JUNE 27, 2024

Big Genomic Data in Agriculture – Challenges and Chances

Sommertagung der AG "Landwirtschaftliches Versuchswesen" Soest

Johanna-Sophie Schlüter and Armin O. Schmitt
Breeding Informatics Division

Big Genomic Data

Challenges

1. Extremely large size of the data
2. Lack of phenotype information for some/all individuals
3. Remaining uncertainties regarding individuals

Chances

1. Generate knowledge from genomic information without phenotypes
2. Discovery of private alleles and examination of relationships at large scale
3. Time and memory efficient

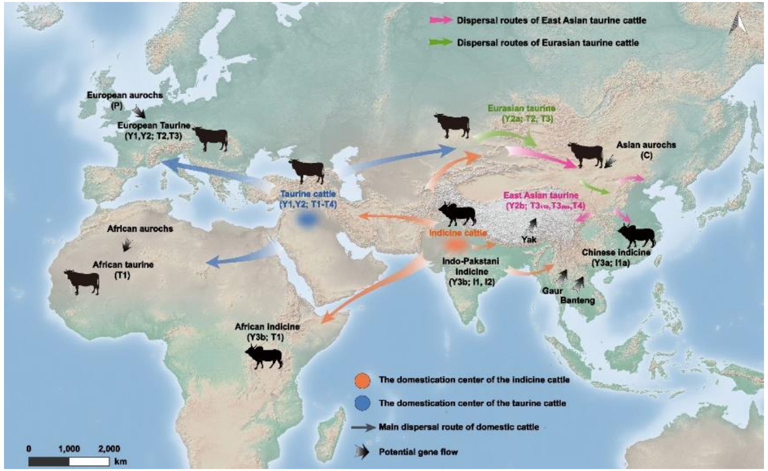
Example Data Set

The 1000 Bull Genomes Project

- Total: 6191 individuals
- ~ 120 million variants
- 30,396 genes (ARS-UCD1.2)

Subset: >= 50 individuals in a breed

- 19 breeds
- 3718 individuals
- ~ 68 million variants
  - MAC >= 1


Dispersal routes of East Asian taurine cattle
Dispersal routes of Eurasian taurine cattle
European aurochs (P)
European Taurine (Y1,Y2; T2,T3)
Eurasian taurine (Y2a; T2, T3)
Asian aurochs (C)
Taurine cattle (Y1,Y2; T1-T4)
East Asian taurine (Y2b; T3, T3, T4)
African aurochs
Indicine cattle
Yak
Chinese indicine (Y3a; I1a)
African taurine (T1)
Indo-Pakstani Indicine (Y3b; I1, I2)
Gaur
Banteng
African Indicine (Y3b; T1)
The domestication center of the indicine cattle
The domestication center of the taurine cattle
Main dispersal route of domestic cattle
Potential gene flow
0
1,000
2,000
km

Domestication routes of taurine and indicine cattle.
Xia et al. (2023)

MAC = Minor allele count

Hayes and Daetwyler (2019); Rosen et al. (2020)

Current Usage

- Imputation and studies on imputation algorithms
- Selection signature detection, ancestry, and diversity
- Few phenotype associations:
  - African animal trypanosomiasis

Limited usage only

Boitard et al. (2016); Pausch et al. (2017);
Jiang et al. (2022); Rajavel et al. (2022)

Phylogenetic Trees

Phylogenetic Trees of Cattle

Dutta et al. (2020):

294 individuals (*B. taurus* & *B. indicus*), ~ 8 million variants

Neumann et al. (2023):

1691 individuals (*B. taurus*), ~ 23 million variants

Chen et al. (2023):

495 individuals (*B. taurus* & *B. indicus*), ~ 67 million variants

How Big is Big Enough?

