Robust estimation of the relationship between DNA copy number and gene expression

Pierre Neuvial

Laboratoire Statistique et Génome
Université d'Évry Val d'Essonne
UMR CNRS 8071 – USC INRA

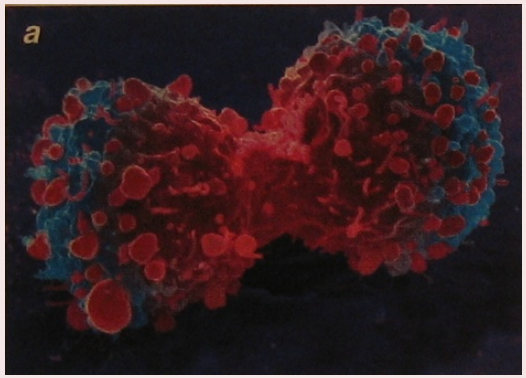
Joint work with Antoine Chambaz and Mark van der Laan

Outline

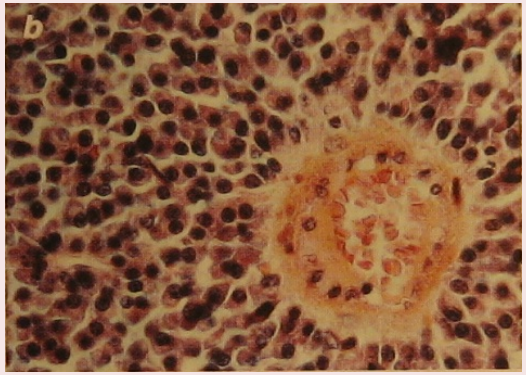
Outline

Caracteristics of tumor cells

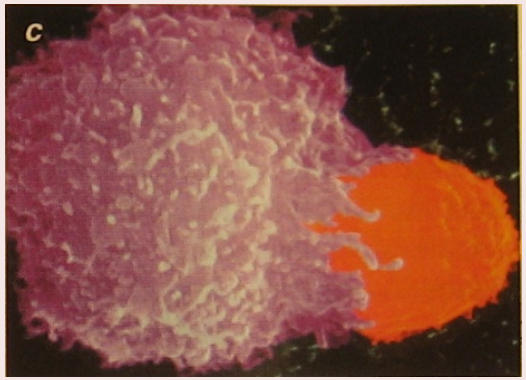
Hanahan & Weinberg (2000)



