Hierarchical modelling of genetic interaction in budding yeast

Darren Wilkinson
@darrenjw
tinyurl.com/darrenjw
School of Mathematics & Statistics,
Newcastle University, UK

Bayes@Lund
Lund University, Sweden
20th April, 2017

Overview

- High-throughput robotic genetic experiments
- Image analysis and data processing
- Stochastic modelling of growth curves
- Hierarchical modelling of genetic interaction
- Elucidating a network of genetic interactions

Joint work with Jonathan Heydari, Keith Newman, Conor Lawless and David Lydall (and others in the "Lydall lab")

Saccharomyces cerevisiae

- *Saccharomyces cerevisiae*, often known as budding yeast, and sometimes as brewer's yeast or baker's yeast, is a single-celled eukaryotic organism
- Eukaryotic cells contain a nucleus (and typically other organelles, such as mitochondria)
- It is useful as a model for higher eukaryotes, having a great deal of biological function conserved with humans
- It is the most heavily studied and well-characterised model organism in biology (eg. first fully sequenced eukaryote)

Synthetic Genetic Array (SGA)

- Possible to obtain a library of around 4,500 mutant strains, each of which has one of the non-essential genes silenced through insertion of a (kanMX) antibiotic resistance cassette and tagged with a unique DNA barcode
- These strains (stored frozen in 96-well plates) can be manipulated by robots in 96-well plates (8×12), or on solid agar in 96, 384 or 1536-spot format
- Synthetic Genetic Array (SGA) is a clever genetic procedure using robots to systematically introduce an additional mutation into each strain in the library by starting from a specially constructed query strain containing the new mutation

Telomeres

- The ends of linear chromosomes require special protection in order not to be targeted by DNA damage repair machinery (bacteria often avoid this problem by having just one chromosome arranged in a single loop)
- Telomeres are the ends of the chromosomal DNA (which have a special sequence), bound with special telomere-capping proteins that protect the telomeres
- CDC13 is an essential telomere-capping protein in yeast
- *cdc13-1* is a point-mutation of *cdc13*, encoding a temperature-sensitive protein which functions similarly to wild-type CDC13 below around $25\,^{\circ}\mathrm{C}$, and leads to "telomere-uncapping" above this temperature

Yeast Lab

- David Lydall's (budding) yeast lab uses a range of high throughput (HTP) technologies for genome-wide screening for interactions relevant to DNA damage response and repair pathways, with a particular emphasis on telomere maintenance
- Much of this work centres around the use of robotic protocols in conjunction with genome-wide knockout libraries and synthetic genetic array (SGA) technology to screen for genetic interactions with known telomere maintenance genes
- Quantitative fitness analysis (QFA) is the term we use for our system of robotic image capture, data handling, image analysis and data modelling

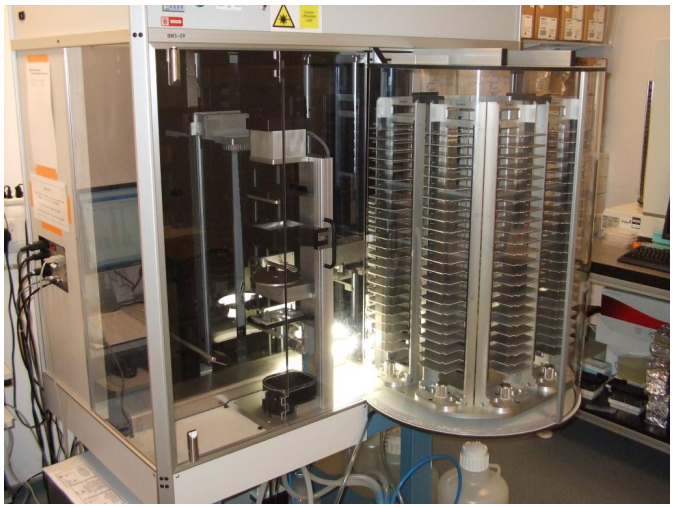
Basic structure of an experiment

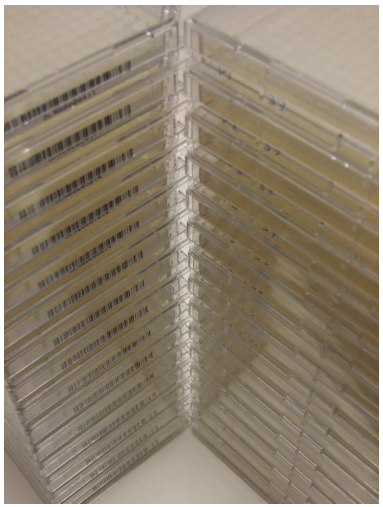
1. Introduce a mutation (such as *cdc13-1*) into an SGA query strain, and then use SGA technology (and a robot) to cross this strain with the single deletion library in order to obtain a new library of double mutants
2. Inoculate the strains into liquid media, grow up to saturation then spot back on to solid agar 4 times
3. Incubate the 4 different copies at different temperatures (treatments), and image the plates multiple times to see how quickly the different strains are growing
4. Repeat steps 2 and 3 four times (to get some idea of experimental variation)
5. Repeat steps 2 to 4 with a "control" library that does not include the query mutation

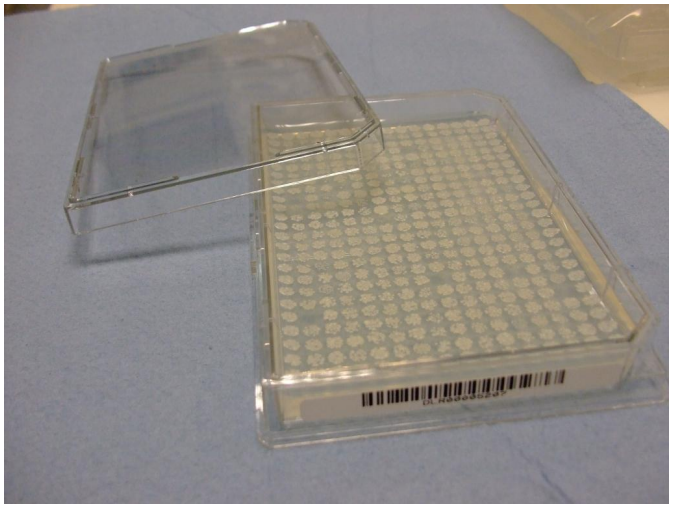
Some numbers relating to an experiment

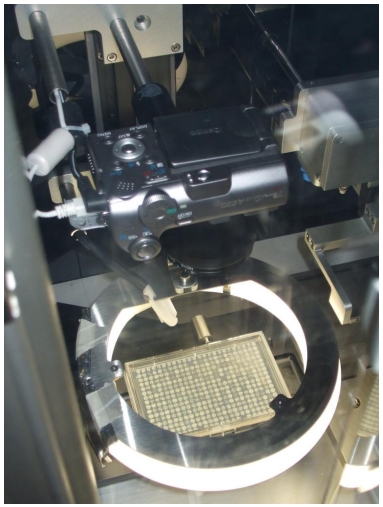
- Initial SGA work (introducing mutations into the query and the library) takes around 1 month of calendar time, and several days of robot time
- The inoculation, spotting and imaging of the 8 repeats takes 1 month of calendar time, and around 2 weeks of robot time
- The experiment uses around £3,000 of consumables (plastics and media)
- The library is distributed across 72 96-well plates or 18 solid agar plates (in 384 format, or 1536 in quadruplicate)
- If each plate is imaged 30 times, there will be around 35k high-resolution photographs of plates in 384 format, corresponding to around 13 million colony growth measurements (400k time series)
- This is big data!

