Graphical Models for Gene Mapping

Fresh Results

Speaker: Dan Geiger
Israel Institute of Technology
Haifa, Israel

Technion, Israel

in collaboration with five Israeli Hospitals,
Microsoft Research, and colleagues

Goals of our Research

- b Explaining biological functions underlying important diseases.
- h Supporting better diagnostic and medical treatments.
- a Developing novel statistical techniques of genetics analysis.
- s Providing the genetics community with advanced analysis tools called **superlink online**.
- s Developping infra structure abilities for high performance computing for geneticists.

Spectrum of Statistical Techniques

Techniques	*Input*
Association studies	Random healthy and affected individuals.
Mapping by Admixture Linkage Disequilibrium	Admixed affected individuals such as African-Americans
Genetic Linkage Analysis	Healthy and affected individuals from a pedigree.

Output: LOCATION OF PREDISPOSING GENES

Admixture Mapping

Inferring Ancestries Effectively & Efficiently in Admixed Populations with Linkage Disequilibrium

In press for the Journal of Computational Biology

Technion, Israel

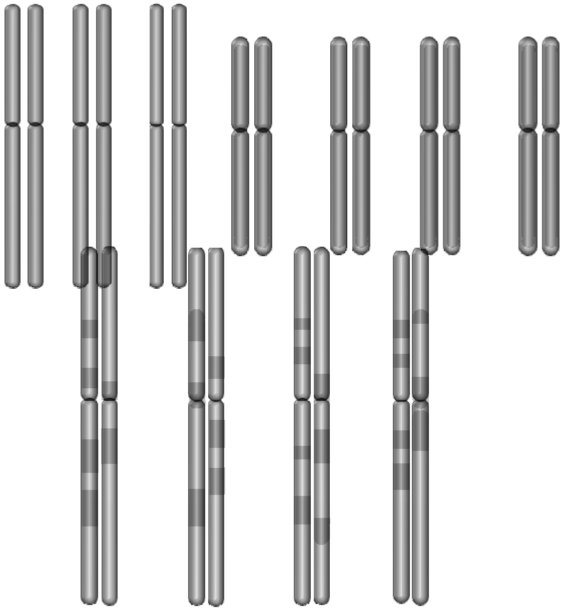
Sivan Bercovici and Dan Geiger

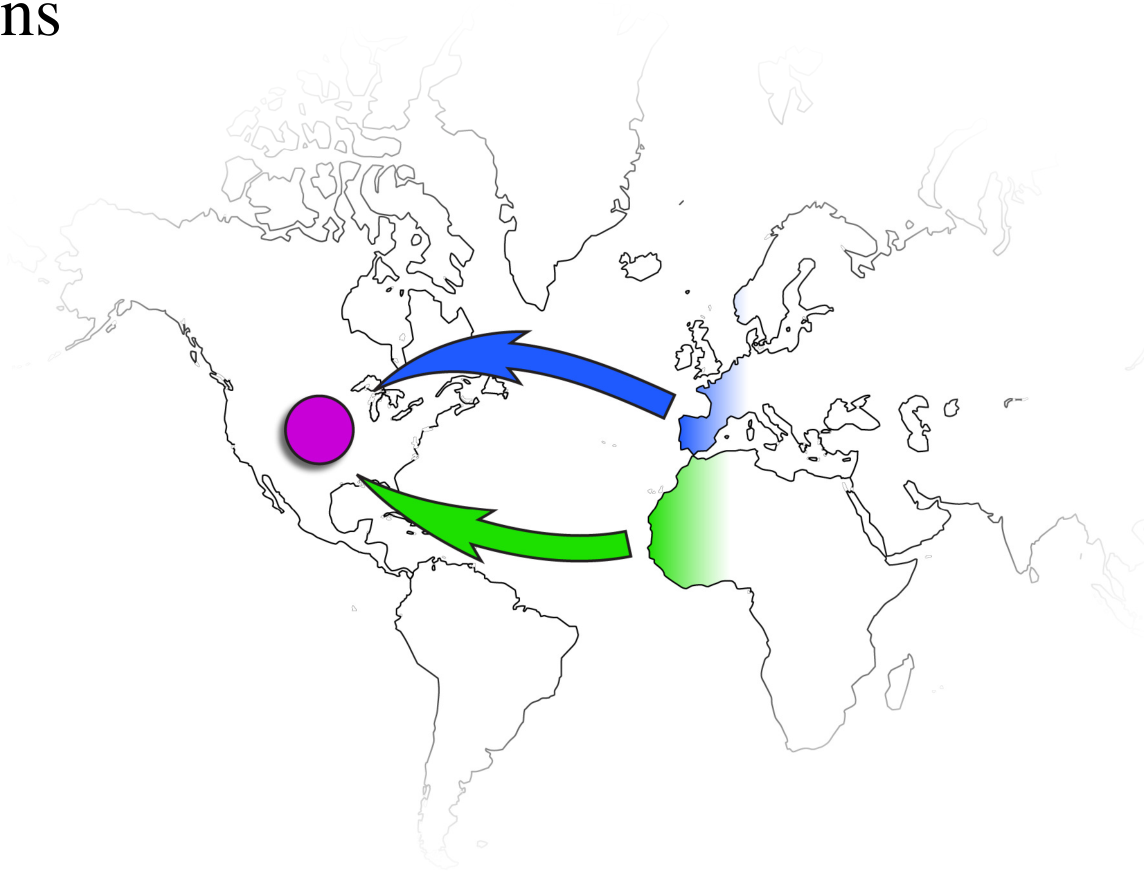
Outline

- **Admixture mapping (MALD)**
- Inference of ancestry
    - Panel construction [Genome Research, Recomb]
    - Ancestry inference [JCB, to appear]

Admixed populations

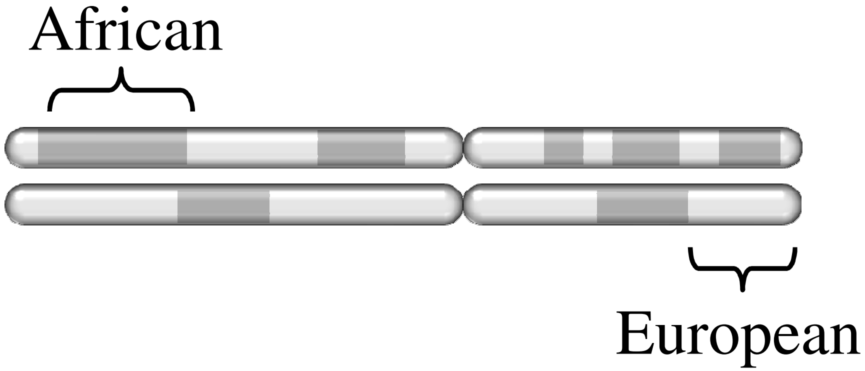
- Individuals originated from several ancestral populations
  - African Americans
  - Latinos

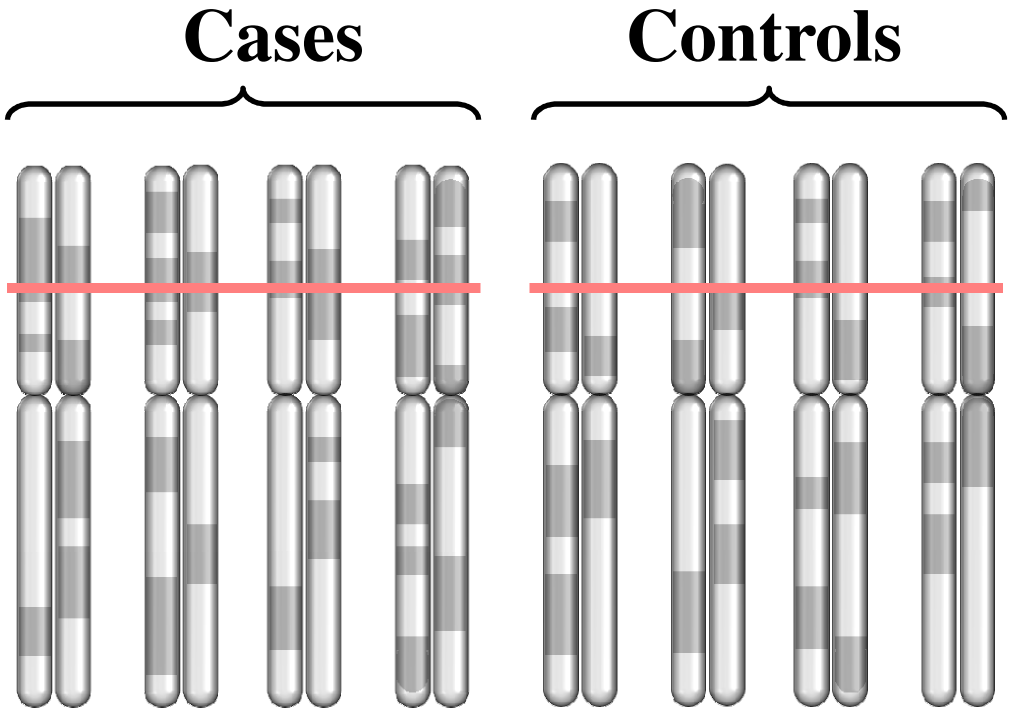




Admixed individual

<u>Admixture</u>
80%, 20%





MALD has three steps

Disease Examples

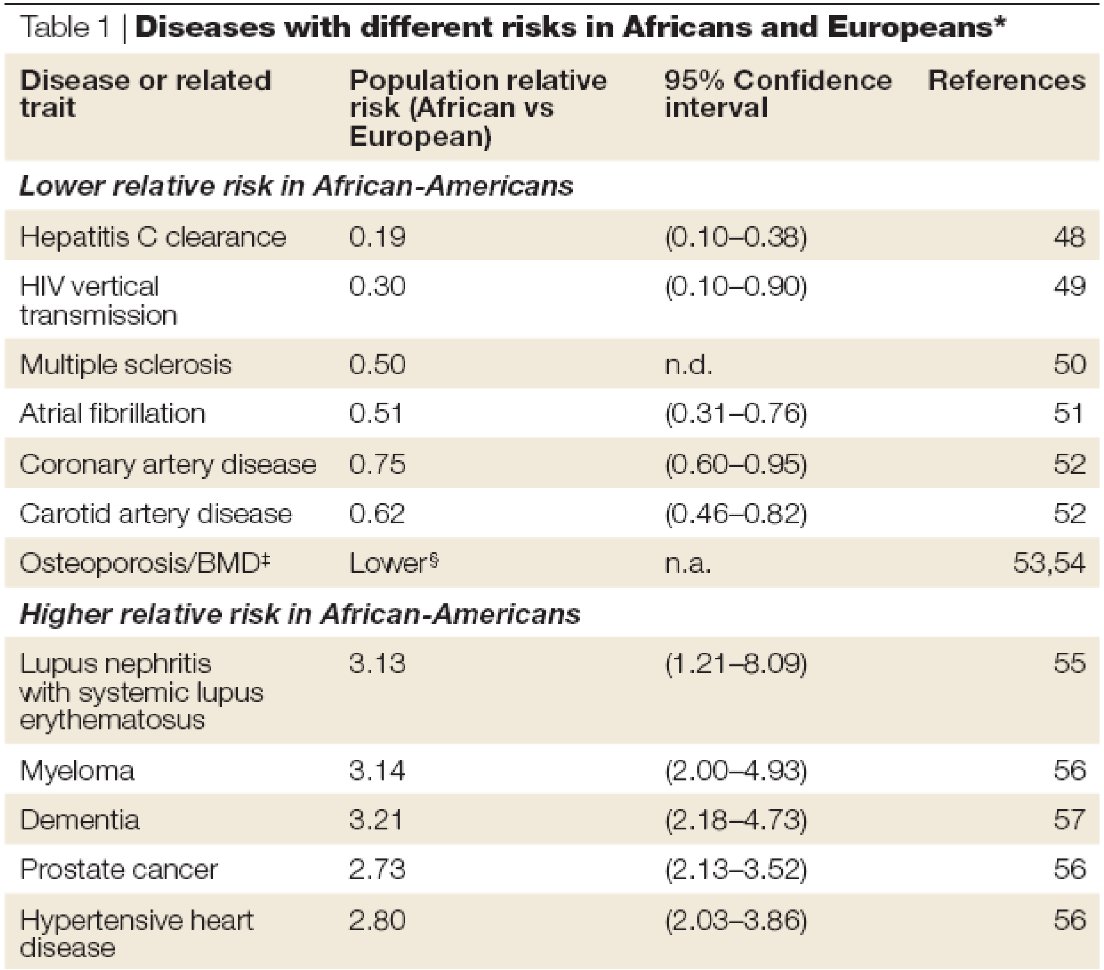


Table 1 | **Diseases with different risks in Africans and Europeans**[*]

Disease or related trait	Population relative risk (African vs European)	95% Confidence interval	References
Lower relative risk in African-Americans			
Hepatitis C clearance	0.19	(0.10–0.38)	48
HIV vertical transmission	0.30	(0.10–0.90)	49
Multiple sclerosis	0.50	n.d.	50
Atrial fibrillation	0.51	(0.31–0.76)	51
Coronary artery disease	0.75	(0.60–0.95)	52
Carotid artery disease	0.62	(0.46–0.82)	52
Osteoporosis/BMD[‡]	Lower[§]	n.a.	53,54
Higher relative risk in African-Americans			
Lupus nephritis with systemic lupus erythematosus	3.13	(1.21–8.09)	55
Myeloma	3.14	(2.00–4.93)	56
Dementia	3.21	(2.18–4.73)	57
Prostate cancer	2.73	(2.13–3.52)	56
Hypertensive heart disease	2.80	(2.03–3.86)	56

"MAPPING BY ADMIXTURE LINKAGE DISEQUILIBRIUM: ADVANCES, LIMITATIONS AND GUIDELINES" (Smith & O'Brien, *Nature Reviews Genetics,* 2005)

Disease Examples

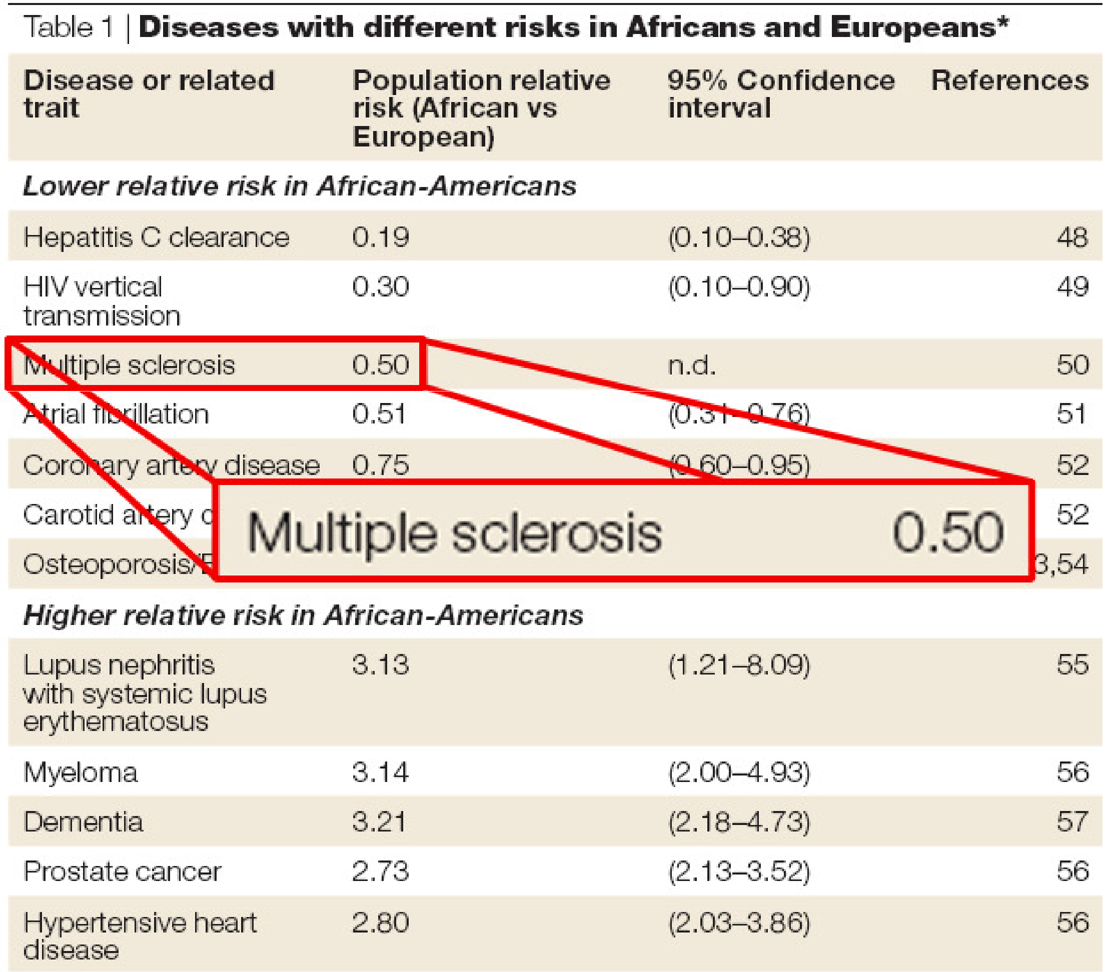


Table 1 | **Diseases with different risks in Africans and Europeans***

Disease or related trait	**Population relative risk (African vs European)**	**95% Confidence interval**	**References**
Lower relative risk in African-Americans			
Hepatitis C clearance	0.19	(0.10–0.38)	48
HIV vertical transmission	0.30	(0.10–0.90)	49
Multiple sclerosis	0.50	n.d.	50
Atrial fibrillation	0.51	(0.31–0.76)	51
Coronary artery disease	0.75	(0.60–0.95)	52
Carotid artery d	[illegible]	[illegible]	52
Osteoporosis/F	[illegible]	[illegible]	3,54
Higher relative risk in African-Americans			
Lupus nephritis with systemic lupus erythematosus	3.13	(1.21–8.09)	55
Myeloma	3.14	(2.00–4.93)	56
Dementia	3.21	(2.18–4.73)	57
Prostate cancer	2.73	(2.13–3.52)	56
Hypertensive heart disease	2.80	(2.03–3.86)	56

