What is knowable about past demography? - The limited scope of genomic data analysis.

Janeesh Kaur Bansal[1], Robert Verity[2], Richard Alan Nichols[1]

School of Biological and Behavioural Sciences, Queen Mary University of London, London UK[1]
Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, London UK[2]

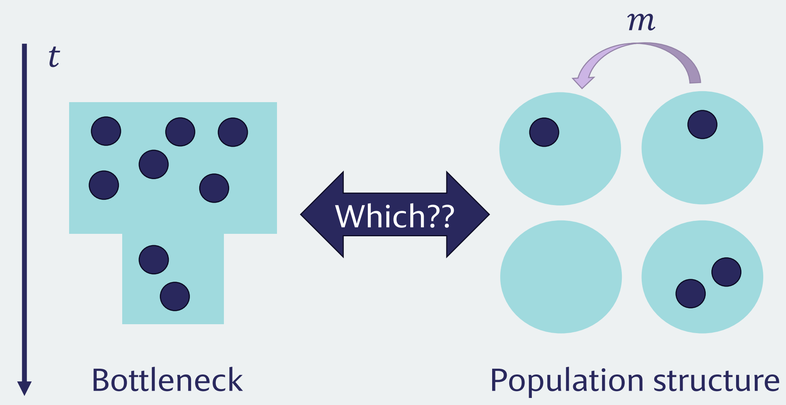
j.k.bansal@qmul.ac.uk
janeeshbansal.github.io

Introduction

Demographic inference tools (e.g. MSMC2) can reveal a population's history. However, the signal produced by the analysis of a **genomic sample from a structured population can be mistaken for a bottleneck** [1, 2].

So, is your inferred population crash actually evidence of population structure?

We provide simple guidelines to help decide.



Methods

ms prime Simulations [3]

ts kit Analysis [4]

MSMC2 Demographic inference [5]

Don't get misled – quantifying the distortions of N_e estimates produced by population substructure

Here we derive simple formulae describing the false history of a recent crash in population size produced by population substructure.

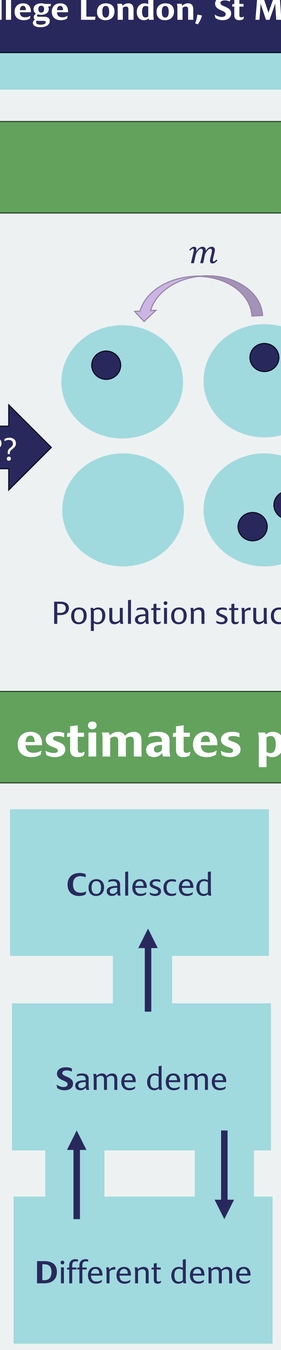
Rate of apparent crash

$$q \approx e^{t\left(-\frac{1}{2N}-\frac{2m}{d-1}\right)}$$

Inflation of N_e estimate

$$N_e = \frac{N}{p}$$

Where $p = \frac{p\left(1-\frac{1}{2N}-2m\right)+(1-p)\frac{2m}{d-1}}{1-\frac{p}{2N}}$