Sample *n* SNPs from virtual matrix

Calculate distance matrix based on *n* SNPs

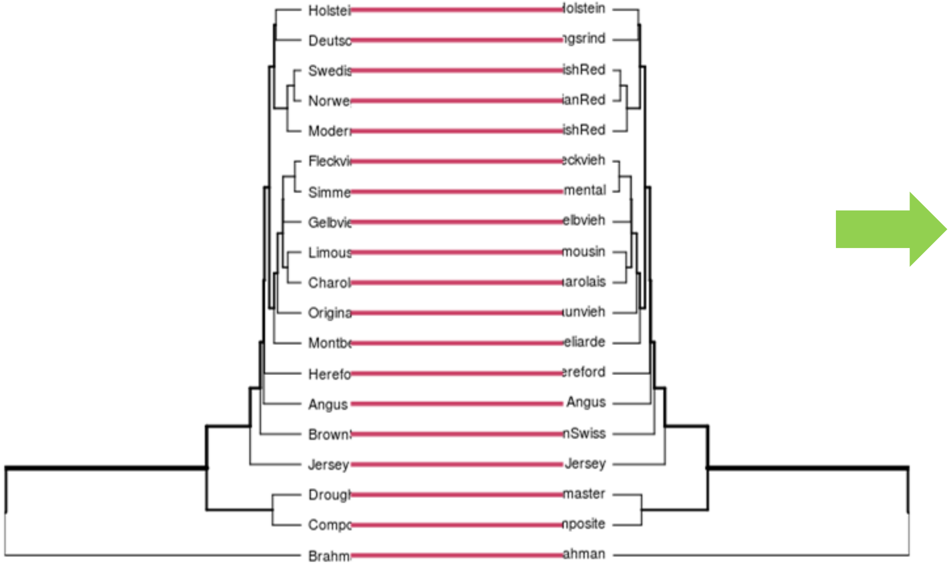
Construction of trees

Determine differences between trees (entanglement scores)

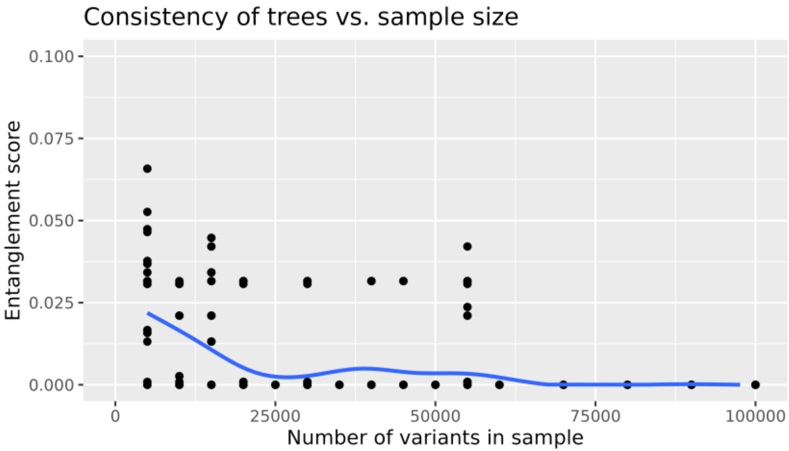
Schmitt (2019) , and using R packages from: Galili (2015); Grueneberg & De Los Campos (2019)

Results

Tanglegram of two phylogenetic trees



Entanglement score = 0



Using R packages from: Galili (2015), Wickham (2016)

