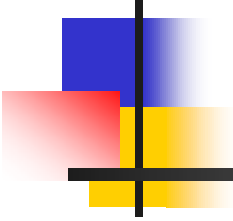


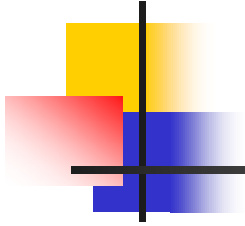
Genotype Error Detection using Hidden Markov Models of Haplotype Diversity



Justin Kennedy, Ion Mandoiu, Bogdan Pasaniuc

CSE Department, University of Connecticut

Outline



n Introduction

n Likelihood Sensitivity Approach to Error Detection

n HMM-Based Algorithms

n Experimental Results

n Conclusion

Genotyping Errors



- n A real problem despite advances in genotyping technology
 - n [Zaitlen et al. 2005] found 1.1% inconsistencies among the 20 million dbSNP genotypes typed multiple times

- n Error types
 - n Systematic errors (e.g., assay failure) detected by departure from HWE [Hosking et al. 2004]
 - n For pedigree data some errors detected as Mendelian Inconsistencies (MIs)
 - n Undetected errors
 - n E.g., if mother/father/child are all heterozygous, any error is Mendelian consistent
 - n Only ~30% detectable as MIs for trios [Gordon et al. 1999]



Effects of Undetected Genotyping Errors

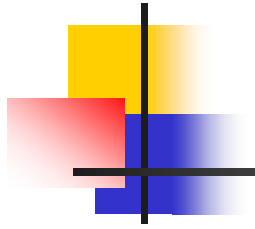
- n Even low error levels can have large effects for some study designs (e.g. rare alleles, haplotype-based)
- n Errors as low as .1% can increase Type I error rates in haplotype sharing transmission disequilibrium test (HS-TDT) [Knapp&Becker04]
- n 1% errors decrease power by 10-50% for linkage, and by 5-20% for association [Douglas et al. 00, Abecasis et al. 01]

Related Work



- n Improved genotype calling algorithms
 - n [Di et al. 05, Rabbee&Speed 06, Nicolae et al. 06]
- n Explicit modeling in analysis methods
 - n [Kruglyak et al 96, Sieberts et al. 01, Sobel et al. 02, Abecasis et al. 02]
 - n Computationally complex
- n Separate error detection step
 - n [Douglas et al. 00, Becker et al. 06]
 - n Detected errors can be retyped, imputed, or ignored in downstream analyses

Outline



n Introduction

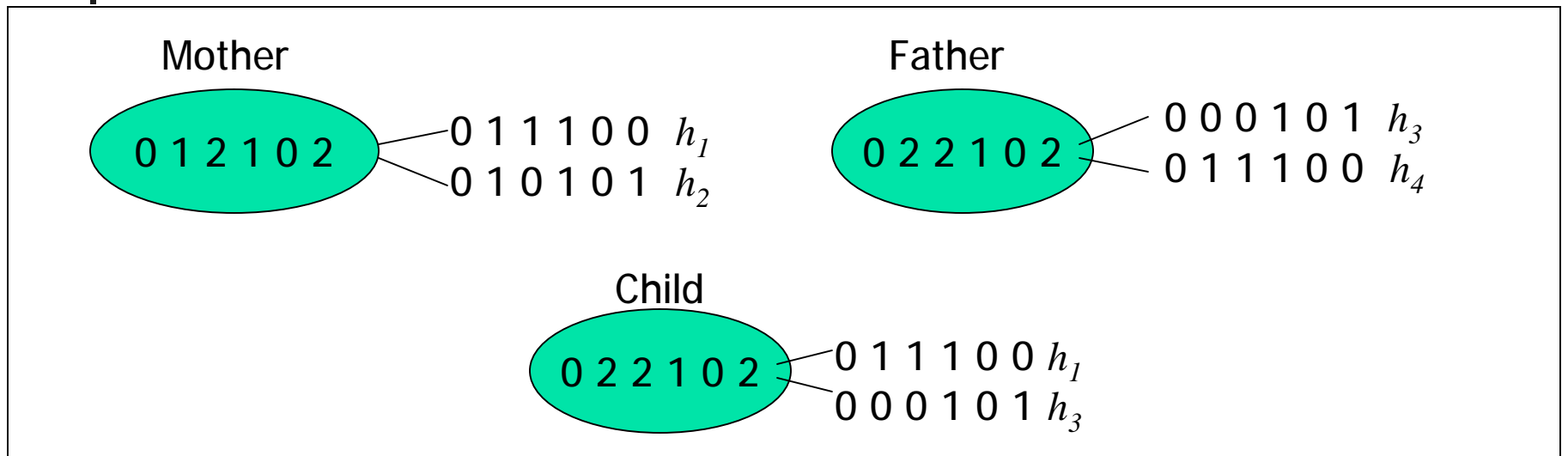
n Likelihood Sensitivity Approach to Error Detection

