

Statistical methods to analyze genomic aberrations
in cancer cells: The case of overlapping ensembles.

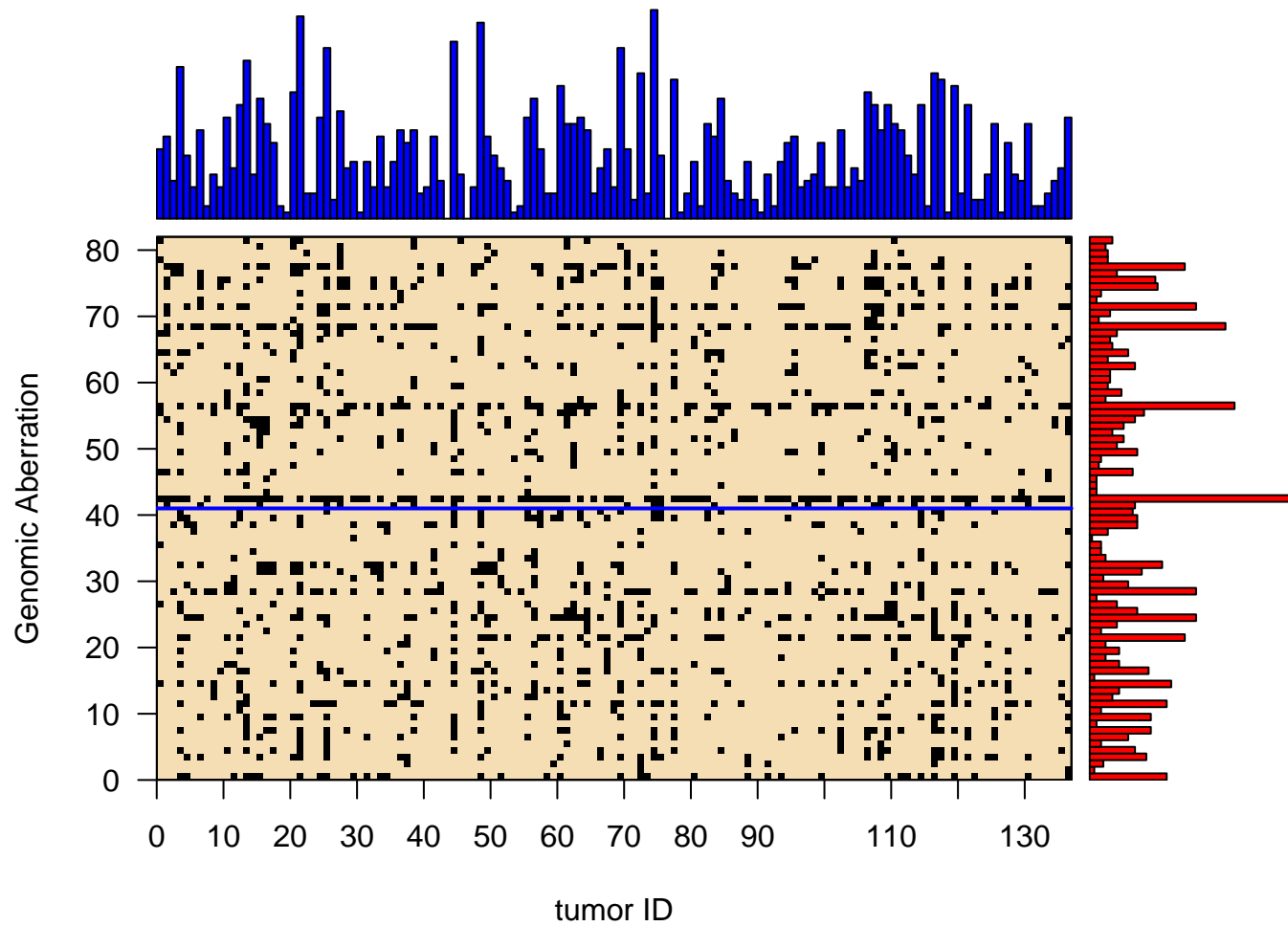
Michael Newton, Hyuna Yang, and David Hastie

IMA, September 2003

COMPARATIVE GENOMIC HYBRIDIZATION

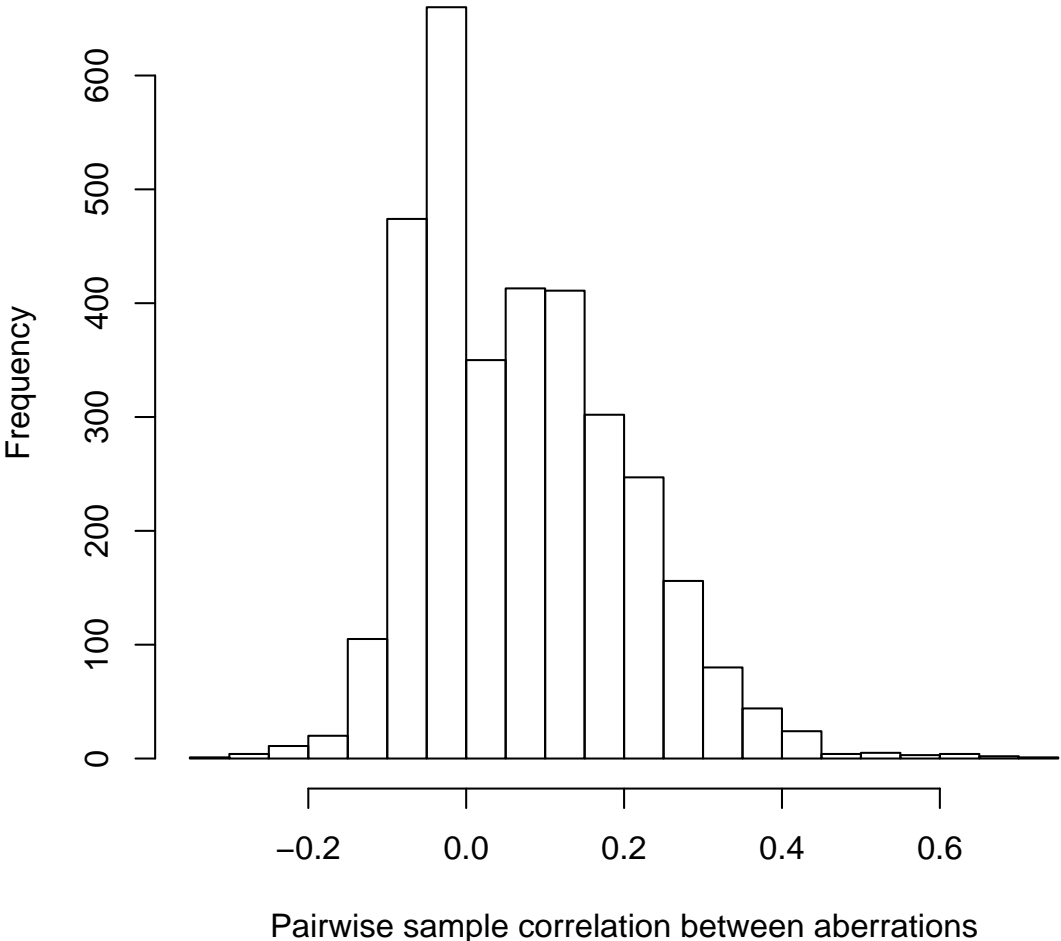
- Kallioniemi, et al., Science, 1992.
- Pinkel, et al., Nature Genetics, 1998.
- Genome-wide DNA copy number variations in tumors
- Competitive hybridization to immobilized DNA of
 1. fluorochrome labeled tumor genomic DNA
 2. differently labeled normal DNA
- Image processing to score chromosomal imbalances

