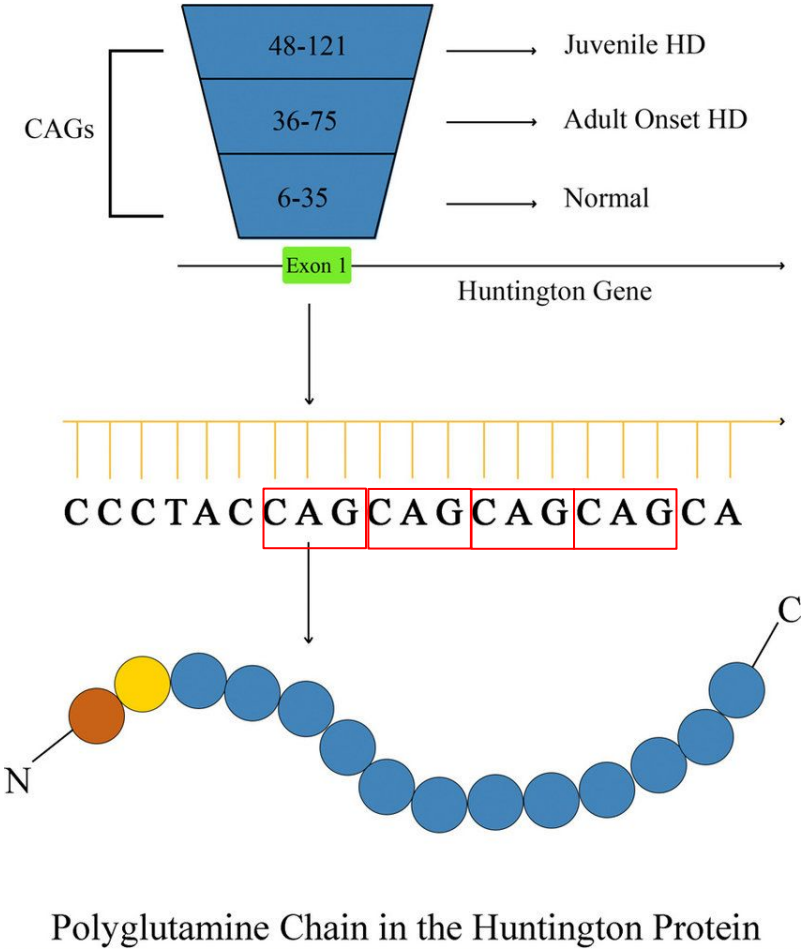
Models for Somatic CAG Repeat Expansion in the Onset and Progression of Huntington's Disease

by Steven Tan, mentored by Bob Handsaker and Seva Kashin

Project Conducted in the Steve McCarroll's Lab, Harvard Medical School

Huntington's Disease

- Huntington's Disease (HD) is an inherited neurodegenerative disease.
- Caused by high number of CAG repeats (36+ repeats) in the Huntingtin (*HTT*) gene.
- Patients with the disease exhibit somatic expansion, where the number of repeats expands individually in each brain cell.
- Leads to more diverse and longer CAG lengths, which can be toxic and lead to neuronal cell death.



Polyglutamine Chain in the Huntington Protein

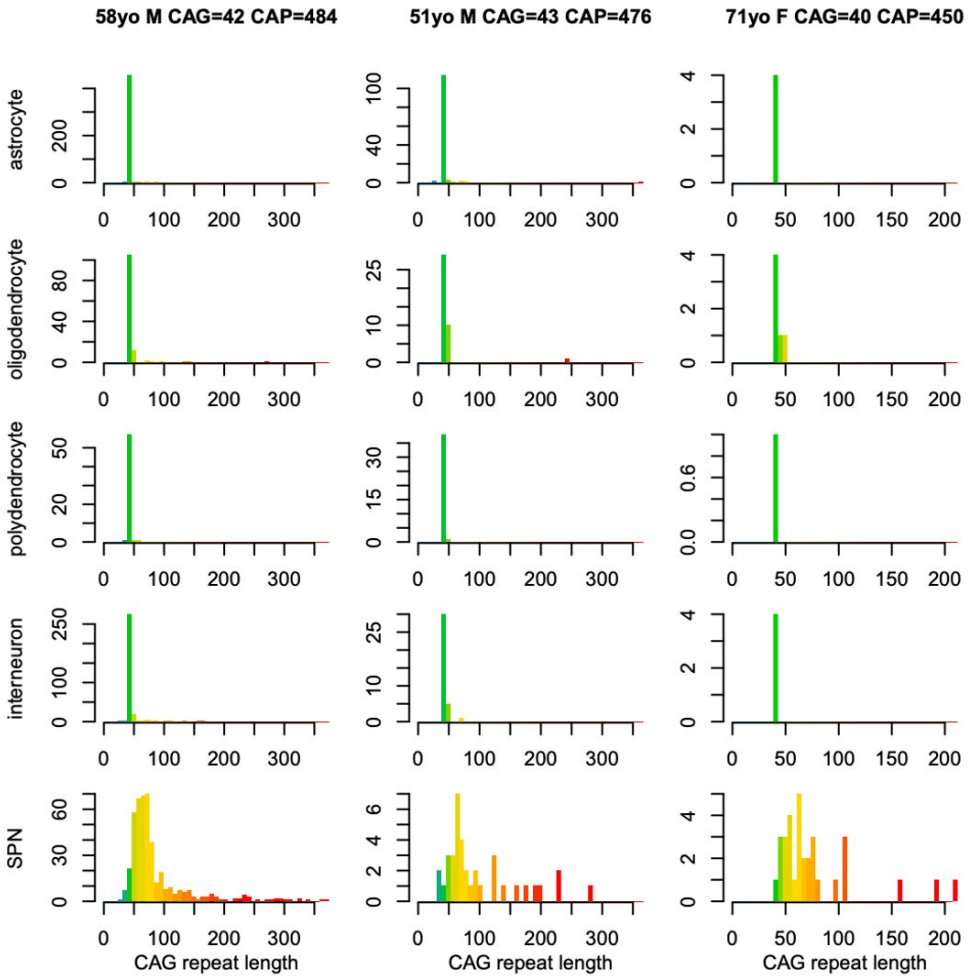
Figure from: https://pubmed.ncbi.nlm.nih.gov/32811395/

The McCarroll Lab collected new biological data:

- Post-mortem brain samples from HD patients.
- Precise CAG measurements from many individual cells.
- Cell type specific.

Somatic expansion occurs mainly in the Spiny Projection Neurons (SPNs), resulting in diverse CAG repeat lengths.

Because expansion is highly cell-type specific, this new data enables the analysis we performed.



astrocytes

oligodendrocytes

polydendrocytes

interneurons

SPNs

Data collected by the McCarroll Lab in collaboration with the McLean Brain Bank

Research Goals

- Somatic expansion is increasingly thought to be the driver of disease progression.
  - Very high numbers of CAG repeat can be toxic and lead to neuronal cell death.
  - Consistent with the fact that cell death is most heavily observed in SPNs.
- Understanding mechanisms behind expansion could be crucial for developing therapeutics.
- How can we create a model that simulates somatic expansion?
- Can such a model help us better understand the expansion mechanism and disease progression?

