biology letters
Molecular evolution

Biol. Lett. (2009) **5**, 397–400
doi:10.1098/rsbl.2009.0104
Published online 25 March 2009

Modelling mitochondrial site polymorphisms to infer the number of segregating units and mutation rate

Michael D. Hendy[1,*], Michael D. Woodhams[1,†] and Andrew Dodd[2,‡]

[1]*Allan Wilson Centre, Massey University, Palmerston North 4474, New Zealand*
[2]*Allan Wilson Centre, Massey University, Auckland 0745, New Zealand*
**Author for correspondence (m.hendy@massey.ac.nz).*
[†]*Present address: School of Information Technologies, University of Sydney, Sydney, New South Wales 2006, Australia.*
[‡]*Present address: Summit plc, Abingdon, Oxfordshire OX14 4RY, UK.*

We present a mathematical model of mitochondrial inheritance evolving under neutral evolution to interpret the heteroplasmies observed at some sites. A comparison of the levels of heteroplasmies transmitted from mother to her offspring allows us to estimate the number N_x of inherited mitochondrial genomes (segregating units). The model demonstrates the necessity of accounting for both the multiplicity of an unknown number N_x, and the threshold θ, below which heteroplasmy cannot be detected reliably, in order to estimate the mitochondrial mutation rate μ_m in the maternal line of descent. Our model is applicable to pedigree studies of any eukaryotic species where site heteroplasmies are observed in regions of the mitochondria, provided neutrality can be assumed. The model is illustrated with an analysis of site heteroplasmies in the first hypervariable region of mitochondrial sequence data sampled from Adélie penguin families, providing an estimate N_x and μ_m. This estimate of μ_m was found to be consistent with earlier estimates from ancient DNA analysis.

Keywords: mitochondrial DNA; mutation rate; segregating units; pedigree analysis; Adélie penguins

1. INTRODUCTION

It has recently been found that estimates of the human mitochondrial mutation rate from pedigree data are many times higher than those estimated from sequence divergence at putatively neutral sites (e.g. Parsons *et al.* 1997; Howell *et al.* 2003; Santos *et al.* 2005). One possible reason for this is that heteroplasmy (variation among different organelle genomes within the same individual) may be maintained for many generations, after the origin by mutation of a new variant, provided that there is not a very small bottleneck in the number of organelle genomes during transmission from parent to offspring (reviewed by Birky 1991). The persistence of heteroplasmic mutations for many generations after their origin may lead to an inflated contribution to a pedigree-based estimate of mutation rate, if they are treated as mutations that have become fixed within individuals. We propose a method for estimating the mutation rate and size of the transmission bottleneck for maternally transmitted mitochondrial genomes. We illustrate this using the dataset of Millar *et al.* (2008) on the Adélie penguin.

Millar *et al.* (2008) sequenced a segment of 344 base pairs in a fast evolving region of first hypervariable region (a region they argued was not under strong selective pressure) from 1931 Adélie penguins (508 families) at Antarctic nesting sites. This was a follow-up to the study of Lambert *et al.* (2002) of ancient and contemporary samples of the same region of mitochondrial DNA. In the data reported in Millar *et al.* (2008), a low proportion of the sites in individual birds exhibited heteroplasmies, where two bases were called on the electropherogram at that site, each with a signal greater than $\theta = 0.23$ of the total signal. Signals below the detection threshold θ, could not be reliably distinguished from background noise and were not reported there. Fifty-five mothers (out of 508) exhibited site heteroplasmies (above θ), with seven having two heteroplasmic sites. Hence, the proportion of sites with an observed heteroplasmy across all mothers was

$$\hat{\beta} = \frac{48 + 2 \times 7}{508 \times 344} = 3.55 \times 10^{-4}. \quad (1.1)$$

In all but three cases, the observed site heteroplasmies persisted above the threshold in both mother and chicks. Site heteroplasmies in a chick were only checked at sites where a heteroplasmy had been observed in its mother.

Following Lambert *et al.* (2002) and Millar *et al.* (2008), we define mutation rate as the rate at which a base substitution is incorporated into all mitochondrial genomes of an individual. Previous studies linking mutation rate with observed site heteroplasmies include two studies (Brandstätter *et al.* 2004; Santos *et al.* 2005) of human populations of up to three generations, where the mutation rate was estimated from assuming that each observed site heteroplasmy represented a transitional substitution. We find most of the substitutions giving rise to an observed site heteroplasmy are subsequently lost, and those that eventually become fixed may persist as an observed site heteroplasmy for many generations and be oversampled.

