2019

Designing Computational Biology Workflows with Perl - Part 2

Esma Yildirim
CUNY Queensborough Community College

Title: Designing Computational Biology Workflows with Perl – Part2
Author/Affiliation: Esma Yildirim / Queensborough Community College
Date: 05/15/2019
Material Type: Lecture
CS + Computational Biology
Software/Equipment Dependencies: An Amazon Web Services (AWS) account, a web browser and a command line interpreter program (e.g. Putty on Windows, Terminal on Linux/MacOSX)
Prior Knowledge Needed (if any): The material covered in “Designing Computational Biology Workflows with Perl – Part1”
Keywords: Gene-sequencing file formats, SNPs, INDELs, Alignment, Variant Calling, Perl
Approximate time needed: 2 hours
Description: This material briefly reintroduces the DNA double Helix structure, explains SNP and INDEL mutations in genes and describes FASTA, FASTQ, BAM and VCF file formats. It also explains the index creation, alignment, sorting, marking duplicates and variant calling steps of a simple preprocessing workflow and how to write a Perl script to automate the execution of these steps on a Virtual Machine Image.

Designing Computational Biology Workflows with Perl – Part 2

The basic steps of a pre-processing workflow after the initial raw data comes out of the gene-sequencer consist of alignment to the reference genome, sorting, marking duplicates and variant calling. Scientists have a plethora of options when selecting tools to do these operations. In this section, we will be looking into these tools and learn how to construct and automate the execution of the steps of a basic preprocessing workflow using Perl scripting language in the Cloud.

1. DNA Structure

A DNA sequence has a double helix structure that looks like a staircase consisting of base pairs. There are four types of bases (nucleotides) in a DNA molecule: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The bases are paired on a sugar-phosphate backbone structure. Adenine base is always paired with Thymine and Guanine base is always paired with Cytosine (Figure 1).

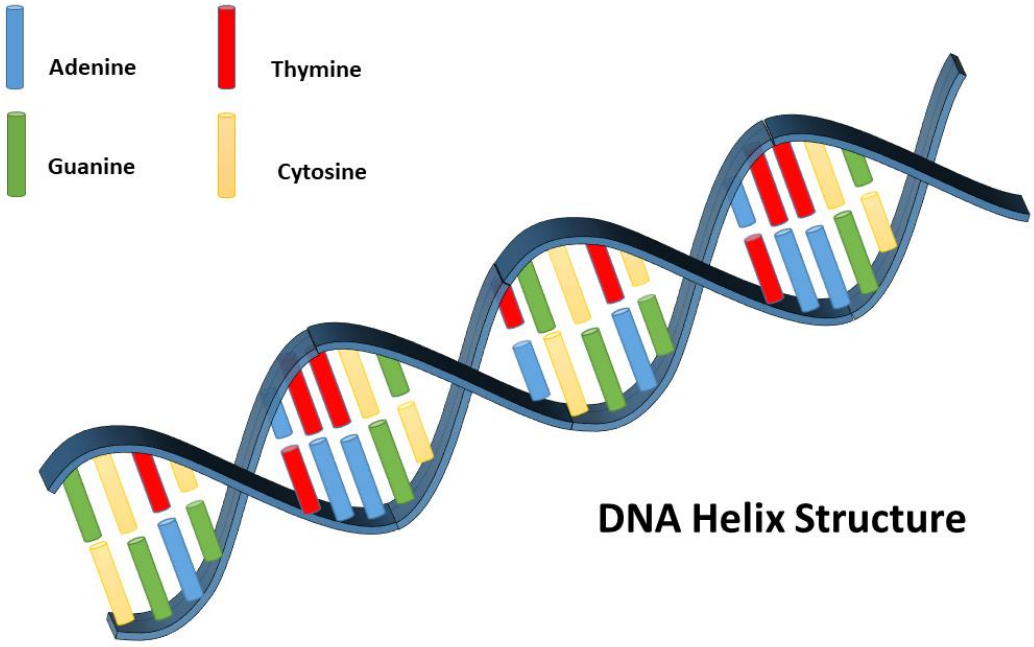

Figure 1. DNA Double Helix Structure

These base pairs (A-T and G-C) cause the two strands of the DNA to have complementary gene sequences. A sample base pair sequence:

