Inferring time-varying genetic network using informative priors

Sophie Lèbre

LSIIT, Theoretical Bioinformatics, Strasbourg,

LICSB - 1st April 2009

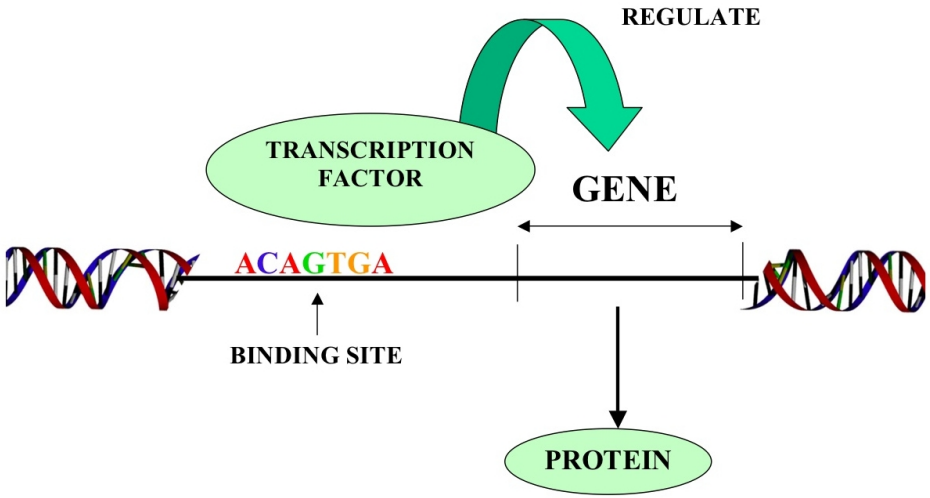
Outline

1. Modelling regulatory networks from gene expression time series
   - $\rightsquigarrow$ Detecting structural changes
   - $\rightsquigarrow$ Using available biological knowledge
2. *Time-varying* network model and inference
   - $\rightsquigarrow$ temporal variation across Drosophila development

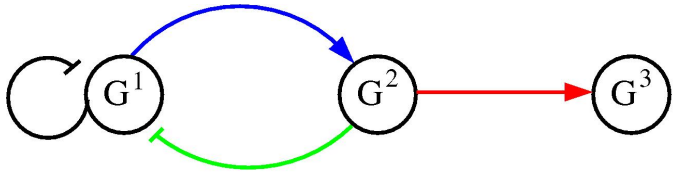
   Joint work with G. Lelandais, F. Devaux, M. Stumpf.
3. Including informative priors
   - $\rightsquigarrow$ edges, degree distribution
4. Effect on Drosophila development network inference

Recovering genes functions?

- Regulatory relationships:



  - up/down regulation
  - retroaction, feedforwards loops...

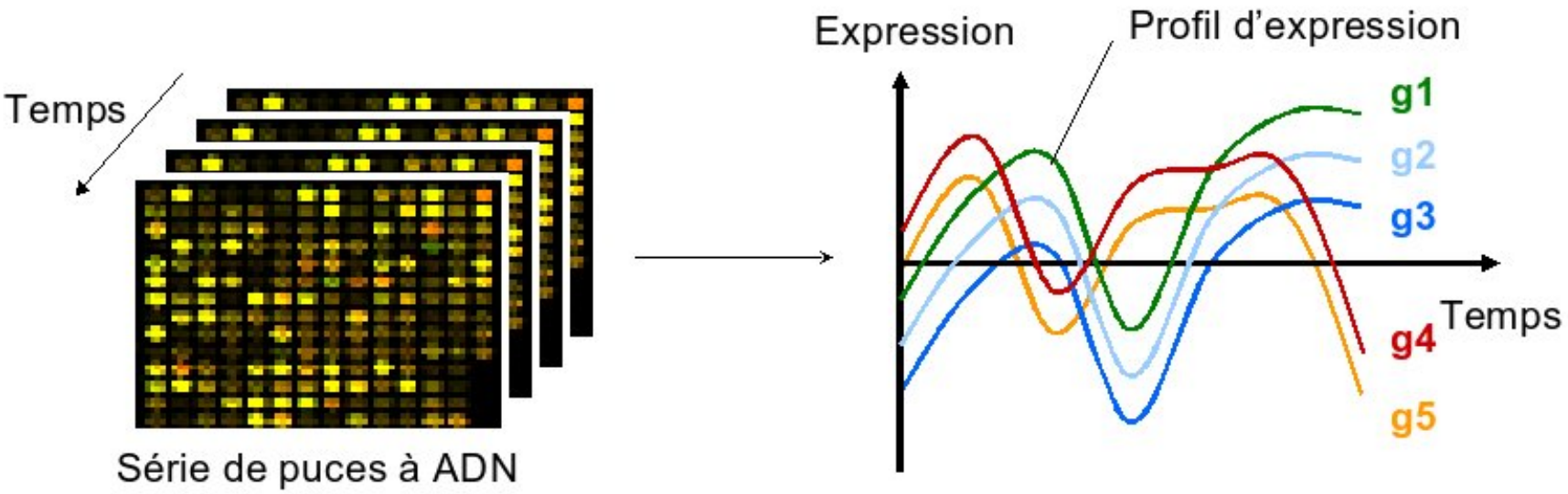


$\Rightarrow$ **Complex dynamic system**

- Objective: identifying this organisation in large scale.

Temporal gene expression data

- Microarrays:
  $\rightsquigarrow$ **simultaneous** expression of **several thousands** of genes.



- Notations: we consider the stochastic process,

$$X = \{X_t^i;\ \forall i \in \{1, ..., p\}, \forall t \in \{1, ..., n\}\}$$

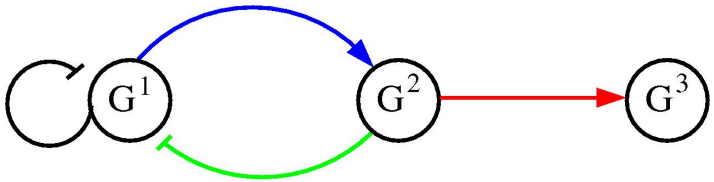
where X_t^i is the expression of gene i at time t,

What information extracting from expression profiles?

- Which genes work together?
- At what instant of the observed process?

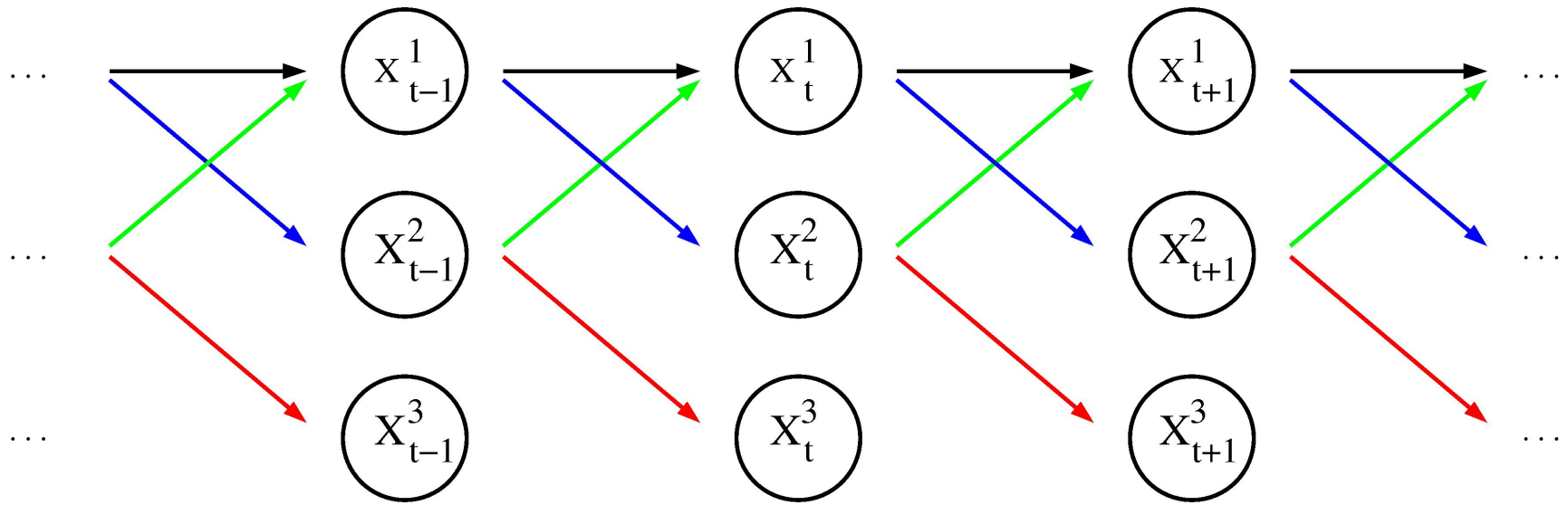
DBN modelling of biological motifs

- A biological motif



- Dynamic Bayesian Networks (DBNs)

  $\rightsquigarrow$ allow to model biological cycles



(Friedman et al. 1998, Murphy and Mian 1999, OpgenRhein and Strimmer 2007)

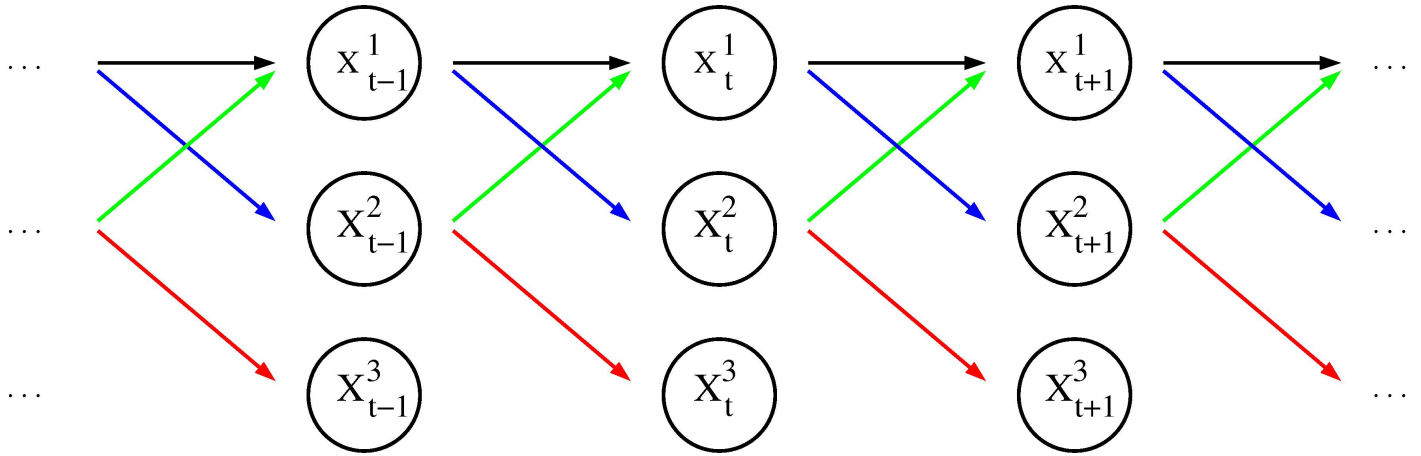
DBN modelling

Assumptions

- $(\mathcal{A}_1)$ 1^{st} *order Markov process*
- $(\mathcal{A}_2)$ *'simultaneous independence' given the past,*

$$\forall t > 1, \forall i, j \in N, \quad X_t^i \perp\!\!\!\perp X_t^j \mid X_{t-1}.$$

- $(\mathcal{A}_3)$ *time homogeneity*



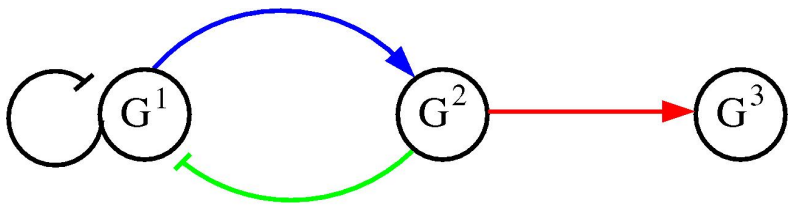
$$f(X) = \prod_{1<t\leq n} f(X_t^1|X_{t-1}^1, X_{t-1}^2)\ f(X_t^2|X_{t-1}^1)\ f(X_t^3|X_{t-1}^2)$$

DBN for a 1^{st} order auto-regressive process: AR(1).

