Novel statistical approaches to explore carcinogenic process on transcriptomic data - from GWAS to post-GWAS

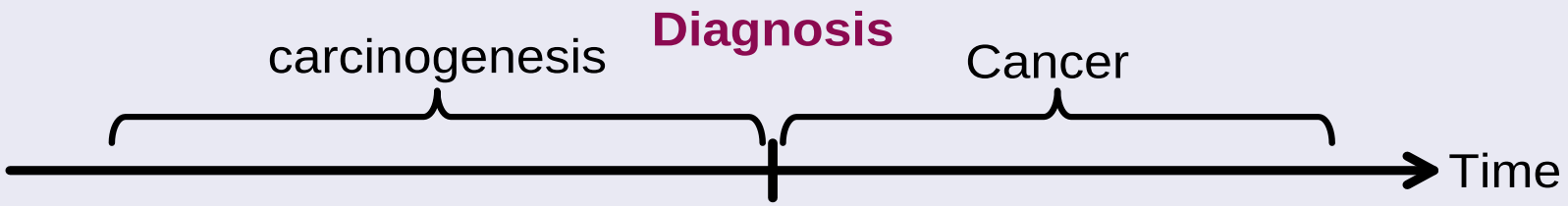
TICE (Transcriptomics In Cancer Epidemiology)
NOWAC (Norwegian Women And Cancer)

Sandra Plancade, University of Tromso (Norway)
Gregory Nuel, University Paris-Descartes
Eiliv Lund, University of Tromso

1st of October 2012

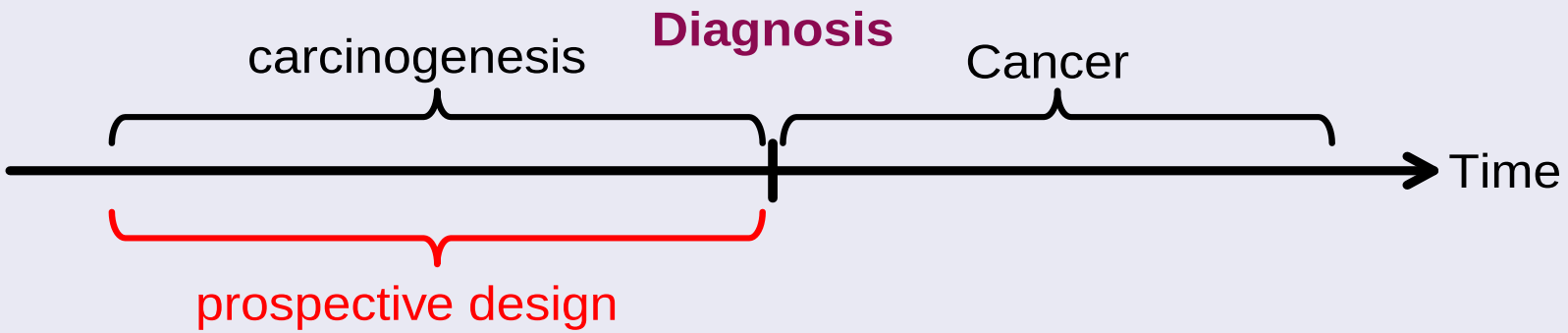
GWAS and post-GWAS designs

GWAS (GenomeWide Association Study)



GWAS and post-GWAS designs

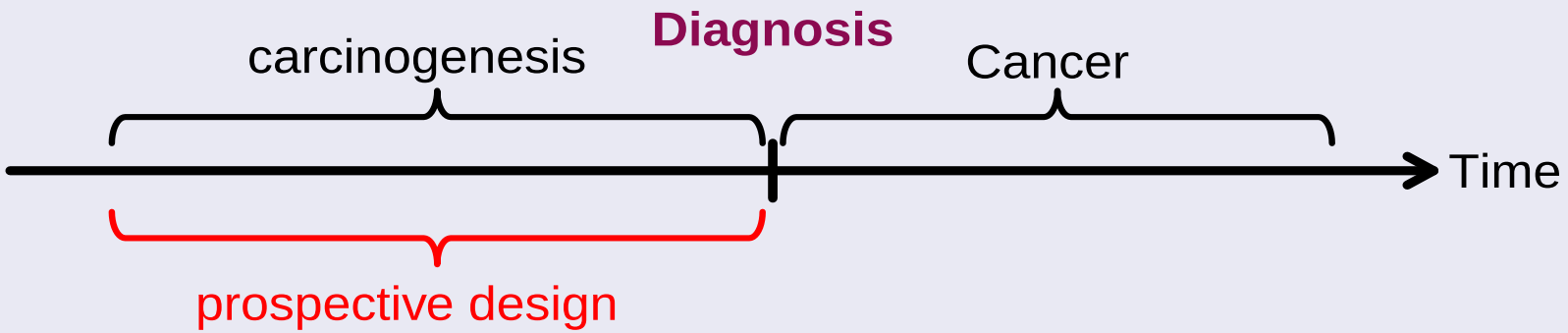
GWAS (GenomeWide Association Study)



- Prospective study: classical epidemiology risk factor (environmental exposures, lifestyle) and genomic data (in particular SNPs)

GWAS and post-GWAS designs

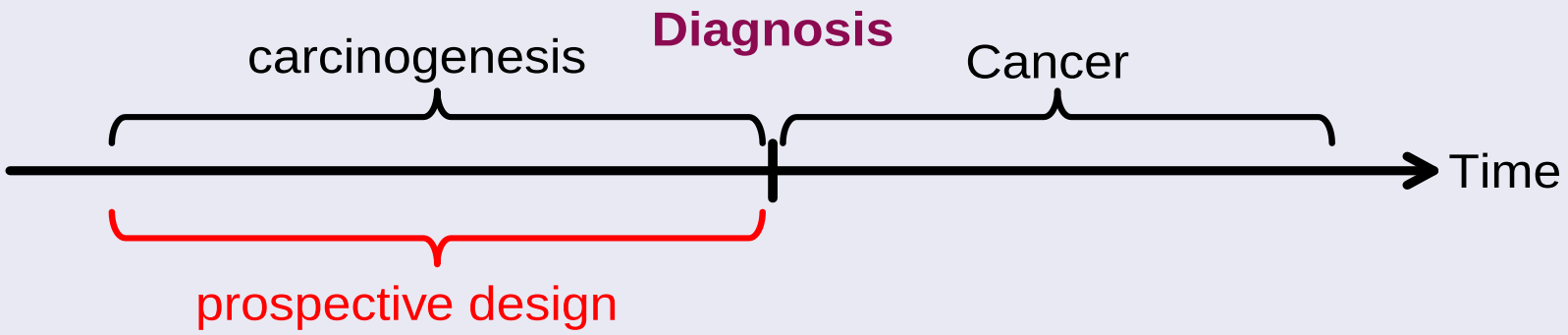
GWAS (GenomeWide Association Study)



- Prospective study: classical epidemiology risk factor (environmental exposures, lifestyle) and genomic data (in particular SNPs)
- Transcriptomic data (gene expression and methylation): at time of diagnosis

GWAS and post-GWAS designs

GWAS (GenomeWide Association Study)



- Prospective study: classical epidemiology risk factor (environmental exposures, lifestyle) and genomic data (in particular SNPs)
- Transcriptomic data (gene expression and methylation): at time of diagnosis

Post-GWAS

Transcriptomic data in a prospective nested CC (case-control) design:

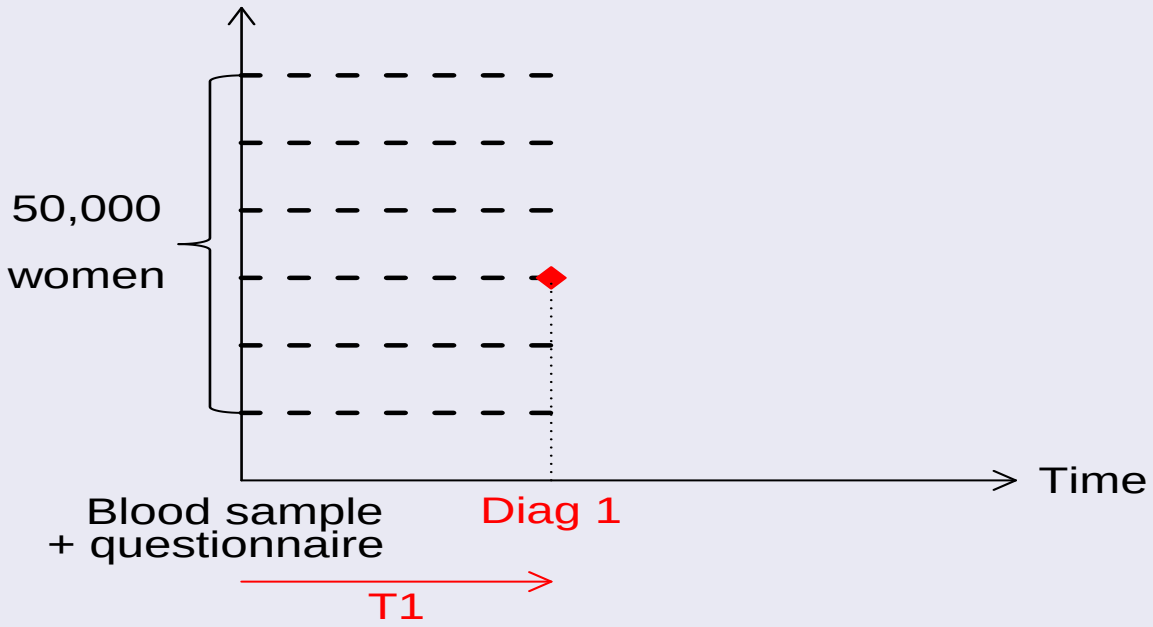
- Hybrid between the prospective and nested CC designs
- Main distinction with prospective GWAS :
  Transcriptomics change over carcinogenic process $\neq$ SNPs are constant.

