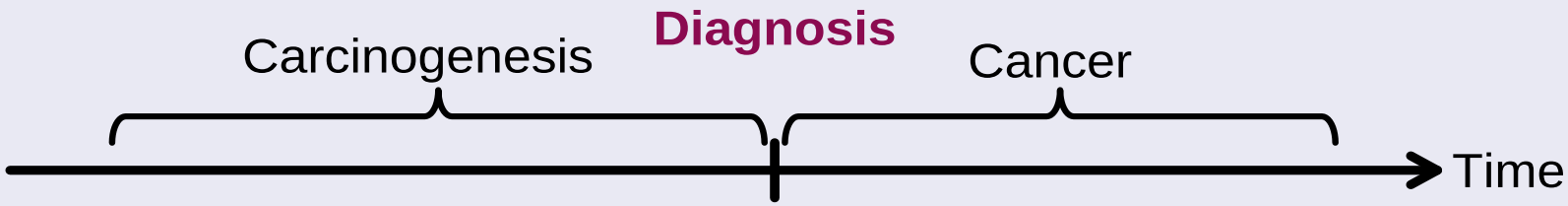
Statistical approaches to explore carcinogenic process on genome-wide transcriptomic data

TICE (Transcriptomics In Cancer Epidemiology)
NOWAC (Norwegian Women And Cancer)

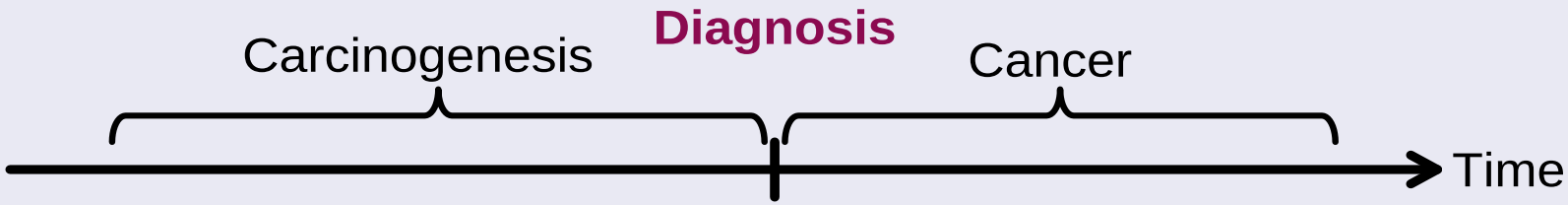
Sandra Plancade, University of Tromsø (Norway)
Gregory Nuel, Université Paris-Descartes
Yoav Benjamini and Marina Bogomolov, Tel Aviv University (Israel)
Eiliv Lund, University of Tromsø

1st of October 2012

Carcinogenesis

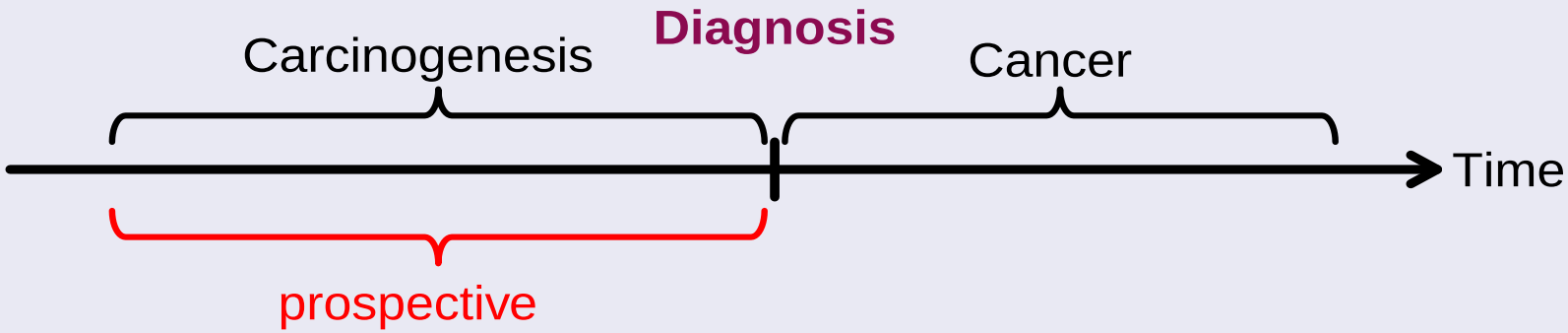


Carcinogenesis



Genome Wide Association Studies (GWAS)

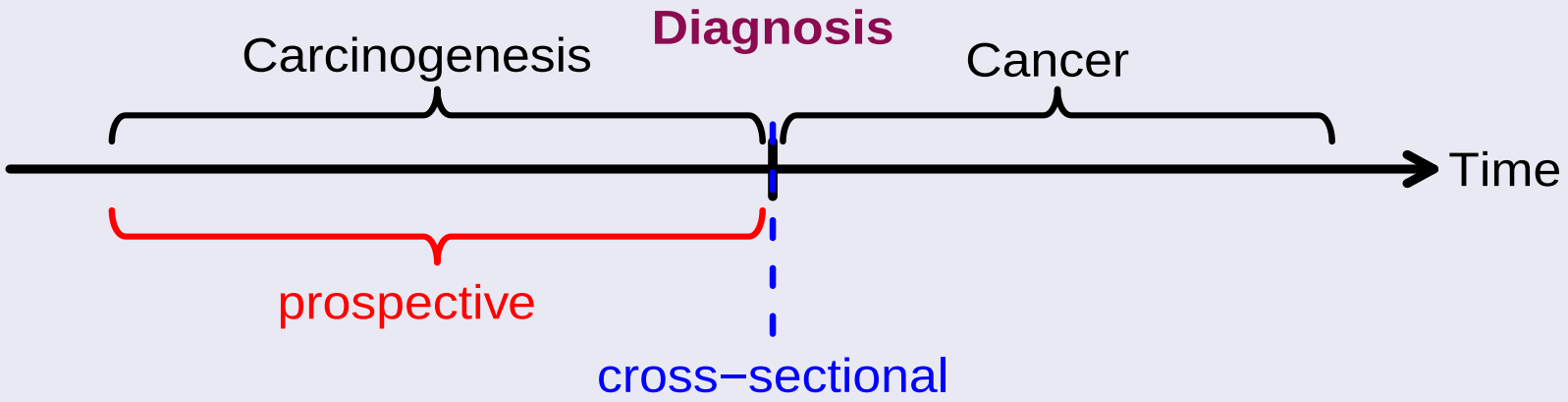
Carcinogenesis



Genome Wide Association Studies (GWAS)

- Prospective study: follow-up.

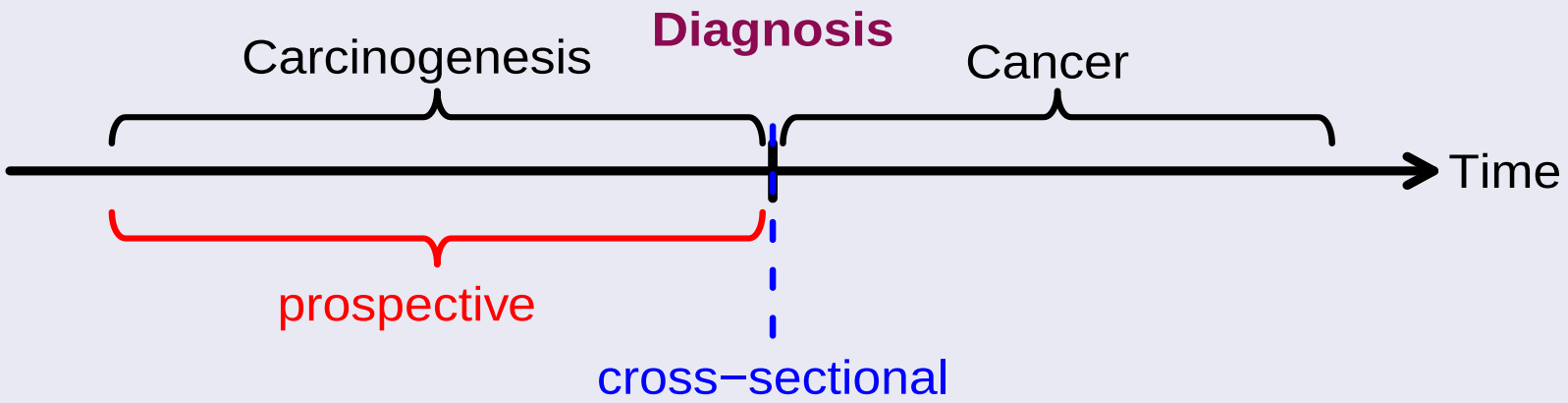
Carcinogenesis



Genome Wide Association Studies (GWAS)

- Prospective study: follow-up.
- Cross-sectional study: observation at a given time.

Carcinogenesis



Genome Wide Association Studies (GWAS)

- Prospective study: follow-up.
- Cross-sectional study: observation at a given time.

$\hookrightarrow$ Our study: prospective design.

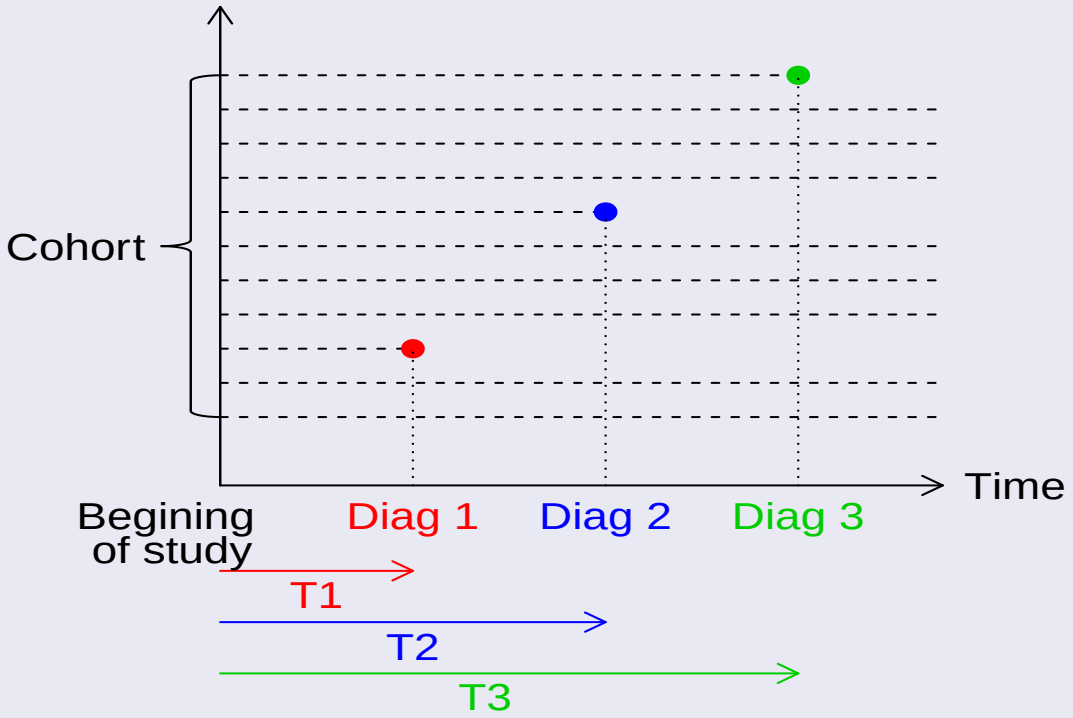
1. Classical prospective GWAS - Nested case-control design
2. Post-GWAS: transcriptomics in a prospective design
3. Statistical approaches for post-GWAS
   - Gene by gene model
   - Latent last-stage model

1. Classical prospective GWAS - Nested case-control design
2. Post-GWAS: transcriptomics in a prospective design
3. Statistical approaches for post-GWAS
   - Gene by gene model
   - Latent last-stage model

