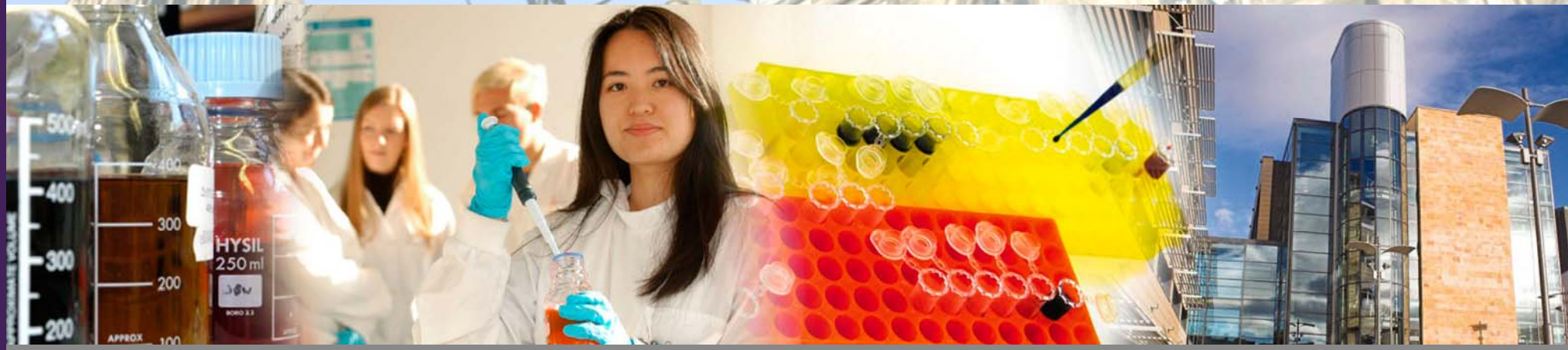
University of Glasgow | Faculty of Biomedical & Life Sciences

LICSB
Learning in Computational Systems Biology

Inference in a probabilistic model of dynamic DNA

April 2009

Catherine Higham, Lord Kelvin 1st year PhD student
Darren Monckton, Dan Haydon, Christina Cobbold and Douglas Wilcox



Probabilistic model

- Stochastic “Birth and Death” process
- Discrete-state continuous-time
- Capture key features DNA mutation mechanism

Rich data set (30,000 *de novo* mutations)

Bayesian framework to calibrate model

Dynamic DNA

Myotonic dystrophy

Initial findings

University of Glasgow

myotonic dystrophy type 1 (DM1)

- Common (1/8000)
- Variable symptoms including
  - Myotonia
  - Wasting limb and facial muscles
  - Cataracts

