Single Marker Analysis
Backcross Population

- simplest association:
  genotype of markers
  phenotype of quantitative trait
- test for trait value differences
  between marker groups
- consider marker i in a backcross population
  genotypes: B_i/B_i versus B_i/b_i
- estimate mean and variance by marker group
- pool variance across groups
- compare using linear model
  anova: group means same or different
  regression: slope zero or not

T-test using anova model

$\tilde{\mu}_1$ = observed trait mean genotype M_i/M_i
$\tilde{\mu}_0$ = observed trait mean genotype M_i/m_i
hypotheses:

$$H_0: \ \mu_1 = \mu_0 \ \text{ and } \ H_1: \ \mu_1 \neq \mu_0$$

test for significance between $\tilde{\mu}_1$ and $\tilde{\mu}_0$:

$$t = \frac{\tilde{\mu}_1 - \tilde{\mu}_0}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}.$$

s^2 is the pooled sampling variance

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}$$

n_1, n_0 are sample sizes by marker class
s_1^2, s_0^2 are variances for each marker class

simple linear regression model

recast same problem using regression:

$$y_j = \mu + bx_j + e_j \quad j = 1, 2, \cdots, n \qquad (1)$$

where

$$\begin{aligned} y_j &= \text{the trait value of individual } j \\ \mu &= \text{mean of the model} \\ x_j &= \begin{cases} 1 & \text{if individual } j \text{ has } M_i/M_i \text{ genotype} \\ 0 & \text{if individual } j \text{ has } M_i/m_i \text{ genotype} \end{cases} \\ b &= \mu_1 - \mu_0 \\ e_j &\sim N(0, \sigma^2) \end{aligned}$$

test hypotheses on slope
$H_0: \ b = 0$ versus $H_1: \ b \neq 0$

$$\hat{b} = s_{xy}/s_x^2 \ \ Var(\hat{b}) = \sigma^2/s_x^2$$

$$t = \frac{\hat{b} - 0}{s/s_x}$$

this t-test is exactly same as the anova t-test

One marker analysis (t test): Mouse data

	Marker class 1			Marker class 0				
Marker	n_1	$\tilde{\mu}_1$	s_1^2	n_0	$\tilde{\mu}_0$	s_0^2	t	P value
1 Hmg1-rs13	41	54.20	111.81	62	47.32	63.67	3.754	0.0001
2 DXMit57	42	55.21	104.12	61	46.51	56.12	4.994	0.000001
3 Rps17-rs11	43	55.30	101.98	60	46.30	54.38	5.231	<0.000001
4 Rps18-rs17	42	55.60	100.69	61	46.25	53.66	5.467	<0.000001
5 DXMit48	43	55.19	105.20	60	46.38	53.60	5.085	<0.000001
6 DXNds1	44	55.43	102.25	59	46.05	50.12	5.538	<0.000001
7 DXMit109	45	55.00	114.05	58	46.22	44.77	5.103	<0.000001
8 Hmg14-rs6	49	54.86	105.83	54	45.70	43.19	5.431	<0.000001
9 DXMit60	50	54.62	106.49	53	45.75	43.88	5.218	<0.000001
10 DXMit16	50	54.68	106.10	53	45.70	43.22	5.306	<0.000001
11 DXMit97	50	54.64	107.05	53	45.74	43.01	5.248	<0.000001
12 Hmg1-rs14	51	53.90	104.61	52	46.29	54.88	4.333	0.000008
13 DXMit3	56	53.50	112.25	47	45.96	41.17	4.266	0.00001
14 Tpm3-rs9	49	53.02	126.06	54	47.37	50.01	3.085	0.001

all markers are significantly associated with QTL
t tests are all very significant

but

not clear how many QTL are on the chromosome
not clear where QTL are located
sample variances differ between marker classes

F_2 population

- three genotypes for marker i
  B_i/B_i, B_i/b_i, b_i/b_i
- two degrees of freedom for hypotheses
- two contrasts of interest
  marker additive effects
  marker dominance effects
- linear models approach
  t tests
  overall anova test with contrasts
  regression test of linear and quadratic effects

