A Virtual Insitute of Statistical Genetics

Rod Ball

Scion
(NZ Forest Research Institute Limited)

International Biometrics Society,
Taupo, NZ. Nov 29–Dec 3 2009

ABSTRACT

The Virtual Institute of Statistical Genetics is a FRST funded research program involving Universities and Crown Research Institutes, that has just completed its first year of operation.

I will describe progress on current projects (Large Datasets and Polyploids) and plans for the next project (Experimental Designs).

Whole genome prediction of genetic values is already being applied to livestock in New Zealand. The VISG large datasets project is developing methods for whole genome association mapping and prediction of genetic values. Currently we are working on a Bayesian Markov chain Monte Carlo (MCMC) method for fitting associations using a multi-category prior (with 3 categories, with separate variance parameters for each category) and block updates for SNP effects. The modelling approach allows for low prior probabilities for non-negligible SNP effects (necessary for a $p << n$ problem), and a non-normal mixture distribution for effect sizes. Special attention is given to algorithms for improving and diagnosing MCMC convergence to avoid problems with existing QTL or whole genome MCMC methods. Also important are algorithms to effectively handle large datasets with currently of the order of 1,000,000 SNP markers genotyped per individual.

A number of important horticultural, crop, and forage species are polyploids. Existing QTL mapping in polyploids is limited to specific marker types and segregation patterns, and inference is limited. The VISG polyploids project is developing methods for QTL mapping in polyploids which make full use of available marker information and enable multi-locus Bayesian inference of the genetic architecture.

Polyploids have 2 or more sub-genomes resulting in (*e.g.* for an allo-tetraploid) 4 or more alleles at each locus each of which could have been inherited from one of 8 grand-parental chromosomes. Markers are rarely fully informative, so that the statistical method needs to contend with considerable and variable amounts of missing information. This is being done by integrating peeling and conditional peeling with a Bayesian QTL mapping method.

Experimental design has been a neglected area in genomics, with even large scale international projects lacking power to detect any but the largest effects with posterior odds greater than 1. The VISG experimental designs project will develop experimental designs with sufficient power to detect genomic associations, with sufficiently high Bayes factor to overcome the low prior odds for genomic associations, and utilising design and analysis options available in various species (*e.g.* clonal replication and spatial analysis).

Virtual Institute of Statistical Genetics (VISG)

- FRST funded project (NZ strengths)
- 93 page proposal (Scion/MapNet)
- 6 statistical genetics projects proposed
- highly rated
  - ‘fantastic team’
  - ‘addressed the domain review’

Virtual Institute of Statistical Genetics (VISG)

- Previously a neglected area in FRST funding
  - gene mapping statisticians and geneticists in different organisations, working in isolation
  - commonality of problems, but mostly lacked resources to develop statistical methods

Virtual Institute of Statistical Genetics (VISG)