Multiple sclerosis 0.50

"MAPPING BY ADMIXTURE LINKAGE DISEQUILIBRIUM: ADVANCES, LIMITATIONS AND GUIDELINES" (Smith & O'Brien, *Nature Reviews Genetics*, 2005)

Disease Examples

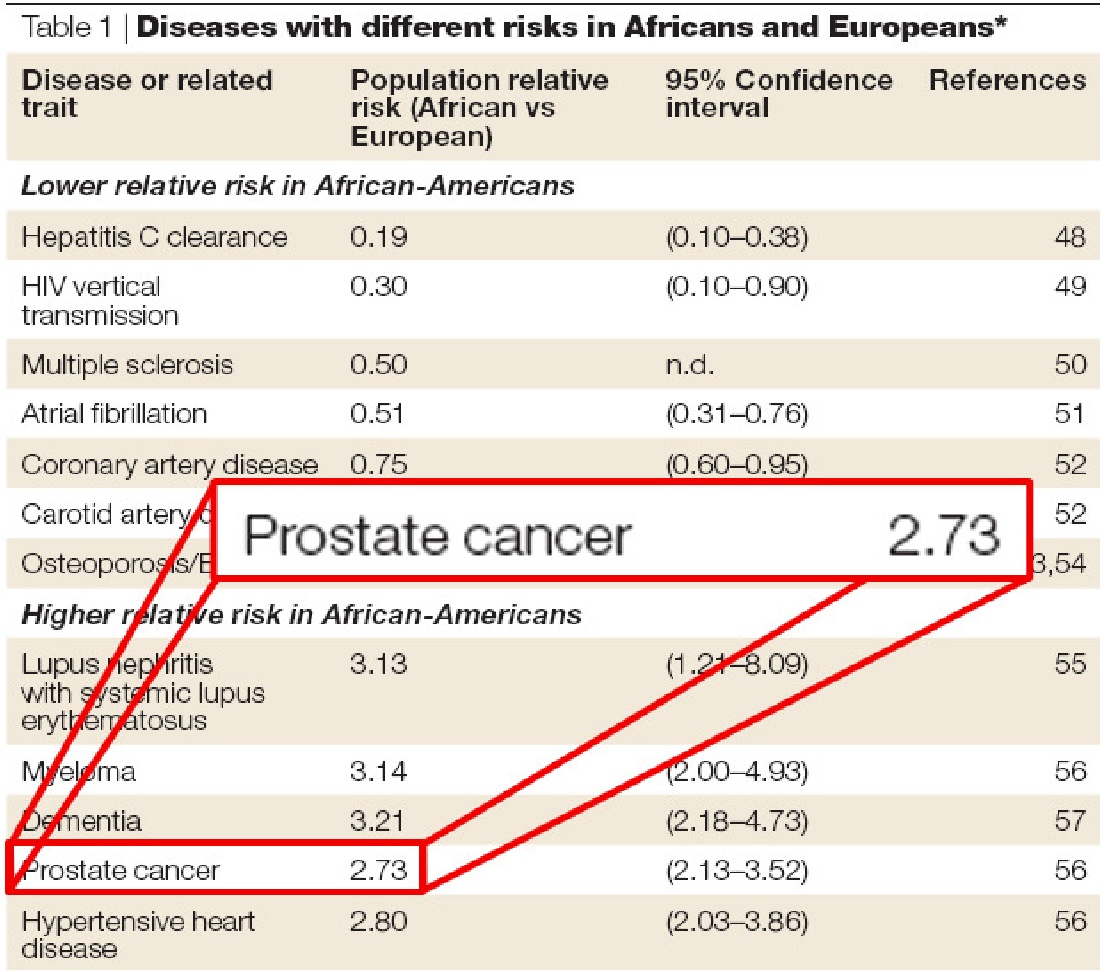


Table 1 | **Diseases with different risks in Africans and Europeans***

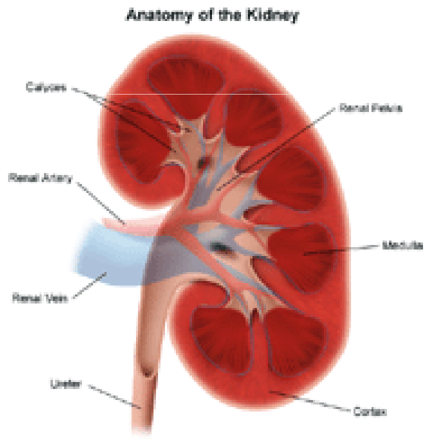
Disease or related trait	Population relative risk (African vs European)	95% Confidence interval	References
Lower relative risk in African-Americans			
Hepatitis C clearance	0.19	(0.10–0.38)	48
HIV vertical transmission	0.30	(0.10–0.90)	49
Multiple sclerosis	0.50	n.d.	50
Atrial fibrillation	0.51	(0.31–0.76)	51
Coronary artery disease	0.75	(0.60–0.95)	52
Carotid artery d			52
Osteoporosis/B			3,54
Higher relative risk in African-Americans			
Lupus nephritis with systemic lupus erythematosus	3.13	(1.21–8.09)	55
Myeloma	3.14	(2.00–4.93)	56
Dementia	3.21	(2.18–4.73)	57
Prostate cancer	2.73	(2.13–3.52)	56
Hypertensive heart disease	2.80	(2.03–3.86)	56

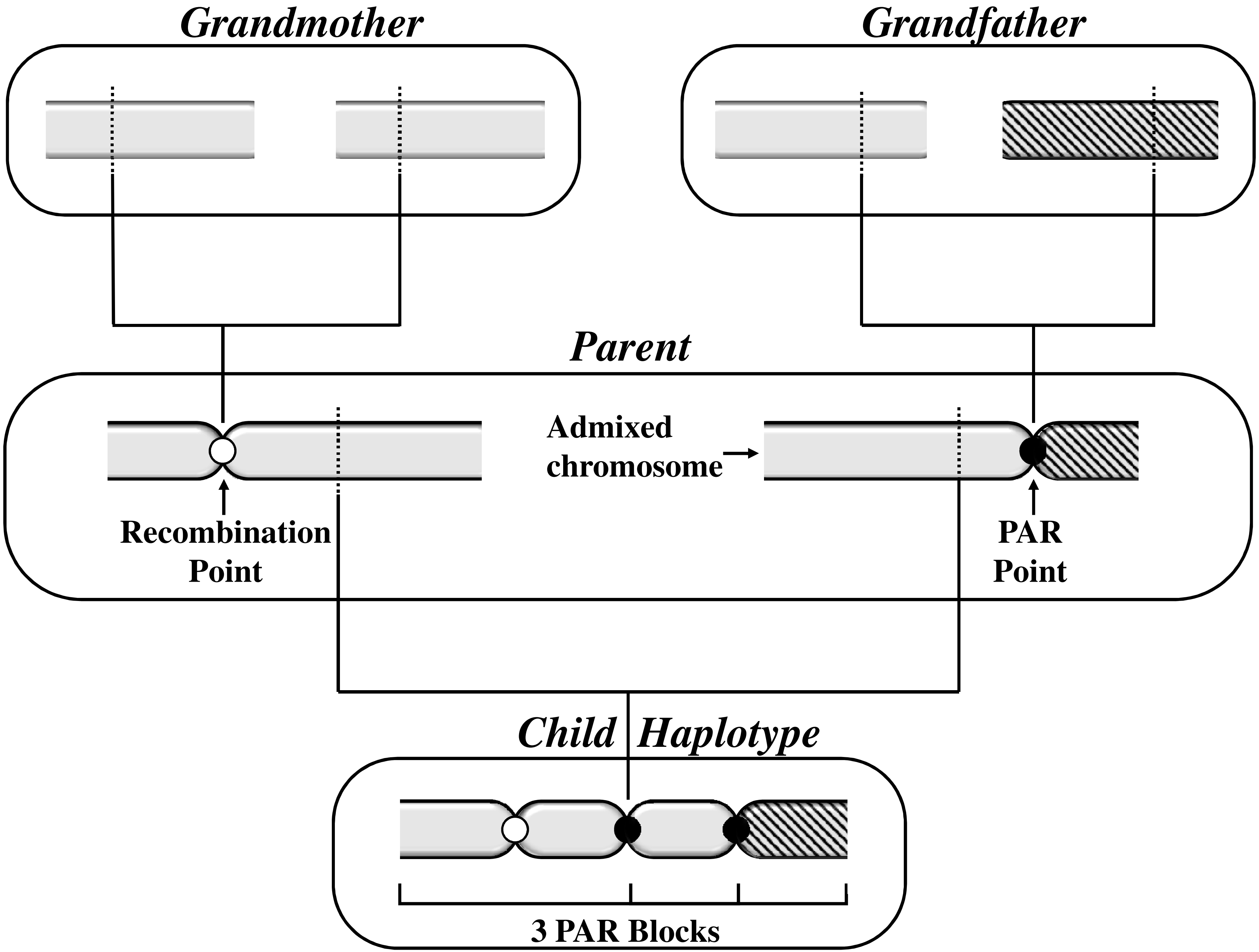
"MAPPING BY ADMIXTURE LINKAGE DISEQUILIBRIUM: ADVANCES, LIMITATIONS AND GUIDELINES" (Smith & O'Brien, *Nature Reviews Genetics*, 2005)

End Stage Renal Disease (ESRD)

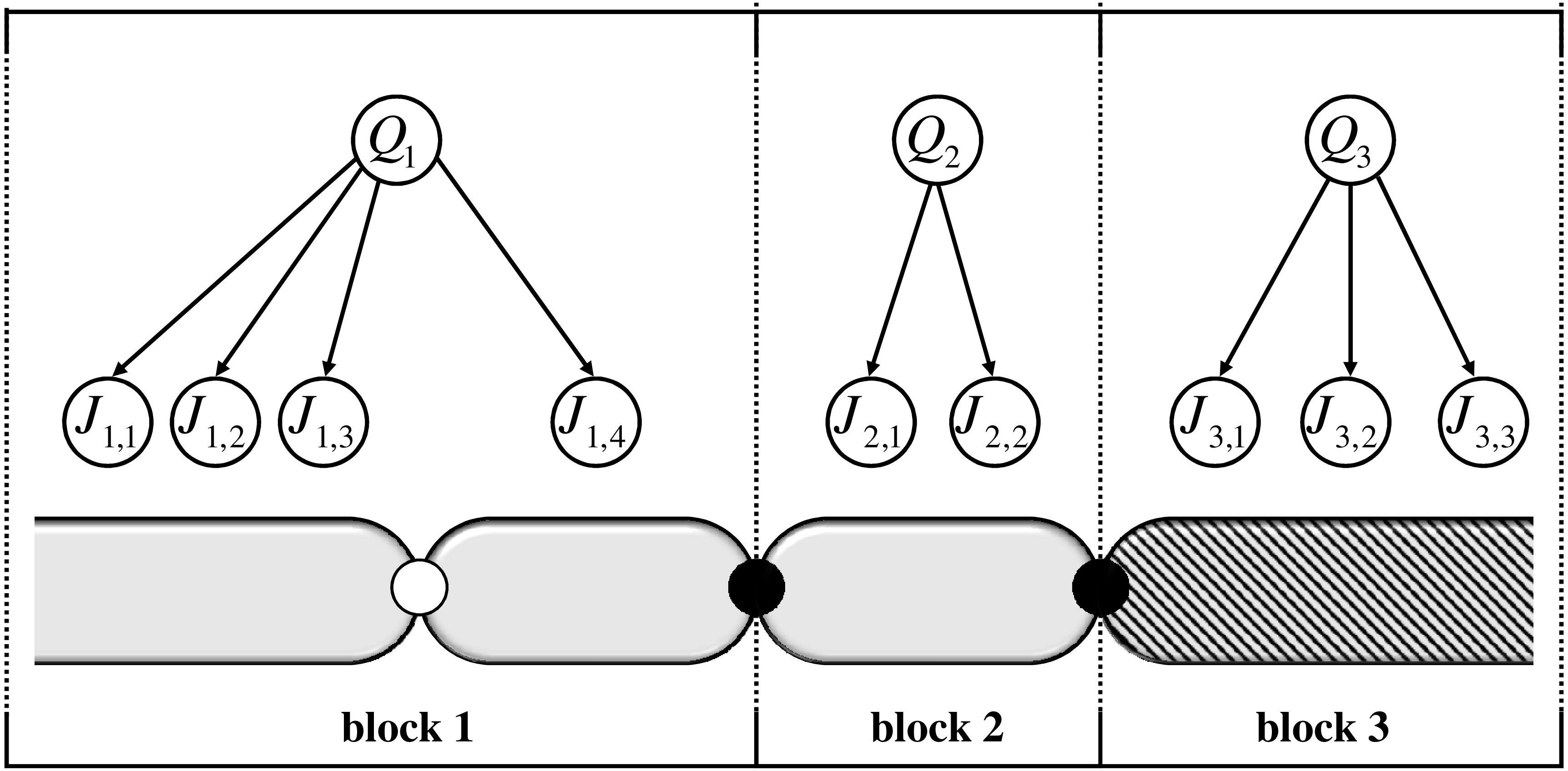
ESRD: causes chronic loss of normal kidney function.
Dialysis: removing waste substances from the blood replacing kidneys.
(http://www.nhlbi.nih.gov/health/dci/Diseases/Cad/CAD_WhatIs.html)

This is a <u>complex disease.</u>
Prevalence: ~0.15% in Israel and the US
ERR = 1.4



