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

By Steven Tan

Mentored by Bob Handsaker[1,2], Seva Kashin[1,2], and Steve McCarroll[1,2]

1 Department of Genetics, Harvard Medical School

2 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard

1 Abstract

Huntington's Disease (HD) is an inherited neurodegenerative disease caused by alleles with 36 or more repeats of the trinucleotide sequence CAG in the huntingtin (HTT) gene. A person with HD inherits an allele with a certain CAG length (> 35) at birth, but somatic expansion within the brain is known to occur throughout their lifetime, resulting in a situation in which individual cells have longer and highly variable numbers of CAG repeats. Somatic expansion is increasingly thought to be a driver of disease onset, as age-at-onset associates with modifier alleles in DNA-repair genes that regulate somatic expansion. Thus, a better understanding of the mechanisms behind CAG repeat expansion could be crucial in revealing novel therapeutic targets. In this study, we adapted a stochastic birth-death model previously used for a different repeat-expansion disease (Myotonic Dystrophy Type 1, or DM1) to model CAG repeat expansion in HD. We made use of a new kind of biological data, in which CAG length has been measured precisely in many individual neurons of the most vulnerable type from post mortem brain samples. We found that single-process models consisting of only one length threshold and rate — models that succeeded in modeling DM1 — were unable to explain all features of repeat expansion data observed in HD patients. Effectively fitting the data required models consisting of two separate processes, suggesting that there may be two distinct biological mechanisms underlying CAG repeat expansion in HD. These processes appear to have differing rates and CAG length thresholds: one at roughly 36 CAGs — a threshold for instability — and another at 70 CAGs, which we hypothesize is a threshold for accelerated expansion. This model deepens our understanding of disease progression and can inform the design of clinical trials for new therapies that target the somatic expansion process.

2 Background

2.1 Huntington's Disease

Huntington's Disease (HD) is one of many diseases known to be caused by inherited alleles with expanded trinucleotide repeats (others include Myotonic Dystrophy Type 1 and Fragile X Syndrome) [1]. HD specifically is caused by inherited alleles with 36 or more CAG repeats in exon 1 of the huntingtin (*HTT*) gene [2]. The disease causes the progressive death of neurons within the brain and is characterized by uncontrolled movements, dementia, and other psychological disturbances. A person only needs to inherit one allele with 36 or more CAG repeats from either parent to contract the disease, and higher CAG lengths above that threshold are associated with earlier age of onset [3]. The average age of onset for HD patients is around 40 years, but for patients inheriting longer alleles it may occur as early as in their juvenile years [4].

Although all cells will start with the same inherited CAG length, patients with the disease exhibit somatic expansion of CAG repeats within certain cell types, in which the number of repeats changes in individual cells over time due to mutations [2]. The instability of repeats generally favors expansions, or increases in CAG repeat length, though contractions may also occur that decrease CAG length [5]. Somatic expansion occurs mostly within neurons, and most abundantly in the Spiny Projection Neurons (SPNs, also called Medium Spiny Neurons or MSNs), and can result in some SPNs reaching extremely high CAG lengths upwards of one thousand [6]. Because neurons are long lived and typically do not regenerate, these changes in CAG length persist in individual neurons throughout the patient's lifetime. These expanded CAG repeats are thought to be toxic and cause cell death, and thus drive disease pathogenesis [7]. This is consistent with the fact the human striatum — whose neuronal population mostly consists of SPNs — is observed to have high amounts of neuronal cell death for patients suffering from HD [8].

2.2 Biology Behind CAG Repeat Expansion

The number of CAG repeats within the DNA sequence of a cell can change over time through mutations. The mechanism behind these mutations is thought to occur through two steps. The first step is the formation of hairpin structures, which are loops of extra DNA that extrude outwards of the main DNA strands [9]. These hairpins form in the repeated CAG segment of the DNA. The second step is when DNA mismatch repair (MMR), which is a biological mechanism that is normally beneficial in fixing mutations, attempts to resolve these hairpin structures [10]. Depending on how they are resolved, the hairpin structure can either be incorporated or removed altogether from the DNA, leading to an increase or decrease in the number of CAG repeats. An increase in CAG length is called an expansion and a decrease is called a contraction.

One way for the DNA hairpin structures to form is through DNA polymerase slippage [11]. Since the brain has high metabolic levels, oxidation occurs frequently in neurons resulting in DNA

damage. Whenever there is damage, DNA damage repair mechanisms such as base excision repair (BER) will cut the damaged segment of the DNA, and DNA polymerase will re-synthesize cut strand based on the other healthy strand [10]. However, the DNA polymerase can slip off during synthesis, and because of the repetitive structure of the CAG repeats, the slipped off DNA strand can attach to the other strand at the wrong spot, shifted over by some number of CAG repeats. This leaves a hairpin extruding out of the DNA. When a new polymerase attaches, it does not recognize the hairpin and will re-synthesize the rest of the DNA strand, leaving the extra loop as a part of the DNA.

In the second step, the DNA mismatch repair (MMR) system will attempt to resolve the hairpins formed. DNA mismatch repair is a biological mechanism that corrects errors and damage within DNA sequences, and generally lowers the rate of mutations. However, in the case of somatic expansion, it may actually worsen the situation by incorrectly repairing these hairpin structures [10]. For example, MMR may cut the strand opposite to the extrusion, pull the hairpin straight, then re-synthesize the original cut strand. In this scenario, the hairpin sequence becomes incorporated into the DNA, resulting in an expansion by the length of the hairpin. Similarly, MMR may cut around the hairpin, remove it, then re-synthesize. Depending on which strand the hairpin was formed (either the template strand or repaired strand), this may result in a correct fix (no change in CAG length) or result in a contraction [10].

Previous studies have shown that genes involved in the DNA MMR system such as *MSH3* are modifier genes for HD and associate with age of onset [12], providing evidence that DNA MMR and somatic expansion drive disease progression.

2.3 Data Used In This Study

For this study, we used a new type of biological data collected by the McCarroll Lab at Harvard Medical School in collaboration with the Harvard Brain Tissue Resource Center at the McLean Hospital. Specifically, the lab devised a method to obtain CAG measurements at a single cell resolution simultaneously with their RNA expression levels, in tissue collected from post-mortem brain samples of HD patients. A transcriptome library is generated for each donor, and by barcoding individual cells, allows for accurate CAG length measurements and determining which cell type each measurement was from through its RNA expression data. This data revealed that somatic expansion was highly cell type specific; it occurred heavily in SPNs but not at all in many other neuronal cell types. The cell type specificity in the data was therefore crucial in allowing us to model somatic expansion in only the SPNs without having the data obfuscated by cell types that did not exhibit expansion. The lab was also able to estimate the fraction of cell death of SPNs by comparing the abundance of SPNs in each HD patient to the abundance seen in controls without the disease. This data was helpful in creating an accurate model accounting for cell death. To our knowledge, cell type specific CAG length measurements and SPN cell death measurements

