Accuracy of imputation from SNP array data to sequence level in chicken

Guiyan Ni[1], T.M. Strom[2], H. Pausch[3], C. Reimer[1],

R. Preisinger[4], H. Simianer[1], M. Erbe[1,5]

[1] Animal Breeding and Genetics Group, Georg-August-University Göttingen, Germany

[2] Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany

[3] Chair of Animal Breeding, Technische Universität München, Freising, Germany

[4] Animal Breeding and Genetics Group, Georg-August-Universität, Göttingen, Germany

[5] Institute of Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany




Introduction

- Array data available for a large number of individuals in many livestock populations
- Whole-genome sequence data
  - Now available due to technical progress in the last years
  - Much higher density than common SNP array panels
  - Still expensive → not possible to sequence all individuals of a population

→ Imputation as key strategy

- Is it promising to impute SNP array data up to sequencing level within a purebred brown layer line?

Data

- 1075 individuals from a commercial brown layer line

Generation	1	2	3	4	5	6	Total
Array data	85	61	66	637	114	112	1,075
Sequence data	22	1	2	-	-	-	25

- Genomic data:
  - Array: Affymetrix Axiom® Chicken Genotyping Array with 580K SNPs
  - Sequence: Illumina HiSeq2000, ~ 8x coverage

- Filtering criteria:

Chromosomes	3	6	28	Total
Array data	35.3K	14.2K	2.9K	52.4K
Sequence data	1164.8K	440.6K	44.3K	1,647.7K

e > 95%

Methods

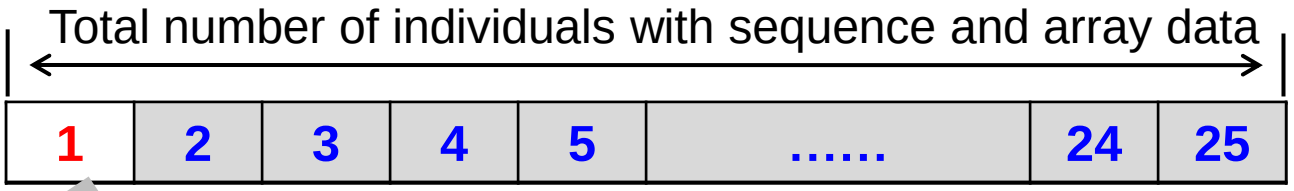
- Imputation programs tested
  - Minimac (Howie et al. 2012)
    - ✓ Applies a hidden Markov model
    - ✓ Needs pre-phased data
      - → phasing done with Beagle 3 (Browning and Browning 2007)
  - FImpute (Sargolzaei et al. 2014)
    - ✓ Applies an overlapping sliding window method
    - ✓ Combines pedigree and linkage disequilibrium information
  - IMPUTE2 (Howie et al. 2009)
    - ✓ Applies a hidden Markov model

Methods

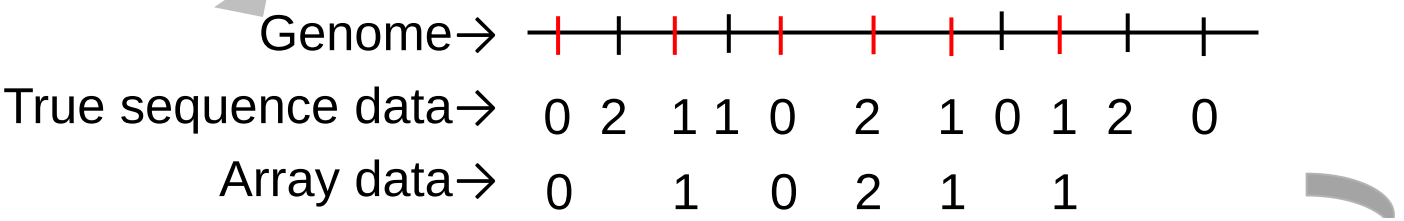
- How well do the imputation programs perform?
- Three different validation strategies
  - Leave-one-out cross-validation
  - Sire-progeny-conflicts
  - Randomly masked SNPs