Need for quantitative predictive tools

LICSB
Learning in Computational Systems Biology

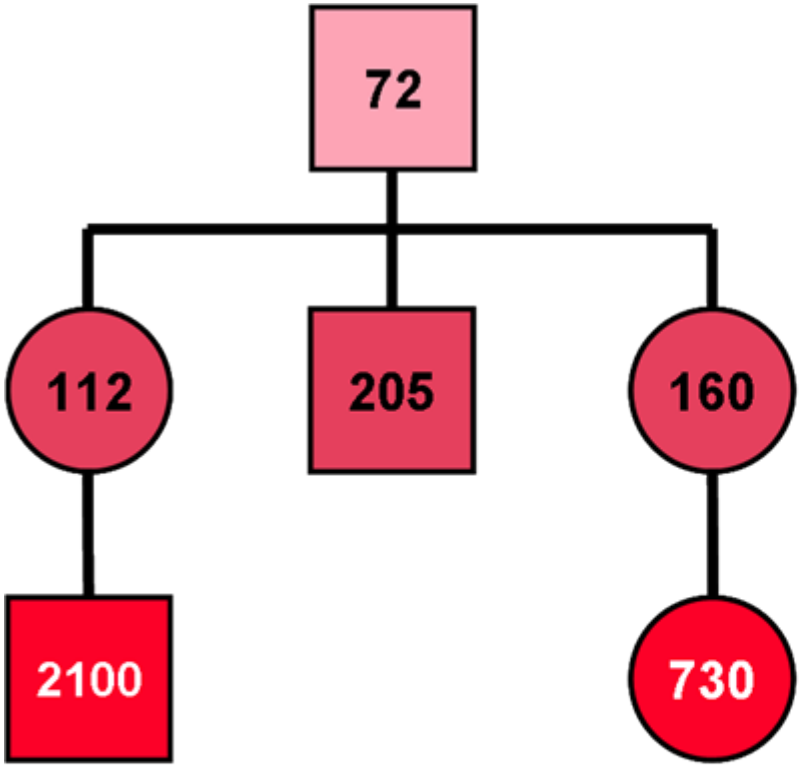
University of Glasgow

CTG repeat expansion in DM1

LICSB
Learning in Computational Systems Biology

University of Glasgow

Anticipation in DM1



late onset

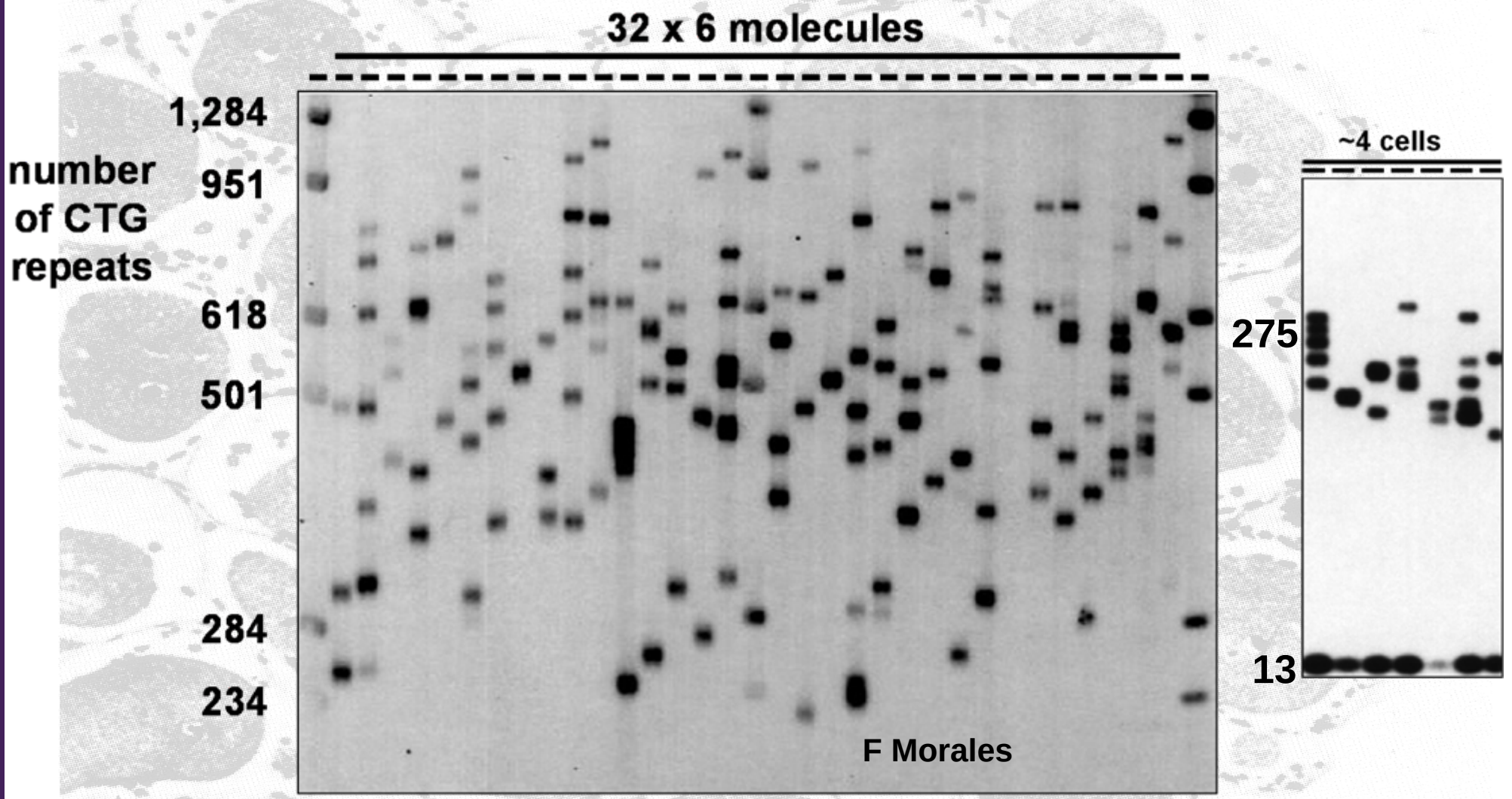
adult onset