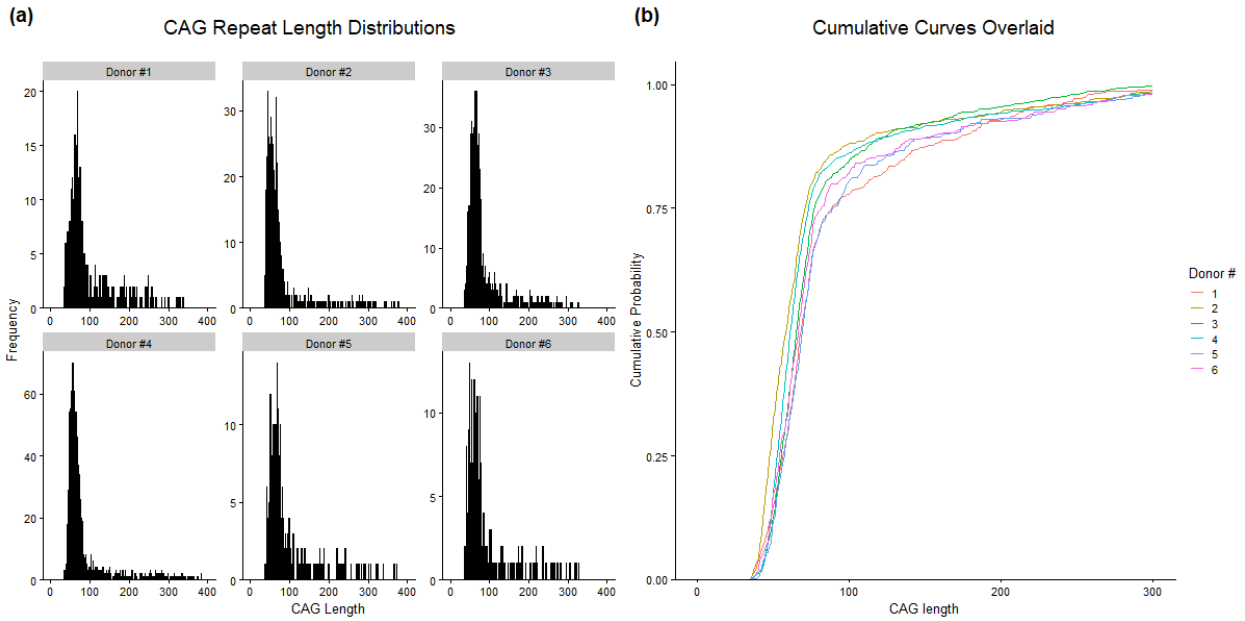

Figure 1: The plots show the CAG repeat length data of SPNs collected from post-mortem brain samples of 6 HD patients. (a) The CAG length distributions for each donor. (b) The empirical cumulative distribution functions (eCDFs) of the CAG lengths for each donor overlaid.

were not previously available, and thus provided us with opportunities for novel statistical models.

For this analysis, we will focus on six HD patients and the CAG length distributions of their SPNs. These patients specifically were chosen because they had high amounts of data and exhibited a noticeable degree of somatic expansion. Shown in Figure 1a and Figure 1b are the CAG length data of SPNs collected from these donors. Table 1 lists each donor's age, inherited CAG length, and fraction of SPN cell death.

2.4 Transcriptional Dysregulation in SPNs

From this data, the lab has also found evidence of cells beginning to exhibit transcriptional dysregulation in their RNA expression data after exceeding 180 CAGs, whereas there was no significant correlation between gene expression and CAG length before that threshold. This evidence suggests that neuronal cell death due to disease pathogenesis mostly occurs in cells with CAG lengths above 180, as transcriptional dysregulation interferes with the production of essential proteins and is likely an early step on the path towards cell death. Along with the data on the fraction of SPN cell death collected for each donor, these later allow us to account for cell death in our models for somatic expansion.

Donor #	Age (A)	Inherited CAG Length (I)	Cell Death Fraction
1	61	42	0.753
2	81	40	0.433
3	51	43	0.734
4	37	43	0.129
5	58	42	0.727
6	48	42	0.778

Table 1: The data used in this study was collected on 6 donors who had HD. Listed in the table for each donor are the age at which they died, the CAG length they inherited at birth, and the fraction of cell death measured for their SPNs.

3 Research Goals

The goal of this research was to create a statistical model for somatic CAG expansion that could generate CAG length distributions similar to the data collected from HD patients. Through these models, we hope to deepen our understanding of somatic expansion and its relationship to Huntington's Disease pathogenesis. We hope to gain insight into the biological mechanism and properties behind somatic expansion, which include:

1. The extent of the bias towards expansions over contractions.
2. How the rate of somatic expansion changes over time.
3. The relationship between somatic expansion and symptom onset.
4. The variability in somatic expansion among individual HD patients.

A deeper understanding of these properties of somatic expansion is important because it could inform the design of therapeutics that target the somatic expansion process, as well as inform the design of clinical trials.

4 Models For Somatic CAG Expansion

To model CAG repeat expansions, we adapted a previous mathematical model that was applied to another triplet repeat expansion disease, DM1 [13]. The DM1 model differs from our work in HD in that they modeled somatic expansion in blood cells, which likely exhibits different behavior than the neurons we modeled in HD. We similarly treat somatic expansion as a stochastic process consisting of both expansions and contractions, which happen at a rate that increases with each CAG length above a certain threshold, and do not happen at all below that threshold. Since we were modeling HD, we used a threshold of 36 CAGs where mutations only occur at or above that threshold, as we know people inheriting 36 or more repeats contract the disease, and we have seen empirically that shorter alleles are stable. We adopted a statistical model that involved simulating each cell through computation, as opposed to the DM1 model which mathematically

calculated the probability distributions. We chose statistical models as they were more flexible and allowed us to model more complex processes without requiring a closed form solution.

4.1 Base Model

The models we implemented simulate each individual cell's CAG length over the patient's lifetime. By simulating many cells independently, we can create a distribution of CAG lengths that can then be fitted to data from patients. To simulate an individual cell, the model assumes mutations will occur randomly at a certain rate. Since the rate of mutations is expected to increase for longer CAG repeats due to there being more sections of DNA subject to mutation, the model calculates the mutation rate as a function of CAG length. In the base model, or simplest model we implemented, this rate increased linearly for each CAG above the threshold. Then, each time a mutation occurs, there is a certain probability that it's an expansion, and otherwise a contraction. These mutations are simulated until the age of death is reached and the final CAG length is used. The base model fits two parameters for each donor:

1. r = rate parameter (used to calculate the mutation rate as a function of CAG length).
2. p = the probability a mutation is an expansion (otherwise a contraction).

We also fix two donor-specific parameters: A = age at death and I = inherited allele length of the donor we are trying to model.

