SNP based parentage verification via constraint non-linear optimisation

Vinzent Boerner and Robert Banks

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

Interbull meeting, 24-28/10/2016, Puerto Varas, Chile

Background

- "historic" parentage verification (PV)

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites,STRs)
  - statistical approaches[Jones et al., 2010]

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches [Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches [Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 $\rightarrow$ Single Nucleotide Polymorphisms

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 $\rightarrow$ Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping $\rightarrow$ genomic selection (GS)

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites,STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 $\rightarrow$ Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping $\rightarrow$ genomic selection (GS)
  - PV still mainly based on STRs

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 $\rightarrow$ Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping $\rightarrow$ genomic selection (GS)
  - PV still mainly based on STRs
  - double genotyping cost for PV and GS

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 $\rightarrow$ Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping $\rightarrow$ genomic selection (GS)
  - PV still mainly based on STRs
  - double genotyping cost for PV and GS
  - "quick and dirty" remedy $\rightarrow$ impute STRs from SNPs

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 → Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping → genomic selection (GS)
  - PV still mainly based on STRs
  - double genotyping cost for PV and GS
  - "quick and dirty" remedy → impute STRs from SNPs
- long-term approach

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites,STRs)
  - statistical approaches[Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 → Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping → genomic selection (GS)
  - PV still mainly based on STRs
  - double genotyping cost for PV and GS
  - "quick and dirty" remedy → impute STRs from SNPs
- long-term approach
 - impute SNPs from STRs (feasible???)

Background

- "historic" parentage verification (PV)
  - based on short tandem repeats (micro satellites, STRs)
  - statistical approaches [Jones et al., 2010]
    - total exclusion (who cannot be mum or dad)
    - partial exclusion (who is most likely mun or dad)
    - etc.
- mid 2000 $\rightarrow$ Single Nucleotide Polymorphisms
  - now dominant marker class for genotyping $\rightarrow$ genomic selection (GS)
  - PV still mainly based on STRs
  - double genotyping cost for PV and GS
  - "quick and dirty" remedy $\rightarrow$ impute STRs from SNPs
- long-term approach
  - impute SNPs from STRs (feasible???)
  - use SNPs for parentage verification directly

