Sequencing Technologies DArT

Dartr Workshop.
March, 2024

DArT diversity arrays technology

Overview

- Genotyping Services
  - DArTseq
    - Complexity Reduction
    - DarTseq Standard
    - DarTseq Methylation
  - Ascertainment Bias
  - Targeted Genotyping
    - DArTag
    - DArTmp
    - DArTcap
  - Sample handling
  - Filtering

Genotyping Service: DArTSeq

- DArTseq is our proprietary genome complexity reduction-based sequencing technology.
- It differs from other methods through its ability to intelligently select the predominantly active – low copy sequence – areas in a genome, which are the ones containing the most useful information. At the same time, DArTseq masks the lesser value, repetitive sequences. It does this through the application of a combination of restriction enzymes to fragment DNA samples in a highly reproducible manner.
- A key component of DArTseq is its complexity reduction step. This involves reducing complexity of the genome by cutting it down to a consistent and reproducible fraction from which markers are discovered and called. This process does not introduce ascertainment bias because we are discovering markers within the (restriction) fragment set without the bias

DarTseq: Characteristics

- DArTseq is designed to provide a consistent genomic representation across samples within and between experiments, even over a decade of operations.
- DArTseq has been used to process tens of thousands of samples from hundreds of organisms, and is capable of co-analysing thousands of samples together, with the great majority of markers called across all of the samples.
- These advantages come from the library processing methods of DArTseq which distinguish it from all other complexity reduced genotyping technologies using restriction enzymes.

DarTseq: Standardised

- Laboratory sample processing is performed in a very high throughput streamlined operation utilising end-to-end pipetting robotics and informatics integration.
- With a Laboratory Information Management System (LIMS). All the physical sample processing steps undertaken in the lab are linked with the data storage and analysis components via tight integration with the LIMS. Sample processing robots have automated data exchange with the LIMS.
- Manual operations have been minimised and plate based sample handling prevents the introduction of sample-tracking mistakes after arrival.
- The analytical components for processing DArTseq data and generating marker calls have been tailored specifically to the unique DArTseq complexity reduction and library construction methods. This results in a streamlined and highly effective data analysis pipeline, enabling rapid generation of high-quality marker data from thousands or even tens of thousands of assays in a single analysis.
- Ability to co-analysis with past services of 14 years despite 5 generations of the illumina sequences. The population of fragments in the genomic representation does not vary across multiple sample submissions despite the many years.

DArTSeq: Standard Examples

- Between 700 to 800 organisms sequenced within the ecological sphere!
- Diversity studies
- Population profiling: Separate or sub-species

DArTSeq: Methylation

- While “regular” DArTseq is capable of detecting methylation variation among SilicoDArT markers, we have also developed a dedicated methylation analysis using the DArTseq platform.
- The assay is designed to look for methylation differences between samples on the fragments of the genomic representation.

DArTSeq: Methylation Examples

- Predict age of animals in a population
- Whole genome scanning to find the potential methylated positions related to aging. Product named dartseqmet. Using a pair of restrction enzymes to cut genome in a specific location. Once genome is cut, we have a set of sequences which contain the cytosine nucletoide with or without methylation. Ratio of methylated over non methylated bases. This ratio is used in age-phenotyped individuals to train a model to predict the age of other samples. The best performing ~50 markers can then be used to perform targeted methylation profiling for future sample profiling.
- Potential use for stress quantification.
- Can also perform whole genome methylation scanning
- Or Targeted methylation profiling

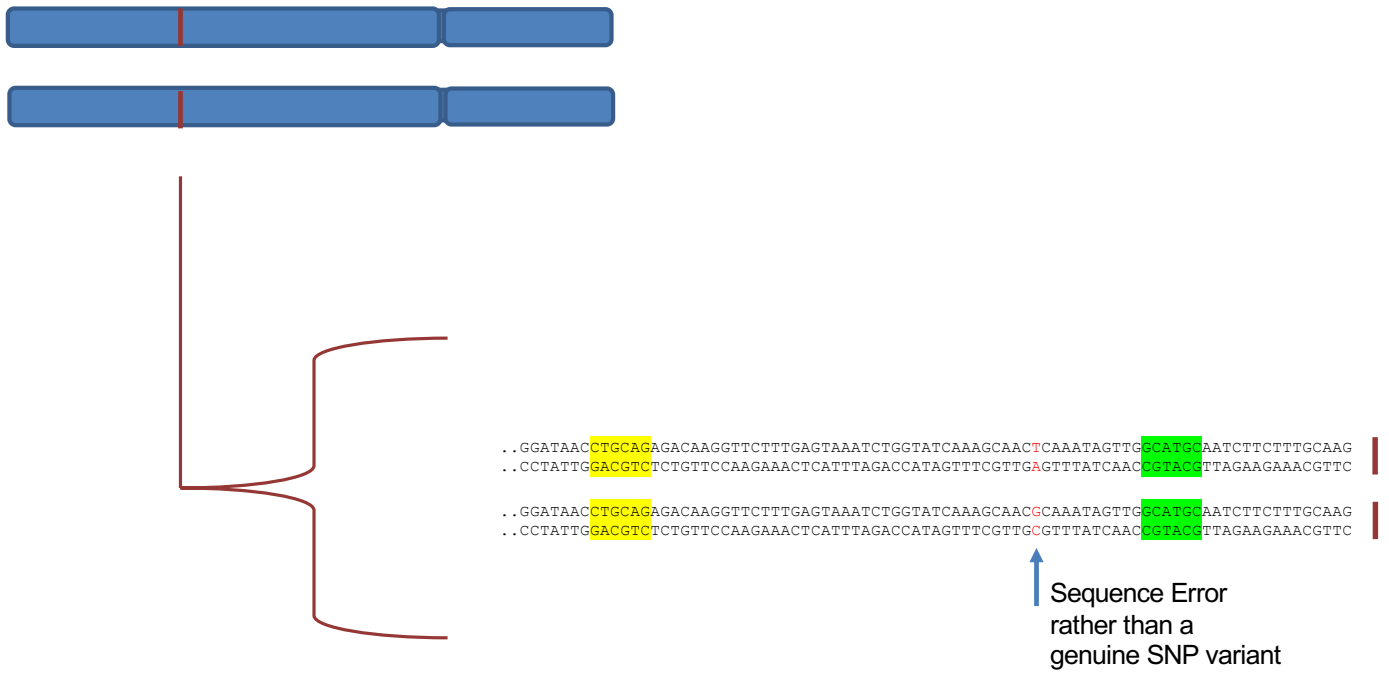
Marker Discovery and Filtering: Rationale

- SNP marker discovery aims to find single base variants which exist at the same locus (location in the genome), and vary between samples and/or the homolgous chromsomes of a single sample.
- The process of marker discovery and filtering aims to distinguish these genuine markers (geniune variants of the same locus) from other classes of variation which may look very similar or sometime identical in the sequencing data. These classes of 'unwanted'variation include:
  - Sequence errors, which include both *sequencing* errors (incorrect base call made by the sequencer) and sequence errors derived from incorrect or unwanted biochemical reactions during library construction.
  - Paralogous sequence pairing, in which very similar or near identical sequences from different loci (locations in the genome) can be paired up incorrectly.
- The marker discovery alogorithm and filtering steps are designed to exclude these later categories of variation and keep only genuine markers as much as possible.
- The next important function of marker filtering is to exclude markers which are genuine, but do not meet the quality requirement for use in downstream applications.

Marker Discovery and Filtering: Rationale

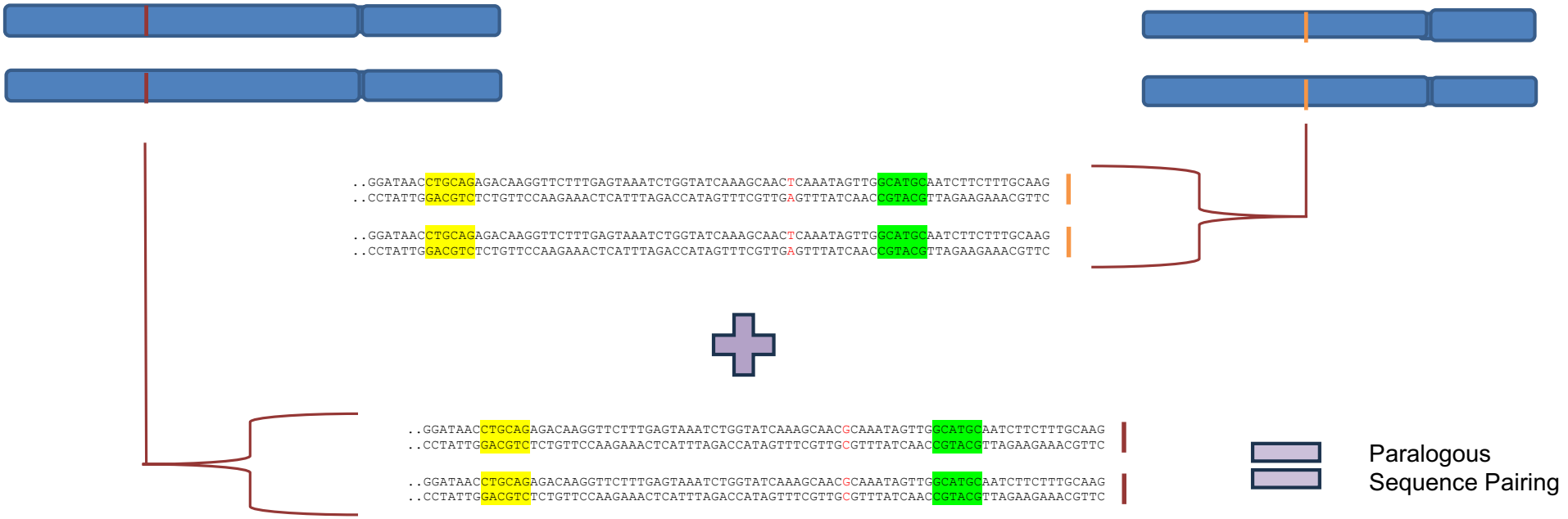
- SNP marker discovery aims to find single base variants which exist at the same locus (location in the genome), and vary between samples and/or the homolgous chromsomes of a single sample.

