

Unix, Perl, and Python

Session 1: Introduction to Unix and LSF

Exercise 3

Goal: Check the quality of a set of Illumina (Solexa) short reads and remove poor-quality reads.

Note: Commands are in the `Courier` font and sometimes appear over multiple lines. Each command should be executed as a single line on `tak`.

1	<p>The X Windows System provides a graphical user interface (GUI) for computers. Mac comes with the program X11, an X windows program. If you are using Mac, open terminal with Go->Utilities->Terminal, and login to <code>tak</code> with <code>ssh -Y username@tak</code> -Y option will allow you to use X windows. For the PC, we use two software products: Xming and PuTTY. Xming is an X Window Server, which can display <code>tak</code> GUI programs on your local desktop. PuTTY is a free SSH client, which allows you to connect to <code>tak</code>. X server (Xming) must be started before setting up a SSH connection with PuTTY. Refer to the website on how to connect your PC to <code>tak</code>: http://jura.wi.mit.edu/bio/education/docs/ssh-sftp.php</p>
2	<p>Download the <code>H3K36me3.fastq.gz</code> from the class site with <code>wget</code> <code>http://jura.wi.mit.edu/bio/education/hot_topics/Unix_Perl_Python/ex/H3K36me3.fastq.gz</code></p>
3	<p>Decompress the downloaded file: <code>gunzip H3K36me3.fastq.gz</code></p>
4	<p>Check the file size. <code>ls -lh H3K36me3.fastq</code> The <code>-h</code> option prints file sizes in more human-readable format Note: This is just 1% of a complete Solexa file.</p>
5	<p>Check the read quality with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/): <code>bsub fastqc H3K36me3.fastq</code> FastQC creates a web-style report called <code>fastqc_report.html</code> in the folder <code>H3K36me3_fastqc</code>. Navigate to your current directory via <code>\\wi-files1</code> Open <code>fastqc_report.html</code> in your browser and check out the data quality. Pay special attention to the graphs under “Per base sequence quality” and “Per sequence quality scores”. What do these graphs show? What might you want to do as a result?</p>
6	<p>Clean the reads with low base-call quality using a quality filter tool from the FASTX Toolkit</p>

	<p>(http://hannonlab.cshl.edu/fastx_toolkit/)</p> <p>One possibility it to keep a read only if it has a quality score of 20 (-q 20) over at least 95% of its length (-p 95). A peculiarity of this program is that it doesn't adjust to different quality score formats, so you need to specify that your quality scores are Sanger style (-Q33).</p> <pre>bsub fastq_quality_filter -q 20 -p 95 -Q33 -i H3K36me3.fastq -o H3K36me3.trimmed.fastq</pre>
7	<p>Check the filtered reads with fastqc again.</p> <pre>bsub fastqc H3K36me3.trimmed.fastq</pre> <p>Check out the new FastQC report (fastqc_report.html in H3K36me3.trimmed_fastqc) to check that the quality of your reads has improved.</p>