Using Random Forests As A Prescreening Tool for Genomic Prediction: Impact of Subsets of SNPs on Prediction Accuracy of Total Genetic Values

Yutao Li, Fernanda Raidan, Bo Li, Zulma Vitezica and Toni Reverter

CSIRO AGRICULTURE & FOOD
www.csiro.au

Why Machine Learning Methods Become Popular in Large Genomic Data Analyses ?

- Dealing with "*Large P and Small N*" problem
- Black–box approaches (No prior knowledge required)
- Taking multiple interactions or correlations among predictor variables (e.g. SNP-SNP interactions) into account
- High prediction accuracy (building training and validation procedures into algorithms)

Knowledge Gap in Genomic Prediction of Total Genetic Values

- **Do non-additive effects captured by machine learning methods contribute to the prediction accuracy of total genetic values ?**

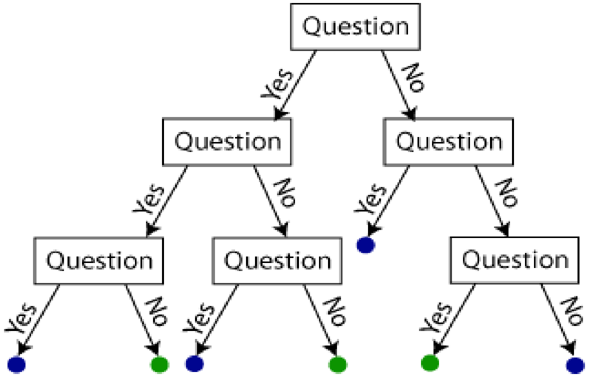
↓

(additive +dominance genetic values)

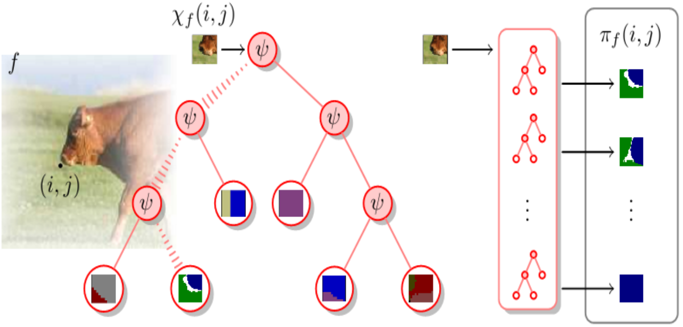
Machine Learning Method – Random Forests (RF)

- **Leo Breiman, *Random Forests*, Machine Learning, 45, 5-32, 2001.**

- **A nonparametric tree-based ensemble machine-learning method for classification or regression of multiple variables.**

