Session 32

An imputation pipeline for cost effective genomic selection in commercial pig breeding

Matthew A Cleveland[1]*, Nan Yu[1], Fernanda Foertter[1], Nader Deeb[1] and John M Hickey[2]

[1]Genus plc, Hendersonville, TN USA

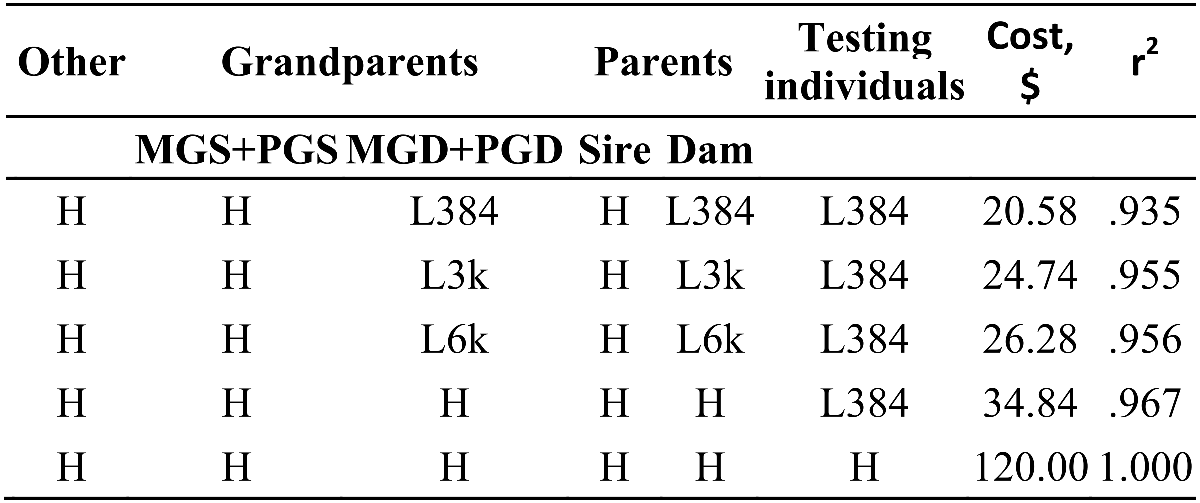
[2]School of Environmental and Rural Science, University of New England, Armidale, NSW Australia

Background

- Genomic selection has the potential for increasing genetic gains in pigs
    - Increased accuracy of selecting replacements
- Genomic approaches require large amounts of data
    - Genotypes increase breeding value accuracy
    - Improvement constrained when dense genotypes on young selection candidates are unavailable
- Cost prohibitive to genotype selection candidates
    - Large number of progeny per litter
    - 120k @ $100 = $12m per year

Cost effective genomic selection

- Strategies for dense and sparse genotyping



Other	Grandparents		Parents		Testing individuals	Cost, $	r^2
	MGS+PGS	MGD+PGD	Sire	Dam			
H	H	L384	H	L384	L384	20.58	.935
H	H	L3k	H	L3k	L384	24.74	.955
H	H	L6k	H	L6k	L384	26.28	.956
H	H	H	H	H	L384	34.84	.967
H	H	H	H	H	H	120.00	1.000

Huang Y, JM Hickey, MA Cleveland and C Maltecca. 2012. *Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost*. Genet. Sel. Evol. 44:25.

- Impute dense genotypes from small SNP panel
- Perform genomic evaluation with complete dense genotypes – gEBV for all selection candidates

Low-density genotyping

- Developed low-density SNP panel to be used across traits and lines

- SNPs selected based on 7 lines
  - Filtering based on genotype coverage, MAF and HW chi-square statistic
  - High confidence in map position
  - Even spacing (based on map position), accounting for LD