n HMM-Based Algorithms

n Experimental Results

n Conclusion

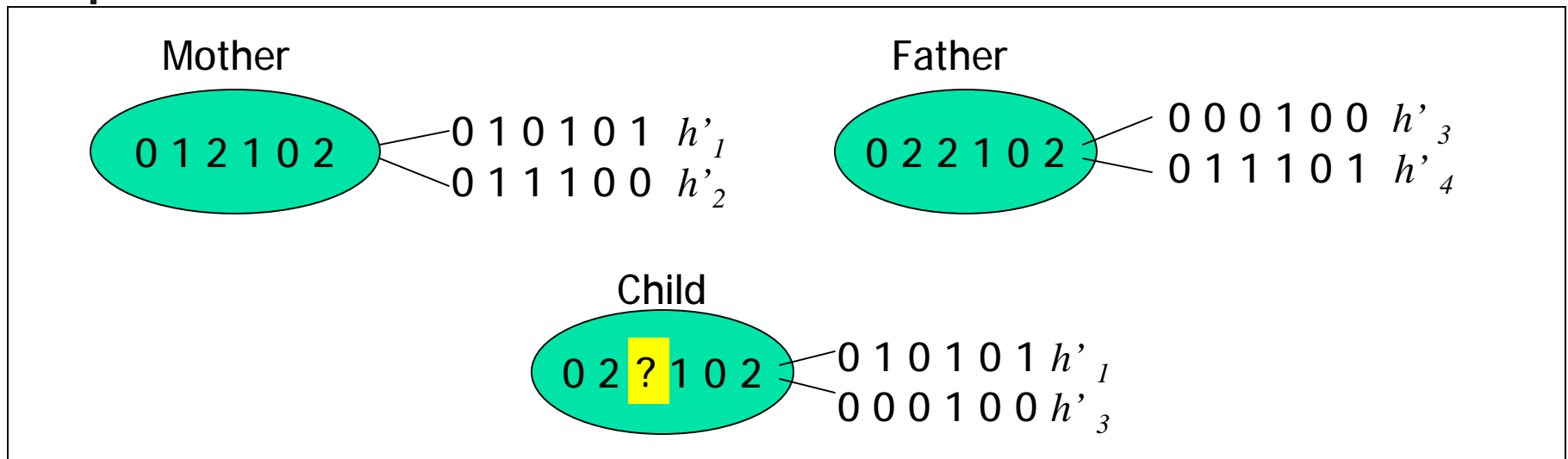
Likelihood Sensitivity Approach to Error Detection [Becker et al. 06]



$$L(T) = \text{MAX } p(h_1)p(h_2)p(h_3)p(h_4)$$

Likelihood of best phasing for original trio T

Likelihood Sensitivity Approach to Error Detection [Becker et al. 06]



$$L(T) = \text{MAX } p(h_1)p(h_2)p(h_3)p(h_4)$$

Likelihood of best phasing for original trio T

$$L(T') = \text{MAX } p(h'_1)p(h'_2)p(h'_3)p(h'_4)$$

Likelihood of best phasing for modified trio T'

Likelihood Sensitivity Approach to Error Detection [Becker et al. 06]