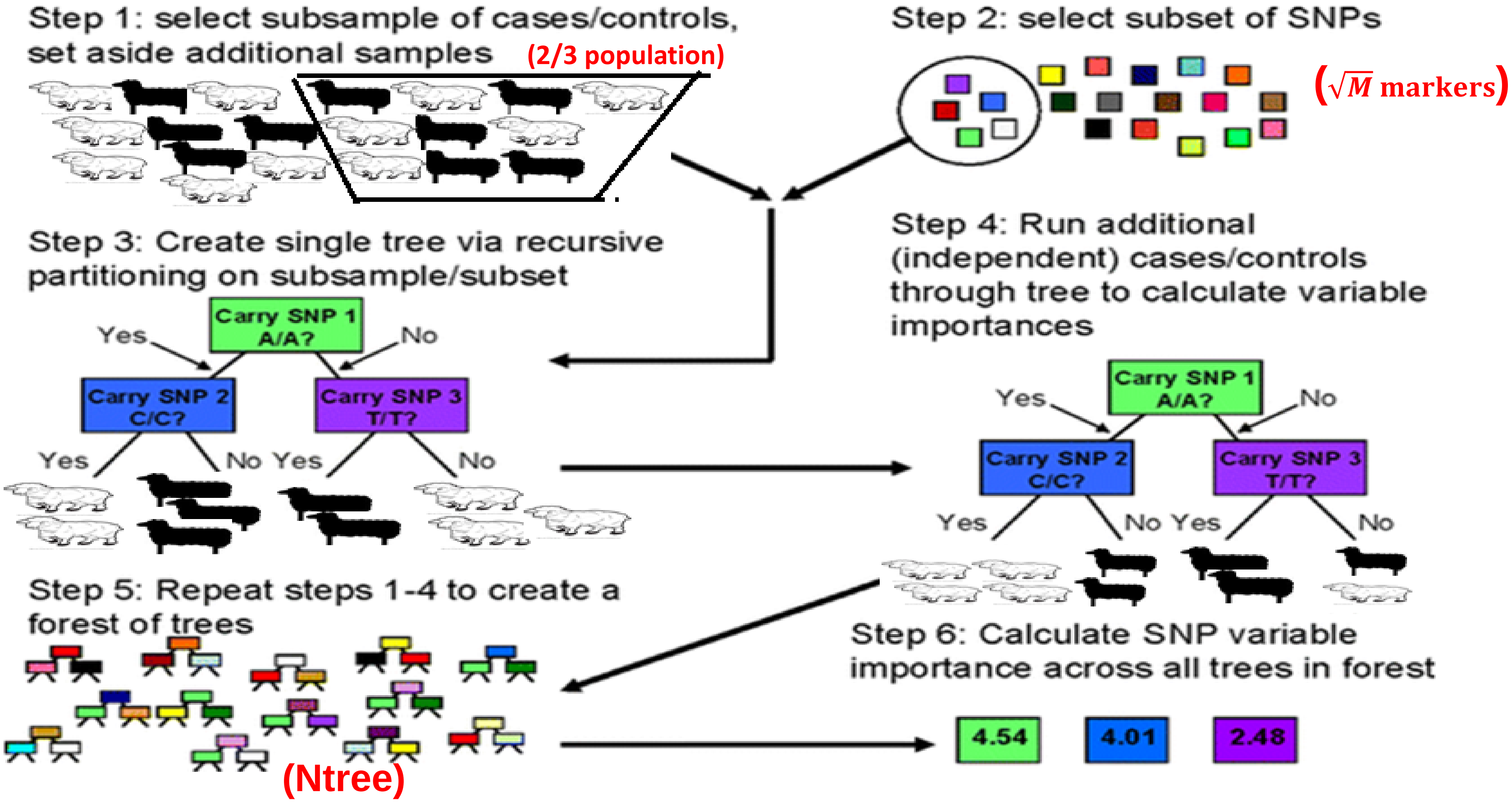


http://shapeofdata.wordpress.com/2013/07/09/random-forests/



http://pdollar.wordpress.com/2013/03/08/structured-random-forests/

Random Forests – How Does It Work?



Step 1: select subsample of cases/controls, set aside additional samples (2/3 population)

Step 2: select subset of SNPs ($\sqrt{M}$ markers)

Step 3: Create single tree via recursive partitioning on subsample/subset

Step 4: Run additional (independent) cases/controls through tree to calculate variable importances

Step 5: Repeat steps 1-4 to create a forest of trees (Ntree)

Step 6: Calculate SNP variable importance across all trees in forest

Nicodemus et al. 2010. Hum Genet 127:441-452.

SNP Variable Importance Value (VIM)

