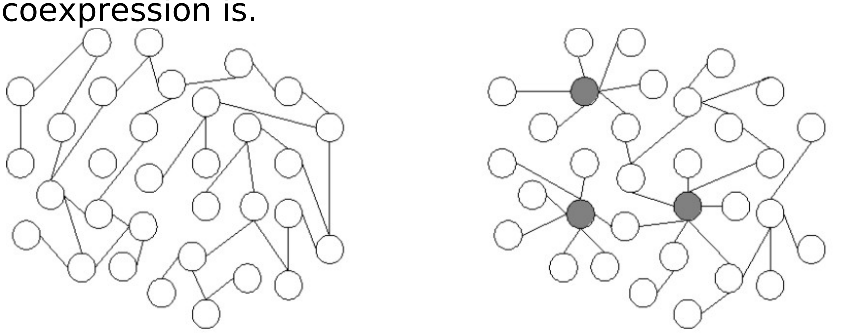
WGCNA MODULE analysis of GENE EXPRESSION DATA

WGCNA is the tool we are going to use in order to carry through this project. Our goal is to find biological hypothesis generated by WGCN (weighted correlation network analysis).

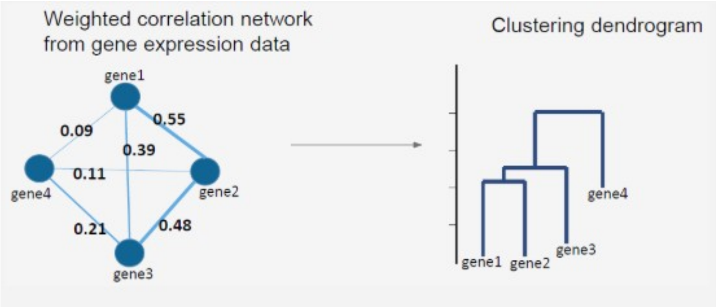
Our project' goal : to find a common biological meaning for a group of genes inside of our COLAUS samples. COLAUS is a gene expression data sampling from the CHUV.
This is real data from real people. The data comes from their LCL which are the lymphocyte cell line.
WGCNA reduces a big data to a few genes of interest. As in the COLAUS set of samples a lot of genes are featured, this is where WGCNA intervenes.
As a metaphor we can use an example, On Zalando, clothes that fit the same body part are grouped in the same category. In WGCNA it is similar, but here no clothes but genes. Indeed genes that have the same expression level are grouped together. In zalando the final few genes are those could be the pieces of clother that also fit your taste and budget. In WGCNA the final few genes are the driver genes, which are the genes that are the most central inside of the gene module and the most coexpressed with our phenotype of interest.
The first way our genes are grouped is inside of a network. A virtual line is created between the genes when the two of them are expressed in the same patterns through our samples. The « line » is weighted, which means it can have various values betwwen 0 and 1 depending on how strong the coexpression is.



(a) Random network (b) Scale-free network

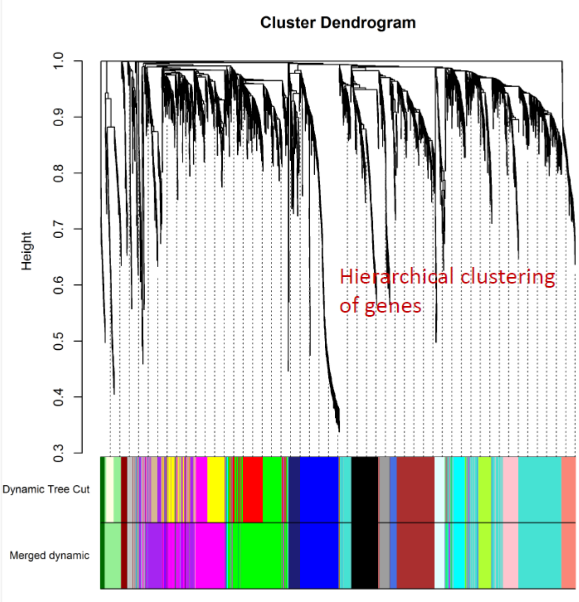
WGCNA uses a mathematical transformation to simplify the process of breaking down the network into modules.
The scale-free-topology is the characteristic of a type of network. As we can see (see figure (b)), this network has few dots with a lot of connection, and many dots with little connection. The network is more readable than a messy, hairball (see figure (a)): Imagine if all the dots would have 5 connections, it would make no sense. None of the dots would be more important than the others and we could not get anything out of it. With the scale-free-network, some genes are more important (the ones with many

connections (Show grey dots)): they are central to the module. With a social example: Imagine a group of people. The one who is friend with everyone has a bigger social influence: he is the “leader” of the group. In order to respect this topology, we adjust the correlation’s modification that has been omitted before.