Modeling Somatic Expansion

The model was inspired by the work of Dr. John Warner and previous models applied for other repeat expansion diseases.

The generative model works as follows:

- Model one cell's CAG length at a time. Assume mutations are a random stochastic process with a certain rate that increases with longer CAG lengths.
- Each mutation has a probability of being an expansion and otherwise contraction, which increase/decrease the CAG length by 1.

Parameters:

- **r** = mutation rate parameter (mutation rate calculated as a function of **r** and CAG length).
- **p** = probability of a mutation being an expansion.

Model Pseudocode

Let **A = age at death**, **I = inherited CAG length**, for the patient we are trying to model. Function below returns the CAG length at death of a single cell.

r = mutation rate parameter

p = probability of expansion

Algorithm 1 Base Model

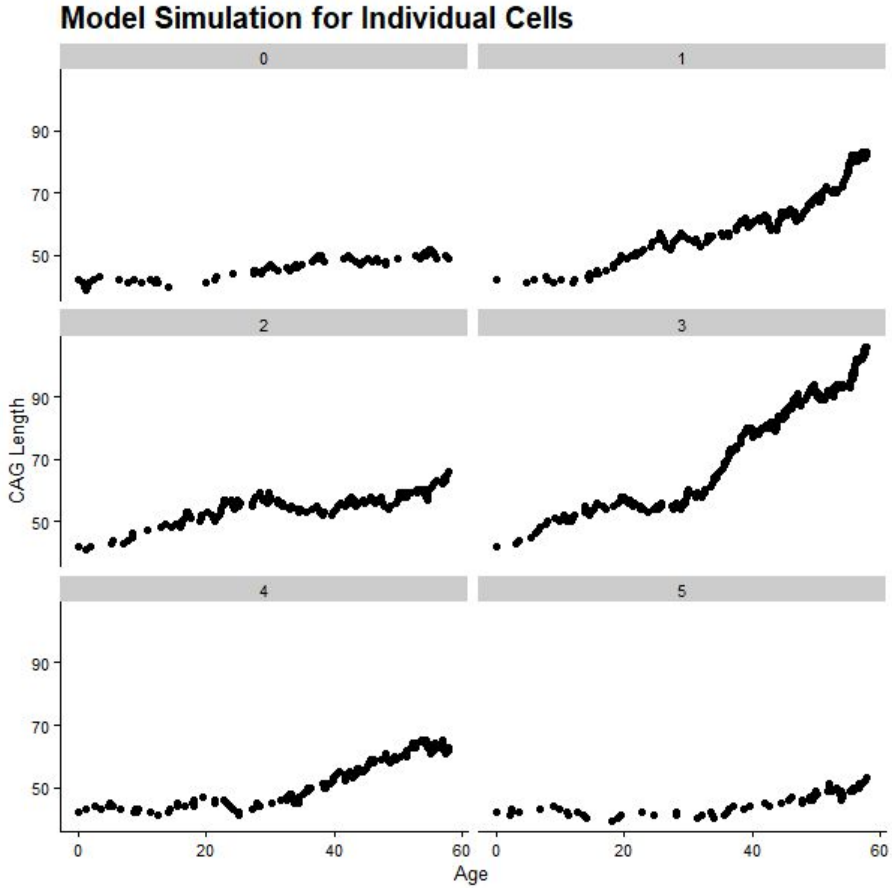
function SIMULATECELLBASEMODEL(r, p, A, I)

- $T \leftarrow 0$ ▷ T stores the current age in years
- $X \leftarrow I$ ▷ X stores the current CAG length
- **while** $T < A$ **do** ▷ repeat process until the age of death is reached
  - **if** $X \leq 35$ **then** ▷ assume somatic expansion not occur 35 CAGs or below
    - **break**
  - **end if**
  - $T_{next} \sim Exp(r \cdot (X - 35))$ ▷ years from next mutation, drawn from an exponential distribution
  - $T \leftarrow T + T_{next}$
  - $u \sim U(0, 1)$ ▷ draw from a uniform distribution
  - **if** $u < p$ **then** ▷ with probability p
    - $X \leftarrow X + 1$ ▷ expansion
  - **else**
    - $X \leftarrow X - 1$ ▷ contraction
  - **end if**
- **end while**
- **return** X ▷ return CAG length at death