Marker Discovery and Filtering: Rationale



- Sequence errors derived from various sources need to be distinguished from genuine variants

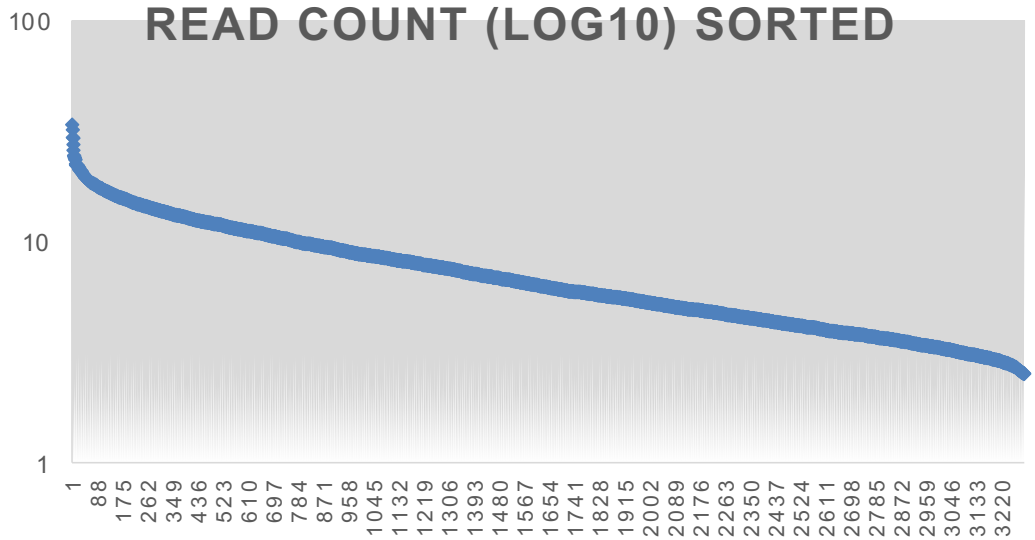
Marker Discovery and Filtering: Rationale



- Sequences derived from different loci may vary by a single base, giving the appearance SNP variation

Marker Discovery and Filtering:

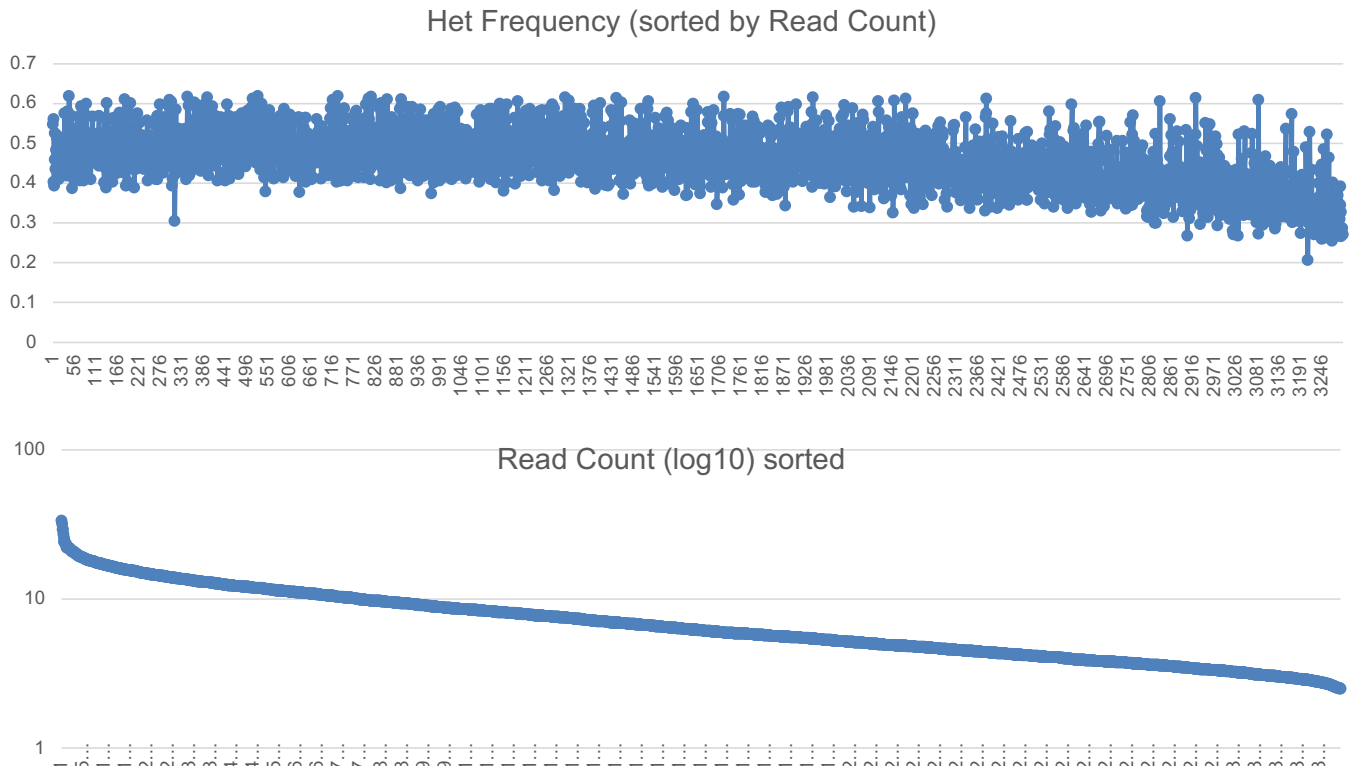
- For a given organism and complexity reduction (Restriction enzyme combination), each fragment tends to show a predictable amplification profile, resulting in read count consistancy across samples, experiments, and across multiple years.



- The DArTseq library construction process differs from other technologies in key aspects which give rise to this highly reproducible profile, with consistent markers discovered and called over tens of thousands of samples.
- The sequencing depth (eg 1.25 or 2.5 million reads per sample) determines how many fragments from this profile will achieve sufficient read count for marker calling.

Marker Discovery and Filtering:

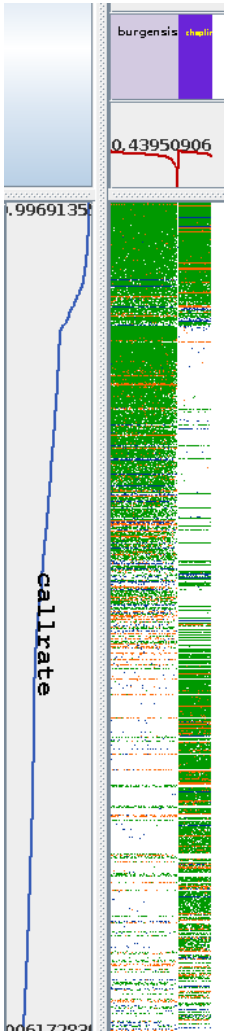
- Heterozygote Calls are depleted at low read counts



Marker Discovery and Filtering:

- Marker filtering is performed on the basis of statistics generated for each candidate marker across the set of samples analysed together.

- There are virtually unlimited possible combinations of samples which could potential be put together for SNP marker calling.
- These may include:
  - Different populations
  - Different sub-species or species
  - Sample groups with very different levels of diversity
  - Sample groups which differ for small or large genome re-arrangements.
  - Sample groups of any of the above types, which also differ from each other in the number of samples per group.
- Marker filtering performed on the combined ‘picture’ may give non-optimal results for some or all groups, yielding a potential source of bias which is inherently complex.
- The context of the required downstream analysis (what question is being asked) is the key to determining how samples should be combined for marker calling.



DArTSeq: Filtering Examples

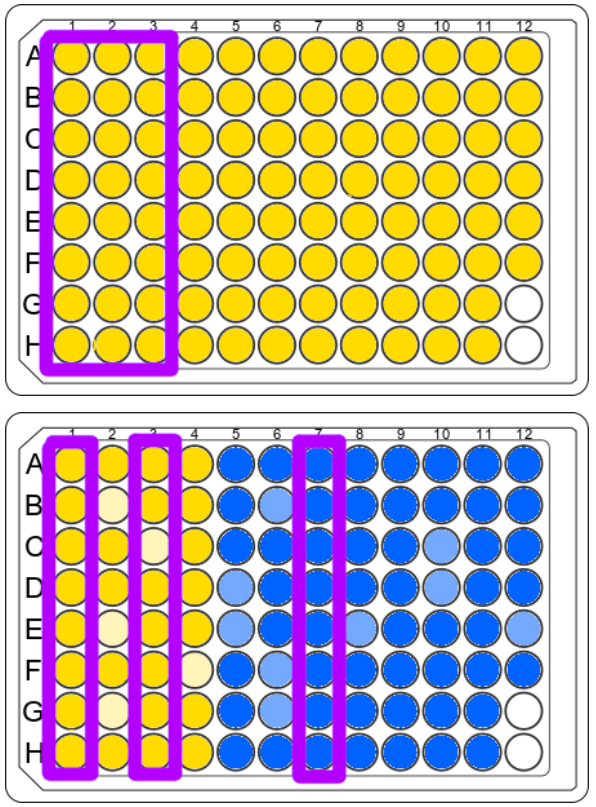
- RepAvg (snps) < 0.86
  - Markers with less than 86% technical replicate consistency will be removed.
  - Too stringent a filter (e.g. < 0.99) will introduce bias, such as markers with high MAF and/or high heterozygosity.
  - On the other hand too relaxed a filter (e.g. < 0.7) could enable non genuine markers to remain.
- AvgCountRef/AvgCountSnp (snps) < 2.5
  - Markers with less an average less than 2.5 counts will be removed.
  - Species dependent
    - Inbreeding organisms (e.g. many plants) < 2.5
    - Out-breeding organisms (e.g. many animals) < 4 or higher
  - Too stringent a filter (e.g. < 20) will usually remove a very large number of markers which were otherwise genuine and very well called

DArTSeq: Filtering Reproducibility

- Reasons for technical reps
  - Quality control of the genotyping process; Full plate and tech rep columns processed by robotics.
  - Supporting high quality markers by focusing on high quality samples (as determined by sample reproducibility)
- Can be useful if a lot of noisy data is introduced by weak structure in low diversity population
- Reproducibility is ideally used to test the quality of markers (and samples) after marker discovery and filtering based on other (non-reproducibility) parameters.
- Hidden or blind replicates will provide a true estimate of error rate in the marker set (take sample quality variation into consideration).

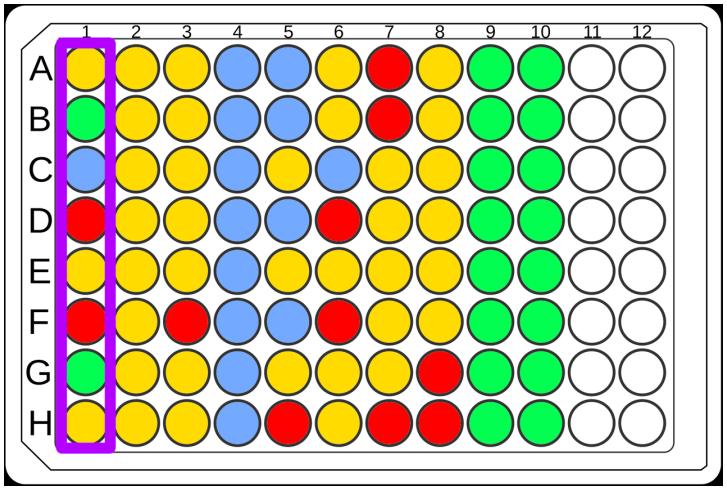
Submission: Plate Assignment and Diversity

- Part of Quality Control a subset of an extract plate is replicated (sampled twiced) to demostrate reproducibilty. Several columns are selected, prioritised as follows:
  - Client has specified which 3 columns (1 to 12) in plate submission comments
  - If multiple material types are present (e.g. dna, tissue) a column of each
  - If a column demostrates poor DNA, the next is prefered
  - Otherwise, the first 3 columns are chosen
- Clients are encouraged to place a mix of diverse representative samples in the first 3 columns to improve quality control coverage



Submission: Optimisation of Restriction Enzymes

- When an organism is to be processed by DarTseq for the first time, a number of restriction enzyme combinations can be tested to find the optimal complexity reduction method for the genome. There are a number of considerations:
  - The standard approach is to genotype 8 samples across the RE combination to be tested.
  - These 8 samples should as closely as possible represent the diversity of samples which are intended to be genotyped in the upcoming orders.
  - These 8 samples should be placed into the first column of a 96-well plate.
  - Other considerations when chossing a complexity method may also apply, particularly consider if the new organism may require co-analysis with an existing organism/product.