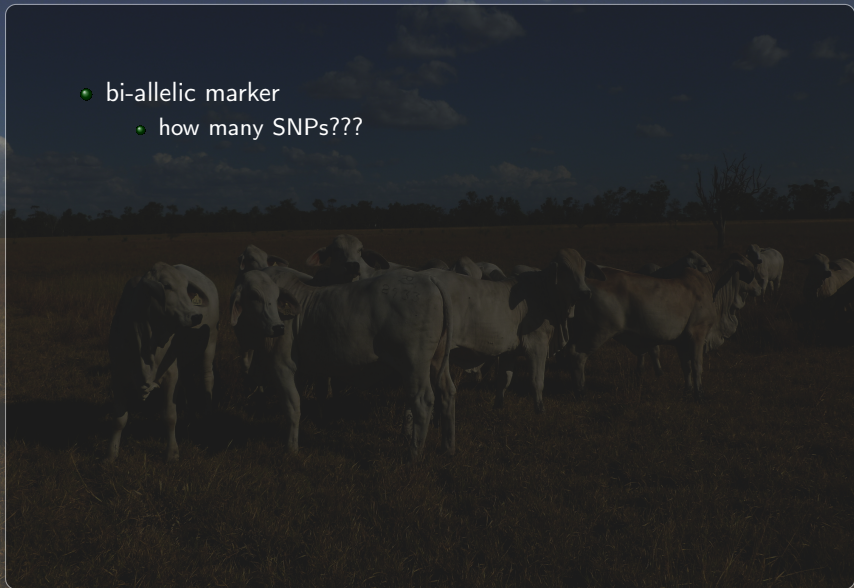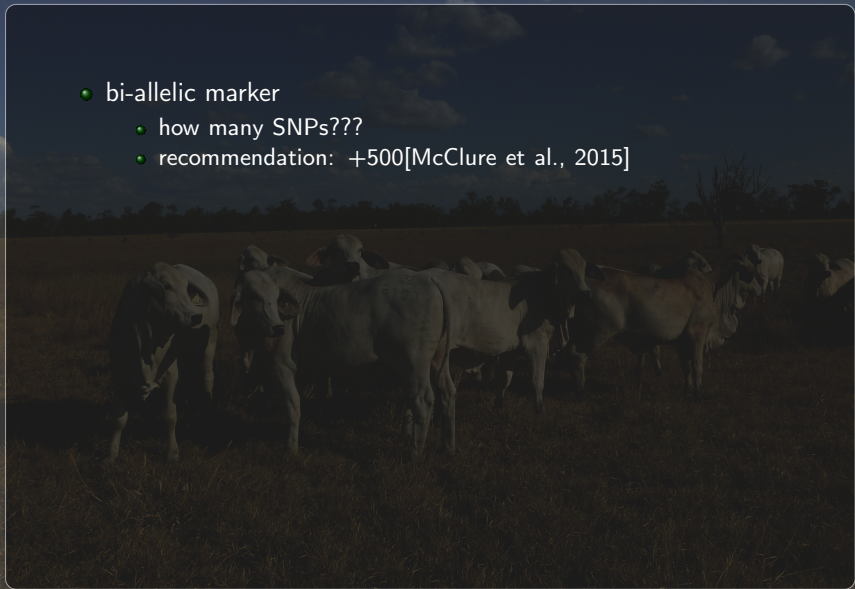# SNP parentage verification $\rightarrow$ not too easy
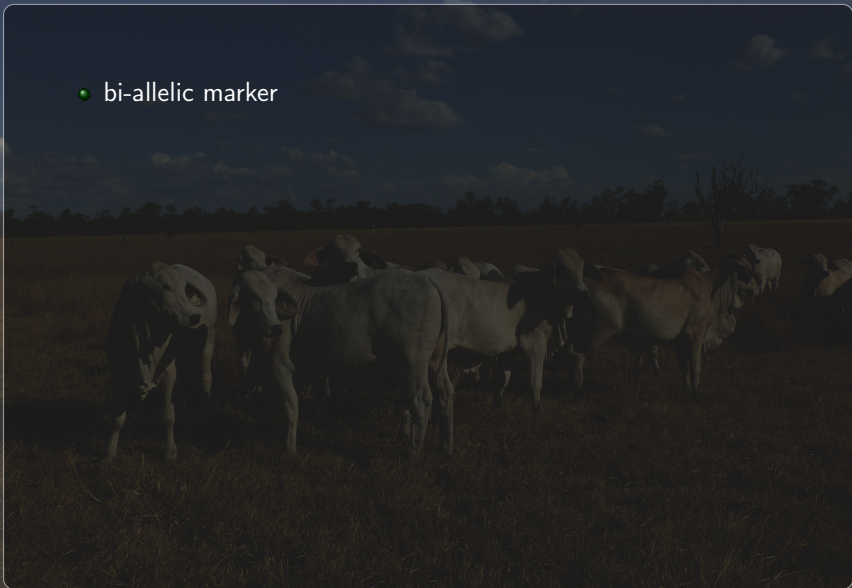
- bi-allelic marker

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability

SNP parentage verification → not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability
- statistical analysis

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability
- statistical analysis
  - exclusion via opposing homozygosity[Wiggans et al., 2009, Hayes, 2011]

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability
- statistical analysis
  - exclusion via opposing homozygosity[Wiggans et al., 2009, Hayes, 2011]
    - "relies" on genotyping errors $\rightarrow$ to be minimised

SNP parentage verification → not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability
- statistical analysis
  - exclusion via opposing homozygosity[Wiggans et al., 2009, Hayes, 2011]
    - "relies" on genotyping errors → to be minimised
    - parents must be among the putative parents → cannot be assured

SNP parentage verification → not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability
- statistical analysis
  - exclusion via opposing homozygosity[Wiggans et al., 2009, Hayes, 2011]
    - "relies" on genotyping errors → to be minimised
    - parents must be among the putative parents → cannot be assured
  - likelihood based method adapted from STRs[Boichard et al., 2014]