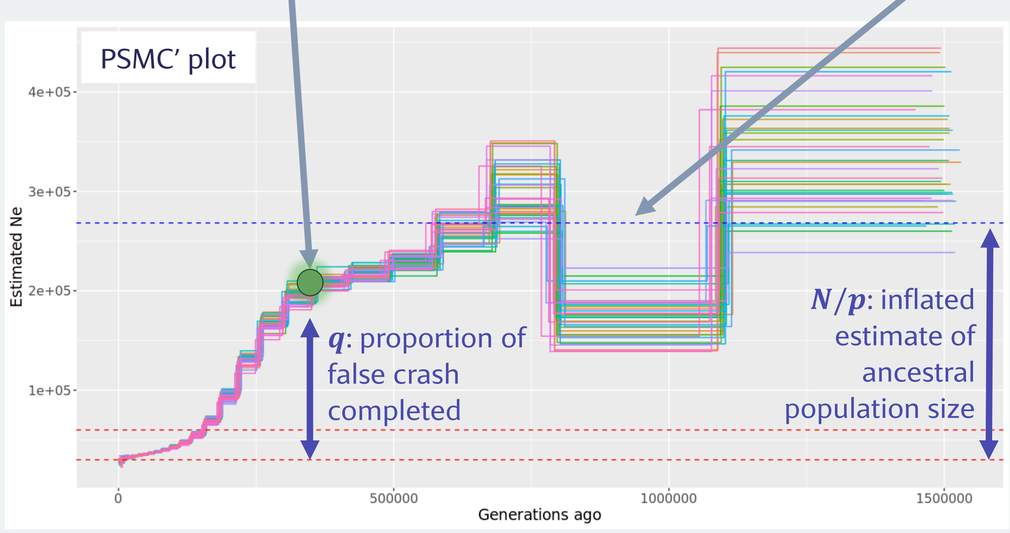


An **ODE** provides the full **shape** of the curve.

$$\frac{dS(t)}{dt} = -2mS(t) - \frac{S(t)}{2N} + \frac{2mD(t)}{d-1}$$

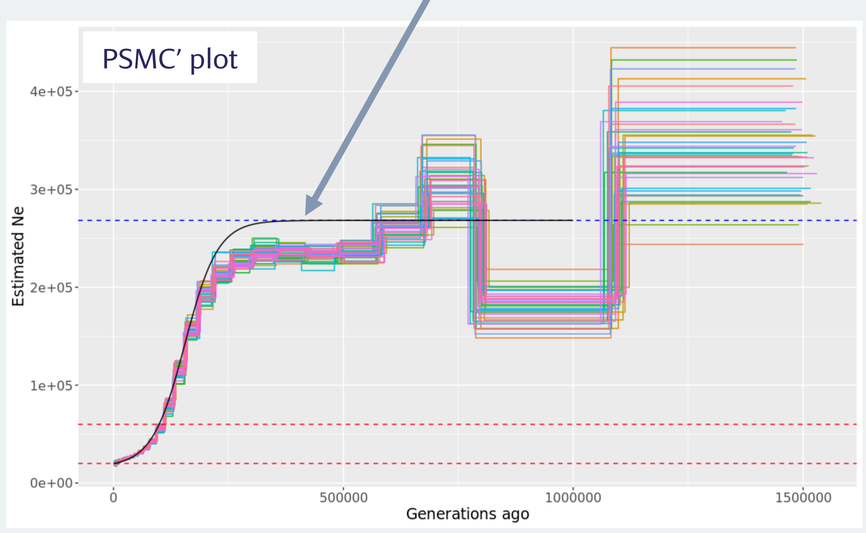
$$\frac{dD(t)}{dt} = 2mS(t) - \frac{2mD(t)}{d-1}$$

$$\frac{dC(t)}{dt} = \frac{S(t)}{2N}$$



Simulation parameters

Number of demes (d)	Population size in deme (N)	Total migration rate (m)
2	30,000	1.07e-6
3	20,000	1.11e-6