Cohort and nested case-control design

Cohort and nested case-control design

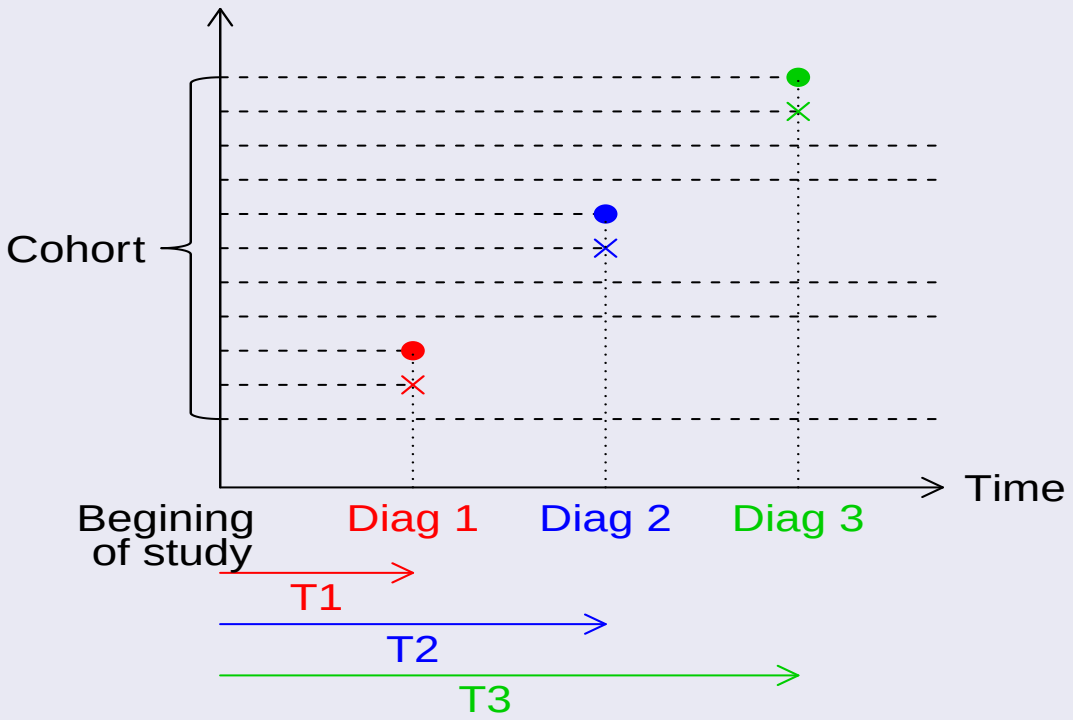
Cohort



- case

Cohort and nested case-control design

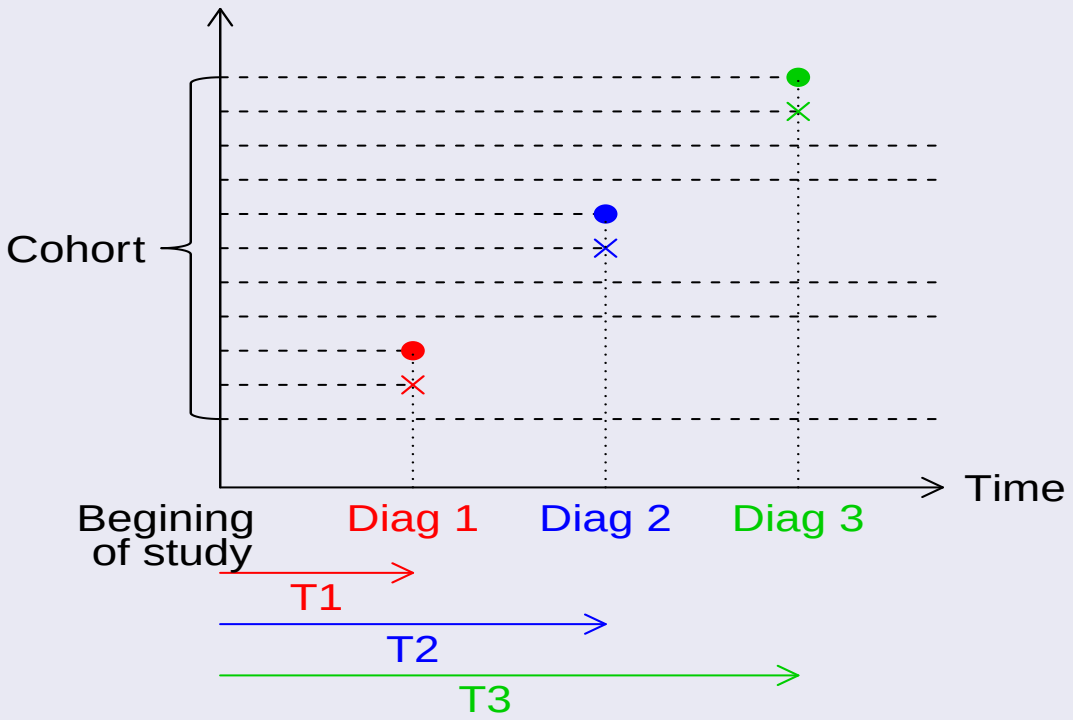
Cohort



- case
- × control

Cohort and nested case-control design

Cohort



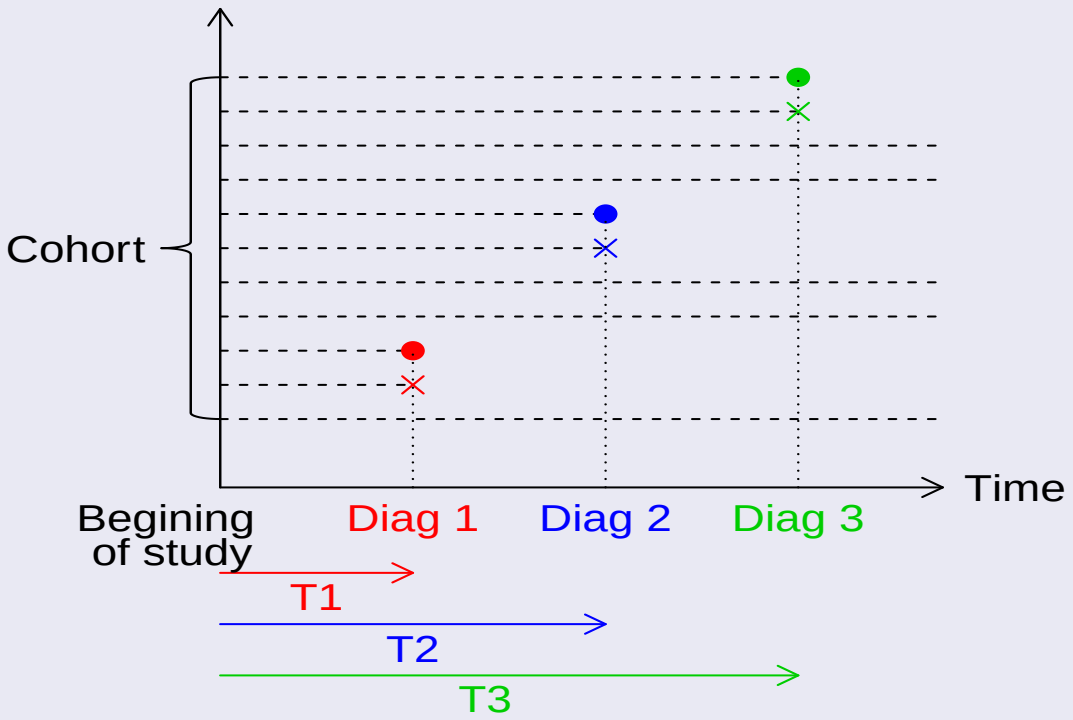
- case
- × control

Data

- Genomics (SNPs)
- Environmental factors

Cohort and nested case-control design

Cohort



- case
- × control

Data

- Genomics (SNPs)
- Environmental factors

GWAS

- Interests: relative risk estimation, prediction.
- Statistical methods: survival analysis model (in particular Cox):

$$\mathbb{P}[\text{Time} \mid \text{genomics, exposures}]$$

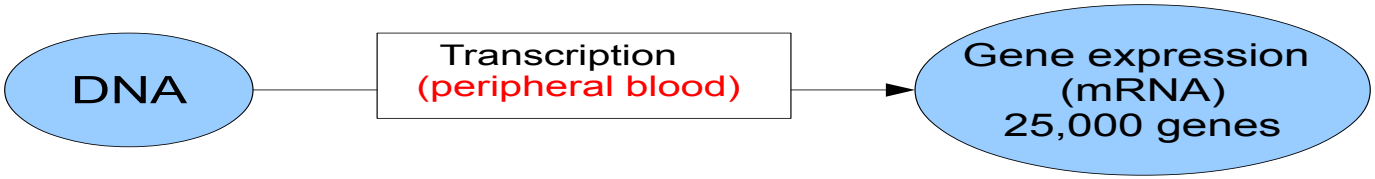
$\hookrightarrow$ Take into account the over-representation of cases.

1. Classical prospective GWAS - Nested case-control design

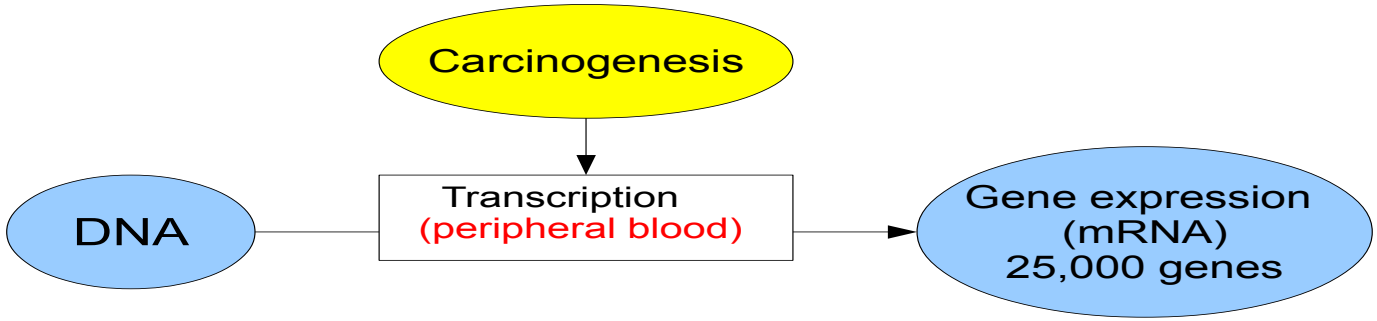
2. Post-GWAS: transcriptomics in a prospective design

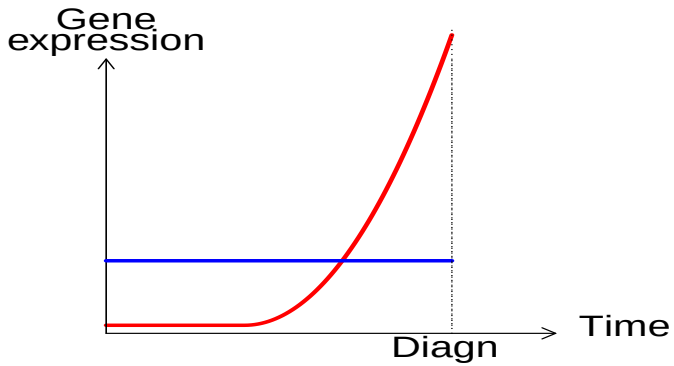
3. Statistical approaches for post-GWAS
   - Gene by gene model
   - Latent last-stage model

Carcinogenesis and transcription in peripheral blood



Carcinogenesis and transcription in peripheral blood

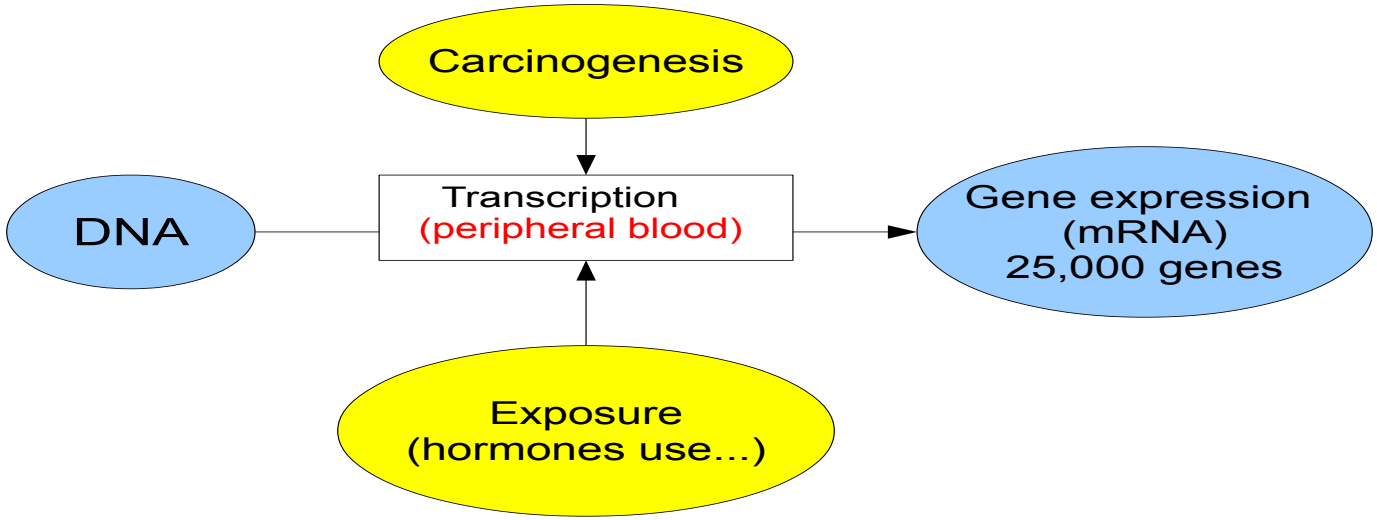


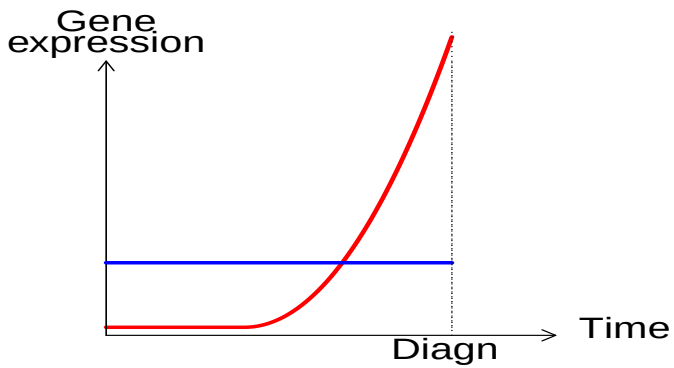


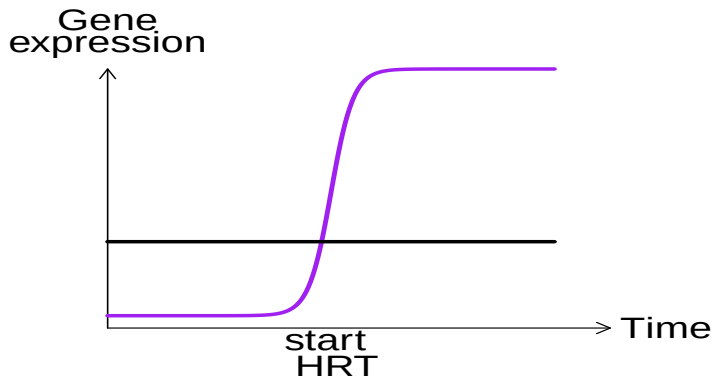
— gene involved in carcinogenesis

— gene non involved

Carcinogenesis and transcription in peripheral blood



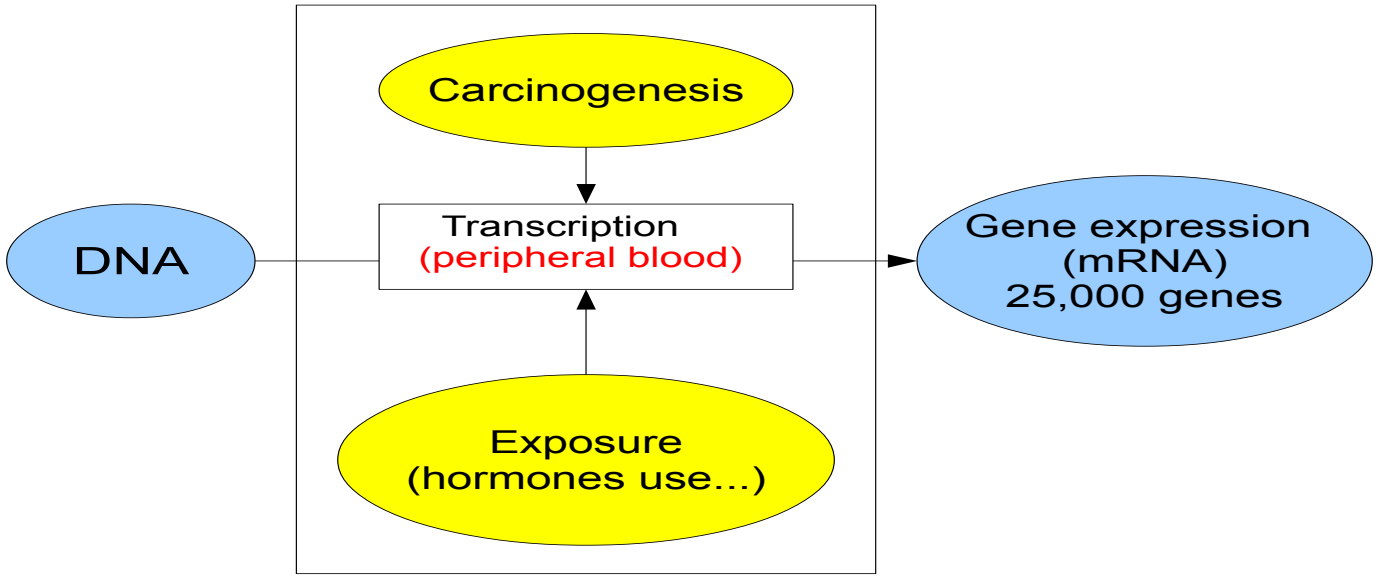


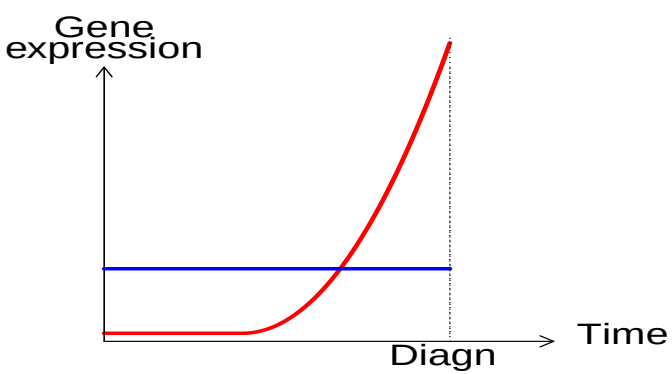


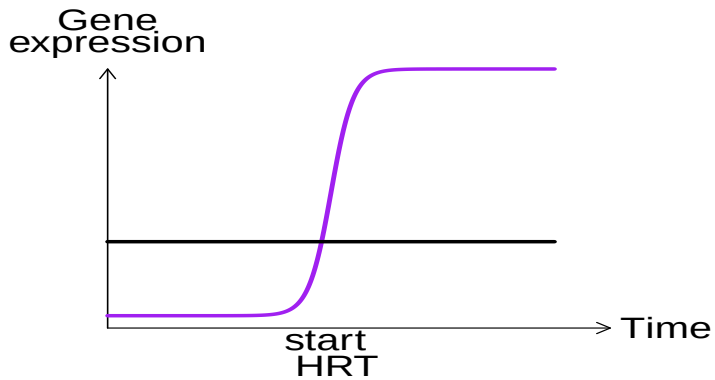
— gene involved in carcinogenesis
— gene non involved