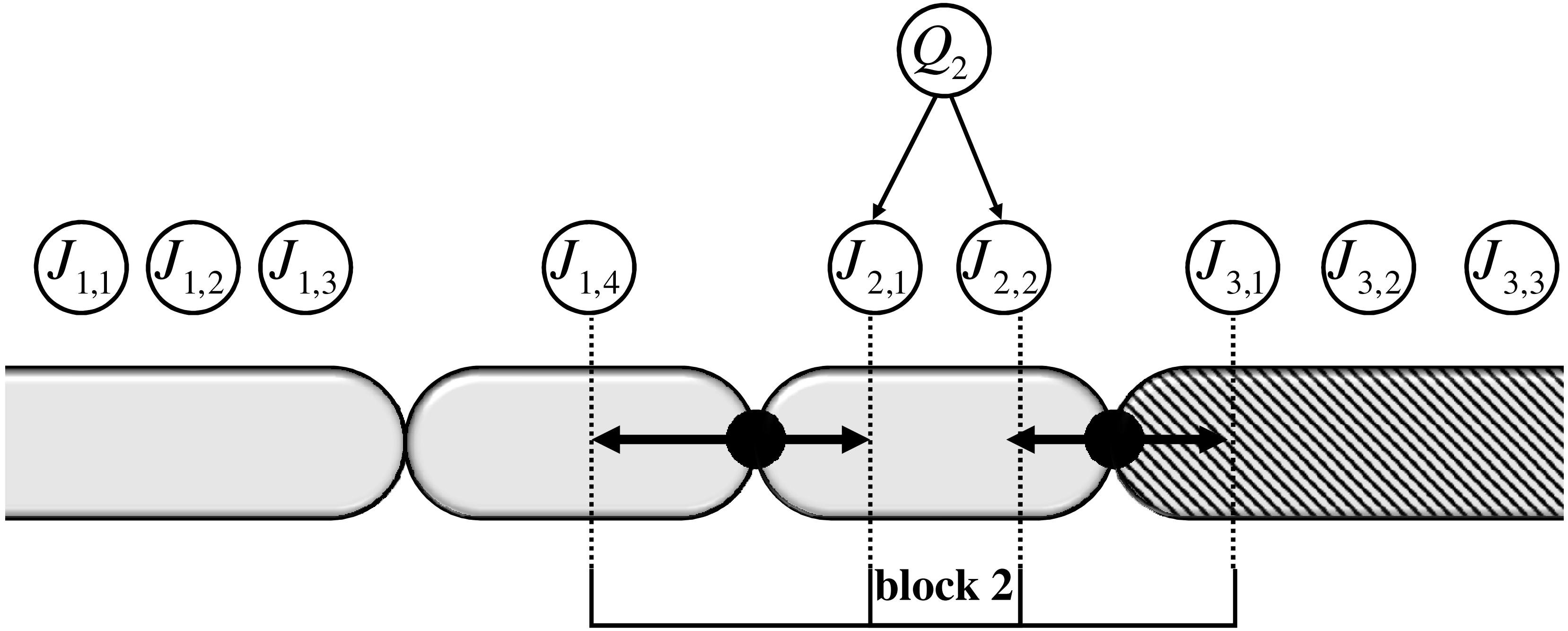

Grandmother
Grandfather
Parent
Admixed chromosome
Recombination Point
PAR Point
Child Haplotype
3 PAR Blocks

Expected Mutual Information (EMI)



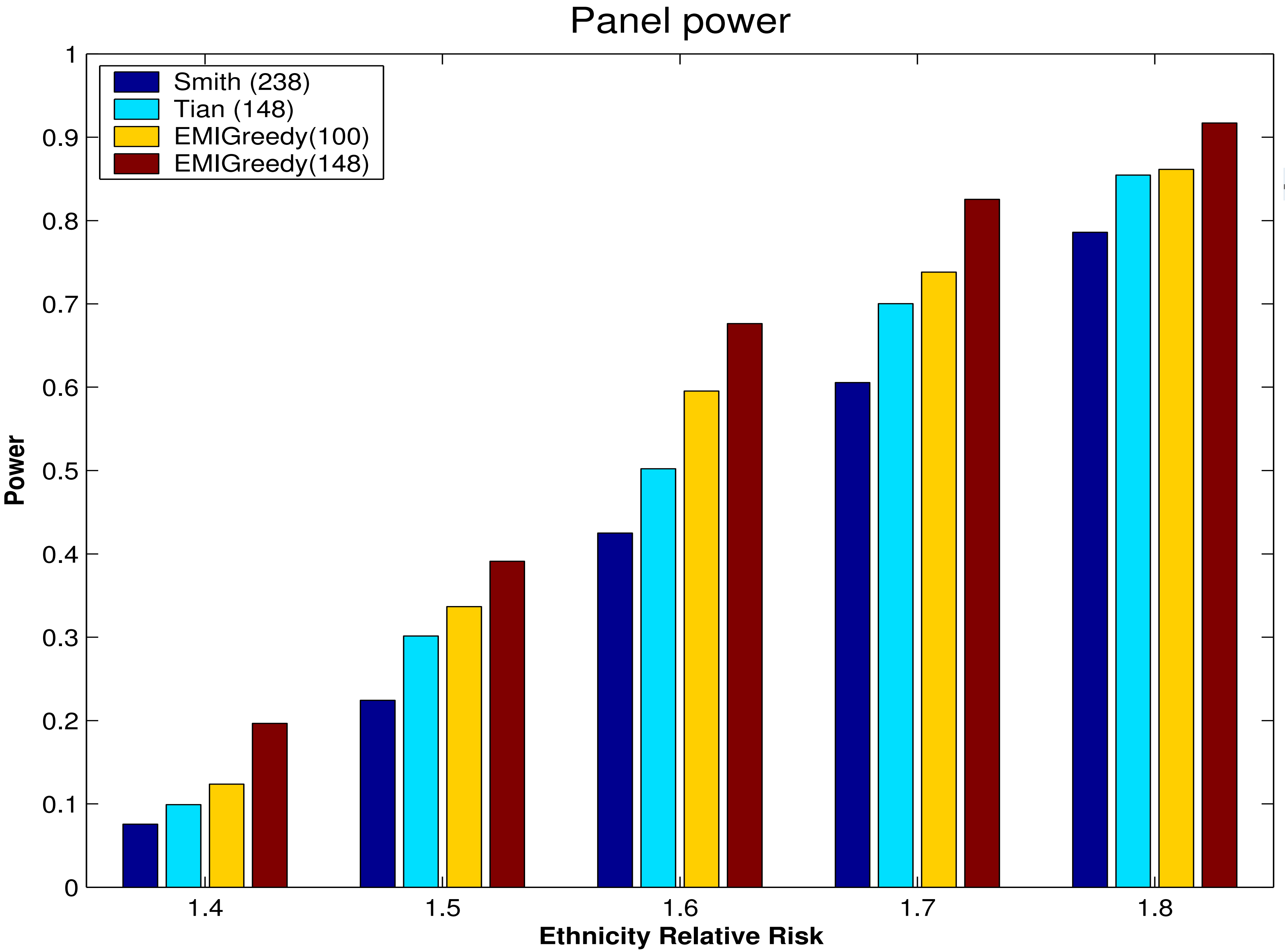
$$\mathbb{E}I(Q_x; J) = \sum_{\pi} P(\pi) \cdot I(Q_x; J|\pi)$$

Computational Shortcut

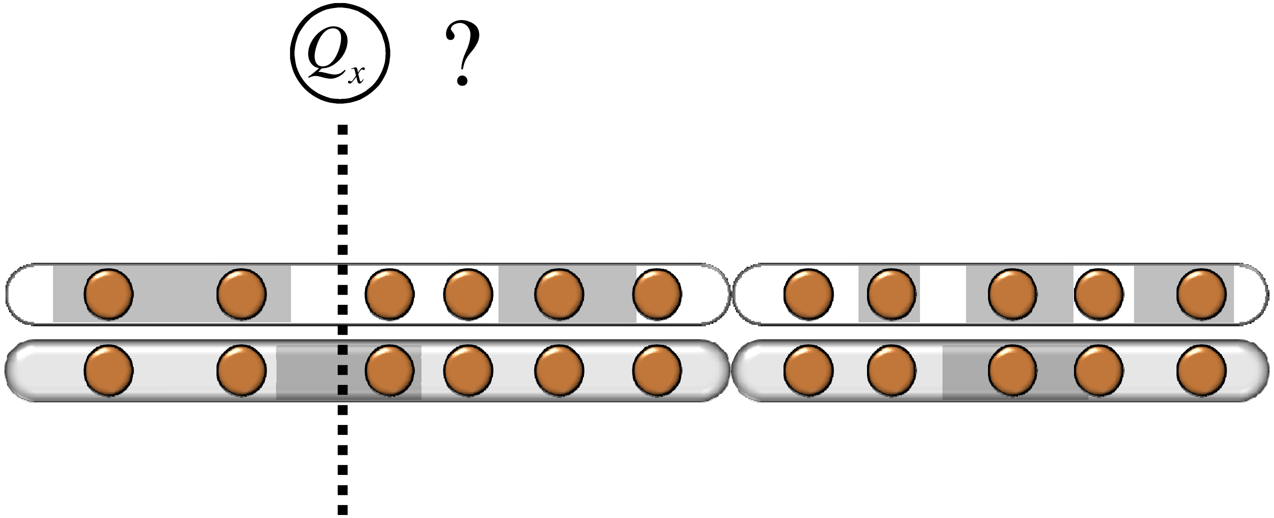


$$\mathbb{E}I(Q_x;J) = \sum_{\pi} P(\pi) \cdot I(Q_x;J|\pi)$$

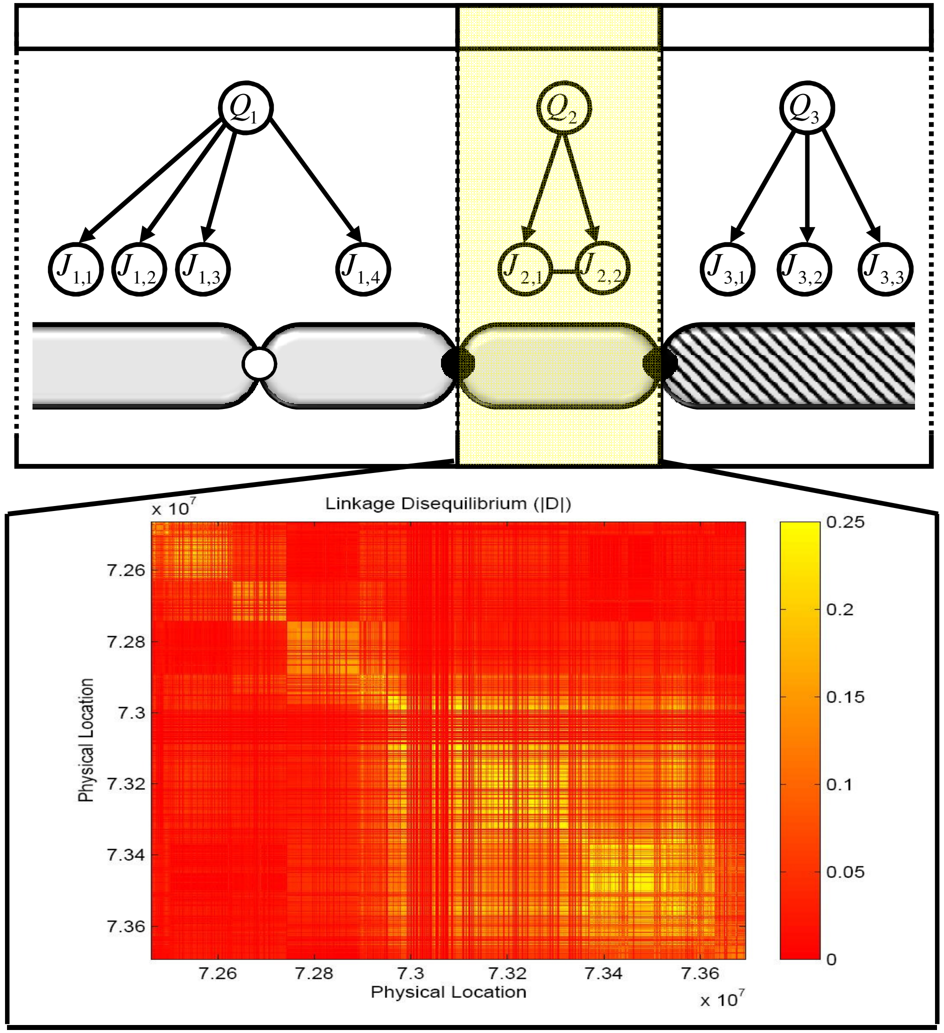
$$EMI(Q_x;J) = \sum_{l \in L} \sum_{r \in R} P_{(l,r)} \cdot I(Q_x;J_{[l,r]})$$


Panel power
Smith (238)
Tian (148)
EMIGreedy(100)
EMIGreedy(148)
Power
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.4
1.5
1.6
1.7
1.8
Ethnicity Relative Risk

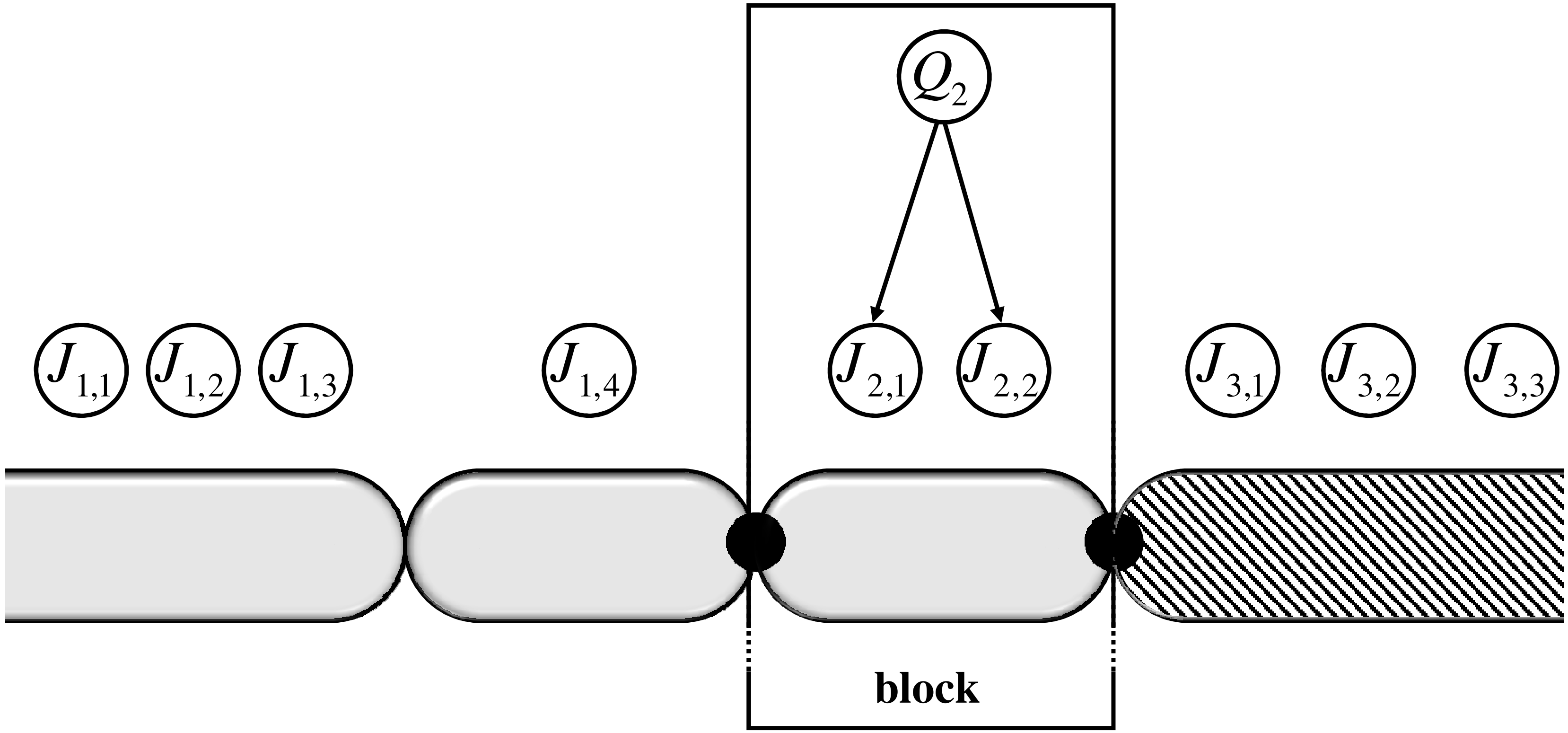
Inferring Ancestry



Linkage Disequilibrium



Ancestry Inference



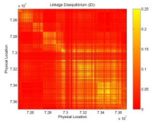
$$P(Q_x|J) = \sum_{\pi} P(Q_x|\pi, J) \cdot P(\pi|J)$$

