Expected coalescence times and segregating sites in a model of glacial cycles

F.F. Jesus[1], J.F. Wilkins[2], V.N. Solferini[1] and J. Wakeley[3]

[1]Departamento de Genética e Evolução, Instituto de Biologia, Unicamp, Campinas, SP, Brasil
[2]Santa Fe Institute, Santa Fe, NM, USA
[3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
Corresponding author: F.F. Jesus
E-mail: flaviaf.jesus@gmail.com/ffjesus@unicamp.br

ABSTRACT. The climatic fluctuations of the Quaternary have influenced the distribution of numerous plant and animal species. Several species suffer population reduction and fragmentation, becoming restricted to refugia during glacial periods and expanding again during interglacials. The reduction in population size may reduce the effective population size, mean coalescence time and genetic variation, whereas an increased subdivision may have the opposite effect. To investigate these two opposing forces, we proposed a model in which a panmictic and a structured phase alternate, corresponding to interglacial and glacial periods. From this model, we derived an expression for the expected coalescence time and number of segregating sites for a pair of genes. We observed that increasing the number of demes or the duration of the structured phases causes an increase in coalescence time and expected levels of genetic variation. We compared numerical results with the ones expected for a panmictic population of constant size, and showed that the mean number of segregating sites can be greater in our model even

when population size is much smaller in the structured phases. This points to the importance of population structure in the history of species subject to climatic fluctuations, and helps explain the long gene genealogies observed in several organisms.

Key words: Refugia, Coalescent theory, Population structure, Pairwise differences

INTRODUCTION

The climatic fluctuations of the Quaternary, with cycles of glacial and interglacial periods, have affected the distribution of several organisms (Webb and Bartlein, 1992; Burnham and Graham, 1999). Many plant and animal species are subject to population reduction and fragmentation, becoming restricted to refugia during glacial periods and expanding again during the interglacials. This affects both population size and structure, in a cyclic fashion.

Demographic changes may be detectable in DNA sequences. In fact, there have been a great number of studies based on sequence or polymorphism data from species that are thought to have undergone these cyclic changes (e.g., Hugall et al., 2002; Lessa et al., 2003; Petit et al., 2003; for reviews, see Hewitt, 2000, 2004). In several of these, the questions of interest are the existence and number of refugia, as well as their importance for current genetic variation.

From the theoretical side there have been numerous studies on the genetic consequences of population size change (e.g., Wright, 1938; Nei et al., 1975; Slatkin and Hudson, 1991) and on the importance of population structure (for review, see Charlesworth et al., 2003). Also, some authors have included subdivision in models of population expansion or colonization (Austerlitz et al., 1997; Ray et al., 2003; Excoffier, 2004). However, none of them have taken into account cyclic changes in both population size and structure. Such a model is necessary if we are to investigate the genetic consequences of cyclic climatic fluctuations.

One question of interest is whether the demographic changes caused by these climatic fluctuations lead to an increase or a decrease in coalescence times and, consequently, in levels of genetic variation. The reduction in population size is expected to reduce the effective population size, mean coalescence time and overall genetic variation. On the other hand, the increase in population subdivision may cause a stretch of the mean coalescence times and an increase in genetic variation.

In the present study, we investigated coalescence times and genetic variation in a model of repeated changes in population size and structure. We used a coalescent approach (Kingman, 1982a,b; Hudson, 1983; Tajima, 1983), and obtained expressions for the expected coalescence time and the number of segregating sites of two sequences, sampled in the present interglacial period. We considered the effects of the opposing forces of population reduction and increased structure, focusing specifically on what conditions result in an overall increase or decrease in mean coalescence times and genetic variation.