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design

Example of post-GWAS design: the NOWAC cohort

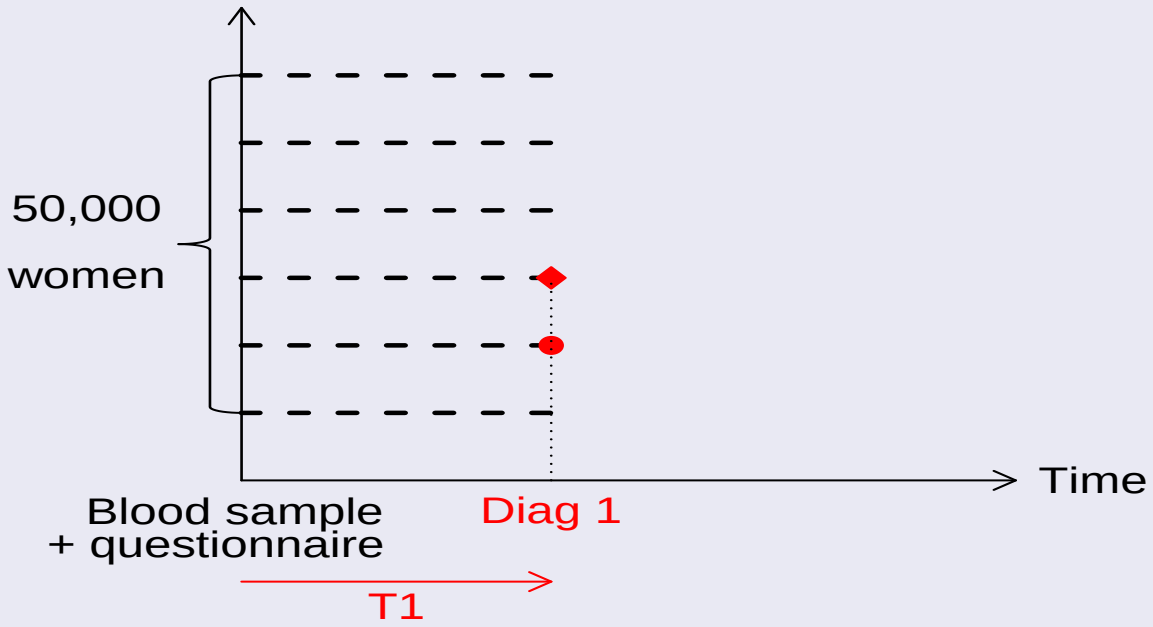
Prospective nested case-control design



◆: case
●: control

Example of post-GWAS design: the NOWAC cohort

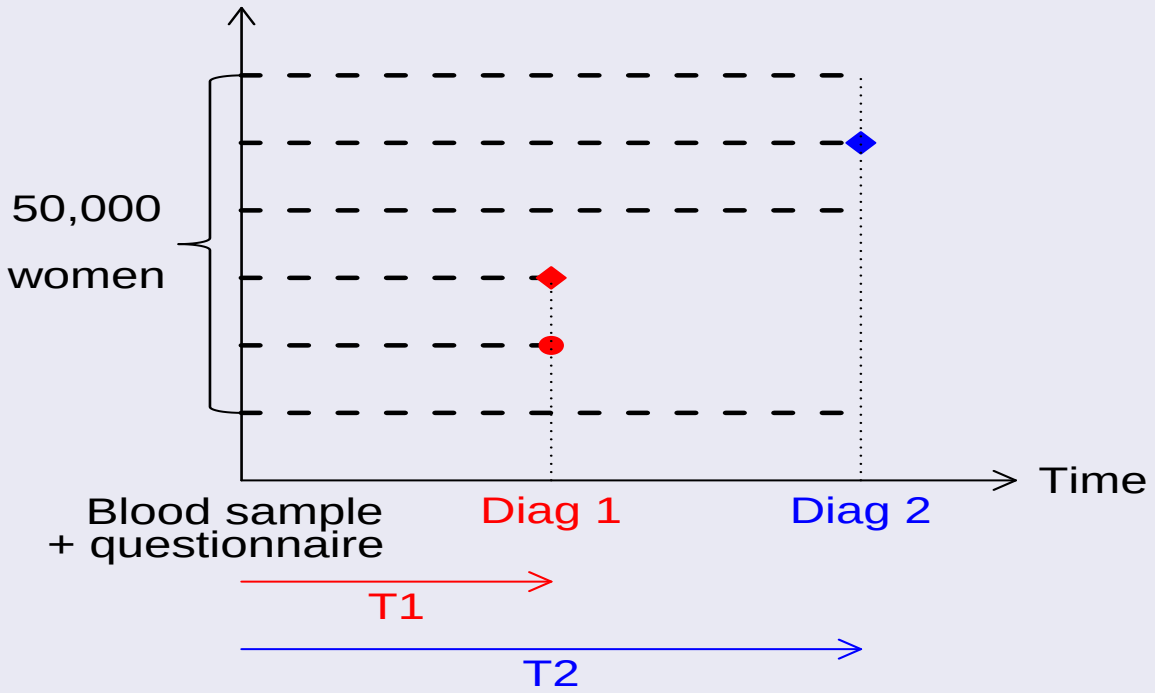
Prospective nested case-control design



- ◆: case
- ●: control

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design

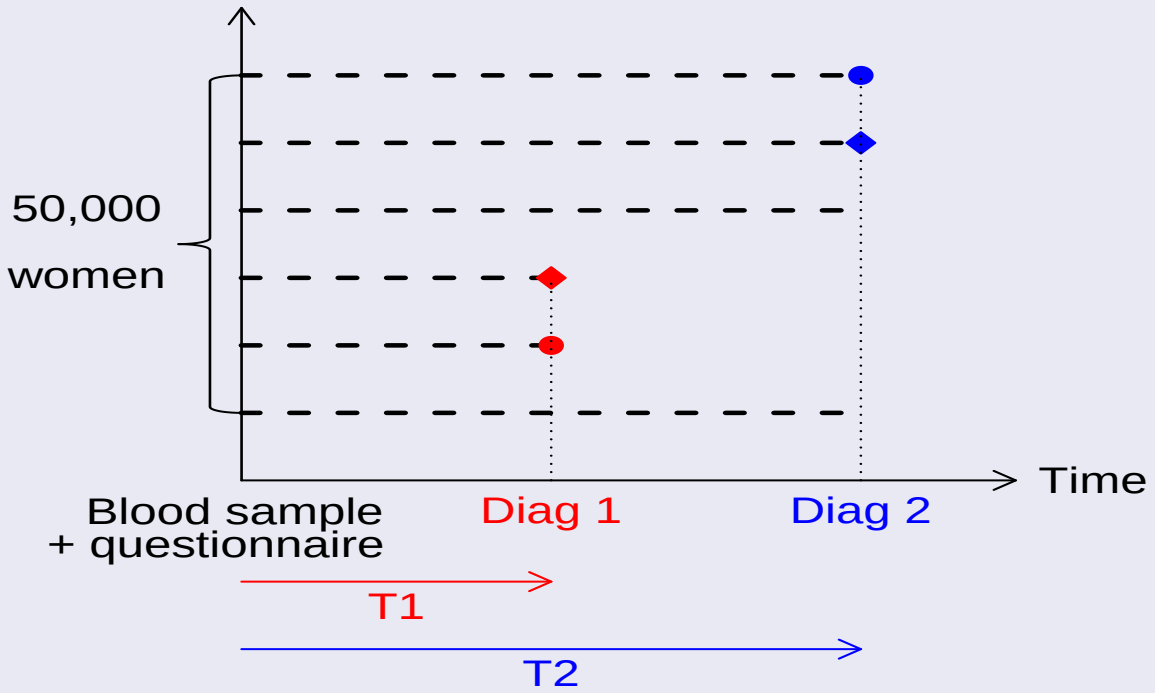


◆: case

●: control

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design

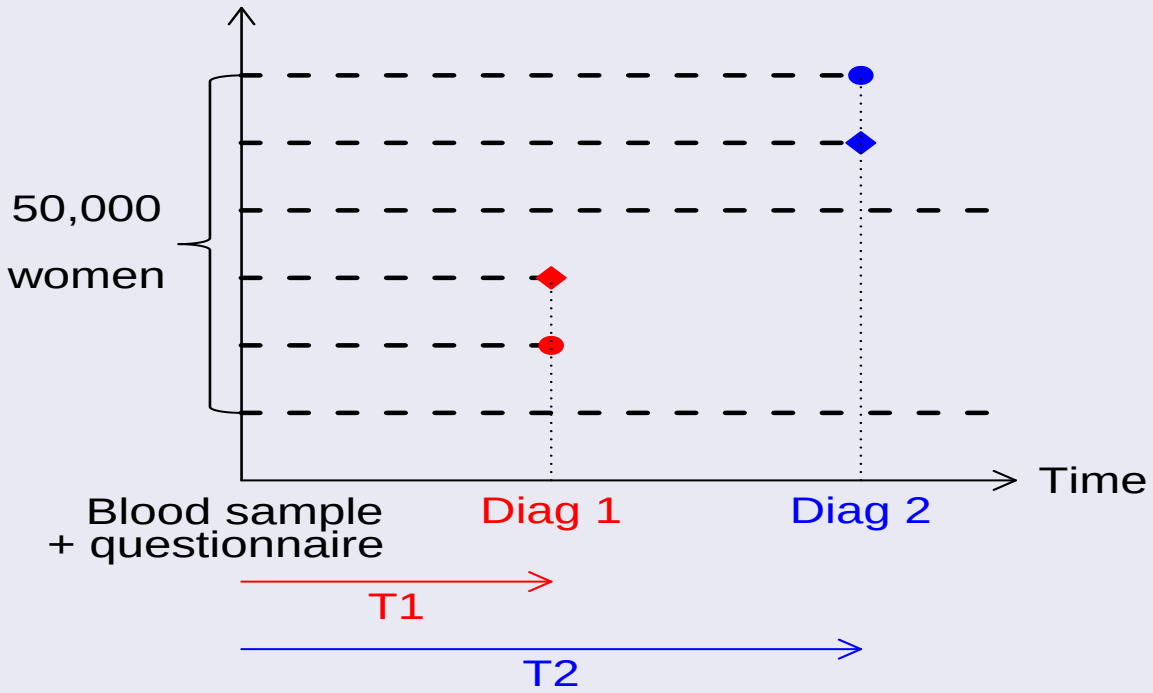


◆: case

●: control

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design

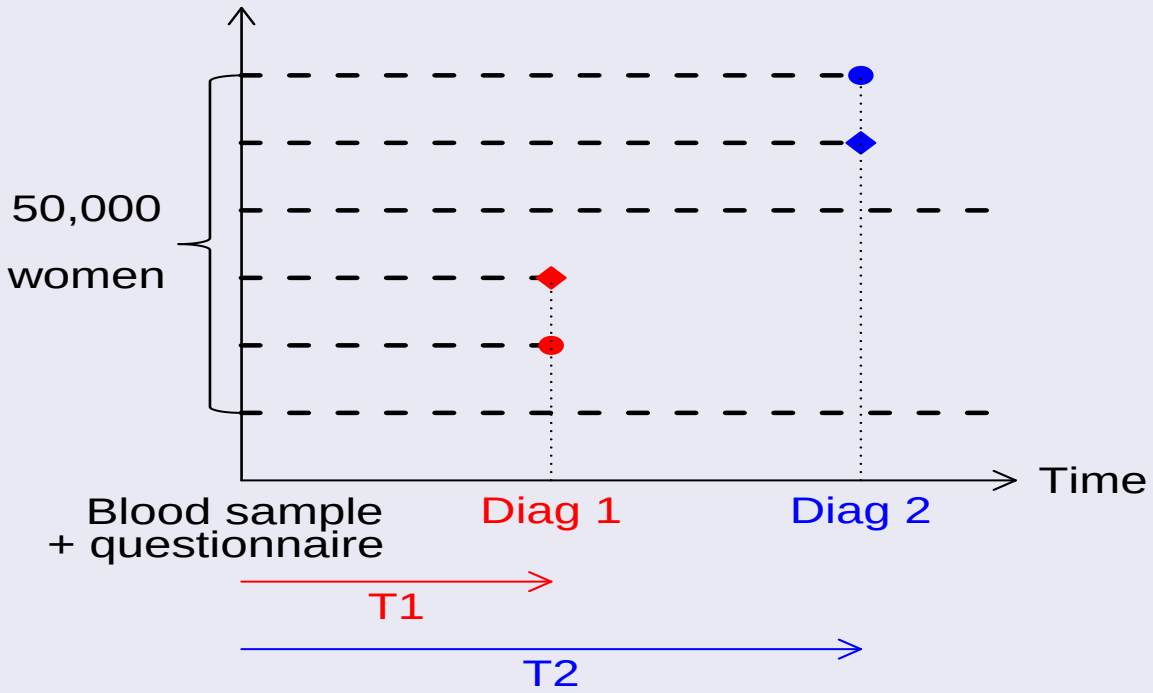


◆: case

●: control

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design



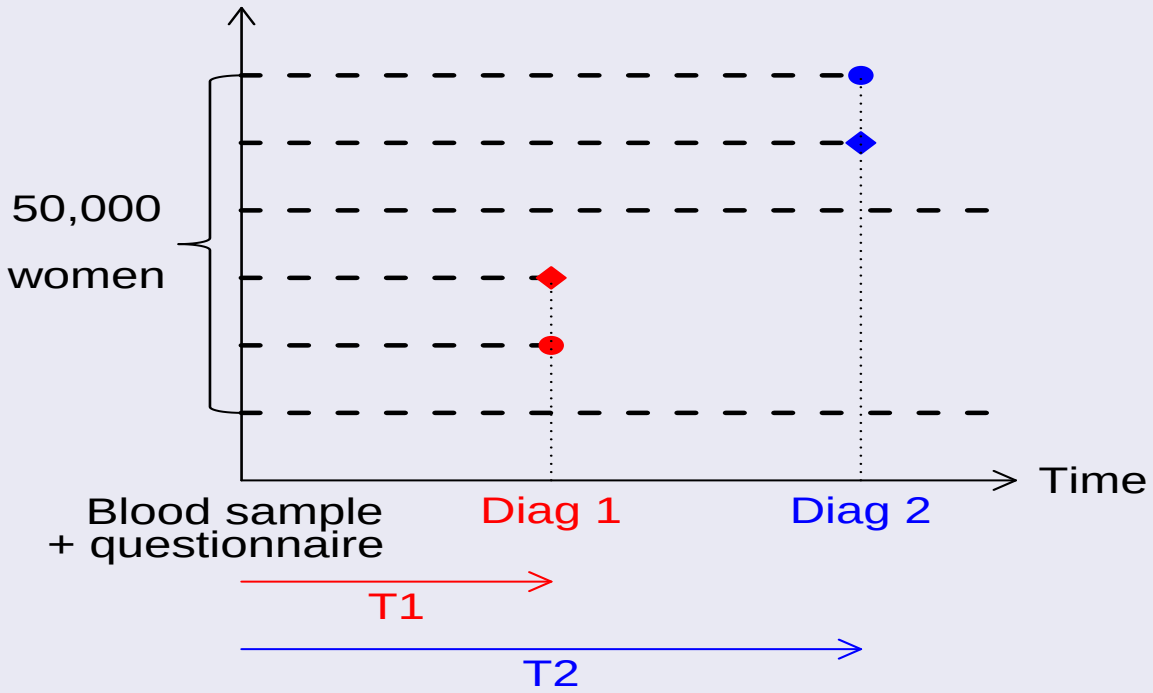
◆: case

●: control

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

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design



