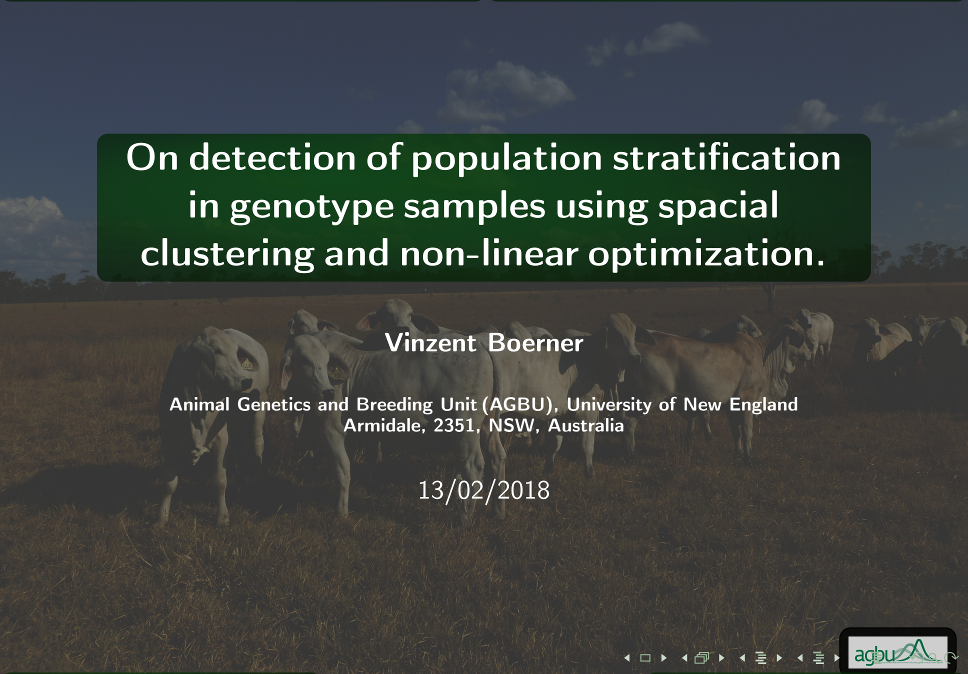


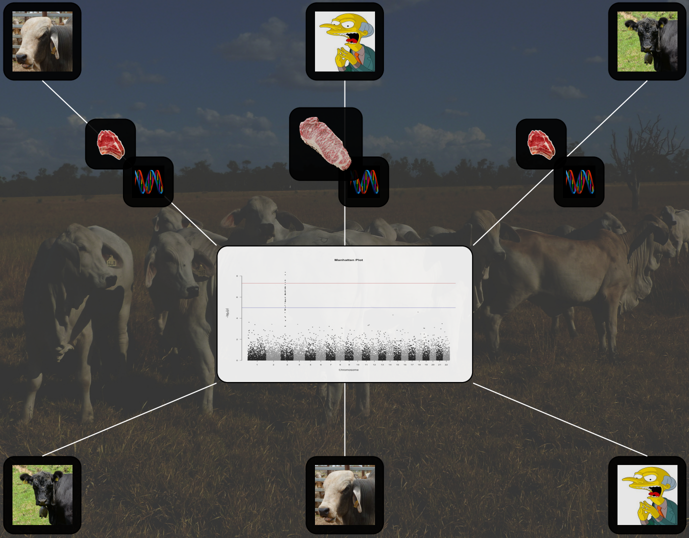
On detection of population stratification in genotype samples using spacial clustering and non-linear optimization.

Vinzent Boerner

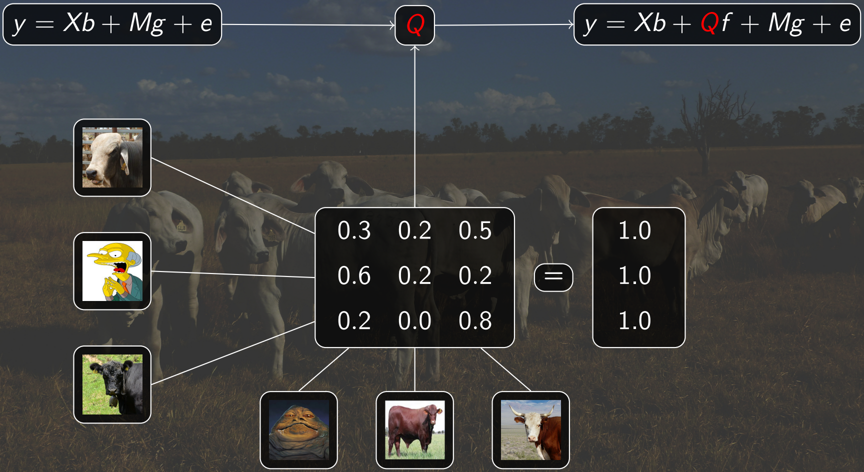
Animal Genetics and Breeding Unit (AGBU), University of New England
Armidale, 2351, NSW, Australia

13/02/2018

Why searching for population stratification



How to account for it??