CGH Aberrations in 137 Invasive Ductal Breast Cancers



From data collected by I. Tomlinson and R. Roylance, Cancer Research UK

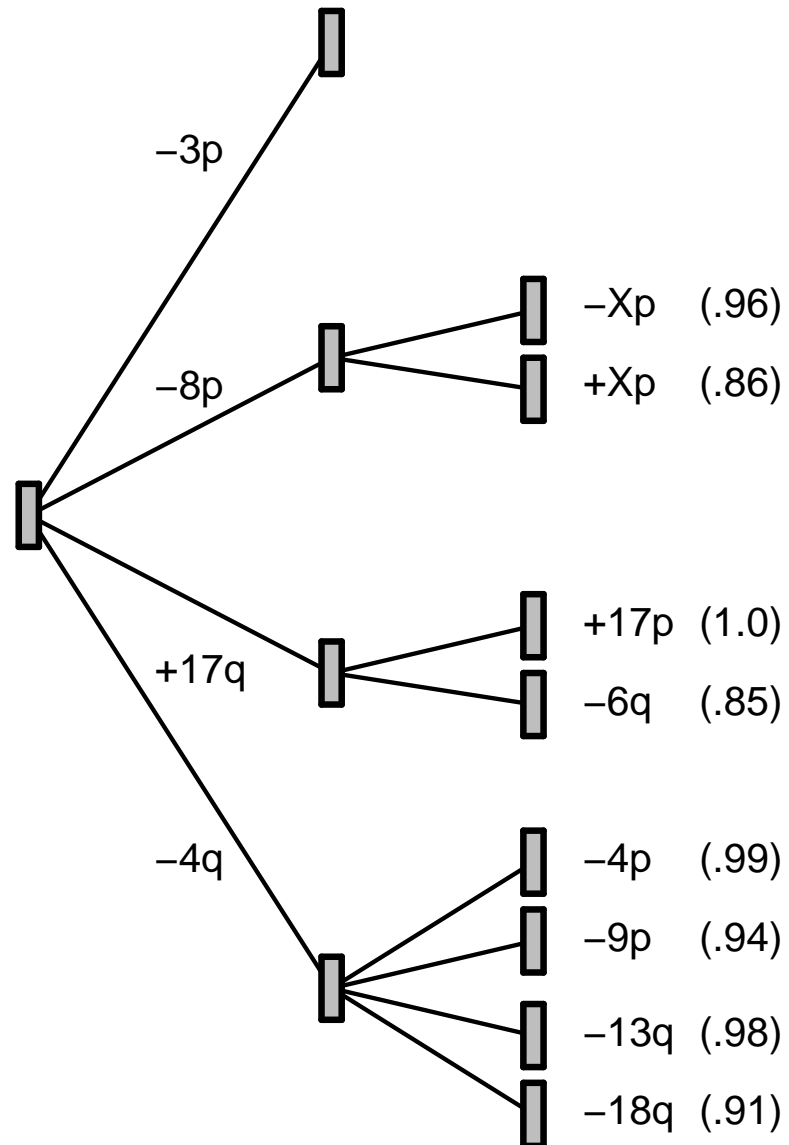
Sample Correlations Are Not Consistent With Independence



Statistical Questions

- test, quantify marginal heterogeneity; identify frequent aberrations; identify neutral aberrations
- test, quantify dependence; identify relevant combinations
- cluster aberrations/tumors
- derive biomarkers

Tree MCMC, RCC data: Estimated Tree



Best $\mathcal{C} \in \mathcal{G}$

$$\{+17q, +Xp, -6q\}$$

$$\{-4q, -8p, -9p\}$$

$$\{-4q, -9p, -18q\}$$

$$\{-8p, -13q, -Xp\}$$

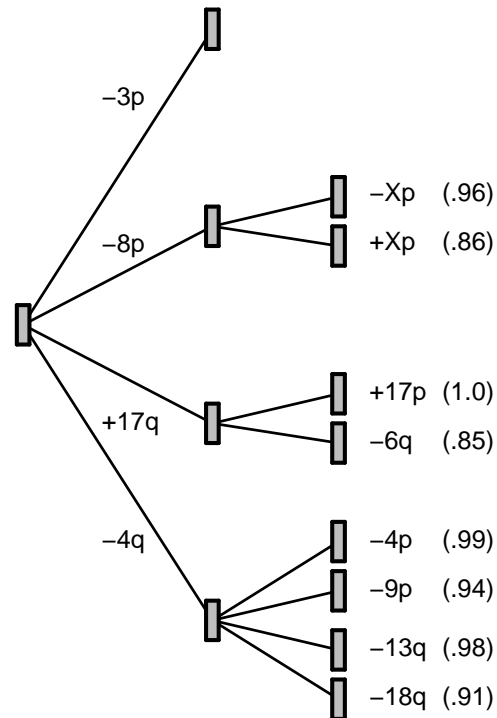
$$\{+17p, +17q\}$$

$$\{+Xp, -8p\}$$

$$\{-4q, -6q\}$$

$$\{-4q, -13q\}$$

$$\{-3p\}$$

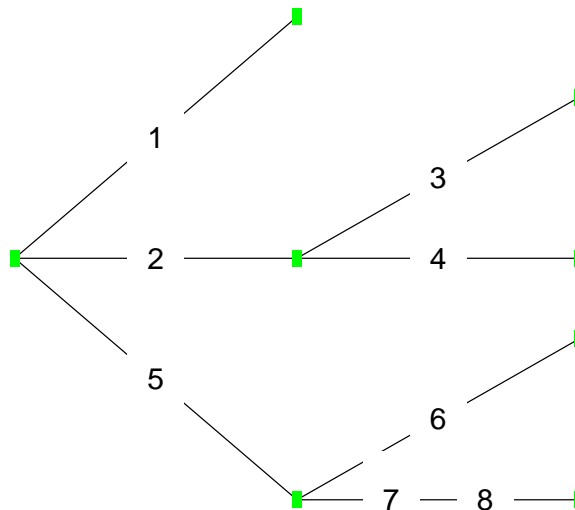


Network Parameter Space: $\mathcal{P} \subset \mathcal{T} \subset \mathcal{G}$

\mathcal{P} . Non-overlapping: $\mathcal{C} = \{C_1, \dots, C_K\}$ is a sub-partition.

\mathcal{G} . General: anything, except no j, k with $C_j \subset C_k$.

\mathcal{T} . Tree-like: Edges e_1, \dots, e_J form a sub-partition. Each edge e has a parent edge $PA(e)$ that is the root or another edge. No loops. Internal node order ≥ 3 . $C_k = \bigcup_{j \in \text{path}_k} e_j$



Instability: $Z_i \sim_{\text{iid}} \text{Bernoulli}(\theta)$ for aberrations $i \in \{1, 2, \dots, n\}$.

Selection:

- Network $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$ of ensembles $C_k \subset \{1, 2, \dots, n\}$
- Open ensemble: $A_k = \bigcap_{i \in C_k} [Z_i = 1]$
- Selection: $\text{SEL} = \bigcup_{k=1}^K A_k$

Measurement Error:

$$[X_i | \mathbf{Z}] \sim \begin{cases} \text{Bernoulli}(1 - \delta) & \text{if } Z_i = 1 \\ \text{Bernoulli}(\gamma) & \text{if } Z_i = 0 \end{cases}$$

IS joint pmf:

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} | \text{SEL})$$

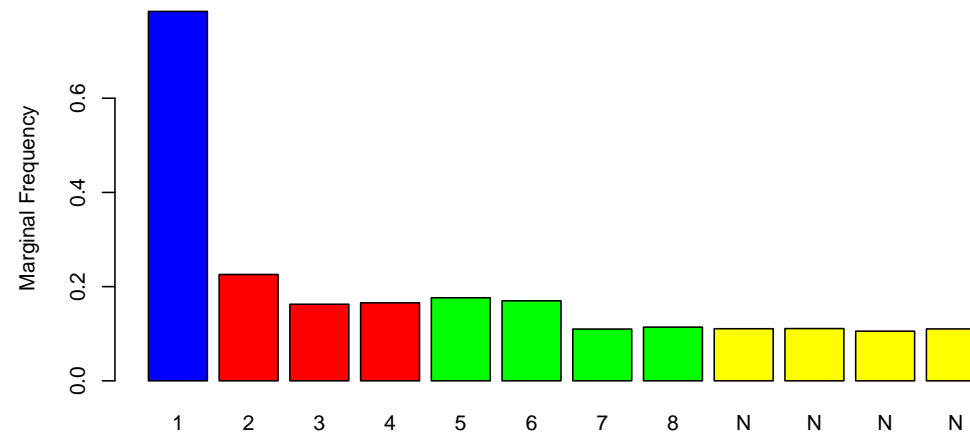
Instability-Selection Model

- snapshot of a dynamic biological system [tumor growth]
- aberrations emerge randomly [genetic instability]
- beneficial aberrations survive [cell-level selection]
- multiple steps, multiple paths

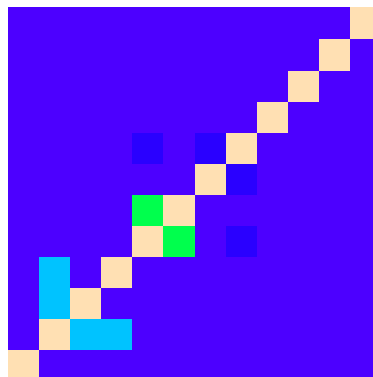
IS Sampling Properties (simulation)

e.g., $n = 12$, $\mathcal{C} = \{\{1\}, \{2, 3\}, \{2, 4\}, \{5, 6\}, \{5, 7, 8\}\}$

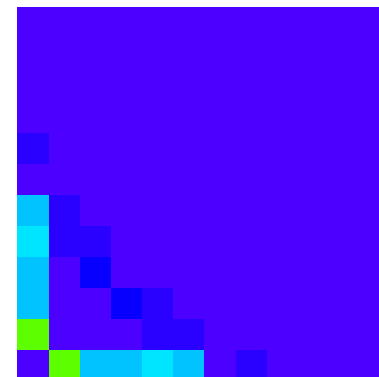
$\theta = 0.1$, $\gamma = \delta = 0.01$



+ve corr



-ve corr



IS Sampling Properties: $\mathcal{C} \in \mathcal{T}$ [Yang]

Assume $\delta + \gamma < 1$; denote $\alpha = E(X_i) = \theta(1 - \delta) + (1 - \theta)\gamma$.

