Arch. Pol. Fish.	Archives of Polish Fisheries	Vol. 10	Fasc. 2	233-240	2002

INFERRING COALESCENCE TIMES FROM GENE TREES IN TWO HYPOTHESIZED SUBPOPULATIONS OF VENDACE (*COREGONUS ALBULA* L.) – A SELECTIVE, ONE-SITE APPROACH

Wojciech Kobus, Paweł Brzuzan

University of Warmia and Mazury in Olsztyn, Poland

ABSTRACT. Data regarding the *Alu*I restriction site polymorphism from a recent study on mitochondrial DNA phylogeographic relatedness of vendace (*Coregonus albula* L.) populations were re-examined using the coalescent method. Restriction site loss at the *Alu*I recognition sequence was modeled as a mutation, and ancestral information such as time to the most recent common ancestor (TMRCA) and age of the mutation were inferred from gene trees assuming the infinitely-many-sites model of mutation. Coalescent trees were simulated under two evolutionary models using the GENETREE program. One model assumed a panmictic population and the other a subdivided one. The mean values of the TMRCAs did not differ between the two cases and were 2.4; this suggested that the most recent, common ancestor of the present vendace might have lived about 720,000 years ago.

Key words: ANCESTRAL INFERENCE, COALESCENT PROCESS, VENDACE (*COREGONUS ALBULA)*, GENE TREES, SAMPLES OF MITOCHONDRIAL DNA

INTRODUCTION

PCR-based RFLP analysis of mtDNA was recently applied to test the genetic and phylogeographic relatedness of five vendace (*Coregonus albula* L.) populations from Poland (Brzuzan et al. 2002). The study revealed two mtDNA groups of vendace which differed in their geographical distribution; one group predominated in western/central (*W*) Poland, whereas the other dominated in the eastern (*E*) part of the country (Fig. 1). The pattern of haplotype distribution and population pairwise divergence estimates both suggested that vendace from the studied geographical region may be derived from two glacial refugia, and that these two lineages shared a common ancestor approximately 0.1-0.5 million years ago. In the present study, the coalescent process was used to model the ancestry of the sample sequences to draw further conclusions about the common ancestor of the two vendace lineages. Although computationally intensive, this approach preserves more information than do pairwise sequence analyses (Tavaré et al. 1997). For example, the mean and standard deviation of the time to the most recent common ancestor (TMRCA) can be

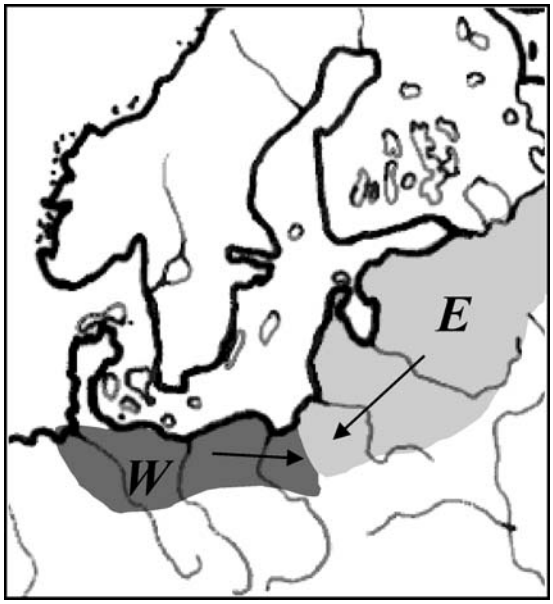

Fig. 1. Map of the location of the hypothetical vendace (*Coregonus albula* L.) subpopulations *W* and *E*. Arrows indicate the likely dispersal of the populations.

found conditional on the gene tree deduced from data assuming an infinitely-many-sites model of mutation (Griffiths and Tavaré 1998). The practical aim of this paper was to illustrate the selective sampling method by Bahlo and Griffiths (2000) through the application of previously published RFLP data.

