmed_fastq_small/**SRR2584863_2**.trim.sub.fastq > data/results/sam/SRR2584863.aligned.sam

bwa mem data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/SRR2589044_1.trim.sub.fastq data/trimmed_fastq_small/**SRR2589044_2.**trim.sub.fastq > data/results/sam/SRR2589044.aligned.sam

(do this for SRR2584866 SRR2584863 SRR2589044)

- **SAM to BAM**

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

samtools view -S -b results/sam/SRR2584863.aligned.sam > results/bam/SRR2584863.aligned.bam

samtools view -S -b results/sam/SRR2589044.aligned.sam > results/bam/SRR2589044.aligned.bam

- **Sort BAM file by coordinates**

(Next we sort the BAM file using the sort command from samtools. -o tells the command where to write the output.

- **Input: "sort" command**

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

samtools sort -o results/bam/SRR2584863.aligned.sorted.bam results/bam/SRR2584863.aligned.bam

samtools sort -o results/bam/SRR2589044.aligned.sorted.bam results/bam/SRR2589044.aligned.bam

- **#Input: "flagstat" command**

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

samtools flagstat results/bam/SRR2584863.aligned.sorted.bam

samtools flagstat results/bam/SRR2589044.aligned.sorted.bam

BCFTOOLS COMMAND FOR VARIANT CALLING

- **Calculate the read coverage of positions in the genome**

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

bcftools mpileup -O b -o data/results/bcf/SRR2584863_raw.bcf -f data/ref_genome/ecoli_rel606.fasta data/results/bam/SRR2584863.aligned.sorted.bam

bcftools mpileup -O b -o data/results/bcf/SRR2589044_raw.bcf -f data/ref_genome/ecoli_rel606.fasta data/results/bam/SRR2589044.aligned.sorted.bam

- **Detect the single nucleotide variants (SNVs)**

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

bcftools call --ploidy 1 -m -v -o results/vcf/SRR2584863_variants.vcf results/bcf/SRR2584863_raw.bcf

bcftools call --ploidy 1 -m -v -o results/vcf/SRR2589044_variants.vcf results/bcf/SRR2589044_raw.bcf

- **(Filter and report the SNV variants in variant calling format (VCF).Filter the SNVs for the final output in VCF format, using vcfutils.pl:**

Input: Filtering with 'vcfutils.pl' command)

vcfutils.pl varFilter results/vcf/SRR2584866_variants.vcf > results/vcf/SRR2584866_final_variants.vcf

vcfutils.pl varFilter results/vcf/SRR2584863_variants.vcf > results/vcf/SRR2584863_final_variants.vcf

vcfutils.pl varFilter results/vcf/SRR2589044_variants.vcf > results/vcf/SRR2589044_final_variants.vcf

- **Explore the VCF format:**

less -S results/vcf/SRR2584866_final_variants.vcf

less -S results/vcf/SRR2584863_final_variants.vcf

less -S results/vcf/SRR2589044_final_variants.vcf

- **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

grep -v "#" results/vcf/SRR2584863_final_variants.vcf | wc -l

grep -v "#" results/vcf/SRR2589044_final_variants.vcf | wc -l

- **Assess the alignment (visualization)**

In order for us to visualize the alignment files, we will need to index the BAM file using 'samtools':

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

samtools index results/bam/SRR2584863.aligned.sorted.bam

samtools index results/bam/SRR2589044.aligned.sorted.bam

- **Viewing with tview**

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

samtools tview results/bam/SRR2584863.aligned.sorted.bam ref_genome/ecoli_rel606.fasta

samtools tview results/bam/SRR2589044.aligned.sorted.bam ref_genome/ecoli_rel606.fasta

- **Visualize the alignment of the reads .**

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

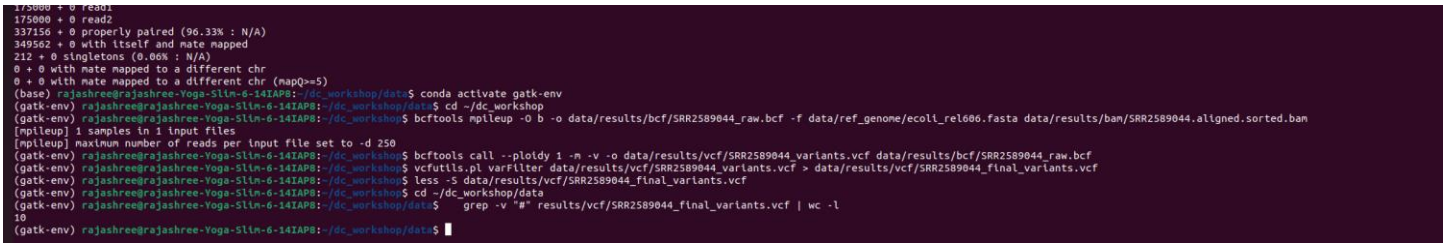
samtools tview ~/dc_workshop/data/results/bam/SRR2584863.aligned.sorted.bam ~/dc_workshop/data/ref_genome/ecoli_rel606.fasta

samtools tview ~/dc_workshop/data/results/bam/SRR2589044.aligned.sorted.bam ~/dc_workshop/data/ref_genome/ecoli_rel606.fasta