Leave-one-out cross-validation

- Within sequenced individuals



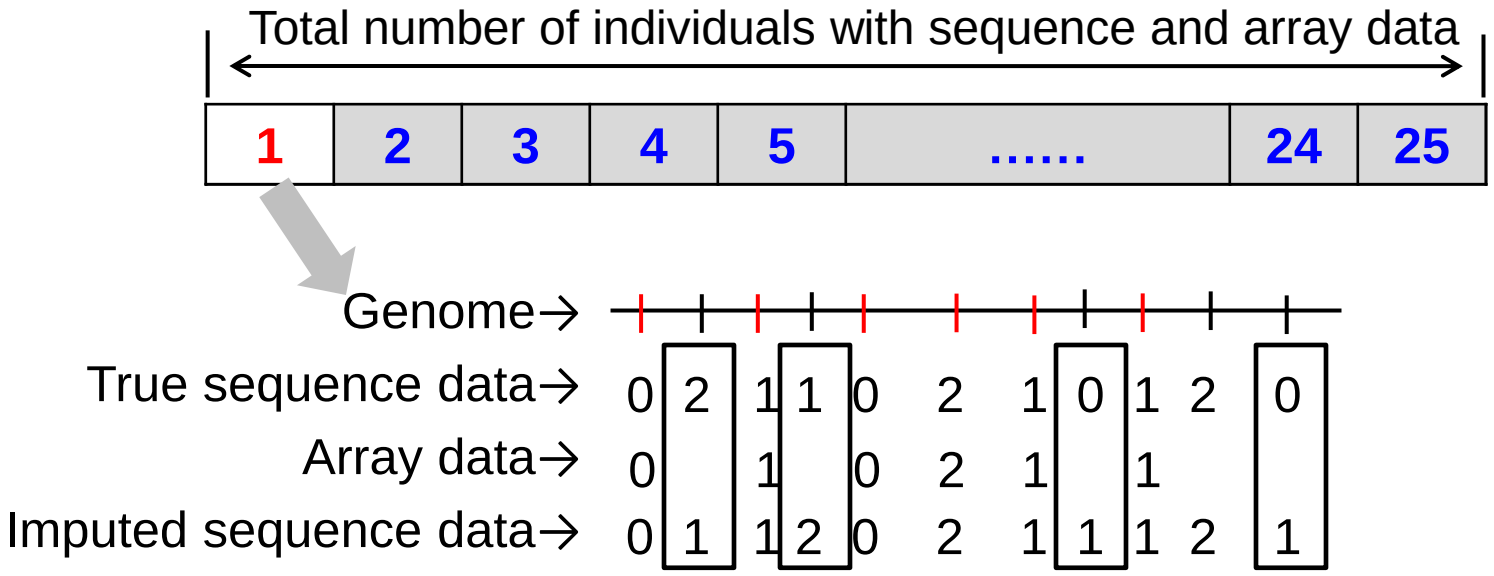
- 1st run



- Impute all other SNP genotypes for individual 1 based on information from the 24 other sequenced individuals

