Weighted estimation of linear mixed models under two-phase sampling for kākāpō genomics data

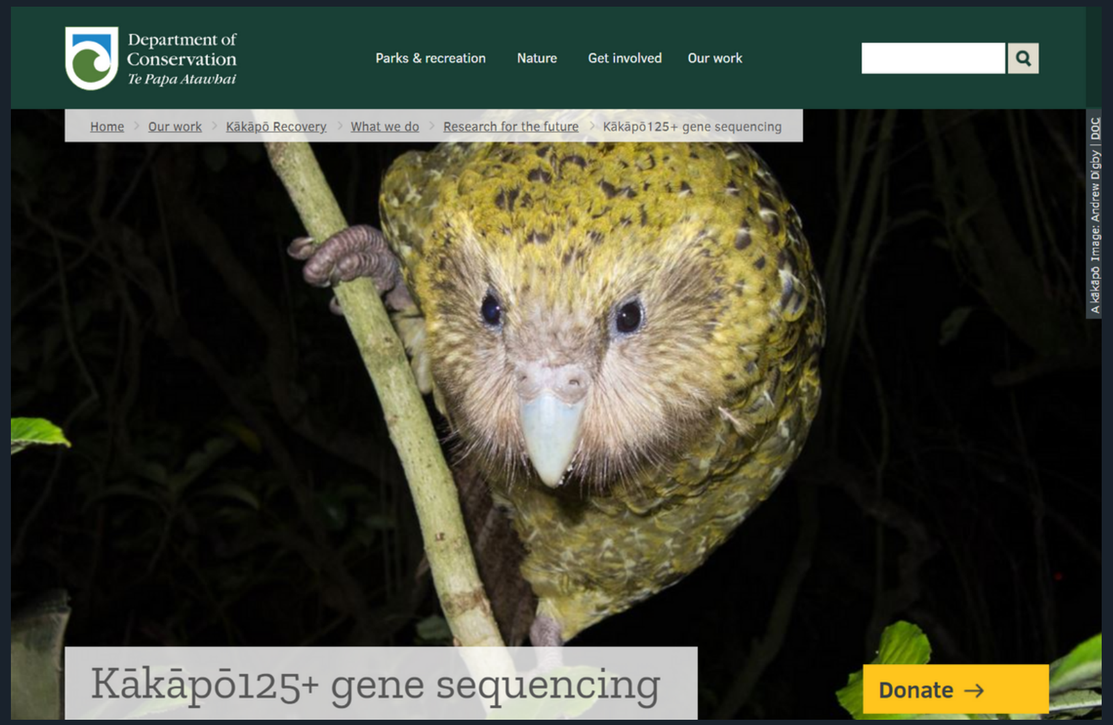
PRESENTED BY
PEI(ZOE) LUO

SUPERVISED BY
PROF THOMAS LUMLEY & DR BEN STEVENSON

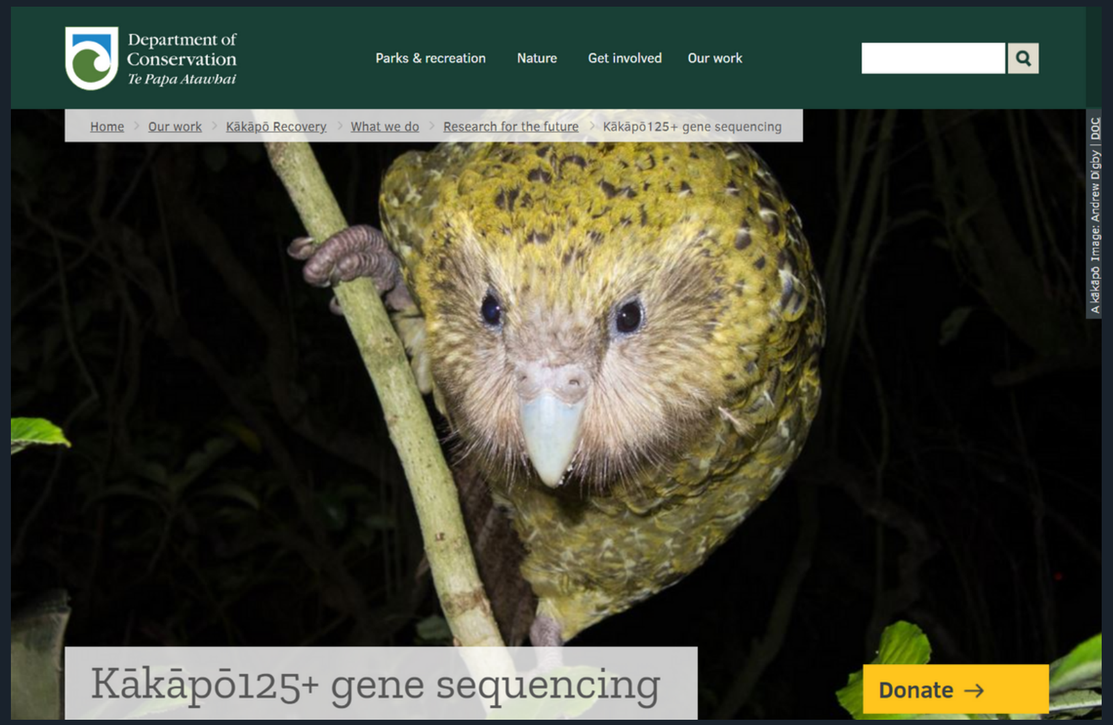
THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Background

- Kākāpō is a critically endangered species in New Zealand, and it is the world's largest, the only flightless, and the only lek-breeding parrot.

