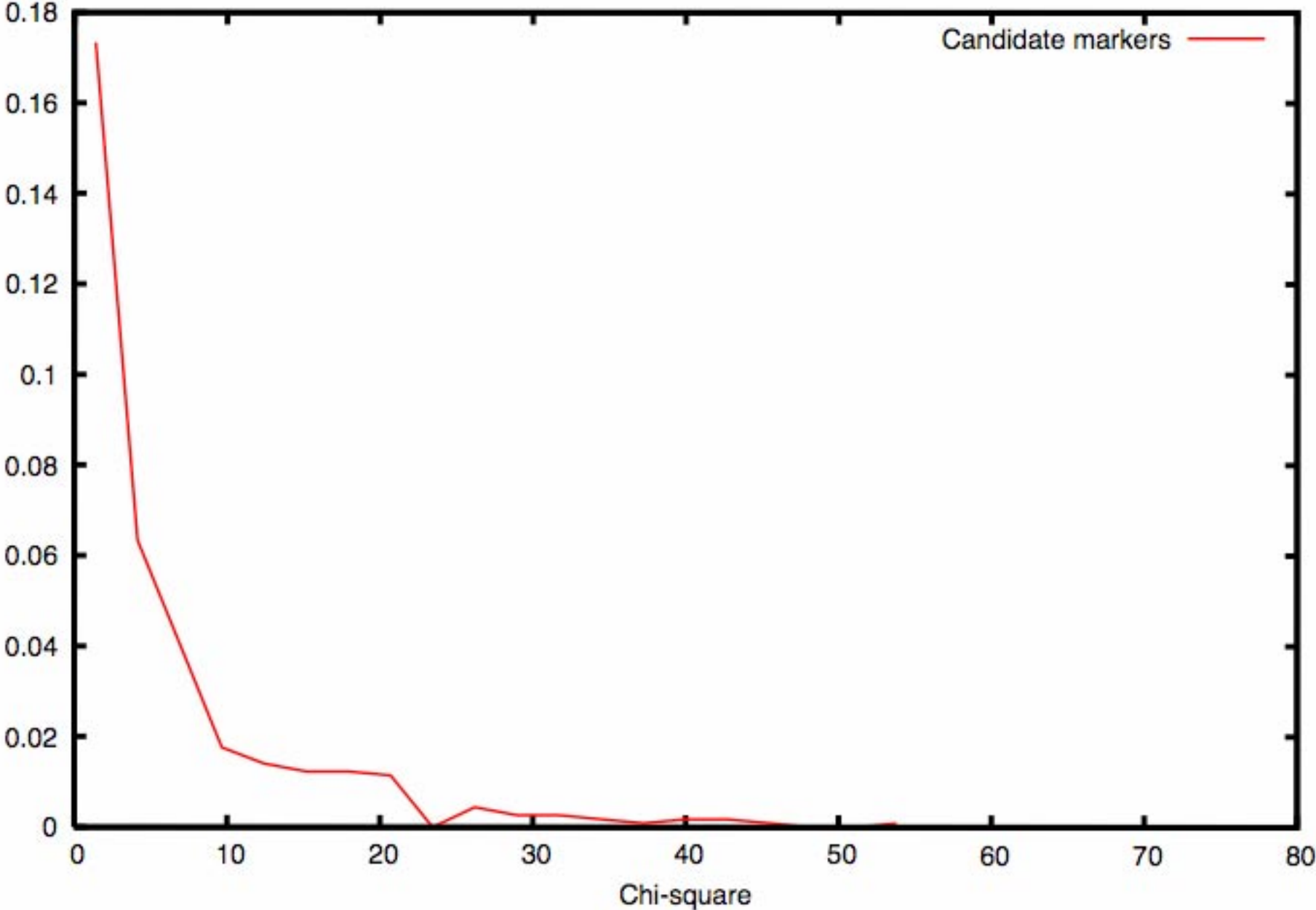
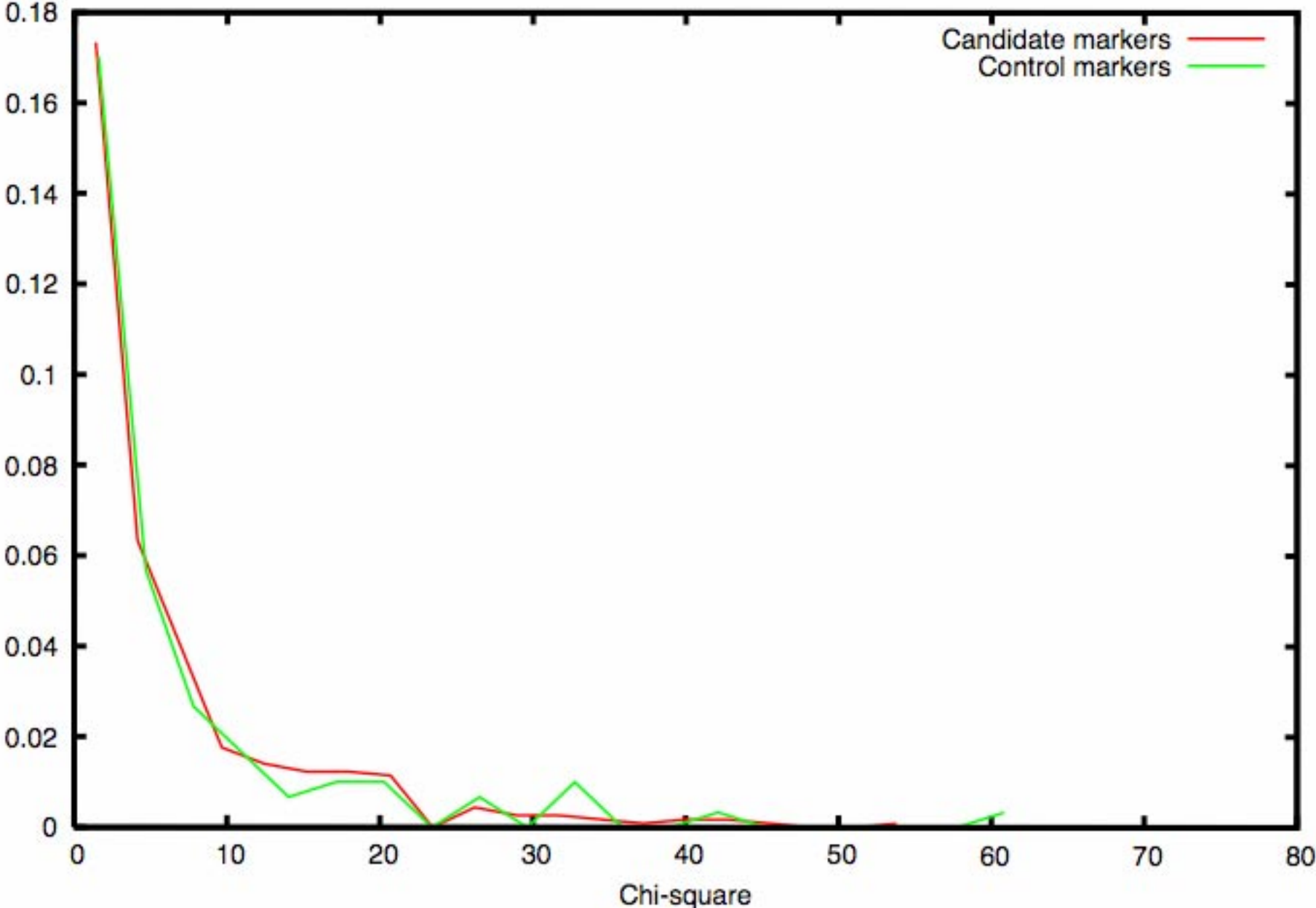