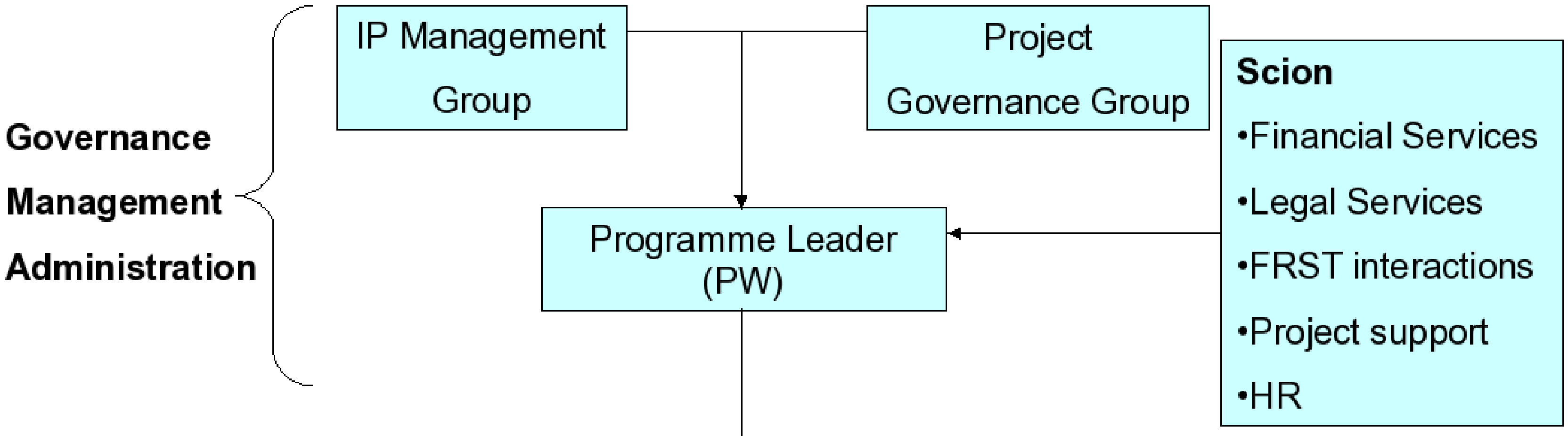
- oversight by Project Governance group (incl. statisticians Bruce Weir, Peter Visscher)

- partnership between geneticists and statisticians

- involvement by NZ research insitutes and Universities

- will develop methods and software (mainly GPL) relevant to NZ end users

VISG STRUCTURE



Governance

Management

Administration

IP Management Group

Project Governance Group

Programme Leader (PW)

Scion

- Financial Services
- Legal Services
- FRST interactions
- Project support
- HR

Science

Teams

Projects	Large Datasets	Polyploids QTL Mapping	Experimental Design	"Project 4"	eResearch Portal
-Statisticians	Rod Ball Ken Dodds Bemoit Auvary Mik Black	Rod Ball Yi Lin Jia Nihal DeSilva	Rod Ball Ken Dodds Bemoit Auvary Mik Black	Sharon Browning Rod Ball	Mik Black
-Geneticists	David Chagne John McCallum Phillip Wilcox Paul Fisher	Gail Timmerman-Vaugh Mark McNeilage Brent Barret	Phillip Wilcox Rod Lea David Chagne PhD student	Gail Timmerman – Vaughan Paul Fisher Geoff Gill David Chagne PhD student	Phillip Wilcox

VISG scientists

Statisticians:
Rod Ball, Scion
Ken Dodds, AgResearch
Benoit Auvray, AgResearch
Mik Black, U. Otago
Sharon Browning, U. Auckland
Nihal De Silva, Plant and Food
Sammie Yilin Jia, Plant and Food

Project Governance Group:
Prof. Bruce Weir, U. Washington
Peter Visscher, QIMR
Elspeth Macrae, Scion
Phillip Wilcox, Scion
Tony Merriman, U. Otago
Gail Timmerman-Vaughan, Plant and Food

Geneticists:
Phillip Wilcox (P/L) Scion
Mark McNeilage, Plant and Food
David Chagné, Plant and Food
Tony Merriman, U. Otago
Geoff Gill, ViaLactia
Gail Timmerman-Vaughan, Plant and Food
John McCallum, Plant and Food
Rod Lea, ESR
Brent Barrett, AgResearch
Paul Fisher, AgResearch

VISG projects

- Large datasets
- Polyploids
- Experimental design (starting this FY)
- One further project to be developed.
- Underpinning methodology and computing.

Large Datasets project

Goal: Whole genome mapping and prediction of genetic value.

Statisticians: Rod Ball(Scion); Ken Dodds(Ag) and Benoit Auvray(Ag).

Geneticists: Phil Wilcox(Scion), Rod Lea(ESR), Tony Merriman(Otago), David Chagné(Plant), Paul Fisher(Ag).