For our study, we follow the maternal ancestry of an individual, tracking the trajectory of a site substitution introduced into the germ line by some ancestor. At each generation, the population of mitochondrial genomes will pass through a bottleneck when the oocyte segregates a small number of its mother's genomes. Lambert *et al.* (2002) argued that this region of the mt-genome is not under strong selective pressure, so our model assumes random drift among the N_x segregating genomes.

This model should be applicable for any region of the mt-genome assumed neutral under selection for any eukaryote species.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2009.0104 or via http://rsbl.royalsocietypublishing.org.

One contribution of 11 to a Special Feature on 'Whole organism perspectives on understanding molecular evolution'.

Received 8 February 2009
Accepted 27 February 2009

2. MATERIAL AND METHODS

Our model focuses on an individual site of the mitochondrial genomes in the maternal ancestry of one individual. When a heteroplasmy (comprising two different nucleotides, generically X, Y) is observed at a site, we presume, this is a consequence of a somatic substitution X→Y or Y→X at that site in a maternal ancestor some g generations prior, inherited by a daughter, which has persisted for g generations.

Consider the maternal ancestry of an individual A_0, with A_1 its mother and A_2 its maternal grandmother, etc. Suppose a nucleotide substitution X→Y has occurred at a site in one genome of a maternal line ancestor A_g ($g\geq 1$), which is inherited by A_{g-1}. Our model shows that the probability of an additional substitution inherited at that site among $A_{g-1}, \ldots, A_0$ is less than 10^{-2}, so we will neglect this possibility. Suppose this site heteroplasmy persists for (exactly) $k\geq 1$ generations, inherited by $A_{g-1}, \ldots, A_{g-k}$, but not by A_{g-k-1}. If $k\geq g$, then A_0 will be heteroplasmic at that site, although not necessarily observable. If $k<g$, $A_{g-k-1}, \ldots, A_0$ are not heteroplasmic at that site, with their genomes either containing all X or all Y.

We propose a model where each oocyte recruits N_x mitochondrial genomes independently from the population of its mother's genomes. (The number N_x is sometimes referred to as the *number of segregating units*. For the Adélie penguins, we estimated N_x to have a 95% confidence interval (CI) of $25<N_x<69$.)

The observed levels of most site heteroplasmies from the blood samples of a mother and her chicks closely agree, which reflects a close agreement in the germ line. We will assume that the variation, after accounting for measurement uncertainty, is due to sampling at the recruiting bottleneck and that the proportions in the blood sample estimate the inherited proportions.

If A_1 (the mother of A_0) contains a site heteroplasmy with the nucleotides Y and X appearing in proportions ϕ and $1-\phi$, then the probability (using a binomial selection with replacement) that A_0 inherits i genomes with allele Y and N_x-i genomes of allele X is

$$\Pr(i \mid N_x, \phi) = \binom{N_x}{i} \phi^i (1-\phi)^{N_x-i}.$$

If A_1 had inherited j copies of allele Y and N_x-j of allele X from her mother A_2, we assume she exhibits the proportions $\phi=j/N_x$ and $1-\phi=1-j/N_x$ of the alleles in the genomes available for inheritance. Hence,

$$p_{N_x}(i,j) = \binom{N_x}{i} \left(\frac{j}{N_x}\right)^i \left(\frac{N_x-j}{N_x}\right)^{N_x-i}, \tag{2.1}$$

is the probability that A_0 inherits i copies of allele Y, and N_x-i copies of allele X, given her mother had inherited j and N_x-j corresponding copies. Let

$$P_{N_x} = [p_{N_x}(i,j)], \tag{2.2}$$

be the matrix of these probabilities, where the N_x-1 rows and columns are indexed by $i,j\in\{1,2,\ldots,N_x-1\}$.

Given that A_g has a somatic mutation in a descendant of one of her N_x founding genomes, the probability that A_{g-1} inherits more than one mutated genome is very small. Hence, we will assume A_{g-1} is heteroplasmic at that site, with proportion $1/N_x$ of its genomes containing Y. If the heteroplasmy is lost g generations later, then A_0 has either all X or all Y at that site. Let $h_{X,Y}$ be the proportion of cases where the heteroplasmy persists and h_X and h_Y be the proportions where the mutation is lost or fixed. The neutral model predicts that as g increases

$$h_{X,Y}\to 0, \qquad h_X \to \frac{N_x-1}{N_x}, \qquad h_Y \to \frac{1}{N_x}.$$

In a simulation study of 10^7 heteroplasmic site histories, we followed the introduction of one mutation until the site heteroplasmy was lost, for $N_x=20$ and for $N_x=40$. We found (table 1) for $\theta=0.23$ that $h_X\approx((N_x-1)/N_x)$ and $h_Y\approx 1/N_x$. Table 1 also gives the average numbers of generations that the site heteroplasmies persist, and are observable. We note that over all histories, the average numbers of generations a site heteroplasmy was observable were almost identical for $N_x=20$ and 40, but the average variation in the levels of a heteroplasmy at a site between a mother and her chick differed significantly.

