ISREC@EPFL
INSTITUT SUISSE DE RECHERCHE EXPÉRIMENTALE SUR LE CANCER

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

PROGRESSION ANALYSIS OF DISEASE APPLIED TO BREAST CANCER RESEARCH

Rachel Jeitziner

20.06.2015

Introduction

Introduction

- During the past few decades, science has made tremendous progress, and the latest technological tools allow assaying tens of thousands of genes simultaneously.
- Hence, we get large volumes of data to search for cancer biomarkers.
- How can we analyze such tons of data ?

How can we analyze such tons of data ?

- **Or** how can we extract some qualitative signal form this data ?
- **And** extract information that is robust in the presence of noise and errors.
- **And** still catching the important features of the data.

Solution… Topological Data Analysis!

- Reduce the complexity of the data.
- Extract some qualitative information.
- Can describe the notion of closeness from one special "**point**" to a "**set**" (for example a gene expressions from one person to the gene expressions of many others, specially a diseased genome to a group of microarrays from healthy tissues).

Solution… PAD

- The method is called Progression analysis of disease (PAD) and is composed of two parts
  - Disease specific genomic analysis (DSGA)
  - and Mapper.

Interesting Articles-DSGA and PDA

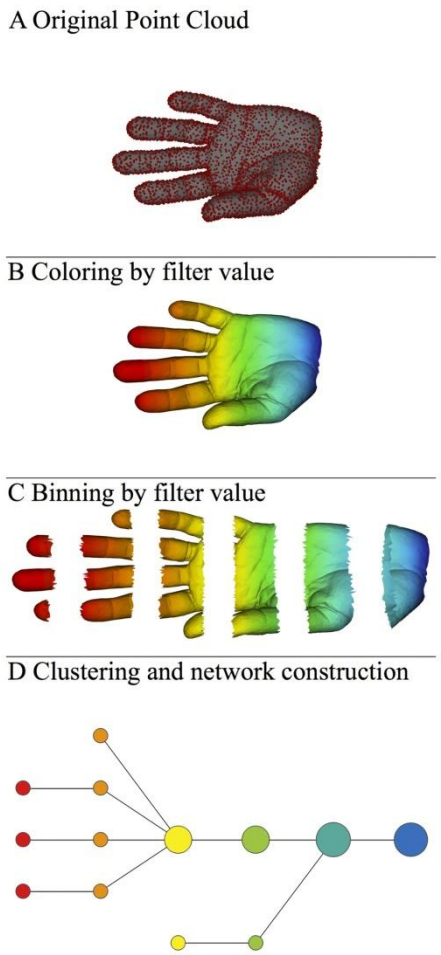
- M. Nicolau R. Tibshirani, A.-L. Borresen-Dale, S. S. Jeffrey, et al., *Disease-specific genomic analysis: identifying the signature of pathologic biology,* Bioinformatics Vol. 23 no. 8 (2007).
- M. Nicolau, A.J. Levine, G. Carlsson, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival,* PNAS Systems Biology vol. 108 no. 17 (2011).

Progression analysis of disease method

MAPPER

- One has to choose a « magnitude function » called f: X → Z, where Z is a reference space. Here, Z is

- and a « distance function » d, which is here the euclidean distance in $\mathbb{R}^3$.
- $x \sim x'$ if $d(x,x') < \varepsilon$.
- This is not yet an equivalence relation, but becomes one under transitive closure, i.e., if $x \sim x_0$ *and* $x_0 \sim x_1$, then we set $x \sim x_1$.
- The algorithm consists in creating ”clusters”, by taking the equivalence classes.