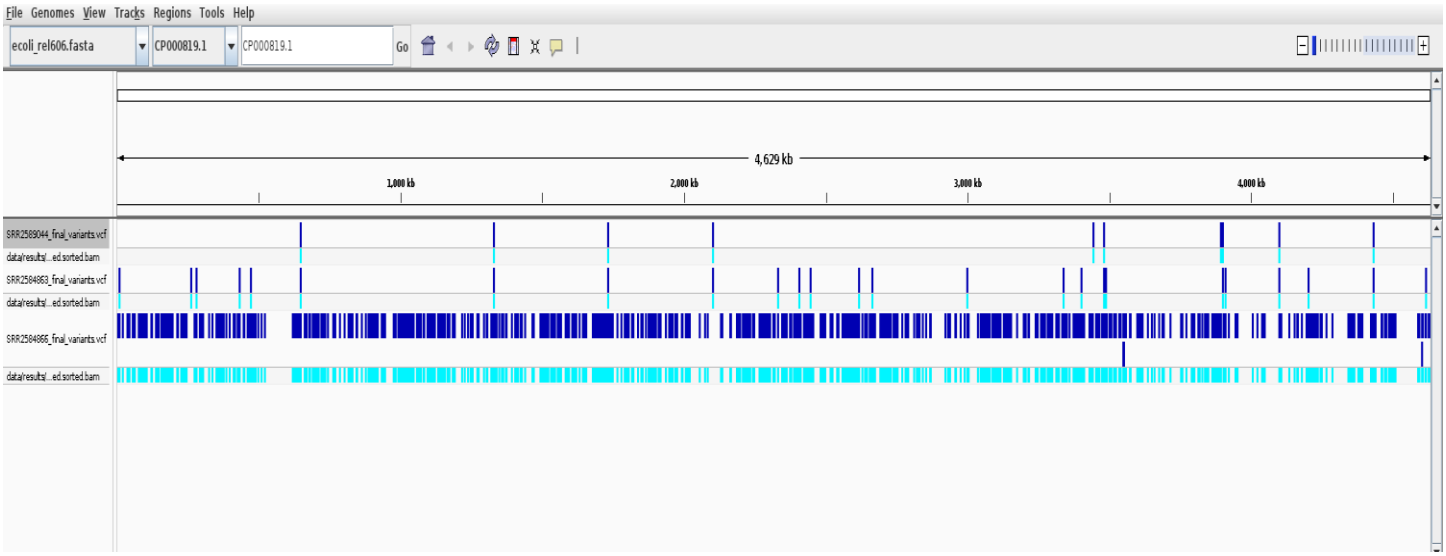
- **Viewing with IGV:**

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

Results and Observation:



The final variants for sample **SRR2589044** is about **10**



The above is the IGV viewer image of three samples against the given reference genome

```
(base) rajashree@rajashree-Yoga-Slim-6-14IAP8:~/dc_workshop/data$ grep -v "#" results/vcf/SRR2584866_final_variants.vcf | wc -l
775
(base) rajashree@rajashree-Yoga-Slim-6-14IAP8:~/dc_workshop/data$   grep -v "#" results/vcf/SRR2584863_final_variants.vcf | wc -l
25
(base) rajashree@rajashree-Yoga-Slim-6-14IAP8:~/dc_workshop/data$    grep -v "#" results/vcf/SRR2589044_final_variants.vcf | wc -l
10
(base) rajashree@rajashree-Yoga-Slim-6-14IAP8:~/dc_workshop/data$
```

This gives how many variants there are thereby aligning the BAM file and VCF files of each sample to reference the genome.

- The sample SRR2589044 has 10 variants.
- The sample SRR2589063 has 25 variants.
- The sample SRR2589066 has775 variants.

  The sample SRR2589066 has evolved the most with variants as time increases compared to other samples of different runtimes.

Evolutionary trends observed in the Ara-3 population:

- Rapid increase in fitness: In the Long-Term Evolution Experiment (LTEE), the Ara-3 population showed a rapid increase in fitness relative to the ancestral strain during early generations. This is likely due to adaptation to the specific growth conditions of the experiment, such as limited glucose availability.

- Diversification and innovation: After around 31,000 generations, the Ara-3 population evolved the ability to use citrate as an additional carbon source, a trait absent in the ancestor. This represents a significant innovation and diversification within the population.
- Coexistence and competition: The emergence of citrate-utilizing (Cit+) bacteria within the Ara-3 population led to a dynamic of coexistence and competition with the non-utilizing (Cit-) bacteria. This can lead to further evolutionary pressures and selection for traits advantageous in this competition.

Implications of E. coli evolution:

- Studying the LTEE and the Ara-3 population provides valuable insights into the mechanisms and dynamics of microbial adaptation, with potential implications for understanding microbial ecology, pathogens, and antibiotic resistance.
- The difficulty of evolving the Cit+ phenotype despite its potential advantage highlights the influence of historical context and genetic constraints on evolutionary events. This emphasizes the non-deterministic nature of evolution and the potential for alternative outcomes under different conditions.
- The LTEE's unique length and detailed data provide a powerful resource for studying evolution in real-time and revealing insights that might be masked in shorter-term experiments.