Bioinformatics Advance Access published August 16, 2016

Systems Biology

Selective Mutation Accumulation: A Computational Model of the Paternal Age Effect.

Eoin C Whelan[1,*], Alexander C Nwala[2], Christopher Osgood[1] and Stephan Olariu[2]

[1]Department of Biology, Old Dominion University, Norfolk, VA, USA, [2]Department of Computer Science, Old Dominion University, Norfolk, VA, USA.

*To whom correspondence should be addressed.

Associate Editor: Prof. Alfonso Valencia

Abstract

Motivation: As the mean age of parenthood grows, the effect of parental age on genetic disease and child health becomes ever more important. A number of autosomal dominant disorders show a dramatic paternal age effect due to selfish mutations: substitutions that grant spermatogonial stem cells a selective advantage in the testes of the father, but have a deleterious effect in offspring. In this paper we present a computational technique to model the spermatogonial stem cell niche in order to examine the phenomenon and draw conclusions across different genes and disorders.

Results: We used a Markov chain to model the probabilities of mutation and positive selection with cell divisions. The model was fitted to available data on disease incidence and also mutation assays of sperm donors. Strength of selective advantage is presented for a range of disorders including Apert's syndrome and achondroplasia. Incidence of the diseases was predicted closely for most disorders and was heavily influenced by the site-specific mutation rate and the number of mutable alleles. The model also successfully predicted a stronger selective advantage for more strongly activating gain-of-function mutations within the same gene. Both positive selection and the rate of copy-error mutations are important in adequately explaining the paternal age effect.

Availability: C++/R source codes and documentation including compilation instructions are available under GNU license at https://github.com/anwala/NicheSimulation.

Contact: ewhel001@odu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As the average age of parenthood becomes more delayed, understanding the effect of parental age on child health becomes more important. (Bray *et al.* 2006). The effect of maternal age has long been acknowledged (Hook 1981), but in recent years the effect of paternal age has been the subject of a great deal of study. Paternal age has been linked to a wide range of traits and diseases, such as spontaneous occurrences of mutations that cause dominant disorders and X-linked diseases (Vogel 1975, Risch *et al*. 1987, Glaser & Jabs 2004). Congenital defects, cancer predisposition disorders, schizophrenia, bipolar disorder, autism and Alzheimer's disease have also been linked to father's age (reviewed in Paul and Robaire 2013).

Due to the larger number of male germline cell divisions compared with the female germline, males produce 3-6 times as many mutations than females throughout evolution (Li *et al.* 1996). Sperm are produced by a continually-dividing population of stem cells and each division represents a chance for replication errors to happen. Penrose (1955) first proposed that replication errors provided an explanation for the observed incidence of genetic diseases with paternal age. In many cases the influence of paternal age is relatively subtle compared with the large scale chromosomal abnormalities characteristic of maternal age effect because point mutations typically have small or no effect on phenotype. However certain substitutions can have devastating effect on those who carry the allele.

Diseases that show a strong paternal age effect, however, are not explained purely by Penrose's copy-error hypothesis and show an exponential increase in incidence with father's age. The mutations responsible

typically display a very specific spectrum of missense substitutions, accumulating faster than the raw mutation rate can account for. Substitutions associated with these disorders present in clumps, indicating a positive selective mechanism of mutation accumulation, as opposed to a high mutation rate or "hot spot" model. Such evidence has so far presented for achondroplasia (Shinde *et al.* 2013), Apert's syndrome (Qin *et al.* 2007, Choi *et al.* 2008), Costello syndrome (Ginnoulatou *et al.* 2013) and Noonan Syndrome (Yoon *et al.* 2013). There is a parallel with the intestinal crypt where mutant cells colonize their niche through selective advantage conferred by their new phenotype (Bozic & Nowak 2013). It is also relevant to cancer etiology as paternal age effect mutations are typically found in tumors (Maher *et al.* 2014).

1.1 The Spermatogonial Stem Cell Niche

Spermatogonial stem cells (SSCs) reside on the basal lamina on the outer edge of the seminiferous tubules within the testes. Spermatogonia are surrounded by much larger Sertoli cells that form the microenvironment for the cells. The spermatogonia divide in cyclical waves and progeny of the stem cells migrate as they divide and differentiate towards the hollow center of the seminiferous tubule (de Rooij & Russel 2000). In humans the active SSCs are comprised of type A_{pale} spermatogonia.

Certain stem cell systems like the colonic crypt have specific arrangements of cells with strictly limited numbers of stem cells (Humphries & Wright 2008). The SSC niche on the other hand lacks obvious repeating structures. However, SSCs are observed to localize to specific areas of the seminiferous epithelium (Yoshida 2008). While spermatogonia can repopulate whole seminiferous tubules that have been depleted by radiation (Shinohara *et al.* 2001), studies of live imaging of stem cells indicate limited migrational capabilities (Klein *et al.* 2010). Additionally, cells that migrate away from the niche are likely to differentiate (de Rooij & van Beek 2013), likely because GDNF distribution is patch-like (Sato *et al.* 2011), creating effective niche limits.

1.2 Motivation & Predictions

The objective of this model was to simulate the accumulation of mutations through positive selection. Several groups have made simulations of mutation accumulation that can present in the signature arrangement but have been limited to FGFR2 mutations causing Apert's syndrome (Yoon *et al.* 2009, Choi *et al.* 2008, Qin *et al.* 2007).