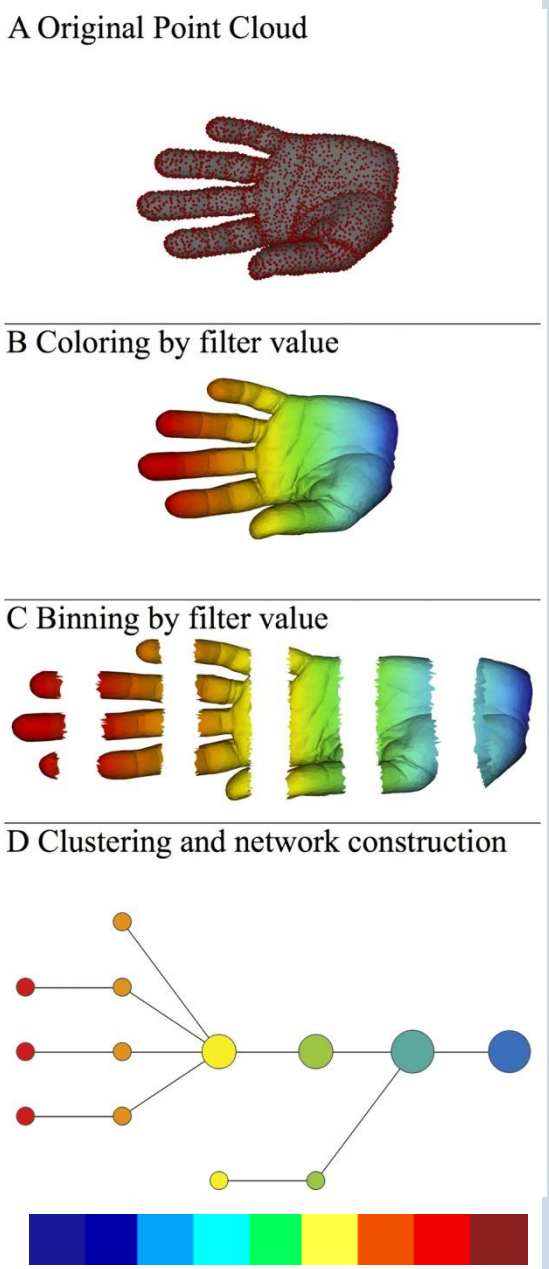
Gunnar et al., *Extracting insights of complex data using topology*, 2013

Mapper

- Choose a covering $U = \{U_\alpha\}$ of Z.
- One applies this clustering to

$$X_\alpha = f^{-1}(U_\alpha).$$

- Take equivalence classes on X_α.
- Hence, we create for every (α,c) a vertex.
- $\{(\alpha_0, c_0), (\alpha_1, c_1), \ldots, (\alpha_k, c_k)\}$ spans a k-simplex if and only if the clusters $X_{(\alpha_0,c_0)}, \ldots, X_{(\alpha_k,c_k)}$ have a non empty intersection.



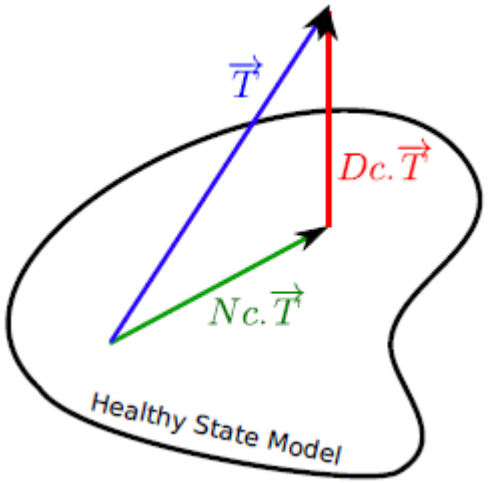
Disease-Specific genomic analysis

- Normal like tissue microarray : **$N_1, \ldots, N_S$**, where **S** is the number of normal samples (S has to be smaller than the number of genes n),
- **GROUP A**
- Diseased tissue microarray : **$T_1, \ldots, T_R$**, where **R** is the number of diseased samples.
- **GROUP B**

Disease-Specific genomic analysis

- We define a subspace $\mathcal{N}$ which represents the normal tissue data

$$\overrightarrow{T} = \overrightarrow{Nc.T} + \overrightarrow{Dc.T}$$



- Now, we have to reduce the size n of genes and every statistical method of size reduction can be used.