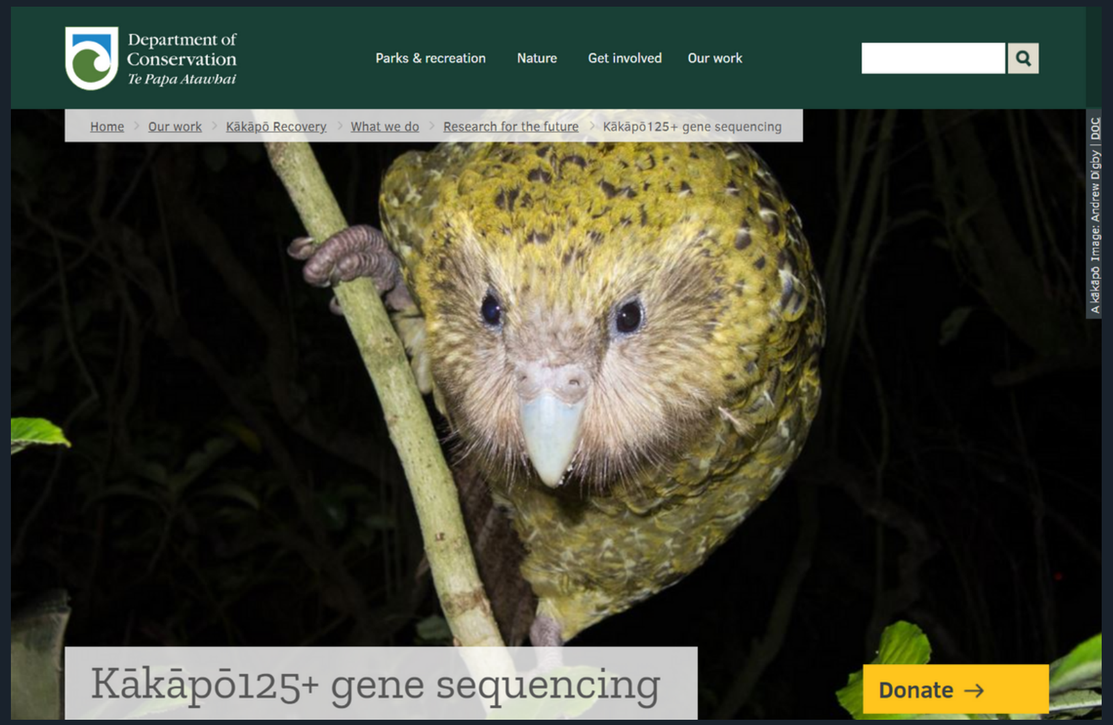


Background



- Kākāpō is a critically endangered species in New Zealand, and it is the world's largest, the only flightless, and the only lek-breeding parrot.
- Whole-genome sequencing has been completed for the entire kākāpō species.
- The DNA sequence data allows the kākāpō recovery team to perform numerous analyses of the kākāpō species providing insights into genetic management, disease, fertility and ageing.