congenital

LICSB
Learning in Computational Systems Biology

University of Glasgow

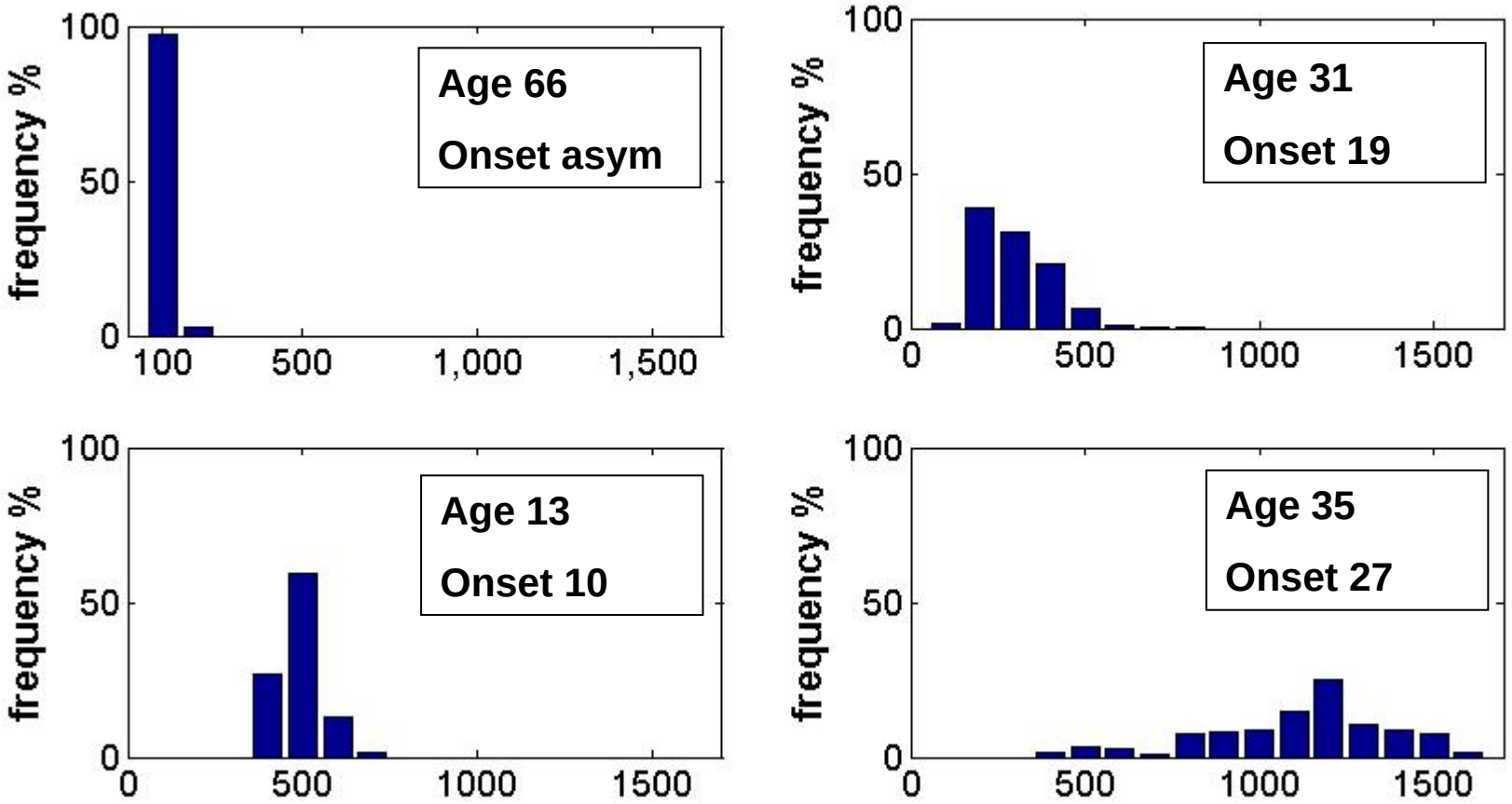
Variation between blood cells from one patient



Small pool analysis

University of Glasgow

Patient repeat length distributions



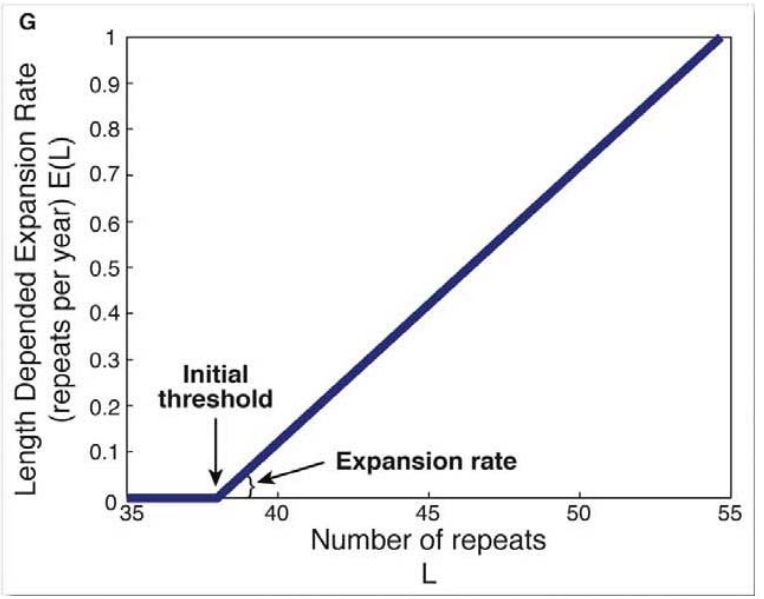
LICSB
Learning in Computational Systems Biology

