Development of a Population Simulator to Optimise Study Designs and Estimate Polygenic Disease Risk/Resilience in Aotearoa New Zealand Māori Populations

Dr Alastair Lamont
Assoc. Prof. Philip Wilcox, Prof. Mik Black

University of Otago, New Zealand

November 24, 2023

Background

As the cost of genotyping decreases, genotype data can be more widely used to improve health outcomes.

- Identifying genetic sites of interest
- Estimating genetic scores

GWAS

Genome-wide association studies (GWAS) are commonly used for this.

- Collect genome-wide on a large group of individuals (10^4 - 10^7)
- Determine which single nucleotide polymorphisms (SNPs) have a significant association with a given trait

Issues with GWAS

Existing datasets appropriate for GWAS often do not include indigenous peoples.

Issues with GWAS

Existing datasets appropriate for GWAS often do not include indigenous peoples.
Indigenous populations such as Māori may have unique genetic architectures, but obtaining sufficiently large datasets is challenging in Aotearoa:

- Substantive cost to generate genotype data
- Reluctance of many Māori to participate due to previous poor experiences

Issues with GWAS

Existing datasets appropriate for GWAS often do not include indigenous peoples.
Indigenous populations such as Māori may have unique genetic architectures, but obtaining sufficiently large datasets is challenging in Aotearoa:

- Substantive cost to generate genotype data
- Reluctance of many Māori to participate due to previous poor experiences

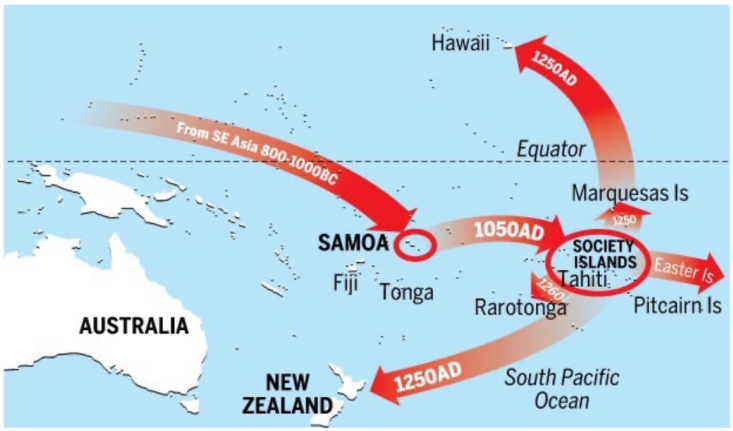
Alternative methods which do not require large genotyped datasets are more practical.

Māori Background

(from Phil Wilcox)

1. Māori – background and context

- **First colonised NZ/Aotearoa ca. 800 years ago**
- Austronesian → **Polynesian language & culture**

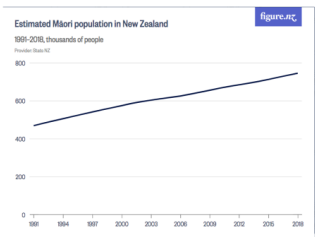


Māori Background

(from Phil Wilcox)

Māori today – general background

- **Population** size: 892 200 in 2022 – see https://www.stats.govt.nz/information-releases/maori-population-estimates-at-30-june-2022/
- **> 140 000 live in Australia(!!), 151 000 overseas (2021)**
- **1/7** NZers of **Maori** descent
- **1/3 under 15** yrs of age
- **Highly urbanised** – approximately 80% live in urban areas
- **Tribal hierarchy:**
  - Approx. 100 iwi (000's – 000 000's members)
  - Several thousand hapu
  - Approx. 20% of Maori do not know tribal ancestry



Home / Kahu

18 per cent of Maori now live overseas

By Simon Collins
29 Nov, 2011 05:30 AM · 3 mins to read

Save Share

Whakapapa

(from Phil Wilcox)

ii. Whakapapa

Whakapapa Maori

Structure, Terminology and Usage

- Key concept in Te Ao Māori
- Both a genetic and social construct
- Literally ‘to place in layers…’ (see https://maaori.com/whakapapa/whakpap2.htm)
- Whakapapa refers to both genealogies AND spiritual, mythological and human ‘purākau’ – stories/histories = ‘flesh’ to the genealogical ‘bones’
- Tātai/Kawai = genealogical component
- Whakapapa widely known (80% of Māori know at least some of their whakapapa)

"Before the coming of the Pakeha [European] to New Zealand with his superior technology, all literature in Maori was oral. Its transmission to succeeding generations was also oral and a great body of literature, which includes haka [dance], waiata [song], tauparapara [chant], karanga [chant], poroporoaki [farewell], paki waitara [stories], whakapapa [genealogy], whakatauki [proverbs] and pepeha [tribal sayings], was retained and learnt by each new generation."

- Sir Timoti Karetu, "Language and Protocol of the Marae [meeting place], in Te Ao Hurihuri, ed Michael King, 1975, Longman Paul, Auckland

UNIVERSITY OF OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

Incorporating Non-Genomic Information

Māori have well-recorded whakapapa (genealogy), that describes the descent of and relationships between all things.
Records of descent go back 35-40 generations, back to Eastern Polynesia.

Incorporating Non-Genomic Information

Māori have well-recorded whakapapa (genealogy), that describes the descent of and relationships between all things.
Records of descent go back 35-40 generations, back to Eastern Polynesia.
Whakapapa can be used to supplement partial genotype information.
Approaches such as ssGBLUP (Aguilar et al., 2010) combine partial genotype and complete relationship information to estimate the correlation in genetic scores.

Study Design

To effectively predict genetic score with these approaches, appropriate study design is needed.

- How many individuals?
- How many of them should be genotyped?
- Which individuals should be genotyped?

Simulation

We are developing population simulations to model genetic structures of Māori communities.

- SLiM 3 simulation software (Haller and Messer, 2019)
- Incorporate estimates of historical details
- Incorporate admixture and effects of European colonisation

Simulation

We are developing population simulations to model genetic structures of Māori communities.

- SLiM 3 simulation software (Haller and Messer, 2019)
- Incorporate estimates of historical details
- Incorporate admixture and effects of European colonisation

These simulations can be used to explore features of study design and analytical methods that lead to optimal prediction.

