Critical Mutation Rate has an Exponential Dependence on Population Size for Eukaryotic-Length Genomes

Elizabeth Aston[1], Alastair Channon[1], Roman V. Belavkin[2], Rok Krasovec[3] and Christopher G. Knight[3]

[1] School of Computing and Mathematics, Keele University, ST5 5BG, UK
[2] School of Engineering and Information Sciences, Middlesex University, London, NW4 4BT, UK
[3] Faculty of Life Sciences, The University of Manchester, M13 9PT, UK
{e.j.aston, a.d.channon}@keele.ac.uk, r.belavkin@mdx.ac.uk, {rok.krasovec, chris.knight}@manchester.ac.uk

Abstract

The critical mutation rate (CMR) determines the shift between survival-of-the-fittest and the survival of individuals with greater mutational robustness (the "flattest"). Small populations are more likely to exceed the CMR and become less well adapted; understanding the CMR is crucial to understanding the potential fate of small populations under threat of extinction. Here we present a simulation model capable of utilising input parameter values within a biologically relevant range. A previous study identified an exponential fall in CMR with decreasing population size, but the parameters and output were not directly relevant outside artificial systems. The first key contribution of this study is the identification of an inverse relationship between CMR and gene length when the gene length is comparable to that found in biological populations. The exponential relationship is maintained, and the CMR is lowered to between two to five orders of magnitude above existing estimates of per base mutation rate for a variety of organisms. The second key contribution of the study is the identification of an inverse relationship between CMR and the number of genes. Using a gene number in the range for *Arabidopsis thaliana* produces a CMR close to its known mutation rate; per base mutation rates for other organisms are also within one order of magnitude. This is the third key contribution of the study as it represents the first time such a simulation model has used input and produced output both within range for a given biological organism. This novel convergence of CMR model with biological reality is of particular relevance to populations undergoing a bottleneck, under stress, and subsequent conservation strategy for populations on the brink of extinction.

Introduction

Fitter genotypes can be outcompeted by genotypes with greater robustness when the mutation rate exceeds a critical mutation rate (CMR); in terms of fitness landscapes, narrow high fitness peaks may be lost, while broader, lower peaks are maintained by a population of reproducing sequences. This so called "survival-of-the-flattest" has been observed in *in silico* evolving systems (Wilke, 2005). CMR has an exponential dependence on population size in both haploid (Channon et al., 2011) and diploid populations (Aston et al., 2013); as population size falls, the CMR above which fitter alleles are lost transitions unexpectedly from near-constant (the previous assumption in evolutionary biology) to drop exponentially for small populations. It has been verified that this model closely reproduces the established mathematical relationship between population size and "error threshold" (ET. No mathematical model has yet been derived for the CMR) (Aston et al., 2013). It is therefore possible that CMRs in small populations could be within the range of biological mutation rates. However, biological organisms typically have lengths and numbers of genes orders of magnitude higher than those used in models of ETs or CMRs, so how relevant such models are to real biological populations remains an open question.

From Artificial to Biological Evolution: Mutation of Genes in Nature

To bridge the gap between artificial and biological evolution it is paramount that, when implemented as a simulation, a model can be given parameter values within the range observed in biological organisms and subsequently output biologically realistic results. The models defined in Aston et al. (2013) used arbitrary values for parameters such as sequence length, selected for their suitability to provide results within a small timeframe.

Whitlock et al. (2003) performed computer simulations to investigate the effects of varying the strength of selection and mutational effects among dimensions. They used a model based on Fisher's model of the geometry of adaptation (Fisher, 1930), but used a hyperellipse in which the strength of selection along any axis was drawn from an exponential distribution. They concluded that changing from a hypersphere to a hyperellipse, and thus introducing dimensions with stronger selection than others, had a negligible effect on their results. It was therefore decided to focus on the parameters of mutation rate, gene length, and gene number; assuming equal strength of selection is not expected to affect the credibility of the results.

