Bayesian estimation of multinomial probabilities with non-unique cell classification: Application to trisomy 21 data

Nokuthaba Sibanda

School of Mathematics, Statistics & Operations Research
Victoria University

2 December 2009

Acknowledgements

Prof Stephanie Sherman, Emory University, Atlanta, USA

A/Prof Eleanor Feingold, University of Pittsburgh, Pennsylvania, USA

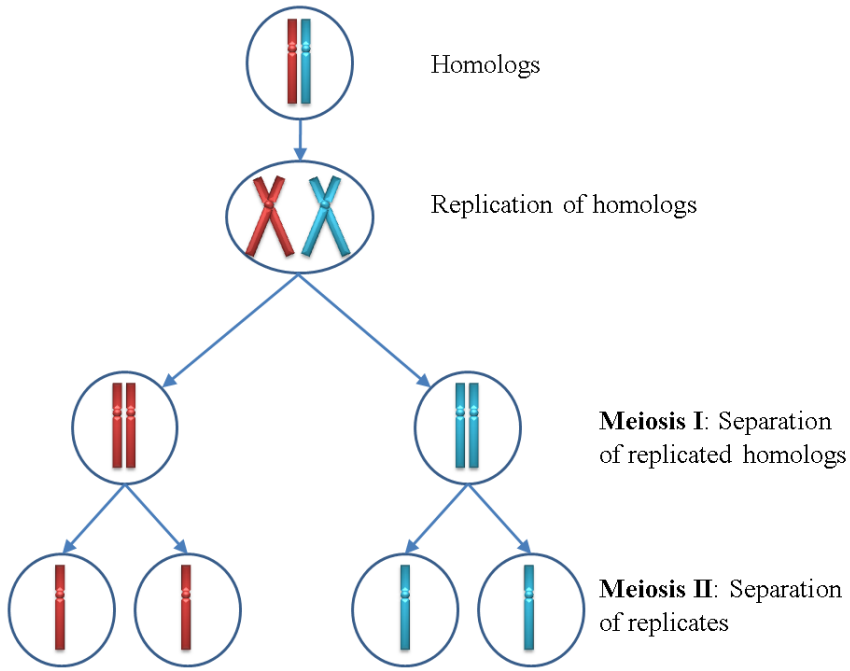
Ray Tobler, Victoria University Wellington

Introduction

- Motivating problem
- Single locus analysis
- Multilocus analysis

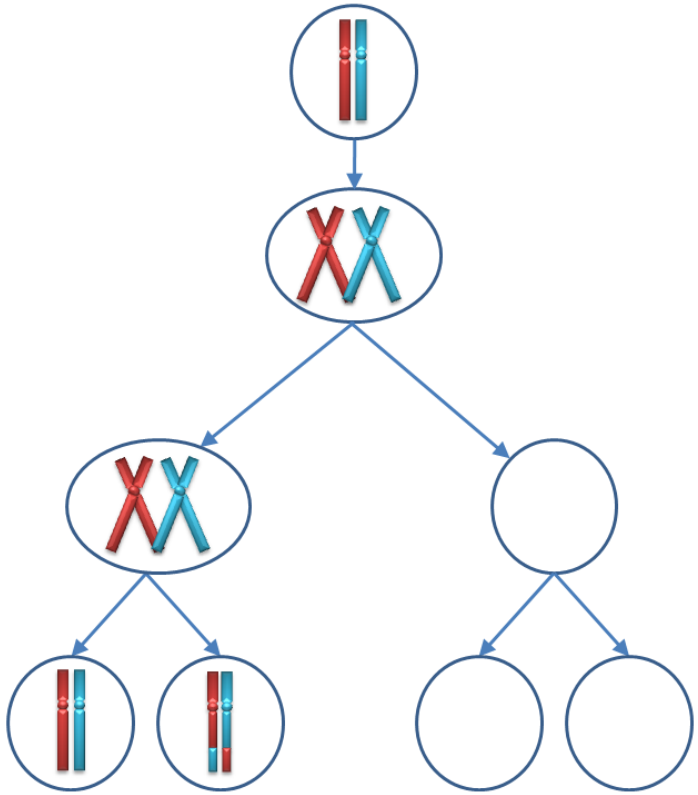
Meiotic nondisjunction

Meiotic nondisjunction is the failure of chromosomes to separate during meiosis. This leads to aneuploid gametes, and subsequently trisomy in the offspring.

