The

Simulation of Genetic Data

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Overview

- Why simulate?
- Gene-dropping: “unconditional”
- Gene-dropping: “rejection sampling”
- Sequential imputation
- Monte-Carlo Markov Chains

Uses of simulation: modelling

If a particular statistical model is complicated, calculating the expected value of a variable in the model may be hard.

It is often easy to simulate the type of data that would be generated under that model, and then record the mean (or variance) of the simulated values.

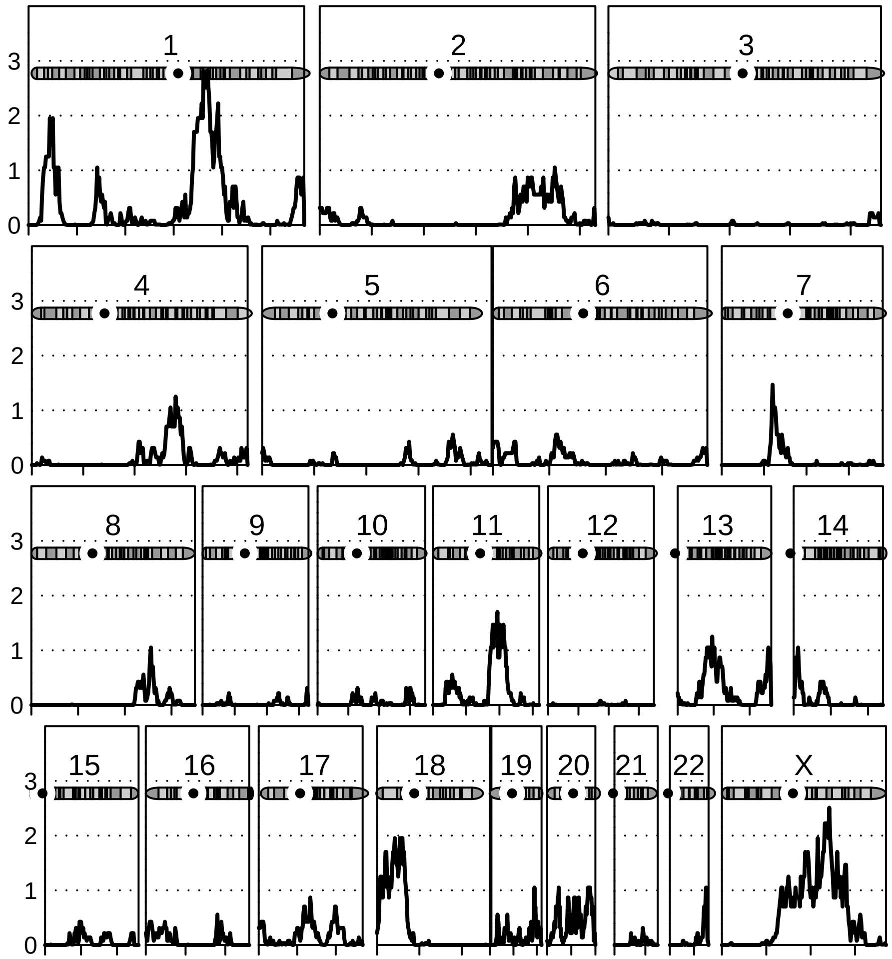
One common genetic application is for tests of assocation within complicated families.

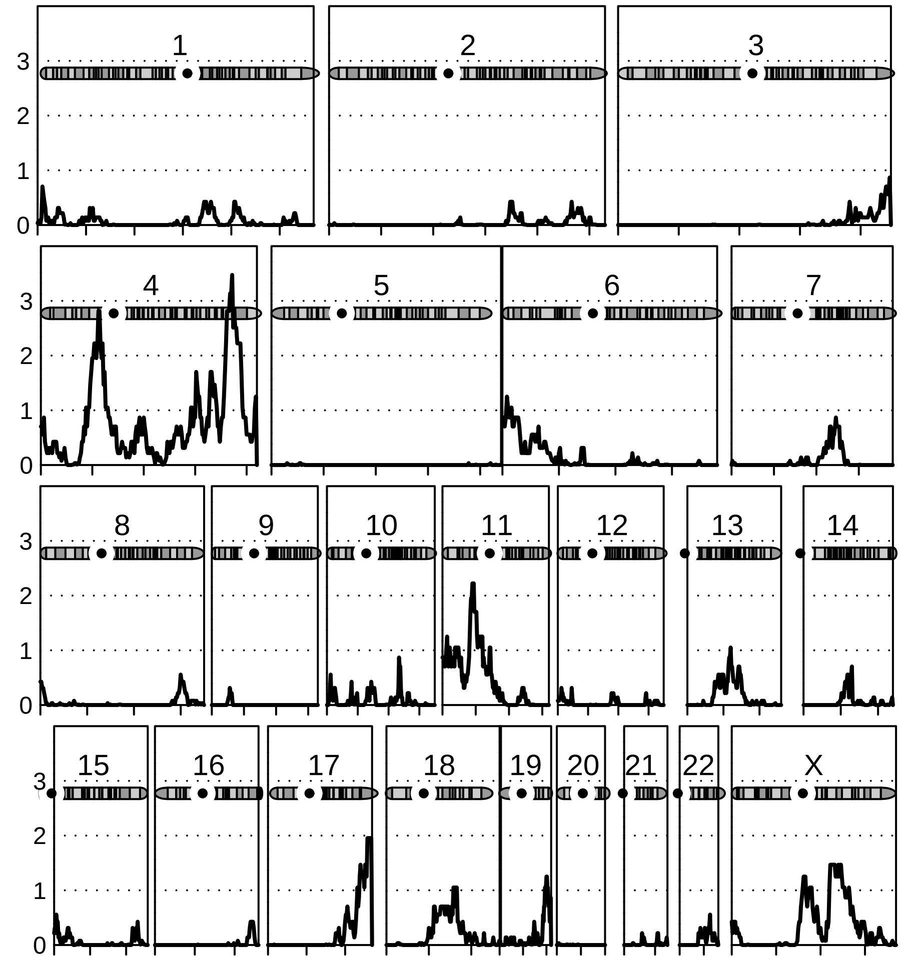
Uses of simulation: Monte-Carlo tests

One has a statistical test for a particular genetic hypothesis, based on complicated family data, and wishes to assign a P-value to it:

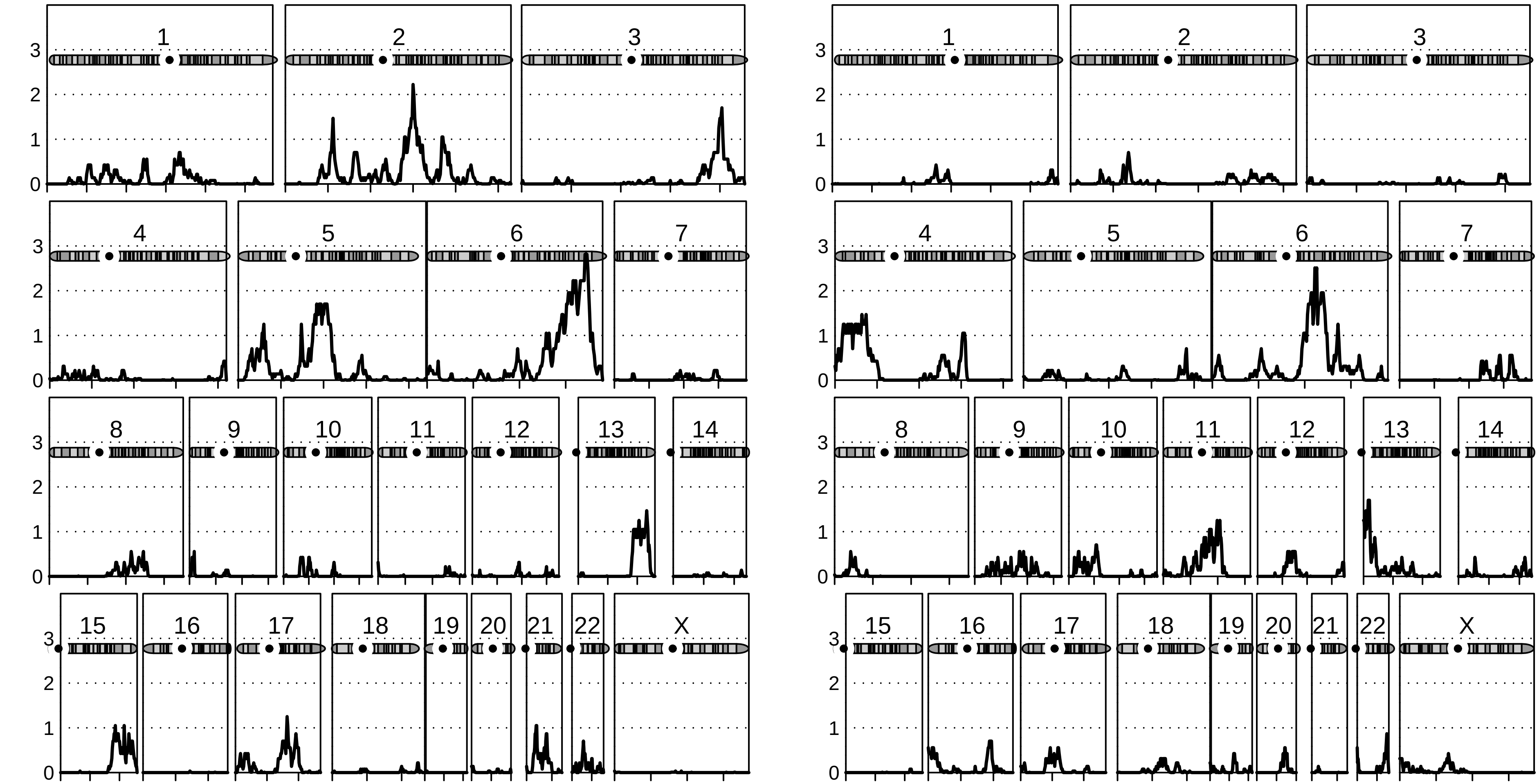
- Calculate your statistic for the observed family
- Simulate data for the same family under the null hypothesis (many times)
- Compare the observed statistic to the distribution of the statistic in the simulations

A common application is to generate genome-wide P-values: the test statistic is the “most significant” result from a genome scan. Many journals will request this.