Background



- Kākāpō is a critically endangered species in New Zealand, and it is the world's largest, the only flightless, and the only lek-breeding parrot.
- Whole-genome sequencing has been completed for the entire kākāpō species.
- The DNA sequence data allows the kākāpō recovery team to perform numerous analyses of the kākāpō species providing insights into genetic management, disease, fertility and ageing.
- One of the major goals of the kākāpō conservation project is to find functional genetic variants that are associated with key traits using genome-wide association studies.

Introduction

Genome-wide association study (GWAS)

A GWAS is a process of finding functional genetic variants by testing hundreds of thousands of genetic variants across the genome in different individuals for association with the trait.

The architecture of complex traits

In many GWA studies, the polygenic model is considered to be the founding principle as it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population.

Linear mixed model (LMM)

Under the polygenic model, linear mixed models can be used to measure the genetic effect of a particular variant on a quantitative trait while accounting the other variants as correlations between related individuals.

What was done for kākāpō?

Whole-genome sequencing of the entire kākāpō population is completed (this is not cost-effective in general...)

Introduction

Genome-wide association study (GWAS)

A GWAS is a process of finding functional genetic variants by testing hundreds of thousands of genetic variants across the genome in different individuals for association with the trait.

The architecture of complex traits

In many GWA studies, the polygenic model is considered to be the founding principle as it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population.

Linear mixed model (LMM)

Under the polygenic model, linear mixed models can be used to measure the genetic effect of a particular variant on a quantitative trait while accounting the other variants as correlations between related individuals.

What was done for kākāpō?

Whole-genome sequencing of the entire kākāpō population is completed (this is not cost-effective in general...)

Introduction

Genome-wide association study (GWAS)

A GWAS is a process of finding functional genetic variants by testing hundreds of thousands of genetic variants across the genome in different individuals for association with the trait.

The architecture of complex traits

In many GWA studies, the polygenic model is considered to be the founding principle as it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population.

Linear mixed model (LMM)

Under the polygenic model, linear mixed models can be used to measure the genetic effect of a particular variant on a quantitative trait while accounting the other variants as correlations between related individuals.

What was done for kākāpō?

Whole-genome sequencing of the entire kākāpō population is completed (this is not cost-effective in general...)

Introduction

Genome-wide association study (GWAS)

A GWAS is a process of finding functional genetic variants by testing hundreds of thousands of genetic variants across the genome in different individuals for association with the trait.

The architecture of complex traits

In many GWA studies, the polygenic model is considered to be the founding principle as it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population.

Linear mixed model (LMM)

Under the polygenic model, linear mixed models can be used to measure the genetic effect of a particular variant on a quantitative trait while accounting the other variants as correlations between related individuals.

What was done for kākāpō?

Whole-genome sequencing of the entire kākāpō population is completed (this is not cost-effective in general...)

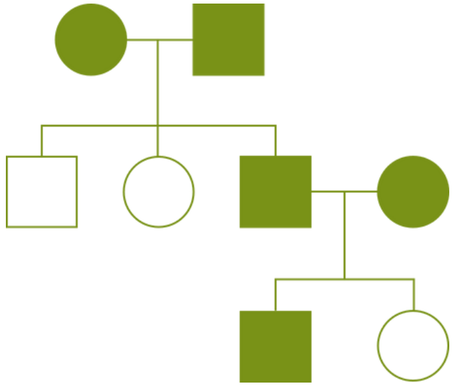
Two-phase sampling design

Two-phase sampling is designed as a cost-saving strategy where the initial sampling of the cohort is followed by a subsampling of the chosen individuals to be resequenced.

Two-phase sampling design

Two-phase sampling is designed as a cost-saving strategy where the initial sampling of the cohort is followed by a subsampling of the chosen individuals to be resequenced.