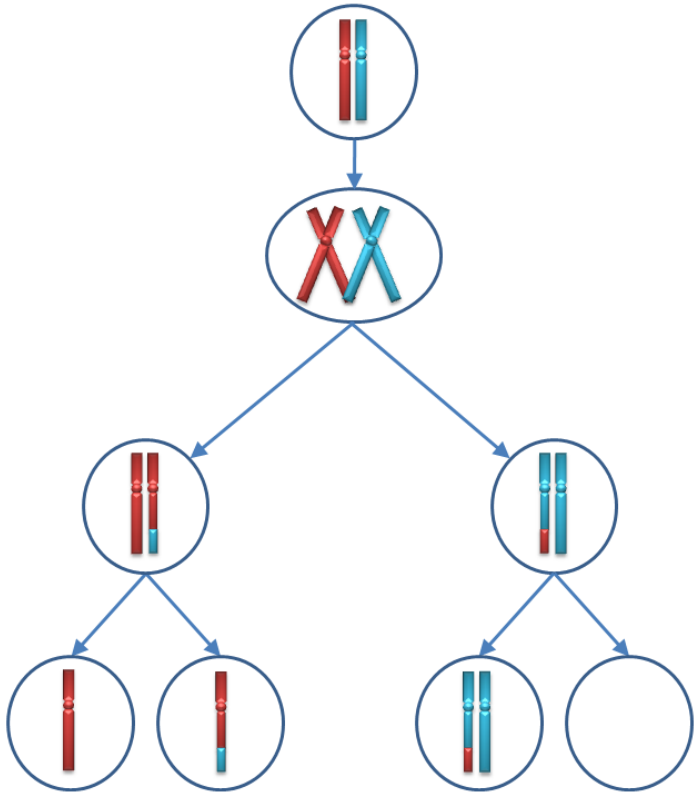


Successful disjunction

Meiosis I nondisjunction



Meiosis II nondisjunction



Motivating problem

Estimation of probabilities of nondisjunction in males and females at the first and second stage of meiosis.

Problem: When looking at genotype data from one locus or more, it may not be possible to determine exactly in which parent and at what stage nondisjunction occurred.

$$m_i = {}^{12}\quad , p_i = {}^{11}\quad , c_i = {}^{112}$$

Parent	Stage	
	I	II
Maternal	✓	✗
Paternal	✓	✓

Motivating problem

Estimation of probabilities of nondisjunction in males and females at the first and second stage of meiosis.

Problem: When looking at genotype data from one locus or more, it may not be possible to determine exactly in which parent and at what stage nondisjunction occurred.

$$m_i = \begin{matrix} 12 \\ 13 \end{matrix}, p_i = \begin{matrix} 11 \\ 12 \end{matrix}, c_i = \begin{matrix} 112 \\ 113 \end{matrix}$$

Parent	Stage	
	I	II
Maternal	✓ ✓	✗ ✗
Paternal	✓ ✗	✓ ✓

Missing data?

The problem can be viewed as a two-way contingency table with missing observations for families where the origin of non-disjunction cannot be uniquely identified.

Method: Imputation and data augmentation (Tanner and Wong, 1987).

Data

Genotype data, from Single Nucleotide Polymorphisms (SNPs) and Short Tandem Repeats (STRs) proximal to the centromere, for a child with trisomy 21 and their mother and father is used.

Data were available at 100 positions on the long arm of chromosome 21 for 350 families with origin of non-disjunction confirmed as follows:

Parent	Stage		
	I	II	Total
Maternal	256 (73.1%)	90 (25.7%)	346 (98.8%)
Paternal	1 (0.3%)	3 (0.9%)	4 (1.2%)

Notation

For a SNP, define
$\mathcal{G} = \{11, 12, 22\}$
$\mathcal{C} = \{111, 112, 122, 222\}$,
where $\{1, 2\} = \{A, G\}$ or $\{T, C\}$.

For a family i, the data available are $\{c_i, m_i, p_i\}$, where $c_i \in \mathcal{C}$, $m_i, p_i \in \mathcal{G}$.

c_i is completely determined by the parental and meiotic stage of nondisjunction.

Notation

For a SNP, define
$\mathcal{G} = \{11, 12, 22\}$
$\mathcal{C} = \{111, 112, 122, 222\}$,
where $\{1, 2\} = \{A, G\}$ or $\{T, C\}$.