Mutation Rates

The mutation rate used in the simulation model is analagous to the biological per base mutation rate (see Table 1). Bac-

Figure 1: **Two-peak fitness landscape**, with one narrow peak of high fitness (Peak 0), and one broader peak of lower fitness (Peak 1). The fitness score is relative, and the width and distance between the peaks are given in terms of Hamming distance. Diagram adapted from Wilke (2005).

terial species were not included due to their use of lateral gene transfer which is not currently included in the simulation model. Viruses, which are known to live very close to the ET (Eigen and Schuster, 1979), were not included due to the complexity and variety of reproduction techniques which include incorporation into a host genome. Nachman and Crowell (2000) obtained an estimate of the average mutation rate per nucleotide by comparing pseudogenes (genes that do not code for proteins or are never expressed) in humans and chimpanzees. Baer et al. (2007) brought together the results of theoretical and empirical studies to list mutation rate estimates in a number of multicellular eukaryotes. Drake et al. (1998) list mutation rate estimates from studies using mutation accumulation and radiation experiments. Lynch (2010a) also lists mutation rates from various sources. Xue et al. (2009) obtained an estimate for the base substitution rate in the human Y chromosome through direct sequencing. Kumar and Subramanian (2002) conducted a computational analysis of 5669 genes from species of placental mammals. Keightley et al. (2009) did whole-genome shotgun sequencing of three mutation accumulation lines of *Drosophila melanogaster*, while Keightley et al. (2014) sequenced two parents and 12 offspring. Denver et al. (2004) provide a direct estimate of the mutation rate from a set of *Caenorhabditis elegans* mutation accumulation lines. Haag-Liautard et al. (2008) and Ossowski et al. (2010) provide estimates using mutation accumulation lines. Durbin et al. (2010) examine variation in the sequence of the human genome. Lynch et al. (2008) provide a mutation rate estimate from complete genome sequencing of *Sacccharomyces cerevisiae.* Lynch (2010b) used existing data to estimate the mutation rate of various eukaryotes.

Genetic Sequences

Derelle et al. (2006), Sharma et al. (2005) and Lewin (2008) list the length of various genes for various biological organisms at between approximately 1000 to 140,000 bp; the sequence length of 30 bp used to produce the results in Aston et al. (2013) is small when compared with the length of genes found in a wide range of natural species.

Aston et al. (2013) used a two-peak landscape, with the height of peak 0 constant at 15 and the radius 2, the height of peak 1 constant at 10 and the radius 5, and the Hamming distance between the peaks set at 10 (Figure 1). If each peak in the two-peak landscape is considered to be a different set of alleles (variant of a gene), estimates of genetic distances between alleles for various genes can be seen to be analogous to the distance between the peaks. They fall within the range of 1 and 56 polymorphisms (Bryan et al., 2000; Ramkumar et al., 2010). Similarly, the number of polymorphisms was estimated to be at most 13 (including non-coding regions) within various human genes studied by Cargill et al. (1999). In both cases the value of 10 used for the distance between peaks in Aston et al. (2013) is close to the range of numbers listed therefore it was decided to keep this number constant. Varying the distance between the peaks may be an interesting future study.

Longer sequences means more bases to potentially mutate each generation, leading to the formation of three hypotheses.

Hypothesis 1

According to the drift-barrier hypothesis, drift prevails over selection to determine mutation if the magnitude of the selection coefficient s is less than $1/Ne$ (where Ne is effective population size). The strength of selection that reduces mutation rate through mutation-selection balance is countered by Ne-dependent genetic drift (Sniegowski and Raynes, 2013). Following this population size dependence, we hypothesise that, for varying gene lengths and numbers, the CMR will vary with population size, and that this will occur in line with the exponential model identified in Aston et al. (2013).

Hypothesis 2

Drake summarized all studies up to 1990 and concluded that the per nucleotide per generation mutation rate u varies inversely with genome size G in microbes (Drake, 1991; Sung et al., 2012). Eigen and Schuster (1979) theoretically determined the ET in terms of selection pressure and sequence length. Using this model, Ochoa et al. (2000) and Ochoa (2006) found that longer sequence lengths lead to lower ETs in genetic algorithms, defined by the equation $p = \frac{ln(\sigma)}{L}$ where p is the ET on a single peak landscape, L is sequence length, and σ is selection strength which is kept constant. Nowak (1992) theoretically determined the ET in terms of the relative fitness of mutant and wild type (a_1 and a_2 respectively, where a_2 is assumed to have the lower fitness) and the sequence length (m). Giving the ET as $1 - q_{crit} = \left(\frac{a_1}{a_2}\right)^{\frac{1}{m}}$, it can be seen that increasing m will decrease the ET. In accordance with this and with Drake

(1991), it is expected that increasing the sequence length will also lower the CMR.