HTP SGA robotic facility







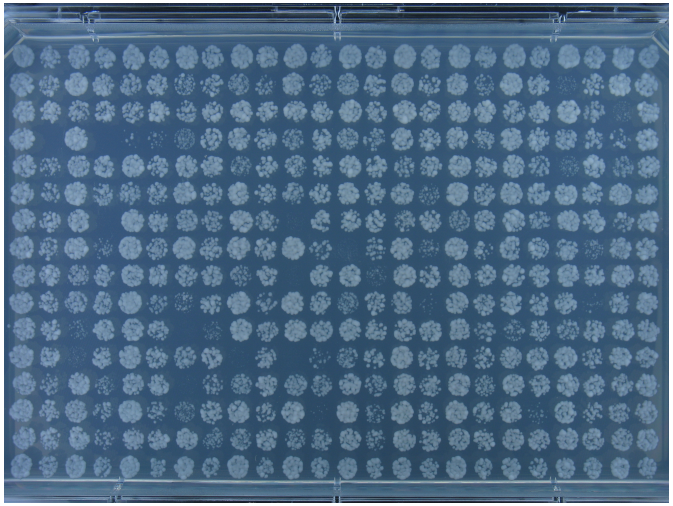


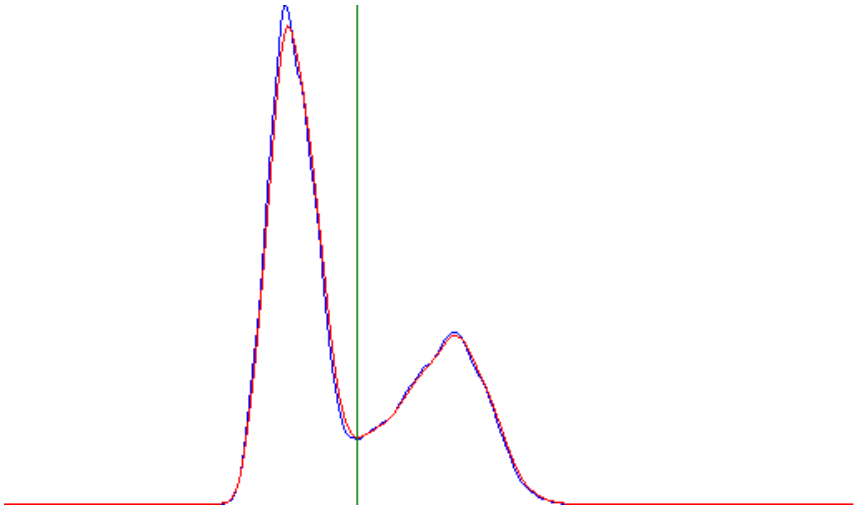
Data analysis pipeline

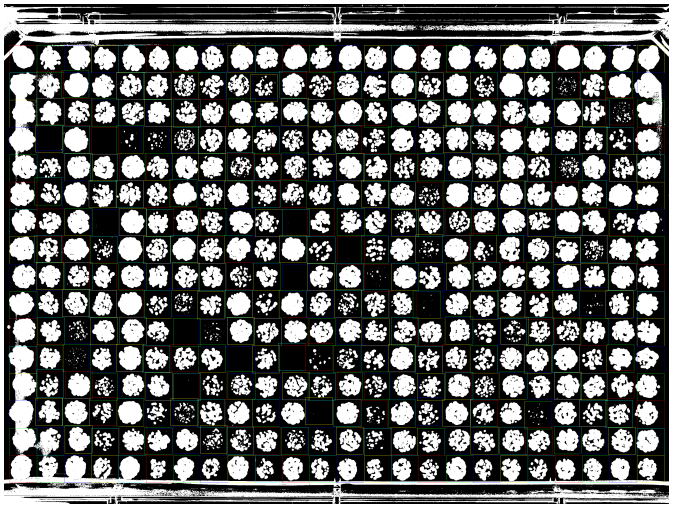
- Image processing (from images to colony size measurements)
- Fitness modelling (from colony size growth curves to strain fitness measures)
- Modelling genetic interaction (from strain fitness measures to identification of genetically interacting strains, ranked by effect size)

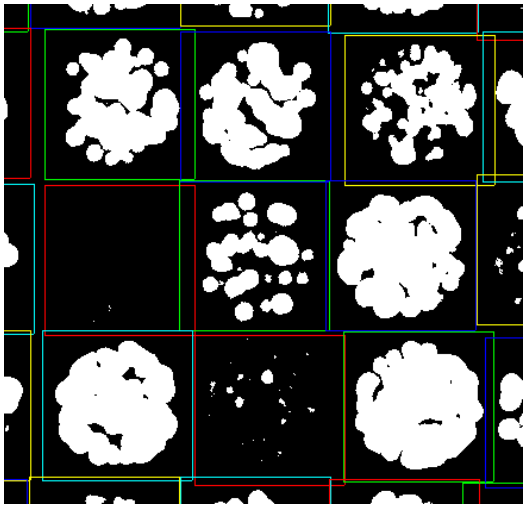
Possible to carry out three stages separately, but benefits to joint modelling through borrowed strength and proper propagation of uncertainty. Not practical to integrate image processing step into the joint model, but possible to jointly model second two stages.

Automated image analysis (Colonyzer)

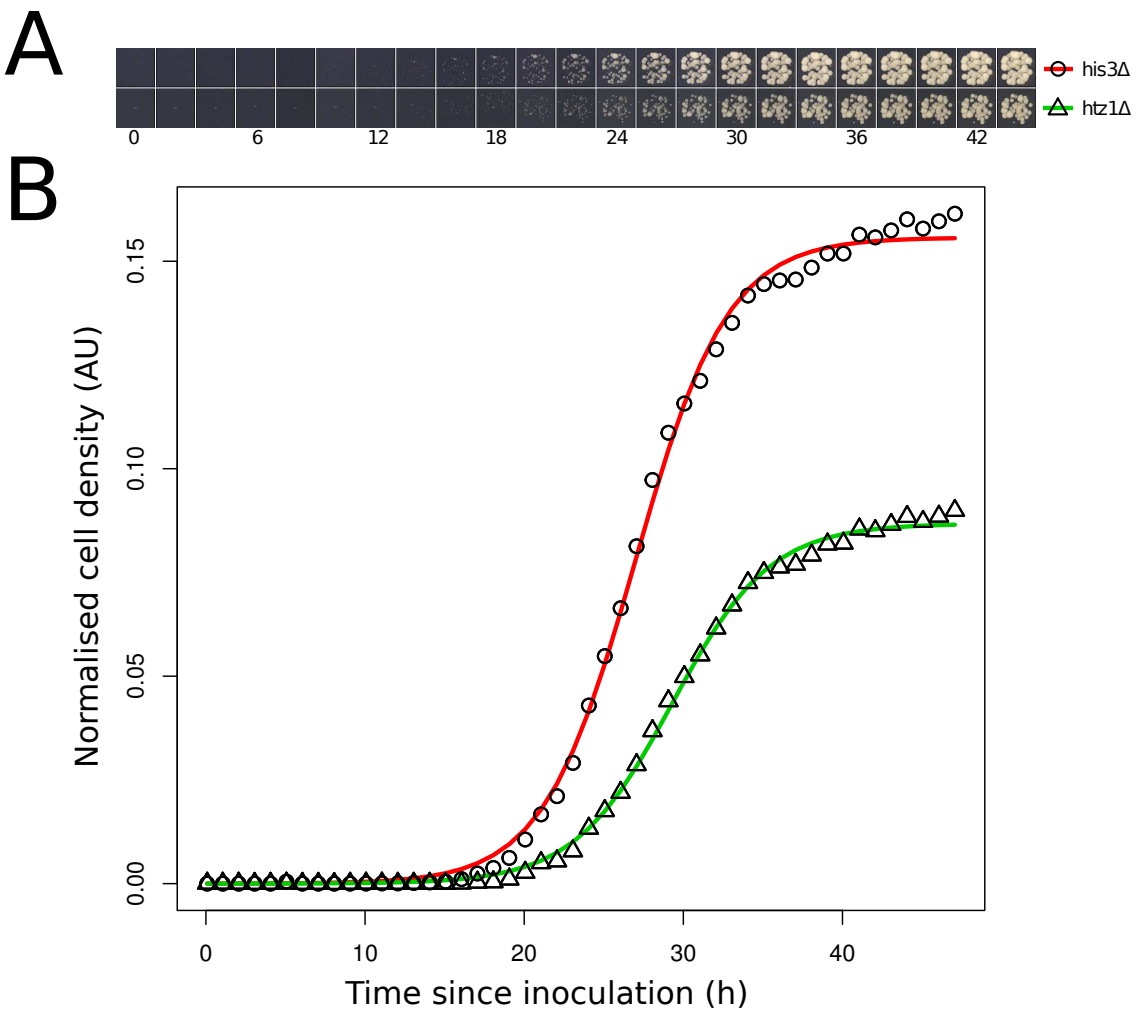








Growth curve



Growth curve modelling

- We want something between a simple smoothing of the data and a detailed model of yeast cell growth and division
- Logistic growth models are ideal — simple semi-mechanistic models with interpretable parameters related to strain fitness
- Basic deterministic model:

$$\frac{dx}{dt} = rx(1 - x/K),$$

  subject to initial condition $x = P$ at $t = 0$
- r is the growth rate and K is the carrying capacity
- Analytic solution:

$$x(t) = \frac{KPe^{rt}}{K + P(e^{rt} - 1)}$$

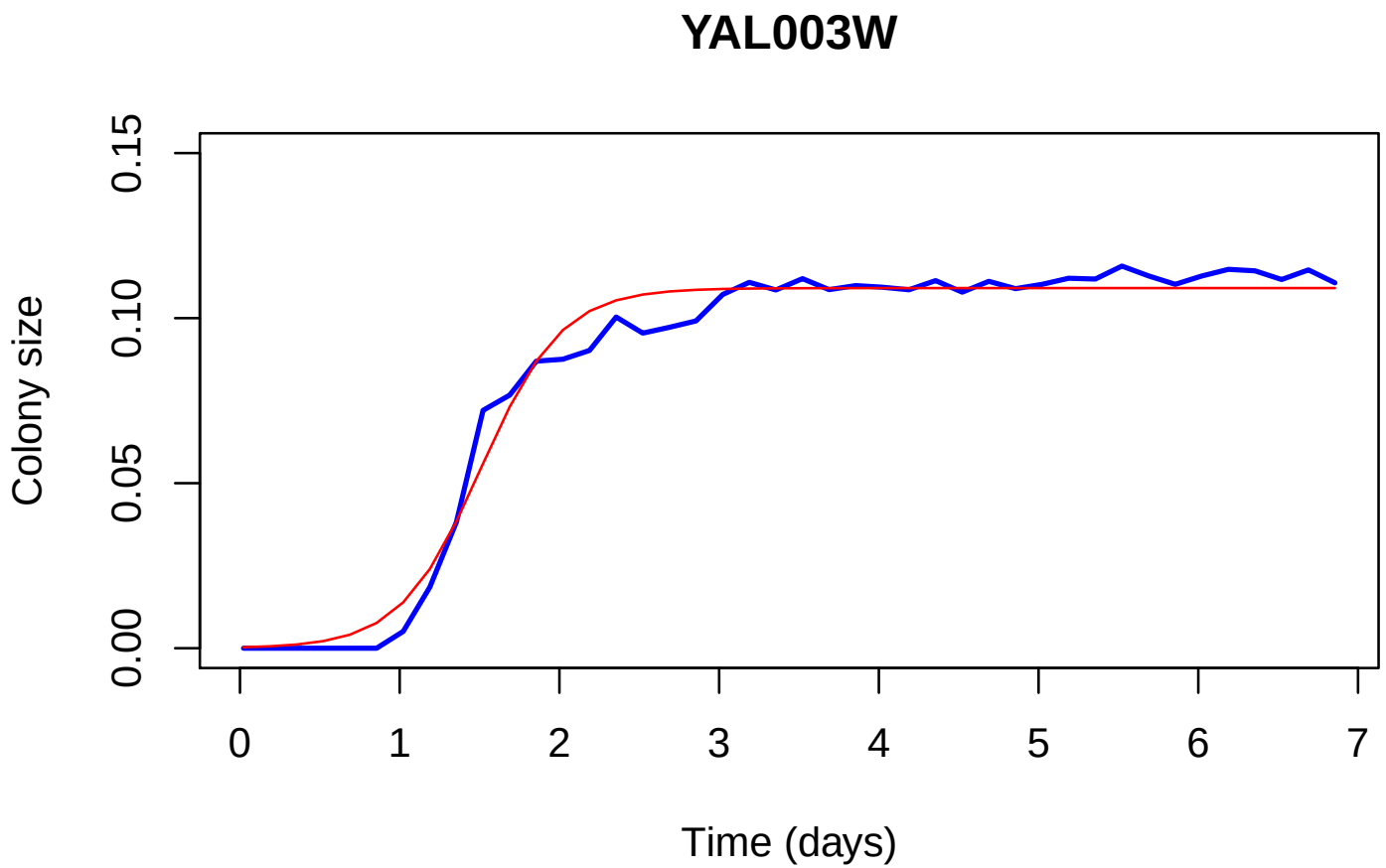
Statistical model

- Model observational measurements $\{Y_{t_1}, Y_{t_2}, \ldots\}$ with

$$Y_{t_i} = x_{t_i} + \varepsilon_{t_i}$$

- Can fit to observed data y_{t_i} using non-linear least squares or MCMC
- Can fit all (400k) time courses simultaneously in a large hierarchical model which effectively borrows strength, especially across repeats, but also across genes
- Generally works well (fine for most of the downstream scientific applications), but fit is often far from perfect...

