Accounting for read depth in the analysis of genotyping-by-sequencing data

Ken Dodds, John McEwan, Timothy Bilton, Rudi Brauning,
Rayna Anderson, Tracey Van Stijn, Theodor Kristjánsson,
Shannon Clarke

AgResearch, Invermay, New Zealand

Genotyping-by-sequencing (GBS)

- Alternative genotyping technology
- Methods here apply to any sequencing-based genotypes
- Reduced representational methods cost-effective
  - Restriction enzyme digests
  - Untargeted regions
  - Barcode to multiplex lanes (48 – 384)
  - No-oligo design/purchase required
  - Reference sequence optional
  - Based on Elshire method (‘GBS’)

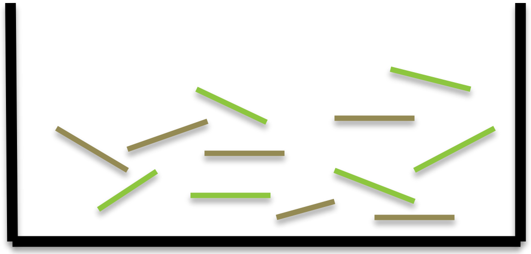
Elshire *et al*
(2011) PLoS ONE

agresearch

GBS SNPs

- True heterozygote (AB)
- Observation probabilities (random model)

# reads (k)	Only A's	A's and B's	Only B's
1	0.5	0	0.5
2	0.25	0.5	0.25
3	0.125	0.75	0.125
4	0.0625	0.875	0.0625
⋮	⋮	⋮	⋮
∞	0	1	0
Inferred genotype	**AA***	**AB***	**BB***



* denotes inferred genotype

Genomic relatedness theory

- $P(\mathrm{AA}^*|\mathrm{AB}) = \ K$
- $P(\mathrm{AB}^*|\mathrm{AB}) \ = \ 1 - 2K$

* denotes inferred genotype

For random sampling:
$K = 1/2^k$
k = depth

- $P(\mathrm{AA}^*) = P(\mathrm{AA}) + P(\mathrm{AB})K$
- $P(\mathrm{AA}) = p^2 + p(1-p)F$

F = inbreeding

- $P(\mathrm{AA}, \mathrm{AA})$ = function of relatedness measures
- …

Genomic relatedness theory

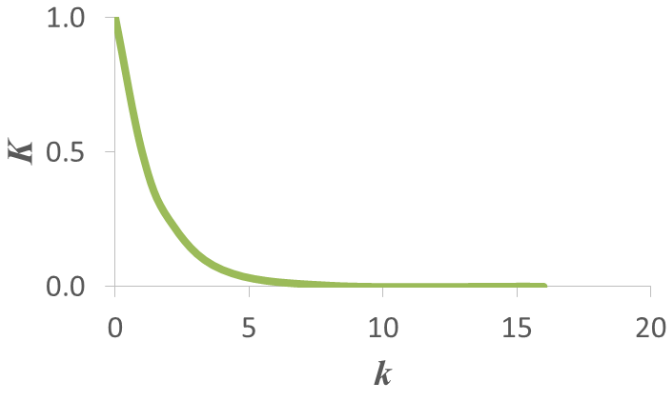
- $\mathbf{M}$: genotypes (0/1/2) (individuals x SNPs)
- Centre using $\mathbf{P}$ (jth column = $2p_j$)
  - $\mathbf{Z} = \mathbf{M}\text{-}\mathbf{P}$
- $\mathbf{G}_1 = \dfrac{\mathbf{ZZ}'}{2\sum p_j(1-p_j)}$
- Numerator for 1 SNP: $z_{ii'} = (x_i - 2p)\,(x_{i'} - 2p)$
  - x_i is the marker score (0, 1, 2) for g_i^*
- $E(z_{ii'}) = 2p(1-p)2\theta$
- $E(z_{ii}) = 2p(1-p)(1 + F_i + 2K_i - 2F_iK_i)$