For a family i, the data available are $\{c_i, m_i, p_i\}$, where $c_i \in \mathcal{C}$, $m_i, p_i \in \mathcal{G}$.

c_i is completely determined by the parental and meiotic stage of nondisjunction.

Is it possible to use c_i, m_i and p_i at given loci to estimate nondisjunction probabilities?

Nondisjunction probabilities

Define events
M_I = maternal meiosis I nondisjunction = E_1
M_{II} = maternal meiosis II nondisjunction = E_2
P_I = paternal meiosis I nondisjunction = E_3
P_{II} = paternal meiosis II nondisjunction = E_4

The aim is to estimate the probability vector $\phi_1 = \Pr(E_1), \phi_2 = \Pr(E_2), \phi_3 = \Pr(E_3)$ and $\phi_4 = \Pr(E_4) = 1 - \sum_{j=1}^{3} \phi_j$.

Single locus exact likelihood

The exact likelihood for a set of n independent families is

$$\begin{aligned} L(\boldsymbol{\phi}|\mathbf{c},\mathbf{m},\mathbf{p}) &= \prod_{i=1}^{n} \Pr(c_i|m_i,p_i,\boldsymbol{\phi}) \Pr(m_i,p_i) \\ &\propto \phi_1^{n_1}\phi_2^{n_2}\phi_3^{n_3}\phi_4^{n_4} \prod_{i=1}^{n-\sum_{j=1}^{4} n_j} \sum_{j=1}^{4} a_{ij}\phi_j, \end{aligned}$$

where n_j is the number of families in which $Pr(c_i|m_i,p_i,E_k) = 0 \ \forall k \neq j$ and $E_j \in \mathcal{E} = \{M_I, M_{II}, P_I, P_{II}\}$.

For example, consider a family with $c_i = 112$, $m_i = 11$ and $p_i = 12$. Then $\Pr(c_i|\boldsymbol{\phi}, m_i, p_i) = \left(\frac{1}{2}\phi_1 + \frac{1}{2}\phi_2 + \phi_3\right)$.

Augmented data likelihood

Introduce a latent variable Z, the event assigned to a family from $\mathcal{E}$.
For the i^{th} family, Z_i is selected with probability $\Pr(Z_i = E_j|\boldsymbol{\phi}, c_i, m_i, p_i) = \Pr(E_j|\boldsymbol{\phi}, c_i, m_i, p_i)$ for a given j.
For example, consider a family with $c_i = 112$, $m_i = 11, p_i = 12$ and $\Pr(c_i|\boldsymbol{\phi}, m_i, p_i) = \left(\frac{1}{2}\phi_1 + \frac{1}{2}\phi_2 + \phi_3\right)$. Here Z_i takes the value E_3, say, with probability $\frac{\phi_3}{\frac{1}{2}\phi_1+\frac{1}{2}\phi_2+\phi_3}$.
The augmented data likelihood is

$$L^*(\boldsymbol{\phi}|\mathbf{Z}, \mathbf{c}, \mathbf{m}, \mathbf{p}) = \prod_{i=1}^{n}\prod_{j=1}^{4} a_{ij} I(Z_i = E_j)\phi_j \quad \propto \quad \prod_{j=1}^{4} \phi_j^{n_j^*},$$

where $\mathbf{Z} = \{Z_i : i = 1, \ldots, n\}$, n_j^* is the number of families for which $Z_i = E_j$.

Posterior distribution

A conjugate prior, the Dirichlet distribution, was used for both likelihood functions. A Dirichlet distribution has density function $\pi(\phi_1, \ldots, \phi_4) \propto \phi_1^{\alpha_1 - 1} \ldots \phi_4^{\alpha_4 - 1}$ for $\sum_{j=1}^{4} \phi_j = 1$ and $\alpha_j > 0,\ j = 1, \ldots, 4$. The resulting posterior densities were

$$\begin{aligned} \pi(\boldsymbol{\phi}|\mathbf{c}, \mathbf{m}, \mathbf{p}) &\propto \prod_{j=1}^{4} \phi_j^{n_j + \alpha_j - 1} \prod_{i=1}^{n - \sum_{j=1}^{4} n_j} \sum_{j=1}^{4} a_{ij} \phi_j \\ \pi^*(\boldsymbol{\phi}|\mathbf{Z}, \mathbf{c}, \mathbf{m}, \mathbf{p}) &\propto \prod_{j=1}^{4} \phi_j^{n_j^* + \alpha_j - 1}. \end{aligned}$$

The terms $(\alpha_j - 1),\ j = 1, \ldots, 4$, in the prior, can be interpreted as the *a priori* expected numbers of families in which only event E_j occurred out of a total of $\sum_{j=1}^{4}(\alpha_j - 1)$ families.

