Hello, my name is Dr. Tirza Doniger, and I work in the Bioinformatics Unit at Bar-Ilan University. I will be guiding you through the practice sessions of Lecture 4.

In this hands-on demonstration we will be using Galaxy to illustrate the analysis of DNA sequencing data. Galaxy is a free, web-based platform for data intensive biomedical research. Its goal is to develop and maintain a system that enables researchers with little or no informatics expertise to perform difficult computational analyses.

Galaxy is freely available for use at the link provided below:
https://usegalaxy.org

OK, let's begin. Click on the "Login or Register" link located at the top of your Galaxy interface. Click on "Register", unless of course you already have an account. This process will only take a few moments of your time, but it will allow you to save and share your data in between sessions. Once you have completed the registration process, you may now login.

As a registered user of the Galaxy main server, you are now entitled to store up to 250 gigabytes of data. Once you are logged in, you will note in the upper right-hand corner a small bar which will always show you how much of your allocated resources you are currently using. If you exceed your disk quota, no further jobs will be run until you have permanently deleted some of your datasets.

The right hand side of the screen contains the history panel. When data is uploaded or an analysis is performed on existing data using Galaxy, each output from those steps generates a dataset. These datasets and the output datasets from subsequent analyses are all stored by Galaxy in the History panel. The left side of the screen contains the list of programs that are available to perform analyses.

We will begin by retrieving the data for today's session. In the left-hand panel, click on the 'Get Data' link followed by Upload File. Then click Paste/Fetch data and enter the URL into the box that appears:
http://bioinfo.lnx.biu.ac.il/files/NA12878_subset_R1.fq.gz

You can leave the type as “Auto-detect”, and simply set the genome to the hg19 genome, that’s the human genome version 19. And then click on start to begin retrieving the data. You may now close this window. Depending on the size of the data it can take several minutes to several hours to complete.

Our dataset, comes from the whole genome sequencing of the subject known as NA12878, a human female, a Utah Mormon with Northern and Western European ancestry. This data is from the 1000 genome project and was generated on an Illumina Hi-Seq Sequencer.

For demonstration purposes, the input file, the fastq file, contains only partial data, those reads mapping to a subsection of chromosome10. In addition, the original data was paired-end and we will only use single-end for the purpose of this demonstration.

The whole raw data set is available at the1000 genome website:
http://www.1000genomes.org/data

Once the upload status turns green, it means the upload is complete. You should now be able to see the file int he Galaxy history panel on the right. The dataset will have a very long name, as it's named after the full URL we retrieved it from. Optionally, once the file is in your History, click on the pencil icon to the right of the file name, then select the Name box and you may change the name to anything you wish or just remove the entire URL and keep the filename. Then click ‘Save’ and the filename will have been edited.

To view the file click on the icon that looks like an eye. This file contains our input reads. It has a .fq which is a fastq extension. Let us begin to understand what a fastq file looks like. These files are the standard output of the Illumina sequencer.

A FASTQ file contains four lines per sequence entry. The first line begins with an '@' character and is followed by a sequence identifier and an optional description. Line 2 is the raw sequence data. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier found in Line 1, although not always, as in the case here. And line 4 encodes the quality per base for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A quality value represents the probability that the corresponding base call is incorrect. This value is encoded in ASCII text in orderto save storage place on computers. There are three different options for fastq file formats in Galaxy: one is fastqillumina, fastqsanger and fastqsolexa. They each differ in the way the qualities are encoded. Older data from Illumina maybe in fastqsolexa format, newer data from illuminais generally in fastqillumina and data downloaded from GEO will generally be in fastqsangerformat. However, Galaxy usually is able to guess based on the data what quality-encoding was used for your data.

Once we have obtained the data, the first step is to verify the quality of the data. In the left-hand panel, click on or under the FASTQ Quality Control. It's already open. And we can now click on the FASTQC read quality reports. Here you will choose the file which is already chosen from our history of the FASTQ file that we are now interested in analyzing. Once we have chosen the correct file we canpress 'Execute'.

OK... You will now note on the right-hand side in the history that these are currently running... You see a clock next to them. When they have completed they too will turn green.

The FastQC program aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware of before doing any further analyses.

As you can see, they have turned green in our history so we now we can begin by looking. They have it as a 'fastQC on data 4 as a web page'. Click on the eye. This would be the output file.
Let's take a detailed look at the FastQC report. On the left is a summary of the quality control categories that are evaluated by the FastQC program. A green check indicates the data passed these QC. A red X or an orange exclamation mark would indicate cause for concern.

So let's begin looking at each category.