Hypothesis 3

We hypothesise that increasing the number of genes (while keeping gene length constant) will further lower the CMR as it will increase the overall sequence length. Gene numbers within biological ranges are expected to lead to CMRs close to the range of mutation rates observed for biological species; it is expected biological organisms will be evolving close to the mutation rate that results in the greatest levels of adaptation.

Simulation Model

The system used a two-peak fitness landscape (Figure 1), with the height of peak 0 constant at 15 and the radius 2, the height of peak 1 constant at 10 and the radius 5, and the Hamming distance between the peaks set at 10 as per Aston et al. (2013). Each individual consisted of one randomly assigned maternal and one paternal sequence of alphabet size 4, and each sequence was split into *n* genes of length *L*. Each gene had an associated target sequence of length *L* corresponding to peak 0 and a target sequence corresponding to peak 1. For example, if *n* is set to 4, there will be target sequences corresponding to peaks $0_1 1_1$, $0_2 1_2$, $0_3 1_3$, and $0_4 1_4$. For simplicity, each peak 0 was set to be all 0s and each peak 1 was randomly generated to be Hamming distance 10 away. Recombination was limited to one event per replication as it was not the focus of the study.

The dominance parameter λ was set to equal a fraction below 1.0 (0.999999999999999 specifically). This sets the relative importance of the maternal and paternal alleles while preventing either allele from drifting neutrally; if λ=1.0 the fitness of only one allele is taken into account, while the other can be anywhere in the fitness landscape. For each individual, the fitness of each of its *n* genes was calculated as the Hamming distance of its maternal and paternal sequences relative to each peak. The fitness values relative to peak 0 were compared with the fitness values relative to peak 1 and the highest of these selected to give a single fitness value for both the maternal and paternal sequences. The resulting maternal and paternal fitnesses were compared and subsequently designated as $f_{\max}$ and $f_{\min}$. The final relative fitness of each gene was calculated as $f = (\lambda \times f_{\max}) + ((1 - \lambda) \times f_{\min})$. The overall fitness of the individual was then taken to equal the minimum fitness out of the *n* genes present. The simulation was run for a range of population sizes to confirm the curves observed in previous experiments (Channon et al., 2011; Aston et al., 2013) are observed as the length and number of genes is increased.

To allow the simulation to complete within a realistic time frame, it was optimised to cease running when any one gene had lost peak 0; this was all the information required to determine the CMR, which was recorded as the mutation rate at which 95% of 2000 runs lost peak 0 within 10,000 generations for any of the possible *n* genes. Launching the simulation for various combinations of parameter values was also optimised to allow the mutation rate being tested for a given gene number to progress to the next mutation rate once 100 out of the possible 2000 runs (corresponding to 5%) have kept peak 0 for the duration of the simulation. Once this threshold has been exceeded, less than 95% of the 2000 runs will have lost peak 0, and the CMR will not have been reached. While this helped significantly with run time, further optimisation will be required in the future; it is currently not feasible to run the simulation for a wide range of population sizes and mutation rates for gene numbers at the upper end of the biological range.

Two approaches were taken when selecting parameter values within the biological range. Firstly, a gene length of 1000 bp was selected as a small yet biologically realistic gene length. This provided a gene length small enough to allow the simulation to run for a range of gene numbers within a realistic time frame. Secondly, based on the information in Table 1 and the value of 2232 bp mean gene size given by Derelle et al. (2006), *Arabidopsis thaliana* (thale cress) was selected as a model organism with a relatively short gene length. It is a plant native to Eurasia, with an effective population size of between 250,000 to 300,000 (Cao et al., 2011), known to contain 25,498 genes encoding proteins from 11,000 families (The Arabidopsis Genome Initiative, 2000). More current estimates of gene number are slightly higher but still within a close range for the purpose of the model (Bevan and Walsh, 2005). The simulation was run for population size 10 with a gene length of 1000 (as per the previous runs) or 2000 (range of *A.thaliana*), but with 25,000 genes to bring the gene number into the range of *A.thaliana*.