Distribution of Chi-square values from population comparison



Distribution of Chi-square values from population comparison



# MCMC methods in genetics

Simon C Heath, CNG

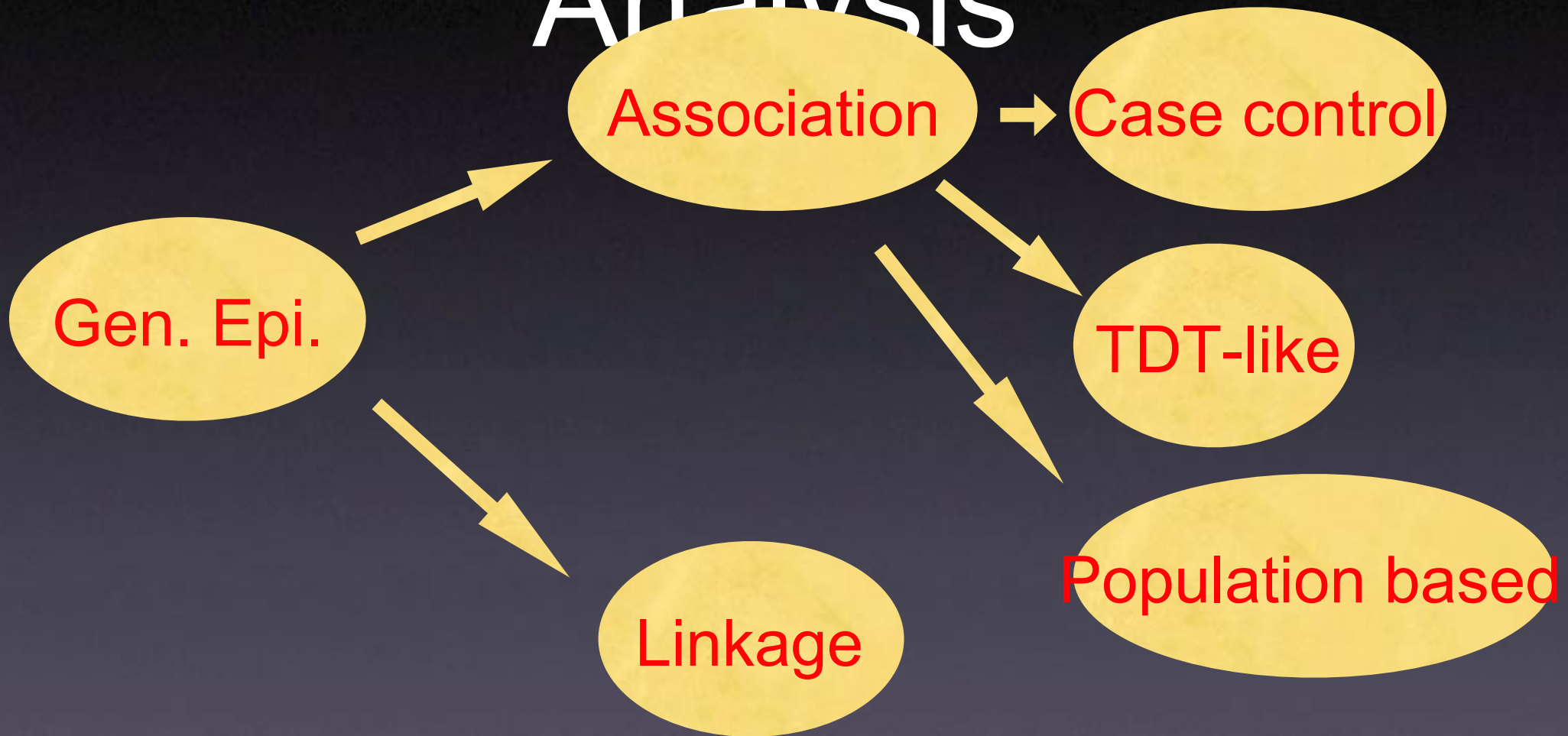
# MCMC

- Sampling based technique which can get approximate results for complex analyses
- Use to extend conventional analyses
  - Larger or more complex datasets
  - More complex models
  - 'Plug and play' analyses