- If $i \in C_k$ for some k , $E(X_i|\text{SEL}) > \alpha$.
- If $i, j \in e$ for some e , $E(X_i|\text{SEL}) = E(X_j|\text{SEL})$ and $\text{cov}(X_i, X_j|\text{SEL}) \geq 0$
- If $j \in e$ and $i \in \text{PA}(e)$ for some e , $E(X_i|\text{SEL}) > E(X_j|\text{SEL})$ and $\text{cov}(X_i, X_j|\text{SEL}) > 0$
- If $i \in e_1$, $j \in e_2$ and e_1, e_2 are edges in different ensembles, $\text{cov}(X_i, X_j|\text{SEL}) < 0$.
- \mathcal{C} is identifiable

IS Sampling Properties: $\mathcal{C} \in \mathcal{G}$ [Hastie]

Again, assume $\delta + \gamma < 1$. Given $\mathcal{C} = \{C_1, \dots, C_K\}$. Let i and j denote two aberrations in $\{1, 2, \dots, n\}$.

- If j never occurs in an ensemble without i , then $E(X_i|\text{SEL}) \geq E(X_j|\text{SEL})$ and $\text{cov}(X_i, X_j|\text{SEL}) \geq 0$.
- If i and j both occur in at least one ensemble but never occur in the same ensemble, then $\text{cov}(X_i, X_j|\text{SEL}) < 0$.
- Suppose that every ensemble contains either i or j or both, and both i and j occur in at least one ensemble without the other. Then $\text{cov}(X_i, X_j|\text{SEL}) < 0$.
- ... ?identifiable? ...

Likelihood

Multiply across tumors.

With data $\mathbf{x} = (x_1, x_2, \dots, x_n)$ from one tumor,

$$P(\mathbf{X} = \mathbf{x} | \text{SEL}) = \underbrace{P(\mathbf{X} = \mathbf{x})}_{\text{easy}} \underbrace{P(\text{SEL} | \mathbf{X} = \mathbf{x}) / P(\text{SEL})}_{\text{harder}}$$

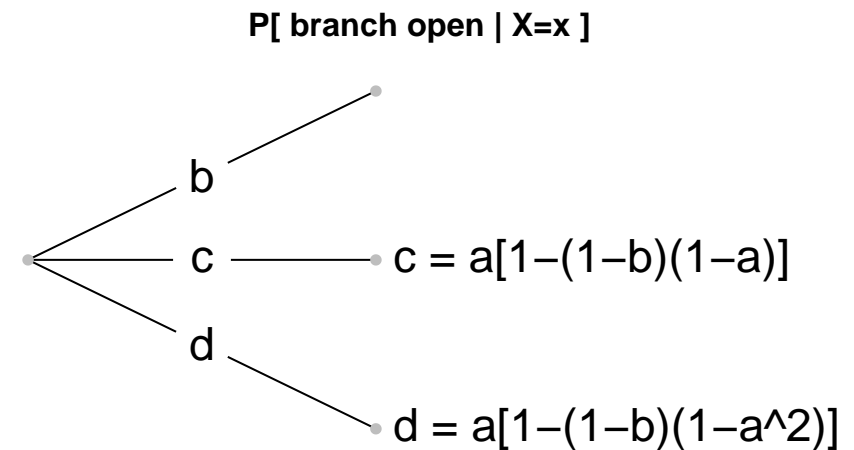
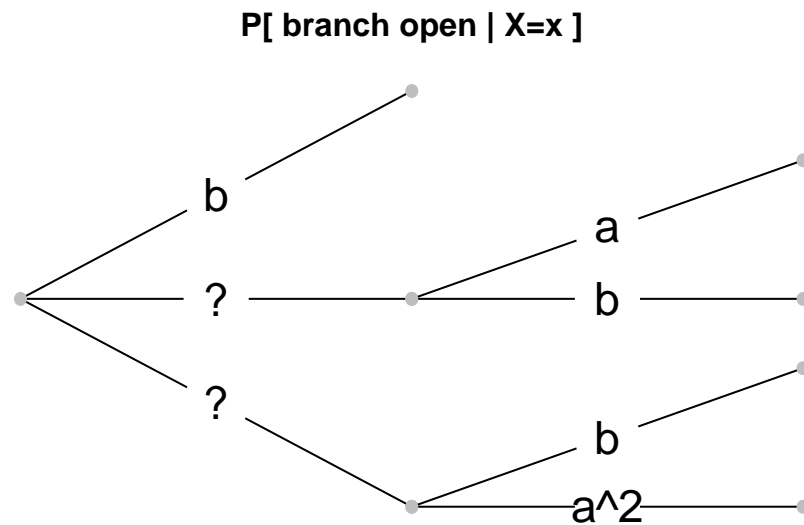
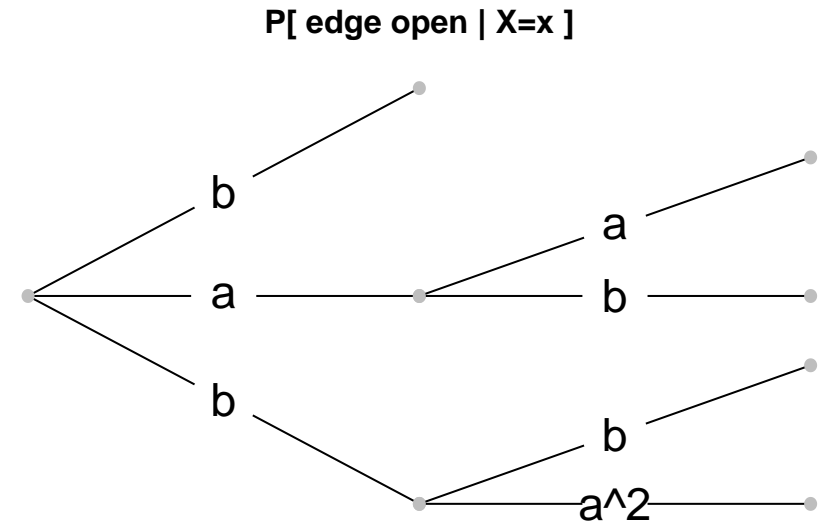
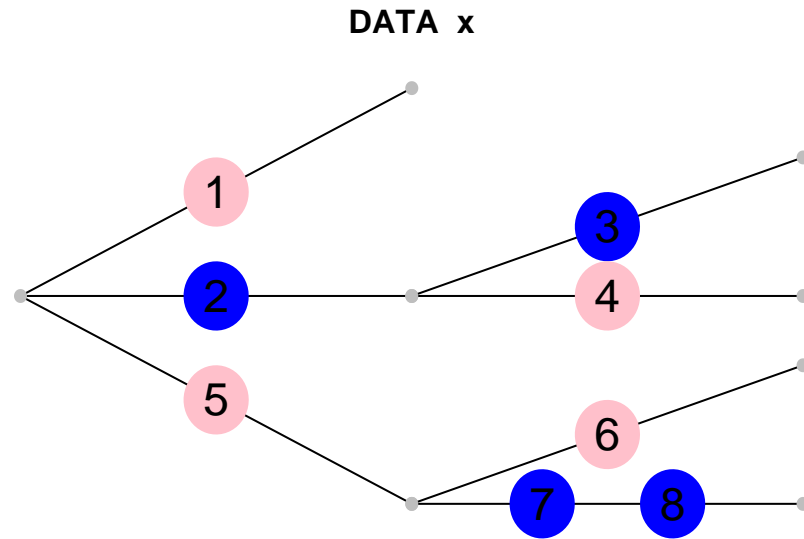
$$P(\mathbf{X} = \mathbf{x}) = \alpha^{\sum_i x_i} (1 - \alpha)^{\sum_i (1 - x_i)}$$