Algorithm 1 describes the process for simulating an individual cells' CAG length over the patient's lifetime. At each step, T_{next} represents the time in years before the next mutation, which is drawn from a random exponential distribution with rate $r \cdot (X - 35)$. The rate is calculated in this way so that for each CAG repeat above 35 CAGs (the threshold above which somatic expansion occurs), there is an increased rate of mutation. Each mutation then has a probability p of increasing by 1 CAG and otherwise decreasing by 1. We then continue this process until the age of death is reached. Thus, each cell is modeled as a stochastic process consisting of both increases and decreases in CAG length.

We also tried models that changed by a variable number CAG repeats each time rather than always changing by one repeat. However, we found that models only changing by one repeat were often able to behave in a similar way to those other models. The previous DM1 model [13] also only considered single CAG repeats, as there has been biological evidence that mutations of single repeats are the most common in repetitive nucleotide sequences [14]. As a result, we decided to use models that change by one repeat each mutation for this study on HD.

4.2 Objective Function

By using the model described, we can simulate many cells to create a list M of CAG lengths, where M_i stores the CAG length of the i-th cell by the time of the patient's death. Similarly, let

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 until 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

the actual CAG distribution from the data of this patient be D, where D_i stores the CAG length of the i-th observation. To quantify how well the model fits the data, we decided to use the Kolmogorov-Smirnov (KS) test statistic. The KS test statistic is calculated as the largest difference between the empirical cumulative distribution functions of two discrete distributions. The statistic ranges between 0 and 1, with a lower value indicating a better fit. Formally, let $F_M(x)$ and $F_D(x)$ be the empirical cumulative distribution functions of M and D respectively. Then, the objective function is calculated as follows:

$$\text{Objective Function Score} = \sup_x(|F_M(x) - F_D(x)|),$$

where sup is the supremum function.

4.3 Fitting the Base Model

Our goal is to find the parameters r and p that generate a CAG length distribution with the least KS test statistic in comparison to the data. To do so, we performed a grid search, which consisted of iterating through an exhaustive list of parameter sets (r, p). Specifically, we iterated through values of r between 0 and 1 and values of p between 0.5 and 1. For each parameter set, we ran the model to generate CAG lengths for $1,000$ cells, and then calculated the KS test statistic between

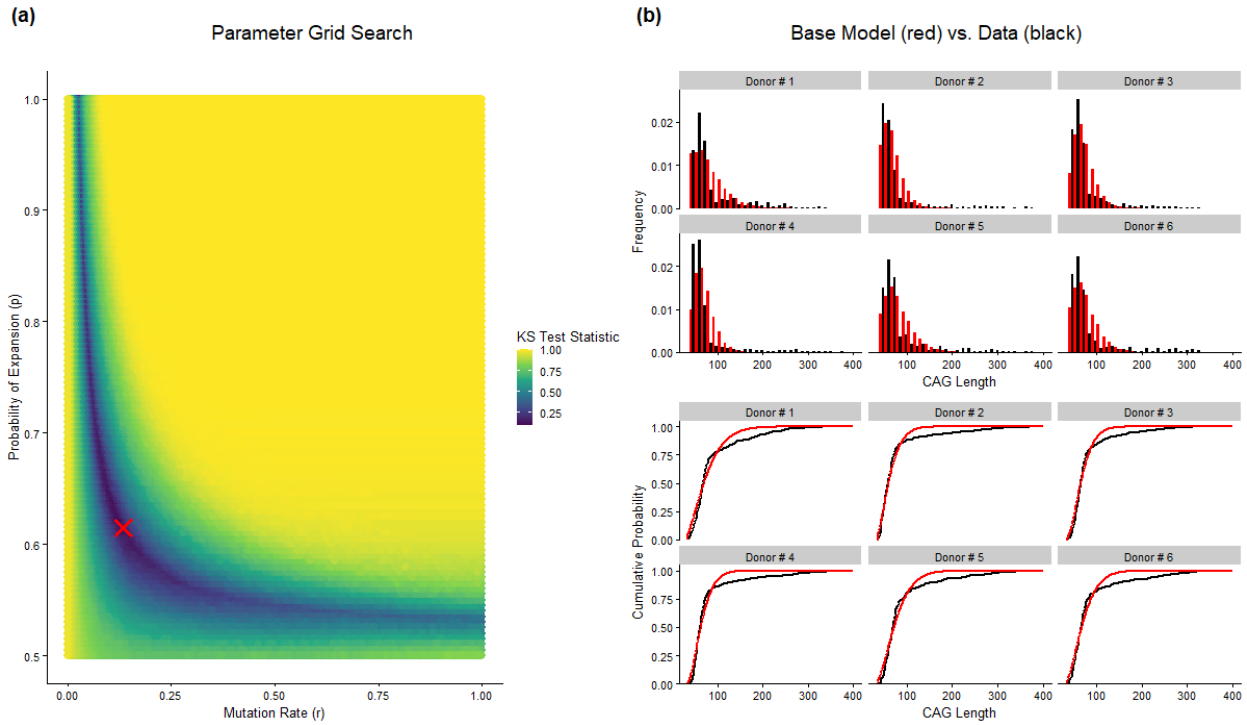

Figure 2: (a) A visualization of the grid search performed for one of the donors (Donor #5), with the two parameters of the model on each axis. Each point represents one parameter pair (r, p) and is colored by the KS test statistic between the distribution generated from the model and the data. The optimum parameters generating the best fit were $r = 0.615$ and $p = 0.135$, which are shown as the red X in the plot. (b) The grid search shown in (a) was repeated to find the optimum parameters for all six donors separately. The CAG distributions generated by the model using those optimum parameters (shown in red) are overlaid with the CAG distributions from the data (shown in black). The top six plots show the histograms and the bottom six plots show the empirical cumulative distribution functions (eCDFs) of the CAG lengths.

those generated lengths the CAG length data. This grid search is visualized in Figure 2*a*. We then ran the grid search separately for each of the six donors to determine the parameters that resulted in the best fit.

Because the model relies on a stochastic process, running the model multiple times can result in varying CAG length distributions. To mitigate this effect, we hoped to simulate a higher number of cells for each parameter set. To keep computation time reasonable while simulating more cells, we found the best fitting parameters in two passes. The initial pass is the grid search we detailed previously, where we only simulated $1,000$ cells for each set of parameters. In the second pass, we took the $5,000$ parameter sets with lowest KS scores from the first grid search and then simulated them again but with $100,000$ cells this time. In this manner, we were able to simulate many more cells for the sets of parameters that were close to being the optimum, and helped reduce computation time. This was applied to each of the six donors, and the optimum parameters for each donor was determined from the second pass. The CAG distributions generated by those best parameters are overlaid with the data they were fit to in Figure 2*b*.

From Figure 2*b*, we see that the model is generally able to fit the shape of the CAG distribution in the data for low CAG values. However, there is a consistent trend in the data of a long tail of CAG values starting at around 70 CAGs that the model fails to capture. This observation, which is consistent across all six donors, suggests that the base model is likely not adequate to explain the underlying biological mechanism.