◆: case
●: control

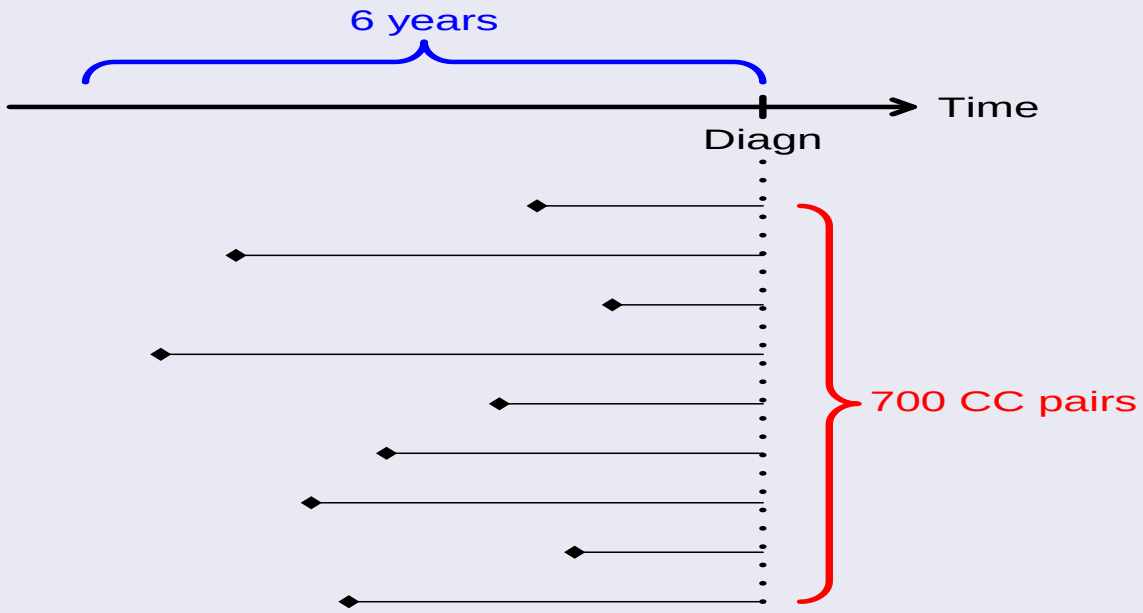
- 6 years of follow-up
- 700 case-control pairs for breast cancer

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.

Example of post-GWAS design: the NOWAC cohort

Prospective nested case-control design



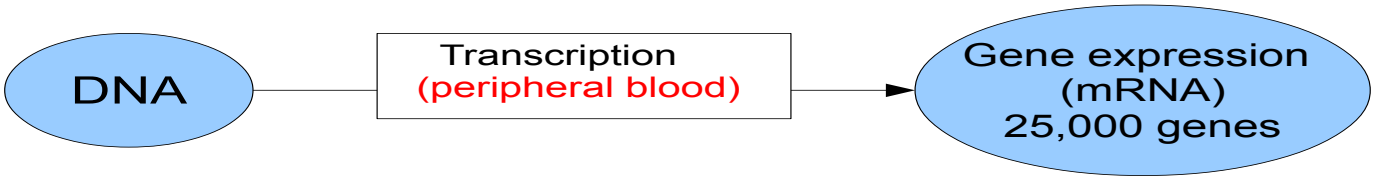
- 6 years of follow-up
- 700 case-control pairs for breast cancer

Data: for each case-control pair i,

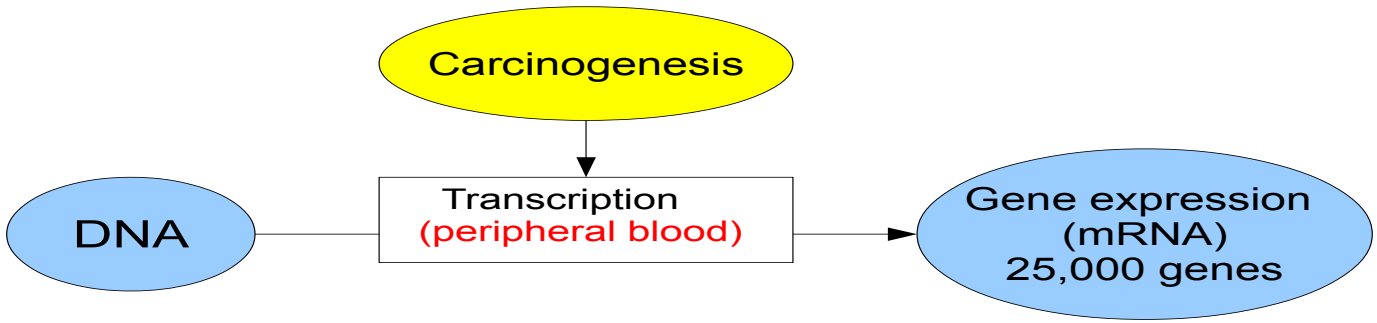
- T_i: Follow-up time.
- $\Delta G_i = \log G_i^{\text{case}} - \log G_i^{\text{control}}$: Difference of gene expression at time T_i before diagnosis (25,000 genes).
- ΔE_i: Exposure of CC pair i at time T_i before diagnosis.

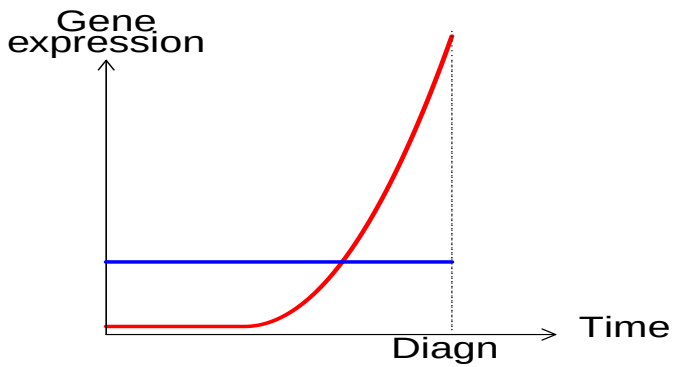
1. Post-GWAS design
2. Exploration of functional changes on gene expression
3. Prospective GWAS and post-GWAS: a different statistical point of view
4. Statistical approaches for post-GWAS: $\mathbb{P}[G|E,T]$