Our primary aim was to estimate the r value for a range of disease causing mutations. Currently the only estimate for an r value comes from Yoon and colleagues, who estimated it to be 0.014 for Apert's syndrome-causing mutations (Yoon *et al.* 2009). We aim to provide estimates for a range of different disease causing loci. We hypothesized that diseases with a more exponential increase with paternal age would have larger r values. We also aimed to determine if differences in incidence rate between alleles of a single gene or between mutations that cause different disorders is due to the underlying mutation rate or to the selective advantage of the particular alleles. Additionally, where different mutations affect the same gene, if there is a stronger activating mutation, we expected a higher r value.

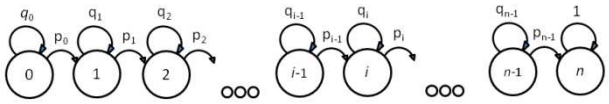
The SSC niche has been the subject of some computer simulations of the normal stem cell niche (de Rooij & van Beek 2013, Ray *et al.* 2014) which model the normal homeostasis of the spermatogonial stem cell niche in terms of spatial arrangement of cell and the cycling of the seminiferous epithelium. Mutation accumulation by positive selection for Apert syndrome has also been simulated looking purely at sequence data. Yoon and colleagues were able to simulate mutation accumulation within the niche to match their mutational data quite closely. In contrast to the simulation presented in the previous work for a single syndrome, in this paper we present a mathematical model that can be used to provide a deeper understanding of mutation accumulation across different genes and disorders.

2 Methods

2.1 Model

We assume that the stem cell niche represents a closed system, with n cells contained within. Of the n cells, each can be in one of two states; mutant or wildtype. With each cell division, a stem cell acquires the mutation with probability p. Stem cells are assumed to divide asymmetrically, one daughter differentiating and ultimately being lost and the other remaining in the niche. With normal divisions, therefore, the number of stem cells will not change regardless of divisions. Mutant cells are different from wildtype in that they have a positive selective probability of r with every cell division, much higher than the probability of mutation. Should a selection event happen, a mutant cell will divide symmetrically to produce two mutant cells. This will increase the niche size above the maximum. To correct this, a random cell is ejected from the niche and lost (all cells are eligible for ejection, including the newly-formed mutant cell).

Figure 1. Representation of the probabilities associated with a niche of n stem cells.



The niche, represented by circles, starts with 0 mutant cells out of n. After a stem cell divides in the niche, the probability p_0 indicates the chance that a mutation occurs and brings the system into state 1 (i.e. one mutant cell in the niche) and q_0 that after the cell division the niche remains in state 0. Once in state 1, the probability of advancing to state 2 has changed to p_1 as the chance of mutation remains but the mutant cells may expand with a selective event. Correspondingly, the chance that a cell division occurs that maintains the niche in state 1 is now q_1. This continues until, ultimately, all of the n cells are mutant, at which point the system remains in state n.

We have modeled the system as a Markov chain (figure 1). The chain has n+1 states, where state 0 represents the niche comprised entirely of wildtype cells and in the final state, n, the cells are entirely mutant. Each niche consists entirely of wildtype cells to begin with and positive selection (r) can only occur once a mutation has first happened to one of the cells. Cells are selected at random from the niche to divide sequentially. State i represents the niche with i mutant cells and n-i wildtype cells. In state i there is within the niche a probability p_i that after a cell division the number of mutant cells will increase by one, and a probability q_i that they will remain the same. The probability of a mutation and a subsequent reversal at the same site is sufficiently small as to be ignored.

Let us imagine that the niche is in state i and a random cell is selected to divide. If the cell selected is wildtype, then with probability p it is transformed into a mutant and returned to the niche, otherwise it is returned as a wildtype. If a mutant cell is selected, it is simply returned to the niche unless a selection event happens with probability r, in which case two mutant cells are returned and subsequently one random cell is lost (all cells including the returned cells are eligible to be lost).

In order to calculate q_i, the probability that after a cell division the system remains in state i, there are therefore three mutually exclusive possibilities with the following probabilities:

(1) A wildtype cell is selected for division, but no mutation occurs:

$$\frac{n-i}{n}(1-p) \tag{2.1}$$

(2) A mutant cell is selected for division and no selection event happens:

$$\frac{i}{n}(1-r) \tag{2.2}$$

(3) A mutant cell divides, a selection event happens and a mutant cell is lost from the niche.

$$\frac{i}{n}r\frac{i+1}{n+1} \tag{2.3}$$

The combined probability q_i is therefore:

$$q_i = \frac{n-i}{n}(1-p)+\frac{i}{n}(1-r)+\frac{i}{n}r\frac{i+1}{n+1} \tag{2.4}$$

With some rearrangement:

$$q_i = 1-p+\frac{i}{n}\left[p-\frac{r(n-i)}{n+1}\right] \tag{2.5}$$

Since $p_i = 1 - q_i$, this can be rewritten as:

$$p_i = \frac{n-i}{n}\left[p+\frac{ir}{n+1}\right] \tag{2.6}$$

The Markov chain produces a matrix, T, with dimensions of $(n+1) \times (n+1)$, where rows indicate the starting state and column denote the final state (i.e. the initial and final number of mutant cells).

$$T = \begin{bmatrix} q_0 & p_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & q_1 & p_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & q_2 & p_2 & \cdots & 0 & 0 \\ 0 & 0 & 0 & q_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & q_{n-1} & p_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{2.7}$$