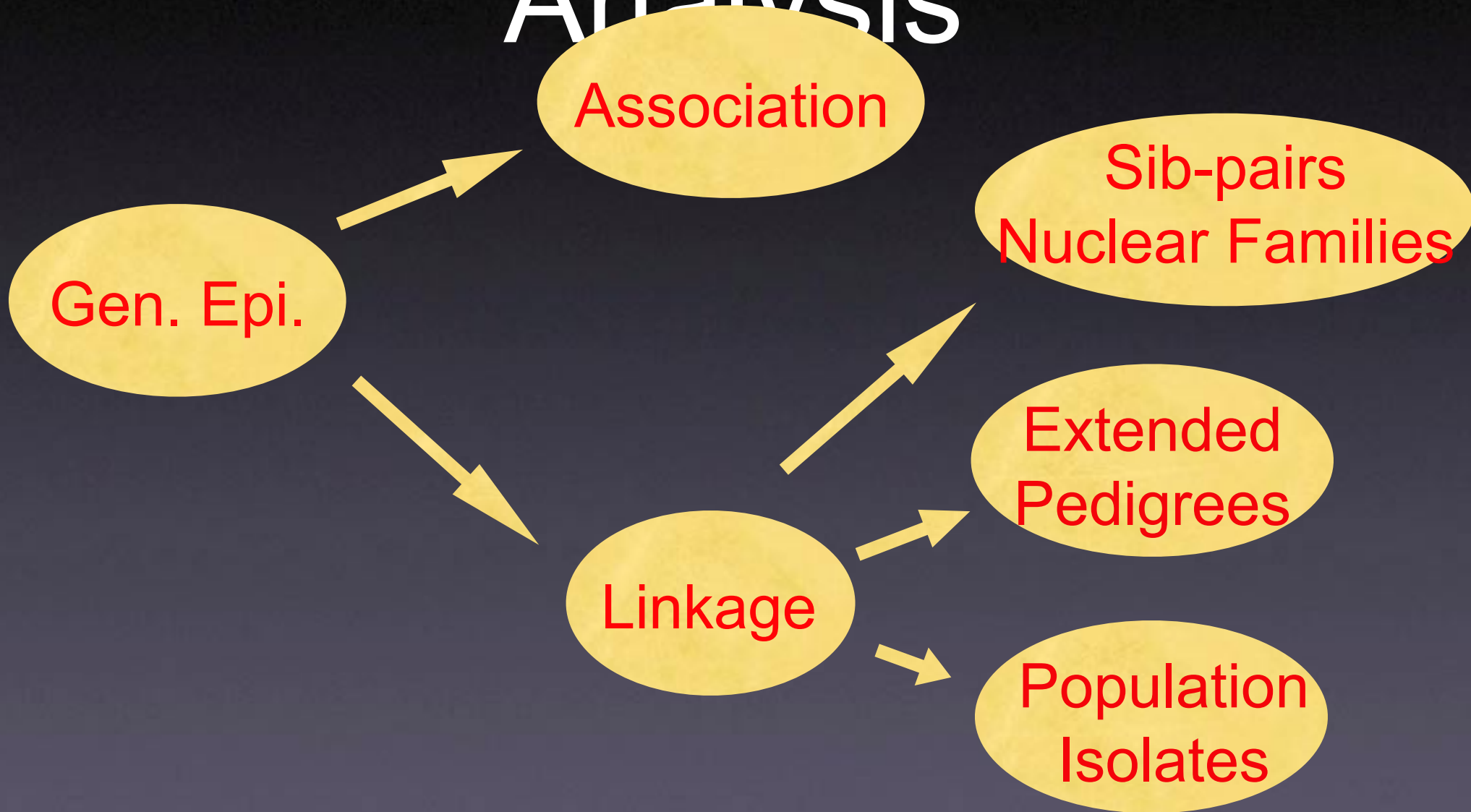
# Genetic Data

- Many forms:
  - Marker data (microsatellite, SNP, ...)
  - Sequence data
  - Gene expression data
- Potentially large amounts

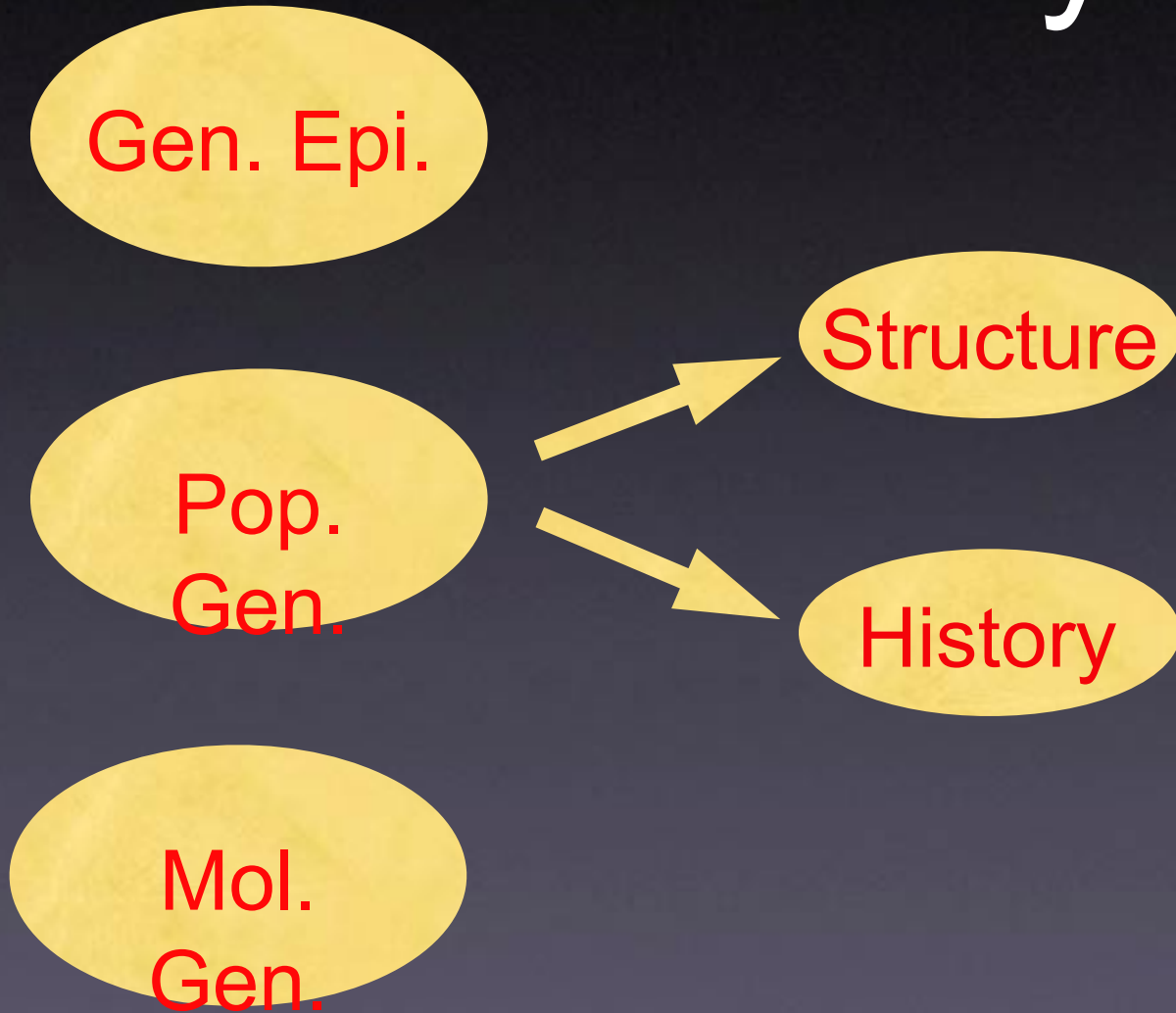
# Genetic Data Analysis



# Genetic Data Analysis



# Genetic Data Analysis



# Genetic Data Analysis

Gen. Epi.

Pop.  
Gen.

Mol.  
Gen.

Structure

Expression



# Complex analyses

- Large amounts of data
- Large amounts of *unobserved* data
- Complex relationships between data items
- Complex models

# Missing Data

- Unobserved/unobservable data
- If all missing data *were* observed, analyses become (relatively) simple
- Dealt with by summing over/finding MLEs for/fixing missing variables

# Linkage Analysis

## (Traditional)

- Fix penetrance parameters, allele frequencies
- Find MLE for parameter  $\theta$
- Sum over ordered genotypes/inheritance patterns

# Linkage Analysis

## (Bayesian)

- Assign priors to penetrance parameters, allele frequencies and  $\theta$
- Sum over ordered genotypes/inheritance patterns, penetrance parameters, frequencies and  $\theta$
- Computationally much harder

# Summation

- Sum over all patterns of missing data which are compatible with observed data
- Algorithms exist which exploit local dependencies amongst variables to reduce computational requirements
- In many cases, summation is not feasible due to (a) complexity of dependencies or (b) mixtures of variable types

# Linkage Analysis

## (Bayesian)

- Assign priors to penetrance parameters, allele frequencies and  $\theta$
- Sum over ordered genotypes/inheritance patterns, penetrance parameters, frequencies and  $\theta$
- Computationally much harder