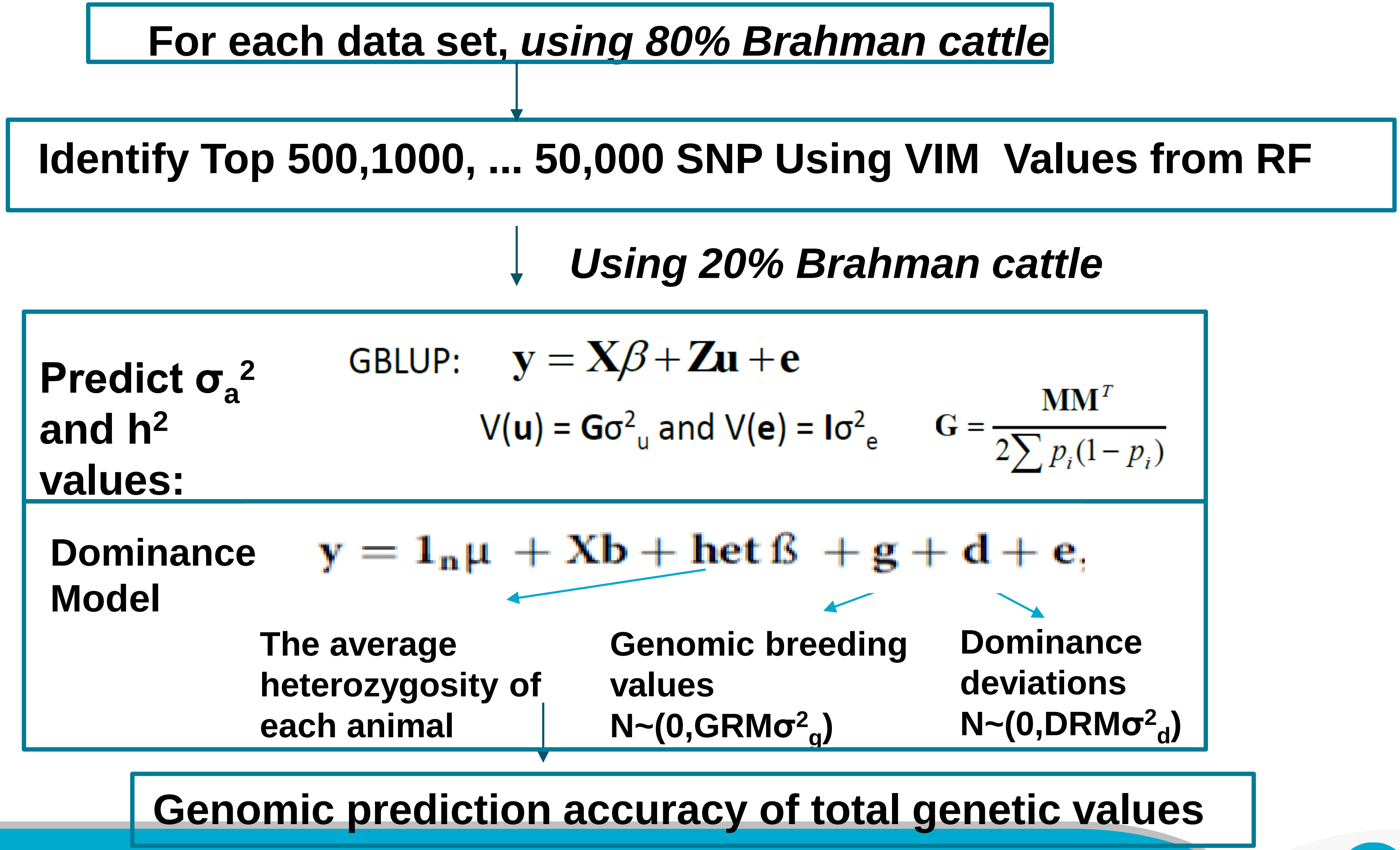
- **RF**: **%IncMSE** (% Increasing in Mean Squared Error when a SNP is not included)

Larger the value, the more important the SNP is

Beef CRC Cattle Dataset

- 2,109 Brahman Cattle
- 651,253 SNPs (MAF > 0.01)
- Phenotype – Yearling Weight (pre-adjusted for average heterozygosity of SNPs, contemporary group and age effects)

Application of RF in Identifying Subsets of SNP for Genomic Prediction



For each data set, *using 80% Brahman cattle*

Identify Top 500,1000, ... 50,000 SNP Using VIM Values from RF

Using 20% Brahman cattle

Predict σ_a^2 and h^2 values:

GBLUP: $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$

$V(\mathbf{u}) = \mathbf{G}\sigma^2_u$ and $V(\mathbf{e}) = \mathbf{I}\sigma^2_e$

$$\mathbf{G} = \frac{\mathbf{MM}^T}{2\sum p_i(1-p_i)}$$

Dominance Model

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Xb} + \mathbf{het}\,ß + \mathbf{g} + \mathbf{d} + \mathbf{e}$$

The average heterozygosity of each animal

Genomic breeding values $N\sim(0,GRM\sigma^2_g)$

Dominance deviations $N\sim(0,DRM\sigma^2_d)$

Genomic prediction accuracy of total genetic values

Variance Estimates from Subsets of SNPs

(RF Selected vs Evenly Spaced vs All SNPs)