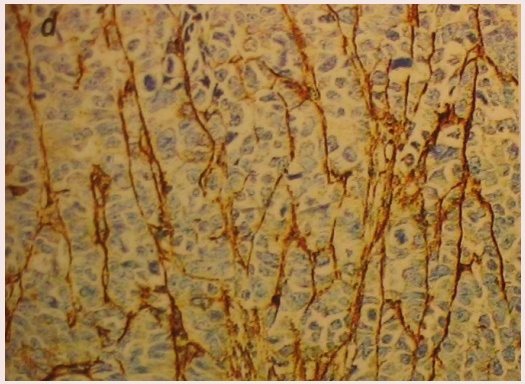
self-sufficiency in growth factors



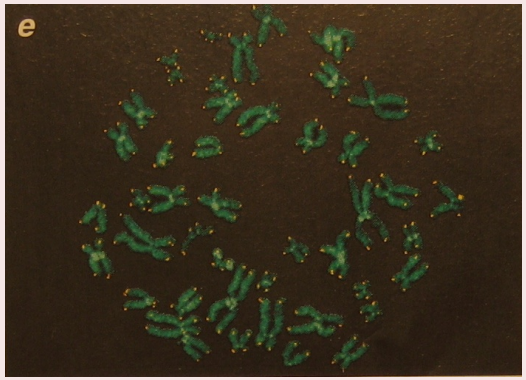
insensibility to anti-growth signals



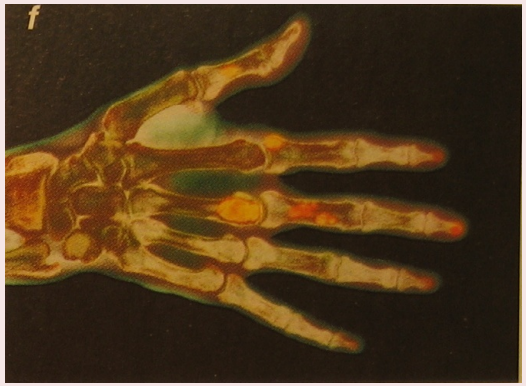
no apoptosis



angiogenesis



limitless replication potential



tissue invasion and metastases

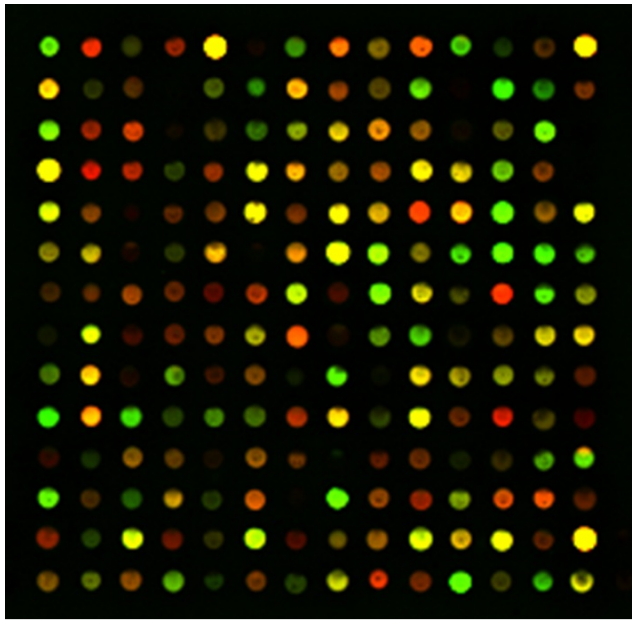
Enabled by **genetic instability** of tumor cells