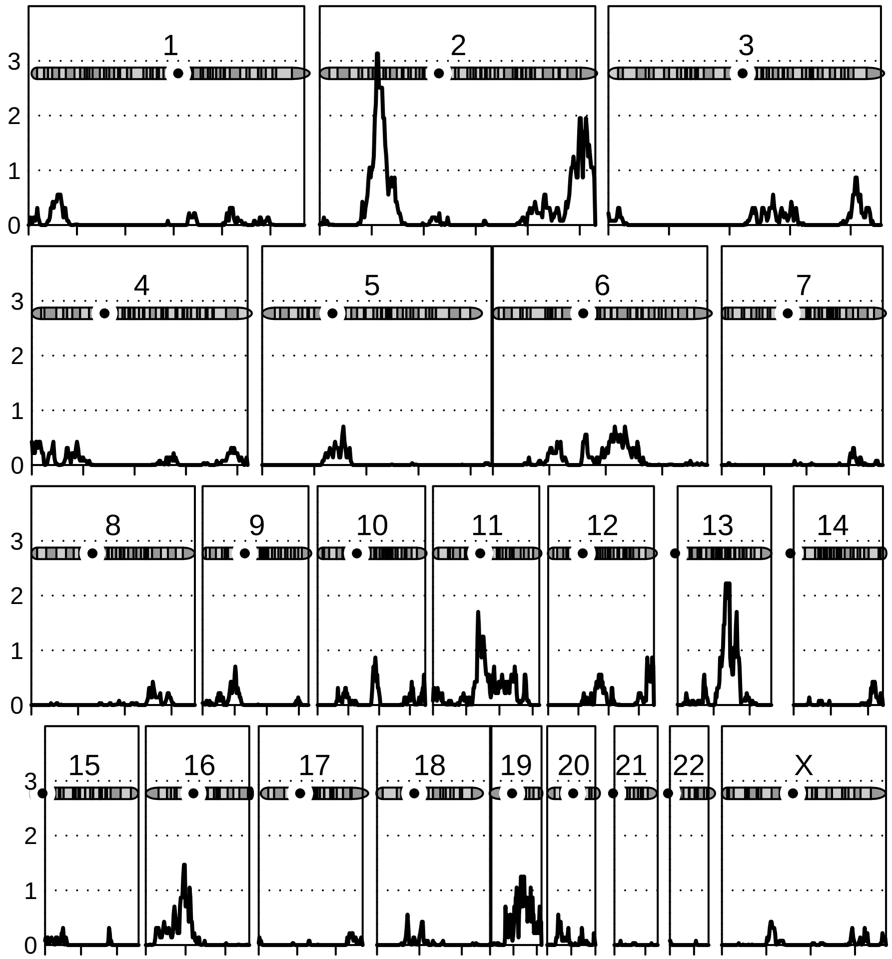

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3

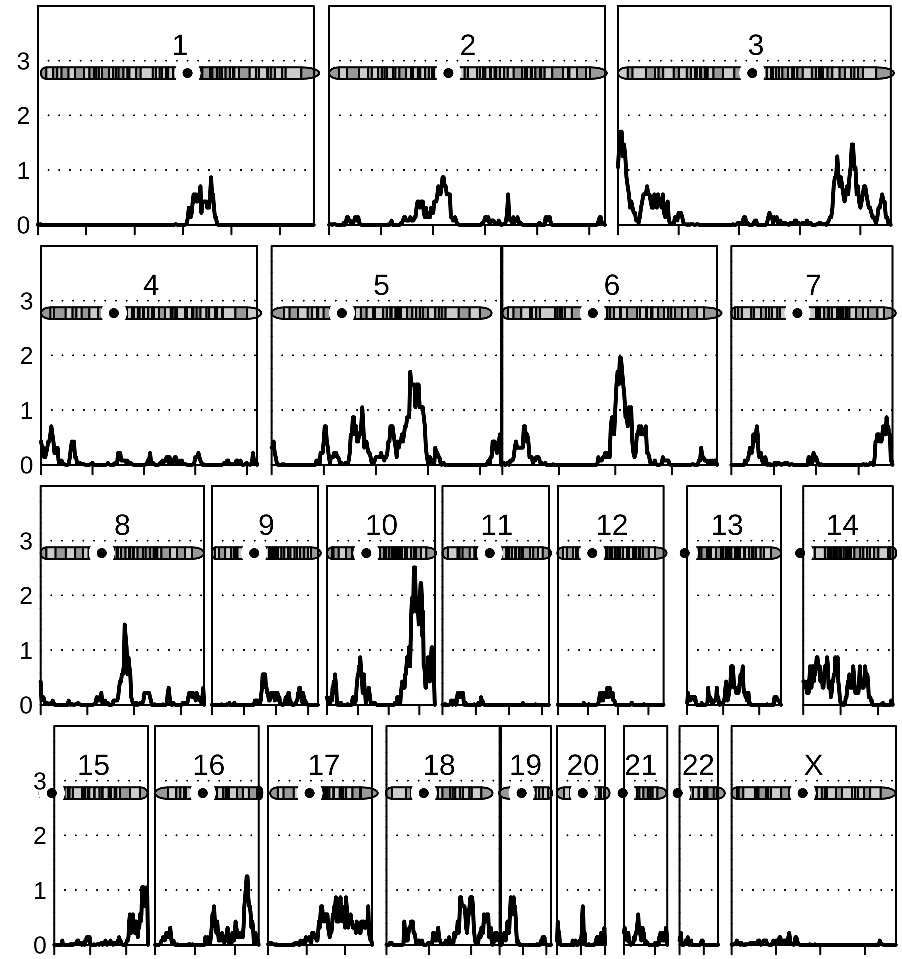

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3

QIMR

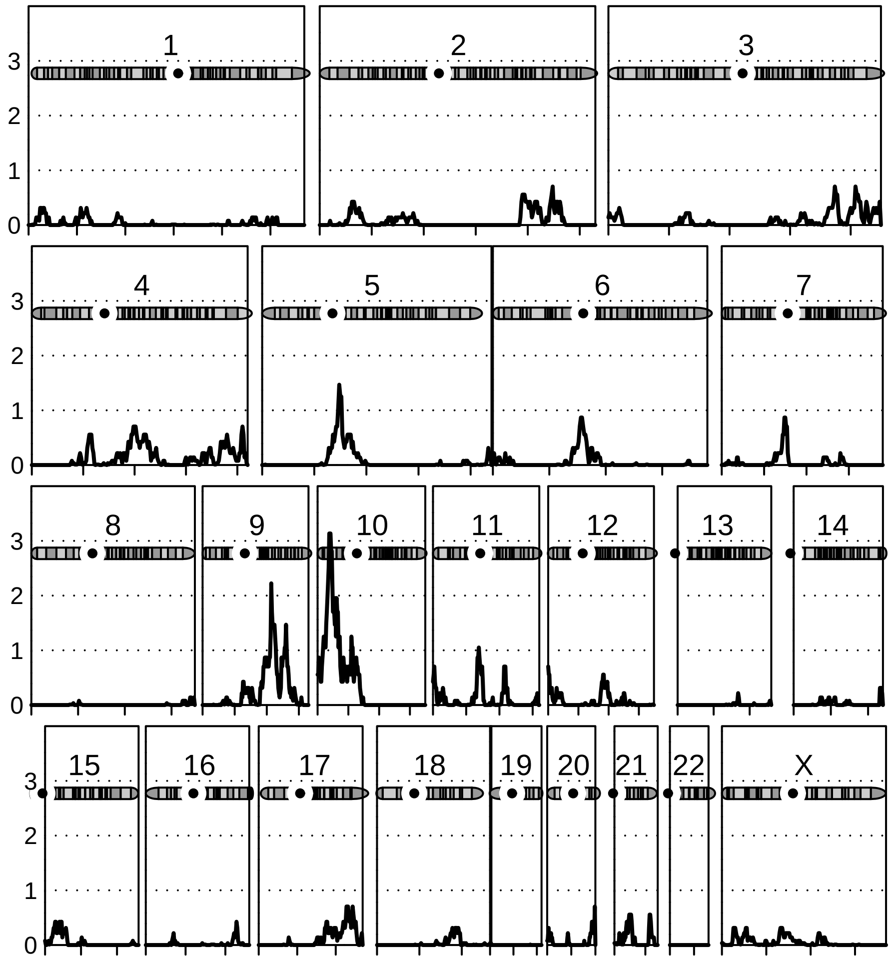

1 2 3
4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 X
0 1 2 3
1 2 3
4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 X
0 1 2 3