———————————————

MODEL AND THEORY

In this model the population undergoes cycles composed of two phases: a panmictic phase, corresponding to the interglacial periods, and a structured phase, corresponding to the glacial periods (Figure 1). We assumed that the species is favored by the interglacials, and therefore has a larger and more continuous population during those times, while having a smaller and more fragmented population during the glacial periods.
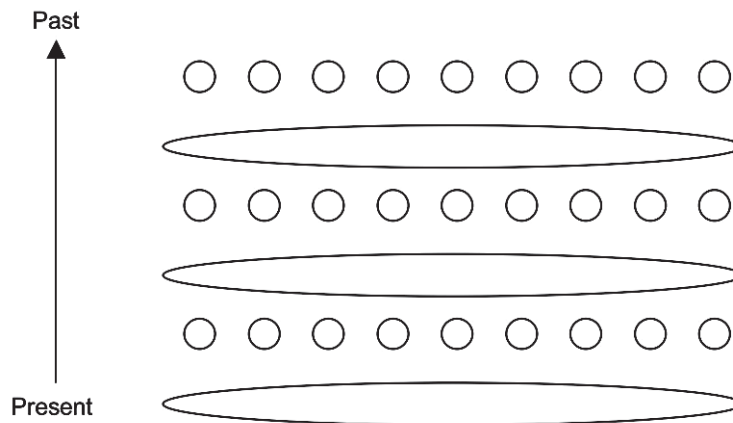

Figure 1. Schematic representation of the model, with alternating panmictic and structured phases corresponding to interglacial and glacial periods, respectively. The current phase is an interglacial.

Reproduction is according to the Wright-Fisher model (Fisher, 1930; Wright, 1931), both in the panmictic phase and within each deme during the structured phase. We considered a haploid organism, but the conclusions can be extended to diploid organisms. We also made the usual coalescent assumption that the population or deme size is much larger than the number of sequences sampled (see Hudson, 1990).

The model has five parameters, which are the population size during the panmictic phases (N_p), the duration in generations of each panmictic phase (t_p), the number of demes in the structured phases (D), the deme size (N_S), and the duration in generations of each structured phase (t_S). We assumed that during the structured phases the population is subdivided into D demes of equal size (N_S), with no migration between them. In the transition from a panmictic to a structured phase, backwards in time, the lineages are randomly distributed among the D demes.

We have derived an expression (see Appendix) for the expected coalescence time (E[coal. time]) of two genes sampled in the present interglacial:

$$E[\text{coal. time}] = \left[\frac{D}{D - (D-1)e^{-t_p/N_p} - e^{-(t_p/N_p + t_S/N_S)}} \right] \quad \text{(Equation 1)}$$

$$\cdot \left\{ N_p.(1 - e^{-t_p/N_p}) + e^{-t_p/N_p}[(1 - 1/D).t_s + (1/D).N_S.(1 - e^{-t_S/N_S})] \right\}$$

We were also interested in levels of genetic variation, and one measure which is largely used is the number of segregating - or polymorphic - sites in a sample of genes. The expectation of this number is a simple function of the mean coalescence time if we assume no intra-genic recombination and a simple mutation model. Assuming an infinite sites mutation model (Kimura, 1969), the expected number of segregating sites in a sample of two genes is, simply:

$$E[S] = 2\mu.E[\text{coal. time}]$$
(Equation 2)

where μ is the mutation rate per sequence per generation, and $E[\text{coal. time}]$ is obtained from Equation 1. The mean number of segregating sites, in this case, is also the mean number of pairwise differences.

RESULTS AND DISCUSSION

We started by studying the influence of each parameter of the structured phases on coalescence times and genetic variation. Since we are interested in habitat reduction during glacial periods, we focused our attention to cases in which both $N_S < N_P$ and $D.N_S < N_P$, i.e., when both the deme size and the total population size in the structured phases - corresponding to the glacial periods - are smaller than the population size in the panmictic phases.

In Figure 2, we plotted the mean number of segregating sites as function of the time (t_S), deme size (N_S) and number of demes (D) of the structured phases. We assumed a mutation rate of $\mu = 10^{-5}$ and fix the parameters of the panmictic phases as $N_P = 100,000$ and $t_P = 10,000$. The other parameter values for each plot are given in the figure legend.
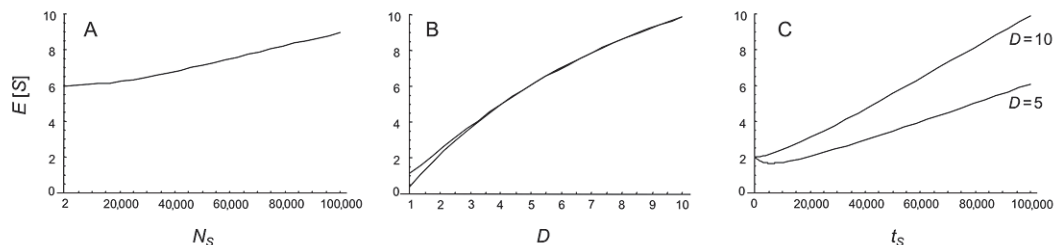