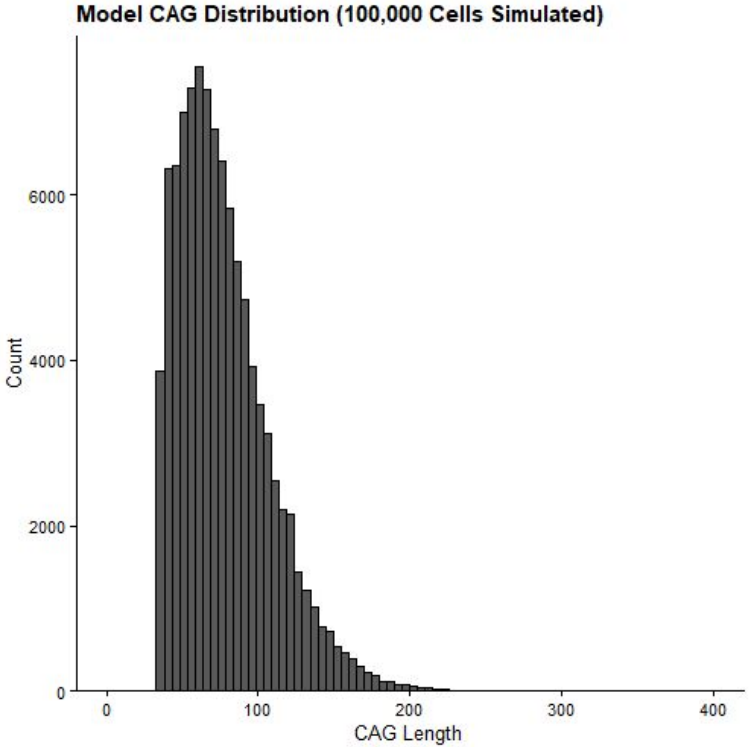
end function

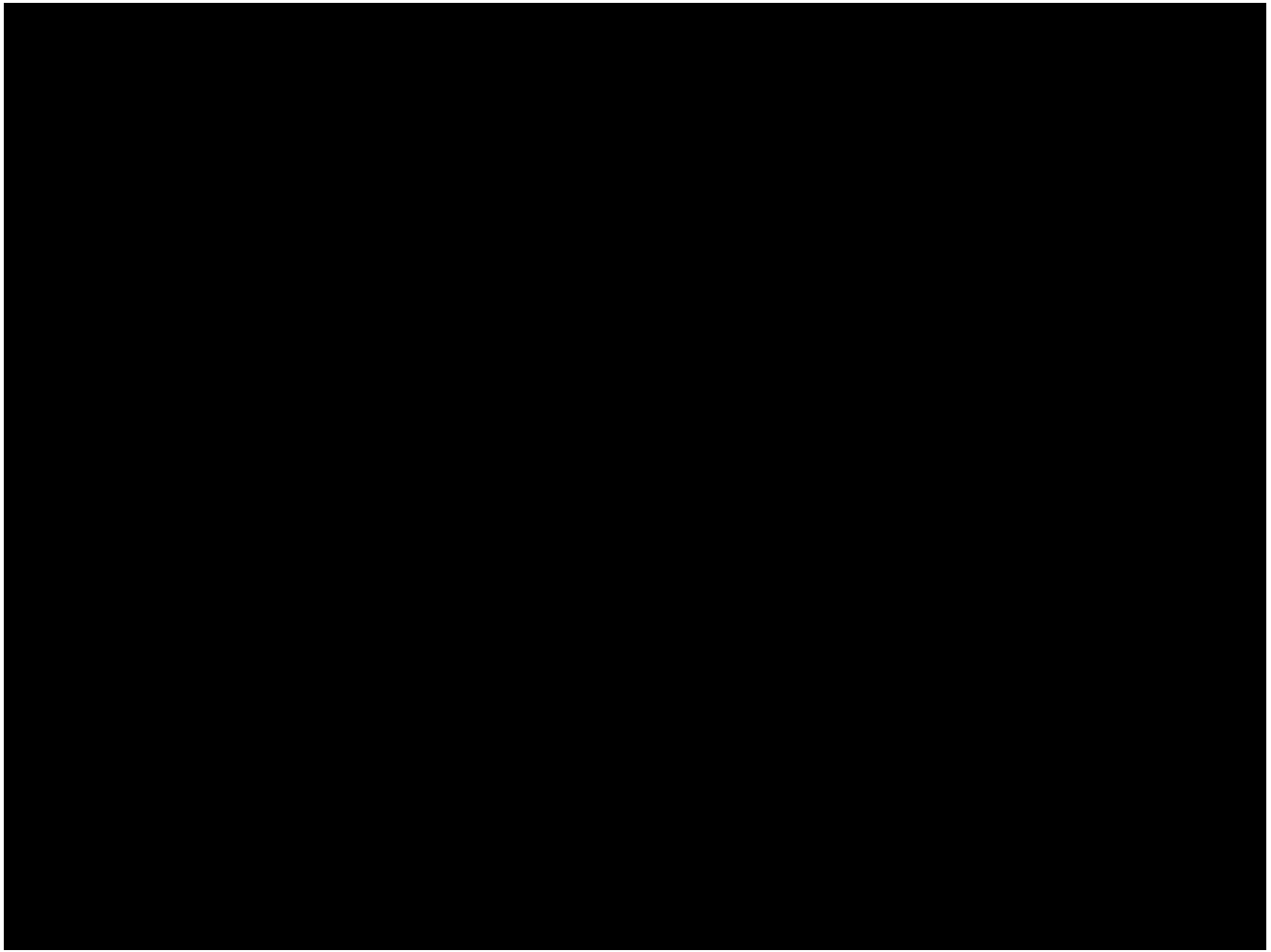
Running the Model

Run with: **r** = 0.135, **p** = 0.615, **A** = 58, **I** = 42



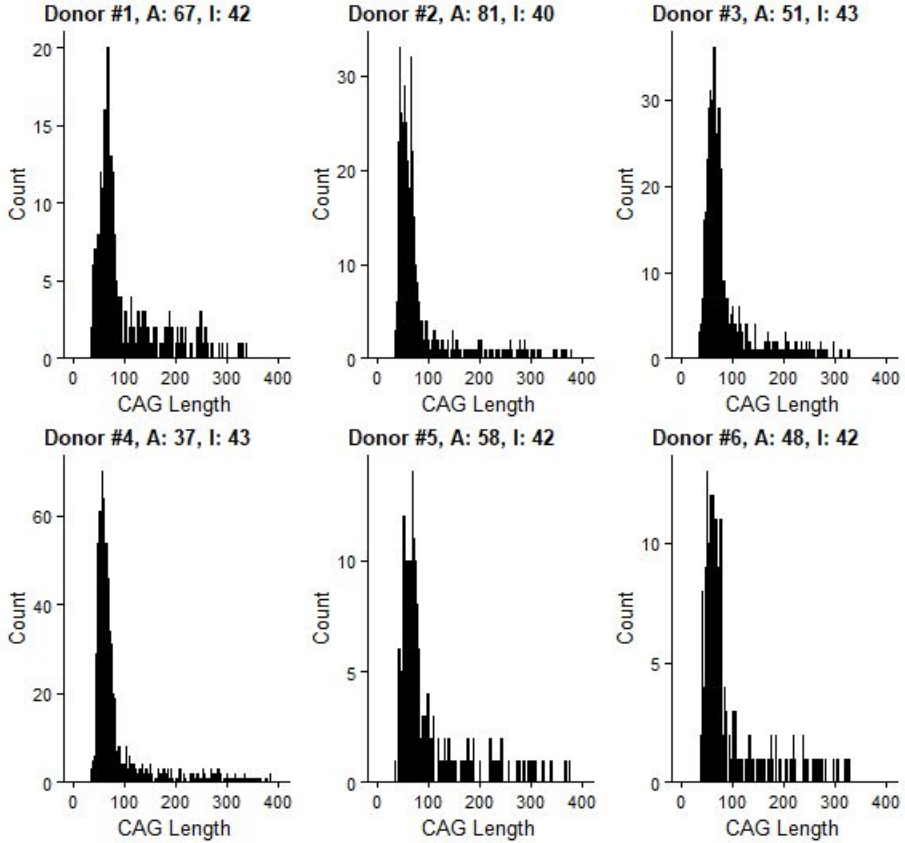
Simulate many cells



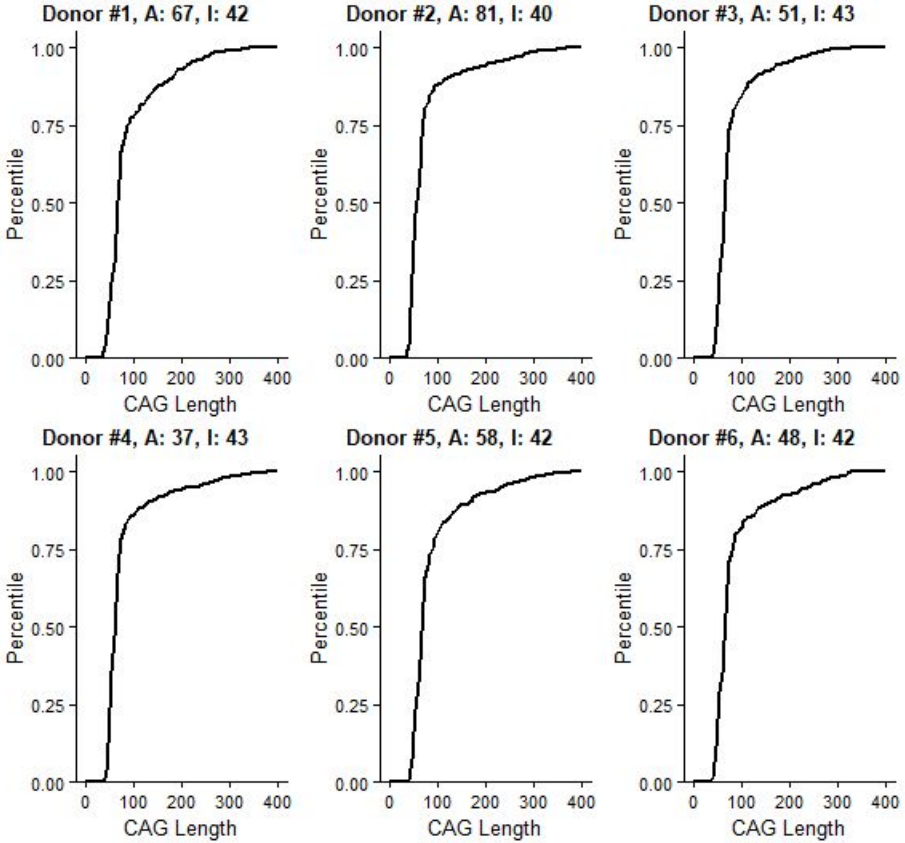


CAG Repeat Data from Post-Mortem Brain Samples

CAG Distributions

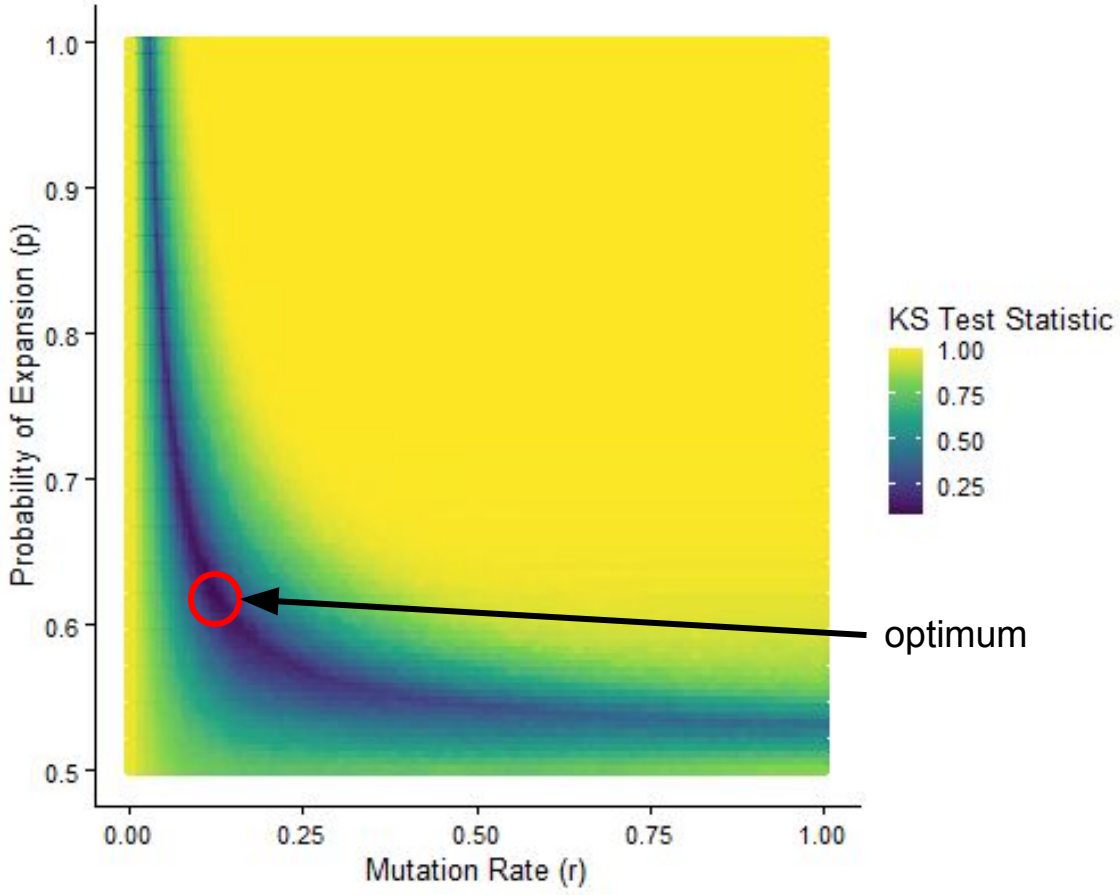