Fitting the logistic curve



Improved modelling of colony growth curves

- Could use a generalised logistic model (Richards' curve) which breaks the symmetry in the shape of "take off" and "landing"

$$\frac{dx}{dt} = rx(1 - (x/K)^{\nu})$$

- This helps, but doesn't address the real problem of strongly auto-correlated residuals
- Better to introduce noise into the dynamics to get a logistic growth diffusion process

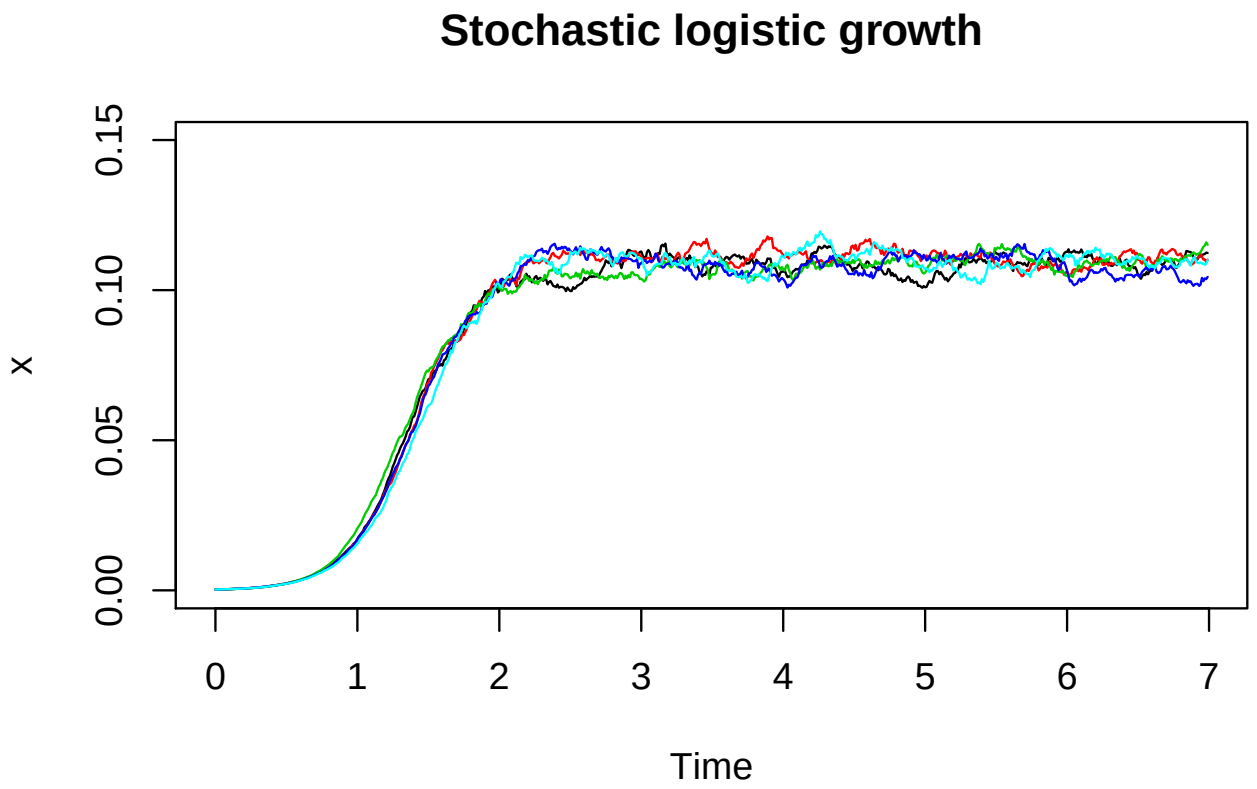
Stochastic logistic growth diffusion

- Well-known stochastic generalisation of the logistic growth equation, expressed as an Itô stochastic differential equation (SDE):

$$dX_t = rX_t(1 - X_t/K)dt + \xi^{-1/2} X_t \, dW_t$$

- The drift is exactly as for the deterministic model
- The diffusion term injects some noise into the dynamics
- The multiplicative noise ensures that this defines a non-negative stochastic process

Sample trajectories from the logistic diffusion



Statistical model

- Model observational measurements $\{Y_{t_1}, Y_{t_2}, \ldots\}$ with

$$Y_{t_i} = X_{t_i} + \varepsilon_{t_i}$$

  where X_{t_i} refers to our realisation of the diffusion process
- Need somewhat sophisticated algorithms to fit these sorts of SDE models to discrete time data
- Standard algorithms would require knowledge of the transition kernel of the diffusion process, but this is not available for the logistic diffusion
- Lots of work on Bayesian inference for intractable diffusions (Golightly & W, '05, '06, '08, '10, '11), but this won't scale to simultaneous fitting of tens of thousands of realisations

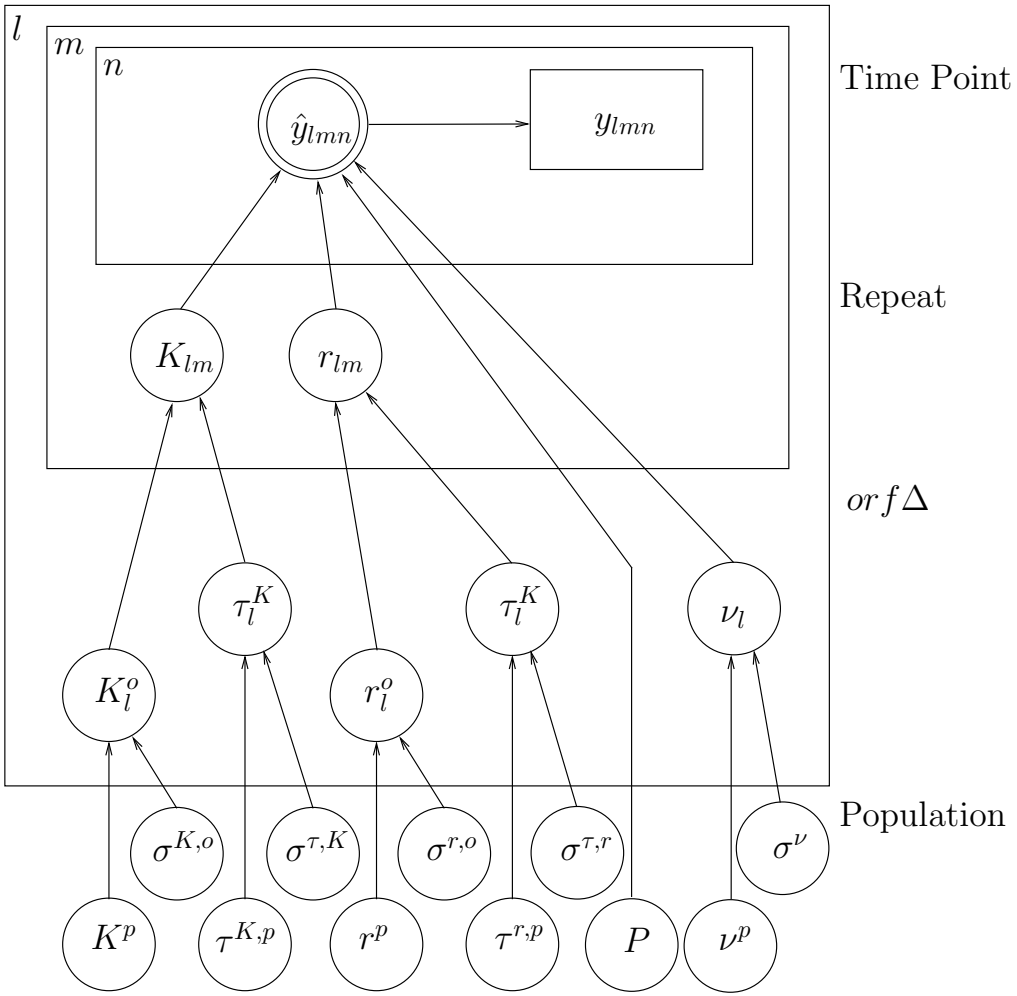
Further simplifications and approximations

- The linear noise approximation (LNA) applied to the log-transformed process is a good model with a tractable transition kernel
- We can implement standard discrete time MCMC methods to estimate model parameters together with the unobserved latent trajectories
- Embedding in a hierarchical model is straightforward
- These methods work fine for hundreds of growth curves, but are still problematic for tens of thousands of growth curves

Integrating out the latent process

- If we are prepared to assume linear Gaussian error on the log scale, we can use Kalman filtering techniques to integrate out the latent process (but this isn't very plausible)
- Alternatively, we could apply a LNA directly to the logistic diffusion (without first transforming), and assume linear Gaussian error on that scale (Heydari et al, 2013)
- This latter approach turns out to be better, despite the fact that the LNA approximation to the true process isn't quite as good
- More important to have a plausible error structure than a super-accurate approximation to the stochastic process

