# Stochastic simulations of gene regulatory networks with `sismonr`

Olivia Angelin-Bonnet, Patrick Biggs & **Matthieu Vignes**

SFS - Massey University

20 Sept 2021
NETBIO - INRAE

# Our research question: why do we observe living organisms...as they are?

- *Central dogma of Molecular Biology*:

  $$\texttt{DNA}\ (\approx \text{gene}) \rightarrow \texttt{RNA}\ (\text{blueprint}) \rightarrow \texttt{Protein}\ (\text{actor})$$

- 'Information flow' about sequence determination (suffers limitations, but mainly OK), not causal state(s) of the observed system
- Instead, in response to environmental conditions, molecular species (DNA, RNA, protein, complexes) interact with each other to impact (govern) the production (expression) of RNA and proteins. Whose tasks can impact other molecules' levels and/or observed system behaviour and structure (phenotypes)
- Feedback loops exist between these different levels

# Modelling with Gene Regulatory Networks (GRN)

- **Goal**: decipher the causal information flow from genotypes to phenotypes. Networks: ideal mathematical representations of these regulatory relationships
- More specifically: describe celuular processes
- Ideally account for space and time scales (e.g. see *VCell* environment, Resasco et al. 2012)
- Deterministic vs stochastic modelling?
- Practically: a model that predicts *entities*' abundance

# Why simulating GRN?

- Real system (usually) unknown
- Yet, we have observations such systems via different omics data sets
- Computational methods are used to reconstruct GRNs from data $\rightarrow$ *GRN inference* topic
- At least two useful frameworks for GRN simulations...
  as long as they are plausible

# Usefulness of GRN simulation tools

1. Evaluate the ability of GRN inference methods to reconstruct a complex system (structure) from observed data: "accuracy" of the method(s); common pathway accross species; needed sample size; signal-to-noise ratio feasibility; type of data: interventional vs observational, temporal vs steady state, heterogeneous biological entitites, missing observations, etc. $\rightarrow$ topic for Olivia's talk

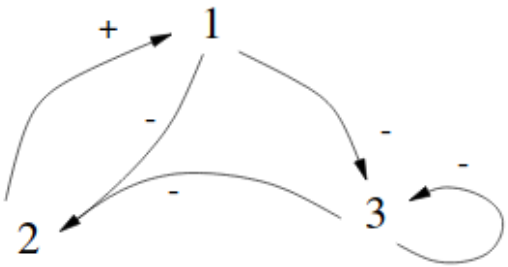
# Usefulness of GRN simulation tools

1. Evaluate the ability of GRN inference methods to reconstruct a complex system (structure) from observed data: "accuracy" of the method(s); common pathway accross species; needed sample size; signal-to-noise ratio feasibility; type of data: interventional vs observational, temporal vs steady state, heterogeneous biological entitites, missing observations, etc. $\rightarrow$ topic for Olivia's talk
2. Predict the behaviour of a known system: from model checking (La Rota et al. 2011) to predicting interventions (Bryce et al. 2010); genomics prediction (Pérez-Enciso and Zingaretti 2019); therapeutic targets (Ma et al. 2019, Fang et al. 2019); personalised medicine (van der Wijst et al. 2018, Gawel et al. 2019).

# GRN simulations, *Un inventaire à la Prévert*

aka a laundry list, de Jong 2002 & Shmulevich and Dougherty 2007

- Directed (or undirected) graphs $\rightarrow$ useful mainly to represent database knowledge

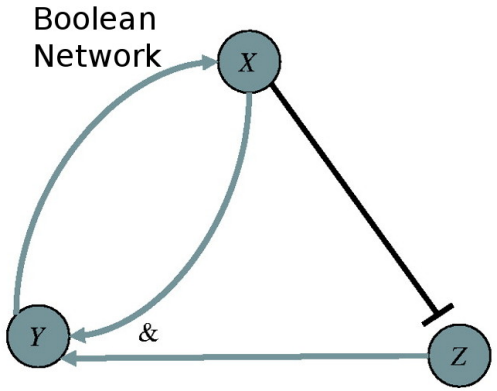



$$V = \{1, 2, 3\}$$
$$E = \{\langle 2, [1, 3], [-, -]\rangle, \langle 3, [1], [-]\rangle, \langle 1, [2], [+]\rangle, \langle 3, [3], [-]\rangle\}$$

# GRN simulations, *Un inventaire à la Prévert*

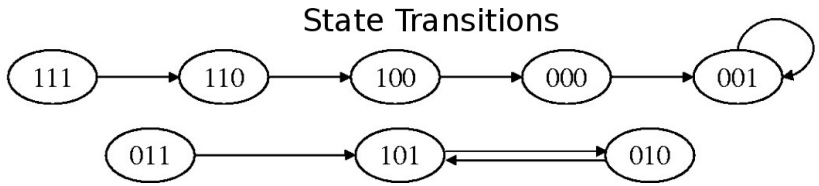
aka a laundry list, de Jong 2002 & Shmulevich and Dougherty 2007