Cumulative Distributions

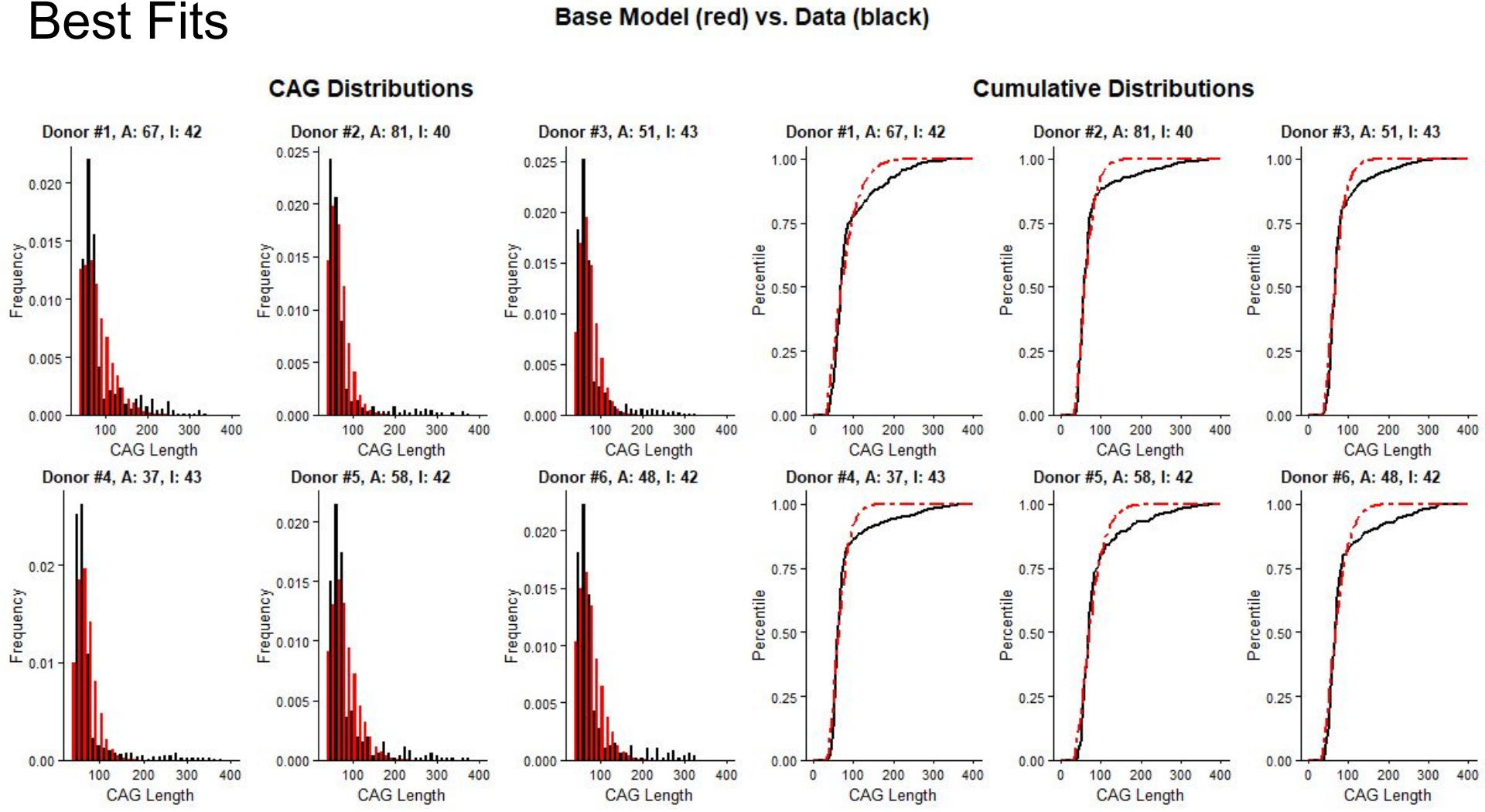


Fitting the Model

- Grid search on (**r**, **p**) parameter pairs, for each generate distribution of CAG lengths and compare to the observed data.
- To evaluate fit between the model and data distributions, we use the Kolmogorov-Smirnov (KS) test statistic.
- Each point represents one set of parameters and is colored by the KS test statistic.