$$\text{Recall } \text{SEL} = \bigcup_{k=1}^K \bigcap_{i \in C_k} [Z_i = 1].$$

$\mathcal{C} \in \mathcal{T}$	$P(\text{SEL} \mathbf{X} = \mathbf{x})$						$P(\text{SEL})$		
	1,0,0	0,1,0	0,0,1	1,1,0	1,0,1	0,1,1	1,1,1	0,0,0	
1	a	b	b	a	a	b	a	b	θ
2	b	a	b	a	b	a	a	b	θ
3	b	b	a	b	a	a	a	b	θ
1,2	ab	ab	b^2	a^2	ab	ab	a^2	b^2	θ^2
2,3	b^2	ab	ab	ab	ab	a^2	a^2	b^2	θ^2
1,3	ab	b^2	ab	ab	a^2	ab	a^2	b^2	θ^2
1,2,3	ab^2	ab^2	ab^2	a^2b	a^2b	a^2b	a^3	b^3	θ^3
1 2	$a + b - ab$	$a + b - ab$	$2b - b^2$	$2a - a^2$	$a + b - ab$	$a + b - ab$	$2a - a^2$	$2b - b^2$	$2\theta - \theta^2$
2 3	$2b - b^2$	$a + b - ab$	$a + b - ab$	$a + b - ab$	$a + b - ab$	$2a - a^2$	$2a - a^2$	$2b - b^2$	$2\theta - \theta^2$
1 3	$a + b - ab$	$2b - b^2$	$a + b - ab$	$a + b - ab$	$2a - a^2$	$a + b - ab$	$2a - a^2$	$2b - b^2$	$2\theta - \theta^2$
1 2,3	$a + b^2 - ab^2$	$b + ab - ab^2$	$b + ab - ab^2$	$a + ab - a^2b$	$a + ab - a^2b$	$b + a^2 - a^2b$	$a + a^2 - a^3$	$b + b^2 - b^3$	$\theta + \theta^2 - \theta^3$
2 1,3	$b + ab - ab^2$	$a + b^2 - ab^2$	$b + ab - ab^2$	$a + ab - a^2b$	$b + a^2 - a^2b$	$a + ab - a^2b$	$a + a^2 - a^3$	$b + b^2 - b^3$	$\theta + \theta^2 - \theta^3$
3 1,2	$b + ab - ab^2$	$b + ab - ab^2$	$a + b^2 - ab^2$	$b + a^2 - a^2b$	$a + ab - a^2b$	$a + ab - a^2b$	$a + a^2 - a^3$	$b + b^2 - b^3$	$\theta + \theta^2 - \theta^3$
1,2 2,3	$b(a + b - ab)$	$a(2b - b^2)$	$b(a + b - ab)$	$a(a + b - ab)$	$b(2a - a^2)$	$a(a + b - ab)$	$2a^2 - a^3$	$2b^2 - b^3$	$2\theta^2 - \theta^3$
1,2 1,3	$a(2b - b^2)$	$b(a + b - ab)$	$b(a + b - ab)$	$a(a + b - ab)$	$a(a + b - ab)$	$b(2a - a^2)$	$2a^2 - a^3$	$2b^2 - b^3$	$2\theta^2 - \theta^3$
1,3 2,3	$b(a + b - ab)$	$b(a + b - ab)$	$a(2b - b^2)$	$b(2a - a^2)$	$a(a + b - ab)$	$a(a + b - ab)$	$2a^2 - a^3$	$2b^2 - b^3$	$2\theta^2 - \theta^3$
1 2 3	$1 - (1 - a)(1 - b)^2$			$1 - (1 - a)^2(1 - b)$			$1 - (1 - a)^3$	$1 - (1 - b)^3$	$1 - (1 - \theta)^3$
$P(x)$	$\alpha(1 - \alpha)^2$	$\alpha(1 - \alpha)^2$	$\alpha(1 - \alpha)^2$	$\alpha^2(1 - \alpha)$	$\alpha^2(1 - \alpha)$	$\alpha^2(1 - \alpha)$	α^3	$(1 - \alpha)^3$	

$$a = P(Z_i = 1|X_i = 1) = \frac{\theta(1-\delta)}{\theta(1-\delta)+(1-\theta)\gamma} \text{ and } b = P(Z_i = 1|X_i = 0) = \frac{\theta\delta}{\theta\delta+(1-\theta)(1-\gamma)}.$$

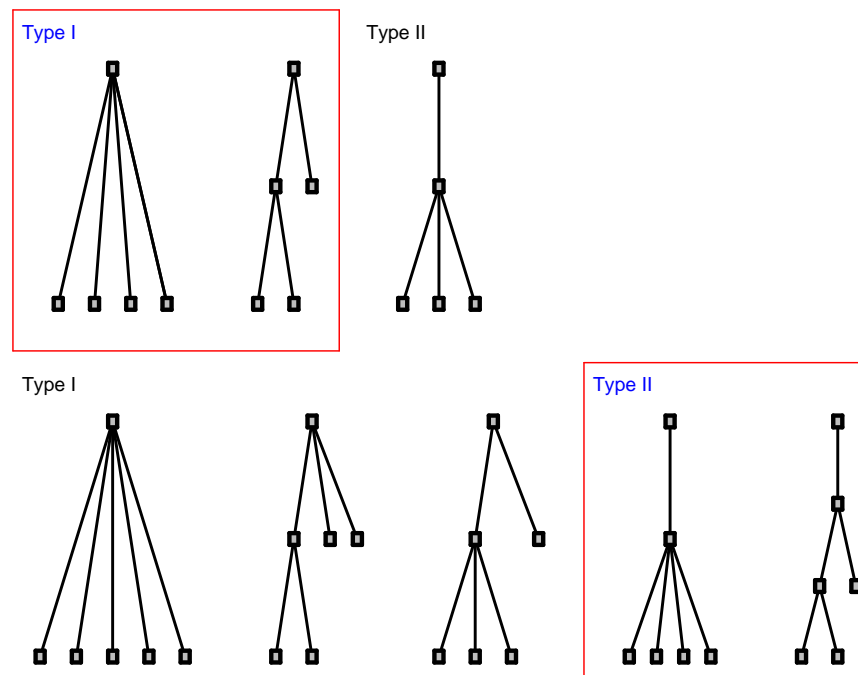
Likelihood Evaluation: $\mathcal{C} \in \mathcal{T}$



$$P(\text{SEL}|X=x) = 1 - (1-b)(1-c)(1-d)$$

Prior: $\mathcal{C} \in \mathcal{T}$

- Uniform on trees, conditional on set of edges.

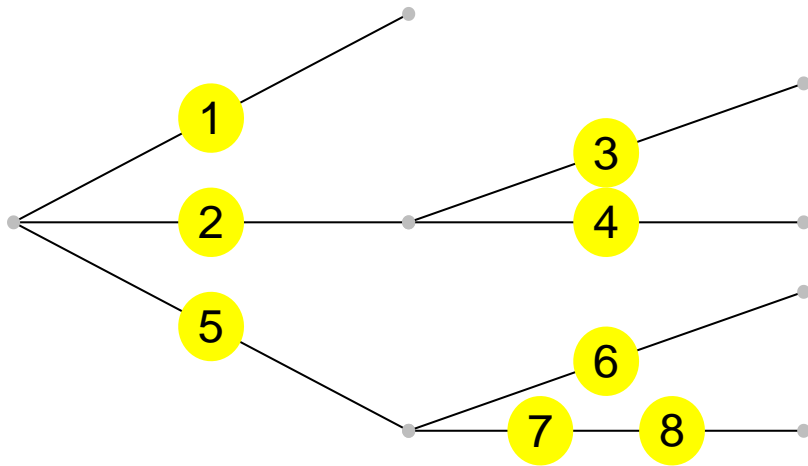


# distinct edges	1	2	3	4	5	6	7	...
# trees	1	1	4	17	116	997	10270	...

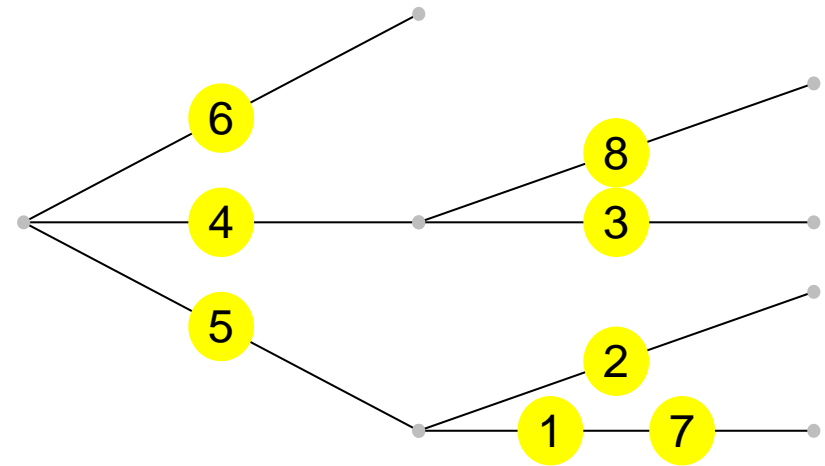
- Polya prior to create edges (partition, neutrality).

