An investigation into the Effects of High Migration on the Ability of PSMC to Infer Population Size History.

# Aims and Objectives

Population size history has been inferred using a range of methods over the years, from using paleontological and archaeological means, to more recent methods using population genetics, based heavily on the amount of genetic variation in a population using sequence data of many individuals within that population. li and durbin [citation], recently proposed a technique using a single individual’s diploid genome to infer the population history of its species using a Pairwise Sequentially Markovian Coalescent (PSMC) model. The applications are vast, but the program specifics are conceptually complicated. However, many have started using pre-existing sequence data to model the demographic history of their species of interest [citation].

PSMC is based on theory that genetic variation within a population increases proportional to population size. This makes sense as there is a bigger gene pool of individuals with their own random changes in DNA. However, when interbreeding between two genetically different species occurs, a large amount of variation is introduced into that population (gene migration), without a notable increase in population size. We have failed to find literature addressing the effects of gene migration on the accuracy of PSMC results. Our aim is to investigate whether PSMC sees this increased genetic variation caused by gene migration and interprets this as a false increase in population size.

To investigate the effects of scenarios such as this, genetic data will be simulated to reflect the genetic variation influx seen in gene migration events. This data will then be given to PSMC to estimate population size history. The accuracy of this estimation will then be quantified and compared to similar PSMC results generated using simulated genetic data that has not be subject to gene migration.

**Why important**

Using genomic data to estimate population size is based on the idea that the amount of variation between individuals in a population is proportional to the number of individuals in that population. In other words, genetic variation reflects the size of the population [citation, maybe not?]. Previous methods have required many individuals from different areas, to be sequenced, limiting the technique either to studies with; a very large budget, due to sequencing cost, or to species that have been extensively sequenced, such as humans. The opportunity to infer population size history from a single genome, allows anyone to model the demographic history of any organism represented in sequence databases such as NCBI. The application of this technique allows us to gain a deeper understanding into what environmental and genetic factors that affected past population bottlenecks and expansions during climatic changes, possibly giving us a way to better understand and predict the long-term effects of the environmental changes that are occurring today. One such study is currently being directed by Dr. Ira Cooke in association with the comparative genomics center at JCU, looking at the signatures of selection within corals. To understanding the genomic factors that allow stress tolerance and survival during events such as bleaching, and water temperature increase many factors including population demographic changes have to be accounted for. Dr. Cooke is therefore developing methods that take into account demographic history, hopefully inferred by PSMC, when studying the genetic factors that might aid us in the preservation of our reefs.

In this study we hope to assess one of the many variables that could affect the results, providing either, insight into a possible limitation of the technique or further strengthening evidence put forward by studies using PSMC.

# Research plan

PSMC explanation

PSMC uses the differences (heterozygote sites) between the two strands of DNA that represent out diploid genome, to estimate population size history. Heterozygote sites within our genome represent the differences in the two sets of genetic codes we receive from out parents, differences that occur due to the random mutation that is bound to happen over time. Using a specified rate of mutation per generation, one can estimate how much time has passed since two genes, that are now slightly different, were identical or *coalesced.* This is often referred to as the time to most recent common ancestor (TMRCA). PSMC will calculate TMRCA for all the heterozygote sites across the genome creating a distribution and compare the frequency of when they coalesced. If many alleles coalesce in a short time span, this indicates that there was a low genetic diversity at that time, which implies a smaller population size.

MS explanation

To generate genetic data that PSMC will use, we will use MS [citation], a program that allows us to simulate the heterozygote sites that would be present in the genome. These genomic characteristics are simulated according to evolutionary parameters, such as rates of mutation and recombination and the changes that would result from demographic events such as migration and bottlenecks. This allows us to test multiple scenarios, set up controls to ensure we can attribute changes in accuracy to specific parameters, and compare the results given to us by PSMC to the “truth” that we specify when simulating the data. Data simulation parameters have previously been calculated to reflect our migration events of interest (Figure 1), Dadi

(<https://onlinelibrary.wiley.com/doi/full/10.1111/mec.14266>).