Ex.



- Low-density genotyping for all individuals in the pedigree.
- Whole-genome sequencing only for individuals coloured in green.

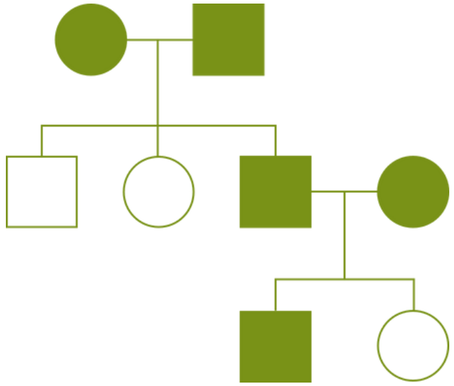
Two-phase sampling design

Two-phase sampling is designed as a cost-saving strategy where the initial sampling of the cohort is followed by a subsampling of the chosen individuals to be resequenced.

Solution:

- Genotype imputation
  - Computationally challenging in situations where the low-resolution genotype has a high error rate.
  - More complicated for endangered species than well-studied species

Ex.



- Low-density genotyping for all individuals in the pedigree.
- Whole-genome sequencing only for individuals coloured in green.

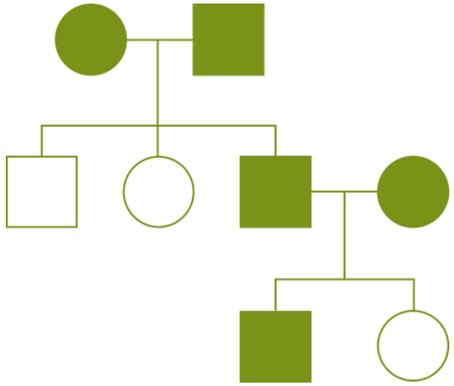
Two-phase sampling design

Two-phase sampling is designed as a cost-saving strategy where the initial sampling of the cohort is followed by a subsampling of the chosen individuals to be resequenced.

Solution:

- Genotype imputation
- Model inference with incomplete data

Ex.



- Low-density genotyping for all individuals in the pedigree.
- Whole-genome sequencing only for individuals coloured in green.

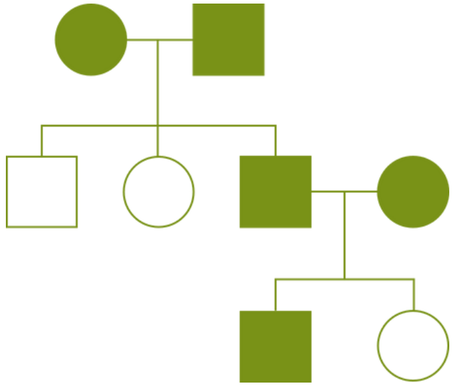
Two-phase sampling design

Two-phase sampling is designed as a cost-saving strategy where the initial sampling of the cohort is followed by a subsampling of the chosen individuals to be resequenced.

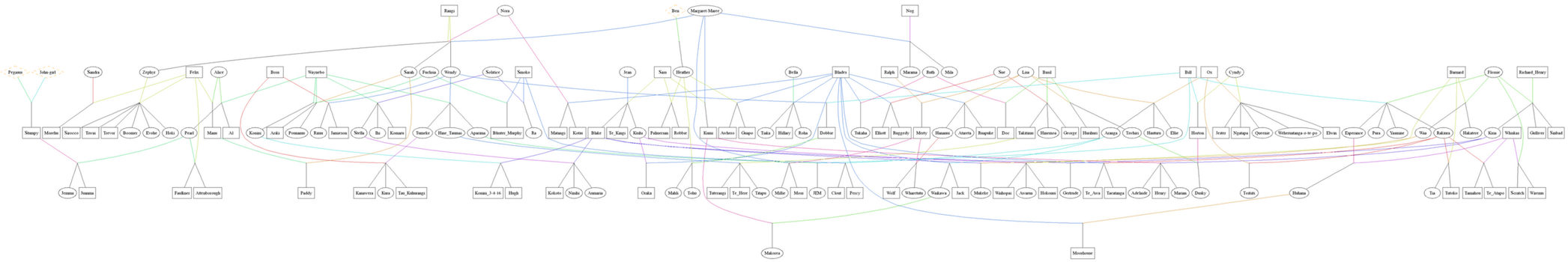
Solution:

- Genotype imputation
- Model inference with incomplete data
  - ***Weighted maximum likelihood estimation approach*** that takes advantage of the fact that the kākāpō population kinship structure is known.