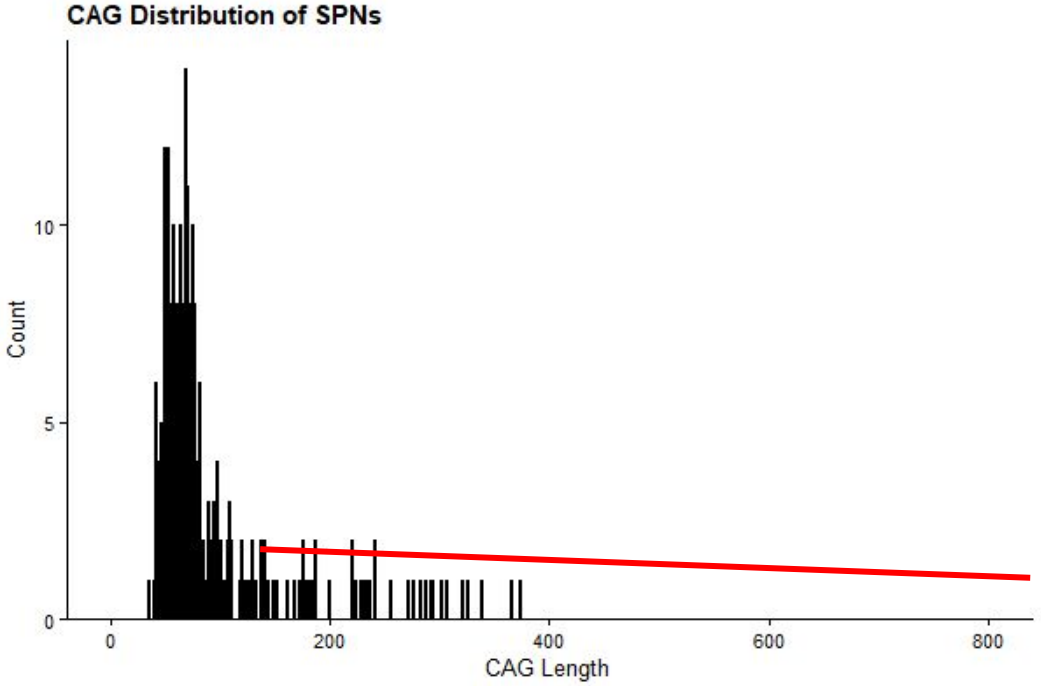


Best Fits


Base Model (red) vs. Data (black)
CAG Distributions
Cumulative Distributions
Donor #1, A: 67, I: 42
Donor #2, A: 81, I: 40
Donor #3, A: 51, I: 43
Donor #4, A: 37, I: 43
Donor #5, A: 58, I: 42
Donor #6, A: 48, I: 42
Frequency
Percentile
CAG Length

Cell Death

- This model doesn't account for cell death, which is pervasive in later stages in HD.
- We often see >70% cell death in the SPNs, and these dead cells generally have much higher CAG lengths.



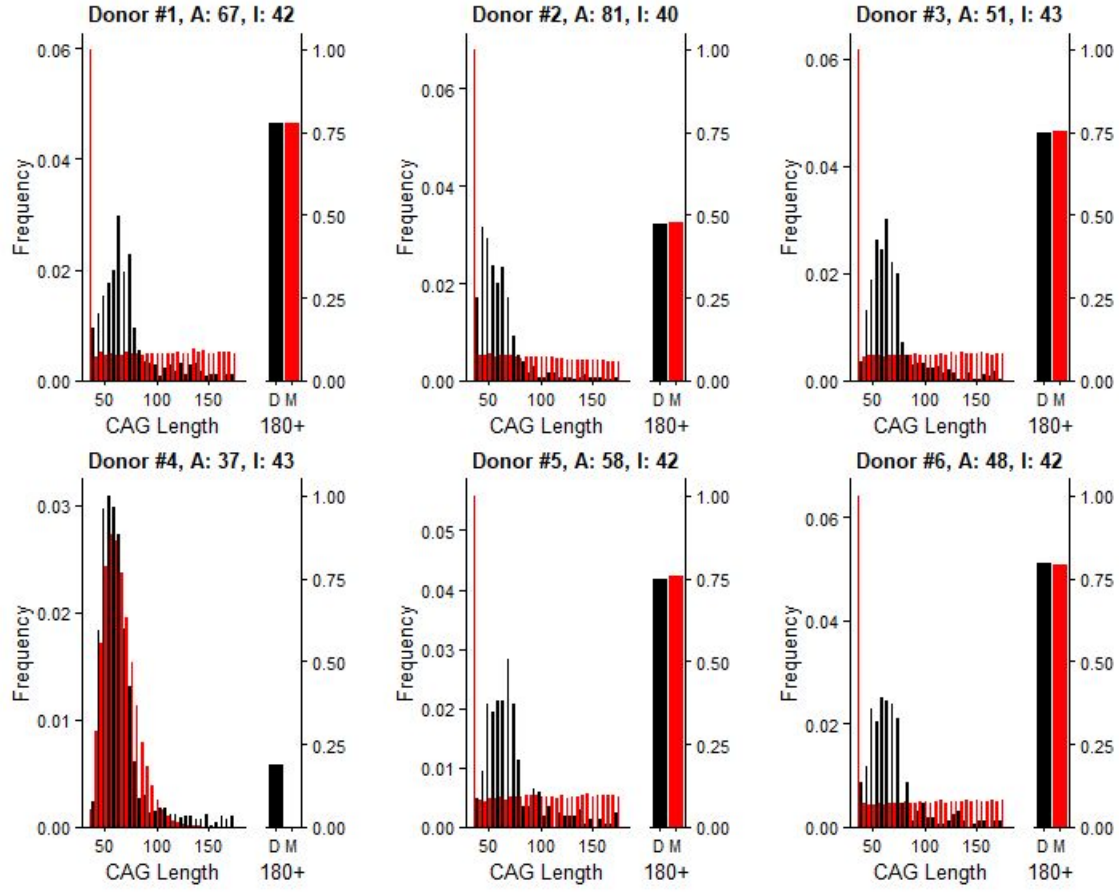
- The data we see isn't representative because cells that have died are not observed.
  - Cells with high CAG lengths are greatly underrepresented. If cells didn't die, many would reach very high CAG lengths that we'd never observe.