Large Datasets project

- Humans: currently to 1M or more SNP marker genotypes available on ‘chip’, 1000s of individuals
- AgResearch sheep genome: currently 50k markers on ‘chip’, $\sim$ 1000 individuals
- Bayesian MCMC methods
- Multi-category prior
- Block updates
- Special attention to MCMC convergence

Polyploids project

Goal: Extend Bayesian multilocus QTL mapping methods to allo-polyploids.

Statisticians: Rod Ball(Scion), Sammie Yilin Jia(Plant) and Nihal DeSilva(Plant).

Geneticists: Gail Timmerman-Vaughan(Plant), Brent Barrett(Ag), John McCallum(Plant), Mark McNeilage(Plant), Geoff Gill(ViaLactica).

Experimental design project

Goal: Effective experimental designs for genome-wide associations and prediction.

Statisticians: Rod Ball(Scion), Benoit Auvray(Ag), Ken Dodds(Ag), Mik Black(Otago), PhD student.

Geneticists: P. Wilcox(Scion), T. Merriman(Otago), D. Chagné(Plant).

Underpinning methodology and computing

Rod Ball, Mik Black, Canterbury, BestGrid, KAREN

- RJMCMC, hybrid sampler
- Parallel computing (e.g. Rmpi, MPI, GPU?).
- R enhancements (e.g. RCArrays, memo-functions, macros)
- netcdf array based database
- Video conferencing, Evo desktop e-research capability
- Sakai portal

Large datasets—the problem

- Whole genome prediction from markers.

- $X_{n \times p}$
  - n large thousands of individuals.
  - $p >> n$ p very large up to $\sim 1,000,000$ SNP markers.

- Can't fit traditional linear model:

$$y = Xb + \text{error} \tag{1}$$

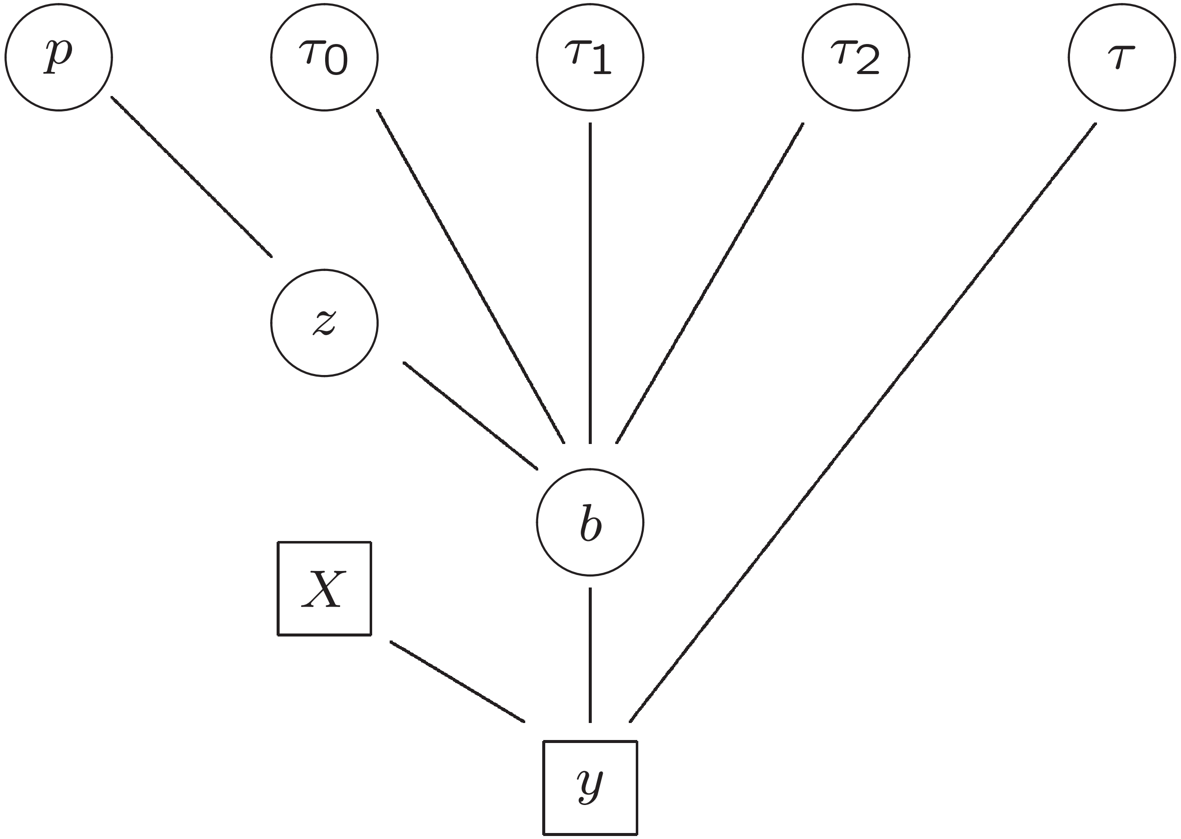
  No degrees of freedom.

