Metafounders

MiX99 course on genomic prediction

Matti Taskinen, Timo Pitkänen, Ismo Strandén
Natural Resources Institute Finland (Luke)

March 9, 2026

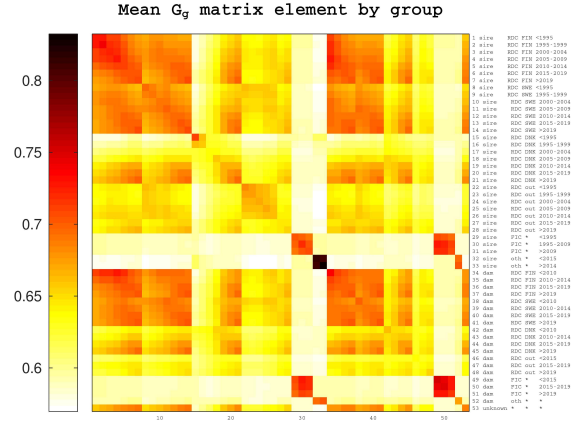
Luke
NATURAL RESOURCES INSTITUTE FINLAND

Contents

- Why metafounders are needed.
- What metafounders are.
- How to estimate metafounder relations.
- With dairy and beef population examples.

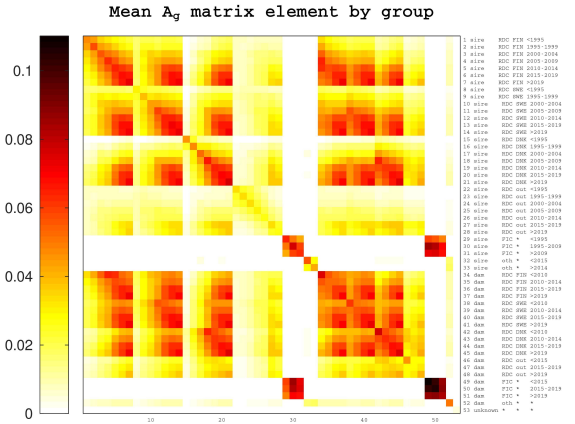
Why: Pedigree relations vs. Genomic relations

- Mean **genomic relations**.
- Nordic Red Dairy Cattle (RDC).
- Grouped in:
  - Sex.
  - Breed.
  - Country of origin.
  - Year classes.



Luke NATURAL RESOURCES INSTITUTE FINLAND

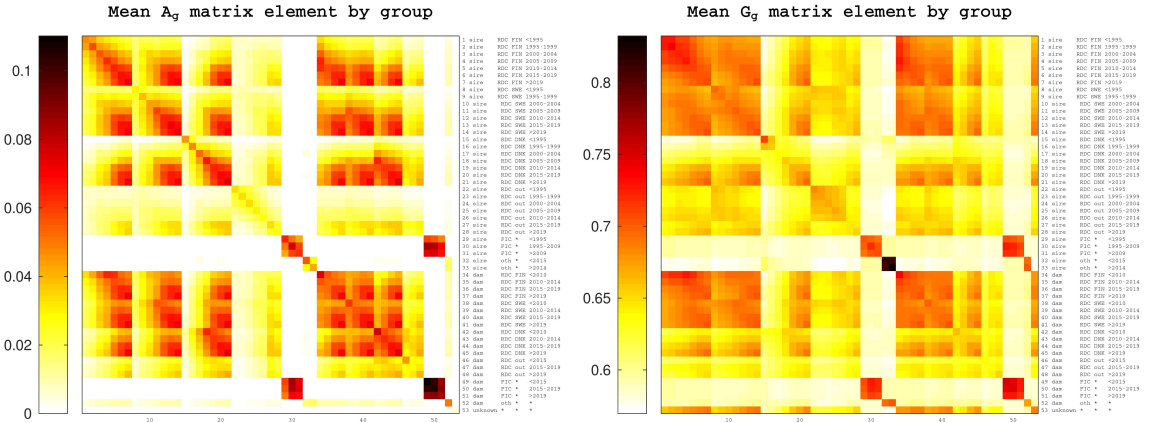
Why: Pedigree relations vs. Genomic relations



- Mean **pedigree relations**.
- Pedigree relations seem to have similar structures as the genomic relations.

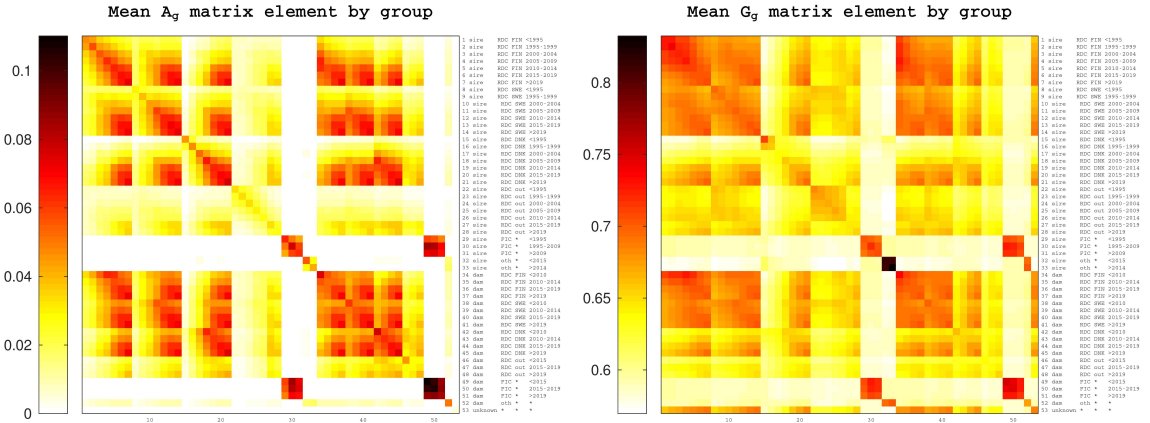
Luke NATURAL RESOURCES INSTITUTE FINLAND

Why: Pedigree relations vs. Genomic relations



- Pedigree relations (left) are much smaller than genomic relations (right).