Growth curve model



Colony fitness

- The results of model fitting are estimates (or posterior distributions) of r and K for each yeast colony, and also the corresponding gene level parameters
- Both r and K are indicative of colony fitness — keep separate where possible
- Often useful to have a scalar measure of fitness — many possibilities, including rK, $r \log K$, or MDR$\times$MDP, where MDR is the maximal doubling rate and MDP is the maximal doubling potential
- Statistical summaries can be fed in as data to the next level of analysis (or, ultimately, modelled jointly as a giant hierarchical model)

Epistasis

From Wikipedia:

- *"Epistasis is the interaction between genes. Epistasis takes place when the action of one gene is modified by one or several other genes, which are sometimes called modifier genes. The gene whose phenotype is expressed is said to be epistatic, while the phenotype altered or suppressed is said to be hypostatic."*
- *"Epistasis and genetic interaction refer to the same phenomenon; however, epistasis is widely used in population genetics and refers especially to the statistical properties of the phenomenon."*

Multiplicative model

- Consider two genes with alleles a/A and b/B with a and b representing "wild type" (note that A and B could potentially represent knock-outs of a and b)
- Four genotypes: aa, Ab, aB, AB. Use $[\cdot]$ to denote some quantitative phenotypic measure (eg. "fitness") for each genotype
- Multiplicative model of genetic independence:

$$[AB] \times [ab] = [Ab] \times [aB] \qquad \text{no epistasis}$$
$$[AB] \times [ab] > [Ab] \times [aB] \qquad \text{synergistic epistasis}$$
$$[AB] \times [ab] < [Ab] \times [aB] \qquad \text{antagonistic epistasis}$$

- Perhaps simpler if re-written in terms of relative fitness:

$$\frac{[AB]}{[ab]} = \frac{[Ab]}{[ab]} \times \frac{[aB]}{[ab]}$$

Genetic independence and HTP data

- Suppose that we have scaled our data so that it is consistent with a multiplicative model — what do we expect to see?
- The independence model $[AB] \times [ab] = [Ab] \times [aB]$ translates to

$$[\text{query} : \text{abc}\Delta] \times [\text{wt}] = [\text{query}] \times [\text{abc}\Delta]$$

- In other words

$$[\text{query} : \text{abc}\Delta] = \frac{[\text{query}]}{[\text{wt}]} \times [\text{abc}\Delta]$$

- That is, the double-mutant differs from the single-deletion by a constant multiplicative factor that is independent of the particular single-deletion
- ie. a scatter-plot of double against single will show them all lying along a straight line

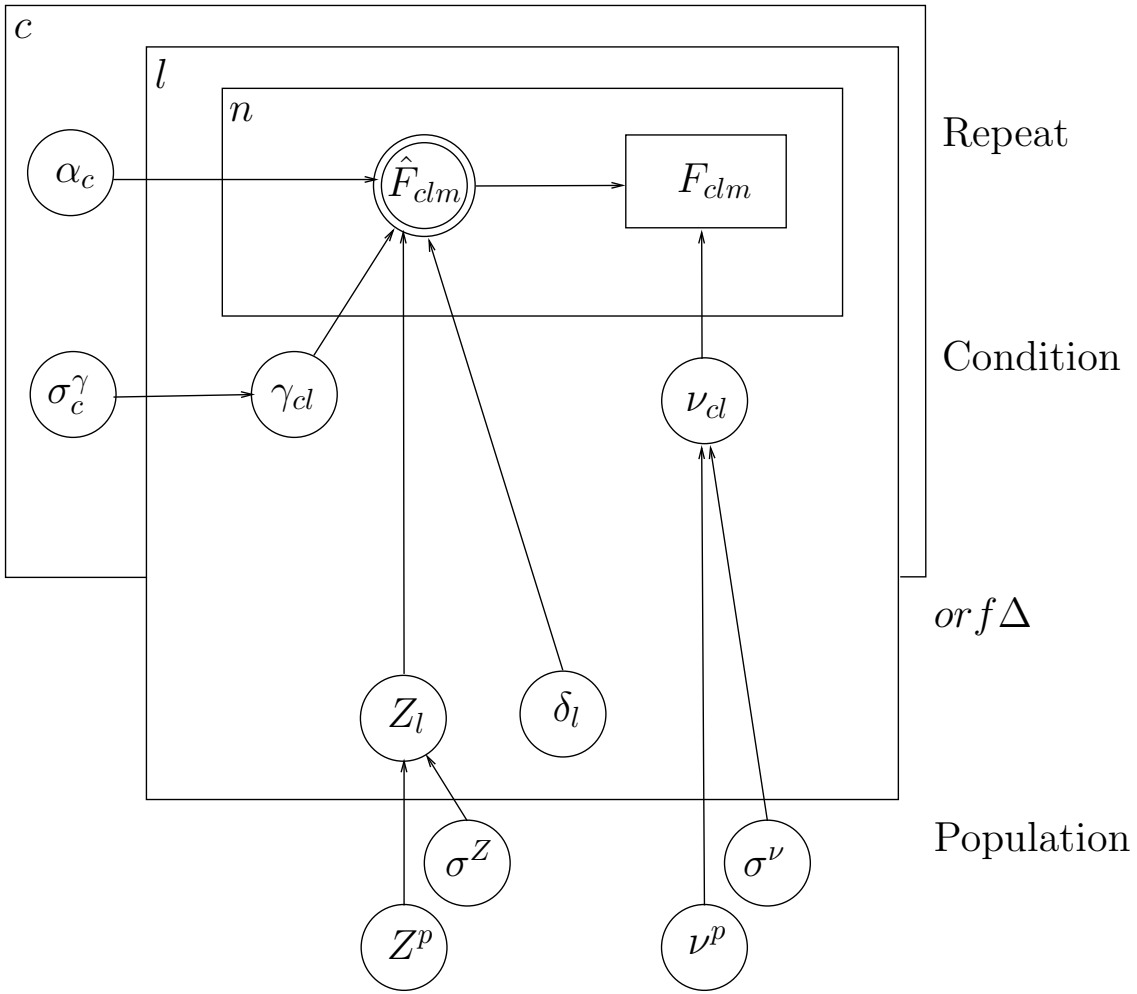
Statistical modelling

- Assume that F_{clm} is the fitness measurement for repeat m of gene deletion l in condition c ($c = 1$ for the single deletion and $c = 2$ for the corresponding double-mutant)
- Model:

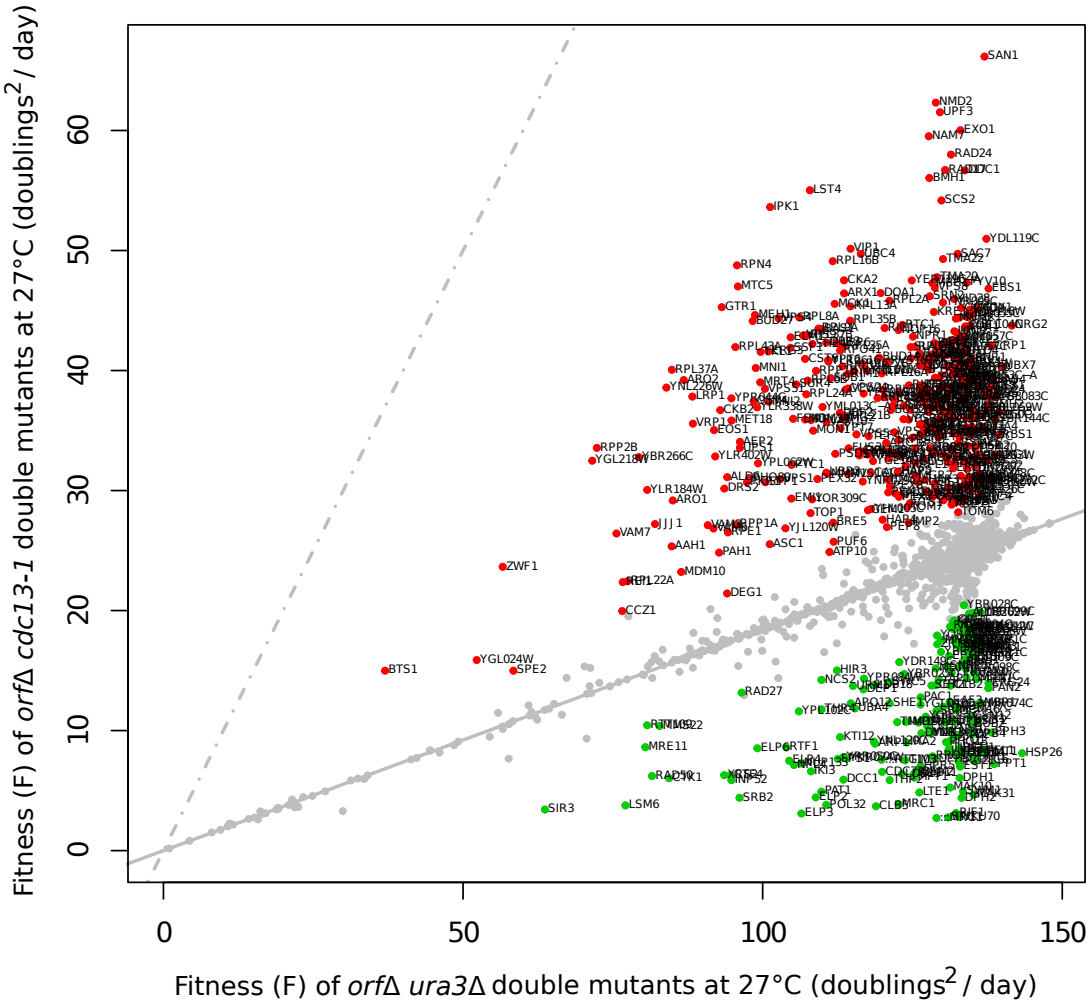
$$F_{clm} \sim N(\hat{F}_{cl}, 1/\nu_{cl})$$
$$\log \hat{F}_{cl} = \alpha_c + Z_l + \delta_l \gamma_{cl}$$
$$\delta_l \sim Bern(p)$$

- δ_l is a variable selection indicator of genetic interaction
- Then usual Bayesian hierarchical stuff...