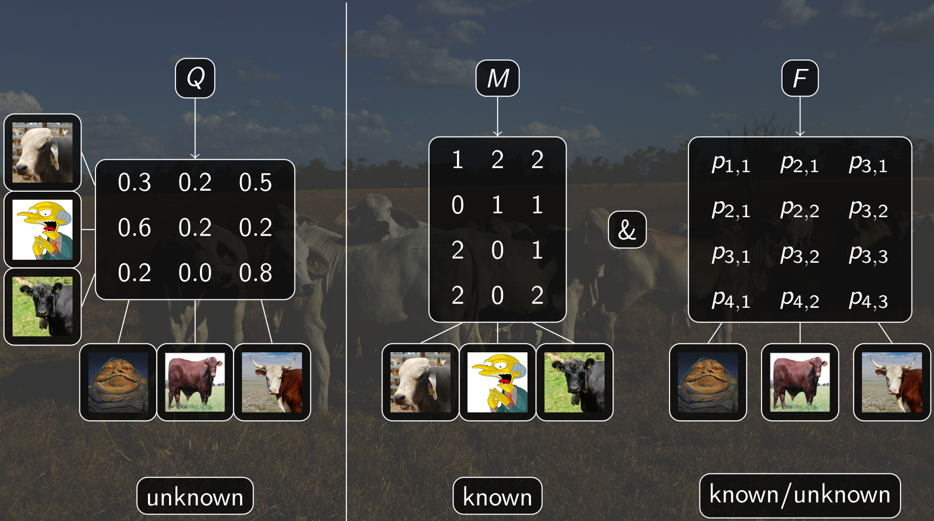


$$y = Xb + Mg + e \longrightarrow Q \longrightarrow y = Xb + Qf + Mg + e$$

$$\begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.0 \\ 1.0 \end{bmatrix}$$

3 putative founder populations

How to get Q??



Q

$$\begin{matrix} 0.3 & 0.2 & 0.5 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{matrix}$$

unknown

M

$$\begin{matrix} 1 & 2 & 2 \\ 0 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 2 \end{matrix}$$

known

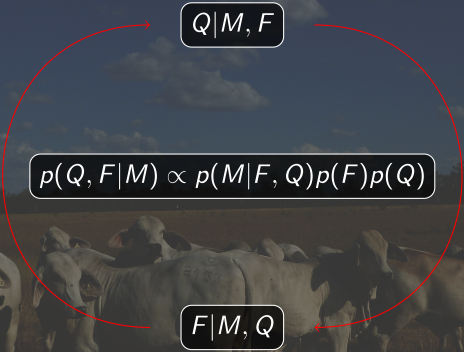
&

F

$$\begin{matrix} p_{1,1} & p_{2,1} & p_{3,1} \\ p_{2,1} & p_{2,2} & p_{3,2} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{matrix}$$

known/unknown

How to get Q if F is unknown: The Loop



$Q|M, F$

$$p(Q, F|M) \propto p(M|F, Q)p(F)p(Q)$$

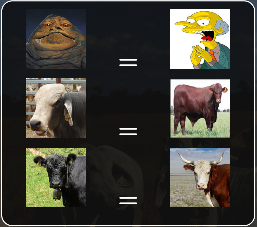
$F|M, Q$

Rotate column vectors in F through R^N until all points presented by columns in M are explained best

ADMIXTURE, FRAPPE, STRUCTURE

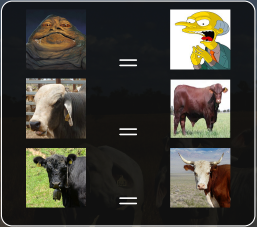
How to get Q if F is unknown: The 2-Step Cascade

Step 1



How to get Q if F is unknown: The 2-Step Cascade

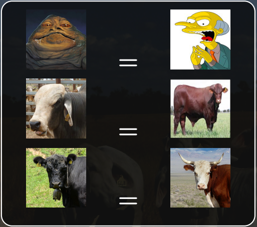
Step 1



point aggregations in R^N=>clouds

How to get Q if F is unknown: The 2-Step Cascade

Step 1

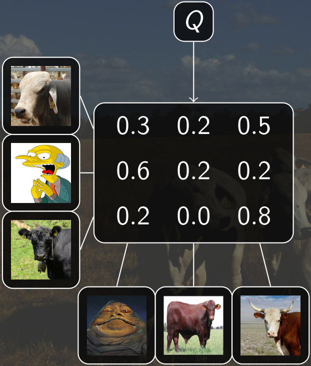


point aggregations in R^N=>clouds

F **from clouds**

$$\begin{array}{ccc} p_{1,1} & p_{2,1} & p_{3,1} \\ p_{2,1} & p_{2,2} & p_{3,2} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{array}$$

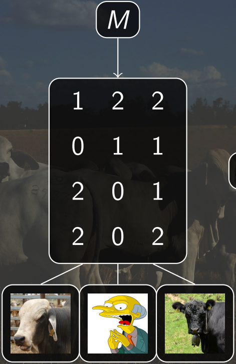
How to get Q if F is unknown: The 2-Step Cascade