- Directed (or undirected) graphs
- Boolean (or logical) regulatory networks




Transitions

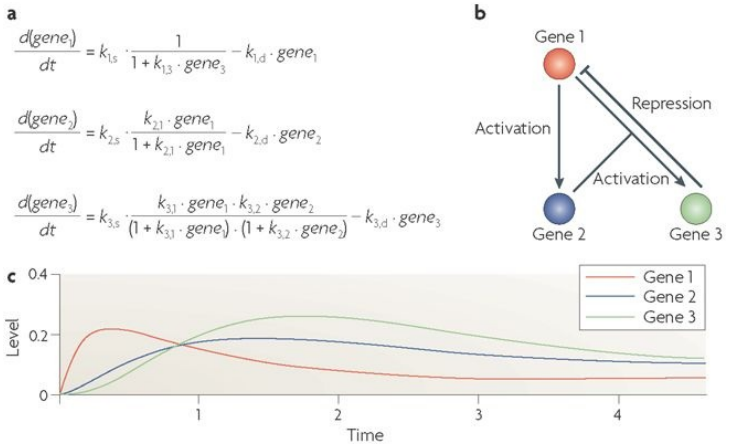
| $t$ | | | $t+1$ | | |
|---|---|---|---|---|---|
| $X$ | $Y$ | $Z$ | $X$ | $Y$ | $Z$ |
| 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 0 |



# GRN simulations, *Un inventaire à la Prévert*

aka a laundry list, de Jong 2002 & Shmulevich and Dougherty 2007

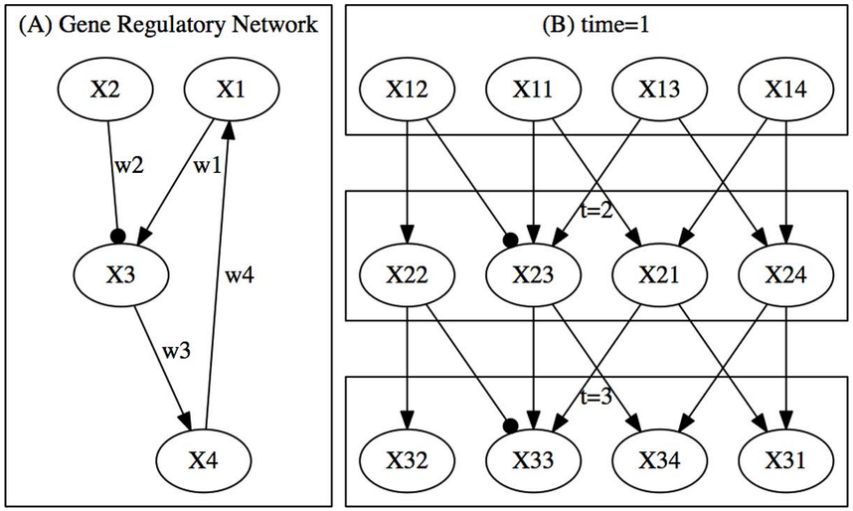
- Directed (or undirected) graphs
- Boolean (or logical) regulatory networks
- Coupled non-linear ordinary differential equations $\frac{dx_i}{dt} = f_i(x)$, $i = 1...p$ reflects production and degradation of all species $i$; can include time-delay; PLDE result from sigmoid approximations by step functions and allow a domain (linear) qualitative analysis



# GRN simulations, *Un inventaire à la Prévert*

aka a laundry list, de Jong 2002 & Shmulevich and Dougherty 2007

- Directed (or undirected) graphs
- Boolean (or logical) regulatory networks
- Coupled non-linear ordinary differential equations $\rightarrow$ piecewise-linear approximating equations or numerical methods
- Bayesian networks $\rightarrow P(X) = \prod_i P(X_i|\mathrm{Pa}(X_i))$ stochastic, but implicit temporal dependency; dynamical Bayesian network still require additional assumptions



# GRN simulations, *Un inventaire à la Prévert*

aka a laundry list, de Jong 2002 & Shmulevich and Dougherty 2007

- Directed (or undirected) graphs
- Boolean (or logical) regulatory networks
- Coupled non-linear ordinary differential equations $\rightarrow$ piecewise-linear approximating equations or numerical methods
- Bayesian networks
- **Stochastic (Master) equation** $\rightarrow$ Gillespie's algorithm with simplifications often necessary (Wilkinson 2018)

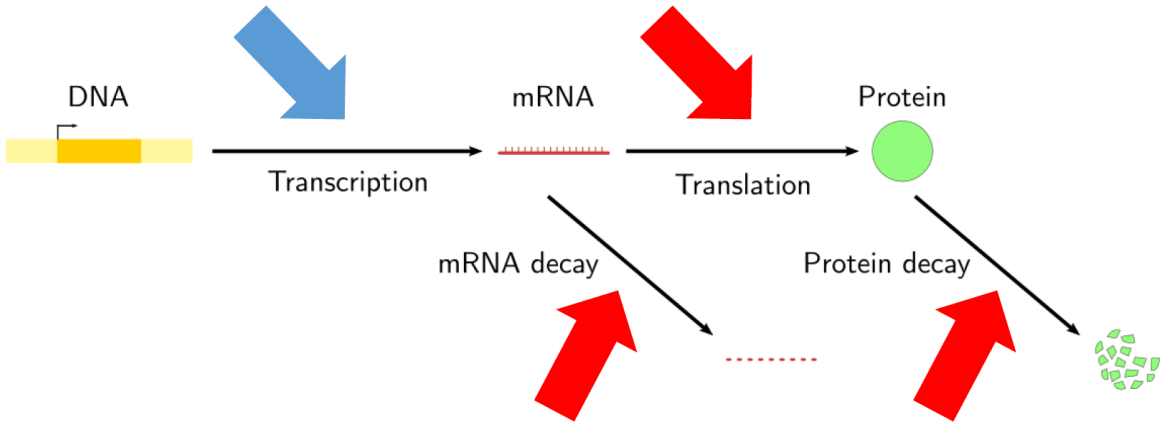
# GRN simulations, *Un inventaire à la Prévert*