SNP parentage verification $\rightarrow$ not too easy

- bi-allelic marker
  - how many SNPs???
  - recommendation: +500[McClure et al., 2015]
- which marker
  - different platforms
  - customisation vs. transferability
- statistical analysis
  - exclusion via opposing homozygosity[Wiggans et al., 2009, Hayes, 2011]
    - "relies" on genotyping errors $\rightarrow$ to be minimised
    - parents must be among the putative parents $\rightarrow$ cannot be assured
  - likelihood based method adapted from STRs[Boichard et al., 2014]
    - difficult to reject putative parents when $N_{SNP} < 100$

General idea

resort to the linear model

$$\text{animals marker} = 0.5 \times \text{mum's marker} +$$
$$0.5 \times \text{dad's marker} +$$
$$\text{mendelian sampling}$$

General idea

resort to the linear model

$$\text{animals marker} = 0.5 \times \text{mum's marker} +$$
$$0.5 \times \text{dad's marker} +$$
$$\text{mendelian sampling}$$

more mathematical: $y = Xb + e$

General idea

resort to the linear model

animals marker = 0.5 × mum's marker+
0.5 × dad's marker+
mendelian sampling

more mathematical: $y = Xb + e$

y : genotype of an animal with one/both parents unknown

X : matrix of column vectors of genotypes from putative parents

b : vector of regression coefficients

e : non-explainable residual

Simple genomic regression

Solve:

$$\arg\min_b f(b) \;=\; y'y - 2y'Xb + b'X'Xb \qquad \rightarrow \qquad \text{simple least square}$$

Simple genomic regression

Solve:

$$\arg\min_b f(b) \quad = \quad y'y - 2y'Xb + b'X'Xb \quad \rightarrow \quad \text{simple least square}$$

Too simple to be useful:

- b's can have any value $\rightarrow$ hard/impossible to interpret
- every b_i can be non-zero $\rightarrow$ who is the parent?

Make it work

1. Step: constraint genomic regression

$$\arg\min_b f(b) \quad = \quad y'y - 2y'Xb + b'X'Xb$$

$$\text{subject to}$$

$$b_i \quad \geq \quad 0 \; \{i = 1, ., N\}$$

$$\sum_i^N b_i \quad = \quad 1$$

Make it work

1. Step: constraint genomic regression

$$\arg\min_b f(b) \;=\; y'y - 2y'Xb + b'X'Xb$$

subject to

$$b_i \;\geq\; 0 \;\{i = 1, ., N\}$$
$$\sum_i^N b_i \;=\; 1$$

2. Step: augment X

- column vectors of genotypes of putative parents
- population allele frequency vector

Make it work

How to solve:

- 2-step iterative non-linear optimisation solver (NLOpt library[Johnson, 2014])

Make it work

How to solve:

- 2-step iterative non-linear optimisation solver (NLOpt library[Johnson, 2014])
  - global minimisation: augmented Lagrangian method

Make it work

How to solve:

- 2-step iterative non-linear optimisation solver (NLOpt library[Johnson, 2014])
  - global minimisation: augmented Lagrangian method
  - local minimisation: method of moving asymptotes

Make it work

How to solve:

- 2-step iterative non-linear optimisation solver (NLOpt library[Johnson, 2014])
  - global minimisation: augmented Lagrangian method
  - local minimisation: method of moving asymptotes

Results evaluation

- b's assign parentage if $> \frac{1}{3} \rightarrow$ assures number of parents ≤ 2

Make it work

How to solve:

- 2-step iterative non-linear optimisation solver (NLOpt library[Johnson, 2014])
  - global minimisation: augmented Lagrangian method
  - local minimisation: method of moving asymptotes

Results evaluation

- b's assign parentage if $> \frac{1}{3} \rightarrow$ assures number of parents ≤ 2
- if both true parents are absent $\rightarrow b$ for population allele frequency vector approaches 1

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle
  - 47702 SNPs extracted from Illumina 50K Bead Chip genotypes

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle
  - 47702 SNPs extracted from Illumina 50K Bead Chip genotypes
  - contained 90 genotyped triplets (sire, dam, progeny)