	Additive Model			Additive + Dominance Model			
	h_a^2	σ_a^2	ACC	h_a^2	σ_a^2	σ_d^2	ACC
RF							
500	0.21	140.3	0.45	0.21	140.1	14.2	0.45
1,000	0.26	171.6	0.49	0.26	171.7	24.2	0.49
5,000	0.39	254.9	0.55	0.39	253.4	58.0	0.55
50,000	0.45	299.0	0.60	0.44	294.0	205.1	0.60
Even							
500	0.04	24.8	0.18	0.03	24.8	0.8	0.18
1,000	0.03	25.8	0.21	0.03	23.5	2.2	0.21
5,000	0.09	58.9	0.24	0.08	58.6	15.7	0.24
50,000	0.34	236.5	0.29	0.34	234.2	55.6	0.29
All SNPs	0.38	259.4	0.44	0.38	258.4	49.9	0.44

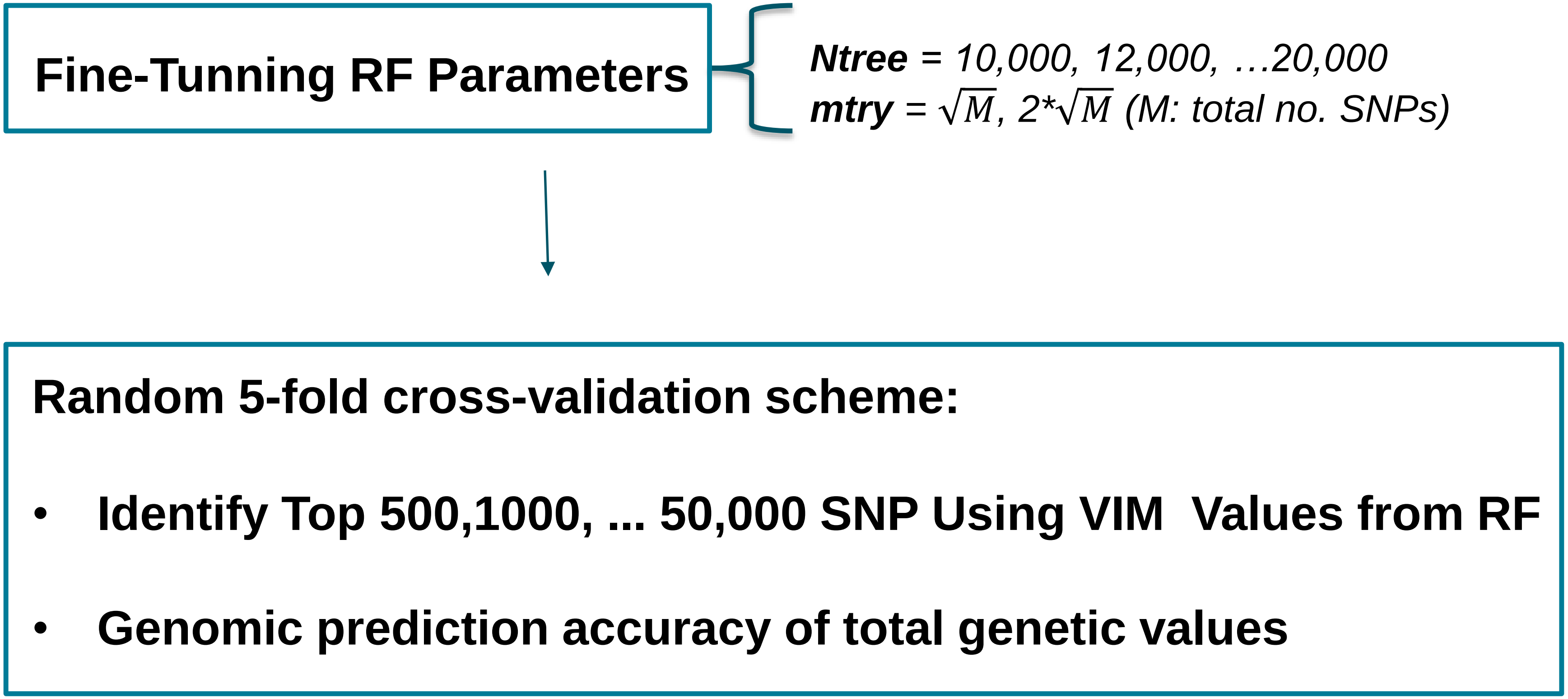
Conclusions

- **Fitting dominance into the genomic model had little impact on the accuracy of genomic prediction of breeding values.**
- **RF has potential to be used as a pre-screening tool for:**

  a) reduction of high dimensionality associated with large genomic data;

  b) identification of subsets of useful SNPs for genomic prediction of breeding values.

