Genomic Aotearoa: Imputation Workshop 2020

Michael Lee

11/09/2020

Background to this workshop: Introductions

MBIE-funded Genomic Aotearoa Project: Primary Industries (Imputation to WGS and GWAS)

Andrew Hess (AgR) & Yu Wang (MU & LIC) Rudiger Brauning (AgR), Shannon Clarke (AgR), Christine Couldrey (LIC), Dorian Garrick (MU), Neville Jopson (AB), Michael Lee (UoO)

- use data from dairy industry (Livestock Improvement Corp.), sheep industry (Beef+Lamb NZ Genetics)
- impute hundreds of thousands of animals to WGS and perform GWAS studies
- Integrate outcomes into genomic predictions - increase genetic gain
- value proposition for GWAS outcomes

What this talk will cover.

- A brief history on genotyping
- Cost of genotyping
- Sequencing technologies
- Fit for purpose - balancing costs and information content
- Value proposition - not all species are created equal
- How can imputation help?

A brief history on genotyping

- isoenzymes - not DNA but was used early.
- microstatellites - PAGE and later capillary sequencers
- RFLPs, RAPDS - restriction enzymes (agarose gels)
- SNPs - Taqman, Seqeunom
- SNP arrays (Illumina Affy)
- Genotype by sequencing (GBS)
- Sequencing

GBS and SNP array

- arrays give same calls - *consistency*, but significant cost to setup (strong price/volume dependency)
- If no SNP arrays available then GBS
- GBS different methods allowing flexibility - e.g. focus on specific parts of the genome
- detect other variation via GBS (e.g. indels, microstats)
- GBS less ascertainment bias cf arrays (depending on their design)
- GBS not (so) reliant on reference genome
- data generation from SNP array less complex; missing data a problem in GBS - need to impute.

What is genomic prediction & selection (GS)

SNP Arrays allow for cost-effective genotyping hence genomic predictions - a current challenge is to extract more value from e.g. sequence.

GS is one of the main reasons genotyping is done in industry - different flavours, but idea the same - use genomic data to provide DNA-based predictions (or enhancement over pedigree) for phenotype

$$\mathbf{y} = X\beta + Zu + e$$

BLUP

Solutions to get breeding values obtained by solving the following set of equations for $\hat{b}$, with assumptions.

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\lambda \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

In BLUP G^{-1} is from Pedigree (A^{-1}) and in GBLUP from genotypes (G^{-1})

Single Step GBLUP

H is a blended pedigree (A) and G matrix.

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \; H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

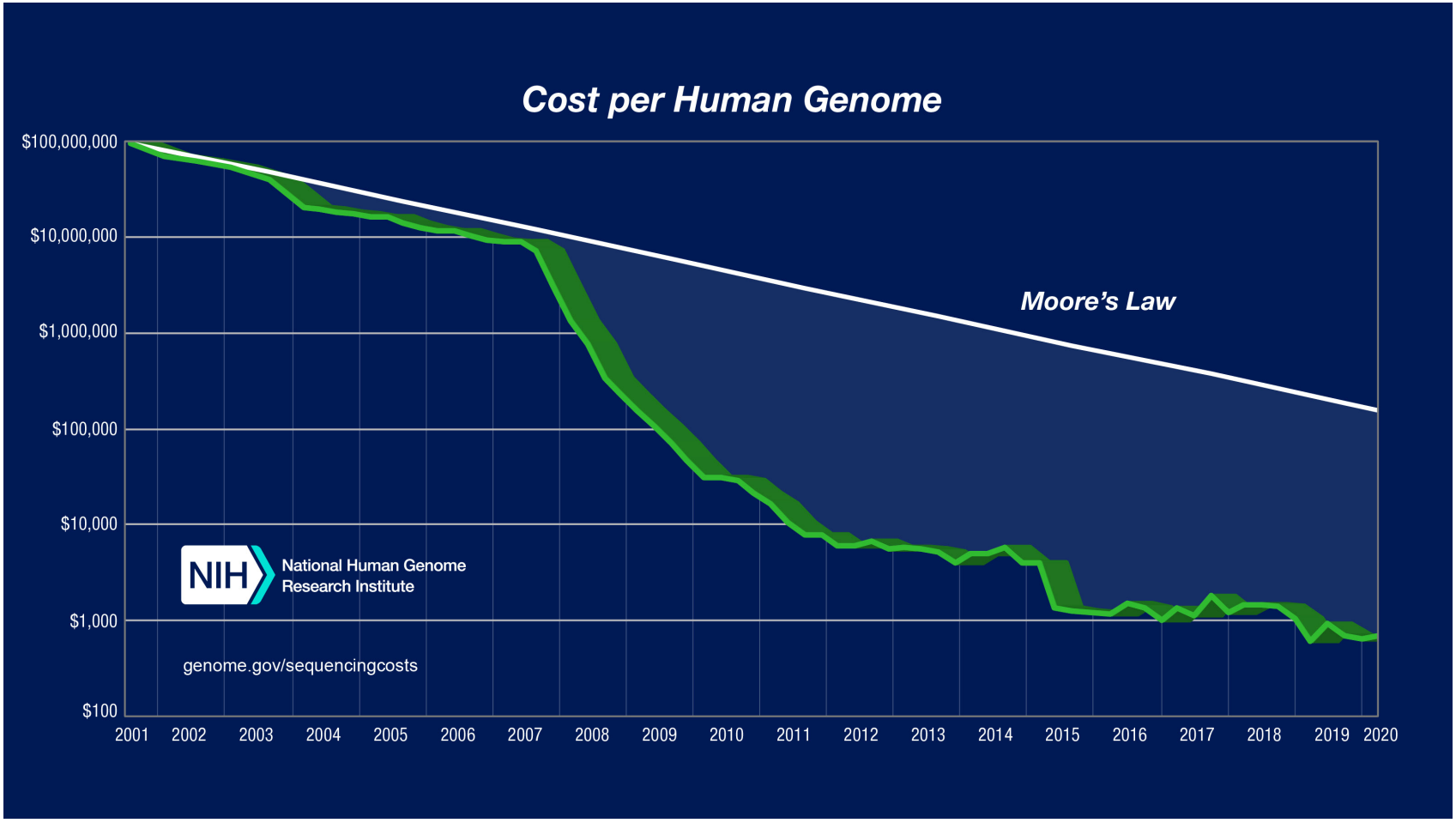
The calculation of H from pedigree (A) and genomic data (G)