Marker i for F_2 Population

marker genotypes	B_i/B_i	B_i/b_i	b_i/b_i
observed trait means	$\tilde{\mu}_2$	$\tilde{\mu}_1$	$\tilde{\mu}_0$
sample sizes	n_2	n_1	n_0
variances	s_2^2	s_1^2	s_0^2

single marker summaries: maize data

	Marker class 2			Marker class 1			Marker class 0		
M	n_2	$\tilde{\mu}_2$	s_2^2	n_1	$\tilde{\mu}_1$	s_1^2	n_0	$\tilde{\mu}_0$	s_0^2
1	43	5.24	2.44	86	4.27	2.93	42	3.11	2.76
2	48	4.82	3.15	89	4.17	3.26	34	3.54	2.84
3	42	5.01	3.23	92	4.14	3.18	37	3.57	2.68
4	44	4.47	2.96	89	4.21	3.36	38	3.99	3.61
5	43	4.57	3.13	87	4.21	3.37	41	3.91	3.28
6	43	4.48	3.03	83	4.06	2.85	45	4.29	4.43
7	44	4.28	3.09	83	4.09	3.01	44	4.44	4.14
8	47	4.36	2.73	81	4.09	3.35	43	4.34	3.93
9	41	4.16	2.36	86	4.21	3.62	44	4.32	3.68
10	40	3.94	2.75	94	4.30	3.68	37	4.34	2.99
11	45	4.25	3.64	89	4.14	3.32	37	4.41	2.97
12	46	4.04	3.25	85	4.25	3.54	40	4.40	2.94

t test for marker additive effect

$$t_1 = \frac{\tilde{\mu}_2 - \tilde{\mu}_0}{\sqrt{s^2\left(\frac{1}{n_2} + \frac{1}{n_0}\right)}}$$

t test for marker dominance effect

$$t_2 = \frac{\tilde{\mu}_1 - \tilde{\mu}_2/2 - \tilde{\mu}_0/2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{4n_2} + \frac{1}{4n_0}\right)}}$$

pooled variance estimate

$$s^2 = \frac{(n_2-1)s_2^2 + (n_1-1)s_1^2 + (n_0-1)s_0^2}{n_2 + n_1 + n_0 - 3}$$

single marker analysis (t test): maize data

	additive		dominance	
M	t_1	P value	t_2	P value
1	6.10	<0.000001	0.38	0.704
2	3.28	0.001	-0.05	0.958
3	3.71	0.0002	-0.57	0.567
4	1.20	0.230	-0.05	0.958
5	1.68	0.093	-0.13	0.897
6	0.46	0.646	-1.18	0.238
7	-0.39	0.698	-0.96	0.337
8	0.06	0.952	-0.92	0.358
9	-0.41	0.684	-0.10	0.920
10	-1.04	0.298	0.57	0.567
11	-0.38	0.704	-0.66	0.509
12	-0.94	0.347	0.09	0.928

t_1 only uses variance pooled from $x = 0, 2$
only t_1 for marker 1, 2 and 3 are significant
could be one or two QTL on the chromosome
QTL effects are largely additive

test can be performed through anova
overall effect (2 d.f.)
selected contrasts

Robustness to Lack of Normality

- t test assumes normal trait distributions within marker classes
- test is quite robust to violations of the assumption
- trait is at best a mixture of normals
  even if normal at true QTL
  mixtures due to segregation of genes
  Doerge (1993): t test generally performs well with mixtures
- departures of the normal distributions hard to detect in segregating population
  most pronounced when means differ

Genetical meaning of the analysis

To understand the relevance of this test to QTL mapping, we need to know what is being tested in genetic terms. Consider a QTL Q linked to a marker B. For a backcross $\frac{BQ}{BQ} \times \frac{BQ}{bq}$, we have

	$\frac{BQ}{BQ}$	$\frac{BQ}{Bq}$	$\frac{BQ}{bQ}$	$\frac{BQ}{bq}$
frequency	$(1-r)/2$	$r/2$	$r/2$	$(1-r)/2$
trait mean	$\mu+a$	$\mu+d$	$\mu+a$	$\mu+d$