Strand 1: ATGGTCGTTAG
Strand 2: TACCTGCAATC

The nitrogen bases can connect through hydrogen bonds (Figure 2.). Adenine and Thymine have two hydrogen bonds in between, while Guanine

and Cytosine have three. On the other ends, each base is connected to the other in line via sugar-phosphate molecules.

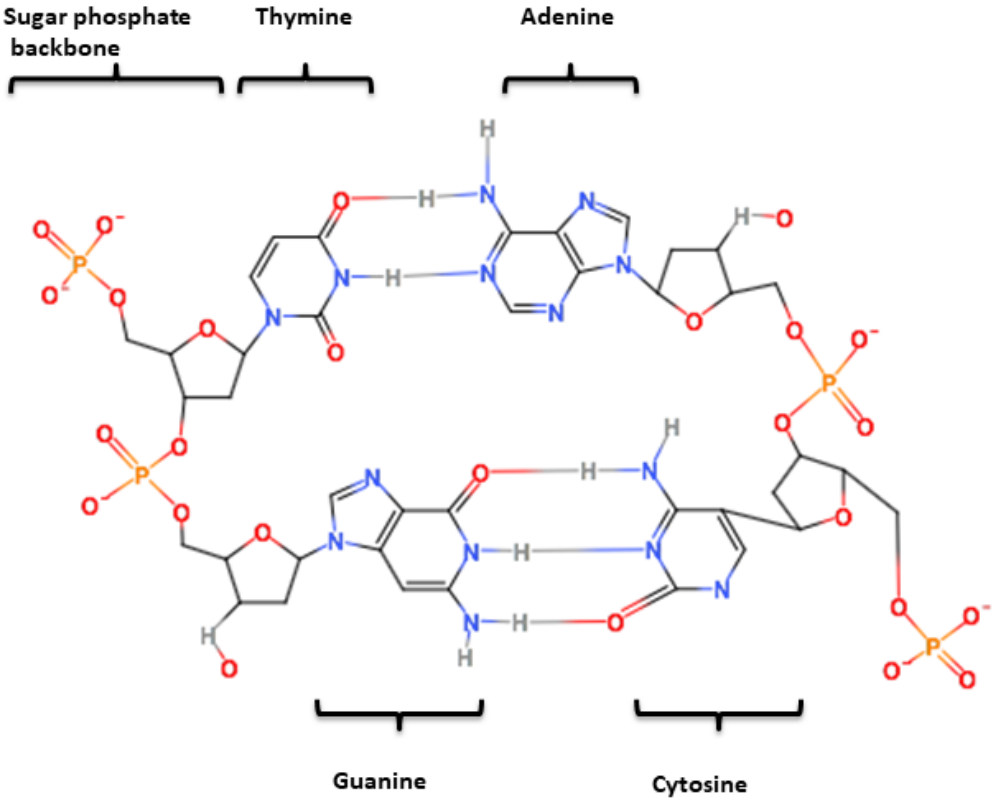

Figure 2. DNA base pair molecules

2. Genetic Variations: SNPs, MNPs, INDELs

Genetic variations represent the differences between a specific species' genome and the reference genome. A genome is a complete set of genes in a cell and a reference genome is the representative sample genome that belong to a species. There are different types of variations; the most common ones are SNPs, MNPs and INDELs.

2.1. Single-nucleotide polymorphism (SNP)

It is a point mutation where one base (nucleotide) is replaced with another when compared to the reference genome DNA strand:

Reference genome strand: AATGGTTGACCTA
Sample strand: AATG**A**TTGACCTA

In the above example, the fifth base in the sample strand is a Adenine(A) base, however in the reference genome, it is a Guanine (G) base. Therefore, G->A is considered a SNP type of mutation (variation).

