Wednesday am

Mapping markers

Mapping genetic Markers

The mapping process can be broken down into three stages:

Are markers linked?

What is marker order?

What are inter-marker distances?

Some definitions

Recombination is a result of chiasmata. A single chiasma will leave two recombinant and two non recombinant chromosomes

Map distance m is the expected number of crossovers = ½ the number of chiasmata in a length of chromosome.

Map distance is linear but recombination fraction can never be >0.5.

Recombination fraction = $(1-p_{(\text{zero chiasmata})})/2$ (Mather)

For a small map distance m,

$p_{(\text{one chiasma})} = 2m$

$p_{(\text{no chiasmata})} = (1-2m)$.

and recombination fraction = map distance

This is not the case at longer distances.

Larger distances

Assume a Poisson distribution of chiasmata in any interval.

In map distance m we require 2m chiasmata

P(no chiasmata) = e^{-2m}

giving

$r = 0.5(1- e^{-2m})$

$m = -0.5(\ln(1-2r)$

This is the Haldane mapping function.

Are markers linked?

Example: A backcross of AaBb to aabb gave:

	AB	Ab	aB	ab	total
observed	27	22	19	32	100
expected	N(1-r)/2	Nr/2	Nr/2	N(1-r)/2	
expected	25	25	25	25	

Chi-squared = 3.92 (3 df) p = 0.270

Three 1df tests:

A:a	0.04	(p-value	0.841)
B:b	0.64	(p-value	0.424)
linkage	3.24	(p-value	0.072)

Two tailed or one tailed tests?

LRT for linkage

Same example:

log likelihood at r = 0.5 = 100 ln(0.5)
= -69.315

log likelihood at r = 0.41 = 41 ln (0.41) + 59 ln(0.59)
= -67.686

LRT = 2*(69.315 – 67.686)
= 3.258

Same as before.

Mapping genetic Markers

Segregation distortion? Use contingency table:

	observed			expected	
	B	b		B	b
A	27	22	A	22.54	26.46
a	19	32	a	23.46	27.54

chi-sq = 3.20 , very close to the previous value.

Assigning markers to linkage groups

Require high significance thresholds to account for multiple testing.

Common is a LOD of 3 – corresponding to a 1-tailed p-value of 0.0001.

Ordering markers

SAR	lowest sum of adjacent recombination coefficients.
SAL	sum of adjacent likelihoods or LODs

But as number of markers rises, it is harder to examine all possible orders. Therefore:

Seriation	add one marker at a time. Test final order with “rippling” and “flipping.”
Branch and bound	mimimises total number of recombinations.

Joinmap approach.

PCA

Simulated annealing

The Joinmap approach

Take a matrix of all pairs of recombination fractions.

Fit a model (the map) to generate predicted pairwise r.f.

Minimise the error sum of squares. Use weighted least squares to account for differences in precision of estimation of pairs.

Gold standard method, but the software is expensive.

Threadmapper http://cbr.jic.ac.uk/threadmapper/

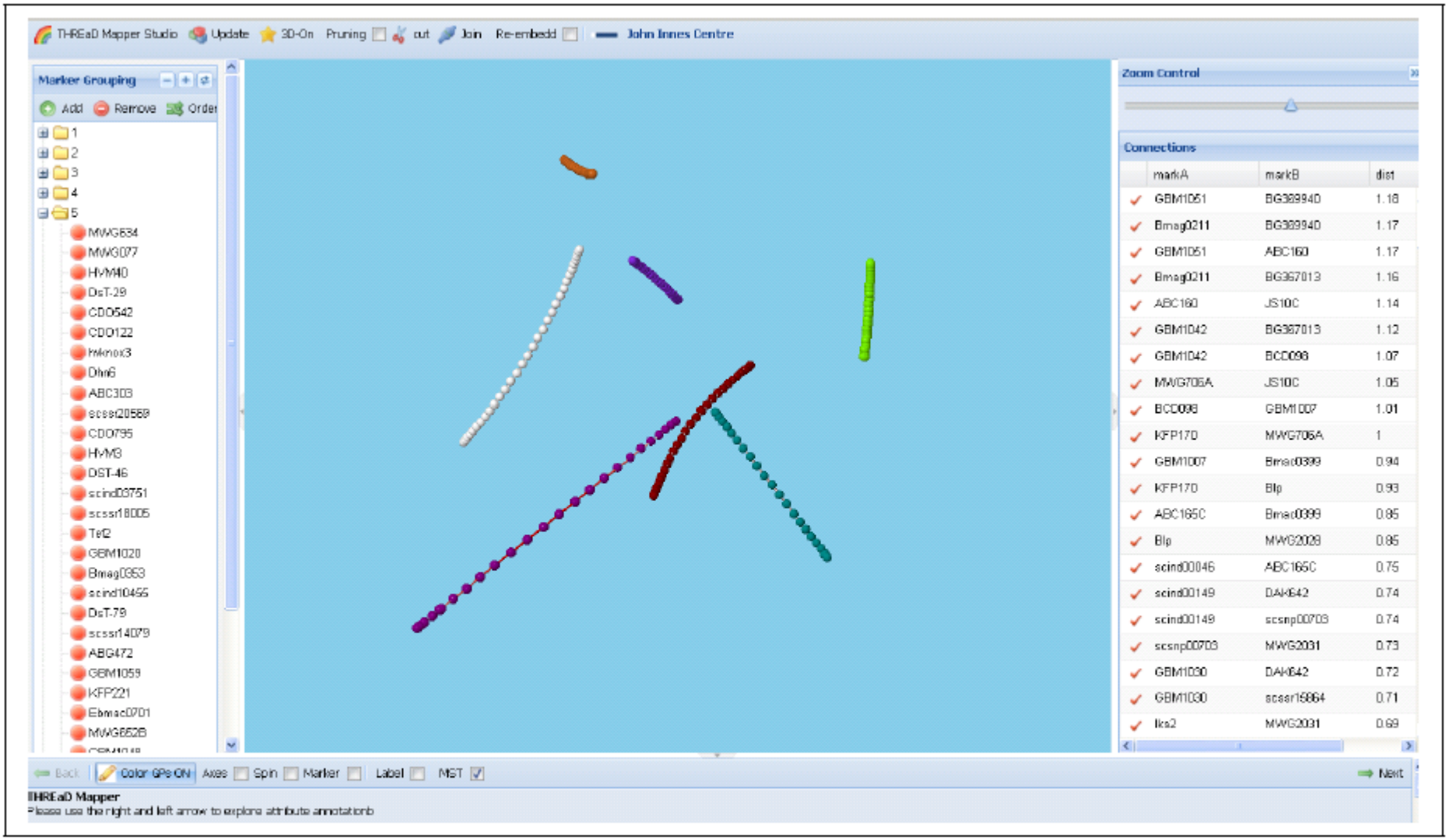


Figure 24: A Spectral embedding of the barley 7 linkage group dataset seen in Figures 21 and 22 within THREaD Mapper Studio. The 7 linkage groups are clearly distinct.