Accounting for Cell Death

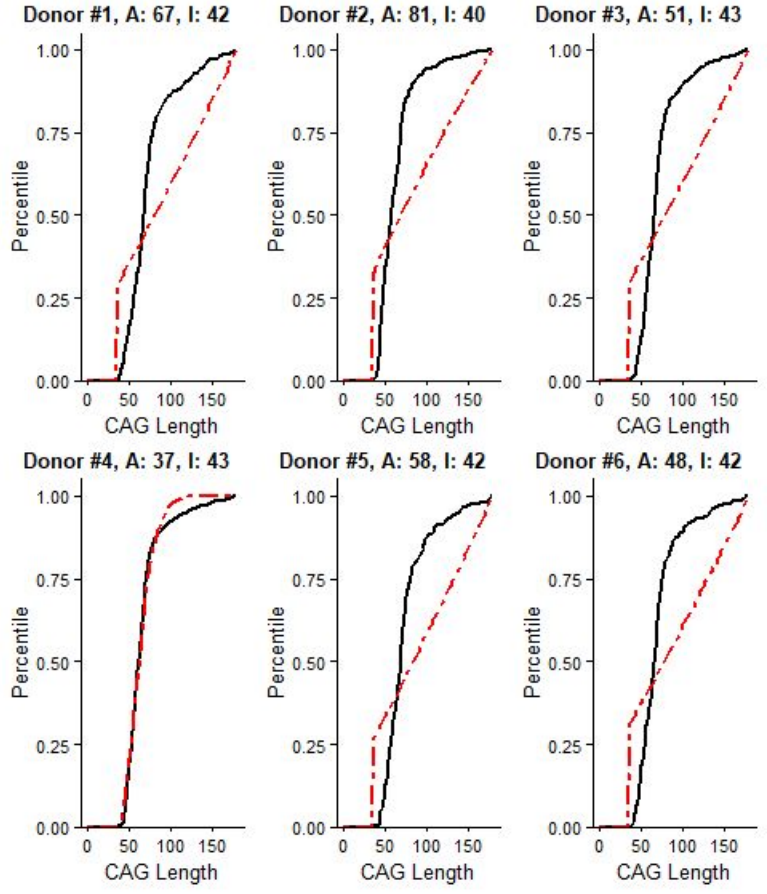
- The lab has found evidence that transcriptional dysregulation of genes in a cell starts happening above ~180 CAGs.
    - So we assume cells that have died are also mostly above 180 CAGs
- In our data we have a certain amount of observed cells above 180 CAGs, but we add back whatever percent of cells that have died to that count.
- We want the model to get a similar proportion of cells above 180 CAGs.
- We modify the objective function to take into account the KS test statistic for the CAG distribution of cells < 180 CAGs, and also the fraction of cells above 180 CAGs in the model and data.

Base Model (red) vs. Data (black), Accounting for Cell Death

CAG Distributions and Fraction of Cells ≥ 180 CAGs



Cumulative Distributions



Need for a Better Model

Base Model unable to fit. Doesn’t get enough cells to a high enough CAG length without compromising the shape.

We tried a number of different models, and one of the most flexible modelled expansion as an exponential process.

Previously the rate was linearly related to the number of CAGs above 35, with this model they’re exponentially related.

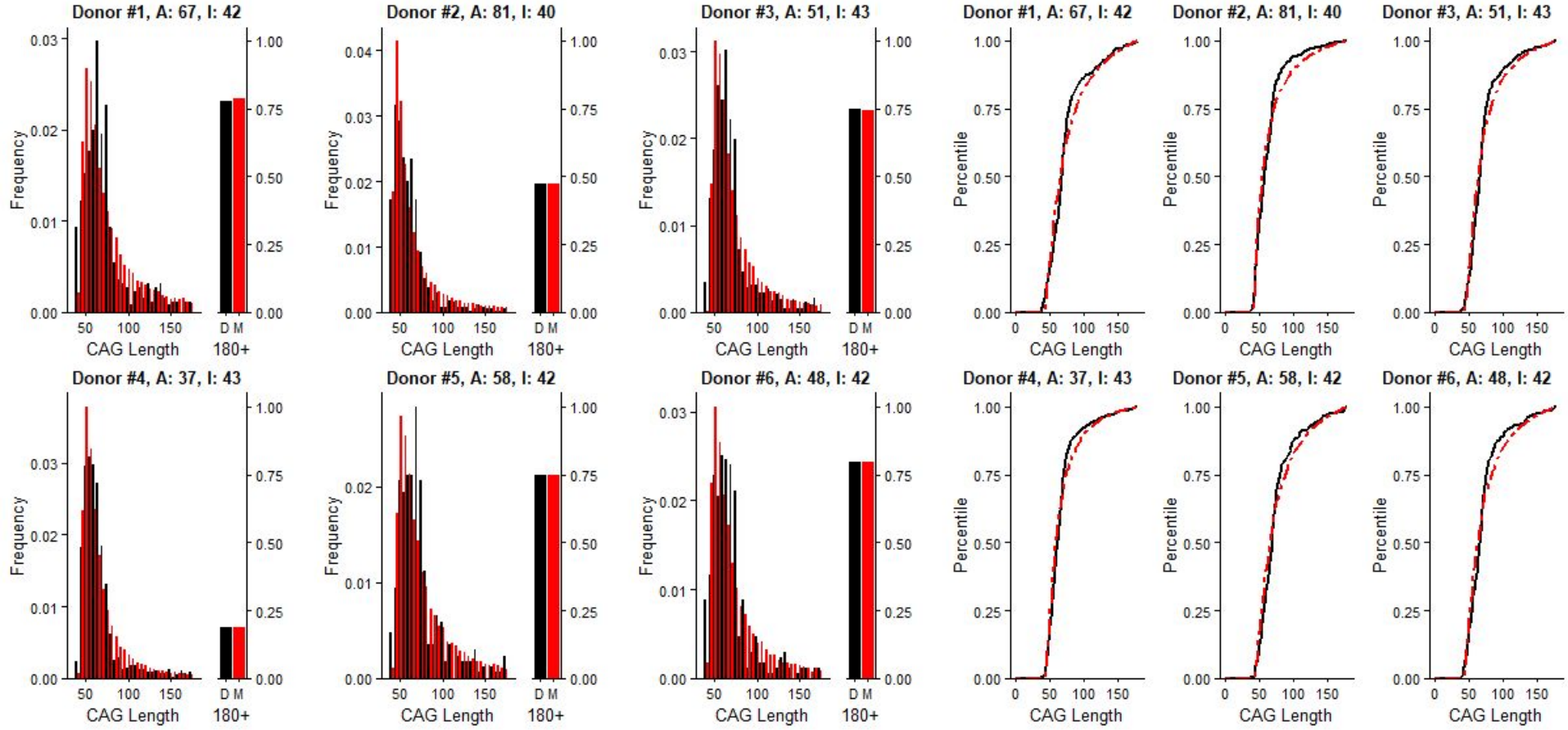
Extra parameter **e**:

$\mathbf{T}_{\mathbf{next}} \sim \text{Exp}(\mathbf{r} * (\mathbf{X} - 35)^{\mathbf{e}})$

Rest of the model is the same. We perform a grid search on triples (**r**, **p**, **e**) and find the best fitting parameters.