Submission: Marker Files

- Current process is providing a template for clients which is filled and then submitted via email

- DArT is currently developing a framework online as part of our upcoming OneDArT release to guide clients in marker file construction and submission

Analysis: Co-analysis and sub-set analysis

- An example of a genotype table with combined analysis of subset of samples with different levels of divergence.
- We will use this to describe appropriate sub-setting of samples for SNP discovery and calling according to the purpose of the analysis and the nature of the samples.

Avoiding Ascertainment Bias

- With DarTseq we can find markers that are significant and vary between the samples of that set which is analysed together.
  - Hence the set of markers to be output vary with the sample set, primarily due to the distribution of polymorphic variants across samples.
- In targeted genotyping we want to ensure that the markers selected for panel construction are derived from the set of diversity we wish to probe when using the panel.
- By selecting meaningful markers with Targeted Genotyping we can ensure visibility of the markers of interest across any sample set.

Targeted Genotyping: DArTag

- Our sequencing technology DArTseq have been used successfully to genotype hundreds of organisms. This method are effective in reporting randomly distributed markers.
- On many occasions, however, there are applications that require genotyping using only a selected marker set. This is targeted genotyping and we have developed three technologies to achieve these results.
- Depending which technology we use there is a variance in their marker density capacity and cost. We will always select the technology for each project based on which is most appropriate and cost effective for each client's needs. We can offer a cost effective solution to practically any application that needs as few as a dozen and up to around 10,000 markers.
- Our Targeted Genotyping uses one of three applications developed by DArT – DArTcap, DArTag and DarTmp.
- An increasing number of organisms have been sequenced with this newer technology. These projects in the ecological domain consist of specific use-cases where targeted genotyping panels are appropriate and where large numbers of samples will be processed.

DArTag: Examples

- Also used with predetermined marker sets for a range of projects
- Koala project:
    - Genetic fingerprinting for individual identification
    - Identification of duplicates
    - Genetic relatedness
    - Sex identification
    - Detection of Chlamydia infection

Targeted Genotyping: DArTcap

- DArTseq analysis is performed initially to discovers markers. A subset of high-value markers of those discovered are chosen to then design "capture probes". Provision of DArTcap services, DArTseq complexity.
- Reduction is performed followed by capture enrichment for the selected marker subset as opposed to direct amplification from the genome as in the other targeted genotyping methods such as DArTag and DArTmp.
- DArTcap includes the benefits of DArTseq complexity reduction whilst reducing the costs per sample and focusing on a defined subset of markers.

DArTcap: Examples

- Used for application in which a defined set of markers (previously determined) needs to be genotyped in new samples
- Population size estimation in Tuna species using Close-Kin Mark-Recapture (CKMR).
- CKMR is is a technique used to infer population demographics by identifying close-kin-pairs (parent-offspring or half-siblings) amongst a collection of sampled animals.

Targeted Genotyping: DArTmp

- DarTmp has a similar technology profile to DarTag, with the primary difference being the number of markers which can be optimally combined into a single panel. Whereas DarTag can combine thousands of markers in a panel, Dartmp is used where several markers up to hundreds of markers are required.
- DarTmp, as with DArTag also provides the greatest cost efficiency when larger numbers of samples are intended to be genotyped with the panel.

DArTmp: Examples

- Smaller panel sizes, ideal for identification including species identification (for example, in fisheries management).

END

Thank you for your attention

I’m sorry to tell you that attention to detail has to start at the very beginning

Renee Catullo

University of Western Australia

Ascertainment bias

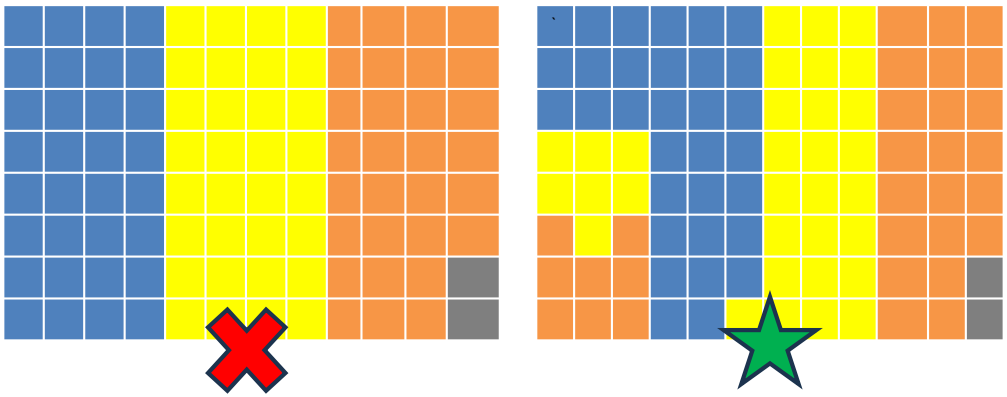
- SNP genotyping data are known to contain bias, normally referred to as ascertainment bias
- Introduced during the genotyping process, because of non-random sampling including:
  - During the selection of populations for novel SNP discovery
  - The number of individuals involved in the discovery panel and selection of SNP markers.
- Can result in changes to inferences of population structure and diversity

ChatGPT!

https://academic.oup.com/g3journal/article/11/9/jkab128/6237890

How to manage but not totally fix

- No perfect complete fix with SNP data
- At the enzyme optimisation stage, ensure you include the full extent of diversity in the optimisation individuals

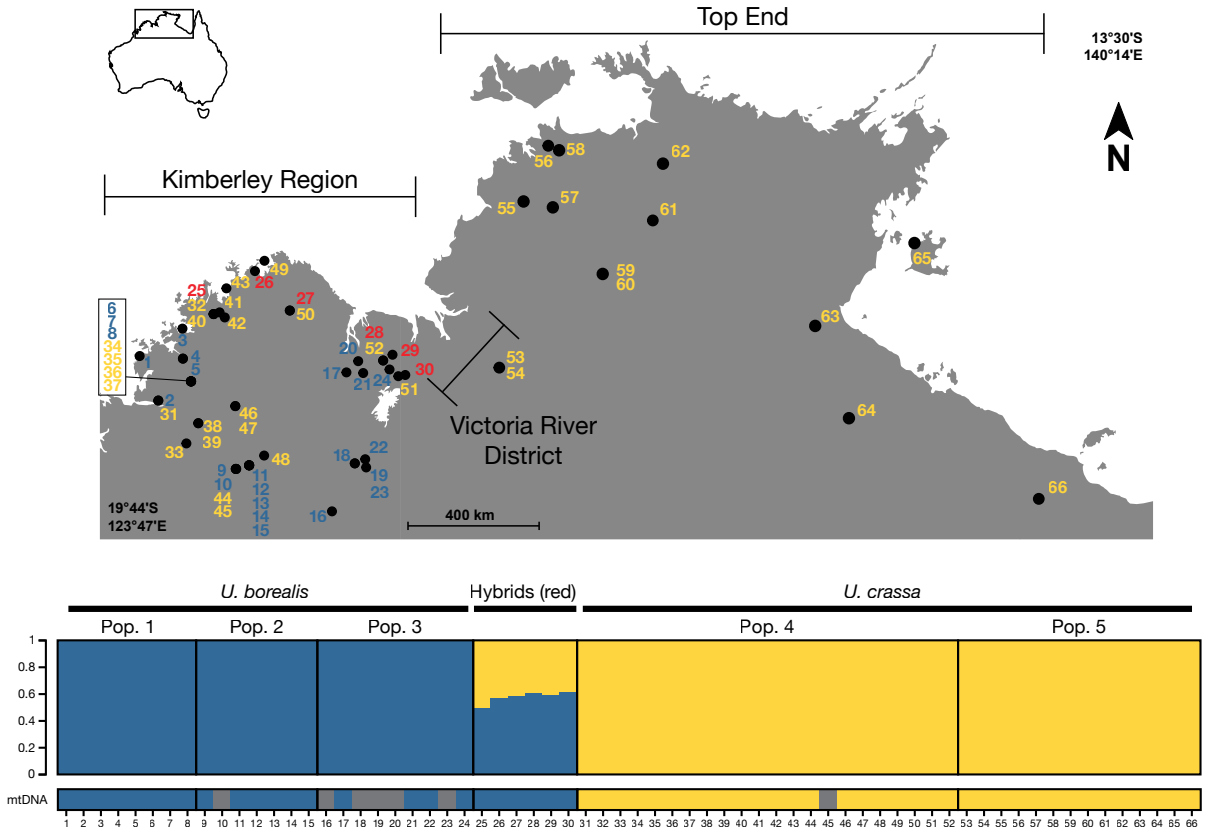


- Diversity Arrays
  - Include the breadth of diversity in the first column of the plate as these are used for enzymes
  - For reproducibility, also do this in the first three rows
  - Get random!
- AGRF
  - Include the breadth of diversity in the individuals provided with greater DNA for the optimisation protocol
  - For reproducibility, duplicate equal numbers of individuals from different groups. This costs money. Do it anyway.