METHOD

- **GROUP A** : $N_1, \ldots, N_S$
- **GROUP B** : $T_1, \ldots, T_R$
- T = Nc.T + Dc.T. Hence, one generates : $Dc.T_1, \ldots, Dc.T_R$.
- The magnitude function f: X-> Z, is defined as

$$f(Dc.T) = \left(\sum |g_r|^l \right)^{k/l}$$

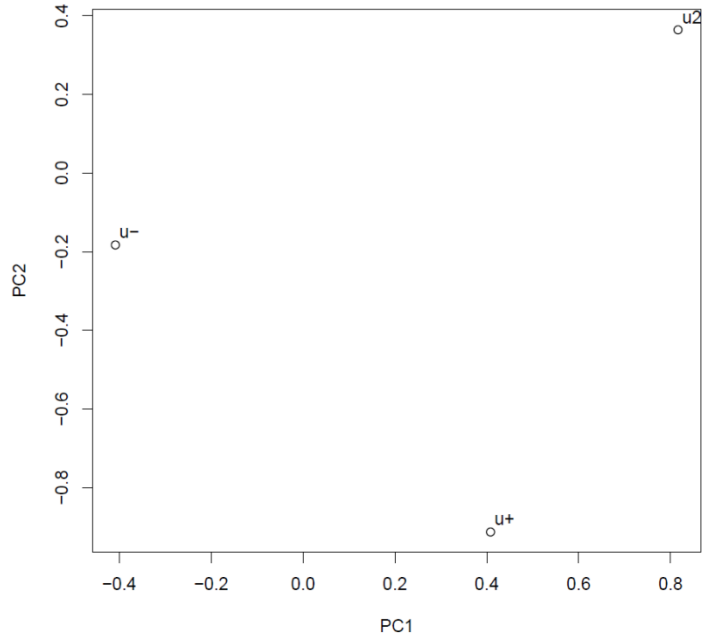
- The distance is the correlation distance

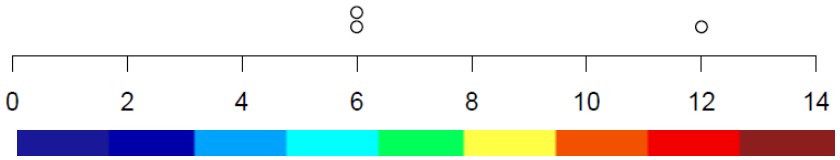
$$d(u,v) = 1 - \frac{(u - mean(u)) \cdot (v - mean(v))}{\|u - mean(u)\| \|v - mean(v)\|}.$$

- Cover ? K intervals with T% overlap.

Toy example : PCA vs Mapper

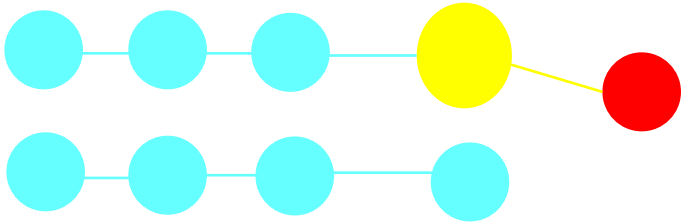
- $u_+ = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, u_- = \begin{pmatrix} -1 \\ -2 \\ -3 \end{pmatrix}, u_2 = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$





$$distance\ matrix = \begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 2 \\ 0 & 2 & 0 \end{pmatrix}$$

