

1 False facts and false views: coalescent analysis
2 of truncated data

3 Einar Árnason,
Institute of Biology, University of Iceland,
Sturlugata 7, 101 Reykjavík, Iceland

email: `einararn@hi.is`

4

Abstract

5 Darwin's dictum on false facts and false views points the way to opening the
6 road to truth via cogent criticism of the published record. Here I discuss a case in
7 which a truncated dataset (false facts) is used for coalescent analysis of historical
8 demography that reaches a foregone conclusion of a bottleneck of numbers (false
9 views).

10 "False facts are highly injurious to the progress of science, for they often en-
11 dure long; but false views, if supported by some evidence, do little harm, for
12 everyone takes a salutary pleasure in proving their falseness; and when this is
13 done, one path towards error is closed and the road to truth is often at the same
14 time opened" (Darwin, 1871, p. 385). Darwin's dictum is in full force and I apply
15 it here to a case where false facts have led to false views hoping to open the road
16 to truth.

17 Ólafsdóttir *et al.* (2014) studied demographic history of Atlantic cod, *Gadus*
18 *morhua*, at Iceland using mtDNA isolated from vertebrae from archaeological
19 sites. They compare their results to already published results from modern times
20 (citing Árnason, 2004). They notice a reduction in haplotype and nucleotide diver-
21 sity in modern times and use coalescent analysis to infer a bottleneck of numbers
22 at 1400–1500 and a marked reduction of effective population size, N_e , in mod-
23 ern times. They use Approximate Bayesian Computation, ABC, to model three

24 population size scenarios evaluated by matches to summary statistics.

25 A key problem of the study of Ólafsdóttir *et al.* (2014) is the handling of the
26 data of the modern samples for which they cite Árnason (2004) which summarizes
27 data from several papers on variation of cytochrome *b* from various localities in
28 the Atlantic ocean. The primary data on Iceland are not in that paper. Árnason
29 *et al.* (2000) published the original primary data on Icelandic cod, a paper not
30 cited by Ólafsdóttir *et al.* (2014).

31 First, the numbers reported in their Table S3 and said to represent "Modern
32 frequency" are not in accordance with the original correct data (Árnason *et al.*,
33 2000; Árnason, 2004) (Table 1). The original data have 519 individuals with
34 23 segregating sites defining 30 haplotypes (Table III of Árnason *et al.*, 2000)
35 whereas Table S3 reports different numbers for common and rare haplotypes and
36 total numbers and omits many haplotypes. There are discrepancies for many but
37 not all haplotypes (Table 1). There also are discrepancies between the numbers
38 for modern times reported in Table 2 of the paper and in supplemental Table S3:
39 sample size of 503 vs 499, number of haplotypes 10 vs 8, with 7 vs 6 segregating
40 sites.

41 Second, Ólafsdóttir *et al.* (2014) do not use all the data of the modern sam-
42 ple (Árnason *et al.*, 2000). They truncate the data by omitting 22 haplotypes, all
43 singleton (17), doubleton (3), one triplet and one quadruplet haplotype. These
44 truncations of the original data result in a dataset of 499 individuals with 8 haplo-
45 types and 6 segregating sites (Table S3). They are false facts. Coalescent analysis
46 in general proceeds by tracing the ancestry of a sample to a common ancestor.

47 By its nature coalescence is sensitive to the size and composition of a sample. If
48 a real sample from a natural population in true fact was both large (as the mod-
49 ern sample *Árnason et al.*, 2000) and had few or no rare alleles (as in Table S3
50 *Ólafsdóttir et al.*, 2014) the genealogy would be characterized by long internal and
51 few or no external branches. There would be a deficiency of low frequency vari-
52 ants and an excess of middle frequency variants. This would be a clear sign of a
53 declining population under coalescence theory (Wakeley, 2009, page 120). Using
54 the truncated data dataset for the Bayesian skyride plot (Minin *et al.*, 2008) under
55 BEAST (Drummond *et al.*, 2012) stacks the odds and *Ólafsdóttir et al.* (2014)
56 reach a foregone conclusion of a population bottleneck and low effective size in
57 the modern times. These are false views.