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle
  - 47702 SNPs extracted from Illumina 50K Bead Chip genotypes
  - contained 90 genotyped triplets (sire, dam, progeny)
- 3 different SNP data sets were formed:

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle
  - 47702 SNPs extracted from Illumina 50K Bead Chip genotypes
  - contained 90 genotyped triplets (sire, dam, progeny)
- 3 different SNP data sets were formed:
  - all SNPs

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle
  - 47702 SNPs extracted from Illumina 50K Bead Chip genotypes
  - contained 90 genotyped triplets (sire, dam, progeny)
- 3 different SNP data sets were formed:
  - all SNPs
  - 500 randomly selected SNPs

Test data set

SNP data

- 4612 genotypes of Australian Angus beef cattle
  - 47702 SNPs extracted from Illumina 50K Bead Chip genotypes
  - contained 90 genotyped triplets (sire, dam, progeny)
- 3 different SNP data sets were formed:
  - all SNPs
  - 500 randomly selected SNPs
  - 100 randomly selected SNPs with minor allele frequency > 0.3

Test data set

Equation data

- $y \rightarrow$ progeny from triplet
- $X \rightarrow$ three different types per y

Test data set

Equation data

- $y \rightarrow$ progeny from triplet
- $X \rightarrow$ three different types per y

X_P

- sire
- dam
- 10 randomly selected animals
- population allele frequency

Test data set

Equation data

- $y \rightarrow$ progeny from triplet
- $X \rightarrow$ three different types per y

X_P

- sire
- dam
- 10 randomly selected animals
- population allele frequency

X_D

- dam
- 11 randomly selected animals
- population allele frequency

Test data set

Equation data

- $y \rightarrow$ progeny from triplet
- $X \rightarrow$ three different types per y

X_P

- sire
- dam
- 10 randomly selected animals
- population allele frequency

X_D

- dam
- 11 randomly selected animals
- population allele frequency

X_R

- 12 randomly selected animals
- population allele frequency

Test data set

Equation data

- $y \rightarrow$ progeny from triplet
- $X \rightarrow$ three different types per y

X_P

- sire
- dam
- 10 randomly selected animals
- population allele frequency

X_D

- dam
- 11 randomly selected animals
- population allele frequency

X_R

- 12 randomly selected animals
- population allele frequency

- randomly selected animals: excluded parents, full sibs and half sibs; re-sampled for every y
- population allele frequency vector was calculated excluding genotypes in X and y

