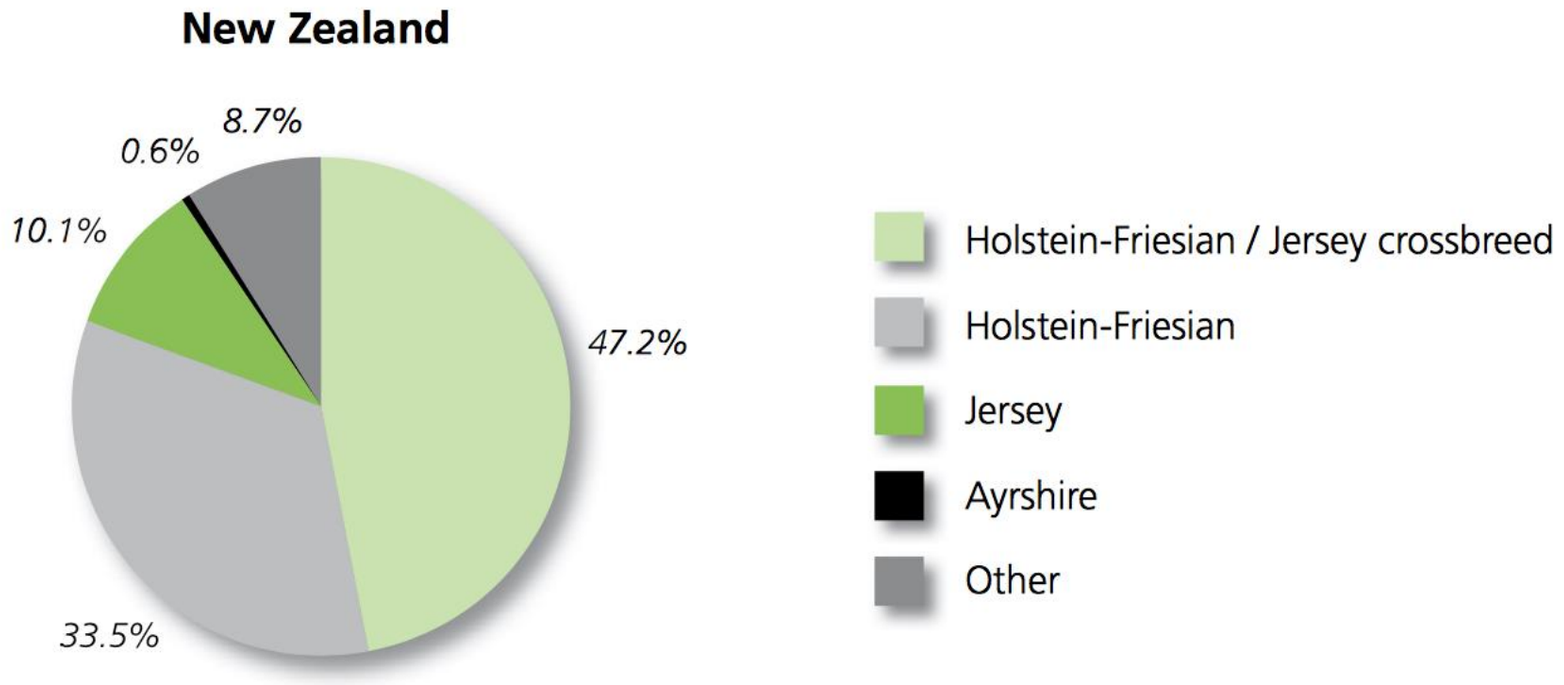
Experiences solving large scale single step marker best linear unbiased models

Bevin Harris
LIC, New Zealand

Background

- Livestock Improvement Genomic Evaluation System
  - Multiple breed population – Jersey, Holstein-Friesian and JxHF crosses