SNP calling is a hypothesis you need to get right before you send a plate for sequencing

- SNPs are explicitly biallelic – sites with more than two alleles are thrown away
- Most SNPs are associated with the deepest divergence in your data, followed by the second, and so on
- Calling SNPs with other species often reduces within-species heterozygosity
- This means if you call SNPs on three species, you will get the least SNPs and potentially incorrectly lowest heterozygosity on your species with the least population structure

Two species


Top End
Kimberley Region
Victoria River District
13°30'S
140°14'E
19°44'S
123°47'E
400 km
N
U. borealis
Hybrids (red)
U. crassa
Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
mtDNA

All

- 1 dataset
- SNPs called for 2 target species AND other outgroup monsoonal species
- Data then filtered to target species

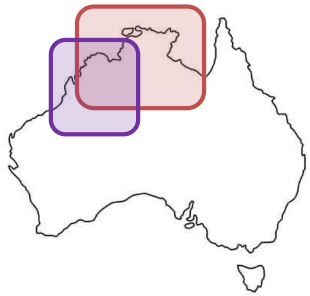
- **3027 SNPs for target species**

B&I

- 1 dataset
- SNPs called for 2 target species

- **3940 SNPs for target species**

B|I



- 2 datasets
- SNPs called on target species independently
- No hybrids in data

- **9,586 SNPs for species 1**
- **13,812 SNPs for species 2**

Filters: 95% call rate by locus, remove singletons, 90% call rate by individual, 99.5% reproducibility, 1 SNP per fragment, read depth 5< & >20

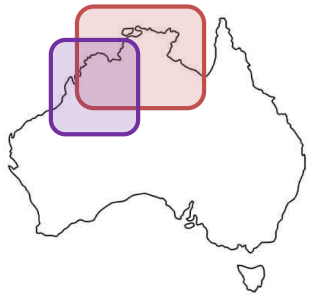
All

Give me SNPs that distinguish between many species, but I will filter them to the ones I care about

B&I

Give me SNPs that distinguish between *U. borealis* and *U. crassa*

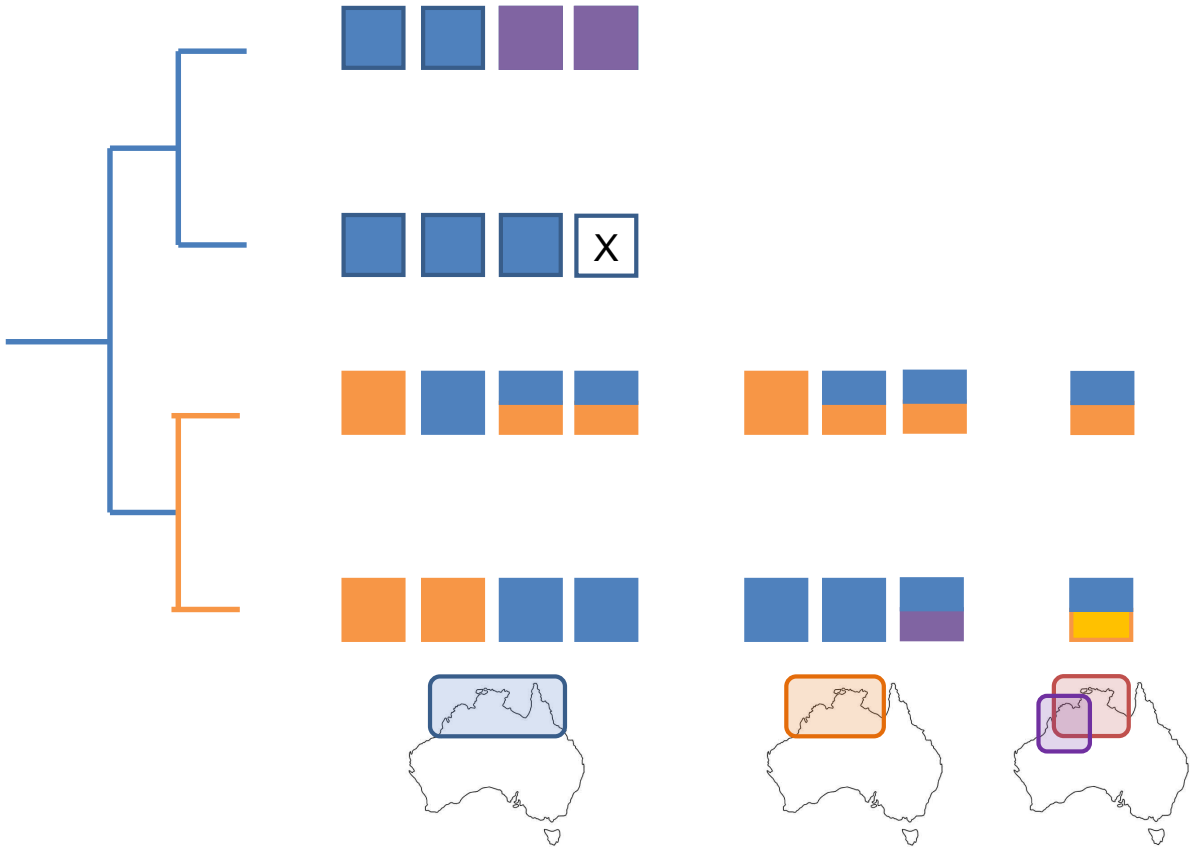
B|I



Give me SNPs that distinguish population structure within *U. borealis* and within *U. crassa*

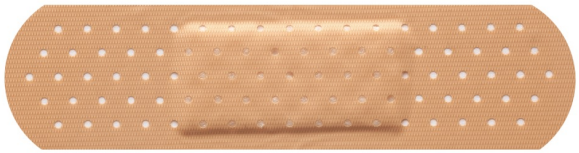
Why are these different questions?

- These limitations are due to calling only biallelic sites
- Underlying diversity differences could mean SNPs are lost at different rates for different populations



Signals you might have called SNPs wrong

- If you included outgroups for a question that doesn't require outgroups
- If you had SNPs called across multiple species and have weirdly low heterozygosity within species
- The fix is not to delete individuals – you need to re-call SNPs on your target group



Addressing biases in SNP calling

Jarrod Sopniewski & Renee Catullo

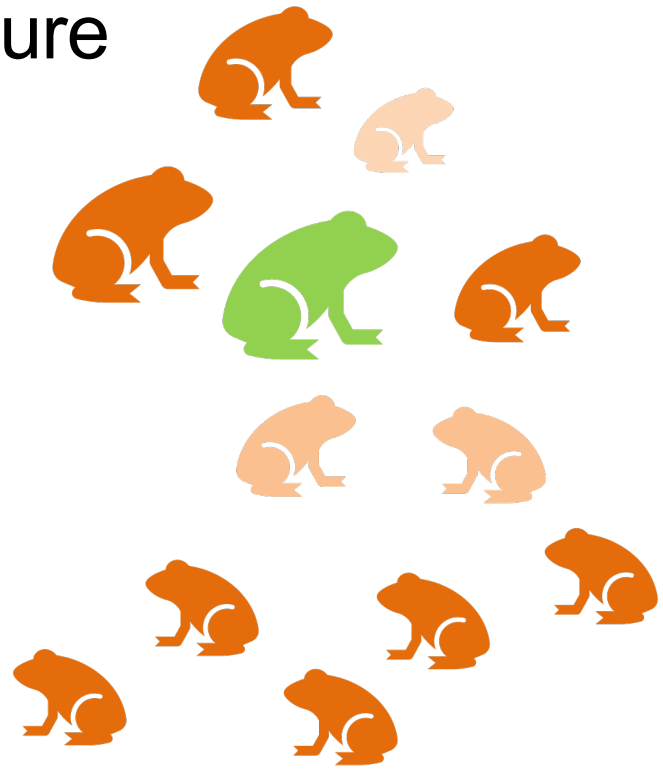
@j_sopniewski

Perth, Whadjuk Noongar boodjar

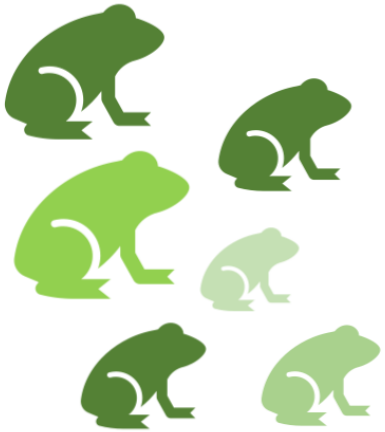
THE UNIVERSITY OF WESTERN AUSTRALIA

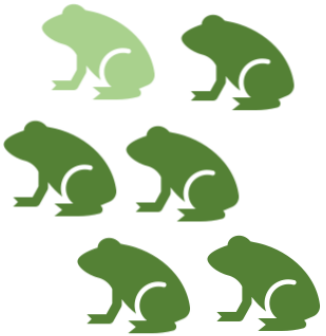
Why do we do pop gen?