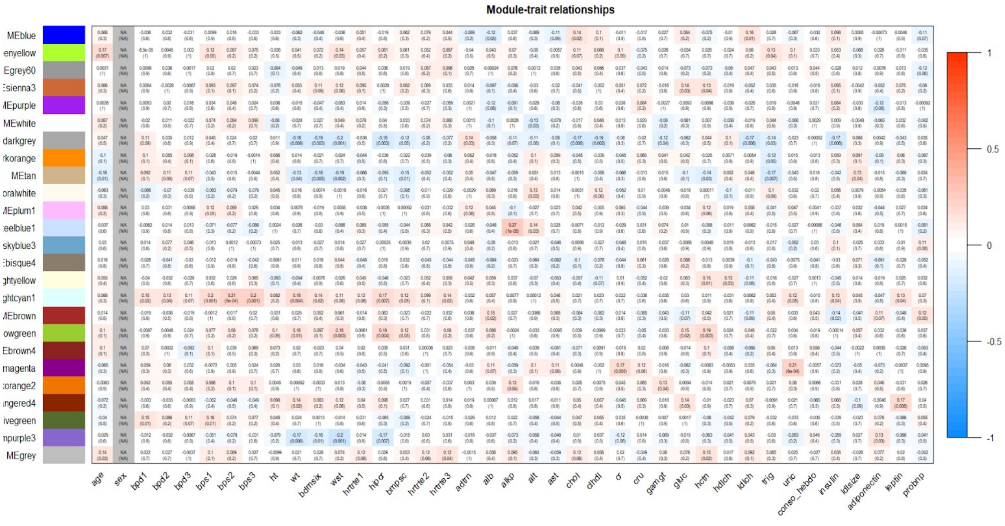
The next step, in order to obtain module, is to pass from a web-network to a dendrogram. A dendrogram is built using the weighted connections, which represent correlations. Let’s see with numbers: the 0.55 correlation between gene 1 and 2 is the highest, so we connect 1 and 2, like a family tree. Then we make it for the second correlation: 0.48 comes next, so we put the gene 3 in connection with 2-3, etc. The lower the genes are in the tree, the more central they are in the module, since wherever we cut, they will be contained in the group.

In our data, we obtain a huge tree with a lot of branches. But there is a problem: this is unusable. In order to make it useful, we have to cut some branches to group some genes. Only after this grouping, we will be able to study the interaction between these genes, and how much they are similar. First, we cut some branches based on characteristics decided by the algorithm. In our case, it is the number of genes: it can be 20 genes per group. We obtain this beautiful colorful line (see picture below). Each colors represents a group. We recreate a dendrogram but this time, we use the groups just created. (This dendrogram is not shown on this slide) In order to further simplify, we proceed to a second cut on this new tree. This time, the algorithm cuts the tree at a fixed height and “group the groups” together. (because yes, remember the first line was group). (Show merged line) As you can see, all the green and red groups are finally just one large green group. Now, the expression of these groups are put in correlation with a phenotype trait.



From the dendrogram shown above, we now organize the data as a matrix (see below). On the left side, we can see the colours from the slide before. They now have a name that we use in the R code. On top of the matrix, there are phenotype traits. In our project, a group of genes is biologically interesting only if its expression correlates with a phenotype of interest. This is one of the ways that WGCNA has to reduce the number of interesting genes to look at. The **grid/table** on the center contains boxes. In every box, there are two value, the top one is a correlation and the bottom one is a p-value. The one that we are going to use the most is the correlation, which show how much the gene module and the phenotype are linked statistically. The colours indicate the value of the correlation and its sign. Blue is a negative correlation, which means that when the phenotype drops, so does the gene expression of this module. If the box is red, it means that the correlation is positive and when the value of the phenotype rises, so does the gene expression. But be careful, correlation is not causality. The phenotype value and the gene expression value (which you cannot see on the table) move at the same time, which means that they are correlated, but if one moves, it does not **cause** the other to move too.