2015/2016

Background

- Genomic Evaluation History
  - 2008 – GBLUP on genotyped sires – multi-breed G matrix
  - 2011 – GBLUP on genotyped sires + cows
  - 2013 – Hybrid Single Step genotyped animals + ancestors – Euclid distance G Matrix
  - 2018 – Full marker single step model – all animals

Marker Model

- Why a marker single step model?
  - Number of SNP markers << number of genotyped individuals
    - Selected sequence SNP with individual weights
  - Marker effects useful for processing animals between evaluations
  - Easier to add an extra polygenic term in model

Marker Model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_g\mathbf{Mm}_g + \mathbf{Z}_n\mathbf{u}_n + \mathbf{Za} + \mathbf{e}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_g & \mathbf{Z}_n \end{bmatrix}$$

$\mathbf{b}$ are the fixed effects

$\mathbf{m}_g$ are the SNP marker effects

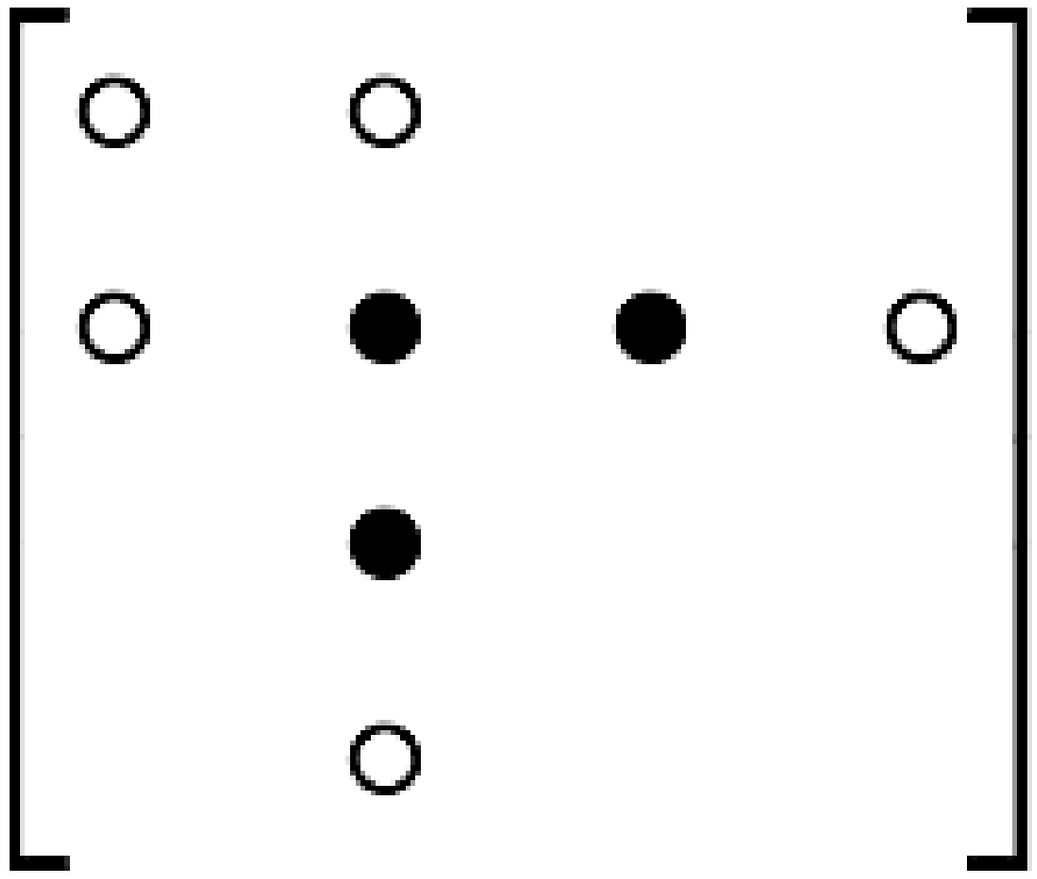
$\mathbf{u}_n$ are the marker breeding values for the non-genotyped individuals