5 intervals / 66 % overlap

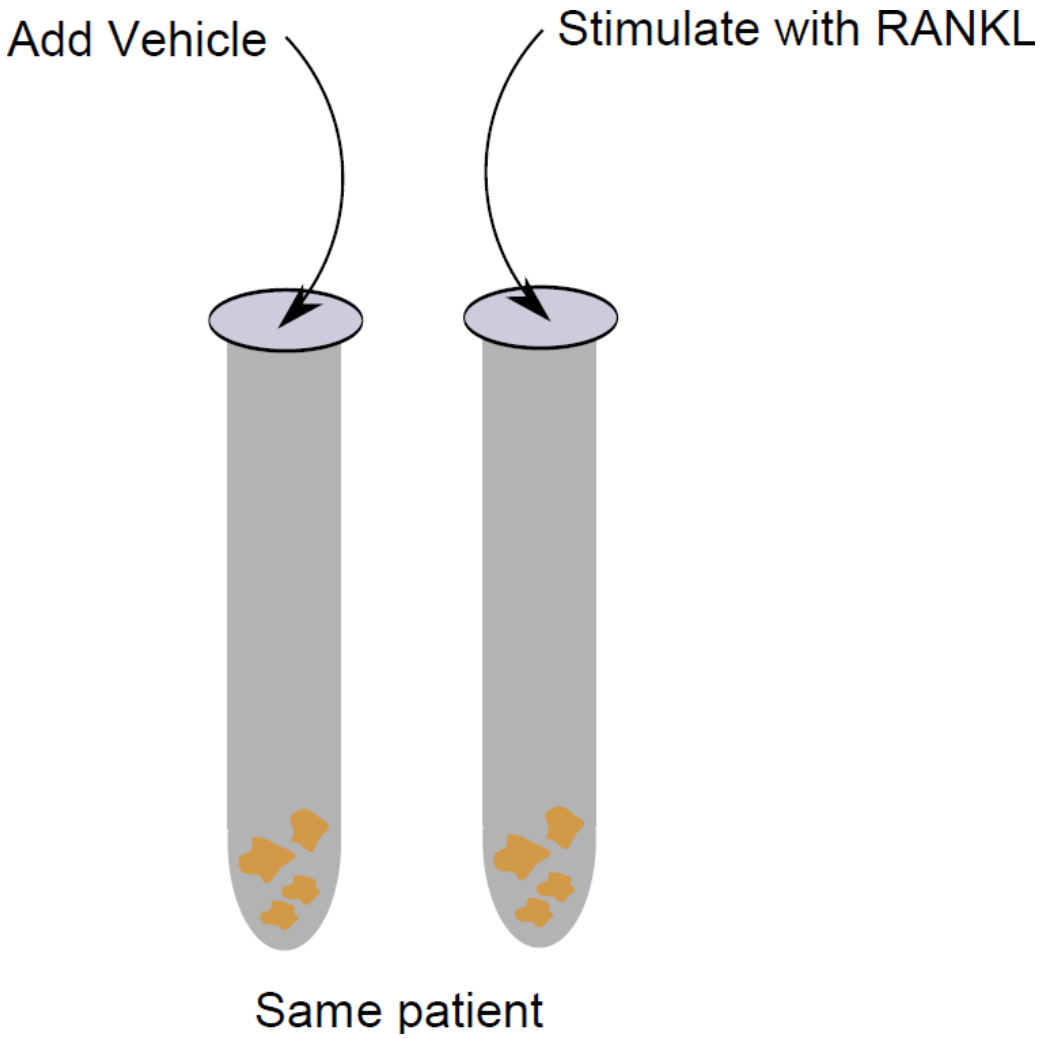


First application of PAD

Background on RANKL

- Receptor activator of nuclear factor kB ligand (**RANKL**) has been shown to be an important protein in mouse mammary gland development and mammary carcinogenesis (M. Beleut et al.)
- Has it the same role in humans ?