Changes in cancer cells at the molecular level

Different levels of biological information

- DNA copy number
- gene expression
- DNA methylation

Quantitative measurements can be obtained from DNA microarrays

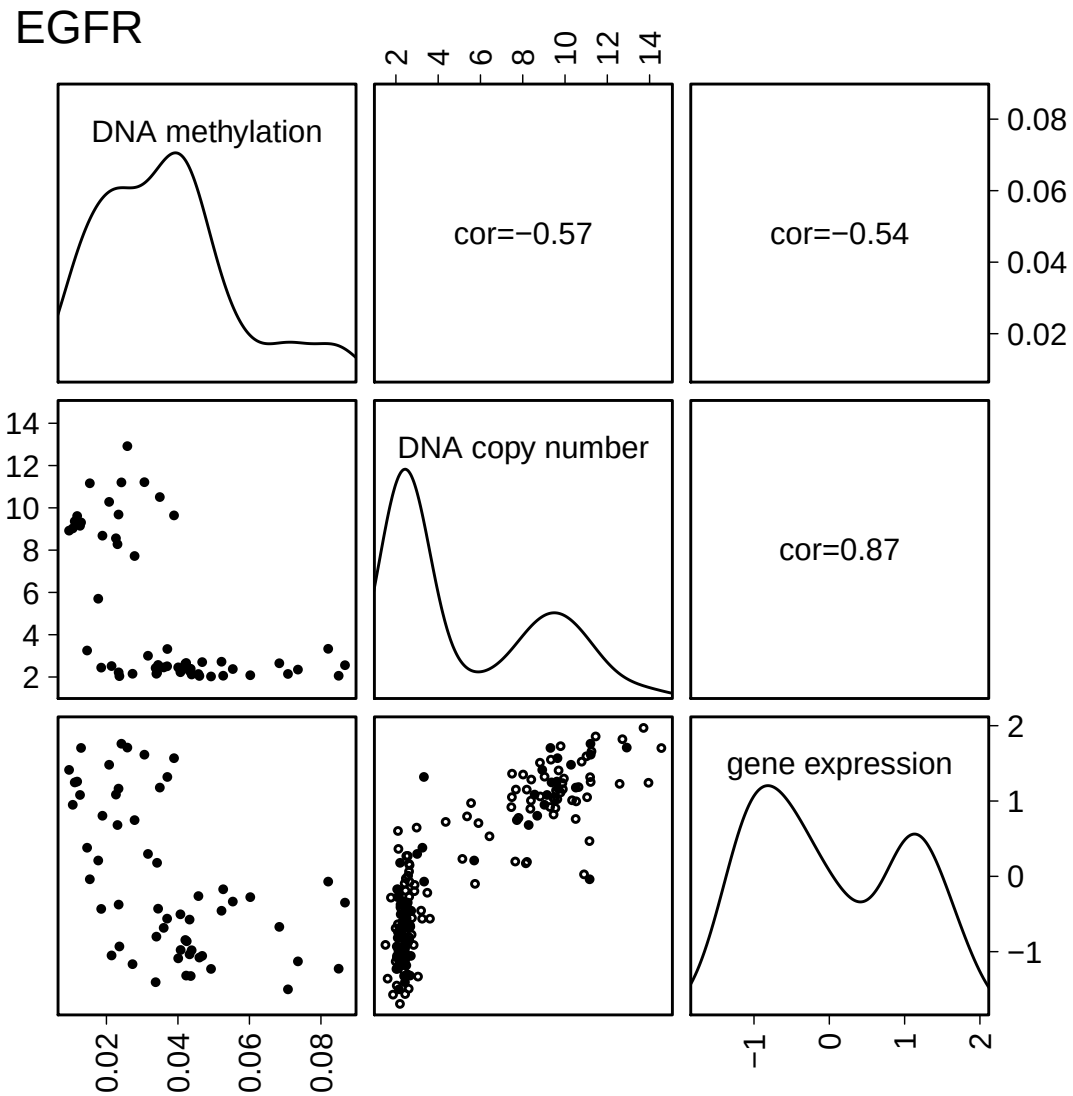


Goal: find genes that **drive** tumorigenesis

- to better understand cancer cells
- to help find new treatments

What gene-level data look like

187 GBM (brain cancer) samples from the Cancer Genome Atlas (TCGA)



Which genes are drivers ?

"Driver genes" are expected to show some **association** between DNA copy number and gene expression

$\Rightarrow$ **Test** for association, and **quantify** it

Methods for genome-wide scanning for gene-level associations

- linear correlations
- differential expression (T-tests) between copy number states
- canonical correlation analyses

Issues with existing methods

- they essentially identify genes that were already known to be implied
- associations may be non linear
- DNA methylation may down-regulate gene expression

Defining "gene-level data"

In the preceding plot:

DNA methylation (W) : proportion of "methylated" signal at a CpG locus in the gene's promoter region.

DNA copy number (X) : smoothed normalized total copy number relative to a set of reference samples.

Expression (Y) : "unified" gene expression level across 3 platforms

Outline

Definition of a parameter of interest

Observation $O = (W, X, Y) \sim P \in \mathcal{M}$ for a given gene:

- W: DNA methylation
- X: DNA copy number; $X = 0$: **copy neutral state** (2 copies)
- Y: gene expression
- $\mathcal{M}$: non-parametric set of all possible data-gen. distributions of O

Parameter of interest (defined for all $P \in \mathcal{M}$)

$$\Psi(P) = \underset{\beta \in \mathbb{R}}{\arg\min}\, E_P\left[(E_P(Y|X,W) - E_P(Y|X=0,W) - \beta X)^2\right]$$

