

# The impact of evolutionary models on the inference of population history

Gunter Weiss\*

## ABSTRACT

A model that describes the evolution of population sequence data can be split conveniently into two independent parts, a mutation and a reproduction model, as long as it is assumed that selection and recombination are absent. Here, we investigate the impact of different mutation or reproduction models on the inference of population history using data from the hypervariable region (HVR) *I* of human mitochondrial DNA utilizing a likelihood ratio statistic based on the mean pairwise differences between DNA sequences and the number of variable sites in a sample. We conduct statistical inference within classes of reproduction models that allow for a deterministic change of population size starting at some point back in time from a population in equilibrium for samples of HVR *I* sequences from Basques, Swedes, and Icelanders. An appropriate mutation model for these data was shown to be the Tamura-Nei model with heterogeneous mutation rates. Comparisons of estimates of population history parameters obtained under this mutation model to estimation results under the widely used infinitely many sites model show that the differences between these estimates are quite irregular. Studying the effects of different demographic scenarios, namely exponential and sudden change in population size, on the estimates of population history parameters revealed a tendency to smaller estimates of the amount of growth and of the time back when the population started to change in size.

---

\*Institute of Zoology, University of Munich, Luisenstrasse 14, D-80333 Munich, Germany, phone: +49-89-5902327, fax: +49-89-5902474, email: gweiss@zi.biologie.uni-muenchen.de

## INTRODUCTION

The hypervariable region *I* (HVR*I*) of mitochondrial DNA represents presumably the most frequently sequenced part of the human genome. The maternal mode of inheritance, the absence of recombination and the high evolutionary rate of HVR*I* sequences make them especially suited for intra-specific studies of our own species. The effort of many groups of molecular geneticists resulted in more than 4,000 sequenced individuals from almost all parts of the world (Handt *et al.*, 1998). The analysis of these new data was greatly influenced by the introduction of ‘coalescent methods’ (Kingman, 1982; Tavaré, 1984; Donnelly and Tavaré, 1995). By describing the evolution of a sample of DNA sequences in terms of stochastic processes, coalescent methods allow to incorporate parametric models of sequence evolution in a population. The appropriateness of these models can be subject to statistical testing and the relevant model parameters can be estimated.

With the increasing amount of available HVR*I* sequence data it became clear that the evolution of this genomic part is more complex than previously assumed: Transversions are less abundant than transitions, transitions between pyrimidines are more frequent than those between purines, and the mutation rate varies extensively among different sites (Kocher and Wilson, 1991; Tamura and Nei, 1993; Wakeley, 1993). Many data analyses have been published, that ignored these facts (Harpending *et al.*, 1993; Sherry *et al.*, 1994; Rogers, 1995; Rogers and Jorde, 1995; Rogers *et al.*, 1996; Wakeley and Hey, 1997). Instead, the infinitely many sites model (Watterson, 1975) that even does not allow for back or parallel mutations was used.

Here, we will study the effects of this simplistic assumption on the estimation of parameters of population history. Furthermore, we will investigate how these estimates are influenced by models of population history that assume different demographic scenarios.

## THEORETICAL BACKGROUND

This section aims to specify the models of population history as well as the mutation models that will be compared in their effects on the inference of population history parameters. The inference method (Weiss and von Haeseler, 1997) will also be recapitulated briefly.

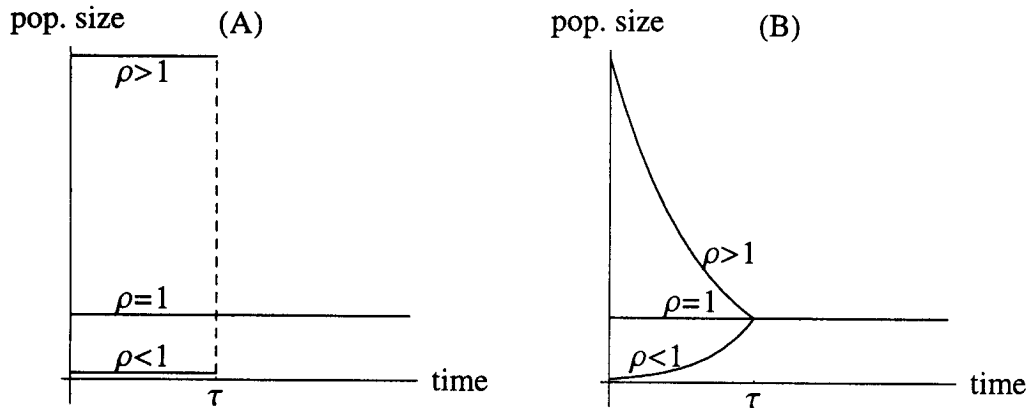


Figure 1: Two models with changes in population size: (A) sudden change model, (B) exponential change model. The x-axis represents time back in the past, the y-axis represents population size.