58 The 1500–1550 and the 1910 samples stand out from the rest (Table S3 *Ólafsdóttir*
59 *et al.*, 2014) and also influence the skyride analysis. The 1500–1550 sample has
60 a relatively large number of haplotypes and segregating sites, a relative evenness
61 in haplotype frequencies giving high nucleotide diversity ($\hat{\pi} = 0.0059$ compared
62 to $\hat{\pi} = 0.0052$ the modern sample *Árnason et al.* (2000), and $\hat{\pi} = 0.0047$ for the
63 truncated data in Table S3 *Ólafsdóttir et al.* (2014)). The 1910 sample has few
64 haplotypes and segregating sites, a relatively high frequency of the most common
65 haplotype and consequently low nucleotide diversity ($\hat{\pi} = 0.0043$). Nucleotide di-
66 versity estimates the scaled effective population size $\theta = 2N_e\mu$ Wakeley (2009).
67 These divergent samples along with the truncated dataset of the modern sample
68 are drivers of the apparent bottlenecks in skyride analysis *Ólafsdóttir et al.* (2014).

69 I have generated distributions of the number of segregating sites, the number

70 of haplotypes and the nucleotide diversity from 1000 random samples of size 36
71 representing the sample size of the 1500–1550 sample and of 1000 samples of
72 size 23 representing the 1910 sample of Ólafsdóttir *et al.* (2014) by random sam-
73 pling from the Árnason *et al.* (2000) dataset. At least 25% of the distributions had
74 a greater number of segregating sites than 6 and a greater number of haplotypes
75 than 8 reported for the 1500–1550 sample in Table S3 (Ólafsdóttir *et al.*, 2014).
76 More than 7% had a higher nucleotide diversity than the 1500–1550 sample. For
77 the 1910 sample 3 out of 1000 had equal or fewer segregating sites than the sam-
78 ple, about 6% had fewer or equal numbers of haplotypes and 25% had a lower
79 nucleotide diversity. Thus these divergent samples are within sampling errors of
80 the modern haplotype frequencies (Árnason *et al.*, 2000). Therefore, the bottle-
81 necks (Ólafsdóttir *et al.*, 2014) are spurious resulting from a combination of the
82 use of the truncated modern-times data and sampling variation in the small ancient
83 samples.

84 There also are internal discrepancies between results given in Table 2 and
85 in supplemental Table S3 of Ólafsdóttir *et al.* own data. For example, Table 2
86 reports 9 haplotypes and 7 segregating sites for the 1500–1550 sample. However,
87 the detailed data reported in Table S3 are 8 haplotypes defined by 6 segregating
88 sites (number of segregating sites can be determined from Table III of Árnason
89 *et al.* (2000) or from Figure 1 of Árnason (2004)). Similarly, there should be 5
90 and not 4 segregating sites in the 1650–1700 dataset and 3 and not 4 segregating
91 sites in the 1910 dataset of Ólafsdóttir *et al.* (2014).

92 Third, ABC analysis in general proceeds by simulating random datasets and

93 selecting a small subset of these that are most similar to the real dataset based
94 on congruence of summary statistics. Ólafsdóttir *et al.* (2014) used number of
95 haplotypes and number of segregating sites and summary statistics based on these
96 in their analysis. Discrepancies in summary statistics described above may bias
97 the selection of the sub-samples of 500 out of a million random datasets. Also
98 they report type I and II errors of 44% and 46% for a scenario of two bottlenecks
99 compared to a scenario of a single bottleneck or a constant population size. The
100 statement that “the ABC analysis supported the scenario of two bottlenecks over
101 the scenario of either a single bottleneck or constant population size. . .” is strange
102 given the very high type I and type II errors rates.