How we are using ms and psmc together

PSMC takes the genome sequence of an individual as input. Then, the genome sequence is allocated into 100 base pair (bp) bins, and each bin is evaluated on whether it contains a heterozygote site. This turns the genomic sequence into a file (.psmcfa) containing K’s and T’s [and Ns?] indicating if a bin is heterozygote or homozygote respectively. Using this intermediary step, we will produce a script that converts the output of MS (positions of heterozygote sites on the genome) into a .psmcfa file, bypassing the need for an actual genome, but providing PSMC with the same data that it would infer from a genomic sequence.

Evaluating the results

PSMC produces results that can be plotted to visualize change in population size over time. By producing MS data that reflects gene migration but no population size change, we can easily evaluate the accuracy of the PSMC results. If PSMC returns completely accurate results we should see a plot with just a straight line, however, even for our negative controls (no gene migration) the PSMC results will most likely fluctuate around our baseline population size. Accuracy will then be calculated based on the amount of divergence from the baseline population size, using by calculating the area between our “true” population size (straight line superimposed on plot) and the line given to us by PSMC. First using data with no migration (negative control) giving us an expected degree of background inaccuracy, then using data simulated with gene migration.

* How to evt plot the .pl plot in R
* Since it produces a plot, does it produce a function to reflect it?

Or

* Use the history2ms.pl function to return the ms commandline that simulates the results produced my psmc
  + That way we can see what parameters are/are not being estimated accurately
  + However this does not give us an overall accuracy rating, and possibly be too complicated to draw any conclustion about accuracy?

Statistical testing will be based on comparing the mean of the two accuracy sampling distributions (With and without migration) to see if we can say there is a statistically significant difference.

Notes:

To investigate the effects of heterozygosity on the accuracy of the PSMC results, the input data must be controlled in a way that allows us to confidently interpret the results in relation to the level of heterozygosity [and more]. To achieve this the input data will be simulated using MS [citation], a program that allows us to control the parameters of evolution and time and produce the resulting genomic sequence. More accurately MS will produce sites of segregation (sites where the two haploid strands differ) that are would be seen in genomes exposed to the same evolutionary parameters and time. MS will also allow us to inflict specific changes in the population demographics, e.g. bottlenecks, so we will know the “true” population history before it is predicted by PSMC giving us the opportunity to compare truth with program estimation under controlled circumstances.

PSMC takes the genome sequence of an individual as primary input. To analyse the heterozygote frequency and locations, the genome sequence is allocated into 100 base pair (bp) bins, and each bin is evaluated on whether it contains a heterozygote allele (site of segregation). This turns the genomic sequence into a file (.psmcfa) containing K’s and T’s [and Ns?] indicating if a bin is heterozygote or homozygote respectively. To avoid using an actual genomic sequence (this would not allow us to control the variables with the same amount of confidence), we will use the output created by MS (heterozygote sites) to directly produce the .psmcfa file, that is used for the remainder of the steps required to infer population size history.

To produce the .psmcfa from the MS output, a small program has to be written to translate the positions of heterozygote sites, given by MS, into information about the characteristics of the sequence of bins.

PSMC uses the genome sequence of an individual to infer the change in population size over time, so to investigate the effects of the different types of genomes (high low hetrozygosity), we have to be able to control the amount of hetrozygosity in the genome sequence. PSMC takes the genome sequence and produces a .psmcfa file, which is a file containing information about the frequency of hetrozygote alleles thoughout the genome. If is created by dividing the geonome up into 100bp bins, if there is site of segregation (difference between the two haploid strands) in the first 100bp the first letter in the .psmcfa file will be a “K” if there is no segregation between the strands in the next 100bp the next letter in the .psmcfa file will be a “T” and so forth.

To control the heterozygosity of the genome we will use the program MS that simulates genomic changes over time given certain information about mutation rate, rate of recombination etc.

To investigate the effects of the parameters in question on the results of PSMC analysis,

* Create and develop scripts to manipulate the parameters of the program
* Run the program many times to give statistically significant results
* Test the