i indexes individuals
j indexes SNPs

p_j allele frequencies

VanRaden 1st method

2θ =relatedness



Genomic relatedness theory: Remarks

- $\mathbf{G}_1 = \frac{\mathbf{ZZ}'}{\boxed{2\sum p_j(1-p_j)}}$

- $E(z_{ii'}) = \boxed{2p(1-p)}2\theta$

- $E(z_{ii}) = \boxed{2p(1-p)}(1 + F_i + 2K_i - 2F_iK_i)$

- $\mathbf{G}_1$ divisor gives the correct expected value
  - Assumes p is known

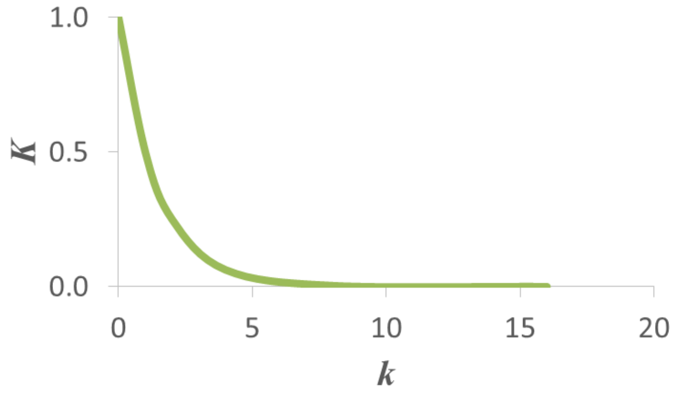
Genomic relatedness theory: Remarks

- $\mathbf{G}_1 = \frac{\mathbf{ZZ}'}{2\sum p_j(1-p_j)}$

- $E(z_{ii'}) = 2p(1-p)2\theta$

- $E(z_{ii}) = 2p(1-p)(1 + F_i + 2K_i - 2F_iK_i)$

- The expected value for different individuals *does not depend on* K
- The expected value for self-relatedness depends on K
  - No information on F when k =1 (discard)
  - Need to take depth into account (k >1)



Genomic relatedness theory

- Missing values
  - Usual method is ‘naïve imputation’ – use $2p_j$
    - Biased downwards
  - Our method: only use SNPs non-missing (for both individuals)
- **K**inship using **G**BS with **D**epth adjustment (KGD)
- Dodds *et al* (2015) BMC Genomics
- R code on github

Example: parentage

- 2203 Atlantic Salmon (pedigree recorded)
- 122 full-sib families, (122 dams, 66 sires, 1177 progeny)
- Genotyped: all sires, 119 dams and 94 full-sib families
- Filter to exclude duplicated regions (HW-.05)
- 30,923 SNPs; mean SNP depth was 7.9 (All)
- 24,899 SNPs; mean SNP depth was 3.3(HW-.05)

Mean relatedness estimates

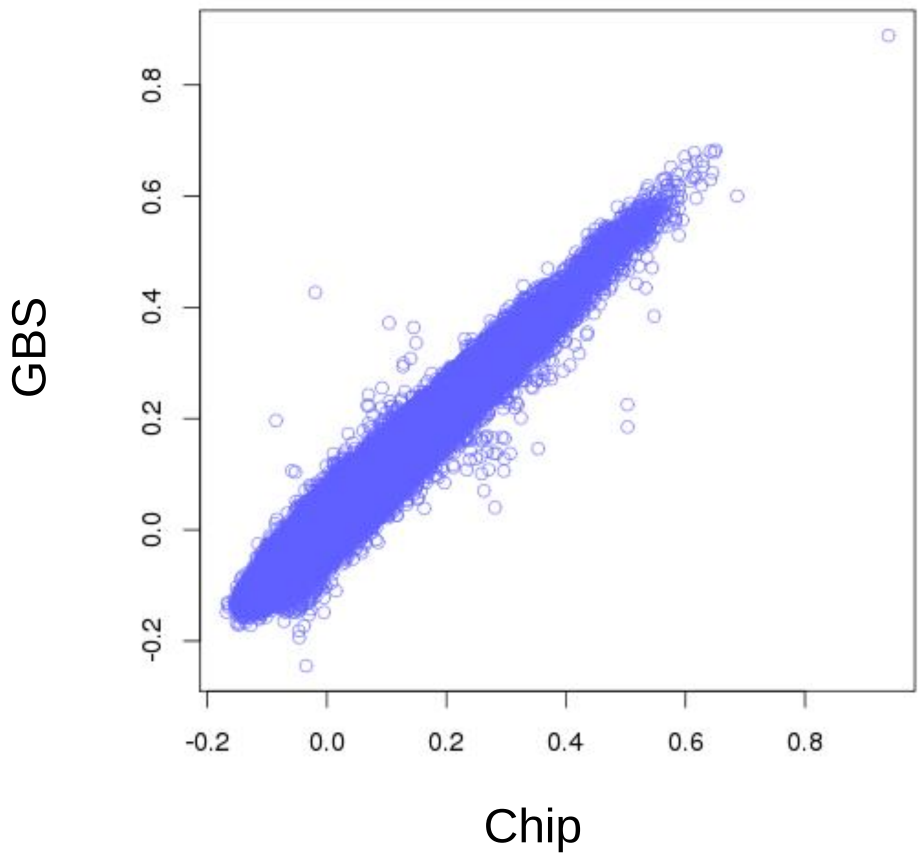
		Relationship group			
Analysis	Number of SNPs	Identity	Full-sibs	Parent-Offspring	Non-sib Offspring
All	30,923	0.739	0.375	0.382	0.109
HW-.05	24,899	1.014	0.454	0.461	0.014