# Dependency structure of 5 linked loci

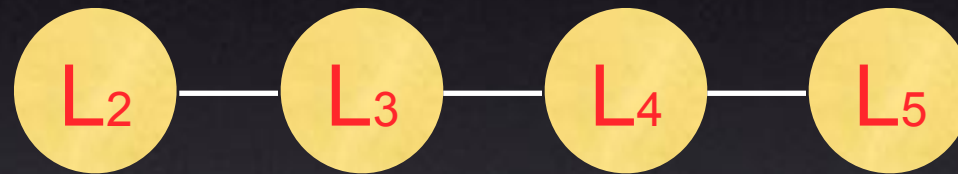


# Dependency structure of 5 linked loci



Generate  $p(S_1, S_2)$   
Integrate over  $S_1$  to get  $p(S_2 | S_1)$

# Dependency structure of 5 linked loci

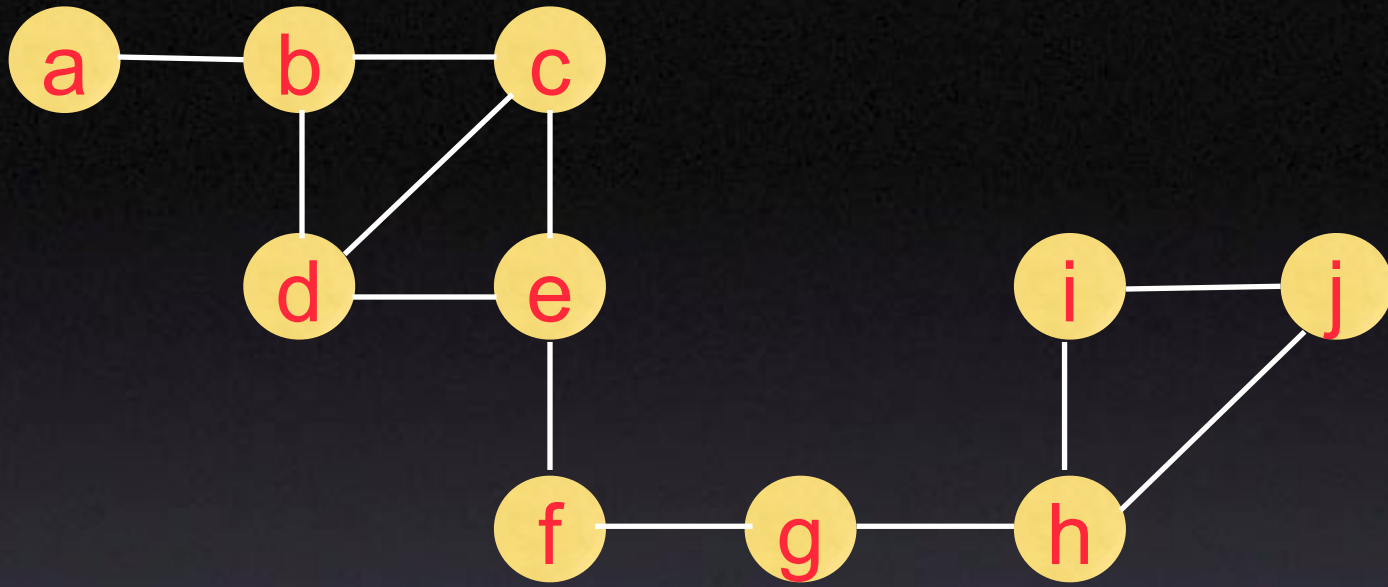


Generate  $p(S_1, S_2)$

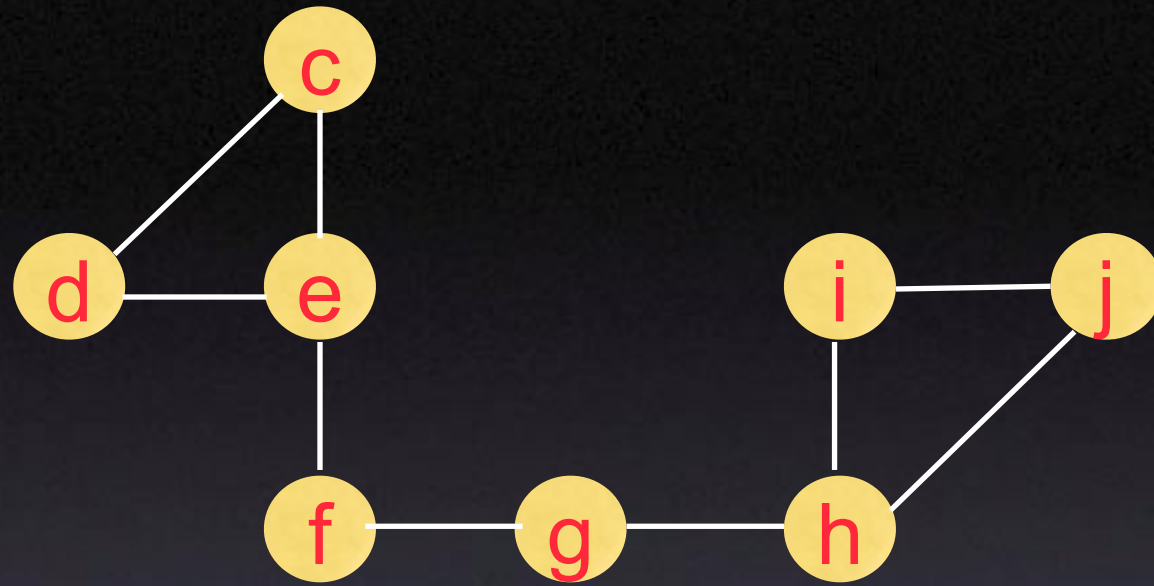
Integrate over  $S_1$  to get  $p(S_2|S_1)$

Generate  $p(S_2, S_3|S_1)$

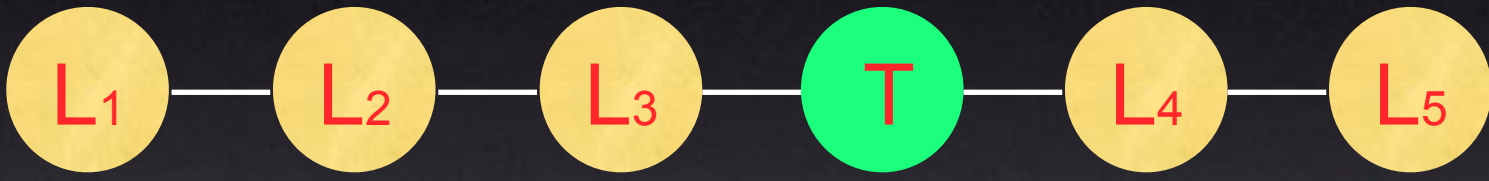
Integrate over  $S_2$  to get  $p(S_3|S_1, S_2)$

