Subsequent steps in a typical RNA-Seq analysis would use a tool such as **DESeq2** (an R package) to set up a differential expression analysis to essentially compare the counts of each transcript/gene between different samples (including replicates) to assign a probability to the observed counts being generated if the gene is NOT differentially expressed between conditions. The DESeq2 package will be the subject of a separate class. For now, we will skip this step and move onto a population scale analysis to complete the circle back to our childhood asthma associated SNPs.

***Section 4*: Population Scale Analysis [HOMEWORK]**

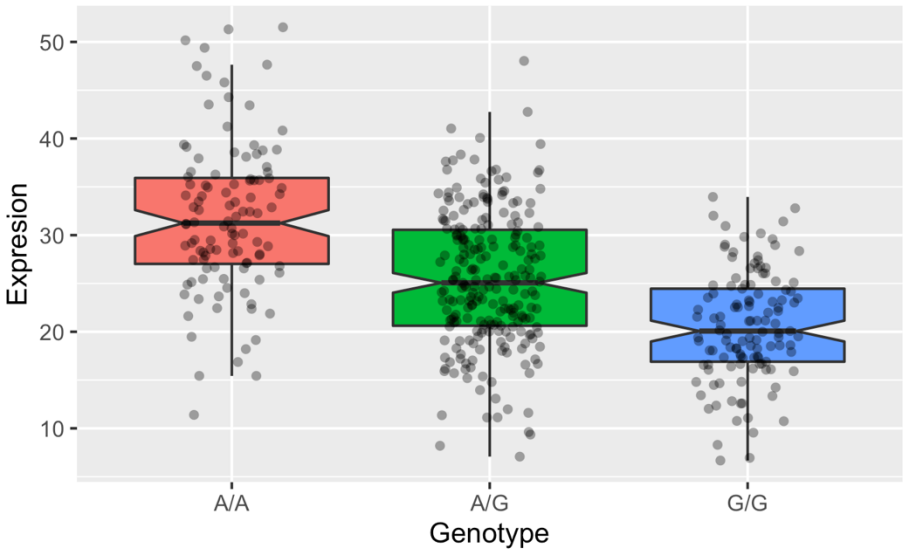
One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (**rs8067378…)** on **ORMDL3** expression**.**

This is the final file you got (https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057.6.txt). The first column is sample name, the second column is genotype and the third column are the expression values.

Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

Q13: *Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.* **Hint**: The **read.table()**, **summary()** and **boxplot()** functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the **boxplot()** function to an R object and examining this object. There is also the **medium()** and **summary()** function that you can use to check your understanding.

***Q14**: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?* **Hint**: An example boxplot is provided overleaf – yours does not need to be as polished as this one.


Expresion
50
40
30
20
10
A/A
A/G
G/G
Genotype

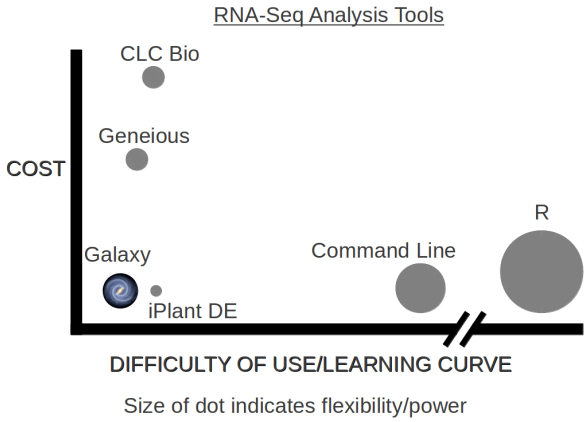
Equipment and Supplies:

This is a bioinformatics tutorial and requires no experimental laboratory equipment or supplies. The workshop could be done with participants bringing their own laptop computers (preferred) or hosted in a computer lab space accommodating 25 participants. Note that if laptops are used it is important that adequate Wi-Fi and power outlets are available. If hosted in a computer lab then guest login access to the computers will be required. No specific software is required beyond a recent web browser (Safari, Chrome or Firefox).

Instructor Notes:

The purpose of this lab session is to introduce a set of tools used in high-throughput sequencing and the process of investigating interesting gene variance in Genomics. High-throughput sequencing is now routinely applied to gain insight into a wide range of important topics in biology and medicine (Soon *et al.* 2013).

In this lab we will use the **Galaxy** web-based interface to a suite of bioinformatics tools for genomic sequence analysis (Afgan *et al.* 2018) Galaxy is free and comparatively easy to use (see Figure opposite for a schematic comparison of some common bioinformatics RNA-Seq analysis methods). Using the public Galaxy server (found at https://usegalaxy.org/) simplifies instructor setup and minimizes the need for dedicated local computing infrastructure.



It is important to note however that using the public Galaxy server, as opposed to a local instance, can result in competition for resources and student jobs being "queued". This will result in variable wait times for job completion that depend on relative server load (i.e. other users demand for resources). This will be most notable during *Section 3 mapping of RNA-Seq reads*. Using the public server, we have observed completion times as little as 20 minutes and as long as one hour for this step. At UCSD we will utilize a custom Galaxy instance for each workshop participant to mitigate potential wait times.

Side-Note: Galaxy was originally written for genomic data analysis. However, the set of available tools has been greatly expanded over the years and Galaxy is now also used for gene expression, genome assembly, epigenomics, transcriptomics and host of other sub-disciplines in bioinformatics.

Description of workshop presentation:

In addition to searching and exploring the major bioinformatics databases OMIM, ENSEMBLE and UCSC participants will use Galaxy's online interface to perform sequence quality assessment, alignment of sequence reads to a reference genome (a.k.a. mapping) and generate a counts table for expressed genes (see: Trapnell *et al.* 2012). All of these steps are performed "in the cloud" using offsite computing resources.

Reference:

- All data files can also be found at: https://bioboot.github.io/bggn213_W19/lectures/#13 Components of Section 2 were adapted from https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise.

1. Verlaan DJ, Berlivet S, Hunninghake GM, Madore A-M, Larivière M, Moussette S, Grundberg E, Kwan T, Ouimet M, Ge B, Hoberman R, Swiatek M, Dias J, Lam KCL, Koka V, Harmsen E, Soto-Quiros M, Avila L, Celedón JC, Weiss ST, Dewar K, Sinnett D, Laprise C, Raby BA, Pastinen T, Naumova AK. Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. **Am J Hum Genet**. 2009 Sep;85(3):377–393. PMCID: PMC2771592

2. Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. **Mol Syst Biol**. 2013;9:640. PMCID: PMC3564260

3. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. **Nucleic Acids Res**. 2018 02;46(W1):W537–W544. PMCID: PMC6030816

4. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nat Protoc**. 2012 Mar 1;7(3):562–578. PMCID: PMC3334321