	QQ		Qq		
	freq	mean	freq	mean	mixture mean
BB	$1-r$	$\mu+a$	r	$\mu+d$	$(1-r)(\mu+a)+r(\mu+d)$
Bb	r	$\mu+a$	$1-r$	$\mu+d$	$r(\mu+a)+(1-r)(\mu+d)$

$$\begin{aligned}\mu_{BB}-\mu_{Bb} &= [(1-r)(\mu+a)+r(\mu+d)] \\ &\quad -[r(\mu+a)+(1-r)(\mu+d)] \\ &= (1-2r)a-(1-2r)d \\ &= (1-2r)(a-d)\end{aligned}$$

multiple QTL linked to marker B_i

(ignoring epistasis)

$$\mu_{BB}-\mu_{Bb}=\sum_{k=1}^{m}(1-2r_k)(a_k-d_k)$$

testing composite parameter
multiple QTL effects *and* recombination frequencies
many QTL may not be linked to the marker
recombination frequency = 0.5

$$H_0:\ \mu_{BB}=\mu_{Bb} \quad \text{and} \quad H_1:\ \mu_{BB}\neq\mu_{Bb}$$

is equivalent to

$$H_0:\ \text{all } r_k=0.5 \quad \text{and} \quad H_1:\ \text{at least one } r_k<0.5$$

we "know" some genes segregate in the population
by experimental design: choice of parent lines
some a_k-d_k are nonzero

can only detect linked QTL

$\tilde{\mu}_{i1}$, $\tilde{\mu}_{i0}$ significantly different
implies marker B_i is linked to one or more QTL

Likelihood analysis

likelihood analysis on a single marker
consider marker B, trait y
single QTL Q linked to marker (r_{BQ})

	$\frac{BQ}{BQ}$	$\frac{BQ}{Bq}$	$\frac{BQ}{bQ}$	$\frac{BQ}{bq}$
frequency	$(1-r_{BQ})/2$	$r_{BQ}/2$	$r_{BQ}/2$	$(1-r_{BQ})/2$
distribution	$N(\mu+\delta,\sigma^2)$	$N(\mu,\sigma^2)$	$N(\mu+\delta,\sigma^2)$	$N(\mu,\sigma^2)$

	QQ	Qq	Distribution of y
BB	$1-r_{BQ}$	r_{BQ}	$(1-r_{BQ})N(\mu+\delta,\sigma^2)+r_{BQ}N(\mu,\sigma^2)$
Bb	r_{BQ}	$1-r_{BQ}$	$r_{BQ}N(\mu+\delta,\sigma^2)+(1-r_{BQ})N(\mu,\sigma^2)$

Likelihood:

$$L(\mu,\delta,\sigma^2,r_{BQ})=$$

$$\prod_{j=1}^{n_{BB}}\left[(1-r_{BQ})\phi\left(\frac{y_{1j}-\mu-\delta}{\sigma}\right)+r_{BQ}\phi\left(\frac{y_{1j}-\mu}{\sigma}\right)\right]$$

$$\times\prod_{j=1}^{n_{Bb}}\left[r_{BQ}\phi\left(\frac{y_{2j}-\mu-\delta}{\sigma}\right)+(1-r_{BQ})\phi\left(\frac{y_{2j}-\mu}{\sigma}\right)\right]$$

where $\phi(z)=\frac{1}{\sqrt{2\pi}}\exp[-z^2/2]$.

hypothesis $H_0 : r_{BQ} = 1/2$
tested by likelihood ratio

$$LR = -2 \ln \frac{L(\hat{\mu}, \hat{\delta}, \hat{\sigma^2}, r_{BQ} = 1/2)}{L(\hat{\mu}, \hat{\delta}, \hat{\sigma^2}, \hat{r}_{BQ})}$$

very specific model and test: assume only one QTL
Q segregating in population
ask whether Q is linked to B or not
if linked, estimate r_{BQ}

Problems of the analysis

This simple analysis captures basic ideas of QTL mapping. There are many problems (McMillan and Robertson 1974; Lander and Botstein 1989) such as:

1. The method cannot tell whether the markers are associated with one or more QTL;
2. The method does not estimate the likely positions of the QTL;
3. The effects of QTL are likely to be underestimated because they are confounded with the recombination frequencies;
4. Because of the confounding effects, the method is not very powerful and many individuals are required for the test.