Context of the experiment



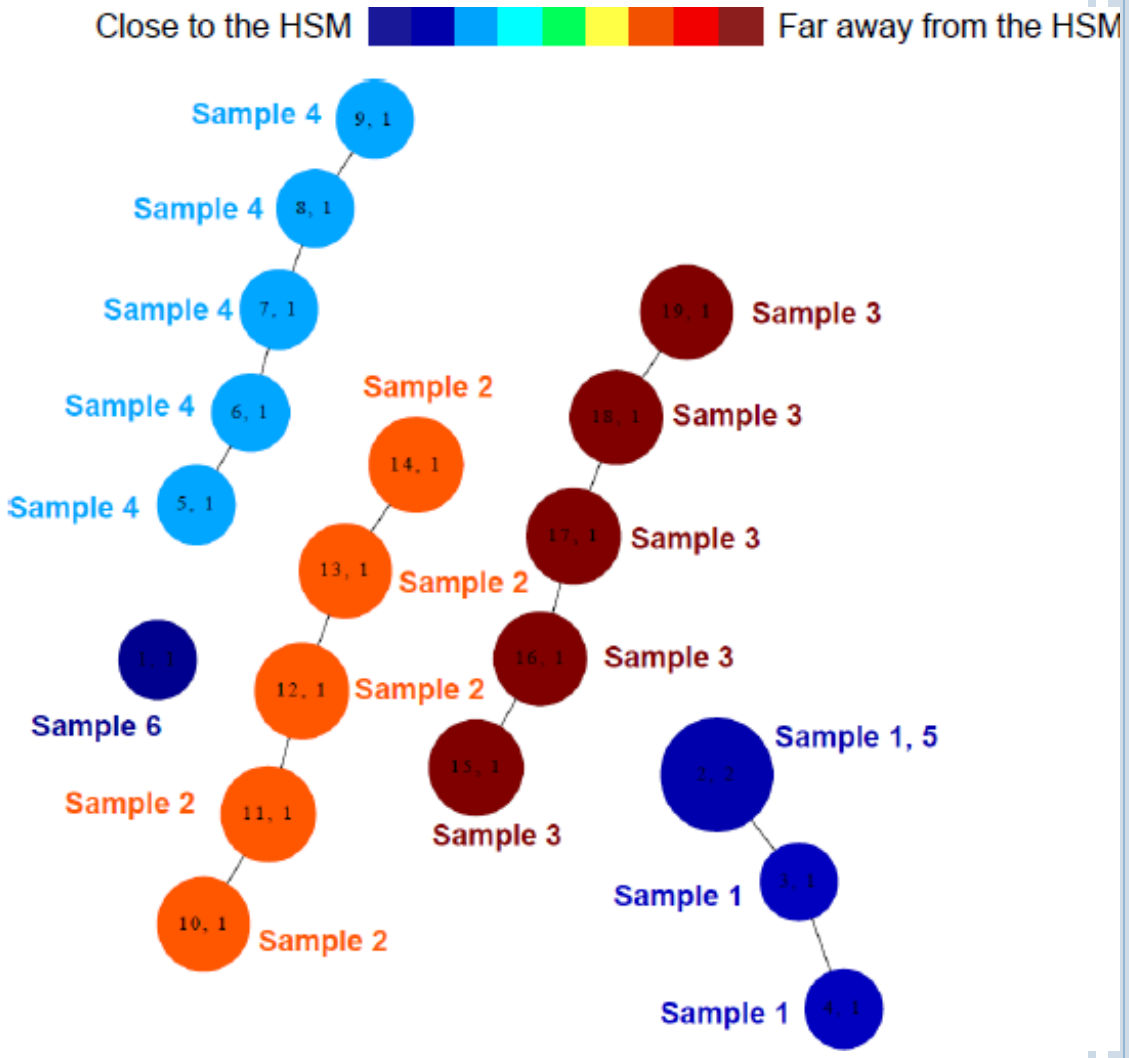
Moderated t-test adjusted with the Benjamini-Hochberg method

- No significant gene !

Topology ?

- Use the DSGA method to find significantly different genes specific to RANKL.
- « Healthy State model » is the control group (GROUP A)
- « Tumour component » is the group where RANKL was added (GROUP B).

Topology ?