Figure 2. Mean number of segregating sites ($E[S]$) as a function of the following parameters of the structured phase: **A.** size of each deme (N_S), with $t_S = 100,000$ and $D = 5$; **B.** Number of demes (D), with $t_S = 100,000$ and $N_S = 10,000$ (lower line) or $D.N_S = 50,000$ (upper line), and **C.** Number of generations (t_S), with $N_S = 10,000$ and $D = 5$ (lower line) or $D = 10$ (upper line). In all cases $N_P = 100,000$; $t_P = 10,000$; $\mu = 10^{-5}$.

As expected, a reduction in population size during the glacial phases will tend to reduce the mean coalescence times and consequently the mean number of segregating sites (Figure 2A). The increase in subdivision, on the other hand, tends to increase the mean coalescence times and the mean number of segregating sites.

In Figure 2B we observe that the mean number of segregating sites increases with the number of demes. We plotted the curve for two different cases: keeping the value of N_S constant or keeping the value of $D.N_S$ constant. In the first case, as the number of demes increases, the total population size ($D.N_S$) also increases. In the second case, the total population size is kept constant, and increasing D means subdividing the population into smaller demes. It is inter-

esting to observe that both curves behave almost identically. Especially, that a greater number of demes, but with smaller size each, will still cause a large increase in coalescence time and genetic variation. We have investigated this with other parameter values, and a similar behavior was observed. This indicates that the number of demes is important in maintaining genetic diversity during periods of population reduction - when migration between them is negligible - even when each deme is of small size.

The duration of the structured phases also influences coalescence time and genetic variation. In Figure 2C we observe that an increase in t_S generally causes an increase in the mean number of segregating sites. In the case where $D = 5$, there is a small range of the plot where the mean number of segregating sites decreases with increasing t_S. However, it increases in most of the plot when $D = 5$, and throughout the plot for larger values of D (shown here for $D = 10$).

Having studied the influence of each parameter of the glacial phases, we now turn to a study of the conditions in which the coalescence times and genetic variation are increased or decreased, in comparison to what is expected for a panmictic population of constant size N_P. With the values chosen ($N_P = 100,000$, $\mu = 10^{-5}$), the expected number of segregating sites would be $E[S] = 2$.

In Figure 3 we plotted $E[S]$ as a function of the number of demes (D) and the total population size during the structured phases ($D.N_S$). Figure 3A is a three-dimensional plot, and Figure 3B is a contour-plot with two levels: above and below $E[S] = 2$. We observe that when $D \geq 2$, the mean number of segregating sites is greater than two, that is, greater than what would be expected in the case where the population is always panmictic with constant size N_P. Even when the total population size and the deme size in the glacial phases are quite small (Figure 3B, bottom), still the expected genetic variation is increased if the population is subdivided into two or more demes.
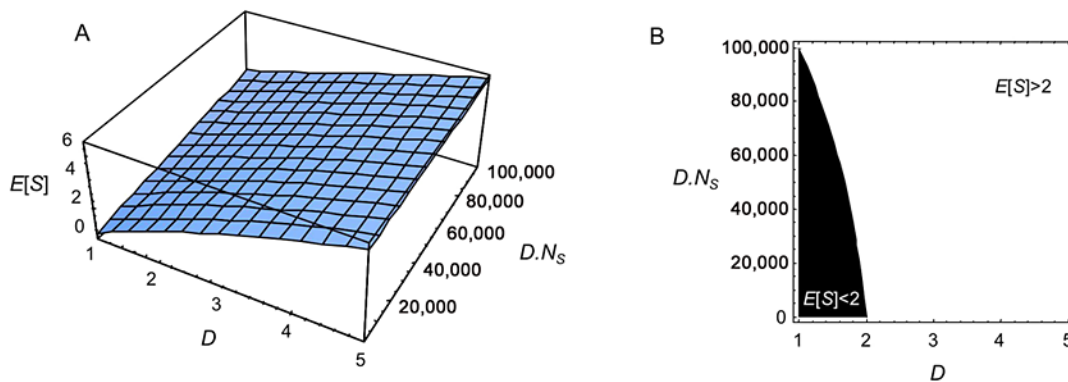

Figure 3. Mean number of segregating sites ($E[S]$) as a function of the number of demes (D) and the total population size ($D.N_S$) during the structured phases, with $N_P = 100,000$; $t_P = 10,000$; $t_S = 100,000$, and $\mu = 10^{-5}$. **A.** Three-dimensional plot; **B.** Contour-plot with the region where $E[S]<2$ in black and $E[S]>2$ in white.

