Verification test June 9th 2020

1. Explain the different linkage methods you have learnt for hierarchical clustering.

Linkage methods are used to compute the distances between clusters in the context of hierarchical clustering. There are 3 main linkage methods:

- **minimum or single**, where the distance between two clusters is defined as the distance between the two nearest points within the clusters;

- maximum or complete, where the distance between two clusters is computed as the distance between the two furthest points within the clusters;

-average, most used method. It considers all the samples or genes inside the two groups. The distance is computed as the sum of the distances between all the genes within the groups divided for the number of genes in the first cluster * the number of genes in the second cluster.

2. Illumina beadchip 450K for whole genome methylation assay: what is the chemistry to distinguish methylated or not methylated CpG?

The beadchip 450k is an array (a platform) from Illumina to compute the level of methylation of the DNA. After the DNA extraction, the bisulfide conversion must be performed. It is a chemical reaction that converts the all non-methylated C in U. After this procedure, the PCR can be done in order to amplify our DNA.

There are 2 main chemistry approaches to detect the methylation level: infinium 1 and infinium 2. The main difference is the beads type.

1. Infinium 1: two types of beads, one for the non-methylated DNA and another for the methylated DNA. The probes attached to the beads (50 bp more or less) terminate with A (not methylated bead) and the other with G (methylated). If the hybridization occurs, the level of methylation can be detected performing a single base extention: a complementary nucleotide will be added after the target nucleotide.

2. Infinium II: only one type for both methylated and not methylated CpG. Beads with both probes for methylated and not methylated DNA. The last base of the probes ends one nucleotide before the C or T that we are looking for. According to the base (single base extention) that will be added. We can discriminate the methylation condition by measuring the color of the fluorescence (A and T are red and G and C are green).

3. When is it right to apply Mann Whitney U test to find gene differentially expressed? Give an example of a possible biological question and experimental design.

The Mann Whitney U test is a statistical test that does **not** assume normal distribution. It is similar to the Wilcoxon sign-rank test, however it works on unpaired data. Using the specific formula ($U = n1n2 + n1(n1+1)/2 - R1$) We get the U and we compare it with the precalculated table to obtain the p-value. n1 = number of data from the first sample, n2 = data from the second sample and R1 is the

sum of ranks from the largest group. If my Uis > than the U in the precomputed table, according to a specific alpha, I accept the Null Hypotesis.

It can be used if we don't know the distribution of my data and we use ranks.

4. What is the MA plot? Why and when is it useful?

The MA plot is one of the best plot to look at our data from an "alternative" point of view. It is usefil to understand if our data have to be normalized . The y axes (M) is the log in base 2 of the ratio Cy5/Cy3 (trated/control, my results), while in the x axes (A) the average value of log in base 2 Cy5+Cy3 (treated+control/2), so the average of intensity.

If our data are correctly normalized, the resulting curve is equal to log1, so parallel to the x axes. It means the our data are independent from the average of the intensity (as they should be).

In general, MA plot is a useful tool to verify if there are problems in balancing between red and green fluorescence. It is also useful to look at the variability around the regression line. Since M axes is the ration between Cy5 and Cy3, if data are correctly normalized, the data that are above a given threshold of M (for example 2 fold change) can be the ones up-regulated or down-regulated.

In order to normalize our data, we can perform the intra-array normalization using a global normalization approach (it assumes that the dyes are related by a constant factor --> R = k*G) or/and the inter-array normalization. In the last case we have make replicates or we have to use the same reference on each array. The inter-array normalization can be reached in 3 steps: 1) scaling, 2) cenetering and 3 normalization.