Super-large scale genomic evaluations

J. Ødegård[1,2] and T. H. E. Meuwissen[2]

[1] AquaGen AS, P.O. Box 1240, N-7462 Trondheim, Norway

[2] Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Aas, Norway

AquaGen

Large-scale genomic analyses:

- Original single-step model (ssGBLUP), Legarra et al. (2009), Christensen & Lund (2010)
  - Complexity increases with number of genotyped animals
  - Inverse genomic relationship matrix (GRM) must be computed prior to the analysis
- Single-step marker effects model (ssMEM), Fernando et al. (2016)
  - No need for inverse GRM
  - Complexity depends on number of loci
- Populations of limited Ne
  - Limited number of haplotypes
- Genomic data can be approximated by a smaller number of principal components

AquaGen

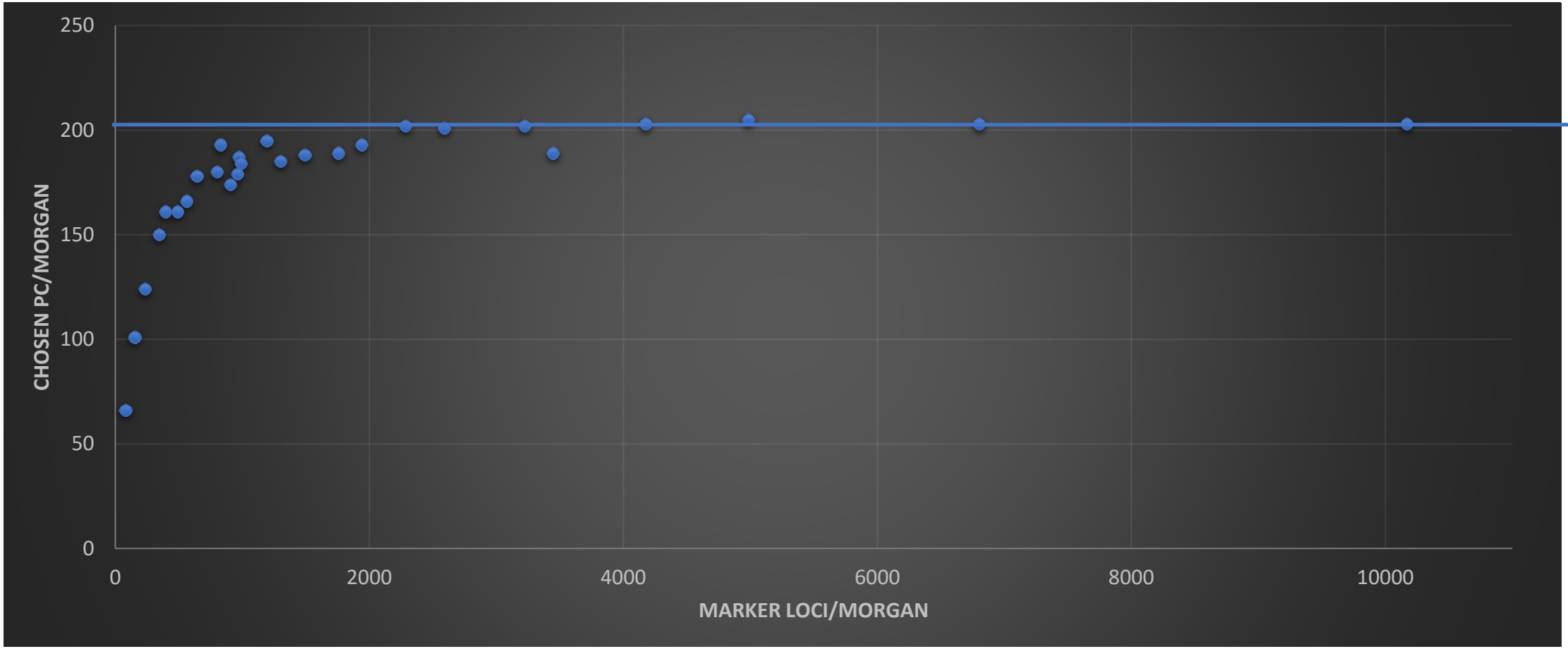
Future developments:

1. No. of genotyped animals increases fast
2. SNP denisity increases to HD (~600k) to sequence data
3. All animals genotyped
   - Removes need for single-step methods
   - But: in the near future genomic evaluations should include ss-method

AIM: develop genomic evaluations that can handle millions of animals and millions of SNP and allow for nongenotyped animals

AquaGen

Principal components explaining >99% of variance (Ne = 500, N = 10,000)

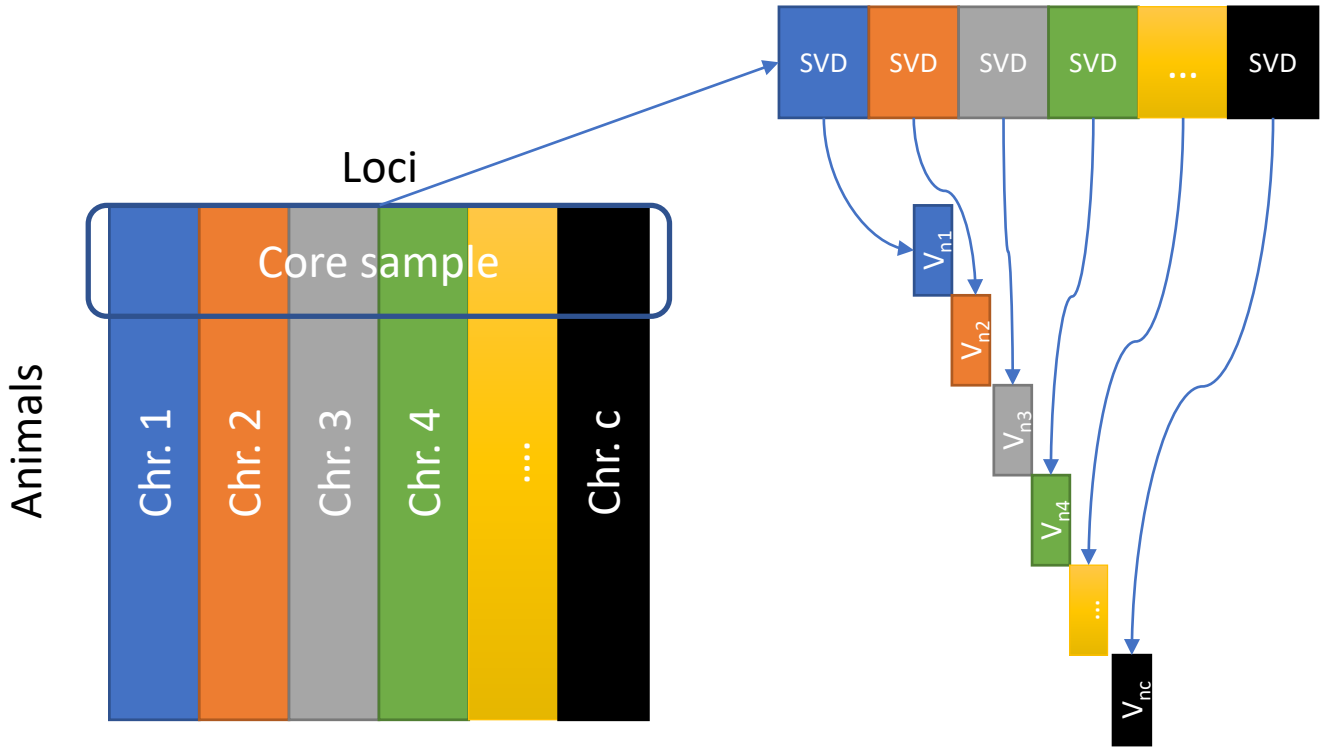


Singular value decomposition (SVD) of genomic data

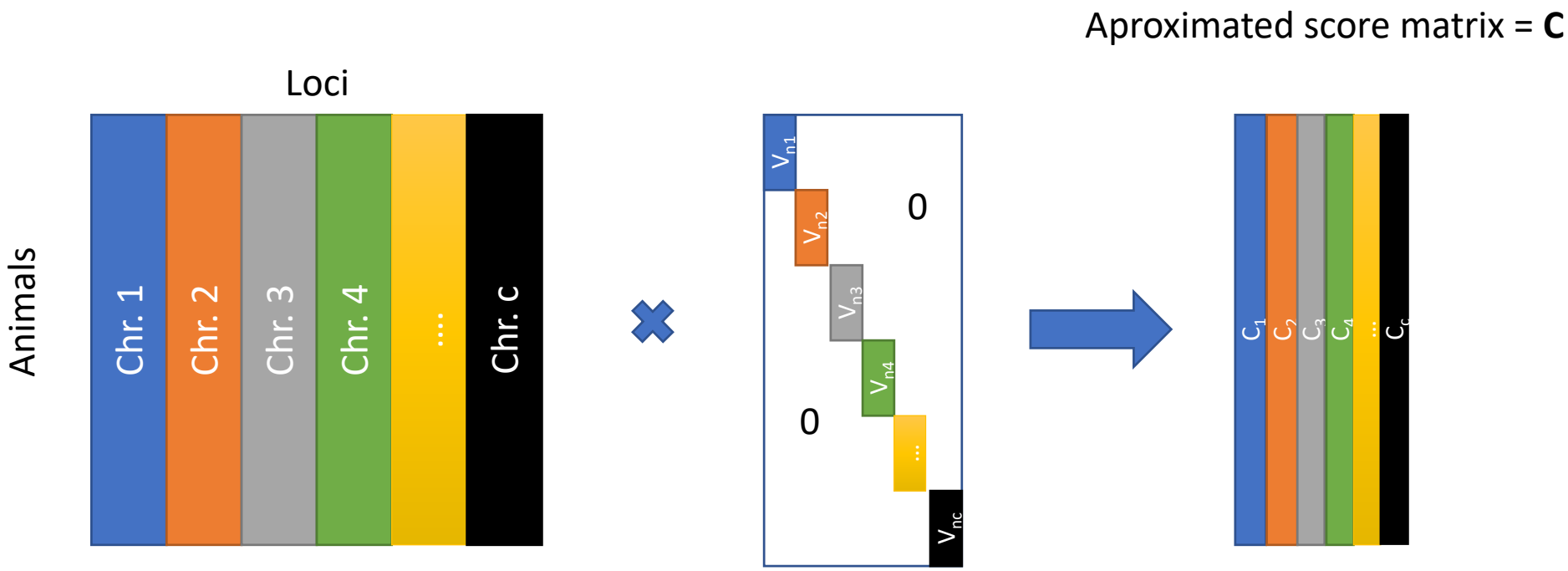
- SVD of $N \times k$ (centered) genotype matrix
  - $\mathbf{M} = \mathbf{USV}'$
  - $\mathbf{U}$ =eigenvectors of $\mathbf{MM}'$ (orthonormal), **U'U = I**
  - $\mathbf{V}$ =eigenvectors of $\mathbf{M}'\mathbf{M}$(orthonormal), **V'V = I**
  - $\mathbf{S}$ is a diagonal matrix (square root of eigenvalues)
- Principal component ridge regression model
  - $\mathbf{y} = \mathbf{Mb} + \mathbf{e} = \mathbf{Ts} + \mathbf{e}$
    - $\mathbf{s} = \mathbf{V}'\mathbf{b}$ (principal component regression coefficients)
    - $\mathbf{T} = \mathbf{US}\ (= \mathbf{MV})$ (score matrix)
- Dimension reduction, include the first q principal components
  - $\mathbf{M} \approx \mathbf{U}_q\mathbf{S}_q\mathbf{V}_q{}'$
  - $\mathbf{T} = \mathbf{U}_q\mathbf{S}_q\ (= \mathbf{MV}_q)$
- Performing SVD is demanding for large datasets