- Need specialised models and algorithms.

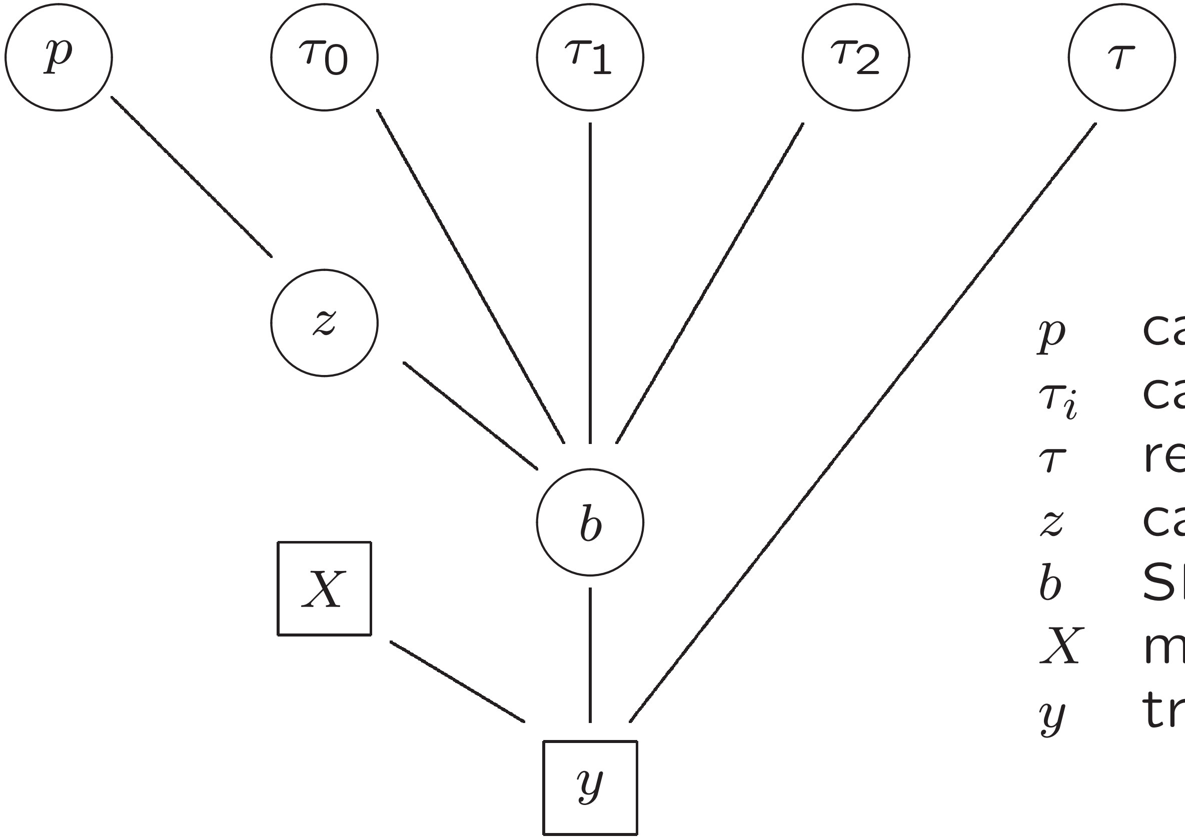
Modelling

- Bayesian model selection approach.
- Models and algorithms for large data.
- Multi-category prior.
- Block updates for effects.