aka a laundry list, de Jong 2002 & Shmulevich and Dougherty 2007

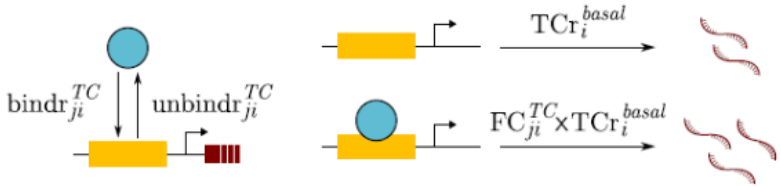
- Directed (or undirected) graphs
- Boolean (or logical) regulatory networks
- Coupled non-linear ordinary differential equations $\rightarrow$ piecewise-linear approximating equations or numerical methods
- Bayesian networks
- **Stochastic (Master) equation** $\rightarrow$ Gillespie's algorithm with simplifications often necessary (Wilkinson 2018)

- Note: some methods use *prior* information: global features, local motifs, white-/black-lists...

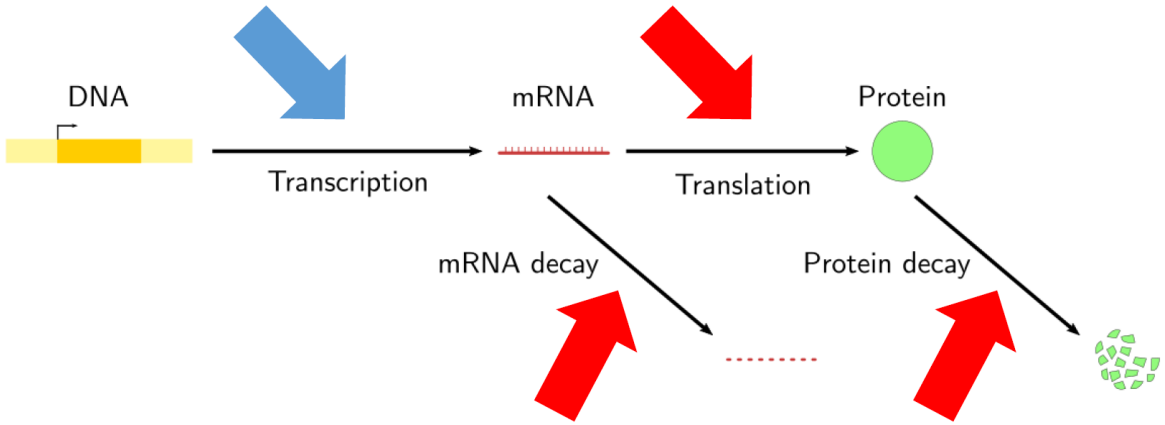
# Start with transcriptional regulations



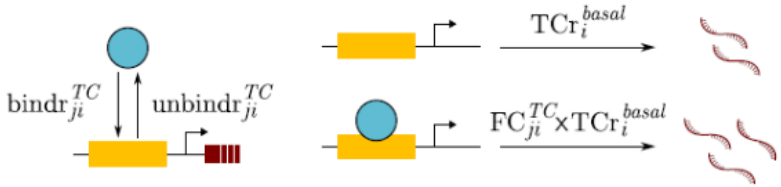

- Transcriptional regulations



# Start with transcriptional regulations



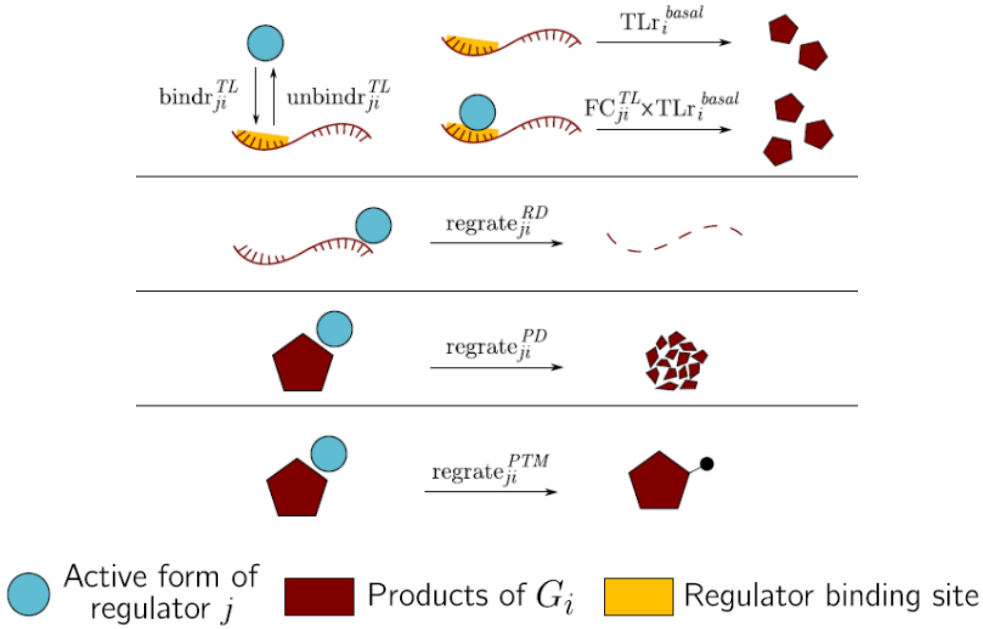

- Transcriptional regulations




- And beyond...