1000 Genomes Data

Rather than simulating from scratch, we are using existing genotype data to represent ancestral West Polynesian individuals. This data is from the 1000 Genomes Project, publicly available at `https://www.internationalgenome.org/home`

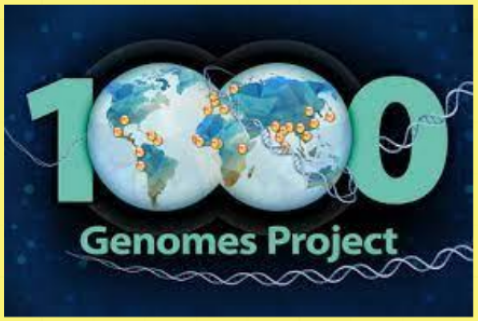
1000 Genomes Data

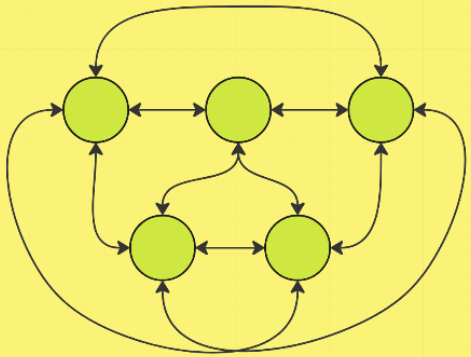
Current simulation details:

- In testing/development phase
- 2 chromosomes (19 and 22)
  - increase recombination rate tenfold, otherwise would have few recombinations
- Two populations
  - 103 Chinese Han ('West Polynesian')
  - 91 British ('European')

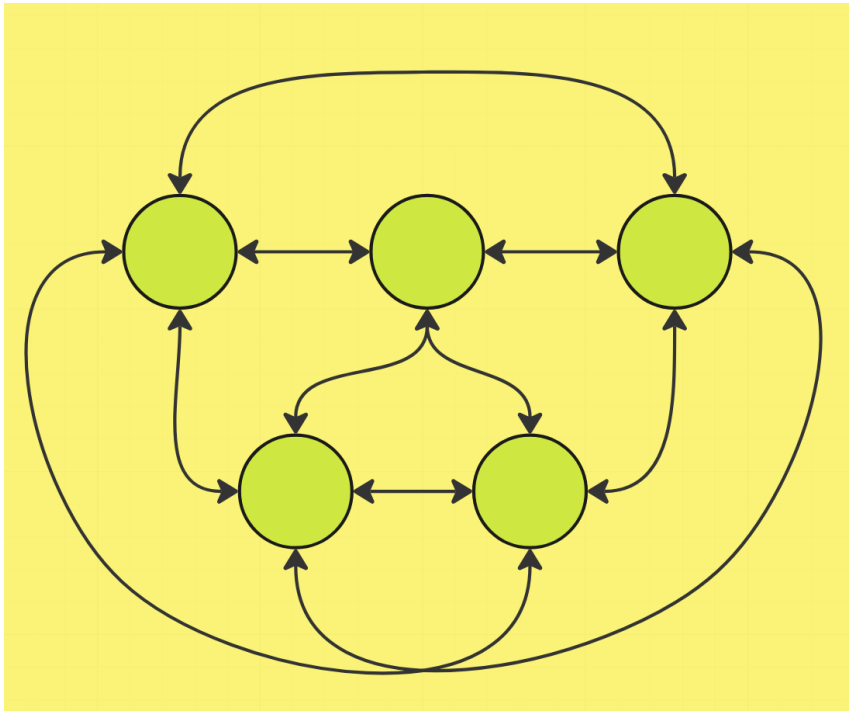
Workflow Diagram

These individuals are the progenitors of an Eastern Polynesian simulation from which individuals will emigrate to Aotearoa.





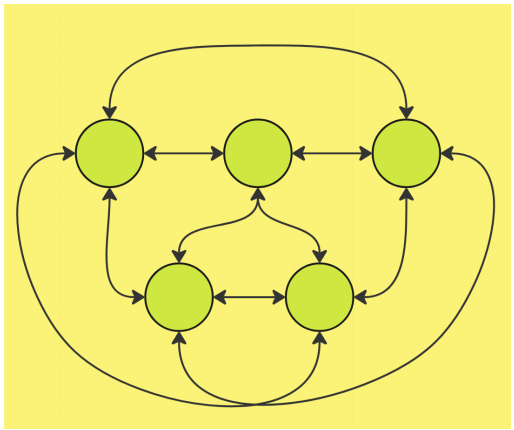
East Polynesian Simulation



Individuals inhabit 5 different islands.

- Each has a stable population size of 600.
- Marriages can only happen between individuals on the same island.
- Each tick 1% of individuals migrate to a different island.

East Polynesian Simulation

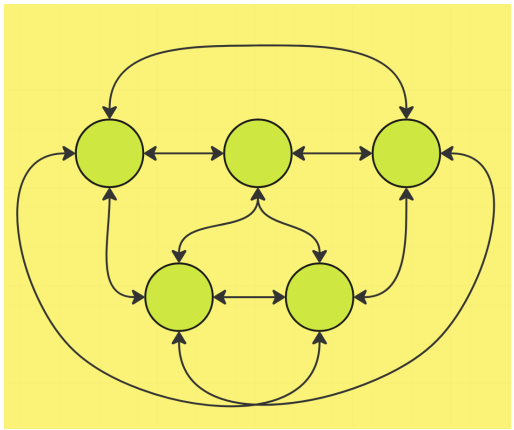


Initial population details:

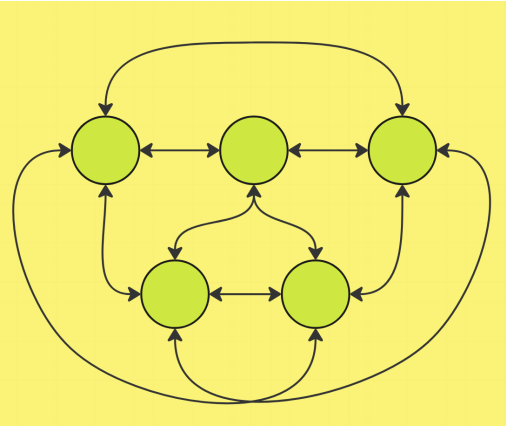
- 103 individuals.
- 120 ticks of 25 years each: 3000 years total.
- Simulation has overlapping generations.
- Reducing tick interval (and increasing the number of ticks) will give more nuance.
  - Increases computational burden.
  - Is it better to reduce tick length, or have more individuals / chromosomes?

East Polynesian Simulation

Individuals can have 4 possible ages: 0, 25, 50, 75.



East Polynesian Simulation

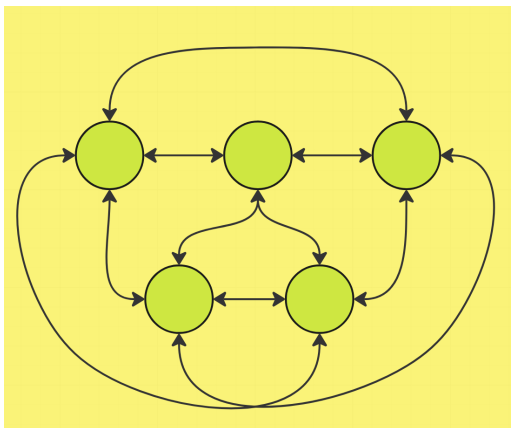


Individuals can have 4 possible ages: 0, 25, 50, 75.

The probability of survival until the next tick depends on age.

