

# A whole of genome approach to QTL identification

Simon Diffey

December 7, 2006

## Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

## Motivation

- Develop an approach that allows appropriate statistical modelling of the phenotypic data.
- Recognises that QTL identification is essentially a model selection problem (Broman and Speed 2002).

## Outcome

- A whole of genome approach to QTL identification that avoids the issues associated with multiple testing.
- Allows the accommodation of non-genetic sources of variation in the model building process.

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTLs as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

## What are QTL?

- Genes or genomic regions that influence traits of interest.
- It is of interest to find the number, genomic region, and the genetic effect of QTL.

## QTL and plant breeding.

- QTL determination is a first step in the development of molecular markers and these markers are then used to indicate whether desirable genes or genomic regions exist in experimental breeding lines (hereafter called lines).
- Molecular marker technology is one of the Grain Research and Development Corporations (GRDC) largest single investments and is part of an overall strategy to provide grain growers with improved varieties quicker.

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

## Genotypic data.

- This involves genotyping or obtaining marker scores for each line.
- In a doubled haploid (DH) population this requires the lines to be scored according to the parental allele present at a set of molecular markers.

## Phenotypic data.

- Obtaining phenotypic data typically involves experiments with large numbers of lines.
- This will usually result in more than one observation on some or all of the lines.

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

## Standard approach

- Many statistical software packages for QTL analysis only allow one phenotypic observation for each line.
- Data from lines that have been replicated are either averaged, or replicates are examined separately.

Genotypic data					Phenotypic data			
line	m1	m2	m3	...	obs	line	rep	y
1	1	1	-1	...	1	1	1	.
2	-1	1	1	...	2	2	1	.
3	-1	-1	-1	...	3	3	1	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
$l$	1	-1	1	...	.	$l$	1	.
					.	1	2	.
					.	10	2	.
					.	33	2	.
					n	57	2	.



- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

## Alternative

- Why not expand the genotypic data to match the phenotypic data?

Genotypic data					Phenotypic data			
line	m1	m2	m3	...	obs	line	rep	y
1	1	1	-1	...	1	1	1	.
2	-1	1	1	...	2	2	1	.
3	-1	-1	-1	...	3	3	1	.
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$l$	1	-1	1	...	.	$l$	1	.
1	1	1	-1	...	.	1	2	.
10	-1	-1	-1	...	.	10	2	.
33	1	1	1	...	.	33	2	.
57	1	1	-1	...	n	57	2	.



- This would allow the non-genetic sources of variation to be accommodated in a statistical model.

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

## Standard QTL methods

- Interval mapping (Lander & Botstein 1989) or regression methods (Haley & Knott 1992, Martinez & Curnow 1992).
- Interval mapping proceeds by conducting a genome scan by stepping along the genome at regularly spaced intervals.
- At each of these steps the evidence for the existence of a putative QTL is considered.

## Some issues

- By stepping along the genome at regularly spaced distances the problem of multiple testing arises i.e tests are correlated if intervals lie on the same chromosome.
- If non-genetic effects are included in the statistical model these can change from one step to another.
- Interval mapping implementations that involve genome scans and accommodate non-genetic sources of variation can be time-consuming.

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

- A more natural approach would be to include all marker information simultaneously and at the same time have the ability to accommodate non-genetic sources of variation.

## A regression approach to interval mapping

Verbyla, A.P., Cullis, B.R., Thompson, R. (2006) The analysis of QTLs by simultaneous use of the full linkage map. *Theoretical and Applied Genetics*, accepted.

$$y|g = X\tau + Z_g g + Zu + \epsilon$$

$\tau$  is a  $t \times 1$  vector of fixed effects.

$u$  is a  $b \times 1$  vector of random effects assumed  $N(\mathbf{0}, \sigma^2 \mathbf{G}(\gamma))$ .

$\epsilon$  is the residual vector assumed  $N(\mathbf{0}, \sigma^2 \mathbf{R}(\phi))$ .

The vector of genotypic effects  $g$  is the main point of interest.