MCMC sampling

A Metropolis-Hastings sampler combined with a change of variable method was used to generate samples from $\pi(\boldsymbol{\phi}|\mathbf{c}, \mathbf{m}, \mathbf{p})$.

Gibbs sampler steps were used to sample $\mathbf{Z}$ and $\boldsymbol{\phi}$ from their full conditional distributions.

SNP rs2259403, 13.62Mbp from centromere on chromosome 21q

305 families with origin of non-disjunction as follows:

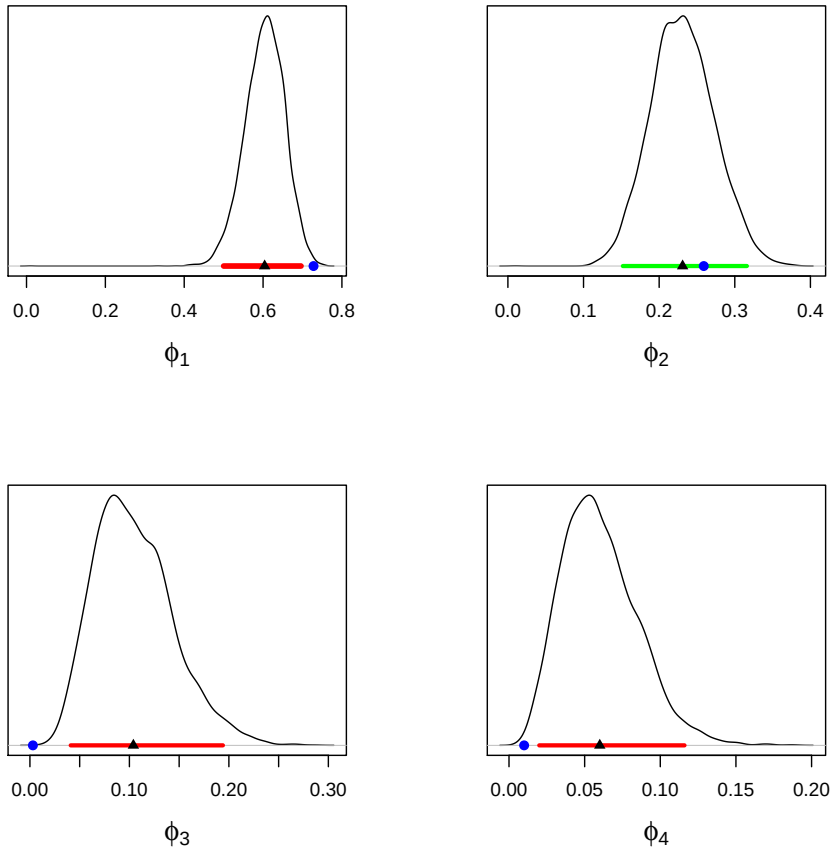
Parent	Stage		Total
	I	II	
Maternal	222 (72.8%)	79 (25.9%)	301 (98.7%)
Paternal	1 (0.3%)	3 (1.0%)	4 (1.3%)

Estimation using SNP rs2259403

Uniform prior with $\alpha_1 = \ldots = \alpha_4 = 1$

	Value (%)	Exact likelihood estimate		Augmented data likelihood estimate		
		Posterior mean (%)	95% Credible interval(%)	Mean no. assigned E_j	Posterior mean (%)	95% Credible interval(%)
ϕ_1	73.1	68.5	(58.7, 77.2)	210	68.5	(58.6, 77.1)
ϕ_2	25.7	26.0	(17.6, 35.0)	79	26.0	(17.7, 34.8)
ϕ_3	0.3	3.8	(0.1, 11.8)	11	3.8	(0.1, 12.1)
ϕ_4	0.9	1.7	(0.04, 6.0)	4	1.7	(0.05, 5.8)

Estimation using SNP rs2259403



Marginal posterior distributions and 95% posterior credible intervals for prior with $\alpha_1 = \ldots = \alpha_4 = 5$ applied to augmented data likelihood.