Modelling Approach “Birth and Death” process

Kaplan et al, *PLoS Computational Biology* 2007 proposed probabilistic model based on expansion only

- Fitted model to mean length and age of onset for several diseases using standard fitting tools

Length dependent expansion



Kaplan et al. 2007

Expansion = “Birth”

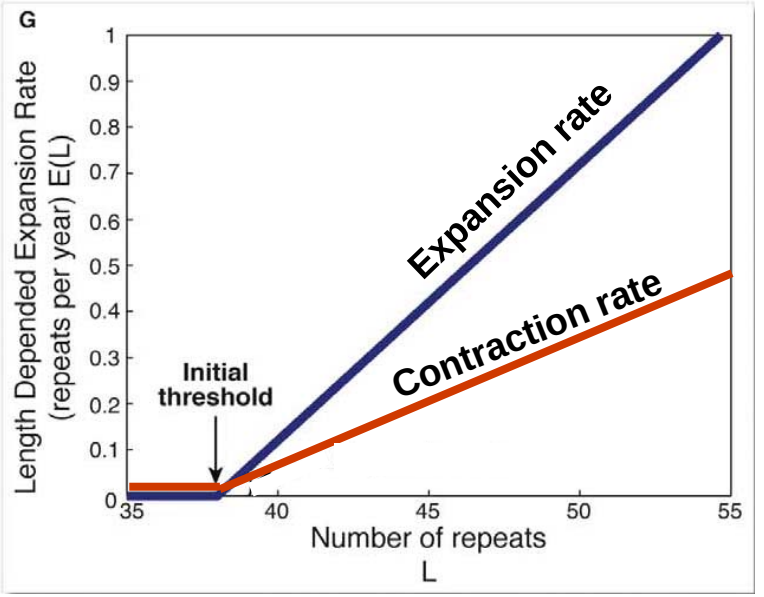
LICSB
Learning in
Computational Systems Biology

Modelling Approach "Birth and Death" process

Kaplan et al, *PLoS Computational Biology* 2007 proposed probabilistic model based on expansion only

We are generalising this model to include contraction

- Evidence of contractions
- Use Bayesian framework to evaluate this hypothesis
- Calibration using the whole dataset and not key statistics