Trial summary

9 equations were solved for for every single y

	X_P	X_D	X_R
SNP_{Full}			
SNP_{500}			
SNP_{100}			

- threshold for b's assigning parentage: $> \frac{1}{3}$

Results
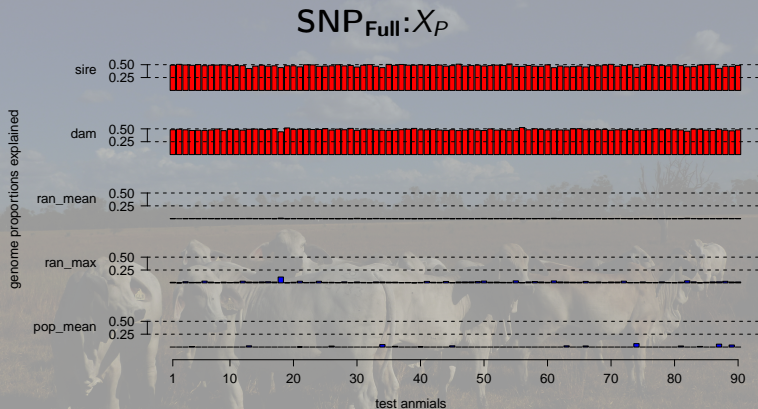
- all parents correctly identified
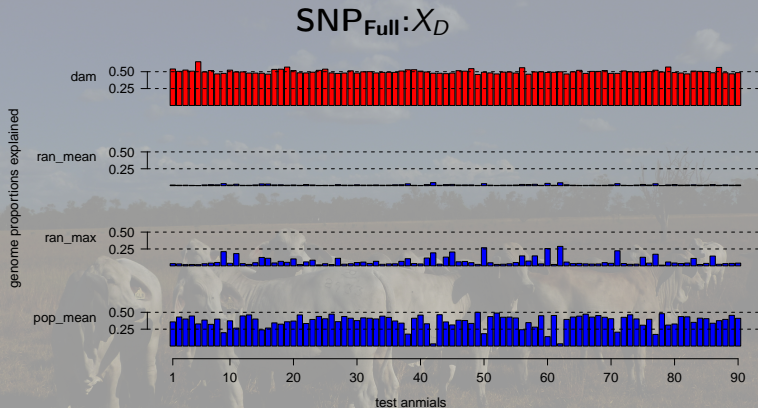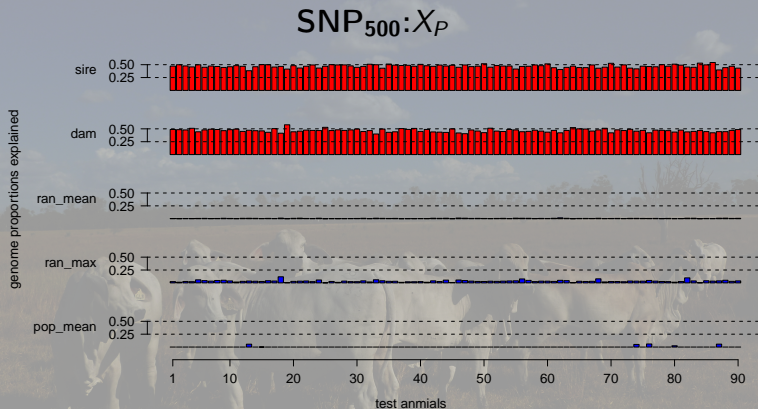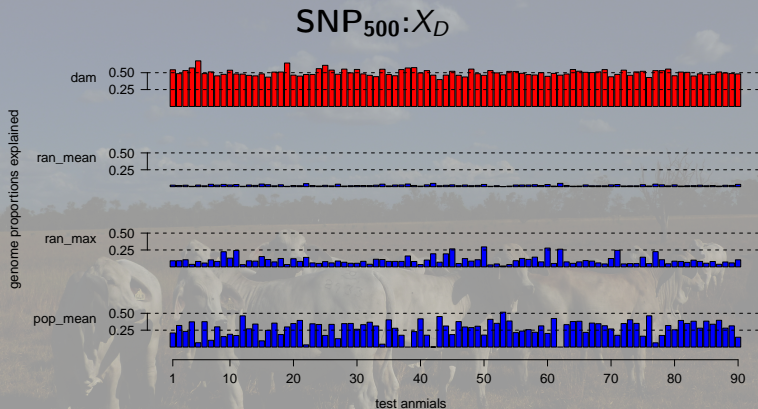
Results

$$\text{SNP}_{\textbf{Full}}:X_D$$

- all dams correctly identified
- parentage correctly rejected for all random animals

