DRMCRL

Group Meeting

Stephen Pederson

Dame Roma Mitchell Cancer Research Laboratories,
The University of Adelaide

25th August, 2020

Introduction

Bioinformatics Hub

- Co-ordinator Bioinformatics Hub, 2014 - 2020
- Currently an ECR: PhD (2008-2018)
- Main areas of expertise:
  - `R`, `bash`, LaTeX
  - Transcriptomics and Statistics

Why such a long PhD?

- Detect alternate transcript usage using whole-transcript microarrays
- Built a Bayesian model which ran a MCMC process (written in C)
- Implemented as an R package

Why such a long PhD?

- Pre-RStudio, Pre-`tidyverse`, Pre-`RMarkdown`, Pre-`Rcpp`
- Model significantly "under-performed" on real data
- Broke the software with no version control strategy
  - MCMC process ran for a week on 4-cores
  - Failure was a random failure to complete (probably C memory issues)
  - Took 6 months to debug
- *No bioinformatics community able to help*

Why such a long PhD?

- Tutored for School of Mathematical Sciences
- Picked up large amounts of work as a musician
- Landlord died $\Longrightarrow$ no place to live
- Bioinfosummer 2013 (Terry Speed)

Formation of the Bioinformatics Hub

An application was made to the IDRF

1. Prof David Adelson: Head, School of Biological Sciences; Chair of Bioinformatics
2. Prof Gary Glonek: Head, School of Mathematical Sciences
3. Prof Mike Wilkinson: Head, School of Agriculture, Food & Wine
4. Prof Julie Owen: Head, School of Paediatrics & Reproductive Health

Others named with a key interest: *Prof Alan Cooper, Prof Wayne Tilley, Prof Sarah Robertson etc*

Formation of the Bioinformatics Hub

- Run training workshops, seminars, journal clubs etc
- Manage a hot-desking facility
- Provide consultations and advice
- Develop formal courses and subjects
- Apply for external research funding
- Build a *critical mass*

Interestingly, this is not a core-facility

Formation of the Bioinformatics Hub

- Run training workshops, seminars, journal clubs etc
- Manage a hot-desking facility
- Provide consultations and advice
- Develop formal courses and subjects
- Apply for external research funding
- Build a *critical mass*

Interestingly, this is not a core-facility

Formation of the Bioinformatics Hub

- Directors:
  -  David Adelson (*Biol. Sci.*)
  -  Gary Glonek (*Mathematical Sciences*)
- Funded by DVCR to recruit Level B6-C3 Academic for 2014
- Administered within School of Biological Sciences (BS)

Formation of the Bioinformatics Hub

- I commenced in March 2014: 10 month contract at Level A4

Why?

- Not attractive for external recruitment
- Unemployed bioinformaticians are unicorns
- I was the "best they could get"

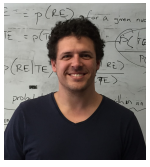
Formation of the Bioinformatics Hub

- I commenced in March 2014: 10 month contract at Level A4

Why?

- Not attractive for external recruitment
- Unemployed bioinformaticians are unicorns
- I was the "best they could get"

2015

- My contract was renewed for 2015
-  Jimmy Breen appointed to RRI $\Longrightarrow$ co-located
-  Hien To appointed as second staff member (2015-2018)
-  Rick Tearle appointed to Davies Research Centre (Roseworthy)

In Later Years

-  Nathan Watson-Haigh (2018-2020)
-  Mark Armstrong (2019-2020)
- Additional 12 staff members (2015-2020)

Our Track Record

Over our existence:

- $8.5m in grant funding
- >1400 distinct individuals through workshops
- Individual support for >360 postgraduate students (MPhil/PhD)
- Co-authored 84 publications + 2 software packages
- Established a Bioinformatics Undergraduate Major
- Very strong and supportive bioinformatics community

Our Track Record

Journal Club on Steve's Birthday, 2019

Publication Highlights

nature > letters > article

MENU

nature

Published: 08 March 2017

Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus

Laura S. Weyrich, Sebastian Duchene, Julien Soubrier, Luis Arriola, Bastien Llamas, James Breen, Alan G. Morris, Kurt W. Alt, David Caramelli, Veit Dresely, Milly Farrell, Andrew G. Farrer, Michael Francken, Neville Gully, Wolfgang Haak, Karen Hardy, Katerina Harvati, Petra Held, Edward C. Holmes, John Kaidonis, Carles Lalueza-Fox, Marco de la Rasilla, Antonio Rosas, Patrick Semal, Arkadiusz Soltysiak, Grant Townsend, Donatella Usai, Joachim Wahl, Daniel H. Huson, Keith Dobney & Alan Cooper -Show fewer authors

Nature **544**, 357–361(2017) | Cite this article

6767 Accesses | **157** Citations | **2291** Altmetric | Metrics

Publication Highlights

PNAS
Proceedings of the National Academy of Sciences of the United States of America

RESEARCH ARTICLE

A comprehensive genomic history of extinct and living elephants

Eleftheria Palkopoulou, Mark Lipson, Swapan Mallick, Svend Nielsen, Nadin Rohland, Sina Baleka, Emil Karpinski, Atma M. Ivancevic, Thu-Hien To, R. Daniel Kortschak, Joy M. Raison, Zhipeng Qu, Tat-Jun Chin, Kurt W. Alt, Stefan Claesson, Love Dalén, Ross D. E. MacPhee, Harald Meller, Alfred L. Roca, Oliver A. Ryder, David Heiman, Sarah Young, Matthew Breen, Christina Williams, Bronwen L. Aken, Magali Ruffier, Elinor Karlsson, Jeremy Johnson, Federica Di Palma, Jessica Alfoldi, David L. Adelson, Thomas Mailund, Kasper Munch, Kerstin Lindblad-Toh, Michael Hofreiter, Hendrik Poinar, and David Reich

PNAS March 13, 2018 115 (11) E2566-E2574; first published February 26, 2018
https://doi.org/10.1073/pnas.1720554115

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved January 24, 2018 (received for review November 24, 2017)

Other Highlights

- Return on Investment over 6.5 years: \$4.4/dollar
- Improved PhD Completion rates (ECMS, Sciences, HMS)
  - 66% with no Hub engagement vs 75% with ($p = 0.0091$)
- Establishment of UoA as a national player in bioinformatics
- Student engagements strongly biased towards women:
  - ECMS: 41.7% Vs 22.0% (12 students)
  - HMS: 61.9% Vs 59.5% (89 students)
  - Sciences: 59.2% Vs 48.3% (260 students)

Weaknesses

- Strategic Vision authored and submitted to Exec Dean of Sciences in 2018
- Rejected in it's entirety with no further discussion
- Included reimbursement for teaching courses
- In 2019 he approved a separate bioinformatics position inside School of BS (still unappointed)

Weaknesses

- Income
  - Money from grants awarded
  - Money for course delivery & student supervision
- No expectation of reimbursement from DVCR
- Short-term contracts made leading grant applications challenging
  - I'm too junior
- Lack of single "high-profile" project
- How to limit access whilst providing open-access?