— gene linked to HRT
— gene non-linked to HRT

Carcinogenesis and transcription in peripheral blood







— gene involved in carcinogenesis
— gene non involved

— gene linked to HRT
— gene non-linked to HRT

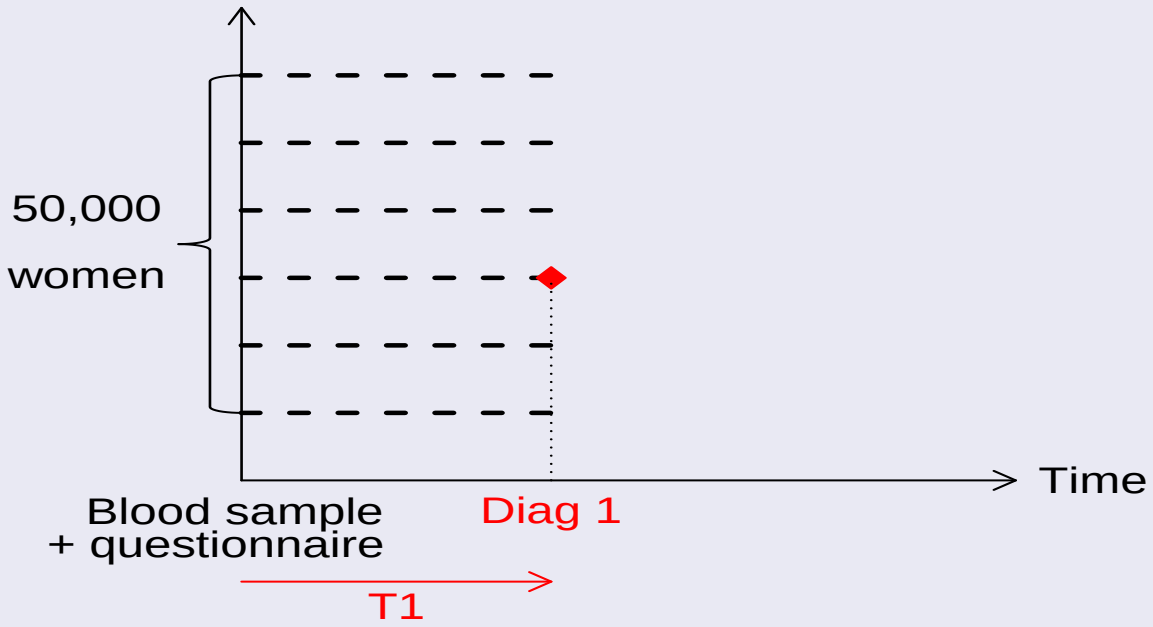
The NOWAC cohort

The NOWAC cohort

Prospective nested case-control design

The NOWAC cohort

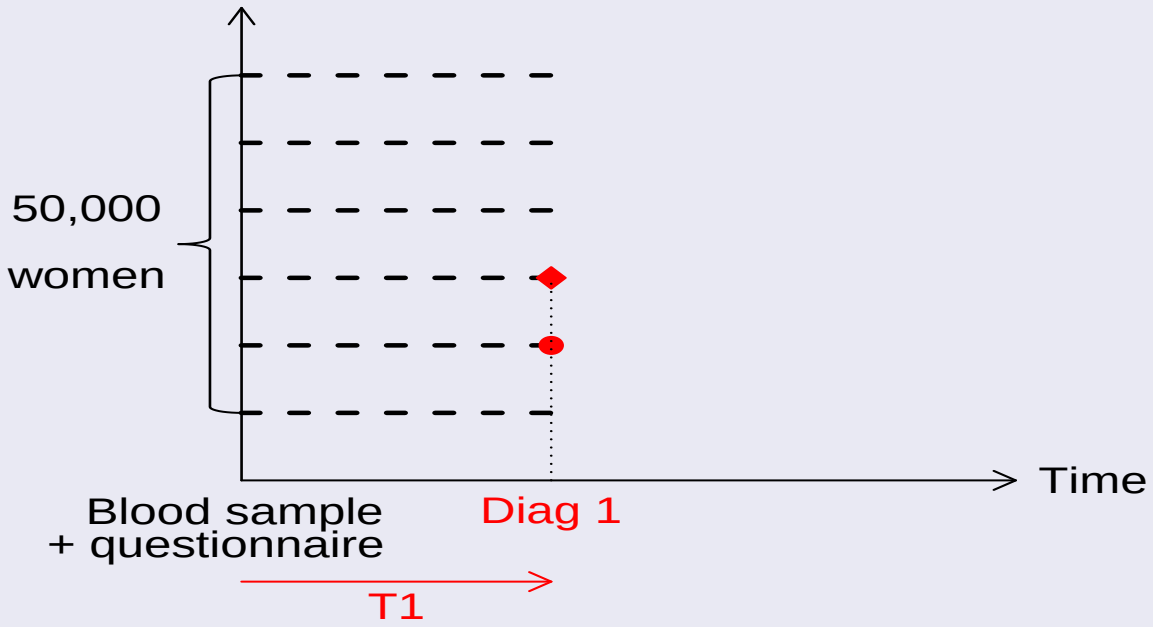
Prospective nested case-control design



◆: case
●: control

The NOWAC cohort

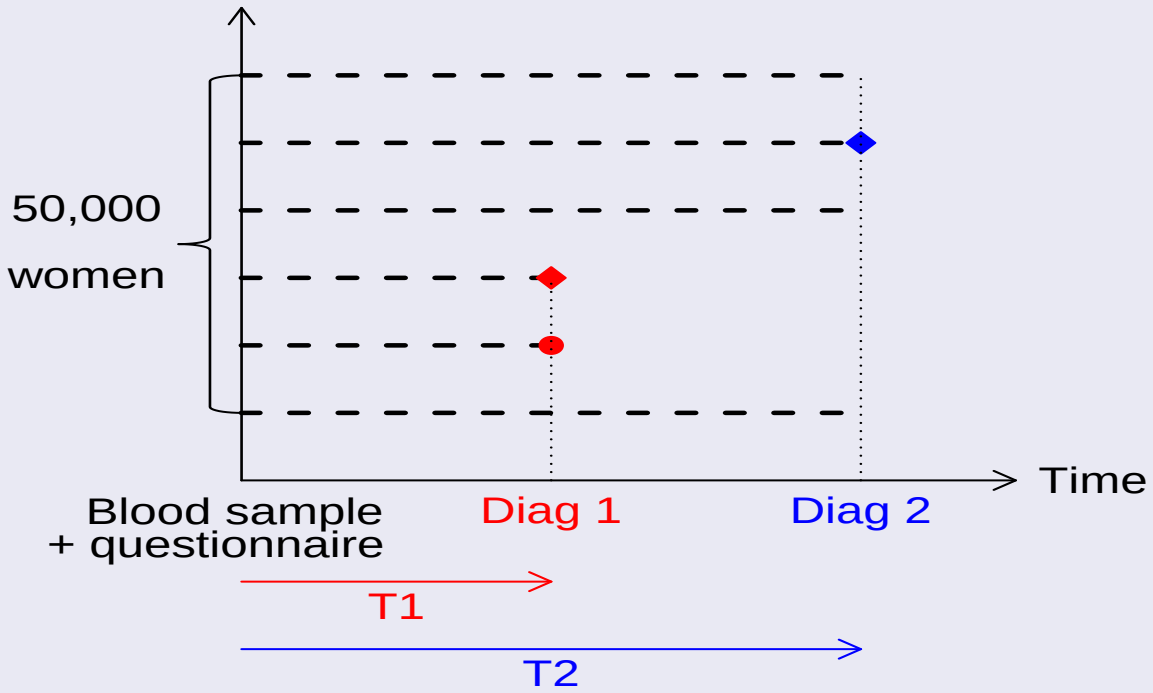
Prospective nested case-control design



◆: case
●: control

The NOWAC cohort

Prospective nested case-control design

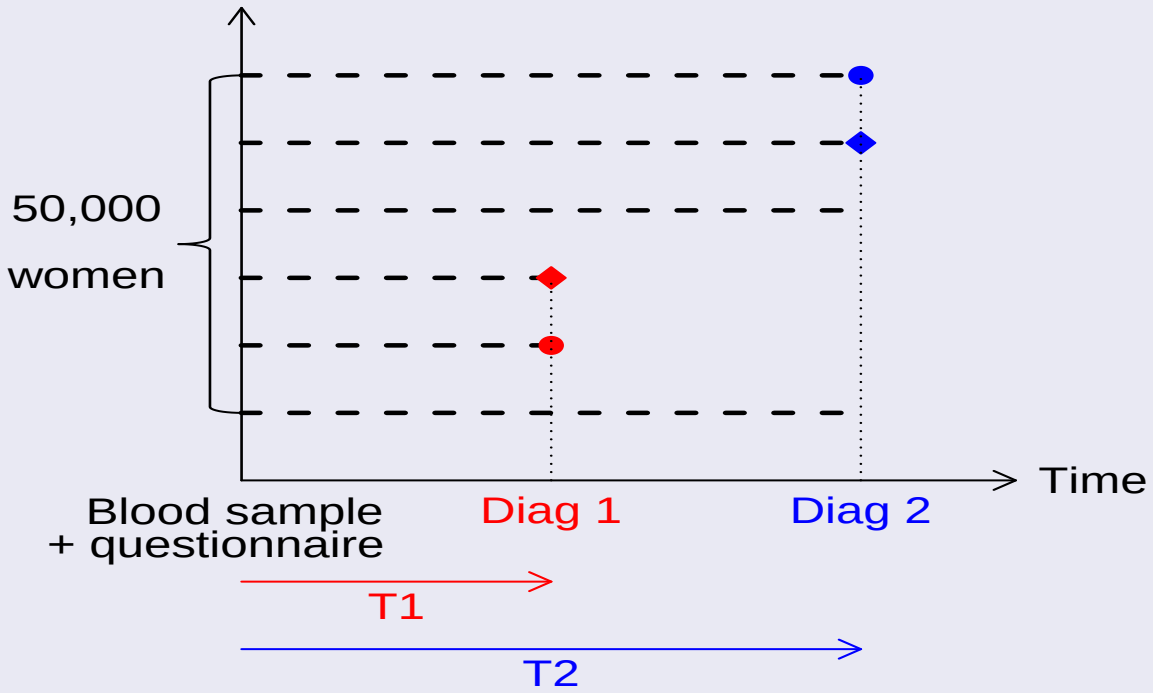


◆: case

●: control

The NOWAC cohort

Prospective nested case-control design

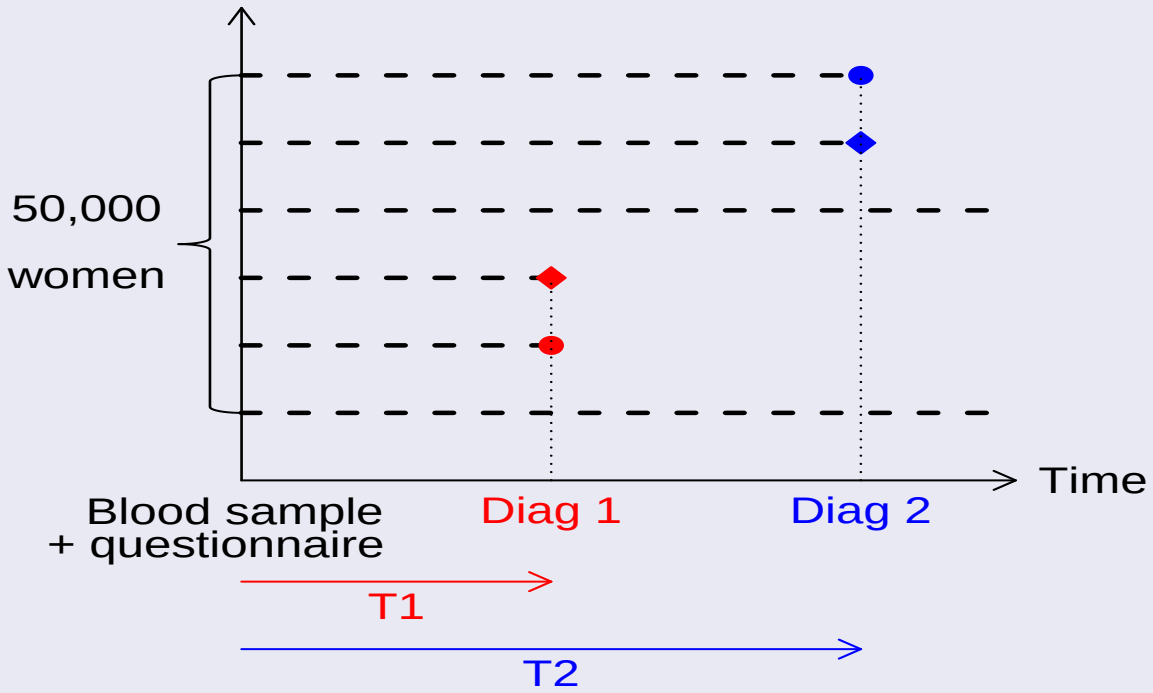


◆: case

●: control

The NOWAC cohort

Prospective nested case-control design

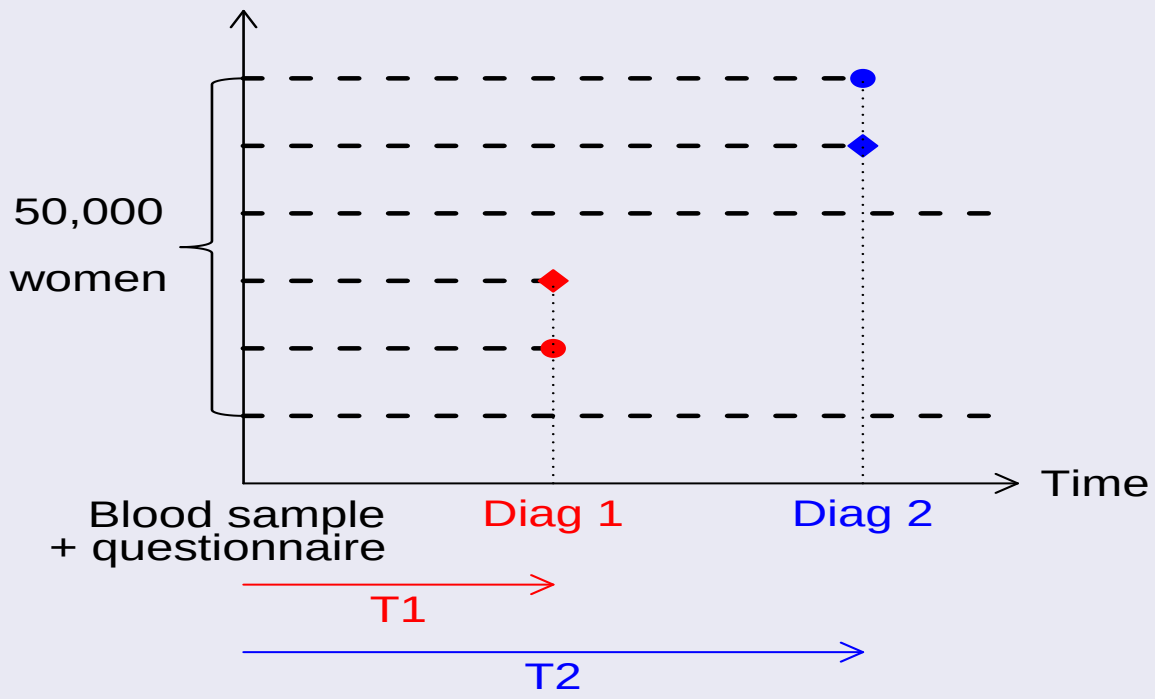


◆: case
●: control

- 6 years of follow-up
- 700 case-control pairs for breast cancer

The NOWAC cohort

Prospective nested case-control design



◆: case
●: control

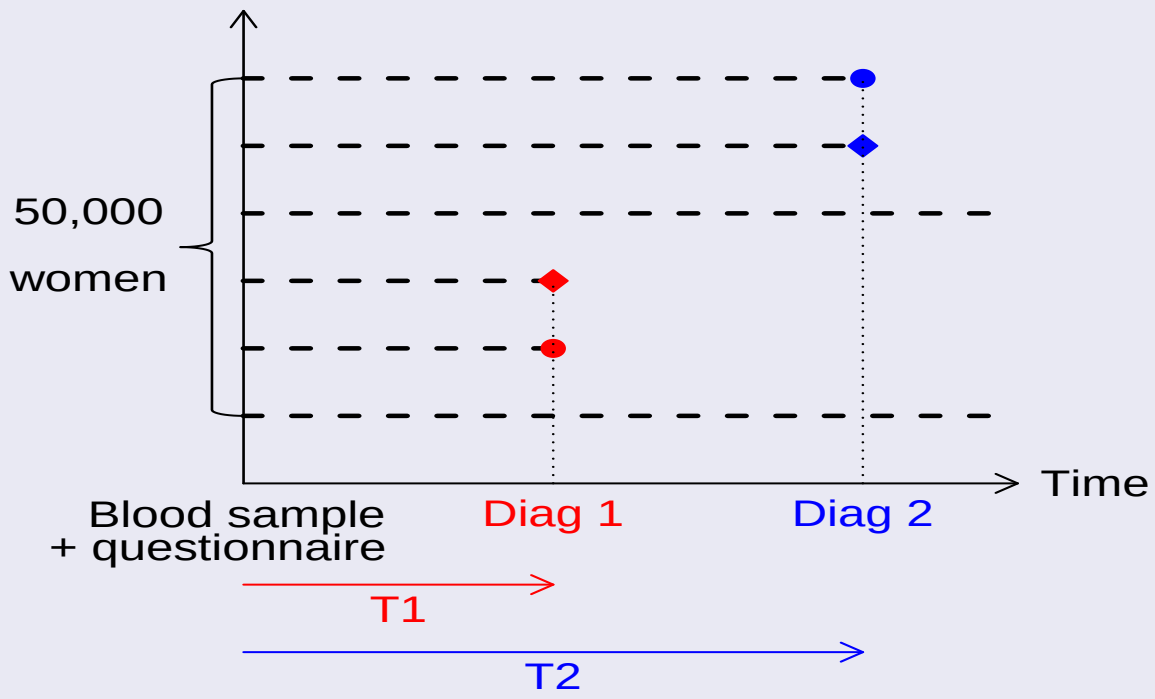
- 6 years of follow-up
- 700 case-control pairs for breast cancer

Data: for each case-control pair i,

- T_i: Follow-up time.
- $\Delta G_i = \log G_i^{\text{case}} - \log G_i^{\text{control}}$: Difference of gene expression at time T_i before diagnosis (25,000 genes).
- ΔE_i: Exposure of CC pair i at time T_i before diagnosis.