Kaplan et al. 2007

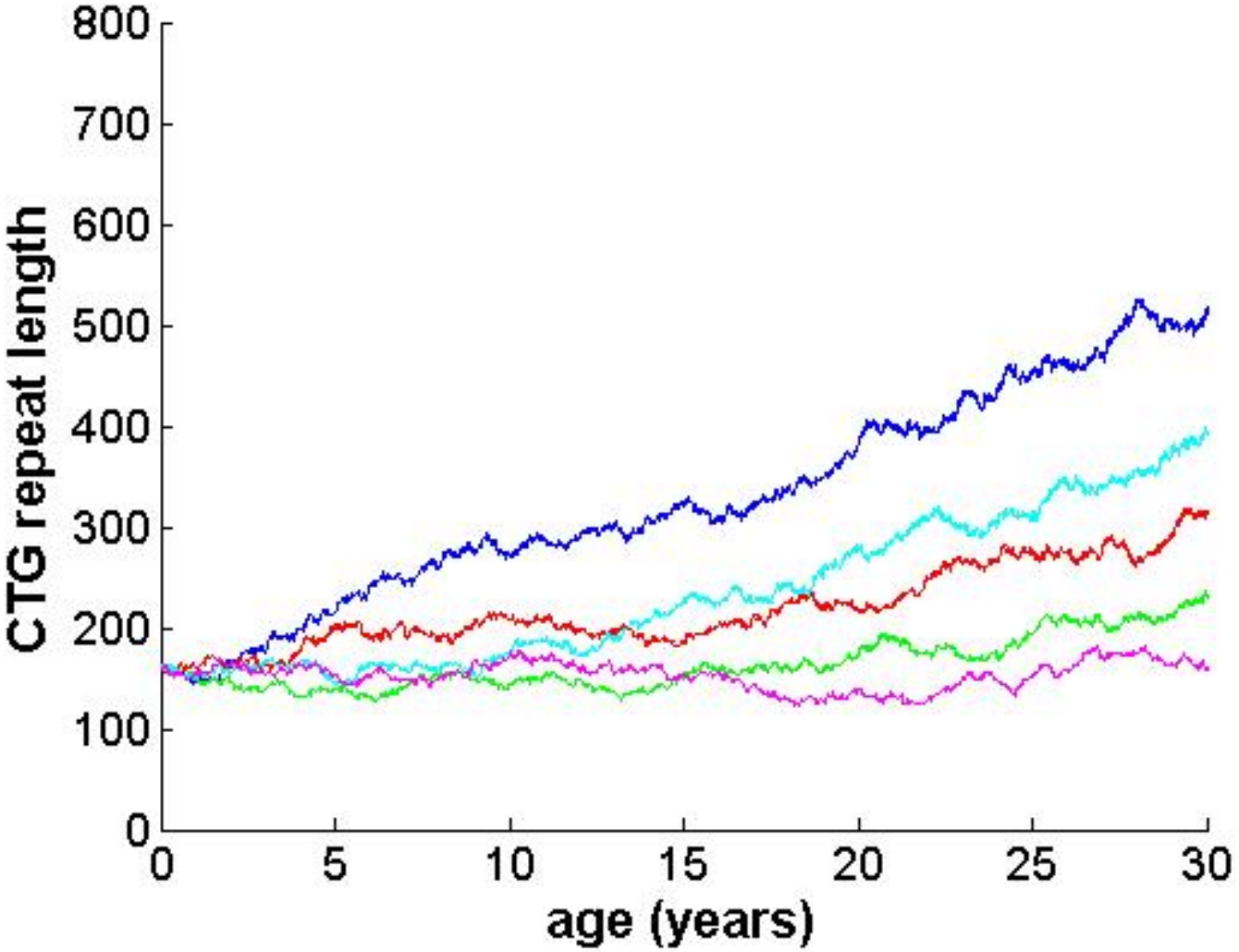
Expansion = "Birth"

Contraction = "Death"