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

If there is a single QTL then the genetic effect for genotype  $i$  can be decomposed as

$$g_i = q_i a + p_i \quad \implies \quad g_i = \sum_{k=1}^c \sum_{j=1}^{m_k-1} q_{k;ij} a_{k;j} + p_i$$

$a$  is the size of a putative QTL.

$p_i$  is the residual or polygenic effect assumed  $N(0, \sigma^2 \gamma_g)$ .

$q_i$  is **unknown** but is either 1 or  $-1$  depending on the parental allele at the QTL.

$$y|g = X\tau + Z_g(Qa + p) + \epsilon$$

$$y|g = X\tau + Z_g Qa + Z_g p + \epsilon$$

where  $Q = [Q_1, Q_2, \dots, Q_c]$  is a matrix of unknown QTL scores.

$$y|g = X\tau + Z_g Qa + Z_g p + \epsilon$$

## Problem

- $q_i$ 's unknown.
- Do know the marker scores bounding an interval ( $m_L, m_R$ ).
- Can calculate a recombination fraction ( $r$ ) between two markers.

Replace  $q_i$  by its' expected value given two flanking markers.

$$E(q_i | m_L, m_R, r) = \lambda(m_L, m_R, r, r_{LQ}) \quad \text{Whittaker et. al. 1996}$$

$$E(Q | m_L, m_R, r) = M^g \Lambda$$

$M^g$  is a  $l \times m$  matrix of marker scores for each line and all  $m$  markers.  
 $\Lambda$  is a block diagonal matrix whose non-zero elements are functions of  $r$  and  $r_{LQ}$ .

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

$$d_{LQ} \sim U(0, d)$$

$$y|g = X\tau + Z_g M^g \Lambda a + Z_g p + \epsilon$$

Assume that a QTL can occur at any location within an interval.

$$d_{LQ} \sim U(0, d)$$

$d$  is the distance between two flanking markers.

$d_{LQ}$  is the distance between the left flanking marker and the QTL.

The distance  $d_{LQ}$  can be calculated using a distance measure such as Haldane's mapping function.

$$d_{LQ} = -\frac{1}{2} \log(1 - r_{LQ})$$

The  $d_{LQ}$  or  $r_{LQ}$  need to be integrated out to form a marginal distribution. However this is analytically intractable.

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

# Whole genome interval mapping

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

Replace  $\Lambda$  by its' expected value

$$\Lambda_E = E(\Lambda)$$

$\Lambda_E$  is a function of known variables. The full data model is

$$\mathbf{y}|g = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{M}^g\Lambda_E\mathbf{a} + \mathbf{Z}_g\mathbf{p} + \boldsymbol{\epsilon}$$

$$\mathbf{y}|g = \mathbf{X}\boldsymbol{\tau} + \mathbf{M}_E\mathbf{a} + \mathbf{Z}_g\mathbf{p} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

where  $\mathbf{M}_E = \mathbf{Z}_g\mathbf{M}^g\Lambda_E$  is a  $n \times (m - c)$  matrix.

The columns of  $\mathbf{M}_E$  are called **derived interval markers** and all intervals are included in a single analysis.

$$\mathbf{y}|g \sim N(\mathbf{X}\boldsymbol{\tau}, \sigma^2\mathbf{H}_E)$$

where  $\mathbf{H}_E = \mathbf{R} + \gamma_a\mathbf{M}_E\mathbf{M}_E^T + \gamma_g\mathbf{Z}_g\mathbf{Z}_g^T + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

Intervals on the linkage map can be classified into two groups.

## Intervals that do not contain a QTL

- Large in number.
- Size of QTL effect will be small.

## Intervals that contain a QTL

- Small in number.
- Size of QTL effect will reflect the presence of a QTL.

When a QTL is present in an interval the size of QTL effects represent outliers in comparison to the majority of intervals.

A method for detecting outliers can therefore be used to detect QTL's. The method used is the **alternate outlier model** (Cook et. al 1982, Thompson 1985, Gogel 1997).

A **score** based statistic is used for QTL detection.