THE COALESCENT

A detailed introduction to coalescent theory is presented in Nordborg (2001). Briefly, underlying a sample of DNA sequence data is a complex pattern of dependencies which reflects the ancestral relationships among sequences. In the absence of recombination, these relationships can be represented by a genealogical tree for which each tip corresponds to a sequence at the present time. Moving towards the root of the tree corresponds to going backwards in time, and branches "coalesce" when the corresponding DNA sequences last had a common ancestor. The root of the tree represents the MRCA of all the sequences in the sample.

Under the standard coalescent model, one unit of "coalescent" time is interpreted as $N/\sigma 2$ generations where $\sigma 2$ denotes the variance in the number of copies of the

sequence in the next generation. Mutations in the standard coalescent occur along the branches of the tree at the points of a homogeneous Poisson process with a rate of $\theta/2$. Due to coalescent time rescaling, this corresponds to a mutation rate of $\mu \equiv \theta/(2N)$ per locus per generation in a population of N sequences. It should be noted that N is used for the number of chromosomes which can be passed on to the next generation. For mtDNA data, N sequences correspond to $2N$ individuals - N males and N females.

Griffiths and Tavaré (1998) show that in a panmictic population the expected age of a mutation at a site which subtends b mutant copies and n-b non-mutant copies in a sample of n genes is as follows:

$$2 \binom{n-1}{b} \sum_{j=2}^{n-b+1} \binom{n-j}{b-1} \times \frac{n-j+1}{n(j-1)} \tag{1}$$

The expected TMRCA is as follows:

$$2\left(1-\frac{1}{n}\right) 2 \binom{n-1}{b} \sum_{j=2}^{n-b+1} \binom{n-j}{b-1} \times \frac{1}{j(j-1)} \tag{2}$$

The probability distribution of the number of mutant copies of a gene at a site known to be segregating is as follows:

$$b \Big/ \sum_{j=1}^{n-1} j^{-1}, \qquad b = 1,\ldots,n-1. \tag{3}$$

Ancestral inferences from coalescent gene trees can be drawn using particular DNA sites (Bahlo and Griffiths 2000). This selective site sampling is appropriate in sequence data which has just been sequenced for particular sites and can be done to consider the geography of particular mutant sites. For example, in subpopulations A, B suppose samples of n_A, n_B have b_A, b_B mutant copies at a site. Then the gene tree from this single site is as follows:

$$\begin{array}{llll} \boldsymbol{A} & \boldsymbol{b_A} & : & \mathbf{1\ 0} \\ \boldsymbol{A} & \boldsymbol{n_A\text{-}b_A} & : & \mathbf{0} \\ \boldsymbol{B} & \boldsymbol{b_B} & : & \mathbf{1\ 0} \\ \boldsymbol{B} & \boldsymbol{n_B\text{-}b_B} & : & \mathbf{0.} \end{array}$$

DATA

In the present paper, we applied the above method by employing previously published data of the *Alu*I restriction site polymorphism in d-loop mtDNA for 97 and 56