Leave-one-out cross-validation

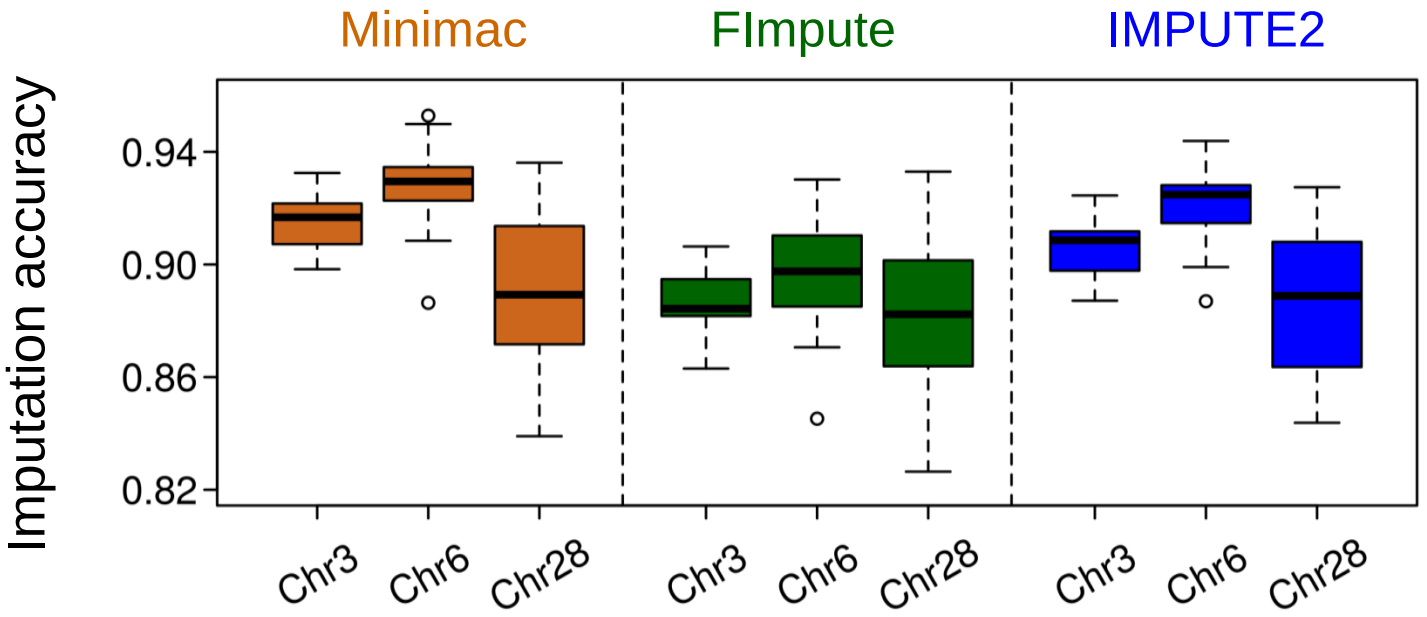
- Within sequenced individuals



- Calculation of correlation between true and imputed sequence data (except array positions)
- Repeat until each individual has been imputed once

Leave-one-out cross-validation

- Within sequenced individuals



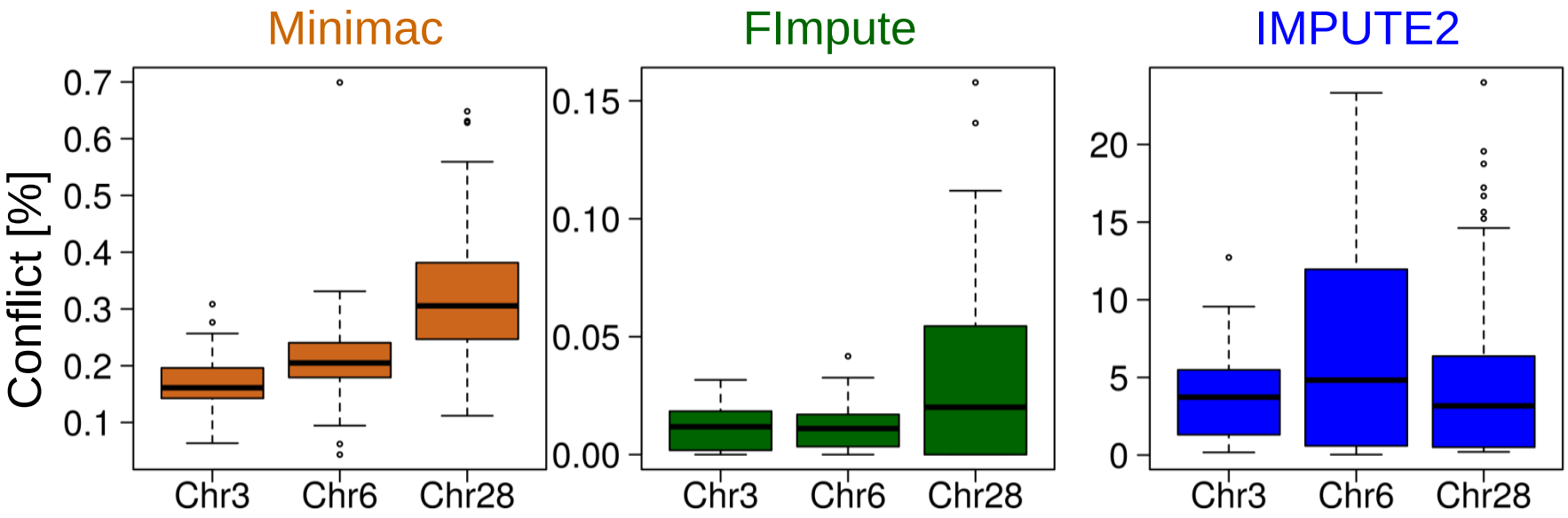
- Imputation accuracy within sequenced individuals was high (~0.9) with all imputation packages
- Performance of FImpute slightly worse than the one of Minimac and IMPUTE2