4.4 Accounting for Cell Death

The previous model did not account for cell death, which plays a major role in the HD and should therefore be something we account for in our models. From Table 1, we see that the fraction of SPN cell death by the time of an HD patients death ranged from 13% to as high as 78% in these donors. Because cells that have died will on average have much higher CAG lengths, the data that we observe is not representative of the true CAG length distribution. As mentioned previously, the McCarroll Lab has found evidence that cells that have died due to disease pathogenesis are likely above 180 CAGs. As a result, high CAG lengths are almost definitely underrepresented in our data, as one could imagine there being a large fraction of cells that have reached 180+ CAGs, but because they have died are not observed in the data.

Suppose the fraction of SPN cell death in a donor was F_{death}, and that we observed N SPNs in the data. Then we can say that there are roughly $N_{dead} = N \cdot \frac{F_{death}}{1-F_{death}}$ cells that have died and thus not observed. If we assume that all of the dead cells have above 180 CAGs, then our previous model is not even close to being accurate. Since F_{death} is often 0.7 or higher, the model should then get more than 70% of the cells above that 180 CAG threshold to be consistent with cell death.

To account for these cells that have died, we modified the objective function to consist of 2 parts: the fit of the CAG distributions with cells strictly less than 180 CAGs in the model and data, and the difference in fraction of cells ≥ 180 in the model and data (after including dead cells). We decided to use this approach as opposed to adding parameters in the model to simulate cell death because we wanted to minimize assumptions about the CAG lengths of the dead cells. The only assumption we are making is that a vast majority of the cells that have died are above 180 CAGs.

More formally, let

1. D_{180} = list of CAG lengths strictly less than 180 CAGs from the data.
2. $F_{D180}(x)$ = empirical cumulative distribution function of D_{180}.
3. M_{180} = list of CAG lengths strictly less than 180 CAGs generated from the model.
4. $F_{M180}(x)$ = empirical cumulative distribution function of M_{180}.
5. f_D = fraction of cells above 180 CAGs in the data (including unobserved dead cells). It is calculated by adding N_{dead} to the count of observed cells above 180 CAGs, and then

dividing by $N + N_{dead}$.

6. f_M = fraction of cells above 180 CAGs in the model.

Then, the objective function quantifying how well the model fits the data is as follows:

$$\text{Objective Function Score} = \sup_x(|F_{M180}(x) - F_{D180}(x)|) \cdot \frac{1}{2} + |f_D - f_M| \cdot \frac{1}{2}$$

We reran the base model using this improved objective function. Shown in Figure 3*a* are the best fits generated by the base model accounting for cell death, and Table 2 shows the fraction of cells above 180 CAGs generated by the base model in comparison to the data. We can see that the base model now completely fails to fit the data. In an attempt to get enough cells past 180 CAGs, the model must turn up the rate parameter very high. But due to its lack of flexibility, it ends up compromising the shape of the distribution. The model CAG distributions look almost like uniform distributions, with the exception of a spike at 35 CAGs, which is a result of stopping when a cell goes below 36 CAGs. Donor #4 is the only donor that avoided this issue, as the fraction of cell death was lower and the model did not have to increase the rate by much. We saw that without cell death, the base model did not fit well, and after including cell death, it fits even worse, further indicating that the base model is not an accurate model for somatic expansion.

4.5 Power Model

We tried multiple other models, some of which calculated the number of CAG repeats to change by as a function of repeat length rather than always changing by one repeat, and others that calculated the mutation rate as some non-linear function of CAG length. Out of these models, we found a model calculating the mutation rate as a power function of CAG length to be the most flexible while having relatively few parameters. We therefore decided to focus on this power model over the others. The model reflects the fact that power-law relations are often found in natural systems.

In the base model, the mutation rate was calculated by $r \cdot (X - 35)$, which indicates a linear relationship between the mutation rate and the number of CAGs above the threshold of 35. For the power model, we calculate the mutation rate as $r \cdot (X - 35)^k$. This adds an extra parameter k, which acts as the exponent and results in the mutation rate being a power function of the number of CAG repeats above 35. Otherwise, the rest of the model remains the same.

We then performed a grid search on triplets of parameters (r, p, k), and found the optimum parameters for each of the six donors using two passes as we did with the base model.

From Figure 3*b*, we can see that the power model fits the data more closely than the base model. The model generally found values of k (the exponent parameter) above two to fit best, as that allows a faster increase in rate that helps get more cells into that longer CAG range while