2.2. Multi-nucleotide polymorphism (MNP)

When multiple consequent SNP mutations occur in the sample strand, compared to a reference strand, the mutation is called a MNP:

Reference genome strand: AATGGTTGACCTA
Sample strand: AA**CAA**TTGACCTA

In the above example, the 3rd, 4th, and 5th base sequence CAA is considered a MNP mutation because, the original sequence in the reference genome is TGG.

2.3. Insertion/Deletion (INDEL)

Insertions may occur when single/multiple nucleotide sequences are introduced in the sample strand but they do not exist in the reference genome strand:

Reference genome strand: AATGGTTGACCTA
Sample strand: AA**CAACC**TGGTTGACCTA

In the above example, the CAACC sequence in the sample strand does not exist in the reference genome.

Deletions occur when single/multiple nucleotide sequences are missing in the sample strand when compared to the reference genome:

Reference genome strand: AATGGTTGACCTA
Sample strand: AATGGTT _ ACCTA

In the above example, after the 7th Thymine (T) nucleotide there should have been a Guanine (G), but it is missing in the sample strand.

3. Steps of a Preprocessing Gene-sequencing Workflow

The revolutionary “short read” method [1] divides DNA into small samples, replicates them and reads those strands in a massively parallel high-throughput fashion. However, these short reads need to be matched into a reference genome before we can identify genetic variations in the gene structure. In this section, we will be discussing alignment, sorting, marking duplicates and variant calling steps of a pre-processing workflow.

3.1. Alignment

A reference genome is a completely sequenced, representative genome of a certain species. Alignment process matches the short reads to their exact position in the reference genome. In Figure 3, multiple short reads are aligned to their exact positions in the reference genome strand. In this figure, the depth or coverage of a base point is decided based on the number of reads this point exist in. For example, the Guanine base in the 7th position of the

reference genome appears in 2 short reads, therefore its depth is considered 2x. Another example is the Thymine base in the 12th base position in the reference genome appear in 5 short reads, resulting in a 5x depth.