Ex.



- Low-density genotyping for all individuals in the pedigree.
- Whole-genome sequencing only for individuals coloured in green.



The model

$$y = X\beta + \epsilon, \epsilon \sim N(0, \Xi)$$

Parameter

y Quantitative trait values

X Genotypes

The model

Fixed effect

$$y = X\beta + \epsilon, \epsilon \sim N(0, \Xi)$$

Random effect

Parameter

y Quantitative trait values

X Genotypes

The model

Fixed effect

$$y = X\beta + \epsilon, \epsilon \sim N(0, \Xi)$$

Random effect

Genetic random effect

Environmental random effect

Parameter

y Quantitative trait values

X Genotypes

The model

Fixed effect

$$y = X\beta + \epsilon, \epsilon \sim N(0, \Xi)$$

Random effect

Genetic random effect

Environmental random effect

$$\Xi = \sigma^2(h\Phi + (1-h)I)$$

Parameter

- y Quantitative trait values
- X Genotypes
- h Heritability
- Φ Pairwise relatedness

Heritability

The proportion of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population.

Estimated population log-likelihood

$$\ell = -\frac{1}{2}\log|\Xi| - \frac{1}{2}(y - X\beta)^T \Xi^{-1}(y - X\beta)$$

Assumption:

The observations are sampled in a way that they are representative of the whole population.

For non-random sampling:

- Full likelihood
  - *Advantage:* It properly accounts for covariance structure and the missing mechanism.
  - *Disadvantage:* It can be very complicated to construct. For related individuals, we need to deal with the fact that individuals in the pedigree who were not sampled have unobserved genotypes.

Estimated population log-likelihood

$$\ell = -\frac{1}{2}\log|\Xi| - \frac{1}{2}(y - X\beta)^T \Xi^{-1}(y - X\beta)$$

Assumption:

The observations are sampled in a way that they are representative of the whole population.

Sample weighted log-likelihood

$$\ell = -\frac{1}{2}\log|\Xi| - \frac{1}{2}\sum_{i,j}\frac{R_{ij}}{\pi_{ij}}(y_i - X_i\beta)^T(\Xi^{-1})_{ij}(y_j - X_j\beta)$$