Age in Years (ticks)	Survival Probability
0 (0)	0.8
25 (1)	0.9
50 (2)	0.5
75 (3)	0

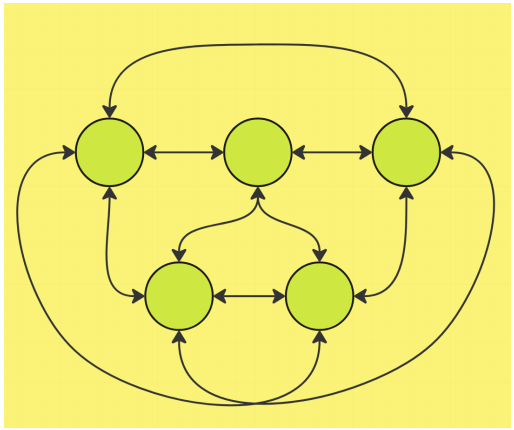
East Polynesian Simulation



Marriage details:

- Only possible for individuals of age 25 or 50.
- No selection - all eligible partners have an equal probability.
- Number of children has a Poisson distribution with $\lambda = 2$.

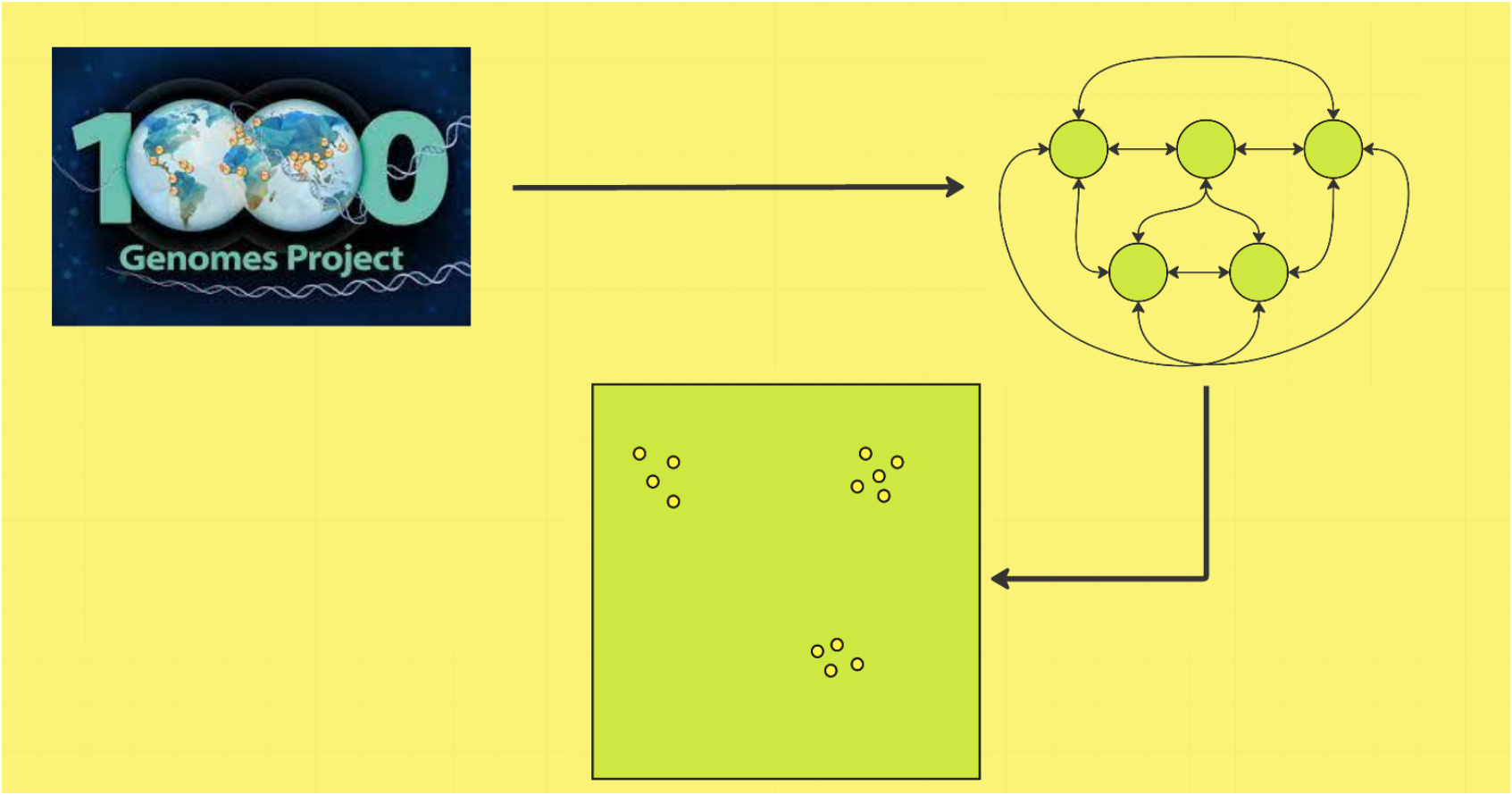
East Polynesian Simulation



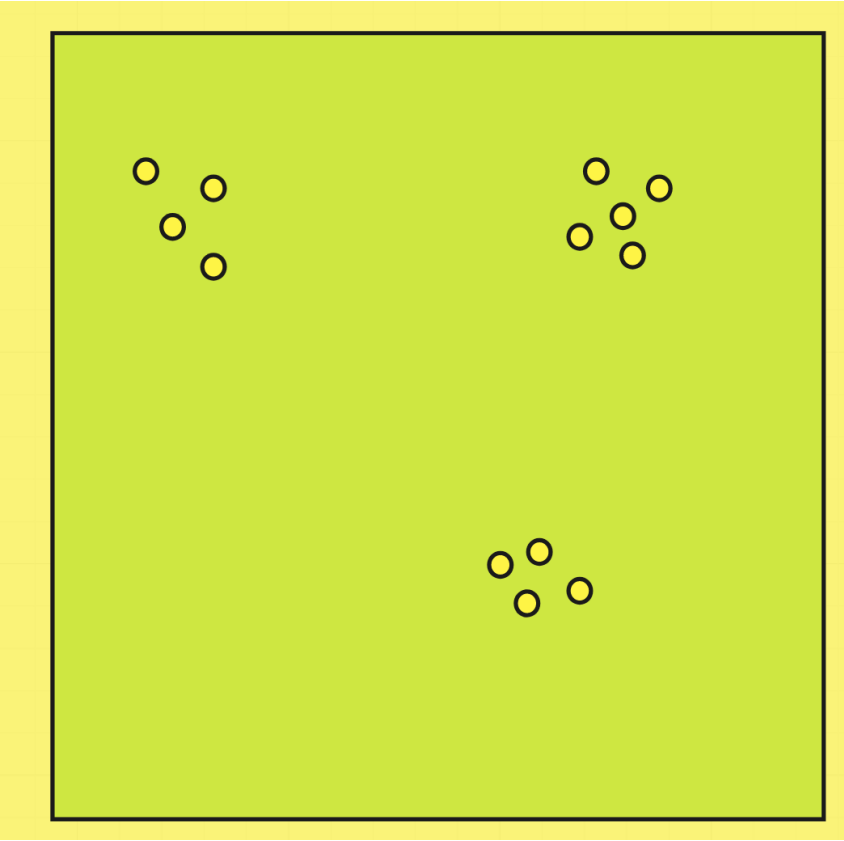
Emigrants to Aotearoa are chosen from East Polynesian individuals who are alive after the final tick.