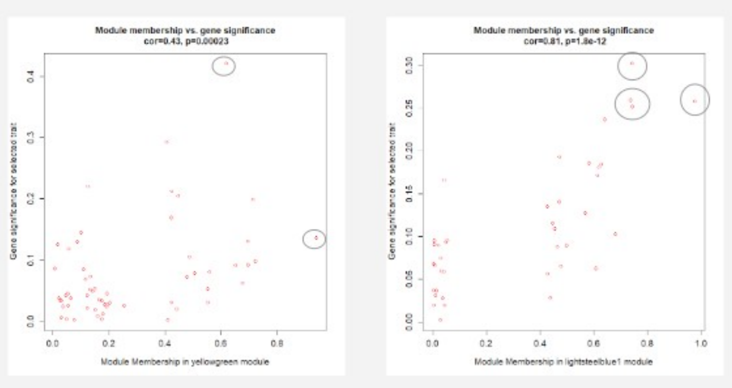
We do not look at the p-value as closely as to the correlation, we just need it to be low enough to be considered significant. Here is our final matrix.



Then, as our next step, we take the module that we chose, the corresponding phenotype, and make a plot out of it. As you notice, on top of the graph is another correlation. This is a different one than the one in the matrix before. What the graph shows is how interesting a specific gene is concerning this phenotype (eye colour ≠ cholesterol level) depending on how much this gene is centered inside of the module (need to explain better).

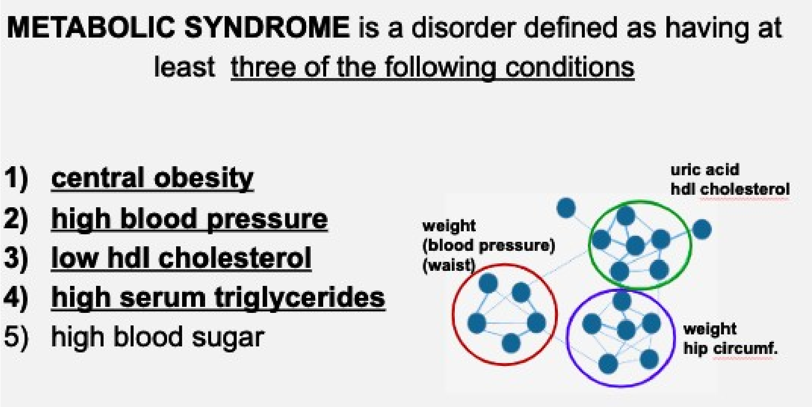
Those graphs show that the genes that are very “centered” also correspond a lot to the phenotype trait. These are the so called “driver genes”. In order to know which module we want to investigate even further, they are going to help us. We are looking for some obvious driver genes, such as in the plot on the right. the one on the left is still usable, but less clear.

The values on top of the graph are again a correlation and a p-value. The correlation is between the average expression value of each gene in each sample and how central it is. The p-value is used to show how reliable that correlation is (not really useful, as for a high number of samples the p.value is always low).



Then, we found three useful hypothesis that we chose to investigate when we researched our phenotypes-module correlations. In order to find those topic, we used gene ontology enrichment technology, which is using an

online database that groups the genes inside our modules into subsets. Those subsets each had a metabolic function, and those are the ones we are going to talk about.
Below is what we found and the common theme :



METABOLIC SYNDROME is a disorder defined as having at least three of the following conditions

1) **central obesity**
2) **high blood pressure**
3) **low hdl cholesterol**
4) **high serum triglycerides**
5) high blood sugar

During our investigations, some phenotypes-module correlations emerged from WGNA that were related to metabolic syndrome, which is defined as having at least three of these conditions.
The underlined ones are the list of disorders/conditions that we found while investigating our phenotypes. We found every conditions except for the high blood sugar in our phenotypes (the correlation between glucose in the blood and the gene modules that were tested in WGCNA was below our threshold).

Our investigations lead us to 3 Hypotheses
Hypothesis 1: Potassium channel could be involved in obesity, based on common interactions.
Hypothesis 2 : We found differents types of prodocadherin and olfactory recepetor in a module “weight” that emerged from WGCNA. Our hypothesis is that the olfactory repceptor genes cooperate with prodocatherin gamma gene in human obesity
Hypothesis 3: in a gene module, we found several types of keratins that were highly expressed. Our hypothesis here is that this specific type might be used as a cancer biomarker.

In more details, here are the explanations :
Concerning the firts hypothesis, we investigated the module darkmagenta, with the phenotype trait uric. Uric acid is a marker of metabolic syndrome and inflammation. It has been associated with the hypertension and cardiovascular disease. Some of the main genes from my module were coding for potassium channel, such as **KCNU1**, **KCNF1** and **KCA10**. Therefore, I investigated the possible link between potassium transport and obesity and according to this publication, it turns out there is one.

