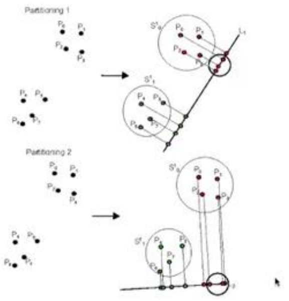
Data analysis – lesson 10

Prof. Calogero

Principal component analysis (PCA)



When I do data reduction, given 10 samples, I have 10 dimensions, but only a small part of the data participates to their visualization. When I do the PCA, I try to find the best projection of the data in order to represent most of the variance; once the 1st one has been found, the others are orthogonal to each other. The 1st dimension is totally arbitrary; for each dataset I use a different arbitrary dimension. The final data reduction will represent some features of the data in the space. Given a dataset, I can represent a subset of the data on one dimension and another subset on another dimension: I transform data in something different in order to represent them in the space. This approach allows to make data compact. In the slide there is a large group of data → I do a random permutation of data: I partition my dataset in many subsets, which are randomly selected → each of the subsets is dimensionality reduced on the basis of a randomly chosen dimension. Many random dimensions are chosen and used to transform data → for each block, I try to put together the data according to their similarity: each representation of data is going to be compared with the others, in order to select the most similar cells. This approach is a sort of clustering, and it is called meta-clustering (I create a similarity matrix using the same data that have been projected on different spaces). This concept is very similar to the kernel: I make multiple projection on data, and in SIMLR I select the projection that better represents the data; in this case, I make a lot of random projections, and I try to find out what are the elements that make the representations very similar to each other. → meta-clustering: I try to put all the clusters together. Since the blocks derive from a unique dataset, they may overlap because some elements are similar to other elements present in other clusters. Finally, I will reconstruct the presence of common overlaps based on the overlapping among clusters. Given 4 clusters, I put all together and obtain a final meta-cluster that resumes the features of each single cluster. I pull together the information coming from different parts of the representation to obtain a full scheme that resumes all the features.

After clustering…

When I do bulk RNA seq, I do dimensionality reduction → differential expression analysis → I obtain a list of genes that may characterize a subset of cells or biological events that I try to explain from the molecular point of view. In scRNAseq, different clusters represent different types of cells. In some cases (such as in a heterogeneous mixture of cancer cells), I don't look at subsets of different cells, but at subsets of the same cell type: the cluster will represent the transcriptional profile that characterizes cancer cells. My aim is to understand why there are new subpopulations and if there are specific markers characterizing them; they are also used to highlight the most important genes that characterize and differentiate clusters and to isolate specific subsets of cells from the overall context. The final set of genes can be used for cell sorting.

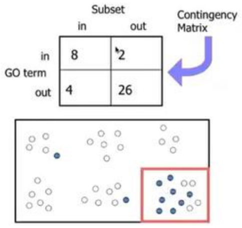
Detection of cluster-specific genes

COMET is one of the most effective tools to identify set of genes that change (better discriminate) among clusters. COMET is an extension of the HyperGeometric test that is done in GO analysis. GO is a computer-useful representation of gene functionality: molecular function, biological properties and cellular location. The task of GO is to help in subsetting the list of genes in order to depict the most useful for the evaluation of the biological effects I am studying. For example, I have 300 genes that are differentially expressed → how can I look at the data? An approach might be Omics-net, which creates networks of differentially expressed genes with some up-level modulators or regulators. GO gives me the possibility to select subsets of characteristics that are more related to the biological problem that I am studying; I can subset genes according to features that are important for me during my experiments. To do so, I need a p-value to evaluate the quality of the final results: the p-value derives from the HyperGeometric test.

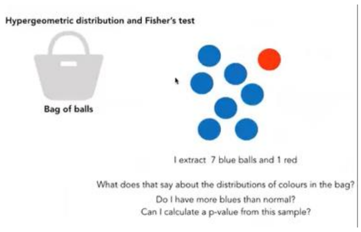
GO ANALYSIS

The dots represented here are all the genes I am studying during a differential expression analysis. The colours represent different GO terms, to which genes belong.

1) Identify the differentially expressed genes.
2) Overlay a GO term to the set of differentially expressed genes, and evaluate if this term is enriched or not in the dataset. Example: I take the blue term, which is present in my cluster of differentially expressed genes and in some non-differentially expressed genes. I need a p-value that represents the statistical significance of what I see. This can be done in an easy way: I have a contingency matrix made of the "subset" and "GO term" variables. There are 8 differentially expressed genes that also make part of the GO term, 2 genes that are part of the GO term only, 4 genes that are not part of the GO term and 26 genes that do not belong to the GO term and are not differentially expressed. Given the matrix, I can apply the HyperGeometric rule present in the slide and obtain the p-value that indicates what is the probability that this event is due to chance. Small probability = the GO term is enriched in my sample. In the slide, the p-value is very small.



		Subset	
		in	out
GO term	in	8	2
	out	4	26



Hypergeometric distribution and Fisher's test

I extract 7 blue balls and 1 red

What does that say about the distributions of colours in the bag?
Do I have more blues than normal?
Can I calculate a p-value from this sample?