Q

$$\begin{matrix} 0.3 & 0.2 & 0.5 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{matrix}$$

unknown

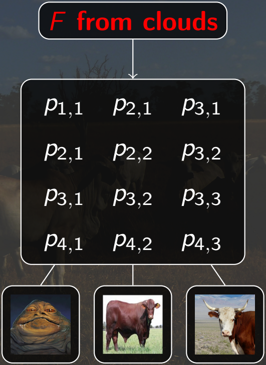


M

$$\begin{matrix} 1 & 2 & 2 \\ 0 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 2 \end{matrix}$$

known

&

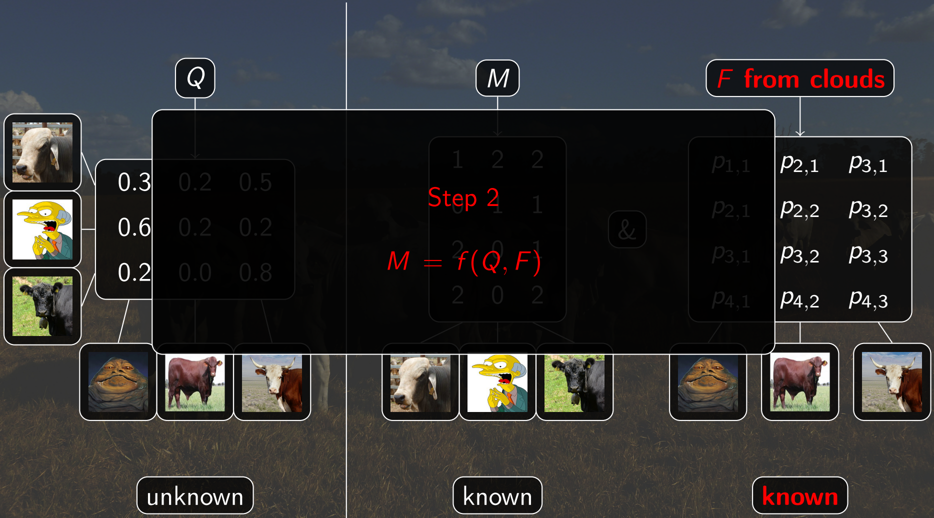


F from clouds

$$\begin{matrix} p_{1,1} & p_{2,1} & p_{3,1} \\ p_{2,1} & p_{2,2} & p_{3,2} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{matrix}$$

known

How to get Q if F is unknown: The 2-Step Cascade



Q

M

F from clouds

$$\begin{matrix} 0.3 & 0.2 & 0.5 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{matrix}$$

$$\begin{matrix} 1 & 2 & 2 \\ 0 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 2 \end{matrix}$$

&

$$\begin{matrix} p_{1,1} & p_{2,1} & p_{3,1} \\ p_{2,1} & p_{2,2} & p_{3,2} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{matrix}$$

Step 2

$$M = f(Q, F)$$

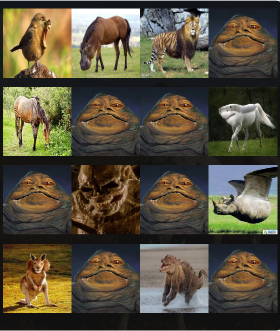
unknown

known

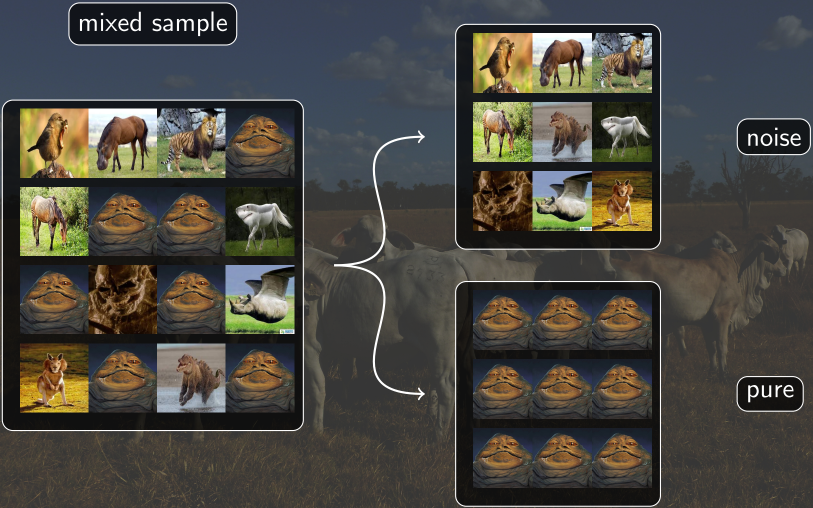
known

Step 1: Cloud detection by cluster analysis

mixed sample



Step 1: Cloud detection by cluster analysis

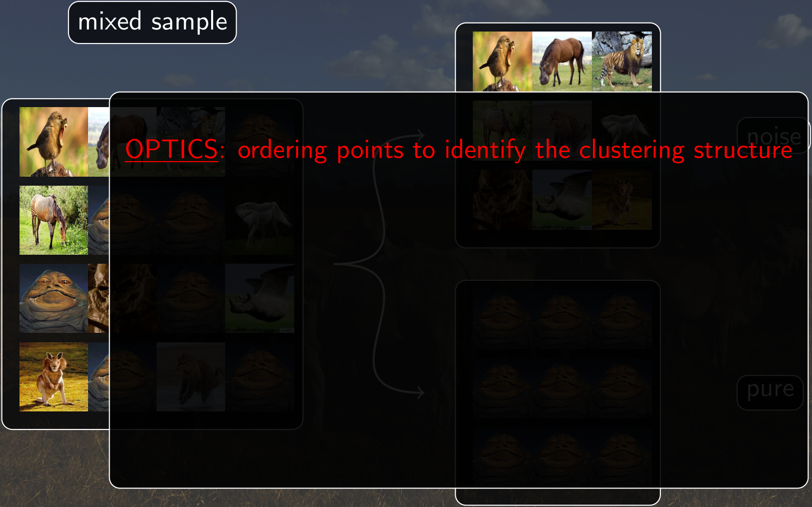


mixed sample

noise

pure

Step 1: Cloud detection by cluster analysis



OPTICS: ordering points to identify the clustering structure

Step 1: Cloud detection by cluster analysis

mixed sample

OPTICS: ordering points to identify the clustering structure

ingredients:

- distance matrix D
  - $M = U\Sigma V$
  - $D = \sqrt{\Sigma V_{,1:K} V'_{,1:K} \Sigma}$
- minimum number of neighbouring points to establish a cluster => prior knowledge

Step 2: genome composition by a linear model

$$M = FQ' + E$$

=>

$$min(E'_{(:,i)}E_{(:,i)})$$

subject to

$$\sum_i Q_{(i,:)} = 1$$

$$Q_{(i,j)} \geq 0$$

- E: SNP×animals matrix of non-explainable residual
- constraints require non-linear optimisation solver

Step 2: genome composition by a linear model

$$M = FQ' + E \quad => \quad min(E'_{(:,i)}E_{(:,i)})$$

subject to

$$\sum_i Q_{(i,:)} = 1$$

$$Q_{(i,j)} \geq 0$$

- E: SNP×animals matrix of non-explainable residual
- constraints require non-linear optimisation solver

Constrained Genomic Regression (CGR, aka “BREEDCOMP”)

- non-linear optimisation solver

global augmented Lagrangian

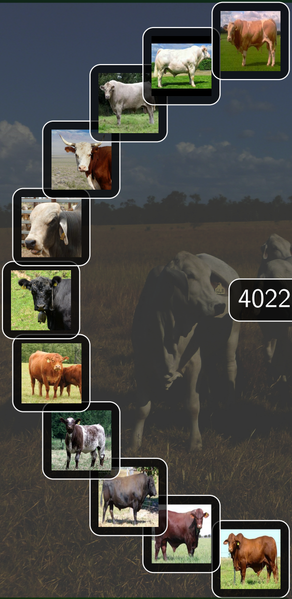
local method of moving asymptotes

Data

Data

Simmental	337
Charolais	899
Murray Grey	316
Hereford	1,500
Angus	1,473
Limousin	1,395
Shorthorn	1,126
Wagyu	1,497
Santa Gertrudis	1,474
Droughtmaster	130
Brahman	1,492
	11,639

Data



4022 SNP common across many panels

Data

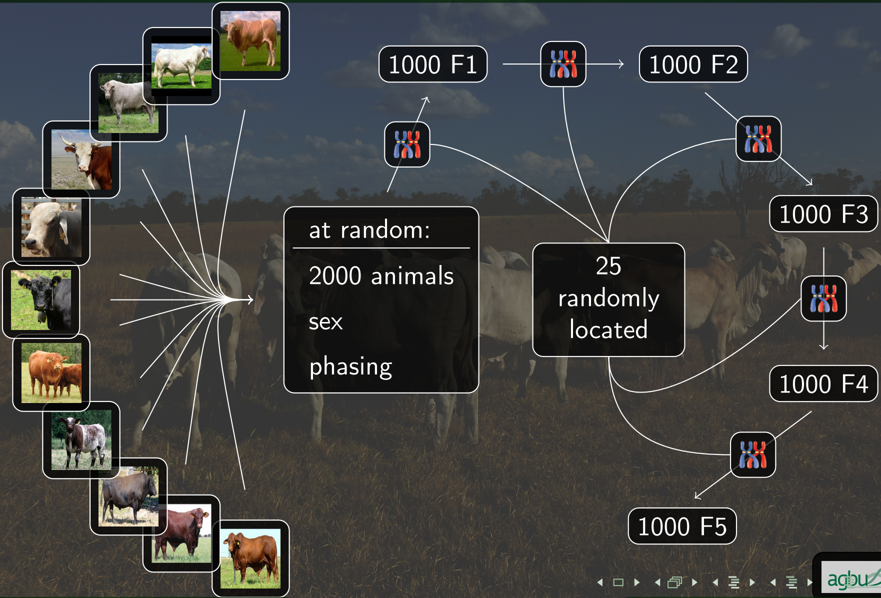
at random:

2000 animals

sex

phasing

Data



at random:

2000 animals

sex

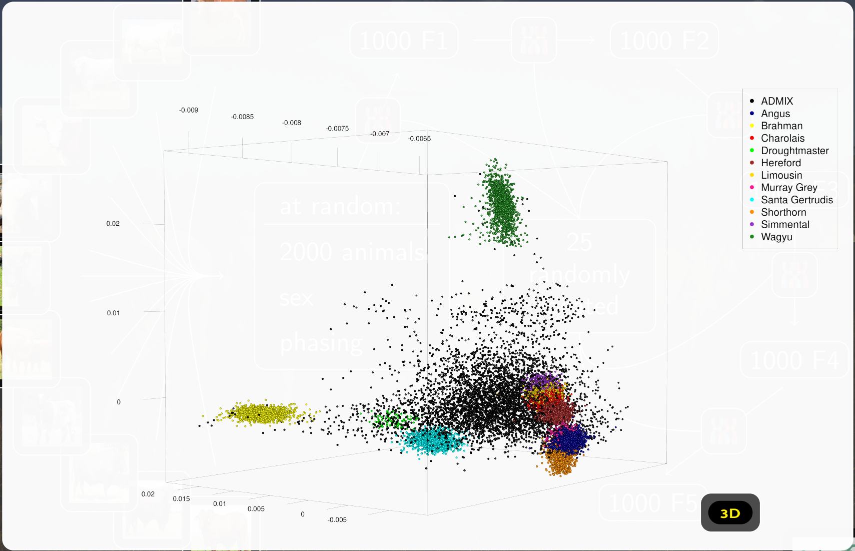
phasing

Data

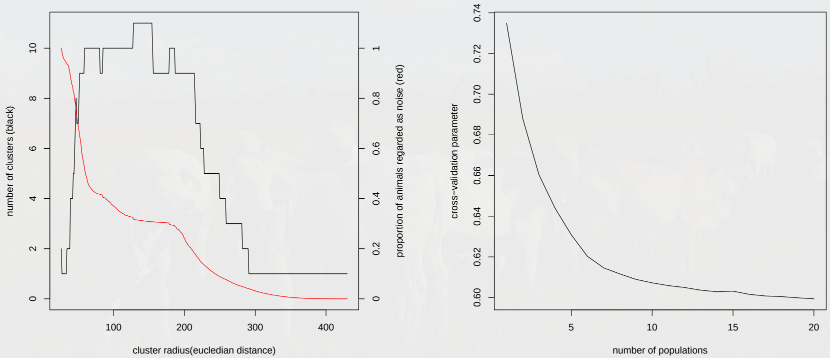
11,639 true pure breed animals

5,000 artificial cross breed animals

Data


-0.009
-0.0085
-0.008
-0.0075
-0.007
-0.0065
0.02
0.01
0
0.02
0.015
0.01
0.005
0
-0.005
● ADMIX
● Angus
● Brahman
● Charolais
● Droughtmaster
● Hereford
● Limousin
● Murray Grey
● Santa Gertrudis
● Shorthorn
● Simmental
● Wagyu
3D
1000 F1
1000 F2
F3
1000 F4
1000 F5
at random:
2000 animals
sex
phasing
25
randomly
...ted

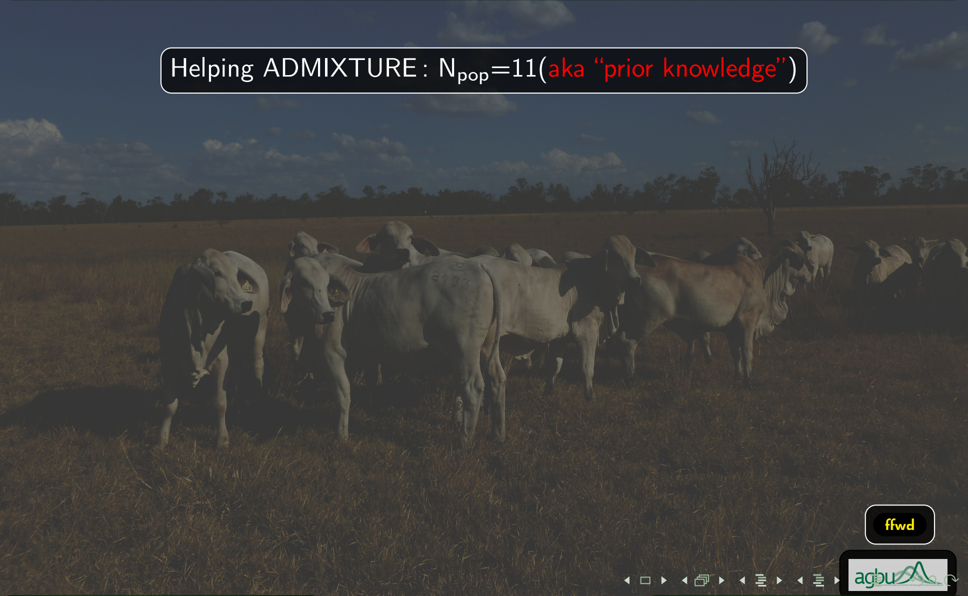
Population recovery: Number of founder populations



(a) OPTICS N_{clusters} (black) and N_{crosses} (red)

(b) ADMIXTURE cross-validation

Population recovery: allele frequency



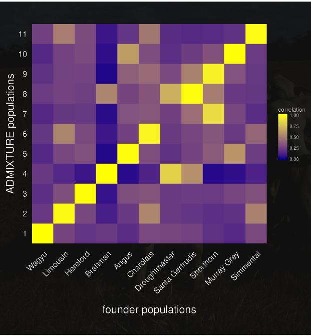
Helping ADMIXTURE : N_{pop}=11(aka “prior knowledge”)

ffwd

Population recovery: allele frequency

Helping ADMIXTURE : N_{pop}=11(aka "prior knowledge")

heatmaps of allele frequency correlations

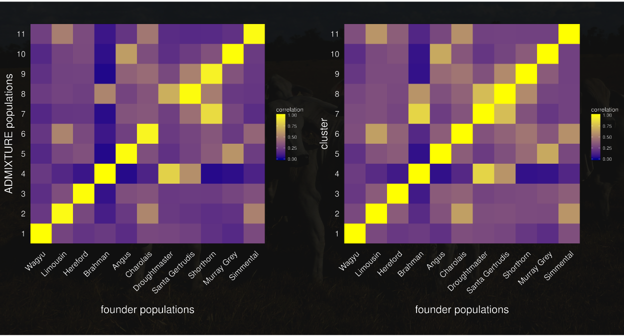


ffwd

Population recovery: allele frequency

Helping ADMIXTURE : N_{pop}=11(aka “prior knowledge”)