# "All biological features great and small"

- Post-transcriptional regulations

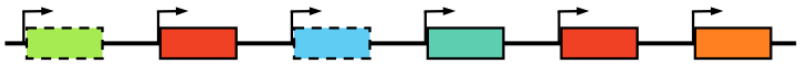


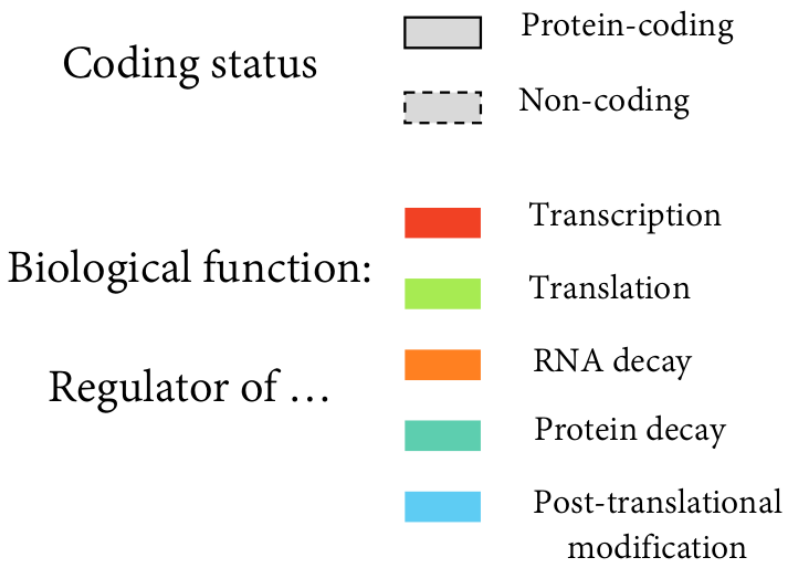

- Impact of genetic mutations: affects rates, binding or product activity.
- Ploidy of the organism.