Results

Increasing the gene length decreses the CMR in line with the exponential model

Figure 2 shows the CMR for two of the sequence lengths studied; increasing sequence length decreases the CMR in a single-gene-per-individual diploid *in silico* evolving system modelled on the biological process of meiosis, while maintaining the exponential relationship with population size presented in Aston et al. (2013). *A.thaliana* has a gene length of 2232 bp (Derelle et al., 2006), the average gene length in humans is 27 kbp (Lewin, 2008), the upper bound for the usual gene length range for flies and mammals is 100 kbp (Lewin, 2008), and the longest gene in the collagen family is 132.83 kbp (Sharma et al., 2005); the length of genes present in biological species varies greatly. Table 1 was used

Table 1: **Mutation rates for various eukaryotic species.** Mutation rate estimates are specified as the number of times a single base will mutate spontaneously. If a timeframe (per generation, per cell division) is specified this is listed in the *Unit* column. * refers to mutation rates used for reference in Figure 3.

Species	*Genome size (Mbp)*	*Mutation rate*	*Base/genome*	*Unit*	*Source*
Human	3080	1.00E-08 - 2.50E-08 *	Per base	Per generation	Nachman and Crowell (2000); Durbin et al. (2010); Lynch (2010a)
Human	3080	1.75E+02	Per genome	Per generation	Nachman and Crowell (2000)
Human	3080	5.00E-11 - 6.00E-02	Per base	Per cell division	Drake et al. (1998); Lynch (2010a)
Human	3080	1.60E-01	Per genome	Per cell division	Drake et al. (1998)
Human (Y chromosome)	58	3.00E-08	Per base	Per generation	Xue et al. (2009)
Human, chimpanzee	3080	3.00E+00	Per genome	Per generation	Baer et al. (2007)
Drosophila melanogaster	120	4.65E-09 - 6.20E-08	Per base	Per generation	Haag-Liautard et al. (2008); Keightley et al. (2009); Lynch (2010a); Keightley et al. (2014)
Drosophila melanogaster	120	9.90E-01 - 1.20E+00	Per genome	Per generation	Baer et al. (2007); Haag-Liautard et al. (2008)
Drosophila spp.	120	7.00E-02	Per genome	Per generation	Baer et al. (2007)
Drosophila melanogaster	120	1.30E-10 - 3.40E-10	Per base	Per cell division	Drake et al. (1998); Lynch (2010a)
Quail, chicken	1050	4.90E-01	Per genome	Per generation	Baer et al. (2007)
Sheep, cow	2870	9.00E-01	Per genome	Per generation	Baer et al. (2007)
Old World Monkey		1.90E+00	Per genome	Per generation	Baer et al. (2007)
Mouse, rat	2640	9.10E-01	Per genome	Per generation	Baer et al. (2007)
Mouse	2640	1.80E-10	Per base	Per cell division	Drake et al. (1998)
Mouse	2640	1.10E-08	Per base	Per generation	Drake et al. (1998)
Saccharomyces cerevisiae	12.1	3.30E-10	Per base	Per generation	Lynch (2010a)
Saccharomyces cerevisiae	12.1	3.30E-10	Per base	Per cell division	Lynch et al. (2008)
Average mammalian		2.20E-09 *	Per base	Per genome/year	Kumar and Subramanian (2002)
Mammalian upper bound		2.61E-09	Per base	Per genome/year	Kumar and Subramanian (2002)
Caenorhabditis elegans	100	8.40E-09 - 2.10E-08	Per base	Per generation	Denver et al. (2004); Haag-Liautard et al. (2008); Lynch (2010b)
Caenorhabditis elegans	100	2.90E+00	Per genome	Per generation	Lynch et al. (2008)
Arabidopsis thaliana	157	7.10E-09 *	Per base	Per generation	Ossowski et al. (2010)
Arabidopsis thaliana	157	6.50E-09	Per base	Per generation	Lynch (2010b)

to identify a known mutation rate for *A.thaliana*, the average mammal, and humans as 7.1×10^{-9}, 2.2×10^{-9}, and 2.5×10^{-8} respectively (per base, per generation). Figure 3 shows each of these mutation rates plotted against their respective gene lengths, along with the maximal CMR produced when the simulation was run with sequence lengths of 2000, 27000, 100000, and 150000. The maximal CMR represents the value at which each curve has levelled out (e.g., Figure 2), applicable to the range of population sizes normally expected for each species without threat of extinction (where population size refers to a local population rather than the total number of individuals globally). It was taken to be the CMR at population size 1000. Note the log scale used for the mutation rate as this enables the difference between the curves and the biological mutation rates to be seen clearly.

While Figure 3 is promising in that none of the biological mutation rates are higher than the respective CMRs produced by the simulation, both sets of mutation rates are between two and five orders of magnitude from each other.