heatmaps of allele frequency correlations

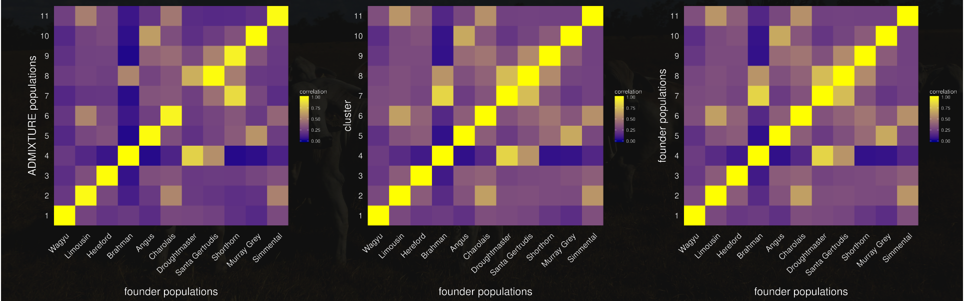


ffwd

Population recovery: allele frequency

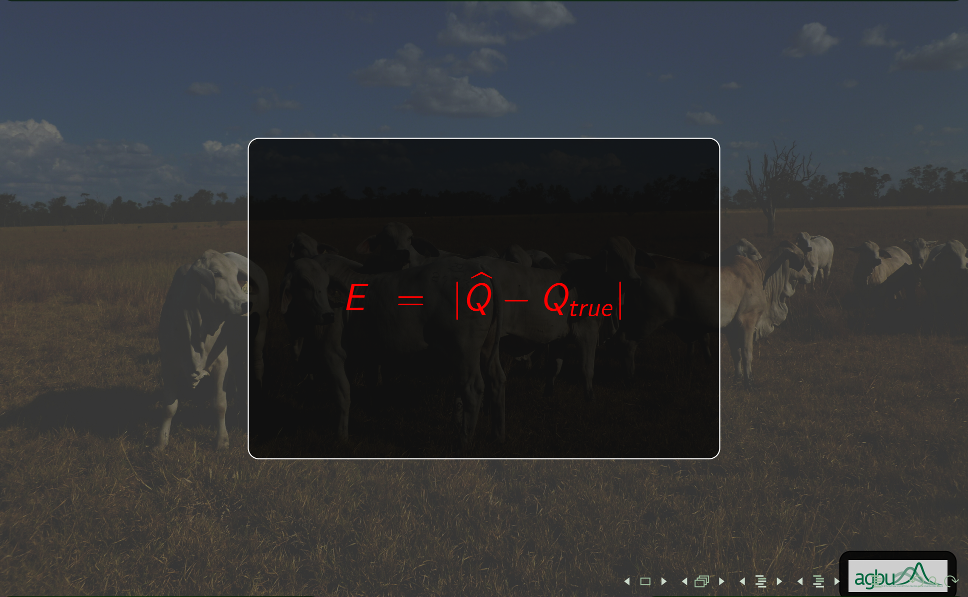
Helping ADMIXTURE : N_{pop}=11(aka “prior knowledge”)