Why: Pedigree relations vs. Genomic relations

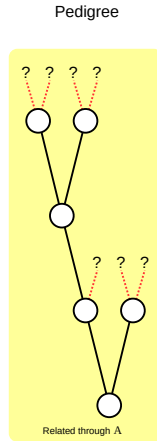


- Pedigree relations (left) are much smaller than genomic relations (right).
- Goal for metafounders: make this difference smaller.

Why: Incomplete pedigree relations

- In traditional breeding value models individuals are related through **pedigree relationship matrix** $\mathbf{A}$.

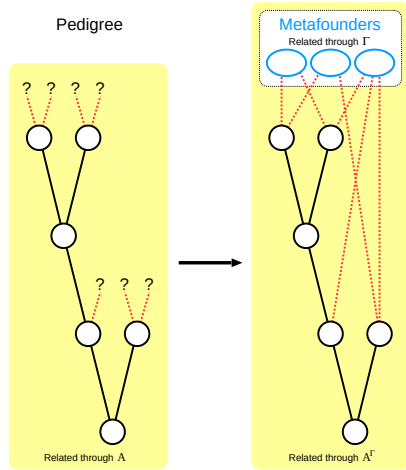
- Pedigree relationships are incomplete:
  - Some parents are unknown.
  - Oldest are assumed to be unrelated.



Why: Supplementing pedigree with genomic data

Metafounders concept:

- Unknown parents replaced with "metafounder" **ancestor groups**.
- Metafounders are assumed to be related through so called **Gamma matrix Γ**.
- Γ matrix needs to be estimated from the genomic information.



Why: Γ "corrected" pedigree relations ($\mathbf{A}^{\Gamma}$ matrix)

- Pedigree relations can be efficiently expressed using very **sparse matrices**.

Pedigree relations:

$$\mathbf{A}^{-1} = \mathbf{L}\mathbf{D}\mathbf{L}', \quad \mathbf{L} = \mathbf{I} - \frac{1}{2}\mathbf{P}$$

Why: $\mathbf{\Gamma}$ "corrected" pedigree relations ($\mathbf{A}^{\mathbf{\Gamma}}$ matrix)

- Pedigree relations can be efficiently expressed using very **sparse matrices**.
- Metafounder "corrected" pedigree relations, $\mathbf{A}^{\mathbf{\Gamma}}$ matrix, has similar very efficient form.

Pedigree relations:

$$\mathbf{A}^{-1} = \mathbf{L}\mathbf{D}\mathbf{L}', \quad \mathbf{L} = \mathbf{I} - \frac{1}{2}\mathbf{P}$$

$\mathbf{\Gamma}$ "corrected" pedigree relations:

$$(\mathbf{A}^{\mathbf{\Gamma}})^{-1} = \mathbf{L}_{\mathbf{\Gamma}} \begin{bmatrix} \mathbf{\Gamma}^{-1} & 0 \\ 0 & \mathbf{D}_{\mathbf{\Gamma}} \end{bmatrix} \mathbf{L}'_{\mathbf{\Gamma}}$$

Why: $\mathbf{\Gamma}$ "corrected" pedigree relations ($\mathbf{A}^{\mathbf{\Gamma}}$ matrix)

- Pedigree relations can be efficiently expressed using very **sparse matrices**.
- Metafounder "corrected" pedigree relations, $\mathbf{A}^{\mathbf{\Gamma}}$ matrix, has similar very efficient form.
- $\mathbf{\Gamma}$ matrix needs to be invertible.

Pedigree relations:

$$\mathbf{A}^{-1} = \mathbf{L}\mathbf{D}\mathbf{L}', \quad \mathbf{L} = \mathbf{I} - \frac{1}{2}\mathbf{P}$$

$\mathbf{\Gamma}$ "corrected" pedigree relations:

$$(\mathbf{A}^{\mathbf{\Gamma}})^{-1} = \mathbf{L}_{\mathbf{\Gamma}} \begin{bmatrix} \mathbf{\Gamma}^{-1} & 0 \\ 0 & \mathbf{D}_{\mathbf{\Gamma}} \end{bmatrix} \mathbf{L}'_{\mathbf{\Gamma}}$$

How many metafounder groups?

- Unknown parents in pedigree are **classified** to metafounder groups.
- Number of metafounder groups depends on:
  - Variation in population.
  - Amount of genomic contributions.
  - Old unknown parent groups.

Grouping examples:

- Country of origin.
- Breed.
- Sex.
- Year classes.
- . . .

How many metafounder groups?

- Unknown parents in pedigree are **classified** to metafounder groups.
- Number of metafounder groups depends on:
  - Variation in population.
  - Amount of genomic contributions.
  - Old unknown parent groups.
- Enough metafounders to separate subpopulations.

Grouping examples:

- Country of origin.
- Breed.
- Sex.
- Year classes.
- . . .

How many metafounder groups?

- Unknown parents in pedigree are **classified** to metafounder groups.
- Number of metafounder groups depends on:
  - Variation in population.
  - Amount of genomic contributions.
  - Old unknown parent groups.
- Enough metafounders to separate subpopulations.
- But metafounders may need to be distinguishable.

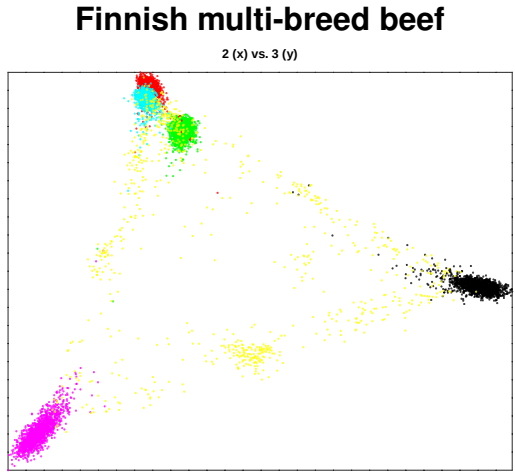
Grouping examples:

- Country of origin.
- Breed.
- Sex.
- Year classes.
- . . .