Increasing the number of genes produces biologically realistic CMRs

As increasing gene length has been seen to decrease CMR, increasing the number of genes was also expected to decrease CMR. The simulation model was run with a minimal yet biologically realistic gene length of 1000, with gene number doubling from n=1 up to n=8192. The CMR was recorded as the mutation rate at which 95% of 2000 runs lost peak 0 for any of the possible n genes. Figure 4 shows the CMR decreases by up to three magnitudes as gene number increases from 1 to 8192, bringing the CMR to within an order of magnitude of the biological mutation rates listed in Table 1. Curve fitting using R showed the results follow quadratic curves; these can be seen to become closer as population size is increased, indicating the decrease in the rate of change of CMR with increasing population size seen in Figure 2.

Population size 10 was also run with 25,000 genes of length 1000 or 2000 to bring the gene number to within the correct range for *A.thaliana* (The Arabidopsis Genome Initiative, 2000; Bevan and Walsh, 2005). Increasing the gene length decreased the CMR further to within an order of magnitude of the per base per generation mutation rate for *A.thaliana* which is given as 7.1×10^{-9} (Table 1). Figure 4 shows per base mutation rate estimates for *A.thaliana*, *C.elegans*, and *D.melanogaster* taken from Table 1, each of which are within an order of magnitude of the simulation results for 25,000 genes for population size 10. It is notable that the genome size estimates for multicellular eukaryotes used in Figure 4 are based on numbers of protein coding genes. Protein coding sequences account for a relatively small proportion of the total genome length in such organisms (1.2% in humans (Consortium, 2012)), but much more of the sequence is functional at some level, probably at least

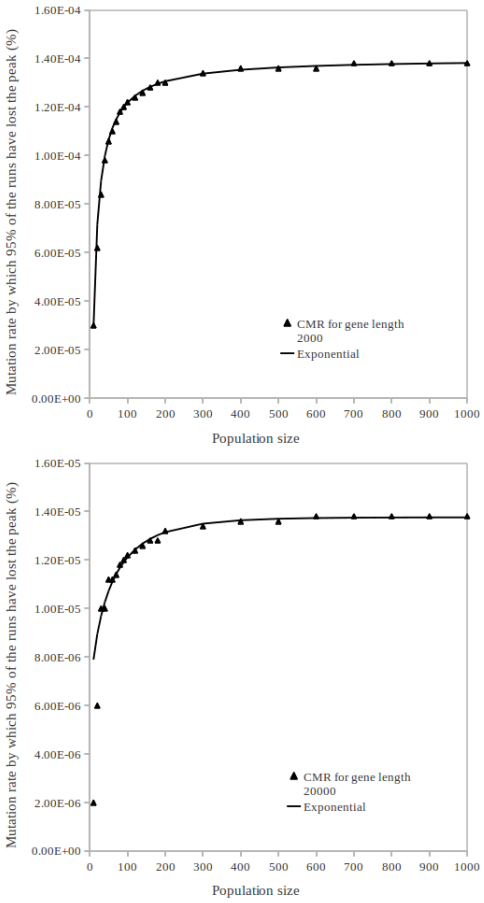

Figure 2: **CMR when the GA was run for one gene with a sequence length of 2000 and 20000.** The exponential line was obtained using the equation $y = A - B * e^{-((N/C)^D)}$(with N being population size), and the parameters determined by curve-fitting using R with a least squares method.

9% (Ward and Kellis, 2012), with estimates of up to 80% in humans (Consortium, 2012; Lu et al., 2015) (albeit this last figure is likely to be a substantial over-estimate (Graur et al., 2013)). This means that the genome size at which these biological mutation rates are plotted in Figure 4 is a minimal estimate, the true value being substantially, perhaps an order of magnitude, higher, therefore putting their observed mutation rates closer to the CMRs estimated by simulation.