- A single individual is sampled at random.
- Further individuals from that island are sampled based on relationships.
- Probability of inclusion increases with relatedness to the initial individual.
- This is repeated until 100 related individuals are sampled (Murray-McIntosh et al., 1998).

Workflow Diagram



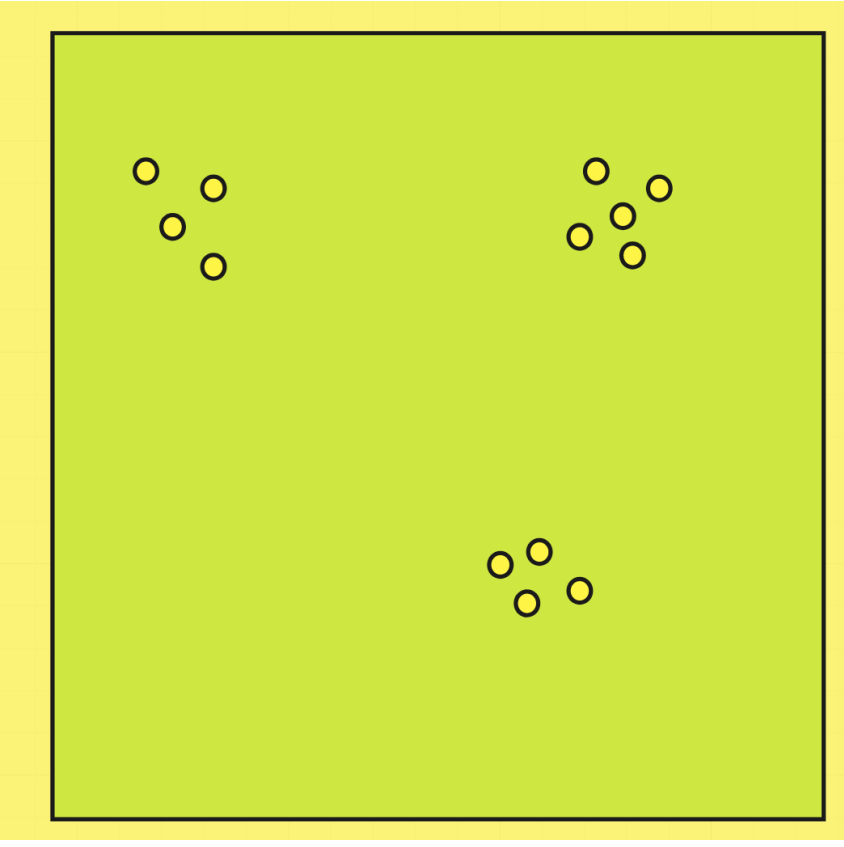
Aotearoa Simulation



Aotearoa is represented as a 2d space which individuals inhabit.

- Marriage probability decreases with distance.
- Distance can be considered either physical or abstract.

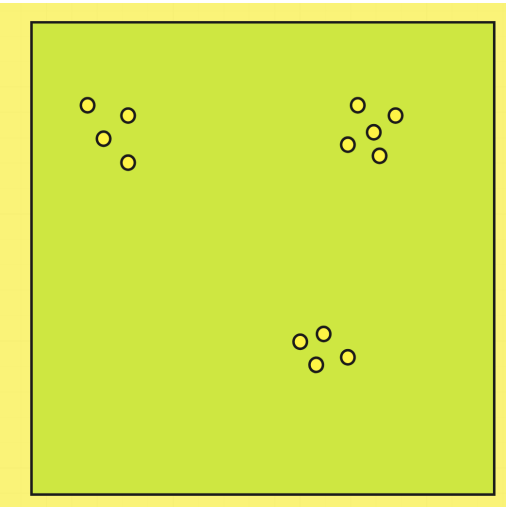
Aotearoa Simulation



Aotearoa is represented as a 2d space which individuals inhabit.

- Marriage probability decreases with distance.
- Distance can be considered either physical or abstract.
- Spatial simulation generates clusters of individuals that change over time.
  - These represent abstract iwi and hapu.

Aotearoa Simulation



Initial population details:

- Inhabitation of Aotearoa is estimated at around 1250CE (Bunbury et al., 2022).
- The stable population size is set to 1000.
- Tick length and survival probabilities are unchanged.

Aotearoa Simulation

European arrival occurs from the 1825 tick onwards.

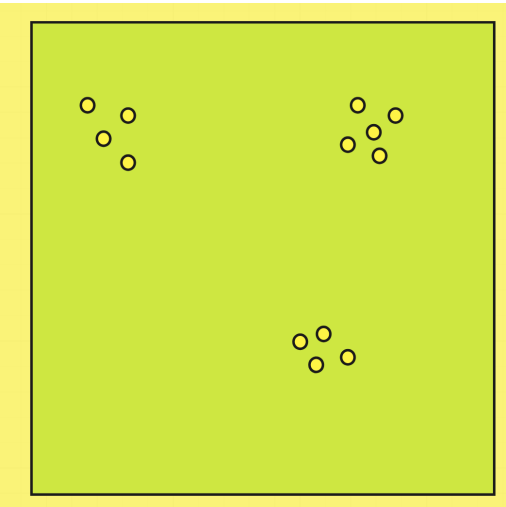
- Māori population size has sharp decline, bottlenecks at 1825 and 1900 due to musket wars and disease.
- The population then increases to present day.

Aotearoa Simulation

European arrival occurs from the 1825 tick onwards.

- Māori population size has sharp decline, bottlenecks at 1825 and 1900 due to musket wars and disease.
- The population then increases to present day.
- Europeans (1000 Genomes Project 'British') are modelled as a separate subpopulation.
- Europeans can then migrate into Aotearoa subpopulation.
  - This allows for constant gene flow without limiting computational performance.

Aotearoa Simulation



The surviving Aoteraroa subpopulation at 2025 (tick 32) is recorded as simulation output. Complete whakapapa is also tracked.

Trait Simulation

Turning simulated genotypes into phenotypes needs several more details.

- Marker effect distribution: iid Normal $(0, 10^{-8})$.
- No environmental effects.
- Residuals are iid Normal $(0, \sigma^2)$ with σ^2 chosen to give $h^2 = 0.5$.
- Parental phenotypes are observed.
- Child phenotypes are unobserved.

The entire simulation process (starting from the initial West Polynesia genotypes) was repeated 40 times.