Sire-Progeny-Conflicts

- Sire-progeny pairs
  - 134 pairs with sequenced sire and genotyped progeny available (1-44 progenies/sire)
  - Comparison of sire's sequence and progeny's imputed sequence
- What must not appear due to Mendelian rules?
  - Opposite homozygous genotypes in sire-progeny pairs
- Calculation of the percentage of SNPs with sire-progeny-conflict for all sire-progeny pairs

Sire-Progeny-Conflicts

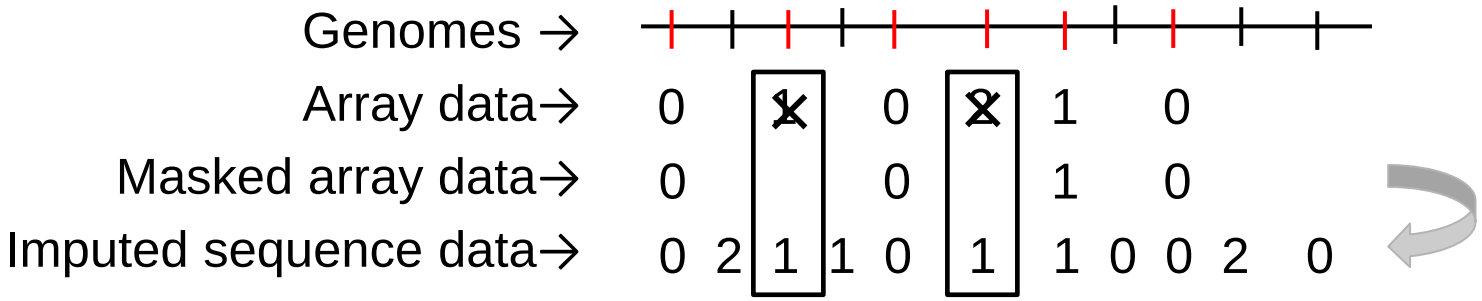
- Within sire's sequence data and progenies' imputed sequenced data



- FImpute (on average 0.01%) outperformed Minimac and IMPUTE2
- Minimac better (0.11%) than IMPUTE2 (2.5%)

Randomly masked SNPs

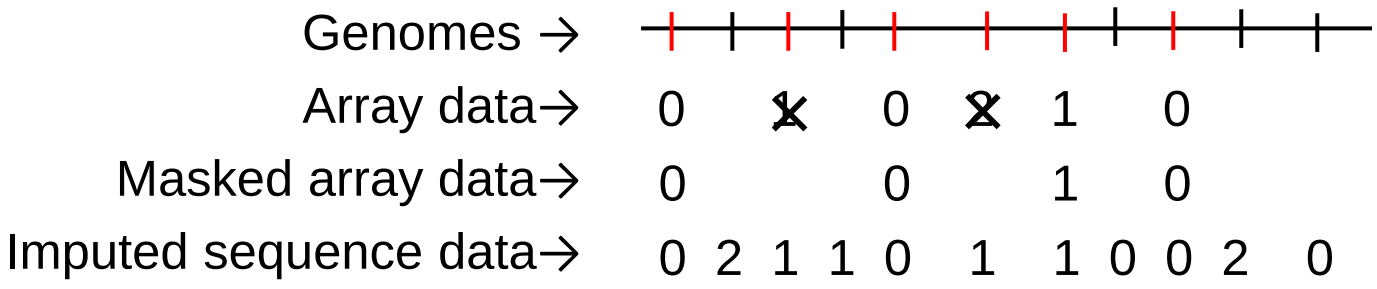
- Within 1075 genotyped individuals



- Select some SNPs in array data randomly
- Assume these SNPs to be unknown→ masked array data
- Impute up to sequence level based on information from 25 sequenced individuals
- For the masked SNPs: calculate correlation between imputed and true array data either within SNP or per individual