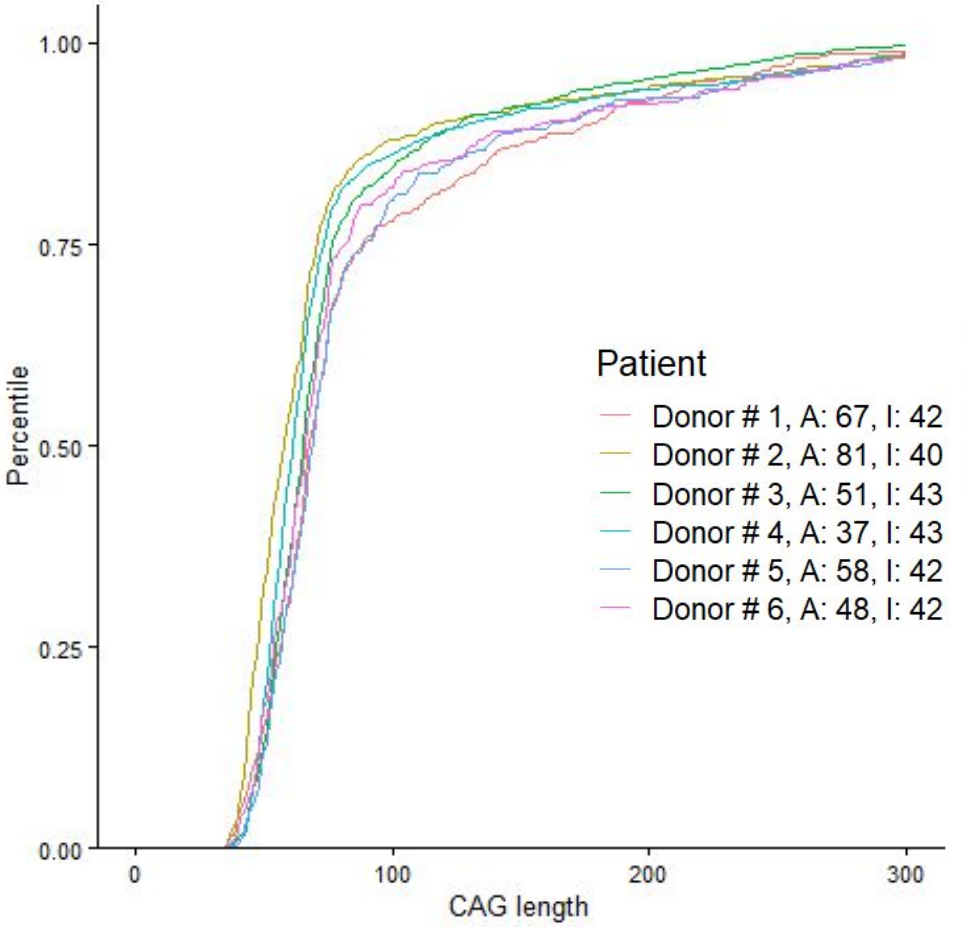
Exponential Model (red) vs. Data (black), Accounting for Cell Death

CAG Distributions and Fraction of Cells ≥ 180 CAGs



Cumulative Distributions

Two Biological Processes



CDFs take similar shape even with big age difference. Consistently see a sharp bend at ~70 CAGs, which is where the long tail starts.

This suggests that there may be 2 biological processes, one starting at 36 CAGs, the second starting at ~70 CAGs. Second process could explain the tail we observe.

Two Process Model

This inspires a 2 process model.

Parameters:

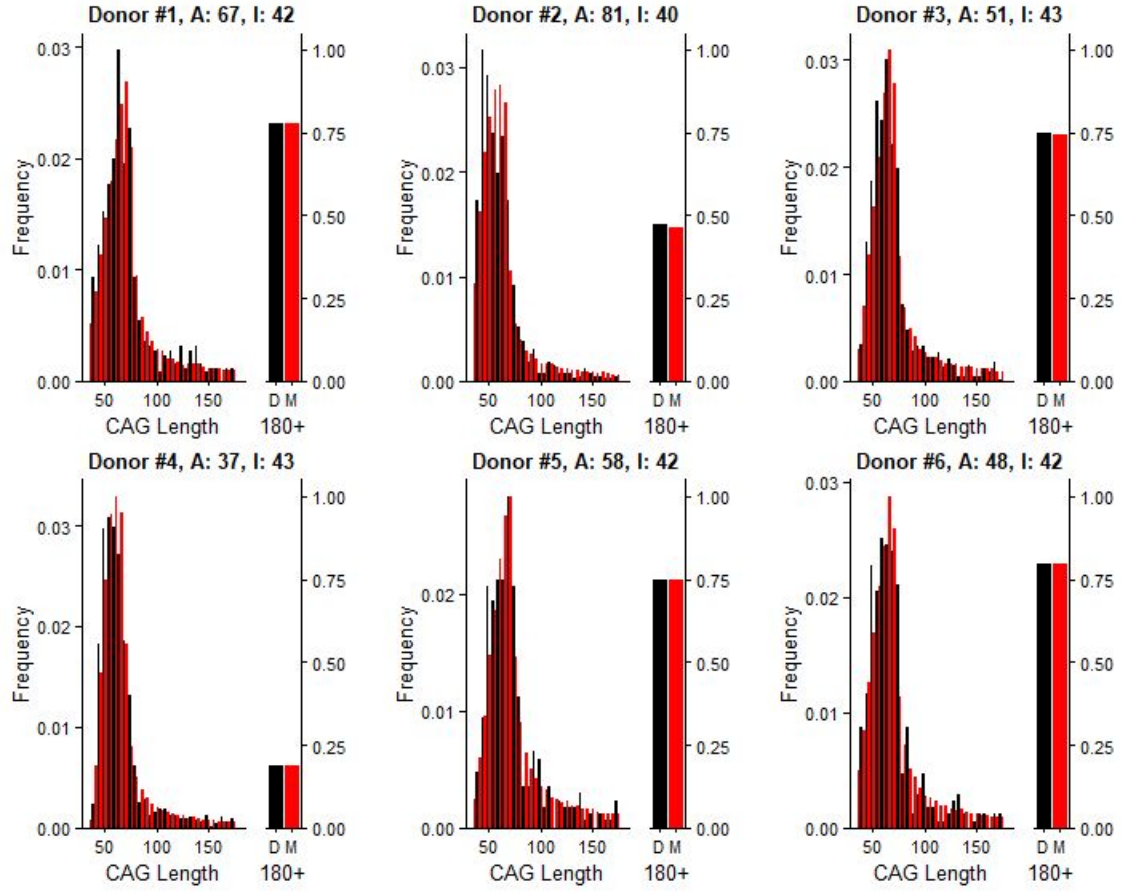
- $\mathbf{r_1}$ = rate of process 1
- $\mathbf{r_2}$ = rate of process 2
- **p** = probability of expansion for both processes
- $\mathbf{t_2}$ = threshold for process 2

The second process operates in a very similar way to process 1, with its own separate threshold and rate.