- In a **semi-parametric model** where $E_P(Y|X,W) = E_P(Y|X=0,W) + \beta X$, we have $\Psi(P) = \beta$.
- By contrast, $\Psi : \mathcal{M} \to \mathbb{R}$ is defined **universally**
- $\Psi(P)$ is a non-parametric variable importance measure of the "effect" of X (continuous) on Y (continuous) accounting for W

Comment on the parameter of interest

Let $\theta(P)(X, Y) = E_P(Y|X, W)$, then

$$\Psi(P) = corr(X, r_P(X, W)) \sqrt{\frac{E_P[r_P(X, W)^2]}{E_P[X^2]}},$$

where $r_P(X, W) = \theta(P)(X, W) - \theta(P)(0, W)$

Case where X is binary

If $X \in \{0, 1\}$, then

$$\Psi(P) = E_P[(\theta_P(1, W) - \theta_P(0, W))h(W)]$$

with weight $h(W) = P(X = 1|W)/P(X = 1)$

Targeted maximum likelihood methods: motivation

Goal: estimate a parameter $\Psi(P)$ from observations arising from a distribution P. Ψ **is known**.

Naive strategy

1. Estimate P using $\hat{P}$
2. Plug-in: $\Psi(\hat{P})$

Our target parameter is $\Psi(P)$, not P !

- $\hat{P}$ aims at balancing bias and variance for the whole distribution
- $\Psi(\hat{P})$ **does not** balance bias and variance for $\Psi(P)$

Targeted maximum likelihood estimation (TMLE)

From an initial estimate P_n^0:

1. Create a model $P_n^0(\varepsilon)$ parametrized by $\varepsilon \in \mathbb{R}$ whose score is the **efficient influence curve** of Ψ at P_n^0
2. Estimate ε using maximum likelihood: ε_n^0
3. Update accordingly: $P_n^1 = P_n^0(\varepsilon_n^0)$

Repeat as many times as necessary... hence **final estimate** $P_n^\star$

Statistical properties

P_0: true distribution of O

Consistency (double robustness)

TMLE is consistent if one of the following conditions holds:

- $\theta(P_n^\star)(0,\cdot)$ consistently estimates true $\theta(P_0)(0,\cdot)$
- $E_{P_n^\star}(X|W)$ and $P_n^\star(X=0|W)$ consistently estimate $E_{P_0}(X|W)$ and $P_0(X=0|W)$

Asymptotic normality

Under the same conditions, TMLE is asymptotically Gaussian

We can compute asymptotic p-values and thus **rank genes**

Outline

Simulation strategy

Assumptions:

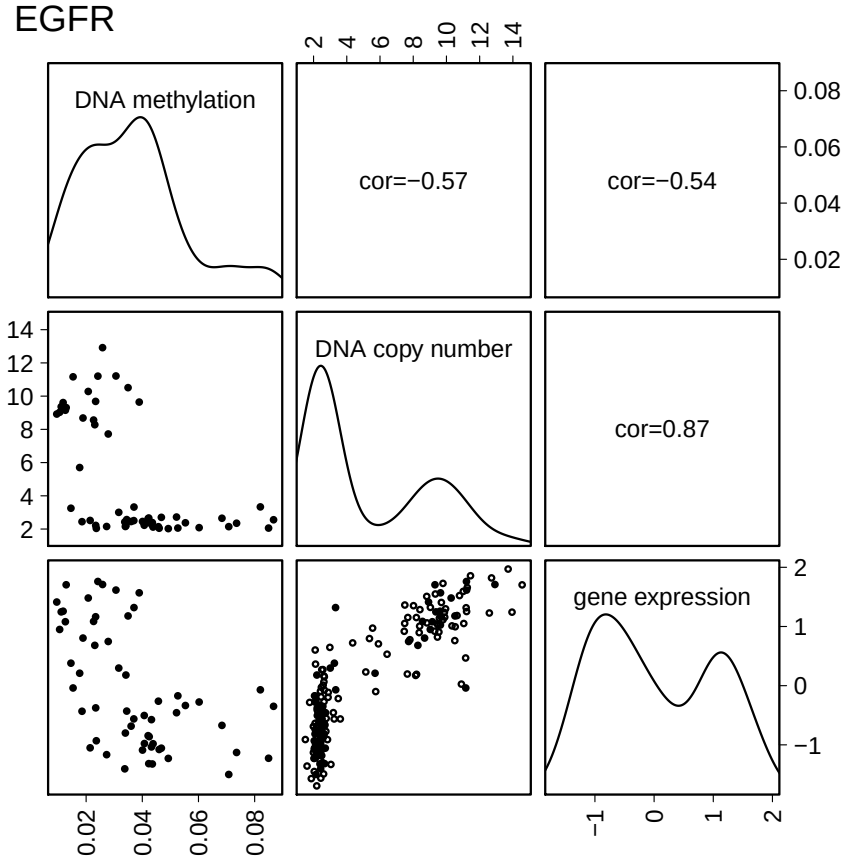
- up to 3 copy number classes: normal regions, and regions of copy number gains and losses
- in normal regions, expression is negatively correlated with methylation
- in regions of copy number alteration, copy number and expression are positively correlated