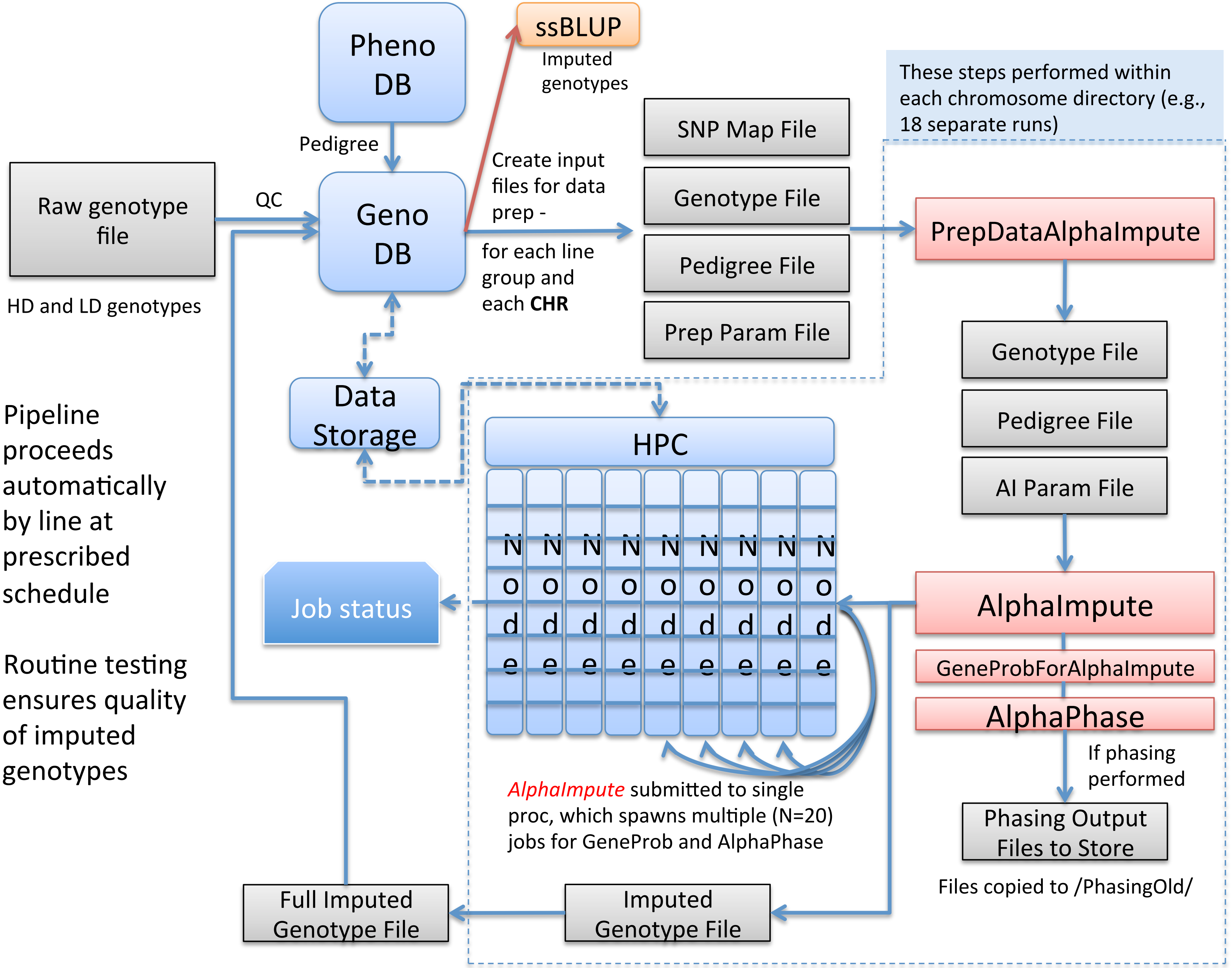
- Final panel <450 SNPs

Imputation

- AlphaImpute[1]
  - Combines simple phasing rules, long-range phasing, haplotype libraries, segregation analysis and recombination modeling
  - Imputes genotypes for all loci at the highest genotype density for animals in the analysis
  - Imputed genotypes are the sum of fully imputed alleles or the sum of allele probabilities

- Automated pipeline
  - Extract raw data, prepare files, impute genotypes, upload probable genotypes

[1] Hickey JM, BP Kinghorn, B Tier, JHJ van der Werf and MA Cleveland. 2012. *A phasing and imputation method for pedigreed populations that results in a single-stage genetic evaluation method.* Genet. Sel. Evol. 44:25.



Pheno DB

Pedigree

ssBLUP

Imputed genotypes

Raw genotype file

HD and LD genotypes

QC

Geno DB

Create input files for data prep -

for each line group and each **CHR**

SNP Map File

Genotype File

Pedigree File

Prep Param File

These steps performed within each chromosome directory (e.g., 18 separate runs)

PrepDataAlphaImpute

Genotype File

Pedigree File

AI Param File

AlphaImpute

GeneProbForAlphaImpute

AlphaPhase

If phasing performed

Phasing Output Files to Store

Files copied to /PhasingOld/

Data Storage

HPC

Node Node Node Node Node Node Node Node Node

Job status

AlphaImpute submitted to single proc, which spawns multiple (N=20) jobs for GeneProb and AlphaPhase