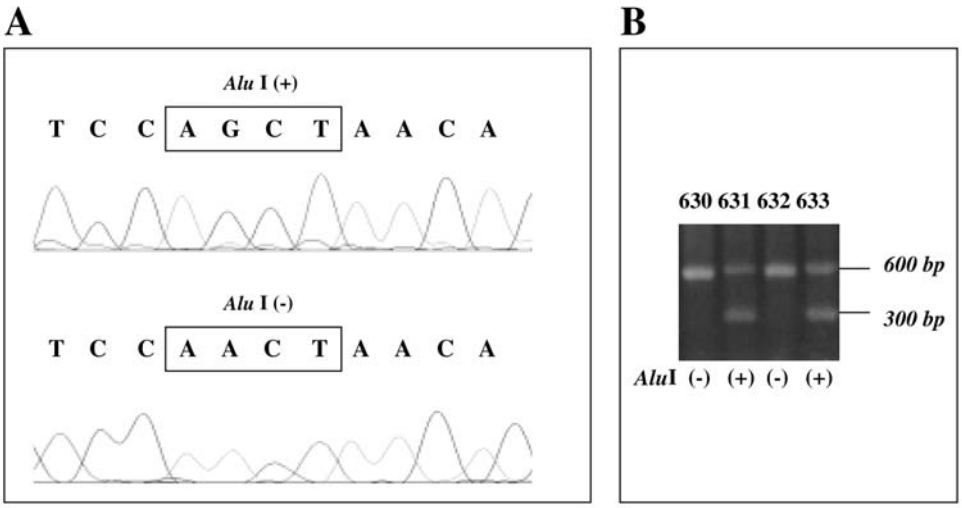

Fig. 2. Comparison of mtDNA sequences carrying (+) the *Alu*I recognition sequence or not (-). (A) Comparison of the d-loop mtDNA sequences (positions 262 to 272) which have a site that is recognized by the restriction enzyme (box in the upper fluorogram), and that without the restriction site (down). The underlying mutation is the G→A substitution. (B) Photograph of a gel showing the *Alu*I restriction banding pattern. The PCR-amplified entire d-loop (fragment size of about 1200 bp) from fish 630 through 633 were digested with the restriction enzyme. The presence of the restriction site (+) in samples 631 and 633 is evidenced by two additional co-migrating bands, each of about 300bp.

vendace individuals representing western (W) and eastern (E) populations from northern Poland, respectively. Inferences which can be drawn under these circumstances must assume that site gain or loss at the *Alu*I recognition sequence, 5'-AGCT-3', is associated with the occurrence of only one nucleotide mutation. To ascertain this, an approximately 300-bp fragment of mtDNA d-loop region carrying the *Alu*I recognition site was sequenced for two vendace individuals and compared with data from two fish lacking the restriction site. PCR conditions, primers used and DNA sequencing procedures followed those described in Brzuzan and Ciesielski (2000). Indeed, it was found that the restriction site loss was attributed to a single nucleotide G→A substitution (Fig. 2A). Hence, the *Alu* I restriction site loss was modeled as a mutation, and we assumed the site to be diallelic, present (+) or absent (-) on the d-loop mtDNAs of 153 vendace previously surveyed with the restriction enzyme (Fig. 2B).

THE SOFTWARE AND SIMULATION MODELS

We used the GENETREE 9.0 program which is described in Bahlo and Griffiths (2000). GENETREE implements the computation technique of Griffiths and Tavaré

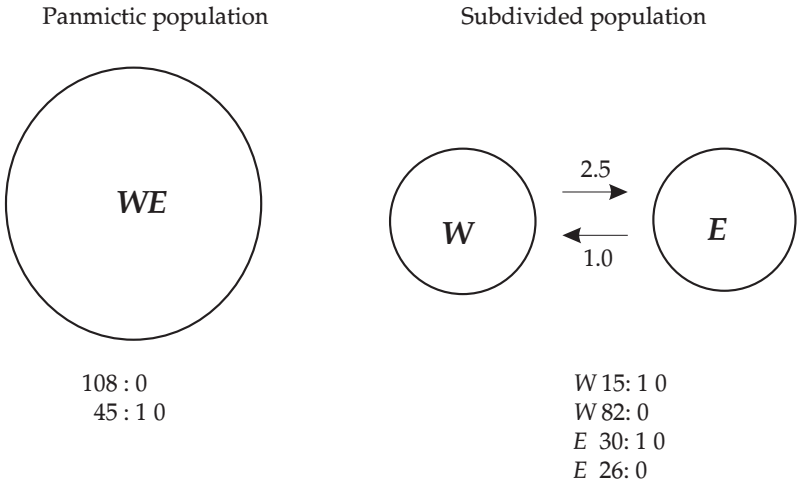

Fig. 3. Schematic depiction of two alternative evolutionary scenarios to explain the haplotype variation at the *Alu*I restriction site. Gene trees from this single site are shown for each model.

