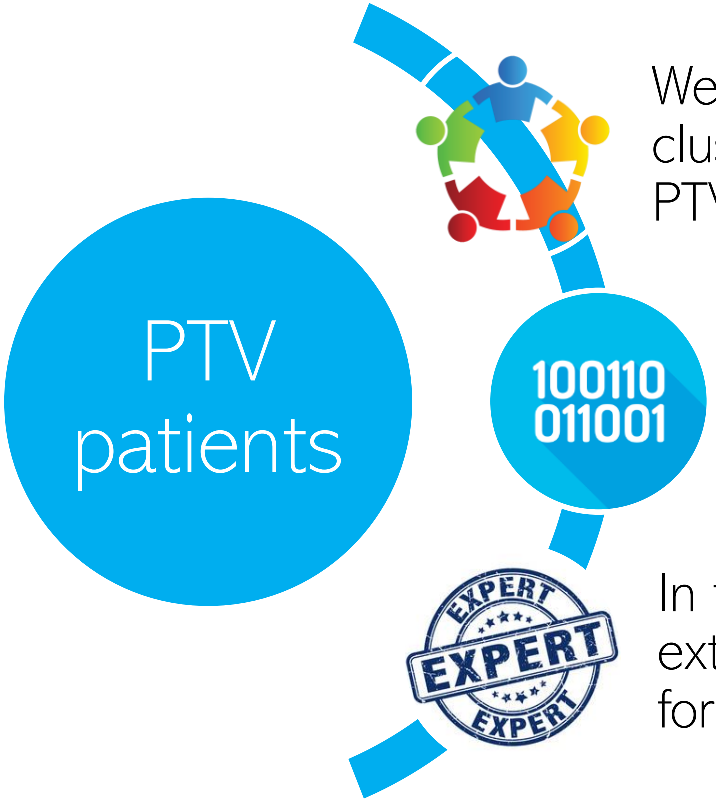
STALICLA R&D

Clustering PTV patients with ASD

Gabriela Martínez
August 2019

Our objectives…



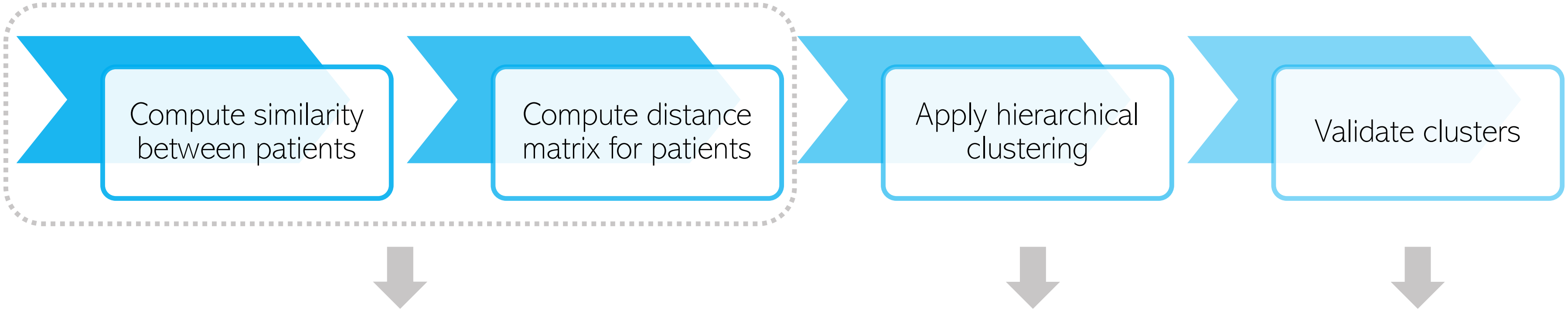
We want to find clusters of similar PTV patients

We are also interested in finding correlated pathways

In the end: we need external validation for these clusters

Roadmap for patients clustering

Statistical results for clustering always need to be validated by experts:



Compute similarity between patients

Compute distance matrix for patients

Apply hierarchical clustering

Validate clusters

This **distance matrix** serves as input for any clustering algorithm and it is the most important step in the pipeline

It facilitates the **visualization** of grouped patients

If this stage fails, clustering needs to be reviewed and recomputed

How are two patients considered “similar”?

The first step for clustering is to define how similarity will be defined.

Jaccard’s distance

Two patients will be “**closer to each other**”, or “**similar**” when they share mutations in the same pathways. For every pathway and pair of patients, the following matrix of similarity is computed:

	0	1
0	M_{00}	M_{10}
1	M_{01}	M_{11}

Unmatching mutations

Common mutations

Unmatching mutations

$$d = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

The more mutations in the same pathways are shared by two patients, the more similar they are!

This measure is a probability: the closer to 1, the more similarity.