Sample 1: 1nmol/l **Sample 2**: 20nmol/l **Sample 3**: 18nmol/l
Sample 4: 2nmol/l **Sample 5**: 1nmol/l **Sample 6:** 0.7nmol/l

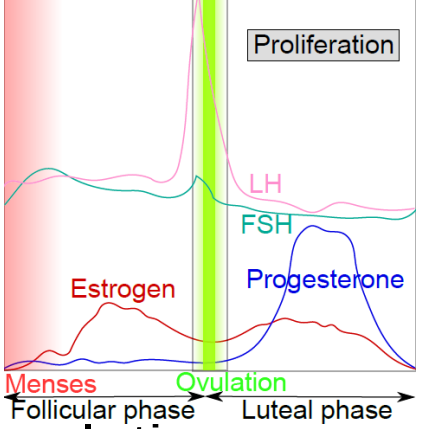
Second application of PAD

Hormones implicated in Breast Cancer

- Epidemiological studies showed that there is an increase of the breast cancer risk of 1.24 for women taking oral contraceptives.
- There is no significant risk anymore 10 years (or more) after stopping use.

Breast cancer and hormonal contraceptives, Collaborative group on Hormonal factors in Breast Cancer,1996.

Contraceptives-Mechanism



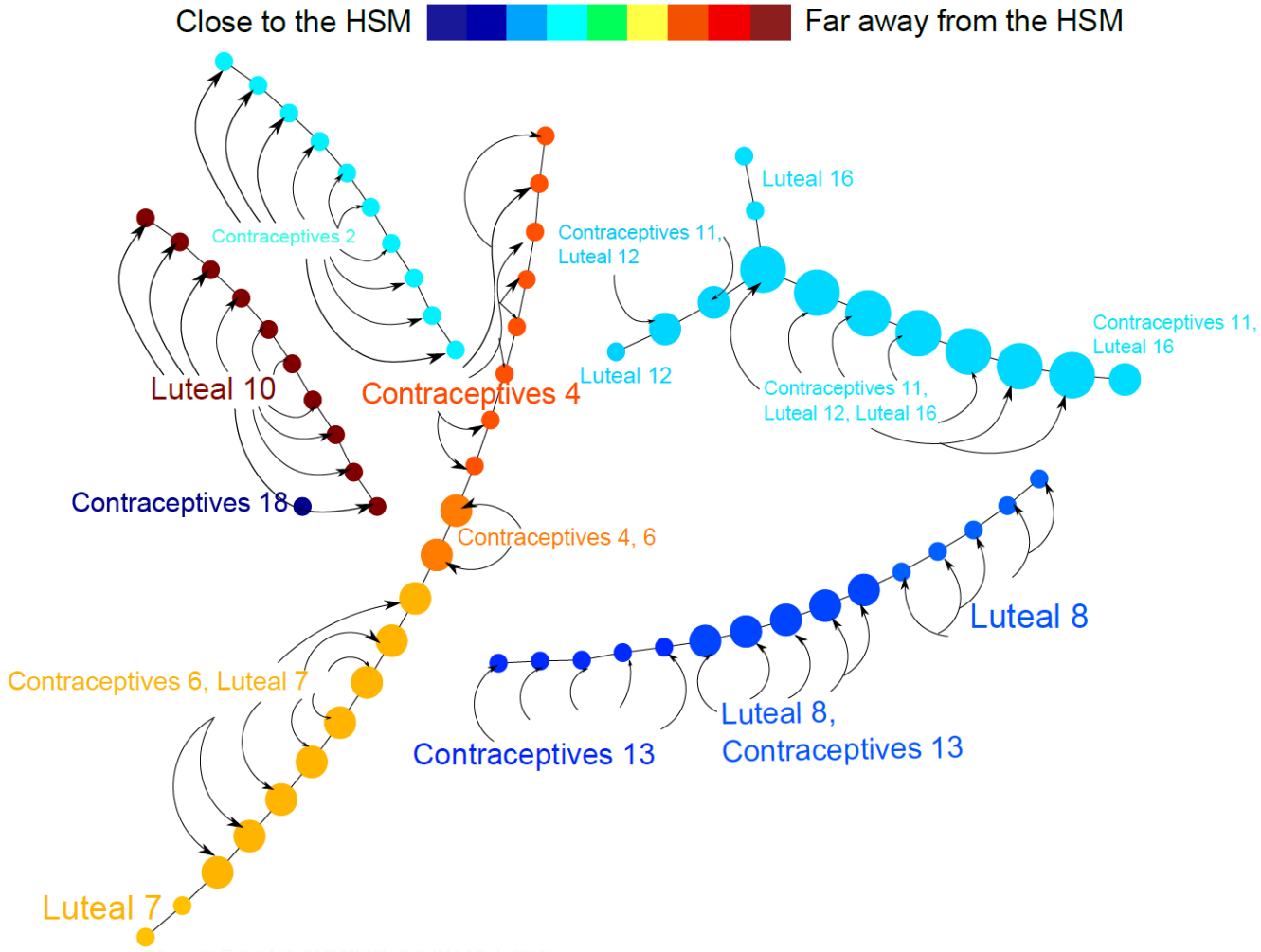
- The idea of oral contraceptives is to block ovulation by introducing a low level of estrogen.
- This has the consequence that the luteal phase is abrogated.
- However, we do not know the action of progestins, which are meant to be replacing the progesterone peak.

