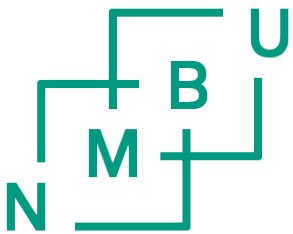
# Development of and Imputation with a SNP map derived from the latest reference genome sequence

Xijiang Yu

Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences

IHA NMBU, Ås, Norway
August, 2018



# Outlines

# Backgrounds

- High and low density chips are often used together for economic reasons
  - Missing genotypes can be imputed.
- NSG are now trying to adopt genomic selection strategy
- 4,204 Norwegian white sheep were genotyped during the past year
  - 826 genotyped with $600k$ (HD) chips
  - 3,378 genotyped with $8k$ (LD) chips

# Imputation problems

- The imputation concordance rate is only $\sim 71\%$
  - Using the genotypes and linkage maps from our genotyping company.
  - Randomly mask $< 100$ ID in the HD results
- The problem may be because of:
  - Too few LD loci $(7,327 : 606,006 \approx 1 : 82.7)$
    - Previous work: $8k \Rightarrow 15k \Rightarrow 600k$, still of $< 90\%$
  - The linkage maps may need to be upgraded.

# Why the linkage maps can be an issue?

- Different chips may be based on different versions of the reference
  - Some shared SNP are of different chromosome locations on LD and HD maps
- Sheep SNP names may be from different name systems
- Quite a few SNP duplicates

# Outlines

# My algorithm to construct such a map

- Index the reference
  - E.g., ATGCATGC $\Rightarrow$ ATGC:1,5; CATG:4; GCAT:3 TGCA:2
  - Note the indices are sorted for faster later searches.
  - Index on every 50bp sequences
- Hash all the 50bp segments into integers to save memory
- Look up the initial 50bp hash of a SNP sequence from the index
  - If found, match the rest of the sequence to confirm.
  - Each SNP sequence was searched in 8 ways.

# Other concerns

I feel thin... sort of stretched, like butter scraped over too much bread.