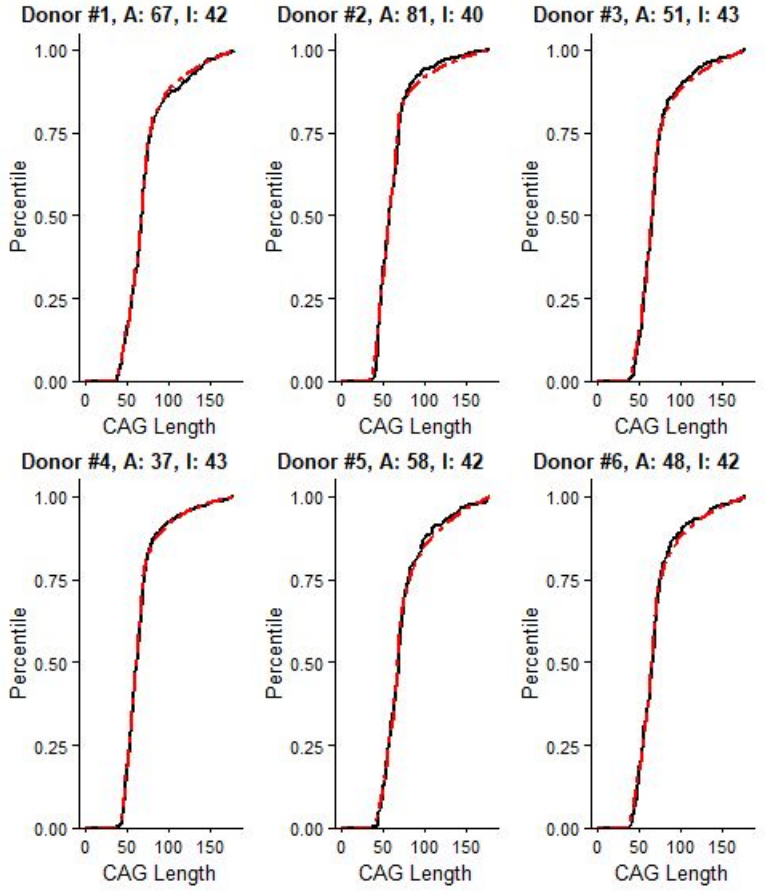
We run a grid search on 4-tuples ($\mathbf{r_1}$, $\mathbf{r_2}$, **p**, $\mathbf{t_2}$).

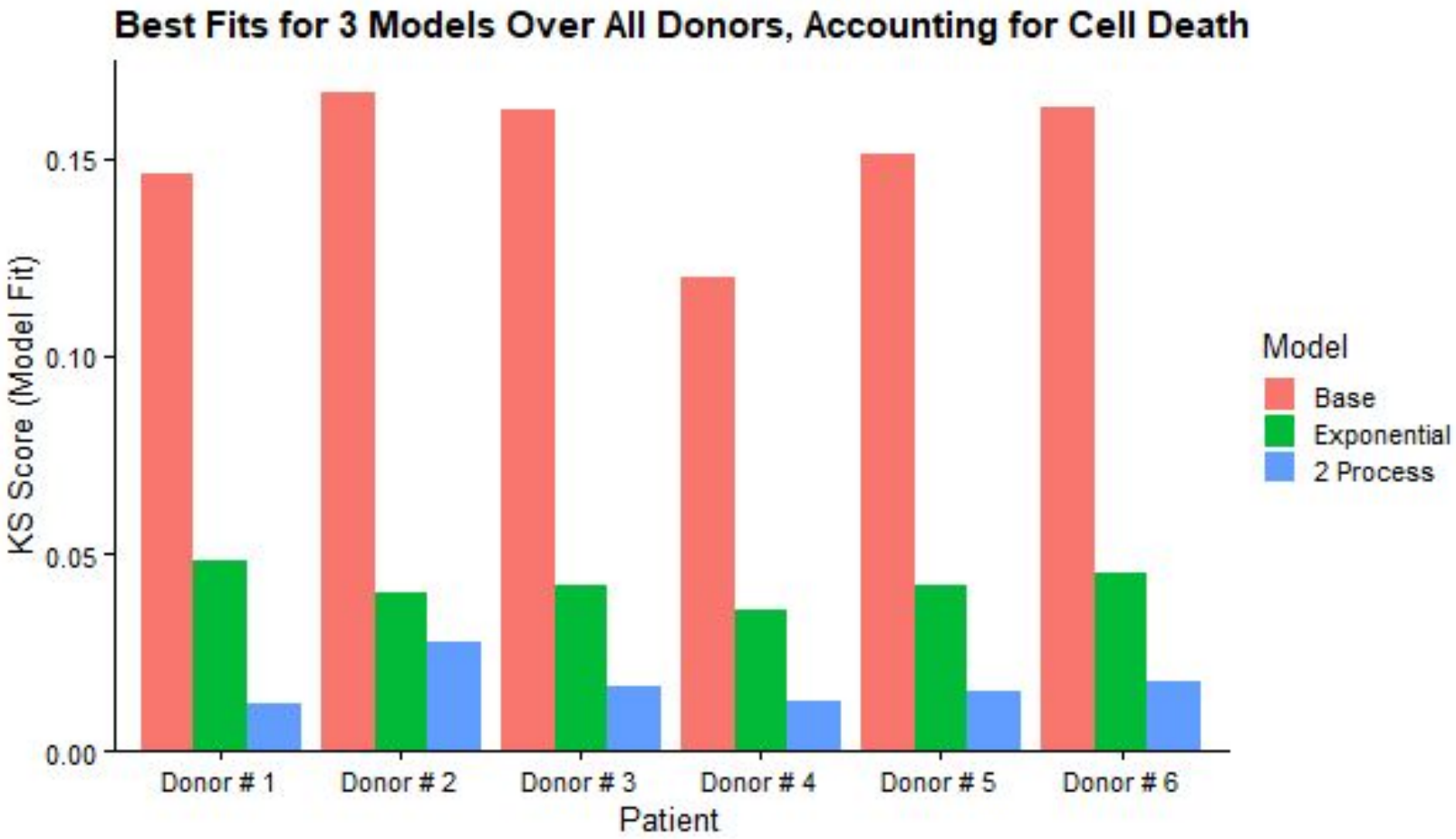
2 Process Model (red) vs. Data (black), Accounting for Cell Death

CAG Distributions and Fraction of Cells ≥ 180 CAGs



Cumulative Distributions




Best Fits for 3 Models Over All Donors, Accounting for Cell Death
KS Score (Model Fit)
0.15
0.10
0.05
0.00
Donor # 1
Donor # 2
Donor # 3
Donor # 4
Donor # 5
Donor # 6
Patient
Model
Base
Exponential
2 Process

Conclusions and Future Work

- The rate of somatic expansion is likely not linear, but grows extremely fast in comparison to CAG length.
- Statistical models with 2 processes are able to explain the data better than single process models.
    - Indicates the possibility of 2 biological mechanisms behind somatic expansion which have differing rates and thresholds — at 35 CAGs and ~70 CAGs.
- Impact of the Model:
    - Deepens our understanding of the molecular mechanisms that drive somatic expansion, as it allows us to see the progression of CAG lengths over time.
    - May inform the design of clinical trials or therapeutics targeting somatic expansion, which patients may benefit from.

Acknowledgements

Bob Handsaker

Seva Kashin

Steven McCarroll

The McCarroll Lab

John Warner (from CHDI)

CHDI FOUNDATION