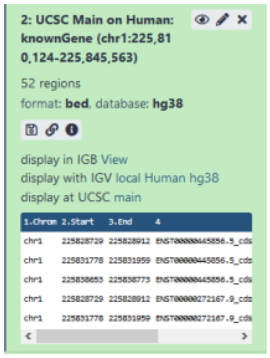
Galaxy

Galaxy is an open-source, easy user interface tool that integrates both command line computational tools with database information to allow for fast/simplified access to genomic analysis and visualization. Galaxy's biggest benefit is its ability to convert the various forms/files found in different databases into a few much more manageable/manipulable formats, allowing those with less coding experience access to the same tools they might not have had the ability to use. It is maintained by the "Galaxy Community" and to this day allows users to access innumerable tools to aid in genomic analysis. Galaxy was used extensively in the Bioinformatics: Tools for Genome Analysis course, both in the research project and examinations.

1. Beginning steps of learning how to use Galaxy involved much use in between Galaxy and the UCSC Main Table Browser, from which I could retrieve BED/WIG files from the UCSC server to analyze and manipulate on Galaxy. The first retrieval involved RefSeq Genes from hg38, retrieving a lookup from the EPHX1 gene region, examining both the coding exons found in this region as well as known SNPs (commonSNPS147).

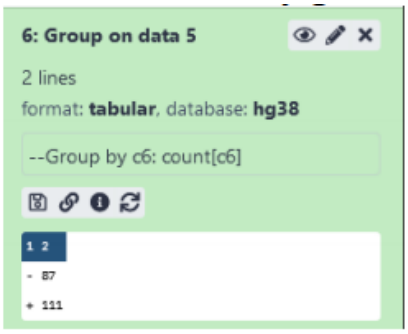


52 coding exons were ID'd by USCS table browser and shown on Galaxy.

202 Common SNPs were ID'd.

2. Next, using both UCSC Main Table Browser and Galaxy, the ENCODE region ENm008 was pulled and the BED file submitted to Galaxy, with a specified interest in the genes that would be found in that region. 198 genes in total were discovered, and the GROUP tool was utilized to organize

those by which strand they were on.



From the Group tool, it was found that 87 genes from the search are on the minus strand and 111 are on the plus strand.

3. Galaxy can also be used to analyze the unique regions in comparing two files by intersecting the two. The workflow titled: 'AG Workflow constructed from history "Unit 7-8 HW"' demonstrates the use of the intersecting too when analyzing two BED files, in this case one containing known exons in a region, and the other CpG island markers in the same region. It was found that in the region of 5q31, there are 542 different exons that intersect a CpG region, and 200 CpG islands that intersect exons. CpG islands are known to be associated with the promoter region of genes, and their association with so many exons in this region could indicate a potential number of genes.

4. For the research project in Bioinformatics: Tools for Genome Analysis course, Galaxy was used on multiple occasions to aid our findings for research into the Human POT1 gene and its variants in order to further our understanding of the possible implications of the rs75932146 SNP. Rs75932146 was reportedly only found in the NM_015450.3 transcript of the POT1 gene, with the POT1 gene having 5 transcript variants. It was important to know the exact location of our SNP in relation to this transcript, as the others had varying amounts of exons. Using Galaxy and UCSC main table browser, it was confirmed that this transcript had 19 exons.

chr7	124822385	124824074	ENST00000357628.8	0	-
chr7	124825251	124825357	ENST00000357628.8	0	-
chr7	124827213	124827305	ENST00000357628.8	0	-
chr7	124829253	124829342	ENST00000357628.8	0	-
chr7	124835278	124835414	ENST00000357628.8	0	-
chr7	124840972	124841178	ENST00000357628.8	0	-
chr7	124842806	124842963	ENST00000357628.8	0	-
chr7	124846941	124846998	ENST00000357628.8	0	-
chr7	124851871	124851951	ENST00000357628.8	0	-
chr7	124852971	124853138	ENST00000357628.8	0	-
chr7	124858956	124859112	ENST00000357628.8	0	-
chr7	124863349	124863640	ENST00000357628.8	0	-
chr7	124870910	124871041	ENST00000357628.8	0	-
chr7	124892265	124892380	ENST00000357628.8	0	-
chr7	124897164	124897212	ENST00000357628.8	0	-
chr7	124898260	124898374	ENST00000357628.8	0	-
chr7	124915573	124915646	ENST00000357628.8	0	-
chr7	124928814	124928999	ENST00000357628.8	0	-
chr7	124929793	124929825	ENST00000357628.8	0	-