Model and Study Design Testing

There are many potential study designs that can be tested.

- Genotype parents vs genotype children?
- Genotype phenotyped individuals?
- Genotype individuals with family history of trait?
- etc.

Model and Study Design Testing

There are many potential study designs that can be tested.

- Genotype parents vs genotype children?
- Genotype phenotyped individuals?
- Genotype individuals with family history of trait?
- etc.

For now, we have tested a straightforward question: what is the value of whakapapa?

- How does whakapapa, or whakapapa + partial genomic perform compared to full genomic?

Models

We have fitted three models:

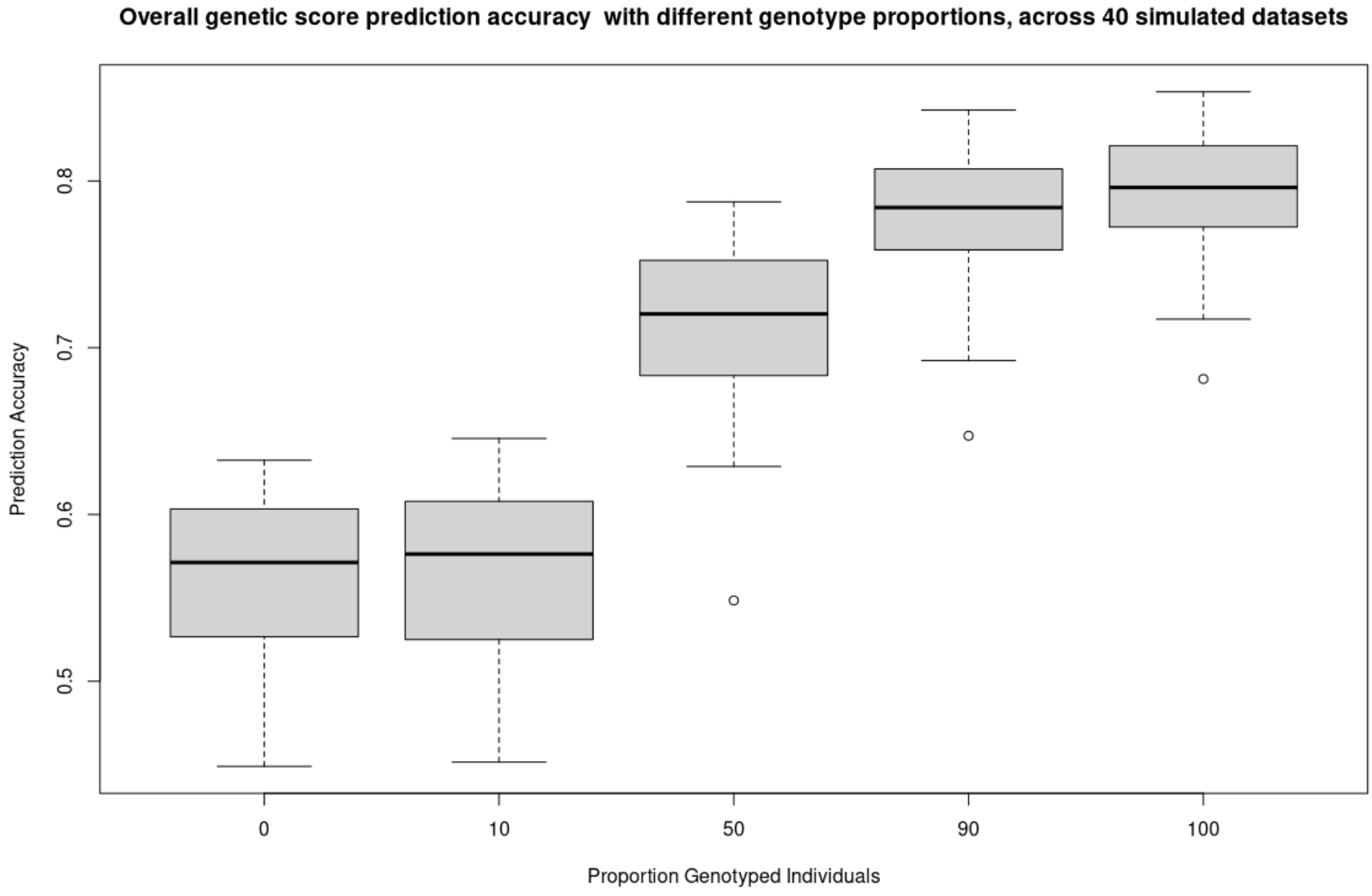
- BLUP, Henderson (1973): whakapapa only.
- GBLUP, VanRaden (2008): full genomic.
- ssGBLUP, Aguilar et al. (2010): whakapapa and partial genomic.

Three different levels of random genotyping were tested with ssGBLUP: 10%, 50% and 90% of the sample genotyped.