How does the HyperGeometric rule work? I have a container filled of balls; the balls are the genes, and the labels represent the different colours of the GO terms. The bag represents all the genes present in the genome. The 8 genes present in the slide have been extracted from the bag, and I would like to know what is the probability and the p-value of having this output (7 blue balls + 1 red ball) by chance: given a bag of balls, what is the p-value that characterizes this event by chance?

First of all, I need the full representation of the balls and their frequency on the total number of balls. The position of the red ball during the random extraction of the balls is casual. How do I calculate the probability? Let's start from the 1st blue ball: there are 8 blue balls out of 40 → the probability is 8/40. Now there are 7 balls → 7/39 is the probability. The probability to obtain the red ball is 5/33 (because I already extracted the blue balls). However, the probability of this event is identical for any other structure where the red ball is present in different positions during the selection of balls → the overall probability of having 7 blue and 1 red ball is bigger than the probability of the single event, because they are summed. How can I move from a probability to a p-value? I can use a coin, whose size has the same probability to appear. If I throw it for the 1st time: 50% head and 50% tail. 2nd time → 50 and 50 and so on. The representation of throwing the coin twice will have the following outcome: 2 heads, 2 tails, a head and a tail, a tail and a head. To calculate the number of events that give 2 heads, I have to divide the times in which HH occurred by the total number of outcomes (which is 4) → ¼. If I look at the probability of having TH or HT = 2/4.

How do I convert the probability in p-value? Definition of p-value = probability that random chance generated the data, or something else that is equal or rarer. Example: I what to transform the probability of 0.25 into a p-value: I have to sum the probability of the event I am considering to the probability of an equal event (the TT event, which has a probability of 0.25 too). I also have to add something that is rarer than my events, but in this case nothing is rarer → 0.25 + 0.25 = 0.5.

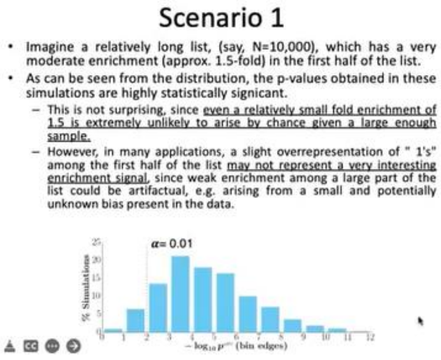
To understand the meaning of "rarer" events, let's make another example: I throw my coin 5 times. What is the p-value to have 4 H and 1 T? 5/32 → equal event: 5/32 (given by 4T and 1 H); rarer events: HHHHH and TTTTT → overall p-value = 0.375.

If I imagine to apply these concepts to the initial dataset, the final p-value is 0.01. I use the probability of the event in order to calculate the p-value in order to see a final output configuration.

What differences characterize the HyperGeometric test with the XL version? The HG test is used to evaluate if there is an enrichment of once in the upper part of a vector: I represent data as a vector of 1 and 0, according to the event being characteristic or not. The representation of 1 and 0 can be completely random → no enrichment. If the "1" events are localized in the upper part of the vector I am considering, there is enrichment. The way I represent data is a binary version of the genes in a cell. Idea: I have my cluster, I take my gene → is this gene more expressed in my cluster with respect to another cluster? 1 if it is, 0 if it is not.

Example: I study TP53 and I have 3 clusters, each one made of 3 cells → my vector is made of 9 elements. The HG test tries to evaluate if the presence of two 1 in the upper part of the representation is significant respect to the other clusters, or if the 1 are there by chance.

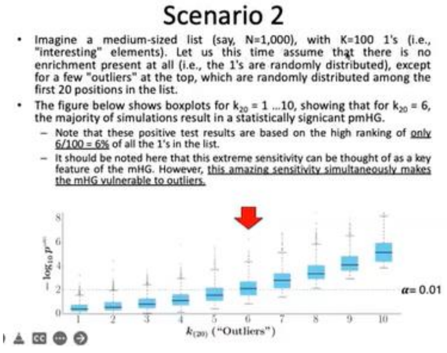
Example: 20 items, with an accumulation of 1s in the upper part of the vector → is this accumulation due to chance of not? The HG test tests all the possible cut-offs to define the p-value, selecting the one able to give the best p-value. The cut-off considers all the genes. The best p-value that I get has a threshold equal to 6: if I consider a block of 6 elements, the representation of 1 will be significantly enriched in the upper part of the vector.



Scenario 1

- Imagine a relatively long list, (say, N=10,000), which has a very moderate enrichment (approx. 1.5-fold) in the first half of the list.
- As can be seen from the distribution, the p-values obtained in these simulations are highly statistically signicant.
  - This is not surprising, since even a relatively small fold enrichment of 1.5 is extremely unlikely to arise by chance given a large enough sample.
  - However, in many applications, a slight overrepresentation of " 1's" among the first half of the list may not represent a very interesting enrichment signal, since weak enrichment among a large part of the list could be artifactual, e.g. arising from a small and potentially unknown bias present in the data.