Mapping functions again.

Three loci:

$$r_{ab} = r_{ac}(1-r_{bc}) + (1-r_{ac})r_{bc}$$

$$= r_{ac} + r_{bc} - 2r_{ac}r_{bc}$$

Not linear but $(1-2r_{ab}) = (1-2r_{ac})(1 – 2r_{bc})$

and

$$\ln(1-2r_{ab}) = \ln(1-2r_{ac}) + \ln(1 – 2\, r_{bc})$$

$$-1/2\, \ln(1-2r_{ab}) = -1/2\, \ln(1-2r_{ac}) + -1/2\, \ln(1 - 2\, r_{bc})$$

$-1/2\, \ln(1-2r)$ is the Haldane mapping function.

when r is small:

$-1/2\, \ln(1-2r) \sim r$ (remember the maths?)

Mapping functions again.

$$r_{ab} = r_{ac} + r_{bc} - 2r_{ac}r_{bc}$$

This relationship, which assumes that recombination in each of the two intervals is independent is often found to fail.

Quantify the failure by the “coefficient of coincidence” c.

$$r_{ac} = r_{ab} + r_{bc} - 2cr_{ab}\theta_{bc}$$

c = 1 gives the Haldane mapping function

c → 0 when r → 0 and c = 1 at r = 0.5 gives the Kosambi mapping function, is generally thought to fit data better.

Some software uses one function, some the other.

Spreadsheet provided to convert from one to the other.

Mapping genetic Markers

The effect of errors

Errors mimic recombination and increase map length.

Detection:

Check out double recombinants. (K&M suggest any within 15cM)

Drop a genotype at a time and measure the effect on map length or likelihood. (Not strictly a LRT.)

Mapping populations

Backcross

F2 population

Inbred lines and doubled haploids derived from F2 or F1.

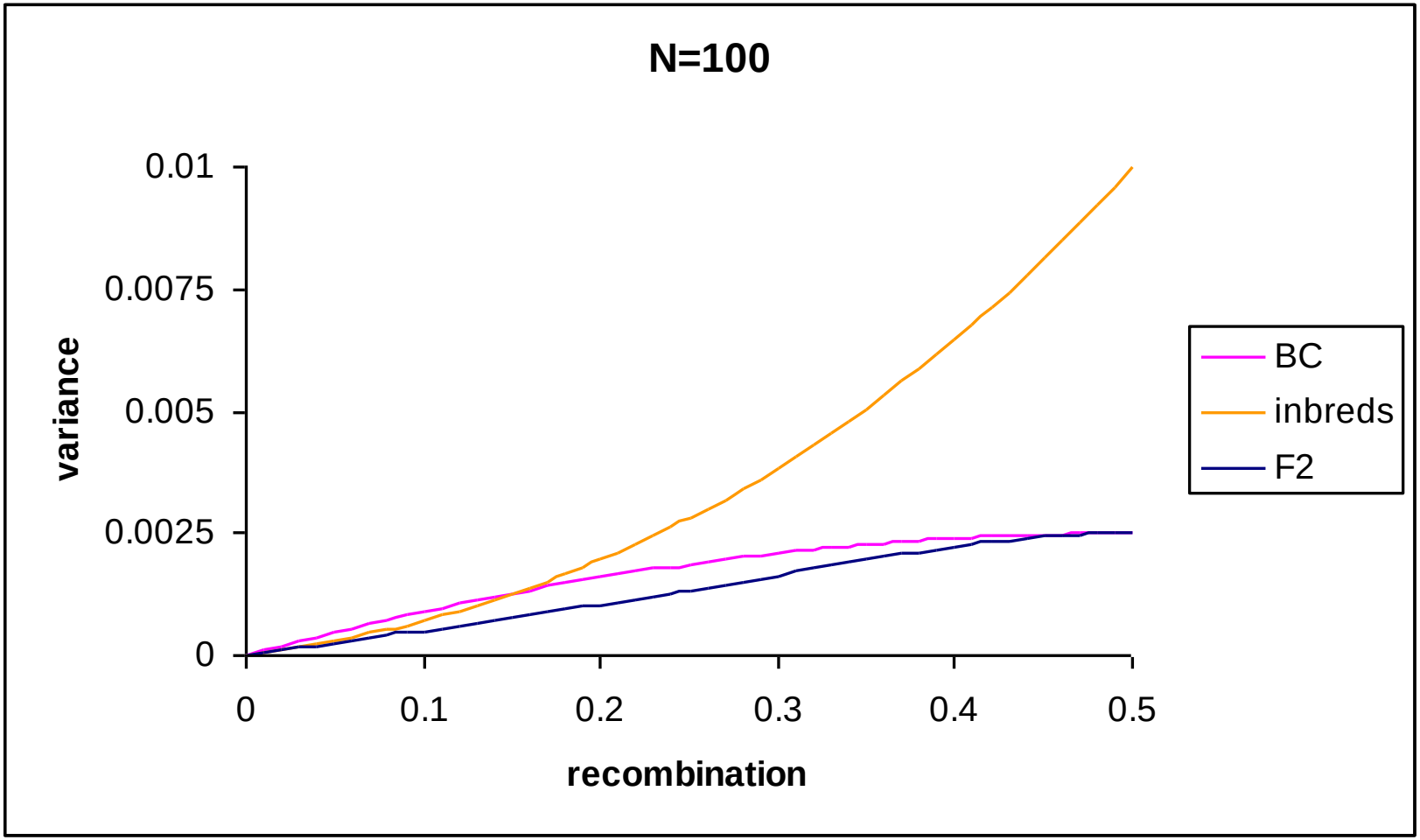
Inbred lines and DH derived from a backcross.

Full-sib families = 4 way crosses.