Population structure

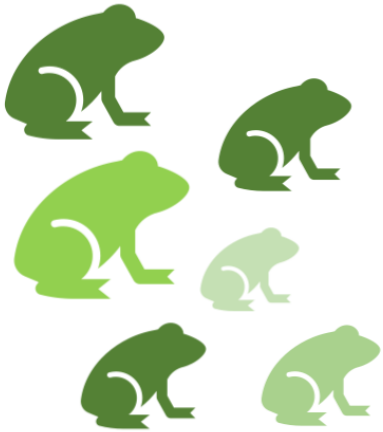


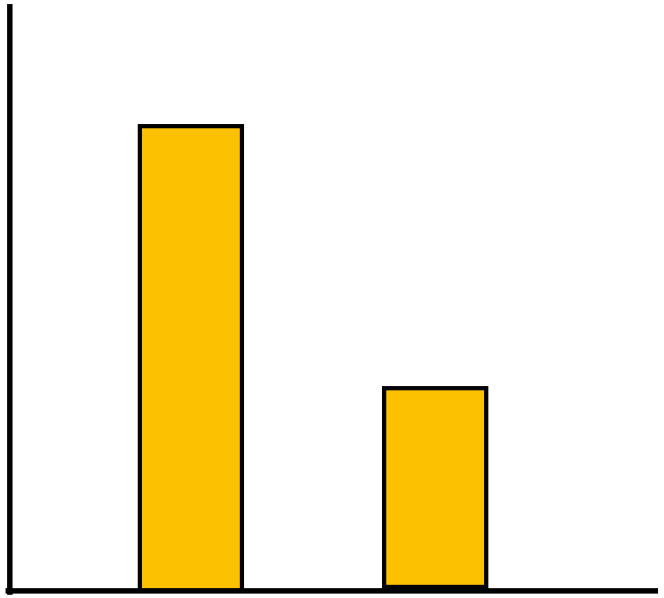
Heterozygosity

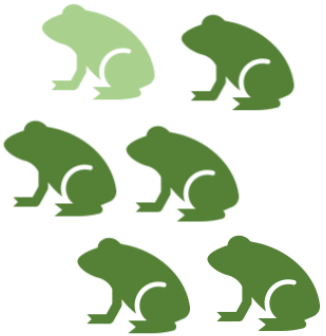




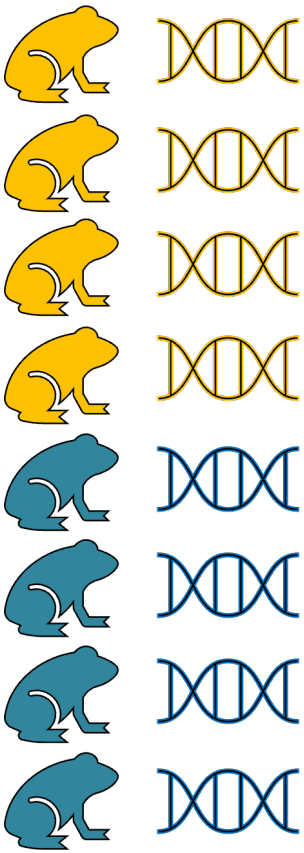
Heterozygosity

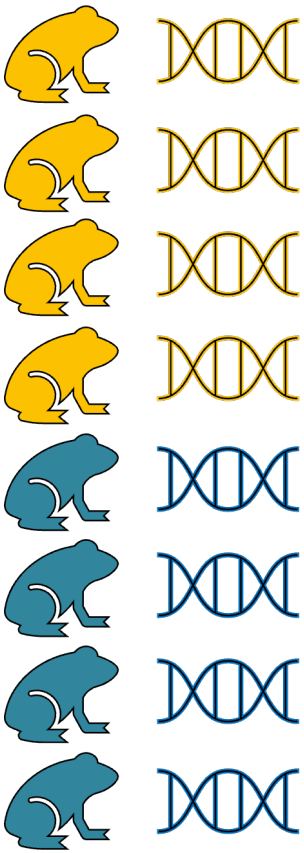


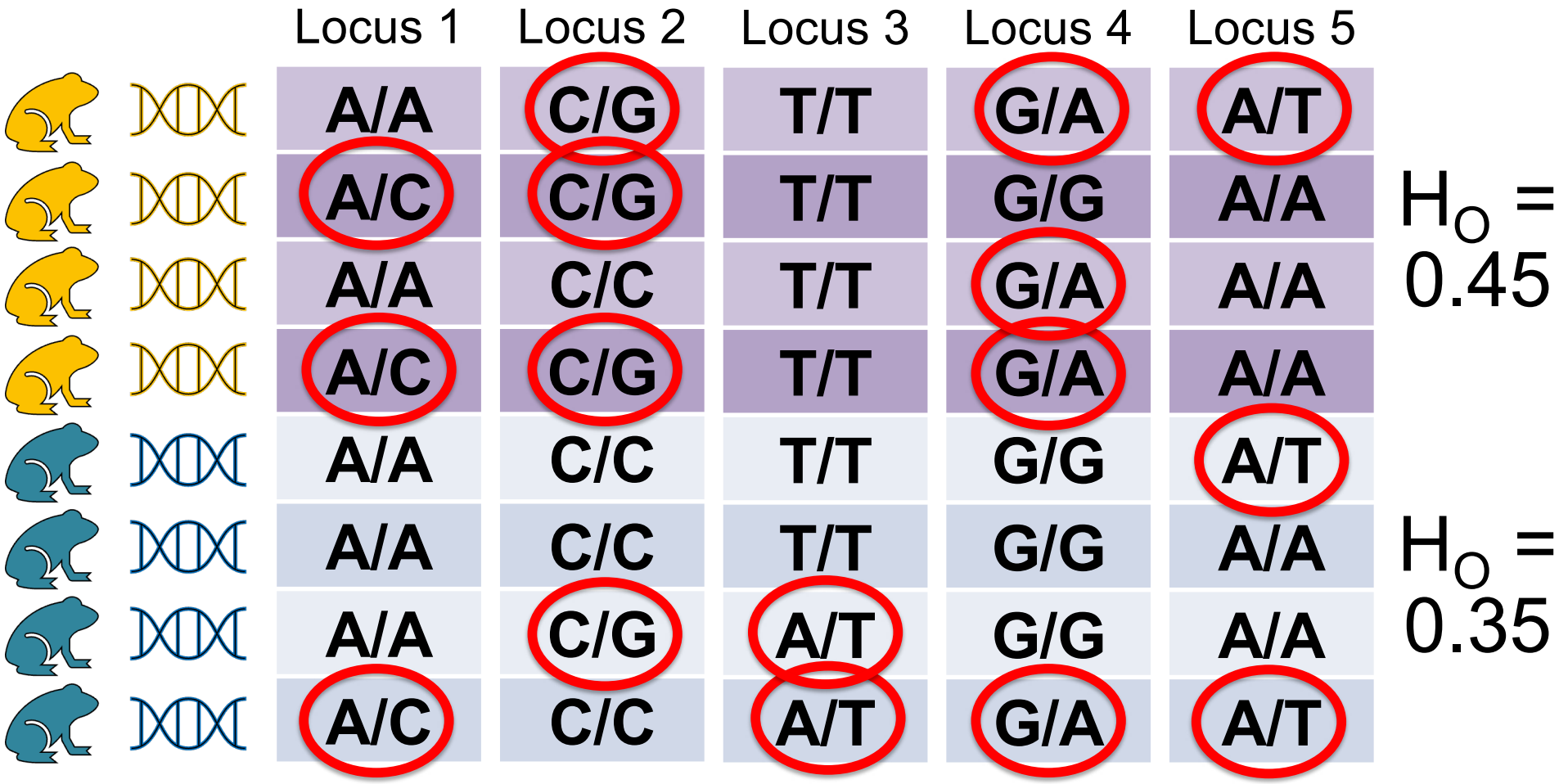


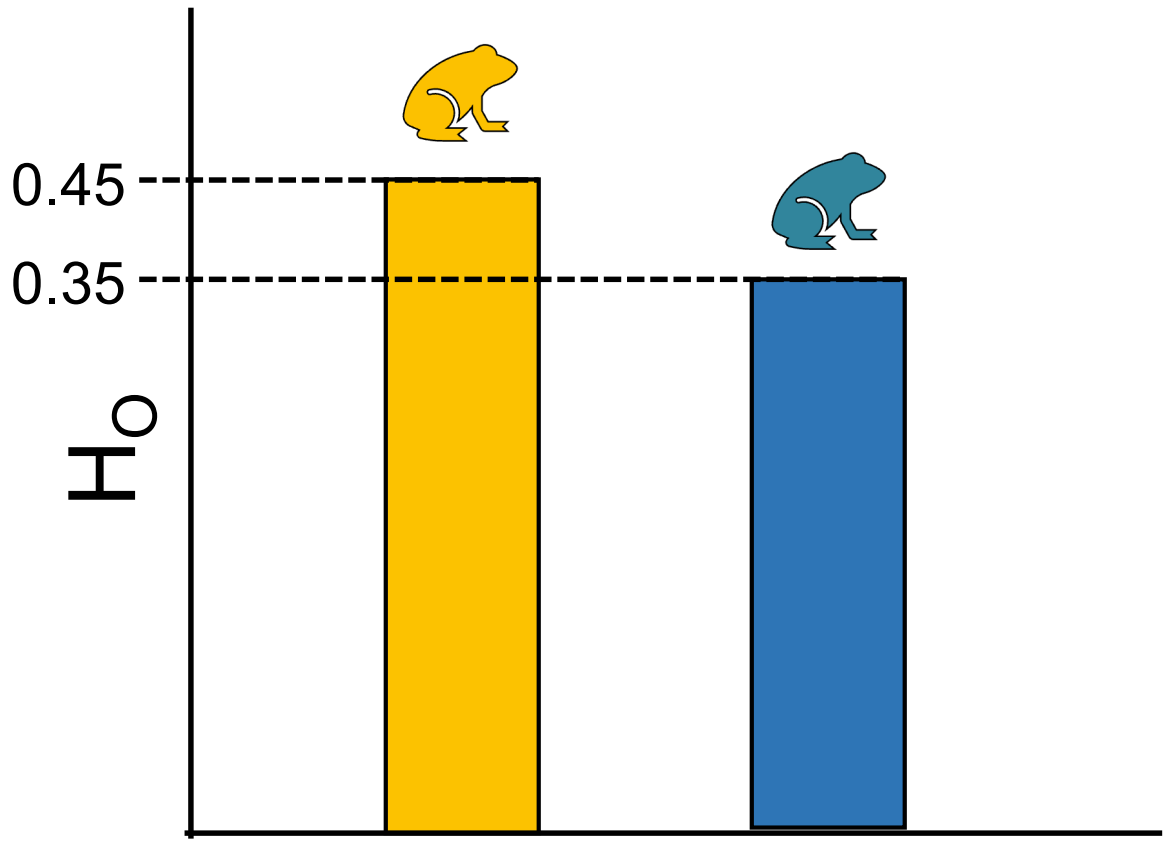


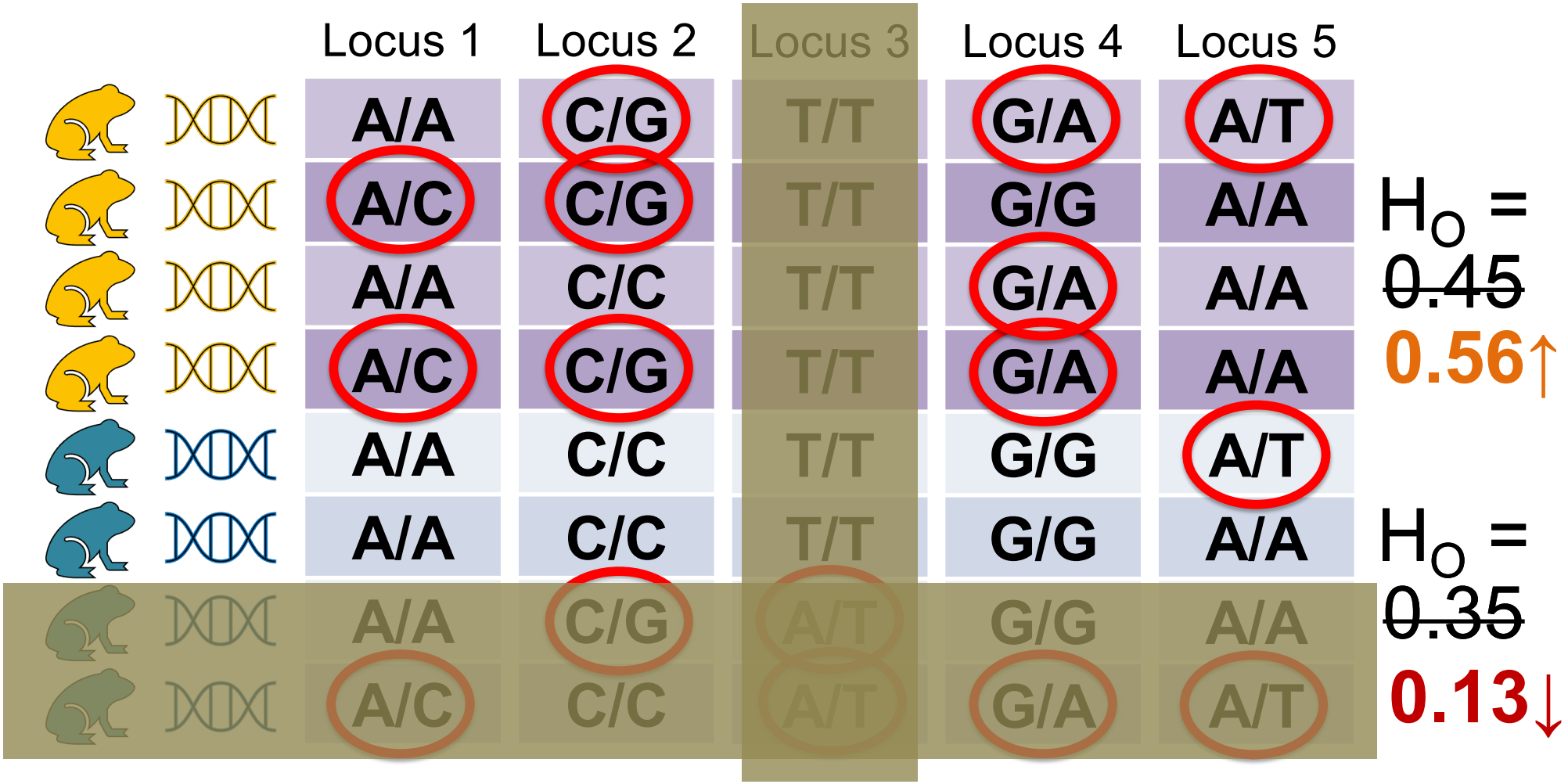
How do we do it?