**Models of population history:** We will investigate two expansion models where a Wright-Fisher population at equilibrium change size at a certain time  $\tau$  in the past to the current population size. The two models analysed are distinct in the way population size alters. The *sudden change* model (Rogers and Harpending, 1992) assumes that the population size ‘jumps’ instantaneously at a time  $\tau$  to the current population size. Apart from that it remains constant in the period before and after this event (Fig. 1(A)). The second model assumes a continuous *exponential change* from the population size in the past to the current population size (Fig. 1(B)).

Both models are defined by three parameters:  $\theta = 2N_0\mu$  is the population parameter at equilibrium in the past. Here,  $N_0$  denotes the initial effective population size and  $\mu$  is the mutation rate per sequence and generation.  $\tau$  is the time when the population size started to change, where  $\tau$  is measured in units of  $1/\mu$  from the present. Finally,  $\rho$  defines the ratio of current to initial population size. A value of  $\rho$  larger than one indicates an increase in population size looking forward in time,  $\rho < 1$  is tantamount with a population size decrease. We get the basic model of constant size as special cases of these models by setting  $\rho$  equal to one. Then the time parameter  $\tau$  remains unspecified.

**Mutation models:** The Tamura-Nei model with  $\Gamma$ -distributed rates (TN+ $\Gamma$ ) (Tamura and Nei, 1993) classifies mutations into transversions, pyrimidine transitions and purin

parameter	identifier	mean
base frequency of A	$\pi_A$	32.7%
base frequency of C	$\pi_C$	33.8%
base frequency of G	$\pi_G$	11.4%
base frequency of T	$\pi_T$	22.1%
transition/transversion	$\kappa$	15.5
pyrimidine/purine trans.	$\xi$	1.85
rate heterogeneity	$\alpha$	0.26

Table 1: Estimated parameters of the Tamura-Nei model with  $\Gamma$ -distributed rates for HVR1 sequences of human mtDNA (Weiss and von Haeseler, 1997).

transitions. It allows for site-to-site rate variation according to a  $\Gamma$ -distribution. Since the scaling of the mutation rate is done with the population parameter  $\theta = 2N\mu$  in our context, this mutation model can be described by six parameters (three base frequencies, a transition/transversion parameter  $\kappa$ , a pyrimidine/purine transition parameter  $\xi$  and the shape parameter  $\alpha$  of the  $\Gamma$ -distribution). Thereby, the second parameter of the  $\Gamma$ -distribution is set equal to  $\alpha$  yielding a distribution with expectation equal to one. Then,  $\mu$  can be regarded as the mean mutation rate of a HVR1 sequence. It has been shown that this model is superior to simpler models to describe the evolution of HVR1 sequences (Weiss and von Haeseler, 1997). Using the large amount of available HVR1 data (Handt *et al.*, 1998), the parameters of the TN+ $\Gamma$ -model were estimated by exploiting the maximum likelihood routines implemented in the tree reconstruction package *PUZZLE* 3.1 (Strimmer and von Haeseler, 1996). Table 1 summarizes the estimation results.

Due to its mathematical simplicity, the so-called *infinitely many sites model* ( $\infty$ -sites) (Watterson, 1975) is widely used to describe the generation of different sequences in a population. This model does not allow for parallel or back mutations, since it assumes that a new mutation takes place at a position that has not mutated in any other sequence before. The number of sites that mutate per generation is a random variable that follows a Poisson distribution with mean  $\mu$ , where  $\mu$  denotes the mutation rate per sequence and generation. While the infinitely many sites model serves as a useful approximation when substitution rates are low and the probability that a mutation occurs twice at a particular nucleotide site in any sequence is negligible, it is inappropriate to model the

complex mutation mechanism of HVR/I sequences. Actually, almost every HVR/I data set from a single human population violates the assumptions of this model.

**Inferring population history parameters:** Recently, a simulation based method to estimate population history parameters under complex mutation models was introduced (Weiss and von Haeseler, 1997). Here, we will briefly describe the rationale of this method. Given a mutation model (TN+ $\Gamma$  or  $\infty$ -sites) and a class of population histories (exponential or sudden change), the evolution of a sample of sequences is fully characterized by the three parameters  $\theta$ ,  $\tau$ , and  $\rho$ . In principle, these parameters could be estimated by maximising the likelihood function given the full data  $\mathcal{D}$ , i.e.  $lik(\theta, \tau, \rho|\mathcal{D})$ . Here,  $\mathcal{D}$  consists of the alignment of the  $n$  sequences. Unfortunately, it is difficult to write down this likelihood function explicitly, even under simple models. In addition, the elegant Markov chain recursion method (Griffiths and Tavaré, 1994a, 1994b) to approximate the likelihood requires enormous computation time when applied to complex models of sequence evolution. Therefore, we replace the data  $\mathcal{D}$  by two simple summary statistics, the mean pairwise sequence difference in the sample  $k$  and the number of variable positions  $s$ , and evaluate the likelihood  $lik(\theta, \tau, \rho|k, s)$  by simulations based on coalescent theory (Weiss and von Haeseler, 1997). The relationship between  $k$  and  $s$  (implicitly used here and explicitly exploited in Tajima's  $D$ -statistic (Tajima, 1989)) proved to be useful to detect demographic signals (Aris-Brosou and Excoffier, 1996; Watson *et al.*, 1996; Weiss and von Haeseler, 1997), if neutrality of evolution in the genomic region studied (Kimura, 1983) is accepted *a priori*.