# 1. Generating the network

```
mystsem <- createInSilicoSystem(G = 10, PC.p = 0.7)
```

All sorts of genes

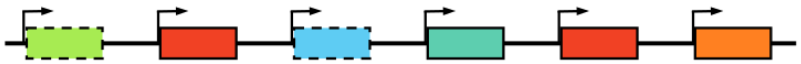


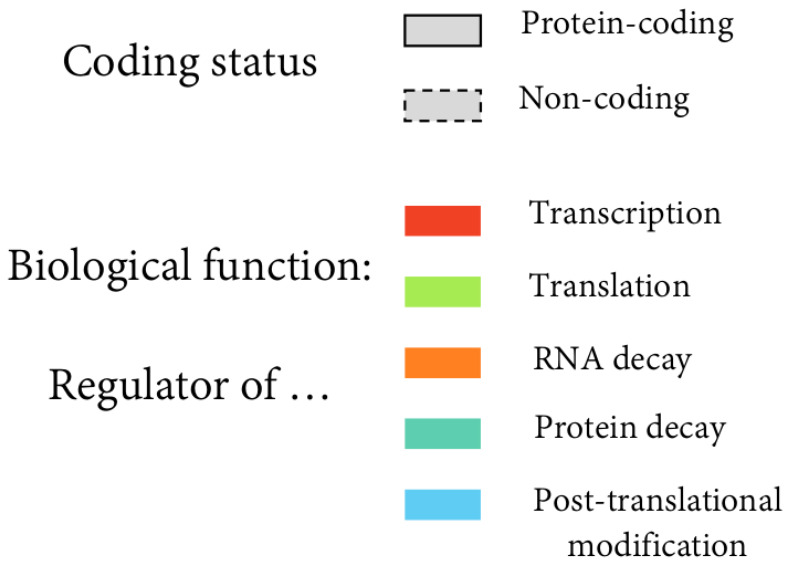


# 1. Generating the network

```
mystsem <- createInSilicoSystem(G = 10, PC.p = 0.7)
```

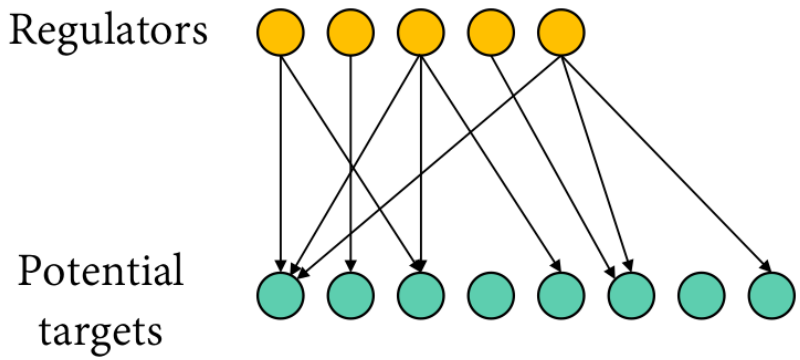
All sorts of genes






Creating the GRN

For each regulation class

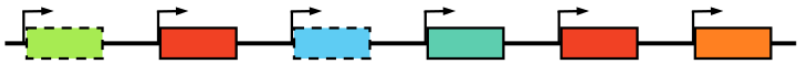


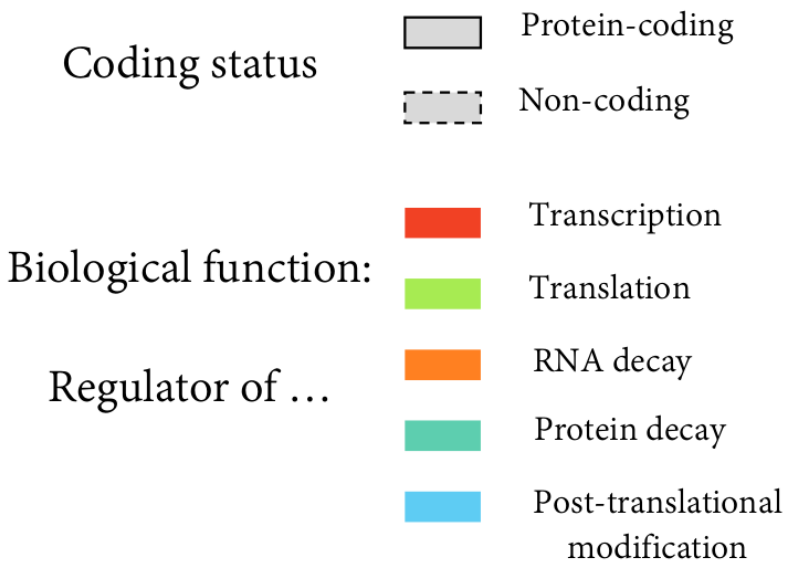

(with a given degree distribution)

# 1. Generating the network

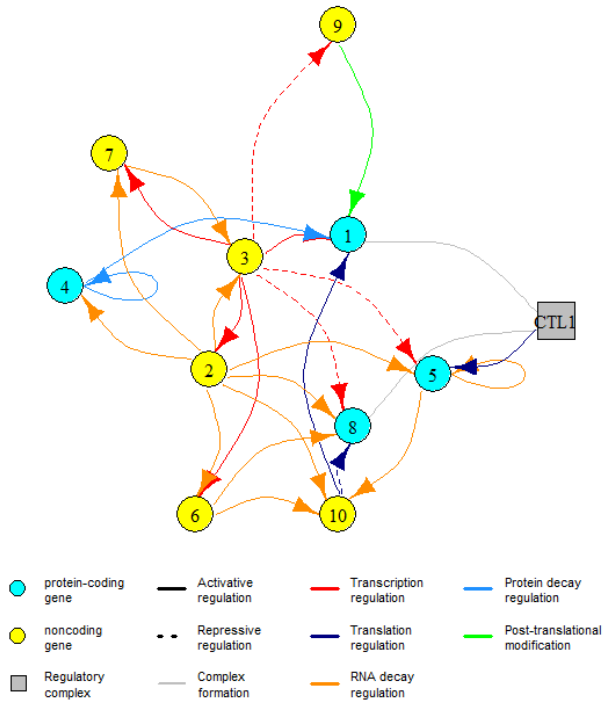
```
mystsem <- createInSilicoSystem(G = 10, PC.p = 0.7)
```

All sorts of genes



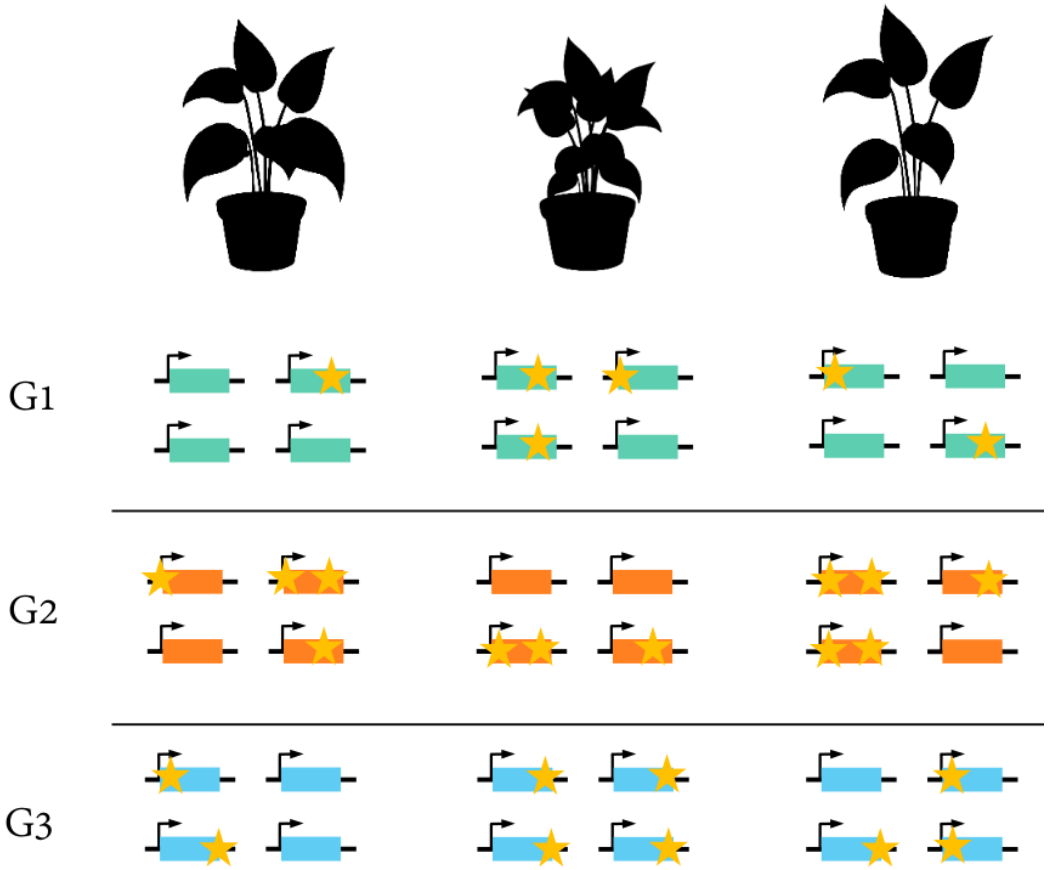



`plotGRN(mysystem)`



# 2. Creating genetically diverse *in silico* individuals

```
mypop <- createInSilicoPopulation(3, mysystem, ngenevariants = 5, ploidy
= 4)
```