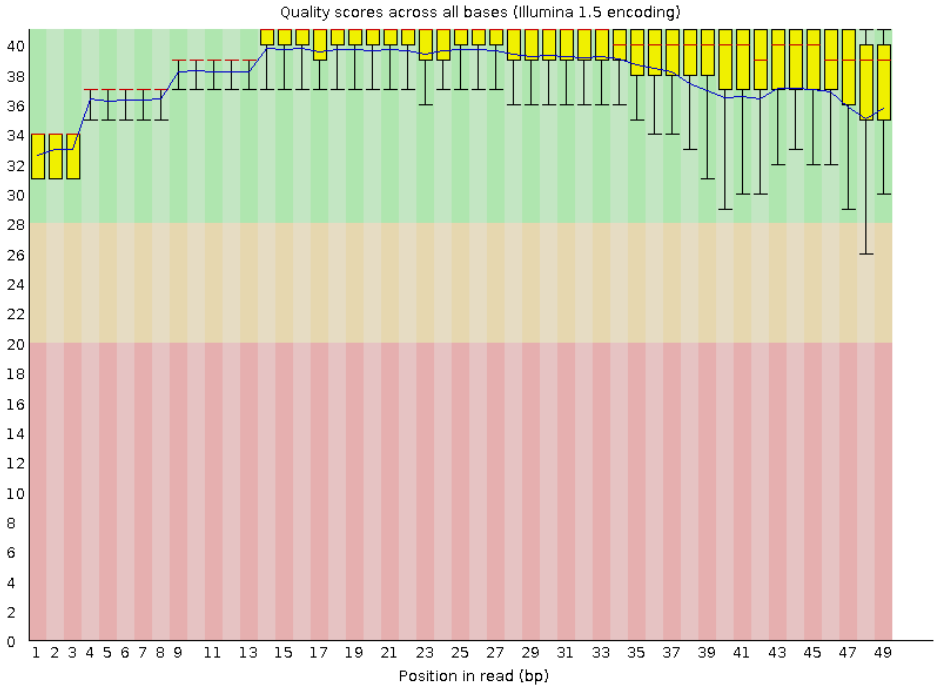
5. Using Galaxy for quality score analysis of NGS reads presented a unique opportunity to work with data that would have otherwise been too "messy" for myself to use. The following is a workflow that was used on multiple occasions to better analyze *C. elegans* reads. First the reads were uploaded to Galaxy in their own history → the reads were then checked for quality using FASTQC → FASTQ Groomer tool used to changed the encoding type to Illumina 1.9 → Trimmomatic used to remove low quality scores from the read → These were then aligned to a reference gnome via the BWA tool.

Example from assignment(see unit 9 & 10 reading summaries):

Part 2

Upload the attached `HW5_Part2_sample_NGS_data.fq.gz` file to Galaxy. This downsampled file is from a NGS experiment on *C. elegans* (genome version WS220/ce10).

a. (0.25 pts) Run **FASTQC** and submit the boxplot of the quality scores. How would you describe the quality of these data?



a.

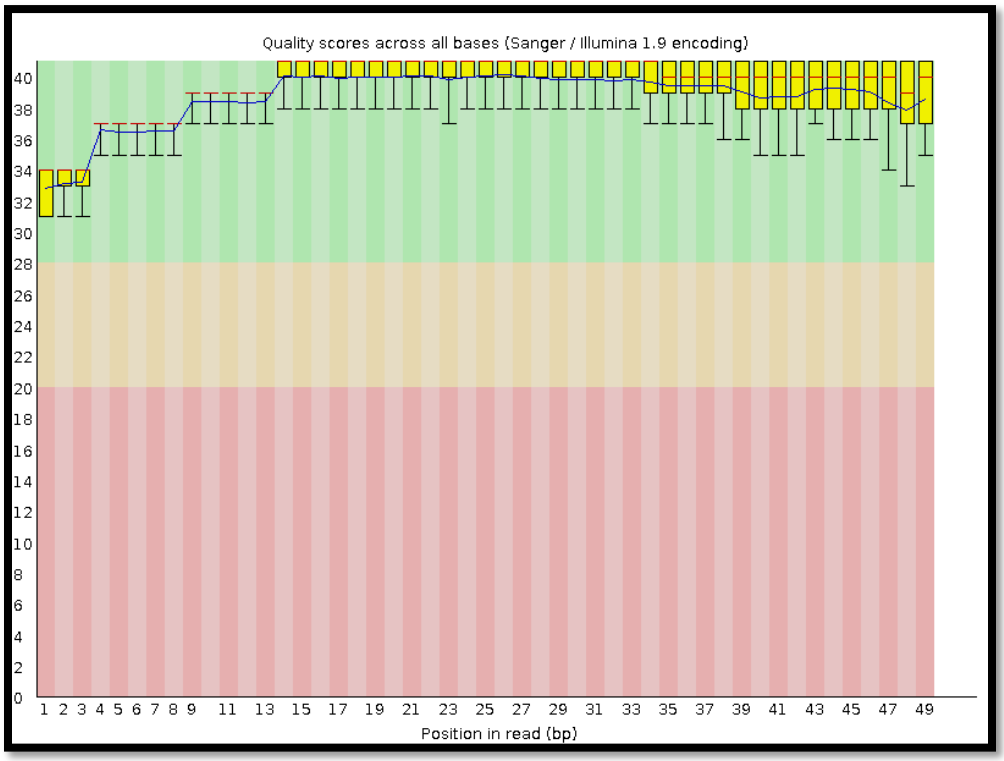
b. (0.25 pts) What phred encoding scheme does this data use? How long are the reads? How many reads are in the file?

Basic Statistics

Measure	Value
Filename	HW5_Part2_sample_NGS_data_fq_gz.gz
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	20000
Sequences flagged as poor quality	0
Sequence length	49
%GC	46

a.
b. Illumina 1.5 encoding scheme
c. Reads are 49bp long

a. From the Trimmomatic results



- A similar method is used in the "Glimmer and Other Command Line Tools" PDF.

There are many further applications of this information that Galaxy can provide the user. After alignment with the reference genome using tools such as BWA, variants can be found in the sequence using the FreeBayes tool and annotated via the VCFfilter tool. Using these tools in conjunction allows the user to see ways in which their own sequence read differs from the reference, and presents an opportunity to identify novel SNPs.

Galaxy is a truly useful tool in the bioinformatic field as a simple method in gathering/manipulating information to analyze specific datasets. My experience with it in the JHU Individualized Genomics and Health program has allowed me to gain experience in working with these larger datasets, and apply that experience to my own research.

Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. *Nucleic Acids Research*. Volume 46, Issue W. Pages W537–W544, doi:10.1093/nar/gky379