GBM data used as a baseline for simulation:

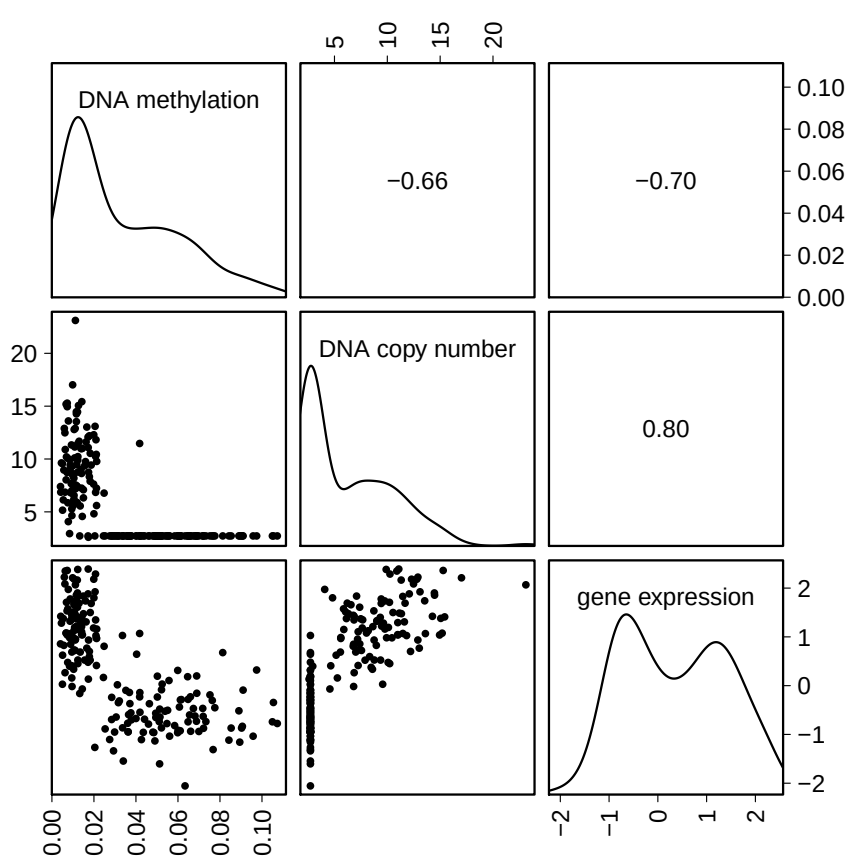
Sample name	Methylation	Copy number	Expression
TCGA-02-0001	0.05	2.72	-0.46
TCGA-02-0003	0.01	9.36	1.25