This matrix gives the probability of moving from one state to another in one step i.e. one cell division within the stem cell niche. The element T_{ij} gives the probability of starting in state i (i.e. a niche containing i mutant cells) and after one cell division in the niche ending in state j with j mutant cells. In order to model the progression of multiple cell divisions within the niche, the matrix can simply be raised to the power of the number of cell divisions. T^2 will give a matrix that provides probabilities for two steps (i.e. two cell divisions within the niche) and T^K will provide probabilities for traversing in K steps. In the final matrix, therefore, the element T^K_{ij} represents the probability that a niche starting with i mutant stem cells will end with j mutant stem cells after K cell divisions. Note that K is total cell divisions occurring amongst any of the cells in the niche, not the average number of divisions per cell, which would be K/n.

For an individual, the number of steps required, K, is a factor of the rate of cell division d (divisions per year per cell), the age of the individual a in years and the number of cells per niche n.

$$K = nda \tag{2.8}$$

The matrix T^K was for any given age solved computationally, by generating a matrix T and calculating values for each cell and then raising the matrix to the power K. Since the model assumes every individual starts with 0 mutant cells, the only relevant part of the final solved matrix is the top row.

$$\begin{matrix} State & 0 & 1 & 2 & 3 & \dots & n-1 & n \\ P_S & \Big[\ P_0 & P_1 & P_2 & P_3 & \dots & P_{n-1} & P_n\ \Big] \end{matrix} \tag{2.9}$$

For a given state s, the value P_s denotes the probability that starting with 0 mutant cells, after K cell divisions, the niche will have s mutant cells. The average number of mutants per niche (assuming sufficient number of replications), M, can be calculated by summing all of the final state values multiplied by the corresponding probabilities.

$$M = \sum_{s=0}^{n} sP_s \tag{2.10}$$

M/n gives us a single proportion of mutant to wildtype cells at a given age a. Within an individual person many niches will deviate dramatically from the average value, but since the number of niches within a single individual can be assumed to be high (see section 2.3), but all contribute sperm equally, we can assume that the overall proportion of mutant sperm to wildtype sperm as a whole will tend towards to M/n.

This can be proven as follows. There are N niches each with n stem cells. Each niche has a different number of mutant cells, u_1, u_2, etc. Each stem cell contributes an equal number of sperm, for simplicity we assume one sperm per stem cell but the following is true for any value of sperm produced per stem cell division as the proportion will remain constant.

The total number of mutant cells U is:

$$U = \sum_{i=0}^{N} u_i \tag{2.11}$$

Dividing by the total number of stem cells over all the niches (Nn) gives the proportion of mutant stem cells over total cells, which is equivalent to the mean proportion of mutant cells per niche:

$$\frac{U}{Nn} = \frac{\overline{X}_U}{n} \tag{2.12}$$

Where $\overline{X}_U$ is the mean number of mutant cells per niche for a given number of niches N. Taken to the limit:

$$\lim_{N\to\infty}\left(\frac{\overline{X}_U}{n}\right) = \frac{M}{n} \tag{2.13}$$

Therefore using the ideal average value M/n from a single niche is an accurate measure of mutant sperm proportion for the entire individual providing N is large and all niches contribute sperm equally.

2.2 Confirmation of Model Design by Simulation

To test the mathematical model, we designed a simulation alongside it to emulate the progression of a single stem cell niche. Early versions of our simulation were much more complex and attempted to simulate all the SSCs within an individual, but simulation of a single niche is sufficient to test the model, particularly averaged over a large number of repeats

(see equation 2.13). Simulation of the stem cell niche was designed in C++. Script design simulated a stem cell population at a niche level independent of the Markov chain model (supplementary algorithm 1).

The simulation progressed by selecting a random cell. If the selected cell was wildtype it became mutant with probability p and if the selected cell was mutant it underwent a selective event with probability r. Selective events represented a symmetric division and added a mutant cell to the niche. As a consequence a random cell was then ejected and lost from the niche. This process was repeated and the simulation was allowed to run for a specific number of cell divisions sufficient to represent a human reproductive lifespan ($n \times d \times 80$ years). The results from each run were then averaged over a hundred thousand repeats of the simulation.

The simulation was tested against the mathematical expression by applying equation 2.10 with the same parameters and age values as the simulation. The simulation tended towards the values provided by the model and showed perfect agreement with sufficient replication (supplementary figure 1).

2.3 Parameters

Mutation probability, *p*. Rahbari *et al.* (2016) estimated the mutation frequency per nucleotide per germline cell division at 4×10^{-11} calculated by sequencing multi-sibling families. This mutation rate is close to that calculated with phylogenetic data (Lynch 2010) and point mutations on the Y-chromosome (Helgason *et al.* 2015). Rahabari and colleagues also noted little variation of mutation rate with paternal age, which allowed us to assume p is a constant value regardless of age. This is the baseline mutation probability per site, before accounting for elevated mutational frequency due to CpG sites or multiple disease-causing alleles within a single gene.