Extended/ mixed pedigrees

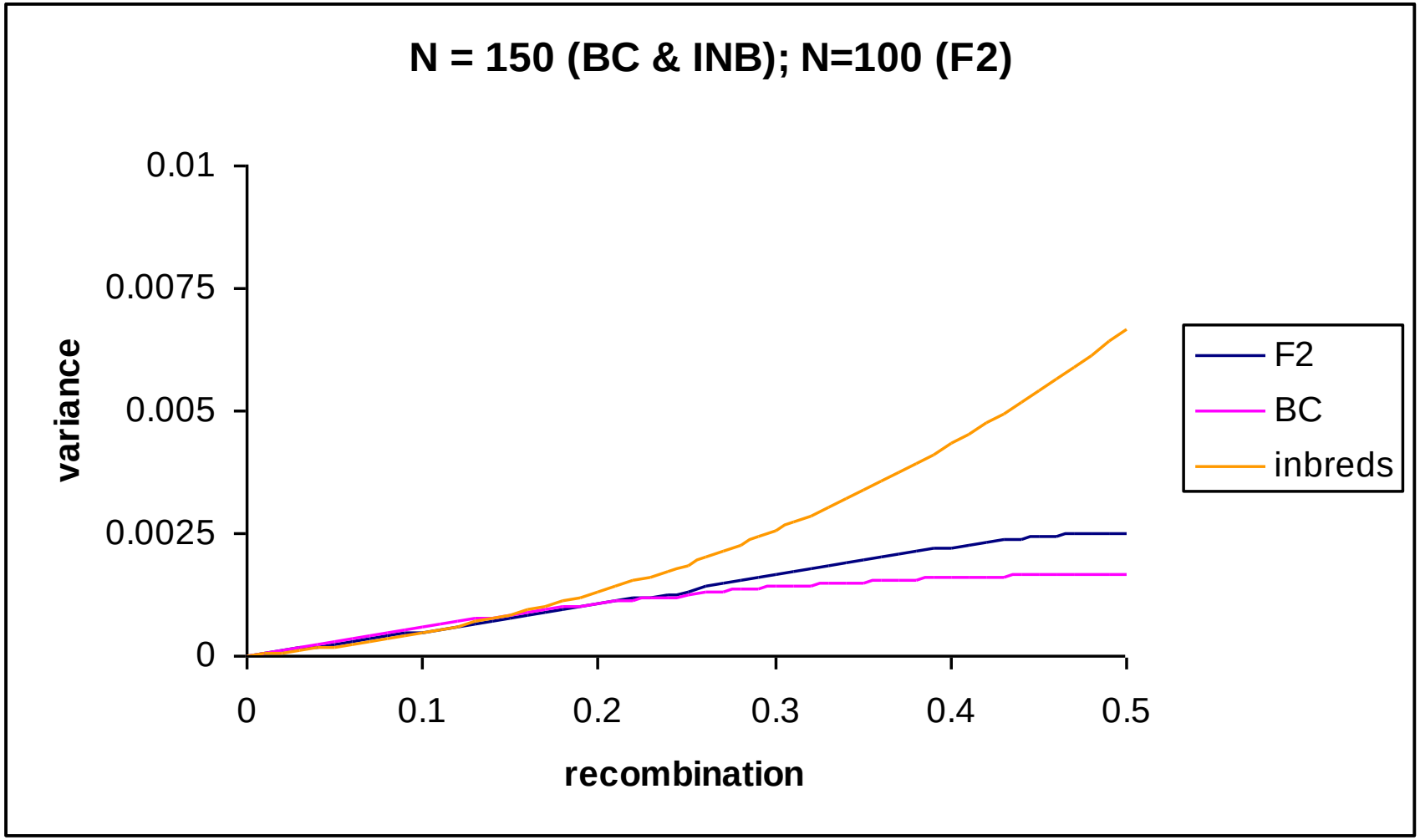
Mapping populations

Co-dominant markers.



Mapping populations

Rough equivalence with >pop size for BC and inbreds.



SSD versus DH lines

DH populations are exactly like a backcross for power and precision.

Map expansion on inbreeding:

Observed recs / total for inbreds = R

Observed recs / total for DH = r

$R = 2r/(1+2r)$

$r = R/2(1-R)$

If r is small, then $R \sim 2r$:

there is roughly 2x as much recombination in small intervals for inbred lines.

Take care when transferring maps and software.

Mapping populations

How many markers do we need?

Very easy to simulate, even in Excel.

Complicated formulae or simulate.

We need more to create the map than we need to map QTL!

Mapping populations

Finally

The finished map is an approximation.

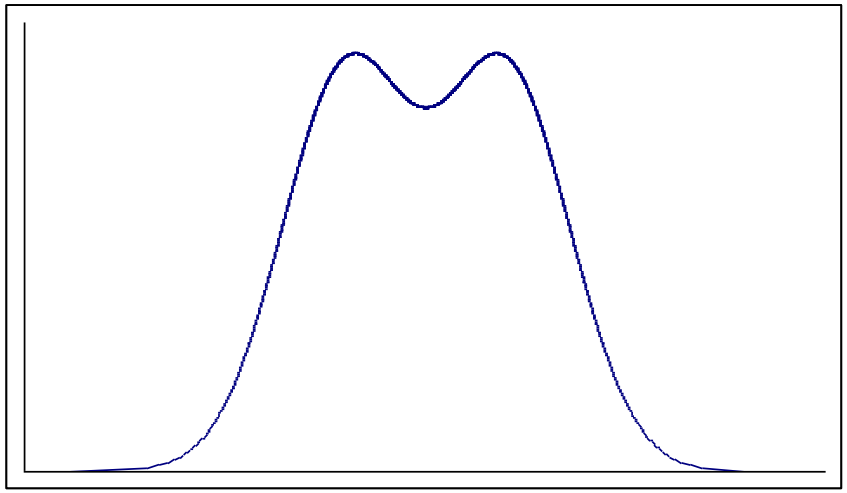
The markers are unlikely to be ordered correctly.

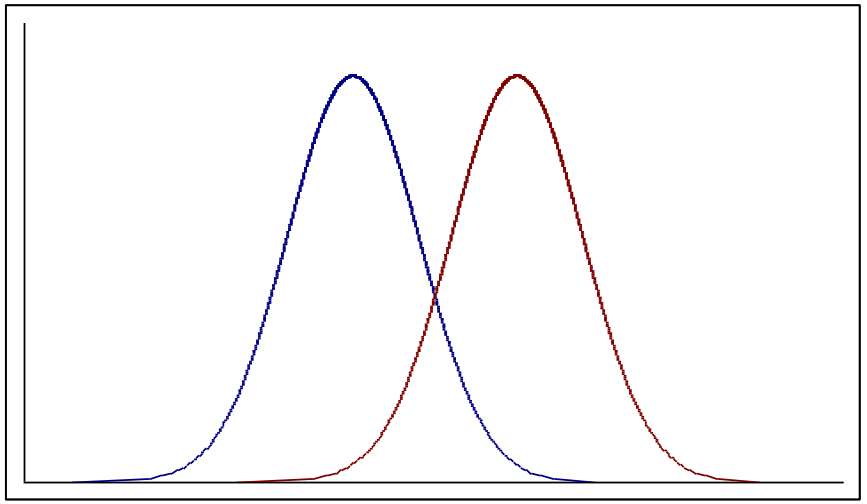
Intermarker distances are estimates only.