Challenges

In 6.3 years

- Biological Sciences: 4 Heads of School
- Agriculture, Food & Wine: 4 Heads of School

SAGC / Our Demise

- Current DVCR didn't believe he should fund us
  - Gave additional money to faculties for these projects
  - HMS unambiguously supportive whilst Sciences "receive no benefit at all from the Bioinformatics Hub"
  - Sciences subsequent internal review found we were beneficial & vital infrastructure requiring funding
- 3 positions provided by UofA as "in-kind" for SAGC funding
  - 1xDVCR, 1xHMS, 1xSciences
  - DVCR re-negotiated this down to the HMS position only

My role

- I was ready to move on this year
- There was a position explicitly for me at SAGC
- Being spread across literally everything is exhausting
- Being funded on the whims of DVCR, Exec Deans etc is not ideal
- Looking to focus more deeply both *biologically* and *bioinformatically*
- Take full advantage of my "ECR window"

Enrichment Analysis

RNA Seq Data

- RNA Seq data can suffer from multiple biases
  - Longer genes will produce more reads $\Longrightarrow$ higher expression
  - Higher expression $\Longrightarrow$ more chance of being detected as DE
- Can also suffer from GC bias (PCR)
- Conventionally assumed these factors are constant across samples

Enrichment Testing

- Two primary types of enrichment testing
  1. Enrichment within one set of genes (DE) compared to another set of genes (not DE)
  2. Ranked-list based approaches
- Both can be susceptible to GC, length or other biases

Fisher's Exact Test

Gene-set	Not DE	DE
Non-coding	623	13
Protein Coding	9670	378
(Totals)	10293	391

- Is the proportion of genes in our gene set higher in DE vs Not DE?
- We use Fisher's Exact Test (Hypergeometric Distribution)

Fisher's Exact Test

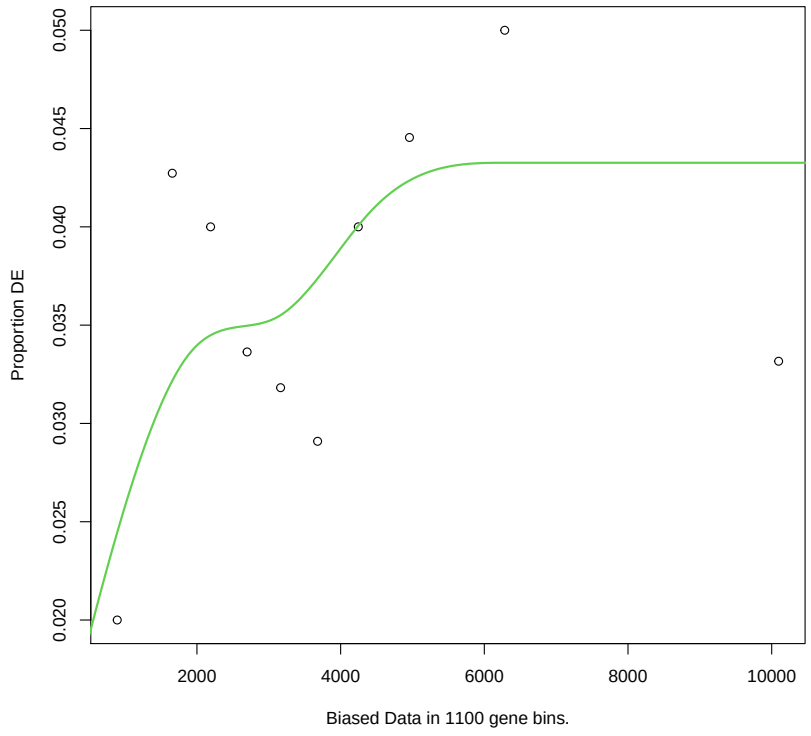
Gene-set	Not DE	DE
Non-coding	623	13
Protein Coding	9670	378
(Totals)	10293	391

- We would expect $\frac{9670}{10293} \times 391 = 0.940 \times 391 = 367$
- Significantly greater than this would be considered enrichment
- Fisher's Exact Test gives $p = 0.02$

Fisher's Exact Test

- Fisher's Exact Test assumes equal probability of sampling each gene
- If we have bias in our sampling $\implies$ spurious results
- We can imagine the probability of sampling red or blue balls from a bag
- If balls are different sizes, we would have sampling bias
- *Does gene length impact the probability of a gene being DE?*

Sampling Bias



Shorter genes appear less likely to be DE

Sampling Bias

- Wallenius' Non-Central Hypergeometric Distribution allows for sampling with bias
- Implemented in the package goseq
- Given that shorter genes are less likely to be DE for this dataset we should use this here
- For the same 2×2 table we get $p = 0.14$

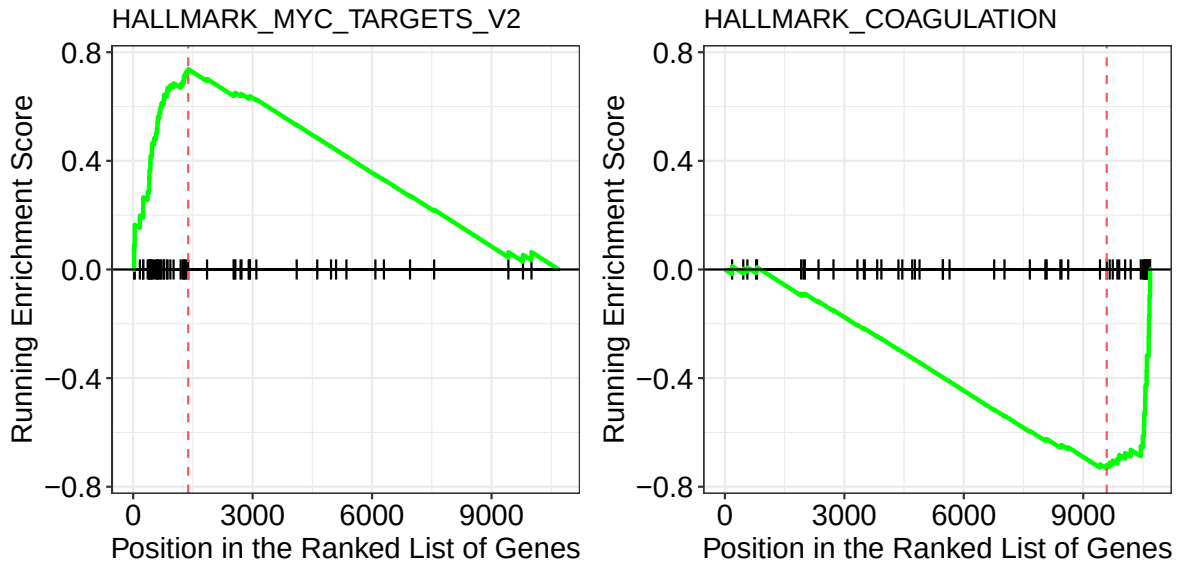
Sampling Bias

Does it really matter?