Mother

0 1 2 1 0 2

Father

0 2 2 1 0 2

Child

0 2 ? 1 0 2

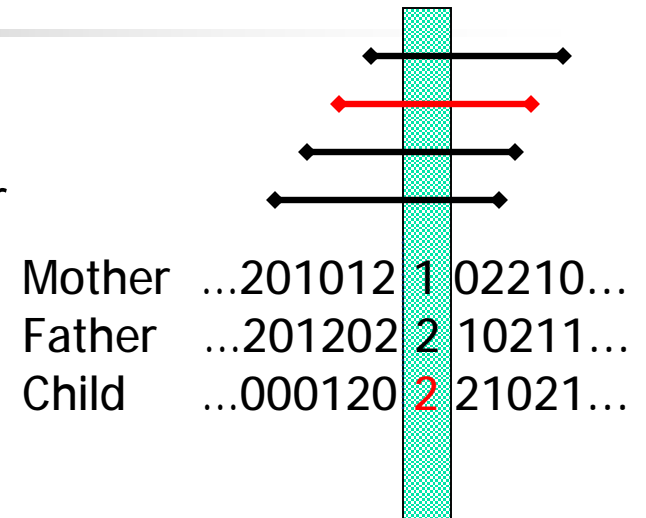
§ Large change in likelihood suggests likely error

§ Flag genotype as an error if $L(T')/L(T) > R$, where R is the detection threshold (e.g., $R=10^4$)

Implementation in FAMHAP [Becker et al. 06]

n Window-based algorithm

- n For each window including the SNP under test, generate list of H most frequent haplotypes (default H=50)
- n Find most likely trio phasings by pruned search over the H^4 quadruples of frequent haplotypes
- n Flag genotype as an error if $L(T')/L(T) > R$ for *at least one window*

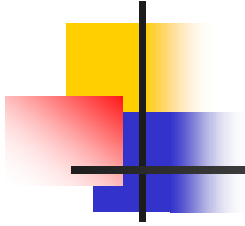




Limitations of FAMHAP Implementation

- n Truncating the list of haplotypes to size H may lead to sub-optimal phasings and inaccurate $L(T)$ values
- n False positives caused by nearby errors (due to the use of multiple short windows)
- n **Our approach:**
 - n HMM model of haplotype diversity è all haplotypes are represented + no need for short windows
 - n Alternate likelihood functions è scalable runtime

Outline



n Introduction

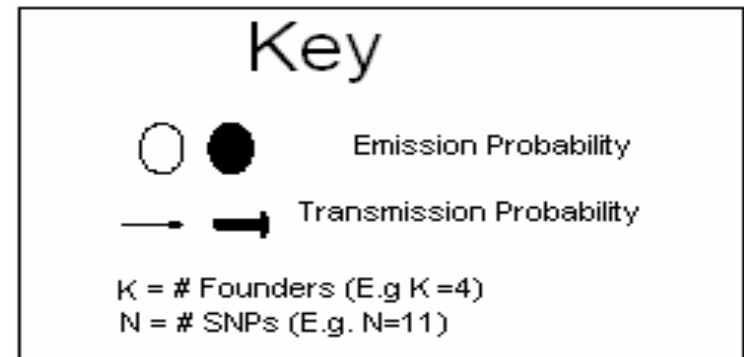
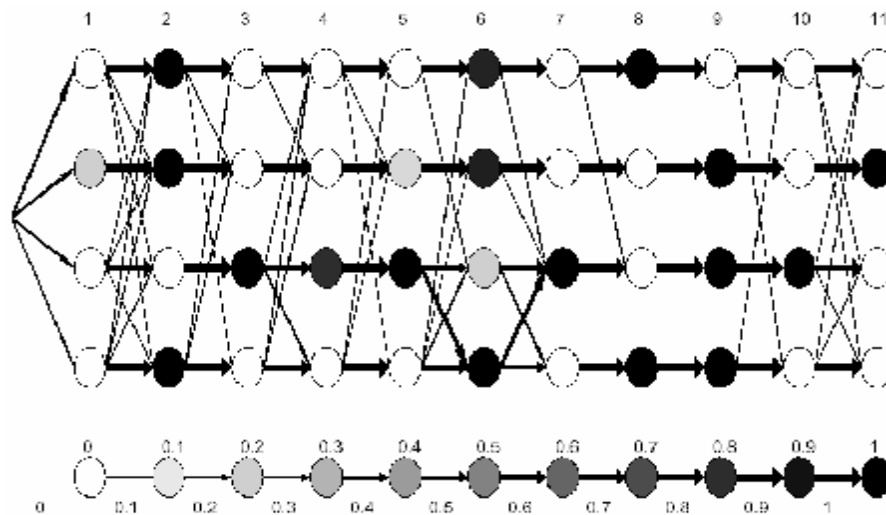
n Likelihood Sensitivity Approach to Error Detection

n HMM-Based Algorithms

n Experimental Results

n Conclusion

HMM Model



- n Similar to models proposed by [Schwartz 04, Rastas et al. 05, Kimmel&Shamir 05]
- n Block-free model, paths with high transition probability correspond to “founder” haplotypes