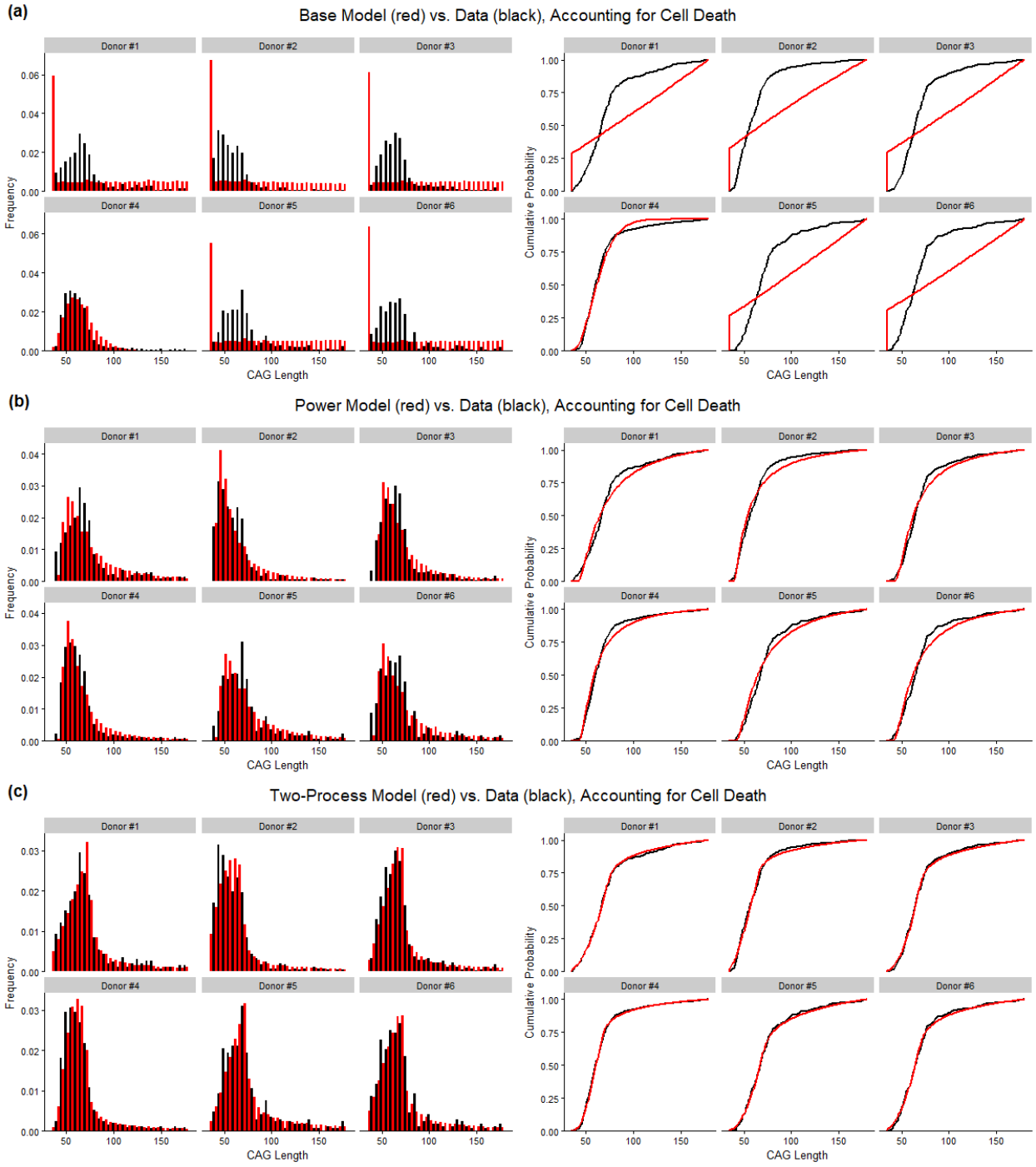

Figure 3: The CAG distributions generated by the models accounting for cell death overlaid with the patient data they were fit to. Only CAG lengths up to 180 CAGs were plotted because we suspect most cells that have died are above this threshold. The models were fit on the CAG lengths up to 180 CAGs, as well as the fraction of cells above that threshold (including the estimated fraction of dead cells in the data). See Table 2 for information on cell death estimation. (a) The model fits for the base model. The base model does not fit well, resulting in almost a uniform distribution with a spike at 35 CAGs for the cells that went below the threshold. (b) The model fits for the power model, which fits better than the base model but still consistently fails to capture the shape near 70 CAGs. (c) The model fits for the two-process model. The two-process model fits well and is able to capture features of the data the other models struggled to.

Donor #	Data	Base Model	Power Model	Two-Process Model
1	0.778	0.779	0.785	0.777
2	0.471	0.476	0.474	0.465
3	0.749	0.755	0.743	0.744
4	0.187	0.00	0.186	0.186
5	0.749	0.759	0.750	0.746
6	0.797	0.790	0.798	0.795

Table 2: Comparison of the fraction of cells above 180 CAGs in the observed data and in the three models. The data column lists for each donor the fraction of cells above 180 CAGs, which consists of both the observed cells and dead cells assumed to have CAG lengths above that threshold. The fraction of cells above 180 CAGs is also shown for the distributions generated by the three models (Base Model, Power Model, Two-Process Model). All of the models were generally successful in getting a similar amount of cells past 180 CAGs consistent with the data.

maintaining the distribution shape of lower CAG lengths. However, the model still does not accurately capture the CAG lengths at around 70 CAGs. In the data, we see a sudden drop in frequency at roughly 70 CAGs for all donors, but in the model we only see a gradual drop in frequency at the threshold. This is reflected in the eCDFs for all donors, as the model eCDFs are consistently below the data eCDFs at around 70 CAGs. Another discrepancy is that the peak of the histograms in the power model are consistently to the left of the peaks in the data. These consistent inaccuracies suggest that the power model may still not accurately represent the somatic expansion process.

4.6 Theory of Two Biological Processes

The previous models all fail to explain certain features in observed data, specifically the long tail of CAG lengths that seems to suddenly start at around 70 CAGs (see Figure 1*a* in Figure 1*b*). Specifically in Figure 1*b*, there is a sharp turn that happens in the eCDFs for all donors at around 70 CAGs, after which the cumulative curves begin to flatten out to create a long tail. This feature remains surprisingly consistent among donors despite them having quite a large variation in terms of age. This feature of the data suggests that there may actually be two biological processes driving CAG expansion. One of the processes would occur as expected, with a threshold of 36 CAGs and that causes the initial somatic instability. But the data suggests that there could be a second process with a threshold around 70 CAGs that causes expansions at a much faster rate. Such a mechanism could explain why there is such a distinct tail of CAG lengths starting at that point for all donors.

Though the biological mechanisms behind expansion are poorly understood, there is evidence supporting a two-process model in previous literature. Hairpins can form from polymerase slippage, and also whenever there is strand separation and re-annealing. Polymerase slippage generally results in smaller hairpins while strand separation/re-annealing can result in very large hairpins. Based on the two-process model, polymerase slippage could constitute the first process

with a threshold of 36 CAGs, while strand separation and re-annealing would constitute the second process with a threshold of roughly 70 CAGs. Further, there has been evidence suggesting an increase in helical tension at 64 CAGs not seen at 54 CAGs [15], which could boost the frequency of hairpin formations around this second process threshold we hypothesize.

4.7 Two-Process Model

These findings inspired us to try a model that consists of two processes. Because the mechanism for resolving hairpins is similar for both mechanisms (through DNA mismatch repair complexes), we kept the same probability of expansion for both processes. We instead added parameters for the rate and threshold for the second process. The model parameters are as follows:

1. r_1 = rate parameter for process 1
2. r_2 = rate parameter for process 2
3. p = probability of expansion for both processes
4. t = threshold for process 2

Process 2 is modeled in the same way as process 1, but has a differing rate parameter and an activation threshold of t CAGs rather than the fixed activation threshold of 35 CAGs set for process 1. These processes are assumed to occur independently. For this model we did not calculate mutation rate as a power function of excess CAG length; each process resembles the original base model. The algorithm for this model is shown in Algorithm 2.

In Algorithm 2, T_1 and T_2 refer to the time from the next mutation event from process 1 and 2 respectively. The minimum between these times is when the next mutation event will happen, after which it will be an expansion with probability p and contraction otherwise.

We performed a grid search on the set of parameters, now quadruplets of (r_1, r_2, p, t), and found the best fitting parameters. The fits with lowest objective function scores are shown in Figure 3*c*.

From Figure 3*c*, we see that the two-process model is able to match features of the data that the power and base model could not capture. It is able to generate that long tail of CAG lengths through the second process, as well as match closely with the data in the histogram, where the peaks match. For a more quantitative analysis on the fits, Figure 4 shows the objective function scores for each model over all donors. We see that the power model consistently outperforms the base model, and that the two-process model consistently outperforms both other models for all donors. The high performance of the two-process model is consistent with our hypothesis that there may be two biological processes underlying somatic expansion with differing rates and thresholds. However, this increased fit could also be affected by the fact that the two-process model has the most parameters out of all three models: four parameters as opposed to the power

Algorithm 2 Two-Process Model

function SimulateCellTwoProcessModel(r_1, r_2, p, t, 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
    $T_1 \sim Exp(r_1 \cdot (X - 35))$ ▷ years until a mutation from process 1
    $T_2 \sim Exp(r_2 \cdot \max(X - t, 0))$ ▷ years until a mutation from process 2
    $T \leftarrow T + \min(T_1, T_2)$ ▷ increment time to when the first mutation happens
    $u \sim U(0, 1)$
    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

model which has three parameters and the base model which has two parameters. This is one of the current limitations of our model. In the future, we plan on moving the model into a likelihood based framework that would allow more statistically grounded model comparisons through the Akaike Information Criterion (AIC).