Efficient Inference

$$\boxed{P(Q_x|J) = \sum_{\pi} P(Q_x|\pi, J) \cdot P(\pi|J)}$$

$$\downarrow$$

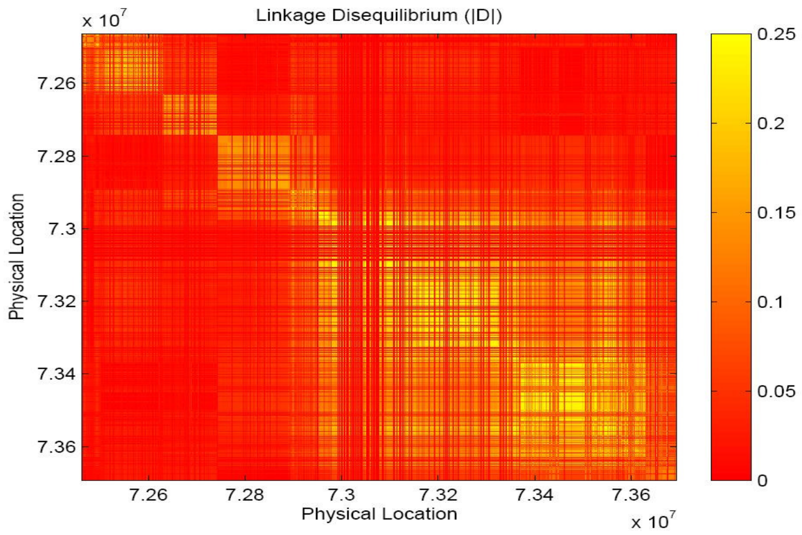
$$\boxed{\begin{aligned} &P(Q_x|J) = \frac{1}{P(J)} \sum_{l \in L} \sum_{r \in R} \\ &\underbrace{P(J_{l,r}|Q_x, \pi_{l,r})} \cdot P(\pi_{l,r}) \cdot P(Q_x) \cdot P(J_{.,l}) \cdot P(J_{r,.}) \end{aligned}}$$

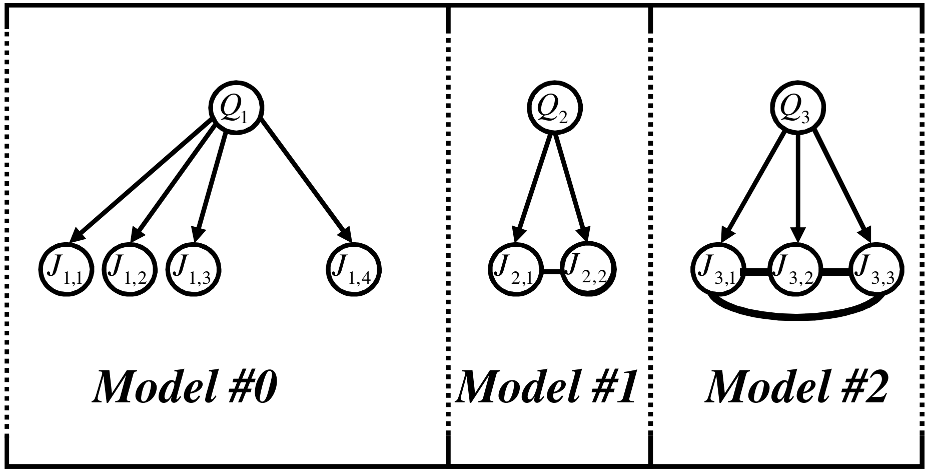


Most Probable Ancestry

$$\hat{Q} = \operatorname*{argmax}_{Q,\pi} P(Q, \pi | J)$$

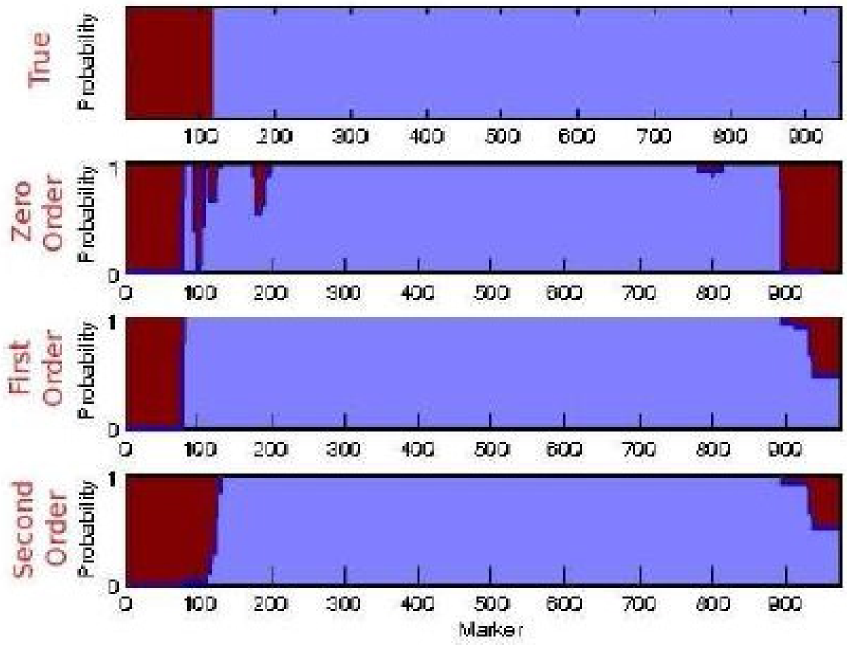
Linkage Disequilibrium Models





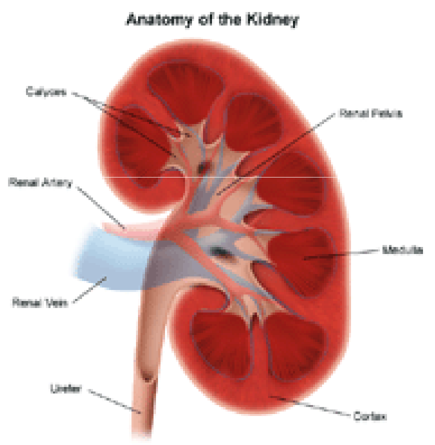
Results (error %)

Method	0	1	2
Post	4.4655 %	0.6 %	0.24 %
MAP	4.11 %	0.29 %	0.16 %



End Stage Renal Disease (ESRD)

ESRD: causes chronic loss of normal kidney function.



<u>RESULT</u>: At Karl Skorecky's lab we scanned merely ~400 affected and were able to locate a suspect gene for ESRD.

Ancestry Inference - Summary

- Probabilistic framework for ancestry inference
  - Better choice of markers
  - Supports realistic LD models
  - Efficient

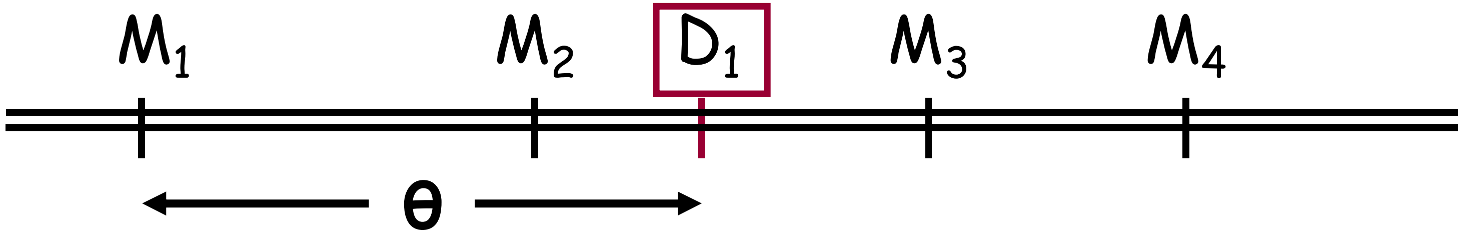
SPEEDING UP HMM ALGORITHMS FOR GENETIC LINKAGE ANALYSIS VIA CHAIN REDUCTIONS OF THE STATE SPACE

To be presented at ISMB 2009

Microsoft Research

Dan Geiger, Christopher Meek & Ydo Wexler

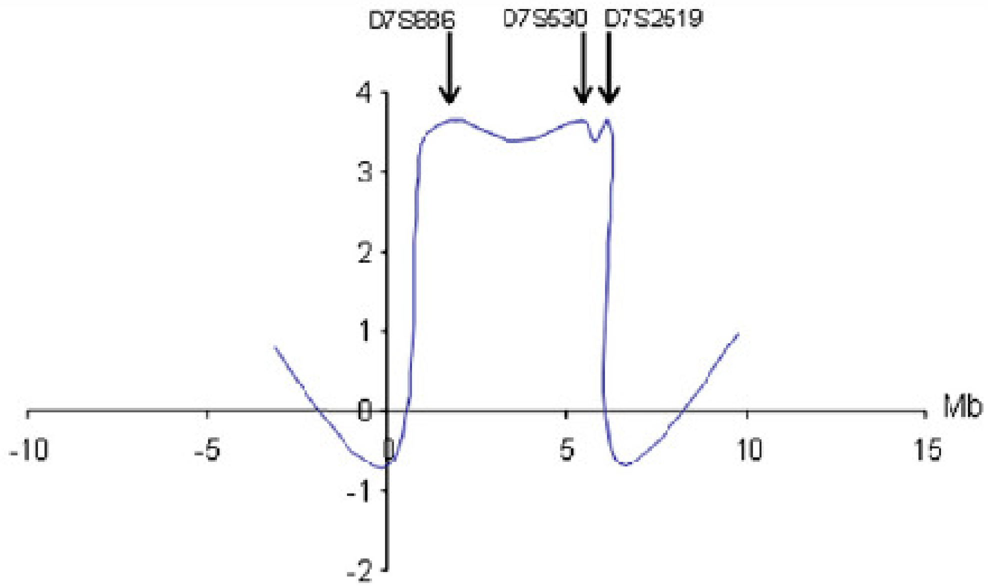
The basic gene mapping principle



Find the location θ that maximizes the LOD score (main computational goal):

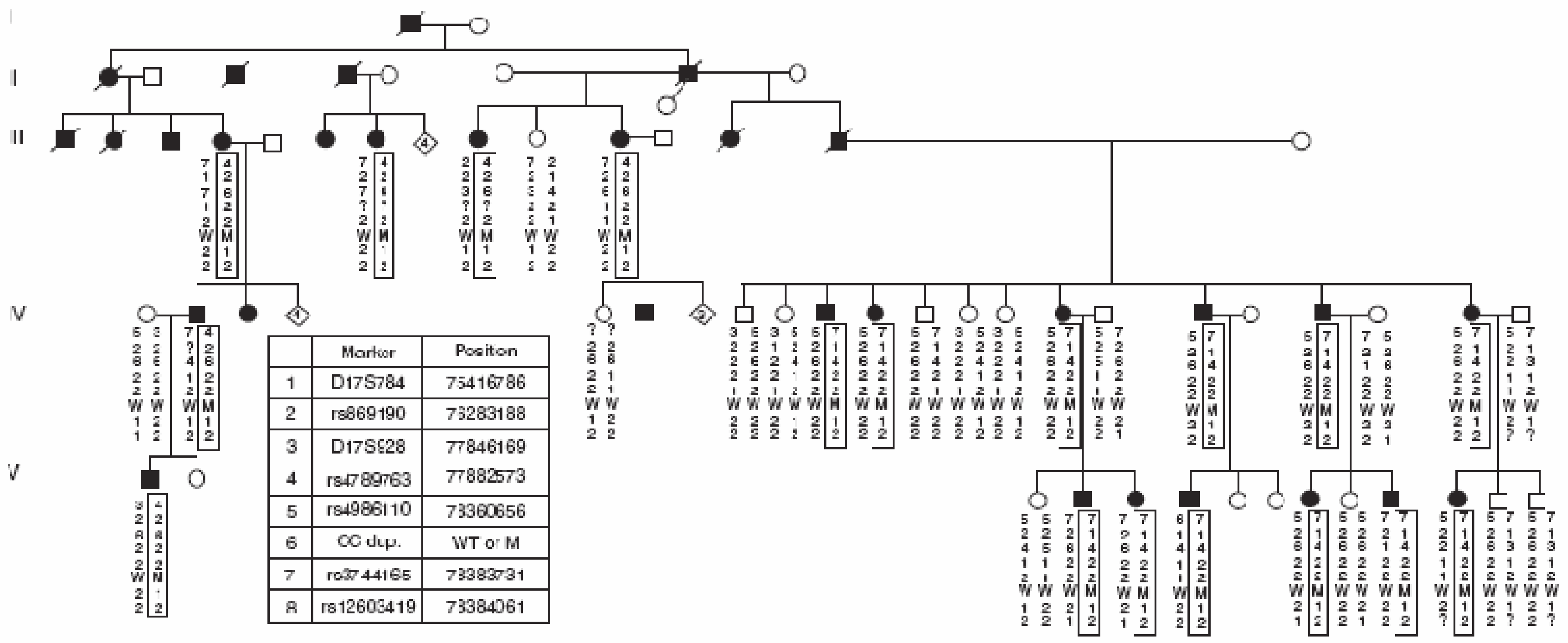
$Z(\theta)=\log_{10}$ [Pr(data|θ) / Pr(data| no linkage)].

Typical Results of Analysis



The American Journal of Human Genetics *82*, 1114–1121, May 2008

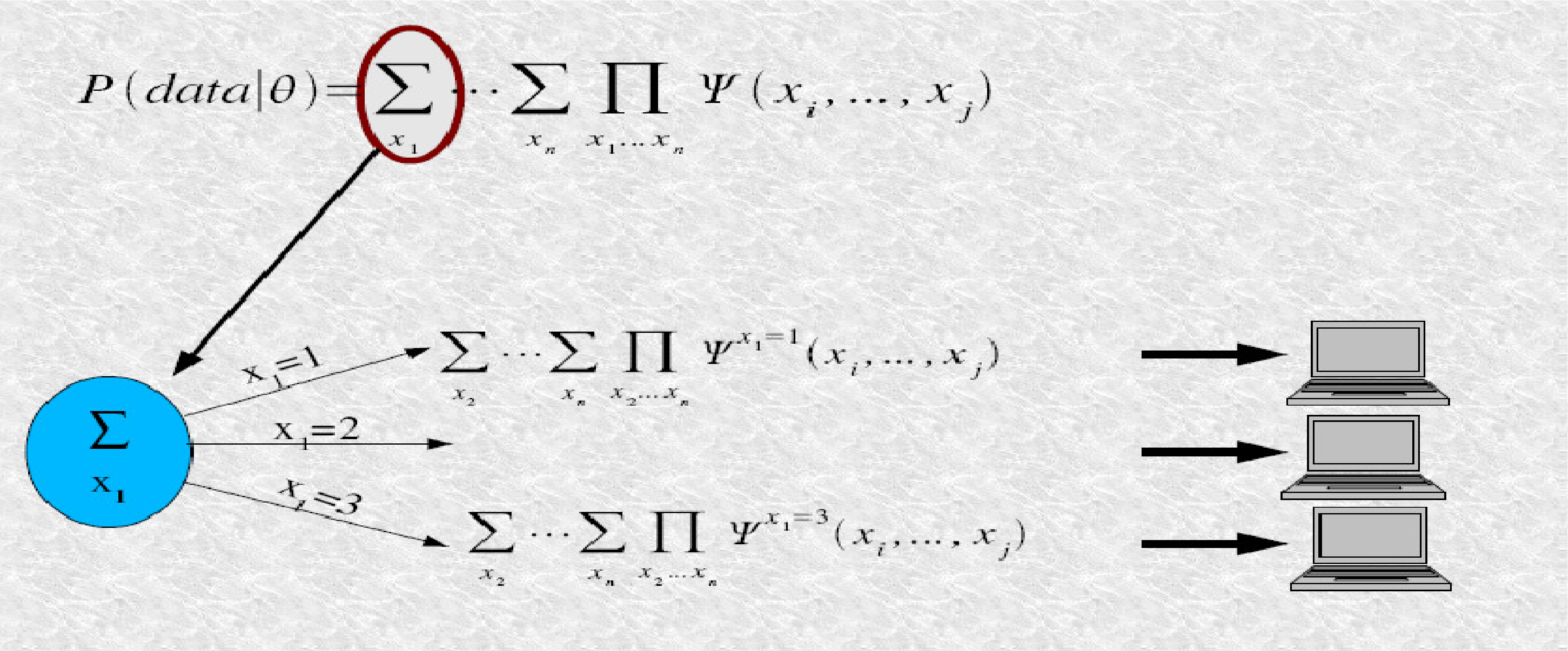
Family Pedigree



	Marker	Positon
1	D17S784	75416786
2	rs869190	73283188
3	D17S928	77846169
4	rs4789763	77882573
5	rs4986110	73360656
6	CC dup.	WT or M
7	rs3744165	73383731
8	rs12603419	73384061