The parameter set  $(\hat{\theta}, \hat{\tau}, \hat{\rho})$  that maximises the likelihood serves as a point estimate. By comparing the maximum likelihood value, i.e.  $lik(\hat{\theta}, \hat{\tau}, \hat{\rho}|k, s)$ , to the likelihood value of a different parameter set, say  $(\theta_0, \tau_0, \rho_0)$ , one can additionally discriminate between plausible population scenarios by using the likelihood ratio:  $lik(\theta_0, \tau_0, \rho_0|k, s)/lik(\hat{\theta}, \hat{\tau}, \hat{\rho}|k, s)$ .

## THE IMPACT OF THE CHOICES OF MODELS

To illustrate the effects of different models of DNA sequence evolution on the inference of population history we applied our method to HVR/I data from three European populations, namely the Basques (Bertranpetit *et al.*, 1995), the Swedes (?), and the Icelanders (Sajantila *et al.*, 1995). Sample sizes  $n$ , mean pairwise sequence differences  $k$ , and numbers of variable positions  $s$  of these data sets are given in Table 2. For each of these

Population	$n$	$k$	$s$
Basques	45	3.24	32
Swedish	32	4.57	38
Icelander	39	5.03	32

Table 2: Sample sizes  $n$ , mean pairwise sequence differences  $k$ , and numbers of variable positions  $s$  of the analysed data sets.

Population	Model	$\hat{\theta}$	$\hat{\tau}$	$\hat{\rho}$	max lik
Basque	TN+ $\Gamma$ & exp. change	1.0	2.25	100	0.0156
	$\infty$ -sites & exp. change	0.5	2.25	100	0.0165
	TN+ $\Gamma$ & sudden change	0.5	1.5	100	0.0164
	$\infty$ -sites & sudden change	2.0	1.25	10	0.0103
Swedes	TN+ $\Gamma$ & exp. change	1.0	2.5	1000	0.0153
	$\infty$ -sites & exp. change	1.0	2.5	100	0.0121
	TN+ $\Gamma$ & sudden change	1.0	2.0	100	0.0155
	$\infty$ -sites & sudden change	3.0	1.5	10	0.0077
Icelander	TN+ $\Gamma$ & exp. change	0.5	4.25	100	0.0120
	$\infty$ -sites & exp. change	2.0	3.5	10	0.0095
	TN+ $\Gamma$ & sudden change	2.0	2.5	10	0.0113
	$\infty$ -sites & sudden change	1.5	2.75	10	0.0113

Table 3: Parameter estimates under different models for the Basque, Swedish, and Icelandic data. See text for explanation.

data we determined the likelihood values on a grid of parameter combinations, whereby  $\theta$  was set equal to multiples of  $1/2$ ,  $\tau$  to multiples of  $1/4$  (mutational units), and  $\rho$  equaled integer powers of 10.

Table 3 show the most probable parameter sets  $(\hat{\theta}, \hat{\tau}, \hat{\rho})$  we inferred by this method under the four different evolutionary models resulting from the combination of one of the mutation models with one of the models of population history. The estimation results between models of exponential and sudden change differ mainly in the estimate of  $\tau$ , which is always less in the case of a sudden expansion scenario. The differences between the two mutation models are more irregular. The estimation results deviate in one, two, or even all three parameters with a tendency to smaller estimates of  $\tau$  and  $\rho$  under the  $\infty$ -sites model. Interestingly, comparison of the maximal likelihood values under the two mutation models show that with one exception (Basques + exp. change) the TN+ $\Gamma$ -model gets equal or more support from the data. This observation indicates that the TN+ $\Gamma$ -model is more appropriate to model the evolution in HVR/I sequences in agreement with earlier results based on phylogenetic tree reconstruction (Weiss and von Haeseler, 1997).