Stem cells per niche, *n*. SSCs are not organized in regular repeating structures with defined cell numbers like the colonic crypt (Humphries & Wright, 2008) and this makes estimating the niche size difficult. While initial models of the SSC niche did not have discrete compartmentalization, recent research has shown preferential clustering of SSCs to specific regions of the seminiferous basal lamina (de Rooij & Griswold 2012, Yoshida *et al.* 2007). To estimate the number of SSCs per niche, we turned to studies in mice, where spermatogenesis has been reconstituted in sterile mice by transplantation of SSCs bearing a reporter gene. From Shinohara *et al.* (2001), adult mice generated at least 108 colonies per testis. Given 35,000 stem cells per mouse testis (Tegelenbosch & de Rooij DG 1993), this amounts to 324 stem cells per niche. Scaling up to human testes by weight, assuming the same number of stem cells per niche, gives us approximately 3 million individual niches, which fulfills the requirement of equation 2.13. The model also assumes the number of SSCs remains constant throughout life. In reality, the number of stem cells declines with age (Paul & Robaire 2013). Assuming attrition occurs evenly among mutant and wildtype stem cells, this stochastic loss will not affect the proportion of mutant to wildtype sperm over many niches. The caveat to this is that it is possible that the mutant cells, rather than a proliferative advantage, are granted some form of resistance to the age-based attrition. Finally, as hypothesized by Yoon and colleagues (Yoon *et al.* 2009), non-proliferating A_{dark} spermatogonia may activate as reserve stem cells and replace losses (including lost mutant cells) with wildtype cells. This would cause a "dip" in the mutation frequency, irrespective of gene as fresh wildtype cells are introduced into the system. This would be informative to model across disorders but is beyond the scope of this paper.

Stem cell divisions per year, *d*. Human spermatogenesis results in one stem cell divisions per spermatogenic cycle of the A_{pale} spermatogonia, so once every 16 days (de Rooij & Russel, 2000), although lower estimates exist (Tomasetti & Vogelstein, 2015). The model selects cells to divide randomly rather than in waves, however the odds of the same cell being selected repeatedly is low and the results are averaged over a large number of niches, the effect of this is negligible and allows us to avoid tracking individual cells.

Selection pressure, *r*. This is the probability that when a pre-existing mutant cell divides it will do so symmetrically and self-renew. The model fits a best-fit curve for the optimal r value to fit the data. Note that in normal steady-state division, SSCs may divide symmetrically into A_{paired} spermatogonia where both daughter cells remain stem cells or both differentiate (de Rooij & Griswold 2012) for which there is some evidence (Klein *et al.* 2010). For simplicity we have assumed, as earlier models have done (Yoon *et al.* 2009), that each stem cell divides asymmetrically in normal homeostatic cell division rather than a balance of differentiating and self-renewing divisions.

2.4 Fitting the Model to Mutation Data

In order to match our model to existing paternal age effect data, we started with birth incidence of various genetic diseases. The larger number of younger parents versus older ones is accounted for by looking at Observed/Expected values, the number of births for a given age category divided by the expected number of births assuming the total number of disease-affected children were distributed to each age category proportional to births in that population.

Using census data from 1966 USA birth data (Vital Statistics of the United States, 1966, U.S. Department of Health, Education and Welfare) to estimate a number of births per age category, C_a, (as per Risch *et al.* 1987), we used the M/n fraction of mutant-to-wildtype sperm for the given age category to produce a number of disease-affected births for that category, simulated mutant births (S).

$$S = \frac{M}{n} C_a \tag{3.1}$$

By dividing the number of fathers in the age category by the total population and then multiplying this proportion by the total number of simulated mutant births, we can calculate the predicted mutant births (P) assuming simulated mutant births are distributed proportional to the paternal age distribution.

$$P = \frac{C_a}{C_t} \sum S \tag{3.2}$$

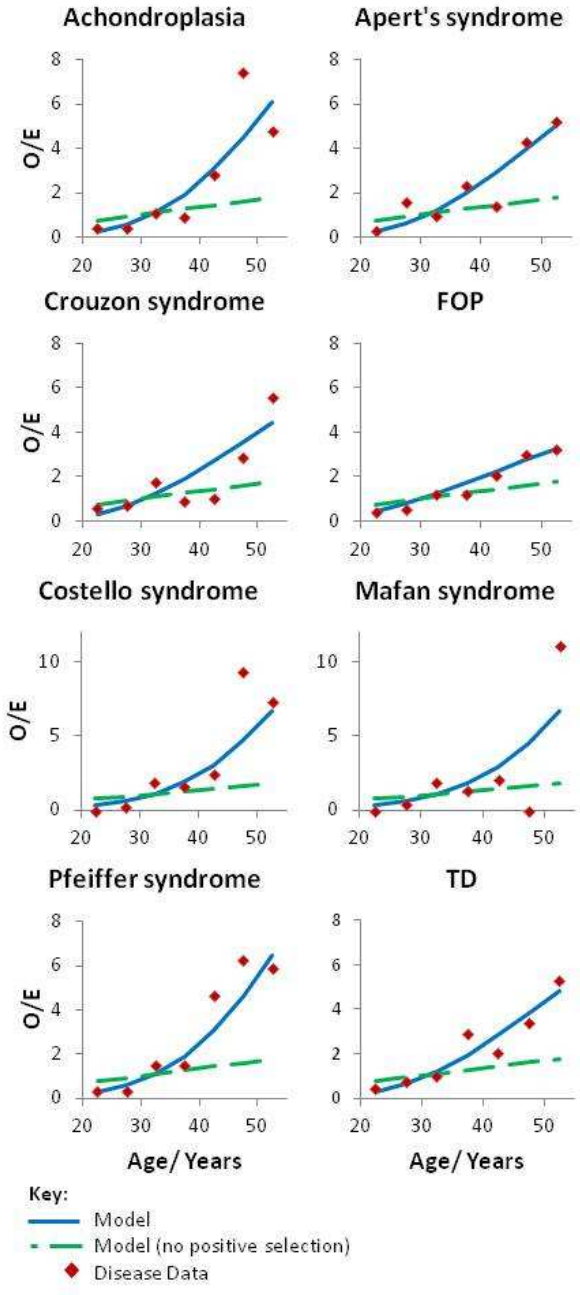
Simulated/Predicted is therefore directly comparable to the Observed/Expected data. In order to fit the S/P data to the existing O/E values, the strength of selection, r, had to be empirically determined. A script was generated in R (R Development Core Team, 2008, URL http://www.R-project.org.) that would match the O/E values according to the following algorithm:

(1) For a given value of r, calculate a single S/P value for the median age of the following age categories: 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54. In each case the average age for the category was used to generate the S/P value.

(2) Calculate the difference between each S/P value and the corresponding O/E value. Take the sum of squares (SOS) of these differences. .

(3) The process was optimized for *r* using R's *optim* function by searching for the *r* value with the lowest SOS score.

Figure 2. Simulation of observed/expected birth numbers.



FOP – fibrodysplasia ossificans progressiva, TD – thanatophoric dysplasia.

O/E disease data from Risch *et al.* (1987), except for Costello syndrome (O/E ratios calculated from Lurie 1994), achondroplasia and thanatophoric dysplasia from Orioli *et al.* (1995).

We also compared our model to high-throughput sequencing data. In this case, proportional numbers of mutant sperm were compared directly to the model's predicted proportion of mutant stem cells to wildtype cells. The same procedure as above was used, except instead of SOS of O/E-S/P, the number of mutant cells per 10^6 cells was used directly and the SOS between experimental and calculated number of mutants was generated.

3 Results

Figure 2 shows the matched model and disease data graphs for 8 disorders that show a strong paternal age effect. Excluding FOP as an outlier (see discussion), the remaining predicted incidence values correlated significantly with the actual incidence values (Pearson's correlation coefficient = 0.91, $p<0.05$). With those disorders where sequence data is available (Costello, Apert and thanatophoric dysplasia syndromes), *r* values can be compared directly between disease data and sperm mutation rates and show close agreement (TOST equivalence test, $\varepsilon=0.0053$, $p<0.05$). The probabilities of positive selection varied from 0.5% to 1.5% with a mean *r* value of 0.0083 for *r* values from birth data and 0.0094 from sequence data.

Table 1. Strength of positive selection (*r*) for 8 diseases

Disease	Gene	*r* value (sequencing data)[a]	*r* value (birth data)[b]
Achondroplasia	FGFR3	-[c]	0.00741
Apert's syndrome	FGFR2	(C755G) 0.0124 (C758G) 0.0126	0.00888
Costello syndrome	HRAS	(G34A) 0.00526	0.00606
Crouzon syndrome	FGFR2	-	0.00997
FOP	ACVR1	-	0.0135
Marfan syndrome	FBN1	-	0.00517
Pfeiffer syndrome	FGFR2	-	0.00668
TD	FGFR3	(A1948G) 0.0105	0.00937

FOP – fibrodysplasia ossificans progressiva, TD – thanatophoric dysplasia.

[a]Calculated by directly matching mutation incidence to that from sperm DNA sequencing (see figure 3), specific mutation is shown in parentheses.

[b]Calculated from birth incidence rates by making the best fit of O/E curves, with adjusted mutation rates (see figure 2).

[c]Several studies have estimated the mutation rate of achondroplasia in sperm but have been omitted due to concerns of the methodology (Maher *et al.* 2014).

The predicted incidence rate is shown in table 2. The raw incidence assumes the baseline mutation rate *p* of 4×10^{-11}. However, mutation rate varies by location in the genome and the sequence in question and of particular interest are CpG sites. These are particularly mutable as cytosine in CpG sites is a methylation site and can spontaneously deaminate to thymine (Lynch 2010). In order to account for the increased mutation chance, the probability of CpG-specific alleles mutating was multiplied by a factor of 15 for a transition or 5 for a transversion (Nachman & Crowell 2000). Additionally, a number of paternal age effect disorders are caused by multiple mutations at a variety of loci and by looking at disease incidence of the disorder as a metric, we include all mutations that contribute to that disease phenotype. In order to simulate this in the model, the aggregate probability of mutation, p_a, is the probability of any

of the mutations occurring that give rise to the phenotype, with the formula:

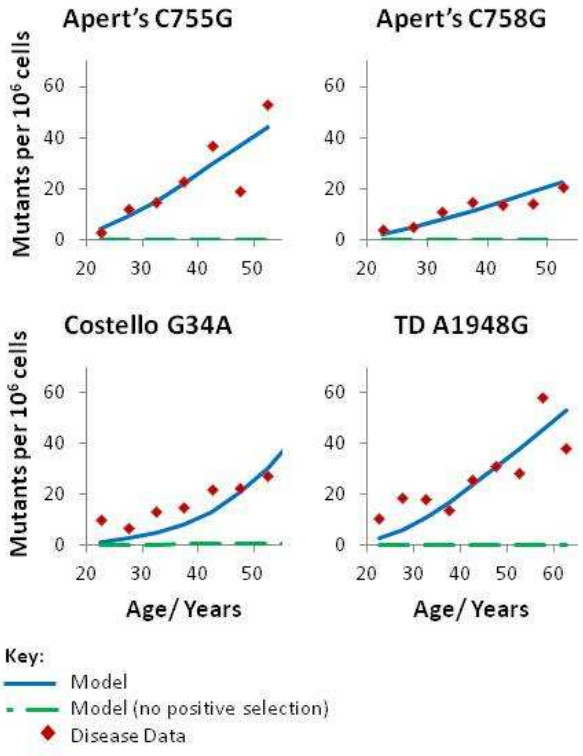
$$p_a = 1 - \left((1 - p)^a \right) \tag{3.3}$$

Where p is the baseline mutation probability and a is the number of potential mutation sites that can produce mutant alleles.

4 Discussion

We hypothesized that one of the two Apert's syndrome-causing mutations (C755G) has a higher incidence than the other (C758G) because the former occurs at a CpG dinucleotide, which has a higher mutability due to spontaneous deamination. Our results support this hypothesis as both of the Apert mutations have a very similar r value with 0.0124 for C755G compared with 0.0126 for C758G. These two distinct amino-acid substitutions (S252W and P253R) appear therefore to have the same selective effect on the cell and the increased incidence of S252W is purely because of increased mutability at this site. The projected incidence rate is very sensitive to the model parameters and particularly the mutation rate. For example, achondroplasia has a high incidence rate relative to the other diseases (1 in 27,000, see table 2). The computed r value on the other hand was middle of the range, which failed to account for the high incidence rate when the baseline value for p was used. Once the value of p was increased to the level of a C→T transition, the predicted incidence agreed well. So the site-specific mutation rate accounted for the relatively high incidence rate of this disease.

Figure 3. Rates of mutants per million sperm with age.



Apert syndrome (Yoon *et al.* 2009), Costello syndrome (Giannoulatou *et al.* 2013), Thanatophoric dysplasia (TD) adapted from Goriely *et al.* (2009).

The predicted incidence rates, after accounting for the number of alleles and the mutation rate, present close to the actual values, with the exception of fibrodysplasia ossificans progressiva. This disease is anomalous as it is a very rare disease (one in 2 million births) but it is predicted to have a high incidence rate as it is caused by a transition at a single CpG site (Shore *et al*. 2006). While the rates of substitution vary by location in the genome, including the rates at CpG sites (Mugal & Ellegren 2011, Fryxell & Moon 2005), it is unlikely the substitution rate could be low enough at this point, even if unmethylated. It might be explained by a very low selective advantage but the projected r value is high (0.0135), producing an O/E curve similar to achondroplasia. The low birth prevalence is also not explained by low survival to term of affected offspring as FOP does not show severe symptoms until later in life or by any variation of expression of the mutant allele as it shows complete penetrance (Petrie *et al.* 2009) so the low incidence rate remains unexplained.

Table 2. Incidence Rates of 8 Diseases

Disease	Raw Predicted Incidence Rate	Adjusted Predicted Incidence Rate	Literature Incidence rate	Reference	Alleles
ACH	1 in 400,000	1 in 27,000	1 in 26,000	[1]	1
Apert	1 in 700,000	1 in 130,000	1 in 100,000	[2]	2
Costello	1 in 2,300,000	1 in 160,000	1 in 300,000	[3]	14
Crouzon	1 in 600,000	1 in 37,000	1 in 60,000	[4]	16
FOP	1 in 340,000	1 in 23,000	1 in 2,000,000	[5]	1
Marfan	1 in 3,400,000	1 in 68,000	1 in 70,000	[6]	50
Pfeiffer	1 in 1,600,000	1 in 130,000	1 in 100,000	[7]	12
TD	1 in 720,000	1 in 60,000	1 in 40,000	[8]	12

ACH – achondroplasia, FOP – fibrodysplasia ossificans progressiva, TD – thanatophoric dysplasia.

Raw predicted incidence rates calculated with a baseline mutation rate of 4×10^{-11}. Adjusted rates account for variation in mutation rate at CpG dinucleotides and the number of mutable alleles that cause the disease phenotype. Alleles denote the number of most common genetic variants that comprise at least 95% of cases of the disease. (Online Mendelian Inheritance in Man, 2016, URL: http://omim.org/). Incidence rates of ACH is a mean value between 0.36 and 0.6 per 10,000 after accounting for 20% of ACH cases being inherited from an affected parent. Sources: [1] Faruqi *et al.* 2014, [2] Blank 1960, [3] Lurie 1994, [4] Helman *et al*. 2014, [5] Hüning & Gillessen-Kaesbach 2014, [6] Lynas 1958, [7] Vogels & Fryns 2006, [8] Connor *et al*. 1985.

The mutations causing thanatophoric dysplasia and achondroplasia both cause constitutive activation of the *FGFR3* but TD mutations activate the receptor more strongly, leading to a more severe phenotype (Naski *et al.* 1996, Bonaventure *et al.* 2007). We can therefore predict the r value for TD to be higher than that for achondroplasia, which is confirmed by our model (TD r = 0.0105, achondroplasia r = 0.00741). Note that both sets of data were taken from one study (Orioli *et al.* 1995) in order to ensure that they are comparable.