MCMC: $\mathcal{C} \in \mathcal{T}$

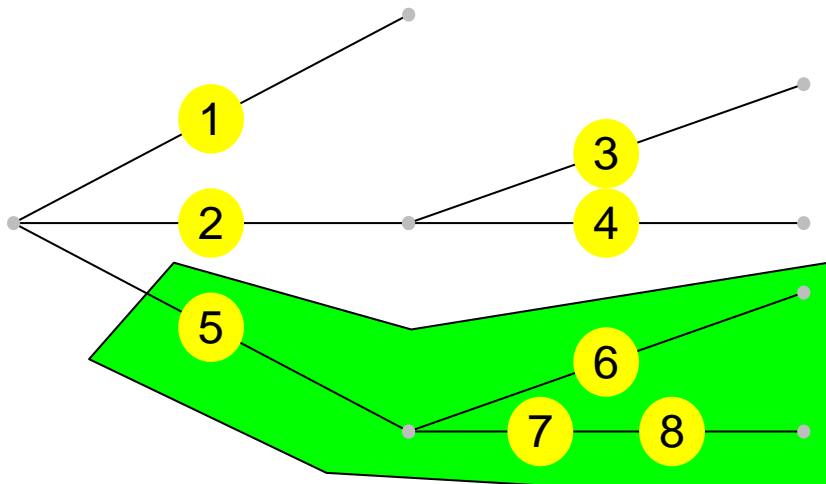
Current State



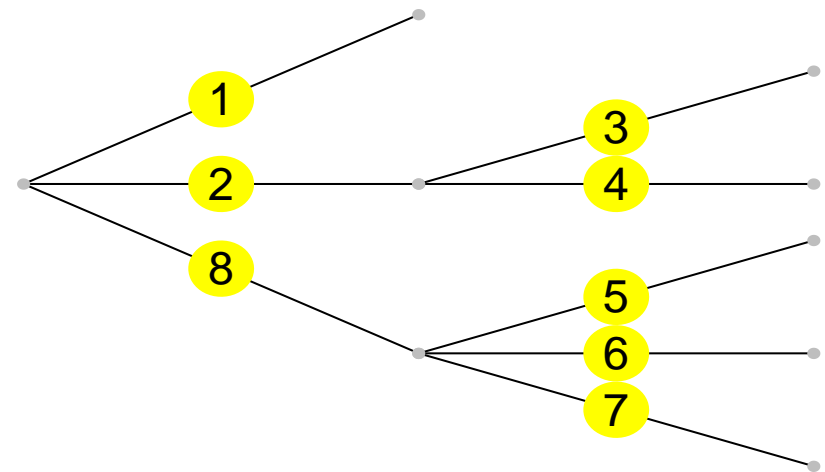
Proposed State by Permutation



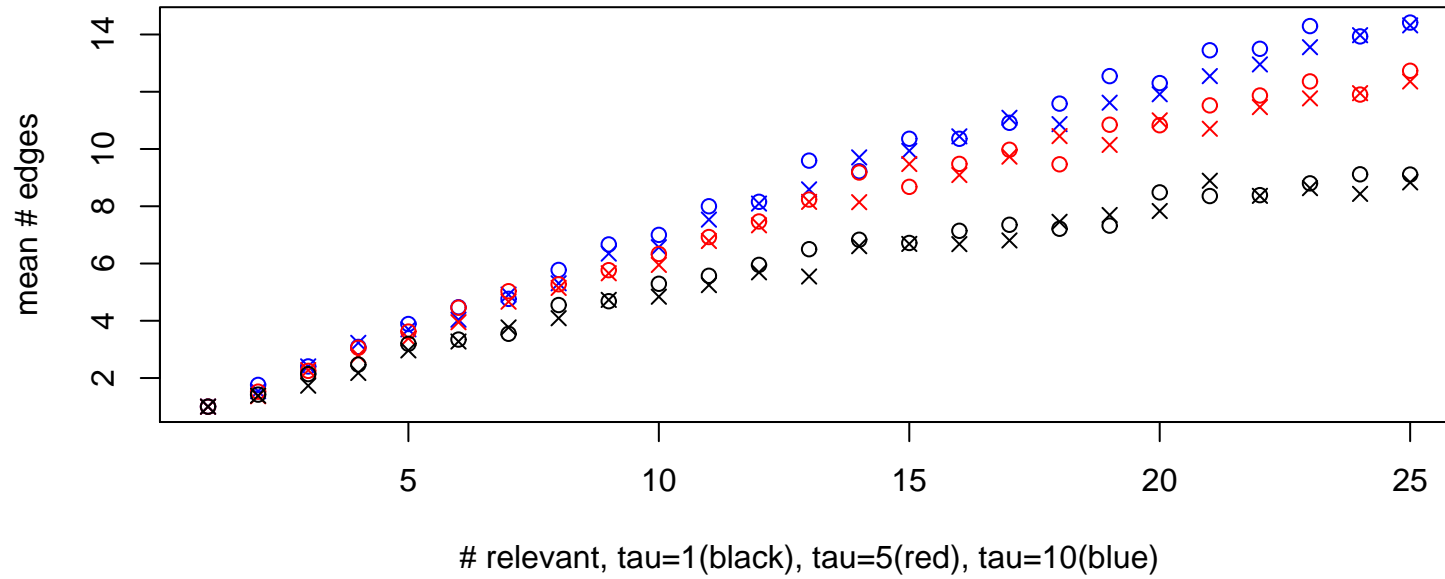
Select Branch



Dissolve, Reform Edges and Branch

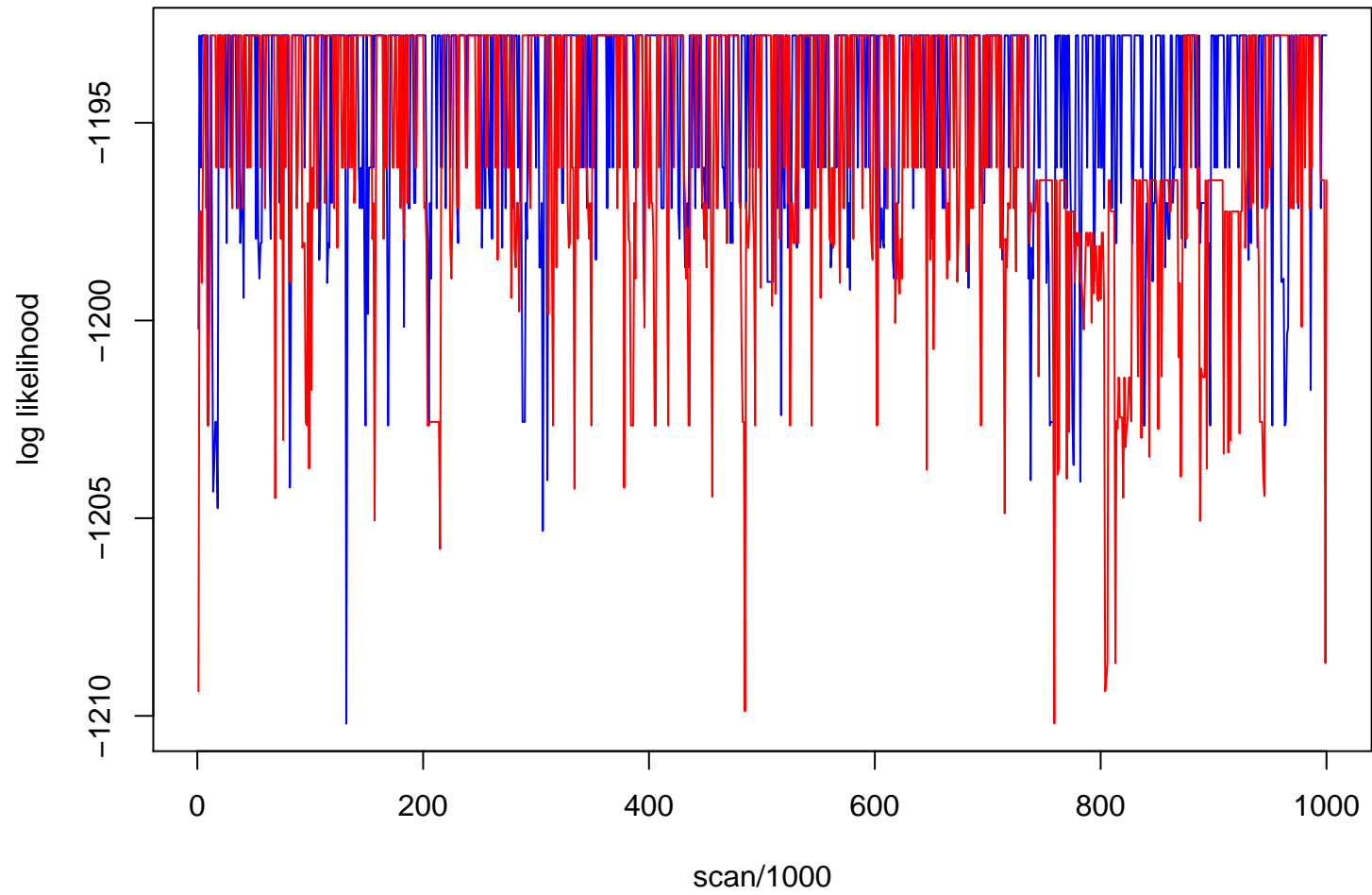


Code check



Tree MCMC, RCC data, prefiltered

Trace plot: log likelihood



Likelihood Evaluation: $\mathcal{C} \in \mathcal{G}$

Inclusion-Exclusion for $P(\text{SEL}) = P(\bigcup_{k=1}^K A_k)$:

$$\sum_{k=1}^K P(A_k) - \sum_{k \neq l} P(A_k \cap A_l) + \dots + (-1)^{K-1} P(A_1 \cap \dots \cap A_K)$$

An efficient algorithm (Thanks J. Kadane.)