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

Steps involved in the QTL detection process are

- Develop a model for the phenotypic data.
- Fit this model with and without random regression effects for the size of QTL.
- Use a REMLRT to determine if random QTL effects are significant.
- If significant process continues otherwise process terminated.
- Use a score based to statistic to find the chromosome most likely to contain a QTL. Then find the interval on that chromosome most likely to contain a QTL.
- Selected interval is transferred to fixed effects part of the model.
- Repeat process until REMLRT for random QTL effects is not significant.

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

The implementation of this method in R is built on the *samm* and *qtl* libraries. Four primary functions are involved

*ready.qtl.samm*

- Merges the derived interval markers into the phenotypic data file.

*samm.find.qtl*

- Implements the QTL detection process.
- Arguments include baseline linear mixed model and Type I error rate.

*summary.qtl*

- Summary function which provides z-ratios and LOD scores for intervals that were found to have putative QTL's.

*plot.qtl*

- Summary function which graphs the chromosome's on which putative QTL's were found and highlights the identified interval.

# Example - DH wheat population

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

## Field phase

- 190 DH lines from a WW2449 / CHARA cross grown in a field trial according to a grid-plot design.
- Single plots of each DH line and multiple plots of 4 check varieties and the two parental lines.
- 264 plots in a 22 row  $\times$  12 column array.

## Milling phase

- 222 field plots were milled.
- 58 of 222 grain samples were duplicated, leaving 164 singles.
- 280 samples milled 10 per day for 28 days.

## Extensibility phase

- 60 of the mill samples were duplicated leaving 220 singles.
- 340 flour samples tested as 10 per day for 34 days.
- Duplicated samples assigned to 2 blocks of days (1-17 and 18-34).



# Example - QTL model

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTLs as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

Decomposition	Model term
eblock	eblk
eblock.eord	
millday	mday
millday.millord	
line	
intervals	grp('qtls')
res line	line
res plot	row.col
res mill	mday.mord
res	units

## Inside the *samm.find.qtl* function

```
> m.QTL <- samm(ext ~ 1 + vtype + lfc, random=~ar1v(col):ar1(row) +
+                  mday + mday:mord + eblk +
+                  line + idv(grp('qtls'),15:312),
+                  data=CW$samm.ready.data)
```

# Example - QTL detection

Overview

Introduction

The data

Data issues

Data issues cont.

Analysis issues

A better way?

Genetic model

$E(q_i | m_L, m_R, r)$

$d_{LQ} \sim U(0, d)$

Whole genome interval mapping

QTL's as outliers

QTL detection process

R implementation

Example - DH wheat population

Example - phenotypic data

Example - QTL model

Example - QTL detection

Example - QTL detection cont.

Acknowledgements

## R *samm.find.qtl* function call

```
> QTL.ext <- samm.find.qtl(baseModel=samm(ext ~ 1 + vtype + lfc,  
+                                     random=~id + ar1v(fcol):ar1(frow) +  
+                                     mday + mday:mord + eblk,  
+                                     data=CW$samm.ready.data),  
+                               TypeI=0.01, parentData=CW)
```

