A single-step principal component ridge regression model for 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

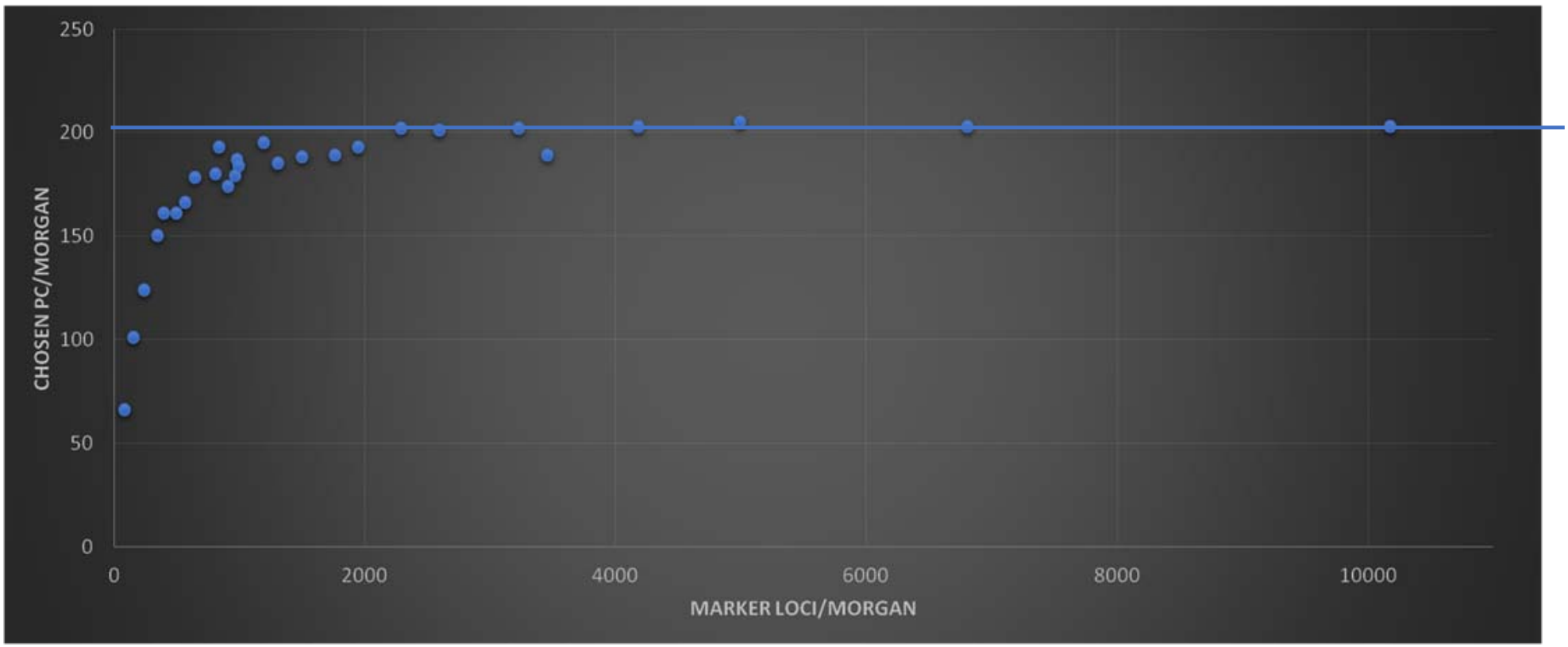
AquaGen

Large-scale genomic data

- 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

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

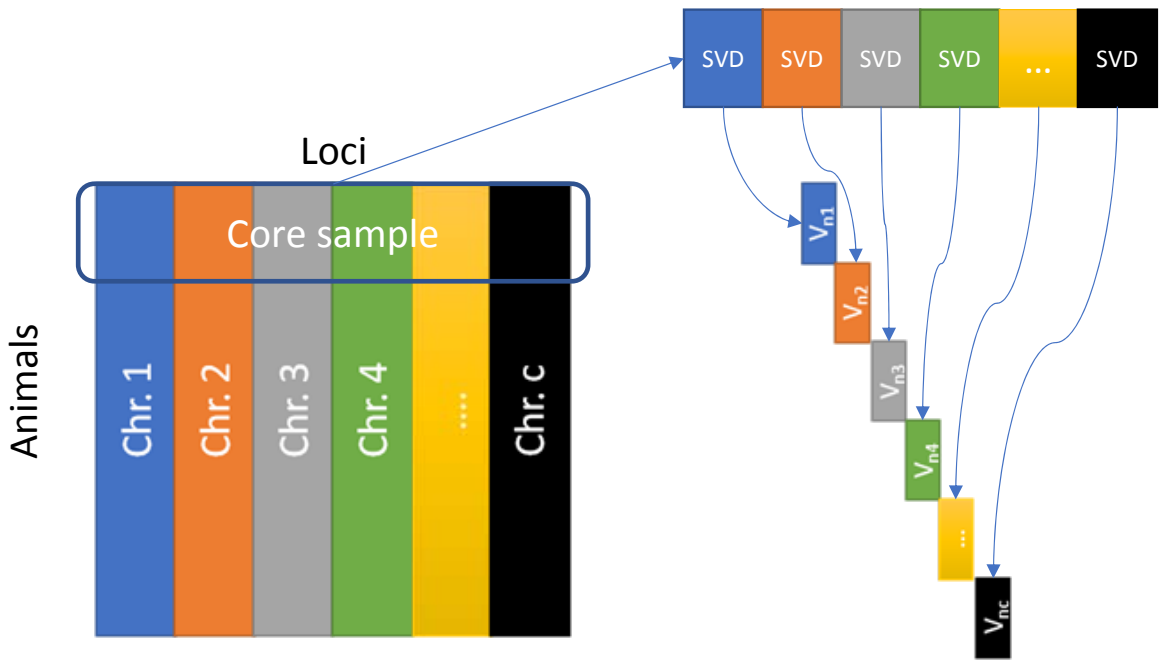


Singular value decomposition (SVD) of genomic data

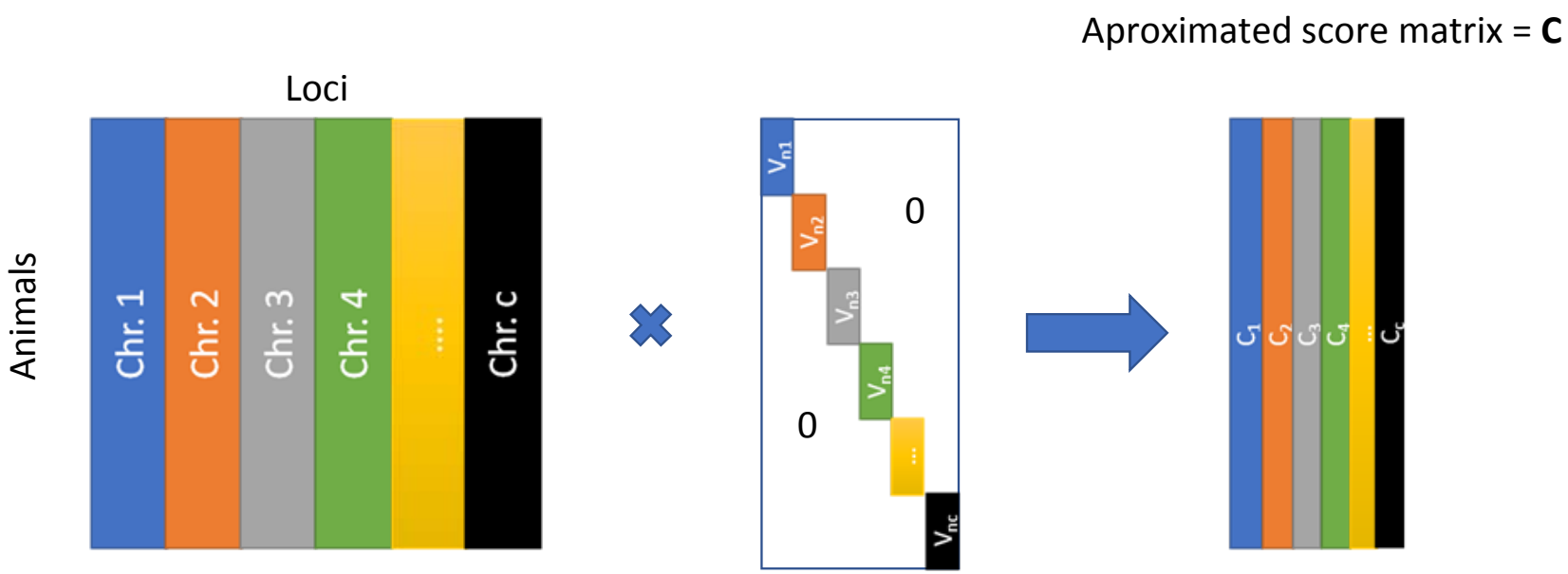
- SVD of $N \times k$ (centered) genotype matrix
  - $\mathbf{M} = \mathbf{USV}'$
  - $\mathbf{U}$ =eigenvectors of $\mathbf{MM}'$ (orthonormal), **U'U = I**
  - $\mathbf{V}$ =eigenvectors of $\mathbf{M}'\mathbf{M}$(orthonormal), **V'V = I**
  - $\mathbf{S}$ is a diagonal matrix (square root of eigenvalues)
- Principal component ridge regression model
  - $\mathbf{y} = \mathbf{Ts} + \mathbf{e}$
    - $\mathbf{s} = \mathbf{V}'\mathbf{b}$ (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:
  - $\mathbf{A}^{22}\hat{\mathbf{M}}_2 = -\mathbf{A}^{21}\mathbf{M}_1$
  - Total genotype matrix (genotyped and ungenotyped) is:
  - $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \hat{\mathbf{M}}_2 \end{bmatrix}$
- ssMEM:
  - $\mathbf{y} = \mathbf{ZMb} + \mathbf{Z}_2\boldsymbol{\epsilon} + \mathbf{e}$
    - where $\epsilon \sim N\left(\mathbf{0}, \left(\mathbf{A}^{22}\right)^{-1}\sigma_a^2\right)$
- ssMEM equations:
  - $\begin{bmatrix} \mathbf{M}'\mathbf{Z}'\mathbf{ZM} + \mathbf{I}\rho\lambda & \mathbf{M}'\mathbf{Z}'\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{ZM} & \mathbf{Z}_2'\mathbf{Z}_2 + \mathbf{A}^{22}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}'\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \end{bmatrix}$
  - 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:
  - $\mathbf{A}^{22}\hat{\mathbf{C}}_2 = -\mathbf{A}^{21}\mathbf{C}_1$ ($\mathbf{C}_1$ = 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:
  - $\mathbf{y} = \mathbf{ZCs} + \mathbf{Z}_2\boldsymbol{\epsilon} + \mathbf{e}$
- ssPCRR equations:
  - $$\begin{bmatrix} \mathbf{C}'\mathbf{Z}'\mathbf{ZC} + \mathbf{I}\rho\lambda & \mathbf{C}'\mathbf{Z}'\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{ZC} & \mathbf{Z}_2'\mathbf{Z}_2 + \mathbf{A}^{22}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{s}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}'\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \end{bmatrix}$$
- 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

Conclusions

- 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)
- 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

AquaGen

Acknowledgements

The Research Council of Norway

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

AquaGen