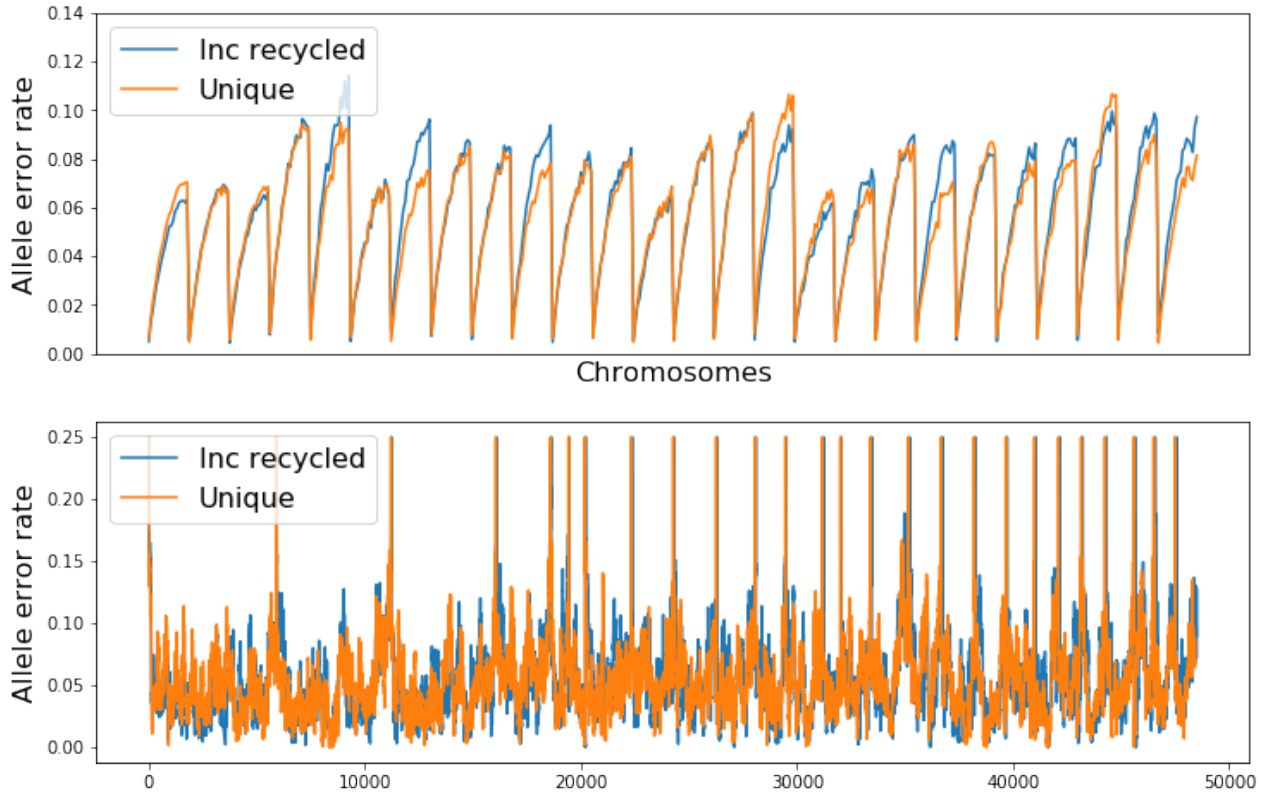
Bilbo Baggins / J.R.R. Tolkien

# Include as many shared LD loci as possible

- Using SNP flanking sequences instead of their probes
- Many sequences were matched many where in the reference
  - Recover them if possible
- After data cleaning, $\mathrm{LD}_{\mathrm{shared}} : \mathrm{HD} \approx 1 : 114.5$

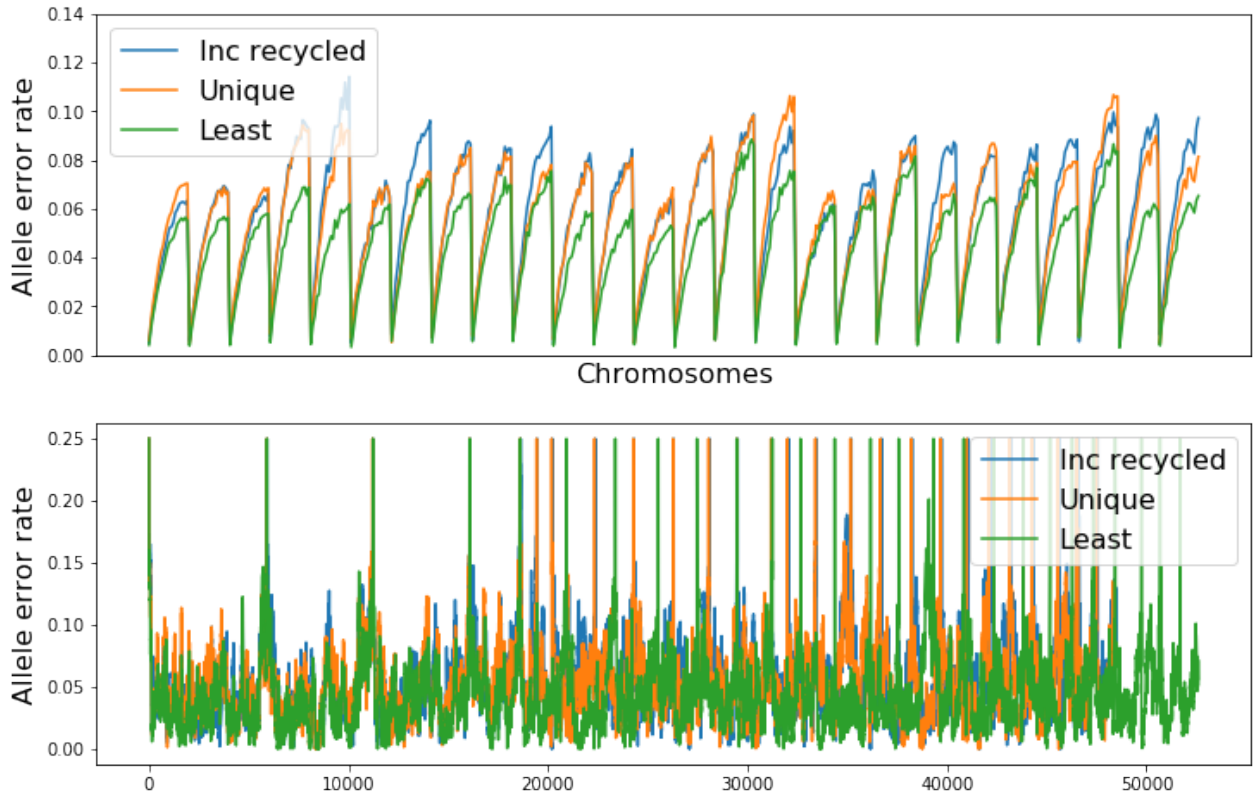
# Outlines

# Accuracy with recycled SNP




- Only recycle SNP on chromosome 1, 3, 13, 14, 16, 17, 21, 24.

# Final test imputation results vs the precious



# Conclusions

- Major
  - Concordance rate increased from 71% to 95%+ with the new map
  - A fast algorithm can finish the map within a few hours.
- Minor
  - Beagle 5 gives better results than beagle 3.3.2
  - Removing imputation results on free chromosome ends can further improve accuracy.