We have the 'Basic Statistics' which includes a filename, what file type, the encoding; it detected the encoding of the file; how many sequences overall, there are 230,282 sequences. If any of the sequences has been flagged by the illumina machine as poor quality that would be indicated here, and followed by the sequence length, how long are

each read. Here each read is 148 bases. Now if there was a range or a variety of lengths it would be indicated here with a range from the smallest to the largest, but because they only indicate one length that means that all the sequences are the same length in this file. The percent GC is the overall GC-content of the file.

The next category is the ‘Per Base Sequence Quality’. This view shows an overview of the range of quality values across all bases at each position in the FastQ file.

For each position in the read a BoxWhisker type plot is drawn. The elements of the plot are as follows:

- The red line that you see is the median value in that position.
- The yellow box represents the inter-quartile range between 25 to 75 percent.
- The upper and lower whiskers represent the 10% and 90% points.
- The blue line that’s snaking through across all the positions represents the mean quality at each position.

The y-axis shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into: very good quality scores, that would be in the green area; calls of reasonable quality, that would be in the orange area; and calls of poor quality in the red area.

Now just so you know, in general almost all the sequencing technologies the chemistry degrades with increasing read length, thus it is quite often that we see that the quality of the reads will degrade as the run progresses. So we may see base calls falling into the orange or red area towards the end of a read. And in order.. You can trim off that part if the quality does degrade below a reasonable level. Generally the acceptable rule of the thumb is that if the read is above 20 that is considered an acceptable quality for the read.

The next category would be the ‘per tile sequence quality’. Now for each read we know their x,y coordinate on the tile, on the illumina chip that they are actually sequenced on and occasionally you will have an area of the chip or tile that has bad or poor imaging or some kind of other problem. And this allows us to see if there is any area on the tile that perhaps because of where the sequence was sequenced from might have caused the problem. As you can see here everything looks - the quality of this tile looks good all over.

The next category is the ‘Per Sequence Quality Score’. Here we are looking at the overall quality score over the entire sequence. This just gives us an idea if there are a subset of your sequences have a low overall quality value. It could be because they were poorly imaged, like they were maybe on the edge of the tile or things like that. But we should see that if there is such a group of reads that have an overall low quality that it’s really the minority and that most of the reads as you can see here, have a high overall quality of 36-37. And that is the mean sequence quality score.

Next up we’re looking at the ‘Per Base Sequence Content’. This looks at each position in the read and what’s the proportion of each base at that position. Now, here you see an overall flat line at each position which indicates that the proportion is overall evenly distributed or really that the distribution reflects the genomic base pair content of the genome.

Here for example... I will show you an example where we will see sequencing reads that still contain a primer. And in this case you would note that the unusual distribution of the base pair content at the beginning of the read indicates that there's still primer sequence attached to the sequence. In this case before alignment we would need to trim the sequence in order to be able to align them. And hopefully after trimming we would see a similar distribution to our, to the sequences that we have in our current library.

OK. Following that we have the ‘per sequence GC content’. Again, this is measuring the GC content across the whole length of each sequence and it compares it to a modeled GC theoretical distribution. Now, in a random library you would expect to see a normal distribution such that the peak would correspond to the GC content of the underlying genome so if you would notice that's a an unusual shaped distribution that could indicate maybe there's a contamination in your library or maybe there's some kind of bias subset of sequences that are causing an unusual distribution.

Next we have the ‘per base and content’. So, if a sequencer is unable to determine with sufficient confidence what base it is, it'll replace that base with an N, and so we would hope that there are a few or no Ns in our library, as we can see here, it looks across all positions for ends and finds zero percent Ns across our library.

Next we have the ‘sequence length distribution’. Illumina sequencers generate uniform sequence lengths but other sequencers can generate wildly varying lengths and you know

within libraries that are uniform lengths there are pipelines that will trim poor quality bases so you may get a distribution of sequence lengths; Here all the sequences are 148 bases.

Following we have the ‘sequence duplication levels’. This is to give us an idea of how diverse our library is, and in a properly diverse library, as you see in this case, most of the sequences will be toward the left of the graph. In the context of over enrichment or contaminations you might see Peaks coming on at the on the right hand side of the graph.

Next, FastQC looks for over-represented sequences. In our case there are no over-represented sequences. Now, generally in a normal high throughput library which contains a diverse set of sequences generally there are no individual sequences that are overrepresented. And over-represented sequence could indicate either some highly biologically significant finding or it may indicate that your library is not as diverse as you expected or there some kind of contamination. Again, important information to know.

Finally, using a list of commonly used adapters, FastQC checks to make sure that your library is free of adapter content. Generally after FastQC if the reads were found to contain some kind of primer sequence or low-quality bases we may first need to trim or remove some of the sequences... But once we've established that the sequence quality of the library is adequate, we may now proceed to the next stage aligning the sequences to the genome which we will visit in the next segment.