Wednesday pm

Mapping traits

Detecting major genes





Observed distribution is a mixture of underlying distributions

Detecting major genes

Fit mixture models

Likelihood of observation z_i=

= (probability that z_i is in group 1) x (the pdf of z_i given that it is in group 1)

+ (probability that z_i is in group 2) x (the pdf of z_i given that it is in group 2)

$$l_{z_i} = p_1 \phi_{1 z_i} + p_2 \phi_{2 z_i}$$

Over all observations:

$$l_{\underline{z}} = \prod_{i=1}^{n} l_{z_i}$$

Single markers - likelihood

$P_{QQ|MM} = (1-r)^2$

$P_{Qq|MM} = 2r(1-r)$

$P_{qq|MM} = r^2$

$P_{QQ|Mm} = r(1-r)$

$P_{Qq|Mm} = (1-r)^2 + r^2$

$P_{qq|Mm} = r(1-r)$

$P_{QQ|mm} = r^2$

$P_{Qq|mm} = 2r(1-r)$

$P_{qq|mm} = (1-r)^2$

Single markers - likelihood

These allow us to write down, just as for the mixture model, likelihood (QTL state | marker class):

Eg for an MM individual:

$$l_{z_i} = (1-r)^2 \phi_{QQ} + 2r(1-r)\phi_{Qq} + r^2 \phi_{qq}$$

and over all individuals $\prod_{i=1}^{n} l_{z_i}$

Single markers - ANOVA

Just test for a difference in means between the marker classes

These are

$$\frac{\mu_{MM} - \mu_{mm}}{2} = a(1-2r) = a'$$

$$\mu_{Mm} - \frac{(\mu_{MM} + \mu_{mm})}{2} = d(1-2r)^2 = d'$$

Can’t distinguish between large effects and close linkage.

Selective genotyping and BSA

Easier to understand and test for than to write down the expected values (but this has been done).

Can approximate and model using truncated mixture models.

Multiple Marker Methods: Maximum Likelihood

Flanking markers gametes

$$p_{M_1QM_2} = (1-r_1)(1-r_2)/2$$
$$p_{M_1Qm_2} = (1-r_1)r_2/2$$
$$p_{M_1qM_2} = r_1r_2/2$$
$$p_{M_1qm_2} = r_1(1-r_2)/2$$
$$p_{m_1QM_2} = r_1(1-r_2)/2$$
$$p_{m_1Qm_2} = r_1r_2/2$$
$$p_{m_1qM_2} = (1-r_1)r_2/2$$
$$p_{m_1qm_2} = (1-r_1)(1-r_2)/2$$

Multiple Marker Methods: Maximum Likelihood

Flanking markers gametes → zygotes

Eg:

$$P_{M_1M_1QQM_2M_2} = \left[(1-r_1)(1-r_2)/2\right]^2$$

As map of markers is known, only a single recombination fraction need be estimated.

Multiple Marker Methods: work on means

$$\frac{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}{2} \approx a(1-2r_1r_2) \approx a$$

after which, substituting back into single marker expectations:

$$r_1 = \frac{1}{2}\left(1 - \frac{\mu_{M_1M_2} - \mu_{m_1m_2}}{2a}\right)$$

Since r_1r_2 will be small.

Multiple Marker Methods: Kearsey & Hyne 1994

$$\frac{\mu_{M_iM_i} - \mu_{m_im_i}}{2} = a(1-2r_i)$$

1) Assume a location for the QTL.

2) Regress the difference in means for each marker on $(1\text{-}2r_i)$ Fix the intercept at zero.

3) If the QTL is located correctly, the slope will be an estimate of a and the error SS will be small.

4) If the slope is located incorrectly, the SS will be large.

5) Slide the assumed location of the QTL along the chromosome to minimimise the SS

Simple, but has never caught on.

Software exists – QTL café.

Interval mapping by least squares regression

Works on pairs of adjacent markers.

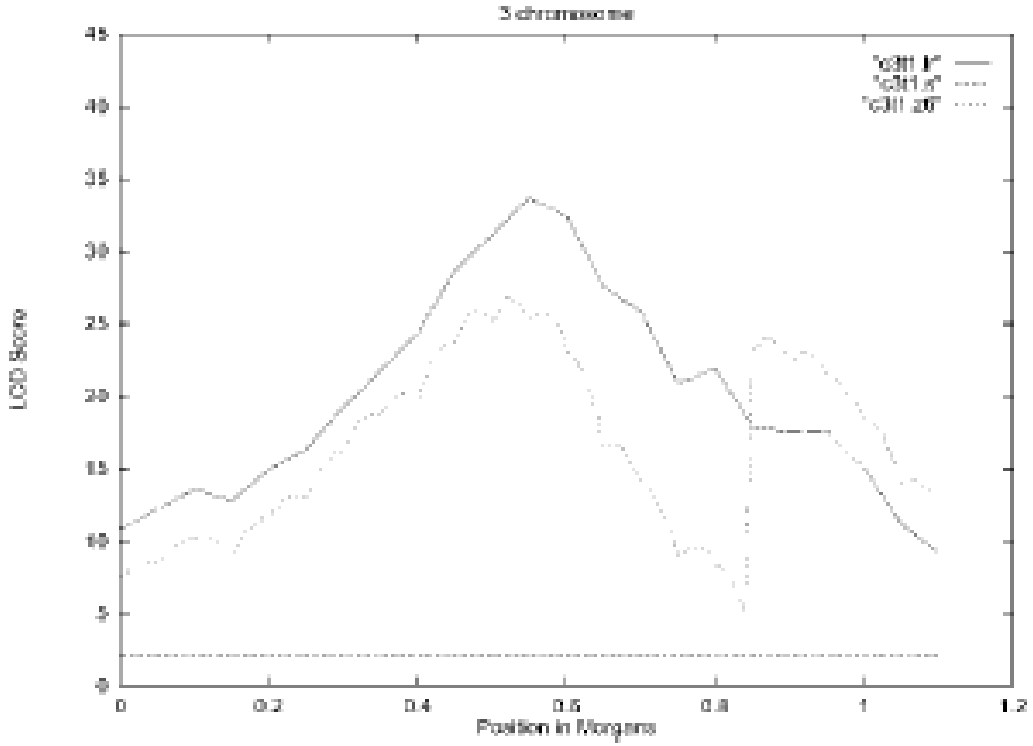
Write down the expected value of the mean for each (two locus) marker class as a function of QTL effect and recombination fractions. This is complicated. Eg, for $M_1M_1M_2M_2$ for an F2:

$$\mu_{M_1M_1M_2M_2} = \mu + a\left[\frac{(1-r_1)^2(1-r_2)^2 - r_1^2 r_2^2}{(1-r_{12})^2}\right] + d\left[\frac{2r_1r_2(1-r_1)(1-r_2)}{(1-r_{12})^2}\right]$$

Do this for all marker classes then regress the phenotype on the coefficients given in the square brackets. The regression coefficients give values of u, a and d. Vary r to get the minimum error SS.

Widely used, very popular, quicker than ML estimation

How many QTL might we detect?