- AR(1) process: $\forall t \geq 1,\ X_t = AX_{t-1} + B + \varepsilon_t,\quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$

$$\begin{bmatrix} X_t^1 \\ \cdot \\ \cdot \\ X_t^i \\ \cdot \\ \cdot \\ \cdot \\ X_t^p \end{bmatrix} = \begin{bmatrix} a_{11} & \cdot & \cdot & \cdot & a_{1j} & \cdot & \cdot & a_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & \cdot & \cdot & \cdot & a_{ij} & \cdot & \cdot & a_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & \cdot & \cdot & \cdot & a_{pj} & \cdot & \cdot & a_{pp} \end{bmatrix} \begin{bmatrix} X_{t-1}^1 \\ \cdot \\ \cdot \\ \cdot \\ X_{t-1}^j \\ \cdot \\ \cdot \\ X_{t-1}^p \end{bmatrix} + \begin{bmatrix} b_t^1 \\ \cdot \\ \cdot \\ b_t^i \\ \cdot \\ \cdot \\ \cdot \\ b_t^p \end{bmatrix} + \begin{bmatrix} \varepsilon_t^1 \\ \cdot \\ \cdot \\ \varepsilon_t^i \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_t^p \end{bmatrix}$$

- Example:



$$A = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{pmatrix}$$

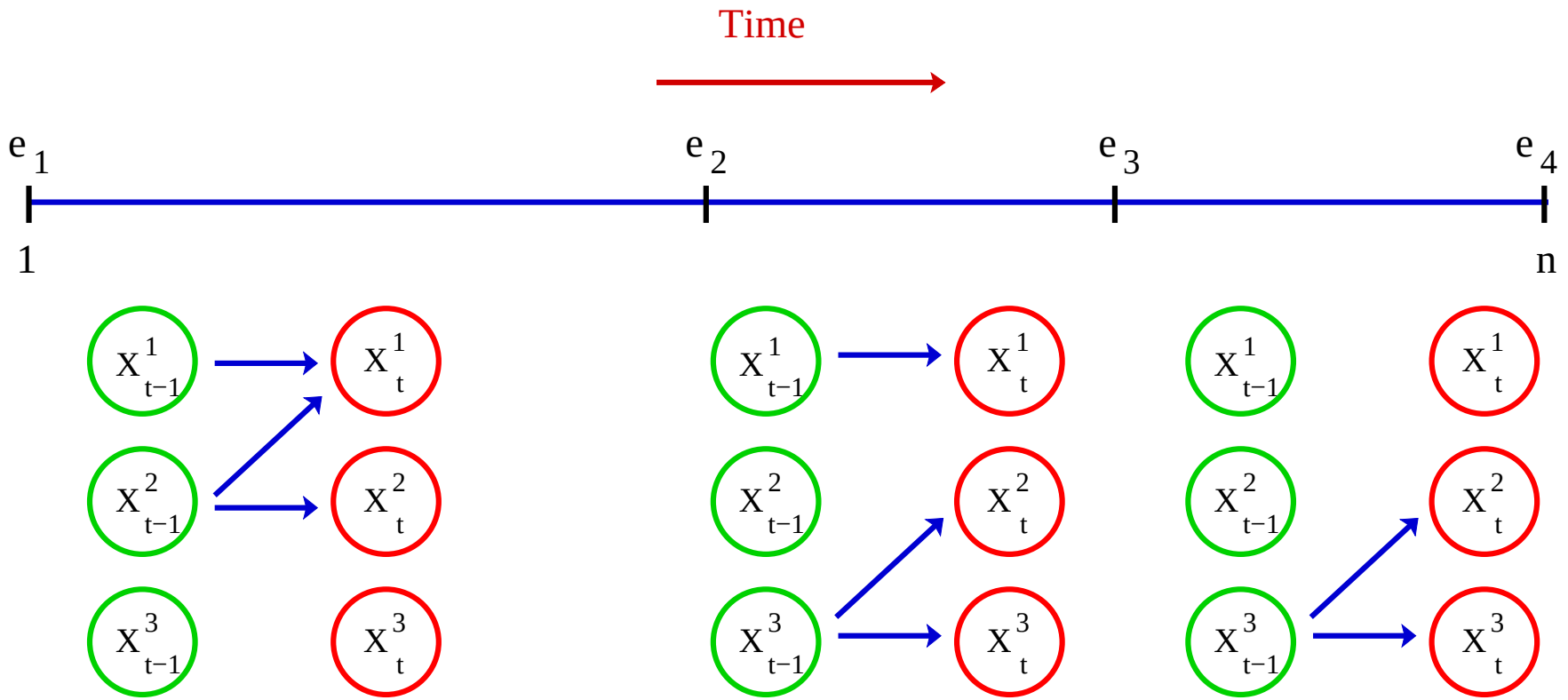
Outline

1. Modelling regulatory networks from gene expression time series
   - $\rightsquigarrow$ Detecting structural changes
   - $\rightsquigarrow$ Using available biological knowledge
2. *Time-varying* network model and inference
   - $\rightsquigarrow$ temporal variation across Drosophila development

   Joint work with G. Lelandais, F. Devaux, M. Stumpf.
3. Including informative priors
   - $\rightsquigarrow$ edges, degree distribution
4. Effect on Drosophila development network inference

Time-varying dynamic Bayesian network model

- Introducing changepoints allowing for network structure change.



$$X_t^1 = b^{10} + b^{11} X_{t-1}^1 + b^{12} X_{t-1}^2$$

$$X_t^1 = b^{10} + b^{11} X_{t-1}^1$$

$$X_t^1 = b^{10}$$

Time-varying DBN model

p genes - n time points - k changepoint positions

For each gene i, $(1 \leq i \leq p)$,

- a **changepoint vector** $\xi^i = (\xi^i_1, ..., \xi^i_{h-1}, \xi^i_h, ..., \xi^i_{k^i}) \subseteq \{2, ..., n\}$
- in each phase h, (for all $\xi^i_h \leq t < \xi^i_{h+1}$),
  - a set of s^i_h **parents** $\tau^i_h = \{j_1, ..., j_{s^i_h}\} \subseteq \{1, ..., p\}$
  - and a set of **parameters** $\theta^i_h = ((b^{ij}_h)_{j \in \{0,...,q\}}, \sigma^i_h)$,

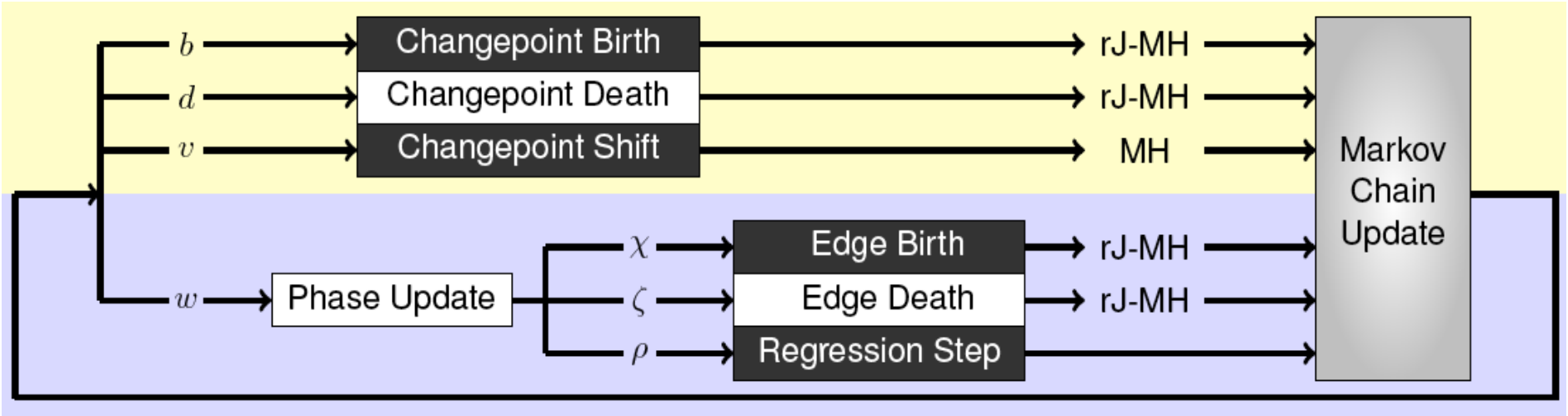
define the regression model,

$$X^i_t = b^{i0}_h + \sum_{j \in \tau^i_h} b^{ij}_h X^j_{t-1} + \varepsilon^i_t, \qquad \varepsilon^i_t \sim \mathcal{N}(0, \sigma^i_h).$$

$\rightsquigarrow$ Unknown dimension

Inference: 2 steps reversible jump MCMC.

- Outline of the tvDBN procedure:



$\rightsquigarrow$ Reversible jump MCMC: Green (1995).
$\rightsquigarrow$ Model selection: Andrieu and Doucet (1999).

2 steps **embedded** *reversible jump MCMC*

- Changepoints priors
  - number of changepoints $k \sim \mathcal{P}(\lambda)$
  - changepoints position $\xi|k \sim \mathcal{U}niform$
- 4 moves : Birth (b_k), Death (d_k), Position shift (η_k), Regression model update (π_k).

$$b_k + d_k + \eta_k + \pi_k = 1$$

$$b_k = c \ \min\left\{1, \frac{\mathbb{P}_{\bar{k}}(k+1)}{\mathbb{P}_{\bar{k}}(k)}\right\}, \ d_k = c \ \min\left\{1, \frac{\mathbb{P}_{\bar{k}}(k-1)}{\mathbb{P}_{\bar{k}}(k)}\right\}, \ \eta_k = \frac{1}{2}(b_k + d_k).$$

$\rightsquigarrow$ c small

- Birth: $\xi^+ = \xi \cup \xi^\star$
  - proposal ratio: $\frac{d_{k+1}}{b_k} \frac{q(\xi^\star|\xi^+)}{q(\xi^\star|\xi)} = \frac{(n-1)p-k}{\lambda}$
  - Jacobian=1
  - posterior distribution ratio: $\frac{\mathbb{P}(k+1,\xi^+,s^+,\tau^+|y)}{\mathbb{P}(k,\xi,s,\tau|y)} \quad \perp\!\!\!\perp b, \sigma$

$\Rightarrow$ Acceptance ratio $R_{k,k+1}(\xi, \xi^+) = \frac{(n-1)p-k}{\lambda} \frac{\mathbb{P}(k+1,\xi^+,s^+,\tau^+|y)}{\mathbb{P}(k,\xi,s,\tau|y)}$.