Example: GRM

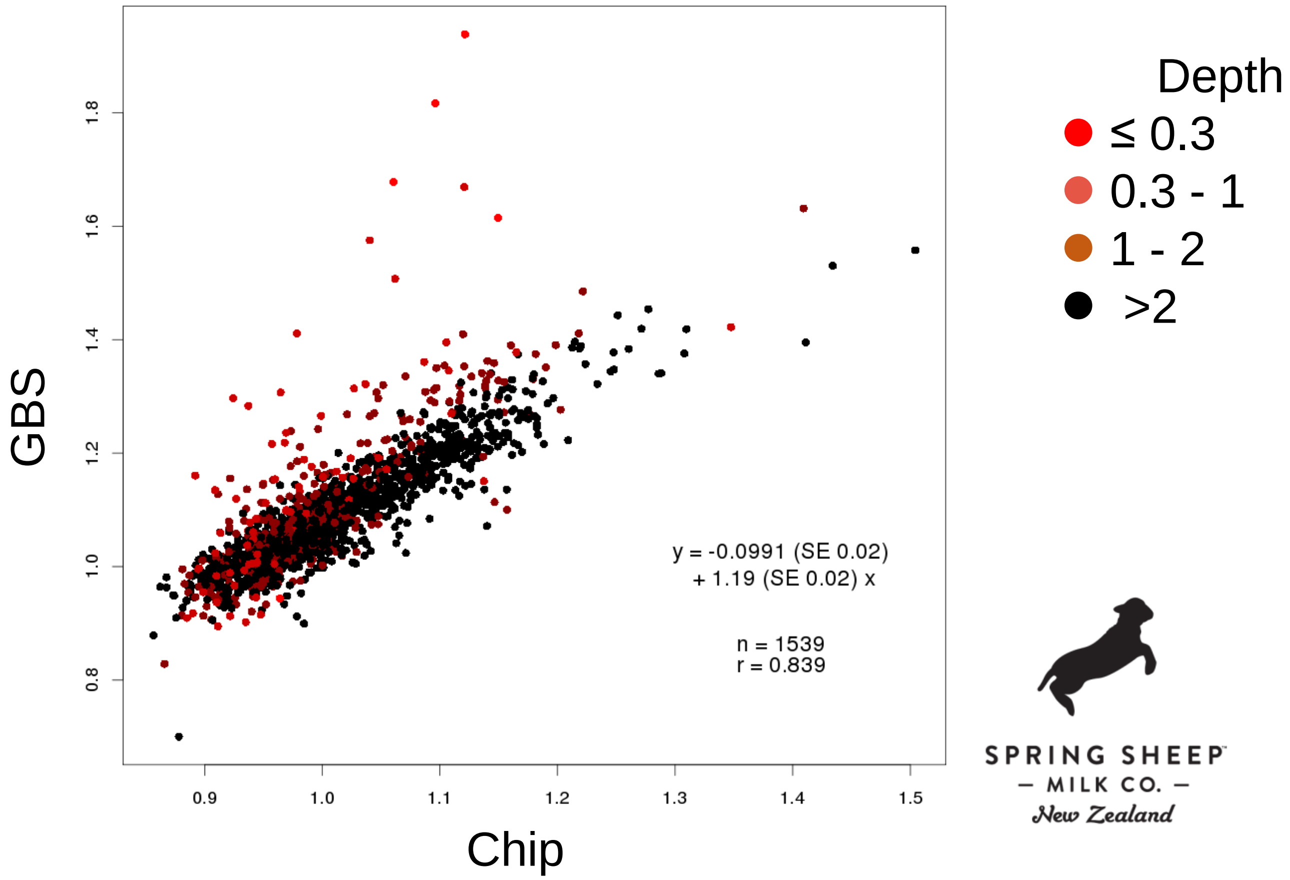
Dairy Sheep
Also 15k chip genotyped

Data courtesy of Suzanne Rowe, AgResearch

Off-diagonal comparisons