The NOWAC cohort

Prospective nested case-control design



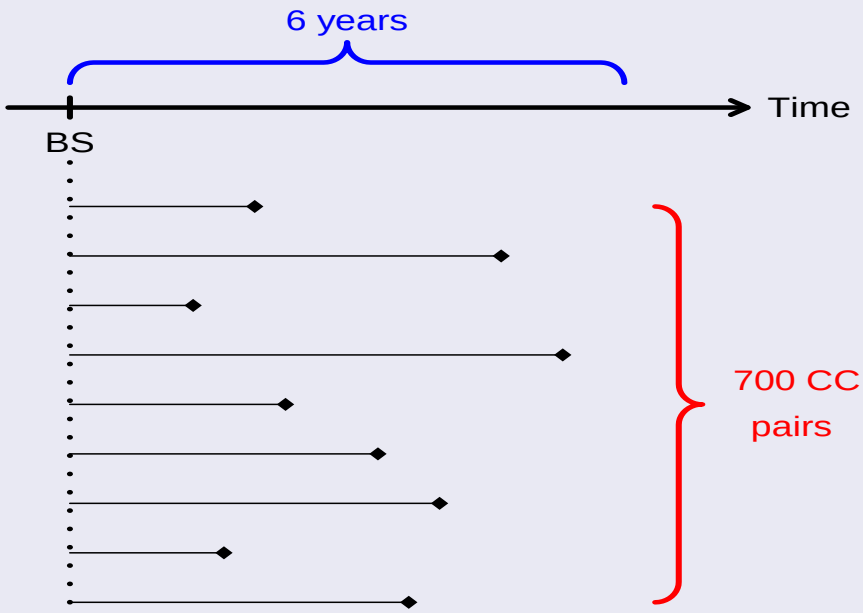
◆: case
●: control

- 6 years of follow-up
- 700 case-control pairs for breast cancer

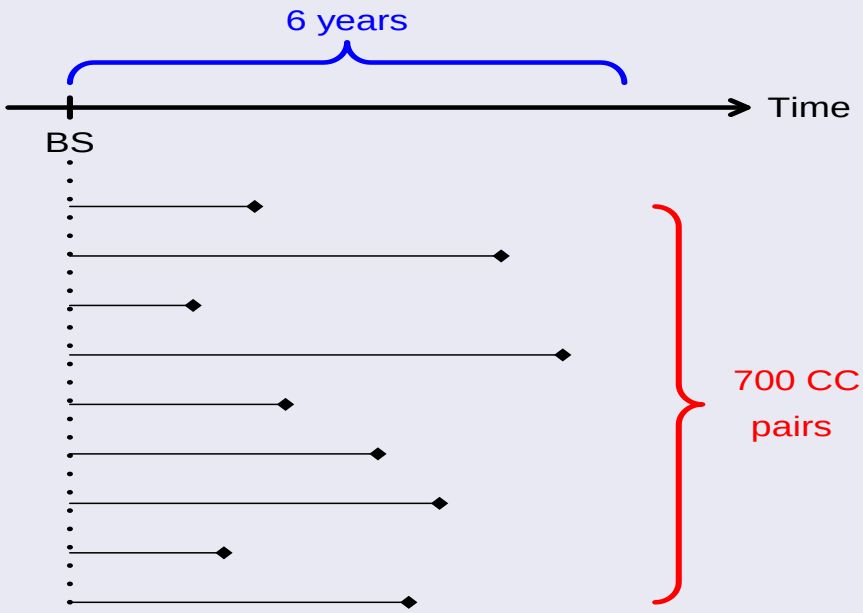
Data: for each case-control pair i,

- T_i: Follow-up time.
- $\Delta G_i = \log G_i^{\text{case}} - \log G_i^{\text{control}}$: Difference of gene expression at time T_i before diagnosis (25,000 genes).
- ΔE_i: Exposure of CC pair i at time T_i before diagnosis.

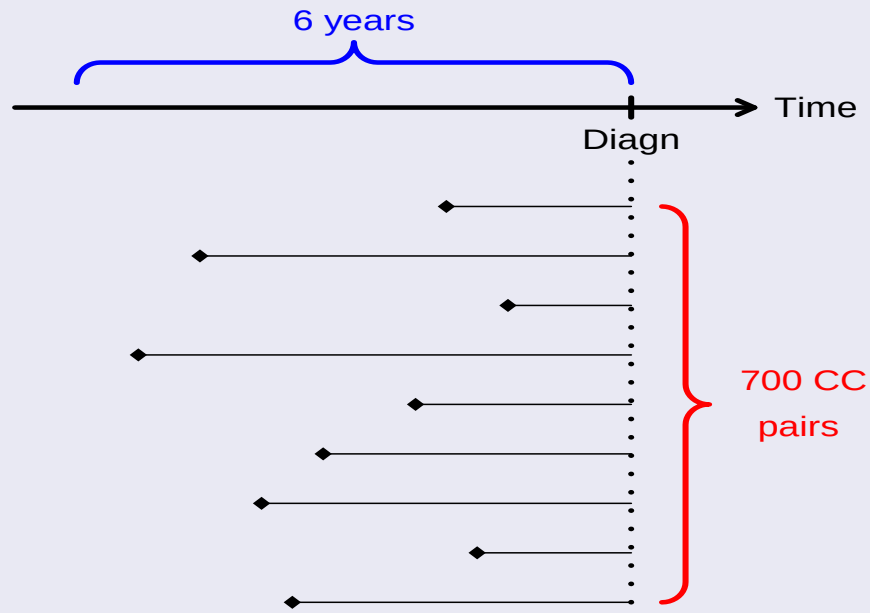
Nested case-control design



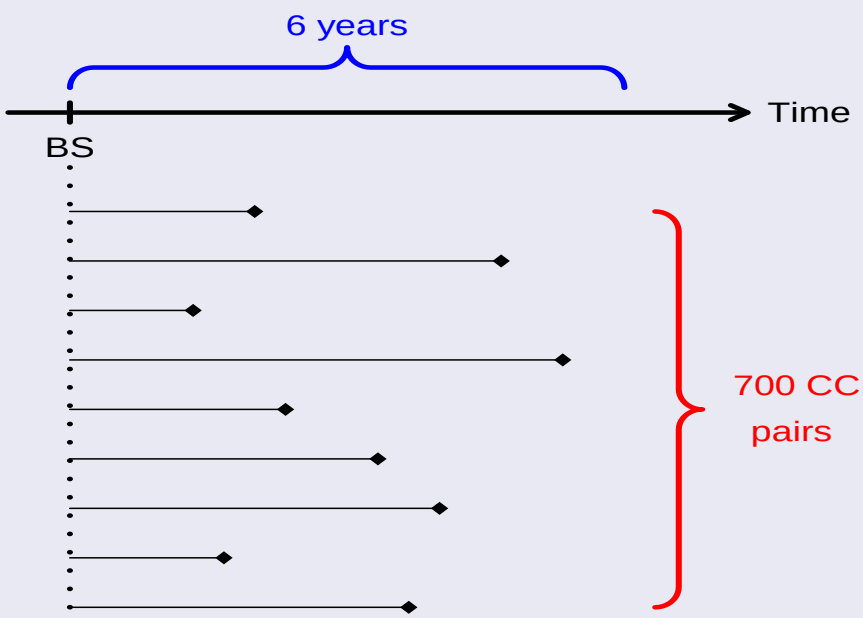
Nested case-control design



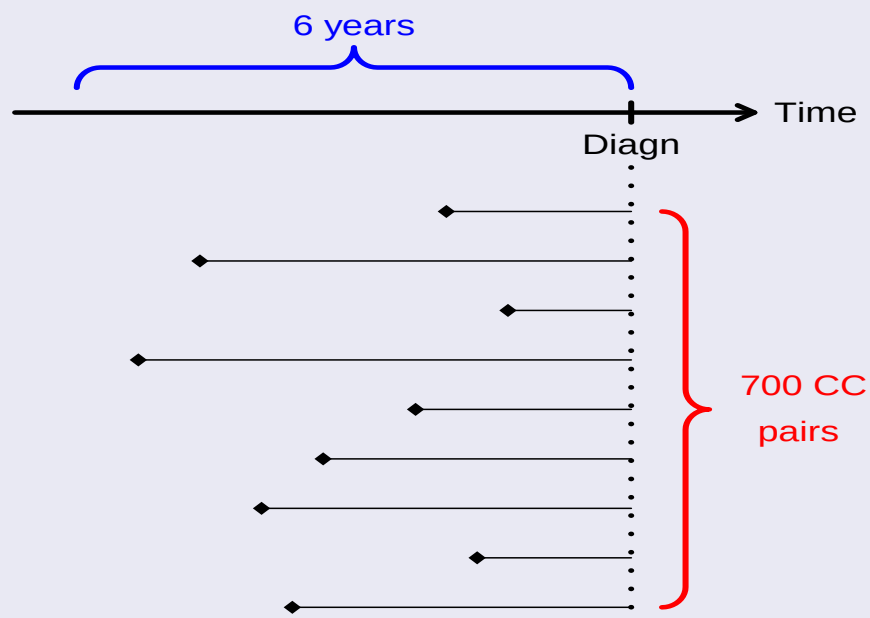
Alternative point of view



Nested case-control design



Alternative point of view



- Measurements of gene expression between 0 and 6 years before diagnosis
- Only one measurement by case-control pair.
- Explore the changes in gene expression 6 years before diagnosis.

Prospective GWAS and post-GWAS: a different statistical point of view

Prospective GWAS and post-GWAS: a different statistical point of view

Prospective GWAS

$$\mathbb{P}[T|G, E] \quad \text{with}$$

- T: time to diagnosis
- E: exposures
- G: genomic data (constant over time).

Genomics: risk factors for cancer

Goal: risk estimation and prediction.

Prospective GWAS and post-GWAS: a different statistical point of view

Prospective GWAS

$\mathbb{P}[T|G, E]$ with

- T: time to diagnosis
- E: exposures
- G: genomic data (constant over time).

Genomics: risk factors for cancer

Goal: risk estimation and prediction.

Post-GWAS

$\mathbb{P}[G|T, E]$ with

- T: time to diagnosis
- E: exposures
- G: transcriptomic data (depend on T)

Transcriptomics: biomarkers of carcinogenesis

Goal: study of change in gene expression during carcinogenesis.

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} + \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} + \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

- Stratified coefficients:

$$\beta = \begin{cases} \beta_1 & \text{if} \quad T_i \leqslant t_0 \\ \beta_2 & \text{if} \quad T_i > t_0 \end{cases}$$

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} \; + \; \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

- Stratified coefficients: $\beta = \begin{cases} \beta_1 & \text{if} \quad T_i \leqslant t_0 \\ \beta_2 & \text{if} \quad T_i > t_0 \end{cases}$

$\hookrightarrow$ Penalization selects the most differentially expressed genes in each strata.

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} \; + \; \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

- Stratified coefficients: $\beta = \begin{cases} \beta_1 & \text{if} \quad T_i \leqslant t_0 \\ \beta_2 & \text{if} \quad T_i > t_0 \end{cases}$

$\hookrightarrow$ Penalization selects the most differentially expressed genes in each strata.

- More generally: $T \sim \lambda(t|G,E,\beta(T))$

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} \; + \; \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

- Stratified coefficients: $\beta = \begin{cases} \beta_1 & \text{if} \quad T_i \leqslant t_0 \\ \beta_2 & \text{if} \quad T_i > t_0 \end{cases}$

$\hookrightarrow$ Penalization selects the most differentially expressed genes in each strata.

- More generally: $T \sim \lambda(t|G,E,\beta(T))$
  $\hookrightarrow$ Not directly interpretable.

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} \; + \; \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

- Stratified coefficients: $\beta = \begin{cases} \beta_1 & \text{if} \quad T_i \leqslant t_0 \\ \beta_2 & \text{if} \quad T_i > t_0 \end{cases}$

$\hookrightarrow$ Penalization selects the most differentially expressed genes in each strata.

- More generally: $T \sim \lambda(t|G,E,\beta(T))$
  $\hookrightarrow$ Not directly interpretable.
  $\hookrightarrow$ Association between gene expression and no-carcinogen exposures?

Cox model in post-GWAS.

- Cox (proportional hazard) model: $\lambda(t|G,E) = \lambda_0(t)\exp\left(\langle\beta,(G,E)\rangle\right)$
- Partial likelihood for nested CC:

$$L(\beta) = \prod_{i \text{ CC pair}} \Big(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\Big)^{-1} \; + \; \text{pen}(\beta)$$

$\hookrightarrow$ The follow-up time disappears = simple logistic regression.

- Stratified coefficients: $\beta = \begin{cases} \beta_1 & \text{if} \quad T_i \leqslant t_0 \\ \beta_2 & \text{if} \quad T_i > t_0 \end{cases}$

$\hookrightarrow$ Penalization selects the most differentially expressed genes in each strata.

- More generally: $T \sim \lambda(t|G,E,\beta(T))$
  $\hookrightarrow$ Not directly interpretable.
  $\hookrightarrow$ Association between gene expression and no-carcinogen exposures?

Summing up

- Survival analysis for nested CC: genes that discriminate cases and controls.
- Our goal: genes that discriminate "long" and "short" follow-up times.