Genetic interaction model



Genetic interaction results

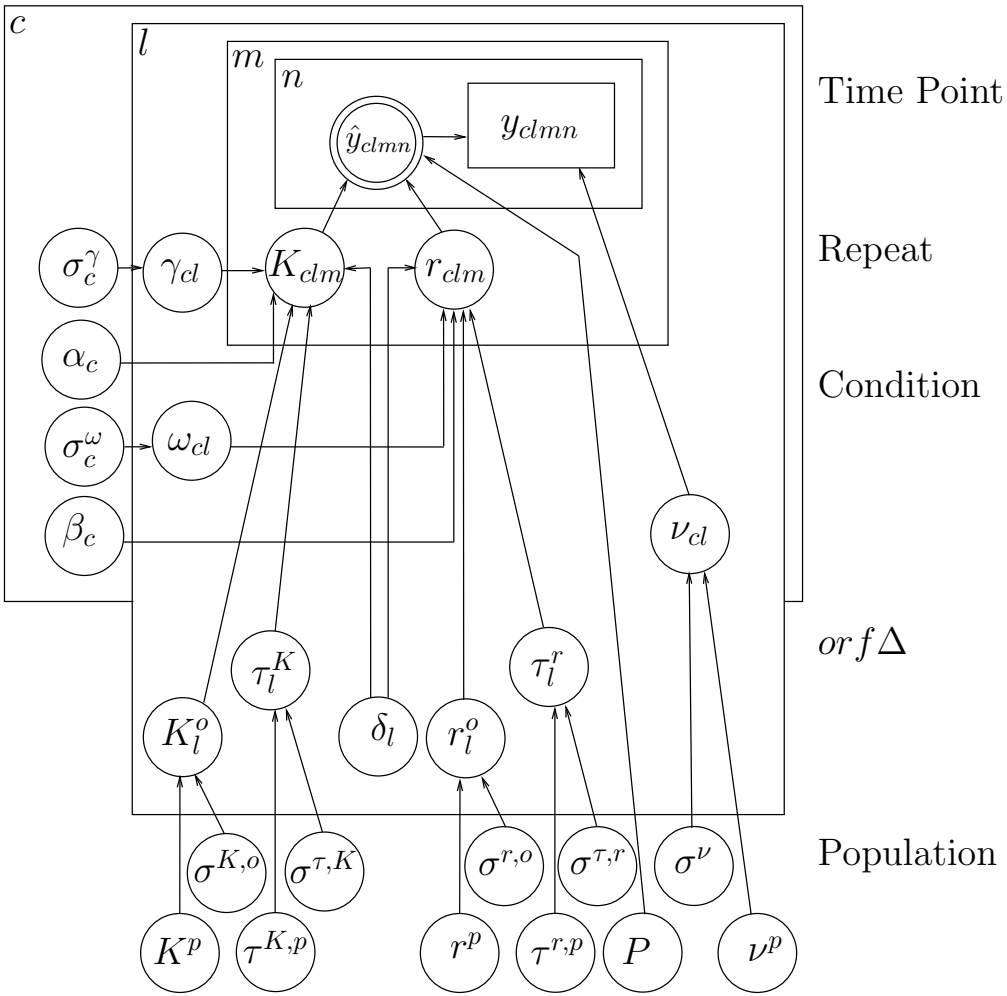


Joint modelling of growth curves and genetic interaction

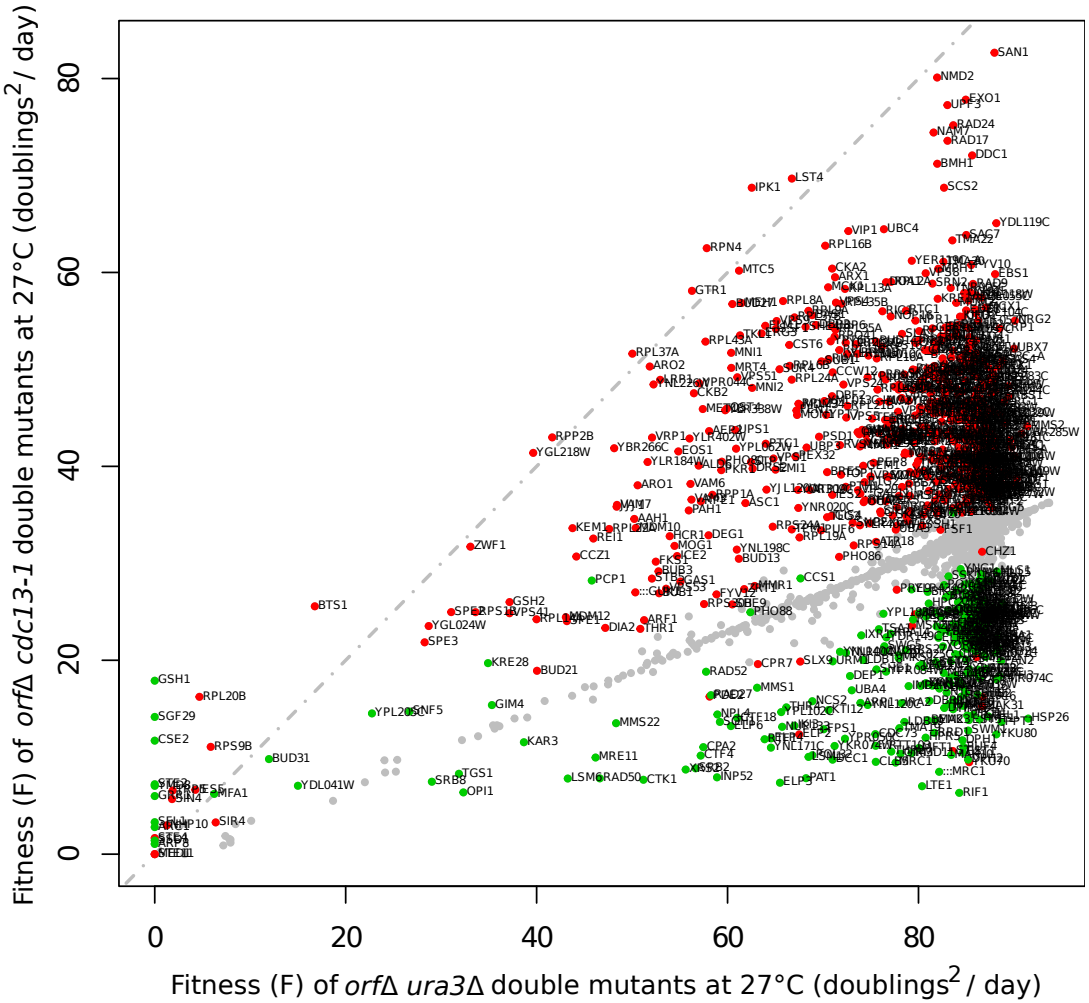
- We can integrate together the hierarchical growth curve model and the genetic interaction model into a combined joint model
- This has usual advantages of properly borrowing strength, proper propagation of uncertainty, etc.
- Also very convenient to avoid requiring a scalar measure of "fitness"
- If y_{clmn} is the colony size at time point n in repeat m of gene l in condition c, then

$$
\begin{aligned}
y_{clmn} &\sim N(\hat{y}_{clmn}, 1/\nu_{cl}) \\
\hat{y}_{clmn} &= X(t_{clmn}; K_{clm}, r_{clm}, P) \\
\log K_{clm} &\sim N(\alpha_c + K_l^o + \delta_l \gamma_{cl}, 1/\tau_{cl}^K) \\
\log r_{clm} &\sim N(\beta_c + r_l^o + \delta_l \omega_{cl}, 1/\tau_{cl}^r)
\end{aligned}
$$

Joint model


c
l
m
n
$\hat{y}_{clmn}$
y_{clmn}
σ_c^γ
γ_{cl}
K_{clm}
r_{clm}
α_c
σ_c^ω
ω_{cl}
β_c
ν_{cl}
τ_l^K
τ_l^r
K_l^o
δ_l
r_l^o
$\sigma^{K,o}$
$\sigma^{\tau,K}$
$\sigma^{r,o}$
$\sigma^{\tau,r}$
σ^ν
K^p
$\tau^{K,p}$
r^p
$\tau^{r,p}$
P
ν^p
Time Point
Repeat
Condition
$or f \Delta$
Population

Joint model results



MiniQFA

Not just a single query, but an "all-by-all" experiment for a (small) subset of the genome

- We have run an experiment with 150 gene knockouts (in both wt and cdc13-1 backgrounds) on a single plate, using each in turn as the query mutation, to map out the full set of pairwise interactions for the subset
- Requires an extension and some modification of the current hierarchical models
- Will allow an in depth study of the prevelance of genetic interaction, and allow consideration of alternative notions of genetic interaction and epistasis, which could distinguish between "direct" and "indirect" interactions