heatmaps of allele frequency correlations



ffwd

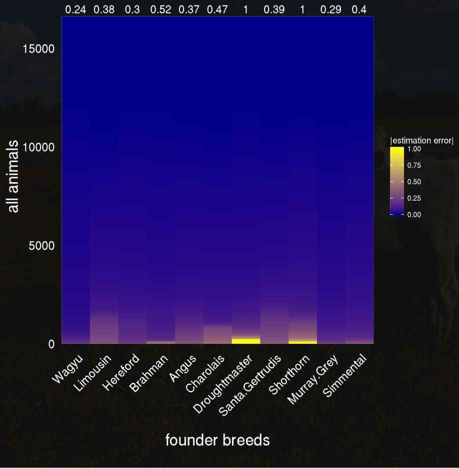
Genome proportion recovery



$$E = |\widehat{Q} - Q_{true}|$$

Genome proportion recovery

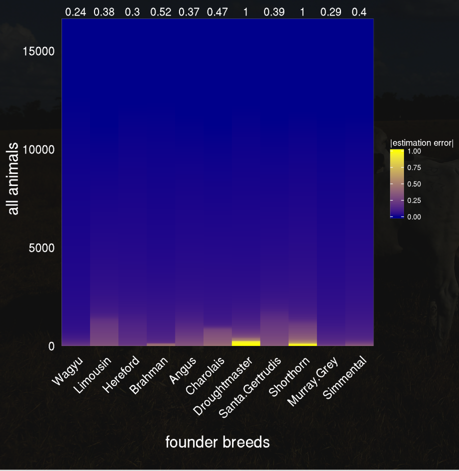
ADMIXTURE



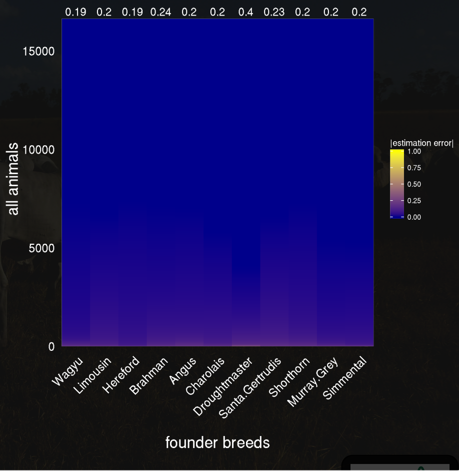
CGR

Genome proportion recovery

ADMIXTURE



CGR



Speed

Founder population recovery

ADMIXTURE

- 18 **hours**
- N_{pop}=1,...,20

OPTICS

- 22 **seconds**
- 405 cluster solutions

Speed

Founder population recovery

ADMIXTURE

- 18 **hours**
- N_{pop}=1,...,20

OPTICS

- 22 **seconds**
- 405 cluster solutions

Breed proportion estimation

ADMIXTURE

- 45 **minutes** with N_{pop}=11

CGR

- 30 **seconds**

Conclusion

- Loop approach (ADMIXTURE) => caution
  - number of populations is unknown
    - may fail to detect number of populations
    - subsequently may wrongly assign genome proportions
  - number of populations is known
    - may wrongly assign genome proportions
  - F is known (“supervised”)
    - may wrongly assign genome proportions (Boerner, AAABG 2017)
- OPTICS => fast and precise
  - detection of point aggregations => pure-bred animals, stabilised crosses
  - detection of noise => cross-bred animals
  - cluster allele frequencies reflect founder allele frequencies
  - relies on point aggregations
- CGR => fast and precise
  - requires good estimate of allele frequencies

Acknowledgement

DNA submitting breeders

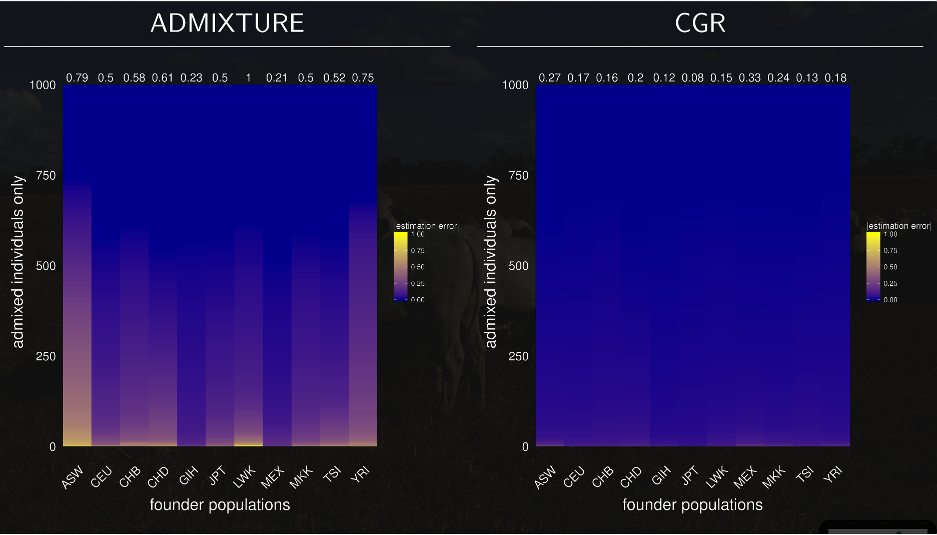
Meat and Livestock Australia

Animal Genetics and Breeding Unit (AGBU)

A joint venture of the University of New England and the NSW Department of Primary Industry

download CGR: http://turing.une.edu.au/∼agbu-admin/BESSiE/

Supervised genome proportion recovery human data set



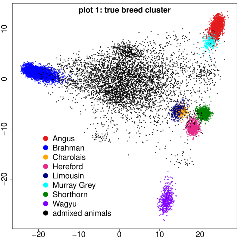
ADMIXTURE

CGR

Population recovery: OPTICS vs. ADMIXTURE

back

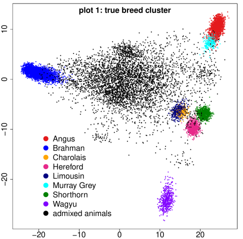
2. vs 3. singular vector, 8 cattle breeds

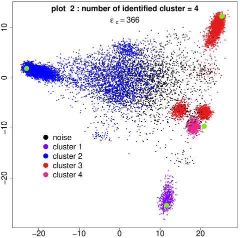


Population recovery: OPTICS vs. ADMIXTURE

back

2. vs 3. singular vector, 8 cattle breeds

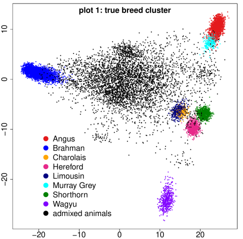


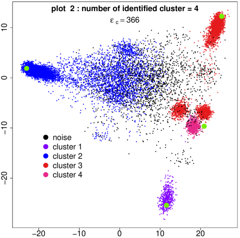


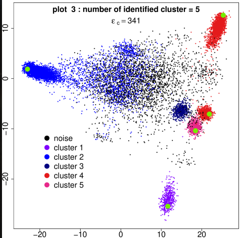
Population recovery: OPTICS vs. ADMIXTURE

back

2. vs 3. singular vector, 8 cattle breeds



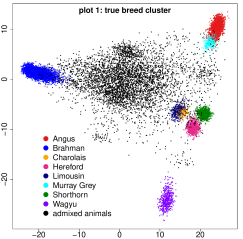


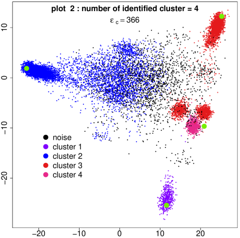


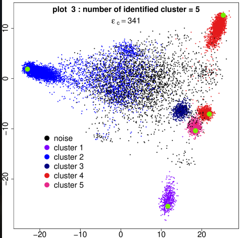
Population recovery: OPTICS vs. ADMIXTURE