Pairwise sampling indicator

Pairwise sampling probability

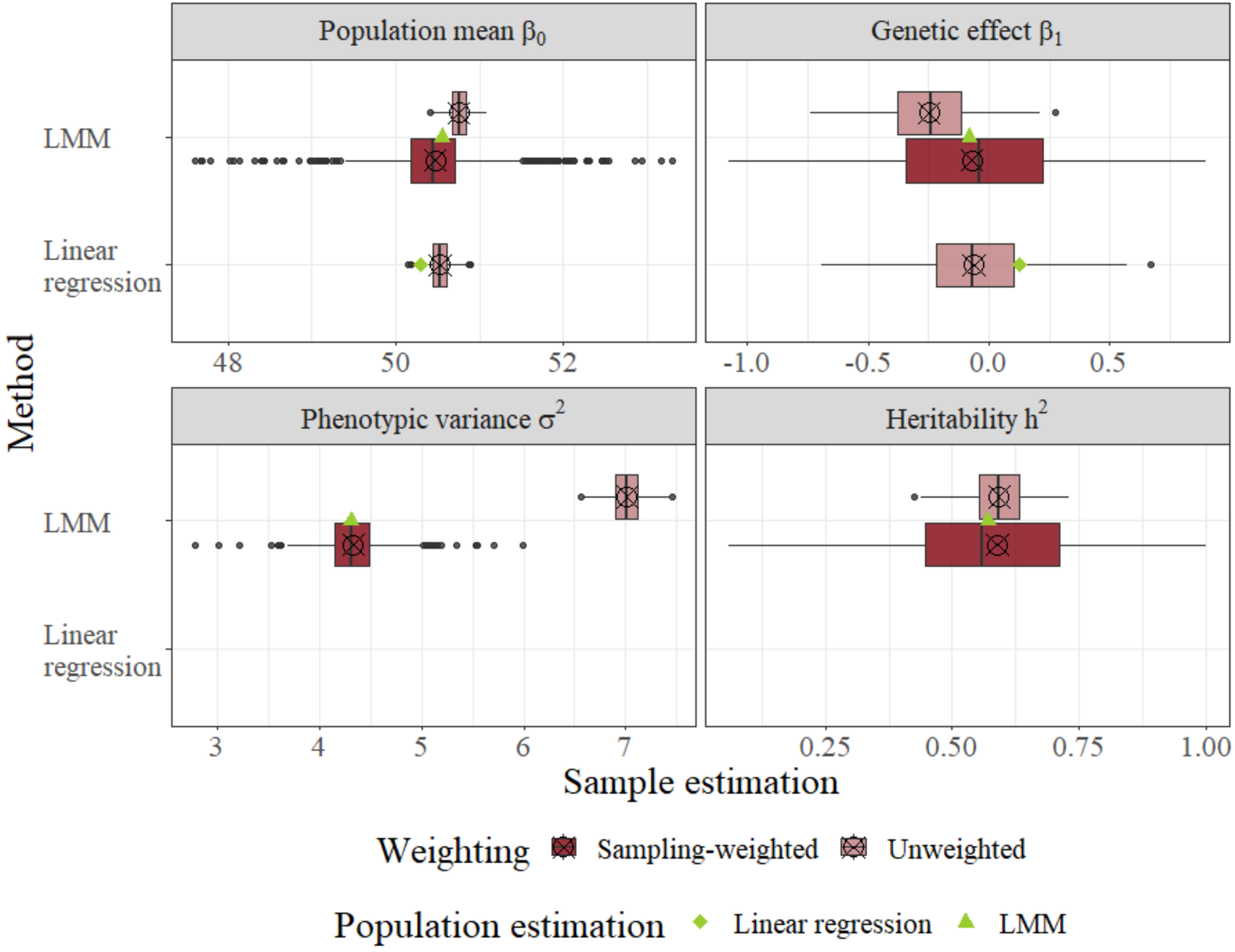
Fitting LMM to kākāpō egg length data under outcome-dependent sampling

The kākāpō egg length data

- Contains 104 kākāpō.
- High level of inbreeding (complicated correlation structure).

Outcome-dependent sampling

1. Always sample the individuals from the two 15% tails of the phenotype distribution;
2. Random sampling from the rest of the individuals.