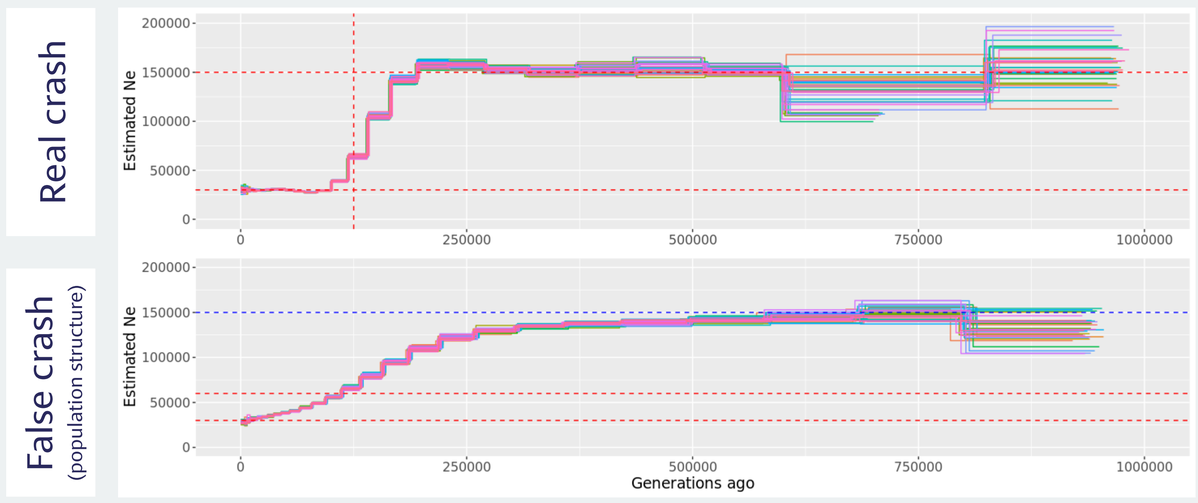


The simple rules and the black curve accurately describe the coalescent rates (equivalent to $1/2N_e$) in the simulations of the subdivided populations that generated these data – so **mathematical models and simulations agree**. However, notice that the algorithm for inferring N_e shows **systematic deviations** from this curve.

More promisingly, the estimates from the **recent past can be more readily interpreted**.

OVERALL: Any one curve could be the result of population structure displaying changes in N_e, however an evolutionary geneticist can make use of their **knowledge of biologically reasonable parameter values** (m, d, N) to distinguish between the different interpretations.

Distinguishing between bottleneck and substructure



Formally, this idea can be implemented by **Approximate Bayesian Computation** to distinguish between the two cases. It will require **informative priors** based on the timing and size of population decline and subdivision for the specific model system.

Conclusions

The effect of population substructure has straightforward effects on the **shape** of the N_e curve in recent time.

Inference tools **overestimate the ancestral coalescence rate** – but why?

Idea in action – **ABC and informative priors**

References

[1] Mazet O, Rodríguez W, Chikhi L. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. Theoretical population biology. 2015 Sep 1;104:46-58.
[2] Wakeley J. Nonequilibrium migration in human history. Genetics. 1999 Dec 1;153(4):1863-71.
[3] Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS computational biology. 2016 May 4;12(5):e1004842.
[4] Kelleher J, Thornton KR, Ashander J, Ralph PL. Efficient pedigree recording for fast population genetics simulation. PLoS computational biology. 2018 Nov 1;14(11):e1006581.
[5] Schiffels S, Wang K. MSMC and MSMC2: the multiple sequentially markovian coalescent. InStatistical population genomics 2020 (pp. 147-165). Humana.

Acknowledgments

Thank you to my supervisors Richard Nichols and Bob Verity. Also, Matteo Fumagalli for helpful suggestions and discussions. Finally, many thanks to the QMUL IT research team for the support and maintenance of Apocrita.