Discussion

Aston et al. (2013) showed that population size influences the CMR that can be tolerated before fitter individuals are outcompeted by those that have a greater mutational robustness in both haploid and diploid artificial populations, a result which has now been demonstrated to have relevance beyond artificial systems. Gene lengths given in Derelle et al. (2006), Sharma et al. (2005) and Lewin (2008) show that the

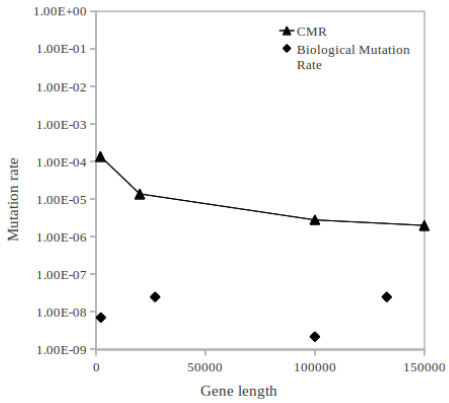

Figure 3: **Maximal CMR plotted alongside biological per base per generation mutation rates for one gene with varying sequence lengths.** Maximal CMR is the CMR recorded for population size 1000 in the simulation, representing the point at which the curve for each gene length has levelled out (e.g., Figure 2). The biological mutation rates were taken from Table 1 for eukaryotic species with comparable gene lengths to the sequence lengths used in the simulation.

sequence length of 30 used in Aston et al. (2013) is significantly smaller than the length of genes observed in biological species. The mutation rates in Table 1 are also many orders of magnitude lower than the CMR reported in Aston et al. (2013). Hypothesis 1 stated that the CMR will always have an exponential dependence on population size, while hypothesis 2 stated increasing the sequence length will lower the CMR; Figure 2 supports these hypotheses as it shows that increasing the gene length by a factor of 10 decreases the CMR by a factor of 10, with each gene length resulting in an exponential fall in CMR with decreasing population size. A change in order of magnitude can also be seen in the biological values. For example, the mean gene length of *A.thaliana* is given as 2232 bp (Derelle et al., 2006), while the average gene length of humans is just over 10 times longer at 27 kbp (Lewin, 2008). In Table 1, the per base per generation mutation rate for *A.thaliana* is given as 7.1×10^{-9}, while the per base per generation mutation rate for humans is an order of magnitude higher at 2.5×10^{-8}. Figure 3 shows that, when compared with mutation rates for biological species with comparable gene lengths, the CMR is always higher as expected. This is a key contribution of the study as it indicates the CMR exhibits a comparable relationship with gene length as previously determined for the ET (Nowak, 1992; Ochoa et al., 2000; Ochoa, 2006); the consistency between the known results for the ET and our new results for the CMR increases confidence in the study.

While it is clear that increasing sequence length decreases the CMR, the CMR remains between two and four orders of magnitude higher than the biological mutation rates (Figure 3). Hypothesis 3 stated that increasing the number of genes

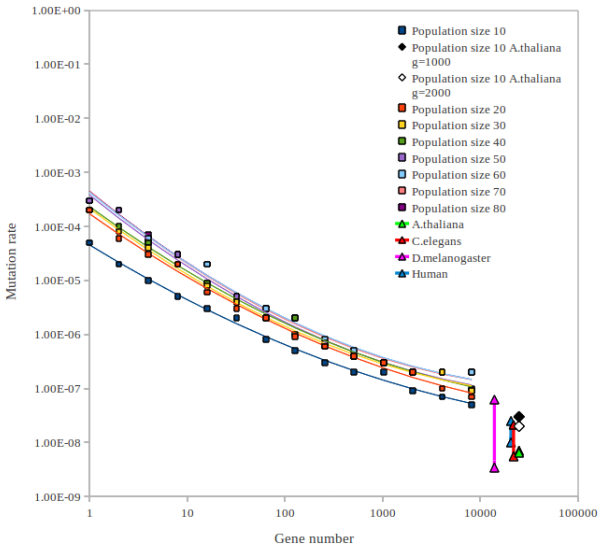

Figure 4: **CMR plotted alongside gene number for varying population sizes.** Data are shown for population sizes 10 to 80 with results plotted on a log log scale. Gene length was kept constant at 1000, while gene number was doubled from 1 up to 8192. The corresponding quadratic lines were obtained by curve-fitting using R (specifically Qmod <- lm(log(ydata)~ log(xdata) + I(log(xdata)^2))). Population sizes shown represent the steep part of the curve in Figure 2 before it levels out. Population size 10 was also run with 25,000 genes, correct range for *A.thaliana*. Gene length was set to 1000 to match the other runs or 2000 to bring it closer to *A.thaliana*'s gene length. For reference, the range of per base mutation rates from Table 1 is shown for *A.thaliana*, *C.elegans* (nematode worm), *D.melanogaster* (fruit fly) and humans (with gene number estimates from The Arabidopsis Genome Initiative, (2000), Nam and Bartel (2012), Ashburner and Bergman (2005), and Consortium (2012) respectively). The mean gene size of *A.thaliana* is 2232 bp (Derelle et al., 2006), the median gene length for *C.elegans* is ~1700 b (Cutter et al., 2009), the average gene length for *D.melanogaster* is 1130 b, and for humans 27 kb (Lewin, 2008).