$\mathbf{a}$ are the additive polygenic effects

$\mathbf{e}$ is the random residual

Solving MME Equations

- Single step marker mixed model (Fernando et. al, 2016)



○ Moderate to high density

● Very high density

Solving MME Equations

- Traditional mixed model applications
- Pre-conditioned conjugate gradient (PCG) with iteration on data method of choice
- Diagonal precondition matrix which is easy to invert
  - Reduce condition number of MME
  - Cluster the eigenvalues of MME
  - Improves convergence speed

Solving MME Equations

- Single step marker mixed model
  - Just in time solving within the PCG algorithm. only need vector by matrix products
  - Use “imputation on the fly” using Cholesky decomposition (Matilainen et al., 2016)
  - Sparse matrix storage — no iteration on data
  - Block precondition matrices

Solving MME Equations

- Block precondition matrix

$$\begin{bmatrix} (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1} & & \\ & (\mathbf{K1})^{-1} & \\ & & (\mathbf{K2})^{-1} \\ & & & (\mathbf{K3})^{-1} \end{bmatrix}$$

Eigen Decomposition

Cholesky Decomposition

$$\mathbf{K1} = \mathbf{M}_g'\mathbf{Z}_g'\mathbf{R}^{-1}\mathbf{Z}_g\mathbf{M}_g + \mathbf{I}\alpha + \mathbf{M}_g'\mathbf{A}^{gn}(\mathbf{A}^{nn})^{-1}\mathbf{A}^{ng}\mathbf{M}_g\lambda$$

$$\mathbf{K2} = \mathbf{Z}_n'\mathbf{R}^{-1}\mathbf{Z}_n + \mathbf{A}^{nn}\lambda$$

$$\mathbf{K3} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{A}^{-1}{}_{\kappa}$$