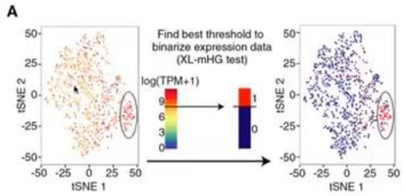
However, in COMET, XL-mHG is used to compensate the errors derived from this kind of representation, and they depend on the scenario: they depend on the domain of knowledge in which I insert the problem. For example, the 1st scenario I see here is characterized by the fact that I have 10.000 N and there is a moderate enrichment (1.5-fold) in one of the first half of the list: in 5000 elements there are more 1s than the other 5000 elements. Is this meaningful in my biological domain? Generally not. The scenario tells me that having an enriched half of data is not meaningful; I can have a significant p-value if my upper part is enriched in 1s.



Scenario 2

- Imagine a medium-sized list (say, N=1,000), with K=100 1's (i.e., "interesting" elements). Let us this time assume that there is no enrichment present at all (i.e., the 1's are randomly distributed), except for a few "outliers" at the top, which are randomly distributed among the first 20 positions in the list.
- The figure below shows boxplots for k_{20} = 1 ...10, showing that for k_{20} = 6, the majority of simulations result in a statistically signicant pmHG.
  - Note that these positive test results are based on the high ranking of only 6/100 = 6% of all the 1's in the list.
  - It should be noted here that this extreme sensitivity can be thought of as a key feature of the mHG. However, this amazing sensitivity simultaneously makes the mHG vulnerable to outliers.

2nd scenario: I have 1000 elements having 100 1s. the fitst 20 values are all 1s: there is a sort of high representation of 1s in the very top part of the vector. How can this be meaningful biologically? The X and L parameters that are used in the test are used to correct this scenario. They make the test more robust with respect to the biases that I can get. L represents the cut-offs of the tests I do (that, as I saw before, are established for each element), and it is the limit number of elements I test: in this case, I establish a cut-off of 4 and estimate the p-value. If, within the elements, there is a significant p-value, this means that there is enrichment. The parameter I decide is up to me. Instead of testing everything, I test 4 elements → if I get a significant p-value I stop here, otherwise I go on. L refers to limiting the cut-off to a specific fraction of the model. X, instead, limits the cut-off to a specific fraction of the model: if I set X = 15, this means that my cut-off is significant in the 15% of the upper part of the model. Then, I test not what happens in the most extreme part of the distribution, but if my p-value is significant or not in the first 15% of my distribution. Playing with these numbers defines the region in which I consider the p-value as significant. If I narrow X very much (6%), I risk to have a big bias. If I extend X to 15%, the p-value becomes not significant anymore. mHG is a powerful test that allows me to look at the enrichment of the top part of the distribution, but it is too biased. X and L parameters allow to moderate the bias. In COMET, L = 3 and X = 15. I look at a relative small top part of the data in order to define my representation. Let's see how the system works.



- An illustration of the binarization procedure applied by COMET to each gene in a cluster-specific manner via the non-parametric XL-mHG test
- For each gene, an expression threshold of maximal classification strength for the given cluster is annotated with the XL-mHG test.
- The XL-mHG P-value measures the significance of the chosen threshold index.
- This threshold index is then matched to an expression cutoff which is used to binarize gene expression values.

On the left there is the log2TPM or log2CPM of the data; the circle represents the cluster. On the basis of the distribution of the expression, I define a cut-off of 8: each gene having a log2CPM = 8 will be associated to 1 and all the others to 0 → I will get many red points in my cluster and a few red points out of it. Using the X and L parameters, I calculate the p-values associated to this partition. Now, I can define 2 parameters: the true-positive probability of the marker associated to my gene and the true-negative probability of the marker not associated to the cluster. I can define the threshold to select the optimal partition that will give the highest number of true + and true - that characterize the cluster and all the rest. If all my cells are red and there is any blue cell: TP=1; TN=0. In other cases, 90% of cells belonging to the cluster are red → TP and the 10% are outside → TP too. Given the prevalence of red cells in the cluster, the system is well balanced. COMET was generated to identify the genes that are characteristic of the clusters I am working with.

In the example, gene A labels my cluster and some other cells outside → gene A is not so good. Gene B labels the cells in my cluster and some genes outside too. If I make the intersection of A and B, my cluster get labelled only. The limit of COMET is the identification of 4 genes that, taken together, are able to select a cluster with high quality. What is interesting is that, when clusters are similar, COMET fails: it tries to find differences, but in presence of similarity they are so tight that they are not found. The system works well in presence of relatively large differences among clusters.

[bulk: select cells that change between a healthy and pathological condition. Using an unbiased approach allows to see also unknown populations.]

Researchers showed that I can obtain a nice separation of cell types looking to a limited number of genes that characterize the different cells. A small number of markers allowed a nice cell separation; the system used to separate cells was COMET.