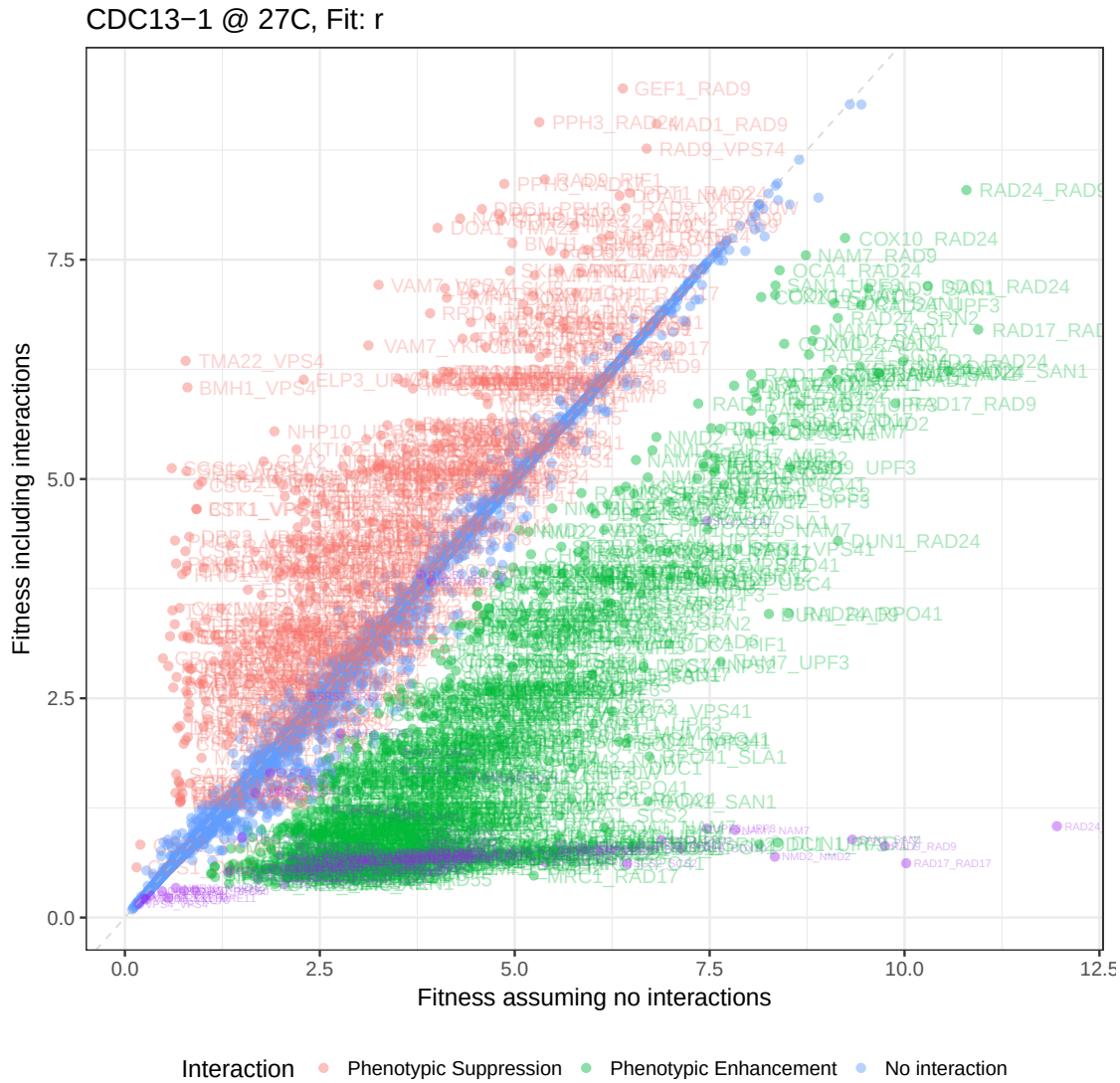
MiniQFA model

- Initial modelling approach based on two stages
- Interaction model:

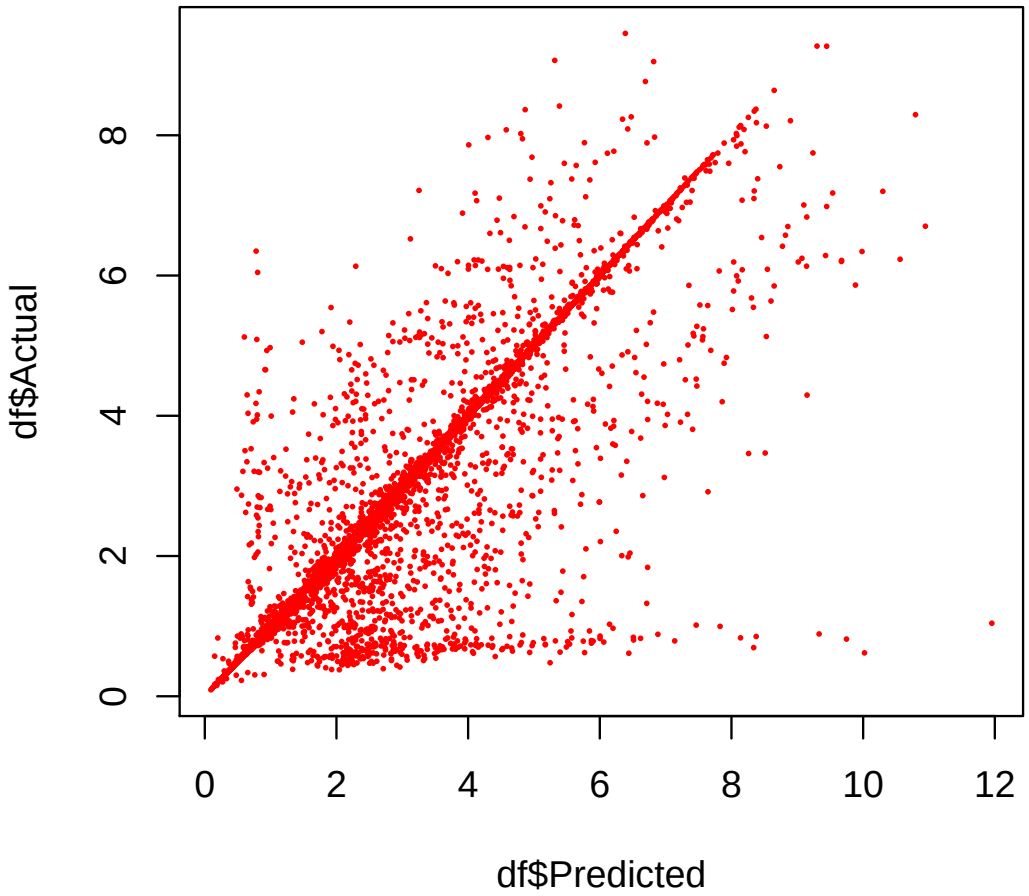
$$F_{ll'm} \sim N(\hat{F}_{ll'm}, 1/\nu)$$
$$\log \hat{F}_{ll'm} = \mu + Z_l + Z_{l'} + \delta_{ll'}\gamma_{ll'}$$
$$\delta_{ll'} \sim \text{Bern}(p)$$

- A version assuming symmetry (eg. $\delta_{ll'} = \delta_{l'l}$, and similar for $\gamma_{ll'}$), and another which doesn't, another with t errors, ...
- Using preliminary data, only around 10% of the $\binom{150}{2} \simeq 11\text{k}$ potential interactions are being selected
- Starting to think about implementation of the "direct" versus "indirect" model...

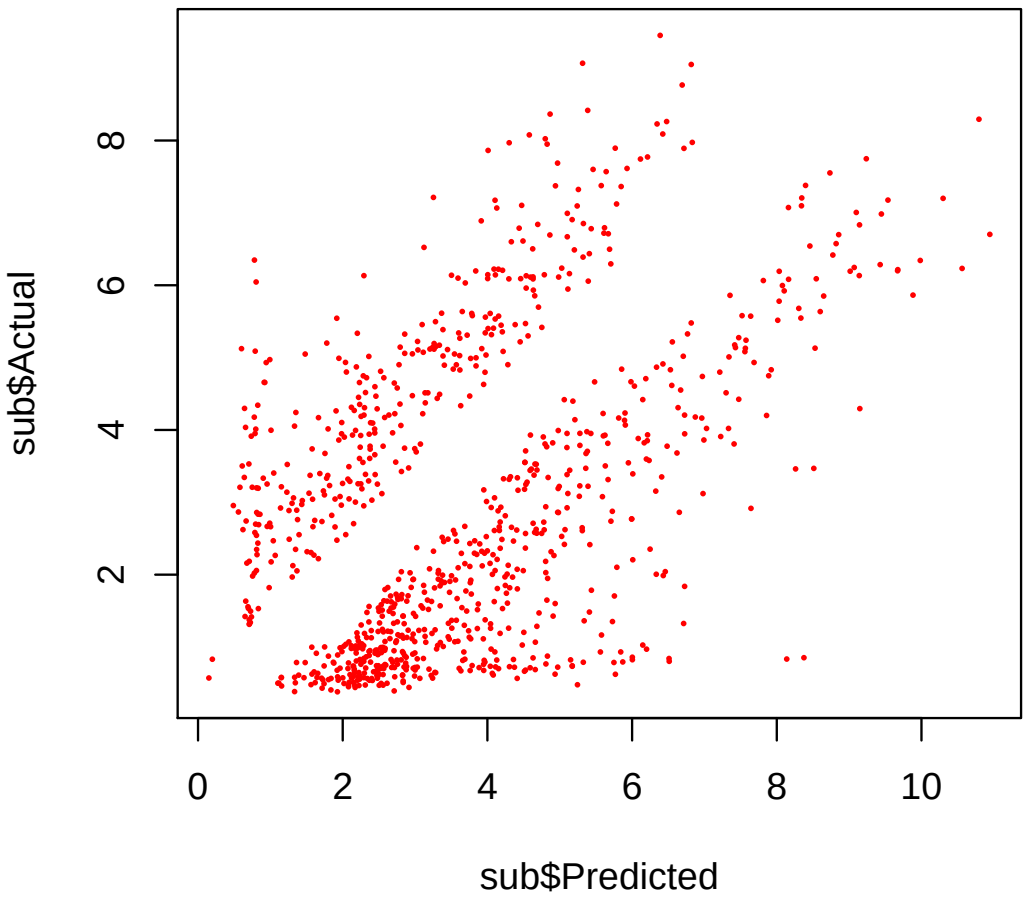
Preliminary MiniQFA results (cdc13-1 at 27C)