STR D21S215, 13.72Mbp from centromere

STR with 10 allelles - more informative.

291 families with origin of non-disjunction as follows:

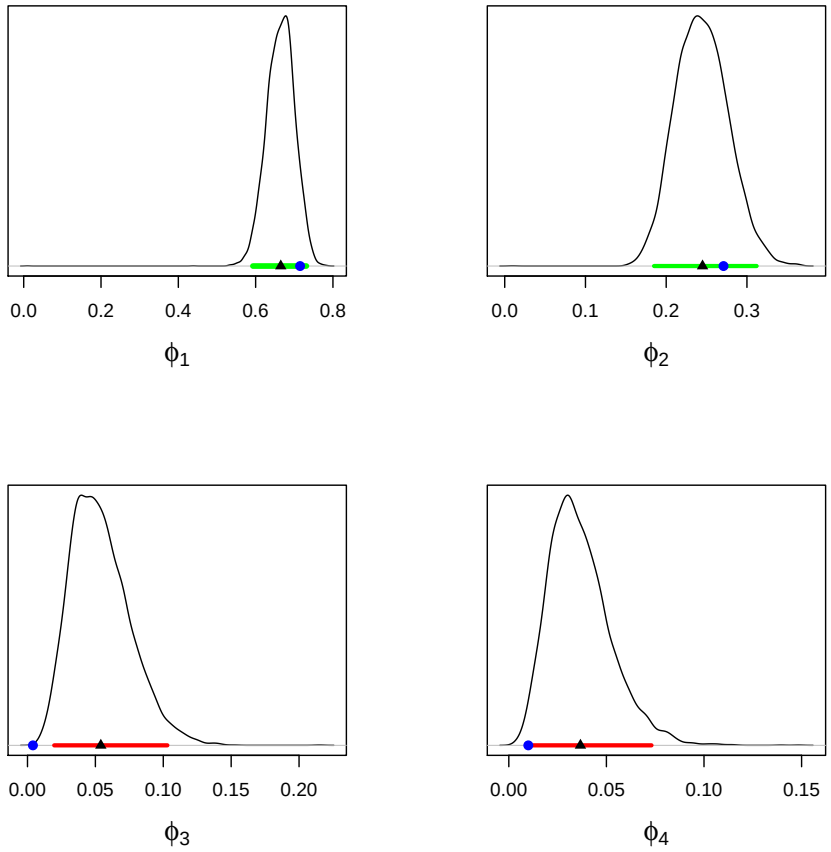
Parent	Stage		Total
	I	II	
Maternal	208 (71.5%)	79 (27.1%)	287 (98.6%)
Paternal	1 (0.4%)	3 (1.0%)	4 (1.3%)

Estimation using STR D21S215

Uniform prior with $\alpha_1 = \ldots = \alpha_4 = 1$

	Value (%)	Exact likelihood estimate		Augmented data likelihood estimate		
		Posterior mean (%)	95% Credible interval(%)	Mean no. assigned E_j	Posterior mean (%)	95% Credible interval(%)
ϕ_1	**73.1**	71.3	(64.3, 77.8)	191	71.4	(64.3, 77.9)
ϕ_2	**25.7**	25.7	(19.4, 32.5)	70	25.5	(19.2, 32.4)
ϕ_3	**0.3**	2.0	(0.08, 6.0)	29	1.9	(0.08, 5.7)
ϕ_4	**0.9**	1.1	(0.03, 3.6)	14	1.1	(0.04, 3.7)

Estimation using STR D21S215



Marginal posterior distributions and 95% posterior credible intervals for prior with $\alpha_1 = \ldots = \alpha_4 = 5$ applied to augmented data likelihood.

Multilocus analysis

Use information from multiple loci simultaneously.
Incomplete reporting for different families at different loci e.g.

FamilyID	Locus	A1	A2	A3	Position (bp)
4710028	rs3126383	1	2	2	13657361
4710028	D21S215	167	171	173	13719788
4710052	rs2259403	1	2	2	13615252
4710052	rs3126383	1	1	2	13657361
4710079	rs2259403	1	1	2	13615252
4710079	rs3126383	1	1	1	13657361
4710079	D21S215	166	168	168	13719788

Use as much of the available information as possible

- If data available at 2 or more loci, select the most informative locus.
- Use data augmentation to estimate $\phi_1, \ldots, \phi_4$.