There are no large QTL in this plot: driven by variation in gene density and recombination fraction

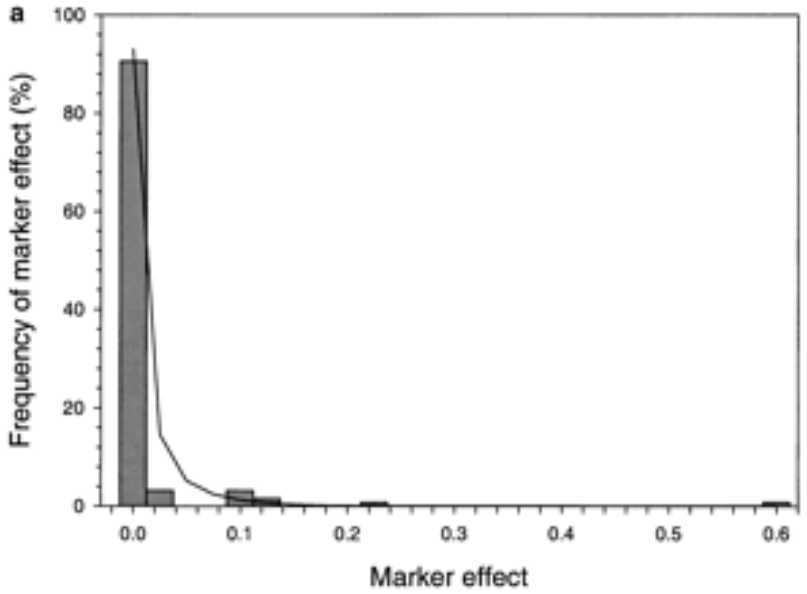
A more sophisticated version of “ghost QTL.”

The Beavis effect

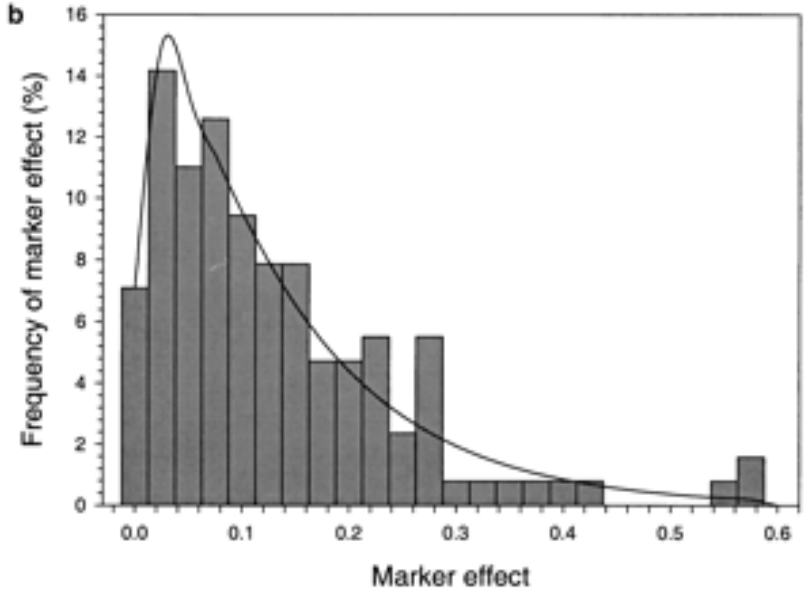
With multiple QTL of small effect, some get lucky and are detected.

These are genuine QTL, but their estimated effect is massively overestimated.

What is the distribution of QTL effects?



Bayesian analysis, effects allegedly unbiased



Single marker regression, showing the Beavis effect.

Detecting multiple QTLs :Composite Interval Mapping

Really just interval mapping with covariates.

The covariates can be other markers (as surrogate QTL).

Same problems as with any modelling exercise – what to include and what not.

Suggested is to:

include no more than $2\sqrt{n}$ unlinked markers

include the nearest neighbouring pair too.

Multiple QTL mapping

Scans multiple intervals simultaneously.

In practice there is a limit to how many you can scan at a time, so some form of pre-selection is required, or you stick to low numbers of intervals.

R/QTL scans all pairs of intervals.

Mapping in half sib families

Simplest case:

We map using loci heterozygous in the common (usually female) parent.

Depending on the (unknown) male marker genotype, It is not always possible to distinguish maternal alleles – in which case that data point cannot be used.

A t-test for a difference between the maternal marker classes is then a test for linkage.

The expected value for the difference is: $(1-2\theta)(a + (q-p)d]$

The square of this term (ignoring E) is $(1-2\theta)^2 Va_{qtl}$

Mapping in half sib families

Combine over families:

As the square of this term (ignoring E) is $(1-2\theta)^2 Va_{qtl}$

Straight forward extension to pooling over multiple half-sib families though an analysis of variance. (Needs the maternal parents to be heterozygous for the QTL.

Mapping in full sib families

More complex than half sibs, but principles are the same:

test for differences in marker classes.

Fully informative markers are best:

M1M2 x M1M2: heterozygous progeny are not informative.

Can combine analysis over families using a nested ANOVA.:

See Lynch & Walsh for details.

Many small pedigrees: Use human genetics software: Merlin

Mapping in multi-founder experimental populations

NAM: Nested association mapping

MAGIC: Multi parent advanced intercross

Increased diversity means more QTL can be mapped.

Increased precision

An alternative to association mapping (tomorrow.

Nested Association Mapping

Developed for maize by Ed Buckler, Cornell.

Crosses of the form: AxB, AxC,AxD…AxZ

Analyse within crosses: linkage information

& between cross: association

High precision.

Results may depend on the choice of “A”