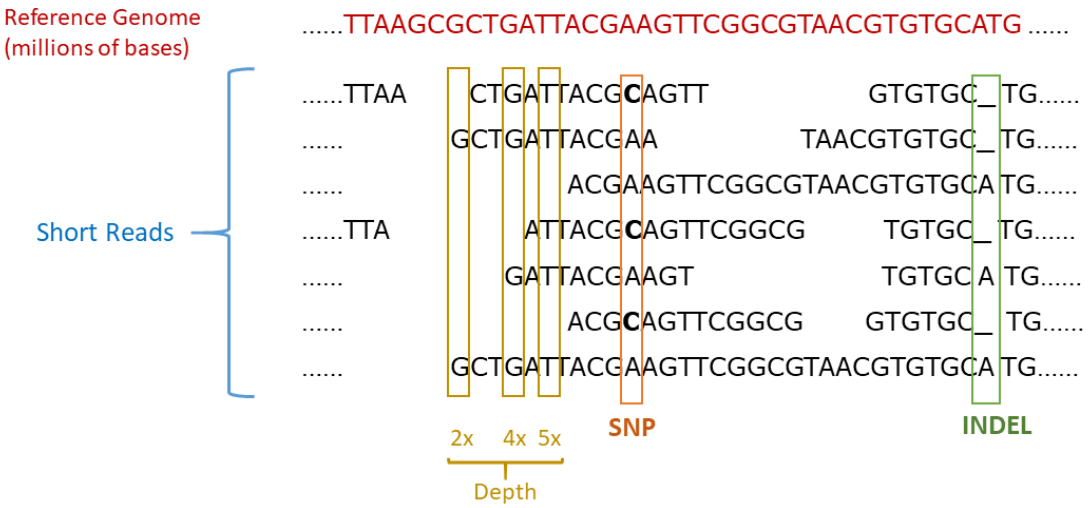

Figure 3. Alignment

Figure 4 presents the first few lines of a FASTA file, which is a very commonly used file format to store the reference genome. This example reference genome is from E.coli K12 strand microorganism.

```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTT
GCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGT
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCT
GAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCG
CCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGC
TGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGT
CAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCG
CTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATAC
CGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGC
ATTTCCAATCTGAATAACATGGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGG
CGGCGCGCGTCTTTGCAGCGATGTCACGCGCCCGTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGA
ATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTGCGAGCTGAACGGGCAATGCAGGAAGAGTTC
TACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGG
TAGGTGATGGTATGCGCACCTTGCGTGGGATCTCGGCGAAATTCTTTGCCGCACTGGCCCGCGCCAATAT
```

Figure 4. E.coli Strand K12 reference genome FASTA file excerpt

The raw short read data is generally stored in another format called FASTQ. They usually come in pairs to represent the double strand of DNA structure.

Figure 5 shows an excerpt from a sample from National Center for Biomedical Information database for E.coli microorganism. The name of the sample is SRR1770413. The length variable shows the number of bases in each read followed by the read sequence

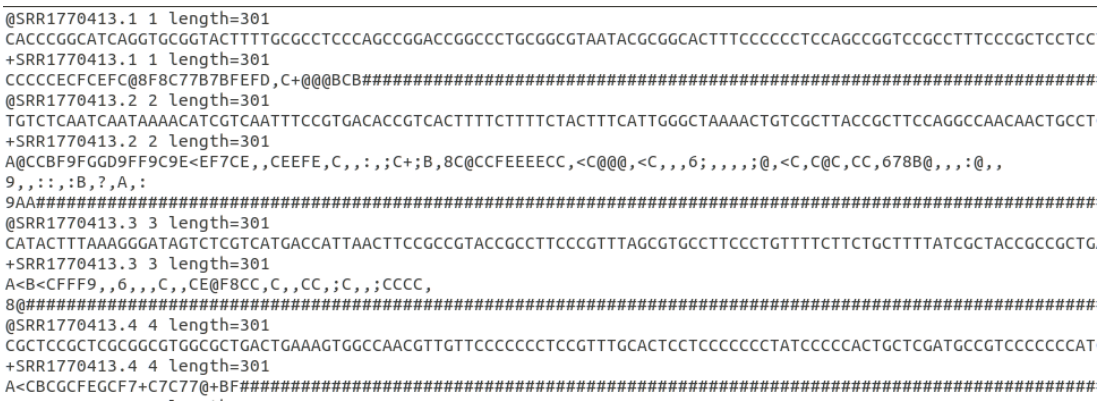

Figure 5. SRR1770413 E.coli microorganism FASTQ file excerpt

The tools we are going to use for the alignment step are **samtools** and **bwa**. These tools are already installed and ready to run in the Virtual Machine Image provided with this course.

We will start with generating some index files from the reference genome FASTA file. From the Terminal program, running the following command will create an index file with ".fai" extension so that several other tools can access certain regions in the FASTA file faster:

$samtools faidx /home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta

The argument "faidx" tells **samtools** to create the index, while the argument "/home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta" is the entire path of the FASTA file the index will be created from. After the index is created, "ls –l" command can be used to list files that ends with ".fai" extension to make sure that it is created.

$cd /home/ubuntu/input/
$ ls -l *.fai
-rw-r--r-- 1 ubuntu ubuntu 29 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.fai

A set of other index files is used by **bwa** (Borrow-Wheeler Aligner) alignment tool. To generate these files, the following command can be used:

$/home/ubuntu/bwa/bwa index /home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta

Again, the argument "index" tells **bwa** to create the index while the next argument "/home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta" is the entire path of the FASTA file, the index files will be created from. The resultant

index files have “.fasta.amb”, “.fasta.ann”, “.fasta.bwt”, “.fasta.bwt”, “.fasta.sa” and “.fasta.pac” extensions. After they are created, “ls –l” command can be used to list files that ends with “.fasta.” in their file names:

```
$ ls -l *.fasta.*
-rw-r--r-- 1 ubuntu ubuntu      12 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.amb
-rw-r--r-- 1 ubuntu ubuntu      98 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.ann
-rw-r--r-- 1 ubuntu ubuntu 4641732 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.bwt
-rw-r--r-- 1 ubuntu ubuntu      29 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.fai
-rw-r--r-- 1 ubuntu ubuntu 1160415 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.pac
-rw-r--r-- 1 ubuntu ubuntu 2320880 Apr 30 11:34 E.coli-str.K-12substr.MG1655.fasta.sa
```

Now that the indices are created, it is time to start the alignment process. The alignment process requires the FASTA file path, the FASTQ file path pair and several other parameters. The output is in a text file format called SAM file format. We will be using samtools again to convert that SAM file into a binary format called BAM file. Here is the entire command:

```
$ /home/ubuntu/bwa/bwa mem -t 2 -R '@RG\tID:K12\tSM:K12' \
   /home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta \
   /home/ubuntu/input/SRR1770413_1.fastq     \
   /home/ubuntu/input/SRR1770413_2.fastq \
   | samtools view -b ->    /home/ubuntu/input/SRR1770413.raw.bam
```

In the above command:

- –t 2 tells bwa to run on 2 processors. If your virtual machine is running on a 4-core processor then this argument can be given as –t 4.
- The IDs are important. '@RG\tID:K12\tSM:K12' portion should be written with care. Here, K12 is the only changeable parameter in this string. It gives the reads the read group name K12 and the sample name K12. You can give any name you want.
- Next comes the FASTA file path /home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta
- The FASTA file path is followed by the FASTQ file path pair: /home/ubuntu/input/SRR1770413_1.fastq /home/ubuntu/input/SRR1770413_2.fastq.
- The \ character at the end of each line allows you to go to another line in the command line. If you write the entire command in one line then \ is not needed.
- The output of the command up until this point is a SAM file. The | character allows us to give this output file to samtools to create the binary BAM file named /home/ubuntu/input/SRR1770413.raw.bam

The execution of this command might take a lot of time. Alignment is a costly process.

Once this is done, a raw aligned BAM file will be created. To see that the file is created properly “ls –l” command can be used:

```
$ ls -l *.raw.bam
-rw-r--r-- 1 ubuntu ubuntu 282074721 Apr 29 14:42 SRR1770413.raw.bam
```

The set of these commands can be automated using a Perl Script:

```
#perl script
$ref = "/home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta";
$fastq1 = "/home/ubuntu/input/ SRR1770413_1.fastq";
$fastq2 = "/home/ubuntu/input/ SRR1770413_2.fastq";
$sample = "/home/ubuntu/input/ SRR1770413";
$cmd = "samtools faidx $ref";
system($cmd);
$cmd = "/home/ubuntu/bwa/bwa index $ref";
system($cmd);
$cmd = "/home/ubuntu/bwa/bwa mem -t 4 -R \'\@RG\\tID:K12\\tSM:K12\' $ref
$fastq1 $fastq2 | samtools view -b - > $sample.raw.bam ";
system($cmd);
```

- $ref, $fastq1 and $fastq2 variables are used to hold the paths of the FASTA and FASTQ files.
- $sample variable contains the prefix name of the FASTQ files to be used as a prefix for the output files.
- $cmd variable is used to hold the command string. The only thing different in the commands is that variable names are used instead of file names and file paths and also we need an additional \ for ‘, @, \ characters so that Perl does not see them as special characters.
- system() function is used to run these commands from inside a Perl script.

Use **vim** command to copy and paste the above code into a file called myscript.pl and then use **perl** command to run the script:

```
$ vim myscript.pl
$ perl myscript.pl
```

3.2. Sorting and Marking Duplicates

After the short reads are matched against the reference genome, duplicate reads must be removed as they will make it difficult to identify SNPs and INDELs. Errors will propagate to all duplicates and the percentage of variations in the reads for the mutation to be considered as SNP or INDEL might be affected. Before marking duplicate operations, the reads should be sorted. Figure 6, shows an example for an error present in the duplicates and after marking of duplicates.