Results

SNP$_{\textbf{Full}}$:X_R

- parentage correctly rejected for all random animals

Results

$SNP_{500}:X_P$

- all parents correctly identified

Results

$\text{SNP}_{500}{:}X_D$
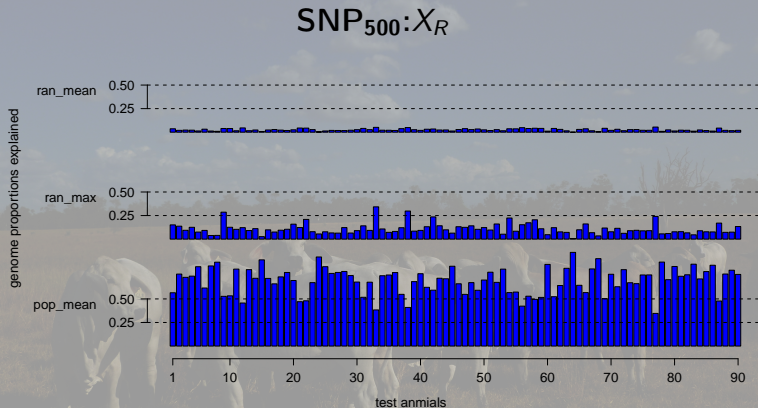
- all dams correctly identified
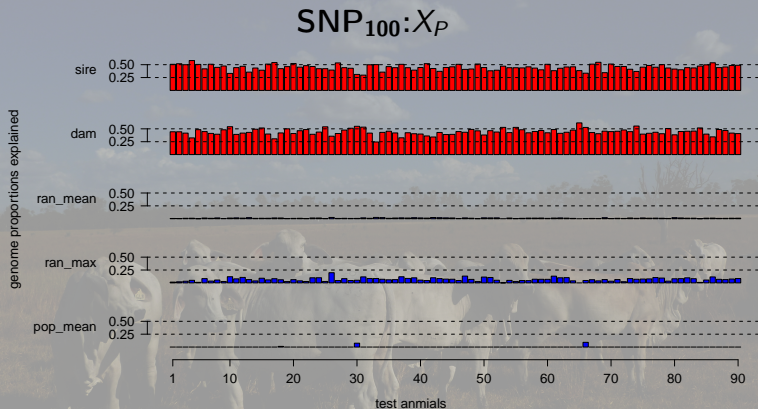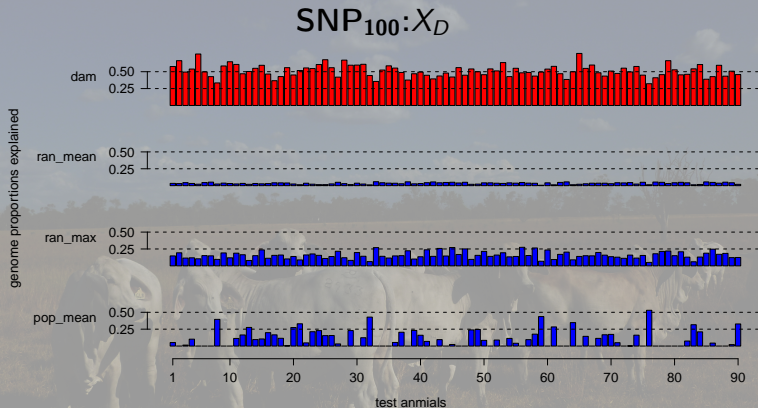- parentage correctly rejected for all random animals

Results

- parentage correctly rejected in 179 of 180 cases

Results

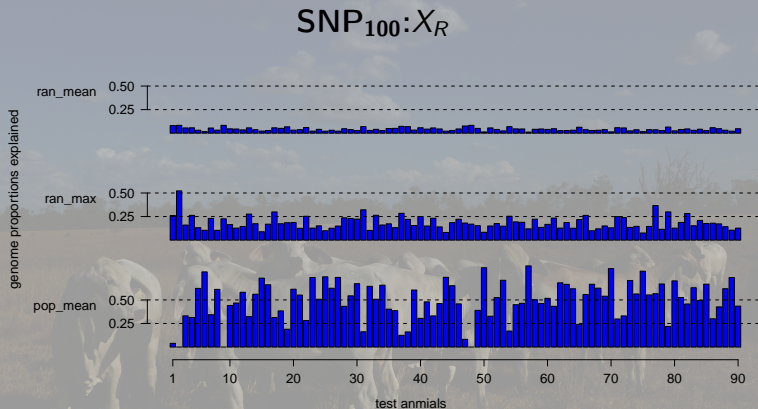
- parents correctly identified in 173 of 180 cases
- parentage correctly rejected for all random animals

Results

- dams correctly identified in 89 of 90 cases
- parentage correctly rejected for all random animals

Results

SNP$_{100}$:X_R

- parentage correctly rejected in 178 of 180 cases

Results summary

power of assignment:

$$\frac{\text{number of correct assignments}}{\text{number of true parents}}$$

	X_P	X_D	X_R
SNP_{Full}	1	1	-
SNP_{500}	1	1	-
SNP_{100}	0.96	0.99	-

Results summary

power of assignment:

$$\frac{\text{number of correct assignments}}{\text{number of true parents}}$$

power of exclusion:

$$1 - \frac{\text{number of wrong assignments}}{\text{number of possible parents}}$$