Principal Component Analysis of genomic data

Princical Component Analysis of genomic data can be used to check that:

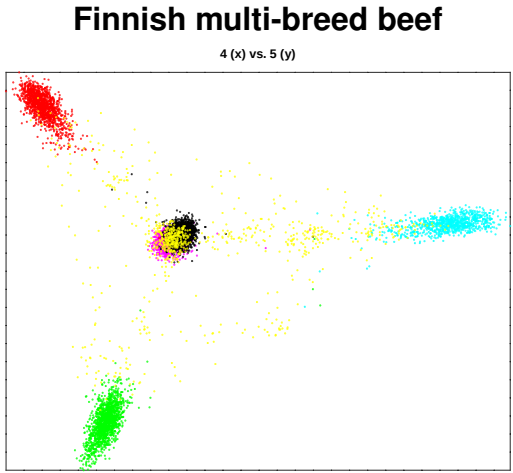
- Breeds can be separated.



Principal Component Analysis of genomic data

Princical Component Analysis of genomic data can be used to check that:

- Breeds can be separated.

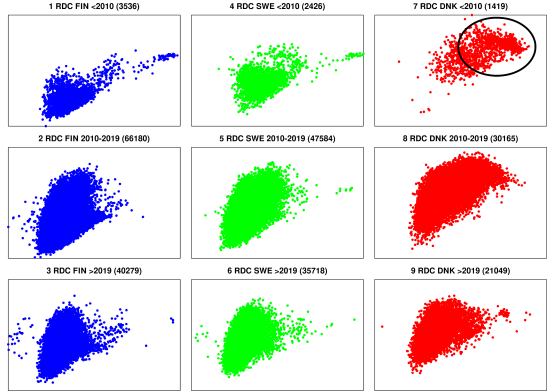


Principal Component Analysis of genomic data

Princical Component Analysis of genomic data can be used to check that:

- Breeds can be separated.
- There are differences between countries and year classes.

Nordic Red Dairy Cattle (RDC)

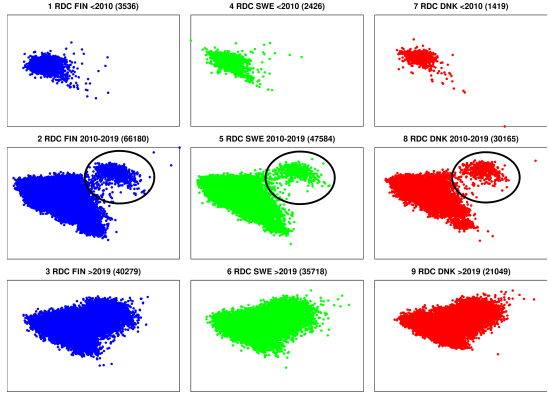


Principal Component Analysis of genomic data

Princical Component Analysis of genomic data can be used to check that:

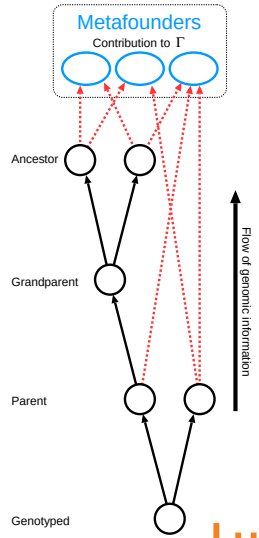
- Breeds can be separated.
- There are differences between countries and year classes.

Nordic Red Dairy Cattle (RDC)



Flow of genomic information to metafounders

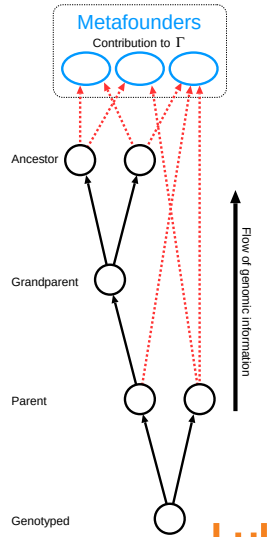
Metafounder relations estimated from pedigree using **base population allele frequencies**:



Flow of genomic information to metafounders

Metafounder relations estimated from pedigree using **base population allele frequencies**:

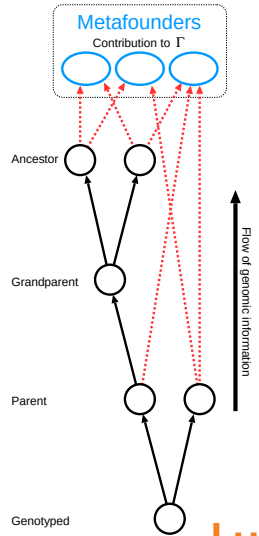
- Genomic information can be contributed from genotyped individuals to allele frequencies of their parents.



Flow of genomic information to metafounders

Metafounder relations estimated from pedigree using **base population allele frequencies**:

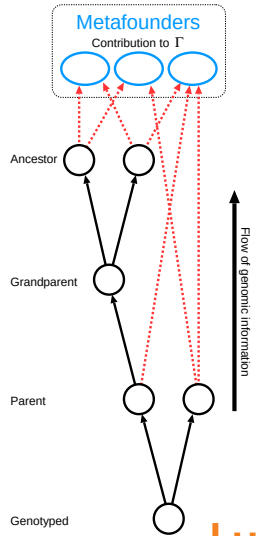
- Genomic information can be contributed from genotyped individuals to allele frequencies of their parents.
- Then from parents to grandparents and other ancestors.



