Estimates of heterozygosity from single nucleotide polymorphism markers are context dependent and often wrong

Jarrod Sopniewski[1] and Renee Catullo[1]

[1]The University of Western Australia

April 16, 2024

Abstract

Heterozygosity is frequently used to describe variation in genetic diversity amongst populations and is often estimated using single nucleotide polymorphisms (SNPs). However, methods of calculating heterozygosity from SNPs have been shown to be affected study design and filtering parameters, reducing their utility and comparability across studies. Though solutions have been proposed to account for identified problems, in our own data, we continued to see inconsistent results. Here, we aimed to further improve methods of reducing inconsistency in these results, specifically by investigating how sample size and missing data thresholds influenced autosomal estimates of heterozygosity (heterozygosity calculated from across the genome, i.e., both fixed and variable sites). We also investigated how the exclusion of tri- and tetra-allelic sites, which is generally standard practice in such studies, could affect eventual estimates of heterozygosity. Across three distinct taxa (a frog, *Litoria rubella*; a tree, *Eucalyptus microcarpa*; and a grasshopper, *Keyacris scurra*) we found autosomal heterozygosity estimates to be affected by samples size when missing data is not allowed and show that this is partly due to the exclusion of tri- and tetra-allelic loci. We also show that the biases introduced by these factors are not consistent between species, or even populations, with higher levels of actual heterozygosity tending to result in larger adverse effects. We propose a modified framework for calculating heterozygosity to reduce these inherent issues and highlight the need for further development in methods such that tri- and tetra-allelic sites can be included in the calculation of population genomics statistics.

Hosted file

`Main_text_file.docx` available at https://authorea.com/users/703602/articles/723148-estimates-of-heterozygosity-from-single-nucleotide-polymorphism-markers-are-context-dependent-and-often-wrong

A) Keyacris scurra

B) Eucalyptus microcarpa

C) Litoria rubella

A) Keyacris scurra

B) Eucalyptus microcarpa

C) Litoria rubella

4

A) Keyacris scurra

B) Eucalyptus microcarpa

C) Litoria rubella

A)

B)

A) Keyacris scurra

| Wallendbeen | Hall | Goulburn | Cooma |

'True' Ho = 0.00396
'True' Ho = 0.00366
'True' Ho = 0.00357
'True' Ho = 0.0021

Population
- Wallendbeen
- Hall
- Goulburn
- Cooma

B) Eucalyptus microcarpa

| Western Victoria | Southern NSW | Central NSW | Central Victoria |

'True' Ho = 0.00612
'True' Ho = 0.00609
'True' Ho = 0.00603
'True' Ho = 0.00593

Population
- Western Victoria
- Southern NSW
- Central NSW
- Central Victoria

C) Litoria rubella

| Kimberley | Southeast | Central |

'True' Ho = 0.00761
'True' Ho = 0.00468
'True' Ho = 0.00246

Population
- Kimberley
- Southeast
- Central

Calculated biallelic heterozygosity relative to 'true' heterozygosity

Sample size (n)

7