- If dancing around $p \leq 0.05$, then *yes*
- For enrichment testing, the best results are indisputible ($p \ll 0.05$) and for these, *not so much*

Ranked List Approaches

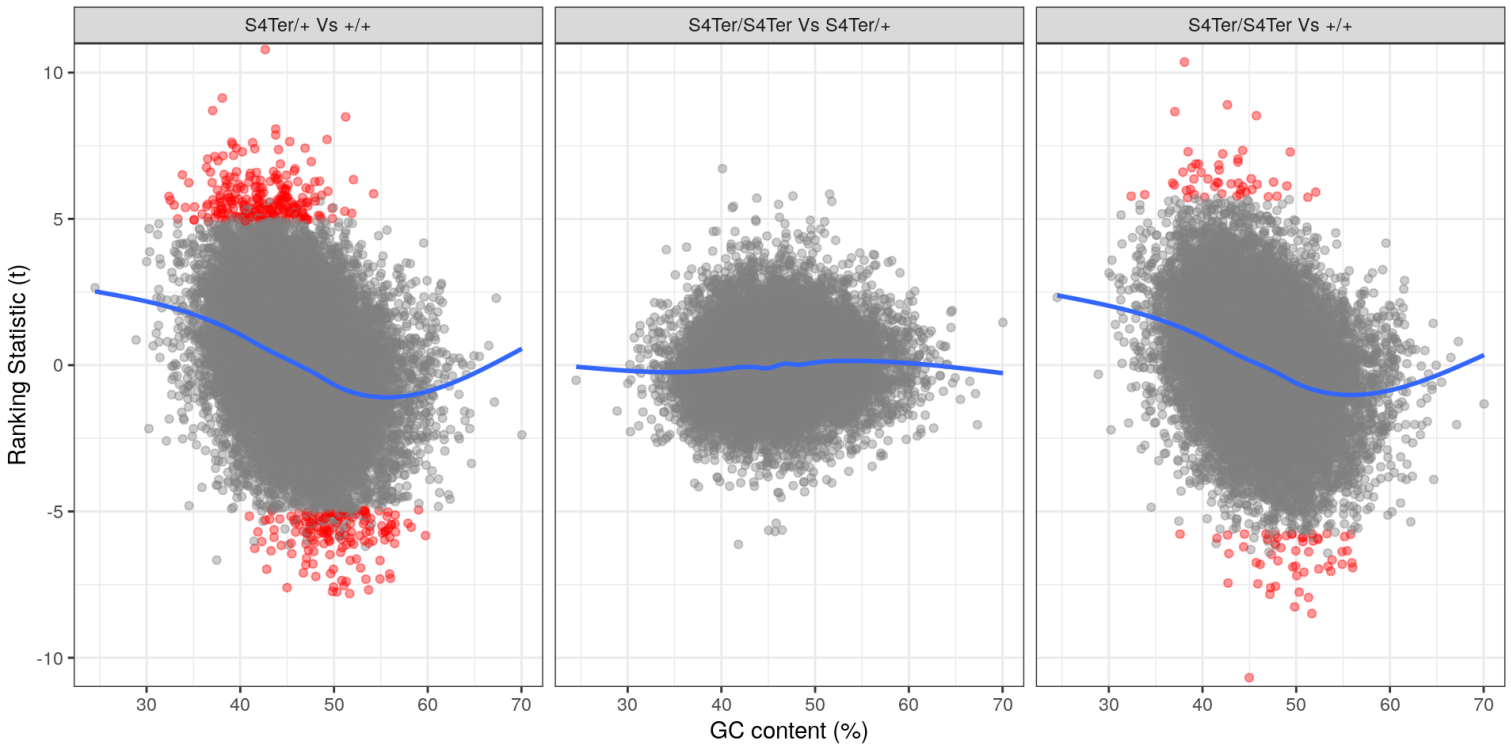
- Gene lists are ranked based on a statistic (t, p, logFC etc)
- Walking down the list detects any unexpected clustering of genes from a gene set at either end



GC and Length bias

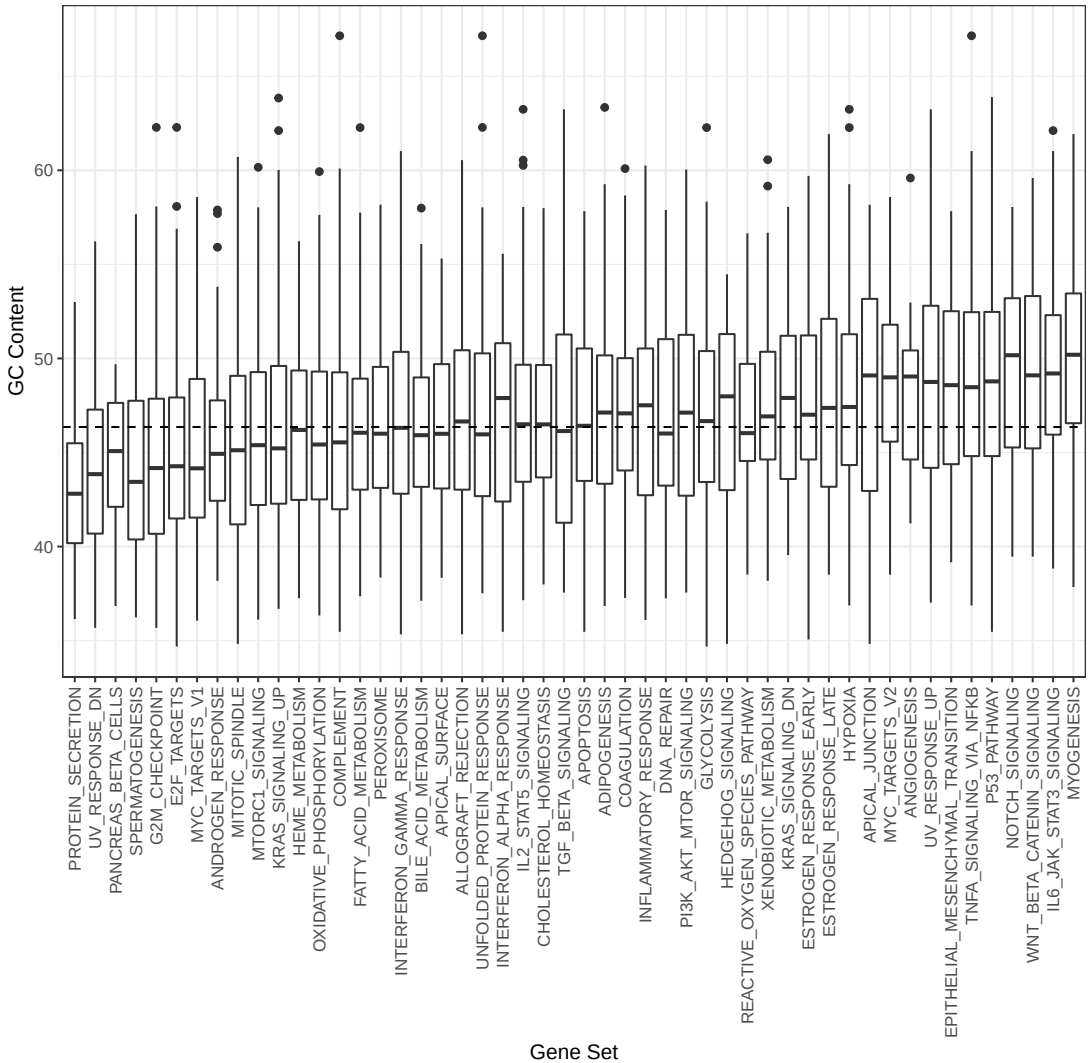
- Generally, GC & length bias within a dataset was assumed to be trivial
- This has been shown to be violated more often than thought (*Mandelboum et al., "Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias"*)
- Can have non-trivial impacts when using ranked lists

