



Genome-wide association mapping

Andrew George
CMIS & Food Futures Flagship
AASC2011, Cairns

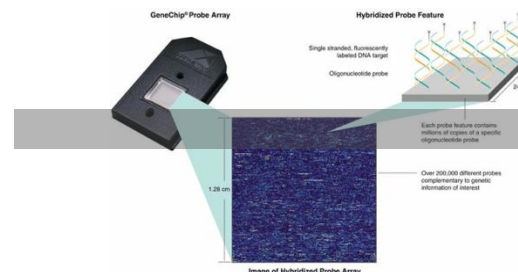
Background

- Aim: to identify the genes that are controlling a trait
- New interest in association mapping

Refinement tool

fine mapping
candidate gene studies

High-throughput genotyping technology



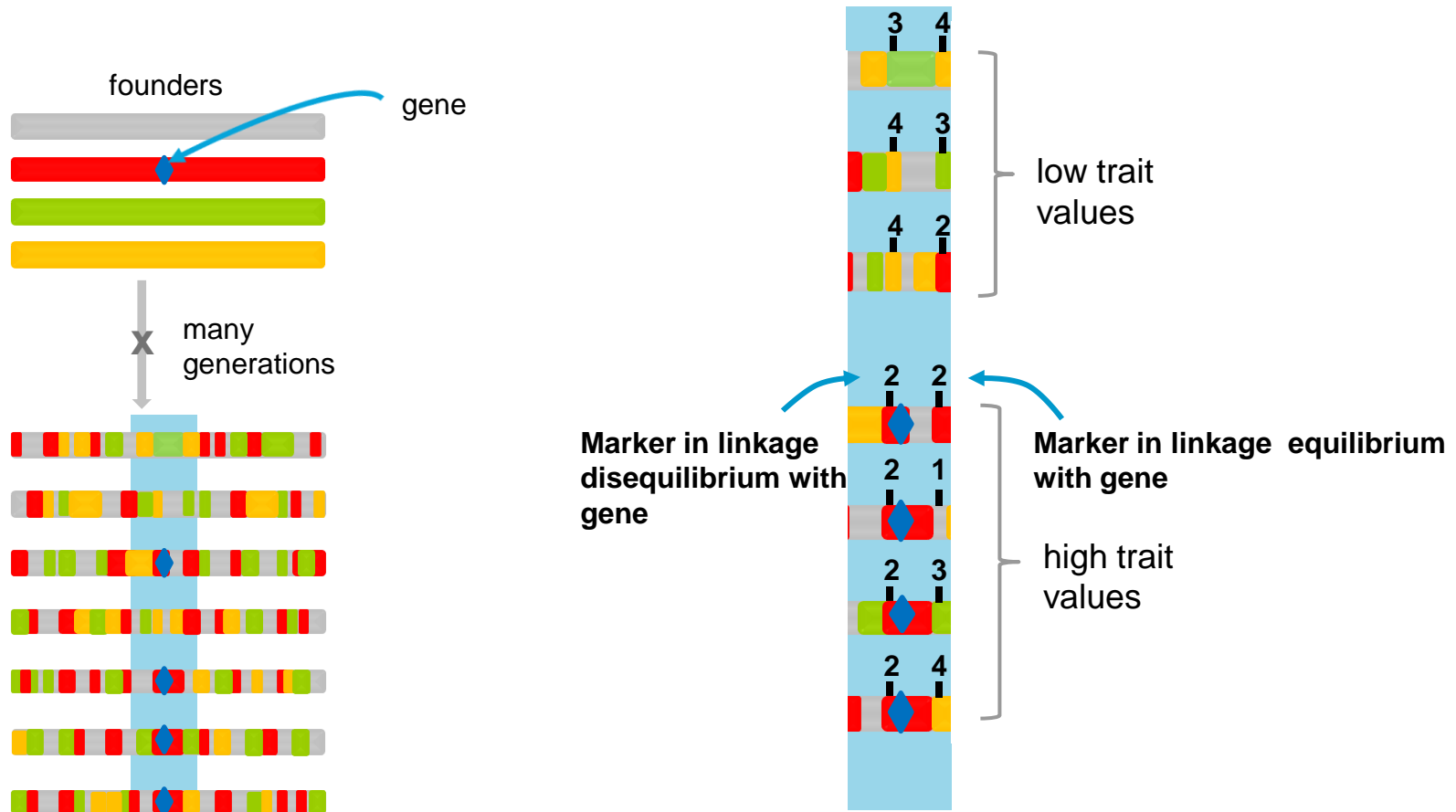
Discovery tool

genome-wide mapping

Background (Cont.)

- How

- Relies on the gene of interest being in linkage disequilibrium with an observed marker locus.