## DISCUSSION

A careful model selection is essential to the estimation of parameters. While it is difficult to judge whether a model of exponential or sudden expansion is more realistic (or better less unrealistic), it is clear that the TN+ $\Gamma$ -model is more appropriate to model the evolution in HVR/I sequences than the infinitely many sites assumption. It has been argued that the “error introduced by the infinite sites model” (Rogers, 1992) is of minor importance to the estimation of population history parameters, because “mitochondrial mismatch analysis is insensitive to the mutational process” (Rogers *et al.*, 1996). In contrast, substantial effects of different mutation processes on various genetic diversity measures have been reported (Lundstrom *et al.*, 1992; Bertorelle and Slatkin, 1995; Aris-Brosou and Excoffier, 1996; Tajima, 1996). Here, we showed that the effects of a oversimplified mutation mechanism on the estimation of population history parameters from HVR/I sequence data are quite irregular and therefore difficult to correct. If one would use the full data rather than two summary statistics as proposed here, these effects are expected to be even more severe.

Therefore, the analysis of HVR/I sequence data using the TN+ $\Gamma$ -model is recommended, even though this analysis is computationally more expensive compared to those

using simpler mutation models.

Acknowledgement:

I am grateful to the organizers (A. v. Haeseler, N. Takahata, M. Uyenoyama) of the workshop and the DFG for financial support. Especially, I would like to thank Antti Sajantila and Svante Pääbo for providing unpublished sequence data.

## Bibliography

- Aris-Brosou, S., and Excoffier, L. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol.* **13**:494–504.
- Bertorelle, G., and Slatkin, M. 1995. The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. *Mol. Biol. Evol.* **12**:887–892.
- Bertranpetit, J., Sala, J., Calafell, F., Underhill, P. A., Moral, P., and Comas, D. 1995. Human mitochondrial DNA variation and the origin of Basques. *Am. J. Hum. Genet.* **59**:63–81.
- Donnelly, P., and Tavaré, S. 1995. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**:401–421.
- Griffiths, R. C., and Tavaré, S. 1994a. Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* **46**:131–159.
- Griffiths, R. C., and Tavaré, S. 1994b. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond.* **B 344**:403–410.
- Handt, O., Meyer, S., and von Haeseler, A. 1998. Compilation of human mtDNA control region sequences. *Nucleic Acids Res.* in press.
- Harpending, H. C., Sherry, S. T., Rogers, A. R., and Stoneking, M. 1993. The genetic structure of ancient human populations. *Current Anthropology* **34**:483–496.
- Kimura, M. 1983. *The neutral theory of molecular evolution*. London: Cambridge University Press.

- Kingman, J. F. C. 1982. The coalescent. *Stoch. Proc. Applns.* **13**:235–248.
- Kocher, T. D., and Wilson, A. C. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and protein-coding regions. In: *Evolution of life: fossils, molecules and culture* (Osawa, S. and Honio, T., eds.) pp. 391–413. Springer Verlag Tokyo.
- Lundstrom, R., Tavaré, S., and Ward, R. H. 1992. Modelling the evolution of the human mitochondrial genome. *Math. Biosci.* **112**:319–335.
- Rogers, A. R. 1992. Error introduced by the infinite sites model. *Mol. Biol. Evol.* **9**:1181–1184.
- Rogers, A. R. 1995. Genetic evidence for a pleistocene population explosion. *Evolution* **49**:608–615.
- Rogers, A. R., Fraley, A. E., Bamshad, M. J., Watkins, W. S., and Jorde, L. B. 1996. Mitochondrial mismatch analysis is insensitive to the mutational process. *Mol. Biol. Evol.* **13**:895–902.
- Rogers, A. R., and Harpending, H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**:552–569.
- Rogers, A. R., and Jorde, L. B. 1995. Genetic evidence on modern human origins. *Human Biology* **67**:1–36.
- Sajantila, A., Lahermo, P., Anttinen, T., Lukka, M., Sistonen, P., Savontaus, M.-L., Aula, P., Beckman, L., Tranebjaerg, L., Gedde-Dahl, T., Issel-Tarver, L., DiRienzo, A., and Pääbo, S. 1995. Genes and languages in europe: an analysis of mitochondrial lineages. *Gen. Res.* **5**:42–52.
- Sherry, S. T., Rogers, A. R., Harpending, H. C., Soodyall, H., Jenkins, T., and Stoneking, M. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. *Human Biology* **66**:761–775.
- Strimmer, K., and von Haeseler, A. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**:964–969.

- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- Tajima, F. 1996. The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* **143**:1457–1465.
- Tamura, K., and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**:512–526.
- Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.* **26**:119–164.
- Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.* **37**:613–623.
- Wakeley, J., and Hey, J. 1997. Estimating ancestral population parameters. *Genetics* **145**:847–855.
- Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., and Pääbo, S. 1996. mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.* **59**:437–444.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**:256–276.
- Weiss, G., and von Haeseler, A. 1997. Inference of population history using likelihood ratio statistic. *Submitted to Genetics*.