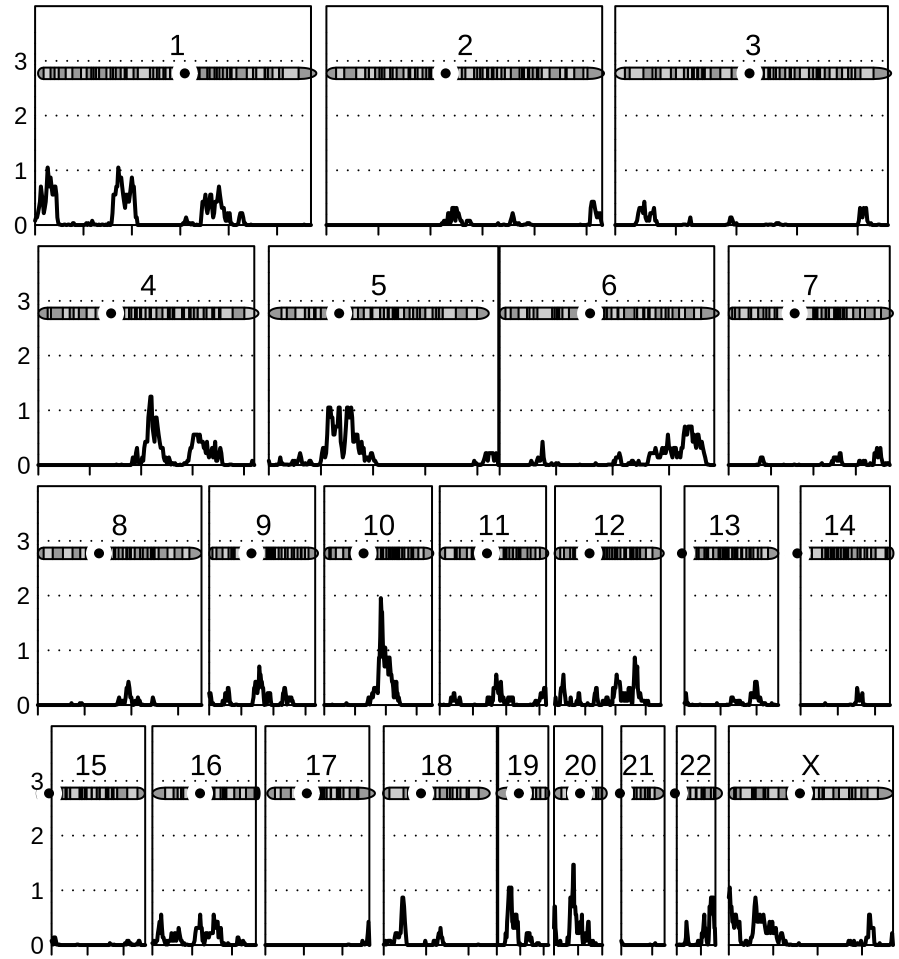
QIMR


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3

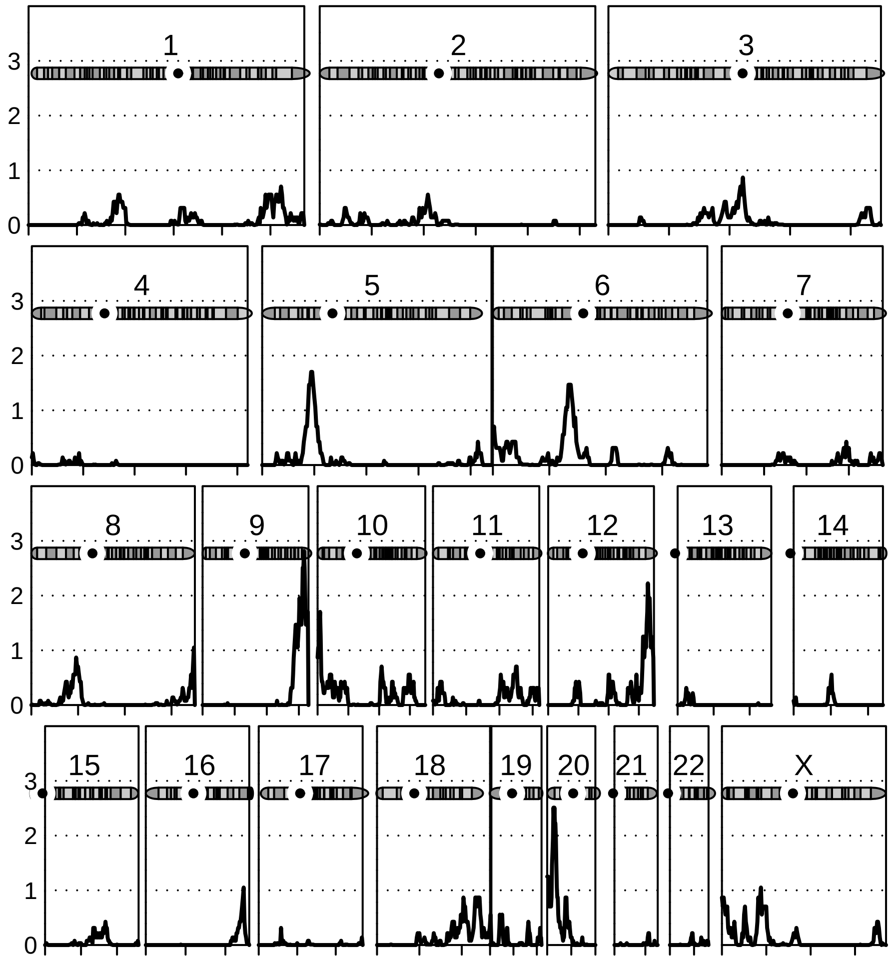

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3

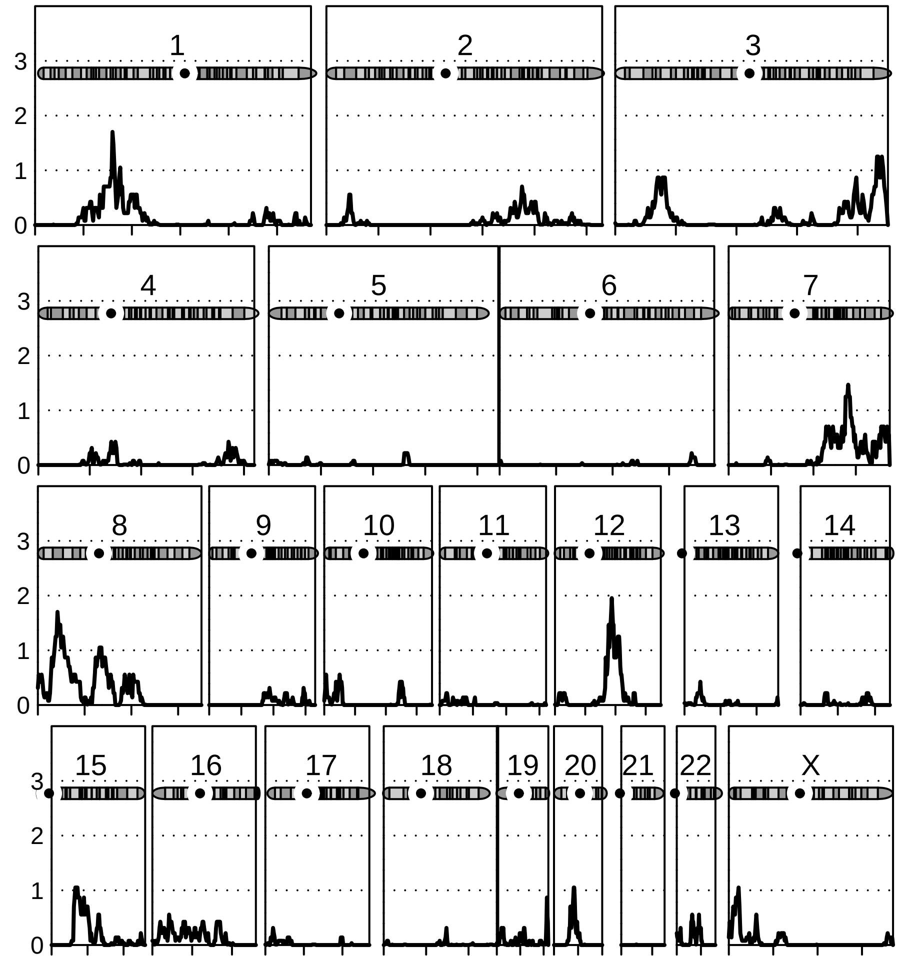
QIMR


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
0
1
2
3


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
0
1
2
3

QIMR


0
1
2
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X


0
1
2
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X

QIMR

Uses of simulation: Power calculations

To evaluate the power of a complicated statistical test

- Simulate data for the same family under the alternative hypothesis (many times)
- Count how often the statistic is significant

Uses of simulation: Checking the robustness of a test

We may wonder if a particular test is robust in the face of *violation of its assumptions*. For example, our twin models all assume the trait or liability is multivariate normally distributed. We can simulate data where this is not correct, and see if the Monte-Carlo P-values agree with the asymptotic P-values.

Gene-dropping: simulating the founders

Gene-dropping is the method used to simulate a codominant marker in a family.

Pedigree founder genotypes are first generated by multinomial sampling from the measured population genotype frequencies.

Assuming Hardy-Weinberg Equilibrium, genotype frequencies can be calculated from allele frequencies:

So we draw two alleles for each person, using the allele frequencies as the probability of choosing each type of allele.

Gene-dropping: simulating the nonfounders

We simulate childrens' genotypes by randomly drawing one allele from each parental genotype (they are equally likely).

And simulate childrens' childrens' genotypes by same process…

Until the pedigree genotypes are completely filled in.

A monozygotic twin always receives the same genotype as his twin.

