Why we need to convert fastq files to bam files

Fastq files contains information for the millions or billions of clusters within the read. In order for the data to be useful the clusters must be aligned so that it turns into one sequence that represents the sample. Bam files are the standard way to represent these files where all the clusters are aligned into one sequence. Additionally, if you want to perform data analysis on the sample such as for a mutation almost all software only takes in a bam file to process.

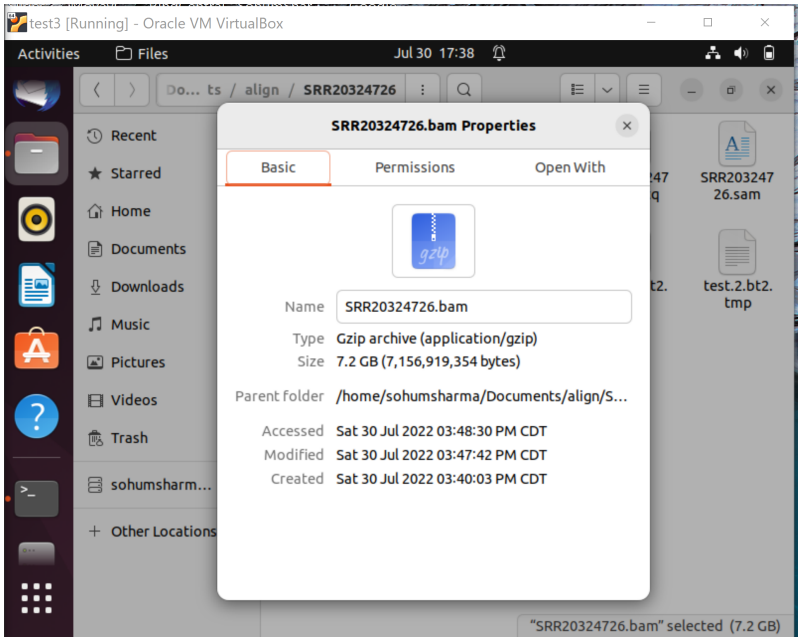
Aligning Human Exome FASTQ File Into BAM file

I used an Ubuntu virtual machine to align a fastq file into a bam file. Here are the steps I took to align a human exome fastq file:

- Get the fastq file for the human exome
  - Commands
    - Sudo apt-get install sra-toolkit
    - Prefetch SRR20324726
    - Fastq-dump –split-files SRR20324726
      - Since this is a paired read –split-files is added to separate the pair into 2 fastq files
  - Link to human exome file chosen: https://www.ncbi.nlm.nih.gov/sra/SRX16357953[accn]
- Install anaconda in order to install bowtie2 and samtools
  - commands
    - wget https://repo.anaconda.com/archive/Anaconda3-2020.02-Linux-x86_64.sh
    - bash Anaconda3-2020.02-Linux-x86_64.sh
- Install bowtie2 which would be used to align the fastq file to create a sam file
  - conda install -c bioconda bowtie2
- Install samtools to convert the sam file from bowtie2 into a bam file
  - conda install -c bioconda samtools
- Download reference genome for bowtie2 to use
  - Go to https://console.cloud.google.com/storage/browser/_details/genomics-public-data/resources/broad/hg38/v0/Homo_sapiens_assembly38.fasta;tab=live_object?pli=1 and click the download button. This will download a file named resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta
- Bowtie2 cannot work with the .fasta version of the reference genome, therefore we must run the following command in the same directory of the .fasta file previously downloaded to build it
  - Command
    - Bowtie2-build resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta bowtie2
  - The bowtie2 at the end is just the prefix added to all new files created and could therefore be anything. I will keep it as bowtie2 to make it clear what the files are for

- Run bowtie2 to align the fastq file
    - Command
        - Bowtie2 -x /home/sohumsharma/Documents/align/bowtie2 -1 /home/sohumsharma/Documents/alignls/SRR20324726/SRR20324726_1.fastq -2 /home/sohumsharma/Documents/alignls/SRR20324726/SRR20324726_2.fastq -S SRR20324726.sam
    - -x specifies the path to the folder containing the build reference genome. The /bowtie2 at the end is not a directory but rather the index used when building the .fasta file
- Convert the .sam file to a .bam file so that it could be passed to different programs for analysis
    - Command
        - Samtools view -b SRR20324726.sam > SRR20324726.bam
    - Absolute paths for the .sam and .bam (where it should be created) must be provided if it is not in the current directory
    - The -b tells samtools a .bam file should be outputted

Final result:



Extracting BAM file for 16s rRNA Gene

Link to source: https://www.ncbi.nlm.nih.gov/sra/SRX16688219[accn]

Prerequisites: I already have sra-toolkit, anaconda, bowtie2, and samtools installed. I also have the reference genome already built from the exome file

I ran the following commands in sequence:

- *prefetch SRR20665657*
- *fastq-dump –split-files SRR20665657.sra*
- *bowtie2 -x /home/sohumsharma/Documents/align/bowtie2 -1 /home/sohumsharma/Documents/align/SRR20665657_1.fastq -2 /home/sohumsharma/Documents/align/SRR20665657_2.fastq -S SRR20665657.sam*
- *samtools view -b SRR20665657.sam > SRR20665657.bam*

Link to BAM file I created for the 16s rRNA gene:
https://drive.google.com/file/d/1jOt7zoZi0kSg0_-TnS1x8hhPt4Kd0MPV/view?usp=sharing

Challenges

Many of the files I used were extremely large, so I kept running out of space on my Linux virtual machine. In the end, I had to allocate half (500 gigabytes) of the disk space on my computer for the virtual machine to handle all the files. Another challenge I ran into was decided which alignment method to use. I first found STAR but there wasn't enough documentation and explanations on how it works so I ended up using bowtie2 as it was more clear and smoother. I also had trouble with the bowtie2-build command on the reference genome where the terminal crashed when I ran the command. I searched through every possible solution on the internet, but none of them worked. I then thought through every possible reason and eventually decided to increase the RAM allocated to the virtual machine from 4096 mb to 16384 mb (the recommended was 1024 when I created the virtual machine). This ended up solving the problem and the command was able to run. It did take a long time, however, so I had to keep the computer running overnight.