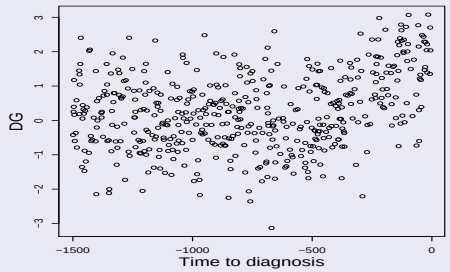
1. Classical prospective GWAS - Nested case-control design

2. Post-GWAS: transcriptomics in a prospective design

3. Statistical approaches for post-GWAS
   - Gene by gene model
   - Latent last-stage model

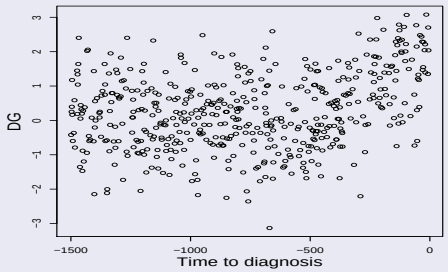
Gene-by-gene model

For each gene



Gene-by-gene model

For each gene



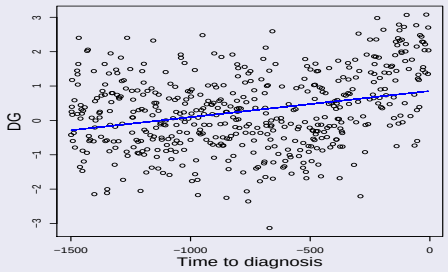
Correlation $(T, \Delta G_g)$

- Spearman test

+ multiple testing

Gene-by-gene model

For each gene



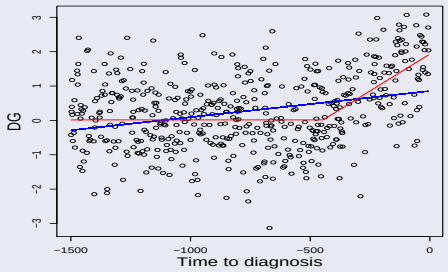
Correlation $(T, \Delta G_g)$

- Spearman test
- Linear model

+ multiple testing

Gene-by-gene model

For each gene



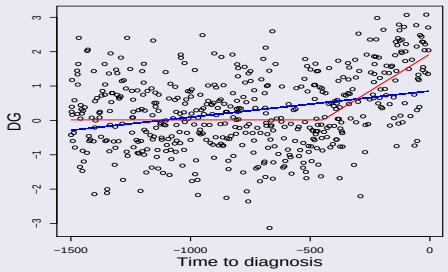
Correlation $(T, \Delta G_g)$

- Spearman test
- Linear model
- "Hockey-stick"
- ...

+ multiple testing

Gene-by-gene model

For each gene



Correlation $(T, \Delta G_g)$

- Spearman test
- Linear model
- "Hockey-stick"
- ...

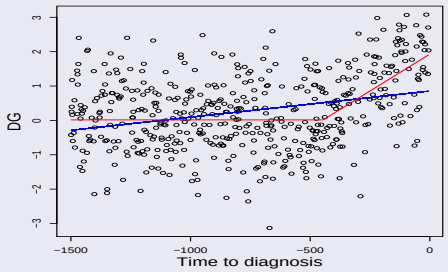
+ multiple testing

Correct for exposures

$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g \Delta E_i + \varphi(T_i | \alpha_2^g) + \varepsilon_{i,g}$$

Gene-by-gene model

For each gene



Correlation $(T, \Delta G_g)$

- Spearman test
- Linear model
- "Hockey-stick"
- ...

+ multiple testing

Correct for exposures

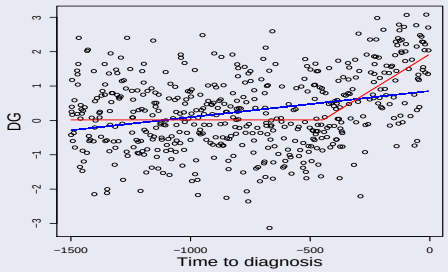
$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g \Delta E_i + \varphi(T_i | \alpha_2^g) + \varepsilon_{i,g}$$

General model

$$\Delta G_{i,g} = \Psi(T_i, \Delta E_i | \Theta_g) + \varepsilon_{i,g}$$

Gene-by-gene model

For each gene



Correlation $(T, \Delta G_g)$

- Spearman test
- Linear model
- "Hockey-stick"
- ...

+ multiple testing

Correct for exposures

$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g \Delta E_i + \varphi(T_i|\alpha_2^g) + \varepsilon_{i,g}$$

General model

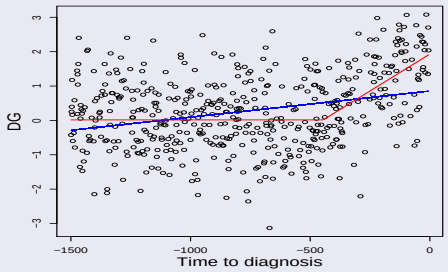
$$\Delta G_{i,g} = \Psi(T_i, \Delta E_i|\Theta_g) + \varepsilon_{i,g}$$

Flexible

- Cross-effect: cancer driven by exposures
- Hierarchical testing: pathways of genes ...

Gene-by-gene model

For each gene



Correlation $(T, \Delta G_g)$

- Spearman test
- Linear model
- "Hockey-stick"
- ...

+ multiple testing

Correct for exposures

$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g \Delta E_i + \varphi(T_i|\alpha_2^g) + \varepsilon_{i,g}$$

General model

$$\Delta G_{i,g} = \Psi(T_i, \Delta E_i|\Theta_g) + \varepsilon_{i,g}$$

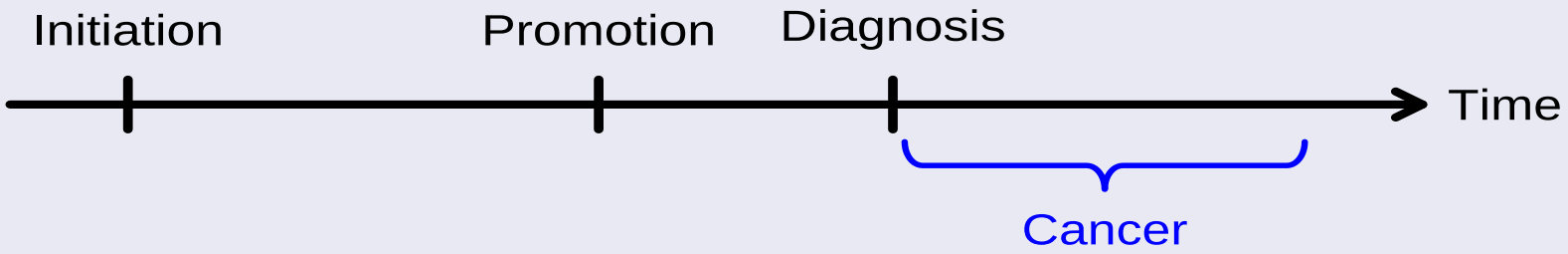
Flexible

- Cross-effect: cancer driven by exposures
- Hierarchical testing: pathways of genes ...

No account for individual dynamics

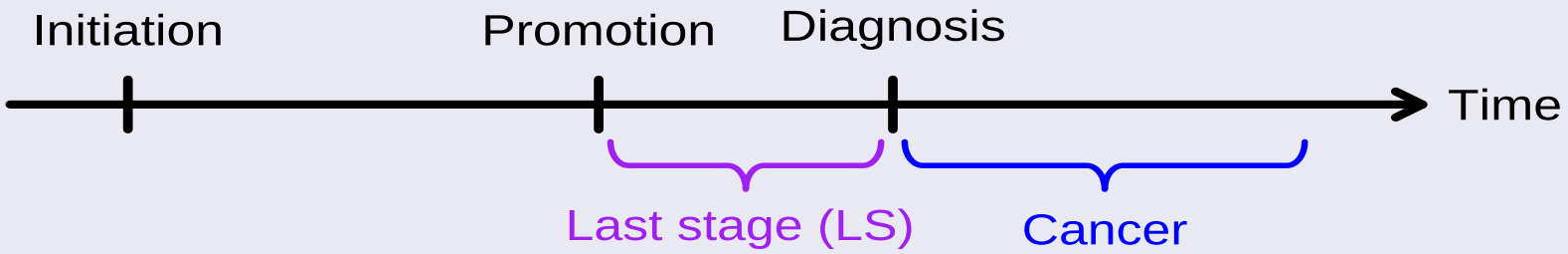
Latent last-stage model

Multi-stage model of carcinogenesis



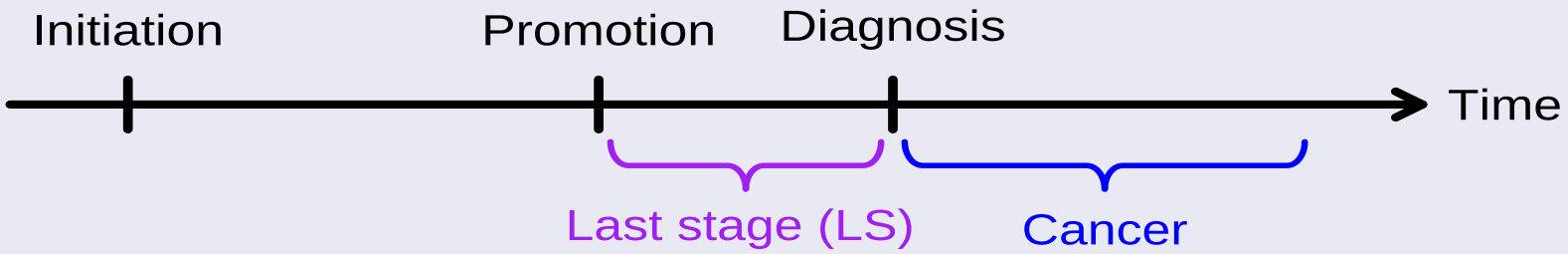
Latent last-stage model

Multi-stage model of carcinogenesis



Latent last-stage model

Multi-stage model of carcinogenesis

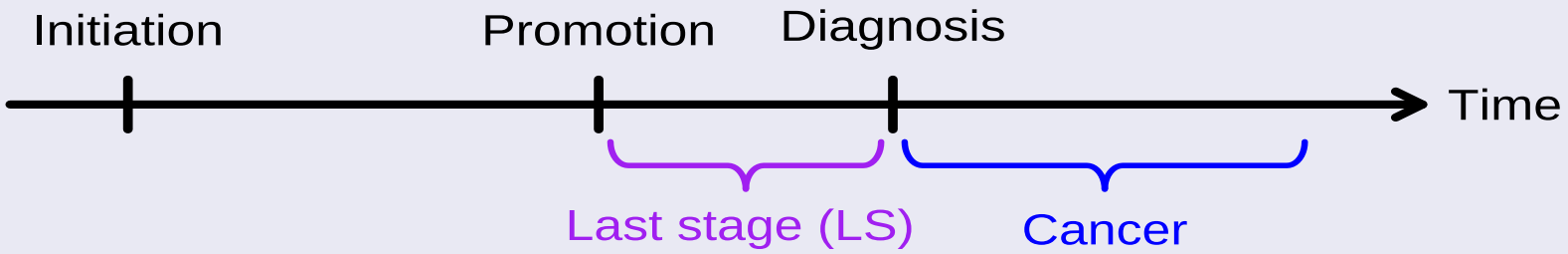


Multi-stage model and gene expression

- Last stage: genes involved in carcinogenesis over/under express.
- Random last stage length.

Latent last-stage model

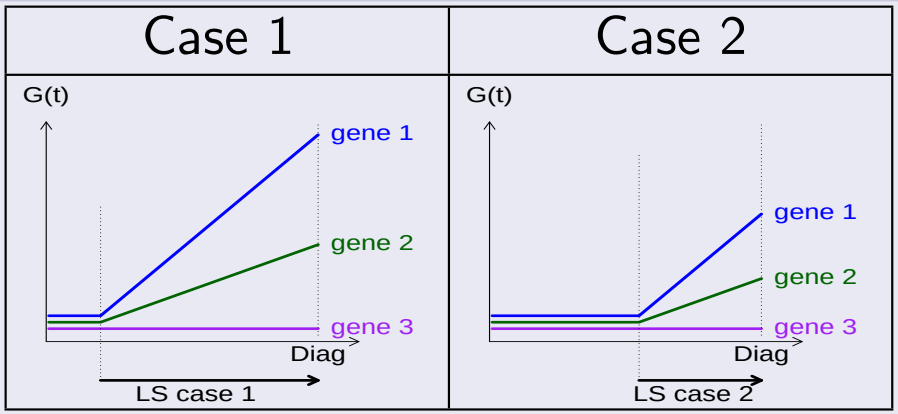
Multi-stage model of carcinogenesis



Multi-stage model and gene expression

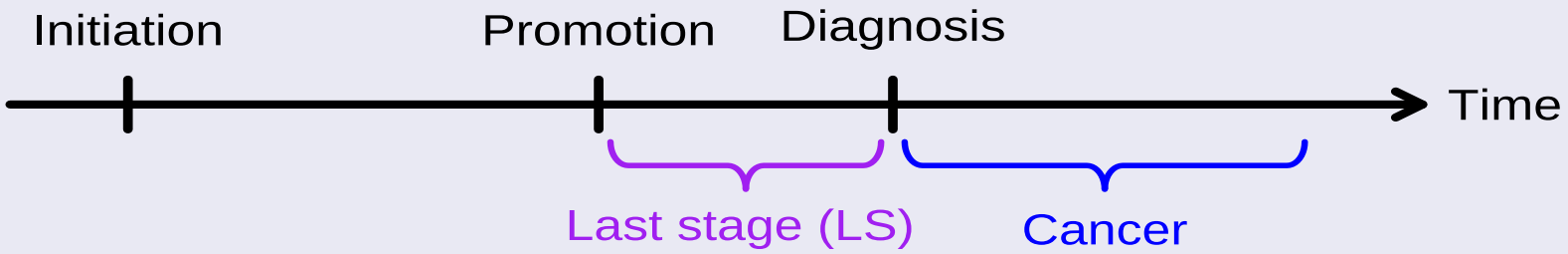
- Last stage: genes involved in carcinogenesis over/under express.
- Random last stage length.

Model 1



Latent last-stage model

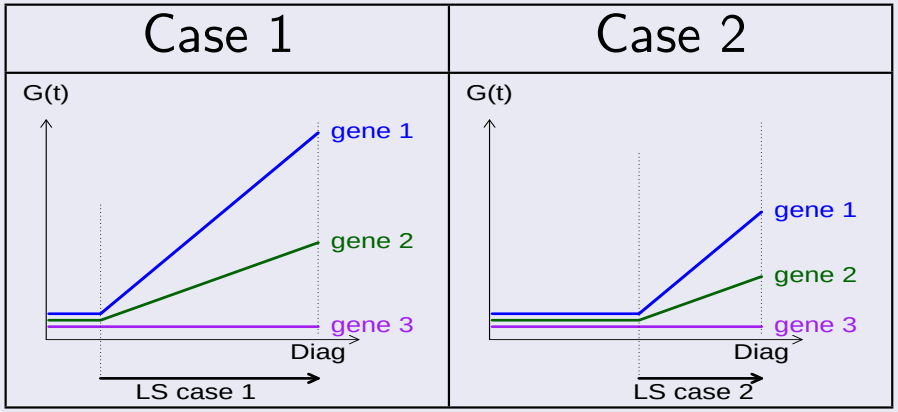
Multi-stage model of carcinogenesis



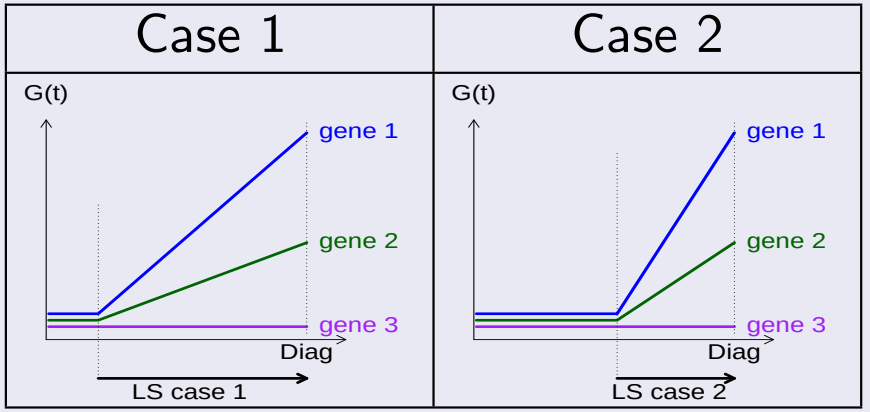
Multi-stage model and gene expression

- Last stage: genes involved in carcinogenesis over/under express.
- Random last stage length.