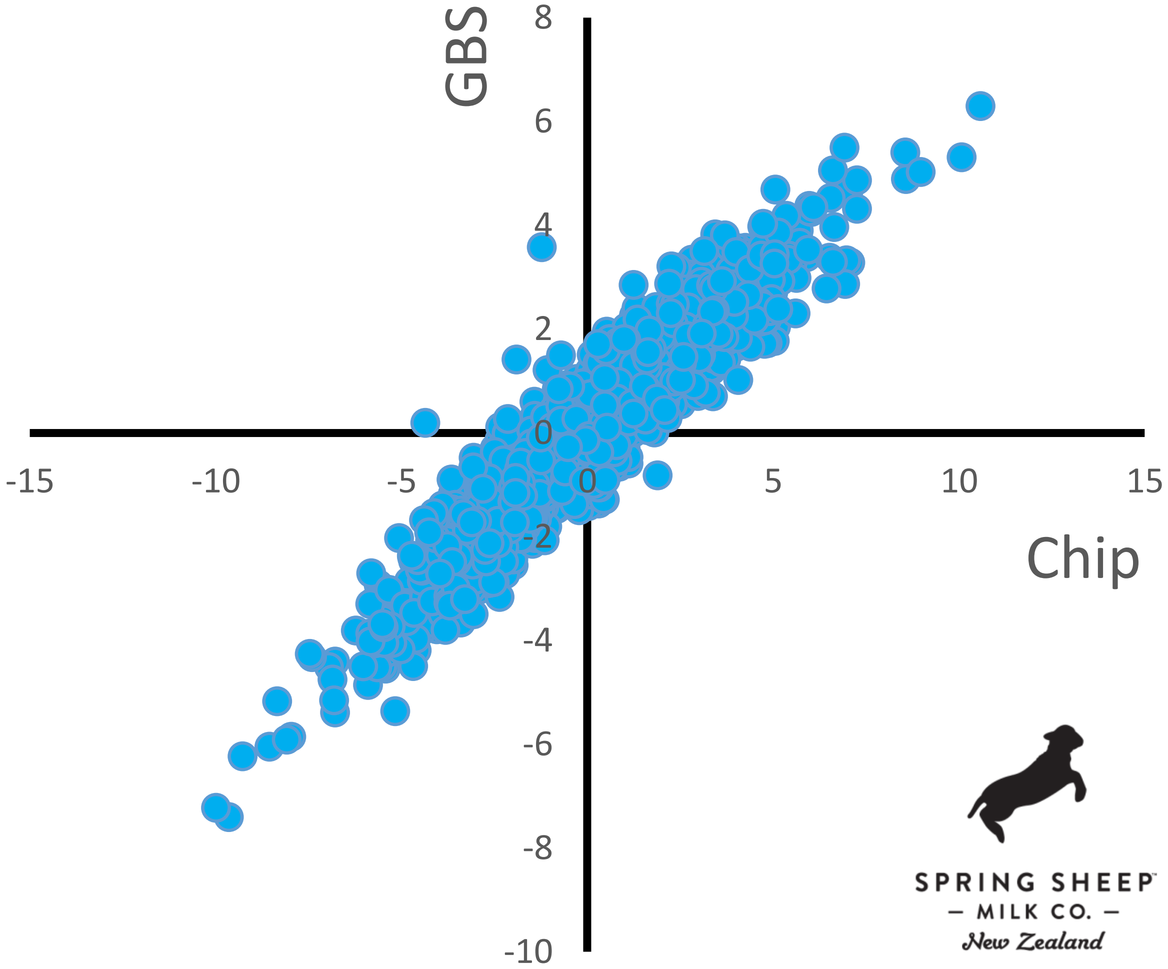
Diagonal comparisons



SPRING SHEEP
— MILK CO. —
New Zealand

Example: Genomic Prediction

- GBLUP Breeding values
- Milk (kg)