(1998) which simulates gene trees conditional on their topology implied by the mutation pattern in the sample of DNA sequences. The program approximates likelihood surfaces with respect to θ, mutation rate, under the infinite-sites mutation model, and *M*, migration matrix, under the island model. Our models are shown in Fig. 3 and both assume a constant effective female population size, *N* (which is going back in time) with either random mating (one panmictic population) or the population splitting (two subdivided populations). The importance of both western and eastern refugium in Europe has been previously argued for whitefish *Coregonus* sp. (Bernatchez and Dodson 1994), brown trout *Salmo trutta* sp. (Osinov and Bernatchez 1996) and Eurasian perch *Perca fluviatilis* L. (Nesbø et al. 1999). For the latter model (subdivided population) an H-option implemented in the GENETREE program allowed the likelihood estimation of the parameter of backward (in time) migration rate from population *W* to *E*, M_{WE}, to be 2.5, while the other migration rate, M_{EW} (from subpopulation *E* to *W*), was assumed to be 1.0 (Fig. 3). The estimated means and standard deviations of both the TMRCA and the age of the mutation at the *Alu*I restriction site come from their empirical distributions. Our simulation results are based on 100,000 replicate runs, with a fixed θ of 0.0 in the command line of GENETREE.

RESULTS AND DISCUSSION

The information that was computed from the data are presented in Table 1.

TABLE 1

Simulation results of the *Alu*I restriction site data under two hypotheses shown in Fig. 3

Parameter	Population		
	Panmictic	Subdivided	
Estimated likelihood of the tree, (SE)	2.22×10^{-2} (5.74×10^{-5})	4.34×10^{-4} (2.42×10^{-6})	
Mean TMRCA (SD)	2.41 (1.30)	2.41 (1.28)	
Age of mutation (SD)	1.02 (0.93)	1.00 (0.89)	
		W	**E**
Sample MRCA distribution in subpopulations	-	0.16	0.84
Mutation distribution in subpopulations	-	0.24	0.76

On the issue of comparing panmictic and subdivided populations, the gene trees from both models appear to be equivalent and yield almost the same coalescent estimates. In comparison with a panmictic population, the estimated likelihood of the tree was smaller for subdivided populations. The difference in the likelihood values is because subdivision with smaller migration rates allows for a longer ancestry, and more trees need to be examined to maximize the likelihood function. The mean and standard deviation of the TMRCA were not different between the two cases and were 2.4. Standard deviations for the TMRCA are large, with values typically one half of their respective means (Table 1). To convert coalescent time (T) to real time (t) in years, we can use the formula $t = NTG$, where G is the generation time in years. Hence, for clades associated with regions with larger or smaller long-term N, the MRCA may be considerably older or younger, respectively (e.g. Nesbø et al. 1999). We applied G=3, and N =100,000 for the female ancestral population size. Hence, the MRCA of the two present vendace lineages is suggested to have lived about 720,000 years ago. Similarly, by applying these values to the coalescent age of the mutation at the *Alu*I restriction site, we estimated it to have occurred some 300,000 years ago. Regardless of exact timing, our results may support the view of Brzuzan et al. (2002) that the studied vendace populations date back to an ancestral population which may have lived in the middle Pleistocene. An old MRCA of the DNA sample studied here is also in agreement with high levels of allozyme diversity observed in previous studies of

vendace (e.g. Vuorinen and Luczynski 1991). It should be noted that the TMRCA and age of mutation estimated here are possibly longer than in reality, because many other potential sites which are not known to be segregating were not typed.

In our data set, the MRCA of the whole population (*WE*) is most likely to be in subpopulation *E* (probability of 0.84; Table 1). A similar message may be read from information on the mutation distribution in the two subpopulations (Table 1). The distribution indicates a likely geographic history of lineages and in which subpopulation the mutation occurred. Mutation at the *Alu*I recognized site examined here has a probability of 0.76 of occurring in subpopulation *E* and only 0.24 of occurring in subpopulation *W*.

We can conclude that this method, which allows analyses of data which is collected at particular sites only, may be applicable as a pilot inferential tool in those studies where the time to the most recent common ancestor and ages of mutation are to be computed. Furthermore, questions about individual sites, such as mutation distribution and MRCA distribution, may be of particular interest geographically.

ACKNOWLEDGMENTS

The study was supported by the University of Warmia and Mazury, Project No. 080302.205.