4.8 Computational Performance

Running the grid search to fit the models was computationally intensive, as we wanted to search a wide range and fine grained set of parameters. For each set of parameters, we also wanted to simulate as many cells as possible to mitigate the effects of the randomness in the stochastic process. In addition to running multiple models for six donors, the computational costs became very large.

Originally, the models were coded in R, but we later moved all of the computationally heavy parts (the grid search and running the model) into Java, which gave nearly a 30x speed boost. We then used R to create the plots from the outputs generated by the model in Java. To further speed up computation, we parallelized the algorithms and utilized the high-performance compute cluster at the Broad Institute. When running each grid search, we broke the parameters sets to be

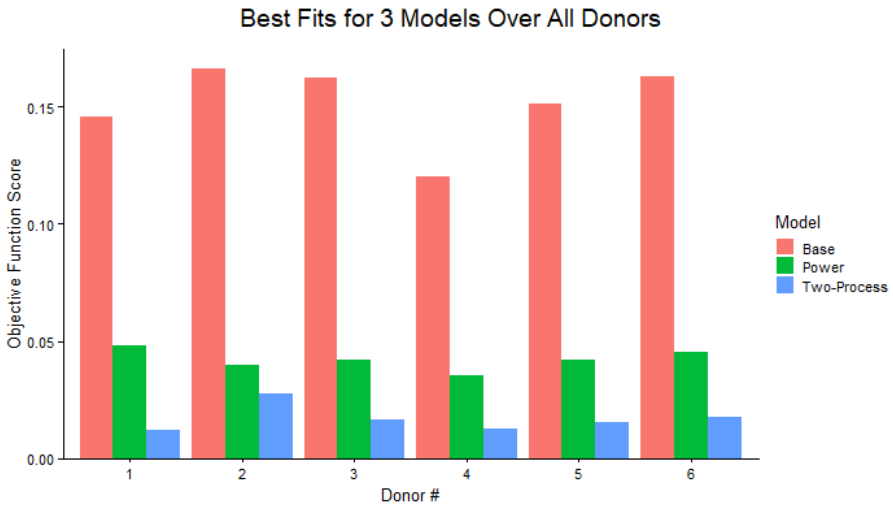

Figure 4: The lowest objective function for each model for each donor. The power model consistently scores better than the base model, and the two-process model consistently fits better than both the base and power model for all donors.

simulated into roughly 50 jobs and submitted them simultaneously to the computer farm. These combined efforts were able to speed up the computation. In the end, the power and two-process models took roughly $3-4$ hours to run, whereas the base model took less than 1 hour.

5 Applications of the Model

With the two-process model, which we believe is an accurate model for somatic expansion, we can now analyze the model behavior to provide insight into the somatic expansion process. The advantage of the model is that it provides a view of the CAG length distribution at any point in time and shows the dynamics of the somatic expansion process, whereas the laboratory data can only provide information at the time of the patient's death. The model can help us better understand the variability of somatic expansion across donors, the extent of bias towards expansions over contractions, the rate of somatic expansion over time, and when age of onset occurs. These insights could help enable therapeutics that target somatic expansion, and could also inform the design of clinical trials.

5.1 Parameter Variability Across Donors

We can use the model to analyze parameter differences between donors to gain insight into how somatic expansion may vary between people. We start by looking at the parameter values that generated the best fits for the two-process models in all donors. Shown in Figure 5a are the four different parameters: r_1, r_2, p, t. The middle of each vertical bar gives the parameter value that resulted in the best fit, whereas the upper and lower points of each bar give the highest and lowest parameter values that generated a fit with an objective function score within 0.01 of the

optimum. We see that the r_1 parameter was generally much lower than the r_2 parameter, indicating that the first process is slower. We also see that the p parameter is very stable, hovering around 0.68 for all donors, indicating that somatic instability consists of more expansions, but also a relatively high amount of contractions. The t parameter is also relatively consistent, generally varying around 70 CAGs, which is consistent with our original hypothesis that there may be a second process occurring at roughly 70 CAGs. This variability within parameters may reflect biological variation between donors; it's possible that donors with favorable genetic modifiers may have a lower r_1, for example. On the other hand, the consistency of the p parameter at around 0.68 may indicate that the degree of bias towards expansion is universal among HD patients. In the future, we hope to pair these parameters generated by the model with other information on the donors such as if they have genetic modifiers, to analyze if the parameter values may reflect those individual differences.

5.2 Dynamics of Somatic Expansion

Another piece of important information we can attempt to gather from the model is the speed at which cells expand in CAG length. From our analyses, the power model and two-process model fit far better than the base model, indicating that the rate of CAG expansion grows much faster than a linear process. To get a better sense of how many years it takes cells to increase through certain CAG length ranges, we analyzed individual cells from the best fitting two-process models on all donors. Figure 5*b* shows the average time in years it took cells to cross certain CAG ranges. We see that as the CAG lengths increase, the rate of somatic expansion increases at a superlinear rate. Whereas it took on average 38.5 years (averaged over all donors) for cells to go from their inherited CAG length to 70 CAGs, it only took on average 6.02 years to cross a larger range of 70 to 120 CAGs, and 2.38 years for cells to cross an even larger CAG range of 180 to 500. If we assume that a cell that has reached 180 CAGs has started to exhibit transcriptional dysregulation, and that a cell that has reached a length of 500+ CAGs is almost certainly dead, then the model indicates that it only takes around 2 years for a cell that has started to exhibit transcriptional dysregulation to die. This gives us more hope in therapeutics that target somatic expansion, as it indicates the possibility of slowing or stopping somatic expansion even in advanced stages of disease progression. At any point, only a small fraction of cells will be in a CAG length range high enough that might result in cell death to occur soon; so as long as we can slow down the somatic expansion for the cells in lower ranges (i.e. 40 to 70 CAGs) before they reach the hypothesized second process where expansion is much faster, the therapeutic could potentially protect a large majority of the surviving SPNs.