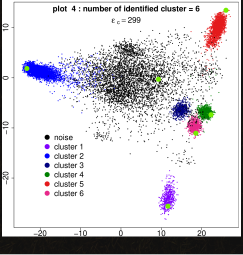
2. vs 3. singular vector, 8 cattle breeds

back





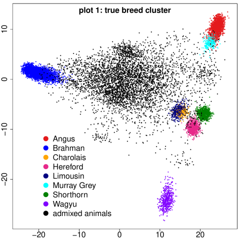


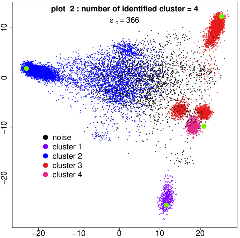


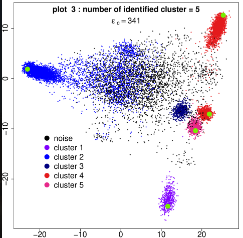
Population recovery: OPTICS vs. ADMIXTURE

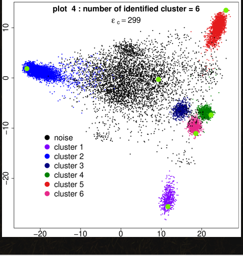
2. vs 3. singular vector, 8 cattle breeds

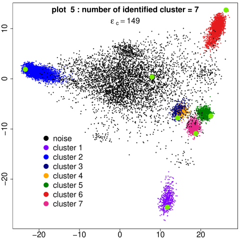
back







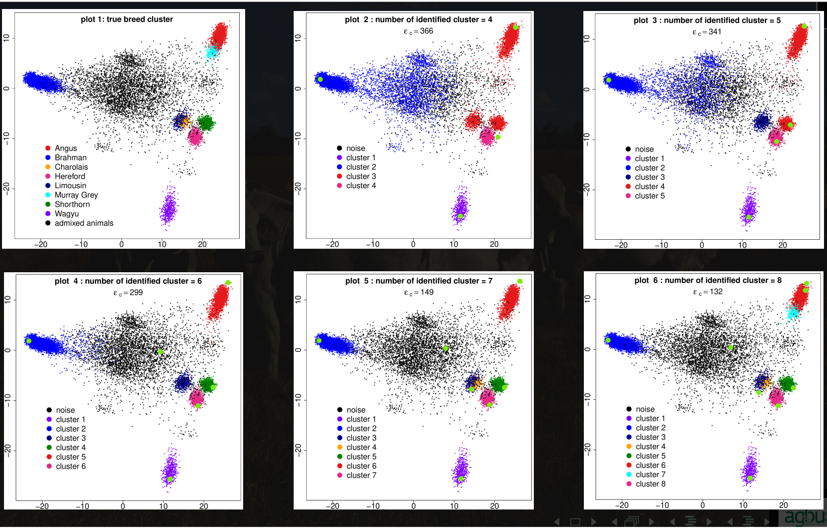


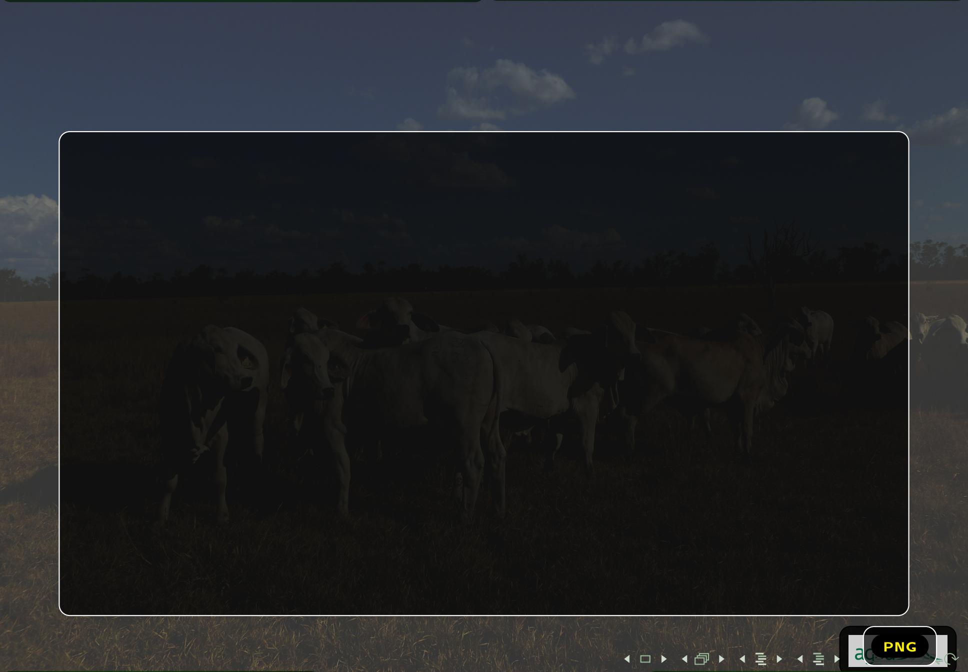


Population recovery: OPTICS vs. ADMIXTURE

2. vs 3. singular vector, 8 cattle breeds

back





PNG