LICSB
Learning in Computational Systems Biology

Evolution of repeat length

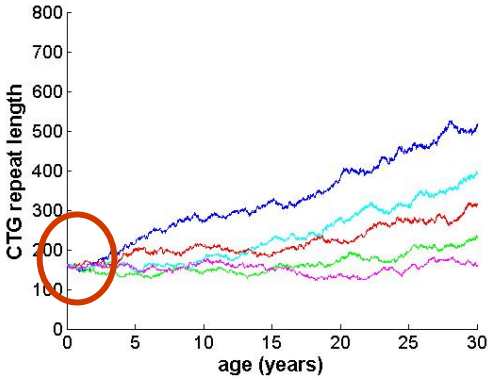
Using biologically realistic parameters

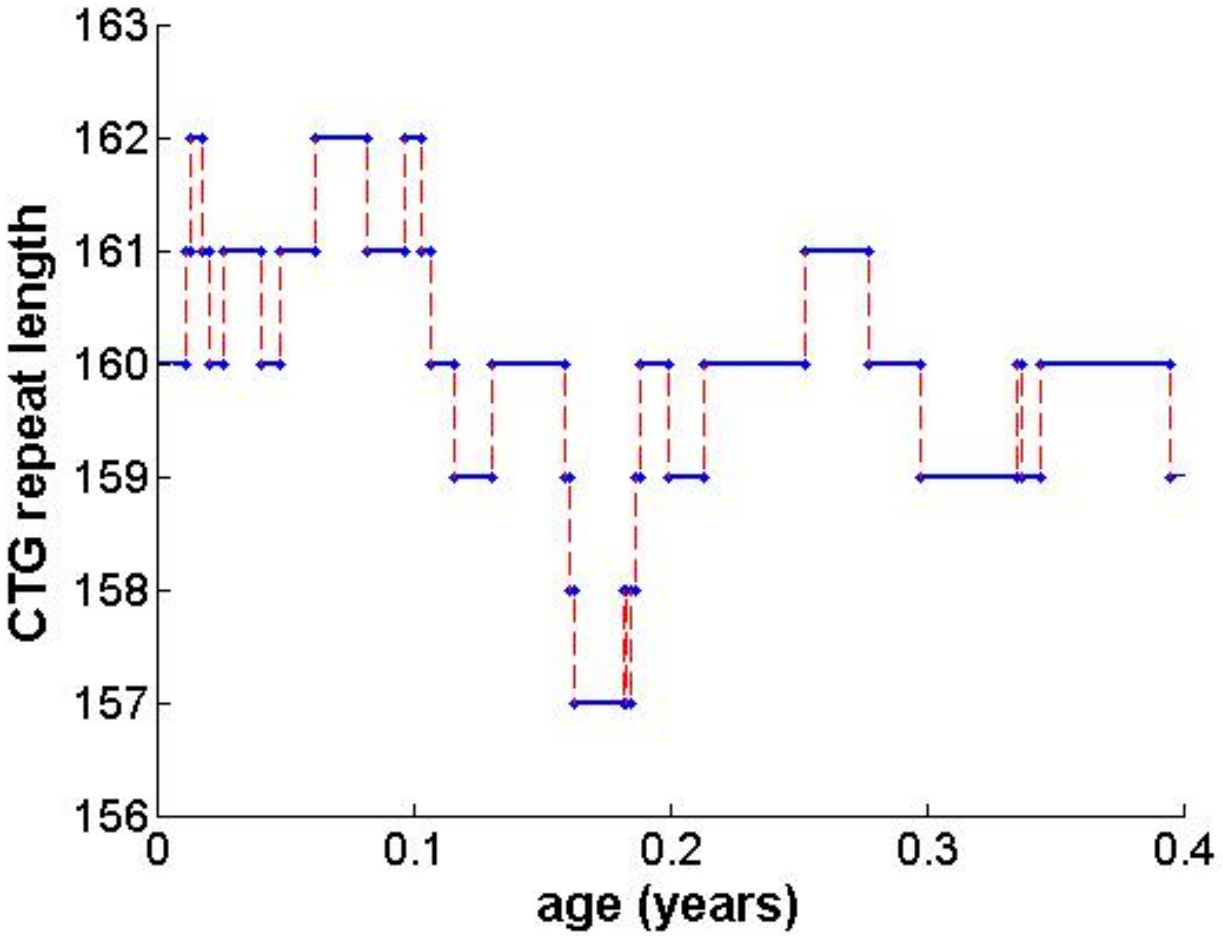


LICSB
Learning in Computational Systems Biology

Evolution of repeat length