These findings of the model are also consistent with a previous study comparing young adults who had HD but had not reached age of symptom onset (preHD) with controls who did not have HD [16]. The study found that there was no significant evidence of cognitive or psychiatric

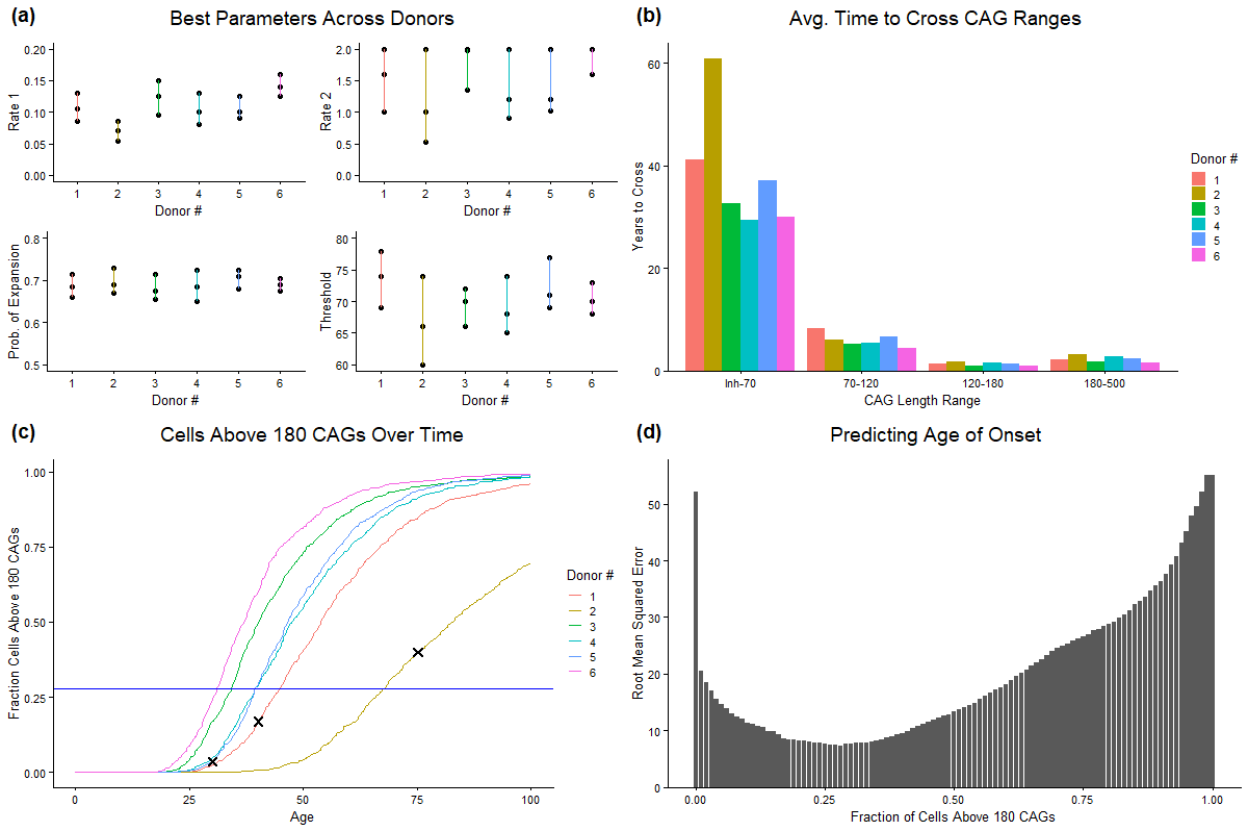

Figure 5: Inferences based on the two-process model. (a) The middle dot on each vertical bar shows the parameters that resulted in the best fit. The upper and lower points of each bar give the highest and lowest parameter values that generated a fit with an objective function score within 0.01 of the optimum. (b) Average time it took for a cell to cross each of the CAG ranges between donors. The first range Inh-70 shows the average time it took for a cell to go from its inherited length to 70 CAGs. (c) Curves show the fraction of cells above 180 CAGs over time, based on the CAG lengths generated by the model. Each X marks the actual age of neurological onset for the donor and the fraction of cells above 180 at that point based on the model. Age of onset data was only available for three of the six donors. The blue line shows the fraction that best predicts age of onset which was found in part (d). (d) Predicting age of onset by the fraction of cells above 180 CAGs. A fraction of 0.28 resulted in the lowest root mean squared error of 7.39 years when predicting age of onset.

impairment for the preHD participants, and that there were no differences in brain imaging measurements other than slightly smaller pumaten volumes for preHD patients. Our model is consistent with these results as it suggests that the rate of somatic expansion is very slow in the early stages of disease and that few if any cells have reached a toxic CAG length at that stage. The model indicates that preHD participants who had not reached the age of symptom onset would have relatively little somatic expansion, which could explain why there was no significant evidence of disease phenotypes.

5.3 Cell Death and Age of Onset

We can also use the model to attempt to relate somatic CAG expansion with the timing of symptom onset. Because we suspect the onset of symptoms generally begins when a certain amount of the SPNs have died, we would like to predict what that fraction of cell death is. In our model, since we assume that the fraction of cells above 180 CAGs is a proxy for cell-biological pathology and death, we can attempt to predict age of onset based on when the fraction of cells above 180 CAGs first reaches some threshold. Out of the six donors used in this study, we have data on the age of neurological symptom onset for three of the donors.

Figure 5*c* shows the fraction of cells above 180 CAGs over time for each donor. These curves were generated based on the two-process model that fit best for each donor, as the model provides a prediction of the CAG length distribution at any point in the patient's life. The three X's on the plot show the three data points we have on age of neurological onset. Donor #1 has an age of onset of 40, donor #2 has an age of onset of 75, and donor #5 has an age of onset of 30. The position of X's on the y-axis show the fraction of cells above 180 CAGs the model predicts them to have by that age.

To find a fraction of cells above 180 CAGs that best predicts age of onset, we performed a brute force search on values between 0 and 1. For each fraction, we found the earliest age at which the donor reached that fraction of cells above 180 CAGs and used that as the predicted age of onset. We then compared the predicted age of onset to the actual age of onset from the data, using root mean squared error (RMSE) as our metric. Figure 5*d* shows the RMSE for each fraction we tested, and we found that a fraction of 0.28 gave the lowest RMSE of 7.39 years. That best fraction of 0.28 is drawn in as the blue line in Figure 5*d*.

Because not all cells above 180 CAGs have died, when 0.28 of the cells are above 180 CAGs, a smaller fraction, closer to somewhere between 0.20 to 0.25 of cells might be dead. This gives us a general estimate that age of onset happens when roughly 20% to 25% of their SPNs have died.

This prediction was quite noisy, especially given that the lowest RMSE was 7.39 years. One source of noise comes from the age of onset data, as the age of onset is highly dependent on when the patient decides to get checked, and the true age of onset will always be earlier than when they are diagnosed. Further, for older patients, it may take even longer for them to realize they have symptoms as they could be attributed to aging. We also only had three data points on age of onset, which was very limited for making accurate predictions.

Despite the prediction being noisy, it is still helpful in determining a general range at which age of symptoms begins, and helps us understand the degree of somatic expansion that has occurred by that time. In the future, we will work to incorporate more data to make this prediction more robust.