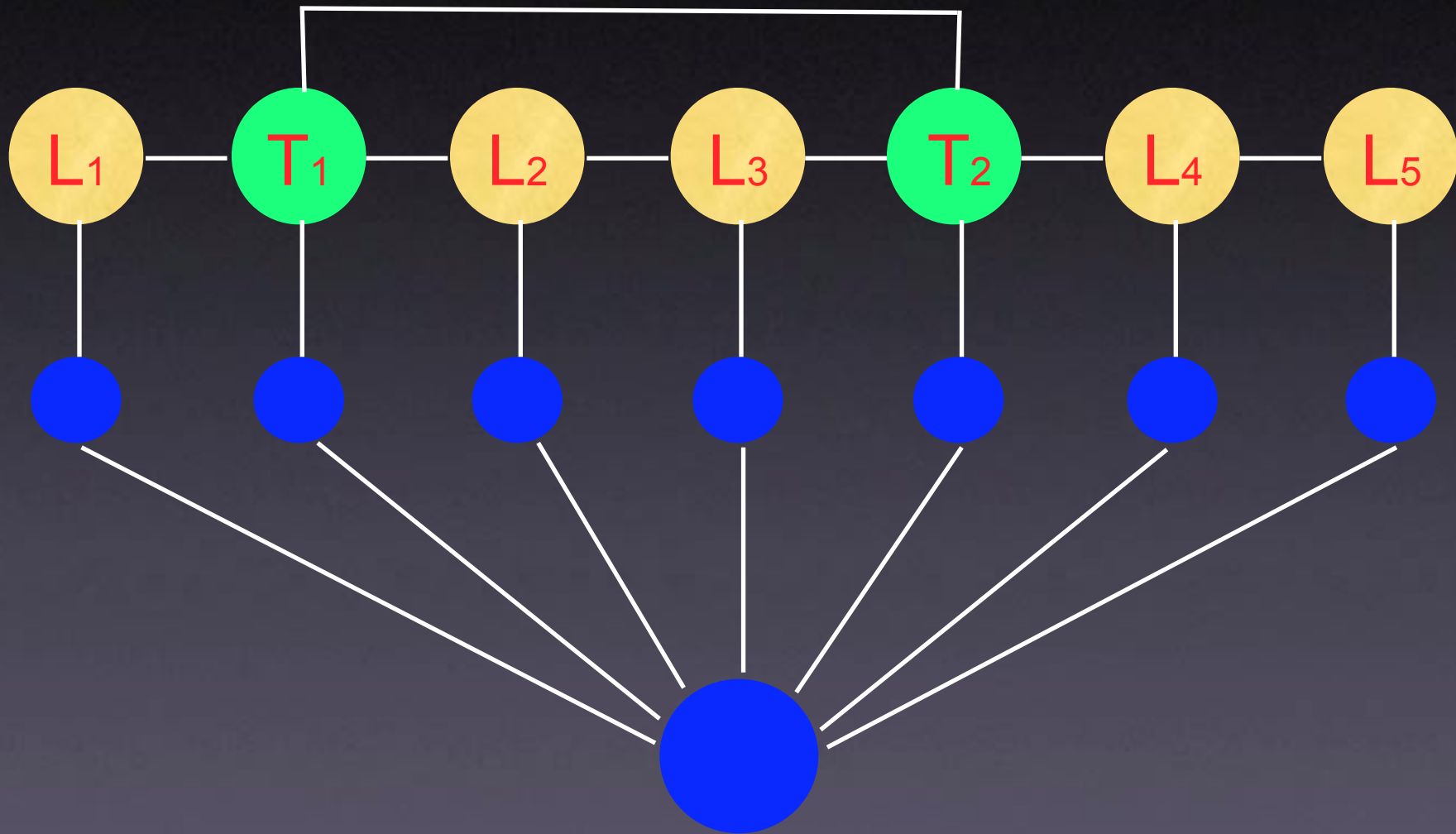


$a \rightarrow b$   
 $b \rightarrow c, d$



a→b  
b→c,d  
c,d→e  
e→f  
f→g  
g→h



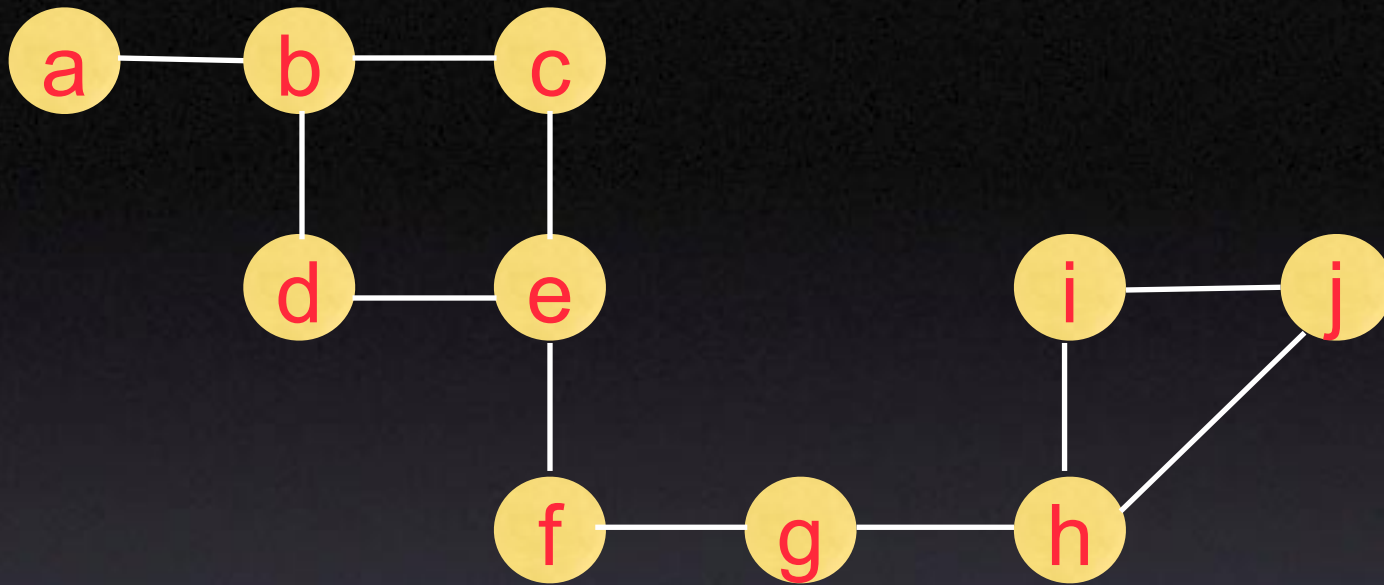


# MCMC

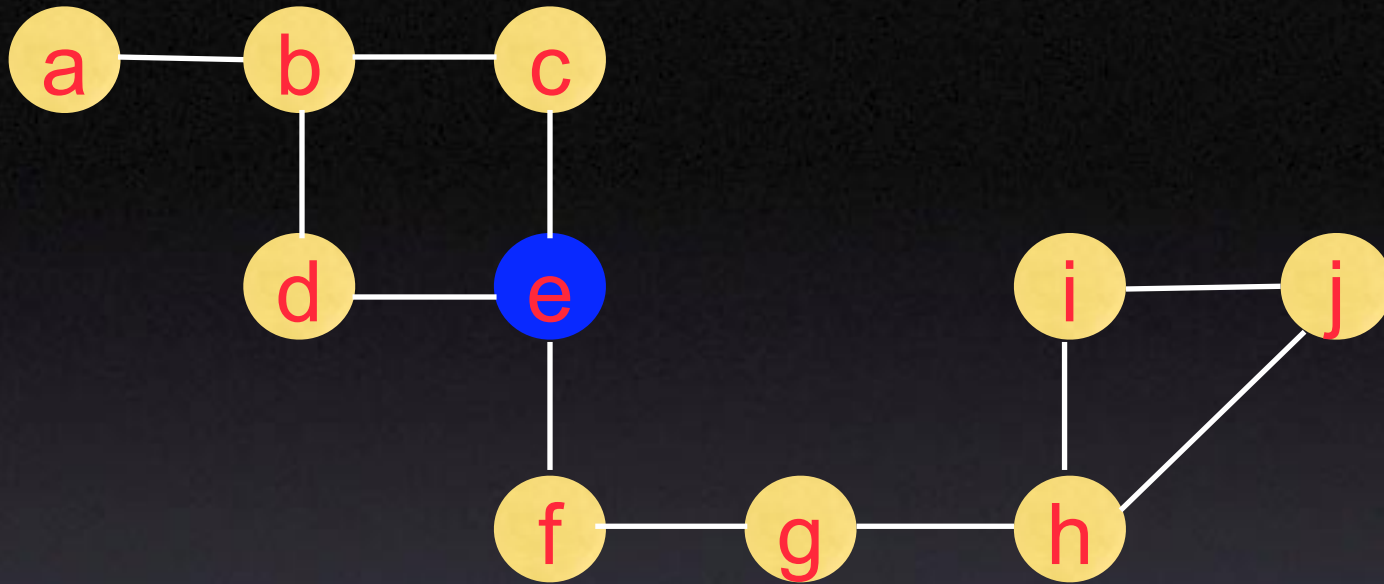
- Technique for producing samples from complicated high-dimensional distributions
- Can be used to perform an approximate summation over missing variables
- Extendible to complex systems

# MCMC

- Produce samples of node states rather than summing over all possible states
- Focus on high probability states
- MC estimates of likelihoods
- Joint posterior distributions of arbitrary sets of variables