AquaGen

Chromosome-wise SVD on a core sample



Chromosome-wise SVD on a core sample



Single-step marker effects model (ssMEM)

- Fernando et al. (GSE 2016, 48:96)
- Compute expected genotypes for non-genotyped animals by solving:
  - $\mathbf{A^{22}}\widehat{\mathbf{M}}_\mathbf{2} = -\mathbf{A^{21}M_1}$
  - Total genotype matrix (genotyped and ungenotyped) is:
  - $\mathbf{M} = \begin{bmatrix} \mathbf{M_1} \\ \widehat{\mathbf{M}}_\mathbf{2} \end{bmatrix}$
- ssMEM:
  - $\mathbf{y} = \mathbf{ZMb} + \mathbf{Z_2}\boldsymbol{\epsilon} + \mathbf{e}$
    - where $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \left(\mathbf{A^{22}}\right)^{-1}\sigma_a^2\right)$
- ssMEM equations:
  - $\begin{bmatrix} \mathbf{M'Z'ZM} + \mathbf{I}\rho\lambda & \mathbf{M'Z'Z_2} \\ \mathbf{Z_2'ZM} & \mathbf{Z_2}'\mathbf{Z_2} + \mathbf{A^{22}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{M'Z'y} \\ \mathbf{Z_2}'\mathbf{y} \end{bmatrix}$
  - where $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$

AquaGen

Single-step principal component ridge-regression (ssPCRR)

- Compute expected scores for all non genotyped animals by solving:
  - $\mathbf{A^{22}\hat{C}_2 = -A^{21}C_1}$ ($\mathbf{C_1}$ = approx. scores of genotyped)
  - Total score matrix (genotyped and ungenotyped) is now: $\mathbf{C} = \begin{bmatrix} \mathbf{C_1} \\ \hat{\mathbf{C}}_2 \end{bmatrix}$
- ssPCRR model:
  - $\mathbf{y = ZCs + Z_2\epsilon + e}$
- ssPCRR equations:
  - $$\begin{bmatrix} \mathbf{C'Z'ZC + I}\rho\lambda & \mathbf{C'Z'Z_2} \\ \mathbf{Z_2'ZC} & \mathbf{Z_2'Z_2 + A^{22}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{s}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{C'Z'y} \\ \mathbf{Z_2'y} \end{bmatrix}$$
- Genotyped EBV:
  - $\hat{\mathbf{a}}_1 = \mathbf{C_1}\hat{\mathbf{s}}$
- Ungenotyped EBV
  - $\hat{\mathbf{a}}_2 = \mathbf{C_2}\hat{\mathbf{s}} + \hat{\boldsymbol{\epsilon}}$

AquaGen

Simulation study

- Simulated population using QMSim (Sargolzaei and Schenkel, 2009)
  - 30 chromosomes of 100 cM
    - 24,259 SNP marker loci
    - 829 QTL
  - $h^2 = 0.25$
  - $N_e = 500$
  - 20,000 genotyped
  - 100,000 ungenotyped
  - All animals had own phenotype
- Chromosome-wise SVD
  - 2000 core animals
  - Number of chosen components set to explain >99% of genomic variation
- Block-iterative solver
- All analyses were run in a Julia environment (https://julialang.org/)

AquaGen

Performance of models

- If full-scale SVD is performed
  - All models are equivalent and give identical results
- (Chromosome-wise) Reduced-dimension ssPCRR
  - EBV correlation to original ssGBLUP was >0.9999
- Large-scale analysis
  - 4710 PC needed (157 per chromosome)
  - Setting up equation system ~ 4 minutes
  - Solving ~ 3 minutes
- Accuracies:
  - Genotyped: 0.90
  - Ungenotyped: 0.76