Aberration	Ensemble					v^*
	v_1	v_2	v_3	v_4	v_5	
1	1	0	0	0	0	0
2	0	1	1	0	0	1
3	0	1	0	0	0	1
4	0	0	1	0	0	1
5	0	0	0	1	1	1
6	0	0	0	1	0	1
7	0	0	0	0	1	0
8	0	0	0	0	1	0
neutrals	0	0	0	0	0	0

E.g., $P(A_2 \cap A_3 \cap A_4)$

$= P\left(\bigcap_{i:(v_i^*=1)} [Z_i = 1]\right)$

$= \prod_{i:(v_i^*=1)} P[Z_i = 1] = \theta^5$

Prior: $\mathcal{C} \in \mathcal{G}$

- Uniform on K (up to a max)
- Uniform on \mathcal{C} given K , or Gibbs

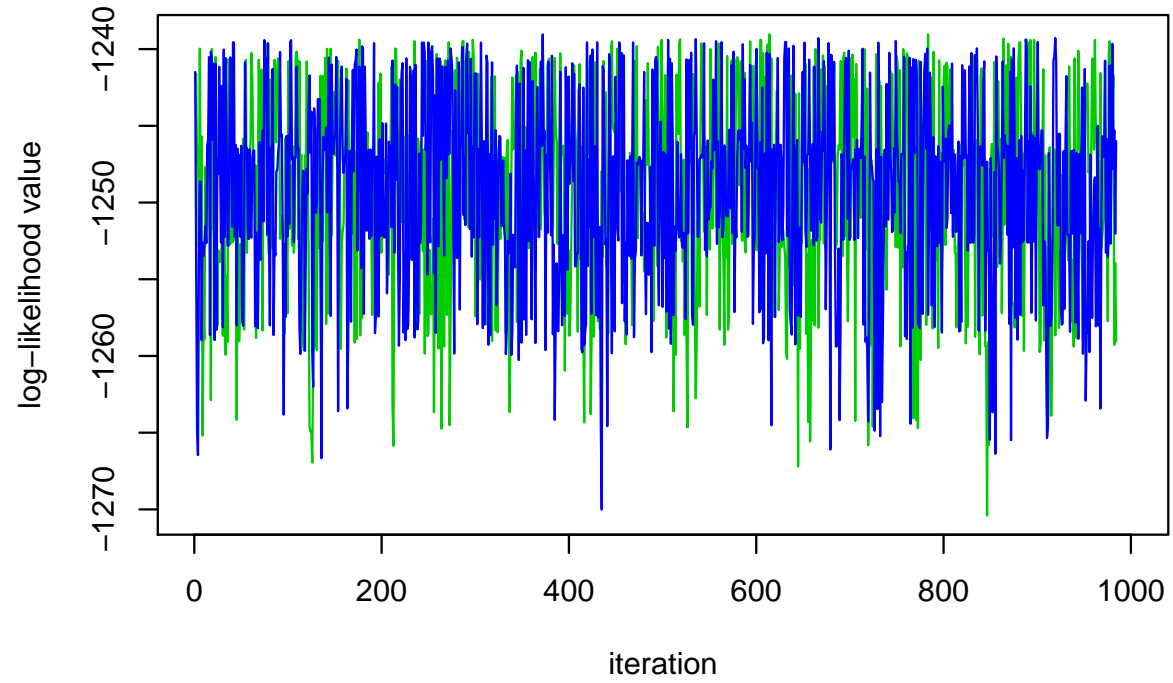
MCMC: add/drop ensembles; randomize matrix elements

	k							
n	1	2	3	4	5	6	...	10
1	1							
2	3	1						
3	7	9	2					
4	15	55	64	25	6	1		
5	31	285	1090	2020	2146	1380	...	2
6	63	1351	14000	82115	304752	759457	...	1067771
7	127	6069	153762	2401910	...			43506231489
8	255	26335	1533504	58089465	...			501425871595264
9	511	111645	14356610	...				2719674203584968630
10	1023	465751	128722000	...				9172837864705015158979
11	2047	1921029	1119607522	...				22524989249381408262409893
12	4095	7859215	9528462944	...				44328073635887914351462953684

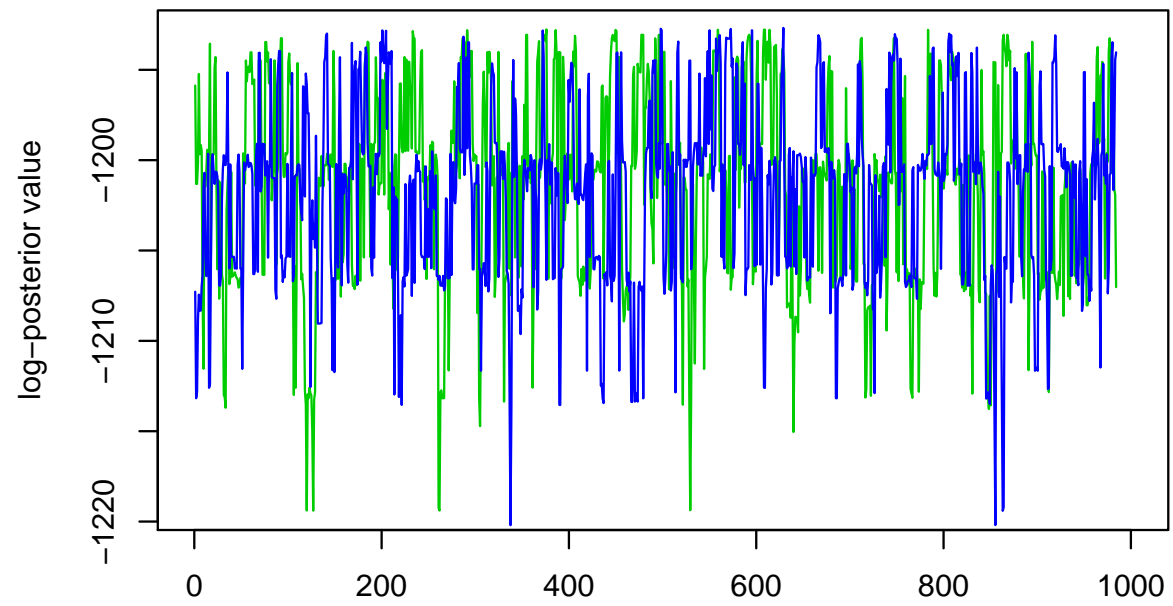
Table 1: Table of number of networks with n aberrations and k ensembles.

$\mathcal{C} \in \mathcal{G}$ MCMC, Renal Cancer Data

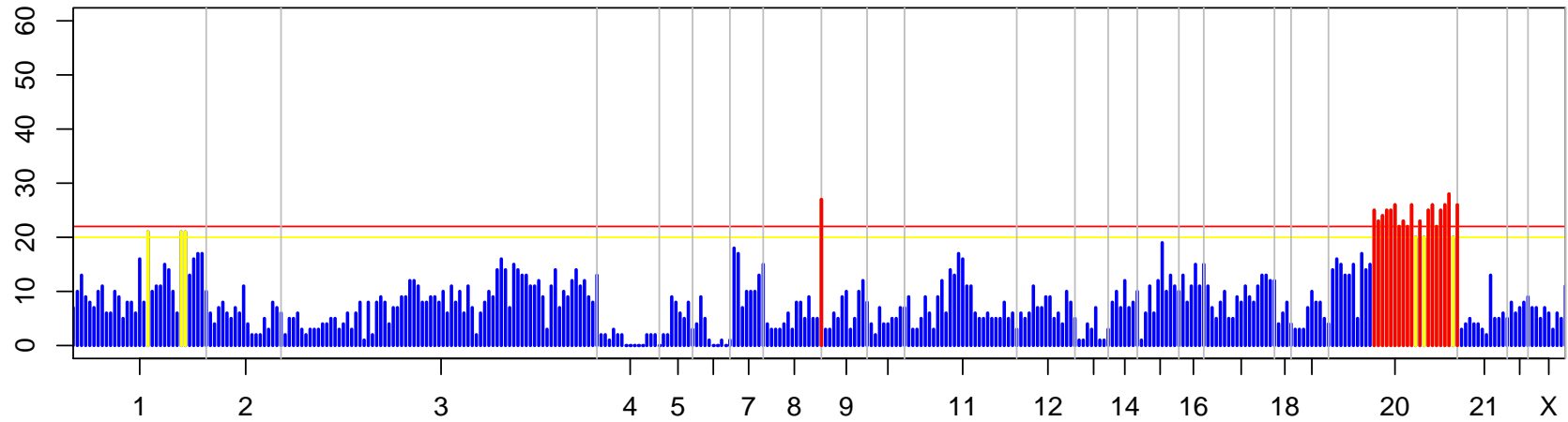
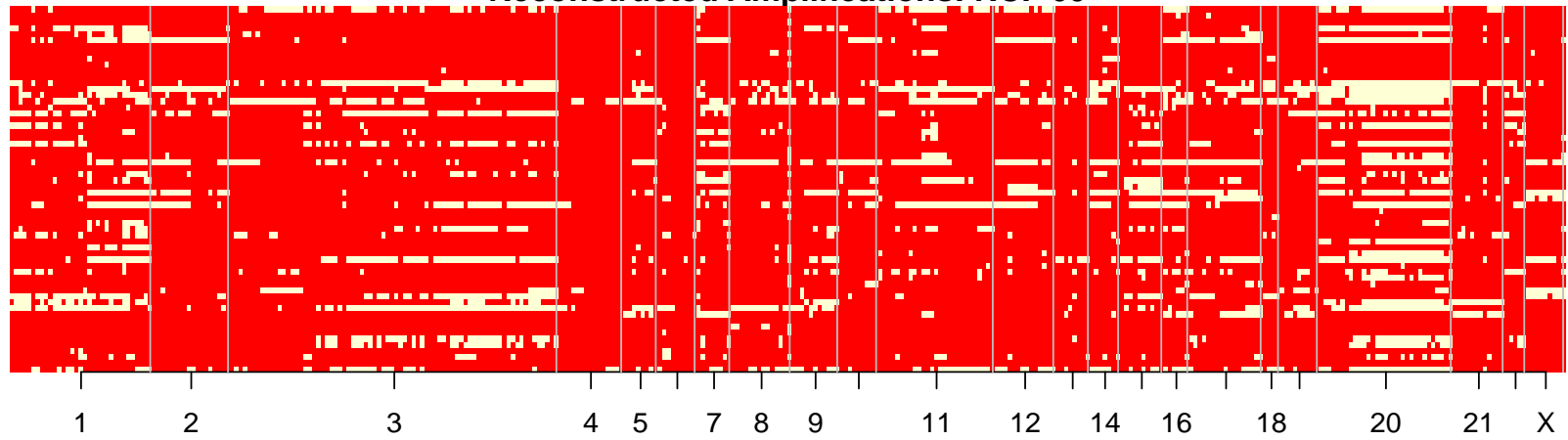
Log-likelihood trace plot