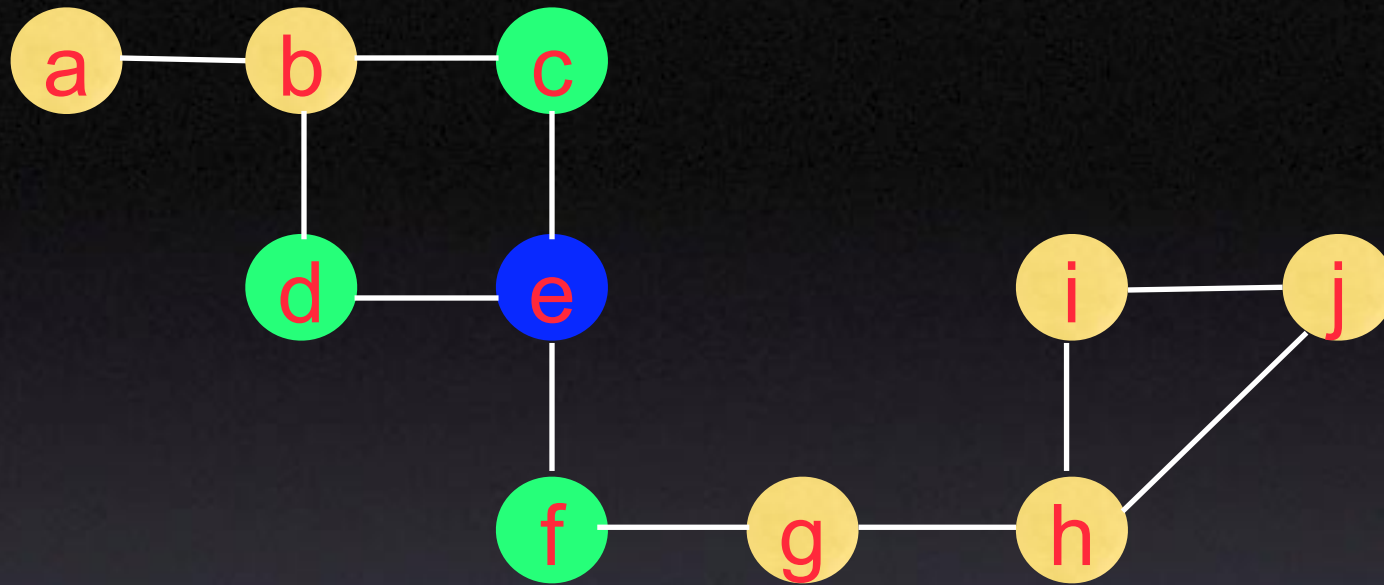


Set initial states of nodes



Set initial states of nodes

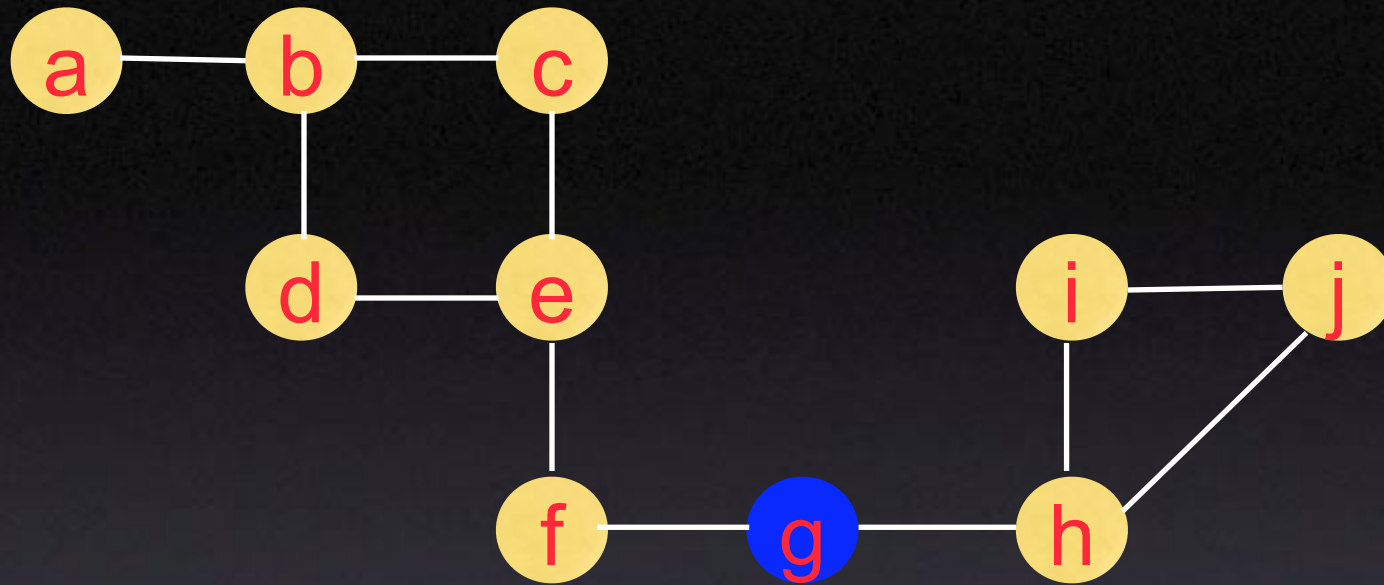
Pick node to update



Set initial states of nodes

Pick node to update

Update conditional on neighbours



Set initial states of nodes

Pick node to update

Update conditional on neighbours

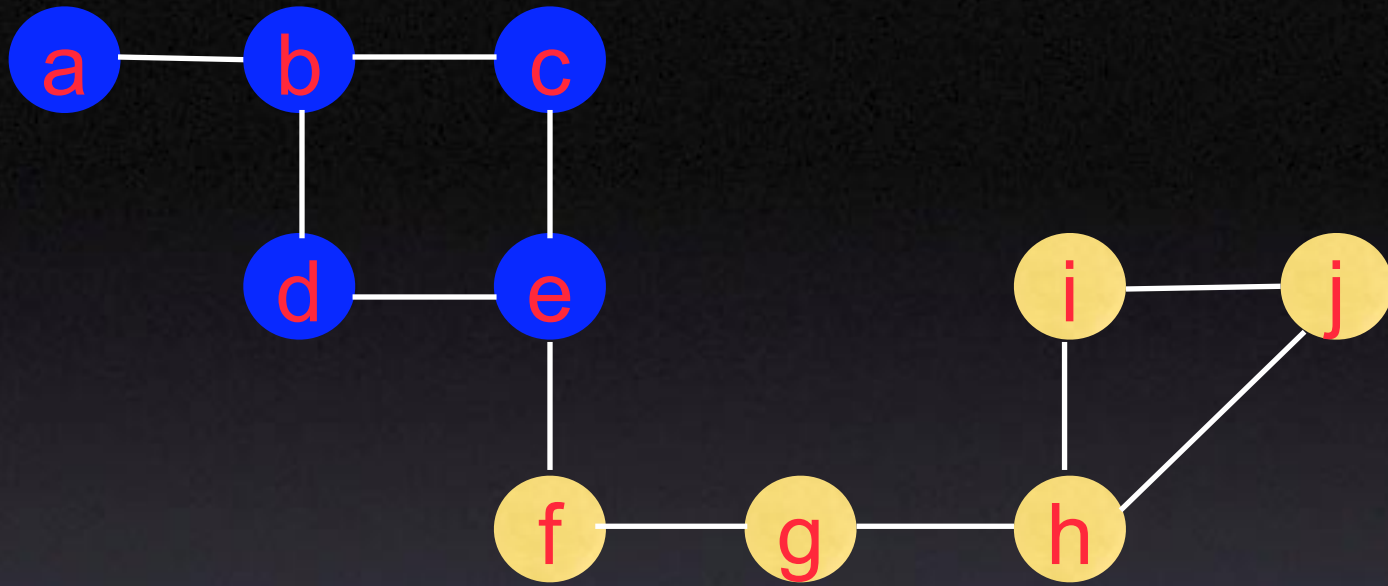
Pick new node to update

# Irreducibility

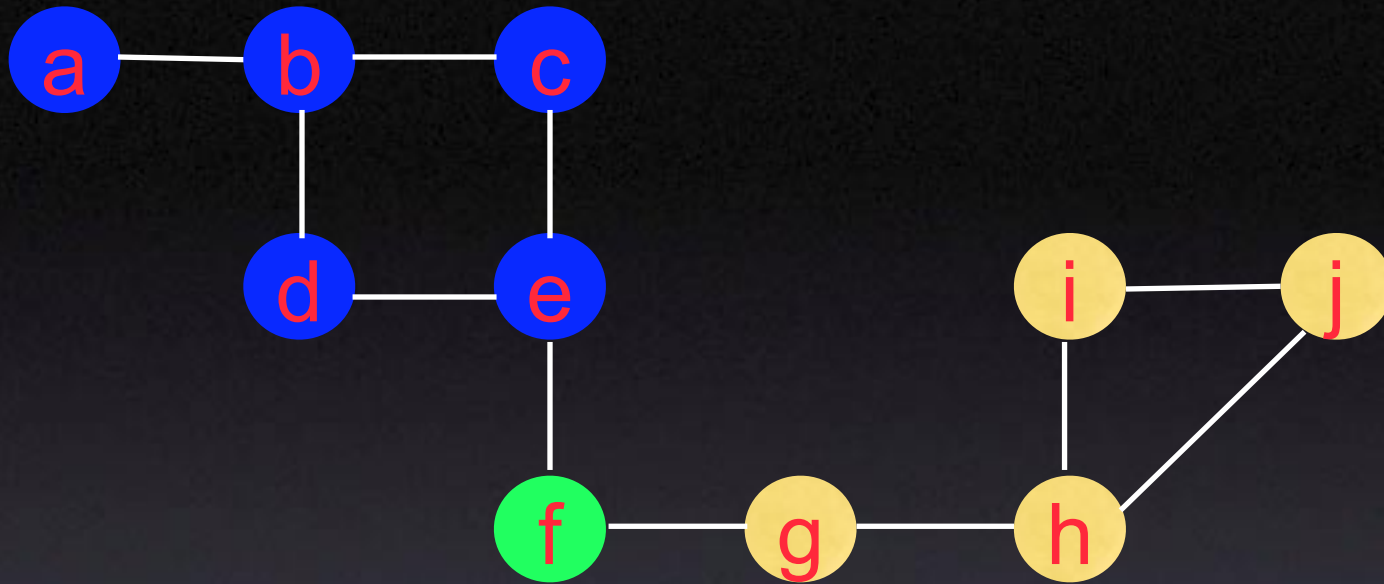
- The sampler should be able to move (eventually) from any possible state to any other possible state
- This can cause problems with single site updates - particularly for family data
- In practice, easy to avoid by updating multiple nodes jointly (except for very complex pedigrees for example)

# Mixing

- A sampler may be irreducible but still have a very low probability of moving between certain states - bad mixing
- A badly mixing sampler will need to be run for many iterations to get good results, and may in fact be practically reducible
- In general, joint sampling of nodes improves mixing (but may be costly)



Set block to update



Set block to update

Update conditional on neighbours

# Applications of MCMC

- Haplotype estimation/analysis
- Population structure determination
- Linkage analysis
- LD mapping using population genetic models
- Contingency tables

# Haplotype estimation

- Set initial frequency estimates  $f$
- For each individual  $i$ , get set of possible haplotype pairs
- Get relative probability for each hap. pair based on  $f$
- Update frequency estimates and repeat
- Get haplotype assignments for each ind.