Randomly masked SNPs

- Within 1075 genotyped individuals



- Number of masked SNPs

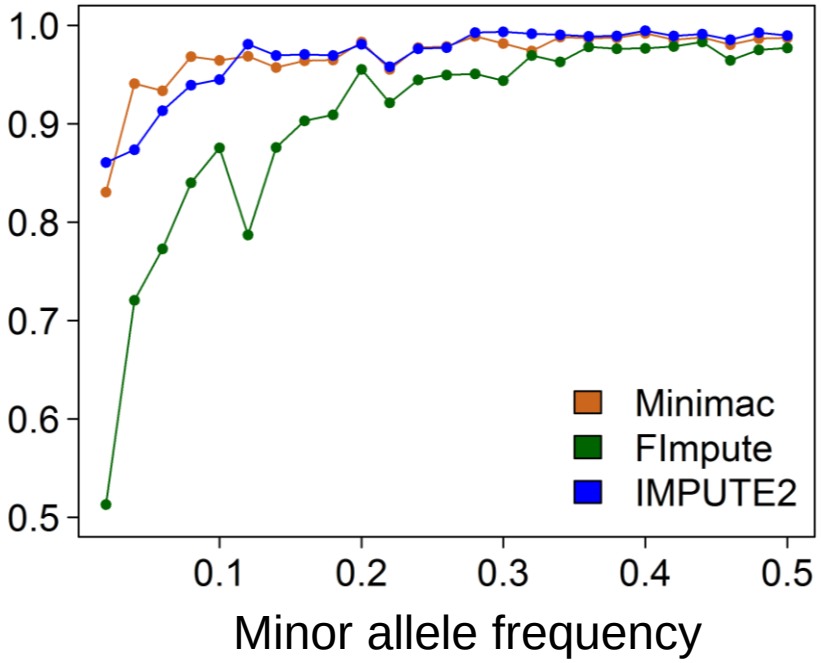
Chr. 3	Chr. 6	Chr. 28
680	270	50

- 5 replicates

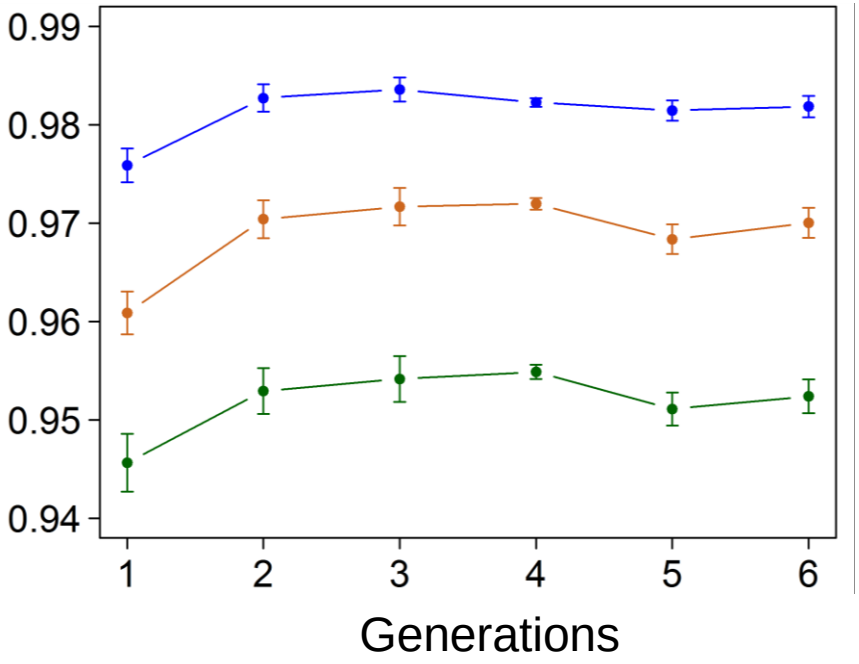
Randomly masked SNPs

- Mean of genotype correlation

per SNP:



per individual:



- Lower imputation accuracy for SNPs with low MAF, especially with FImpute
- High imputation accuracy per individual across several generations

Conclusions

- Imputation accuracy measured as correlation: Minimac and IMPUTE2 performed slightly better than FImpute
- Advantages of FImpute regarding the occurrence of Mendelian inconsistencies
- Imputation accuracy clearly lower for rare than for common SNPs
- Sequence imputation yields reasonably accuracy, even across several generations
  - From a very limited number of sequenced individuals
  - In closed breeding populations

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed - Synergistic plant and animal breeding” (Grant ID 0315528C).

