Analyzing functional data using R Geno ontology analysis
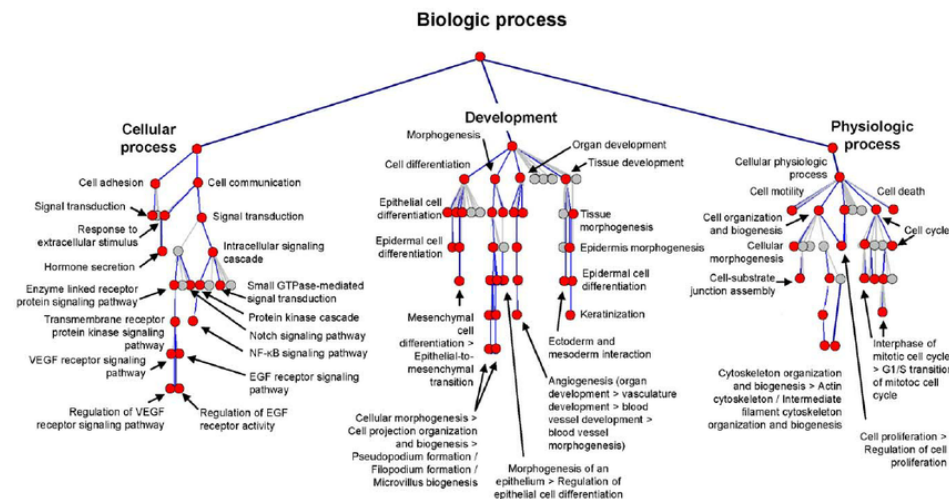
Pablo A.S. Fonseca
05/02/2021

Gene Ontology

Gene Ontology (GO) is a collection of controlled vocabularies describing the biology of a gene product in any organism

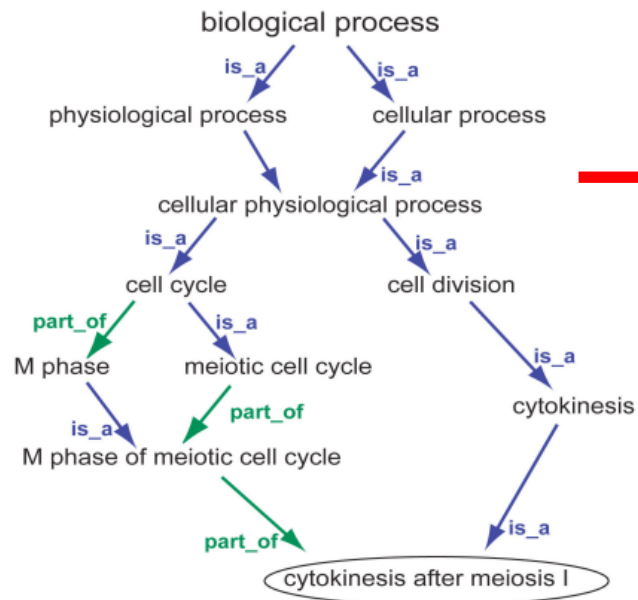The terms are organized in a tree structure

Terms get more detailed as you move down the tree

There are ~25.000 terms divided in three groups: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC)

A gene can be assigned to any of the ontologies (BP, MF and CC) and assigned to several GO terms.

True path rule:
If a gene is member of a term, it is also member of the terms parents

Gene Ontology

This structure is extremely useful to organize and to allows a easy identification of the biological processes related with a gene. Consequently, helping to understand the real function of a gene.

Gene Ontology enrichment

The enrichment analysis is performed to identify if a GO term is related with the candidate genes in list more than expected by chance.

	Our list	Genes in the Genome	
In GO term	N_{11}	N_{12}	$N_{\bullet 1}$
not in GO term	N_{21}	N_{22}	$N_{\bullet 2}$
	$N_{1\bullet}$	$N_{2\bullet}$	$N_{\bullet\bullet}$

H_0: The percentage of "In GO term" genes associated with a GO category is proportional to the number on the genes in the genome associated with the same category percentage on the genome: $N_{1\bullet}/N_{\bullet\bullet}$

expected: $N_{\bullet 1}N_{\bullet 1}/N_{\bullet\bullet}$

observed: N_{11}

test $(O-E)^2/E$ (Chi-squared test)

Important:
Several kinds of statistical tests can be performed. This was only a simple example.