Carcinogenesis and transcription



Carcinogenesis and transcription

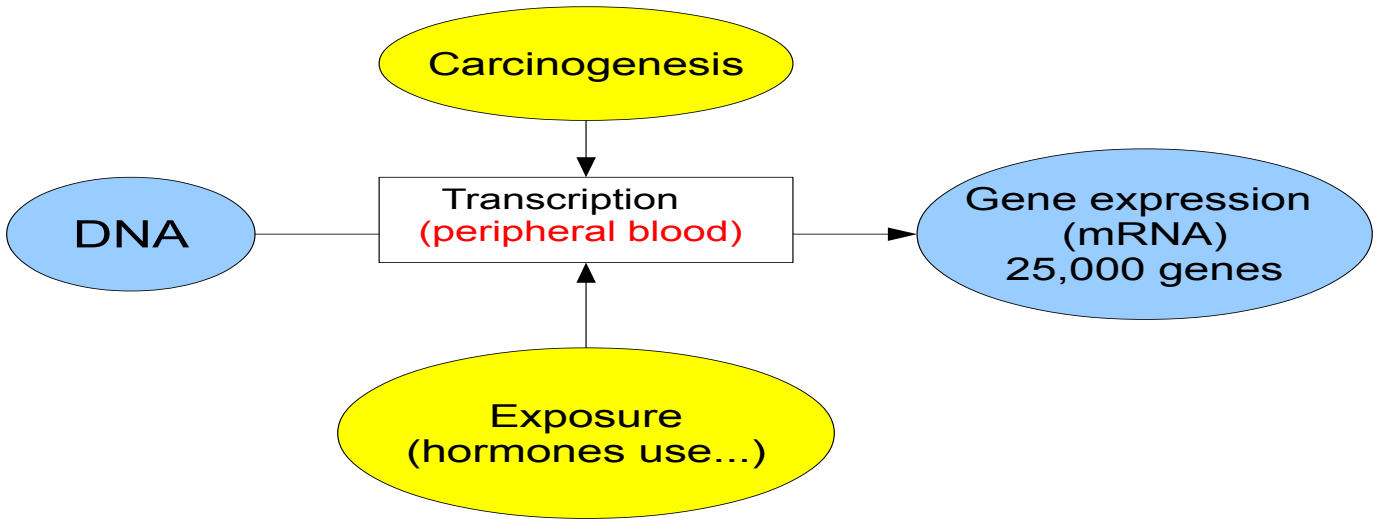


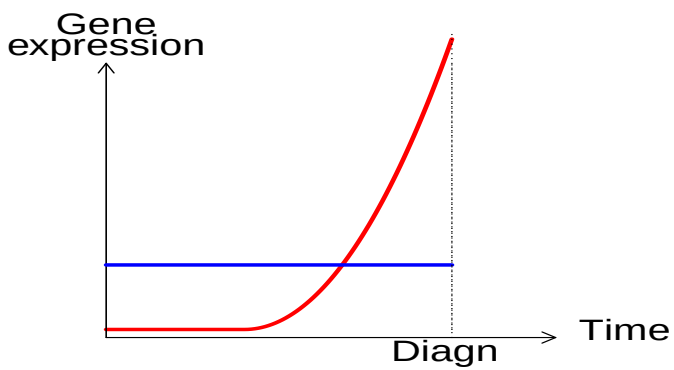


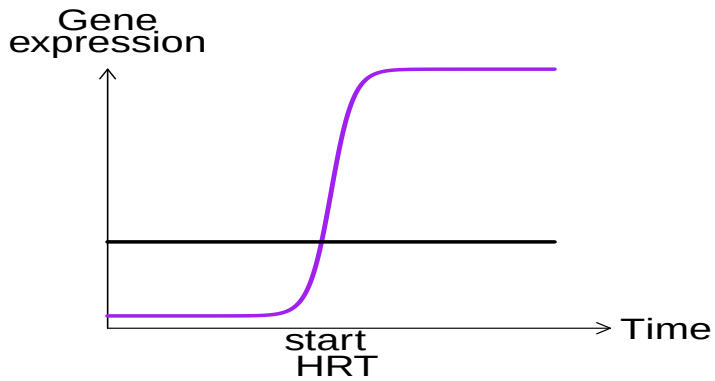
— gene involved in carcinogenesis

— gene non involved

Carcinogenesis and transcription



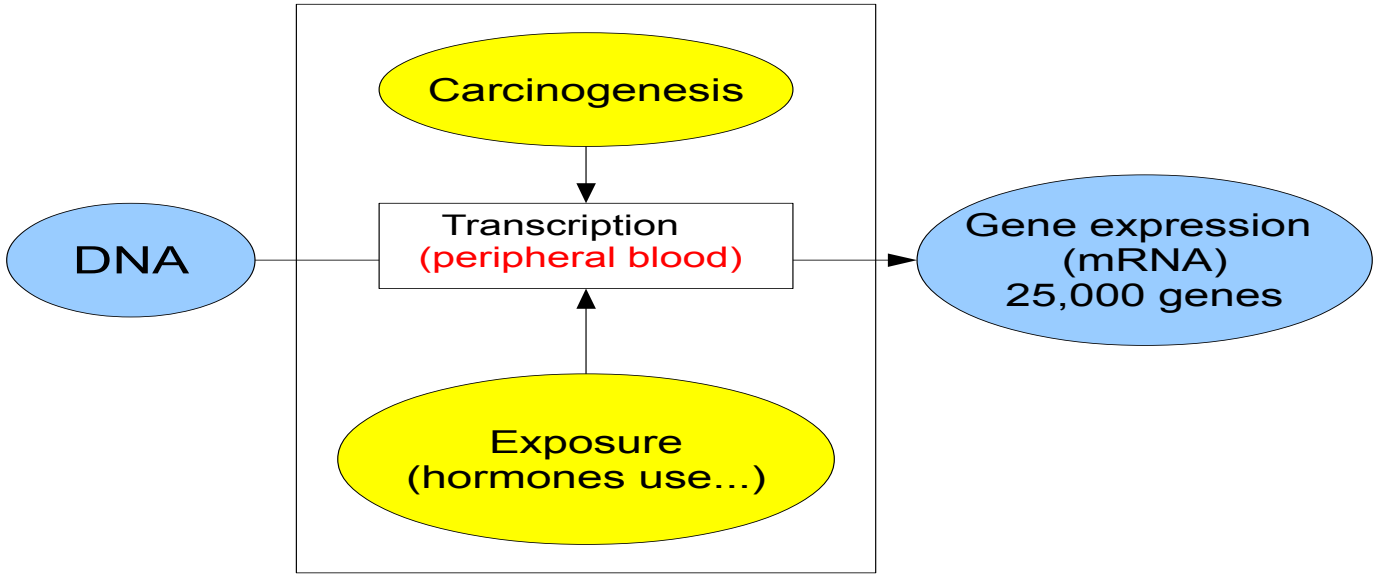


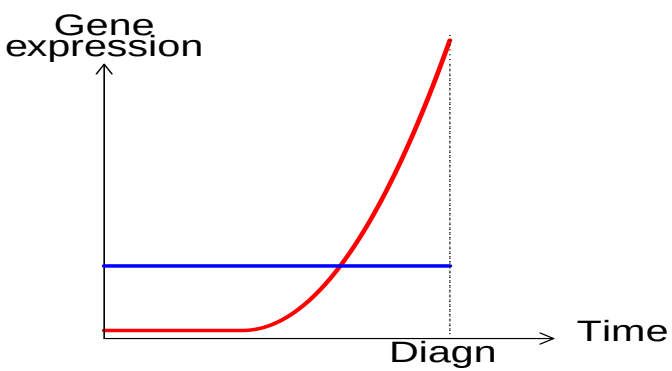


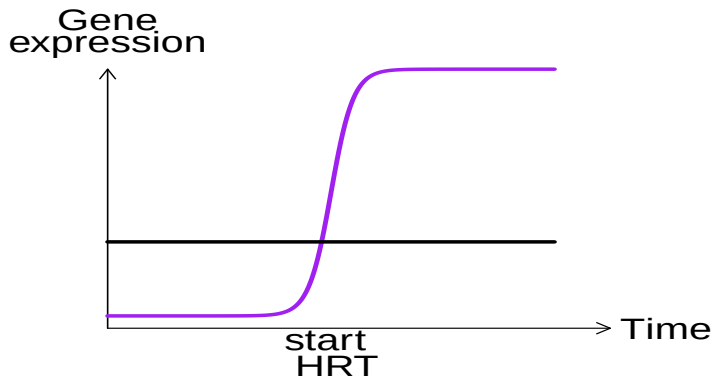
— gene involved in carcinogenesis
— gene non involved

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

Carcinogenesis and transcription







— gene involved in carcinogenesis
— gene non involved

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

1. Post-GWAS design
2. Exploration of functional changes on gene expression
3. Prospective GWAS and post-GWAS: a different statistical point of view
4. Statistical approaches for post-GWAS: $\mathbb{P}[G|E,T]$

Survival analysis models in prospective GWAS

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

- T: follow-up time
- E: exposures
- G: genomic data

Survival analysis models in prospective GWAS

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

- T: follow-up time
- E: exposures
- G: genomic data

Functional changes for post-GWAS

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

- T: follow-up time
- E: exposures
- G: transcriptomic data

Survival analysis models in prospective GWAS

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

- T: follow-up time
- E: exposures
- G: genomic data

Functional changes for post-GWAS

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

- T: follow-up time
- E: exposures
- G: transcriptomic data

What is different?