will lower the CMR. Gene numbers within biological ranges were expected to lead to CMRs close to the range of mutation rates observed for biological species. Consistent with this, Figure 4 demonstrates that when gene length is kept constant, doubling the number of genes leads to a reduction in the CMR at which 95% of runs lose peak 0 for at least one gene. The magnitude of this reduction is variable, but occurs across all population sizes shown in Figure 4. The population sizes shown represent the steepest part of the curve in Figure 2 before it levels out.

It is expected that biological organisms have evolved to mutate below the CMR; mutation in loss of function alleles will have less of an impact on fitness compared with the same level of mutation in a functional allele. This means peaks of lower fitness and greater mutational robustness can be expected to exist in real life fitness landscapes (independent of the potential effect of epistasis). Real biological organisms therefore have the potential to lose higher fitness peaks at the CMR. There is a lower limit on mutation rate as defined by the drift-barrier hypothesis therefore it is expected biological mutation rates will exist somewhere between this lower limit and the CMR. At some point(s) in parameter space biological mutation rates and CMRs will come close; it is expected mutation rates will be just below the CMR in at least some cases.

Figure 4 shows a drop in CMR in the order of three magnitudes as gene number increases from 1 to 8192. This is a key contribution of the study as it brings the CMR to within an order of magnitude of the biological mutation rates listed in Table 1. The decreasing CMR shown in Figure 4 indicated that increasing the gene number further would bring the CMR directly into the range of biological mutation rates. To test this, population size 10 was run with 25,000 genes of length 1000 or 2000 to bring the gene number and length to within the correct range for *A.thaliana*. This decreased the CMR further to within an order of magnitude of the per base per generation mutation rate for *A.thaliana* (Table 1). Figure 4 also shows per base mutation rate estimates for *C.elegans*, humans, and *D.melanogaster* taken from Table 1, all of which are also within an order of magnitude of the simulation results for 25,000 genes. The mutation rates for *A.thaliana* and *C.elegans* are at or below the predicted CMR while *D.melanogaster* is slightly higher but likely to be below the predicted CMR for a population size greater than 10 based on the trend in Figure 4. This is an important contribution; it is a demonstration that, in a system in which an individual's fitness is dependent on the minimum fitness of its n constituent genes, it is possible to input biologically realistic parameter values for a specific organism into the simulation model and produce a CMR within the range of current biological estimates of mutation rate for that organism.

Bringing the CMR into the biological range is a very important step in the development of an *in silico* model to directly model the evolution of biological species. Future work will require further optimisation of the simulation model to increase run time feasibility. The current study had a high level of neutrality due to the small width of the peaks relative to the size of the adapting sequences. Varying the width of the peaks and distance between them provides a potential future study into the effects of neutrality on the CMR. It should also be noted that eukaryotic organisms such as those discussed here have their DNA organised into chromosomes, for example, the five chromosomes of *A.thaliana* (The Arabidopsis Genome Initiative, 2000). This gap in the current model presents a potential for further development of the model and future study of the effect

of recombination on the CMR. Prediction of the CMR for populations of varying sizes will enable identification of the optimum mutation rate, a crucial parameter in the evolution of small populations where CMR is known to vary significantly (Aston et al., 2013); this has the potential to influence understanding of populations undergoing a bottleneck, under stress, and subsequent conservation strategy for populations on the brink of extinction.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council [grant numbers BB/M020975/1, BB/M021106/1, BB/M021157/1]. The dataset underpinning the results is openly available from Zenodo at http://doi.org/bfcv.

References

Ashburner, M. and Bergman, C. M. (2005). Drosophila melanogaster: A case study of a model genomic sequence and its consequences. *Genome Research*, 15:1661–1667.

Aston, E., Channon, A., Day, C., and Knight, C. G. (2013). Critical mutation rate has an exponential dependence on population size in haploid and diploid populations. *PLoS ONE*, 8(12):e83438.