AquaGen

ssPCRR vs. APY

APY

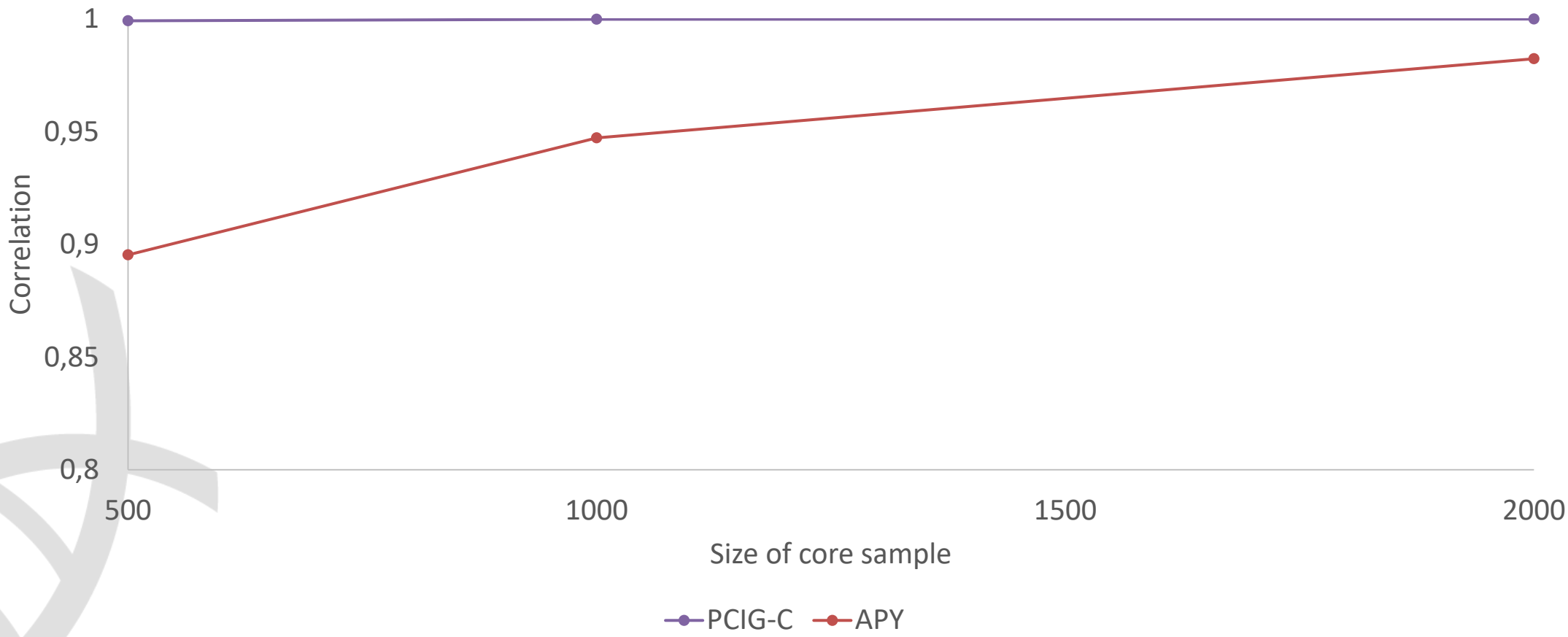
- Utilizes a core sample
- Approximates the (inverse) GRM
- Core BV explain all genetic variation
  - Non-core EBVs are merely linear functions of the core EBVs
  - Core and non-core EBVs assumed equally reliable
  - Smaller cores inflates calculated reliability
- Larger cores needed

ssPCRR

- Utlilizes a core sample
- Approximates genotype matrix
  - No need for inverse GRM
- Alleles (haplotypes) within the core explain all genetic variation
  - All BV are functions of components effects
  - No. of components may exceed core size
  - Core and non-core EBVs not assumed equally reliable
- Smaller cores needed