Bahamondes, Luis, et M. Valeria Bahamondes. « New and Emerging Contraceptives: A State-of-the-Art Review ». *International Journal of Women's Health*, février 2014, 221. doi:10.2147/IJWH.S46811.r

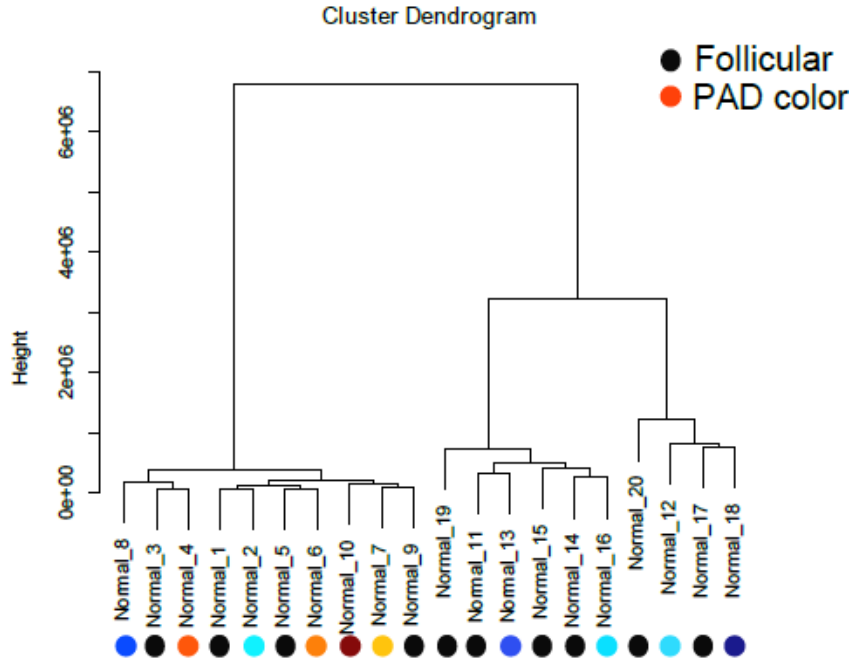
Article

- Pardo I, Lillemoe HA, Blosser RJ, Choi M, Sauder CA, Doxey DK, Mathieson T, Hancock BA, Baptiste D, Atale R, Hickenbotham M, Zhu J, Glasscock J, Storniolo AM, Zheng F, Doerge R, Liu Y, Badve S, Radovich M, Clare SE (2014) *Next-generation transcriptome sequencing of the premenopausal breast epithelium using specimens from a normal human breast tissue bank.* Breast Cancer Res 16: R26.

