Université d'Evry Val d'Essonne **Année universitaire 2022-2023**
M2 Data Science **Statistics for genomic data science in health**

TP 1

1. Load the preprocessed data ("Series Matrix File(s)") from the GEO website for the Magrangeas et al (2016) study on neuropathy in multiple myeloma (GSE66903)
2. Briefly describe the goal of the study. How was performed the genotyping ?
3. Use a spreadsheet (eg Excel) to examine the data structure and load the first 1000 rows of the table into R Studio.
4. What are the sample sizes ?
5. For each SNP and for each individual, calculate the call rate. How many SNPs/individuals should we exclude ?
6. Calculate the Minor Allele Frequencies (MAF). Comment.
7. Perform a Hardy Weinberg disequilibrium test. Comment.
8. Can we apply a χ^2 test for assocation (allelic or genotypic) ? If yes, do it.
9. Apply an exact Fisher test (without the fisher.test function and check the result with the fisher.test function).
10. Perform a logistic regression. Is there a significant association ? Give an Odds Ratio estimation and comment.
11. Redo the logistic regression by changing the genotype reference. How do the results change ?
12. Apply a Cochran-Armitage trend test.
13. Compare the different results. Which statistical test would you choose ?
14. Evaluate the population stratification using the genomic control and a principal component analysis.
15. Implement and apply the Bonferroni, Holm and BH procedures.
16. Check the results using the p.adjust function.
17. Display the most significant SNPs and build a Manhattan plot.