Private Alleles

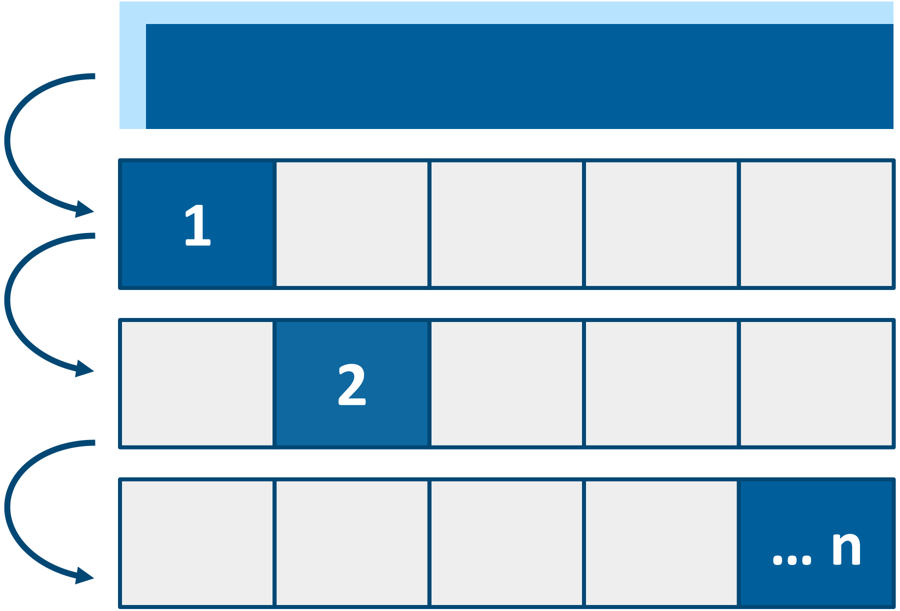
Hopping Window

Constructing virtual matrix

Extract the first *n* genomic variants

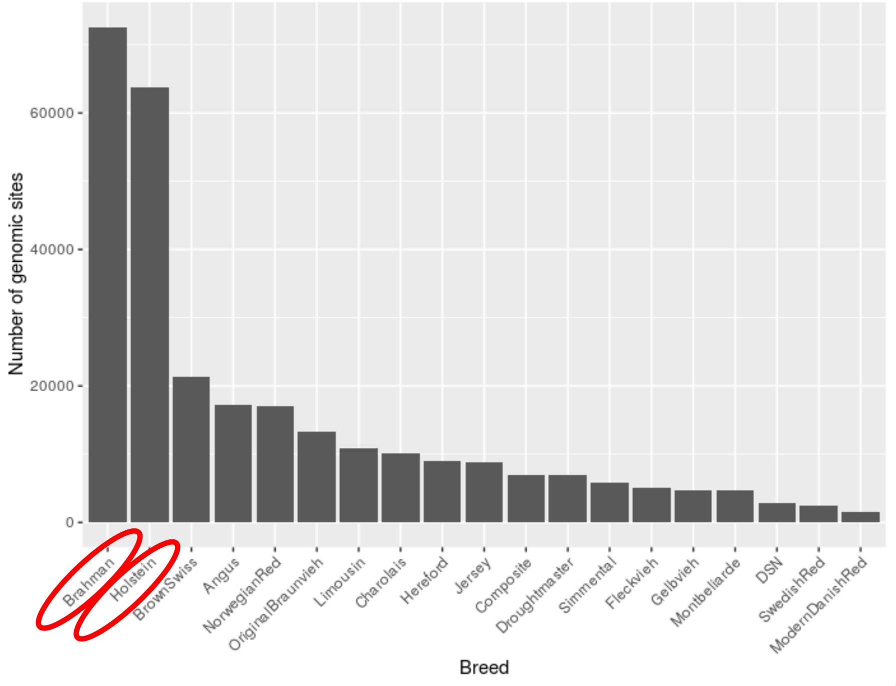
Determine and store private alleles

Go to next window of *n* genomic variants



Using R packages from: Jombart & Ahmet (2011); Kinsella et al. (2011); Kamvar, Brooks & Grünwald (2015); Grueneberg & De Los Campos (2019);

Chromosome 28: 13.6 Million Genomic Sites



- Most breeds only contain few private alleles
- Most private alleles occur in only few individuals of a breed
- Potential:
  - Define set of interesting variants
  - Discriminate efficiently between groups

Conclusions

Big Data is omnipresent in agriculture and readily available

Genomic data by itself is rich in information and often not exploited to its full potential

Neglected data due to missing phenotype information

➡ Insight to relevant variations and relationships is possible using pragmatic approaches

Extremely large data size hinders analysis

➡ For big data it is crucial to identify efficient analysis options

Thank you for your attention!