6 Conclusion

The McCarroll Lab devised a method of collecting accurate CAG measurements from specific brain cell types, allowing us to analyze somatic expansion specifically in the SPNs, which has not been possible before. Through applying and evaluating different models for somatic expansion, we found that previous models that successfully modeled other diseases failed to explain the observed distribution of CAG repeat lengths of SPNs from HD patients. This new data also allowed us to incorporate estimates of cell death into the models, which is an important factor in HD progression. When cell death was included, the original model was even more incapable fitting the observed data, as it was not able to get enough cells to high CAG lengths while maintaining the distribution shape seen in the data. This led us to implement a power model, in which the rate of expansion increased as a power function of the CAG length. Though the power model fit far better, it still consistently failed to capture one feature of the data, which was the distinct long tail of CAG lengths that started at approximately 70 CAGs. We then implemented a model consisting of two processes, and found that it consistently fit better than both other single process models, as it was able to capture these distinctive features of the data.

These models help us understand a few things about somatic expansion. First, they suggest that there may be two biological processes driving somatic expansion, with different CAG length thresholds of 36 CAGs and roughly 70 CAGs. The threshold of 36 CAGs is the initial threshold for somatic instability which has been widely known, but we suspect that there may be a second threshold at around 70 CAGs after which somatic expansion occurs at a much faster rate. Second, these models estimate that roughly 68% of mutations are expansions and 32% are contractions, providing evidence that contractions do occur frequently, but just less than expansions. Third, these models indicate that the rate of somatic expansion increases as a superlinear function of CAG length, and that it may only take 2 years on average for a cell to go from exhibiting transcriptional dysregulation to death. Lastly, these models estimate that age of onset begins when roughly 28% of the cells are over 180 CAGs, which may translate to around $20 - 25\%$ cell death in the SPNs.

These findings may inform the design of therapeutics and help companies prioritize drugs that target the somatic expansion process. The model indicates that the average cell takes around 40 years to go from their inherited length to 70 CAGs, and that somatic expansion begins to speed up greatly after that threshold. This means that if a therapeutic could slow down the somatic expansion process of cells in those lower CAG ranges by even a small fraction, it could potentially prevent a large majority of cells from ever reaching the second threshold and expanding to pathogenically high CAG values within a normal person's lifespan. Further, the idea that cells only begin to die in a very short window of time suggests that cell death is not a slow process caused by the gradual buildup of toxins, but a sudden process only affecting a small group of cells

at any time. This is important because current candidate therapeutic approaches are based on reducing Huntingtin expression levels [17], but it may be that only a small fraction of SPNs are affected by such toxicity at any point in time. Our findings rather point to therapeutics that target somatic expansion as being potentially more effective, as such treatments could be used later in disease progression and still protect a majority of the surviving cells from expanding into toxic CAG lengths. Ultimately, we hope our findings may shift more focus onto developing therapeutics that target the somatic expansion process.

For future work, we hope to find biological evidence supporting the two-process model and the superlinear increase in rate of somatic expansion. There are other labs who are currently developing in-vitro experiments with DNA constructs that can monitor the expansion and contraction events. Future experiments using those systems could suggest biophysical processes consistent with our findings. We also hope to move our model into a likelihood framework that would be more statistically robust.

7 Acknowledgements

I would like to thank my mentors Bob Handsaker, Seva Kashin, and Steven McCarroll, for giving me guidance throughout the research process. I also thank Dr. John Warner from the CHDI Foundation for informing me on the framework behind existing models for somatic expansion. I thank Sabina Beretta at the Harvard Brain Tissue Resource Center and the McCarroll Lab for providing me the data for this study. I also thank all of the Huntington's Disease brain donors and their families, who are generous for providing the donations that were essential for this study. I also thank the MIT PRIMES program for providing me with this opportunity and matching me with the McCarroll Lab.

References

[1] La Spada AR, Paulson HL, Fischbeck KH. Trinucleotide repeat expansion in neurological disease. Ann Neurol. 1994 Dec;36(6):814-22. doi: 10.1002/ana.410360604. PMID: 7998766.

[2] MacDonald, Marcy E., et al. "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes." Cell 72.6 (1993): 971-983.

[3] Langbehn DR, Hayden MR, Paulsen JS; and the PREDICT-HD Investigators of the Huntington Study Group. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. Am J Med Genet B Neuropsychiatr Genet. 2010 Mar 5;153B(2):397-408. doi: 10.1002/ajmg.b.30992. PMID: 19548255; PMCID: PMC3048807.

[4] Kwa L, Larson D, Yeh C, Bega D. Influence of Age of Onset on Huntington's Disease Phenotype. Tremor Other Hyperkinet Mov (N Y). 2020 Jul 9;10:21. doi: 10.5334/tohm.536. PMID: 32775035; PMCID: PMC7394225.

[5] Monckton, Darren G. "The contribution of somatic expansion of the CAG repeat to symptomatic development in huntington's disease: A historical perspective." Journal of Huntington's Disease 10.1 (2021): 7-33.

[6] Laura Kennedy, Elizabeth Evans, Chiung-Mei Chen, Lyndsey Craven, Peter J. Detloff, Margaret Ennis, Peggy F. Shelbourne, Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis, Human Molecular Genetics, Volume 12, Issue 24, 15 December 2003, Pages 3359–3367, https://doi.org/10.1093/hmg/ddg352

[7] Nalavade, R., et al. "Mechanisms of RNA-induced toxicity in CAG repeat disorders." Cell death & disease 4.8 (2013): e752-e752.

[8] Bano, D., et al. "Neurodegenerative processes in Huntington's disease." Cell death & disease 2.11 (2011): e228-e228.

[9] Hou, Caixia, et al. "Incision-dependent and error-free repair of (CAG) n/(CTG) n hairpins in human cell extracts." Nature structural & molecular biology 16.8 (2009): 869-875.

[10] Iyer, Ravi R., and Anna Pluciennik. "DNA mismatch repair and its role in Huntington's disease." Journal of Huntington's Disease 10.1 (2021): 75-94.

[11] Petruska, John, Michael J. Hartenstine, and Myron F. Goodman. "Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease." Journal of Biological Chemistry 273.9 (1998): 5204-5210.

[12] Lee, Jong-Min, et al. "CAG repeat not polyglutamine length determines timing of Huntington's disease onset." Cell 178.4 (2019): 887-900.

[13] Higham, Catherine F., et al. "High levels of somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-frequent expansion and contraction mutations." Human Molecular Genetics 21.11 (2012): 2450-2463.

[14] Xu, Xin, et al. "The direction of microsatellite mutations is dependent upon allele length." Nature genetics 24.4 (2000): 396-399.

[15] Bacolla, Albino, et al. "Flexible DNA: genetically unstable CTG· CAG and CGG· CCG from human hereditary neuromuscular disease genes." Journal of Biological Chemistry 272.27 (1997): 16783-16792.

[16] Scahill, Rachael I., et al. "Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington's disease Young Adult Study (HD-YAS): a cross-sectional analysis." The Lancet Neurology 19.6 (2020): 502-512.

[17] Tabrizi, Sarah J., Rhia Ghosh, and Blair R. Leavitt. "Huntingtin lowering strategies for disease modification in Huntington's disease." Neuron 101.5 (2019): 801-819.