2 steps **embedded** *reversible jump MCMC*

- Generation of an ergodic Markov chain.
- Reversible Markov chain: detailed balance satisfied.
- Equilibrium distribution converge to the desired post-distribution,

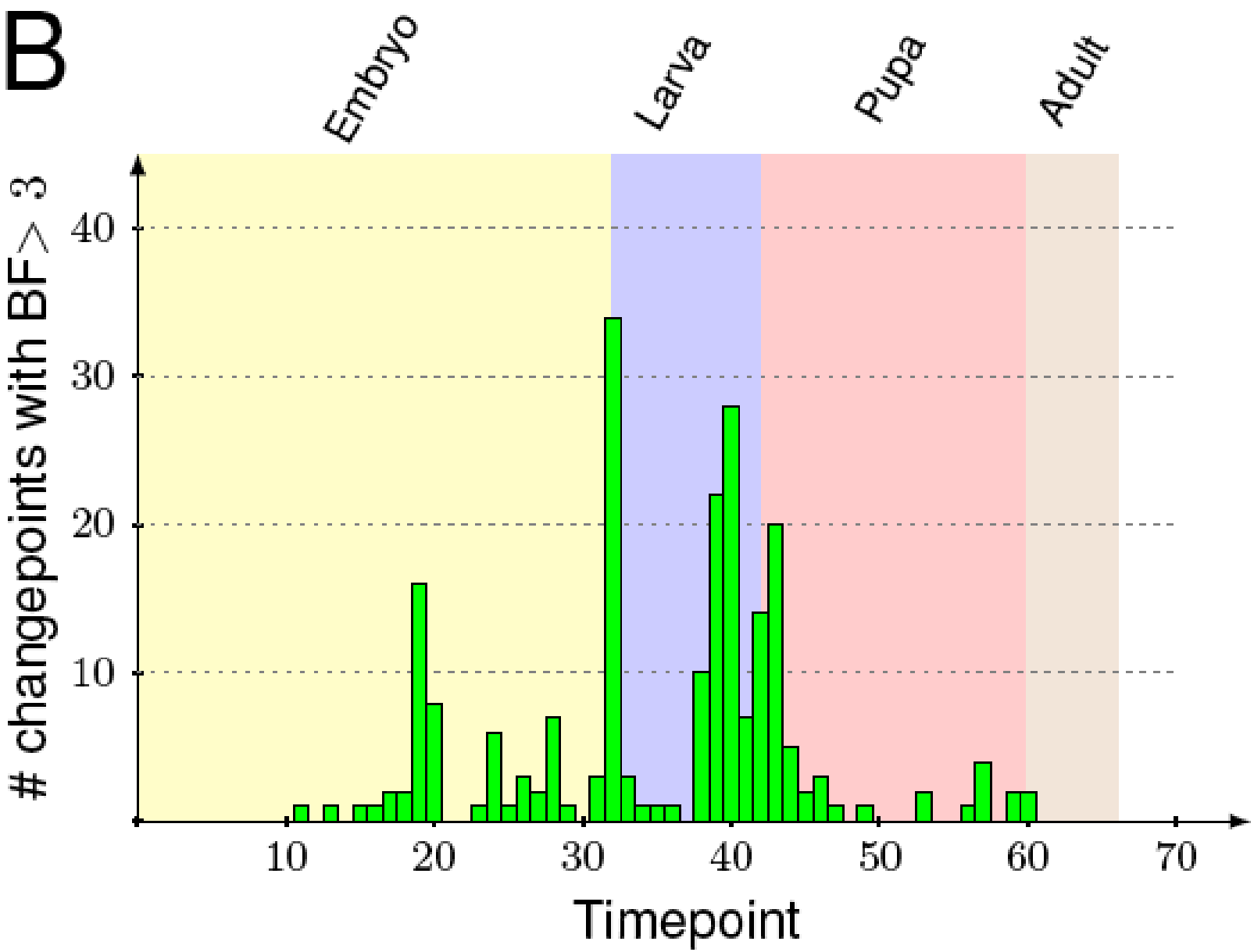
$$\mathbb{P}(k, \xi, s, \tau, \theta | x).$$

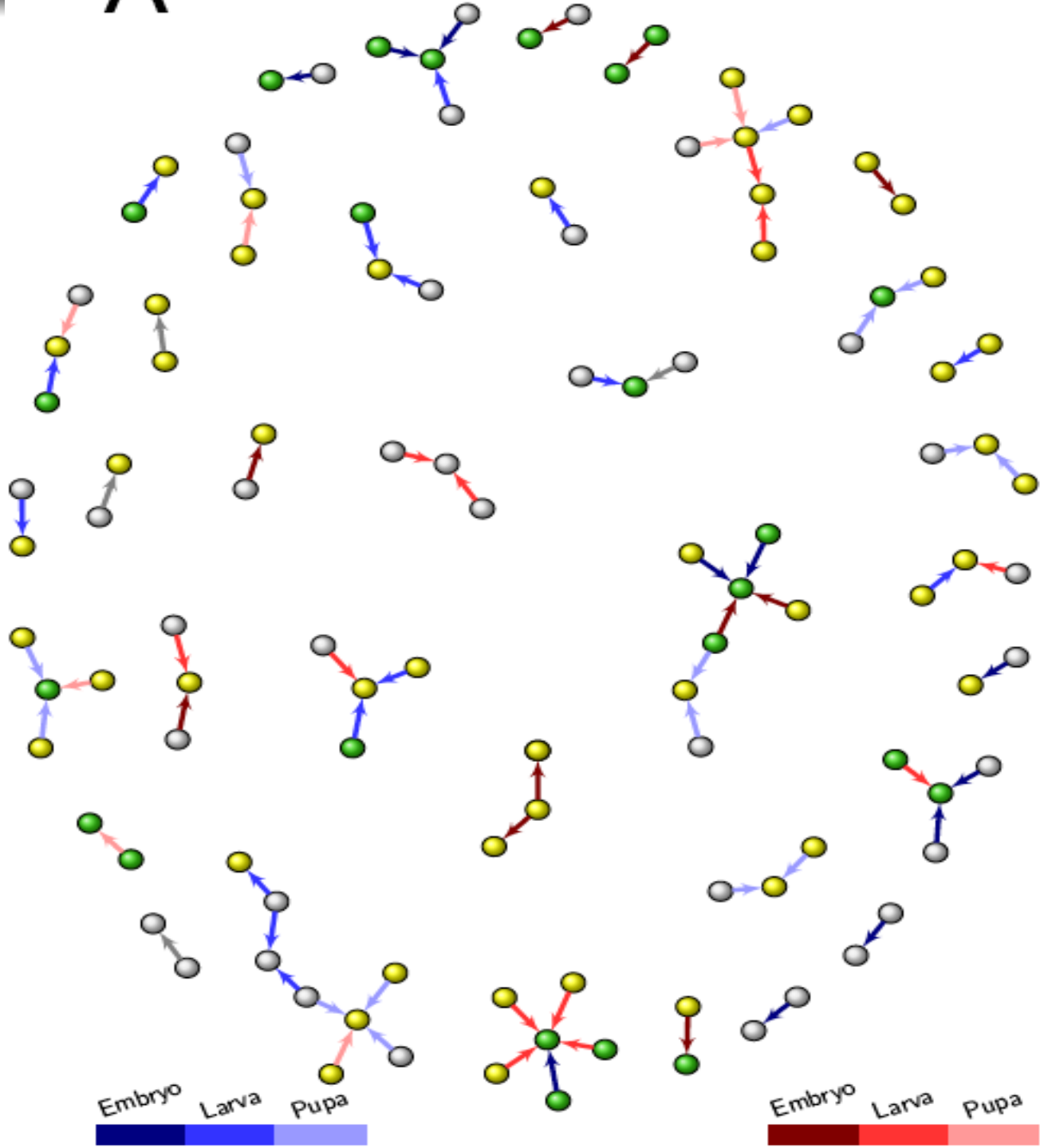
- Bayes Factor (BF) analysis.

Drosophila melanogaster life cycle data

- Data by Arbeitman et al. (2002)
  - gene expression across the whole life cycle of *D. melanogaster*: embryo, larva, pupa and adult.
  - 4028 genes
  - 66 successive time points
- Results
  - 50 genes have strong evidence for being "regulated" during only a fraction of their life (i.e. BF>12)
  - Implying 52 parents/regulators
  - Changepoints tend to cluster at the development transitions (embryo/larva and larva/pupa) + mid-embryo stage.

Changepoints cluster at phase transition and mid-embryo




Embryo
Larva
Pupa
Embryo
Larva
Pupa

Inferred target/parent gene vs Gene Ontology

	Developmental Process	Transcription Activity	Total
Inferred target genes	7	24	50
Inferred parent genes	7	26	52
Total	14	50	

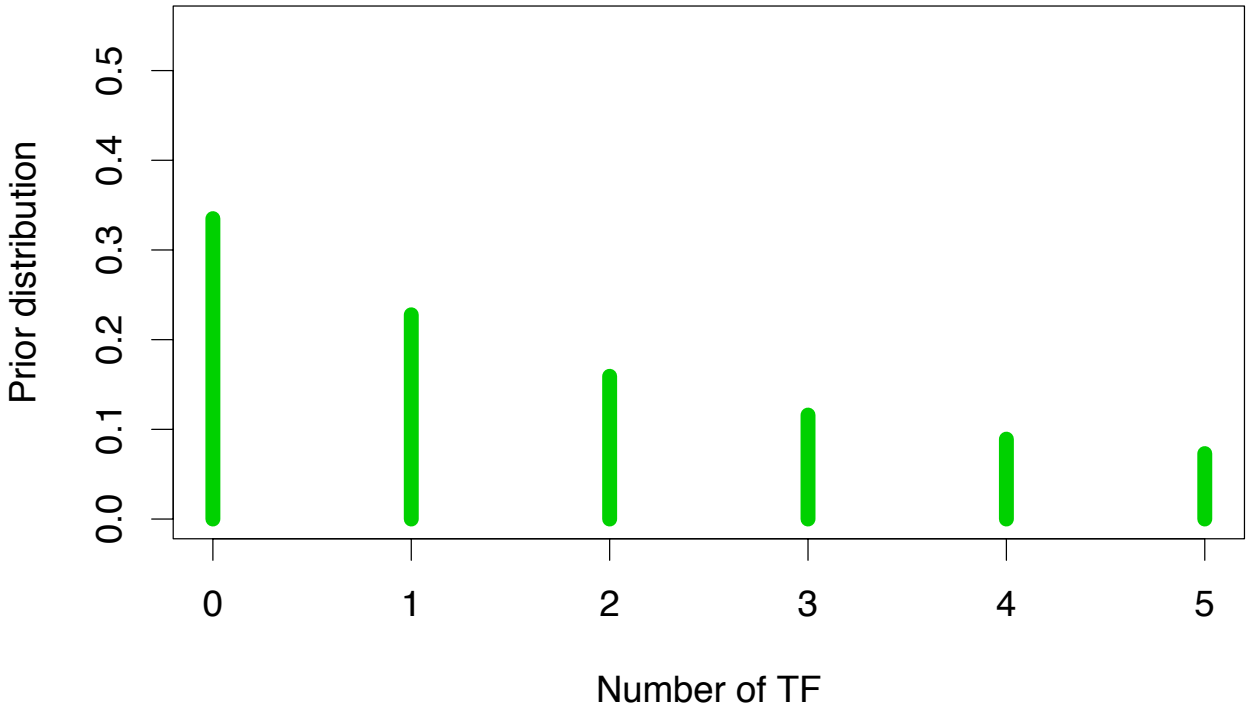
Outline

1. Modelling regulatory networks from gene expression time series
   - $\rightsquigarrow$ Detecting structural changes
   - $\rightsquigarrow$ Using available biological knowledge
2. *Time-varying* network model and inference
   - $\rightsquigarrow$ temporal variation across Drosophila development

   Joint work with G. Lelandais, F. Devaux, M. Stumpf.
3. Including informative priors
   - $\rightsquigarrow$ edges, degree distribution
4. Effect on Drosophila development network inference

Usual priors

- Network sparsity
  - ⇝ number of parents: $s_h^i \sim \mathcal{P}(\lambda)$ where $\lambda \sim \mathcal{IG}(\alpha, \beta)$.