Simulated data set mimics real data set

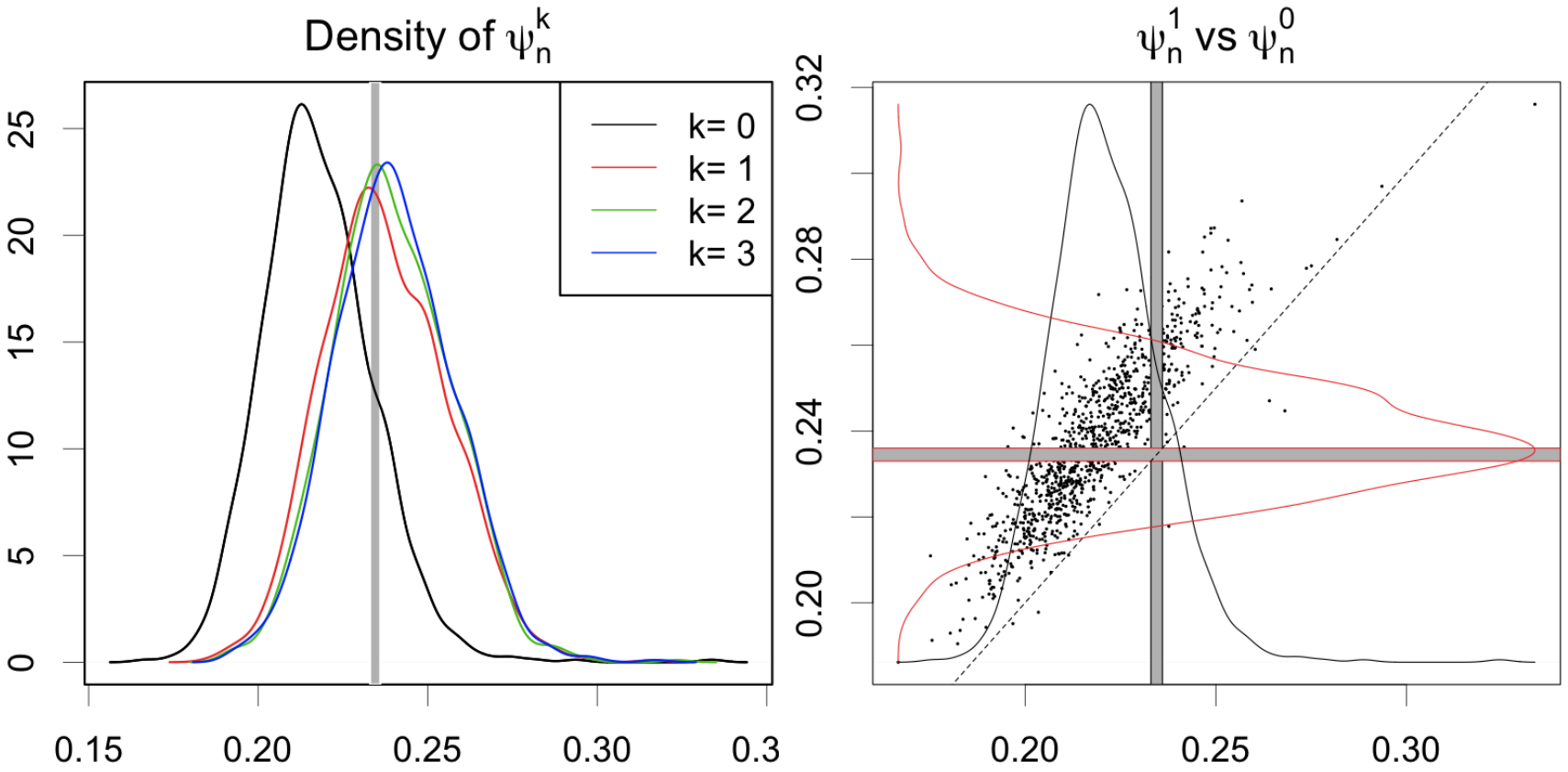
Real data (GBM, n=187)