103 Fourth, the method section of the paper seems to imply that all the molecular
104 work was done in a dedicated ancient DNA laboratory in Canada. However, the
105 supplement states that only DNA isolation was done in dedicated ancient DNA
106 laboratory in Canada and that the rest of the molecular work from PCR ampli-
107 fication to sequencing was done in a lab in Reykjavik where “no previous work
108 on Atlantic cod had taken place”. However, this statement is inaccurate. The
109 post PCR work was actually done in shared facilities where Atlantic cod DNA of
110 modern samples, both mitochondrial and nuclear *Pan I* (Árnason, 2004; Árnason
111 *et al.*, 2009; Eiríksson & Árnason, 2013), has been amplified and sequenced for
112 many years. It is, therefore, not clear how established criteria for ancient DNA
113 work (Cooper & Poinar, 2000) were adhered to.

114 Also, there is no mention of how it was determined that the vertebrae sampled
115 from archaeological sites represent vertebrae from different individuals. How, for

116 example, can we know that the high evenness and high nucleotide diversity of
117 the 1500–1550 sample or the low diversity of the 1910 sample is not pseudo-
118 replication due to sampling multiple vertebrae from the same individual?

119 **References**

120 Árnason E, 2004. Mitochondrial cytochrome *b* DNA variation in the high-
121 fecundity Atlantic cod: Trans-Atlantic clines and shallow gene genealogy. *Ge-*
122 *netics*, **166**, 4, 1871–1885.

123 Árnason E, Hernandez U B, Kristinsson K, 2009. Intense habitat-specific
124 fisheries-induced selection at the molecular *Pan I* locus predicts imminent col-
125 lapse of a major cod fishery. *PLoS ONE*, **4**, 5, e5529.

126 Árnason E, Petersen P H, Kristinsson K, Sigurgíslason H, Pálsson S, 2000. Mito-
127 chondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland
128 and Greenland. *J. Fish Biol.*, **56**, 409–430.

129 Cooper A, Poinar H N, 2000. Ancient DNA: Do it right or not at all. *Science*,
130 **289**, 5482, 1139.

131 Darwin C R, 1871. *The Descent of Man, and Selection in Relation to Sex*, vol. 2.
132 London: John Murray, 1st ed.

133 Drummond A J, Suchard M A, Xie D, Rambaut A, 2012. Bayesian phylogenetics
134 with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.

- 135 Eiríksson G M, Árnason E, 2013. Spatial and temporal microsatellite variation
136 in spawning Atlantic cod, *Gadus morhua*, around Iceland. *Can. J. Fish. Aquat.
137 Sci.*, **70**, 1151–1158.
- 138 Minin V N, Bloomquist E W, Suchard M A, 2008. Smooth skyride through a
139 rough skyline: Bayesian coalescent-based inference of population dynamics.
140 *Mol. Biol. Evol.*, **25**, 1459–1471.
- 141 Ólafsdóttir G Á, Westfall K M, Edvardsson R, Pálsson S, 2014. Historical DNA
142 reveals the demographic history of Atlantic cod (*Gadus morhua*) in medieval
143 and early modern Iceland. *Proc. R. Soc. B*, **281**, 20132976.
- 144 Wakeley J, 2009. *Coalescent Theory*. Greenwood Village, Colorado, USA:
145 Roberts and Company Publishers.

Table 1. Discrepancies in frequencies of haplotypes in data for modern-times. First row is from Table III in Árnason *et al.* (2000) Árnason *et al.* (2000). Second row is truncated data from Table S3 of Ólafsdóttir *et al.* (2014) Ólafsdóttir *et al.* (2014) said to be modern-times data from Árnason (2004) Árnason (2004). Third row is discrepancy added (+) and omitted (–) between the first two rows. Other represents a pool of 22 rare haplotypes omitted in Ólafsdóttir *et al.* (2014).

Data source	Haplotype									Total
	A	D	C	E	G	MI	RI	NI	Other	
Table III in Árnason <i>et al.</i> (2000)	238	80	20	75	62	3	3	8	30	519
Table S3 in Ólafsdóttir <i>et al.</i> (2014)	242	80	20	78	64	3	3	9	0	499
Discrepancy	+4	0	0	+3	+2	0	0	+1	–30	–20