HMM Training



- n Previous works use EM training of HMM based on **unrelated** genotype data
- n Our 2-step algorithm exploits **pedigree info**
 - n Step 1: Infer haplotypes using pedigree-aware algorithm based on entropy-minimization
 - n Step 2: train HMM based on inferred haplotypes, using Baum-Welch

Alternate Likelihood Functions

- Maximum phasing probability $L(T) = \text{MAX } p(h_1)p(h_2)p(h_3)p(h_4)$ is hard to compute
- We use alternate likelihood functions that are monotonic under data deletion & efficiently computable:
 - **Viterbi probability (ViterbiProb)**: the maximum probability of a set of 4 HMM paths that emit 4 haplotypes compatible with the trio
 - **Probability of Viterbi Haplotypes (ViterbiHaps)**: product of total probabilities of the 4 Viterbi haplotypes
 - **Total Trio Probability (TotalProb)**: total probability $P(T)$ that the HMM emits four haplotypes that explain trio T along all possible 4-tuples of paths

Efficient Computation of Viterbi Probability

- n For a fixed trio, Viterbi paths can be found using a 4-path version of Viterbi's algorithm in $O(NK^8)$ time
- n K^3 speed-up by factoring common terms:

$$Pre_1(j; q_1, q'_2, q'_3, q'_4) = \max_{q'_1 \in Q_j} \{V_f(j; (q'_1, q'_2, q'_3, q'_4))\gamma(q'_1, q_1)\}$$

$$Pre_2(j; q_1, q_2, q'_3, q'_4) = \max_{q'_2 \in Q_j} \{Pre_1(j; (q_1, q'_2, q'_3, q'_4))\gamma(q'_2, q_2)\}$$

$$Pre_3(j; q_1, q_2, q_3, q'_4) = \max_{q'_3 \in Q_j} \{Pre_2(j; (q_1, q_2, q'_3, q'_4))\gamma(q'_3, q_3)\}.$$

$$V(j+1; q_1, q_2, q_3, q_4) = E(j+1; q_1, q_2, q_3, q_4) \max_{q'_4 \in Q_j} \{Pre_3(j; q_1, q_2, q_3, q'_4)g(q'_4, q_4)\}$$

Where:

- $E(j+1; q_1, q_2, q_3, q_4)$ = maximum probability of emitting SNP genotypes at locus $j+1$ from states (q_1, q_2, q_3, q_4)
- γ = transition probability

Overall Runtimes

n Viterbi probability

- n Likelihoods of all $3N$ modified trios can be computed within $O(NK^5)$ time using forward-backward algorithm
- n Overall runtime for M trios $O(MNK^5)$

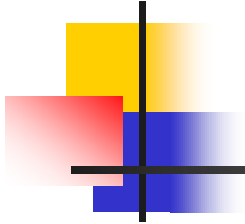
n Probability of Viterbi haplotypes

- n Obtain haplotypes from standard traceback, then compute haplotype probabilities using forward algorithms
- n Overall runtime $O(M(NK^5 + N^2K))$

n Total trio probability

- n Similar pre-computation speed-up & forward-backward algorithm
- n Overall runtime $O(MNK^5)$

Outline



n Introduction

n Likelihood Sensitivity Approach to Error Detection

n HMM-Based Algorithms

n Experimental Results

n Conclusion

Datasets



- n Real dataset [Becker et al. 2006]
 - n 35 SNPs per individual genotype sequence
 - n 551 trios
- n Synthetic datasets
 - n 35 SNPs, 30-551 trios
 - n Preserved missing data pattern of real dataset
 - n Haplotypes assigned to trios based on frequencies inferred from real dataset
 - n 1% error rate, four error insertion models
 - n Random allele
 - n Random genotype
 - n Heterozygous-to-homozygous
 - n Homozygous-to-heterozygous