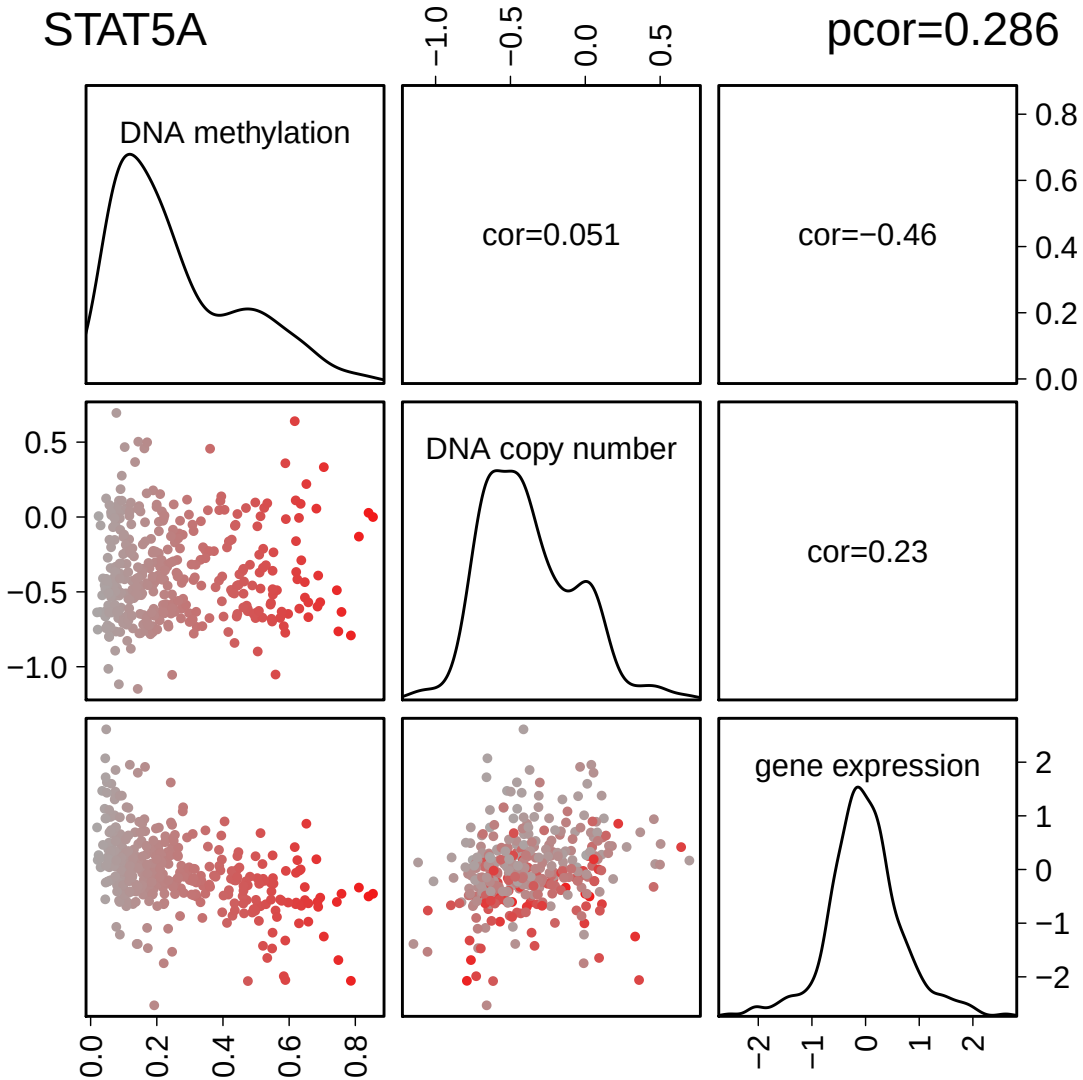
Simulated data (n=200)



Simulated data: TMLE corrects initial estimation



Real data analysis : TCGA OV data set



Thanks

- **Antoine Chambaz**
- Mark van der Laan
- Terry Speed

The Cancer Genome Atlas Research Network