REFERENCES

Bahlo M., Griffiths R.C. 2000 - Inference from gene trees in a subdivided population - Theor. Pop. Biol. 57: 79-95.

Bernatchez L., Dodson J. 1994 – Phylogenetic relationships among Palearctic and Nearctic whitefish (*Coregonus* sp.) populations as revealed by mitochondrial DNA variation – Can. J. Fish. Aquat. Sci. 51 (1): 240-251.

Brzuzan P., Ciesielski S. 2000 - Mitochondrial DNA sequence from an extinct population of European vendace, *Coregonus albula* L. - Folia biologica (Cracow) 48: 151-154.

Brzuzan P., Kozlowski J., Fopp D. 2002 - Genetic structure of Polish populations of vendace (*Coregonus albula*) inferred from mitochondrial DNA - Arch. Hydrobiol. Spec. Issues Advanc. Limnol. 57: 1-10.

Griffiths R.C., Tavaré S. 1998 - The age of a mutation in general coalescent tree - Stoch. Models 14: 273-295.

Nesbø C.L., Fossheim T., Vøllestad A., Jakobsen K.S. 1999 - Genetic divergence and phylogeographic relationships among European perch (*Perca fluviatilis*) populations reflect glacial refugia and postglacial colonization - Mol. Ecol. 8: 1387-1404.

Nordborg M. 2001 - Coalescent theory. In: Handbook of Statistical Genetics (Eds. D.J. Balding, C. Cannings & M. Bishop), Wiley, Chichester: 179-212.

Osinov A.G., Bernatchez L. 1996 - „Atlantic" and „Danubian" phylogenetic groupings of brown trout *Salmo trutta* complex: Genetic divergence, evolution and conservation - J. Ichthyol. 36: 723-746.

Tavaré S., Balding D.J., Griffiths R.C., Donnely P. 1997 - Inferring coalescence times from DNA sequence data – Genetics 145: 505-518.

Vuorinen J., Luczynski M. 1991 - Electrophoretic variation in four Polish populations of vendace (*Coregonus albula* (L.)) - Acta Hydrob. 33: 77-86.

STRESZCZENIE

CZAS KOALESCENCJI SEKWENCJI MITOCHONDRIALNEGO DNA W DWÓCH HIPOTETYCZNYCH SUBPOPULACJACH SIELAWY (*COREGONUS ALBULA* L.)

Przodkowie sielawy (*Coregonus albula* L.) zasiedlającej dzisiaj jeziora Polski pochodzili prawdopodobnie z dwóch odrębnych genetycznie populacji, które przetrwały epokę lodowcową w odizolowanych, nie zamarzniętych schroniskach, a następnie kolonizowały terytorium dzisiejszej Polski podczas ustępowania lodowca (rys. 1).

Dane dotyczące polimorfizmu miejsc restrykcyjnych dla endonukleazy *Alu*I regionu regulacyjnego mtDNA uzyskane w cytowanej pracy opracowano ponownie (rys. 2) i wykorzystano do oszacowania czasu do najbardziej współczesnego wspólnego przodka (TMRCA) obu populacji sielawy. Zastosowano metodę koalescencji z uwzględnieniem modelu nieskończonej liczby podstawień nukleotydowych w sekwencji DNA. Rozkłady TMRCA analizowano przy założeniu dwóch modeli ewolucyjnych używając programu GENETREE (Bahlo and Griffiths 2000). Pierwszy model zakładał istnienie populacji panmiktycznej, a drugi populacji podzielonej (rys. 3). Średnie wartości TMRCA nie różniły się w obu przypadkach i wyniosły 2,4. Sugeruje to, że najbardziej współczesny wspólny przodek sielawy obecnie występujący na terenie Polski mógł żyć około 720.000 lat temu (tab. 1).

CORRESPONDING AUTHOR:

Dr hab. Paweł Brzuzan
Uniwersytet Warmińsko-Mazurski
Zakład Genetyki Ewolucyjnej
ul. Oczapowskiego 5
10-957 Olsztyn
Tel./Fax: +48 89 5234772; e-mail:brzuzan@uwm.edu.pl