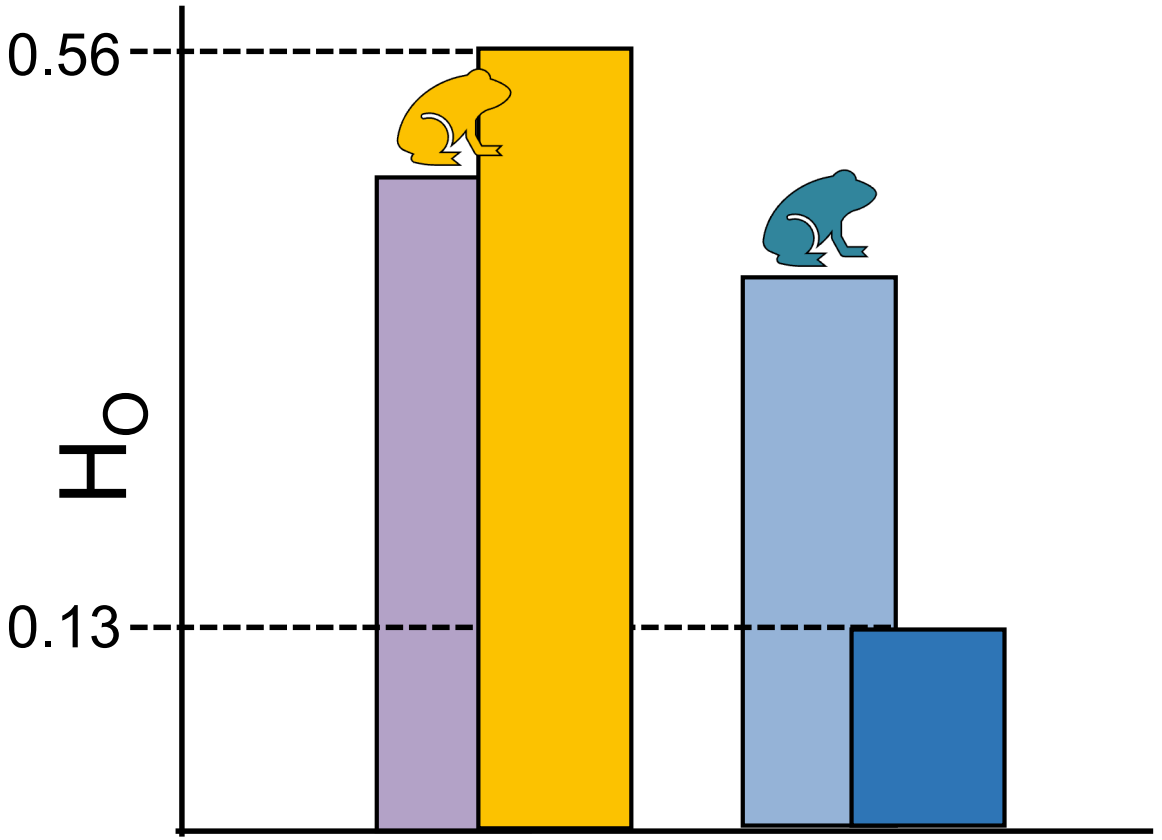





Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
A/A C/G T/T G/A A/T
A/C C/G T/T G/G A/A
A/A C/C T/T G/A A/A
A/C C/G T/T G/A A/A
A/A C/C T/T G/G A/T
A/A C/C T/T G/G A/A
A/A C/G A/T G/G A/A
A/C C/C A/T G/A A/T
H_O = 0.45
H_O = 0.35


0.45
0.35
H^O


Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
A/A C/G T/T G/A A/T
A/C C/G T/T G/G A/A
A/A C/C T/T G/A A/A
A/C C/G T/T G/A A/A
A/A C/C T/T G/G A/T
A/A C/C T/T G/G A/A
A/A C/G A/T G/G A/A
A/C C/C A/T G/A A/T
H_O = 0.45
0.56↑
H_O = 0.35
0.13↓


0.56
H^O
0.13

How has this been tested?

RESEARCH ARTICLE | Full Access

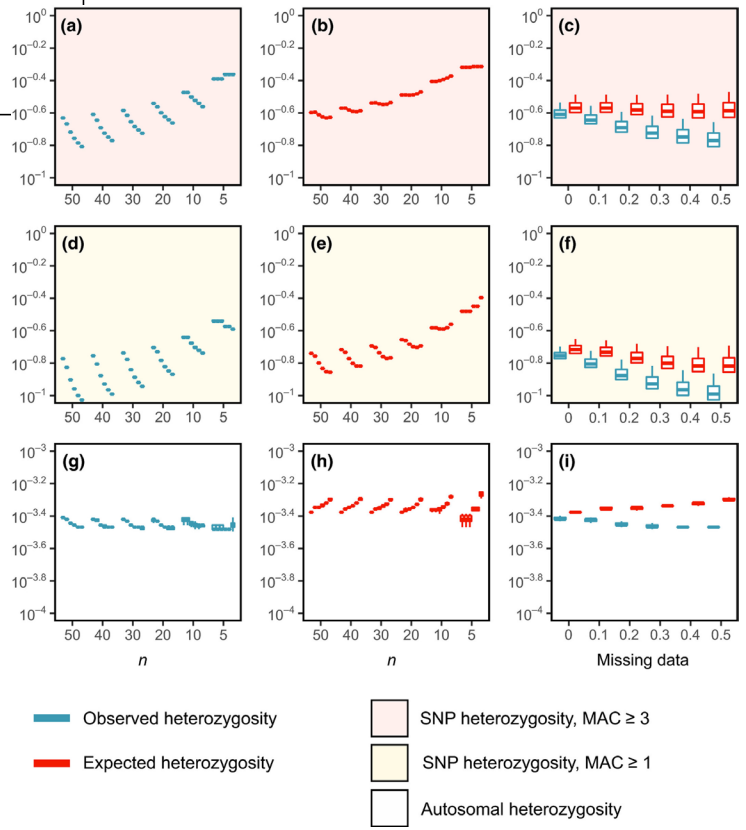
Unbiased population heterozygosity estimates from genome-wide sequence data

Thomas L. Schmidt, Moshe-Elijah Jasper, Andrew R Weeks, Ary A Hoffmann

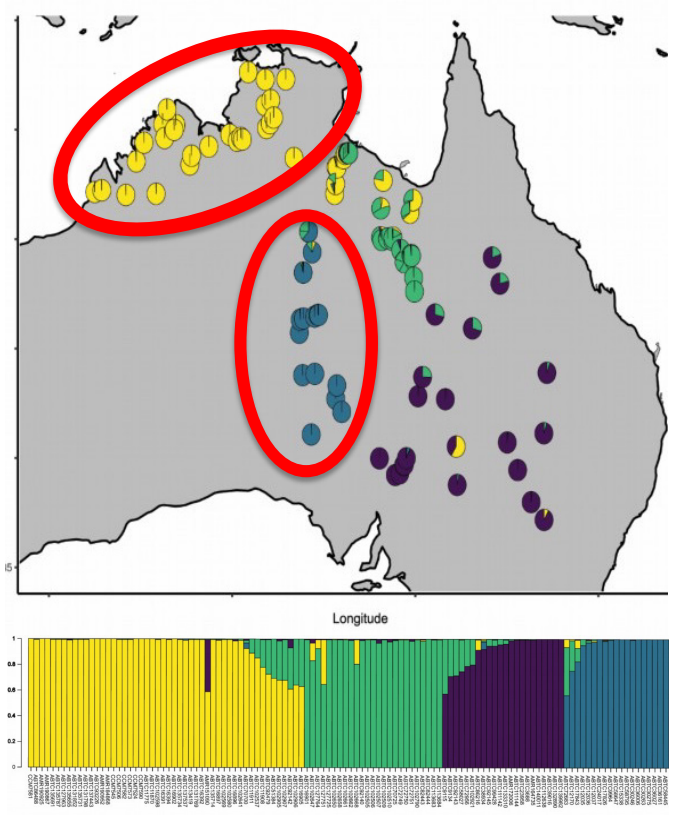
Recommendations from Schmidt et al.:

(a) autosomal heterozygosity should be reported instead of (or in addition to) SNP heterozygosity
(b) sites with any missing data should be omitted
(c) populations should be analysed in independent runs

This being the case, our findings that heterozygosity estimates can be consistent even at low *n* (figure 7a) point to the optimal design for heterozygosity being deep sequencing of a small number of individuals (perhaps 5–10) from each population, rather than shallower sequencing of many individuals.

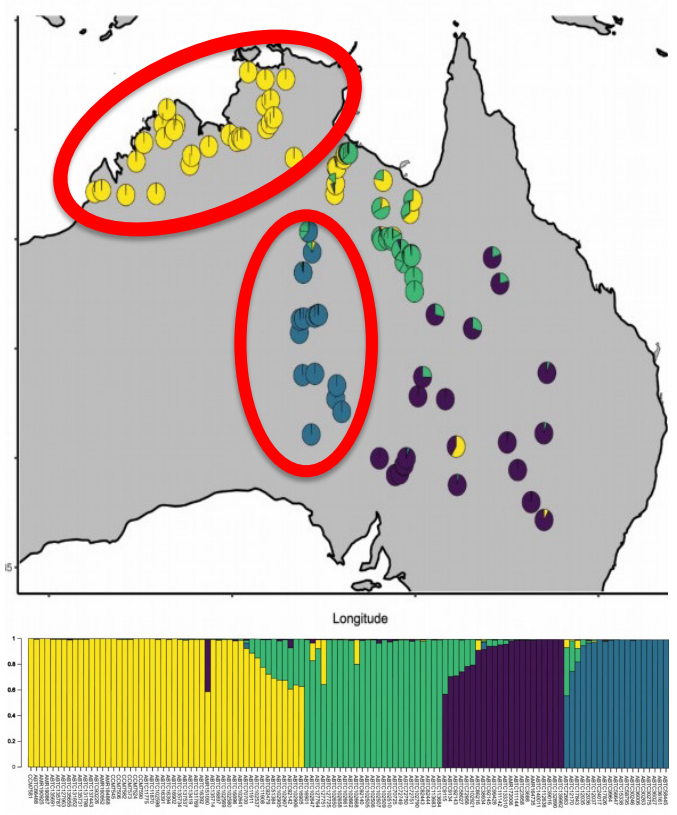


- 2 populations of *Litoria rubella*



Purser et al. (*accepted*)

- Multiple populations of *Litoria rubella*
- DartSeq by Diversity Arrays
- Raw reads assembled in *Stacks 2* (Catchen et al., 2013; Rochette et al., 2019)



Purser et al. (*accepted*)

- Analyse populations collectively or independently?
- Sample size?
  - Ratios?
- Call rate?