Using biologically realistic parameters





LICSB
Learning in Computational Systems Biology

Bayesian framework

Bayes' Theorem links what we can quantify to what we would like to know

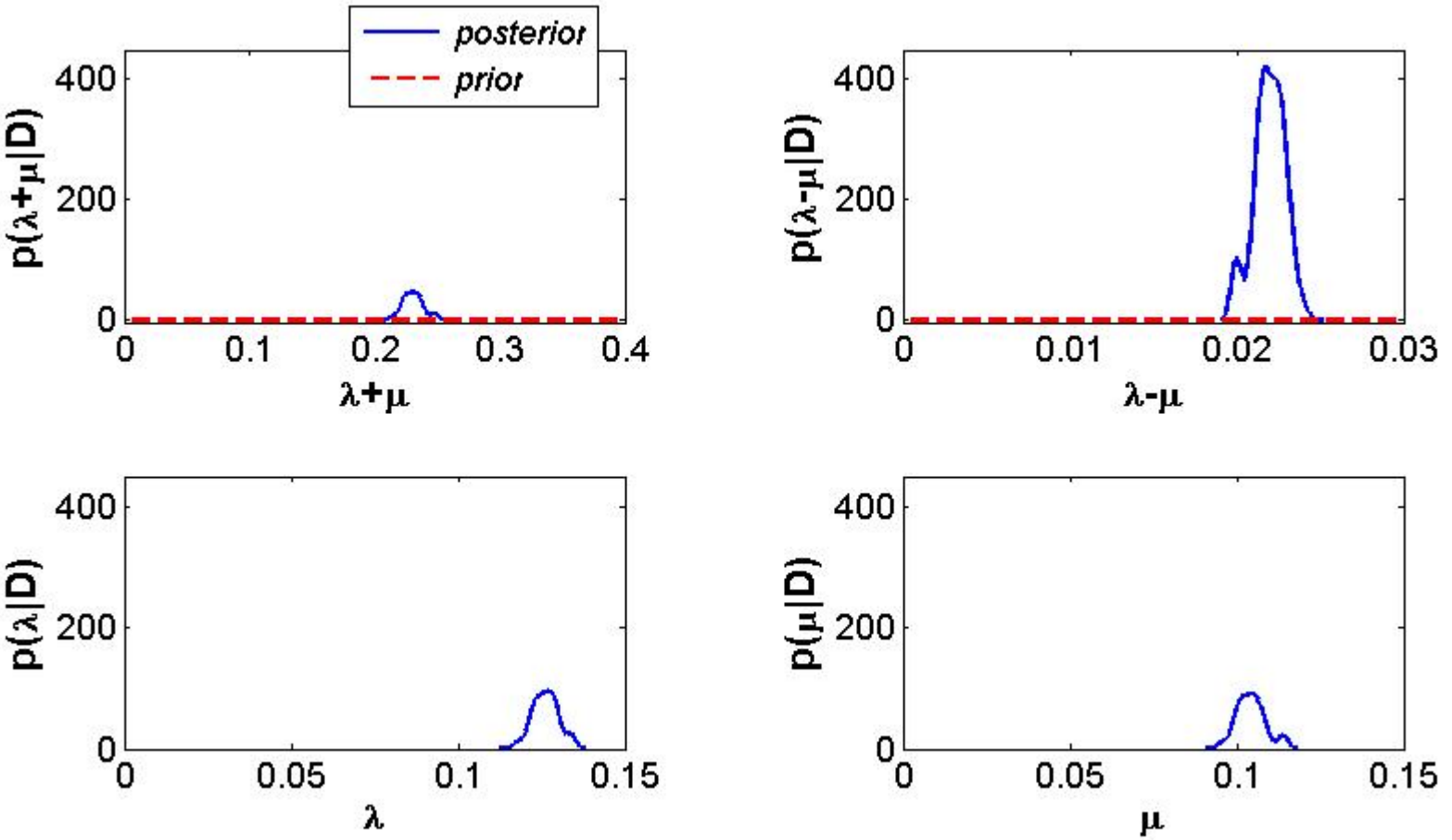
posterior proportional to likelihood x priors