# Haplotype estimation

- Restriction on the number of loci
- Difficult to use prior knowledge of haps.
- Difficult to assess variability of haplotype estimates if  $f$  is being estimated
  - Accounting for variability in estimate of  $f$ , particularly when  $f$  is multi-modal
- Problem if haplotypes used in subsequent study (i.e., association)

# MCMC for haplotypes

- Two sample blocks - haplotype assignments and haplotype frequencies
- At each iteration, perform association test on current haplotype assignments
- 'Average' results of test over iterations

# Population structure

- Attempt to assign individuals to to a group based on genetic data
- Estimate likelihood for models with different numbers of populations
- Inference on pop. membership/no. pops
- MCMC sampler to explore the space of possible group allocations

# Mixed inheritance models

- Trait has monogenic and polygenic (gaussian) components
- Common in QTL analyses
- Summation over both discrete genotypes and gaussian polygenic variables not generally possible
- Block sample discrete and continuous variables to get an MCMC sampler

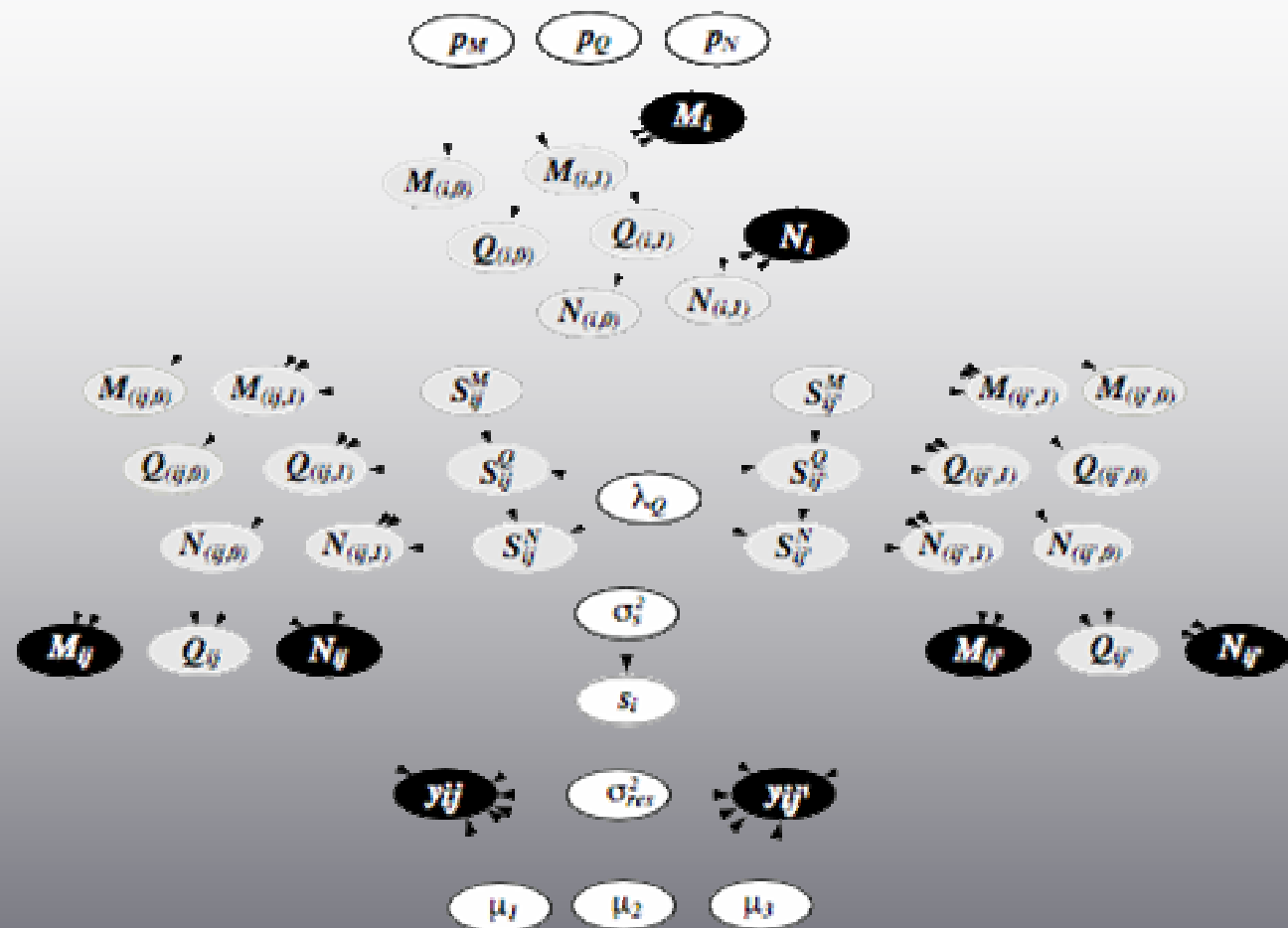


Figure 6: The graphical model taken from Sheehan *et al.* (2002) for the full Bayesian analysis on a half-sib design with one sire  $i$  and two daughters,  $j$  and  $j'$ .

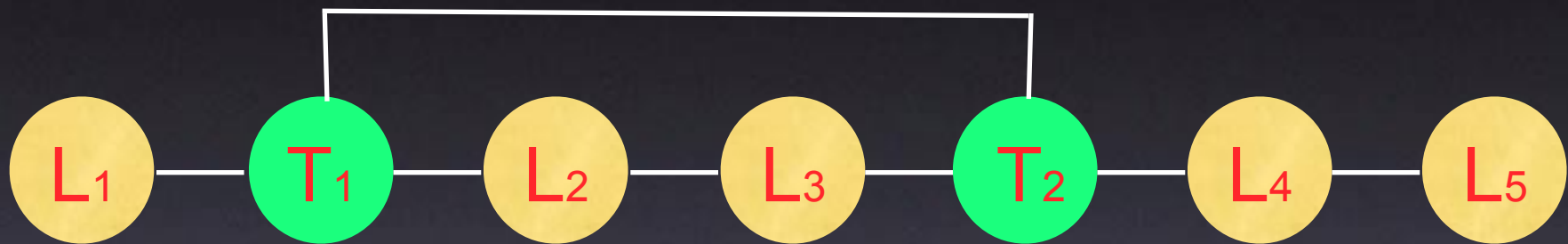
# Large pedigrees

- Exact linkage calculations are limited to either small pedigrees (many markers) or large, simple pedigrees (few markers)
- Large pedigrees with many loci can not be analyzed using exact methods
- Large complex pedigrees can not be analyzed with any number of markers

# Large pedigrees

- For pedigrees simple enough to perform a single locus exact calculation, can update locus by locus
- For pedigrees too complex (too many loops) can update sections of the pedigree - risk of reducibility/poor mixing
- Likelihood estimation/haplotype assignments/IBD estimation possible

# 2 locus models



Using MCMC can get joint distribution of inheritance patterns for both trait loci

# Conclusions

- MCMC is a very flexible and powerful tool
- Extends the range of current analyses, and facilitates more complex analyses
- Allows modular approach to analyses
- Beware of mixing problems!