The Likelihood function

$$P(\mathbf{data}|\theta)=\sum_{x_k}\cdots\sum_{x_3}\sum_{x_1}\prod_{i=1}^{n}P_\theta(x_i\,|\,pa_i)$$

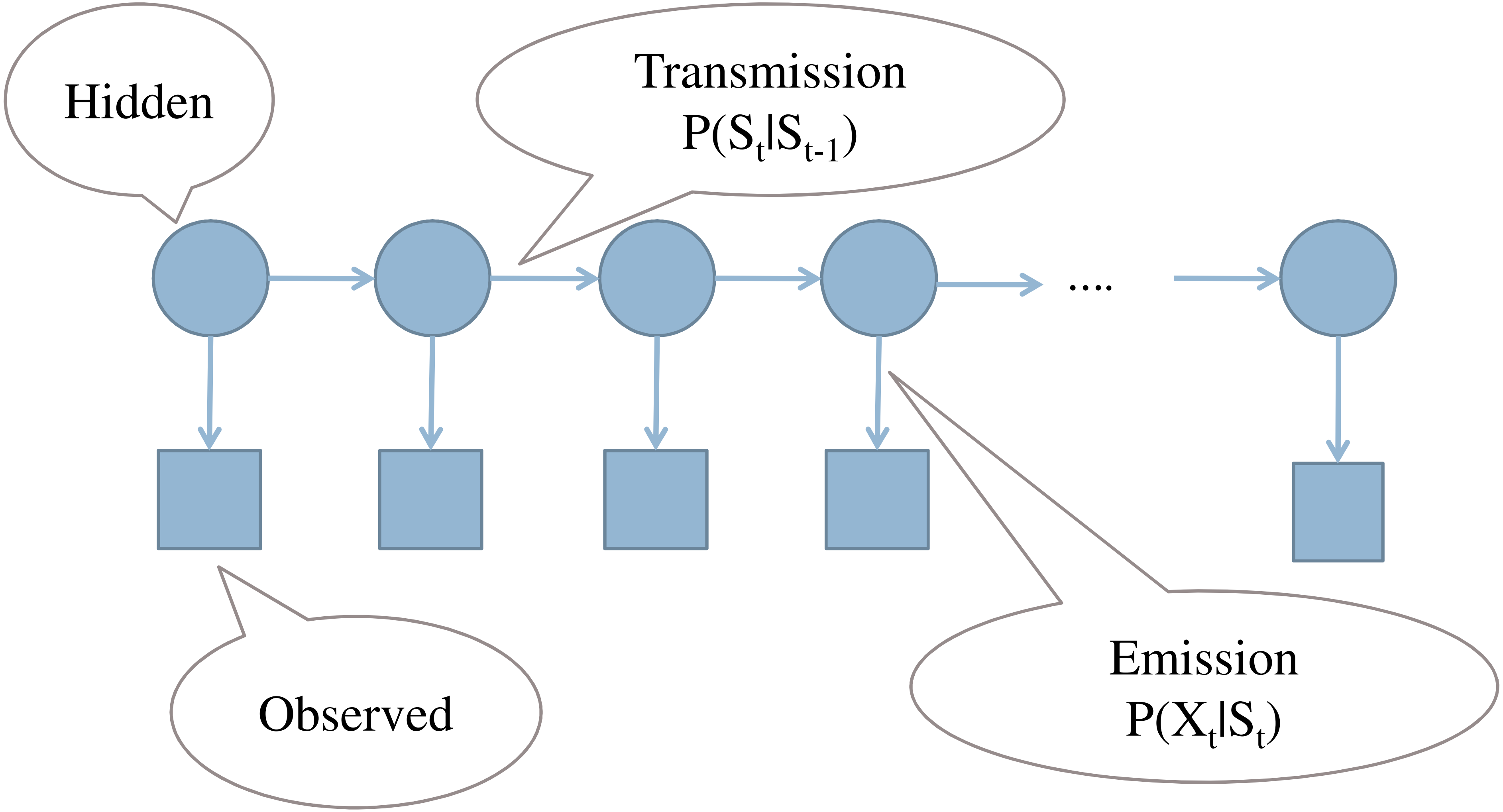


$$P(data|\theta)=\sum_{x_1}\cdots\sum_{x_n}\prod_{x_1\ldots x_n}\Psi(x_i,\ldots,x_j)$$

$$\sum_{x_2}\cdots\sum_{x_n}\prod_{x_2\ldots x_n}\Psi^{x_1=1}(x_i,\ldots,x_j)$$

$$\sum_{x_2}\cdots\sum_{x_n}\prod_{x_2\ldots x_n}\Psi^{x_1=3}(x_i,\ldots,x_j)$$

Hidden Markov Models (HMMs)



Lander Green Algorithm

HMM Computations

- Forward-backward, Viterbi, likelihood of data
  - All take $O\left(L|S|^2 + cL|S|\right)$
- Example (likelihood of evidence):

$$P(data) = \sum_{s_1} P(s_1)P(x_1 \mid S_1 = s_1)\sum_{s_2} P(S_2 = s_2 \mid S_1 = s_1)P(x_2 \mid S_2 = s_2)\cdots$$
$$\cdots\sum_{s_L} P(s_L \mid s_{L-1})P(x_L \mid s_L)$$

- If |S| is large computation is slow
- SOFTWARE: GeneHunter, Alegro, Merlin
- GOAL: reduce the size of S

State space reduction

- Divide states of S into equivalence classes [s]
- Sum over one representative per class
- Example:

$$P(data) = \sum_{[s_1]} P([s_1]) P(x_1 \mid S_1 = [s_1]) \sum_{[s_2]} P(S_2 = [s_2] \mid S_1 = [s_1]) P(x_2 \mid S_2 = [s_2]) \\ \cdots \sum_{[s_L]} P([s_L] \mid [s_{L-1}]) P(x_L \mid [s_L])$$

- Correctness ?

Condition I – Emission Probabilities

- The single slot likelihood given a hidden state s is equal for all states in the class [s]
  - If s, s’ in the same class then

$$P(x_i \mid s) = P(x_i \mid s')$$

$$\Downarrow$$

$$\forall s \in [s] \quad P(x_i \mid [s]) = P(x_i \mid s)$$

Condition II-Transmission Probabilities

- □ Define the transition probability from a state s’ to the class [s] by $P([s] \mid s') = \sum_{s \in [s]} P(s \mid s')$

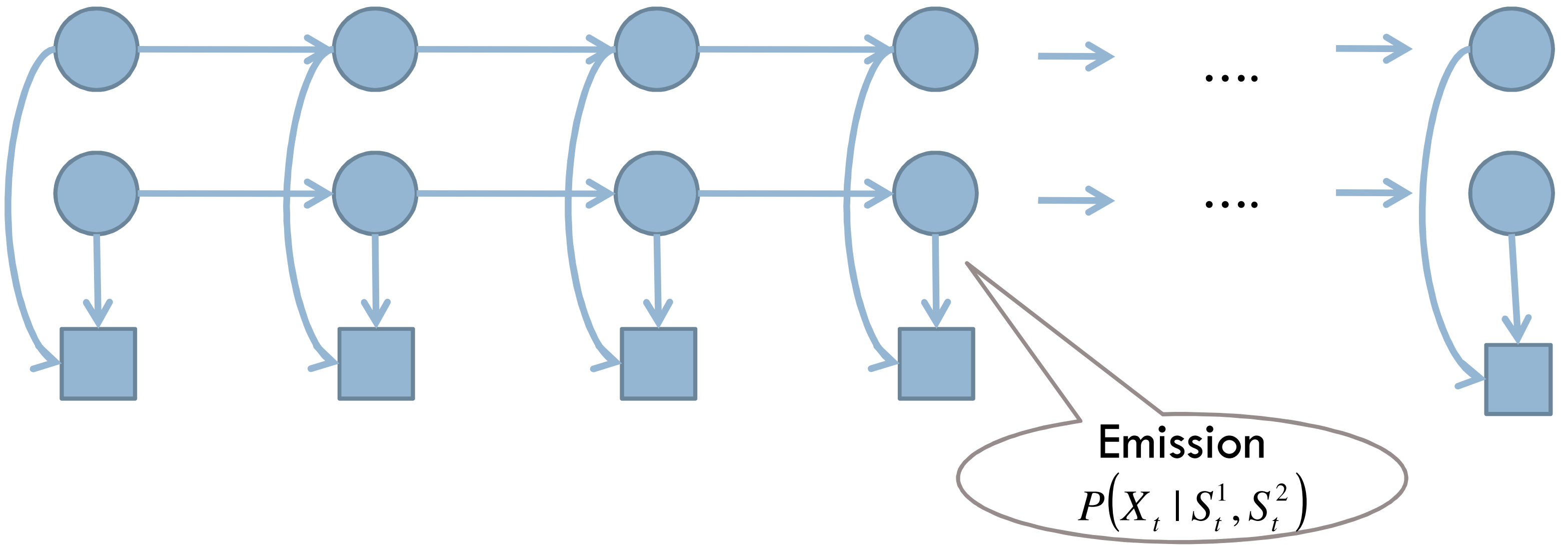
- □ If s’, s” in the same class then $P([s] \mid s') = P([s] \mid s'')$

$$\Downarrow$$

$$P([s] \mid [s']) = P([s] \mid s')$$

- □ Complexity is quadratic in number of classes, not in number of states.

Factorial HMMs



Emission
$P(X_t \mid S_t^1, S_t^2)$

- State-space is now $S_i = (S_i^1, \ldots, S_i^k)$
- Complexity $O\left(L|S|\log|S| + cL|S|\right)$
  - Ghahramani & Jordan

- Homogeneously Factored HMM
  - transition $P_j\left(s_i^j \mid s_{i-1}^j\right)$ is equal for all j

Simplifying assumptions

- Binary variable (selectors)
  - A selector is either ON or OFF
- Symmetric transition – probability to switch states
  - $P\left(s_i^j = 0 \mid s_{i-1}^j = 1\right) = \theta$
  - $P\left(s_i^j = 1 \mid s_{i-1}^j = 0\right) = \theta$

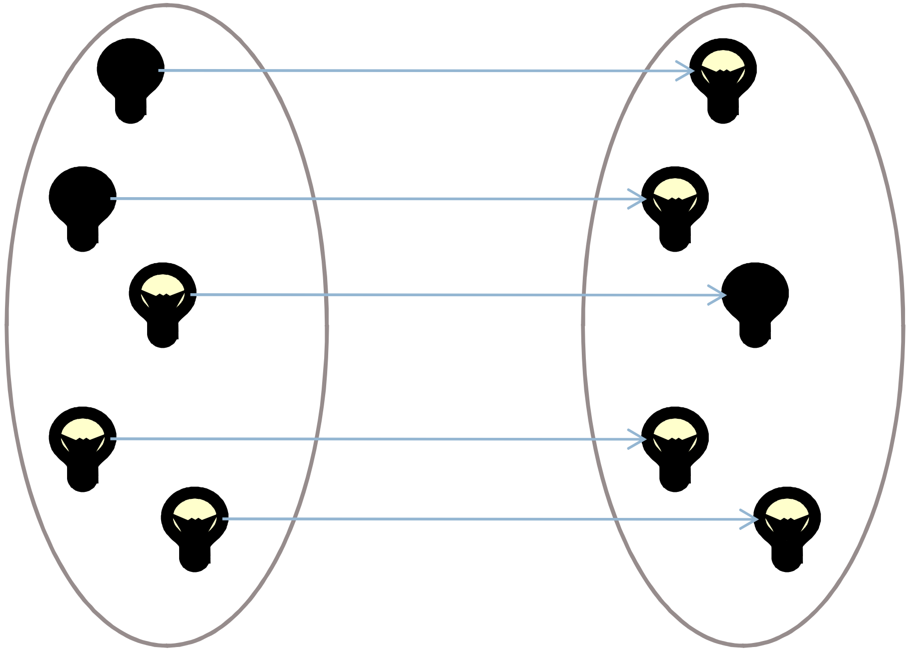
Counting partition

A state space reduction for factored HMMs

- Selectors are grouped together
  - A cluster C with r selectors
  - Equivalence class [j] = all states with j selectors ON
- $c(j,r) = r!/\, j!(r-j)!$ states become one state
- Each cluster r+1 states
- Still factored HMM
- Thm: *Counting Partitions* satisfy Condition II

Example

- We just care how many bulbs are ON
- The probability of getting from 3 bulbs ON to 4 bulb ON doesn’t depend on the bulbs identity



Complexity

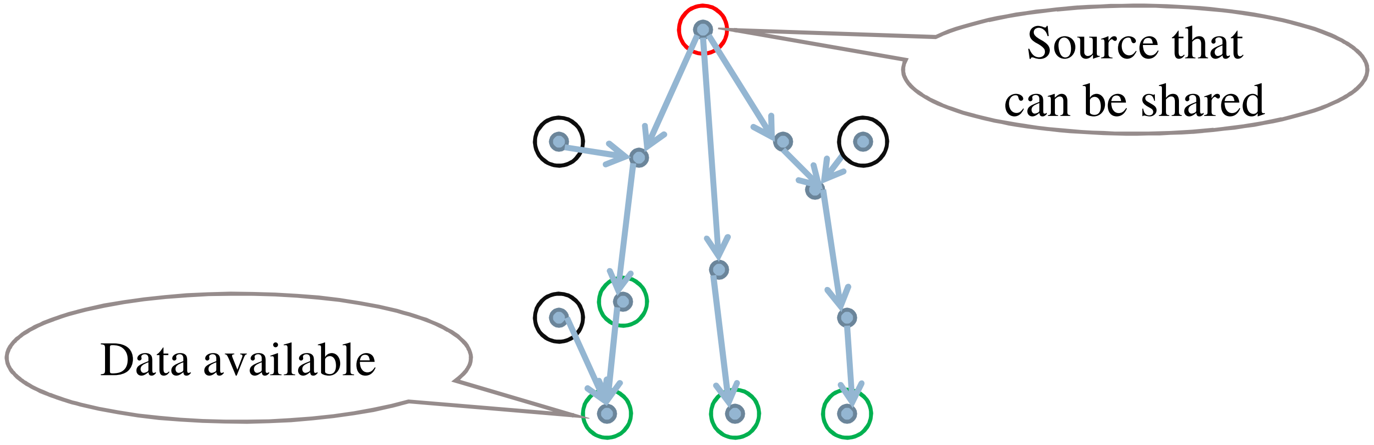
- State space for a cluster reduces from 2^r to r+1
- If all selectors are in one cluster the complexity becomes quadratic in r and linear in the length.
- If each selector has a cluster then no savings.