Model 1

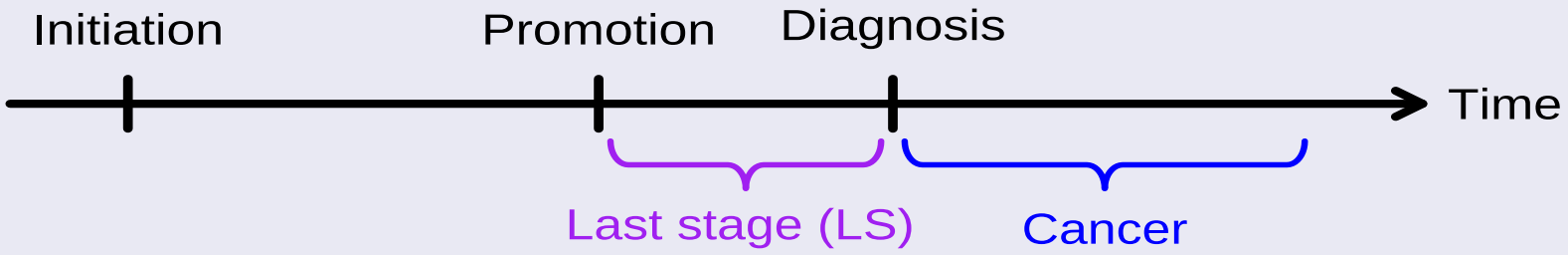


Model 2



Latent last-stage model

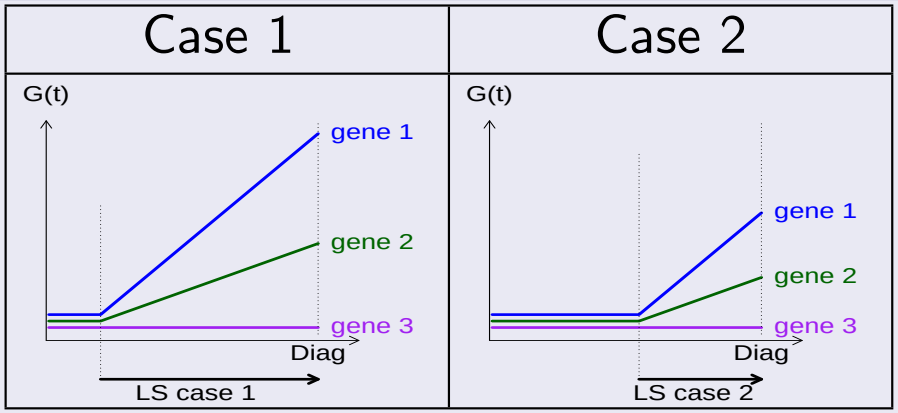
Multi-stage model of carcinogenesis



Multi-stage model and gene expression

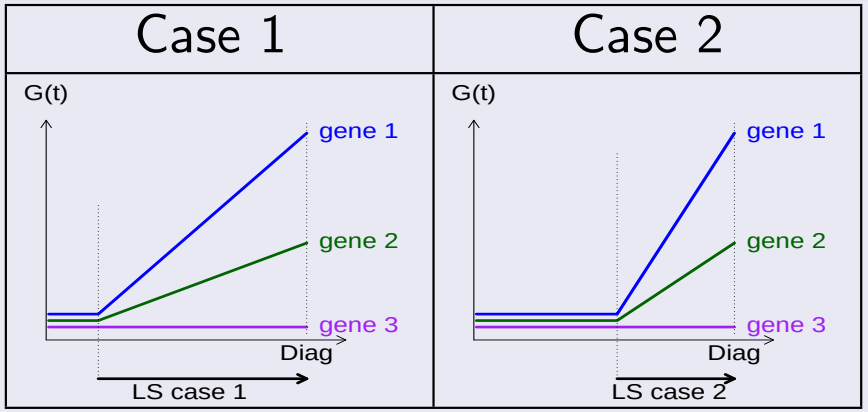
- Last stage: genes involved in carcinogenesis over/under express.
- Random last stage length.

Model 1



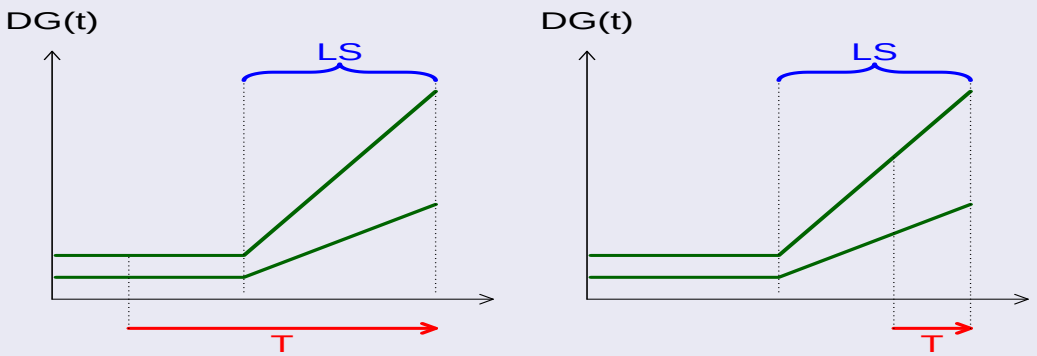
↪ Screening program

Model 2

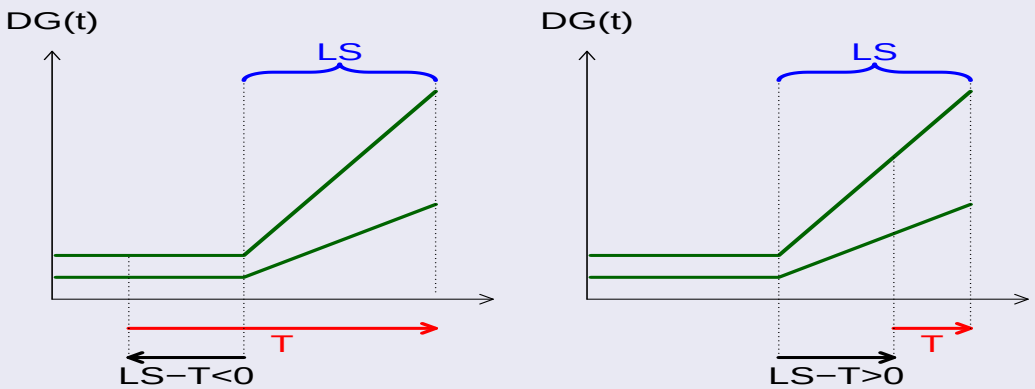


↪ Diagnosis from symptoms

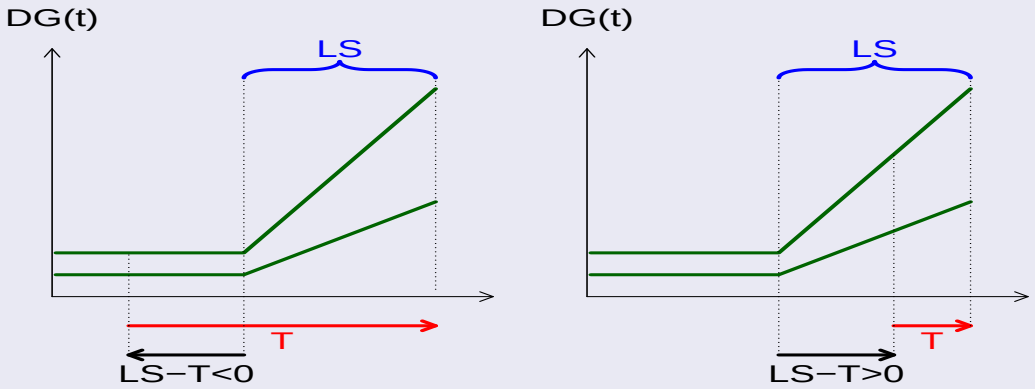
Statistical model 1



Statistical model 1



Statistical model 1

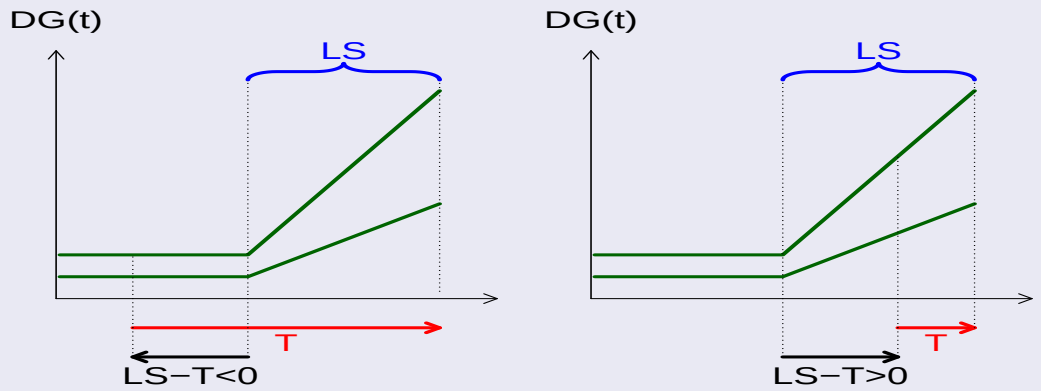


Gene expression

For each case-control pair i and gene g:

$$\Delta G_i^g \;=\; \beta_0^g \;+\; \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i) \mathbb{1}(LS_i > T_i) \;+\; \varepsilon_{i,g}$$

Statistical model 1



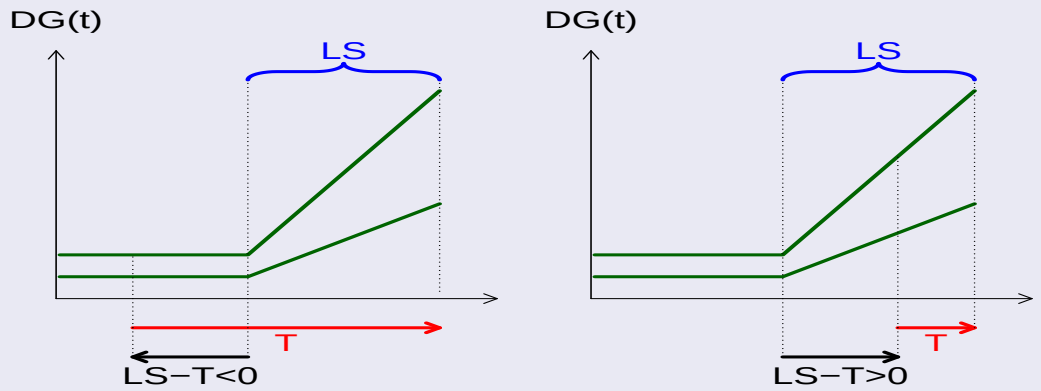
- β_0^g: DE before last stage.

Gene expression

For each case-control pair i and gene g:

$$\Delta G_i^g = \beta_0^g + \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i)_i \mathbb{1}(LS_i > T_i) + \varepsilon_{i,g}$$

Statistical model 1



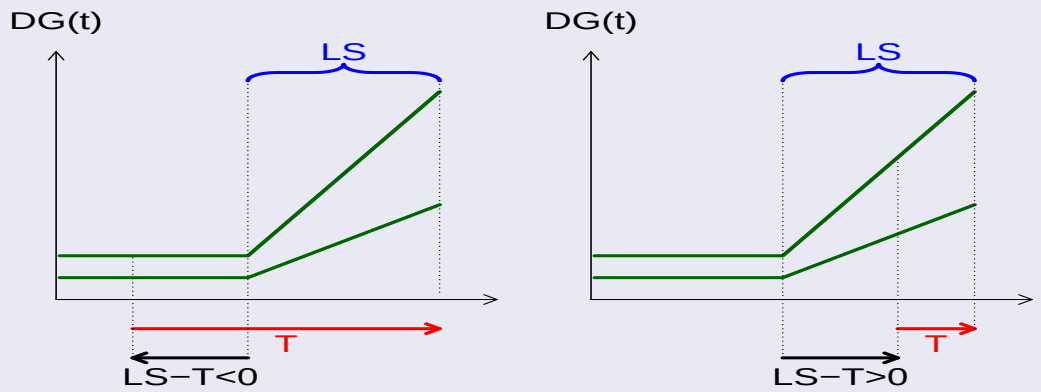
- β_0^g: DE before last stage.
- β_1^g: Exposure effect.

Gene expression

For each case-control pair i and gene g:

$$\Delta G_i^g = \beta_0^g + \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i) \mathbb{1}(LS_i > T_i) + \varepsilon_{i,g}$$

Statistical model 1



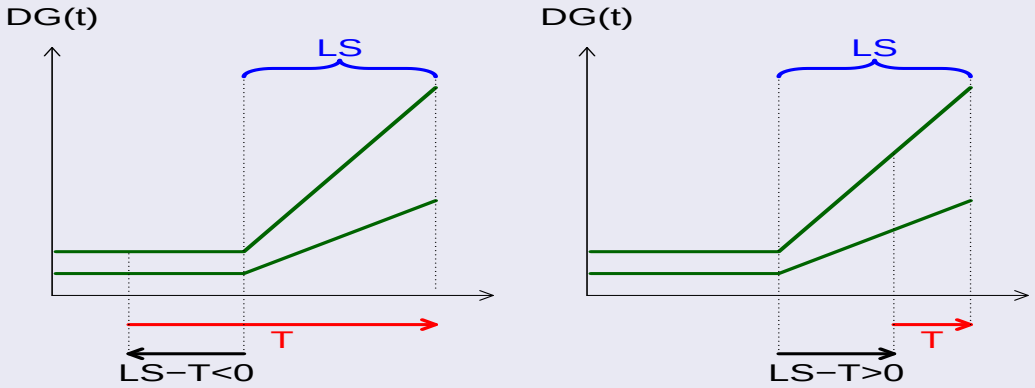
- β_0^g: DE before last stage.
- β_1^g: Exposure effect.
- $\beta_2^g \neq 0$ iif gene g is involved in LS.

Gene expression

For each case-control pair i and gene g:

$$\Delta G_i^g = \beta_0^g + \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i) \mathbb{1}(LS_i > T_i) + \varepsilon_{i,g}$$

Statistical model 1



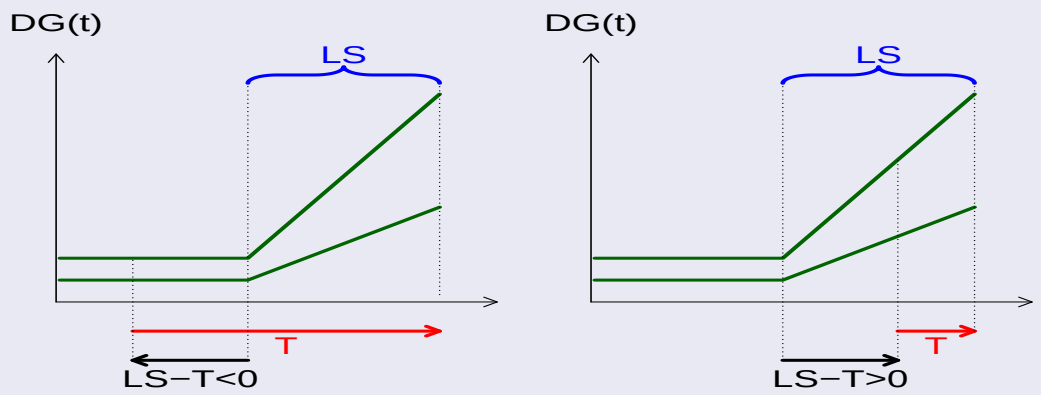
- β_0^g: DE before last stage.
- β_1^g: Exposure effect.
- $\beta_2^g \neq 0$ iif gene g is involved in LS.
- $\varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2)$.

Gene expression

For each case-control pair i and gene g:

$$\Delta G_i^g = \beta_0^g + \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i) \mathbb{1}(LS_i > T_i) + \varepsilon_{i,g}$$

Statistical model 1



- β_0^g: DE before last stage.
- β_1^g: Exposure effect.
- $\beta_2^g \neq 0$ iif gene g is involved in LS.
- $\varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2)$.