Gene-Dropping: an application

For example, testing association between a binary trait and a codominant marker, correctly allowing for the pedigree structure of the data:

Test Statistic:	Ordinary contingency table chi-square test, X^2_{Obs}
Problem:	Usual reference distribution assumes independence of observations
Solution:	Generate correct reference distribution by simulation

Gene-Dropping: algorithm

Estimate marker allele frequencies for complete sample, regardless of trait phenotype

Repeat B times:

a. Simulate founder (parental) genotypes as independent draws from ideal population with observed marker allele frequencies

b. Simulate childrens' genotypes by randomly drawing one allele from each parental genotype

Simulate childrens' childrens' genotypes by same process…

c. If a genotype is missing in the original pedigree, remove it from the simulated pedigree

d. Calculate chi-square test using simulated data X^2_i

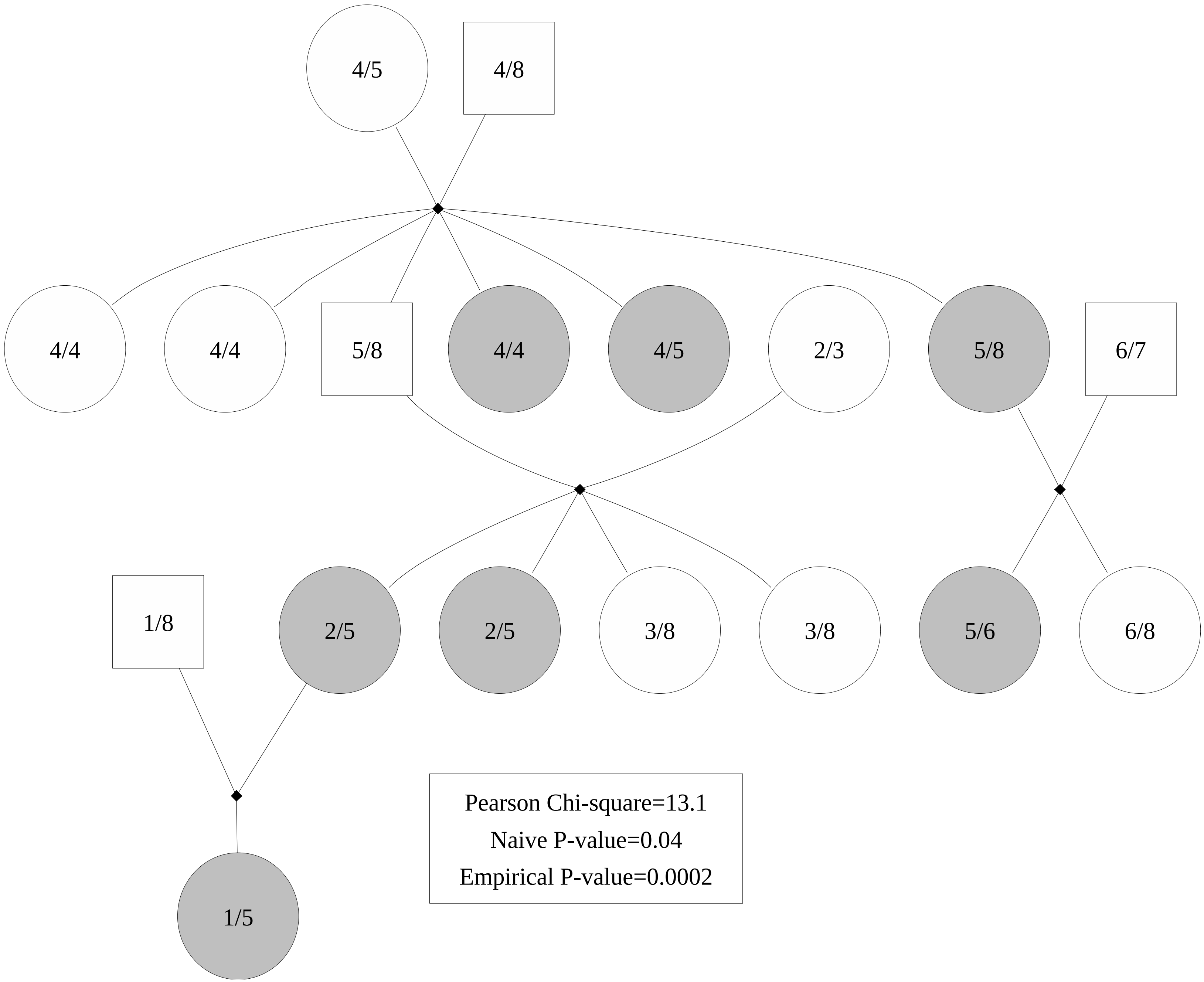
N = How many times $X^2_i \geq X^2_{Obs}$

Empirical P-value = N/B

QIMR

Gene-Dropping: refinements

- The trait values for each pedigree member **do not change** from replicate to replicate, so the effects of *unmeasured* genes are included in the simulation.
- To produce within-family tests, the simulation can skip step Ia. above, so the reference distribution is "Conditional on Parental Genotypes"
- B does not have to be fixed, so that the simulation stops when the P-value is sufficiently accurate (sequential approach).
- The association test as described here capitalizes on linkage, and in a single pedigree is almost purely a test of cosegregation.



Pearson Chi-square=13.1
Naive P-value=0.04
Empirical P-value=0.0002

QIMR

Breast cancer and BRCA1

Hall et al (1990) reported that breast cancer in densely affectd pedigrees was linked to a marker (D17S74) on chromosome 17. In the first pedigree they described, the P-value for linkage using a nonparametric linkage (NPL) test is P=0.023.

If we tabulate allele counts at the marker, we see that the “5” allele is only seen in cases.

D17S74 Allele	1	2	3	4	**5**	6	7	8
Breast Cancer	1	2	0	1	**6**	1	0	1
Unaffected Female	0	1	3	4	**0**	1	0	3

The Pearson $X^2 = 13.1$, df=6, P=0.041

By contrast, the gene-dropping Monte-Carlo P-value is P=0.0002 (this family is segregating the c.2800 AA deletion).