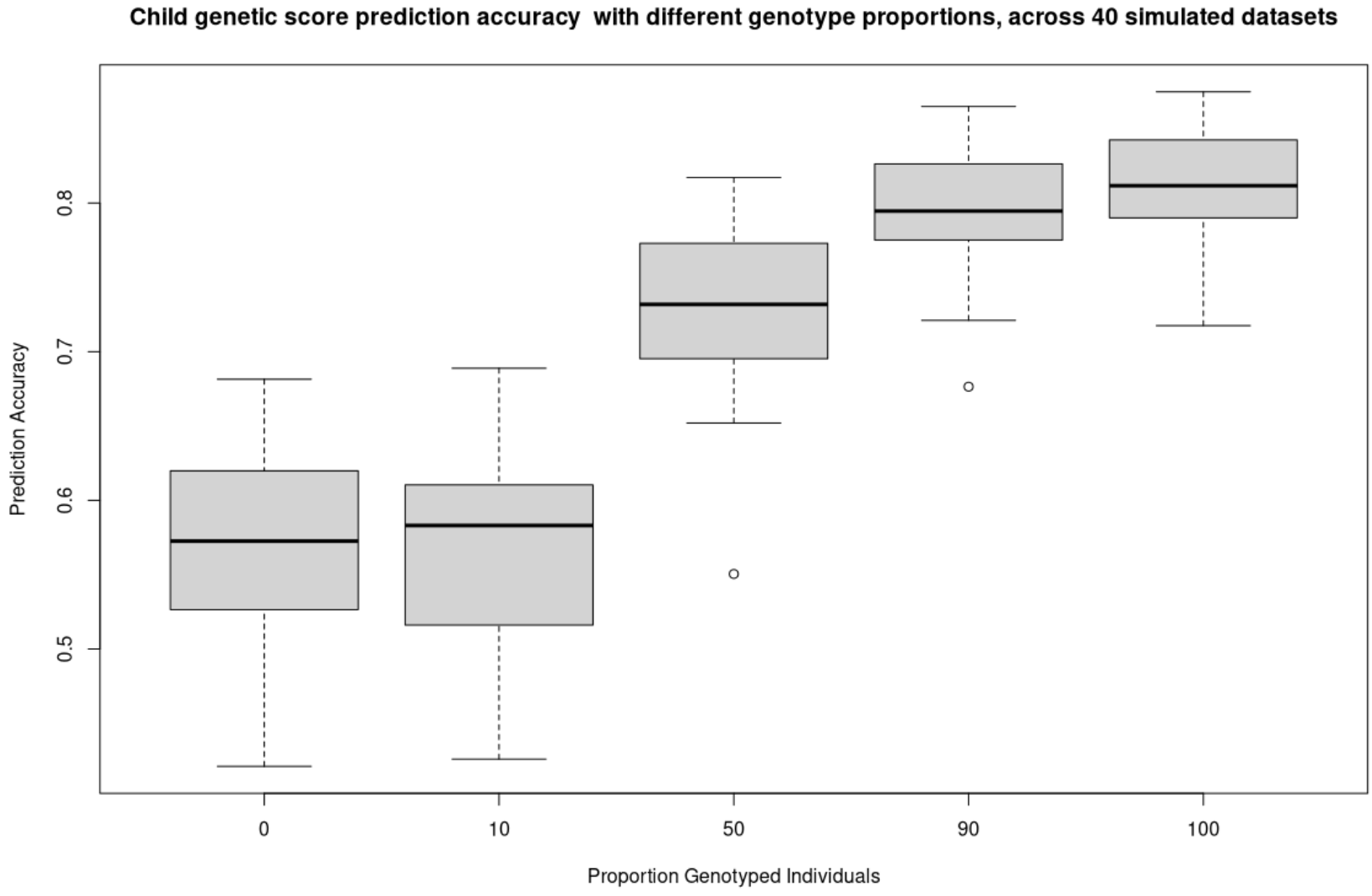
Analyses

Genetic score r^2 were calculated for each dataset, both for the entire sample and for the unphenotyped children only.

Overall Prediction Accuracy



Child Prediction Accuracy



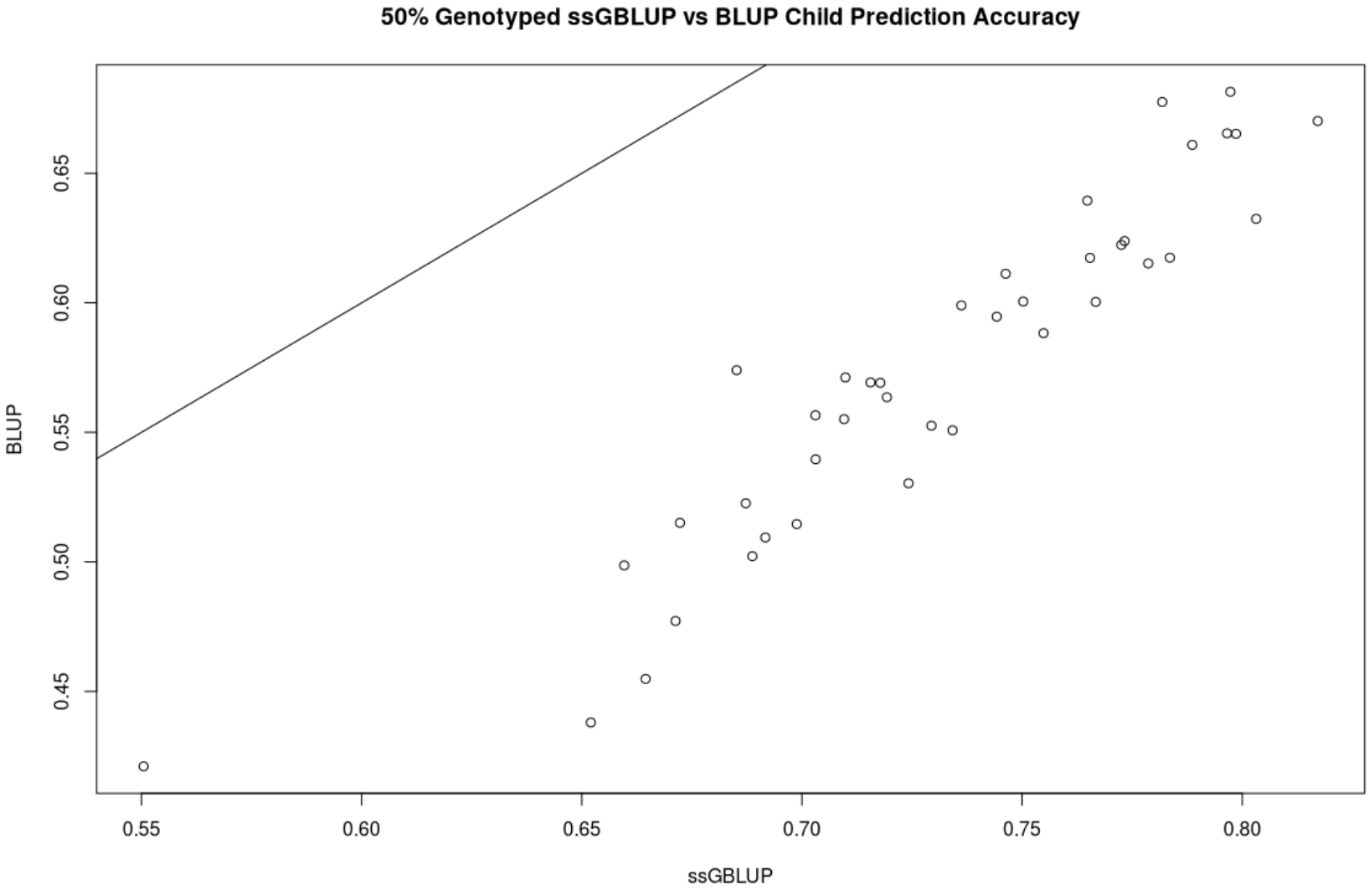
Paired Data

Of course, when comparing two approaches we need to remember that our data is paired.