$$H = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} + A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{pmatrix},$$

Other things you might want to do with the data

- population structure
- estimate genetic parameters (e.g. N_e, in-breeding)
- GWAS
- assign breed
- assign parentage and/or fix parentage
- Gene tests for marker assisted selection
- faults (genetic disease)
- signatures of selection, fixation etc.
- others

Cost of genotyping



https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Sequencing technologies

Short reads make assembly problematic

Table 1: Long read Sequencing

	PacBioSequel	PromethION	IlluminaChromium
Length	10-15	MolSize	NA
Maxlength	>80	MolSize	<100
Cost/Gb	$85	$24	$8-11
Throughput	5-10Gb	0.125-6Tb	0.8-1.8Tb

From TIGs: Van Dijk *etal*. (2018) Only takes a ng of DNA for the Chromium system

Fit for purpose - balancing costs and information content

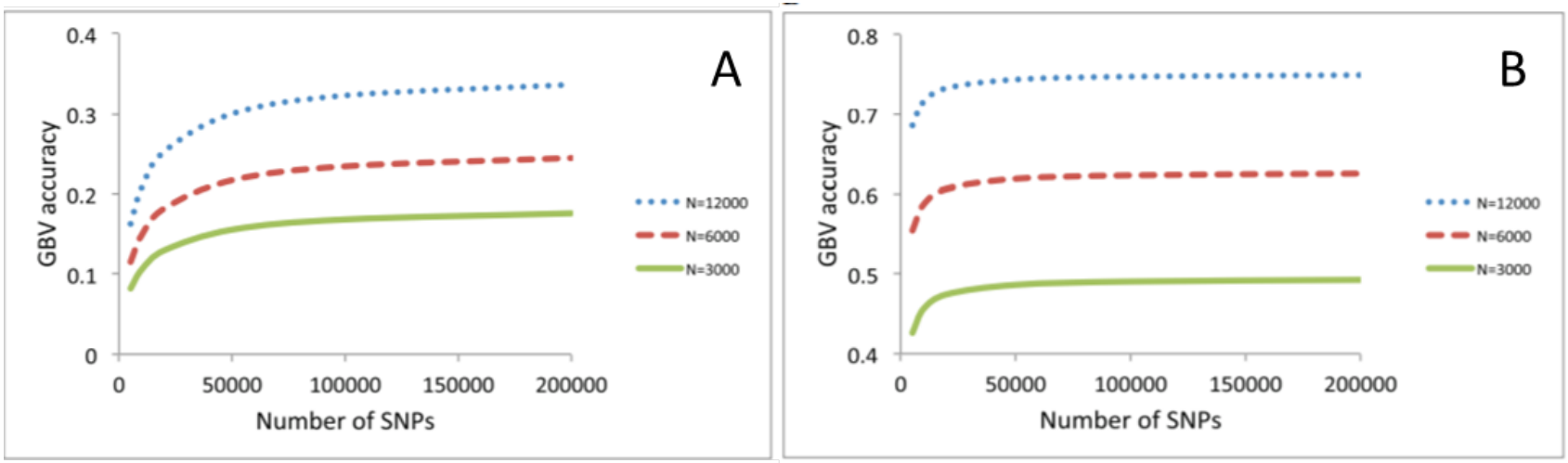
You need a decent reference population - Not all genomes & populations are the same. Some considerations:

- size & complexity (e.g. ploidy, length) of the genome
- structure of population
- effective population size (N_e)
- how well is the variome characterised
- reference genome and tools available (commercial SNP Chips)

ideally characterise the population and plan the reference population.

Why it matters - Genomic prediction accuracy & bias

GBV accuracy as a function of training sample size (n) and SNP density for a trait with h^2 of 0.25.



A:N_e=1,000; B: N_e=10,000; From Lee, Clark & van der Werf, 2018

$$M_e = \frac{2N_eLk}{log(N_eL)}$$

M_e = number of chromosome segments segregating in the population; L =average length of chromosomes; k =no of chromosomes.

Value Proposition

For SNP arrays price depends a lot on volume.

Example from animal breeding

- cost of Low density SNP chip $25
- cost of 50K chip $55
- typically fine imputation accuracy can be highly accurate if reference population is sufficient
- accuracy of imputation depends on structure of population to be imputed - important to have well designed reference population
- 50K density used in genomic selection

For a self-replacing flock genotype: sires with HD or 50K = 15x$55 + progeny with LD = 985x$25 = $25,450 compared with 1000x$55=$55,000

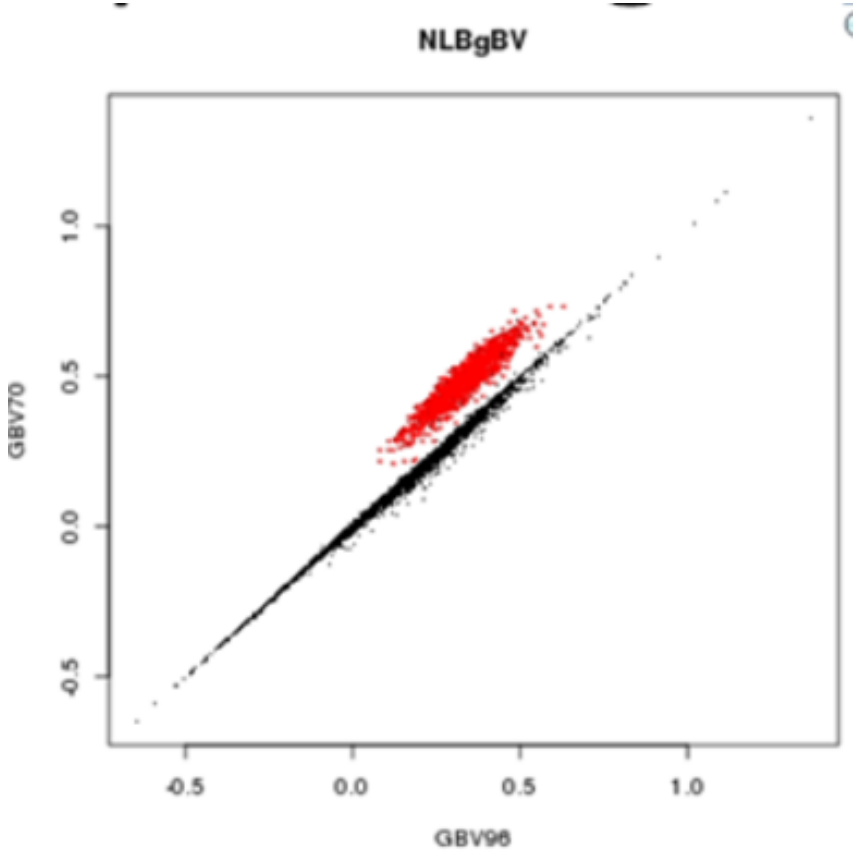
A lot depends on cost to get a decent reference population (e.g. genotyed|sequenced ancestors)

How can imputation help?

- If poorly designed imputation might just give you rubbish making things worse (e.g. poor imputation may decrease accuracy and may bias genomic predictions - making predictions worst than just pedigree)
- depending on what data you have you might be able to make more of your data without further genotyping costs
- if possible can be strategic about establishing a reference population for imputation
- if the price difference is not large for different densities might be better to just pay the higher price for the certainty, then don't have the cost of imputation.
- still a large cost for WGS cf. HD genotyping.
- done well imputation will increase the numbee of markers/individual allowing e.g. LD genotypes to be used in GS or increasing power in GWAS.