Gene-Dropping a trait

We can use gene dropping to simulate genotypes at a codominant locus. How do we simulate a quantitative trait under control of that locus?
This is done by specifying the genetic model:

- For a quantitative trait, the genotypic means and (environmental) variances
- For a binary trait, the penetrances

We then simulate the trait values for each person in the pedigree, drawing from the appropriate random number generator eg normal or binomial.

Gene-Dropping a polygenic trait

How do we simulate a quantitative trait under control of multiple quantitative trait loci?

- Simulate multiple loci, and specify an overall model (pseudopolygenic)
- Simulate breeding values

Under the polygenic model, each individual has a normally distributed "genotype", their breeding value. We can gene-drop then:

- Simulate founder breeding values as random normal deviates
- Simulate children as the average of the parental breeding values plus the effects of the segregation variance (a random normal deviate drawn from $1 - \frac{1}{2}(F_{FA} + F_{MO})$

We then simulate the trait values for each person in the pedigree, drawing from E.

Gene-Dropping: Programs

A large number of different computer programs provide gene dropping

- GASP
- JPAP
- MENDEL
- MERLIN
- MORGAN
- SIB-PAIR
- SIMULATE

Gene dropping and rejection sampling

A further refinement of gene-dropping is to set further conditions on the simulation. For example, we might want to simulate genotypes at one locus conditional on those observed at a linked locus.

One approach to doing this is **Rejection Sampling** (trial and error).

Repeat until have accumulated *B* samples:

Usual gene drop

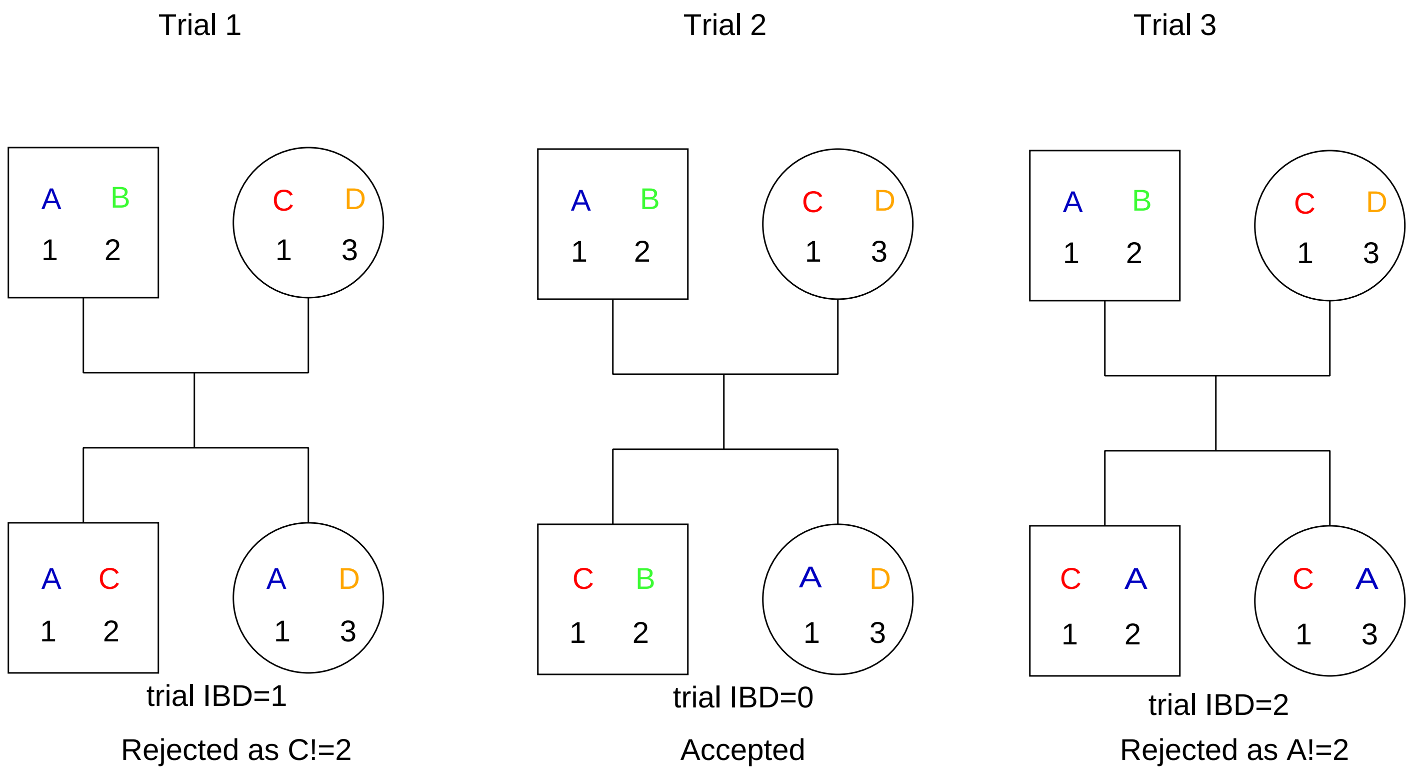
Test if simulated sample meets specified condition

Keep if acceptable

Summary of accepted samples

This works well if the conditions aren't too restrictive.

IBD estimation by rejection sampling



Only 1 in 16 trials will be successful on average, and all the accepted samples will have IBD=0.

QIMR

FBAT by rejection sampling: preliminaries

The FBAT test is a TDT that correctly allows for missing parents. Unaffected offspring are used to impute the missing parental genotype, but only certain informative constellations of parent and child genotypes allow unequivocal imputation. Therefore, the naive TDT statistic applied to such data is biased:

Take a diallelic marker (alleles 1 and 2 with frequencies p and q). The penetrances for the genotypes are $\{f0, f1, f2\}$. Overall, for a backcross $1/2 \times 2/2$, the expected proportion of 1 alleles transmitted to a child is 0.5. If only families where the child is affected are ascertained, the expected proportion of 1 alleles transmitted to a child is f1/f2.