- Omic data are considered as:
  - Risk factor in prospective GWAS.
  - Biomarkers of carcinogenic process in post-GWAS.
- Different goals:
  - GWAS: relative risk estimation.
  - Post-GWAS: analysis of functional changes.

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

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

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

- 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.

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

- 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:

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

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

- 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.

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

- 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: $\lambda(t|G,E,T)$:

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

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

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

  $\hookrightarrow$ 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: $\lambda(t|G,E,T)$:
  $\hookrightarrow$ Not directly interpretable.

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

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

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

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

Limits of survival analysis models in post-GWAS analysis: illustration with Cox model.

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

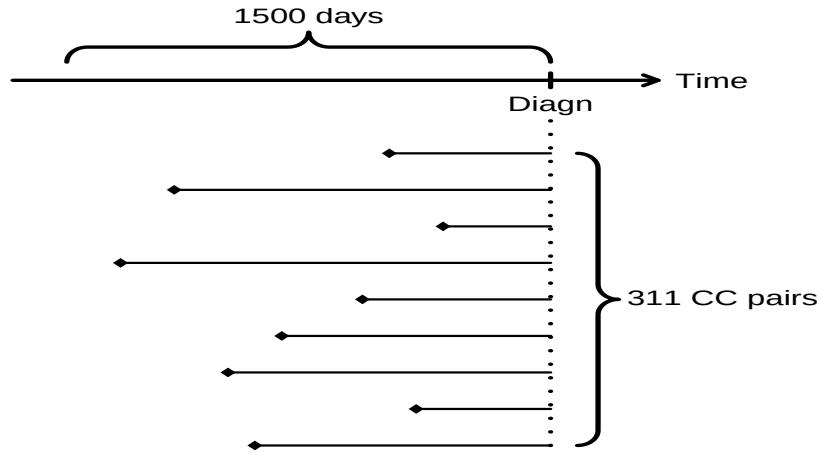
$$L(\beta) = \prod_{i \text{ CC pair}} \left(1 - \exp\left(\langle\beta,(\Delta G_i, \Delta E_i)\rangle\right)\right)^{-1} + \text{pen}(\beta)$$

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

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

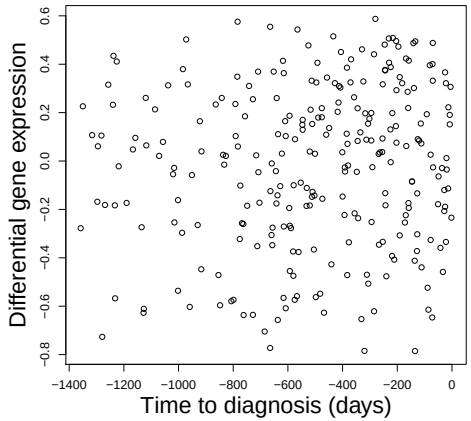
  $\hookrightarrow$ Penalization selects the most differentially expressed genes in each strata.
- More generally: $\lambda(t|G,E,T)$:
  $\hookrightarrow$ Not directly interpretable.
  $\hookrightarrow$ Association between gene expression and no-carcinogen exposures?
- Summing up
  - Survival analysis for nested CC: detect genes that discriminate between cases and controls.
  - Our goal: detect genes that discriminate between "long" and "short" follow-up times.

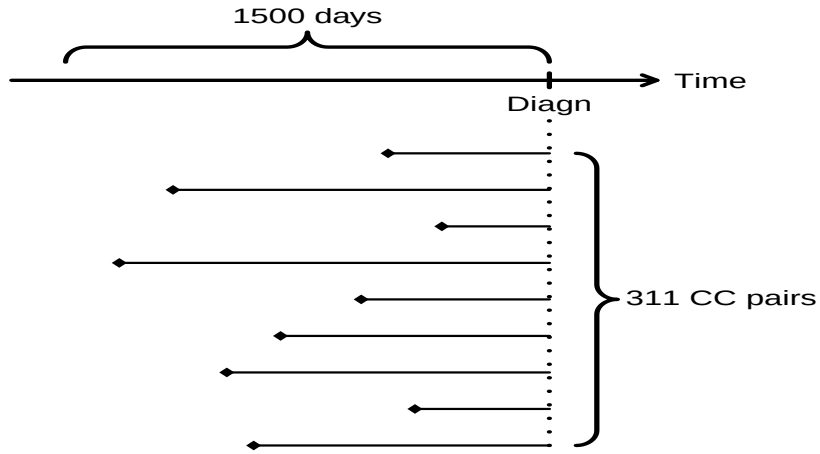
1. Post-GWAS design
2. Exploration of functional changes on gene expression
3. Prospective GWAS and post-GWAS: a different statistical point of view
4. Statistical approaches for post-GWAS: $\mathbb{P}[G|E,T]$