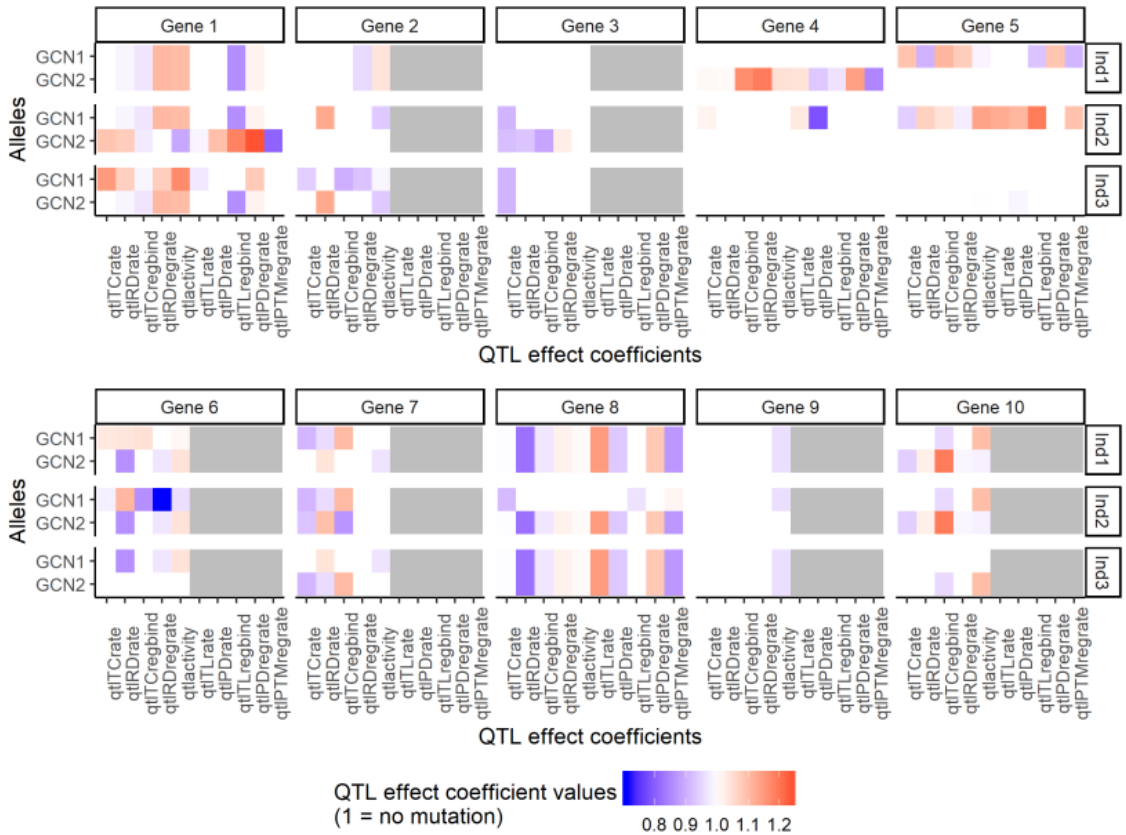



...

(3 individuals, 5 maximum gene variants, ploidy of 4)

# 2. Creating genetically diverse *in silico* individuals

`plotMutations(mypop, mysystem)`



# 3. Generating the stochastic model and numerically solving the Master equation

```
sim <- simulate(Parallel)InSilicoSystem(mysystem, mypop, simtime = 1000,
ntrials = 5)
```

$$\begin{cases} r_1 : DNA_1 \rightarrow DNA_1 + RNA_1 \\ r_2 : RNA_1 \rightarrow RNA_1 + Prot_1 \\ r_3 : DNA_3 + Prot_1 \rightarrow DNA_3^{\text{bound}} \\ r_4 : DNA_3^{\text{bound}} \rightarrow DNA_3 + Prot_1 \\ \dots \end{cases}$$

After init. & until convergence (maximum simul. time)

1. sample time to next reaction $\tau$
2. rates $r_i$ and concentrations$(t)$ $\Rightarrow$ propensities$(t)$ for reactions to occur during $[t, t+\tau]$
3. Sample the next reaction according to propensities
4. Update concentrations$(t+\tau)$

# 3. Generating the stochastic model and numerically solving the Master equation

`sim$Simulation`

| | time | trial | R5GCN2 | P5GCN2 | R7GCN2 | R3GCN1 | R1GCN2 | P1GCN2 | Pm1GCN2 | R9GCN1 | R6GCN2 | R10GCN2 | R1GCN1 | P1GCN1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 6 | 2 | 2 | 14 | 178 | 0 | 2 | 1 | 2 | 13 | 186 |
| 2 | 1 | 1 | 0 | 6 | 0 | 0 | 14 | 0 | 48 | 2 | 0 | 0 | 13 | 0 |
| 3 | 2 | 1 | 0 | 6 | 0 | 0 | 14 | 0 | 14 | 2 | 0 | 0 | 13 | 0 |
| 4 | 3 | 1 | 0 | 6 | 0 | 0 | 14 | 0 | 2 | 2 | 0 | 0 | 13 | 0 |
| 5 | 4 | 1 | 0 | 6 | 0 | 0 | 14 | 1 | 0 | 2 | 0 | 0 | 13 | 0 |
| 6 | 5 | 1 | 0 | 6 | 0 | 0 | 14 | 1 | 1 | 2 | 0 | 0 | 13 | 0 |

...