For each A_k in the ancestry, let n_k be the number of its founding genomes with nucleotide Y at that site. We have assumed that $n_{g-1}=1$. If for $k<g-1$, $n_k=0$ or N_x, the heteroplasmy is lost. Suppose $1\leq n_{k+1}=j<N_x$, then the probability that $n_k=i$, $(1\leq i<N_x)$ is

$$\Pr(n_k = i \mid n_{k+1} = j) = (P_{N_x})_{i,j}.$$

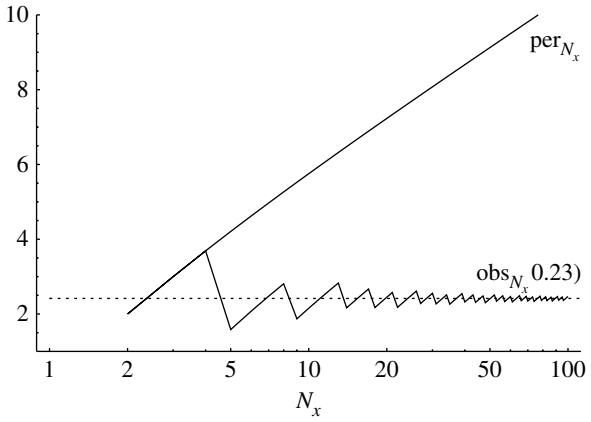

Figure 1. Values of $(Q_{N_x})_{i,1}$ (written as per_{N_x}), the expected number of generations a site heteroplasmy persists, and of $\mathrm{obs}_{N_x}(0.23)$, the expected number of generations the site heteroplasmy are observable with threshold $\theta=0.23$ (plotted on a logarithmic scale). Note that $\mathrm{obs}_{N_x}(\theta)$ converges to the constant value $2\ln(\theta^{-1}-1)=2.417$ (dotted line), while per_{N_x} (solid line) grows logarithmically as N_x increases.

In lemma 1 of the electronic supplementary material, we show that

$$\Pr(n_k = i \mid n_{g-1} = 1) = (P_{N_x})_{i,1}^{(g-k-1)},$$

which is the ith entry of the leading column of $P_{N_x}^{(g-k-1)}$. Assuming that the probability that A_g introduces a somatic mutation into the germ line at a selected site is α, then summing over all generations $g\geq 1$, the probability that A_0 has a site heteroplasmy with $n_0=i$ $(1\leq i<N_x)$ is

$$\Pr(n_0 = i) = \sum_{g\geq 1} \alpha (P_{N_x})_{i,1}^{(g-1)} = \alpha (Q_{N_x})_{i,1}, \tag{2.3}$$

where $(Q_{N_x})_{i,1}$ is the first entry in the ith row of

$$Q_{N_x} = \sum_{g\geq 1} P_{N_x}^{(g-1)} = (I - P_{N_x})^{-1}.$$

As $(Q_{N_x})_{i,1}$ is the expected number of generations that a new site heteroplasmy persists with i copies (figure 1), the expected number of generations a heteroplasmy is observable at that site is

$$\mathrm{obs}_{N_x} = \sum_{\theta\leq i/N_x\leq 1-\theta} (Q_{N_x})_{i,1}.$$

Most site heteroplasmies never reach the detection threshold, those that do, usually persist for many more than obs_{N_x} generations (table 1).

In figure 2, we plot the values of $(Q_{20})_{i,1}$ and $(Q_{40})_{i,1}$, noting that these two distributions are almost identical, and that $(Q_{N_x})_{i,1}$ is closely approximated by $2/i$ within the observed region. (The limit $(Q_{N_x})_{i,1}\to 2/i$ as $N_x\to\infty$ was noted by Fisher and Wright (Ewens 2004, eqn (1.56)).)

