# Ricerca finalizzata GR 2022 – Endometriosi e Cancro Ovarico

## **Piano di analisi dei dati.**

**Fernando Palluzzi, PhD**

# 1. Background and general settings.

From a previous pilot study, three different signatures, of about 100 genes each, were defined:

1. Benign ovarian cyst (NA = 6 RNA-seq samples).
2. Atypical ovarian cyst (NB = 6 RNA-seq samples).
3. Ovarian cancer (NC = 6 RNA-seq samples).

These will be referred to as *prior signatures*.

About 1/3 of the subjects in A share the same clinical/molecular characteristics of C, but not all of them develop a malignant phenotype.

A panel of 1825 genes will be transcriptionally profiled to detect a single molecular signature able to predict the differences among the three groups (primarily signatures A-to-C and, secondarily, B-to-C), and providing a valid mechanism for malignant transformation and cancer outbreak.

# 2. Data analysis: cancer signature detection.

Every signature detection block will be derived from pairwise comparisons, including an experimental group G1 (e.g., group C), a control group G0 (e.g., group A), and their contrast β = log2(G1/G0), referred to as log fold-change (logFC). Each contrast will be evaluated assuming a null hypothesis H0: β = 0, for each gene. A gene with false discovery rate (FDR) < 0.05 is considered as differentially expressed.

For a given contrast (e.g., C vs. A), a set of differentially expressed genes (DEGs) will be calculated, using the R [1] package edgeR [2], based on a negative binomial distribution for (sequencing) count data, and accounting for expression data overdispersion, correlation among gene expressions, and possible deviations from data normality.

The set of DEGs will be compared with the prior signatures of both G1 and G0 groups, and only differentially expressed genes from each prior signature will be retained. The enrichment of DEGs within every prior signature will be evaluated through Fisher’s exact test (with a significance level of 5%). The set of enriched DEGs from every prior signature in a contrast will be merged (i.e., enriched DEGs in G1 prior signature + enriched DEGs in G0 prior signature), yielding a putative molecular signature (PMS) for that contrast.

For each PMS, a structural equation model (SEM) will be generated to represent the covariance network among PMS genes. The covariance structure will be perturbed by an exogenous source, represented by the group variable G = {0, 1} (for instance, group A = 0 and group C = 1). This multivariate linear network will be critical for two key cancer-related features:

1. the causal influence of the phenotype over the observed transcriptional deregulation;
2. the paths of (in)direct perturbation propagation among DEGs.

The former will further assess the PMS, leading to the final cancer-associated signature, and the latter will determine the causal structure (mechanism) underlying malignant transformation. SEM building, fitting, cancer signature and mechanism extraction will be done with the R package *SEMgraph* [3]. This package will also control high-dimensionality, parallelization settings, and fitting heuristics for large-scale genomic models. Possible deviations from normality will be controlled through non-paranormal transformation, with the R package *huge* [4].

# 3. Data analysis: signature validation through random forests.

After the first stage of cancer-associated signatures detection and modelling, a larger dataset (N = 240) will be used to benchmark our ability to predict a given phenotype: disease-free (group A), atypical (group B), and tumoral (group C).

Predictions will be done using random forest classifiers (RFCs) [5], by dividing the input dataset in 5 groups: 4/5 will be used to carry out a 4-fold cross-validation, while the fifth one will be used as an external validation set. The RFC allows to generate bootstrap-derived samples and out-of-bag (OOB) observations: the former will be used to generate independent random trees (classifiers), and the latter will be used as internal on-the-fly validation sets, to generate OOB classification errors. RFC settings include 10000 bootstrap-derived random trees and 3 randomly chosen variables per tree branching (decision).

In addition, the OOB datasets will be used to rank features on the base of classification accuracy, through the mean decrease accuracy index (MDA), and classification entropy, through the mean decrease in Gini impurity (MDG).

RFC performances will be evaluated in terms of: accuracy, F1 score, sensitivity, specificity, positive predictive value, negative predictive value, OOB error, and subject-level Brier scores [6]. Additionally, the top-ranking features will be used to group subjects by similarity, further validating the features (genes) associated with each phenotype.

# 4. Limitations.

The current analysis has two minor limitations. Firstly, the use of a linear multivariate network (SEM) might miss nonlinear interactions that could have a relevant importance on the overall regulatory context. However, this limitation is largely compensated by the use of a machine learning approach for validating and ranking causal inference results. The second limitation could arise due to the limited sample size for the RFC training. The nested cross-validation approach, the intrinsic OOB error estimation, and the convergence with causal inference results (i.e., the ability of the cancer-associated signature to predict phenotypes) should mostly reduce also this limitation, although a larger independent validation set is required in future studies.

# References

1. R Core Team (2021). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Robinson MD, McCarthy DJ, Smyth GK (2010). **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics*, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.
3. Palluzzi F, Grassi M (2021). **SEMgraph: An R Package for Causal Network Analysis of High-Throughput Data with Structural Equation Models**. arXiv preprint arXiv:2103.08332. doi: 10.48550/arXiv.2103.08332.
4. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2012). **The huge Package for High-dimensional Undirected Graph Estimation in R**. *J Mach Learn Res*. 2012 Apr; 13: 1059–1062. PMID: 26834510
5. Liaw A and Wiener M. **Classification and Regression by randomForest**. *R News*. 2002;2(3):18-22.
6. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950;78(1):1-3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

# Further readings

1. Palluzzi F, Ferrari R, et al. **A novel network analysis approach reveals DNA damage, oxidative stress and calcium/cAMP homeostasis-associated biomarkers in frontotemporal dementia**. *PLoS One*. 2017 Oct 11;12(10):e0185797. doi: 10.1371/journal.pone.0185797. eCollection 2017.
2. Benussi A\*, Grassi M\*, Palluzzi F\*, Koch G, Di Lazzaro V, Nardone R, Cantoni V, Dell'Era V, Premi E, Martorana A, Di Lorenzo F, Bonnì S, Ranieri F, Capone F, Musumeci G, Cotelli MS, Padovani A, Borroni B. **Classification Accuracy of Transcranial Magnetic Stimulation for the Diagnosis of Neurodegenerative Dementias**. Annals of Neurology 2020 Mar; 87(3): 394-404. PMID: 1925823. DOI: 10.1002/ana.25677. \* These authors contributed equally.