HMM for linkage analysis

- Individuals have a pair of selectors at each location
- Modeled as a homogenously factored HMM
  - Assumptions (binary, symmetry) hold
- The state space is 2^{2n-f}
  - n is the number of non-founders in the pedigree
  - GeneHunter, Allegro, Merlin (and superlink)
- Fast for small pedigrees, impossible for larger pedigrees

Chain reductions

- Pedigrees that contain many people for which there is no genetic data
  - Recent generations are measured
  - Chains from common ancestors to individuals with data

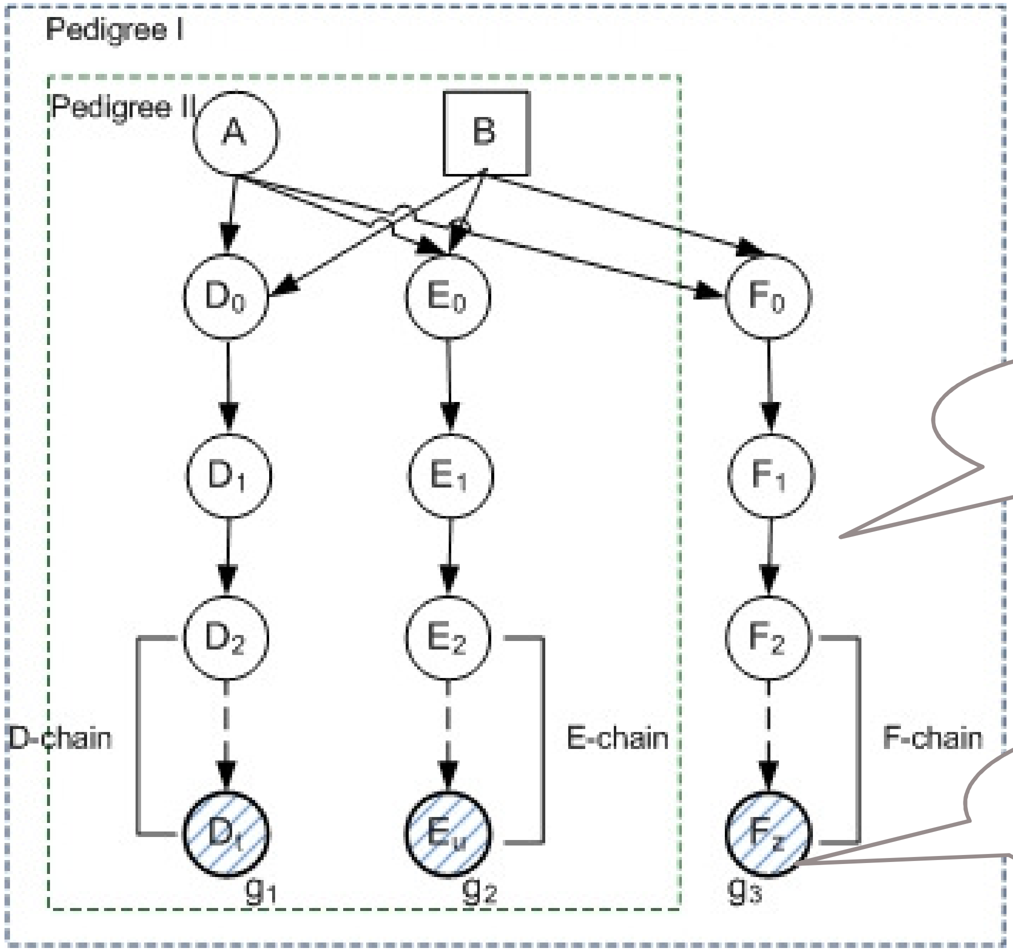


Chain reductions

- Theorem: The selectors for individuals in valid chains can be clustered via the Counting Partition; Condition I is satisfied as well.

Example: g-degree cousins

- 2 founder that matter (4 possible sources)

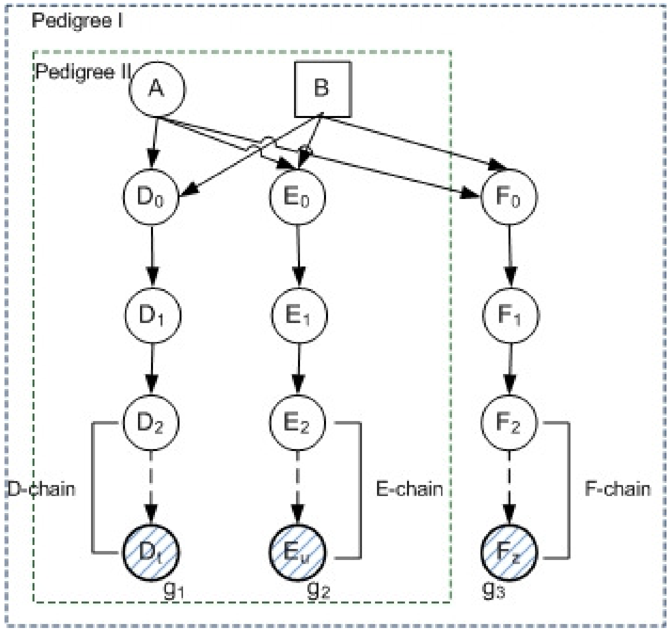


3 chains with no information

Information only for the last generation

Example: g-degree cousins (cont.)

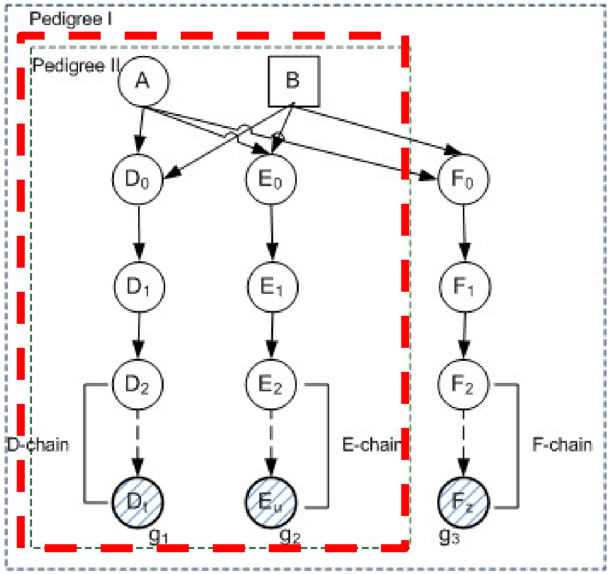
- # informative meioses $4+t+u+z$
  - inheritance vector size $2^{4+t+u+z}$
- New state space $2^7 \cdot t \cdot u \cdot z$



Chain (loop) reductions

- 2 chains that share a common source
  - No other chain out of this source

- The selectors in the 2 chains can be clustered together

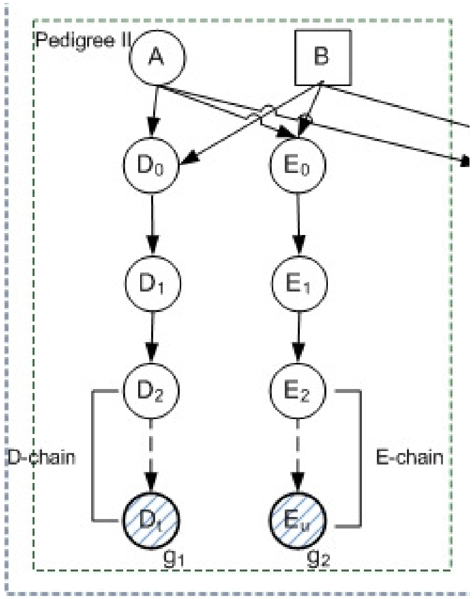


Chain (loop) reductions

- 2 chains that share a common source
  - No other chain out of this source

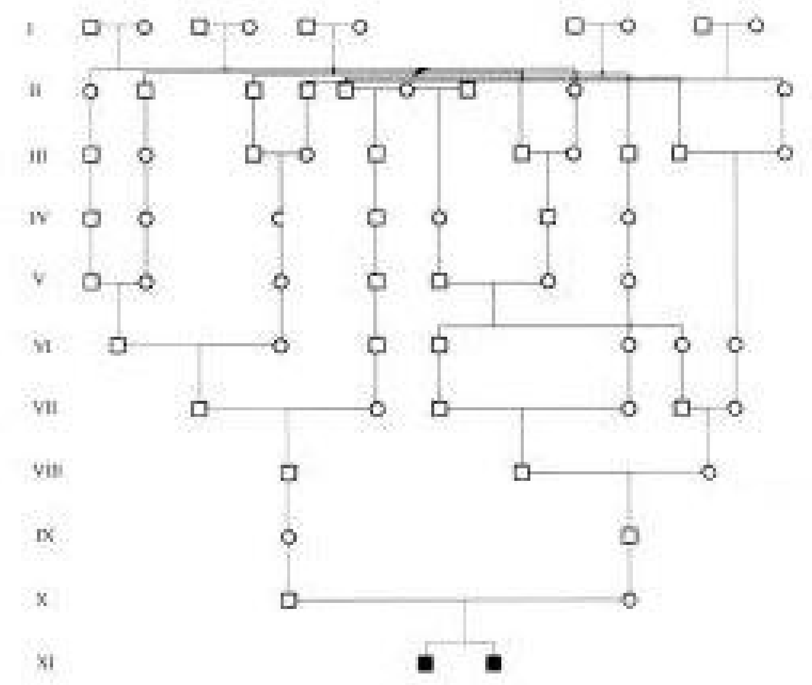
- The selectors in the 2 chains can be clustered together

- We only care whether g_1, g_2 got the same source



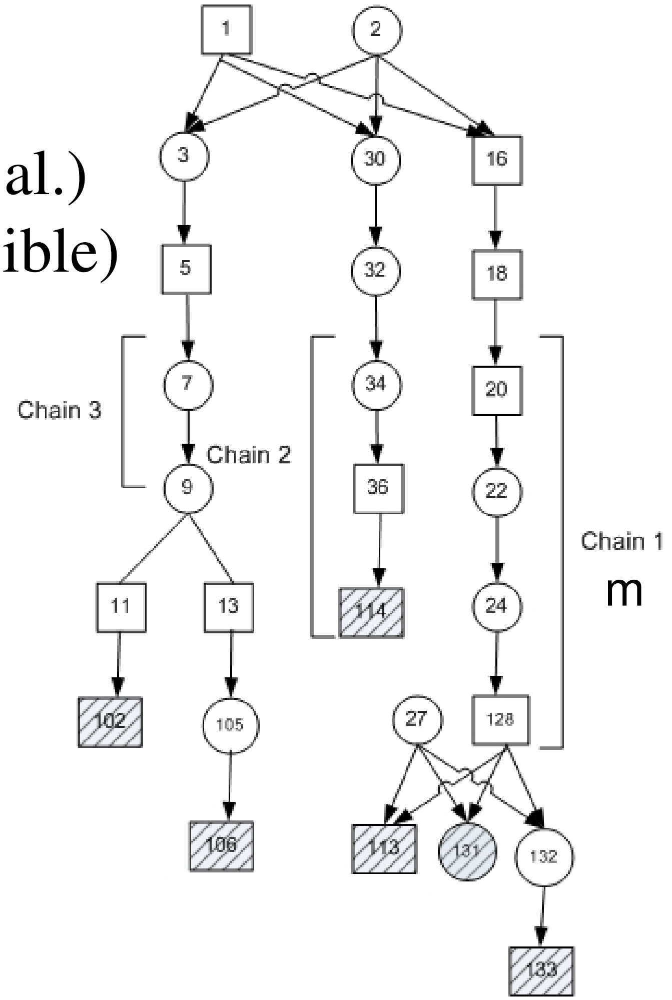
Results

- Pedigree for studying cold-inducing sweating syndrome
  - State space 2^{50} (not feasible)
  - Reduced state space = 2^{32} (still not feasible, but better)



Results

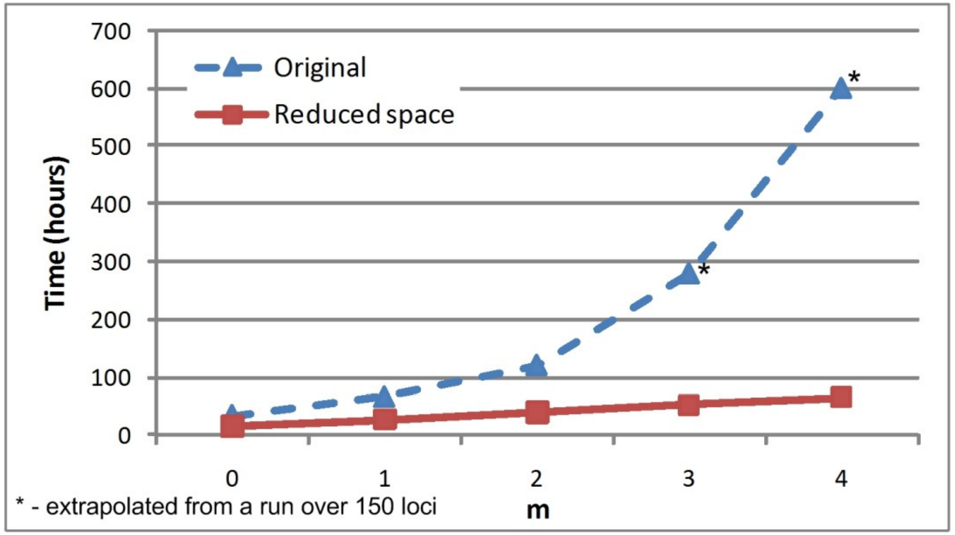
- Pedigree for pituitary adenoma
  - State space 2^{27} (not feasible)
  - Approximations were used (Albers et.al.)
  - Reduced state space = $2^{18}*3*4*5$ (feasible)