LMM-based association mapping

- Linear mixed model approach for association mapping
(Yu *et al* 2006, Zhao *et al.* 2007)

- A mixed linear model is formed for each marker.
- The strength of the marker-trait association is measured through the statistical significance of the fixed marker effect in the model.
- QK approach

$$Y = X\beta_{env} + \overbrace{Q\beta_{subpop}}^{\text{fixed}} + \overbrace{x_j\beta_j}_{\text{jth marker}} + \overbrace{Z_{env}u_{env} + Z_g u_g}^{\text{random}} + e$$

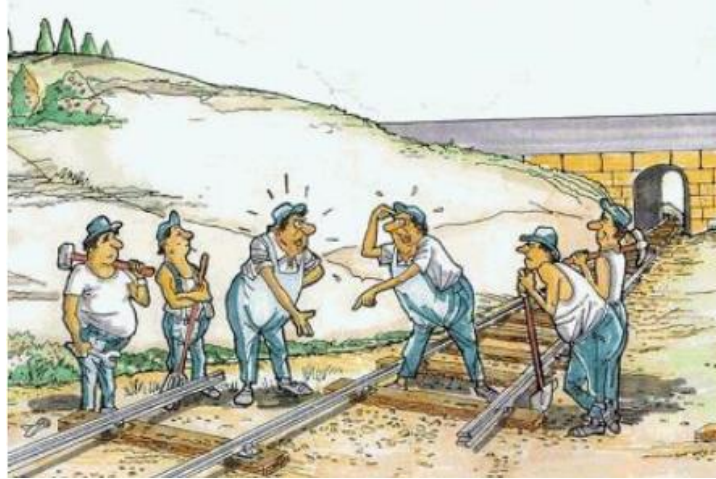
subpopulation familial relatedness

$$u_{env} \sim N(0, \sigma_{env}^2 G) \quad e \sim N(0, \sigma_e^2 R) \quad u_g \sim N(0, \sigma_g^2 K)$$

Pedigree-based
Marker-based

Strength of marker-trait association measured with Wald statistic of β_j

Problems in applying LMM association mapping to real problems



- **Computational**

- How to perform a large number of a single-marker analyses when the LMM is complex.

- **Statistical**

- How “best” to control the Type 1 error when a large number of correlated tests are performed.

Problem – Computational (Cont.)

Two-stage LMM strategy

- First stage captures most of the computational complexity and is performed only once.
- The second stage is performed for each maker.

Residuals approach

Stage 1:
$$Y = X\boldsymbol{\beta}_{env} + Z_{env}\boldsymbol{u}_{env} + e$$

Stage 2:
$$\hat{e} = Q\boldsymbol{\beta}_{subpop} + x_j\boldsymbol{\beta}_j + Z_g\boldsymbol{u}_g + e^*$$

$$e^* \sim N(0, \sigma_{e^*}^2 R^*)$$

Problem – Computational (Cont.)

Predicted variety means approach

$$\text{Stage 1: } Y = X\beta_{env} + x_{variety}\beta_{variety} + Z_{env}u_{env} + e$$

$$\text{Stage 2: } \hat{x}_{variety} = Q\beta_{subpop} + x_j\beta_j + Z_g u_g + e^*$$

$$e^* \sim N(0, \sigma_e^2 R^*)$$

Problem – Computational (Cont.)

Our approach to LMM-based association mapping

$$\text{Stage 1: } Y = X\beta_{env} + Q\beta_{subpop} + Z_{env}u_{env} + Z_g u_g + e$$

$$\text{Stage 2: } Y = X\beta_{env} + Q\beta_{subpop} + x_j\beta_j + Z_{env}u_{env}^* + Z_g u_g^* + e^*$$

$$u_{env}^* \sim N(0, \hat{\sigma}_{env}^2 G) \quad u_g^* \sim N(0, \hat{\sigma}_g^2 K) \quad e^* \sim N(0, \hat{\sigma}_e^2 R)$$

Stage 2 analysis is now a generalized least squares step of general form $Y = X^* \beta^* + e^*$ where $e^* \sim N(0, \hat{\sigma}^2 \hat{H})$