prob(θ|data) is proportional to *prob(data|θ) prob(θ)*

- Let $P_n(t)$ be probability that #CTG repeats is *n* at patient age *t*
- $P_n(t)$ obtained by solving high dimensional master equation
  one ODE for each possible repeat length
- Assuming cell lengths evolve independently
  likelihood is product of $P_n(t)$ over all the data points

- Use Markov chain Monte Carlo (MCMC) to obtain posterior pdfs for parameters

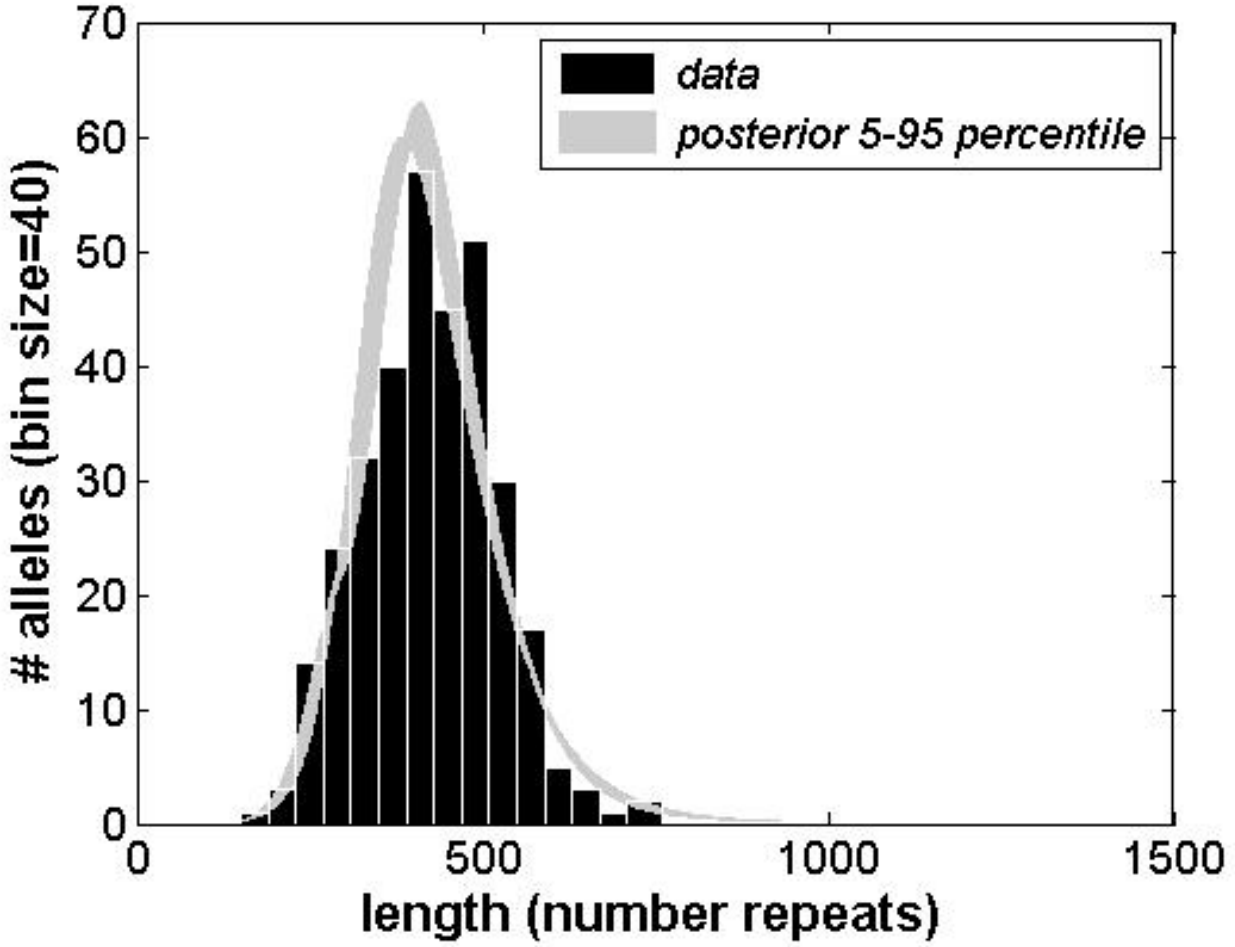
Posterior pdfs



DM1 patient (data from blood cells) age 56

University of Glasgow

Posterior envelope



LICSB
Learning in Computational Systems Biology

Initial findings and future work

Initial findings

- Evidence for contractions
- The observed tendency towards expansion of repeat length is the net result of many more expansion and contraction mutations than previously thought

Future work

- Investigate the full dataset
- Explore other forms for expansion and contraction
- Mathematical modelling will be extended to incorporate new data being generated by our lab

Aims

Improve prognostic information

genetic counselling, age of onset,
severity of disease

Clinical trials

account for variation and lower error bars for drugs

Slow/reverse repeat is therapeutic target

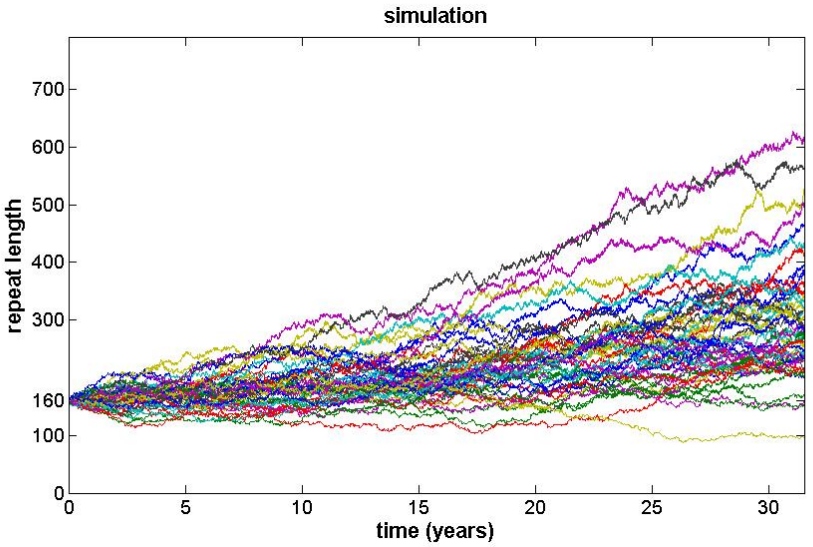
Biological understanding

Shed light mechanism unstable DNA

Stochastic “birth and death” process

suppose repeat length is *n* at time *t*, *λ* is the expansion rate, *μ* is the contraction rate and *a* is the threshold, then at time *t* + *δt*:

- Probability length is $n + 1 \approx \lambda (n-a) \delta t$
- Probability length is $n - 1 \approx \mu (n-a) \delta t$
- Probability length is $n \approx 1 - (\lambda + \mu) (n-a) \delta t$



each coloured line represents the evolution of repeat length in one cell

LICSB
Learning in Computational Systems Biology

Mathematical Analysis

We derive an expression for $P_n(t |\lambda,\mu,a)$ the probability of length n at time t given parameter values for λ, μ and a:

$$P_n(t+\delta t) = P_n(t)\,[1-(n-a)(\lambda-\mu)\delta t] + P_{n-1}(t)\,(n-a-1)\,\lambda\delta t + P_{n+1}(t)\,(n-a+1)\,\mu\,\delta t$$

Dividing by δt and letting $\delta t \rightarrow 0$ gives

$$dP_n(t)/dt = \lambda\,(n-a-1)\,P_{n-1}(t) - (\lambda+\mu)\,(n-a)\,P_n(t) + \mu\,(n-a+1)\,P_{n+1}(t)$$

or equivalently

$$d\mathbf{P_n}/dt = A\mathbf{P_n} \text{ with solution } \mathbf{P_n} = e^{At}\mathbf{P_a}$$

Numerical approach will allow λ and μ to vary with n