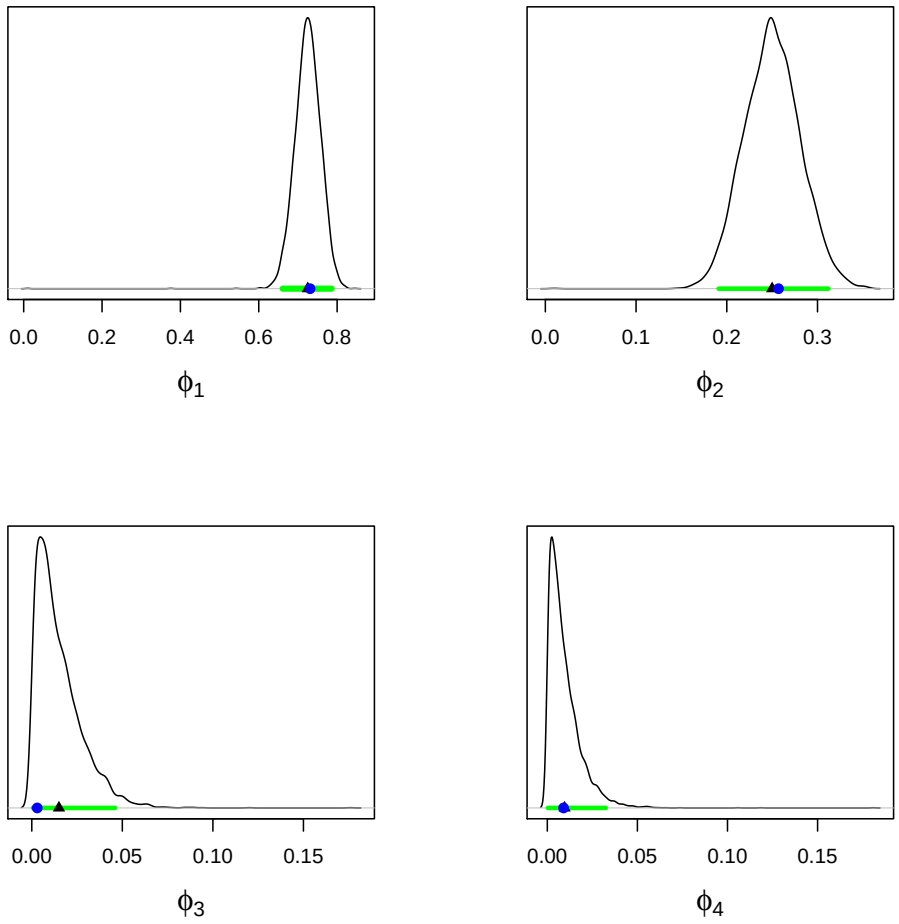
Results for three loci

- Combined information for 2 SNPs and 1 STR closest to the centromere.
- All families had complete genotype reporting (mother, father and child) on at least one of these loci.

Uniform prior with $\alpha_1 = \ldots = \alpha_4 = 1$

	Nondisjunction probabilities			No. of families assigned E_j ($Z_{.j}$)		
	Value (%)	Posterior mean (%)	95% Credible interval(%)	Value	Posterior mean	95% Credible interval
ϕ_1	73.1	72.5	(66.1, 78.6)	256	255	(240, 269)
ϕ_2	25.7	25.0	(19.1, 31.1)	89	88	(74, 102)
ϕ_3	0.3	1.5	(0.06, 4.6)	1	4	(0, 13)
ϕ_4	0.9	1.0	(0.03, 3.2)	3	2	(0, 9)

Results for three loci



Marginal posterior distributions and 95% posterior credible intervals for prior with $\alpha_1 = \ldots = \alpha_4 = 5$.

References

1. Ott J (1991). Analysis of human genetic linkage. 2nd ed. John Hopkins University Press, Baltimore.
2. Smith CAB, Stephens DA (1997). Simple likelihood and probability calculations for linkage analysis. In *Genetic Mapping of Disease Genes*.
3. Swartz T, Haitovsky, Vexler A and Yang T (2004). Bayesian identifiability and misclassification in multinomial data. *The Canadian Journal of Statistics* **32**: 285-302.
4. Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82, 528-550.