m

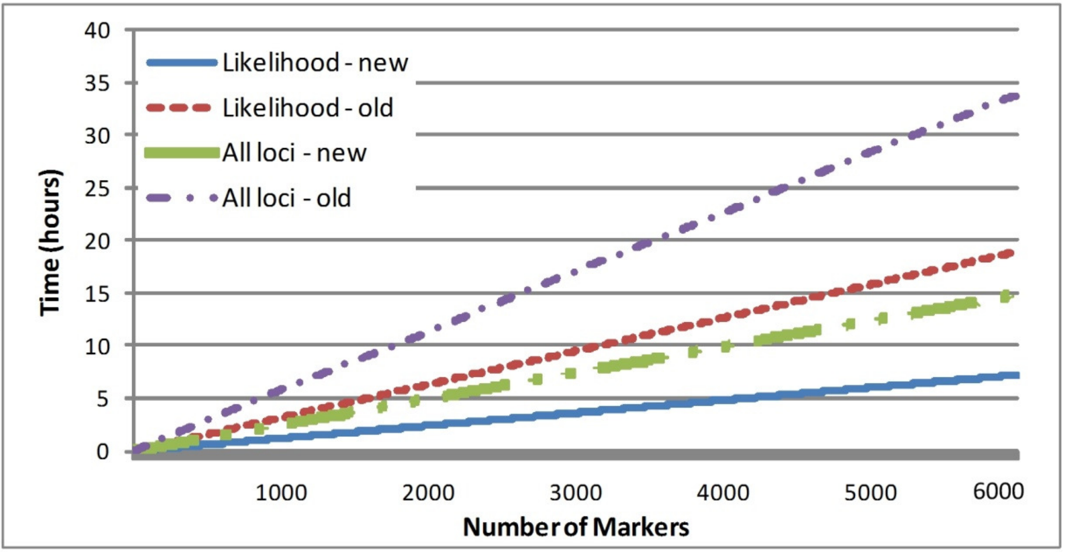
Results

- □ Computed across 6000 loci
- □ Performs as should in theory



* - extrapolated from a run over 150 loci

m=0



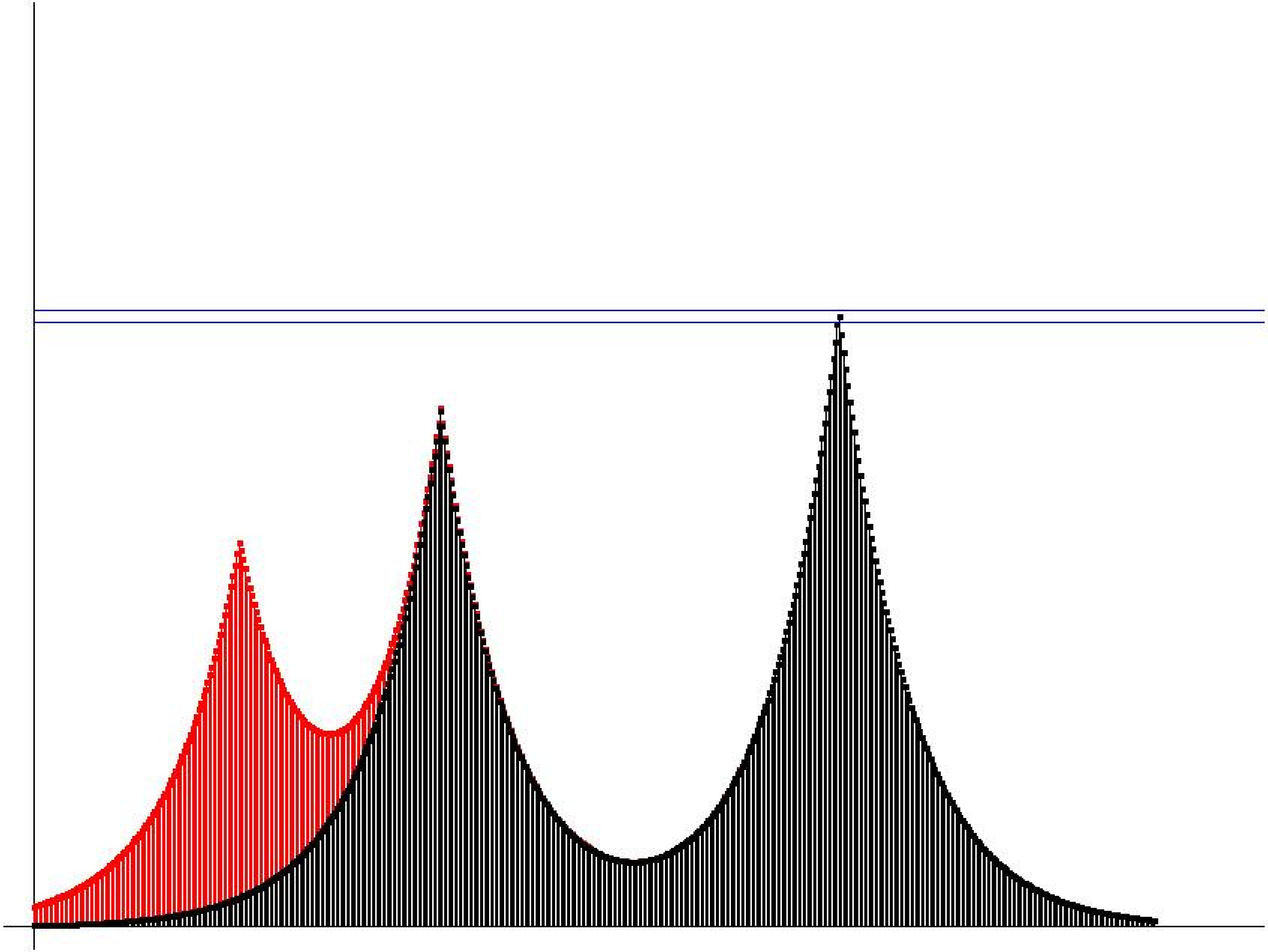
Acknowledgements

Karl Skorecki
Liran Shlush
Alan Templeton
Walter Wasser
Guennady Yudkovsky

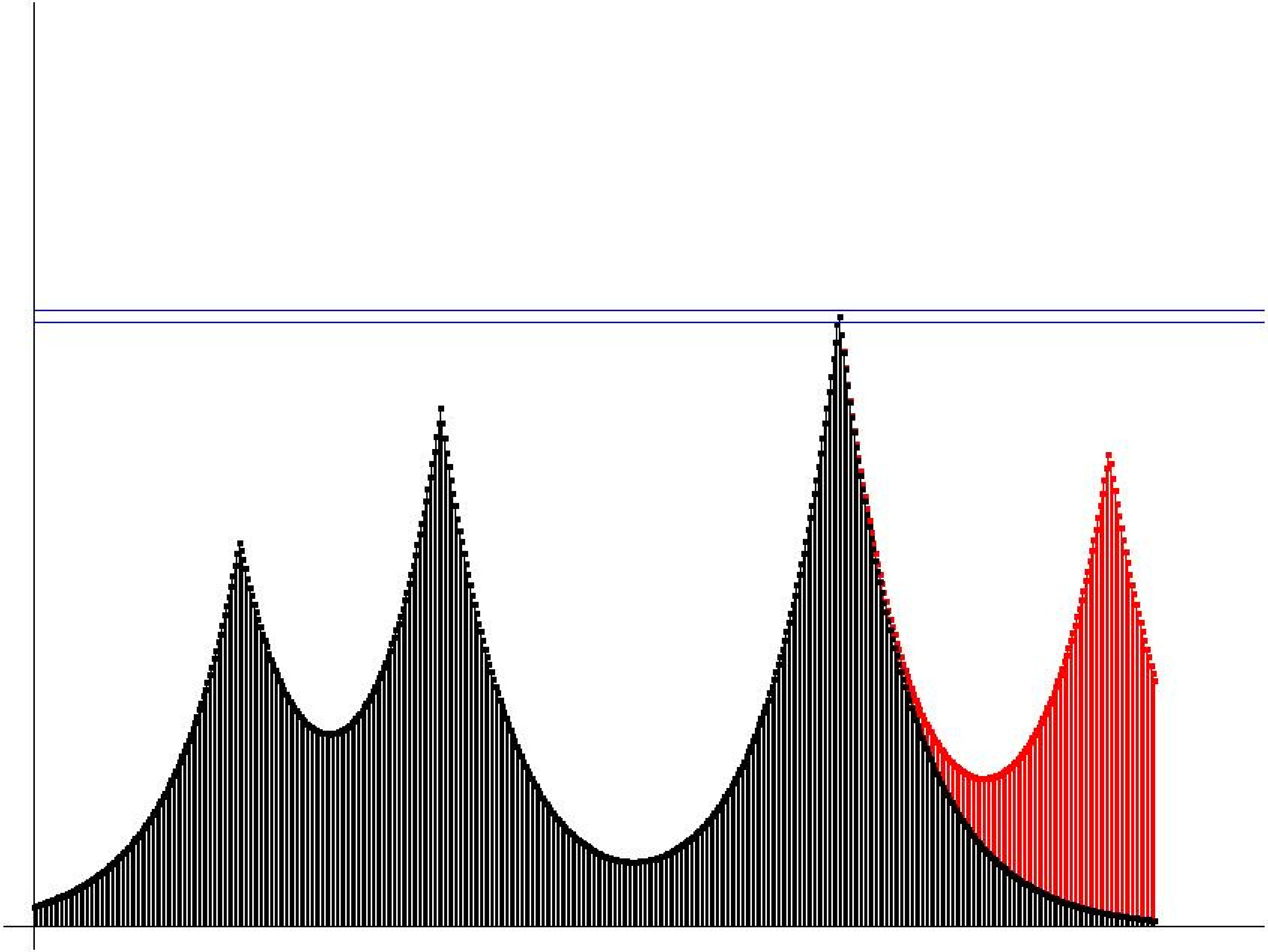
Mark Silberstein
Assaf Schuster

Thank You

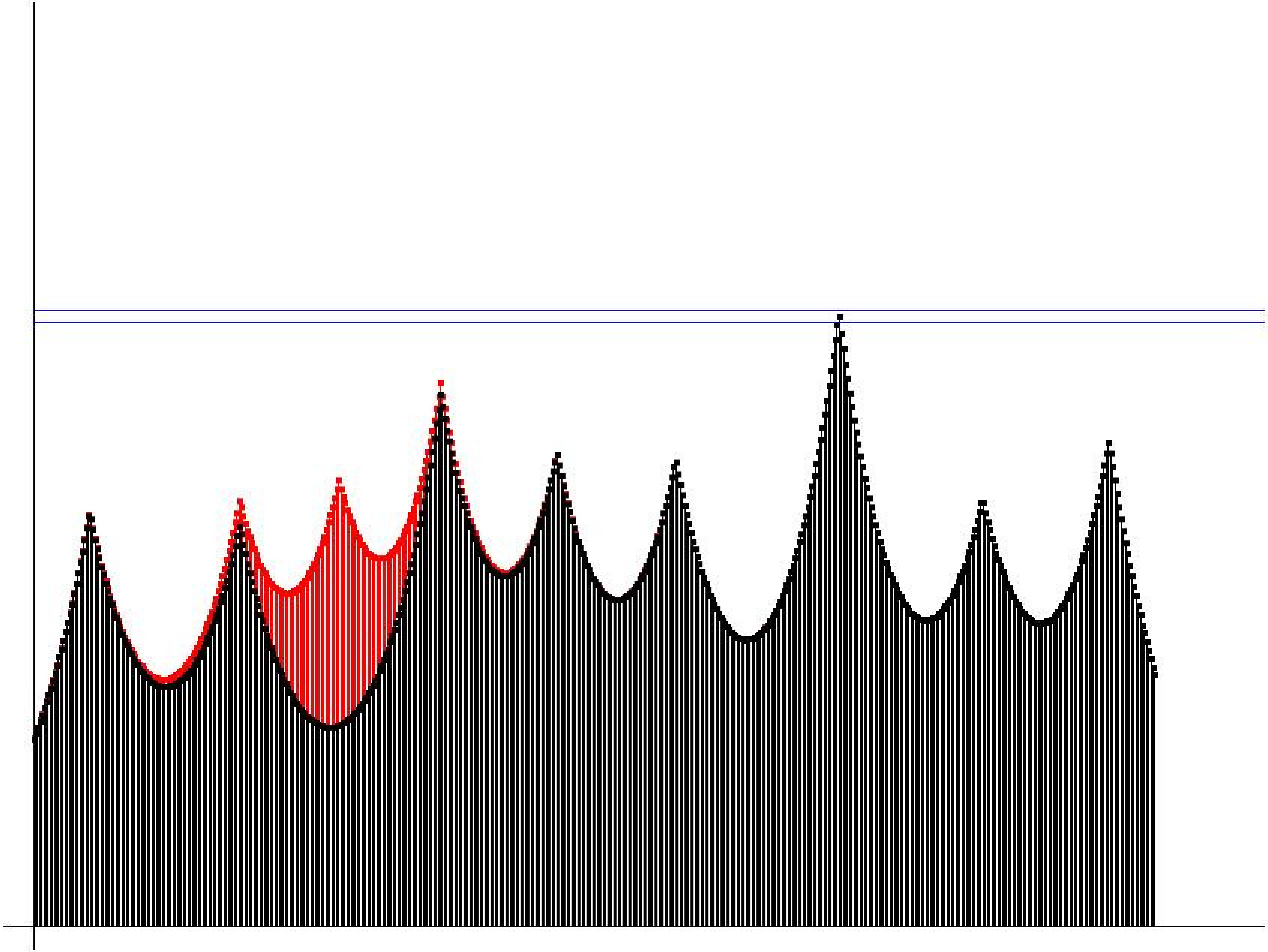
Markers: 4



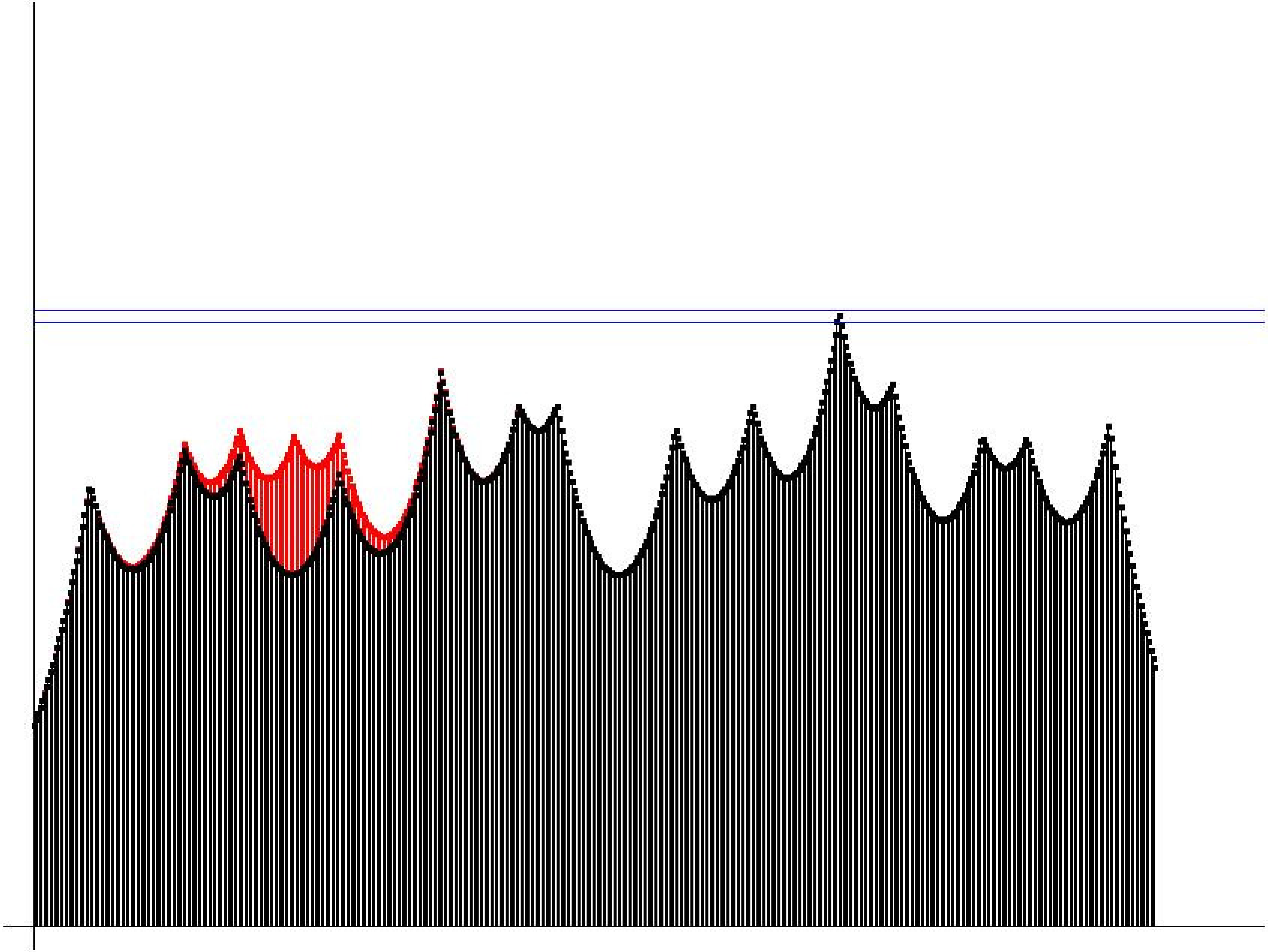
Markers: 5



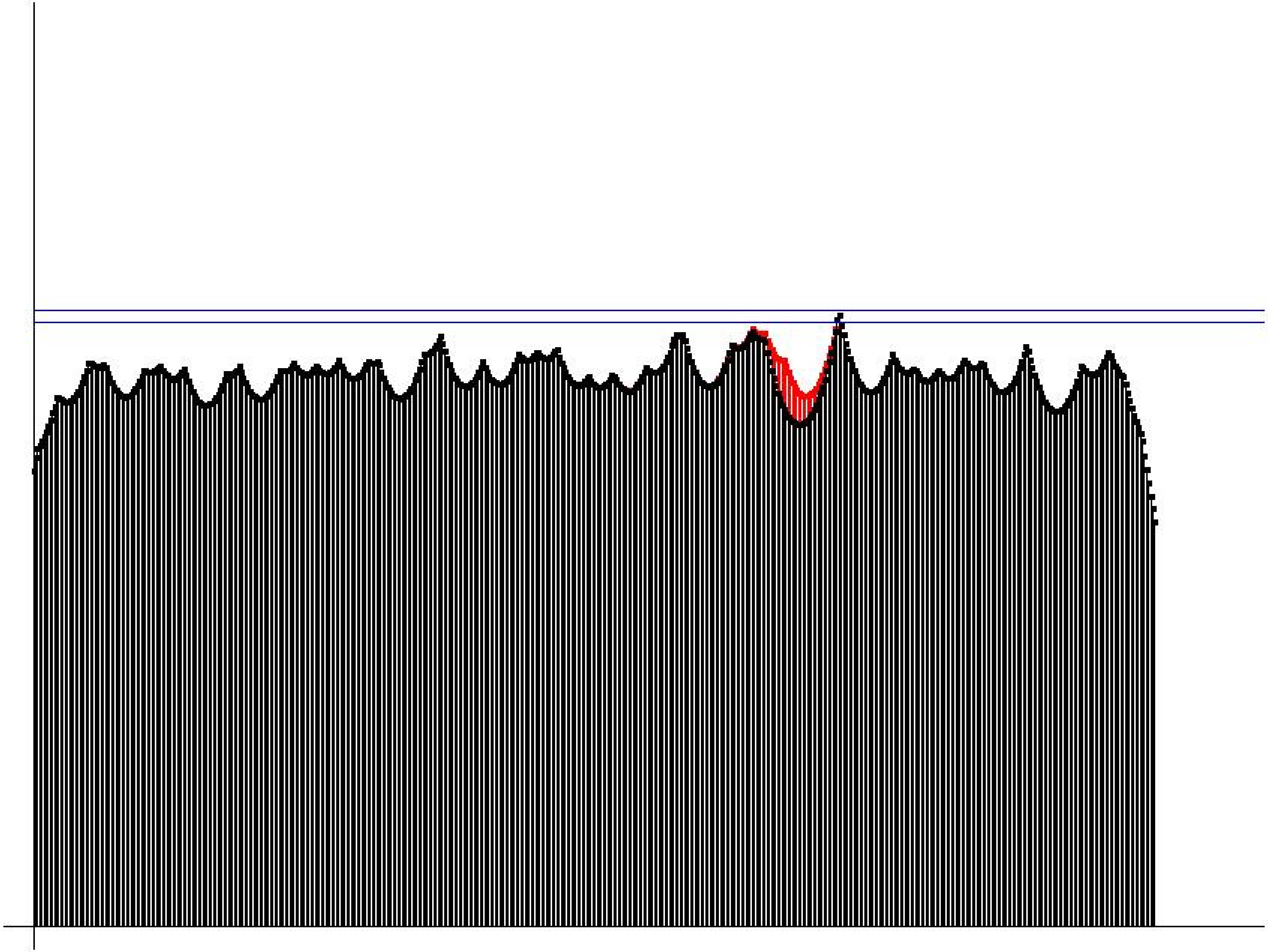
Markers: 10



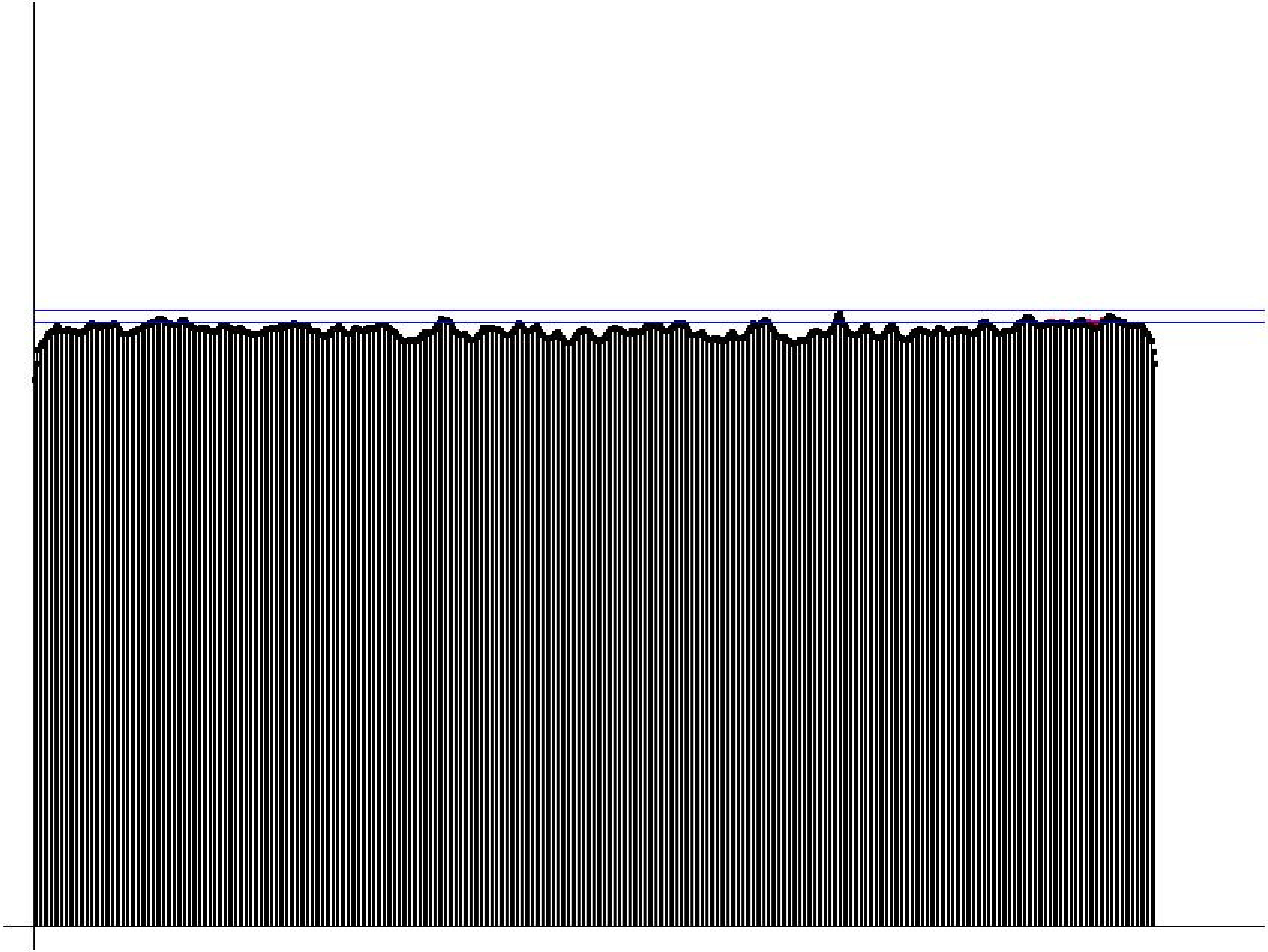
Markers: 16