Gene expression

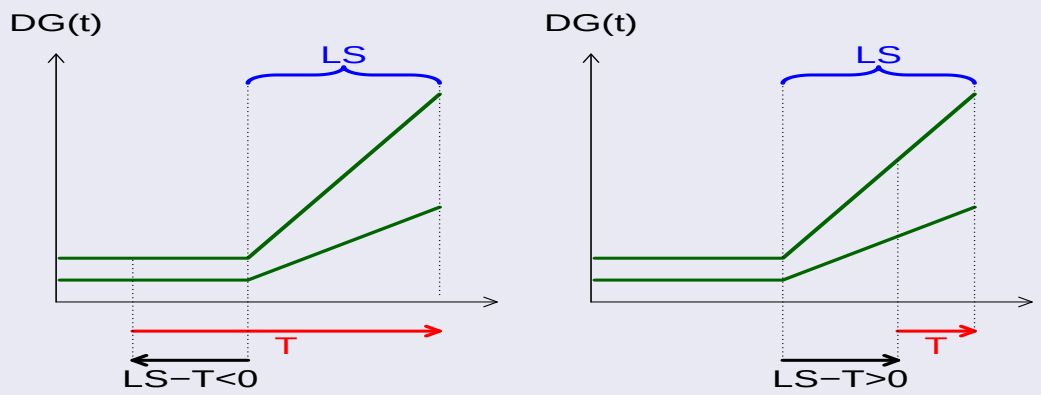
For each case-control pair i and gene g:

$$\Delta G_i^g = \beta_0^g + \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i) \mathbb{1}(LS_i > T_i) + \varepsilon_{i,g}$$

Last-stage length

$$LS_i \sim \Gamma(k, \theta)$$

Statistical model 1



- β_0^g: DE before last stage.
- β_1^g: Exposure effect.
- $\beta_2^g \neq 0$ iif gene g is involved in LS.
- $\varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2)$.

Gene expression

For each case-control pair i and gene g:

$$\Delta G_i^g \;=\; \beta_0^g \;+\; \langle \beta_1^g, \Delta E_i \rangle + \beta_2^g (LS_i - T_i) \mathbb{1}\!\mathrm{l}(LS_i > T_i) \;+\; \varepsilon_{i,g}$$

Last-stage length

$$LS_i \sim \Gamma(k, \theta)$$

$\hookrightarrow$ (k, θ) may depends on the exposures of the case.

$$\begin{cases} k = 1 + \exp(\langle \kappa, (1, E_i^{\text{case}}) \rangle), \\ \theta = \exp(\langle \tau, (1, E_i^{\text{case}}) \rangle). \end{cases}$$

Parameter estimation

Model

$$\begin{cases} LS_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, \Delta E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2) \end{cases}$$

Parameter estimation

Model

$$\begin{cases} LS_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, \Delta E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2) \end{cases}$$

1. Starting point from an heuristic.

Parameter estimation

Model

$$\begin{cases} LS_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, \Delta E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2) \end{cases}$$

1. Starting point from an heuristic.
2. j$^{\text{th}}$ iteration.

$\Theta^{(j)} = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$ $\longrightarrow$ Sample $(LS_i^{(j),1}, \ldots, LS_i^{(j),N})$ from $\mathbb{P}_{\Theta^{(j)}}[LS|\Delta G, T, E^{\text{case}}, \Delta E]$

$\downarrow$

$(\beta^{(j+1)}, \sigma^{(j+1)})$ MLE from $\mathbb{P}_{\beta,\sigma}[\Delta G|\Delta E, LS^{(j)}]$

$(\kappa^{(j+1)}, \tau^{(j+1)})$ MLE from $\mathbb{P}_{\kappa,\tau}[LS^{(j)}|, E^{\text{case}}, T_i]$

$\nwarrow$

Parameter estimation

Model

$$\begin{cases} LS_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, \Delta E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g^2) \end{cases}$$

1. Starting point from an heuristic.
2. j$^{\text{th}}$ iteration.

$\Theta^{(j)} = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$ $\longrightarrow$ Sample $(LS_i^{(j),1}, \ldots, LS_i^{(j),N})$ from $\mathbb{P}_{\Theta^{(j)}}[LS|\Delta G, T, E^{\text{case}}, \Delta E]$

$\downarrow$

$(\beta^{(j+1)}, \sigma^{(j+1)})$ MLE from $\mathbb{P}_{\beta,\sigma}[\Delta G|\Delta E, LS^{(j)}]$
$(\kappa^{(j+1)}, \tau^{(j+1)})$ MLE from $\mathbb{P}_{\kappa,\tau}[LS^{(j)}|, E^{\text{case}}, T_i]$

3. $\widehat{\Theta} = \sum_{j \geqslant \text{burn-in}} \Theta^{(j)}$.

Algorithm SEM

$$\begin{cases} LS_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ \Delta G_i^g = \langle \beta^g, (1, E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

Algorithm SEM

$$\begin{cases} LS_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ \Delta G_i^g = \langle \beta^g, (1, E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

Let $\Theta^{(j)} = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$

Algorithm SEM

$$\begin{cases} LS_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1+\exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ \Delta G_i^g = \langle \beta^g, (1, E_i, (LS_i - T_i)^*) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

Let $\Theta^{(j)} = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$

Simulated expectation

$$\mathbb{E}_{\Theta^{(j)}}[\log \mathbb{P}_\Theta[\Delta G_i, LS_i]] = \sum_{i=1}^{n} \int_{LS_i} \log \mathbb{P}_\Theta[\Delta G_i, LS_i] \mathbb{P}_{\Theta^{(j)}}[LS_i | \Delta G_i]$$

Sample N repetitions of $\{LS_i^{(j)}\}_{i=1:n}$ from distribution $\mathbb{P}_{\Theta^{(j)}}[LS_i|\Delta G_i]$.

$$\begin{aligned} \mathbb{P}_{\Theta^{(j)}}[LS_i|\Delta G_i] &= \frac{\mathbb{P}_{\Theta^{(j)}}[\Delta G_i|LS_i] \cdot \mathbb{P}_{\Theta^{(j)}}[LS_i]}{\mathbb{P}_{\Theta^{(j)}}[\Delta G_i]} \\ &\propto \Pi_{g=1}^{p} \underbrace{\mathbb{P}_{\Theta^{(j)}}[\Delta G_i^g|LS_i]}_{\mathcal{N}(\langle \beta_j^g, (1, E_i, (LS_i - T_i)^*) \rangle, \sigma_g)} \cdot \underbrace{\mathbb{P}_{\Theta^{(j)}}[LS_i]}_{\Gamma(k_j, \theta_j) - T_i} \end{aligned}$$

Maximization

$$\log \mathbb{P}_{\Theta}[\Delta G_i, LS_i] = \log \mathbb{P}_{\beta,\sigma}[\Delta G_i | LS_i]] + \log \mathbb{P}_{\kappa,\tau}[LS_i]$$

Thus

$$(\beta_g^{(j+1)}, \sigma_g^{(j+1)}) = \arg\max \sum_{i=1}^{n} \left(\frac{1}{N} \sum_{\ell=1}^{N} \phi(\Delta G_i^g - \langle \beta_g, (1, E_i, (LS_{i,\ell}^{(j)} - T_i)^*) \rangle) \right)$$

where ϕ is the standard normal density and

$$(\kappa^{(j+1)}, \tau^{(j+1)}) = \arg\max$$
$$\sum_{i=1}^{n} \left(\frac{1}{N} \sum_{\ell=1}^{N} \psi(LS_{i,\ell}^{(j)} | k = 1 + \exp(\langle \kappa, E_i \rangle),\ \theta = \exp(\langle \tau, E_i \rangle) \right)$$

where ψ is the gamma distribution density.

Convergence of the algorithm on simulated data

Simulations

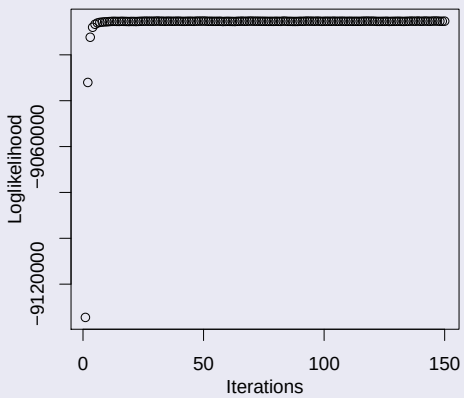
- $n = 150$ pairs, $p = 2000$ genes and $p_0 = 100$ genes involved in the last stage.
- Draw β_1^g, β_2^g from $\mathcal{N}(0, 1)$.
- Draw $(\beta_2^1, \ldots, \beta_2^{p_0})$ from $\mathcal{N}(0, 0.01)$, and $\beta_2^{p_0+1} = \cdots = \beta_2^p = 0$.
- Draw σ from $\chi^2(3)$
- $E =$ binary variable $(0/1)$
- $T =$ uniformly samples in $(0, 800)$
- LS generated with parameters $\tau = c(3, 0.5)$, $\kappa = c(2, 0.5)$.
- ΔG generated from $P_{(\beta,\sigma,\tau,\kappa)}[\Delta G | LS, T, E]$.

Convergence of the algorithm on simulated data

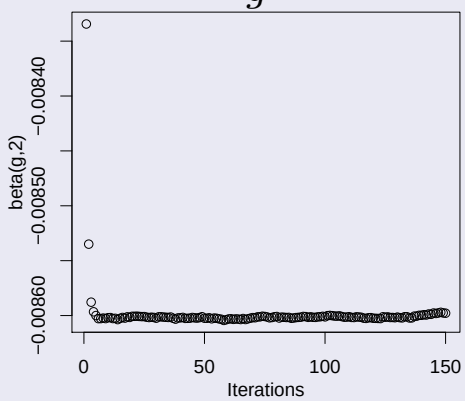
Simulations

- $n = 150$ pairs, $p = 2000$ genes and $p_0 = 100$ genes involved in the last stage.
- Draw β_1^g, β_2^g from $\mathcal{N}(0, 1)$.
- Draw $(\beta_2^1, \ldots, \beta_2^{p_0})$ from $\mathcal{N}(0, 0.01)$, and $\beta_2^{p_0+1} = \cdots = \beta_2^p = 0$.
- Draw σ from $\chi^2(3)$
- $E =$ binary variable (0/1)
- $T =$ uniformly samples in $(0, 800)$
- LS generated with parameters $\tau = c(3, 0.5)$, $\kappa = c(2, 0.5)$.
- ΔG generated from $P_{(\beta,\sigma,\tau,\kappa)}[\Delta G | LS, T, E]$.

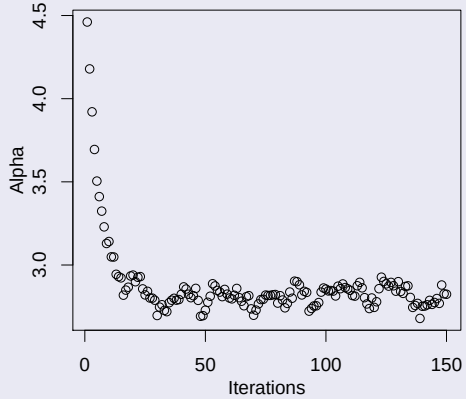
Loglikelihood



β_g^2



τ_1

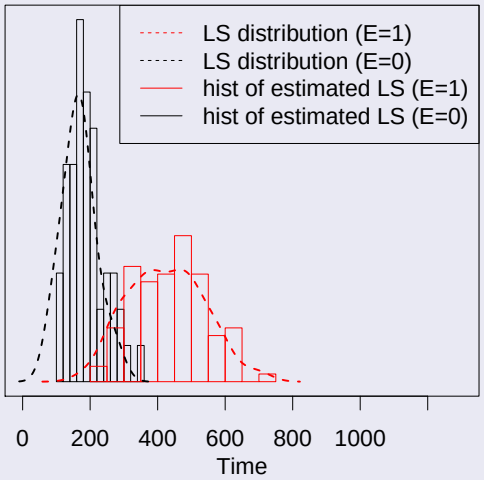


Two quantities of interests

Last-stage length estimation

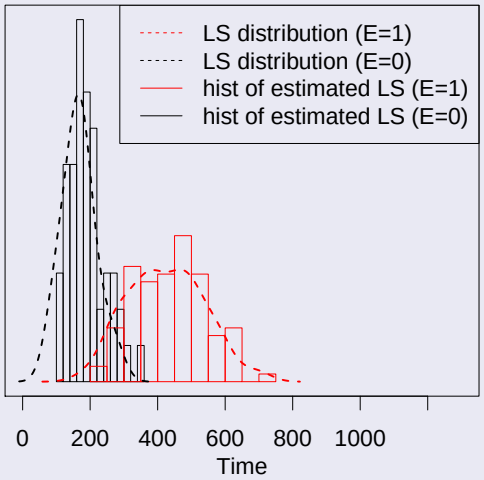
Two quantities of interests

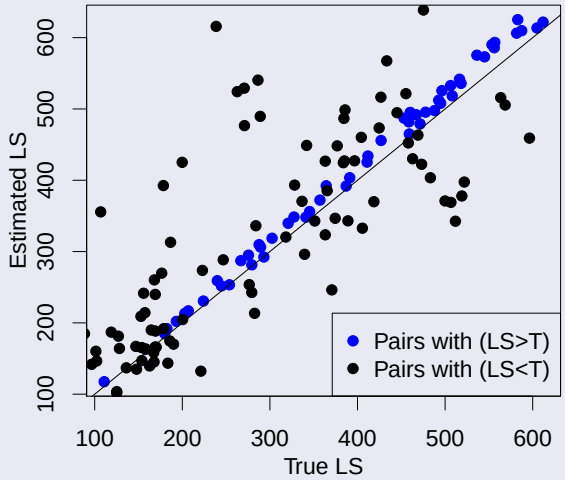
Last-stage length estimation



Two quantities of interests

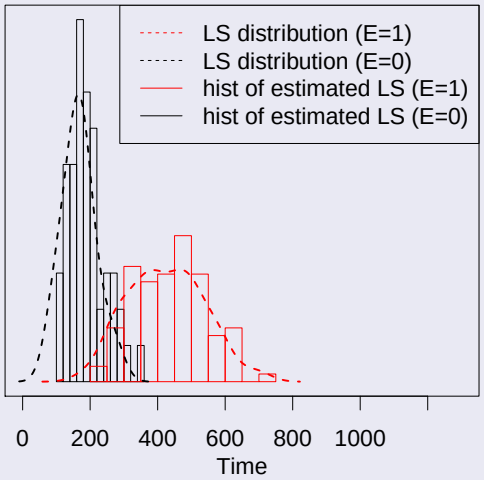
Last-stage length estimation

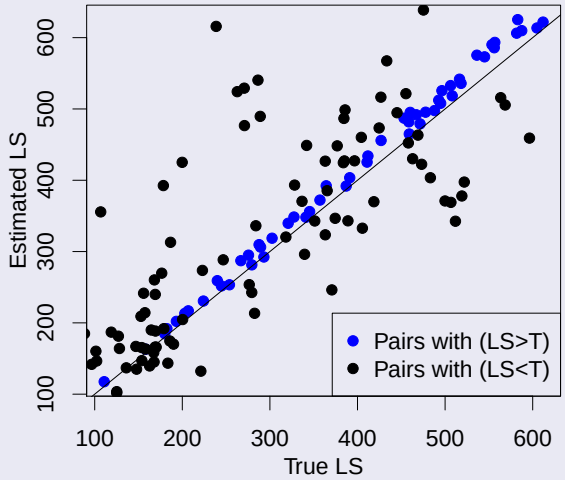




Two quantities of interests

Last-stage length estimation

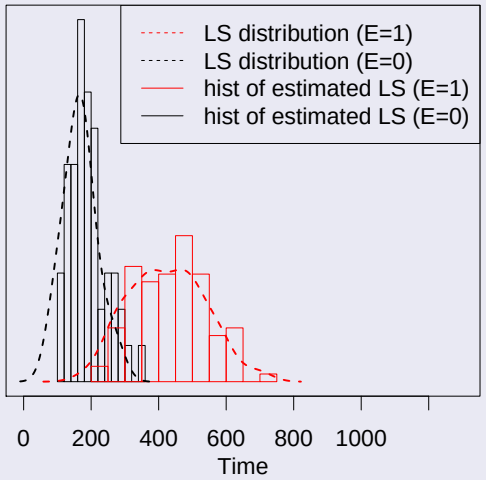


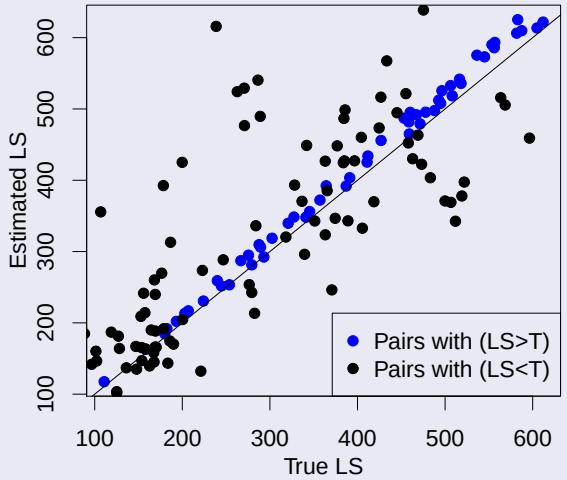


- Discrimination between $E^{\text{case}} = 0$ and $E^{\text{case}} = 1$
- More precise estimation when $LS_i > T_i$.