AquaGen

Correlation to full G matrix based GBLUP (h^2 = 0.5)



AquaGen

Principal component-based inverse GRM (PCIG)

- Invertible GRM
  - $\mathbf{G} \approx \frac{1}{\rho} \cdot \widehat{\mathbf{T}}\widehat{\mathbf{T}}'$ does not have full rank and has thus no inverse
    - The problem can be circumvented by adding a small number to the diagonal
  - $\tilde{\mathbf{G}} = \left(\frac{1}{\rho} \cdot \widehat{\mathbf{T}}\widehat{\mathbf{T}}' + \mathbf{I}\theta\right)$
- Exact inverse by the Woodbury formula:
  - $\tilde{\mathbf{G}}^{-1} = \left(\frac{1}{\rho} \cdot \widehat{\mathbf{T}}\widehat{\mathbf{T}}' + \mathbf{I}\theta\right)^{-1} = \frac{1}{\theta}\left(\mathbf{I} - \widehat{\mathbf{T}}\left(\widehat{\mathbf{T}}'\widehat{\mathbf{T}} + \mathbf{I_p}\rho\theta\right)^{-1}\widehat{\mathbf{T}}'\right)$
  - The only explicit inverse needed is: $\left(\widehat{\mathbf{T}}'\widehat{\mathbf{T}} + \mathbf{I_p}\rho\theta\right)^{-1}$
  - Dimension is number of chosen components (columns in $\widehat{\mathbf{T}}$)
  - Inverse GRM can be produced for any number of animals

AquaGen

Direct calculation of BayesC by SVD

- BayesC prior=> prob π: $b_j \sim N(0,\sigma^2)$ and prob (1- π): $b_j = 0$
- PCRR-MME: $(S^2 + I\lambda)s = T'y$ with $\hat{b} = V\hat{s}$
- PEV of SNP effects:

$$PEV(b_j) = \boldsymbol{V}_{j.}(\boldsymbol{S^2} + \boldsymbol{I}\lambda\ )^{-1}\boldsymbol{V}_{j.}'\sigma_e^2$$

- Effective no of records to estimate SNP effect, n_j:

$$PEV(b_j) = \frac{\sigma_e^2}{n_j + \lambda}$$

$$(n_j + \lambda)\widehat{b_j} = RHS_j$$

AquaGen

Posterior probab. SNP has effect

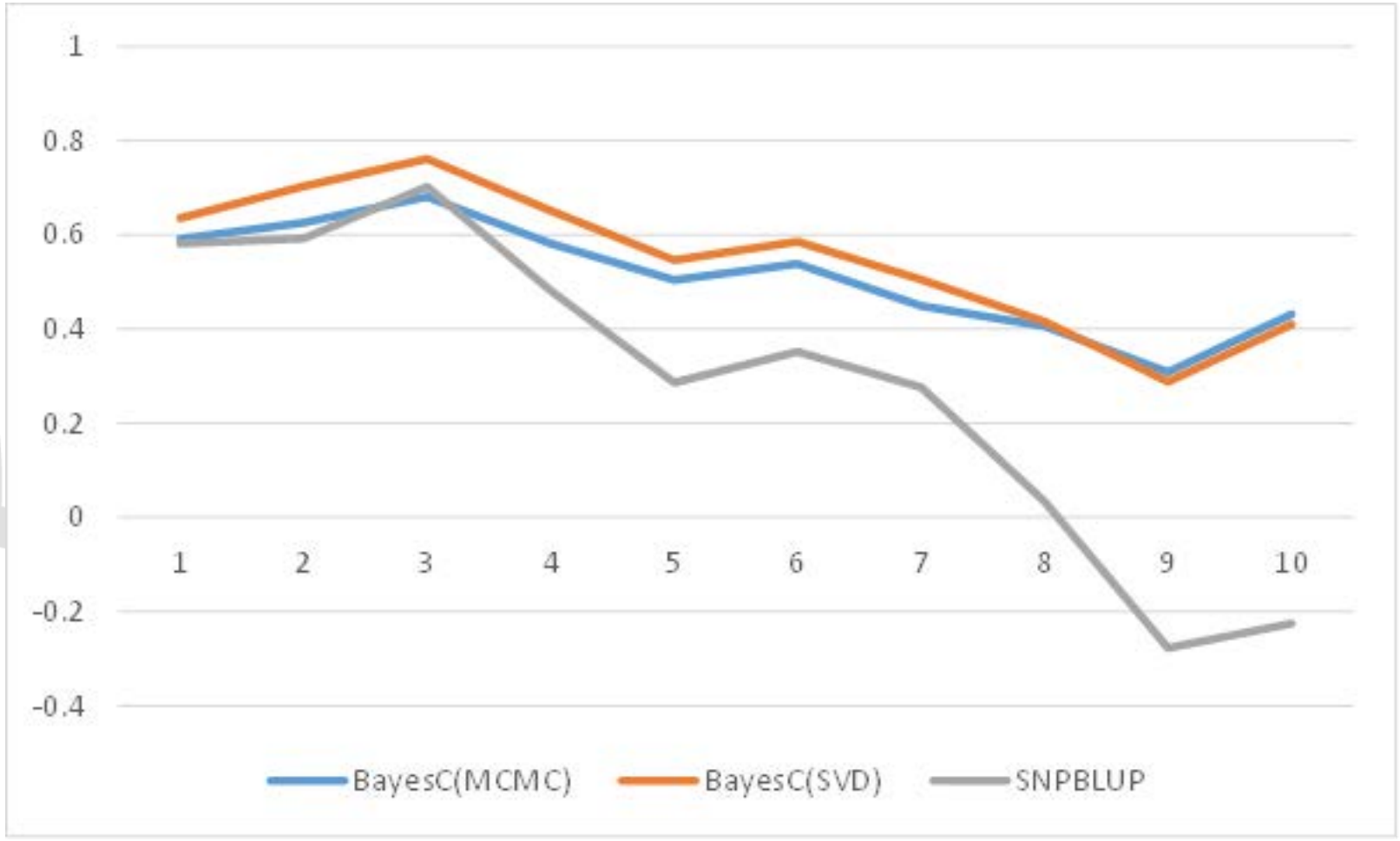
- Log-Likelihood ratio of presence/absence of SNP effect j:

$$LLR_j = \frac{1}{2}[\log(\lambda) - \log(\lambda + n_j) + \frac{RHS_j{}^2}{(n_j+\lambda)\sigma_e^2}]$$

- Log ratio of Priors
  - LRPrior=log[π/(1-π)]
- Log-Ratio of posterior prob = LLR_j +LRPrior
- Weighing SNP effects by their Posterior Probs
  - Use in weighted GBLUP model
  - i.e. direct calculation of BayesC - GEBV

AquaGen

Accuracy of selection over 10 generations



AquaGen

Conclusions

- As no of genotyped animals and SNPchip density increases
  - Cannot have animal based model
  - Cannot have SNP based model
  - Solution : SVD component based model

- Large-scale genomic data from populations of limited Ne
  - Few PC capture nearly all genetic variation
    - << number of loci (dense data)
    - << number of genotyped animals (large N)

AquaGen

Conclusions

- Fast SVD and dimension reduction
  - Smaller core sample
  - Parallell chromosome-wise SVD

- Single-step PC ridge regression (ssPCRR)
  - Very close approximation of the original ssGBLUP EBVs
  - Dimension of equation system greatly reduced
  - No need for inverse relationship matrices of genotyped animals

- Direct calculation of BayesC by SVD
  - Accuracy similar to that of MCMC methods
  - BayesC GEBV more persistent across generations than BLUP-GEBV

AquaGen

Acknowledgements

The Research Council of Norway

Project no. 255297: ”From whole genome sequence to precision breeding”

AquaGen