We show in lemma 2 of the electronic supplementary material, that assuming $Q_{N_x}(i)\approx(2/i)$, the probability β that a site has an observable heteroplasmy is closely approximated by $2\alpha\ln(\theta^{-1}-1)$. Assuming neutral evolution, $1/N_x$ of the substitutions entering the germ line will become fixed in the maternal line of descent, so that the mutation rate can be estimated as

$$\mu_{\mathrm{m}} = \alpha/N_x t \approx \frac{\beta}{2N_x t\ln(\theta^{-1}-1)}, \tag{2.4}$$

where t is the generation time.

3. RESULTS

We have shown that majority of heteroplasmies cannot be observed, leading to undersampling; but that those that do may persist for many generations, leading to oversampling. These two effects do not balance. We have demonstrated that the estimate of the mutation rate from the density of observed heteroplasmic sites is dependent both on the number of segregating units N_x and on the detection threshold θ.

Table 1. Site heteroplasmy histories: results from 10^7 simulations for $N_x=20$ and 40 segregating units. (In more than 99.9% of the histories, the introduced mutation is either lost or fixed within 200 generations, and no site heteroplasmy survived more than 520 generations. Statistics are presented for the cases HX_n (X→Y lost and never observed at $\theta=0.23$), HX_o (lost but observable for at least one generation) and HY (the mutation X→Y is ultimately fixed). We note that the proportions fixed (HY) is close to its expected value of $1/N_x$. g_{av} records the mean number of generations the site heteroplasmy survives, and g_{av}(obs) records the mean number of generations the site heteroplasmy is in the observable range. Note as N_x is doubled, the g_{av}(obs) values double, but the proportions are halved so the mean number of generations for which a site heteroplasmy is observable is approximately constant, close to 2.30 in both cases. In the final row, we note the average mean square difference between the mother/chick pairs over all observed heteroplasmic sites differs for $N_x=20$ and $N_x=40$.)

	$N_x=20$			$N_x=40$		
category	proportion	g_{av}	g_{av}(obs)	proportion	g_{av}	g_{av}(obs)
HX_n	82.47%	3.87	—	91.05%	4.94	—
HX_o	12.54%	24.26	10.27	6.46%	50.06	19.36
HY	4.99%	37.18	20.75	2.48%	77.54	40.93
all	100.00%	8.09	2.32	100.00%	9.66	2.27
obs. M/C	av. diff. = 8.49%			av. diff. = 5.98%		

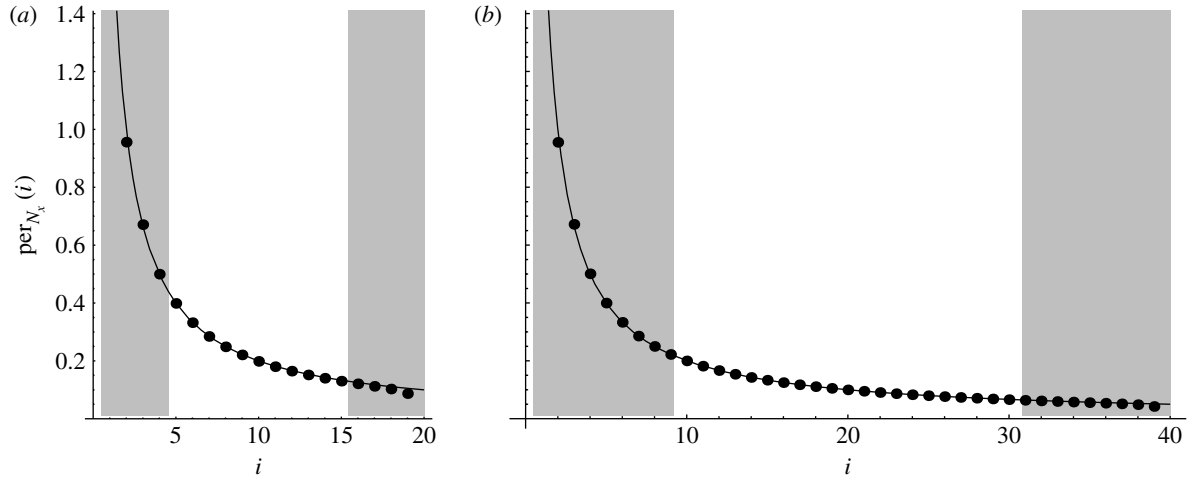

Figure 2. The values of $per_{20}(i)$, (*a*) for $i=2, \ldots, 19$, and $per_{40}(i)$, (*b*) for $i=2, \ldots, 39$, are plotted as dots. The solid lines are the curves $2/i$. The shaded regions are outside the detection threshold for $\theta=0.23$ with the observable region (unshaded) between them. We note that in both observed regions the points $per_{N_x}(i)$ fit the curve $2/i$ closely.