We also plotted the mean number of segregating sites as a function of the duration (t_S) and deme size (N_S) during the structured phases, fixing the number of demes at $D = 2$ (see Figure 4). We observe that, when t_S is large, the mean number of segregating sites is greater than two even with very small values of N_S (Figure 4B, bottom right corner). Paleoclimatic

studies support the notion that the glacial periods lasted much longer (~80,000-90,000 years) than the interglacials (~10,000) (Petit et al., 1999; Jouzel, 2003). Therefore, in species with one generation/year, for instance, such large values of t_S are highly plausible.
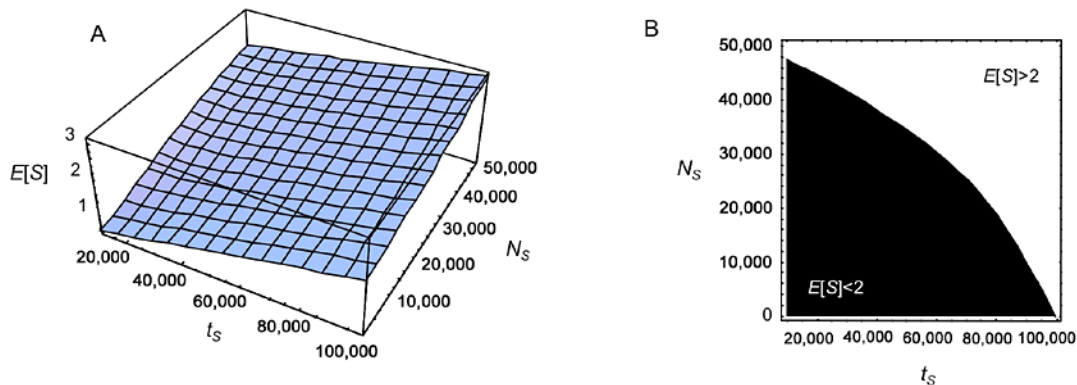

Figure 4. Mean number of segregating sites ($E[S]$) as a function of the time in generations (t_S) and the size of each deme (N_S) during the structured phases, with $N_P = 100,000$; $t_P = 10,000$; $D = 2$, and $\mu = 10^{-5}$. **A.** Three-dimensional plot; **B.** Contour-plot with the region where $E[S]<2$ in black and $E[S]>2$ in white.

Most molecular studies of species which have undergone the cyclic changes of the Quaternary have focused on the genetic consequences of the last cycle only. However, in several species the most recent common ancestor at some loci predates the last glacial period, sometimes tracing back even to the Pliocene (>2 mya) (e.g., Paulo et al., 2001; Hewitt, 2004; Smith and Farrell, 2005). The results obtained here suggest that population subdivision, especially during the long glacial periods, is a possible explanation for these long gene genealogies.

Population subdivision is known to cause an increase in genetic variation at the species level. Based on the island model, for instance, Wright (1951) concluded that a subdivided population is capable of preserving a greater number of alleles than a well-mixed population of the same size. Nei and Takahata (1993), studying mean coalescence times in a finite island-model, found that the effective size of a species can be larger when migration rates are low. In the present study, we have shown that, even when the increase in population structure - by means of subdivision - is accompanied by a simultaneous reduction in population size, it can still cause an increase in expected coalescence time and levels of genetic variation.

The model presented here has some quite unrealistic assumptions, such as the total absence of population subdivision during interglacial periods as well as the absence of migration during the glacial periods. The investigation of a more general model, where these assumptions are relaxed, is the subject of an ongoing study. Nonetheless, the results obtained with the present model are useful in pointing to the importance of population subdivision in the history of species subject to climatic cycles, and offer a potential explanation for some of the observed patterns in molecular data.

ACKNOWLEDGMENTS

REFERENCES

Austerlitz F, Jung-Muller B, Godelle B and Gouyon P (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theor. Popul. Biol.* 51: 148-164.