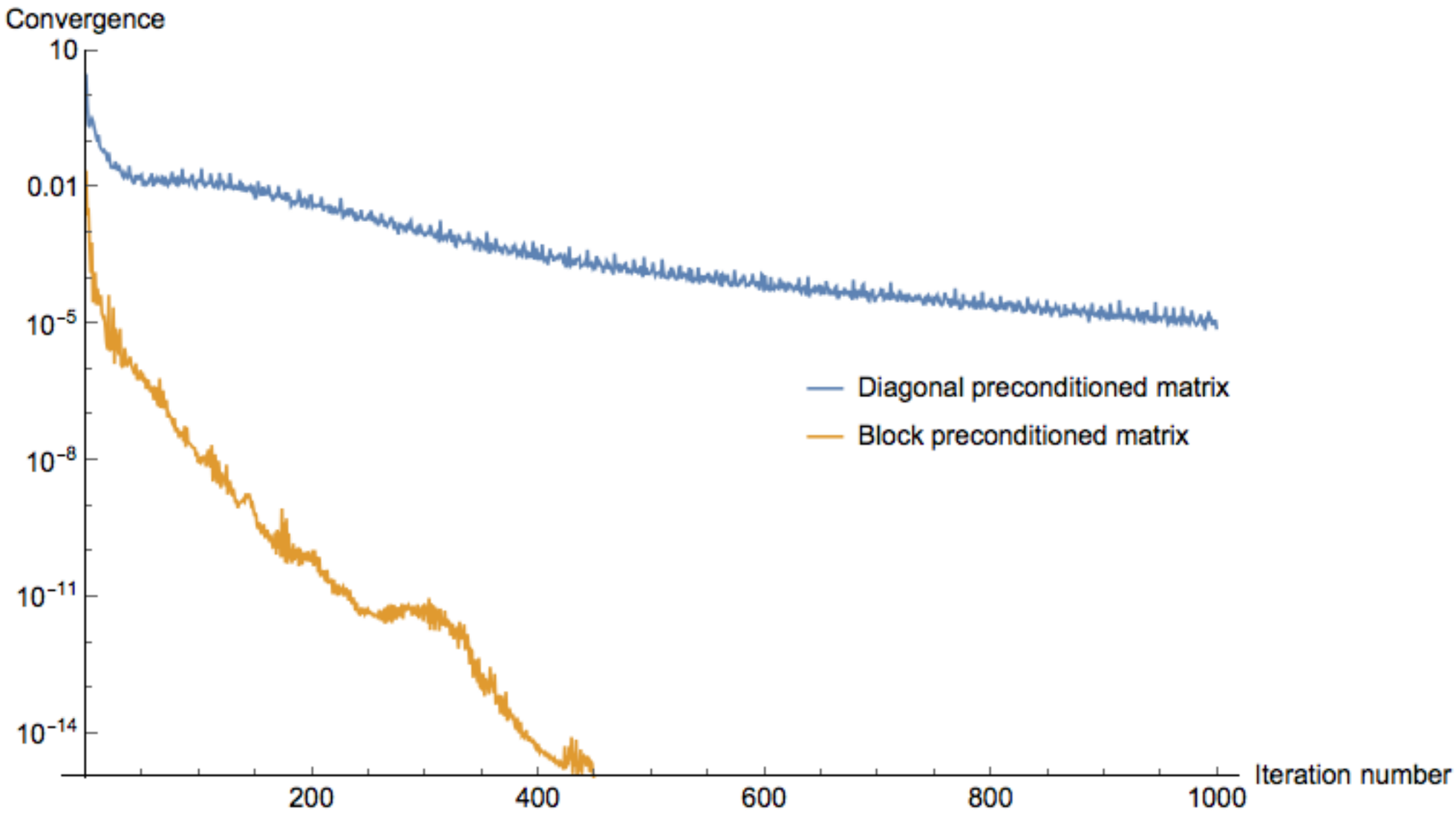
Data

	Animals	Genotypes	Phenotypes	N SNP
Dataset 1	212k	12K	122k	7.7k
Dataset 2	28.3M	105k	1.8M	34.7K
Dataset 3	28.3M	105k	15.9M	34.7K