If only one parent and the proband available, then to be useful, the genotyped parent is heterozygous and child is homozygous. This gives the expected proportion of 1 alleles transmitted to the child as p, not 0.5.

Typed Parent	Child	Untyped Parent	Proportion	
$1/2$	$1/1$	$(1/1)$	$\frac{p^2}{2}$	
$1/2$	$1/1$	$(1/2)$	$\frac{pq}{2}$	$\frac{p(p+q)}{2}$
$1/2$	$2/2$	$(1/2)$	$\frac{pq}{2}$	
$1/2$	$2/2$	$(2/2)$	$\frac{q^2}{2}$	$\frac{q(p+q)}{2}$

FBAT by rejection sampling: a sketch of the algorithm

To estimate the mean and variance of the proportion of alleles transmitted to affected children, Sib-pair uses gene dropping with rejection sampling.

Only families where the missing parental genotypes can be inferred unequivocally are used for the FBAT. The child genotypes used to perform that inference are thus “used up”, and cannot be used to test for transmission distortion.

A gene-drop simulation is rejected if the simulated transmission pattern is not consistent with the child genotypes originally used to infer the missing parental genotype.

Permutation based tests

An alternative approach that you may be familiar with is permutation testing.

Instead of simulating "new" data to obtain a distribution under the null hypothesis, we repeatedly scramble up the existing data to give a null distribution.

For example, in the case of association between trait values and genotype at a codominant marker, we can swap the trait values between different individuals in the dataset, and recalculate the test statistic. We leave the genotypes unchanged.

For family data, the swapping must correctly allow for the structure of the family. This is efficient for simpler pedigrees, such as nuclear families.

In this case, we would swap trait values between sibs or twins within families, or swap values of entire sibships or twin pairs between families.

Sequential Imputation

- An efficient alternative algorithm for conditional simulations
- Especially useful in larger problems
- Uses Maximum Likelihood to calculate probabilities for genotypes to be simulated
- SLINK and SIMLINK are the two commonest programs
- Mainly used for power calculations
- Can “fill in” genotypes for family members who are yet to be genotyped

Monte-Carlo Markov Chain methods

- What are MCMC methods
- MCMC for exact contingency table analysis
- MCMC simulation of unobserved marker genotypes in Sib-pair
- MCMC GLMMs in Sib-pair

A Segue into Monte-Carlo Markov Chains

Rejection sampling becomes inefficient if there are too many side conditions. For a pedigree with many missing genotypes, it could be worse than 1 accepted sample per 10^7.

The correct proportions of the different possible unobserved genotypes "fall out" automatically from a rejection sampler:

Possibility	X_1	X_2	X_3	…	X_i	…	X_n	Total
Count	N_1	N_2	N_3	…	N_i	…	N_n	B
Proportion	p_1	p_2	p_3	…	p_i	…	p_n	1

Sampler	**Proposal mechanism**	**Filter**	**Samples**
Rejection	Gene-drop *every* possibility	Rejection	Independent
MCMC	Proposal based on last *sample*	Surrogate LR	Correlated

MCMC for exact contingency table analysis

$$X_1 \leftrightarrow X_2 \leftrightarrow X_3$$

	X_1		X_2		X_3	
	2	3	2	3	2	3
3	**0**	3	**1**	2	**2**	1
2	2	0	1	1	0	2

A Monte Carlo Markov Chain can be constructed that moves one step at a time between all the legal tables, and the proportion of samples of each table represents the probability of that table.

Proposal: choose two rows (*xorigin* and *xdestination*) and two columns (*yorigin* and *ydestination*). Change the counts (a,b,c,d) by {+1,-1,-1,+1}.

Filter: Calculate the ratio of probabilities of present table to proposed table ($\frac{ad}{(b+1)(c+1)}$).

This ratio is the Metropolis criterion (q).
If $q \geq 1$, always accept the new sample proposal; otherwise accept the new proposal q proportion of times, or keep the old table as the new sample.

QIMR

MCMC simulation of unobserved marker genotypes

We can try and perform a similar trick to visit every legal table of the missing genotypes for a pedigree. The Metropolis criterion is easy to calculate, being the ratio of the likelihoods for the *changed* genotypes:

- founders: the frequency of the genotype in the population
- nonfounders: the probability that genotype would be transmitted from her parents.

The difficult part is defining a proposal mechanism that will eventually visit **every** legal possible genotype, without having a large number of rejections (non-Mendelian proposals).

For diallelic markers, Lange and Matthysse (1989) described a correct method that is fairly efficient.

For multiallelic markers, there is no simple proposal method, but a variety of work-arounds are in use (in programs such as SIMWALK2, LOKI, and MORGAN).

Summarizing MCMC simulations

As opposed to the case of rejection sampling, the simulated samples from a MCMC are not independent. Therefore the effective number of samples is a lot smaller than the observed number.

Batching is one method to estimate correct standard errors of summary statistics from correlated samples. Sib-pair summarizes parameters as means of the simulated values, and the standard errors as the standard deviation of $\sqrt{B}$ subsample means (Jones et al 2005). The interbatch lag-1 serial correlation is calculated as a diagnostic for the appropriate number of values to simulate (Ripley 1987).

An alternative approach is **thinning**, where one retains only every N^{th} sample. The contingency table P-value estimator in Sib-pair sets N to the number of observations in the table.

MCMC programs

- SIMWALK2
- LOKI
- MORGAN

When would we use these programs?

- Multipoint parametric linkage analysis in large pedigrees
- Error checking and haplotyping in large pedigrees
- Estimation of IBD for larger pedigrees: both MENDEL and MERLIN can use SIMWALK2 or LOKI IBD matrices