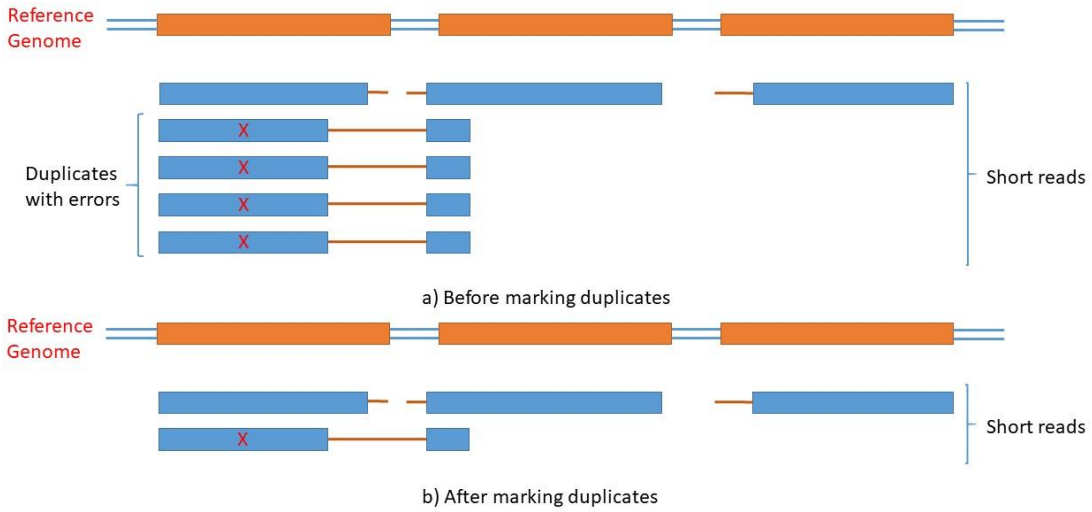

Figure 6. Marking duplicates

The tool for sorting and marking duplicates operations is called **sambamba**. When the raw BAM file is created after the alignment process, it is given to sambamba command as a parameter after the “sort” option:

$/home/ubuntu/sambamba sort /home/ubuntu/input/SRR1770413.raw.bam

This command will create a sorted file with the name SRR1770413.raw.sorted.bam. To list the file, “ls” command can be used:

$ls –l /home/ubuntu/input/*.sorted.bam

For marking duplicates, use the command as:

$/home/ubuntu/sambamba markdup –hash-table-size=4194304 /home/ubuntu/input/SRR1770413.raw.sorted.bam /home/ubuntu/input/SRR1770413.bam

Here, “/home/ubuntu/input/SRR1770413.raw.sorted.bam” is the path of the input file to be marked for duplicates, and “/home/ubuntu/input/SRR1770413.bam” is the output file after the operation. The hash table size is necessary, when the sorted file size is too big. You may list the output file with “ls –l” command:

$ls –l /home/ubuntu/input/*.bam

To automate the process, add the following lines to the end of the previous perl script:

```
$cmd = "/home/ubuntu/sambamba sort $sample.raw.bam";
system($cmd);
$cmd = "/home/ubuntu/sambamba markdup --hash-table-size=4194304 $sample.raw.sorted.bam $sample.bam";
```

system($cmd);

3.3. Variant Calling

The final step in the workflow to identify and visualize the variances is converting the BAM file into a VCF file where SNPs, MNPs and INDELs are listed. A very simple tool called **freebayes** will be used to do this conversion:

$freebayes -f /home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta --ploidy 1 /home/ubuntu/input/SRR1770413.bam /home/ubuntu/input/SRR1770413.vcf

Here, “/home/ubuntu/input/E.coli-str.K-12substr.MG1655.fasta” is the reference genome file, “--ploidy 1” option indicates to look for variations in haploid chromosomes, “/home/ubuntu/input/SRR1770413.bam” is the path of the input BAM file created after marking duplicates, and “/home/ubuntu/input/SRR1770413.vcf” is the path of the output VCF file to be created.