Large datasets hierarchical model



Large datasets hierarchical model



p	cat. proportions
τ_i	cat. precisions
τ	residual precision
z	cat. indicator variables
b	SNP marker effects
X	model matrix (genotypes)
y	trait data

Multi-category prior

Motivation:

- non-normal distribution, approximated by mixture distribution.
- allow for many small effects, keep power to detect larger effects.
- most effects in zero category ($\tau_0 = \infty$) $\Rightarrow$ drop out of the algebra.

$$z_j \in \{0, 1, 2\}, \quad \text{var}(b_j) = \begin{cases} 1/\tau_0 & z_j = 0 \\ 1/\tau_1 & z_j = 1 \\ 1/\tau_2 & z_j = 2 \end{cases} \qquad (2)$$

MCMC sampling

Generate a sample from the posterior distribution by:

- Update one or more parameters at a time from their conditional distribution.
- Repeat for each parameter, and iterate.
- Special attention to improving and diagnosing MCMC convergence, *e.g.*:
  - block updates for b's — blocks corresp. to genome blocks, size corresp. to extent of LD.
  - update from marginals for variance parameters.

Block updates algebra

$$\begin{aligned}
\lambda_j &= \tau_{z_j}/\tau && (3)\\
\Gamma^{-1} &= \mathrm{diag}(\lambda_j) && (4)\\
X'X + \Gamma^{-1} &= R'R && (5)\\
b_0^* &= (y'X + b_0'\Gamma^{-1})R^{-1} && (6)\\
S &= y'y + b_0'\Gamma^{-1}b_0 - b_0^{*'}b_0^* && (7)
\end{aligned}$$

Block updates for b

Sample:

$$b^* \sim N(b_0^*, \tau) \tag{8}$$

backsolve:

$$Rb = b^* \tag{9}$$

Note: Only need to consider columns where $z_j \neq 0$.

Gibbs update for τ_b

(single category case)

$\int \_\_\_ db \rightsquigarrow$

$$[\tau_b \mid \tau_\mu, \tau, y] \propto \tau_b^{A_1+k/2-1} \exp(-B_1\tau_b) \exp(-\frac{S}{2}\tau)|R^{-1}| \qquad (10)$$

Updates for λ_b

- Parameterise in terms of $\lambda_b = \tau_b/\tau$
- Then R, S free of τ.
- $\int \_\_\_ db \rightsquigarrow$

$$[\lambda_b \mid \lambda_\mu, \tau] \propto \lambda_b^{A_1+k/2-1} \exp(-(B_1\lambda_b + S/2)\tau)|R^{-1}| \tag{11}$$

- $\int \_\_\_ d\tau \rightsquigarrow [\lambda_b \mid \lambda_\mu, y] \propto \lambda_b^{A_1+k/2-1}$

$$\times \frac{\Gamma(A_0 + A_1 + A_2 + n_\lambda + \frac{n}{2} - 2)}{(B_0\lambda_\mu + B_1\lambda_b + B_2 + S/2)^{(A_0+A_1+A_2+n_\lambda+\frac{n}{2}-2)}}|R^{-1}| \tag{12}$$

Implementation progress

- sampler with block updates for b's implemented.
- tested on simple dataset and some testing on sheep dataset
- some difficulties with sheep data
- investigating
  - more informative priors/lower prior probability for effects
  - move types for better mixing between categories
  - path sampling

Polyploids—the problem

- QTL mapping–marker trait association in families or pedigrees.
- Allo-polyploids, multiple sub-genomes retaining their identity *e.g.* AABB, AABBCC.
- In diploid plant families can generally infer segregation pattern and linkage phase.
- In polyploids have up to *e.g.* 8 alleles in allo-tetraploids that could have been inherited at a given locus.
- $\Rightarrow$ missing information.