- Incoming edges: Uniformative priors
  - ⇝ set of parents: $\tau_h^i | s_h^i \sim \mathcal{U}niform$

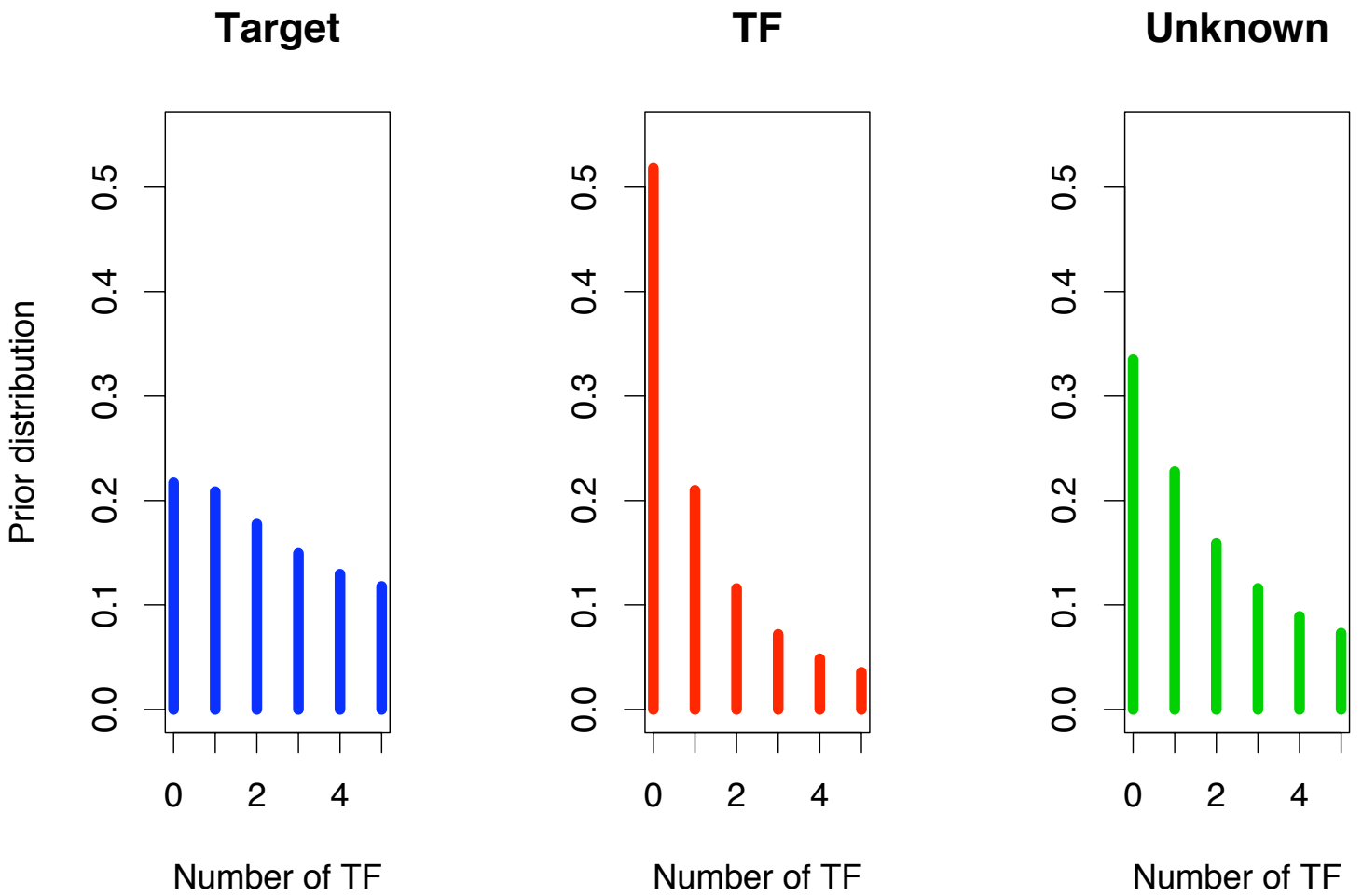
Including informative priors

- Network connectivity
  - $\rightsquigarrow$ Some genes share common properties:
    - Genes implicated in drosophila nervous system development are expected to work together (or muscle, heart, ...)
    - Transcription factor activates the transcription of target genes
  - $\rightsquigarrow$ cluster genes sharing common properties
    $+$ 1 cluster with uninformative priors
- L clusters with specific priors
  - Number of incoming edges
  - Inter cluster connectivity

Gene cluster priors

- Number of incoming edges for cluster l: hyperparameters (α_l, β_l)

$$p(s_h^i) \sim \mathcal{P}(\lambda), \quad \lambda \sim \mathcal{IG}(\alpha_l, \beta_l).$$



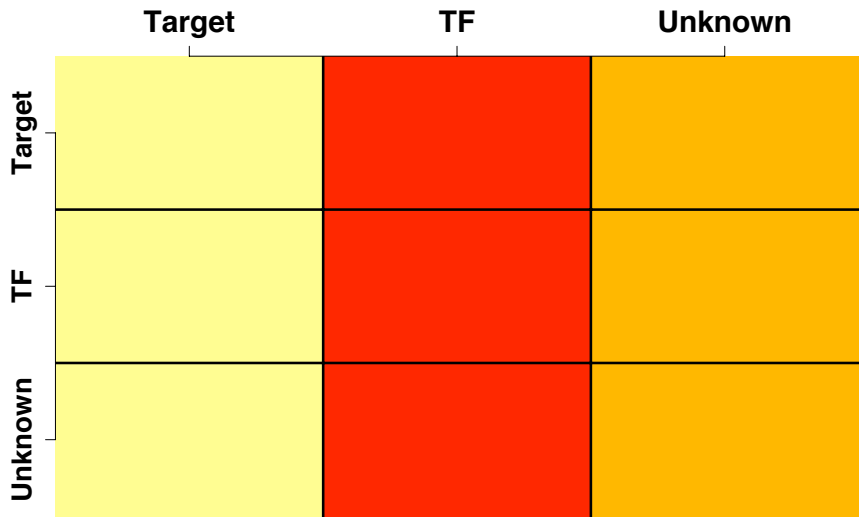
$(\alpha = 1.4, \beta = 0.5)$ $(\alpha = 0.6, \beta = 0.5)$ $(\alpha = 1, \beta = 0.5)$

Gene cluster priors

- Weight of the putative parent genes for target gene of each cluster l

$$\rightsquigarrow \text{ inter cluster connectivity weight matrix } W_{[L \times L]}$$

where $W[l, m]$ is weight for genes of cluster m to be a parent of genes of cluster l



Informative priors

- Possible changepoints priors
  - Some time point position *may* be considered as more probable than others.

  $\Rightarrow$ specific weight for changepoint positions $\omega = (\omega_t)_{t \in N}$

$$p(\xi | k \text{ Changepoints}) = f(\omega)$$

- Remarks
  - Only commonly accepted knowledge is considered.
  - Bayes factor analysis : results significance taking priors into account.

Outline

1. Modelling regulatory networks from gene expression time series
   - $\rightsquigarrow$ Detecting structural changes
   - $\rightsquigarrow$ Using available biological knowledge
2. *Time-varying* network model and inference
   - $\rightsquigarrow$ temporal variation across Drosophila development

   Joint work with G. Lelandais, F. Devaux, M. Stumpf.
3. Including informative priors
   - $\rightsquigarrow$ edges, degree distribution
4. Effect on Drosophila development network inference

Inferred target/parent gene vs Gene Ontology

- 14 genes implicated in drosophila development process:

  Minifly - Anarchist - Quo Vadis - Neurexin IV - Bicoid - smi21F - zfh1
  Glass - Peter Pan - Rbp9 - Quail - Castor - Mip - CG32486

- + Neighbours $\rightsquigarrow$ 36 genes

  $\Rightarrow$ tvDBN inference with informative priors specific to 3 clusters :

  - Cluster 1: Developmental process
  - Cluster 2: Transcription factor activity
  - Cluster 3: Unknown

Inferred target/parent gene vs Gene Ontology

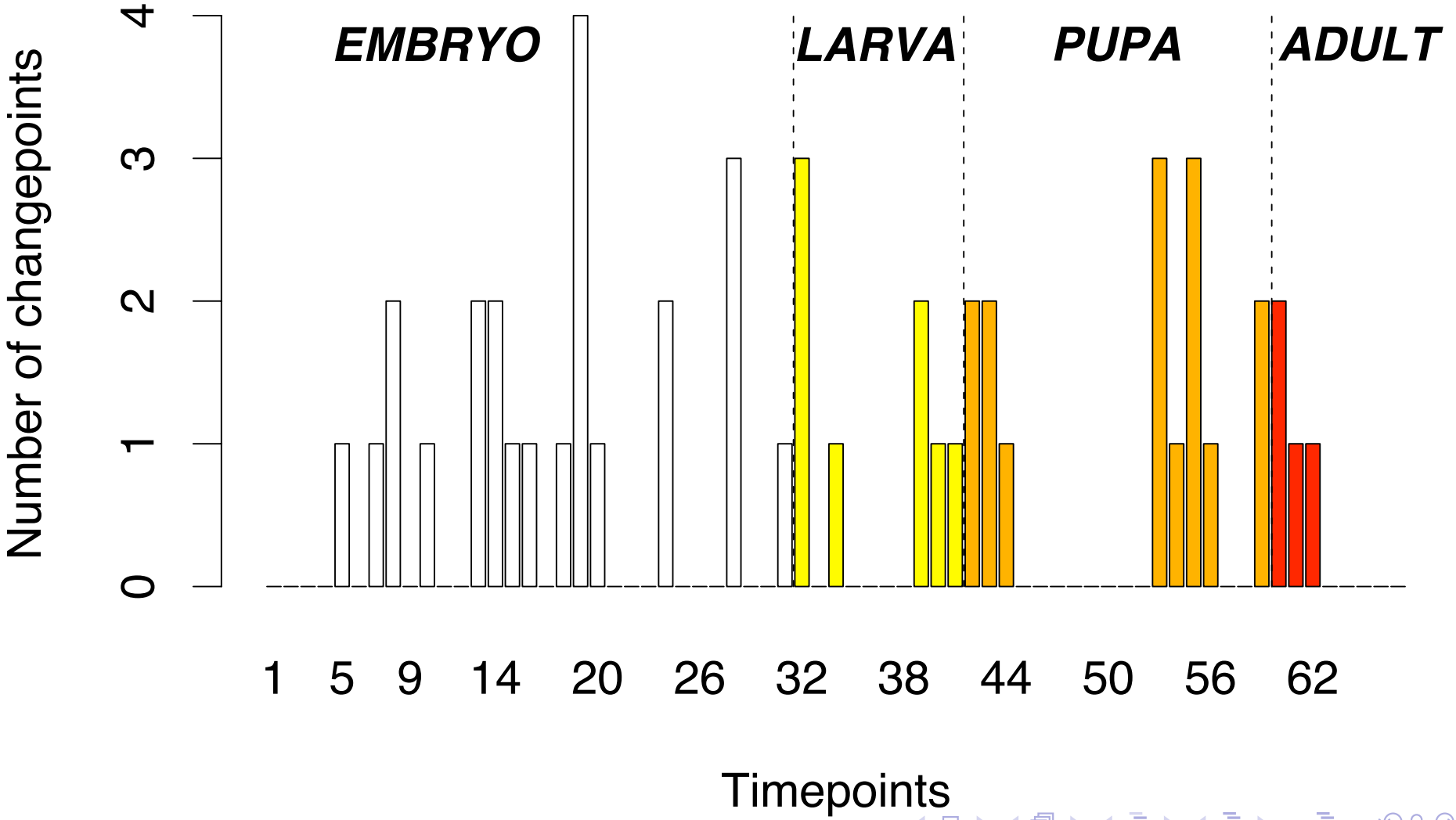
- Uninformative priors:

	Developmental process	Transcription Activity	Unknown
Target genes	10	7	13
Parent genes	10	5	13

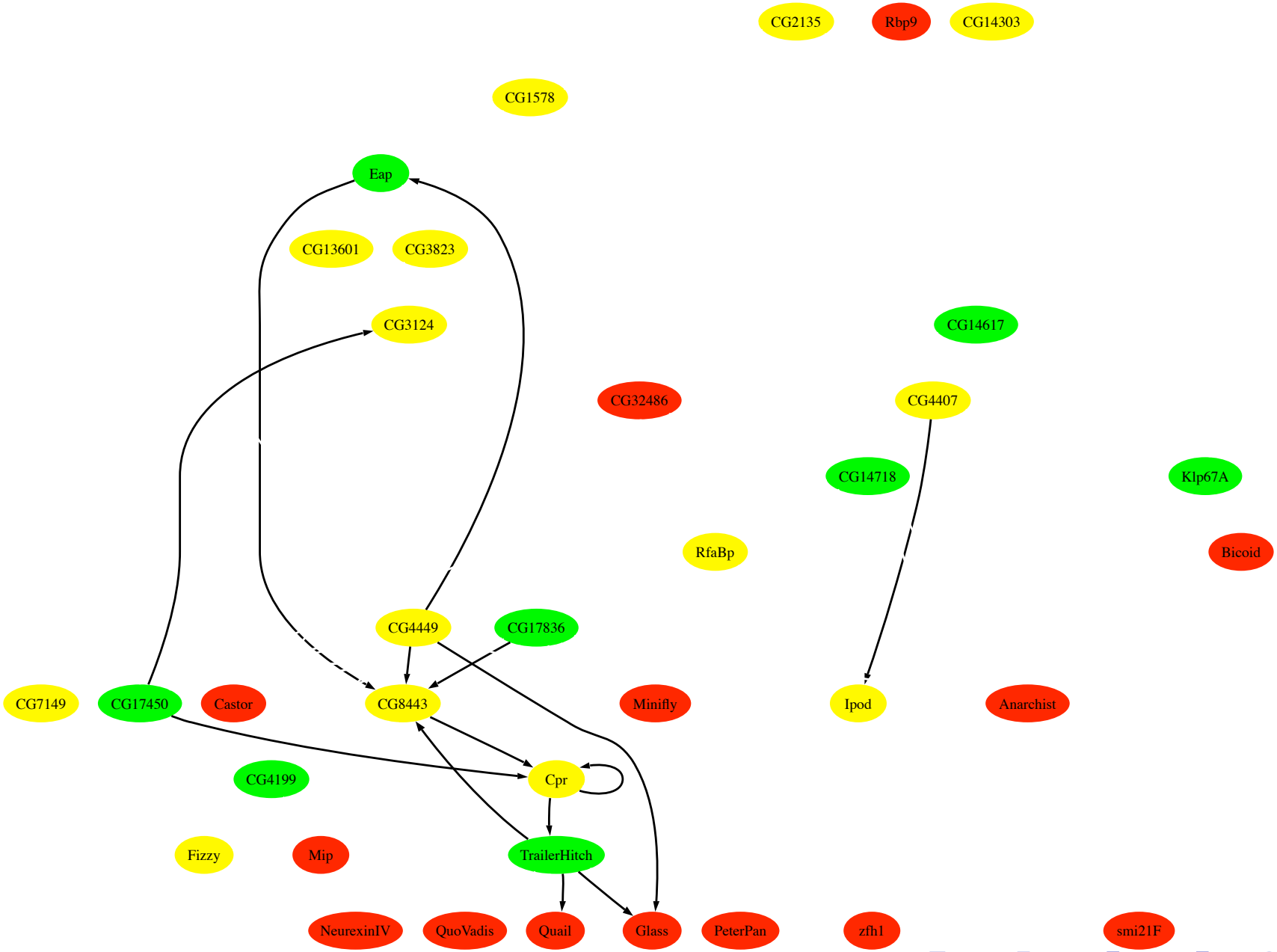
- Informative priors:

	Developmental process	Transcription Activity	Unknown
Target genes	13	7	13
Parent genes	3	8	10

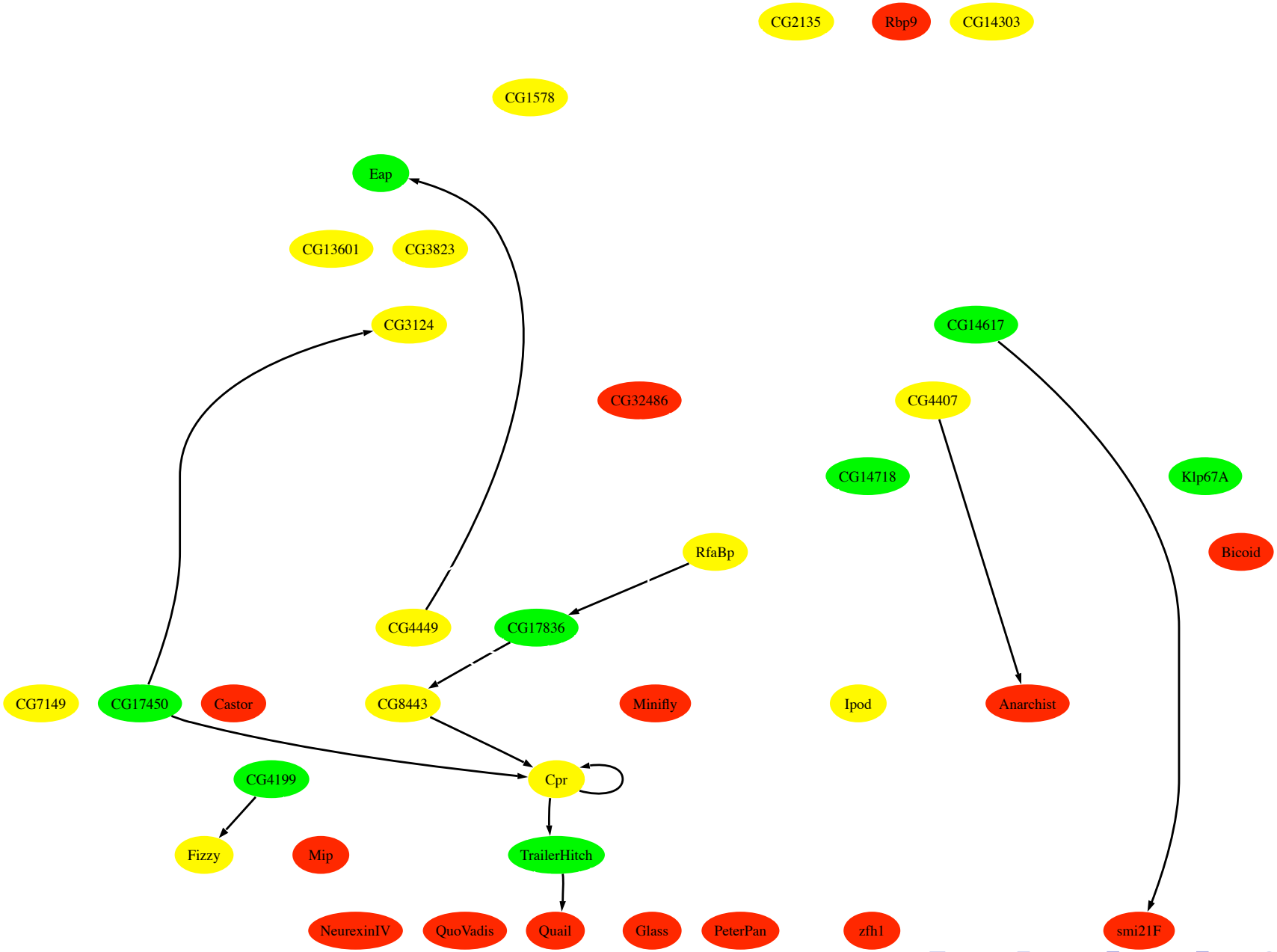
CP distribution (with informative priors on gene clusters)