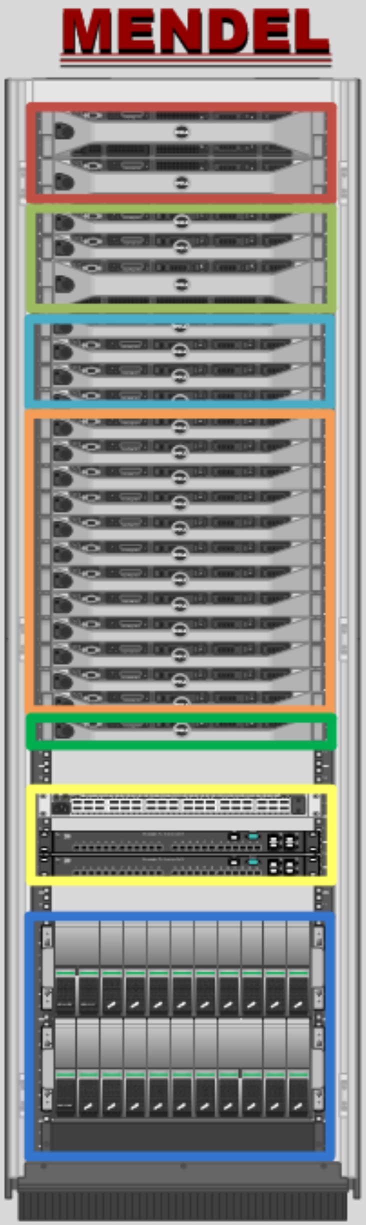
Imputed Genotype File

Full Imputed Genotype File

- ❑ Pipeline proceeds automatically by line at prescribed schedule
- ❑ Routine testing ensures quality of imputed genotypes

High Performance Computing Cluster

Cluster Totals
216 cores
1,664GB RAM
114TB Storage



Redundant Head-nodes
2 with 8 cores & 48GB RAM

Production Compute-nodes
4 with 8 cores & 48GB RAM

Genotype Database
12 cores & 96GB RAM

Panasas ActiveStor 11 pNFS Storage
2 Shelves
Triple Redundant Director Blades
114TB RAW Storage
Mounted on all Cluster Nodes

High Memory Compute-nodes
2 with 8 cores & 96GB RAM
1 with 32 cores & 512GB RAM

All-purpose Compute-nodes
Production and R&D
8 with 8 cores & 48GB RAM
4 with 12 cores & 48GB RAM

Network Connectivity
2 x Brocade TurboIron 24x 10GbE
(Cluster Mgmt & Storage)
1 x Qlogic InfiniBand QDR 32-port (MPI)

Hardware & Software Vendors

DELL
Bright Computing
Scientific Linux
panasas
BROCADE
QLOGIC
intel

Computational considerations

- >5k animals with 60k genotypes in multiple lines (>21k 60k genotypes overall)
  - Dense genotyping continues to grow
- Large number of selection candidates genotyped for low-density panel
  - Following imputation all have 60k genotypes
- Routine pipeline runs
  - <13 hours per line (in serial)
  - Probable genotype files ~600Mb per line

Imputation testing

- Investigate imputation and gEBV accuracy using alternative genotyping scenarios

- Data
  - N=4,579 60k genotyped
  - N=183 full parent/grandparent genotypes; no progeny
  - Three low-density panels: 450, 3k, 6k
  - 33k SNPs after filtering (chr 1-18)

Results: imputation accuracy

- Implementation considering 4 approaches to genotyping close relatives

	Genotyping Scenario						Imputation accuracy		
	Other	PGS+MGS	PGD+MGD	Sire	Dam	Progeny			
	4220	74	108	70	107	184	**450**	**3k**	**6k**
S1	H	H	H	H	H	L	**0.97**	0.99	1.00
S2	H	0	0	H	H	L	**0.95**	0.98	0.99
S3	H	H	0	H	0	L	**0.91**	0.97	0.98
S4	H	H	L	H	L	L	**0.94**	0.99	0.99
S1_r	0	H	H	H	H	L	**0.96**	0.99	0.99
S2_r	0	0	0	H	H	L	**0.92**	0.97	0.97
S3_r	0	H	0	H	0	L	**0.85**	0.94	0.95
S4_r	0	H	L	H	L	L	**0.90**	0.97	0.98

Results : gEBV accuracy

- Calculate gEBV using single-stage evaluation[1]
- Compare gEBV from full dense genotyping to gEBV from low-density genotyping/imputation

		Genotyping Scenario						Imputed gEBV Accuracy		
	N HD Geno.	Other	PGS+ MGS	PGD+ MGD	Sire	Dam	Progeny	450	3k	6k
S1	2519	H	H	H	H	H	L	**0.94**	0.97	0.97
S2	2344	H	0	0	H	H	L	**0.89**	0.95	0.96
S3	2318	H	H	0	H	0	L	**0.87**	0.92	0.93
S4	2318	H	H	L	H	L	L	**0.90**	0.96	0.97
S1_r	323	0	H	H	H	H	L	**0.79**	0.81	0.80
S2_r	148	0	0	0	H	H	L	**0.71**	0.73	0.71
S3_r	122	0	H	0	H	0	L	**0.69**	0.76	0.75
S4_r	122	0	H	L	H	L	L	**0.75**	0.80	0.80

[1] Aguilar et al. , 2009

Summary

- An imputation pipeline has been implemented in routine production

- Continued optimization will stabilize runtimes as data size increases

- Imputation accuracy for very low-density panel was high

- gEBV accuracy was high, but appropriate genotyping strategies are needed

Genus | ABS PIC

Thank you!

“Pioneering animal genetic improvement to help nourish the world.”