Two quantities of interests

Last-stage length estimation



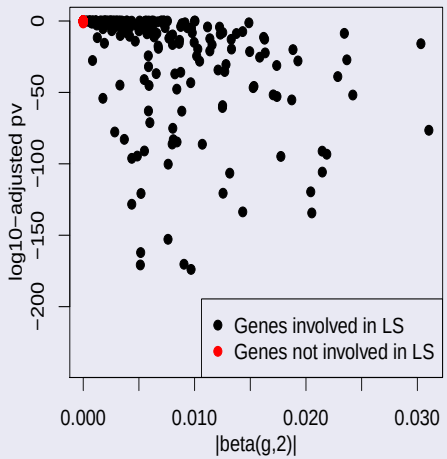


- Discrimination between $E^{\text{case}} = 0$ and $E^{\text{case}} = 1$
- More precise estimation when $LS_i > T_i$.

Detection of genes involved in the last-stage

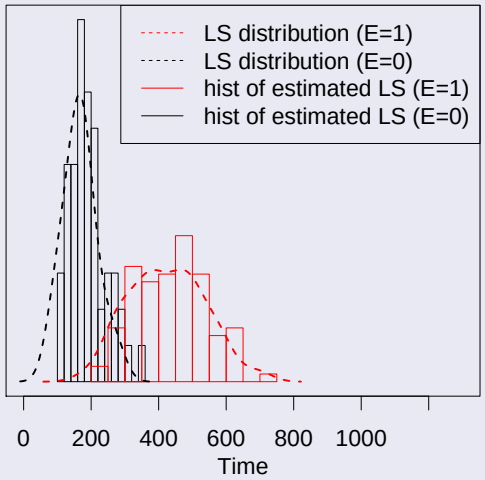
F-test + FDR

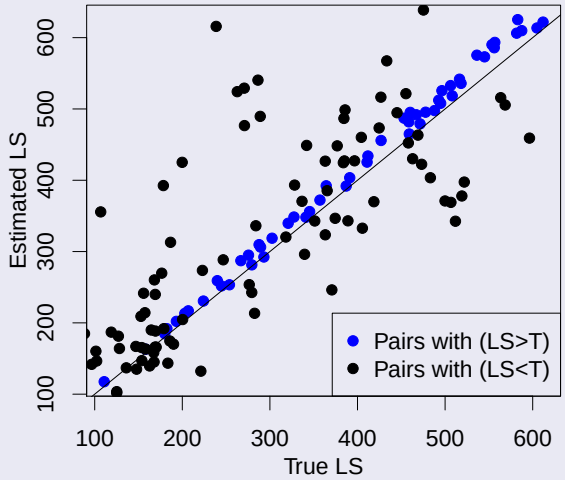
$(\beta_2^g, \log\ \text{pv})$



Two quantities of interests

Last-stage length estimation





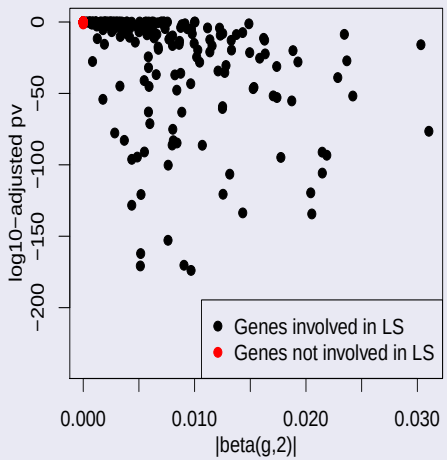
- Discrimination between $E^{\text{case}} = 0$ and $E^{\text{case}} = 1$
- More precise estimation when $LS_i > T_i$.

Detection of genes involved in the last-stage

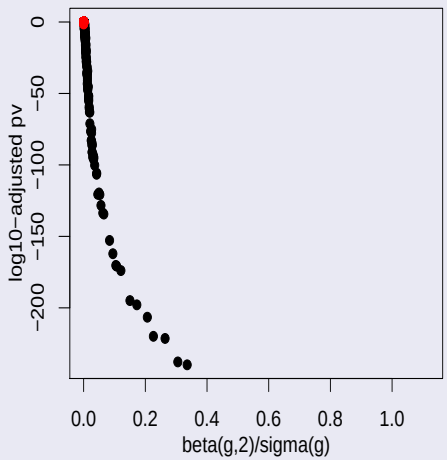
F-test + FDR

- pv depends on signal/noise ratio

$(\beta_2^g, \log\ \text{pv})$

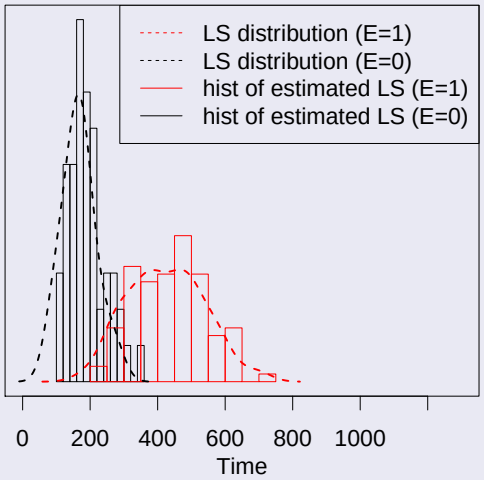


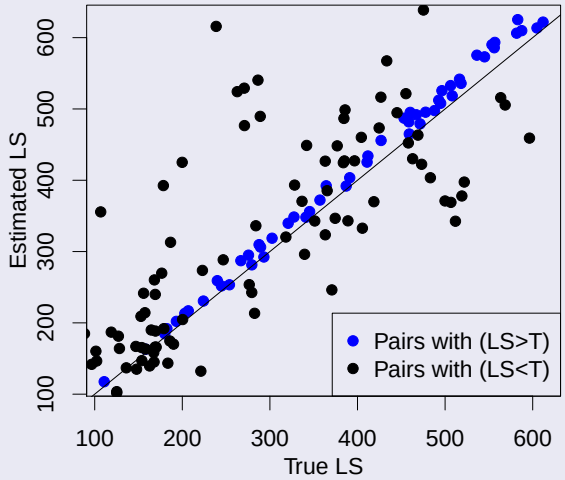
$(\beta_2^g/\sigma^g, \log\ \text{pv})$



Two quantities of interests

Last-stage length estimation





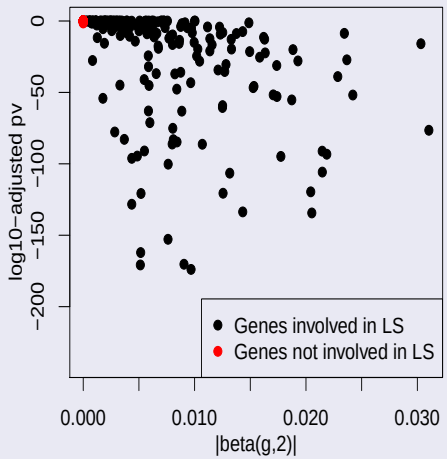
- Discrimination between $E^{\text{case}} = 0$ and $E^{\text{case}} = 1$
- More precise estimation when $LS_i > T_i$.

Detection of genes involved in the last-stage

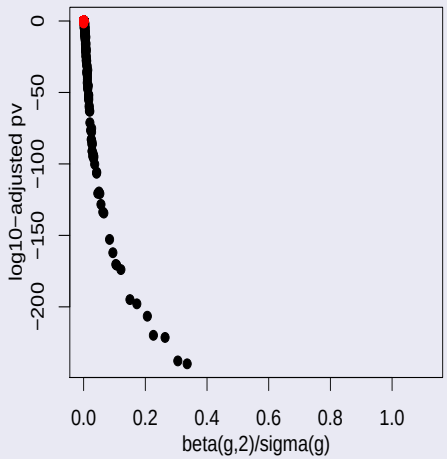
F-test + FDR

- pv depends on signal/noise ratio
- weak impact of exposure + constant effect.

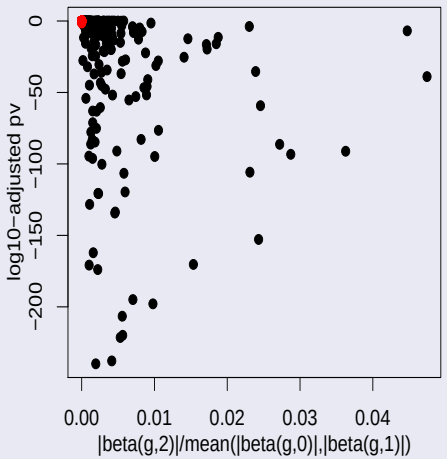
$(\beta_2^g, \log\ \text{pv})$



$(\beta_2^g/\sigma^g, \log\ \text{pv})$



$(\beta_2^g/(\beta_0^g + \beta_1^g), \log \text{pv})$



Comparison with gene-by-gene models

Three tests are compared:

- F-test from LLS model with estimated LS:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 (LS_i - T_i)^* + \varepsilon_{i,g}$$

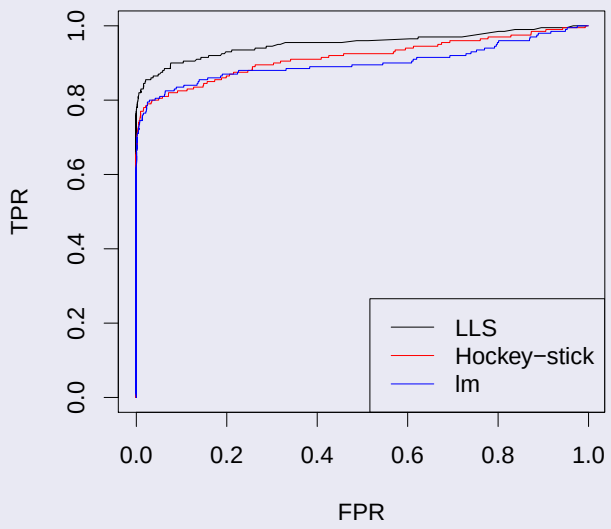
- F-test in linear model:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 T_i + \varepsilon_{i,g}$$

- F-test in hockey-stick model:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 (T_i - t_0)^* + \varepsilon_{i,g}$$

ROC curve



Comparison with gene-by-gene models

Three tests are compared:

- F-test from LLS model with estimated LS:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 (LS_i - T_i)^* + \varepsilon_{i,g}$$

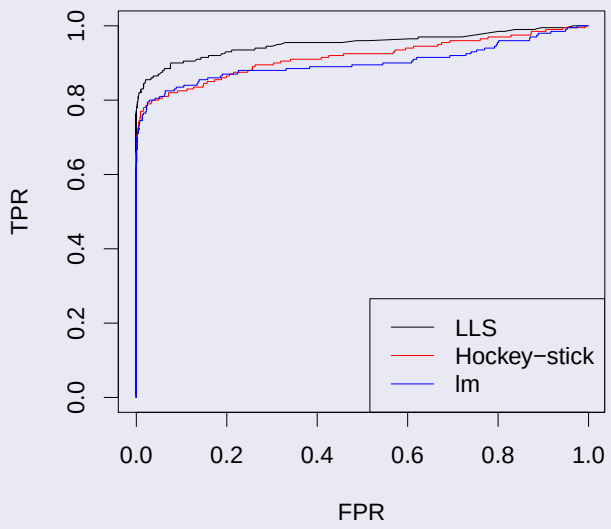
- F-test in linear model:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 T_i + \varepsilon_{i,g}$$

- F-test in hockey-stick model:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 (T_i - t_0)^* + \varepsilon_{i,g}$$

ROC curve



- Slightly better sensitivity for LLS.
  - ◇ data simulated according to LLS model → favorable situation.

Comparison with gene-by-gene models

Three tests are compared:

- F-test from LLS model with estimated LS:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 (LS_i - T_i)^* + \varepsilon_{i,g}$$

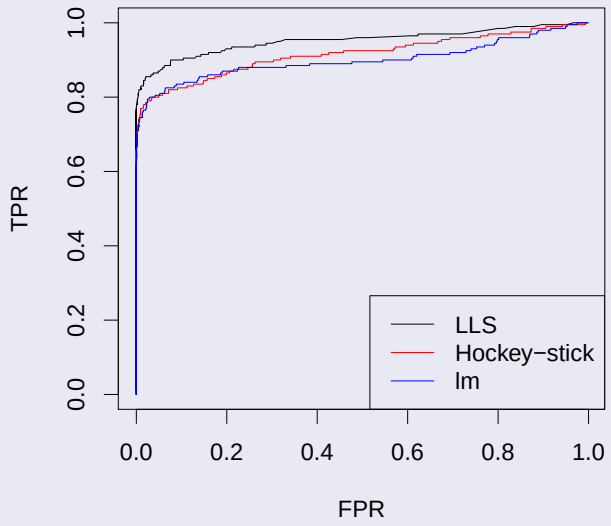
- F-test in linear model:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 T_i + \varepsilon_{i,g}$$

- F-test in hockey-stick model:

$$DG_i^g = \beta_g^0 + \beta_g^1 \Delta E_i + \beta_g^2 (T_i - t_0)^* + \varepsilon_{i,g}$$

ROC curve



- Slightly better sensitivity for LLS.
  - ◇ data simulated according to LLS model → favorable situation.
- LLS model provides additional information about individual dynamics.

What comes next?

What comes next?

Is there signal in blood?

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.
- Whole set of data available soon.

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.
- Whole set of data available soon.

Choice of exposures to correct individual variations.

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.
- Whole set of data available soon.

Choice of exposures to correct individual variations.

- PCA on an independent sample.

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.
- Whole set of data available soon.

Choice of exposures to correct individual variations.

- PCA on an independent sample.

Stratification of cases:

- Stage of cancer (in situ, invasive, metastasic)
- Type of cancer (receptors,...)

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.
- Whole set of data available soon.

Choice of exposures to correct individual variations.

- PCA on an independent sample.

Stratification of cases:

- Stage of cancer (in situ, invasive, metastasic)
- Type of cancer (receptors,...)

Pre-selection of genes:

- A priori biological knowledge
- Analysis at time of diagnosis.

What comes next?

Is there signal in blood?

- At time of diagnosis, a large number of differentially expressed genes in blood.
- Whole set of data available soon.

Choice of exposures to correct individual variations.

- PCA on an independent sample.

Stratification of cases:

- Stage of cancer (in situ, invasive, metastasic)
- Type of cancer (receptors,...)

Pre-selection of genes:

- A priori biological knowledge
- Analysis at time of diagnosis.

Account for dependence between genes.

Conclusion

From GWAS to post-GWAS design:

- New goals: exploration of functional changes on transcriptomic data.
- Novel statistical approaches:

Prospective GWAS: $\mathbb{P}[T|G,E]$ ◆ Post-GWAS: $\mathbb{P}[G|T,E]$

Gene-by-gene model

- Flexible
- Inclusion of biological assumptions.

Latent last stage model

- Validated on simulated data

Require further developments to be applied on data:

- Choice of parametrization and relevant exposures.
- Account for dependence between genes.