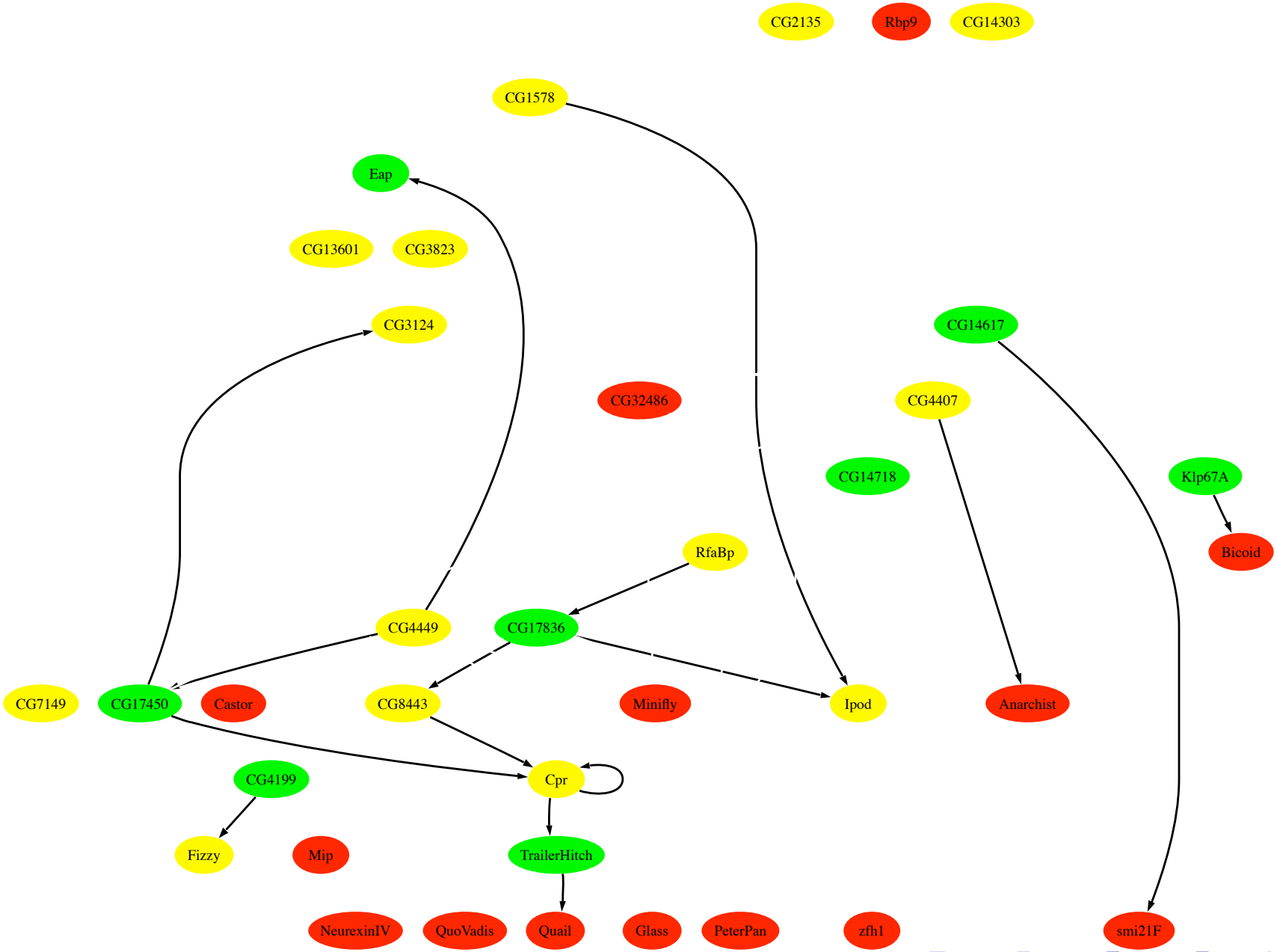
D. melanogaster: Early Embryo



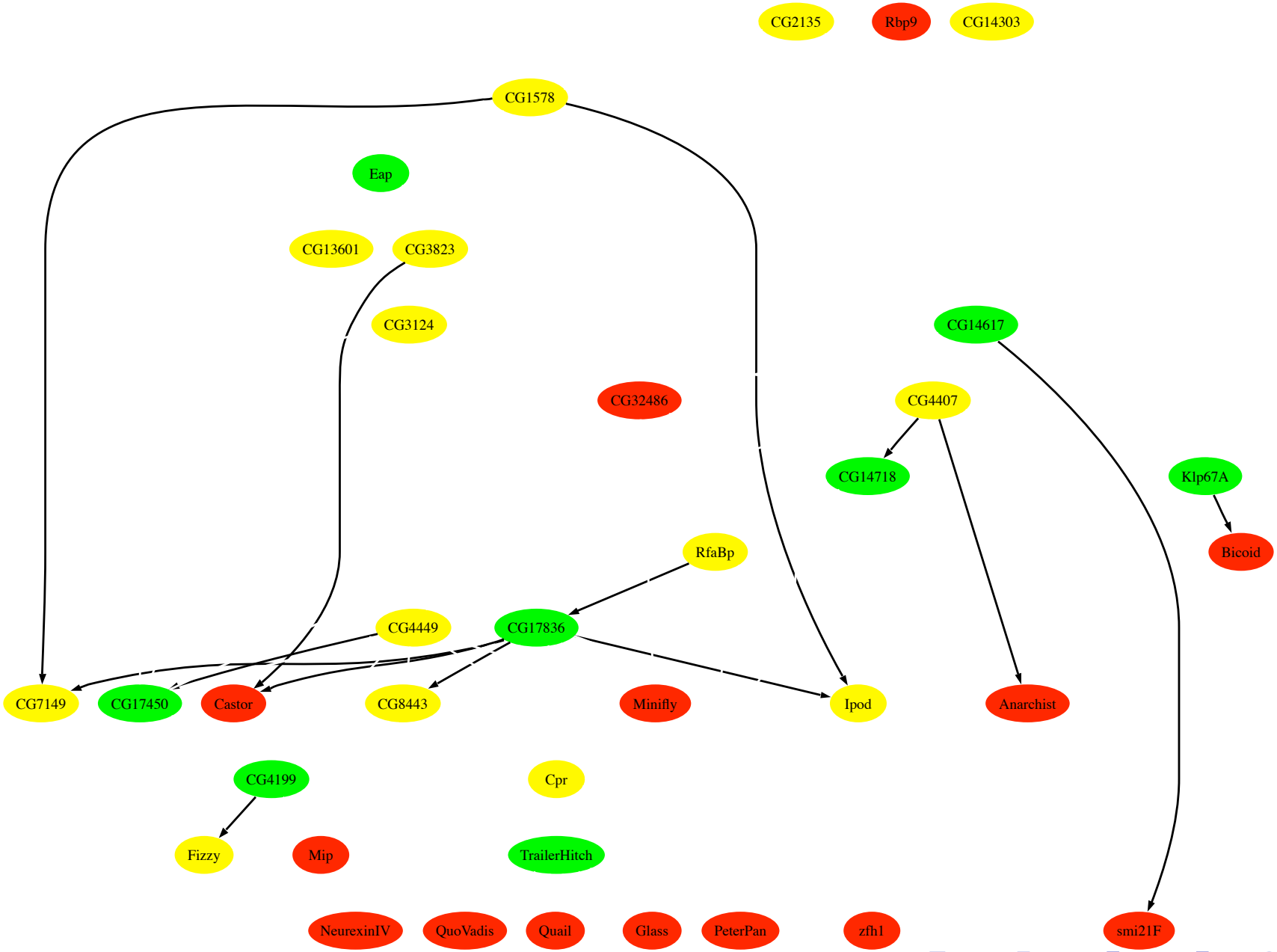
D. melanogaster: End Embryo



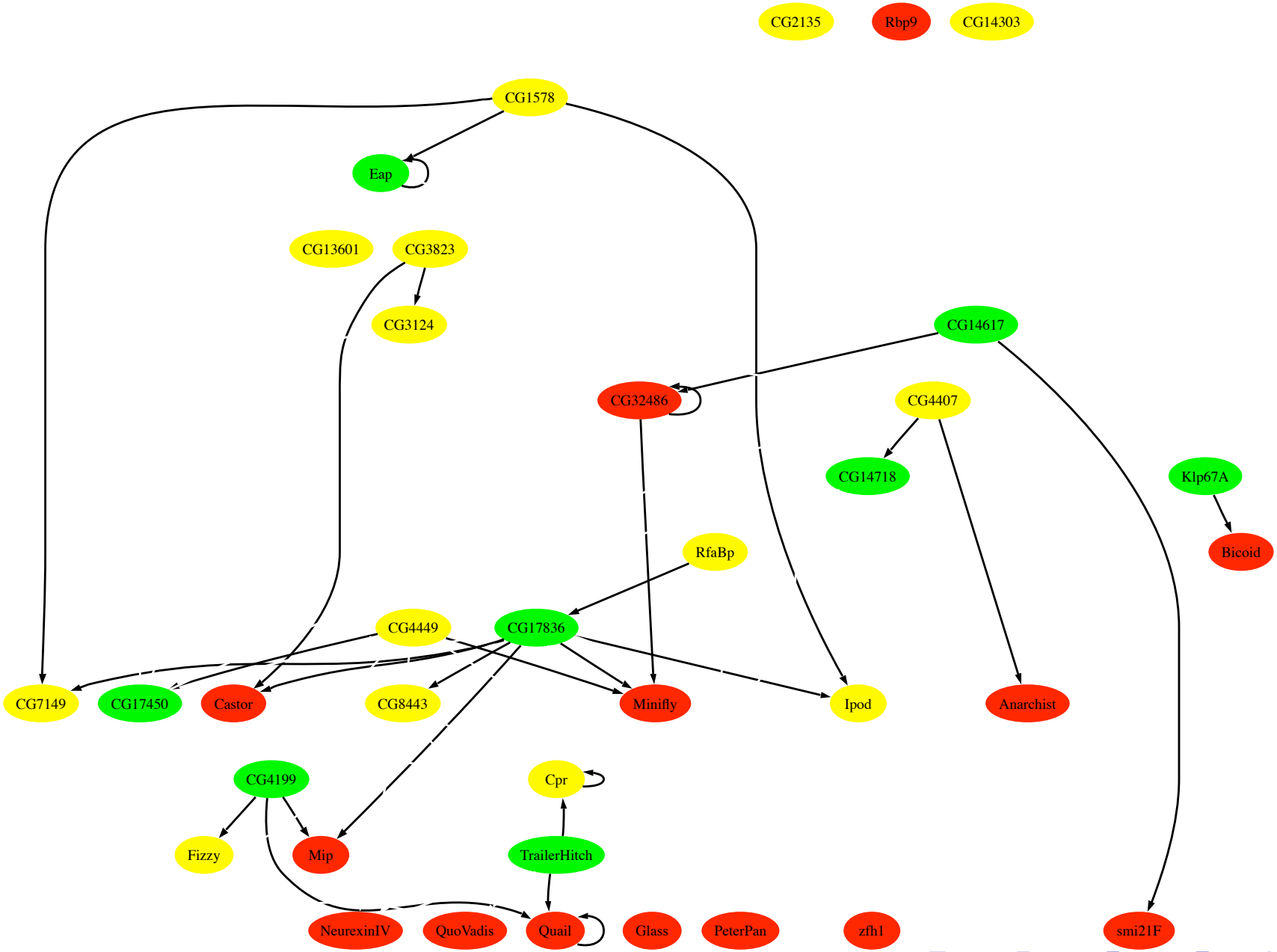
D. melanogaster : Larva



D. melanogaster: Pupa



D. melanogaster: Adult



Perspectives

- Results
  - A new approach allowing to infer time-dependent networks
  - Simultaneous inference of the changepoints position & the models within phases
  - Inclusion of some available biological knowledge

- Future Work
  - $\rightsquigarrow$ Larger scale
  - $\rightsquigarrow$ Cluster: muscle / nervous system / heart /...