Application of RF in Identifying Subsets of SNP for Genomic Prediction of Cattle Live Weight



Fine-Tunning RF Parameters

Ntree *= 10,000, 12,000, …20,000*

mtry *=* $\sqrt{M}$*, 2**$\sqrt{M}$ *(M: total no. SNPs)*

Random 5-fold cross-validation scheme:

- **Identify Top 500,1000, ... 50,000 SNP Using VIM Values from RF**
- **Genomic prediction accuracy of total genetic values**

GRM and DRM Calculations

$$G = \frac{Z_a Z_a'}{2\sum_{k=1}^{m} p_k q_k}$$

- ***Z*$_a$ *(nxm)***

$$\left\{\begin{array}{ll} \mathbf{2-2p_k} & \textbf{(AA)} \\ \mathbf{1-2p_k} & \textbf{(AB) (VanRaden et al., 2008)} \\ \mathbf{-2p_k} & \textbf{(BB)} \end{array}\right.$$

- p_k - menor allele frequence of locus *k*

$$D^* = \frac{Z_d Z_d'}{4\sum_{k=1}^{m} p_k^2 q_k^2}$$

- ***Z*$_d$ *(nxm)***

$$\left\{\begin{array}{ll} \mathbf{2q_k^2} & \textbf{(AA)} \\ \mathbf{2p_k(1-p_k)} & \textbf{(AB) (Vitezica et al., 2013)} \\ \mathbf{-2p_k} & \textbf{(BB)} \end{array}\right.$$

- Matrix D* was the combined with identity matrix I as $\mathrm{D} = 0.95D^* + 0.05I$ to improve numerical stability

Variance Estimates from Subsets of SNPs

RF Selected vs Evenly Spaced vs All SNPs

	Additive Model			Additive + Dominance Model			
	h_a^2	σ_a^2	% Total σ_a^2	h_a^2	σ_a^2	σ_d^2	σ_p^2
RF							
500	0.21 (0.03)	140.3 (22.8)	667.0 (26.5)	0.21 (0.08)	140.1 (22.7)	14.2 (8.6)	668.1 (26.2)
1,000	0.26 (0.03)	171.6 (25.0)	658.8 (26.1)	0.26 (0.03)	171.7 (25.0)	24.2 (11.4)	660.2 (26.2)
5,000	0.39 (0.04)	254.9 (32.7)	658.2 (26.3)	0.39 (0.04)	253.4 (32.5)	58.0 (21.3)	658.9 (26.4)
50,000	0.45 (0.04)	299.0 (38.7)	669.2 (26.7)	0.44 (0.05)	294.0 (38.3)	205.1 (60.8)	670.3 (27. 5)
Even							
500	0.04 (0.02)	24.8 (14.2)	691.0 (26.1)	0.03 (0.02)	24.8 (8.7)	0.8 (1.7)	691. 8 (24.9)
1,000	0.03 (0.02)	25.8 (14.4)	691.6 (25.9)	0.03 (0.02)	23.5 (12.6)	2.2 (5.1)	690.3 (25.5)
5,000	0.09 (0.03)	58.9 (21.2)	703.3 (27.8)	0.08 (0.03)	58.6 (21.2)	15.7 (15.6)	705.7 (28.2)
50,000	0.34 (0.05)	236.5 (38.7)	690.0 (26.5)	0.34 (0.05)	234.2 (45.7)	55.6 (51.8)	689.8 (26.9)
All SNPs	0.38 (0.05)	259.4 (38.4)	680.9 (26.5)	0.38 (0.05)	258.4 (38.2)	49.9 (33.5)	680.5 (26.5)

Genomic Prediction Accuracy of Total Genetic Values

	Additive Model	Additive + Dominance Model
	Acc	Acc
RF		
500	0.45	0.45
1,000	0.49	0.49
5,000	0.55	0.55
50,000	0.60	0.60
Even		
500	0.18	0.18
1,000	0.21	0.21
5,000	0.24	0.24
50,000	0.29	0.29
All SNP	0.44	0.44