Polyploids — Example 1

Example 1. Marker on a polyploid chromosome.

```
A ---------M---------------------------
A ---------m---------------------------

B ---------m---------------------------
B ---------m---------------------------

Marker phenotype: M.
```

Polyploids – Example 2

Example 2. Counter-example: correlation does not imply causation (linkage).

Example marker on a polyploid chromosome.

```
A ---------M1--------M2----------m3---------
A ---------m1--------m2----------m3---------

B ---------m1--------M2----------M3---------
B ---------m1--------m2----------m3---------
```

M1 correlated with M2 on sub-genome A,
M2 correlated with M3 on sub-genome B.

Spurious linkage induced between M1 and M3.

Polyploids – Example 2 (ctd)

Another possible model representing the same data.

```
A ---------M1-----------M2--------------M3---------
A ---------m1-----------m2--------------m3---------

B ---------m1-----------m2--------------m3---------
B ---------m1-----------m2--------------m3---------
```

Same expected correlation between M1 and M2, and between M2 and M3, but further apparent distance.

M3 on the wrong subgenome.

Polyploids QTL mapping previous work

- Considered only single markers (*e.g.* Doerge and Craig 2000) or pairs of markers (*e.g.* Cao *et al.* 2005).
- Single locus *i.e.* test for a single QTL versus no QTL separately at each locus (*e.g.* Cao *et al.* extend interval mapping).
- $\Rightarrow$ Lack the benefits of Bayesian multilocus approach.
- But, flanking markers may not even be informative.

VISG Polyploids Modelling

- Bayesian model selection approach
- BIC method and/or RJMCMC
- Multiple imputation to handle missing information.
- Peeling and conditional peeling along chromosomes to sample from the missing information.

BIC method (Ball; Genetics 2001)

- A Bayesian model selection method

- A non-MCMC, multi-locus QTL mapping method

- considers multiple models representing alternate QTL genetic architectures according to their probabilities

- QTL architectures represented to within the resolution of the marker map by linear regression on subset of selected markers.

BIC method

- avoids selection bias (Miller 1990, Beavis 1994) where the same data is used to select loci, and estimate the size of their effects, due to over-estimated effects being more likely to be selected.

- Missing values estimated by multiple imputation.

BIC method example:

Sample results (Cf. AMP, Table 7.4, p119).

Table 1. Top 10 models for a linkage group with 5 markers.

	markers								
model	M1	M2	M3	M4	M5	k	R^2	prob	cum.p
1	F	T	F	F	F	1	18.6	50.5	50.5
2	F	F	T	F	F	1	17.4	28.0	78.4
3	F	T	T	F	F	2	23.8	9.0	87.4
4	F	T	F	T	F	2	22.7	4.9	92.3
5	F	T	F	F	T	2	21.5	2.7	95.0
6	F	F	F	F	F	0	0.0	1.1	96.1
7	T	F	T	F	F	2	19.6	1.0	97.1
8	T	T	F	F	F	2	18.9	0.7	97.8
9	F	F	T	F	T	2	18.3	0.5	98.4
10	F	F	T	T	F	2	17.6	0.4	98.8
total	2.1	68.5	39.5	5.9	3.7				100.0

Bayesian Model selection/BIC method

Note inference of genetic architecture (to within the resolution of the marker map):

- postprob for number of QTL in a region
- postprob for QTL in the vicinity of a marker
- unbiased estimates of QTL effects—avoid selection bias by considering all models, not just models where the effect is selected or ‘significant’ (unlike interval mapping).