To automate this task, the following lines must be added to the end of the previous perl script after the marking duplicate operations lines:

$cmd = "freebayes -f $ref --ploidy 1 $sample.bam > $sample.vcf";
system($cmd);

IGV is a tool created by the Broad Institute of MIT and Harvard to visualize the variations in genomes[2]. After the reference genome FASTA file and VCF file is uploaded to the program, it can zoom in to specific sections of the sample genome. In Figure 7, when the mouse hovers over the red bar, a small window opens indicating that it was a SNP of T-> G variation.

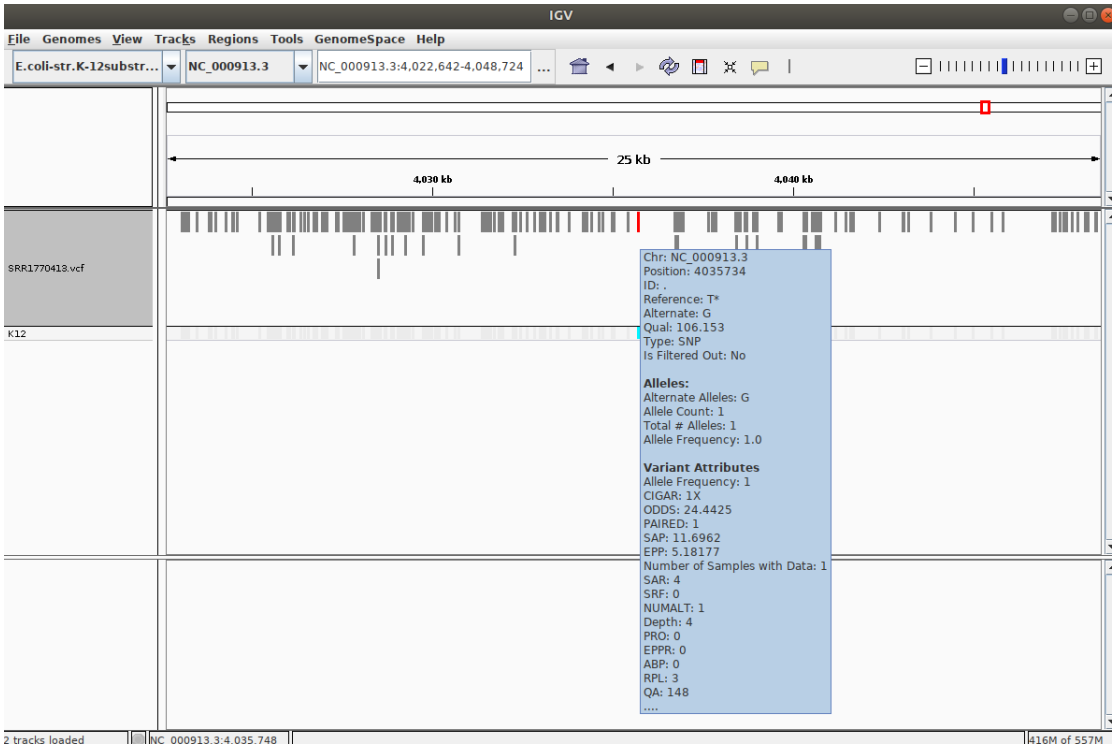

Figure 7. Variant visualization with IGV [2].

In conclusion, alignment, sorting, marking duplicates and variant calling tasks of a preprocessing workflow was automated by using Perl scripting language and run on a VMI. The results were visualized using IGV program where SNPs, MNPs and INDELs can be easily seen.

References:

[1] An introduction to Next Generation Sequencing Technology

[2] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011).

This OER material was produced as a result of the CS04ALL CUNY OER project.