Our estimate for r from sequence data for Apert's syndrome (r = 0.0125) showed good agreement with that estimated by Yoon and colleagues, who estimated r to be 0.014 (Yoon *et al.* 2009), although the value from birth data was lower (r = 0.00888).

The model presented in this paper provides a mathematical understanding of the accumulation of selfish disease-causing mutations. We have successfully predicted the incidence rates of different diseases

based on O/E curves and information of the molecular nature of the mutations and estimates for the strength of selection. The selective advantage granted by these mutations is the most important factor in terms of the exponential increase over time but the site-specific mutation rate and the number of mutable sites plays a key role in how common the disease is at the population level.

Acknowledgements

We thank Dr Mike Stacey and Dr Holly Gaff for informative discussions regarding the model presented in this paper.

Funding

This work has been supported by the Mary Louise Andrews award for Cancer Research (2013).

Conflict of Interest: none declared.

References

Blank CE. (1960) Apert's syndrome (a type of acrocephalosyndactyly)-observations on a British series of thirty-nine cases. *Ann Hum Genet.* **24**:151-64.

Bonaventure J, Horne WC, Baron R. (2007) The localization of FGFR3 mutations causing thanatophoric dysplasia type I differentially affects phosphorylation, processing and ubiquitylation of the receptor. *FEBS J.* **274**(12):3078-93.

Bozic I, Nowak MA. 2013 Cancer. Unwanted evolution. *Science.* **22**;342(6161):938-9

Bray I, Gunnell D, Davey Smith G. (2006) Advanced paternal age: how old is too old? *J Epidemiol Community Health.* **60**(10):851-3.

Choi SK, Yoon SR, Calabrese P, Arnheim N. (2008) A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. *Proc Natl Acad Sci U S A.* **105**: 10143–10148.

Connor, J. M., Connor, R. A. C., Sweet, E. M., Gibson, A. A. M., Patrick, W. J. A., McNay, M. B., Redford, D. H. A. (1985) Lethal neonatal chondrodysplasias in the West of Scotland, 1970-1983, with a description of a thanatophoric, dysplasialike, autosomal recessive disorder, Glasgow variant. *Am. J. Med. Genet.* **22**: 243-253.

Faruqi T, Dhawan N, Bahl J, Gupta V, Vohra S, Tu K, Abdelmagid SM. (2014) Molecular, phenotypic aspects and therapeutic horizons of rare genetic bone disorders. *Biomed Res Int.* **2014**:670842.

Fryxell KJ1, Moon WJ. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol.* **22**(3):650-8.

Giannoulatou E, McVean G, Taylor IB, McGowan SJ, Maher GJ, Iqbal Z, Pfeifer SP, Turner I, Burkitt Wright EM, Shorto J, Itani A, Turner K, Gregory L, Buck D, Rajpert-De Meyts E, Looijenga LH, Kerr B, Wilkie AO, Goriely A. (2013) Contributions of intrinsic mutation rate and selfish selection to levels of de novo HRAS mutations in the paternal germline. *Proc Natl Acad Sci U S A.* **110**(50):20152-7.

Glaser RL, Jabs EW. (2004) Dear old dad. *Sci Aging Knowledge Environ.* **2004**(3):re1.

Goriely A, Hansen RM, Taylor IB, Olesen IA, Jacobsen GK, McGowan SJ, Pfeifer SP, McVean GA, Rajpert-De Meyts E, Wilkie AO. (2009) Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. *Nat Genet.* **41**(11):1247-52.

Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Jagadeesan A, Ebenesersdóttir SS, Kong A, Stefánsson K. (2015) The Y-chromosome point mutation rate in humans. *Nat Genet.* **47**(5):453-7.

Helman SN, Badhey A, Kadakia S, Myers E. (2014) Revisiting Crouzon syndrome: reviewing the background and management of a multifaceted disease. *Oral Maxillofac Surg.* **18**(4):373-9.

Hook EB. (1981) Rates of chromosome abnormalities at different maternal ages. *Obstet Gynecol.* **58**(3):282-5.

Hüning I, Gillessen-Kaesbach G. (2014) Fibrodysplasia ossificans progressiva: clinical course, genetic mutations and genotype-phenotype correlation. *Mol Syndromol.* **5**(5):201-11.

Humphries A, Wright NA. (2008) Colonic crypt organization and tumorigenesis. *Nat Rev Cancer.* **8**(6):415-24.

Klein AM, Nakagawa T, Ichikawa R, Yoshida S, Simons BD. (2010) Mouse germ line stem cells undergo rapid and stochastic turnover. *Cell Stem Cell.* **6**;7(2):214-24.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol.* **5**(1):182-7.

Lurie IW. (1994) Genetics of the Costello syndrome. *Am J Med Genet.* **52**(3):358-9.

Lynas MA. (1958) Marfan's syndrome in Northern Ireland; an account of thirteen families. *Ann Hum Genet.* **22**(4):289-309.

Lynch M. Rate, molecular spectrum, and consequences of human mutation. (2010) *Proc Natl Acad Sci U S A.* **107**(3):961-8.

Maher GJ, Goriely A, Wilkie AO. (2014) Cellular evidence for selfish spermatogonial selection in aged human testes. *Andrology*. **2**(3):304-14.

Mugal CF, Ellegren H. (2011) Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol.* **12**(6):R58.

Nachman MW, Crowell SL. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*. **156**(1):297-304.