Multiple imputation

- jointly analyse multiple copies of the data with independent randomly sampled values for the missing data adjusting the likelihood appropriately

- need to sample from the distribution of missing marker information

- will sample from a set of fully infomative ‘virtual markers’ using a variant of ‘peeling’.

Peeling (Elston and Stewart 1971).

- Missing information problem for diploid human pedigrees because of small family size.

- Peeling—exhaustively evaluates joint or marginal probabilities in a pedigree.

- Feasible for several markers simulataneously.

- A special case of graphical models methods (*e.g.* Lauritzen and Spiegelhalter 1988; Thomas *et al.* 2000).

Peeling

- Summation over progeny in parent progeny triples.
- Remove the progeny from the graph. (It is peeled away).
- Repeat for all progeny.

End result is a function on remaining node or a value (likelihood) for the model in terms of any parameters θ that are conditioned on in the above process.

Reverse the steps (reverse peeling) obtaining a random sample from the distribution.

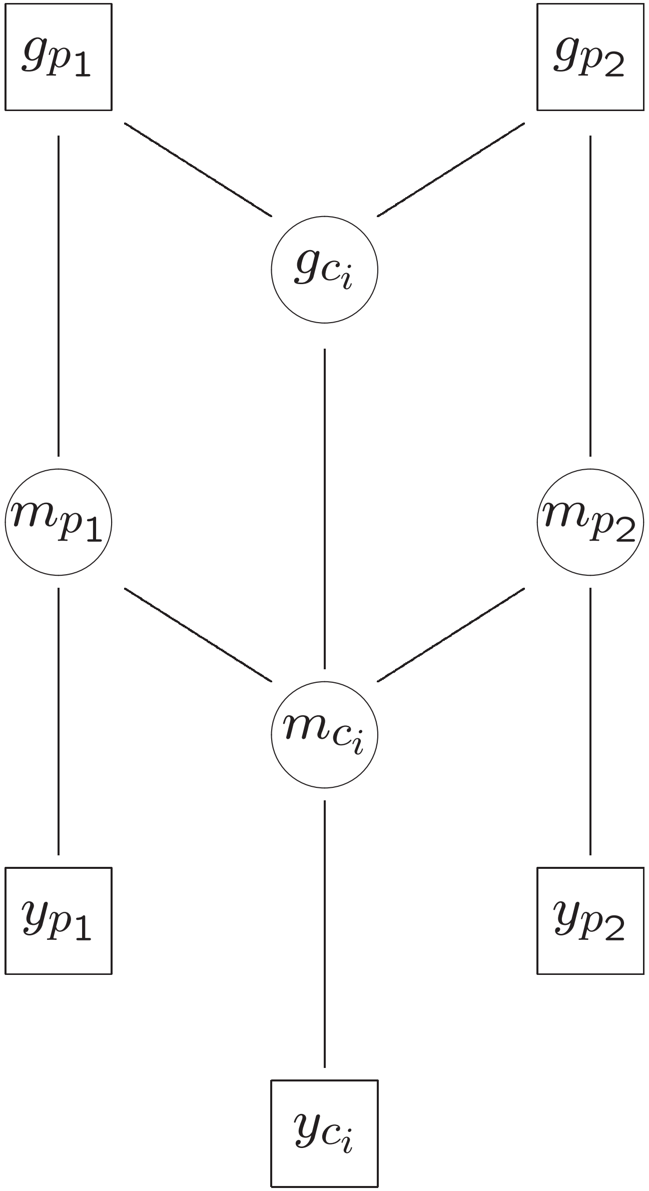
VISG peeling

- Peeling to sample from the virtual marker genotypes at one locus.

- Conditional peeling. Peel sequentially along the genome, conditional on previously sampled values and recombination rates.