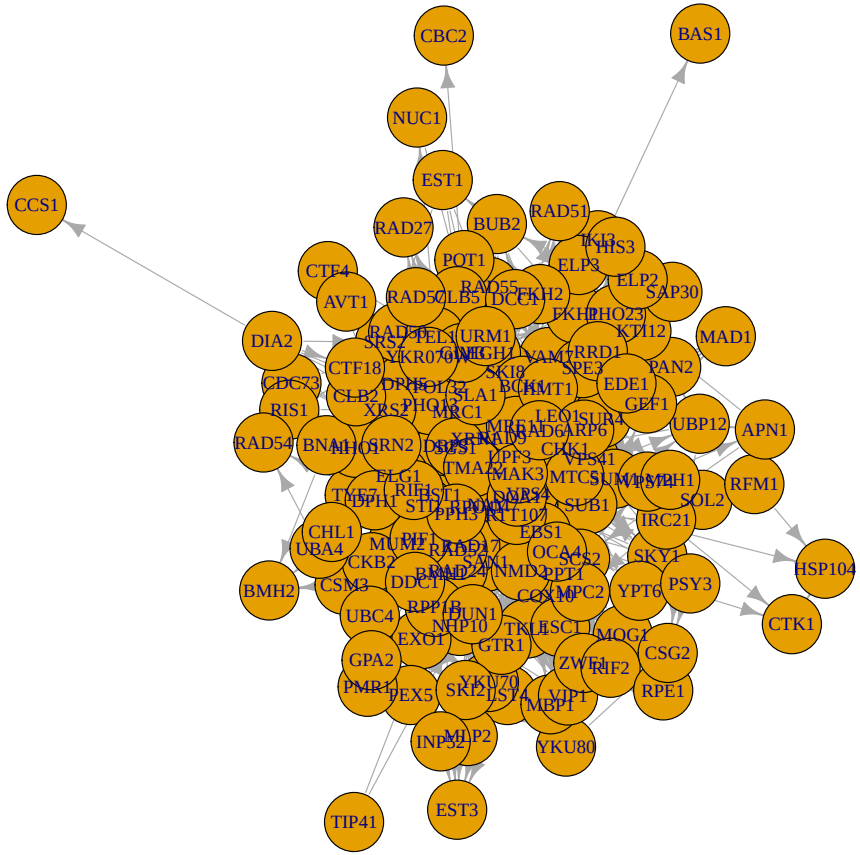
All double deletions



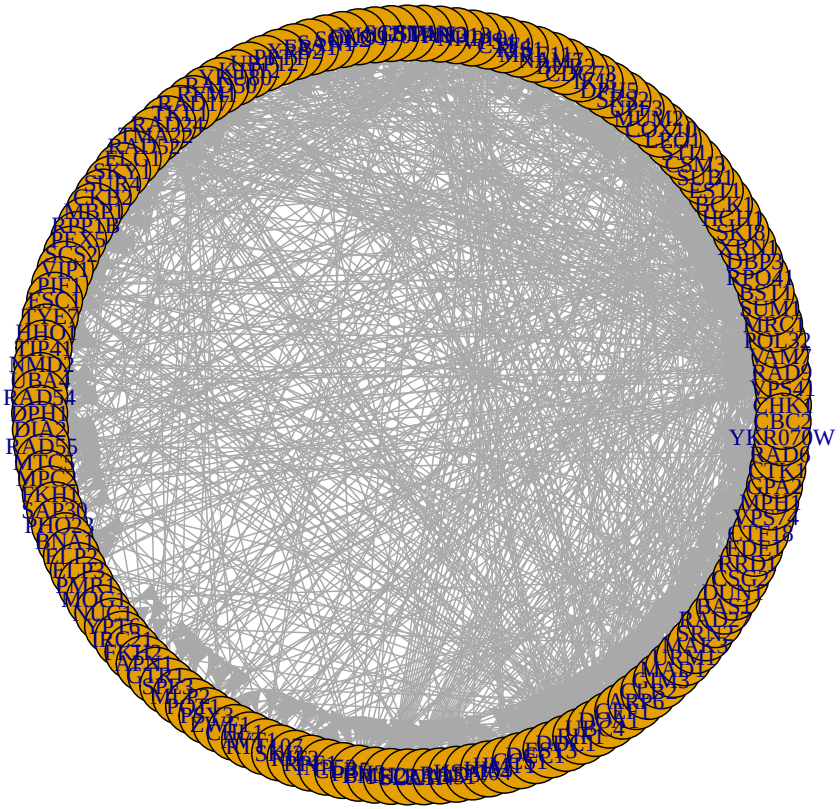
Only the interacting double deletions



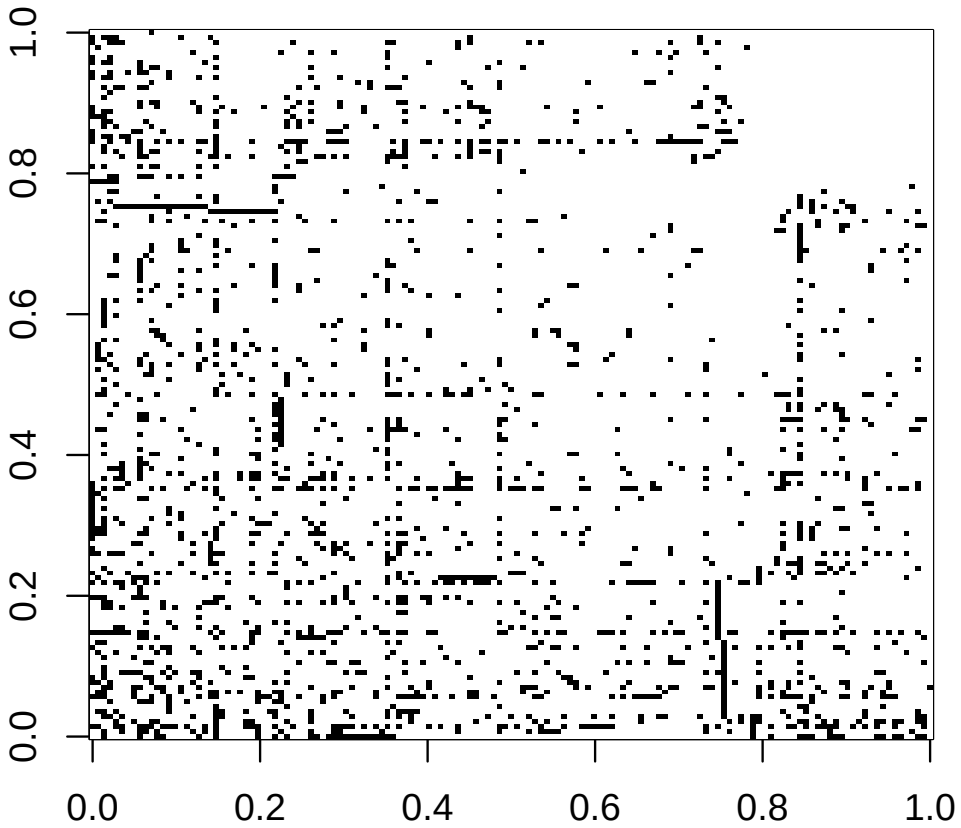
Graph of pairwise interactions



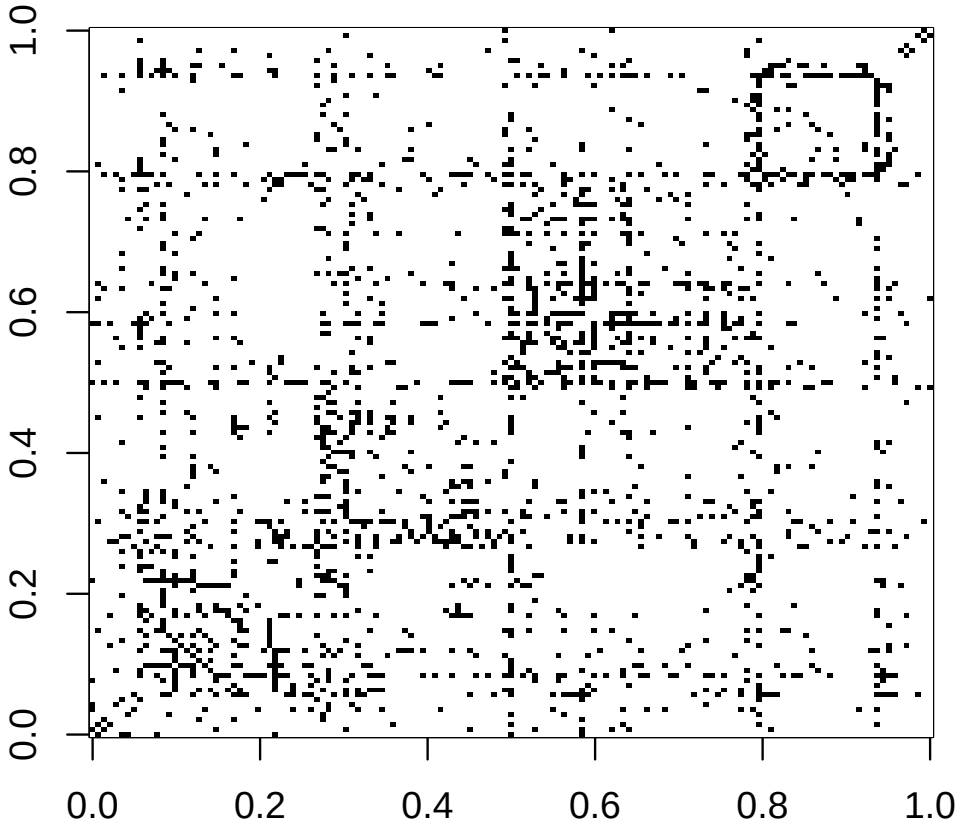
Circle layout



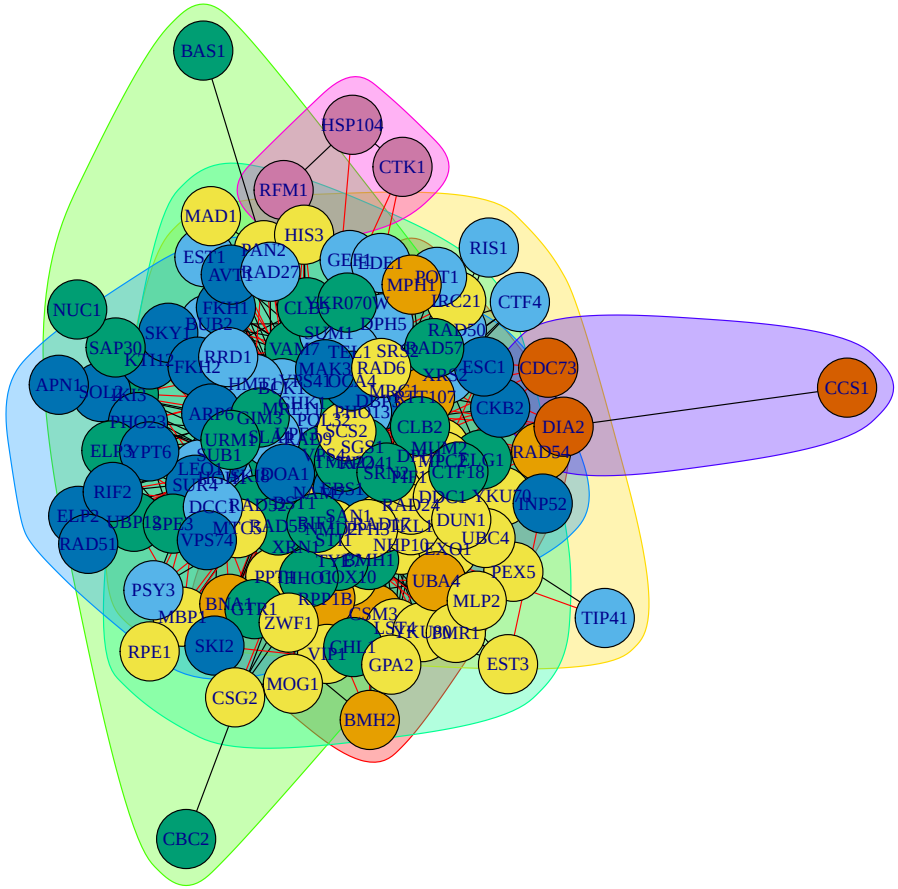
Adjacency matrix



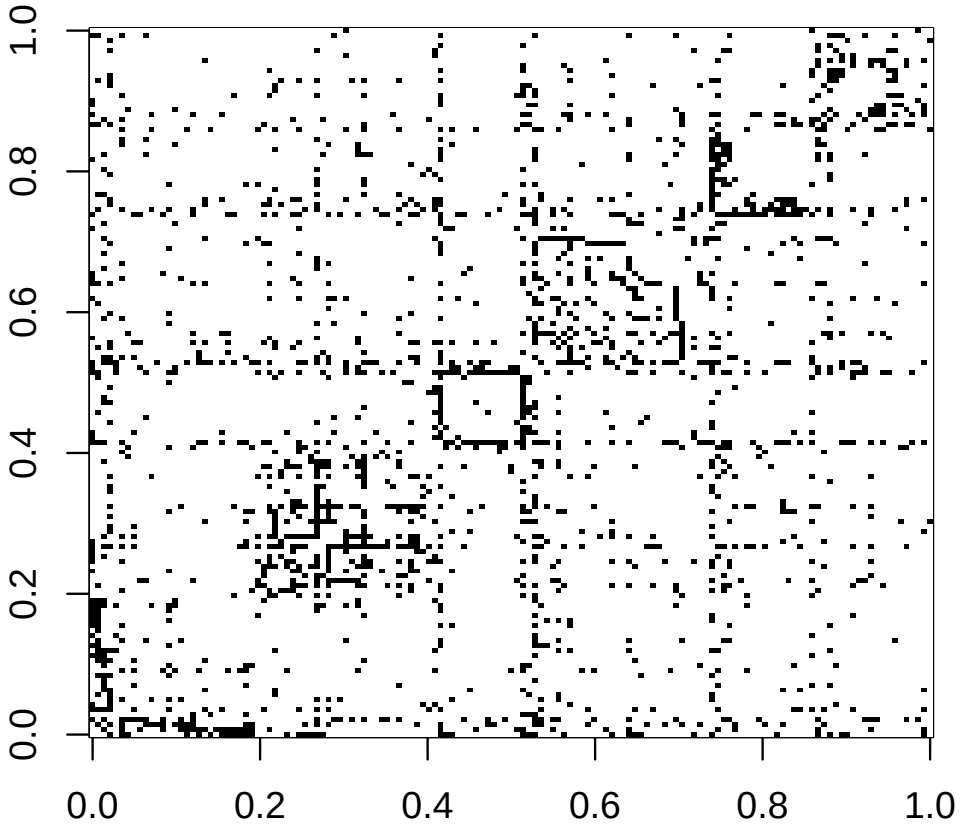
Adjacency matrix ("fast greedy" graph clustering)



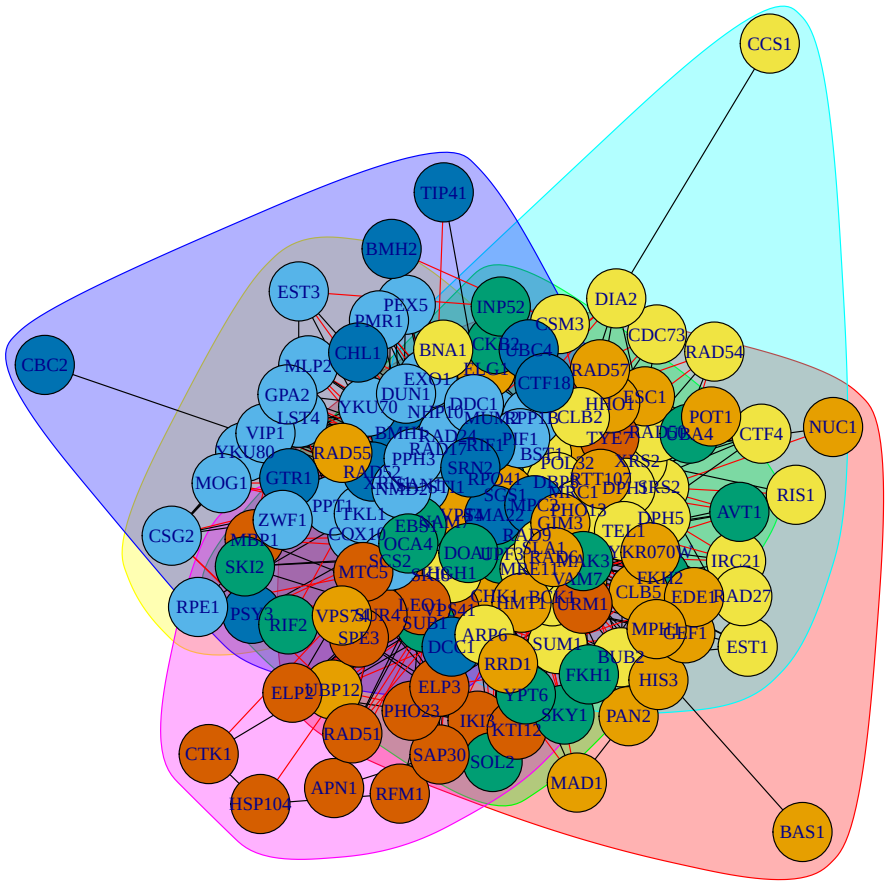
Interaction graph ("fast greedy" graph clustering)



Adjacency matrix ("spin–glass" graph clustering)



Interaction graph ("spin–glass" graph clustering)



"Big data" issues

- Understanding conflict between model and data in big data contexts — does more data demand more complex models?
- Model simplifications and improvements to MCMC (linear Gaussian block updates and proposals, "INLA proposals", 2-block proposals, reparameterisations, GVS, etc.)
- Basic parallelisation strategies (parallel chains, parallellelised single chain)
- Investigation of data parallel strategies (consensus Monte Carlo, etc.)
- Novel representations, interpretations and implementations of Bayesian hierarchical models using ideas from probabilistic programming and strongly typed functional programming languages (Scala, Haskell, Eta, OCaml, ...)

Summary

- Modern bioscience is generating large, complex data sets which require sophisticated modelling in order to answer questions of scientific interest
- Big data forces trade-offs between statistical accuracy and computational tractability