For the Adélie penguin data of Millar *et al.* (2008), with $\theta=0.23$, $t=6.46$ and $\hat{\beta}=3.55\times10^{-4}$, we used Bayesian analysis (see 'Estimating N_x from mother–chick comparisons' of the electronic supplementary material) to obtain a maximum-likelihood estimate of $N_x=36.5$, with 95% CI of $25.0\leq N_x\leq 66.9$. Using equation (2.4), Millar *et al.* (2008) estimated the median value to be $\mu_m=0.55$ s s^{-1}Myr^{-1} (CI $0.29\leq\mu_m\leq0.88$), a value not significantly different from the ancient evolutionary rate of $k=0.86$ s s^{-1}Myr^{-1} (CI $0.53\leq k\leq1.17$) estimated by Lambert *et al.* (2002) for the same region. In their study, Millar *et al.* showed that if they had assumed that each observed heteroplasmy represented a mutation in transition, the estimate of μ would have been increased by a factor of nearly 100.

4. DISCUSSION

This model has assumed neutral evolution for the regions of mitochondria genome under analysis. Whether specific regions are under selective pressure is outside the scope of this study; however, Rand *et al.* (1994), for example, addressed this issue. Our model is applicable wherever the assumption of neutrality can be assumed for pedigree studies of any species where multiple copies of the mitochondrial genome are inherited, and a sufficient number of heteroplasmies are observed.

Accounting for the effects modelled here may illuminate the apparent discrepancies reported between molecular substitution rates and those estimated from pedigree studies, such as for human studies, e.g. Parsons *et al.* (1997), Howell *et al.* (2003) and Santos *et al.* (2005). As these studies do not report observational thresholds, our model cannot be applied directly to their data.

Improvement in the accuracy of determining the relative expression levels of site heteroplasmies (with $\theta\ll1/N_x$) might lead to a direct estimate of N_x. However, it is likely that N_x may vary across the sample, in which case the distribution would not clump around the i/N_x values. We have shown (by simulation) that the model is robust against moderate variations in N_x, provided we take N_x to be an idealized harmonic mean of the individual values in the sample.

We gratefully acknowledge support from the New Zealand Centres of Research Excellence Fund. We thank the following people who assisted in the development of this paper: J. Esins, D. M. Lambert, C. D. Millar, D. Penny, and K. Schliep. We thank the editor and referees for their very helpful comments in improving the presentation.

Birky, C. W. 1991 Evolution and population genetics of organelle genes. In *Evolution at the molecular level* (eds R. K. Selander, A. G. Clark & T. S. Whittam), pp. 112–134. Sunderland, MA: Sinauer. (See also pp. 202–221.)

Brandstätter, A., Niederstätter, H. & Parson, W. 2004 Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed base-calls in the entire human mitochondrial control region. *Int. J. Legal Med.* **118**, 47–54. (doi:10.1007/s00414-003-0418-z)

Ewens, W. J. 2004 *Mathematical population genetics*. New York, NY: Springer.

Howell, N., Smejkal, C. B., Mackey, D. A., Chinnery, P. F., Turnbull, D. M. & Herrnstadt, C. 2003 The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am. J. Hum. Genet.* **72**, 659–670. (doi:10.1086/368264)

Lambert, D. M., Ritchie, P. A., Millar, C. D., Holland, B., Drummond, A. J. & Baroni, C. 2002 Rates of evolution in ancient DNA from Adélie penguins. *Science* **295**, 2270–2273. (doi:10.1126/science.1068105)

Millar, C. D., Dodd, A., Anderson, J., Gibb, G. C., Ritchie, P. A., Baroni, C., Woodhams, M., Hendy, M. D. & Lambert, D. M. 2008 Mutation and evolutionary rates in Adélie penguins from the Antarctic. *PLoS Genet.* **4**, e000209. (doi:10.1371/journal.pgen.1000209)

Parsons, T. J. *et al.* 1997 A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* **15**, 363–368. (doi:10.1038/ng0497-363)

Rand, D. M., Dorfsman, M. & Kann, L. M. 1994 Neutral and non-neutral evolution of *Drosophila* mitochondrial DNA. *Genetics* **138**, 741–756.

Santos, C., Monteil, R., Sierra, B., Bettencourt, C., Fernandez, E., Alvarez, L., Lima, M., Abade, A. & Aluja, M. P. 2005 Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: a model using families from the Azores Islands (Portugal). *Mol. Biol. Evol.* **22**, 1490–1505. (doi:10.1093/molbev/msi141)