Burnham RJ and Graham A (1999). The history of Neotropical vegetation: new developments and status. *Ann. MO Bot. Gard.* 86: 546-589.

Charlesworth B, Charlesworth D and Barton NH (2003). The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Ecol. Evol. Syst.* 34: 99-125.

Excoffier L (2004). Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol. Ecol.* 13: 853-864.

Fisher RA (1930). The genetical theory of natural selection. Clarendon, Oxford, England.

Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* 405: 907-913.

Hewitt GM (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. R. Soc. Lond. B* 359: 183-195.

Hudson RR (1983). Testing the constant rate neutral allele model with protein sequence data. *Evolution* 37: 203-217.

Hudson RR (1990). Gene genealogies and the coalescent process. In: Oxford surveys in evolutionary biology (Futuyma D and Antonovics J, eds.). Vol. 7. Oxford University Press, Oxford, England, 1-44.

Hugall A, Moritz C, Moussalli A and Stanisic J (2002). Reconciling paleodistribution models and comparative phylogeography in the wet tropics rainforest land snail *Gnarosophia bellendenkerensis* (Brazier 1875). *Proc. Natl. Acad. Sci. USA* 99: 6112-6117.

Jouzel J (2003). Past climate (the last 400 ka): from geological times to future climate change. *C. R. Geoscience* 335: 509-524.

Kimura M (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61: 893-903.

Kingman JFC (1982a). The coalescent. *Stochastic Processes Applications* 13: 235-248.

Kingman JFC (1982b). On the genealogy of large populations. *J. Appl. Probab.* 19A: 27-43.

Lessa EP, Cook JA and Patton JL (2003). Genetic footprints of demographic expansion in North America, but not Amazonia, during the Late Quaternary. *Proc. Natl. Acad. Sci. USA* 100: 10331-10334.

Nei M and Takahata N (1993). Effective population size, genetic diversity, and coalescence time in subdivided populations. *J. Mol. Evol.* 37: 240-244.

Nei M, Maruyama T and Chakraborty R (1975). The bottleneck effect and genetic variability in populations. *Evolution* 29: 1-10.

Paulo OS, Dias C, Bruford MW, Jordan WC, et al. (2001). The persistence of Pliocene populations through the Pleistocene climatic cycles: evidence from the phylogeography of an Iberian lizard. *Proc. R. Soc. Lond. B Biol. Sci.* 268: 1625-1630.

Petit JR, Jouzel J, Raynaud D, Barkov NI, et al. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399: 429-436.

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, et al. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* 300: 1563-1565.

Ray N, Currat M, Excoffier L (2003). Intra-deme molecular diversity in spatially expanding populations. *Mol. Biol. Evol.* 20: 76-86.

Slatkin M and Hudson RR (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555-562.

Smith CI and Farrell BD (2005). Range expansions in the flightless longhorn cactus beetles, *Moneilema gigas* and *Moneilema armatum*, in response to Pleistocene climate changes. *Mol. Ecol.* 14: 1025-1044.

Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460.

Webb T and Bartlein PJ (1992). Global changes during the last 3 million years - Climatic controls and biotic responses. *Annu. Rev. Ecol. Syst.* 23: 141-173.

Wright S (1931). Evolution in Mendelian populations. *Genetics* 16: 97-159.

Wright S (1938). Size of population and breeding structure in relation to evolution. *Science* 87: 430-431.

Wright S (1951). The genetical structure of populations. *Ann. Eugen.* 15: 323-354.

APPENDIX

In order to obtain an expression for the mean coalescence time, we made use of the fact that the expectation of a sum of random variables is equal to the sum of the expectations of each random variable. The expected mean coalescence time is the sum of the contributions of each phase, which we will call here X_1, X_2, X_3, etc. We need to obtain the sum of the $E[X_i]$'s.