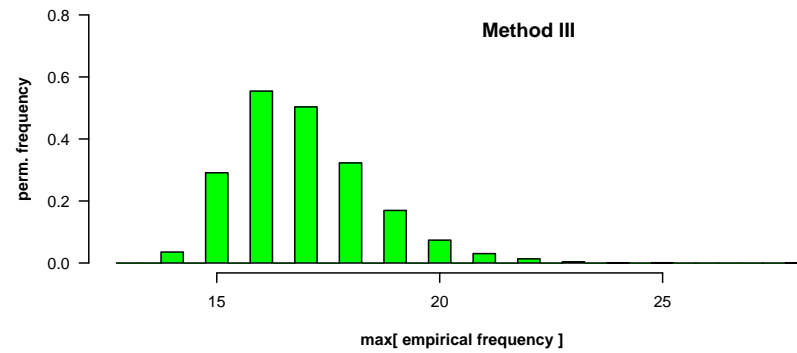
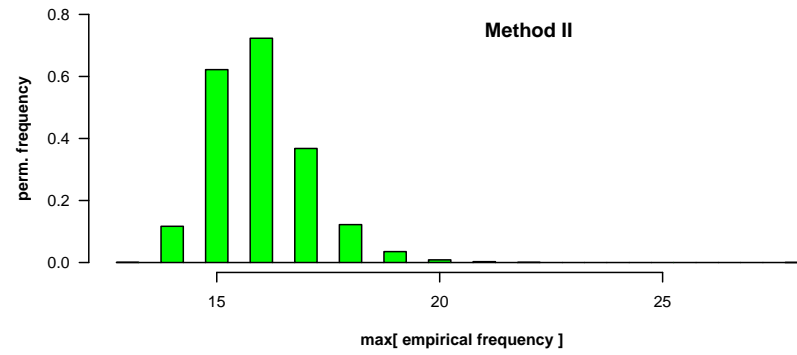
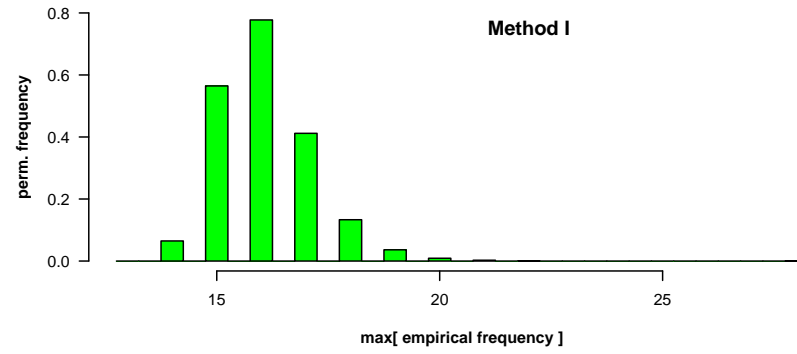
Log-posterior trace plot



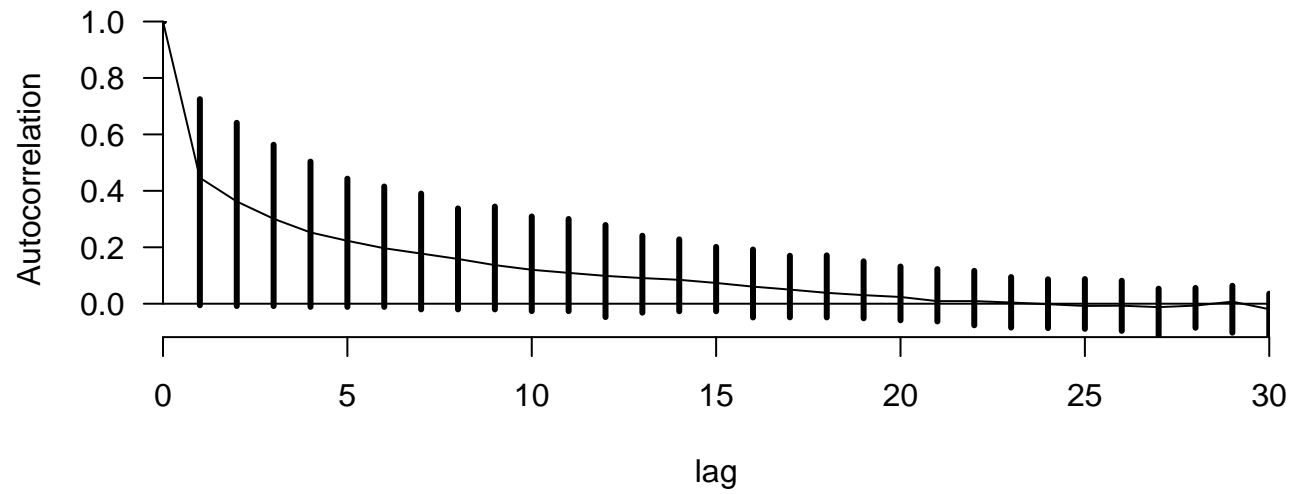
Reconstructed Amplifications: NCI-60



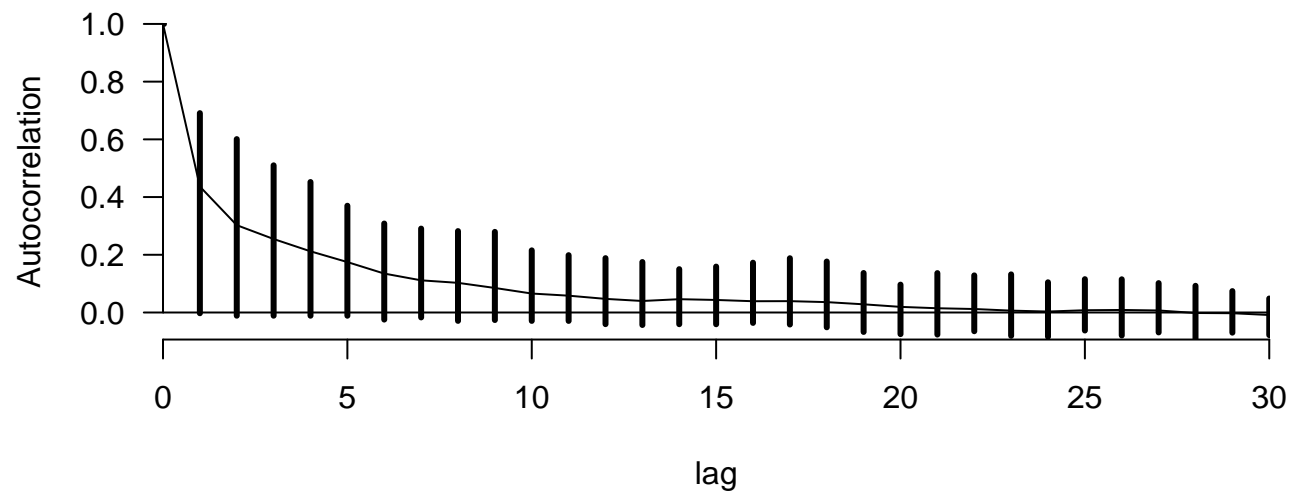
Permutation Distributions for Amplifications: NCI-60