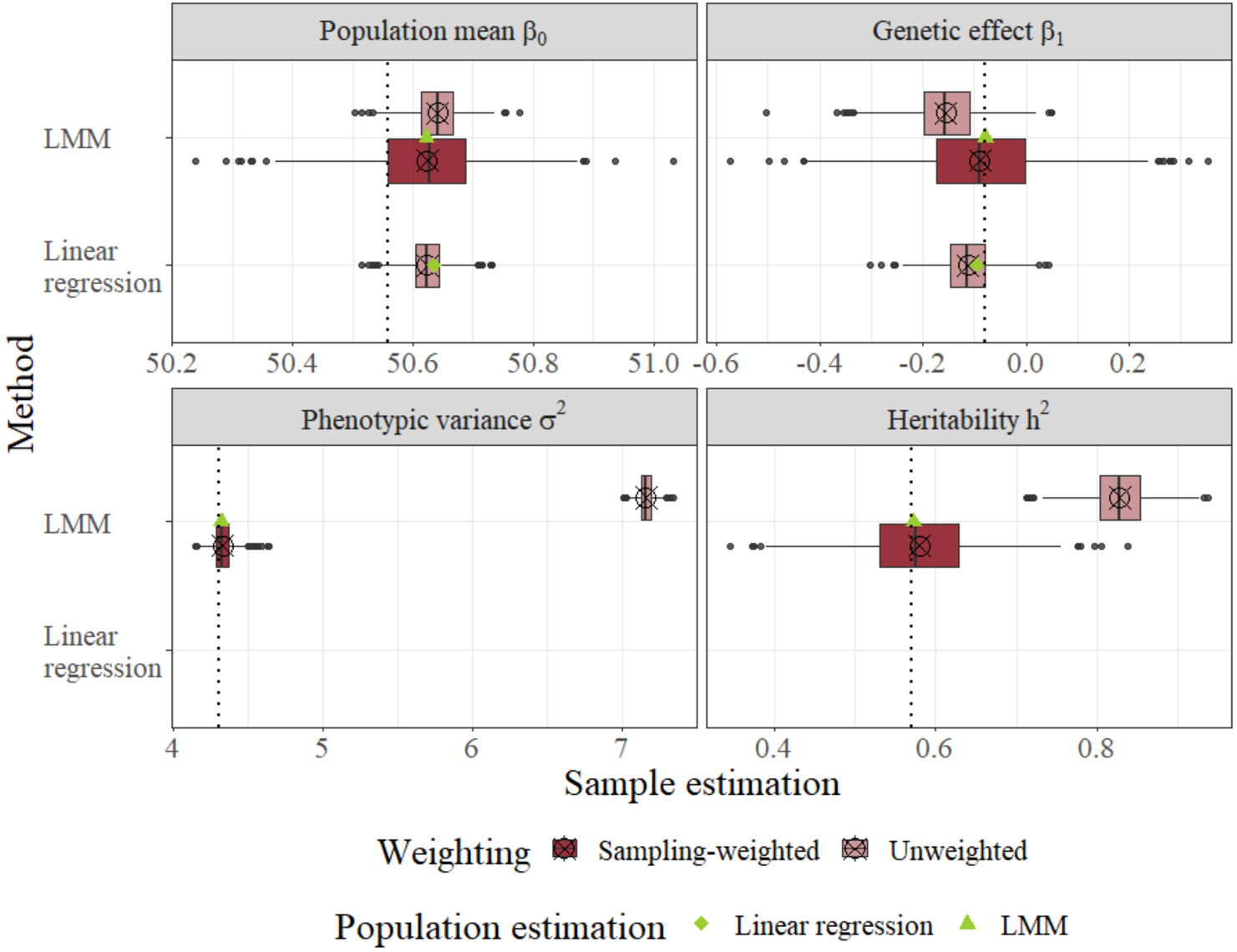
Fitting LMM to simulated nuclear family data under outcome-dependent sampling

The simulated nuclear family data

- Contains 1200 individuals from 300 independent nuclear families, each has two unrelated parents and two children.

Outcome-dependent sampling

1. Always sample the individuals from the two 15% tails of the phenotype distribution;
2. Random sampling from the rest of the individuals.



Consistency of the sample weighted likelihood estimator

Proof:

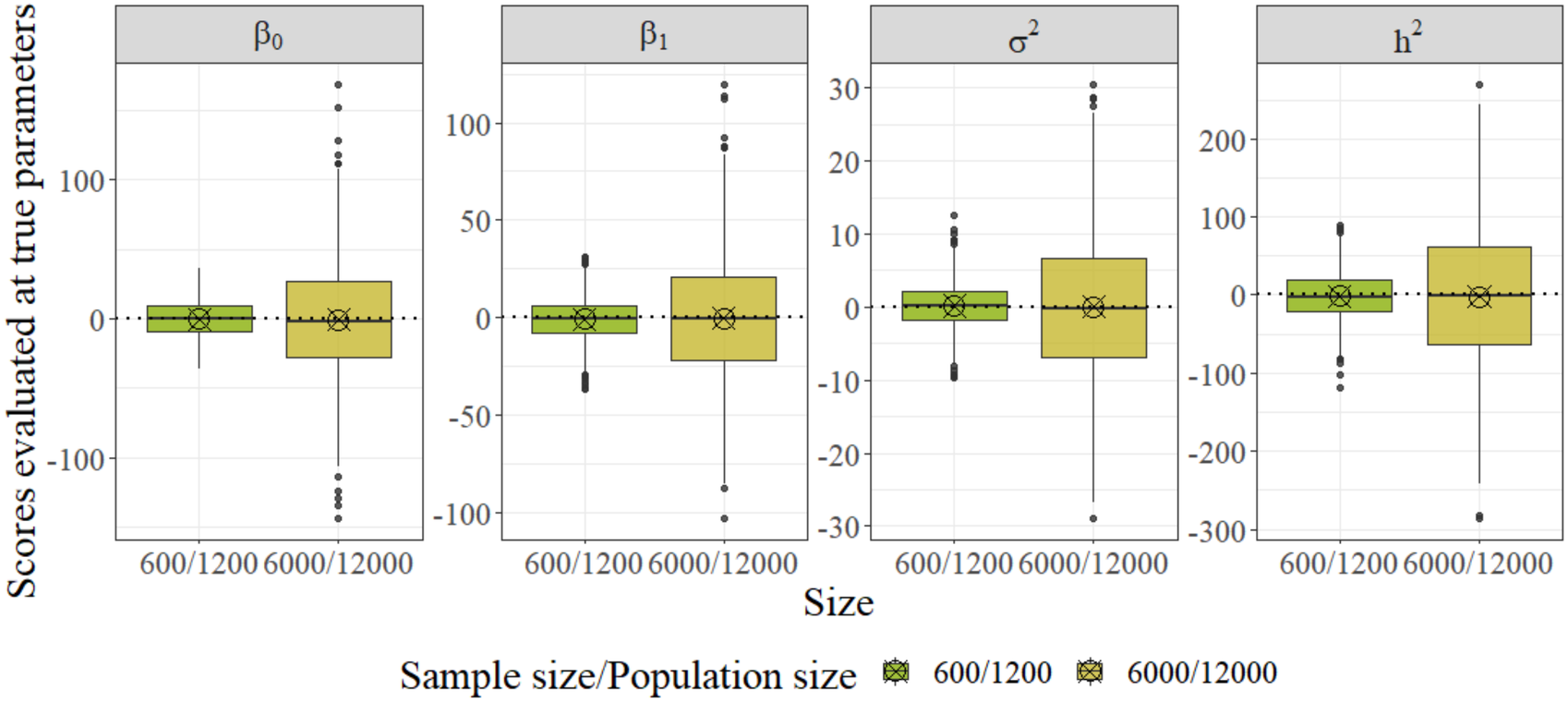
Since $\mathbb{E}_\pi[R_{ij}|Y,\Phi] = \pi_{ij}$, the rest follows from the consistency proof of the population likelihood estimator.

Consistency of the sample weighted likelihood estimator

Proof:

Since $\mathbb{E}_{\pi}[R_{ij}|Y,\Phi] = \pi_{ij}$, the rest follows from the consistency proof of the population likelihood estimator.

For the outcome-dependent sampling design:



Summary

- We proposed a weighted maximum likelihood approach for fitting linear mixed models under two-phase designs, that takes advantage of knowing the population kinship structure, and it is easy to implement.

Summary

- We proposed a weighted maximum likelihood approach for fitting linear mixed models under two-phase designs, that takes advantage of knowing the population kinship structure, and it is easy to implement.

- The proposed method corrects the sampling bias by re-weighting the samples regardless of kinship structure, and provides a consistent sample weighted likelihood estimator.

Summary

- We proposed a weighted maximum likelihood approach for fitting linear mixed models under two-phase designs, that takes advantage of knowing the population kinship structure, and it is easy to implement.
- The proposed method corrects the sampling bias by re-weighting the samples regardless of kinship structure, and provides a consistent sample weighted likelihood estimator.

What's more

- A confidence interval can be obtained using parametric bootstrap method.
- The idea can be extended to generalized linear mixed models under family-based sampling designs.

Thanks for listening!

Email: pluo244@aucklanduni.ac.nz

The **R package WLMM** is available on GitHub:
https://github.com/zoeluo15/WLMM

SOURCE: SARAH MAYBE LITTLE (@SARAHMAYLITTLE)