And we found? A lot of the same things as Schmidt et al.

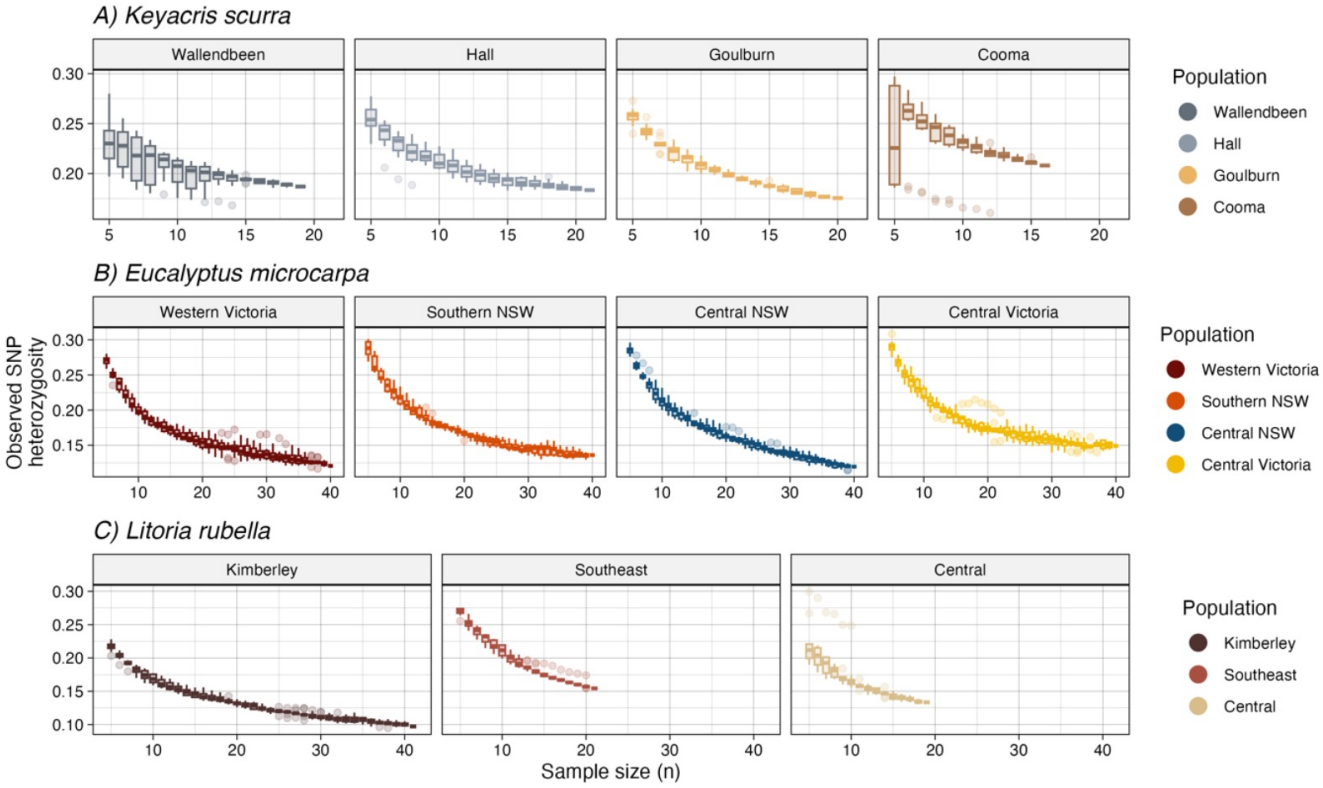
RESOURCE ARTICLE | Open Access

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

Jarrod Sopniewski, Renee A. Catullo

First published: 03 March 2024 | https://doi.org/10.1111/1755-0998.13947

SNP Heterozygosity vs sample size

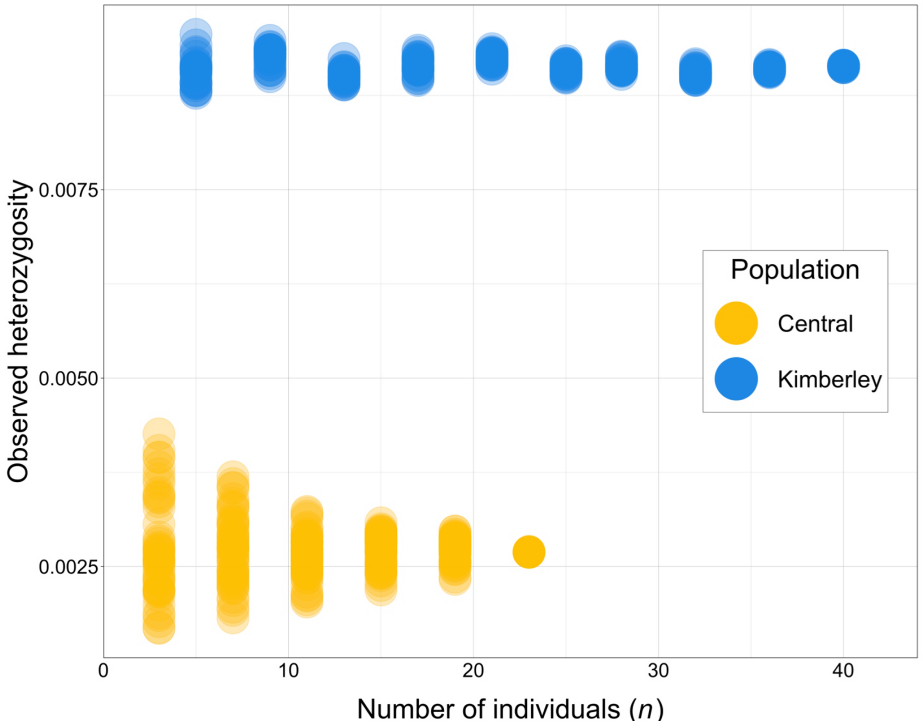


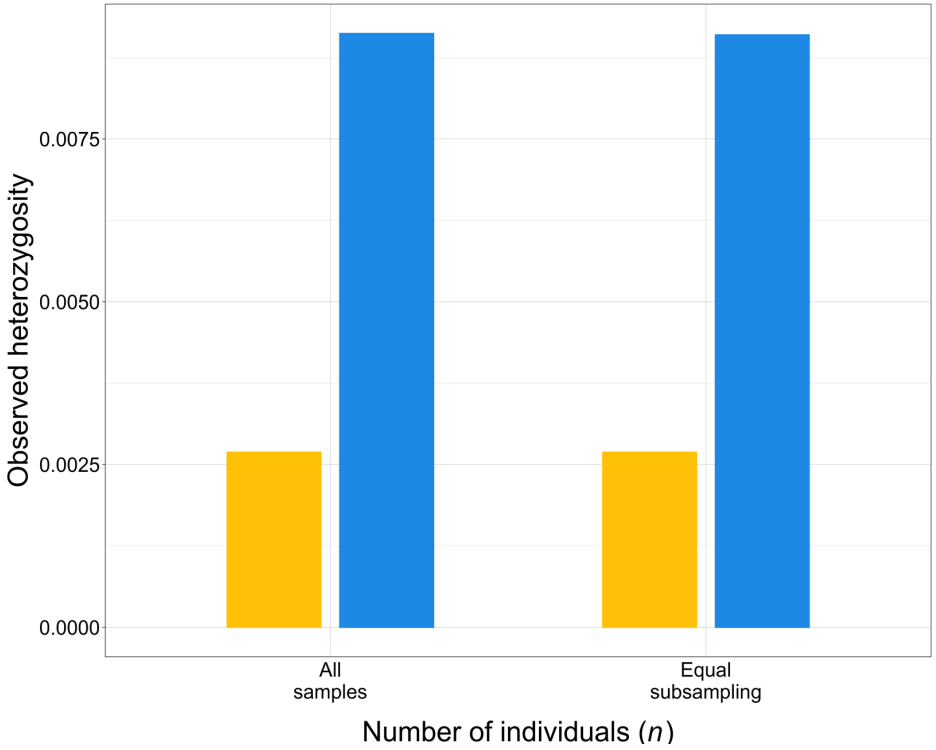
As sample size increases, the number of sites assessed for heterozygosity calculations (i.e., the number of sites with variation in at least one individual) increases.

Proportionally, however, the actual number of polymorphic sites increases at a far lower rate, meaning that estimates of observed heterozygosity decrease dramatically at higher sample sizes.

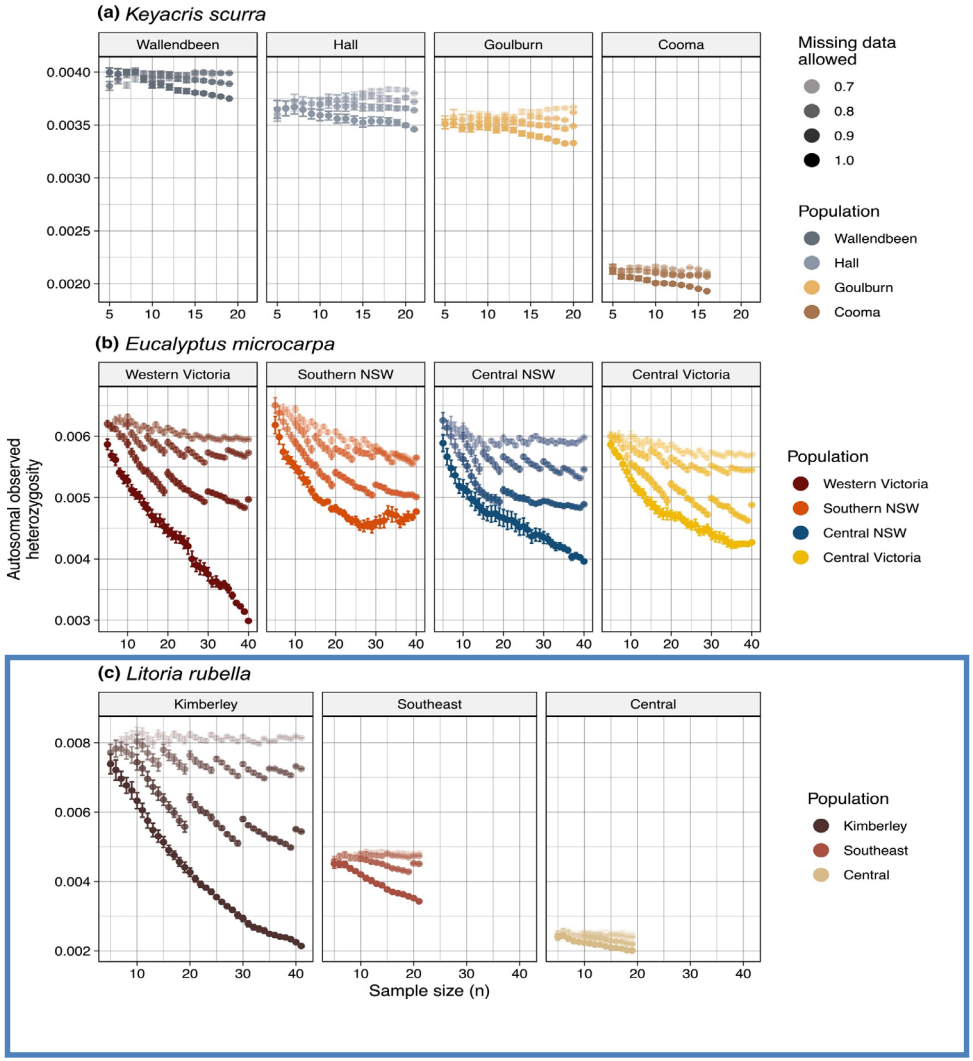
Any fixes?