Flow of genomic information to metafounders

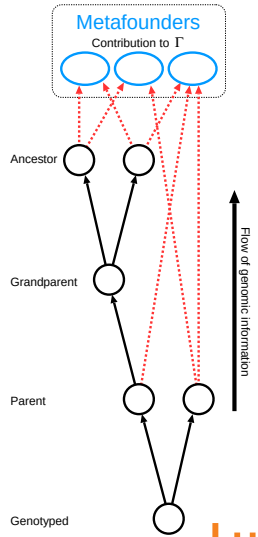
Metafounder relations estimated from pedigree using **base population allele frequencies**:

- Genomic information can be contributed from genotyped individuals to allele frequencies of their parents.
- Then from parents to grandparents and other ancestors.
- Genomic contributions are channeled to metafounders through missing parents in the pedigree.



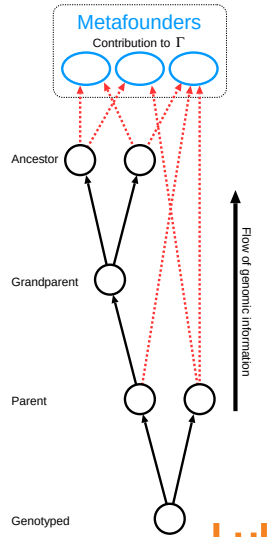
Flow of genomic information to metafounders

- Metafounder relations can be estimated from smaller pedigree of genotyped and their ancestors.



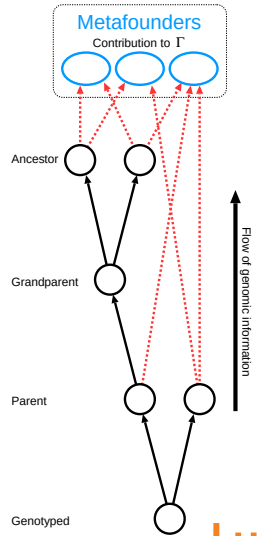
Flow of genomic information to metafounders

- Metafounder relations can be estimated from smaller pedigree of genotyped and their ancestors.
- Enough missing parents and genomic contributions for each metafounder:



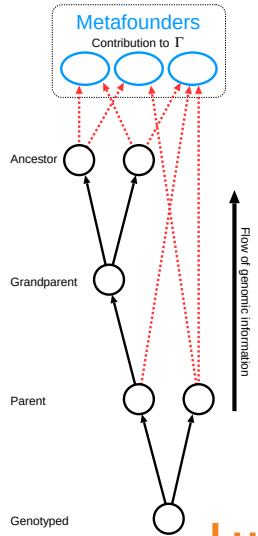
Flow of genomic information to metafounders

- Metafounder relations can be estimated from smaller pedigree of genotyped and their ancestors.
- Enough missing parents and genomic contributions for each metafounder:
  - ► No missing parents $\Rightarrow$ no contributions.



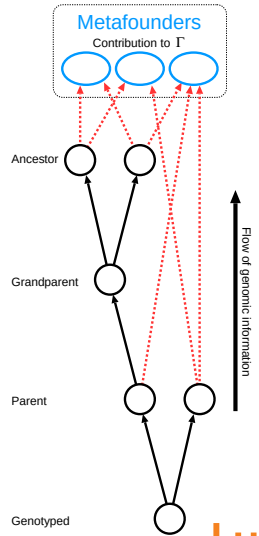
Flow of genomic information to metafounders

- Metafounder relations can be estimated from smaller pedigree of genotyped and their ancestors.
- Enough missing parents and genomic contributions for each metafounder:
  - No missing parents $\Rightarrow$ no contributions.
  - Evaluation pedigree may have missing parents even if none in $\mathbf{\Gamma}$ estimation pedigree!

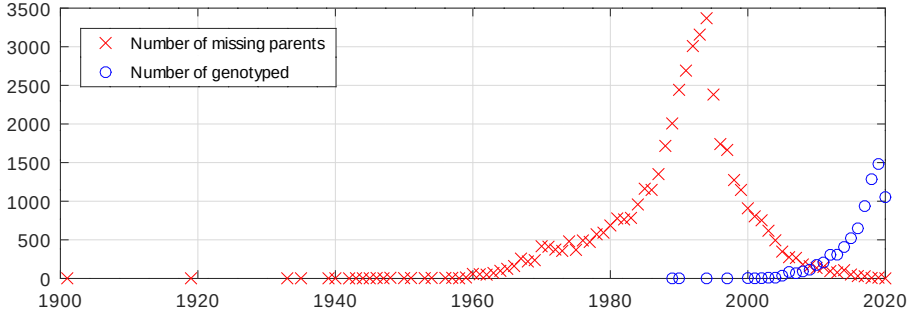


Flow of genomic information to metafounders

- Metafounder relations can be estimated from smaller pedigree of genotyped and their ancestors.
- Enough missing parents and genomic contributions for each metafounder:
  - No missing parents $\Rightarrow$ no contributions.
  - Evaluation pedigree may have missing parents even if none in $\mathbf{\Gamma}$ estimation pedigree!
- Metafounders get genomic contributions also from other group's genotypes:
  - Important for older metafounder groups.