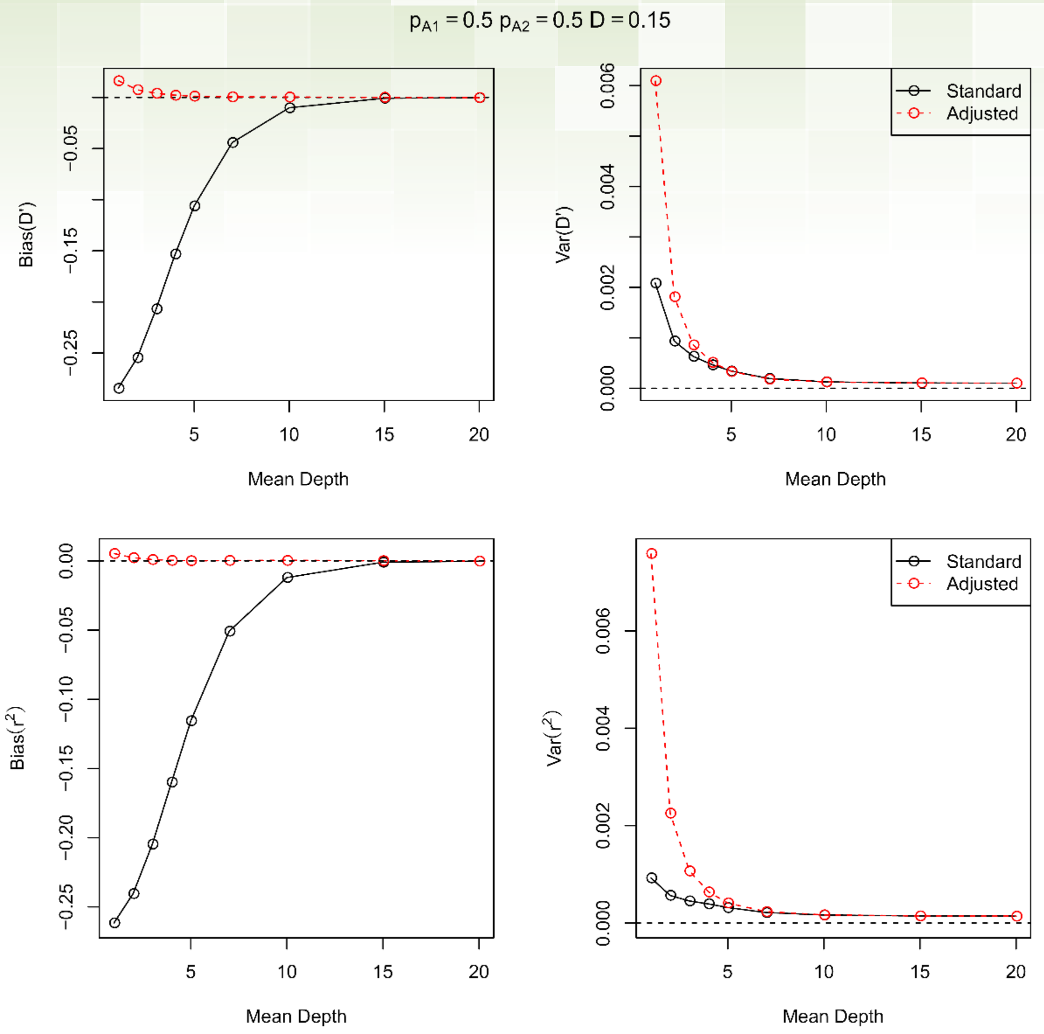


SPRING SHEEP
– MILK CO. –
New Zealand

agresearch

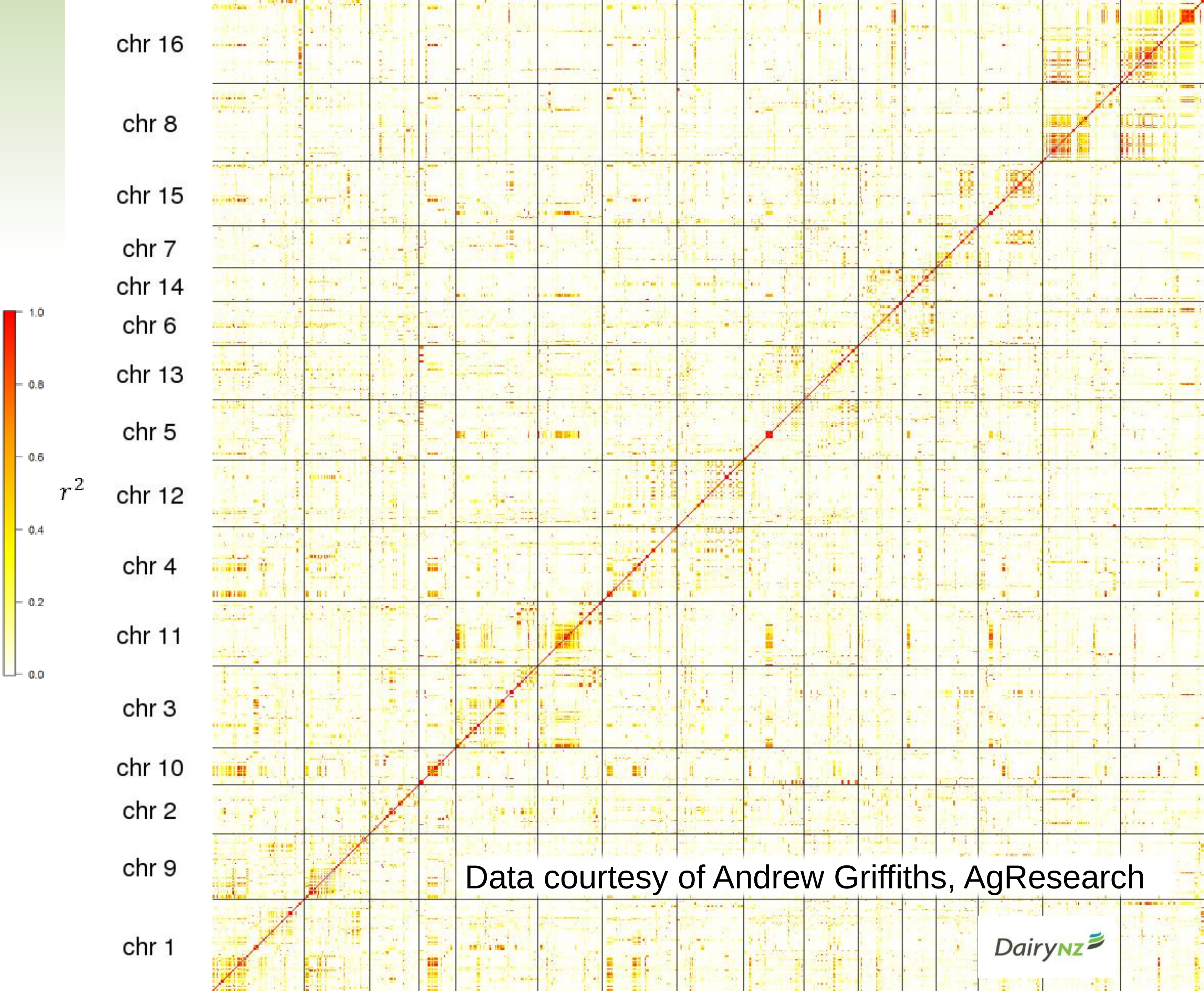
GBS: Linkage Disequilibrium

- Use Likelihood approach
  - Observed genotype combination given LD, read depth
- Simulations
  - True Values:
    - $D' = 0.6$
    - $r^2 = 0.36$
- Standard method results in strong bias
- Depth-adjusted method improves bias but is more variable
- The two methods were similar with high read depth


$p_{A1} = 0.5 \; p_{A2} = 0.5 \; D = 0.15$

LD Estimation: Application

- Genome assembly improvement
  - White clover
  - Allotetraploid
  - Full-sib family
  - Marker pairs in “backcross” configuration



Data courtesy of Andrew Griffiths, AgResearch

Conclusions

- Low depth GBS …
  - Analysis methods to accommodate data type
    - Allelic sampling, rather than genotypes
  - Method for unbiased relatedness estimation
    - Many genetic analyses can be based on relatedness estimates
  - Method for Linkage disequilibrium
  - Enables low depth GBS to be used

Acknowledgements

Genomics for Production & Security in a Biological Economy

Ministry of Business,
Innovation & Employment

Suzanne Rowe (dairy sheep)

Andrew Griffiths (clover)

StofnFiskur

agresearch