GC and Length bias



Both Mutant vs WT comparisons have non-trivial GC issues

GC and Length bias



Managing Bias

- Conditional Quantile Normalisation (CQN) can mitigate this (unless extreme)
- Incorporates a gene and sample specific offset for length and GC bias
  - Rules out `voom` $\Longrightarrow$ Negative Binomial models only
- TMM normalisation is fine for well-behaved data

Incucyte Analysis

Incucyte Data

- Data from the Incucyte is challenging to *analyse* and to *visualise*
- Richard & Jean asked for help in May
- Have a manual import R script running
- Have some viable code for visualisation
- Have some opaque code for fitting statistical models

Incucyte Data

- Plan is to build an R package with:
  - Easy data import
  - Easy data visualisation
  - Viable analytic approaches
- Publish a workflow for visualisation and statistical analysis

Data Parsing

- Code for parsing data works well
- Only parses count/intensity data
- Uses fancy text searching to extract groups and treatments
- Wrote a parser for the `*PlateMap` files over the weekend
- Plan to define an `R` object with metadata + intensities

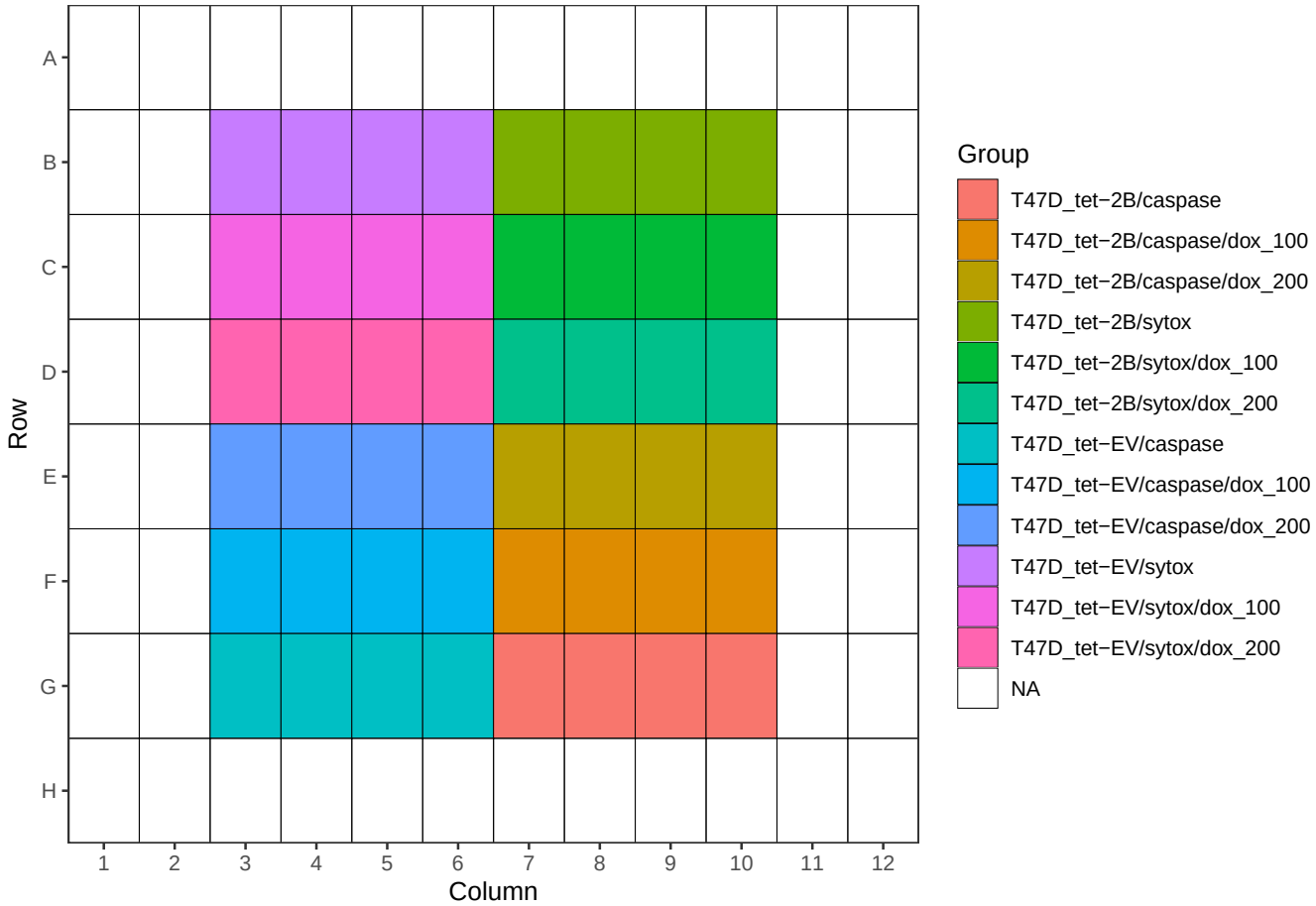
Data Visualisation

- Current visualisation code works, but looks intimidating
- Plan is to write simple plotting functions on my object class