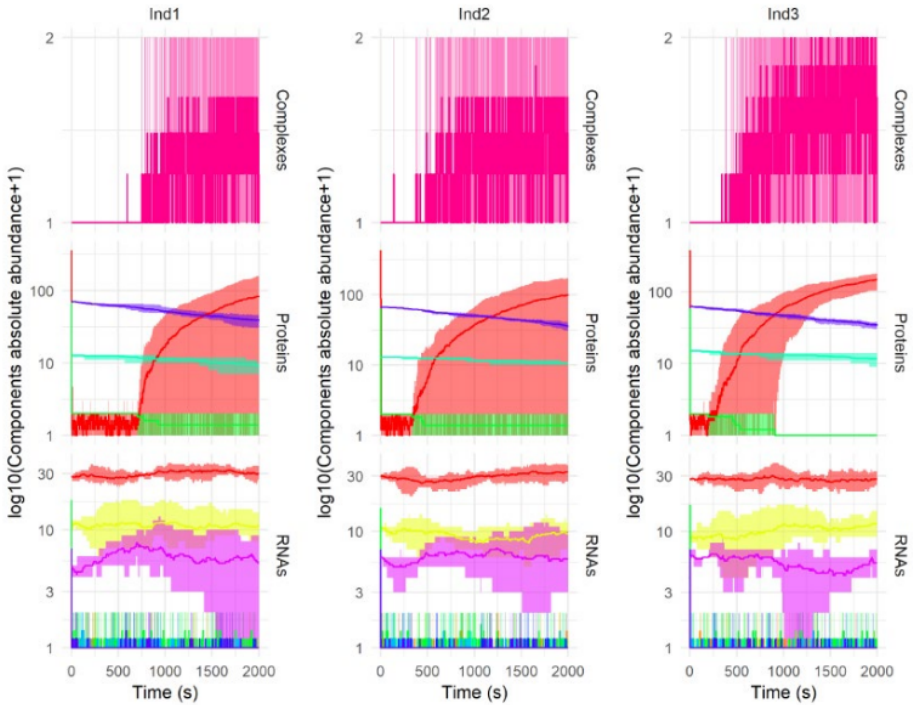
| | R8GCN1 | P8GCN1 | R6GCN1 | R10GCN1 | R2GCN2 | R8GCN2 | P8GCN2 | R5GCN1 | P5GCN1 | R4GCN1 | P4GCN1 | R3GCN2 | R2GCN1 | R9GCN2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 37 | 1 | 2 | 5 | 3 | 33 | 1 | 6 | 8 | 34 | 2 | 5 | 2 |
| 2 | 0 | 37 | 0 | 0 | 5 | 0 | 33 | 0 | 6 | 0 | 1 | 0 | 5 | 2 |
| 3 | 0 | 37 | 0 | 0 | 5 | 0 | 33 | 0 | 6 | 0 | 1 | 0 | 5 | 2 |
| 4 | 0 | 37 | 0 | 0 | 5 | 0 | 33 | 0 | 6 | 0 | 1 | 0 | 5 | 2 |
| 5 | 0 | 36 | 0 | 0 | 5 | 0 | 31 | 0 | 6 | 0 | 1 | 0 | 5 | 2 |
| 6 | 0 | 36 | 0 | 0 | 5 | 0 | 32 | 0 | 6 | 0 | 1 | 0 | 4 | 2 |

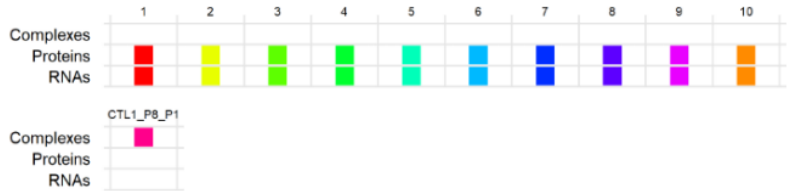
...

| | CTL1_P8GCN1_Pm1GCN2 | CTL1_P8GCN1_Pm1GCN1 | CTL1_P8GCN2_Pm1GCN1 | CTL1_P8GCN2_Pm1GCN2 | Ind |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | Ind1 |
| 2 | 0 | 0 | 0 | 0 | Ind1 |
| 3 | 0 | 0 | 0 | 0 | Ind1 |
| 4 | 0 | 0 | 0 | 0 | Ind1 |
| 5 | 1 | 0 | 1 | 1 | Ind1 |
| 6 | 1 | 0 | 1 | 0 | Ind1 |