More recent suggestion is to have >1 common parent

Sequenced parents & tagging SNPs may identify all variants in each X

MAGIC

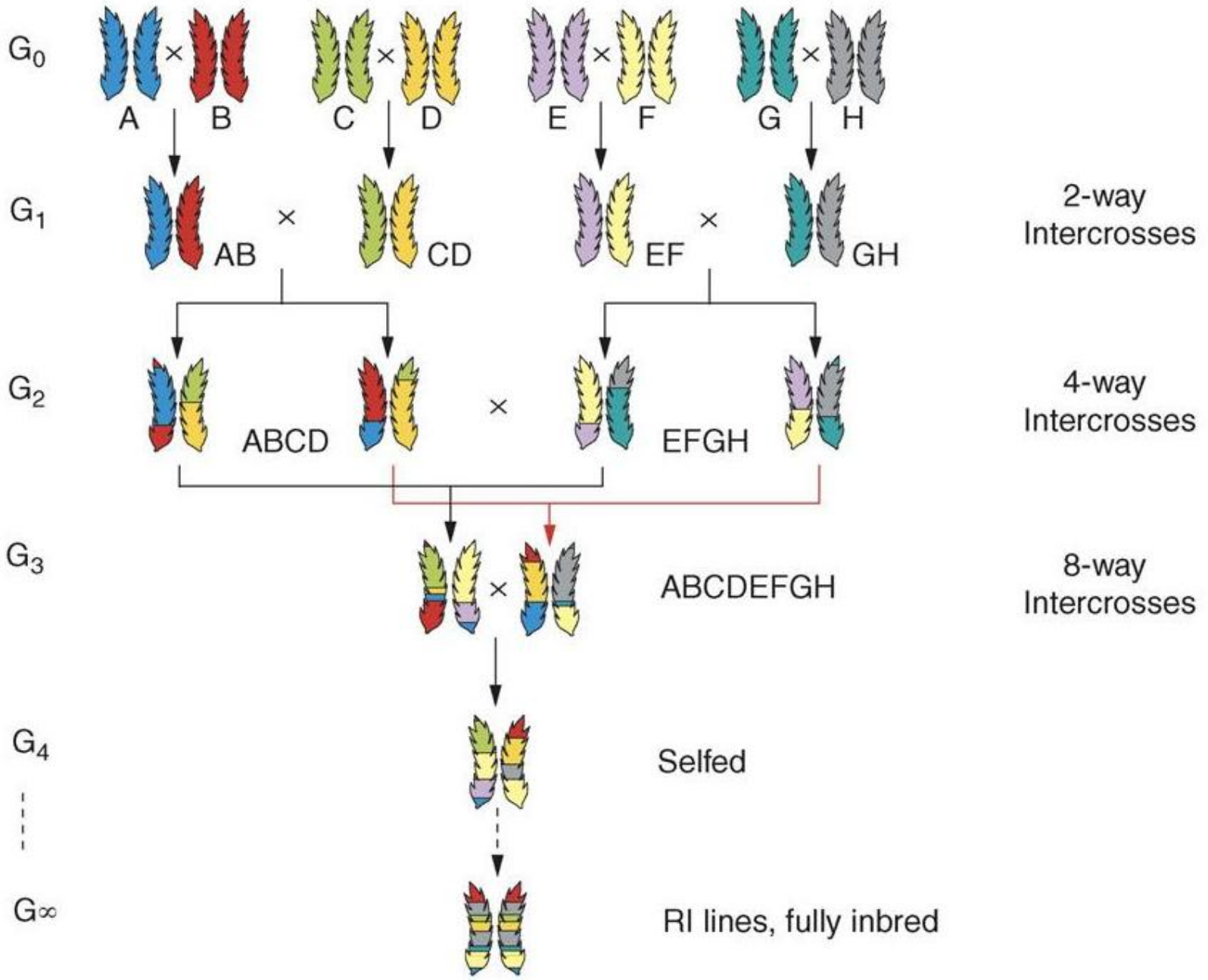
An extension of the Advanced intercross (Darvasi & Soller)

Diverse populations : good for multiple interacting traits and loci.