- Stochastic dynamic models are much more flexible than deterministic models, but come at a computational cost — the LNA can sometimes represent an excellent compromise
- Notions of genetic interaction translate directly to statistical models of interaction
- Big hierarchical variable selection models are useful in genomics, but can be computationally challenging

Funding acknowledgements

Major funders of this work:

- BBSRC (RCUK)
- MRC (RCUK)
- Wellcome Trust
- Cancer Research UK

References


Addinall, S. G., Holstein, E., Lawless, C., Yu, M., Chapman, K., Taschuk, M., Young, A., Ciesiolka, A., Lister, A., Wipat, A., Wilkinson, D. J., Lydall, D. A. (2011) Quantitative fitness analysis shows that NMD proteins and many other protein complexes suppress or enhance distinct telomere cap defects. *PLoS Genetics*, **7**:e1001362.


Heydari, J. J., Lawless, C., Lydall, D. A., Wilkinson, D. J. (2016) Bayesian hierarchical modelling for inferring genetic interactions in yeast, *Journal of the Royal Statistical Society, Series C*, **65**(3):367–393.


Heydari, J. J., Lawless, C., Lydall, D. A., Wilkinson, D. J. (2014) Fast Bayesian parameter estimation for stochastic logistic growth models, *BioSystems*, **122**:55-72.


Lawless, C., Wilkinson, D. J., Addinall, S. G., Lydall, D. A. (2010) Colonyzer: automated quantification of characteristics of microorganism colonies growing on solid agar, *BMC Bioinformatics*, **11**:287.


Wilkinson, D. J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems, *Nature Reviews Genetics*. **10**(2):122-133.


Wilkinson, D. J. (2011) *Stochastic Modelling for Systems Biology, second edition*. Chapman & Hall/CRC Press.