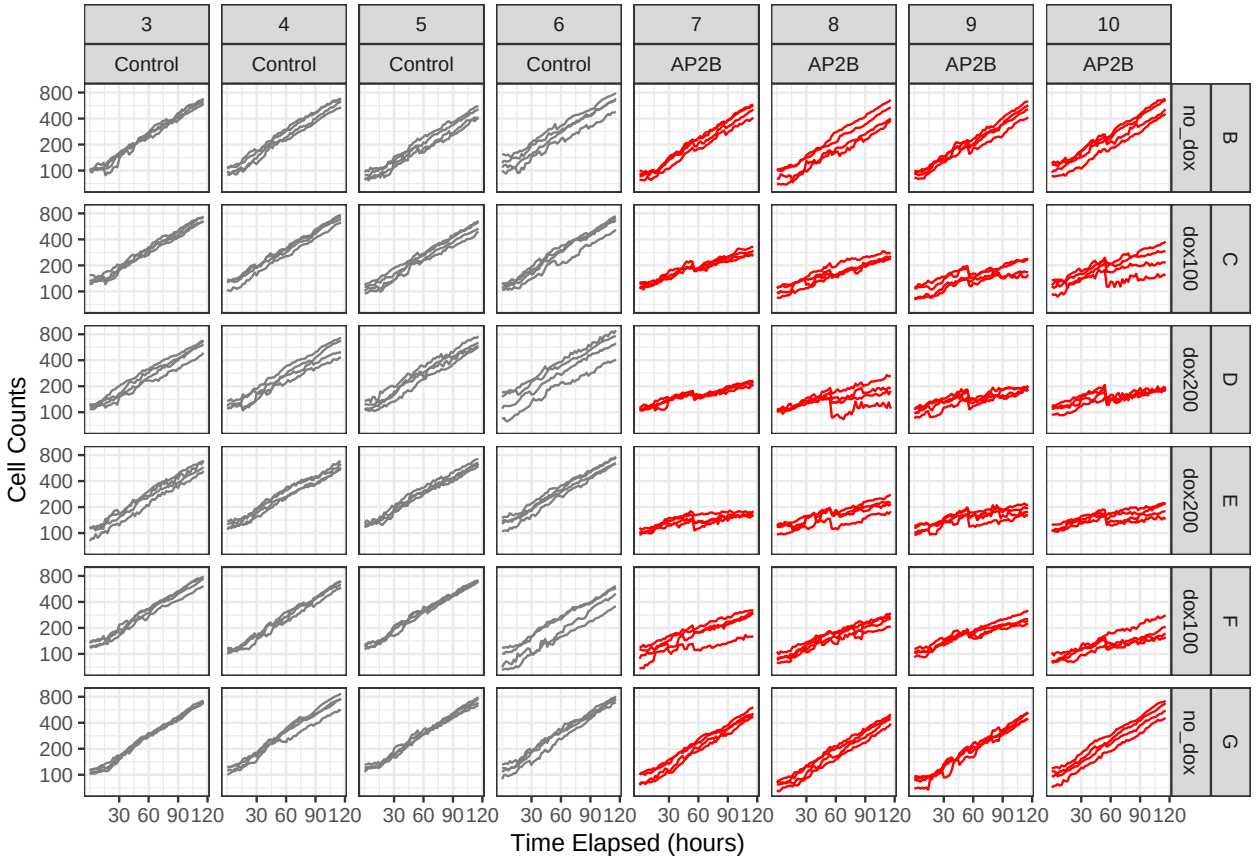
```
plateMap <- parsePlateMap(f = "Vessel 897 T47D 2B
  Kate sytox caspase.PlateMap")
plotPlateMap(plateMap) +
    scale_fill_discrete(na.value = NA) +
    theme_bw() +
    theme(panel.grid = element_blank())

```

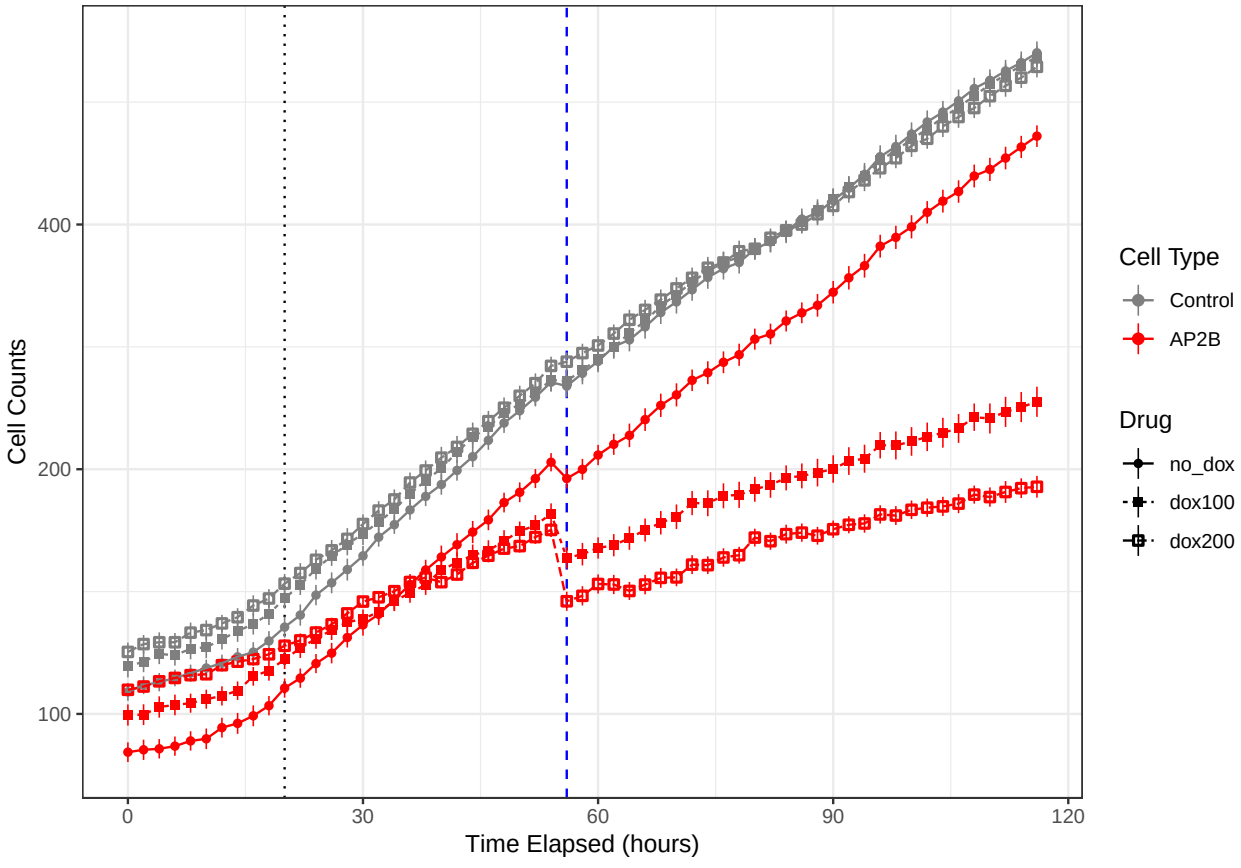
Data Visualisation



Data Visualisation



Data Visualisation



Growth Data: Analytic Challenges

- Cell counts are discrete $\Longrightarrow$ Poisson distributed data
  - This is a type of Generalised Linear Model (GLM)
  - Poisson Distributions measure counts per unit of measurement
  - The rate parameter (λ) models the number of counts/unit
  - Here we have cells/image
- Poisson models for the Incucyte will fit the
  1. starting number of cells (i.e. the intercept) and,
  2. change in the rate of counts as a function of time (i.e. the slope)
- Can be fit for each treatment group, or individual treatments