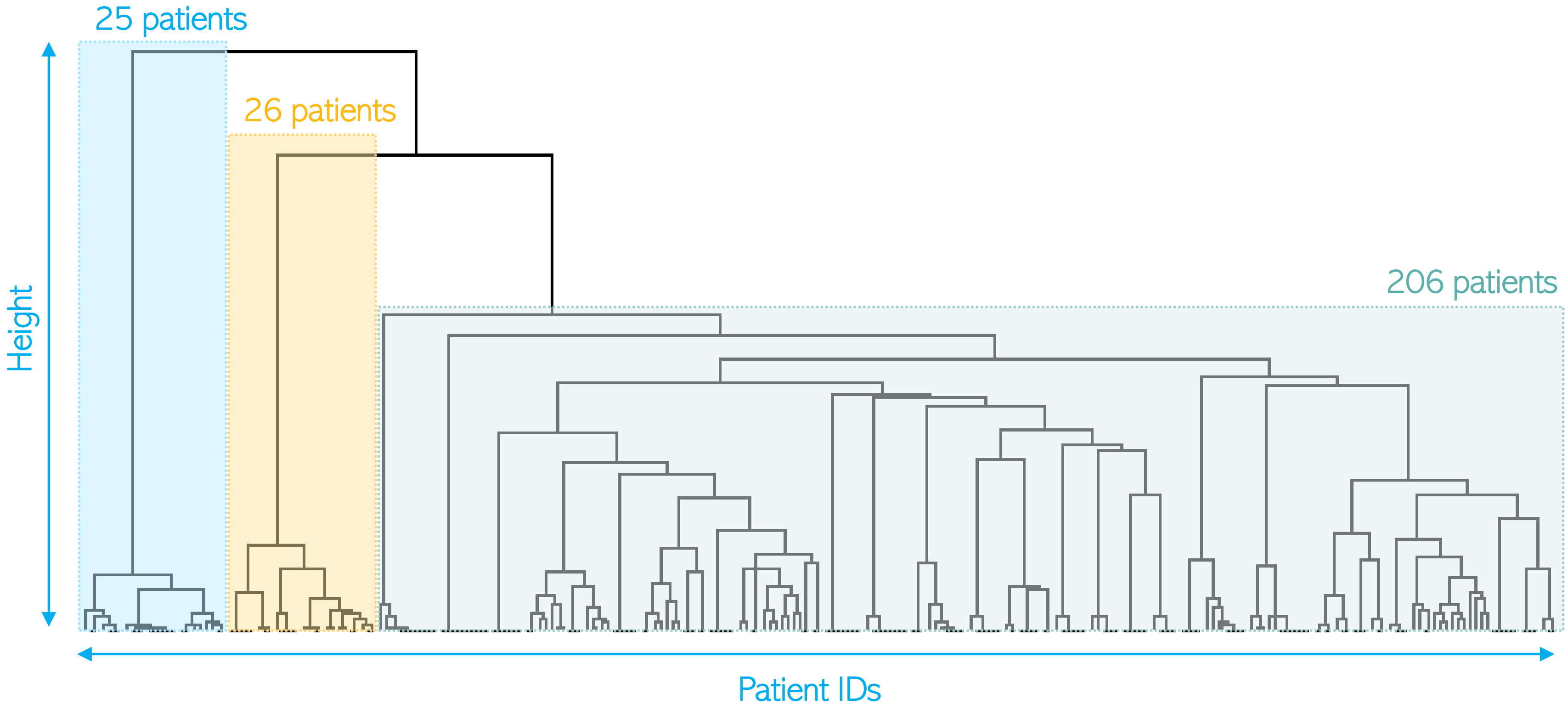
How does our patient similarity matrix look like?

After all computations, we will have a large matrix (257 x 257) of patients and their measure of distance (the closer this number to 1, the more similar these patients are):

	Patient_1	Patient_2	Patient_3	Patient_4
Patient_1	1			
Patient_2	0.95	1		
Patient_3	0.50	0.90	1	
Patient_4	0.30	0.76	0.66	1

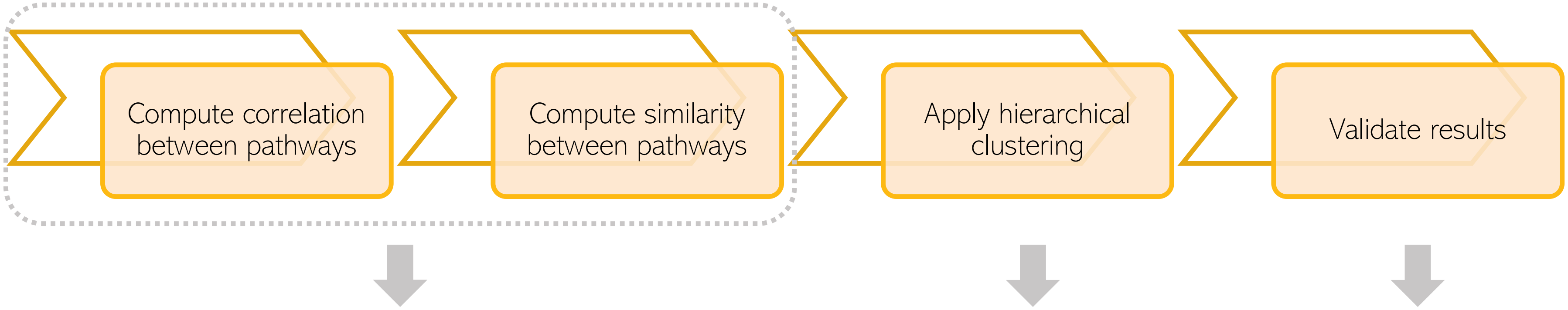
Hierarchical clustering for patients

3 clusters of PTV patients were visually identified. Results are generated for 2,3,4 and 5 clusters in a separate xlsx file.