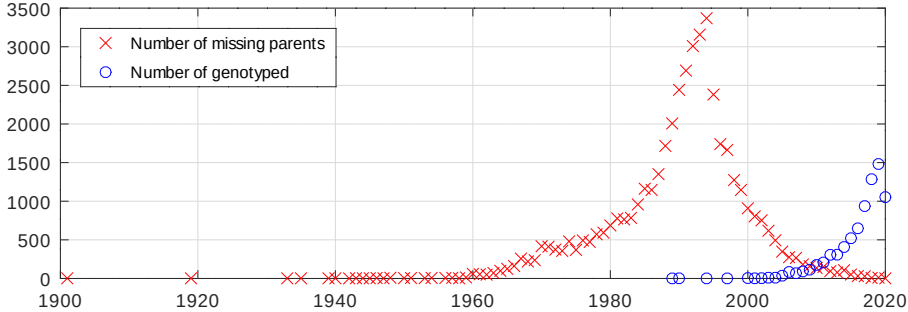


Flow of genomic information to metafounders



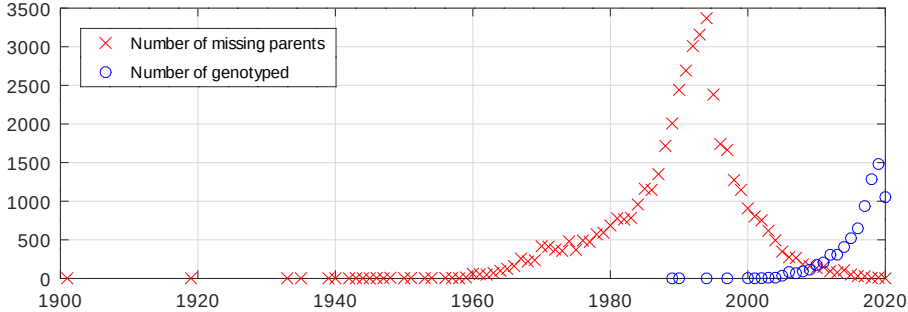
- Genotyped animals (blue) are mostly from last few years.

Flow of genomic information to metafounders



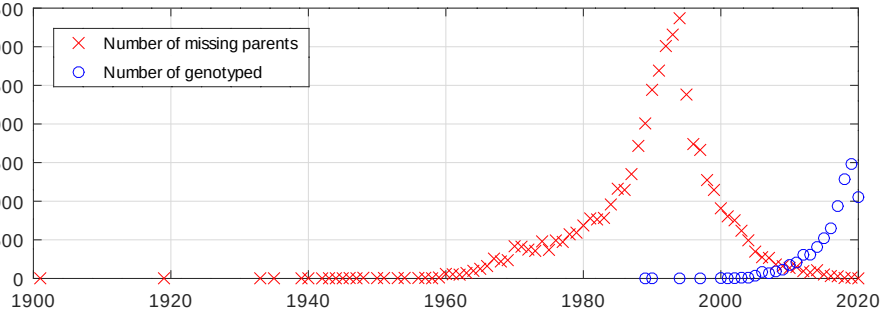
- Genotyped animals (blue) are mostly from last few years.
- Missing parents (red) in pedigree can be quite far in past.

Flow of genomic information to metafounders



- Genotyped animals (blue) are mostly from last few years.
- Missing parents (red) in pedigree can be quite far in past.
- Metafounder relations need to be estimated 30 years ago.

Flow of genomic information to metafounders



- Genotyped animals (blue) are mostly from last few years.
- Missing parents (red) in pedigree can be quite far in past.
- Metafounder relations need to be estimated 30 years ago.
- Note: Avoid **pruning** pedigree to certain depth.

Estimating metafounder relations from genotypes

Estimating metafounder relations:

- MiX99 companion utility: `Bpop`

```
# Calculate base population allele frequencies:
LOG="$(OUTPUTFILE).log"
echo                                                                    >&2
echo "Base polulation allele frequencies (Bpop):"                       >&2
echo " - Pedigree   file: $(basename $PEDIGREE)"                        >&2
echo " - Inbreeding file: $(basename $INBREEDING)"                      >&2
echo " - Genotype   file: $(basename $GENOFILE)"                        >&2
echo " - Output     file: $(basename $OUTPUTFILE)"                      >&2
echo " - Log        file: $(basename $LOG)"                             >&2
$(SCRIPT)/Bpop -CHM -nthr 4 -nospace 12 -F "$INBREEDING" -groups "$NGROUPS" \
       -a "$OUTPUTFILE" "$PEDIGREE" "$GENOFILE" >&2 > "$LOG"
```

Estimating metafounder relations from genotypes

Estimating metafounder relations:

- MiX99 companion utility: `Bpop`
- Pedigree: missing parents replaced with negative metafounder groups.

```
# Calculate base population allele frequencies:
LOG="$(OUTPUTFILE).log"
echo
echo "Base polulation allele frequencies (Bpop):"                    >&2
echo " - Pedigree   file: $(basename $PEDIGREE)"                     >&2
echo " - Inbreeding file: $(basename $INBREEDING)"                   >&2
echo " - Genotype   file: $(basename $GENOFILE)"                     >&2
echo " - Output     file: $(basename $OUTPUTFILE)"                   >&2
echo " - Log        file: $(basename $LOG)"                          >&2
$(SCRIPT)/Bpop -CHM -nthr 4 -nospace 12 -F "$INBREEDING" -groups "$NGROUPS" \
       -a "$OUTPUTFILE" "$PEDIGREE" "$GENOFILE" >&2 > "$LOG"
```

Estimating metafounder relations from genotypes

Estimating metafounder relations:

- MiX99 companion utility: `Bpop`
- Pedigree: missing parents replaced with negative metafounder groups.
- $\mathbf{\Gamma}$ matrix (or inverse): option `-gamma` (`-igamma`).

```
# Calculate base population allele frequencies:
LOG="$(OUTPUTFILE).log"
echo
echo "Base polulation allele frequencies (Bpop):"                    >&2
echo " - Pedigree   file: $(basename $PEDIGREE)"                     >&2
echo " - Inbreeding file: $(basename $INBREEDING)"                   >&2
echo " - Genotype   file: $(basename $GENOFILE)"                     >&2
echo " - Output     file: $(basename $OUTPUTFILE)"                   >&2
echo " - Log        file: $(basename $LOG)"                          >&2
$(SCRIPT)/Bpop -CHM -nthr 4 -nospace 12 -F "$INBREEDING" -groups "$NGROUPS" \
        -a "$OUTPUTFILE" "$PEDIGREE" "$GENOFILE" >&2 > "$LOG"
```

Estimating metafounder relations from genotypes

Estimating metafounder relations:

- MiX99 companion utility: `Bpop`
- Pedigree: missing parents replaced with negative metafounder groups.
- $\mathbf{\Gamma}$ matrix (or inverse): option `-gamma` (`-igamma`).
- Or from base population allele frequencies: option `-a`.

```
# Calculate base population allele frequencies:
LOG="$(OUTPUTFILE).log"
echo                                                                    >&2
echo "Base polulation allele frequencies (Bpop):"                       >&2
echo "  - Pedigree   file: $(basename $PEDIGREE)"                       >&2
echo "  - Inbreeding file: $(basename $INBREEDING)"                     >&2
echo "  - Genotype   file: $(basename $GENOFILE)"                       >&2
echo "  - Output     file: $(basename $OUTPUTFILE)"                     >&2
echo "  - Log        file: $(basename $LOG)"                            >&2
$(SCRIPT)/Bpop -CHM -nthr 4 -nospace 12 -F "$INBREEDING" -groups "$NGROUPS" \
        -a "$OUTPUTFILE" "$PEDIGREE" "$GENOFILE" >&2 > "$LOG"
```

Allele frequencies $\mathbf{F}$:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 & \cdots & \mathbf{F}_n \end{bmatrix}$$

Calculating $\mathbf{\Gamma}$ matrix from $\mathbf{F}$:

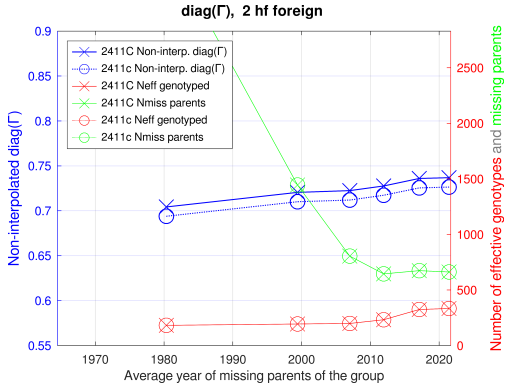
$$\mathbf{\Gamma} = 8 * \mathrm{cov}(\mathbf{F})$$

or

$$\mathbf{\Gamma} = \frac{8}{m} * (\mathbf{F} - \tfrac{1}{2})'(\mathbf{F} - \tfrac{1}{2})$$

Estimating metafounder relations from genotypes

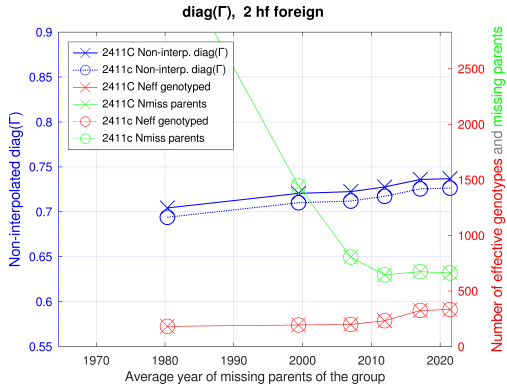
Ensuring good coverage of genomic contributions:



Estimating metafounder relations from genotypes

Ensuring good coverage of genomic contributions:

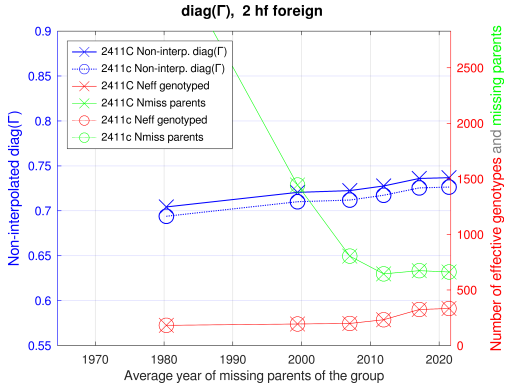
- Number of missing parents (green).



Estimating metafounder relations from genotypes

Ensuring good coverage of genomic contributions:

- Number of missing parents (green).
- Effective genomic contributions (red).

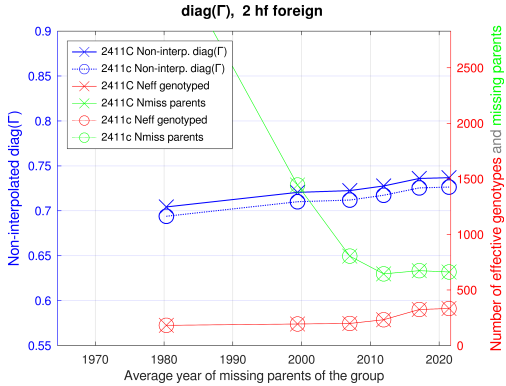



```
Number of genotyped animals,                    n : 15744
Effective number of genotyped animals,      n_m :  9524 9458
Effective number of genotyped contributions, f:  831.7358 175.9585
Proportion of effective N genotyped,      n_m/n :    60.49%   60.07%
Measure of direct genotyped,              f/n_m :     8.73%    1.86%
```

Estimating metafounder relations from genotypes

Ensuring good coverage of genomic contributions:

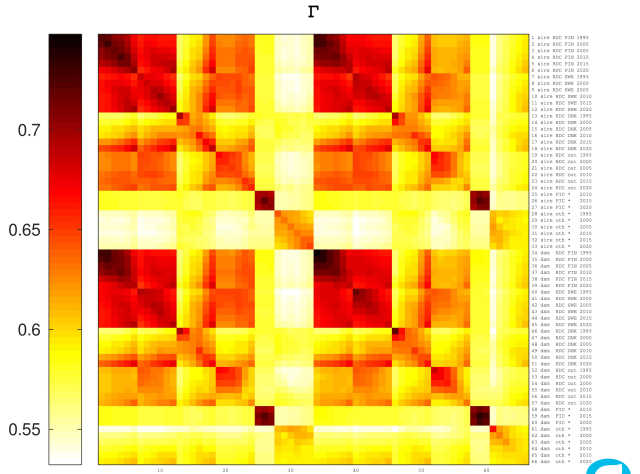
- Number of missing parents (green).
- Effective genomic contributions (red).
- Stability of Γ matrix values (blue).




```
Number of genotyped animals,                    n : 15744
Effective number of genotyped animals,     n_m :  9524 9458
Effective number of genotyped contributions, f:  831.7358 175.9585
Proportion of effective N genotyped,     n_m/n :    60.49%   60.07%
Measure of direct genotyped,             f/n_m :     8.73%    1.86%
```