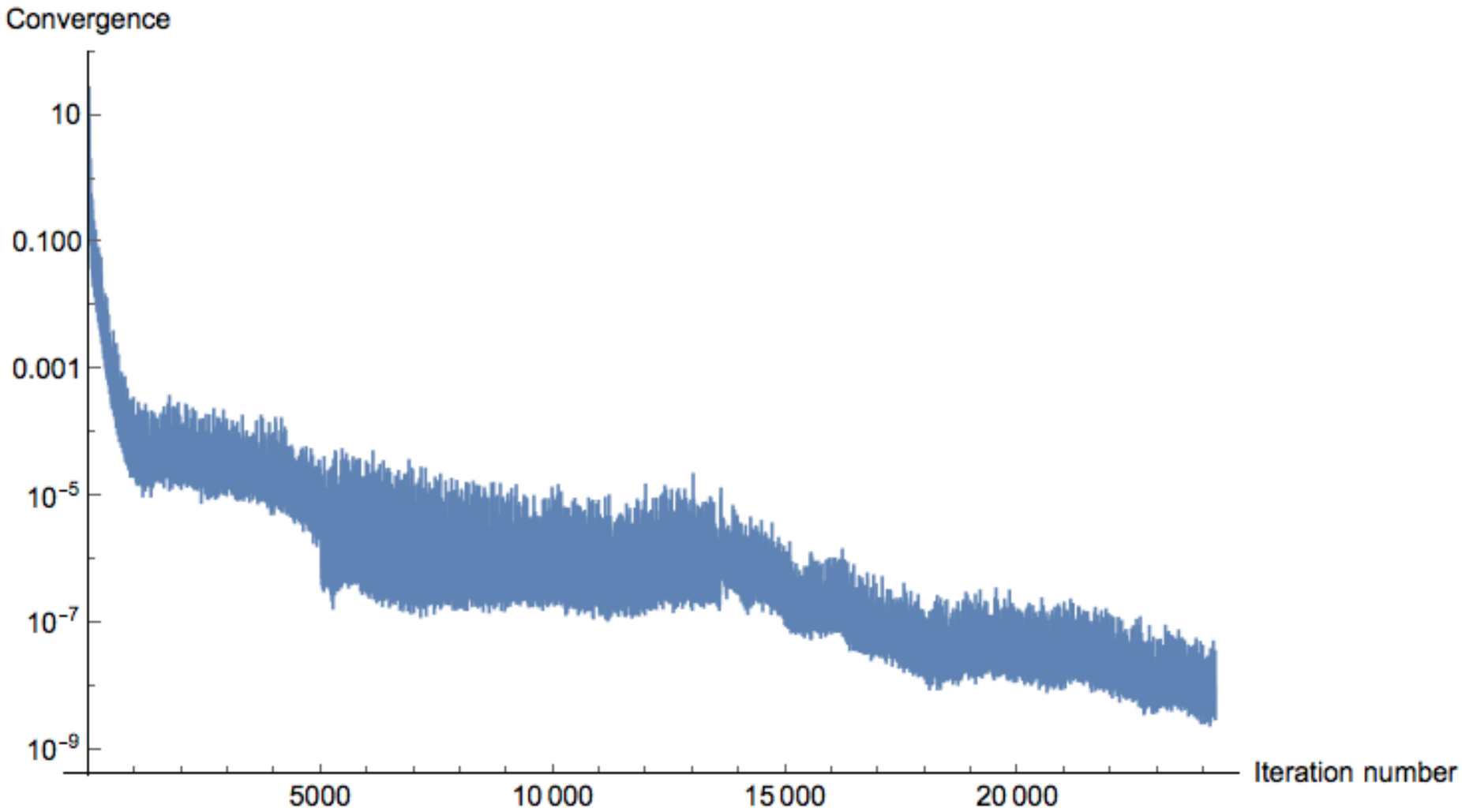
Solving MME Equations

- Pre-computation steps
  - Sparse numerator relationship matrix – Time: < 5m
  - $\mathbf{M}'_g\mathbf{A}^{gn}(\mathbf{A}^{nn})^{-1}\mathbf{A}^{ng}\mathbf{M}_g$
    considerably effort required – 34.7K SNP =18 hours