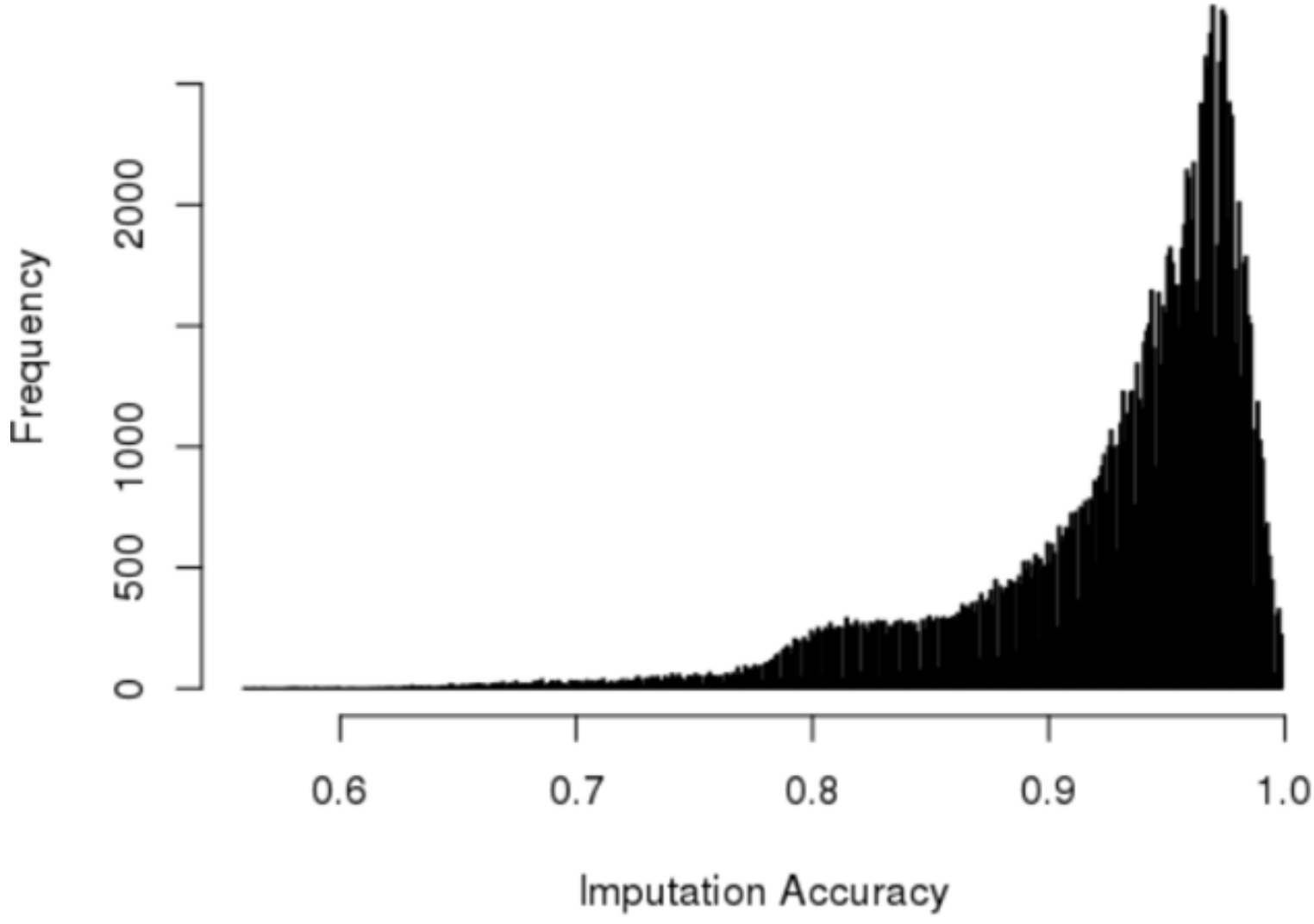
Poor imputation & genomic predictions

Effect of poorly imputed genotypes on genomic predictions -



Imputation Accuracy

Histogram of accuracies estimated by validation (n=164,035) LD to 50K.



Summary

- imputation is a useful and cost-effective method to get more utility out of your data
- results can be rubbish
- heavily dependent on you reference population - ideal if you can design this
- important to understand the process, your population, data and assess outcomes from imputation

Thank You