	X_P	X_D	X_R
SNP_{Full}	1	1	-
SNP_{500}	1	1	-
SNP_{100}	0.96	0.99	-

	X_P	X_D	X_R
SNP_{Full}	1	1	1
SNP_{500}	1	1	0.99
SNP_{100}	1	1	0.99

Discussion

- general performance

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
  - does not rely on true parents being among putative parents

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
  - does not rely on true parents being among putative parents
  - does not rely on genotyping errors

Discussion

- general performance
    - almost perfect assignment for $N_{SNPs} \geq 500$
    - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
    - automatically scales to any input data set
    - fast
        - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
        - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
    - does not rely on true parents being among putative parents
    - does not rely on genotyping errors
- compared to likelihood based methods

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
  - does not rely on true parents being among putative parents
  - does not rely on genotyping errors
- compared to likelihood based methods
  - better performance for exclusion for $N_{SNPs} \leq 100$ (compared to Boichard et al. [2014])

Discussion

- general performance
    - almost perfect assignment for $N_{SNPs} \geq 500$
    - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
    - automatically scales to any input data set
    - fast
        - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
        - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
    - does not rely on true parents being among putative parents
    - does not rely on genotyping errors
- compared to likelihood based methods
    - better performance for exclusion for $N_{SNPs} \leq 100$ (compared to Boichard et al. [2014])
- possible significance test for solutions:

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
  - does not rely on true parents being among putative parents
  - does not rely on genotyping errors
- compared to likelihood based methods
  - better performance for exclusion for $N_{SNPs} \leq 100$ (compared to Boichard et al. [2014])
- possible significance test for solutions:
  - test $e'e|b_i > \frac{1}{3}$ against its empirical distribution constructed from e's when one or both parents were in X

Discussion

- general performance
  - almost perfect assignment for $N_{SNPs} \geq 500$
  - almost perfect exclusion for $N_{SNPs} \leq 100$ (more important)
  - automatically scales to any input data set
  - fast
    - SNP_{50k}: 0.1 real time seconds per animal (INTEL i7 3770)
    - SNP_{500}: 0.003 real time seconds per animal (INTEL i7 3770)
- compared to exclusion via opposing homozygosity
  - does not rely on true parents being among putative parents
  - does not rely on genotyping errors
- compared to likelihood based methods
  - better performance for exclusion for $N_{SNPs} \leq 100$ (compared to Boichard et al. [2014])
- possible significance test for solutions:
  - test $e'e|b_i > \frac{1}{3}$ against its empirical distribution constructed from e's when one or both parents were in X
  - may also allow to reject "tricky" sets of putative parents containing progeny, full sibs or half sibs

Acknowledgements

- This work was funded by Meat and Livestock Australia (Project B.BFG.0050).
- The authors thank various Australian Angus breeders for supplying genotypes.

References

Didier Boichard, L Barbotte, and L Genestout.   Accurassign, software for accurate maximum-likelihood parentage assignment.   In *Proc. 10th. WCGALP, Vancouver, Canada, August*, page np, 2014.

BJ Hayes.   Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *J. Dairy Sci.*, 94(4):2114–2117, 2011.

Steven G Johnson. The nlopt nonlinear-optimization package. 2011. *URL http://ab-initio. mit. edu/nlopt*, 2014.

Adam G Jones, Clayton M Small, Kimberly A Paczolt, and Nicholas L Ratterman.   A practical guide to methods of parentage analysis. *Mol. Ecol. Resour.*, 10(1):6–30, 2010.

MC McClure, J McCarthy, P Flynn, R Weld, M Keane, K O'Connell, M Mullen, S Waters, JF Kearney, Z Kowalski, et al.   Snp selection for nationwide parentage verification and identification in beef and dairy cattle. *ICAR Technical Series*, (19):175–181, 2015.

G. R. Wiggans, T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell.   Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the united states and canada. *J. Dairy Sci.*, 92(7):3431–3436, Jul 2009.   doi: $10.3168/jds.2008\text{-}1758$.   URL http://dx.doi.org/10.3168/jds.2008-1758.