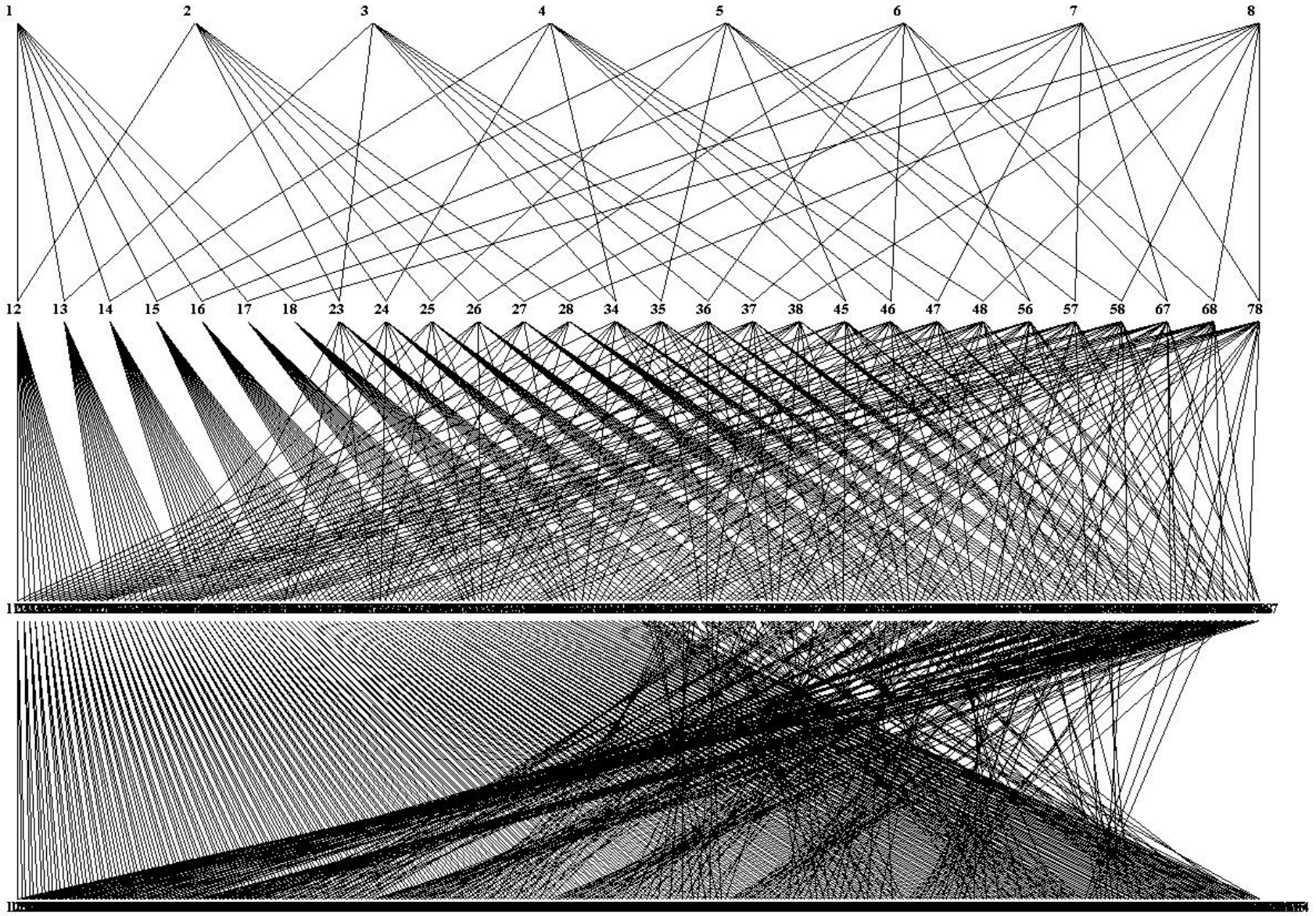
QTL detection in early generations

Fine mapping in later generations

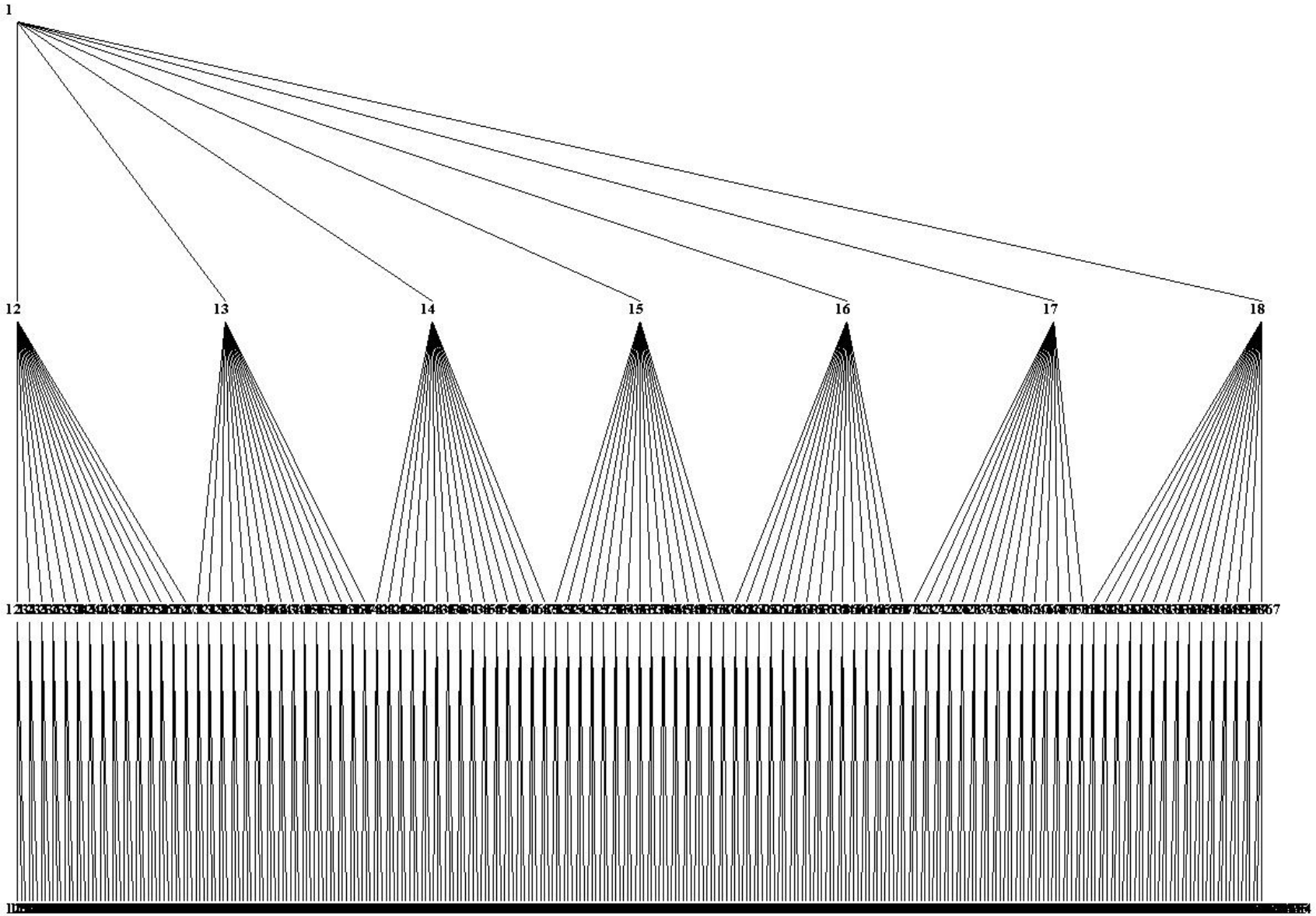
Multi-parent Advanced Generation Inter-Cross (MAGIC)



28→210→315 complete pedigree



28→210→315 descendants of founder 1



Success in mouse

97 traits

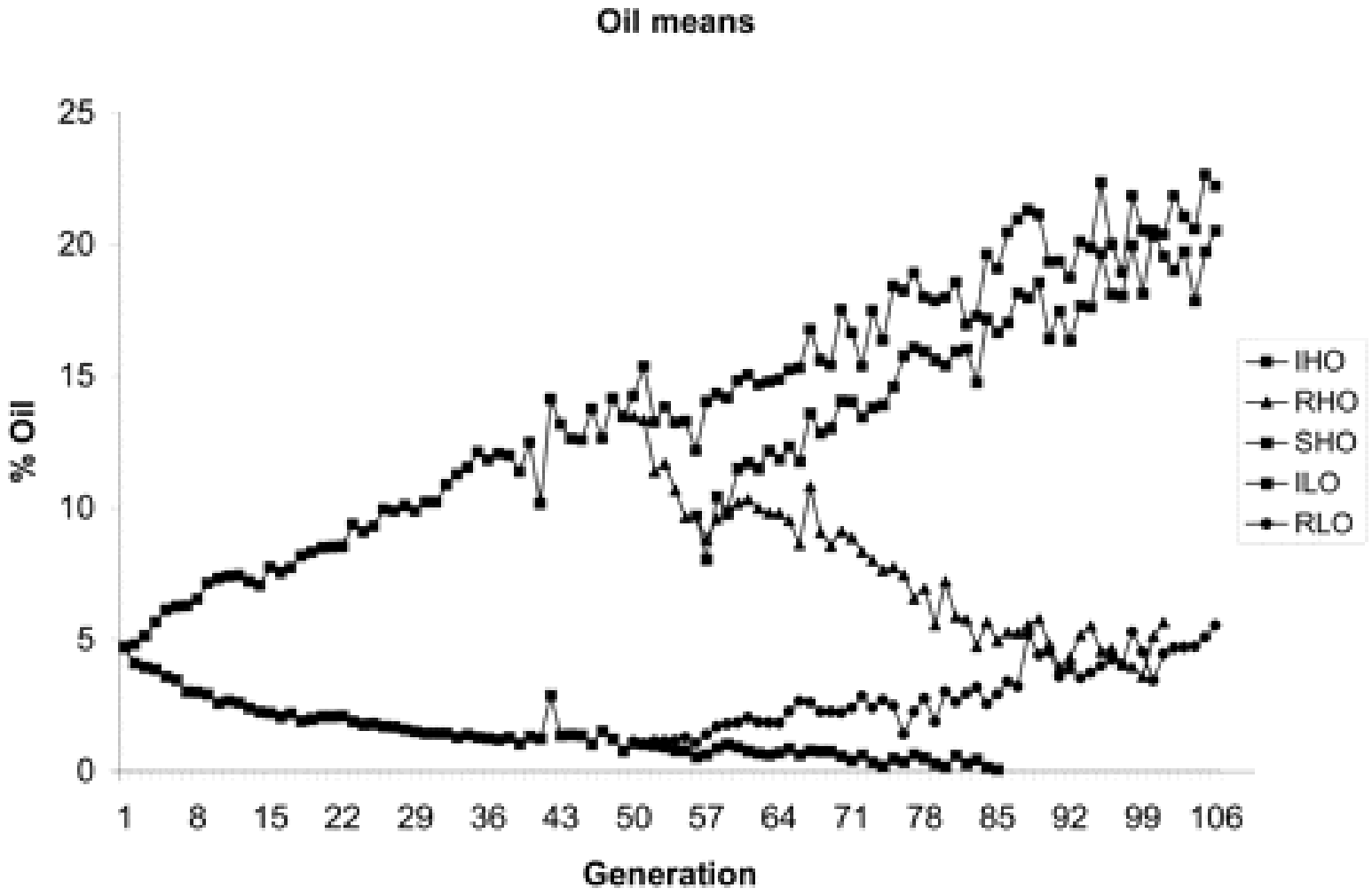
843 QTLs, average 95% confidence interval of 2.8 Mb.