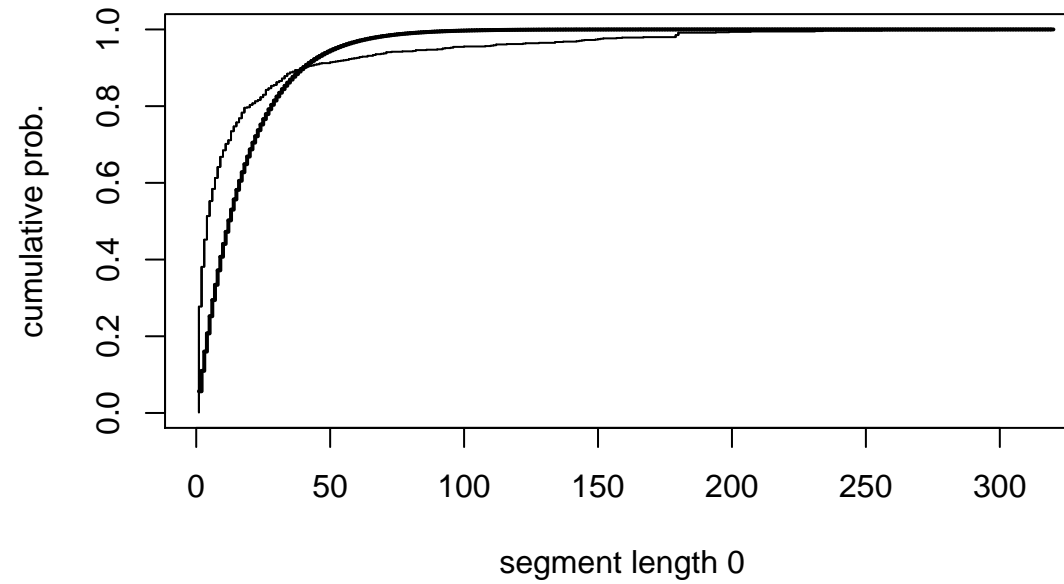
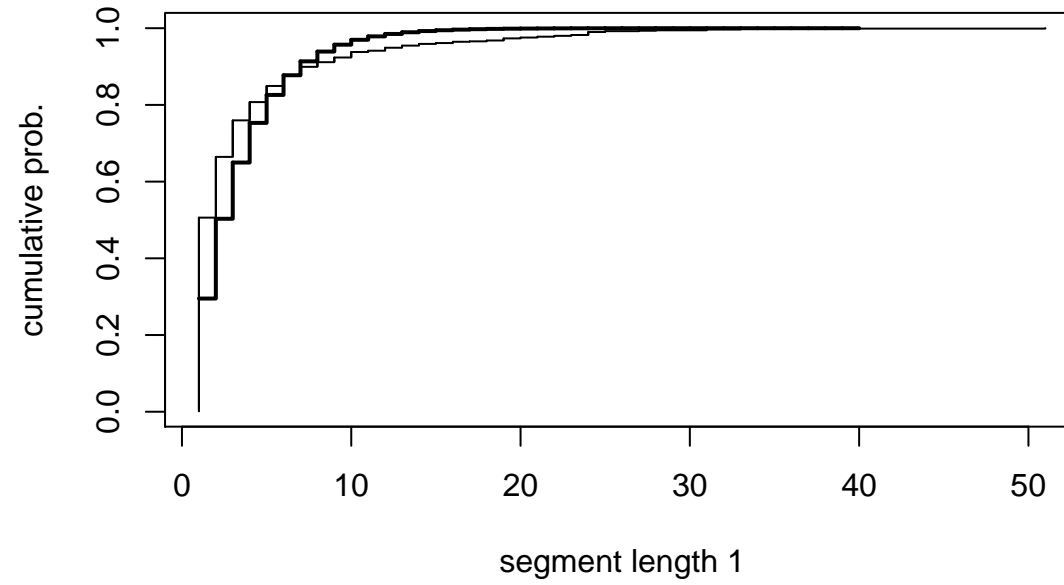
Amplifications



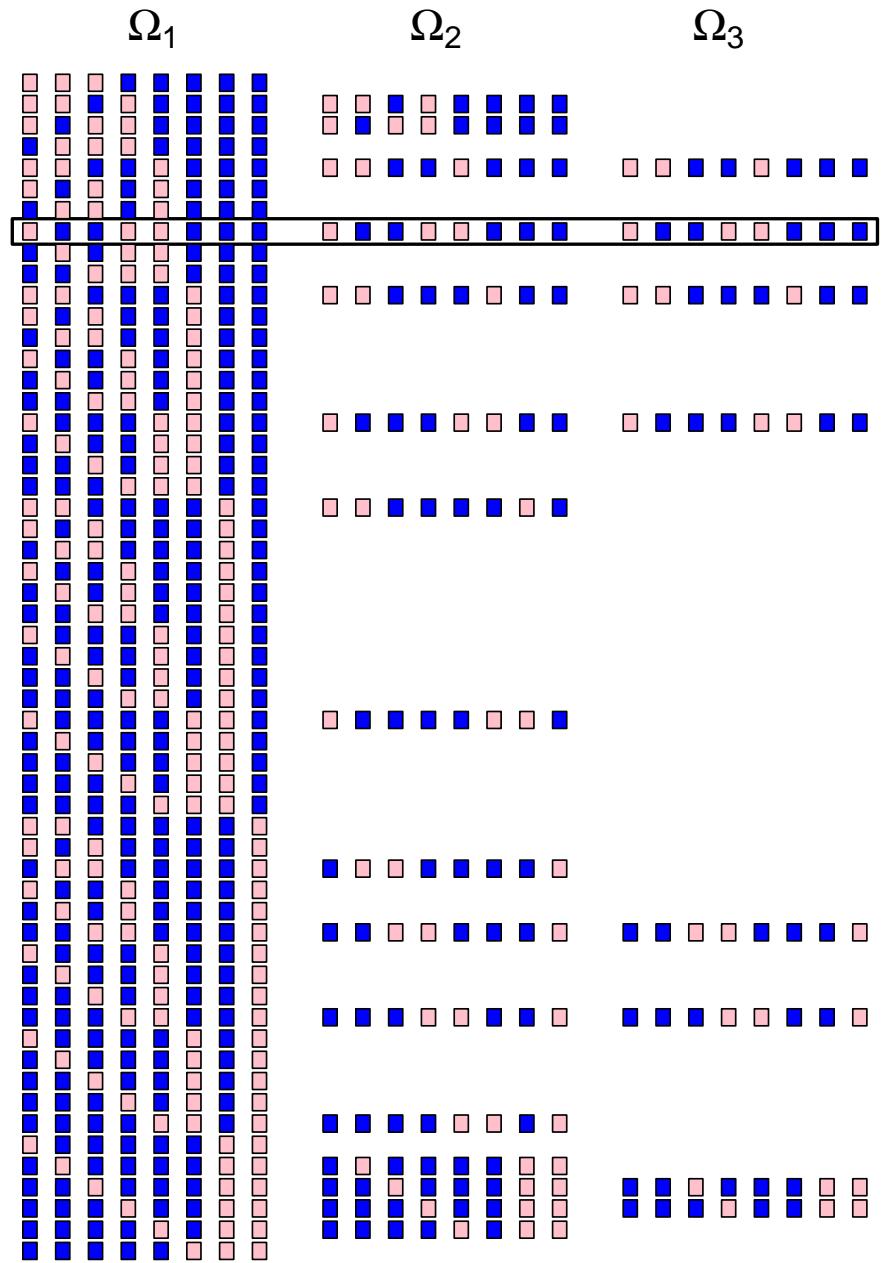
Deletions



Segment Lengths: Amplifications



Binary-sequence spaces: Each column represents the set of possible sequences that one might obtain by shuffling the observed sequence $x = (0, 1, 1, 0, 0, 1, 1, 1)$ under various restrictions. This test sequence x is listed seventh in the list Ω_1 containing all 56 binary sequences having three 0's and five 1's (the drawing code is blue for 1 and pink for 0). The list Ω_2 contains all those sequences sharing the same transition statistics as x . The list Ω_3 contains all those sequences sharing the same segment length set as x . As necessary by the theory, $\Omega_3 \subset \Omega_2 \subset \Omega_1$.



Let

$$X(g) = \# \text{ [tumors with damage at } g\text{]}$$

$$T = \max_g X(g)$$

$$S_m = \text{sufficient stat. on } H_0, \text{ model } m$$

In exact testing of: $H_0 : E[X(g)] = \text{constant}$,

$$P_m [T \geq \text{crit}(s) | S_m = s] \leq \alpha$$

and thus marginally

$$P_m [T \geq \text{crit}(S_m)] \leq \alpha.$$

Conjecture (robustness): For certain reversible, stationary m ,

$$P_m [T \geq \text{crit}(S_{\text{iid}})] \leq \alpha$$