Growth Data: Analytic Challenges

- Data is also captured within each image, within each well
  - This introduces correlations within each image/well
  - Leads to underestimate of standard errors
  - Leads to overestimate of significance
- A nested model should be applied
  - Commonly known as *mixed-effects models*
  - Allows for random variability in *intercepts* between images
  - Potentially allows the same for slopes
- End model is a Generalised Mixed-Effects Model

GLMMs

- GLMMs are notoriously challenging
- To estimate model effects for a linear (Gaussian) model $\Longrightarrow$ simple matrix algebra ($< 1s$)
- To estimate effects for a GLMM:
  - Fit model iteratively using EM-algorithm
  - Can take minutes
  - Requires convergence to be considered valid
  - Nesting of model terms within images can impede convergence

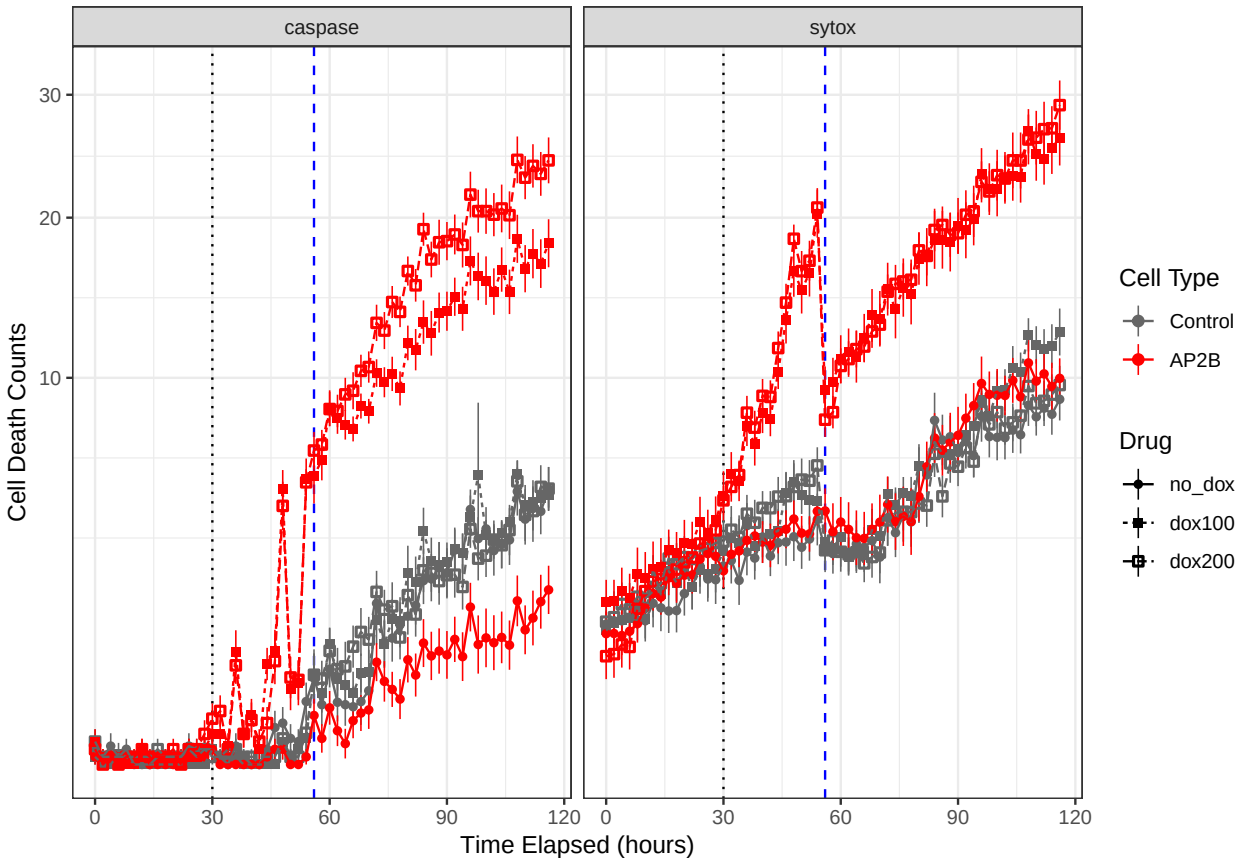
Cell Death: Analytic Challenges

- Cell death is a function of cell growth
- e.g. 10 dead cells could be 10/10, or 10/1000
- Needs to analysed incorporating cell growth
- Cell growth counts are implicitly cumulative
- Cell death counts are transient