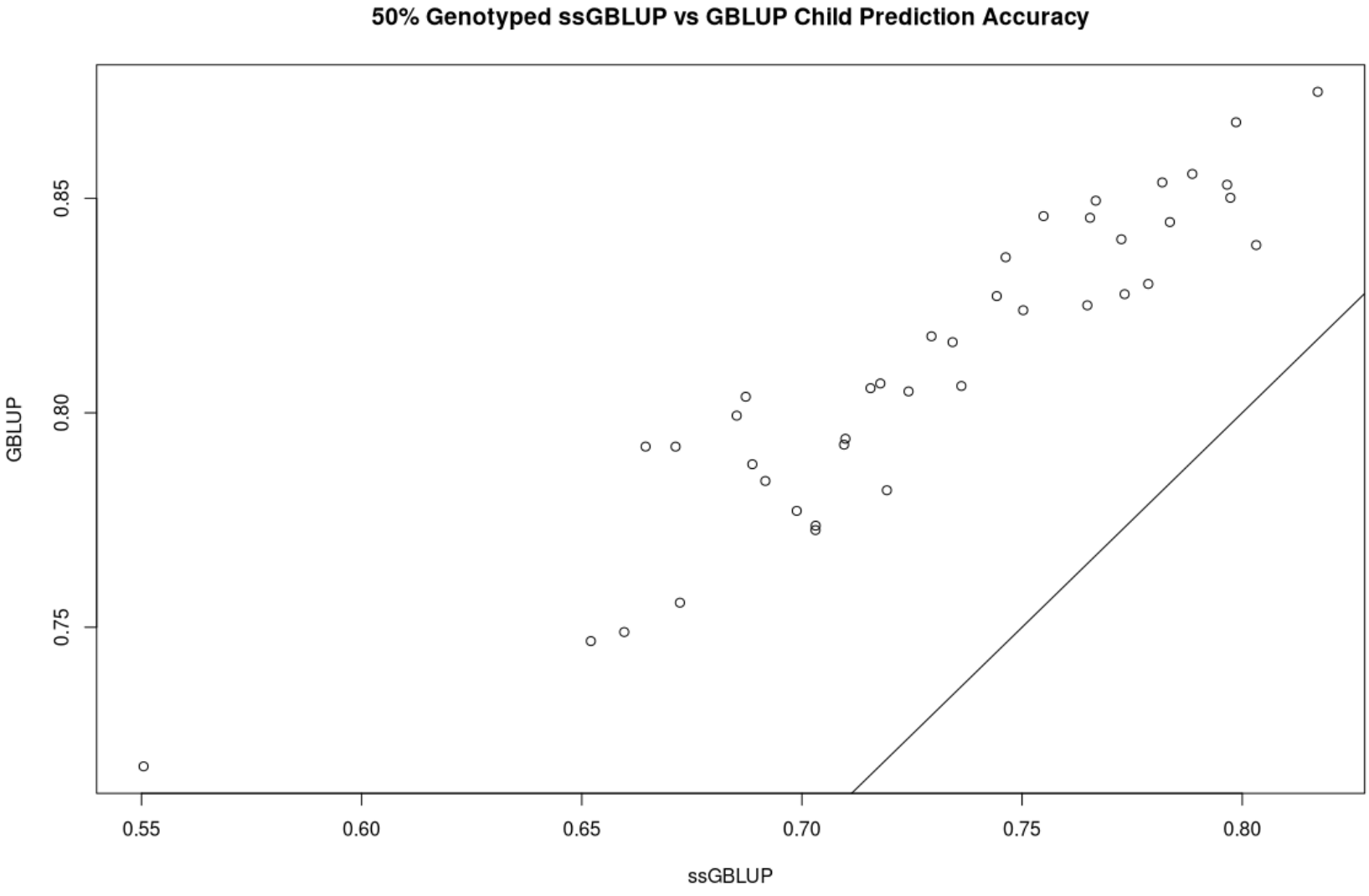
- The same 40 datasets have been fitted to each approach, not independent datasets.

This gives us a better picture of whether one approach is consistently better than another:

Pairwise Comparison of ssGBLUP and BLUP



Pairwise Comparison of ssGBLUP and GBLUP



Results Summary

If we look at the ranges over 40 datasets there is a clear trend, not only for r^2 but also for RMSE.

Model (% Genotyped)	r^2 (mean)	RMSE (mean)
BLUP (0)	0.449-0.633 (0.563)	11.53-15.03 (13.48)
ssGBLUP (10)	0.451-0.646 (0.564)	11.21-14.82 (13.40)
ssGBLUP (50)	0.548-0.787 (0.714)	8.57-12.38 (10.03)
ssGBLUP (90)	0.647-0.842 (0.777)	7.51-10.42 (8.68)
GBLUP (100)	0.681-0.853 (0.791)	7.17-9.76 (8.37)

Conclusions

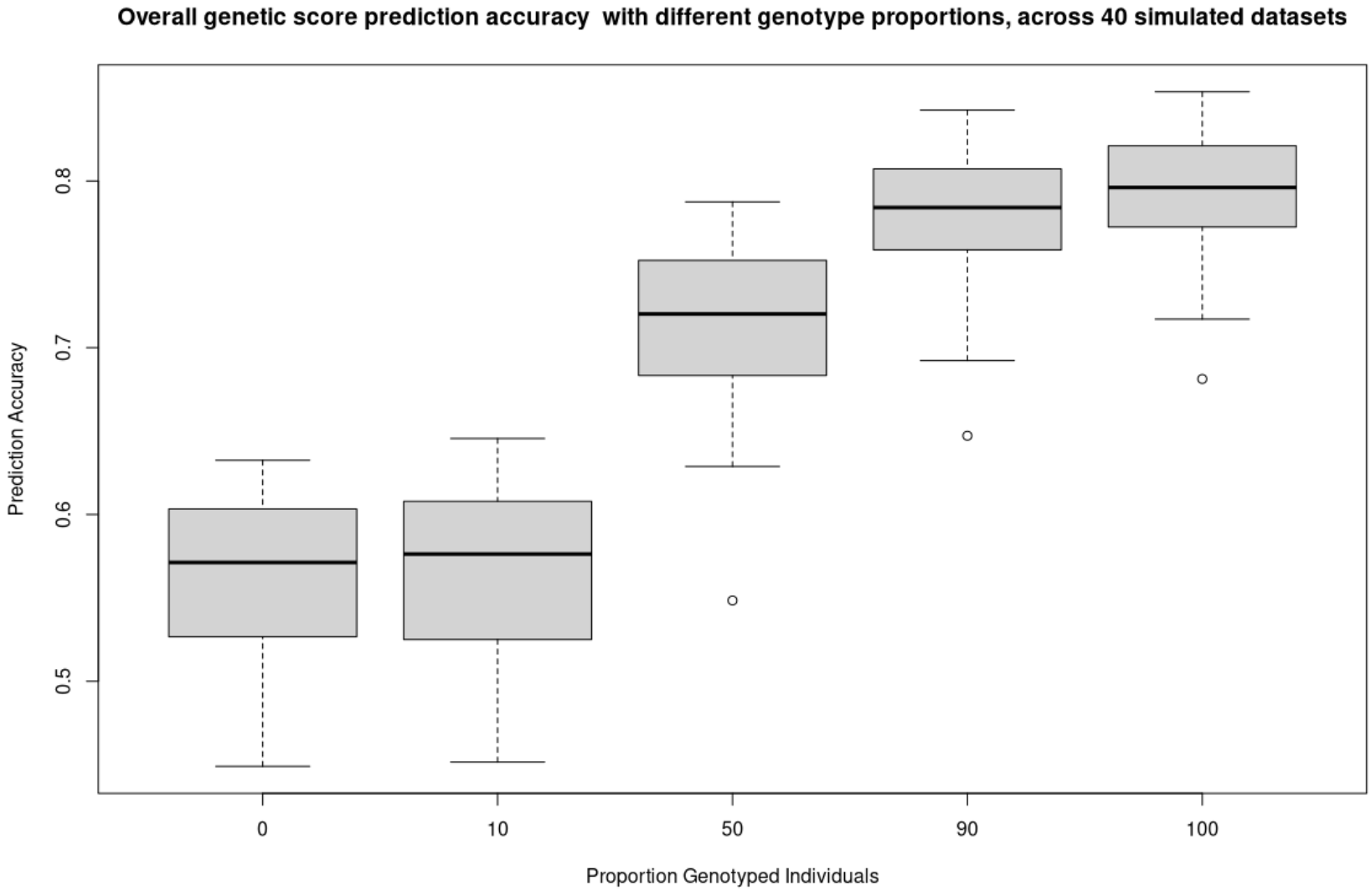
What value does whakapapa alone have?