For the first phase - panmictic - the contribution can be divided into two cases: if coalescence happens already during this phase, and if it does not. If it does not occur, the contribution of this phase for the coalescence time is simply t_P. If it does, the contribution is the number of generations that have passed until the coalescence event. Summing these parts multiplied by their respective probabilities, and simplifying, we obtain:

$$E[X_1] = \tau_p \cdot e^{-\tau_p} + \int_{t=0}^{\tau_p} t \cdot e^{-t} dt = 1 - e^{\tau_p} \qquad \text{(Equation A1)}$$

where $\tau_p = t_p / N_p$. It is important to note that, in this calculation, we used the continuous-time approximation (see Hudson, 1990), with time measured in units of N_p generations. The contribution of this phase in units of generations is, therefore:

$$E[X_1] = N_p \cdot (1 - e^{-t_p / N_p}) \qquad \text{(Equation A2)}$$

For the second phase (structured), the contribution will be equal to zero if the two lineages have already coalesced during the previous phase. If they have not coalesced (probability equal to $e^{-\tau}$), the two lineages will be randomly distributed among the demes. The probability that they will end up in a single deme is 1/D, and in separate demes 1-1/D. Since we are assuming that there is no migration, two lineages in separate demes have zero probability of coalescing. In this case, the contribution for the mean coalescence time is t_S. And in the case where the lineages end up in the same deme, the contribution is calculated in the same way as for the panmictic phase, using t_S and N_S instead of t_P and N_P. Therefore, the contribution of this phase for the mean coalescence time is:

$$E[X_2] = e^{-t_p / N_p} [(1 - 1/D) . t_S + (1/D) . N_S . (1 - e^{-t_S / N_S})] \qquad \text{(Equation A3)}$$

For the third phase, panmictic, we only need the probability that the lineages have not coalesced yet, multiplied by the result of the first phase (Equation A2). The probability that the lineages have not coalesced by the beginning of the third phase is:

$$P[\text{no coal. in 1st or 2nd phase}] = e^{-t_p / N_p} . [1 - (1/D) + (1/D) . (e^{-t_S / N_S})] =$$
$$= e^{-t_p / N_p} . (D - 1 + e^{-t_S / N_S}) / D \qquad \text{(Equation A4)}$$

The contribution of the third phase is then:

$$E[X_3] = [e^{-t_p / N_p} . (D - 1 + e^{-t_S / N_S}) / D] . N_p . (1 - e^{-t_p / N_p}) \qquad \text{(Equation A5)}$$

Likewise, the contribution of the fourth phase is:

$$E[X_4]=[e^{-t_p/N_p}.(D-1+e^{-t_S/N_S})/D].e^{-t_p/N_p}[(1-1/D).t_S+(1/D).N_S.(1-e^{-t_S/N_S})] \quad \text{(Equation A6)}$$

For each odd phase, then, the contribution for the mean coalescence time is of the form:

$$E[X_i]=[e^{-t_p/N_p}.(D-1+e^{-t_S/N_S})/D]^{\frac{i-1}{2}} N_p.(1-e^{-t_p/N_p}); i \text{ odd} \quad \text{(Equation A7)}$$

And in each even phase, the contribution is of the form:

$$E[X_i]=[e^{-t_p/N_p}.(D-1+e^{-t_S/N_S})/D]^{\frac{i-2}{2}}$$
$$. e^{-t_p/N_p}[(1-1/D).t_S+(1/D).N_S.(1-e^{-t_s/N_S})]; i \text{ even} \quad \text{(Equation A8)}$$

Combining both, and summing for each phase, we obtain the mean coalescence time:

$$E[\text{coal. time}]=\left\{\sum_{i=0}^{\infty}\left[e^{-t_p/N_p}.(D-1+e^{-t_S/N_S})/D\right]^i\right\}$$
$$.\left\{N_p.(1-e^{-t_p/N_p})+e^{-t_p/N_p}[(1-1/D).t_S+(1/D).N_S.(1-e^{-t_S/N_S})]\right\} \quad \text{(Equation A9)}$$

The sum present in the previous equation is a geometric series of kind:

$$\sum_{i=0}^{\infty}a^i \quad \text{(Equation A10)}$$

with |a|<1, which converges to 1/(1-a). This allows one more simplification:

$$E[\text{coal. time}] = \left[\frac{D}{D-(D-1)e^{-t_p/N_p}-e^{-(t_p/N_p+t_S/N_S)}}\right]$$
$$.\left\{N_p.(1-e^{-t_p/N_p})+e^{-t_p/N_p}[(1-1/D).t_s+(1/D).N_S.(1-e^{-t_S/N_S})]\right\} \quad \text{(Equation A11)}$$

This is then our Equation 1, presented beforehand in the main text.