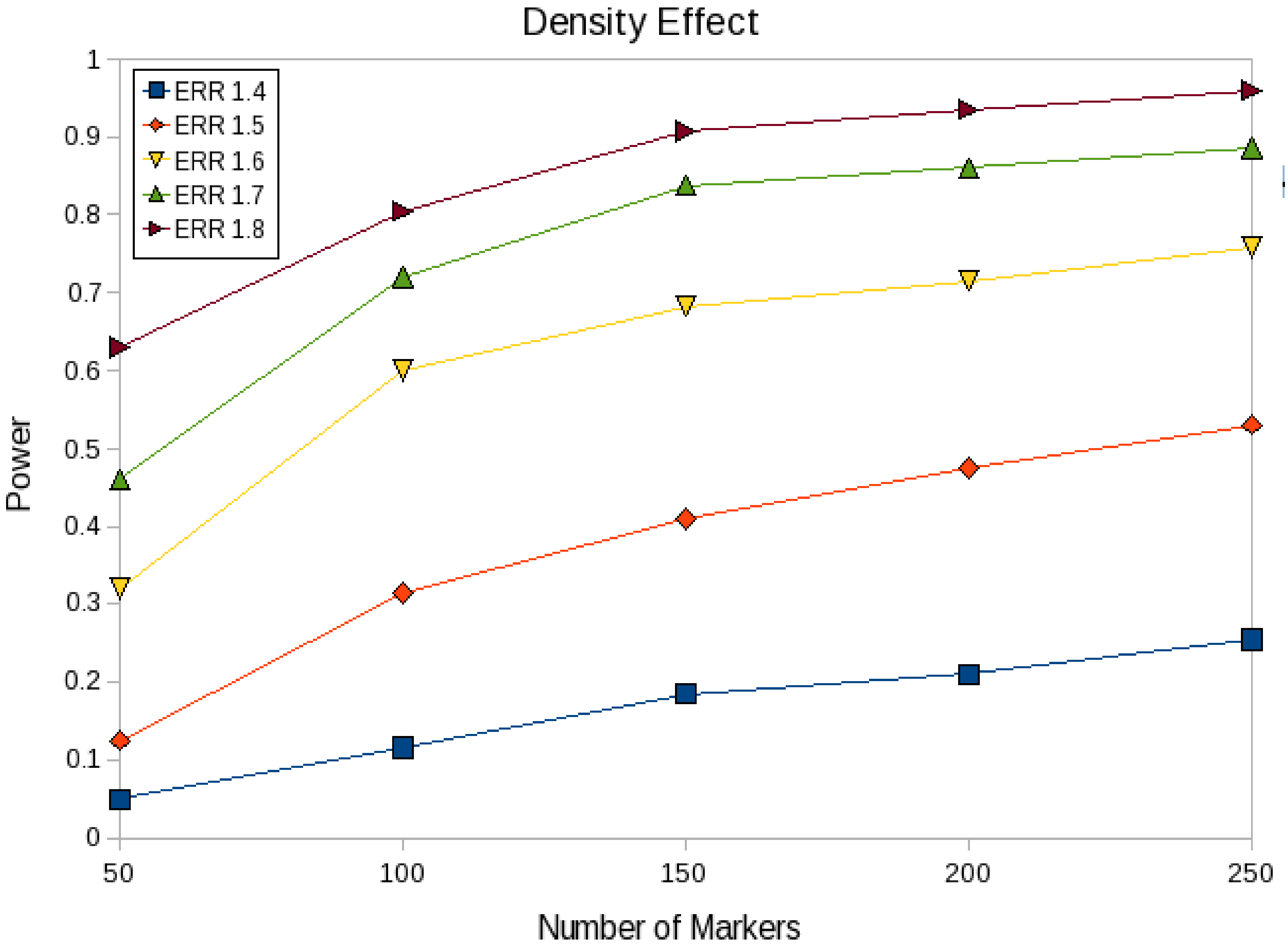


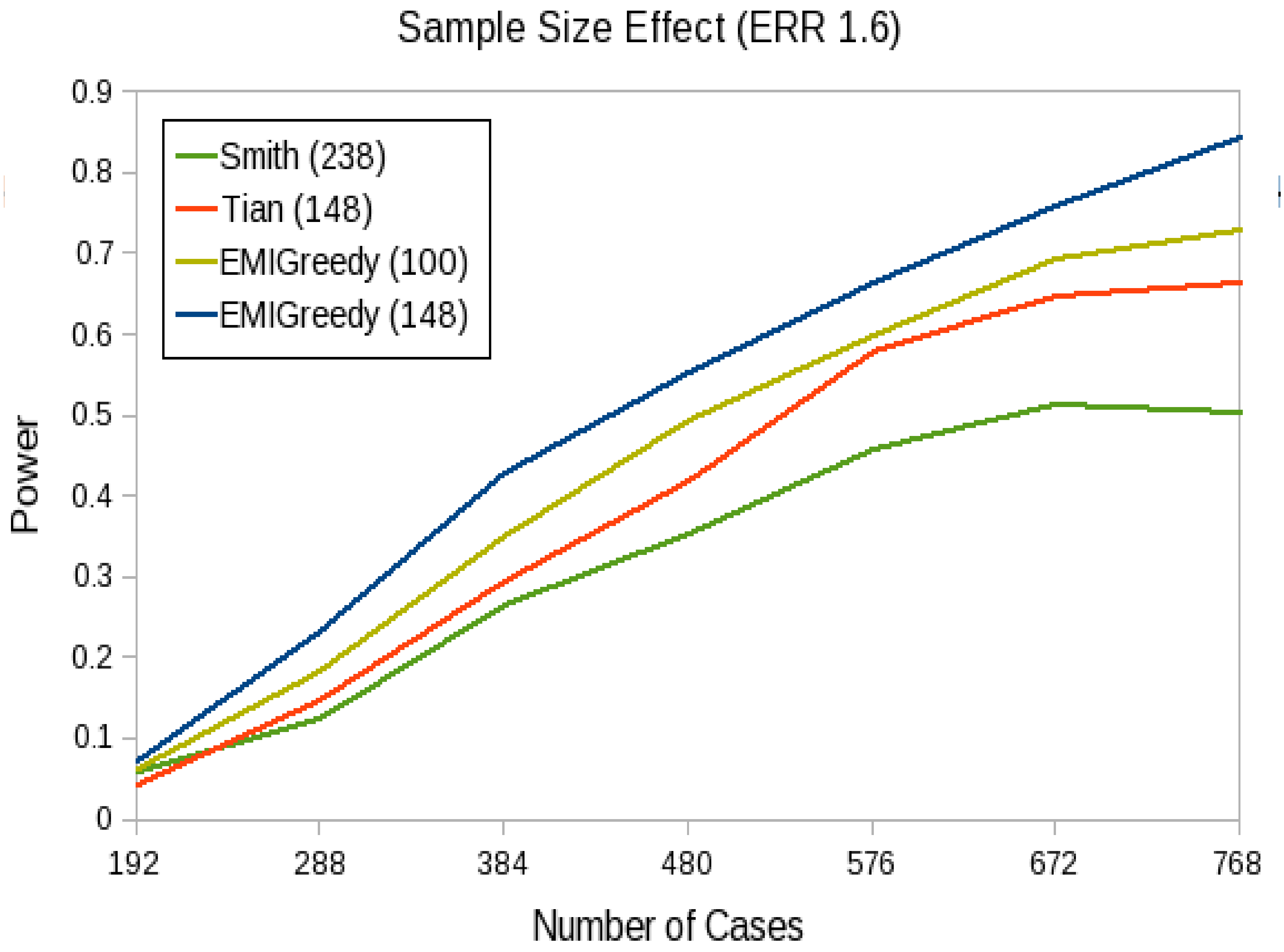
Markers: 43

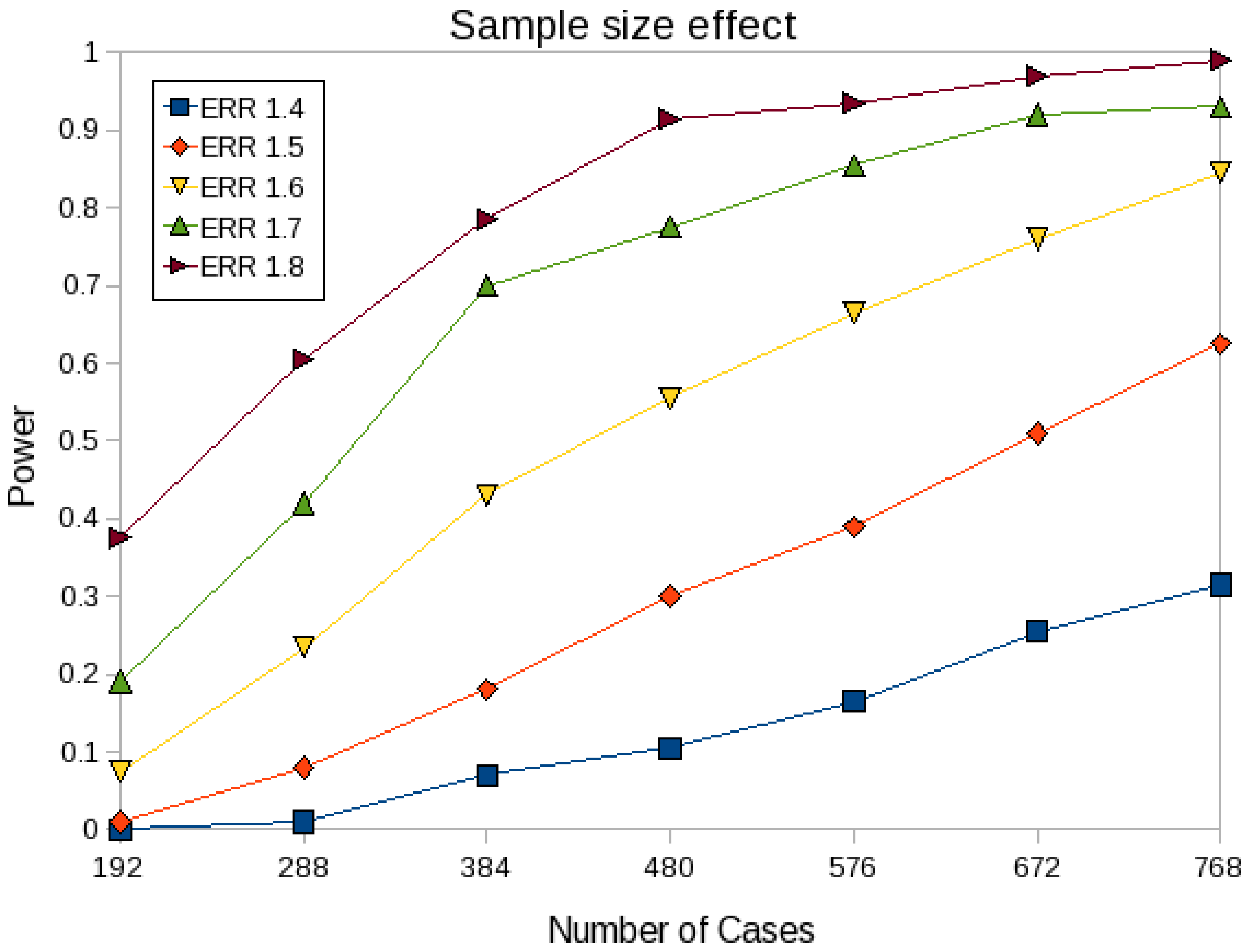


Markers: 130




Density Effect
ERR 1.4
ERR 1.5
ERR 1.6
ERR 1.7
ERR 1.8
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Number of Markers


Sample Size Effect (ERR 1.6)
Smith (238)
Tian (148)
EMIGreedy (100)
EMIGreedy (148)
Power
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
192
288
384
480
576
672
768
Number of Cases


Sample size effect
ERR 1.4
ERR 1.5
ERR 1.6
ERR 1.7
ERR 1.8
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
192
288
384
480
576
672
768
Number of Cases