1500 days
Time
Diagn
311 CC pairs

One gene g

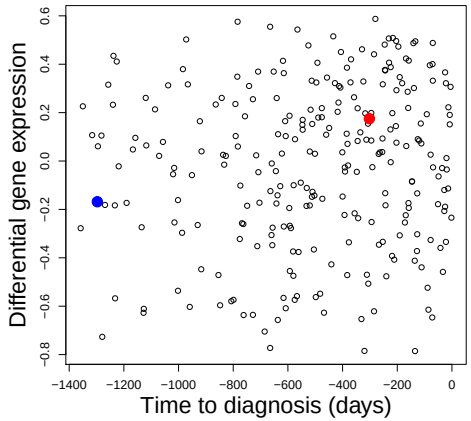
$i = 1 \ldots, 311$ CC pairs

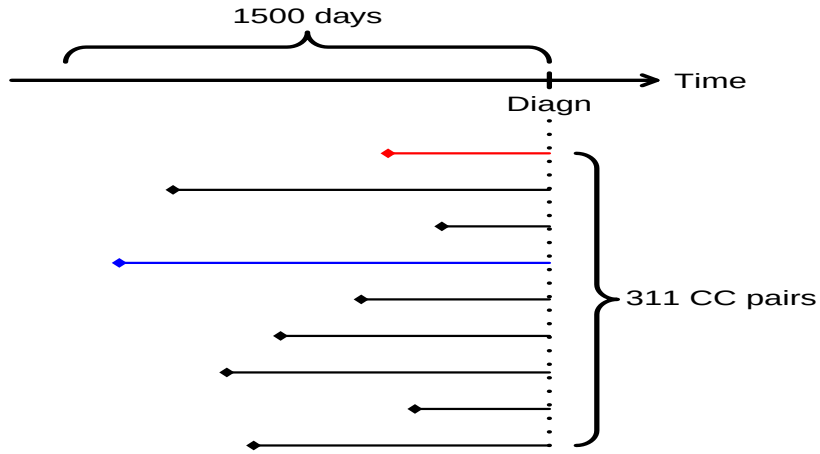




One gene g

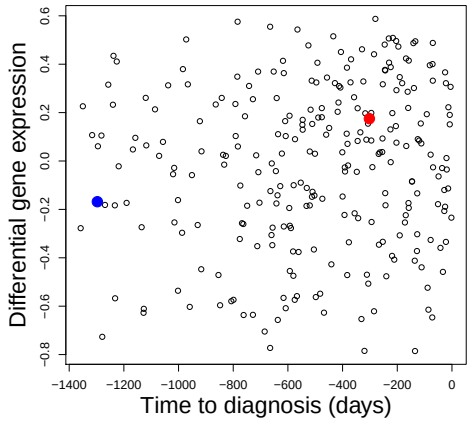
$i = 1 \ldots, 311$ CC pairs



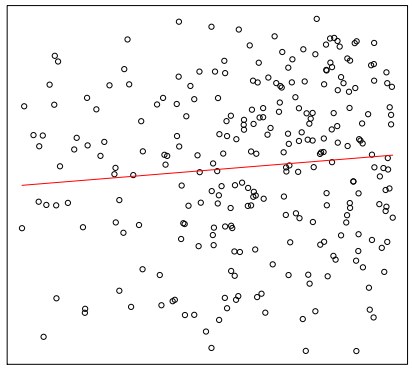


One gene g

$i = 1 \ldots, 311$ CC pairs



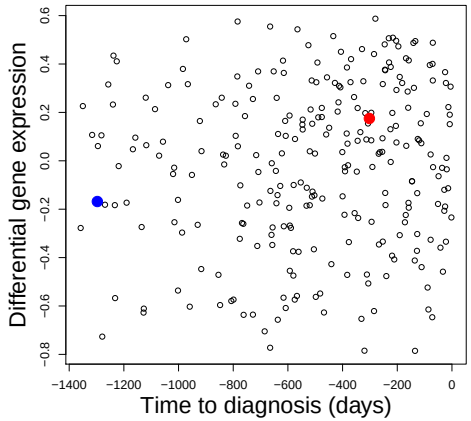
Linear model



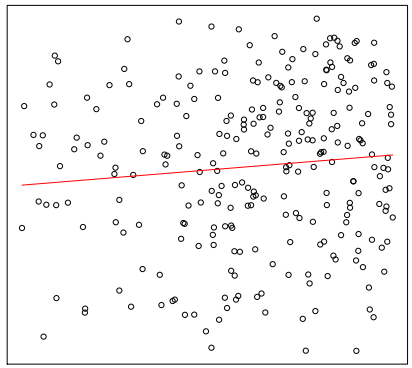
$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

One gene g

$i = 1 \ldots, 311$ CC pairs

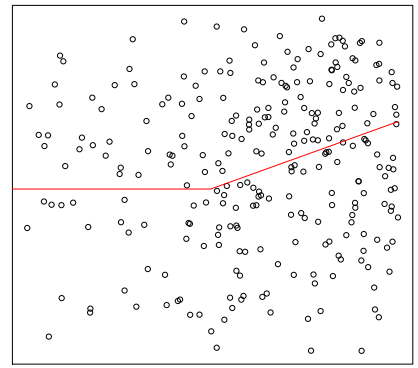


Linear model



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

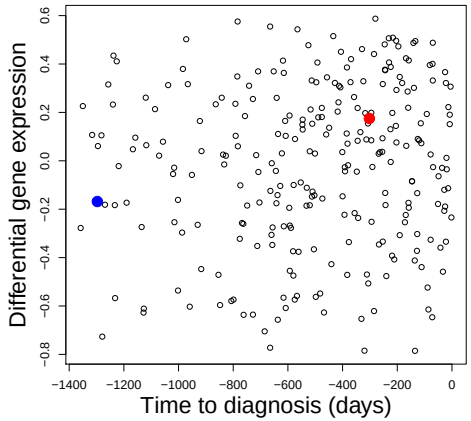
"Hockey-stick"



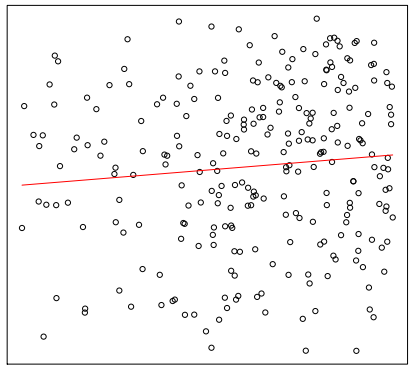
$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g (T_i - t_0) \mathbb{1}(T_I > t_0) + \varepsilon_{i,g}$$

One gene g

$i = 1 \ldots, 311$ CC pairs

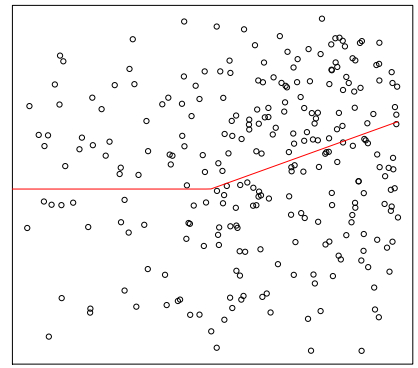


Linear model



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

"Hockey-stick"

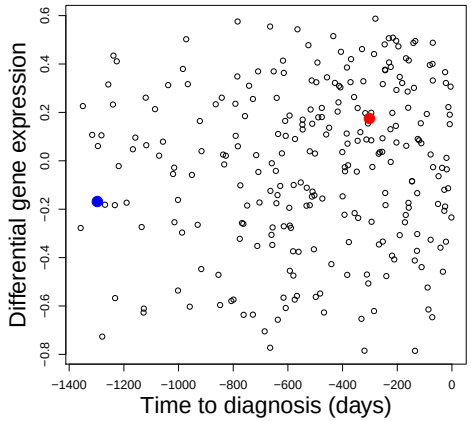


$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g (T_i - t_0) \mathbb{1}(T_I > t_0) + \varepsilon_{i,g}$$

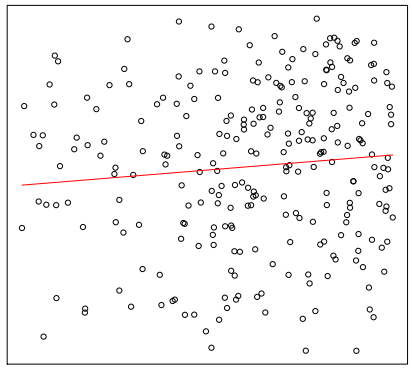
- General model: $\Delta G_{i,g} = f(T_i, \Delta E_i | \Theta_g) + \varepsilon_{i,g}$.

One gene g

$i = 1 \ldots, 311$ CC pairs

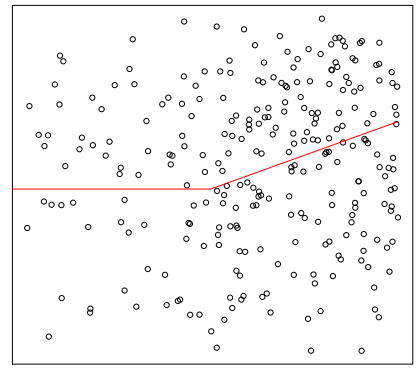


Linear model



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

"Hockey-stick"

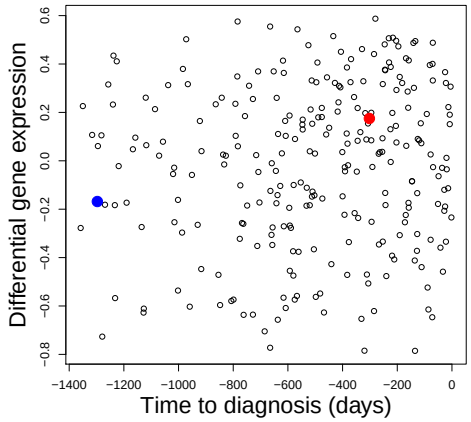


$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g (T_i - t_0) \mathbb{1}(T_I > t_0) + \varepsilon_{i,g}$$

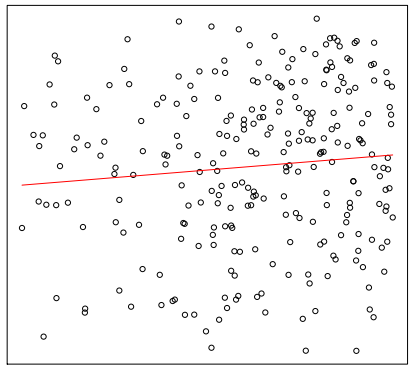
- General model: $\Delta G_{i,g} = f(T_i, \Delta E_i | \Theta_g) + \varepsilon_{i,g}$.
- Testing time-effect for each gene + correction for multiple testing.