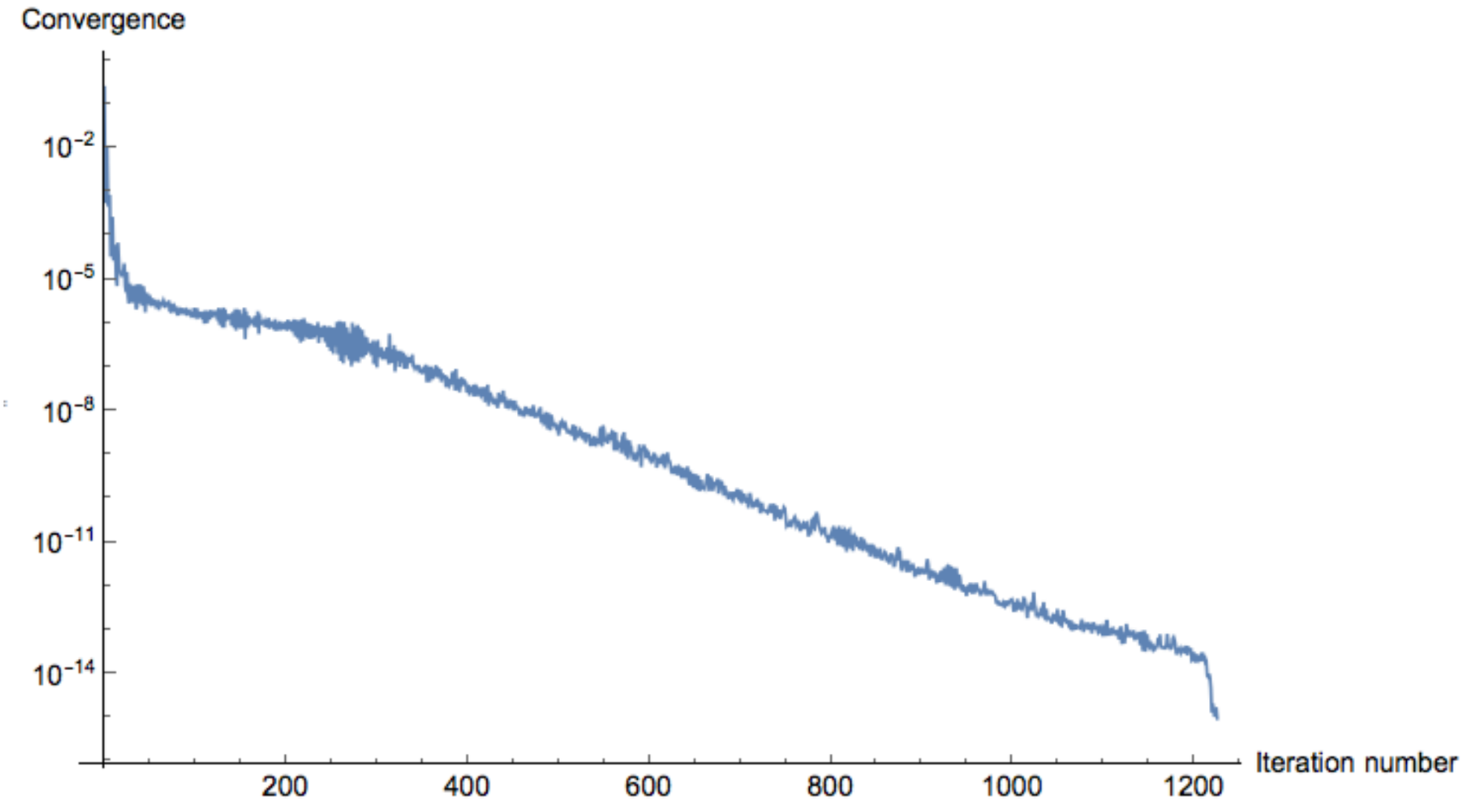
Convergence Dataset 1



Convergence Dataset 2 Diagonal Precondition Matrix



Convergence Dataset 2 Block Precondition Matrix



Data

	Diagonal		Block	
	Iterations	**Time**	**Iterations**	**Time**
Dataset 1	1834	8m9s	476	2m12s
Dataset 2	>30,000	> 7 days	1216	475m01s
Dataset 3	>30,000	> 7 days	1001	368m23s

Considerations

- Problems with a large number of contemporary groups in fixed effects
  - Two blocks in the precondition matrix for the fixed effects
- Large number of SNP – multiple blocks in the precondition matrix
  - Increase in number of iterations

Conclusions

- With the incorporation of SNP markers into large scale genetic evaluation systems the computational efficiency in terms of time and convergence becomes important

- For large datasets, the use of a diagonal preconditioned matrix within a PCG solver may be insufficient to provide convergence within a reasonable time (if at all).

- The use of a block preconditioned matrix in the PCG solver allows convergence

Conclusions

- The downside to the use of a block preconditioned matrix is the time to compute the Eigenvalues/vectors or Cholesky factorizations of the blocks

- This can take in excess of two hours for populations > 25m animals

- Probably still not a major time burden


REGISTER
NOW
ICAR
7–11 February 2018
THE GLOBAL STANDARD
FOR LIVESTOCK DATA
Annual Conference
icar2018.nz
11–16 February 2018
11th WORLD CONGRESS
ON GENETICS
APPLIED TO
LIVESTOCK PRODUCTION
wcgalp.com
AOTEA CENTRE, AUCKLAND, NEW ZEALAND
See you in 2018 in
NEW ZEALAND

Questions?