Conclusions

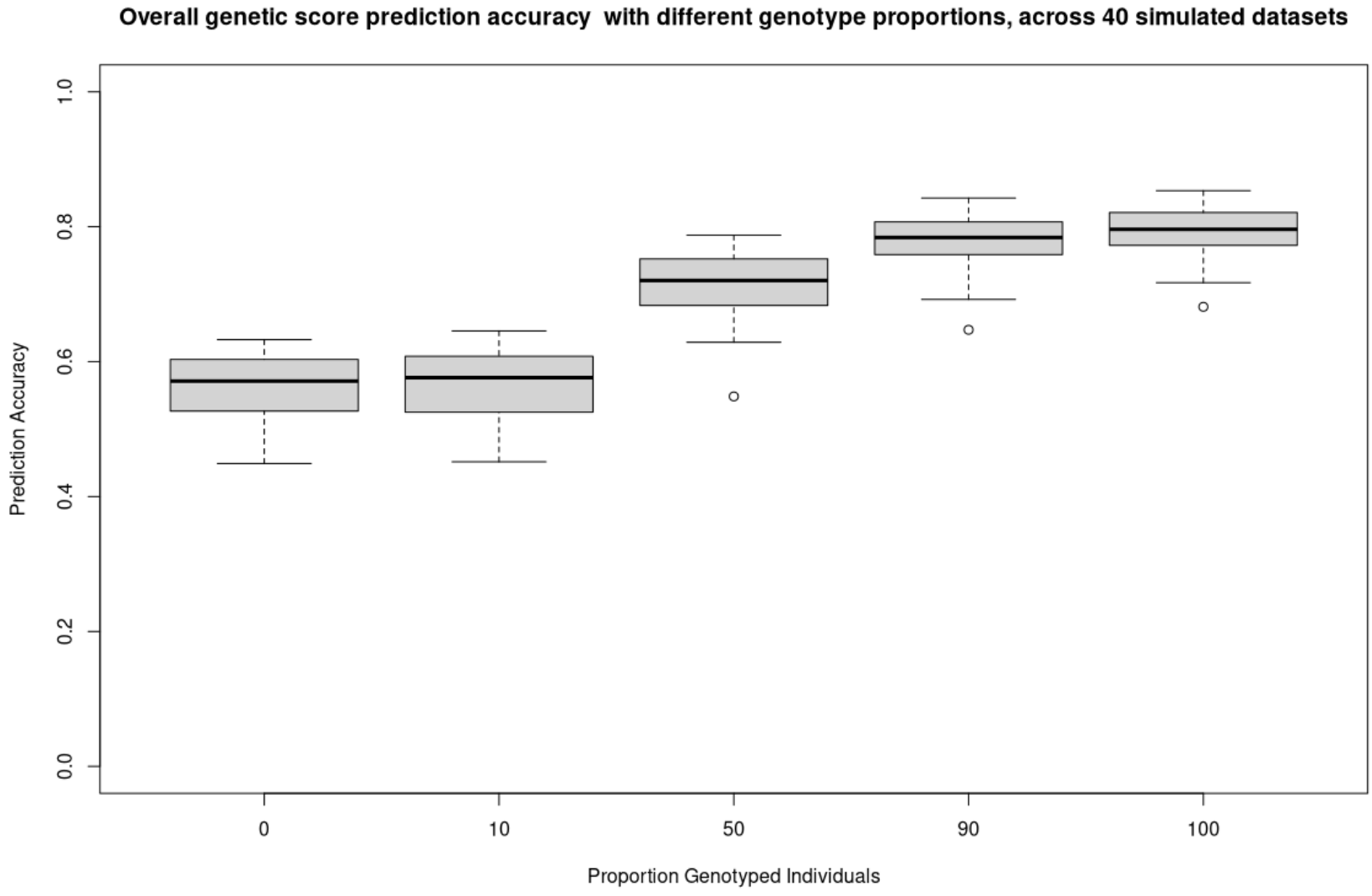
What value does whakapapa alone have?

What value does DNA add?

Overall Prediction Accuracy



Overall Prediction Accuracy



Conclusions

What value does whakapapa alone have?

What value does DNA add?

- Genotyping has a cost!
- Genotyping is also affected by diminishing returns.
- With whakapapa info, full genotyping is likely not optimal.

Conclusions

In which case:

- How many individuals should be genotyped?
- Which individuals should be genotyped to get the most benefit?

Next Steps

We will be testing and investigating this for a range of details:

- heritabilities
- trait expression options
- sample sizes
- historical aspects
- etc.

Next Steps

These results are interim only, with many features are also to be added or further developed:

- Binary traits and disease risk estimation
- More in depth spatial interactions
- Environmental (trait-specific??) variables
- Scaling up to full genome

Acknowledgments

Assoc. Prof. Philip Wilcox,
Prof. Mik Black

MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

Funded by MBIE via 'Beyond Prediction: explanatory and transparent data science'

1000 Genomes Project

UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

University of Auckland

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

University of Otago

References I

Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., and Lawlor, T. (2010), "Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score," *Journal of Dairy Science*, 93, 743–752.

Bunbury, M. M., Petchey, F., and Bickler, S. H. (2022), "A new chronology for the Māori settlement of Aotearoa (NZ) and the potential role of climate change in demographic developments," *Proceedings of the National Academy of Sciences*, 119, e2207609119.

Consortium, . G. P. et al. (2015), "A global reference for human genetic variation," *Nature*, 526, 68.

Haller, B. C. and Messer, P. W. (2019), "SLiM 3: forward genetic simulations beyond the Wright–Fisher model," *Molecular biology and evolution*, 36, 632–637.

References II

Henderson, C. (1973), "Sire evaluation and genetic trends," *Journal of Animal Science*, 1973, 10–41.

Murray-McIntosh, R. P., Scrimshaw, B. J., Hatfield, P. J., and Penny, D. (1998), "Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences," *Proceedings of the National Academy of Sciences*, 95, 9047–9052.

VanRaden, P. (2008), "Efficient methods to compute genomic predictions," *Journal of dairy science*, 91, 4414–4423.