The expansion of white adipocyte tissue depends of cell proliferation, which itself depends of the ion channels functionality. For the potassium channels, they participate in the cell cycle. **All these genes**, which come from the article, are potassium channels and, as you can see, there are involved in the preadipocyte proliferation.

Since the main genes mentioned before (**KCNU1**, **KCNF1** and **KCA10**) are also potassium channels, we tried to find a link between our genes and obesity. Sadly, no articles mentions any involvement of our genes and the cardiovascular disease.

So we looked at the possible interactions of all these potassium channels. The results are here: almost all the potassium channels from the publication interact with one protein, named *disks large homolog 1* (DLG1). It is structural protein, which play a role in cell proliferation (and many other things that are not related with obesity).
One of our main genes = **KCNF1**, also interacts with DLG1, with its PDZ domain.

With the help of this common interaction, we can infer on the possible implication of KCNF1 in the regulation of white adipocyte tissue.

For hypothesis number 2: “Olfactory receptor cooperate with protocadherin gamma genes in human obesity” I’m going to talk about the prodocadherin and the olfactory receptor in the module called weight. Prodacherin are homophilic cell adhesion protein and olfactory receptor are odorant receptor. Their individually function is not the central point. Their interaction is what matters.

We found an article in the litterature which say : that prodocadherin and olfactory receptor cooperate. A lot of protocaherins variants exist. We found 3 PCDH that are in the article and in the module which is represented by the intersection of the 2 circles. Our PCDH are gamma.

In the article they said that it’s mostly the prodocadherin beta that are cooperating with the olfactory receptor. However, our hypothesis is that the gamma’s one are cooperating too and not only the beta prodocadherin.

Concerning the olfactory receptor, we didn’t have the same that are in the article but our hypothesis is the same. We think that the olfactory receptor we found are cooperating with the protocaherin gamma.

The article said that the cooperation between them leads to genetic predispositions to obesity. It make sense
because the module emerged from WGCNA and correlate with the phenotype weight and theses genes are present in it .

This is the reason why our hypothesis is that the prodocadherin gamma and the olfactory receptor are working together and thes leads to a genetic predispositions to obesity.

And finally, concerning our last hypothesis :
Keratin is usually found in humans in the hair or epithelium.
Keratin 24 is coexpressed with a subset of a group of genes involved in colorectal cancer and could be used as a biomarker for this disease
A study showed that it is part of a group if seven genes, CYR61, UCHL1, FOS, FOS B, EGR1, and VIP that were up-regulated in cancer patients during a study. They cause, linked to other tumorigenesis factors (Wnt, PI3K, MAP kinase, hypoxia, G protein-coupled receptor) a disruption of cellular homeostasis in the intestinal mucosa.
Those factors are interesting because in our module, we actually find a lot of G-protein or G protein receptor coding genes and as said previously, those two kind of genes are coupled in this cancer
This study means that K24 could potentially be used as a marker in order to identify colorectal cancer at an early stage.
Finding K 24 in this module is really interesting to us because obesity and weight gain is a known risk factor for colon cancer, and we chose this module to investigate because it correlated highly with weight and hip circumference.

What we have to notice is that k24 is already known for its signalling properties in this case but we did also find a high k39 expression in our gene module. It could mean that k39 is also a signal of colon cancer, as it is often expressed at the same time as k24.
K39 is known nowadays for its cytoskeletal function in human hair. It is also a type I cytoskeletal keratin, as keratin 24. A dysfunction of K39 is involved in splenic diseases, which is also interesting because K24 is also highly expressed in the spleen.
The spleen disease makes the cholesterol and LDL rate rise, which may be seen in weight gain also.

These numerous similarities between keratin 24 and keratin 39 leads us to hypothesize that they both might be used to detect early colon cancer. As only a few studies have been conducted on this topic, this may be worth investigating.

To conclude, in order to go further than we did, here is what is left to be done :

Hypothesis 1: We could test spatially and temporally the interaction between KCNF1 (our main gene) and DLG1 in model organism. We could also knock out the gene and see the resultant phenotype.

Hypothesis 2 : We could test if PCDH gamma and olfactory receptor that are in our module leads to an obesity in model organism

Hypothesis 3: as stated previously, keratin 39 might be a coloncancer indicator, as well as K24. We could use the same mechanism as in the study that was on the slide and test its presence and expression in diseased patients. We could also check if more of the genes or more of the factors overlap with our samples.