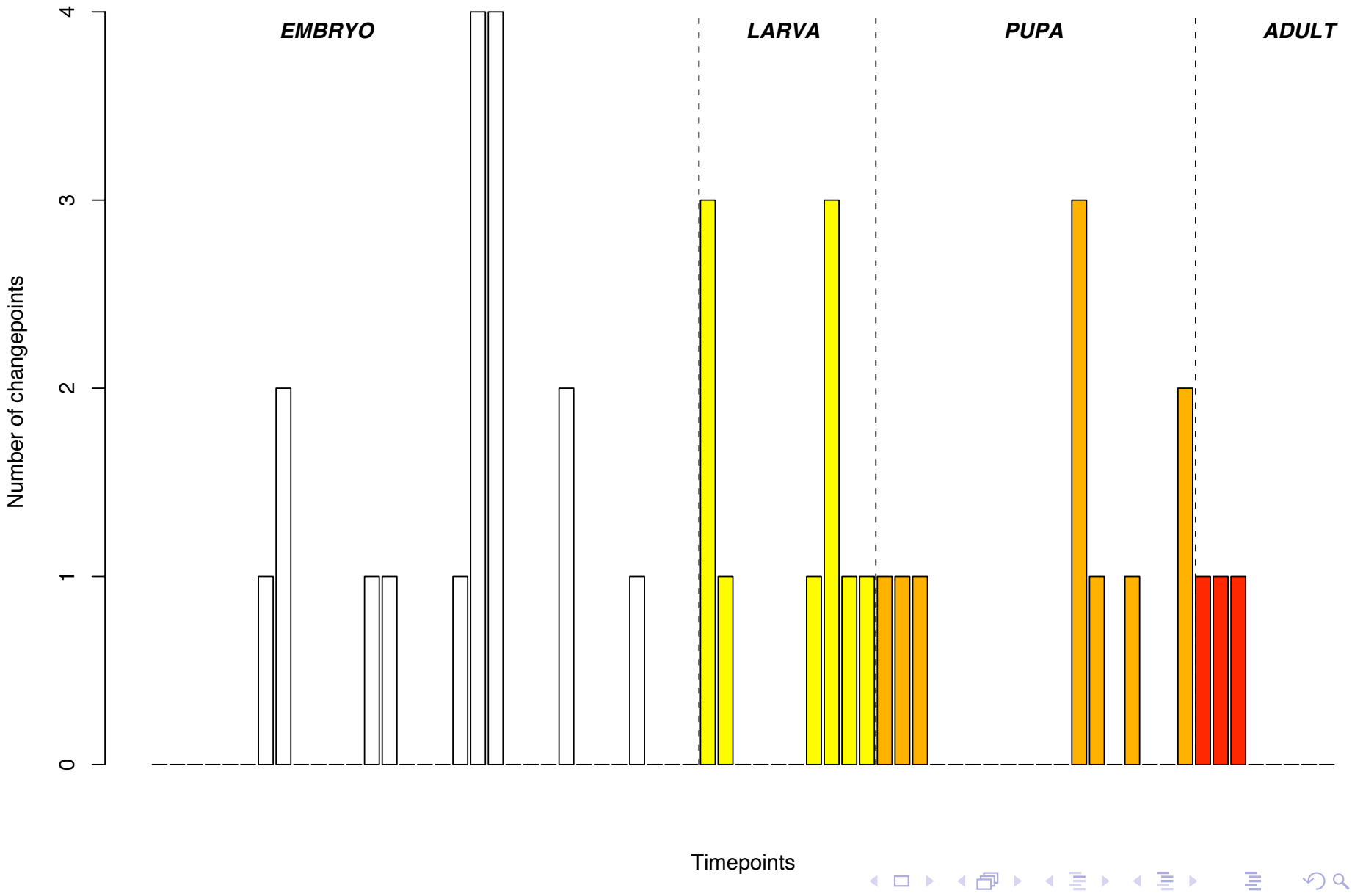
Perspectives

- Results
  - A new approach allowing to infer time-dependent networks
  - Simultaneous inference of the changepoints position & the models within phases
  - Inclusion of some available biological knowledge

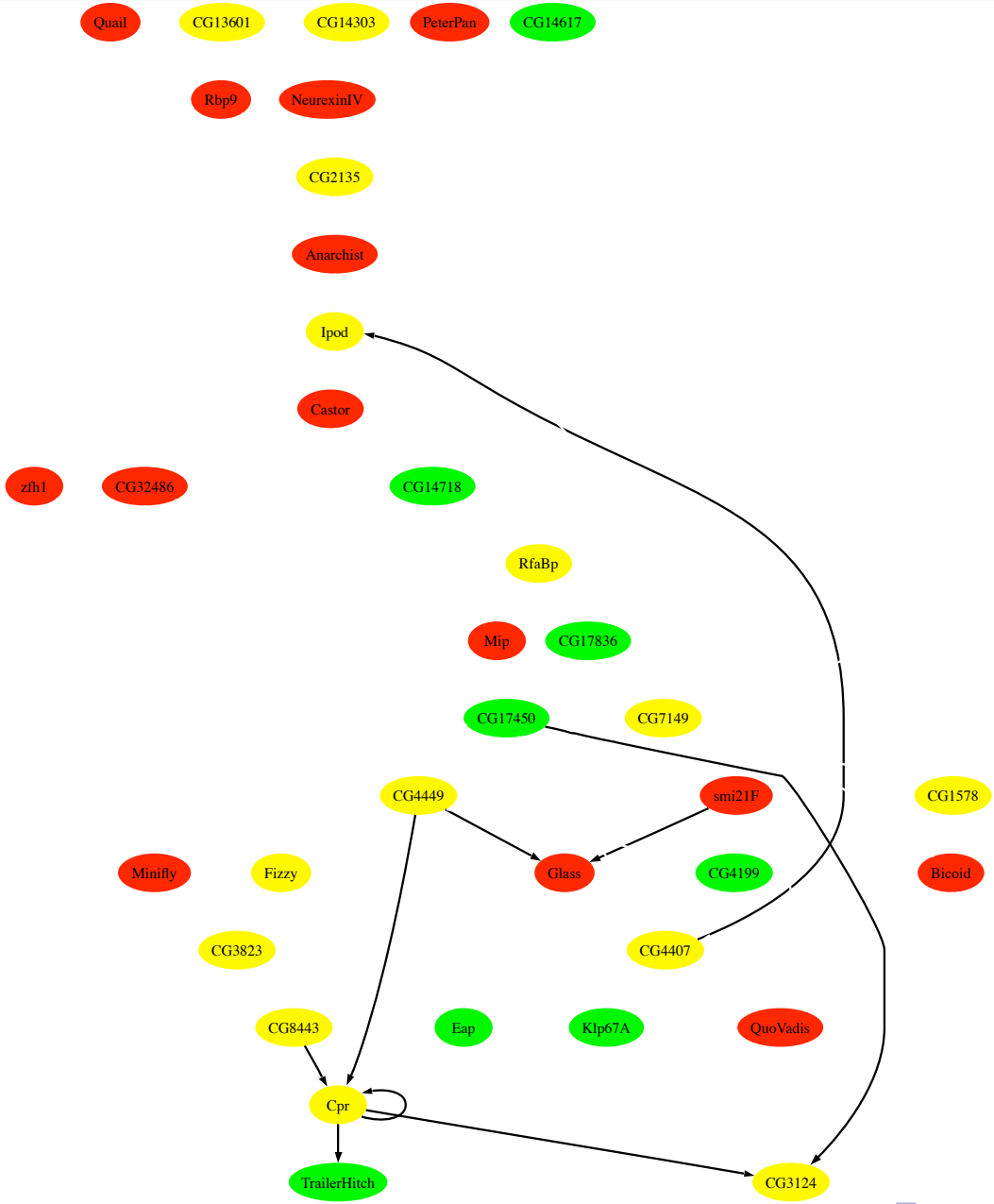
- Future Work
  - $\rightsquigarrow$ Larger scale
  - $\rightsquigarrow$ Cluster: muscle / nervous system / heart /...

Thank you for your attention!

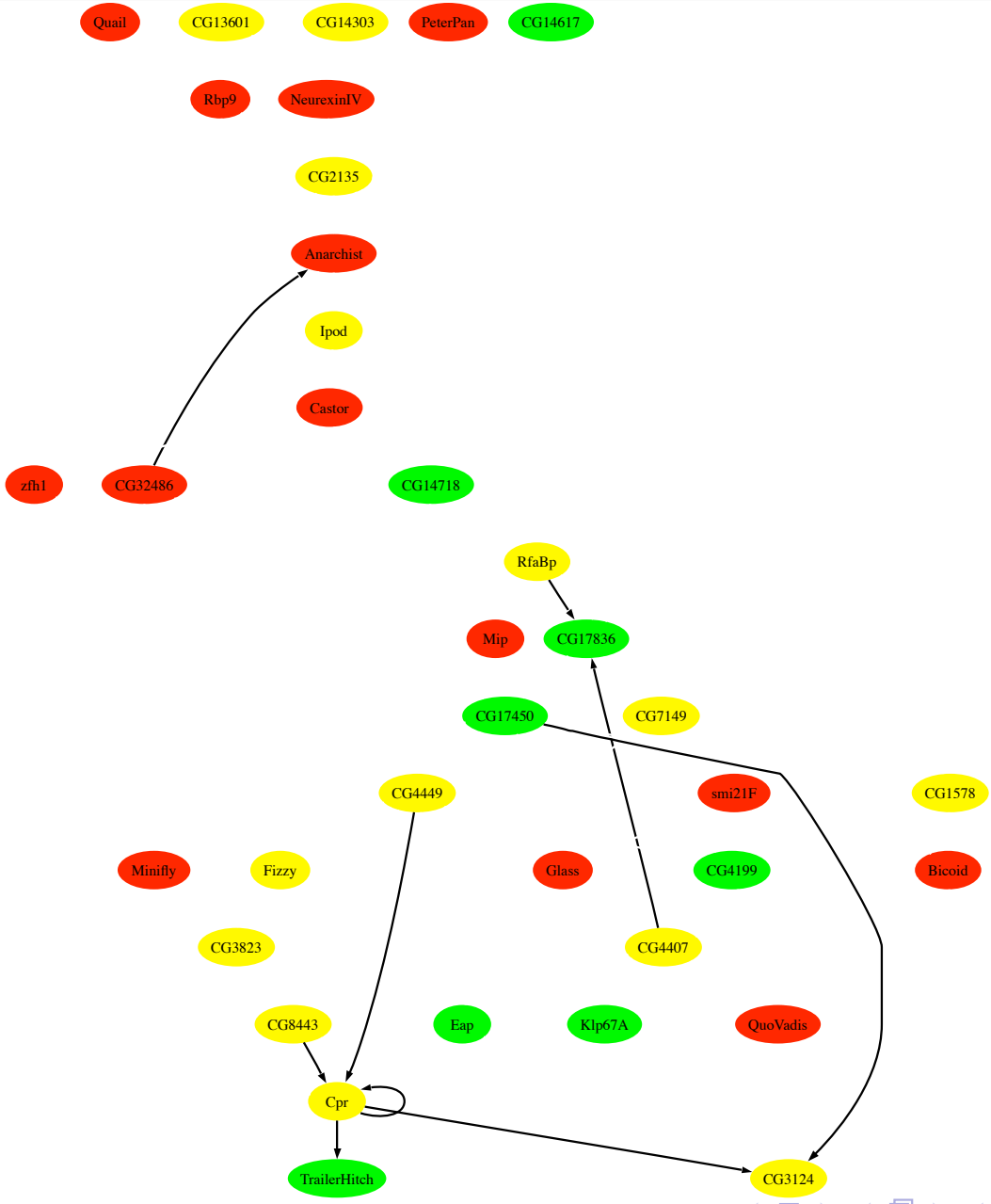
CP distribution (with uninformative priors)