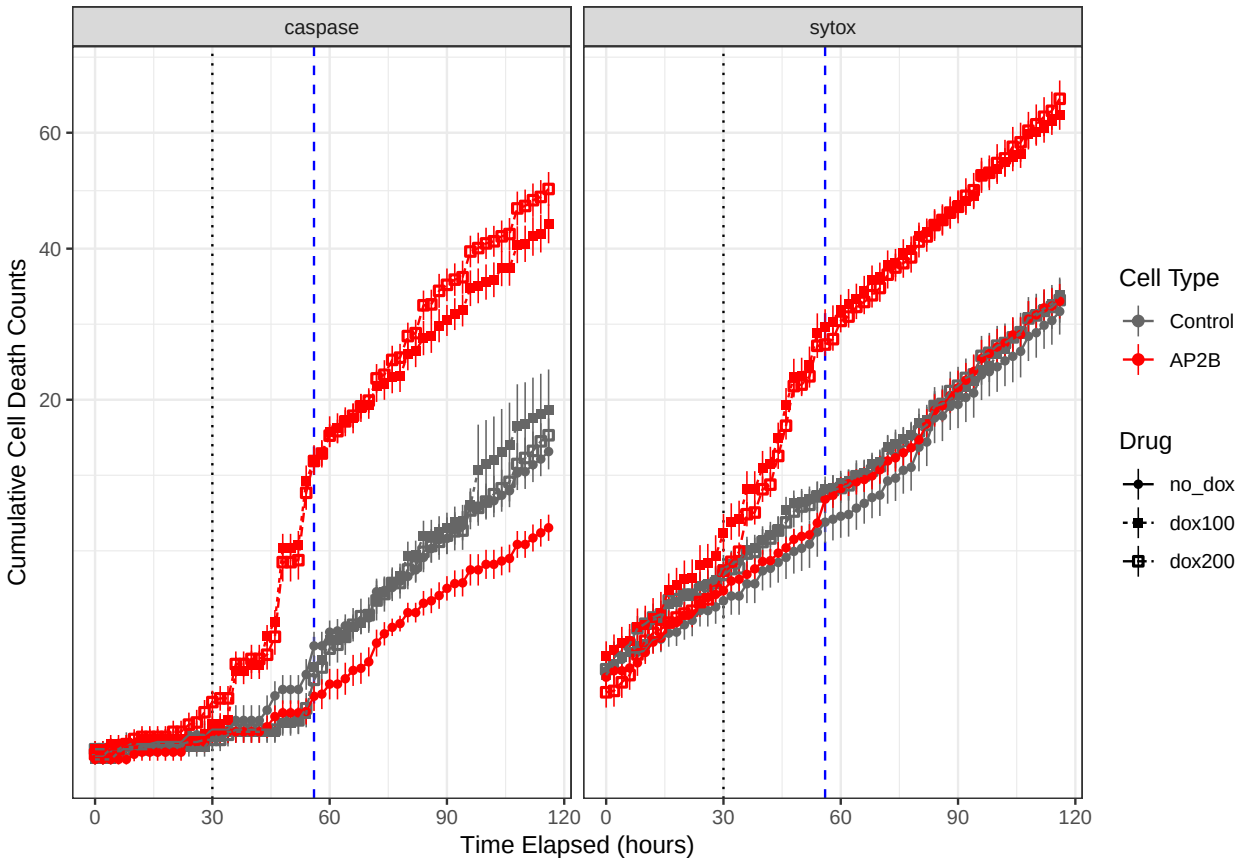
Cell Death: Analytic Challenges

- Do we analyse two cumulative variables?
- Do we analyse two transient variables?
- Do we just take a time-point as a snapshot and ignore this?