Autosomal heterozygosity (Schmidt et al.)

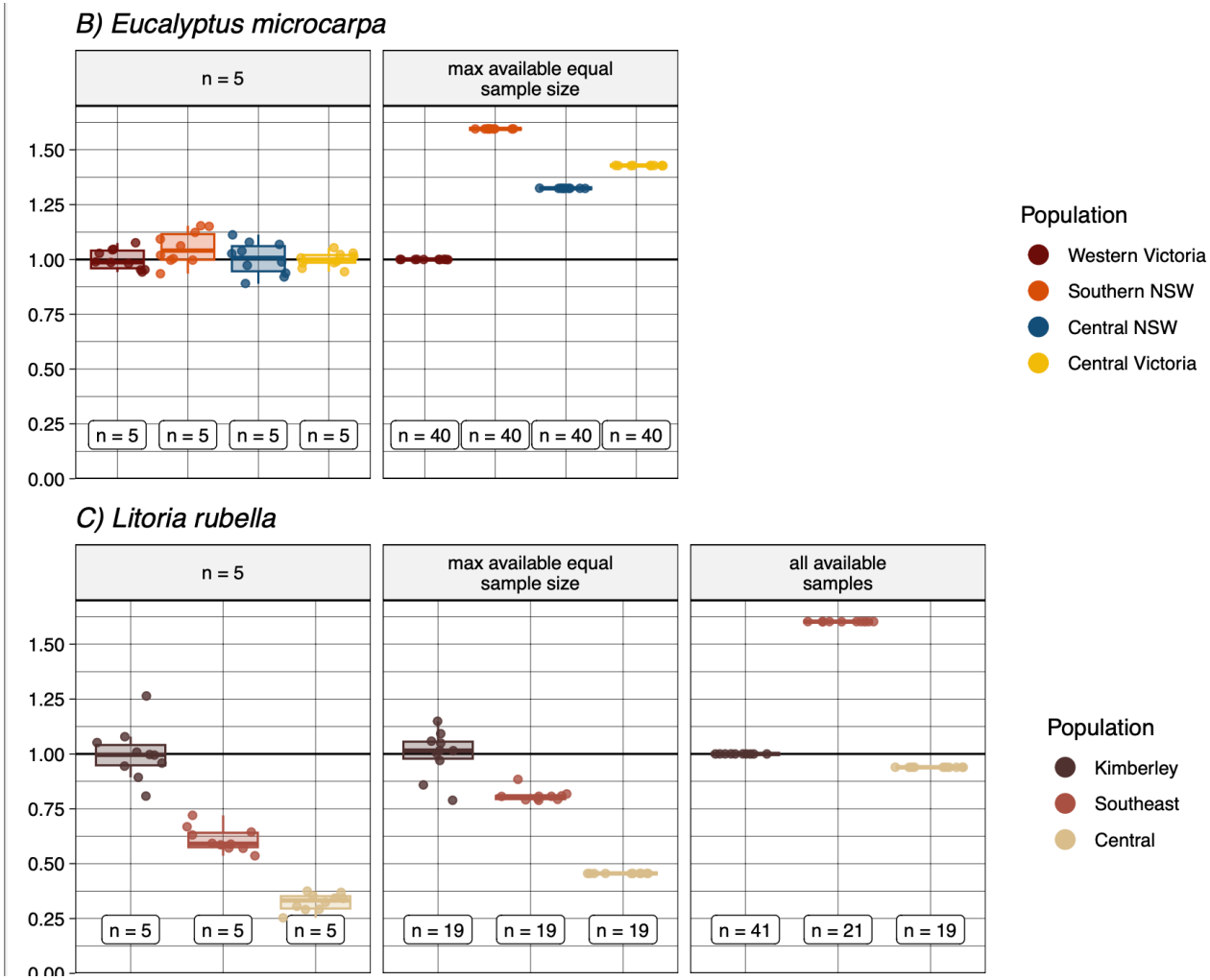




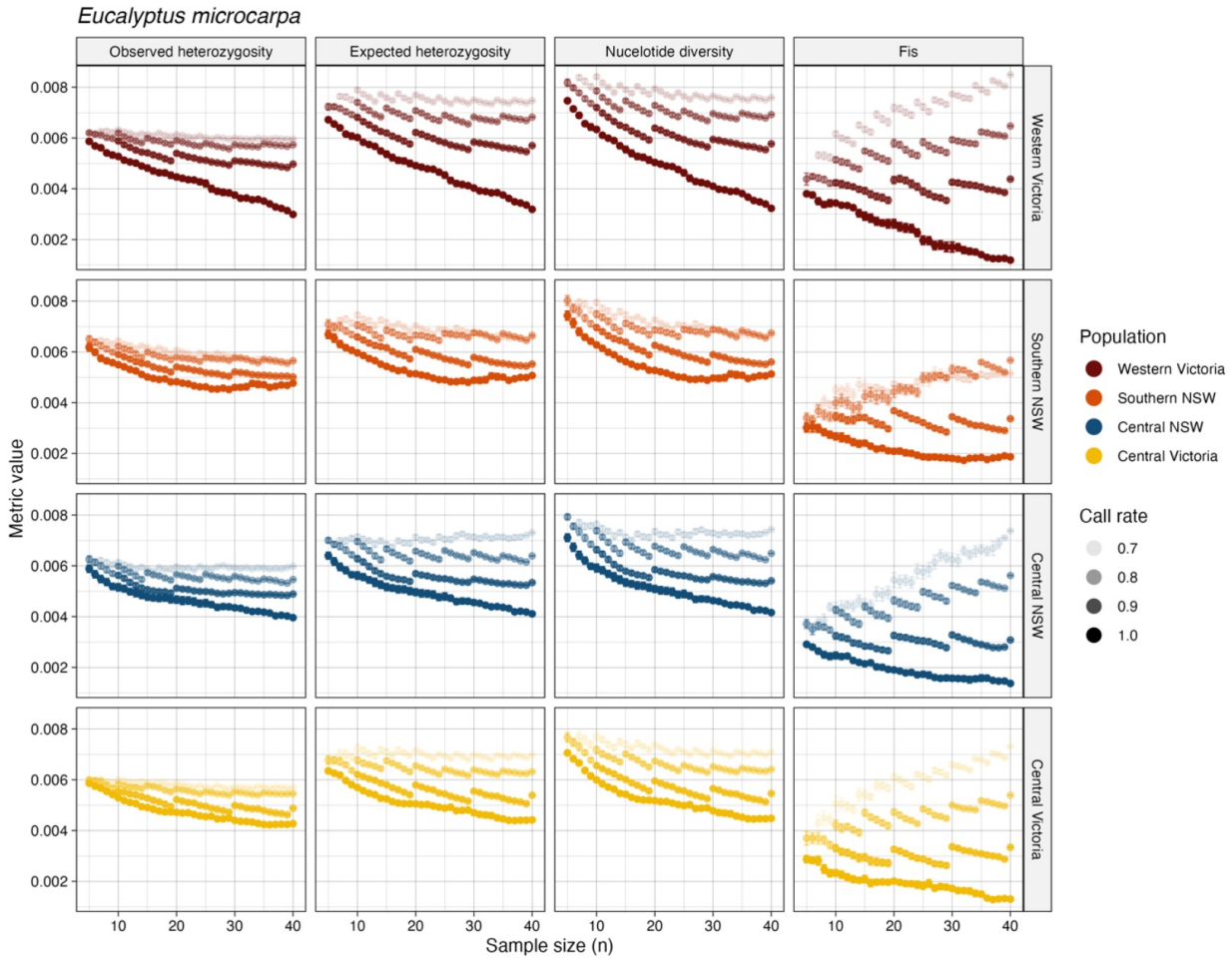
All good?


(a) Keyacris scurra
Wallendbeen
Hall
Goulburn
Cooma
0.0040
0.0035
0.0030
0.0025
0.0020
5
10
15
20
Missing data allowed
0.7
0.8
0.9
1.0
Population
Wallendbeen
Hall
Goulburn
Cooma
(b) Eucalyptus microcarpa
Western Victoria
Southern NSW
Central NSW
Central Victoria
Autosomal observed heterozygosity
0.006
0.005
0.004
0.003
10
20
30
40
Population
Western Victoria
Southern NSW
Central NSW
Central Victoria
(c) Litoria rubella
Kimberley
Southeast
Central
0.008
0.006
0.004
0.002
Sample size (n)
Population
Kimberley
Southeast
Central

Even with Autosomal Ho, relative numbers matter at call rate =1



Other metrics are also impacted by filtering but this is not well explored



Take aways…

- Filtering and study design decisions have consequences!
- There is serious disagreement on exactly what the problem is and how to fix it

TRY:

- Take note of Schmidt et al. and Sopniewski & Catullo
- Make using autosomal heterozygosity standard practice. This means running STACKS or other pipelines on raw data, Estimate autosomal Ho for each population independently.
- Balanced populations, test your call rate filters
- Repeated sampling with smaller numbers can give confidence in estimates
- Filters that preferentially remove heterozygous sites are more impactful at large sample sizes

Always get the raw data

1. Allows you to repeatedly revisit your SNP calling as your understanding of the questions evolves (different individuals/different parameters)
2. I will reject your paper if you don’t publish the raw data on the NCBI sequence read archive or Bioplatforms Australia

Galaxy Australia

- Bioinformatics resources available to through the AAF login (your university)
- Lots of compute power
- Has the Stacks2 pipeline available

https://australianbiocommons.github.io/how-to-guides/stacks_workflows/stacks

https://catchenlab.life.illinois.edu/stacks/manual/