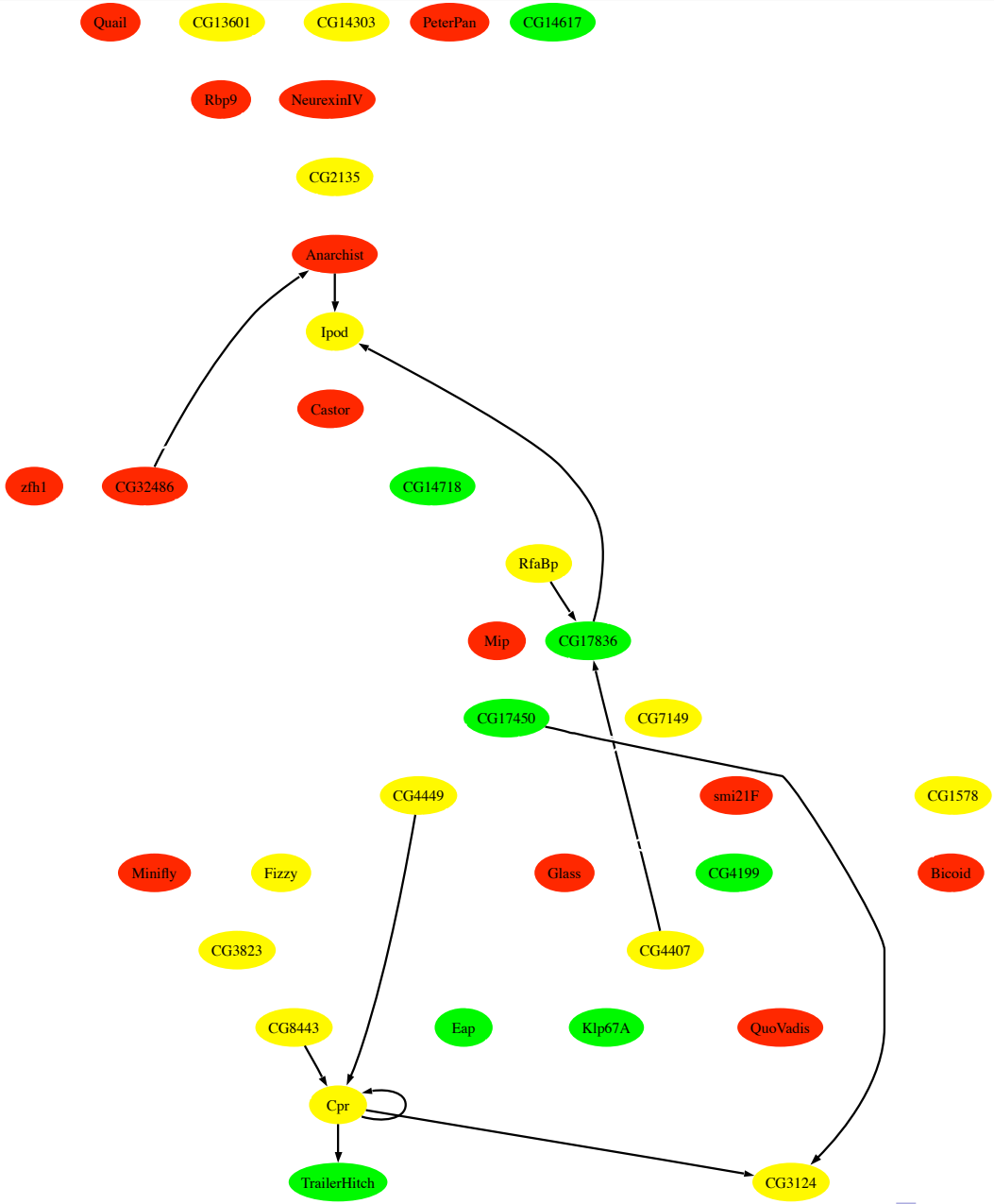
D. melanogaster: Early Embryo



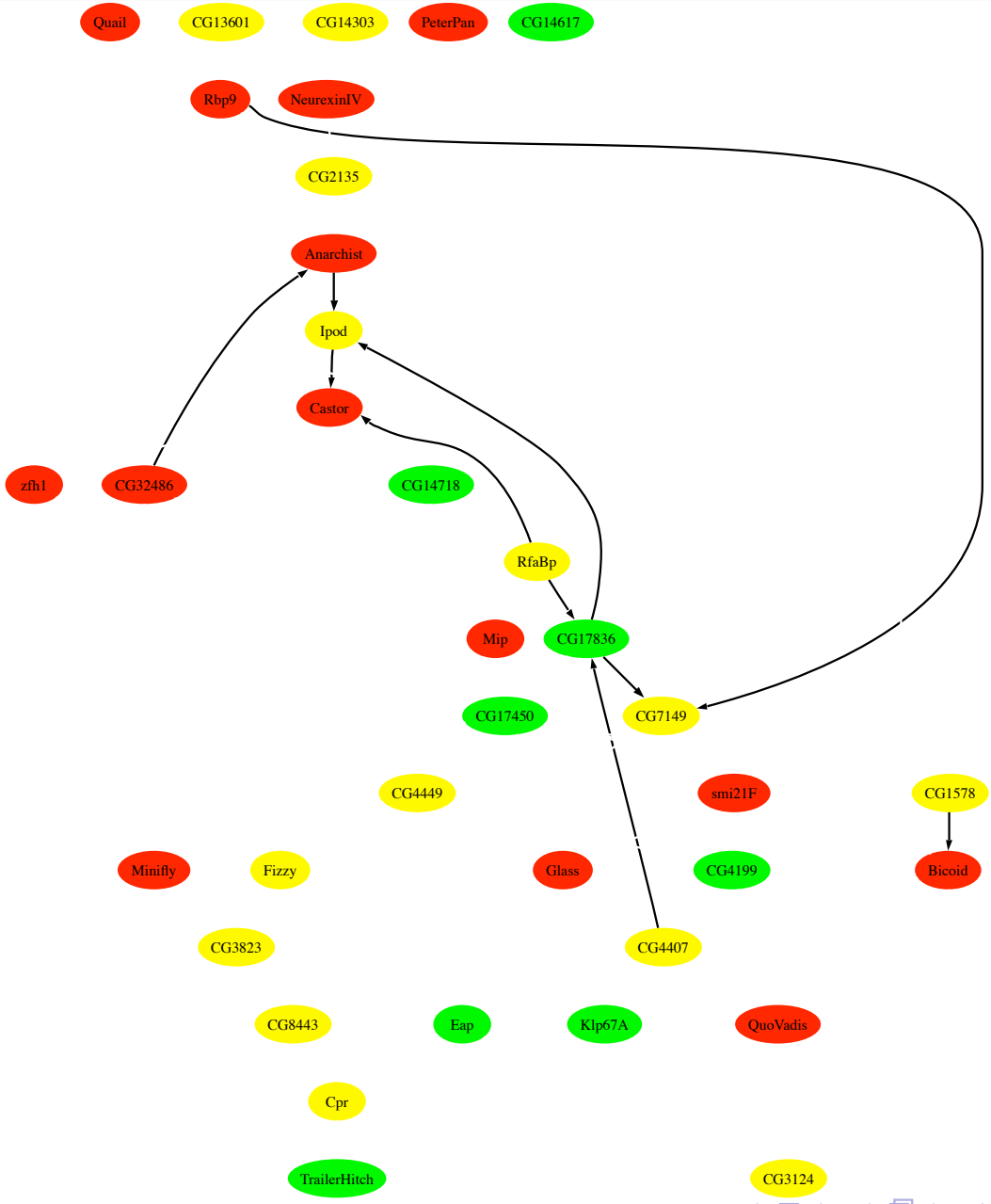
D. melanogaster: End Embryo



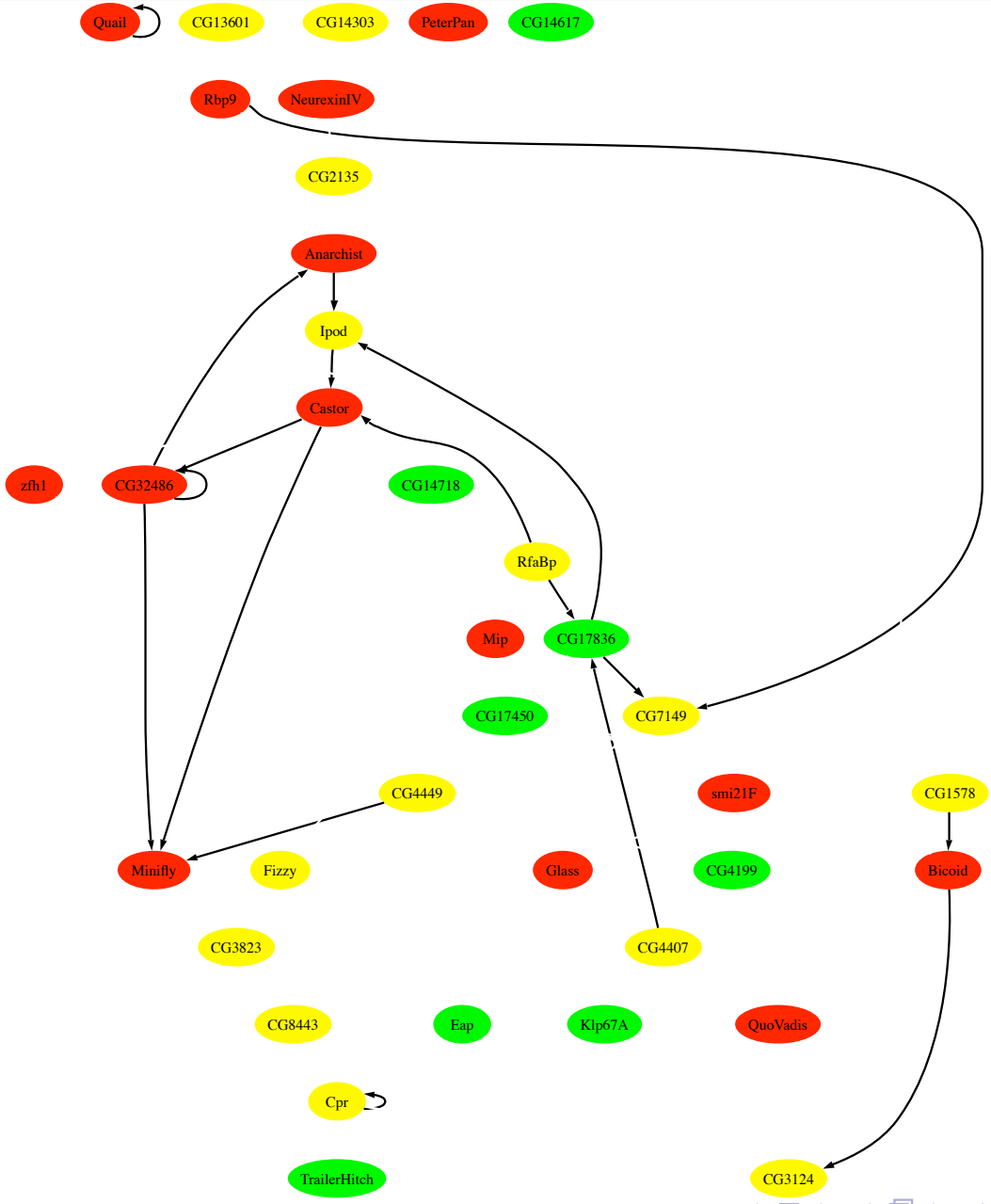
D. melanogaster : Larva



D. melanogaster: Pupa



D. melanogaster: Adult



Unknown dimension

The overall parameter space Θ writes as a finite union of subspaces

$$\Theta = \cup_{k=0}^{\overline{k}} E_k \times \Theta_k$$

where,

$$E_k = \left\{ \xi = \{\xi^i\}_{1 \leq i \leq p};\ \forall i \in P, \xi^i \subseteq \{2, .., n-1\}, |\xi^i| = k^i, \sum_{i=1}^{p} k^i = k \right\}$$

$$\Theta_k = \prod_{i=1}^{p} \prod_{h=0}^{k^i} \left\{ \cup_{s_h^i=0}^{\overline{s}} \{s_h^i\} \times B_{s_h^i} \right\},$$

with $B_0 = \mathbb{R} \times \mathbb{R}^+$, $B_{s_h^i} = \mathbb{R}^{s_h^i+1} \times \mathcal{P}_{s_h^i}(Q) \times \mathbb{R}^+$ for $k \geq 1$
and $\mathcal{P}_{s_h^i}(Q)$ contains all subsets of Q of dimension s_h^i.