Estimated Γ matrix visualized using colors

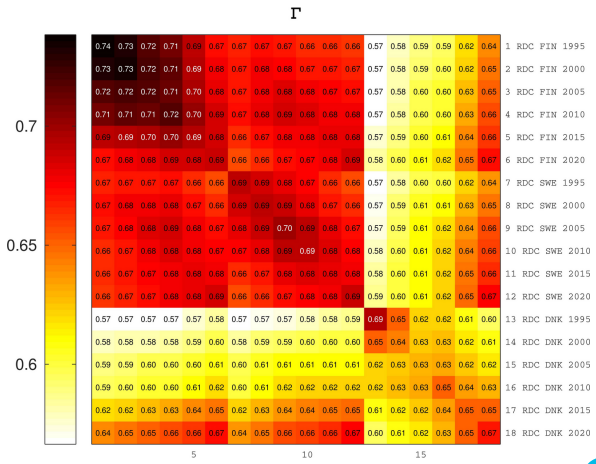
- Nordic RDC.
- Metafounders needed to match old unknown parent groups.
- 66 metafounders:
  - Sires and dams.
  - RDC, Finncattle, and other breeds.
  - 3 Nordic countries and outside.
  - Year classes.



Luke NATURAL RESOURCES INSTITUTE FINLAND

Estimated Γ matrix shows population dynamics

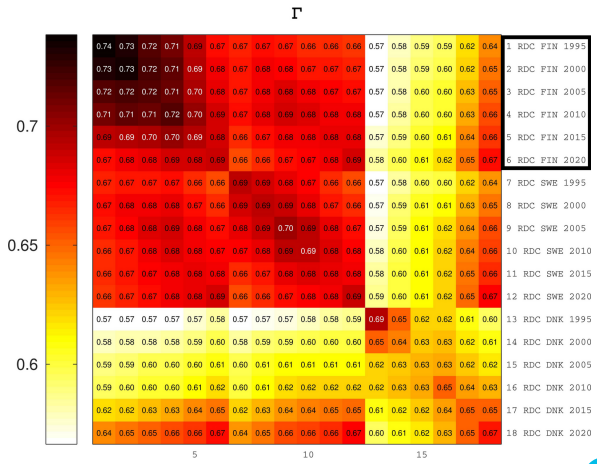
Estimated Γ matrix (partial):



Estimated Γ matrix shows population dynamics

Estimated $\mathbf{\Gamma}$ matrix (partial):

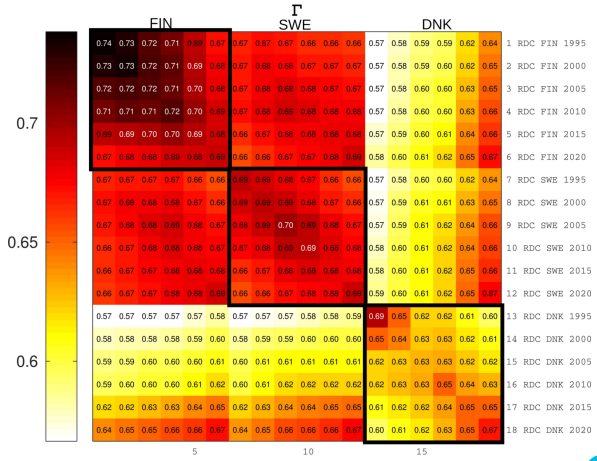
- 6 year classes.



Estimated Γ matrix shows population dynamics

Estimated Γ matrix (partial):

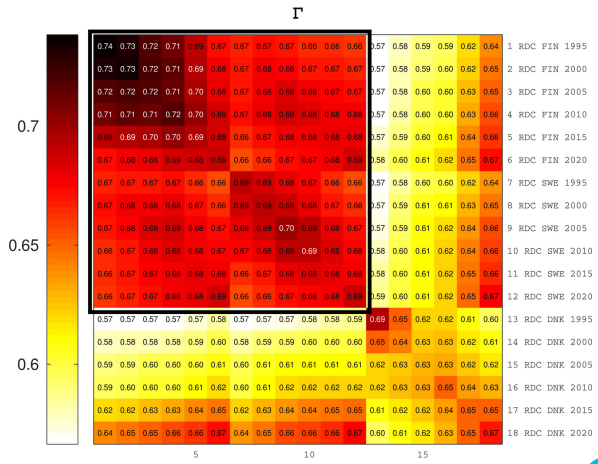
- 6 year classes.
- FIN, SWE, and DNK:



Estimated Γ matrix shows population dynamics

Estimated Γ matrix (partial):

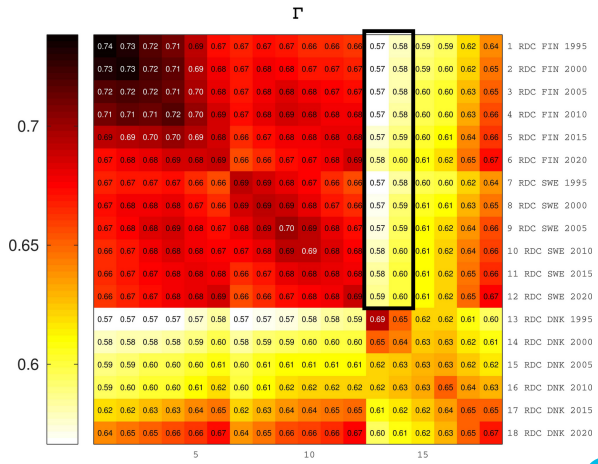
- 6 year classes.
- FIN, SWE, and DNK:
  - FIN and SWE more closely related.