G.Ni personlly thanks the China Scholarship Council for the financial support.

Thanks for your attention.

References

Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.

Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley E a, Gerstein MB, Butte AJ, Ji HP, Snyder M: **Performance comparison of whole-genome sequencing platforms.** *Nat Biotechnol* 2012, **30**:78–82.

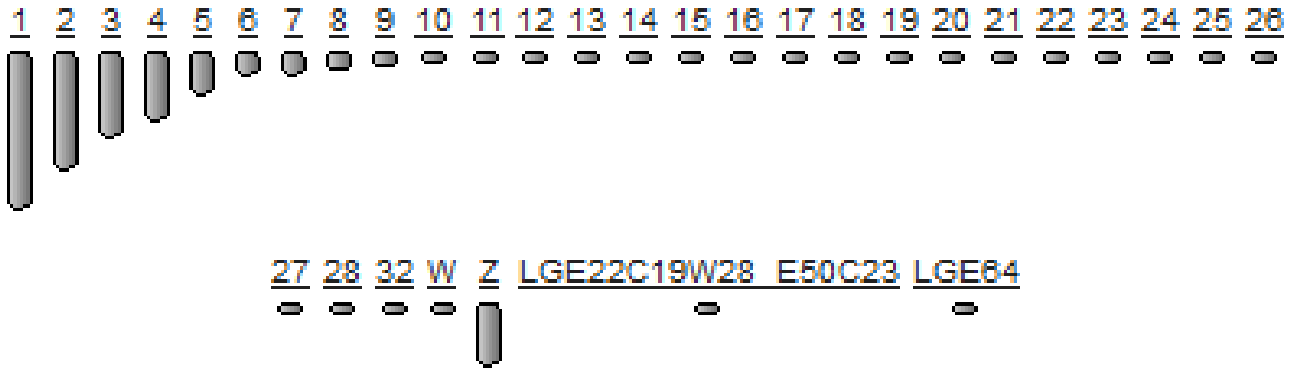
Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: **Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.** *Nat Genet* 2012, **44**:955–9.

Sargolzaei M, Chesnais JP, Schenkel FS: **A new approach for efficient genotype imputation using information from relatives.** *BMC Genomics* 2014, **15**:478.

Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.** *PLoS Genet* 2009, **5**:e1000529

Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81**:1084–97.

Add on Chicken genome

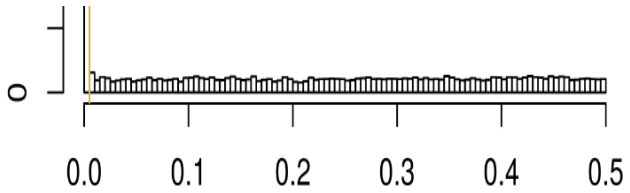


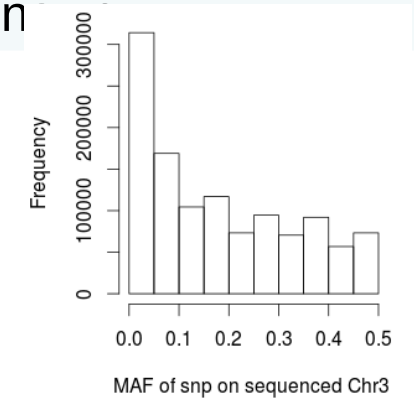
•http://www.ncbi.nlm.nih.gov/genome?term=gallus%20gallus