Wheat quality study

field trial



- multiple sites
- 467 cultivars
- Fields organized as rows x plots



milling process



Samples randomized over mill day and mill order



baking process



Samples randomized over bake day and bake order



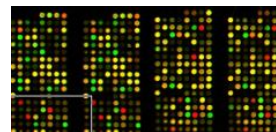
wheat quality traits

- loaf volume,
- loaf score
- loaf colour
- texture
- loaf height

Association Study



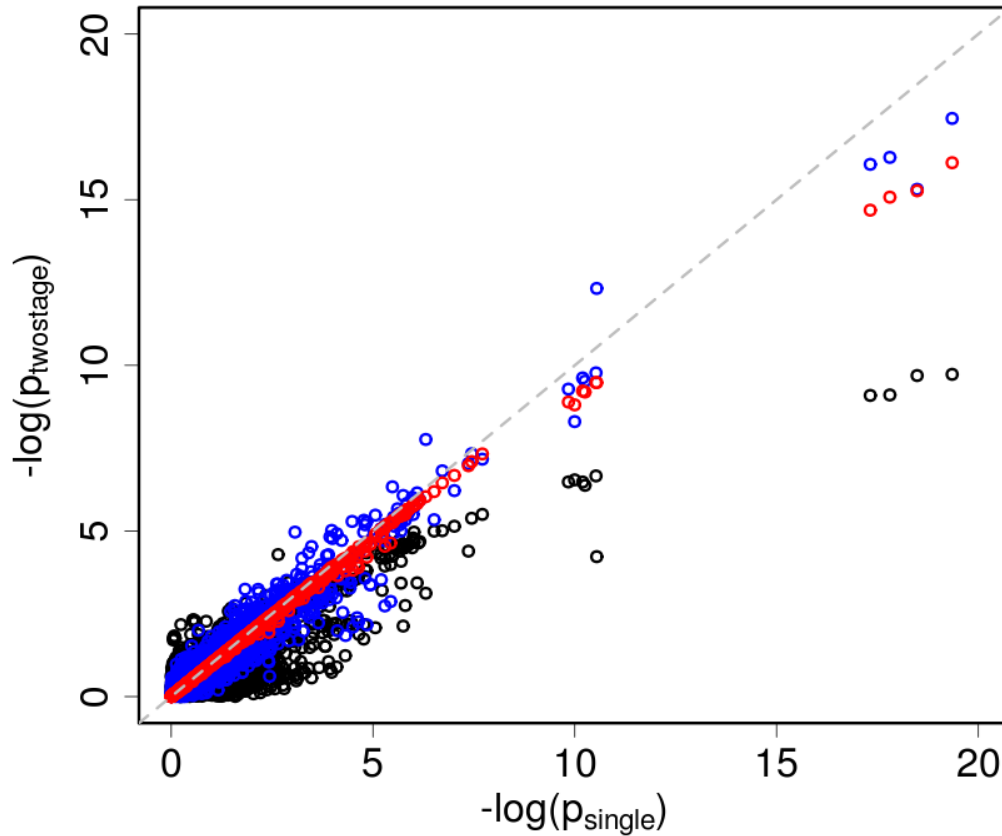
Existing study



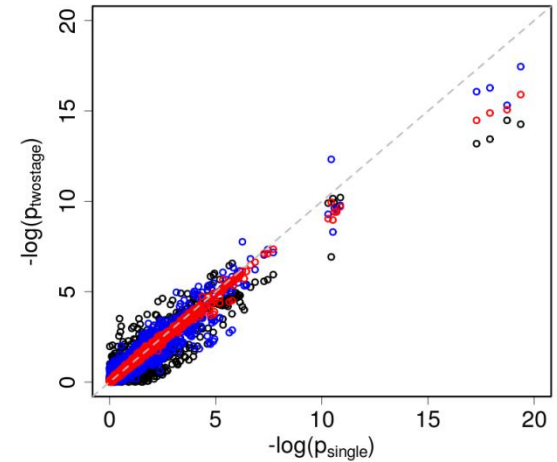
Genome-wide Association Study

Single-stage versus Two-stage analysis approaches

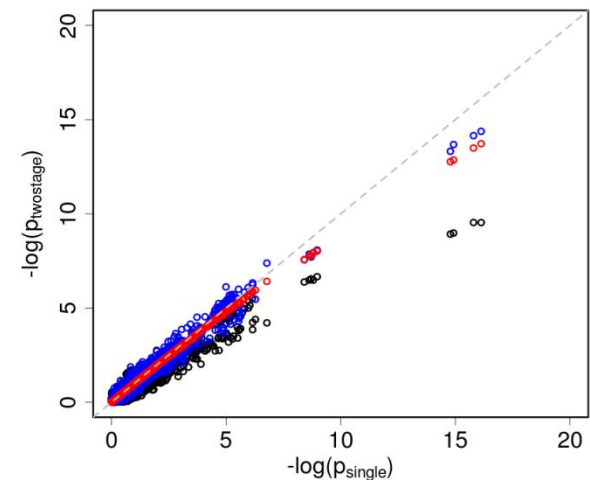
Correct model



Incorrect environmental model



Incorrect genetic model



Mathematics, Informatics, Statistics
Andrew W. George

Phone: 07 3833 5543

Email: andrew.george@csiro.au

Web: www.csiro.au/cmis

www.csiro.au

Thank you

Contact Us

Phone: 1300 363 400 or +61 3 9545 2176

Email: enquiries@csiro.au Web: www.csiro.au