Baer, C. F., Miyamoto, M. M., and Denver, D. R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, 8:619.

Bevan, M. and Walsh, S. (2005). The Arabidopsis genome: A foundation for plant research. *Genome Research*, 15:1632–1642.

Bryan, G. T., Wu, K., Farrall, L., Jia, Y., Hershey, H. P., McAdams, S. A., Faulk, K. N., Donaldson, G. K., Tarchini, R., and Valent, B. (2000). A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene pi-ta. *The Plant Cell*, 12:2033–2045.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., MÜller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–963.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22:231–238.

Channon, A., Aston, E., Day, C., Belavkin, R. V., and Knight, C. G. (2011). Critical mutation rate has an exponential dependence on population size. In *Advances in Artificial Life, ECAL 2011: Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems*.

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Cutter, A. D., Day, A., and Murray, R. L. (2009). Evolution of the Caenorhabditis elegans genome. *Molecular Biology and Evolution*, 26(6):1199–1234.

Denver, D. R., Morris, K., Lynch, M., and Thomas, W. K. (2004). High mutation rate and predominance of insertions in the caenorhabditis elegans nuclear genome. *Nature*, 430:679–682.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A. Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piégu, B., Ball, S. G., Ral, J. P., Bouget, F. Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., and Moreau, H. (2006). Genome analysis of the smallest free-living eukaryote ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences of the United States of America*, 103(31):11647–52.

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences of the United States of America*, 88:7160–7164.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148:1667–1686.

Durbin, R. M., Altshuler, D., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., and Collins, F. S. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

Eigen, M. and Schuster, P. (1979). *The hypercycle*. Springer, New York.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford University Press.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013). On the immortality

of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*, 5(3):578–90.

Haag-Liautard, C., Coffey, N., Houle, D., Lynch, M., Charlesworth, B., and Keightley, P. D. (2008). Direct estimation of the mitochondrial dna mutation rate in drosophila melanogaster. *PLoS Biology*, 6(8):e204.

Keightley, P. D., Ness, R. W., Halligan, D. L., and Haddrill, P. R. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a drosophila melanogaster full-sib family. *Genetics*, 196(1):313–320.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009). Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19:1195–1201.

Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):803–808.

Lewin, B. (2008). *Genes IX*. Jones and Bartlett Learning.

Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K.-H., and Zhao, H. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific Reports*, 5:10576.

Lynch, M. (2010a). Evolution of the mutation rate. *Trends in Genetics*, 26:345–352.

Lynch, M. (2010b). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(37):16013–16015.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., and Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27):9272–9277.

Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304.

Nam, J. and Bartel, D. P. (2012). Long noncoding RNAs in c. elegans. *Genome Research*, 22:2529–2540.

Nowak, M. A. (1992). What is a quasispecies? *Trends in Ecology and Evolution*, 7:118–121.

Ochoa, G. (2006). Error thresholds in genetic algorithms. *Evolutionary Computation*, 14(2):157–182.

Ochoa, G., Harvey, I., and Buxton, H. (2000). Optimal mutation rates and selection pressure in genetic algorithms. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2000)*.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*, 327:92–94.

Ramkumar, G., Biswal, A. K., Madhan Mohan, K., Sakthivel, K., Sivaranjani, A. K. P., Neeraja, C. N., Ram, T., Balachandran, S. M., Sundaram, R. M., Prasad, M. S., Viraktamath, B. C., and Madhav, M. S. (2010). Identifying novel alleles of rice blast resistance genes pikh and pita through allele mining. *International Rice Research Notes*.

Sharma, V. K., Brahmachari, S. K., and Ramachandran, S. (2005). (TG/CA)n repeats in human gene families: abundance and selective patterns of distribution according to function and gene length. *BMC Genomics*, 6:83.

Sniegowski, P. and Raynes, Y. (2013). Mutation rates: How low can you go? *Current Biology*, 23(4):R147–R149.

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(45):18488–18492.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408:796–815.

Ward, L. D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337(6102):1675–8.

Whitlock, M., Griswold, C., and Peters, A. (2003). Compensating for the meltdown: The critical effective size of a population with deleterious and compensatory mutations. *Annales Zoologici Fennici*, 40:169–183.

Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, 5:44.

Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., Asan, MacArthur, D. G., Quail, M. A., Carter, N. P., Yang, H., and Tyler-Smith, C. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology*, 19:1453–1457.