Cell Death: Analytic Challenges



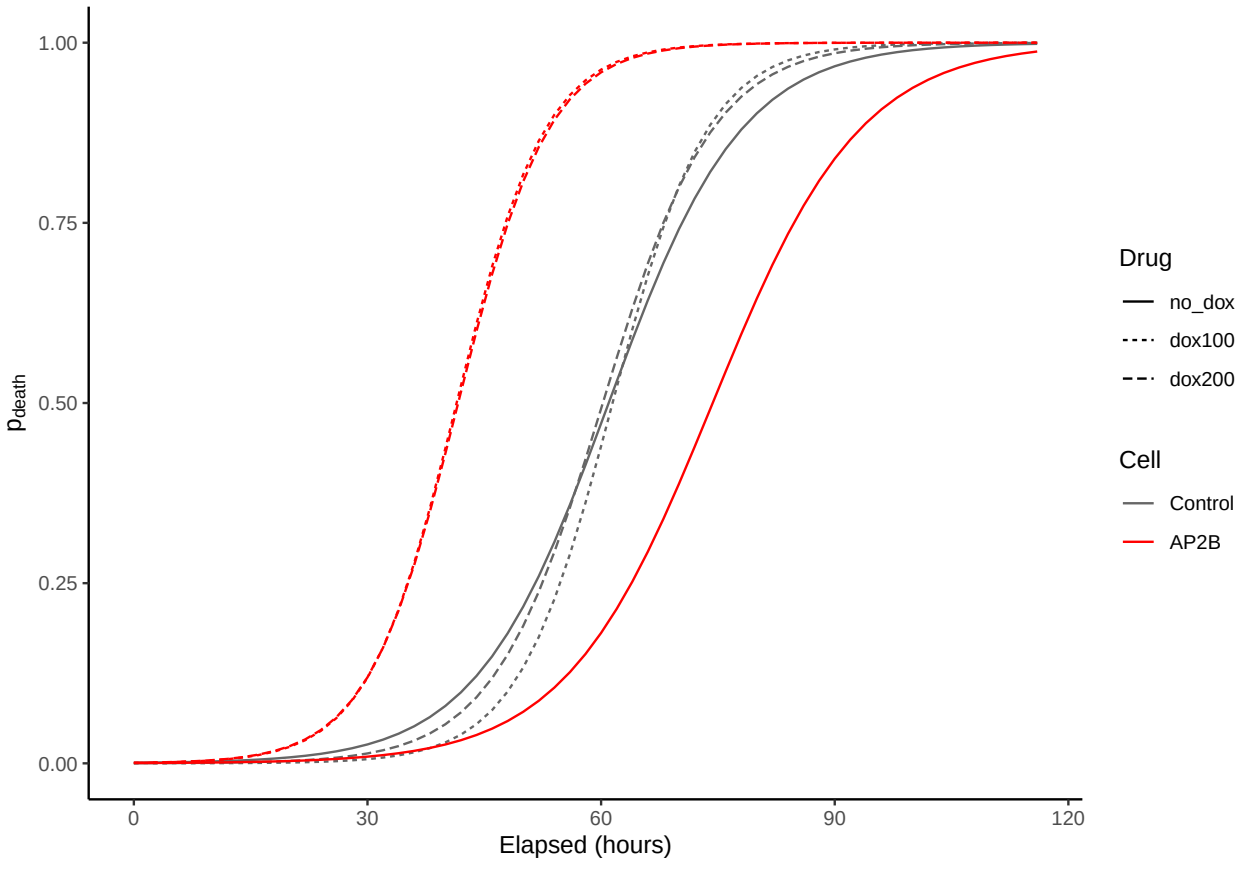
Cell Death: Analytic Challenges



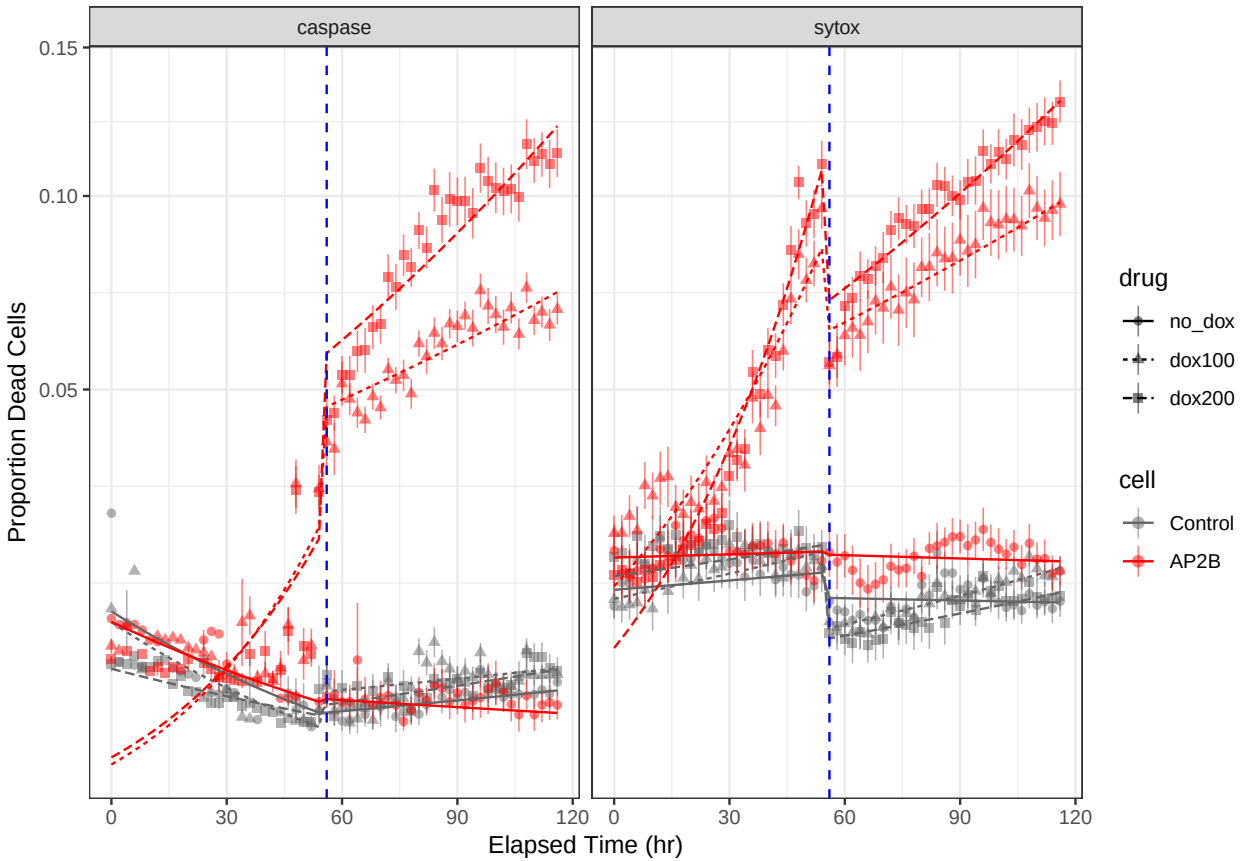
Cell Death: Analytic Challenges

- Data looks like zero-inflated, binomial (i.e. logistic) regression
  1. First we find the probability of observing > 0 dead cells with time
  2. Then we fit the proportion of dead cells
- This models the probability of a binary outcome, e.g. death = success
- Could fit:
  1. The proportion of dead cells at time t (transient)
  2. The total proportion of cells that have died by time t (cumulative)

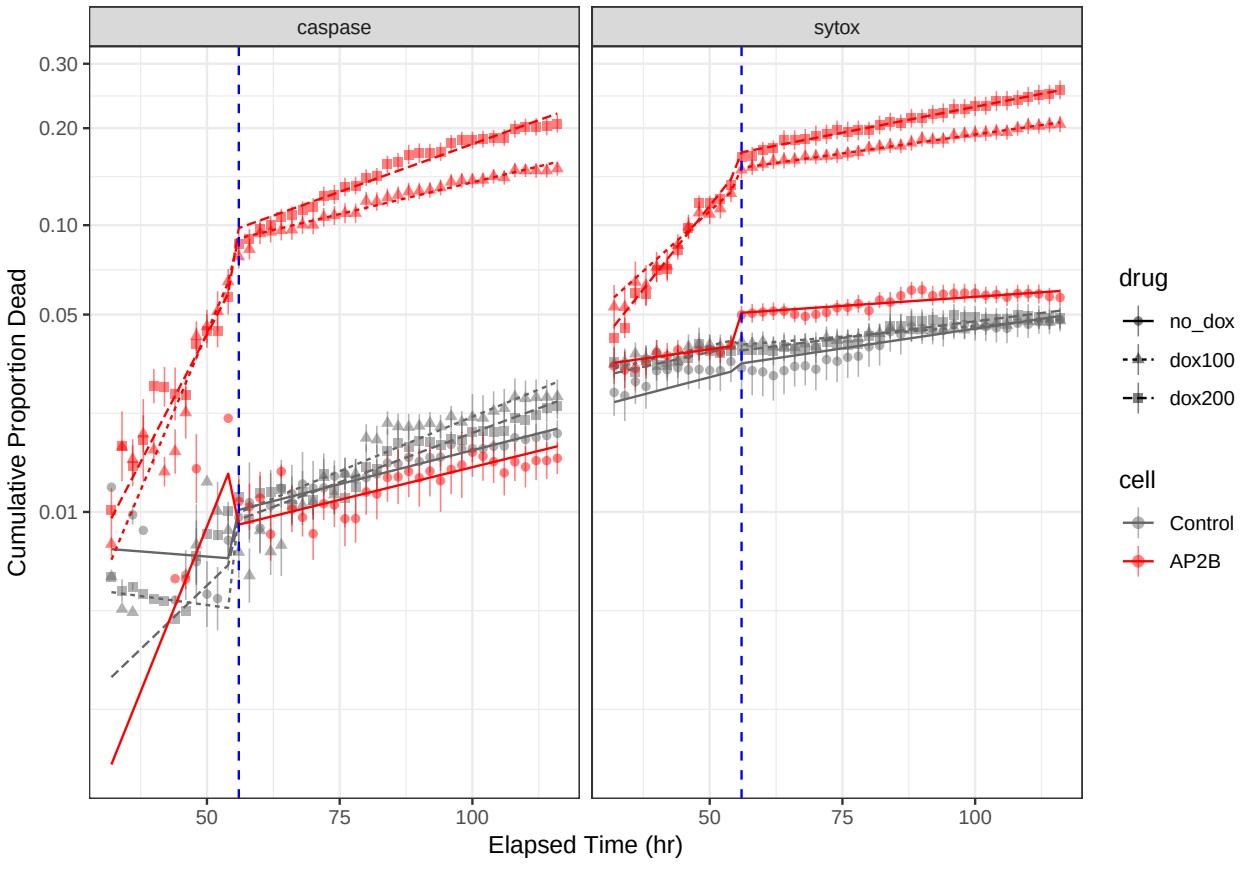
Probability of Any Death



Proportion Dead (Transient)



Proportion Dead (Cumulative)



Cell Death: Analytic Challenges

- Comparison of slopes between groups is *not appropriate*
- Need to analyse the differences in proportion dead
  - Odds Ratio as a function of time?
- This is sampling without replacement $\Longrightarrow$ Hypergeometric, not Binomial (???)
- GLMM convergence issues remain solvable, but delicately balanced