# 3. Generating the stochastic model and numerically solving the Master equation

`plotSimulation(sim$Simulation)`



# Conclusion

## What `sismonr` does

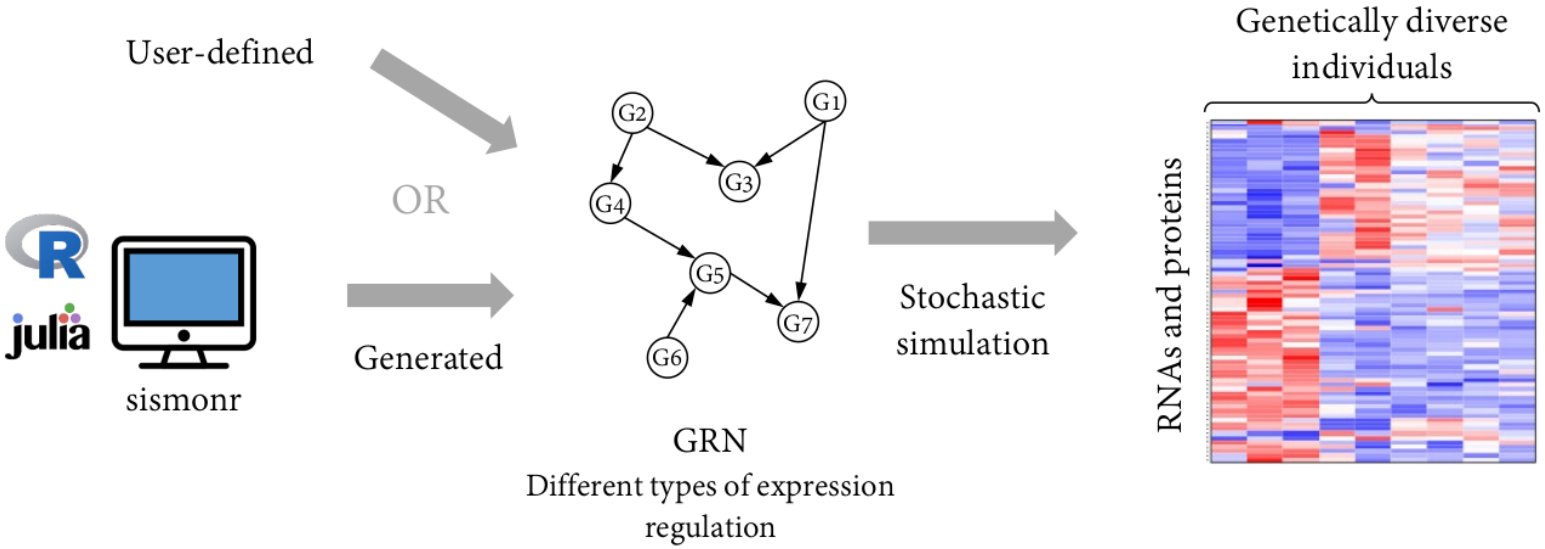
In three steps:

1. *In silico* system creation: genes & GRN
2. *In silico* individual creation: genetic mutation quantitative effects
3. Stochastic simulation of gene expressions for each individual

Full tutorial available at https://github.com/oliviaAB/sismonr and `sismonr` (v2.1) on the CRAN at https://cran.r-project.org/package=sismonr
(Angelin-Bonnet et al. 2020)

# Conclusion

## What `sismonr` does




Full tutorial available at https://github.com/oliviaAB/sismonr and `sismonr` (v2.1) on the CRAN at https://cran.r-project.org/package=sismonr (Angelin-Bonnet et al. 2020)

Thank you so much for your attention.

Any questions?

# Bibliography 1/2


H. de Jong

Modeling and simulation of genetic regulatory systems: a literature review.
*J Comput Biol*, 9(1):67-103, 2002.

I. Shmulevich and E.R. Dougherty
Genomic Signal Processing.
*Princeton University Press*, 2007.

D. Bryce, M. Verdicchio and K. Kim

Planning interventions in biological networks.
*ACM TIST*, 1(2):11, 2010.

C. La Rota, J. Chopard, P. Das, S. Paindavoine, F. Rozier, E. Farcot, C. Godin, J. Traas and F. Monéger

A Data-Driven Integrative Model of Sepal Primordium Polarity in Arabidopsis.
*Plant Cell*, 23(12):4318-4333, 2011.

D.C. Resasco, F. Gao, F. Morgan, I.L. Novak, J.C. Schaff and B.M. Slepchenko

Virtual Cell: computational tools for modeling in cell biology.
*Wiley Interdiscip Rev Syst Biol Med*, 4(2):129-140, 2012.

D.J. Wilkinson
Stochastic modelling for systems biology.
*CRC Press*, 3rd edition, 2018.

M.G.P. van der Wijst, D.H. de Vries, H. Brugge, H.-J. Westra and L. Franke

An integrative approach for building personalized gene regulatory networks for precision medicine.
*Genome Medicine*, 10(1):96, 2018.

# Bibliography 2/2


J. Ma, J. Wang, L.S. Ghoraie, X. Men, L. Liu and P. Dai

Network-based method for drug target discovery at the isoform level.
*Scientific Reports*, 9(1):13868, 2019.

M. Pérez-Enciso and L.M. Zingaretti

A Guide on Deep Learning for Complex Trait Genomic Prediction.
*Genes*, 10(7):553, 2019.

D.R. Gawel, J. Serra-Musach, [...], M. Benson

A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases.
*Genome Medicine*, 11(1):47, 2019.

H. Fang, The ULTRA-DD Consortium, [...], J.C. Knight

A genetics-led approach defines the drug target landscape of 30 immune-related traits.
*Nature Genetics*, 51(7):1082-1091, 2019.

O. Angelin-Bonnet, P.J. Biggs, S. Baldwin, S. Thomsonand M. Vignes

sismonr: Simulation of In Silico Multi-Omic Networks with adjustable ploidy and post-transcriptional regulation in R.
*Bioinformatics*, 36(9): 2938-2940, 2020.