One gene g

$i = 1 \ldots, 311$ CC pairs

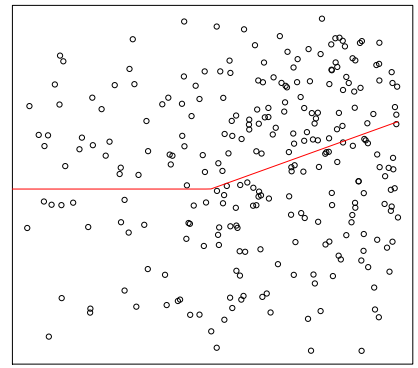


Linear model



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

"Hockey-stick"

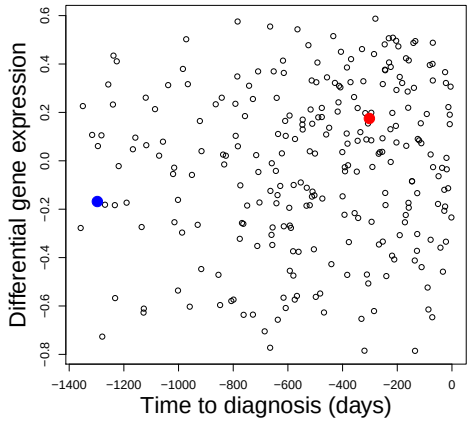


$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g (T_i - t_0) \mathbb{1}(T_I > t_0) + \varepsilon_{i,g}$$

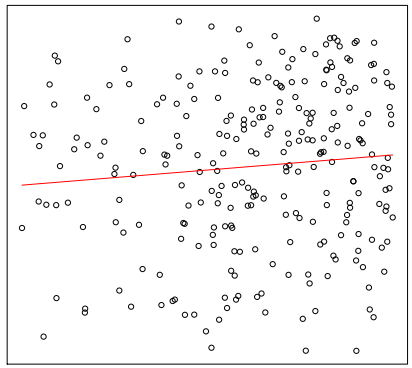
- General model: $\Delta G_{i,g} = f(T_i, \Delta E_i | \Theta_g) + \varepsilon_{i,g}$.
- Testing time-effect for each gene + correction for multiple testing.
- Controls used as reference.

One gene g

$i = 1 \ldots, 311$ CC pairs

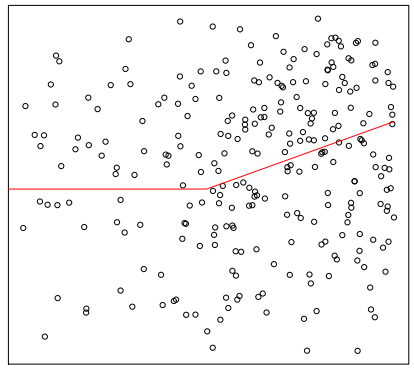


Linear model



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

"Hockey-stick"

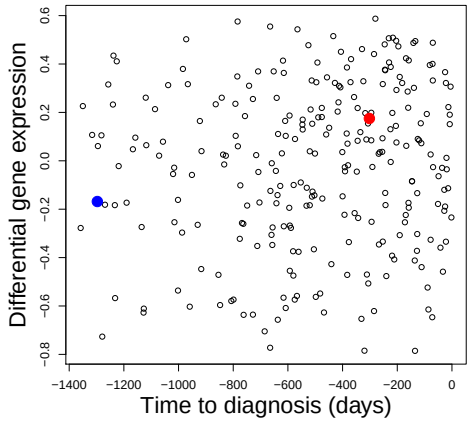


$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g (T_i - t_0) \mathbb{1}(T_I > t_0) + \varepsilon_{i,g}$$

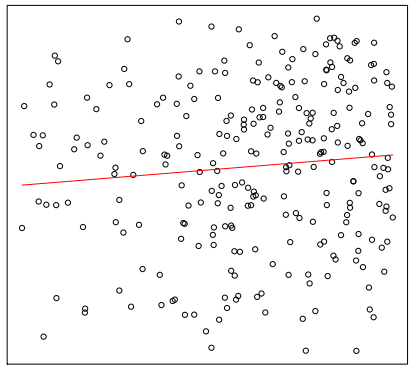
- General model: $\Delta G_{i,g} = f(T_i, \Delta E_i | \Theta_g) + \varepsilon_{i,g}$.
- Testing time-effect for each gene + correction for multiple testing.
- Controls used as reference.
- Flexibility allows to include biological assumptions:
  - Cancer driven by exposures,
  - Paths of genes with hierarchical FDR, ...

One gene g

$i = 1 \ldots, 311$ CC pairs

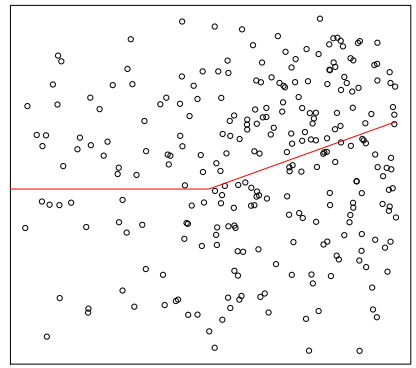


Linear model



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g T_i + \varepsilon_{i,g}$$

"Hockey-stick"



$$\Delta G_{i,g} = \alpha_0^g + \alpha_1^g (T_i - t_0)\mathbb{1}(T_I > t_0) + \varepsilon_{i,g}$$

- General model: $\Delta G_{i,g} = f(T_i, \Delta E_i | \Theta_g) + \varepsilon_{i,g}$.
- Testing time-effect for each gene + correction for multiple testing.
- Controls used as reference.
- Flexibility allows to include biological assumptions:
  - Cancer driven by exposures,
  - Paths of genes with hierarchical FDR, ...
- Latent variable model based on multistage model of carcinogenesis.

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

with LS_i the length of the last stage for case i.

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.
- What's next?

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.
- What's next?
  - Parametrization of the time effect.

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.
- What's next?
  - Parametrization of the time effect.
  - Determinate the exposures which affects gene expression

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics → transcriptomics
  - Different goals:
    relative risk estimation → exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.
- What's next?
  - Parametrization of the time effect.
  - Determinate the exposures which affects gene expression (huge subject!)

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.
- What's next?
  - Parametrization of the time effect.
  - Determinate the exposures which affects gene expression (huge subject!)
  - Stratified with respect to the stages of cancer.

Conclusion

- From prospective GWAS to post-GWAS.
  - Different design:
    genomics $\rightarrow$ transcriptomics
  - Different goals:
    relative risk estimation $\rightarrow$ exploration of functional changes
  - Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - Gene-by-gene model.
  - Latent variable model which accounts for individual dynamics.
- What's next?
  - Parametrization of the time effect.
  - Determinate the exposures which affects gene expression (huge subject!)
  - Stratified with respect to the stages of cancer.
  - etc...

Conclusion

- From prospective GWAS to post-GWAS.
  - ◇ Different design:
    genomics → transcriptomics
  - ◇ Different goals:
    relative risk estimation → exploration of functional changes
  - ◇ Different statistical point of view:
    $\mathbb{P}[T|G,E] \rightarrow \mathbb{P}[G|T,E]$
- Statistical approaches for analysis of functional changes on transcriptomic data:
  - ◇ Gene-by-gene model.
  - ◇ Latent variable model which accounts for individual dynamics.
- What's next?
  - ◇ Parametrization of the time effect.
  - ◇ Determinate the exposures which affects gene expression (huge subject!)
  - ◇ Stratified with respect to the stages of cancer.
  - ◇ etc...

Takk!