Clustering pathways is also a good idea

It may be useful to find which pathways are “**similar**”, as they may lead to related patients’ behavior or symptoms.



Compute correlation between pathways

Compute similarity between pathways

Apply hierarchical clustering

Validate results

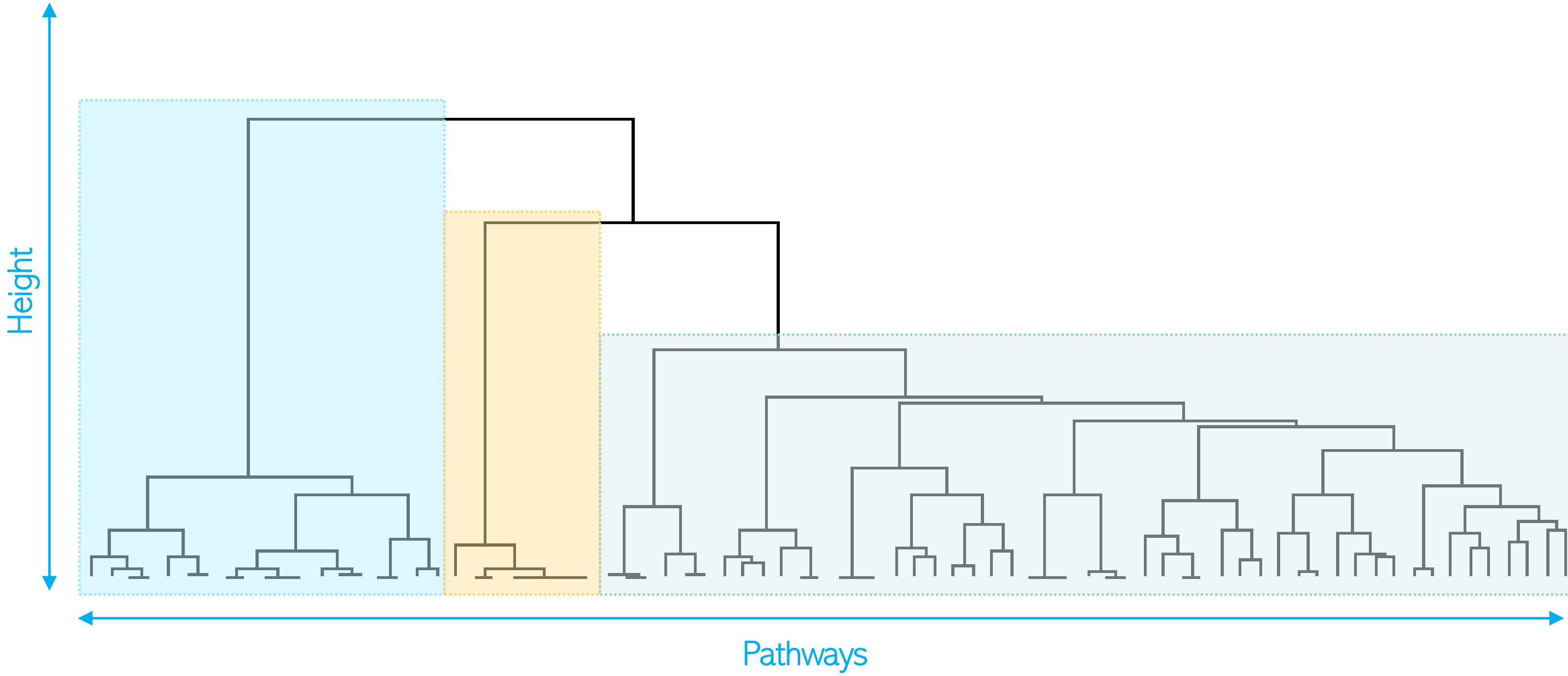
The approach for analyzing similarity between variables (pathways) differs from the one for observations (patients)

It facilitates the **visualization** of potentially akin pathways

Pathways could be grouped to recompute patients’ clusters

Hierarchical clustering for pathways

Dimensionality reduction could be possible by grouping similar pathways into new variables that can improve the natural clustering of the PTV patients.



Conclusions

- Visual examination of hierarchical clustering allowed to explore an initial clue about the natural association of PTV patients.

- The clustering exercise is a cyclic process that needs to be externally validated by knowledge experts before changing the statistical approach.

- Clustering not only patients but also pathways may result in a more powerful feature engineering scope, useful for future clustering processes.

- Other methods, such as k-means and LCA were tested, with almost the same results as the ones obtained with hierarchical techniques,