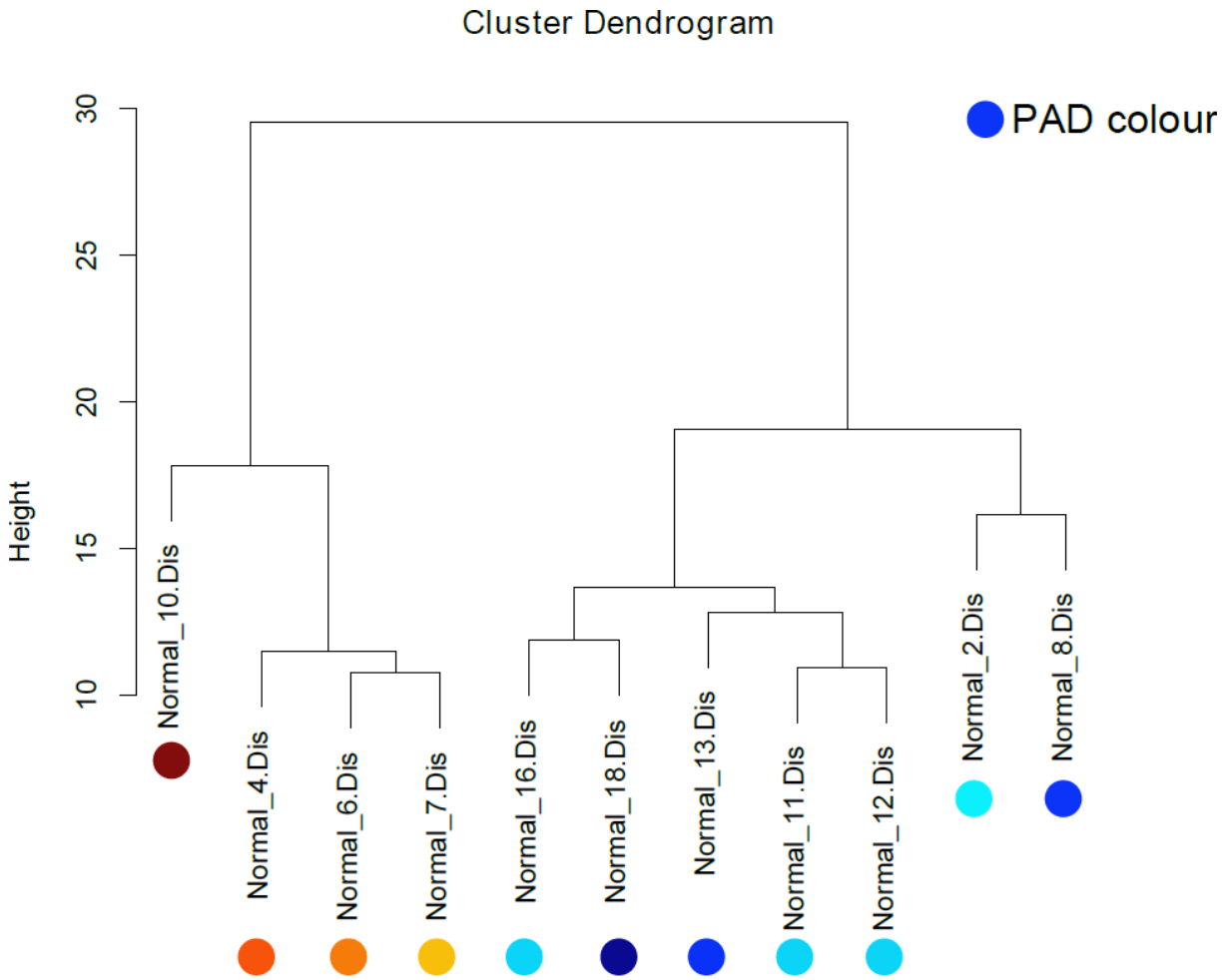
Mapper-Analysis



Standard tools, dendrogram



STANDARD TOOLS, DSGA OUTPUT



Principal Component Analysis