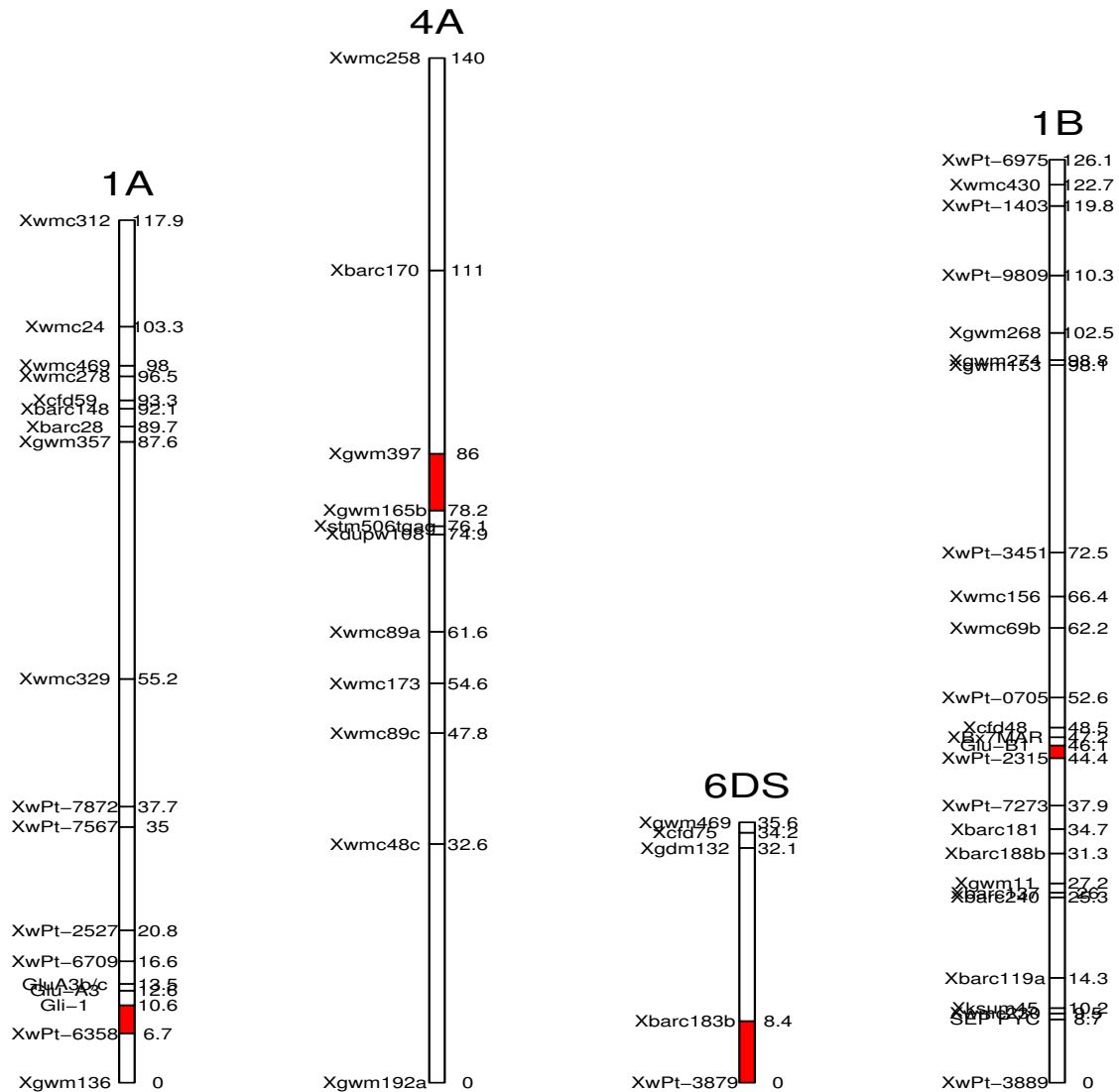
## R *summary.qtl* call and output

```
> summary.qtl(QTL.ext, CW)
```

	bound.markers	L/R	dist(cM)	z.ratio	LOD
C_1A_3	(XwPt-6358, Gli-1)	6.7 / 10.6	5.72	7.10	
C_4A_9	(Xgwm165b, Xgwm397)	78.2 / 86	4.50	4.40	
C_6DS_2	(XwPt-3879, Xbarc183b)	0 / 8.4	3.65	2.89	
C_1B_13	(XwPt-2315, Glu-B1)	44.4 / 46.1	12.07	31.64	

# Example - QTL detection cont.

## R *plot.qtl* output



- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

- Overview
- Introduction
- The data
- Data issues
- Data issues cont.
- Analysis issues
- A better way?
- Genetic model
- $E(q_i | m_L, m_R, r)$
- $d_{LQ} \sim U(0, d)$
- Whole genome interval mapping
- QTL's as outliers
- QTL detection process
- R implementation
- Example - DH wheat population
- Example - phenotypic data
- Example - QTL model
- Example - QTL detection
- Example - QTL detection cont.
- Acknowledgements

- Brian Cullis, Alison Smith, Ari Verbyla
- NSW DPI
- GRDC
- NSW Ag Genomics Centre