Estimated Γ matrix shows population dynamics

Estimated Γ matrix (partial):

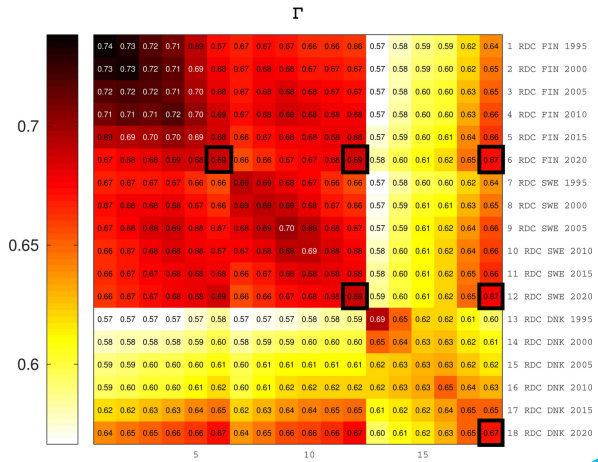
- 6 year classes.
- FIN, SWE, and DNK:
  - ► FIN and SWE more closely related.
- Older DNK less related to FIN/SWE.



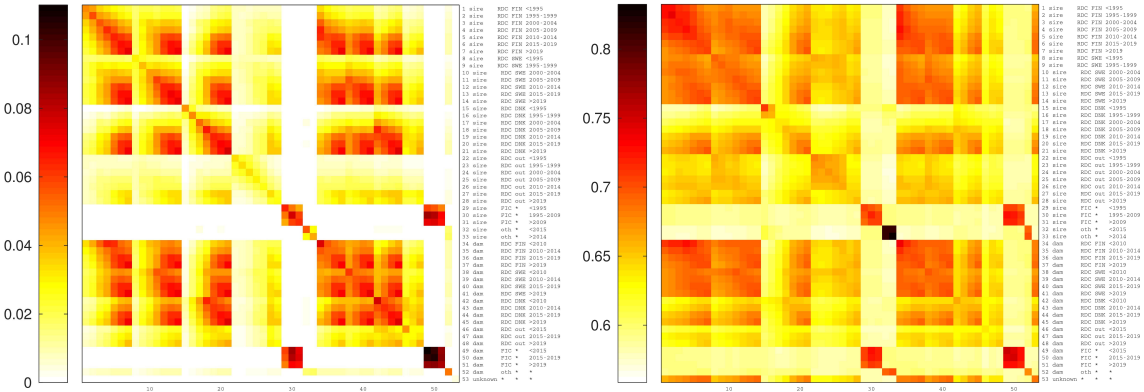
Estimated Γ matrix shows population dynamics

Estimated Γ matrix (partial):

- 6 year classes.
- FIN, SWE, and DNK:
  - ► FIN and SWE more closely related.
- Older DNK less related to FIN/SWE.
- Newest year classes more closely related in all 3e countries.

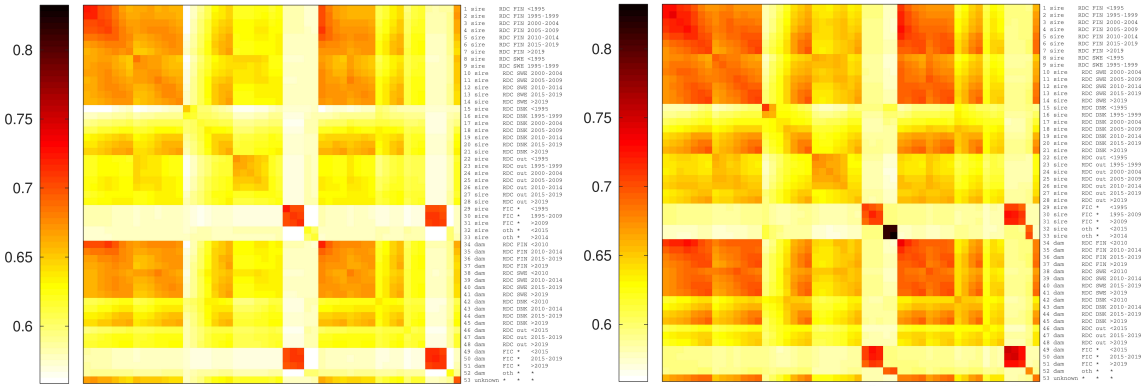


Γ "corrected" pedigree vs. genomic relations



- Average pedigree relations ($\mathbf{A}$, left) were much smaller than average genomic relations (right).

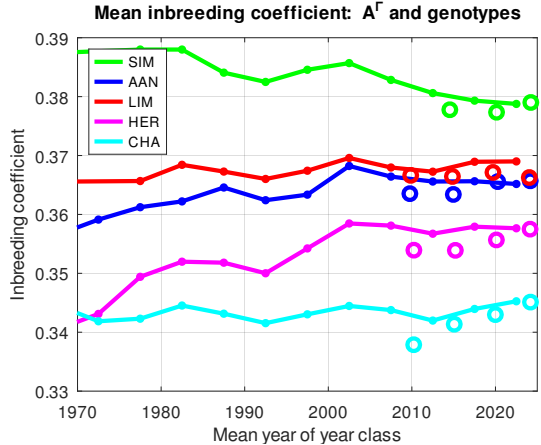
Γ "corrected" pedigree vs. genomic relations



- With metafounders average pedigree relations ($\mathbf{A}^{\mathbf{\Gamma}}$, left) are much closer to average genomic relations (right).

Γ "corrected" pedigree vs. genomic inbreeding

- Finnish multi-breed beef.
- Comparing average **inbreeding coefficients**:
  - Metafounder "corrected" pedigree inbreeding (lines).
  - Genomic inbreeding (circles).
- With metafounders average inbreeding coefficients are close to genomic coefficients.



Matti Taskinen, Timo Pitkänen, Ismo Strandén


Luke
NATURAL RESOURCES
INSTITUTE FINLAND