- Complexities (not yet being addressed)
  - recombination rates may vary between parents (male or female) and sub-genomes
  - recombination distances between markers may also need to be estimated.
  - recombination rates not known $\Rightarrow$ need pairwise peelings

Hierarchical model for allo-tetraploids



y_{p_1}, y_{p_2}	parental marker phenotypes (optional)
y_{p_i}	progeny marker phenotypes
m_{p_1}, m_{p_2}	parental marker genotypes
m_{c_i}	progeny marker genotypes
g_{p_1}, g_{p_2}	parental f.i. virtual markers (given)
g_{c_i}	progeny f.i. virtual markers

Peeling equations for an allo-tetraploid

Joint distribution:

$$\begin{aligned} f(m_{p_1}, m_{p_2}, \ldots) = {} & [m_{p_1}][m_{p_2}][y_{p_1}|m_{p_1}][y_{p_2}|m_{p_2}] \times \\ & \prod_{i=1}^{n} [g_{c_i}|g_{p_1}, g_{p_2}][m_{c_i}|g_{c_i}, m_{p_1}, m_{p_2}][y_{c_i}|m_{c_i}] \end{aligned} \tag{13}$$

where:

$$[y_{c_i}|m_{c_i}] = I(f_2(m_{c_i}) = y_{c_i}) \tag{14}$$

f_2: ‘2nd forgetful function’.

Peeling equations for an allo-tetraploid (ctd.)

Let:

$$\begin{aligned} R_{gci}(g_{c_i}, m_{p_1}, m_{p_2}) &= \sum_{m_{c_i}} [m_{c_i} | g_{c_i}, m_{p_1}, m_{p_2}][y_{c_i} | m_{c_i}] \\ &= I(f_2(c(m_{p_1}, m_{p_2})[g_{c_i}]) = y_{c_i}) \qquad (15) \end{aligned}$$

Joint distribution after ‘peeling’ progeny marker genotypes m_{c_i}:

$$\begin{aligned} f(m_{p_1}, m_{p_2}, g_{c_1}, g_{c_2}, ...) = {} & [m_{p_1}][m_{p_2}][y_{p_1} | m_{p_1}][y_{p_2} | m_{p_2}] \times \\ & \prod_{i=1}^{n} [g_{c_i} | g_{p_1}, g_{p_2}] R_{gci}(g_{c_i}, m_{p_1}, m_{p_1}) \qquad (16) \end{aligned}$$

Peeling for an allo-tetraploid (ctd.)

- Continue, peel back to marginal distribution $[m_{p_1}]$, and sample from that distribution.

- Reverse the process sampling from each value in turn.

Progress

- Three year project workplan written and accepted by the Project Governance Group.
- Peeling equations derived.
- R functions for peeling and conditional peeling implemented by Sammie Jia. Being tested at Plant and Food.
- C functions being developed.

Experimental designs—the problem

Previously many association spurious:

- Use of p-value thresholds which correspond to weak evidence especially with large sample sizes.
- Experiments designed with power to obtain a given p-value $\Rightarrow$ not powerful enough.

Experimental designs—methods

Adapt existing frequentist power calculations (design for p-value threshold equivalent to desired Bayes factor; cf. Ball Genetics 2005 for associations in unstructured populations.)

Extend to other designs *e.g.*

- Humans:
  - case-control studies

- Plants:
  - control spatial variation
  - utilise clonal replication
  - optimal designs

Virtual Institute of Statistical Genetics (VISG)
– PhD Studentships –

1. Beginning Oct 2009. Develop Bayesian methods for design of association mapping experiments.
   - Plant species
   - Human genetics applications.

2. Beginning January 2011. Topic yet to be confirmed.

Contact:	Rod Ball	`rod.ball@scionresearch.com`
or	Phil Wilcox	`phil.wilcox@scionresearch.com`

Acknowledgements

This work was funded by the New Zealand Foundation for Research Science and Technology through a contract with Scion and sub-contracts with AgResearch, Plant and Food Research, ESR, and Otago University.