Naski MC, Wang Q, Xu J, Ornitz DM. (1996) Graded activation of fibroblast growth factor receptor 3 by mutations causing achondroplasia and thanatophoric dysplasia. *Nat Genet.***13**(2):233-7.

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2016.

Orioli IM, Castilla EE, Scarano G, Mastroiacovo P. (1995) Effect of paternal age in achondroplasia, thanatophoric dysplasia, and osteogenesis imperfecta. *Am J Med Genet.* **6**;59(2):29-17.

Paul C, Robaire B. (2013) Ageing of the male germ line. *Nat Rev Urol.* **10**(4):227-34

Penrose LS. (1955) Parental age and mutation. *Lancet.* **269**:312-313.

Petrie KA, Lee WH, Bullock AN, Pointon JJ, Smith R, Russell RG, Brown MA, Wordsworth BP, Triffitt JT. (2009) Novel mutations in ACVR1 result in atypical features in two fibrodysplasia ossificans progressiva patients. *PLoS One.* **4**(3):e5005.

Qin J, Calabrese P, Tiemann-Boege I, Shinde DN, Yoon SR, Gelfand D, Bauer K, Arnheim N. (2007) The molecular anatomy of spontaneous germline mutations in human testes. PLoS Biol. **5**(9):e224.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, Stratton MR; UK10K Consortium, Hurles ME. (2016) Timing, rates and spectra of human germline mutation. *Nat Genet.* **48**(2):126-33.

Ray D, Pitts PB, Hogarth CA, Whitmore LS, Griswold MD, Ye P. (2014) Computer simulations of the mouse spermatogenic cycle. *Biol Open.* **4**(1):1-12

de Rooij DG, Russell LD. (2000) All you wanted to know about spermatogonia but were afraid to ask. *J Androl.* **21**(6):776-98.

de Rooij DG, Griswold MD. (2012) Questions about spermatogonia posed and answered since 2000. *J Androl.* **33**(6):1085-95.

de Rooij DG, van Beek ME. (2013) Computer simulation of the rodent spermatogonial stem cell niche. *Biol Reprod.* **88**(5):131

Risch N, Reich EW, Wishnick MM, McCarthy JG. (1987) Spontaneous mutation and parental age in humans. *Am J Hum Genet.* **41**(2):218-48.

Sarabipour S, Hristova K. (2016) Mechanism of FGF receptor dimerization and activation. *Nat Commun.* **7**:10262.

Sato T, Aiyama Y, Ishii-Inagaki M, Hara K, Tsunekawa N, Harikae K, Uemura-Kamata M, Shinomura M, Zhu XB, Maeda S, Kuwahara-Otani S, Kudo A, Kawakami H, Kanai-Azuma M, Fujiwara M, Miyamae Y, Yoshida S, Seki M, Kurohmaru M, Kanai Y. (2011) Cyclical and patch-like GDNF distribution along the basal surface of Sertoli cells in mouse and hamster testes. *PLoS One.* **6**(12):e28367.

Shore EM, Xu M, Feldman GJ, Fenstermacher DA, Cho TJ, Choi IH, Connor JM, Delai P, Glaser DL, LeMerrer M, Morhart R, Rogers JG, Smith R, Triffitt JT, Urtizberea JA, Zasloff M, Brown MA, Kaplan FS. (2006) A recurrent mutation in the BMP type I receptor ACVR1 causes inherited and sporadic fibrodysplasia ossificans progressiva. *Nat Genet.* 38(5):525-7.

Shinde DN, Elmer DP, Calabrese P, Boulanger J, Arnheim N, Tiemann-Boege I. (2013) New evidence for positive selection helps explain the paternal age effect observed in achondroplasia. Hum Mol Genet. **22**(20):4117-26.

Shinohara T, Orwig KE, Avarbock MR, Brinster RL. (2001) Remodeling of the postnatal mouse testis is accompanied by dramatic changes in stem cell number and niche accessibility. *Proc Natl Acad Sci U S A.* **98**(11):6186-91.

Tegelenbosch RA, de Rooij DG. (1993) A quantitative study of spermatogonial multiplication and stem cell renewal in the C3H/101 F1 hybrid mouse. *Mutat Res.* **290**(2):193-200.

Tomasetti C, Vogelstein B. (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science.* **347**(6217):78-81.

Vogel F, Rathenberg R. (1975) Spontaneous mutation in man. *Adv Hum Genet.* **5**:223-318.

Vogels A, Fryns JP. Pfeiffer syndrome. (2006) *Orphanet J Rare Dis.* **1**:19.

Yoon SR, Qin J, Glaser RL, Jabs EW, Wexler NS, Sokol R, Arnheim N, Calabrese P. (2009) The ups and downs of mutation frequencies during aging can account for the Apert syndrome paternal age effect. PLoS Genet. **5**(7)

Yoon SR, Choi SK, Eboreime J, Gelb BD, Calabrese P, Arnheim N. (2013) Age-dependent germline mosaicism of the most common noonan syndrome mutation shows the signature of germline selection. *Am J Hum Genet.* **92**(6):917-26.

Yoshidaf S, Sukeno M, Nabeshima Y. (2007) A vasculature-associated niche for undifferentiated spermatogonia in the mouse testis. *Science.* **317**(5845):1722-6.

Yoshida S. (2008) Spermatogenic Stem Cell System in the Mouse Testis. *Cold Spring Harb Symp Quant Biol* **73**: 25-32