Add on

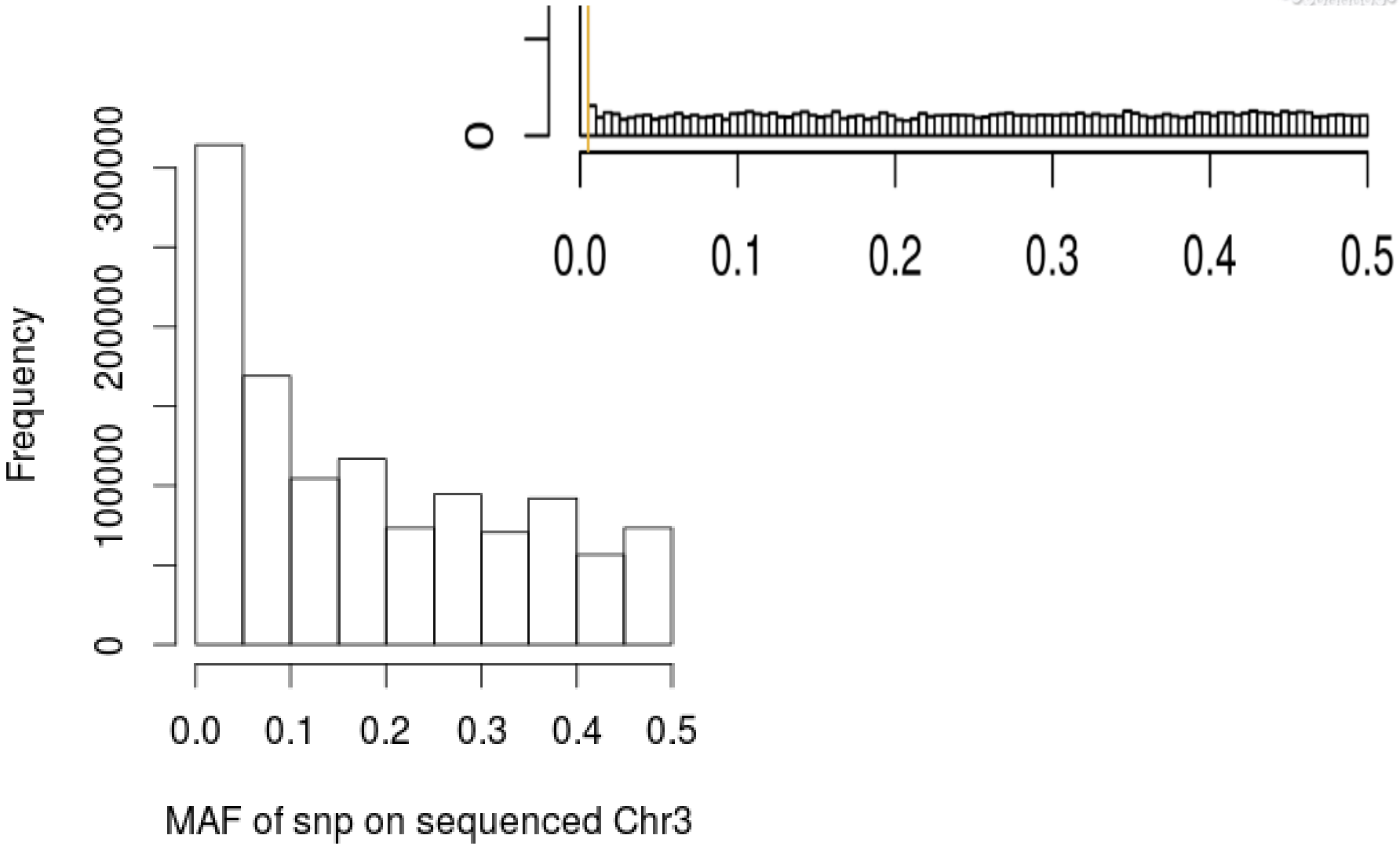
- Array data VS whole-genome sequence data

	Array data	Whole-genome sequence data
DNA variation	Only SNPs	SNPs, indels, CNVs…
Number of variations	Up to the commercial chips design	Up to aliment and detection algorithms, much more than array data
MAF of variations	Similar to Uniform distribution	Similar to gamma distribution
Costs	Relative cheap	Getting cheaper, but still expen





Add on


0
0.0
0.1
0.2
0.3
0.4
0.5
Frequency
0
100000
200000
300000
0.0
0.1
0.2
0.3
0.4
0.5
MAF of snp on sequenced Chr3

- Density of HD data and sequence data
- SNP/Kb

	Chr. 3	Chr. 6	Chr. 28
HD	0.31	0.39	0.58
Sequence	8.66	10.45	7.99