The QTLs contribute to variation in 97 traits, including models of human disease (asthma, type 2 diabetes mellitus, obesity and anxiety) as well as immunological, biochemical and hematological phenotype

4.8 Mb region with QTL for anxiety

- Select conserved regions and compare sequence distribution pattern among founders with that of the QTN
- 14 SNPs identified as functional candidates out of 15,000

“Selection works.”



JW Dudley Crop Sci (2007) 47(S3)S20–S31

Illinois long-term selection experiment (est.1896)

Gen 70 – high and low selections crossed

Hybrid population intermated for 10 generations

50 QTL accounting for 50% of Vg identified by LD mapping

Resolution of 2-3 cM

Laurie et al Genetics 2004, **168**:2141-2155

Sounds like MAGIC

OPEN ACCESS Freely available online

PLoS GENETICS

A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*

Paula X. Kover[1,2]*, William Valdar[3], Joseph Trakalo[3], Nora Scarcelli[2], Ian M. Ehrenreich[4], Michael D. Purugganan[4], Caroline Durrant[3], Richard Mott[3]

1 Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom, **2** Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom, **3** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **4** Center for Genomics and Systems Biology, New York University, New York, New York, United States of America

Abstract

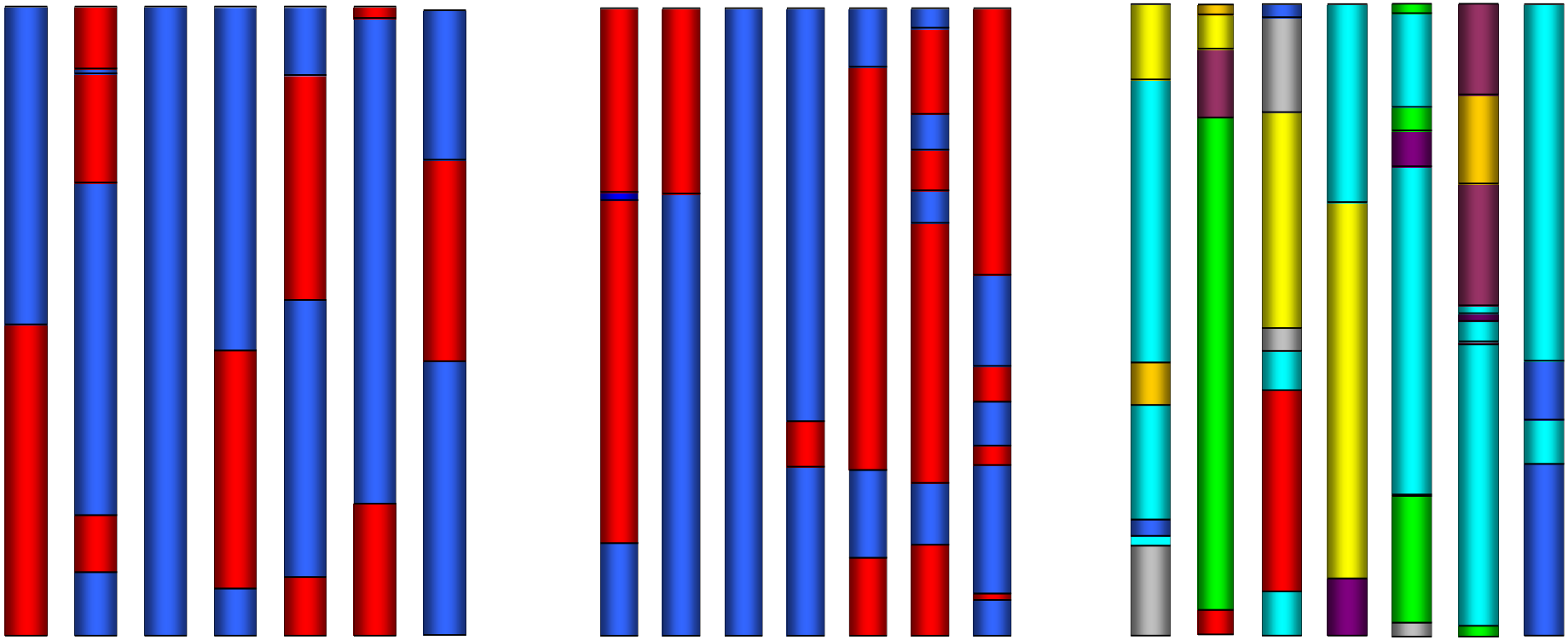
Identifying natural allelic variation that underlies quantitative trait variation remains a fundamental problem in genetics. Most studies have employed either simple synthetic populations with restricted allelic variation or performed association mapping on a sample of naturally occurring haplotypes. Both of these approaches have some limitations, therefore alternative resources for the genetic dissection of complex traits continue to be sought. Here we describe one such alternative, the Multiparent Advanced Generation Inter-Cross (MAGIC). This approach is expected to improve the precision with which QTL can be mapped, improving the outlook for QTL cloning. Here, we present the first panel of MAGIC lines developed: a set of 527 recombinant inbred lines (RILs) descended from a heterogeneous stock of 19 intermated accessions of the plant *Arabidopsis thaliana*. These lines and the 19 founders were genotyped with 1,260 single nucleotide polymorphisms and phenotyped for development-related traits. Analytical methods were developed to fine-map quantitative trait loci (QTL) in the MAGIC lines by reconstructing the genome of each line as a mosaic of the founders. We show by simulation that QTL explaining 10% of the phenotypic variance will be detected in most situations with an average mapping error of about 300 kb, and that if the number of lines were doubled the mapping error would be under 200 kb. We also show how the power to detect a QTL and the mapping accuracy vary, depending on QTL location. We demonstrate the utility of this new mapping population by mapping several known QTL with high precision and by finding novel QTL for germination data and bolting time. Our results provide strong support for similar ongoing efforts to produce MAGIC lines in other organisms.

Citation: Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, et al. (2009) A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. PLoS Genet 5(7): e1000551. doi:10.1371/journal.pgen.1000551

F2 derived

AIC

MAGIC



	F2 & self	AIC & self	8-way MAGIC & self
Prob (no recombination)	0.241	0.128	0.036
# tracts	2.6	3.3	4.7
# founders	2	2	3.5