Supervisors:

Prof. Dr. Armin Schmitt
(University of Göttingen)
Prof. Dr. Mehmet Gültas
(University of Applied Sciences Soest)
Prof. Dr. Michaela Schmitz
(University of Applied Sciences Soest)

Breeding Informatics Division:

Dr. Felix Heinrich
Dr. Thomas Lange
Henry Munroe
Maria Rotärmel
Qian Fei Zhu

References

- Boitard, S., Boussaha, M., Capitan, A., Rocha, D., & Servin, B. (2016). Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds. *Genetics*, *203*(1), 433–450. https://doi.org/10.1534/genetics.115.181594
- Chen, N., Xia, X., Hanif, Q., Zhang, F., Dang, R., Huang, B., Lyu, Y., Luo, X., Zhang, H., Yan, H., Wang, S., Wang, F., Chen, J., Guan, X., Liu, Y., Li, S., Jin, L., Wang, P., Sun, L., … Lei, C. (2023). Global genetic diversity, introgression, and evolutionary adaptation of indicine cattle revealed by whole genome sequencing. *Nature Communications*, *14*(1), 7803. https://doi.org/10.1038/s41467-023-43626-z
- Dutta, P., Talenti, A., Young, R., Jayaraman, S., Callaby, R., Jadhav, S. K., Dhanikachalam, V., Manikandan, M., Biswa, B. B., Low, W. Y., Williams, J. L., Cook, E., Toye, P., Wall, E., Djikeng, A., Marshall, K., Archibald, A. L., Gokhale, S., Kumar, S., … Prendergast, J. G. D. (2020). Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication. *Nature Communications*, *11*(1), 4739. https://doi.org/10.1038/s41467-020-18550-1
- Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718–3720. https://doi.org/10.1093/bioinformatics/btv428
- Grueneberg, A., & De Los Campos, G. (2019). BGData - A Suite of R Packages for Genomic Analysis with Big Data. *G3 Genes|Genomes|Genetics*, *9*(5), 1377–1383. https://doi.org/10.1534/g3.119.400018
- Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, *7*, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024
- Jiang, Y., Song, H., Gao, H., Zhang, Q., & Ding, X. (2022). Exploring the optimal strategy of imputation from SNP array to whole-genome sequencing data in farm animals. *Frontiers in Genetics*, *13*, 963654. https://doi.org/10.3389/fgene.2022.963654
- Jombart, T., & Ahmed, I. (2011). *adegenet 1.3-1* : new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521
- Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, *Volume 2011*, bar030. https://doi.org/10.1093/database/bar030
- Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, *6*. https://doi.org/10.3389/fgene.2015.00208
- Neumann, G. B., Korkuć, P., Arends, D., Wolf, M. J., May, K., König, S., & Brockmann, G. A. (2023). Genomic diversity and relationship analyses of endangered German Black Pied cattle (DSN) to 68 other taurine breeds based on whole-genome sequencing. *Frontiers in Genetics*, *13*, 993959. https://doi.org/10.3389/fgene.2022.993959
- Pausch, H., MacLeod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D., & Goddard, M. E. (2017). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, *49*, 24. https://doi.org/10.1186/s12711-017-0301-x
- Rajavel, A., Klees, S., Hui, Y., Schmitt, A. O., & Gültas, M. (2022). Deciphering the Molecular Mechanism Underlying African Animal Trypanosomiasis by Means of the 1000 Bull Genomes Project Genomic Dataset. *Biology*, *11*(5), 742. https://doi.org/10.3390/biology11050742
- Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., Rowan, T. N., Low, W. Y., Zimin, A., Couldrey, C., Hall, R., Li, W., Rhie, A., Ghurye, J., McKay, S. D., Thibaud-Nissen, F., Hoffman, J., Murdoch, B. M., Snelling, W. M., … Medrano, J. F. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, *9*(3), giaa021. https://doi.org/10.1093/gigascience/giaa021
- Schmitt, A. O. (2022, June 27-30). *Efficient analysis of big genomic data sets on a common PC* [Poster presentation]. IWBBIO 2022 9th international Work-Conference on Bioinformatics and Biomedical Engineering, Gran Canaria (SPAIN).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed. 2016). Springer International Publishing : Imprint: Springer.
- Xia, X., Qu, K., Wang, Y., Sinding, M.-H. S., Wang, F., Hanif, Q., Ahmed, Z., Lenstra, J. A., Han, J., Lei, C., & Chen, N. (2023). Global dispersal and adaptive evolution of domestic cattle: a genomic perspective. *Stress Biology*, *3*, 8. https://doi.org/10.1007/s44154-023-00085-2