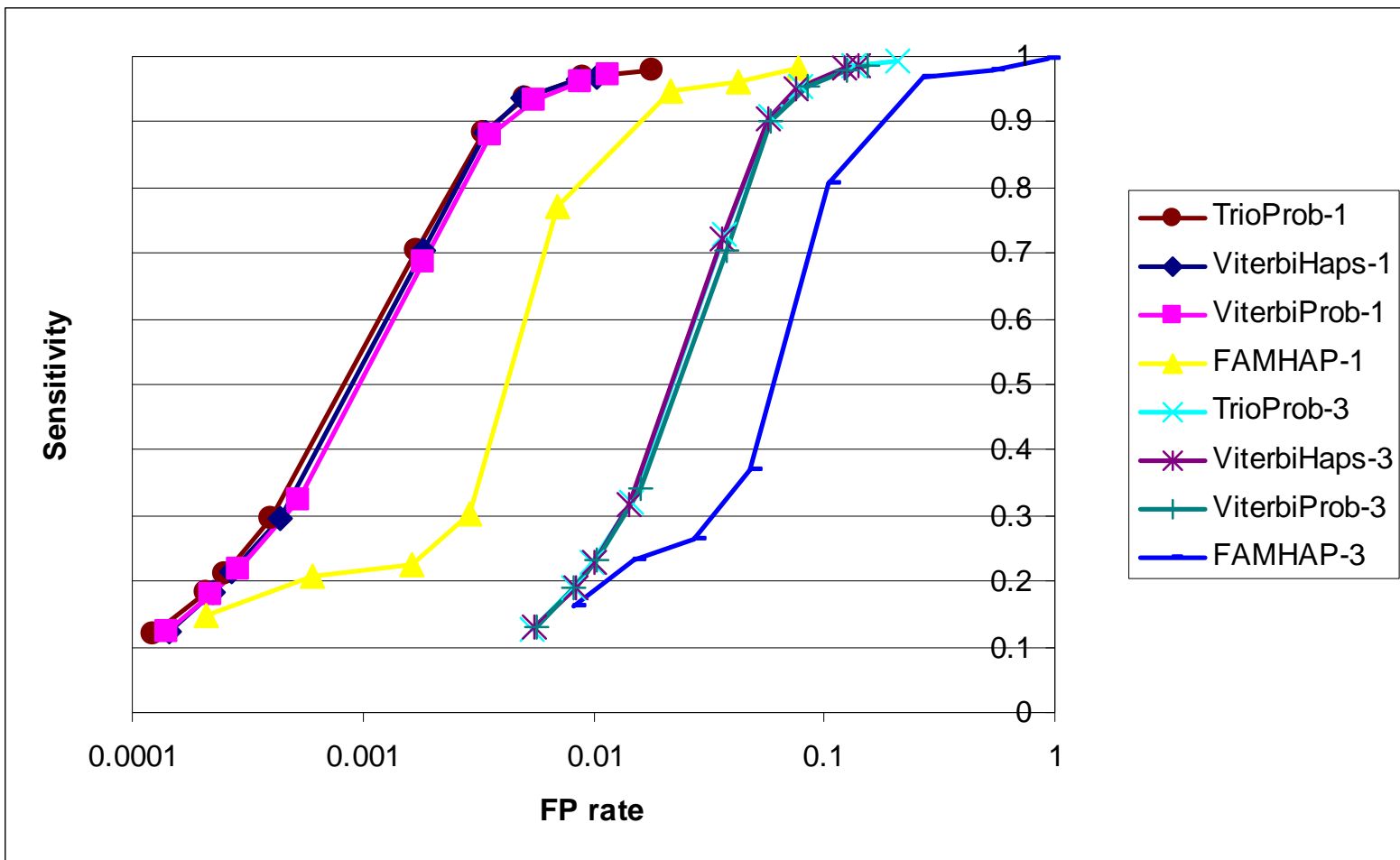
Experimental Setup



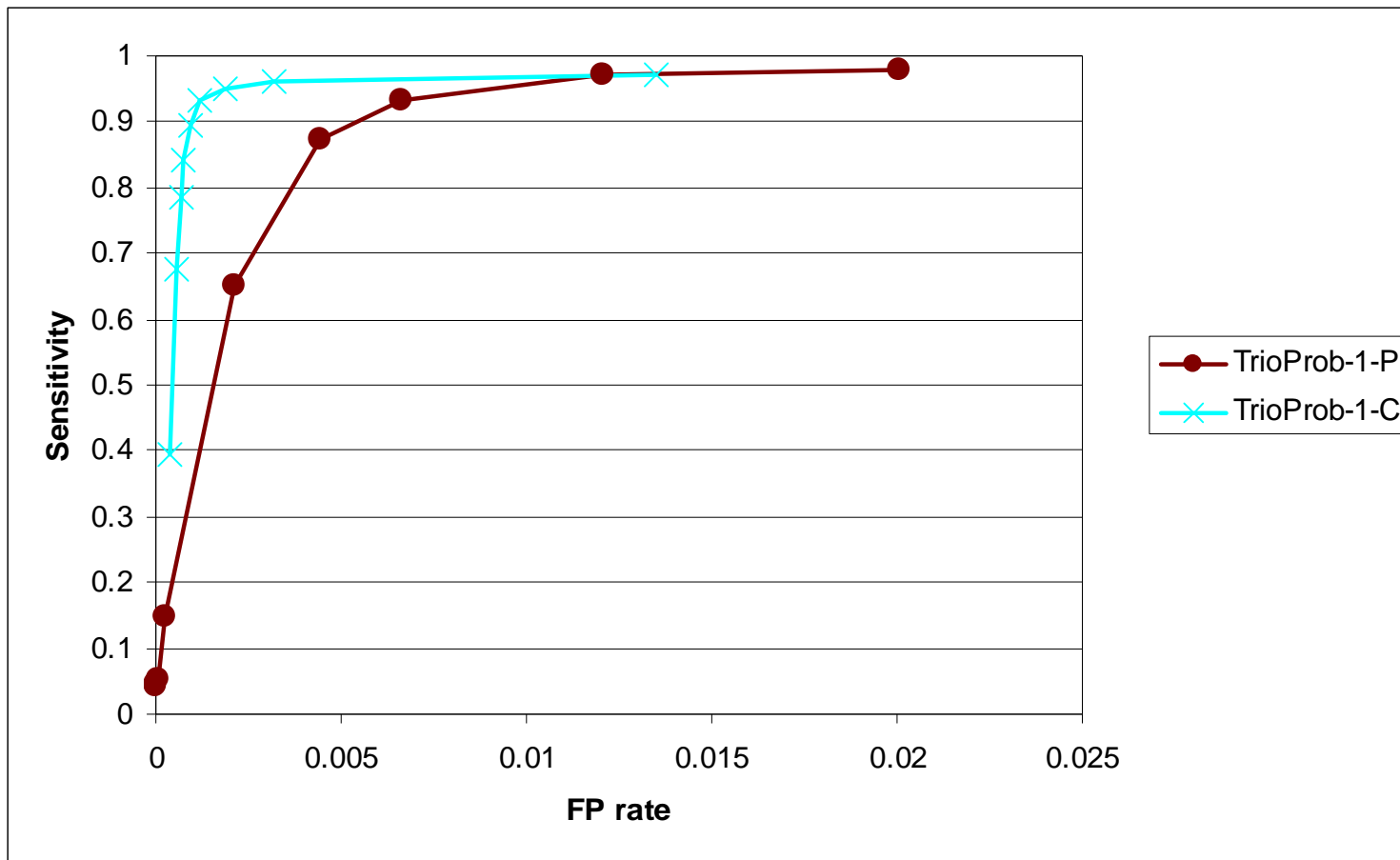
- n Two strategies for handling MIs
 - n Set all three individuals to unknown prior to error detection, or
 - n Set child only to unknown (preserving parents' original data)

- n Two testing strategies
 - n Test one SNP genotype: ViterbiProb-1, ViterbiHaps-1, TotalProb-1
 - n Simultaneously test three SNP genotypes at the same locus: ViterbiProb-3, ViterbiHaps-3, TotalProb-3

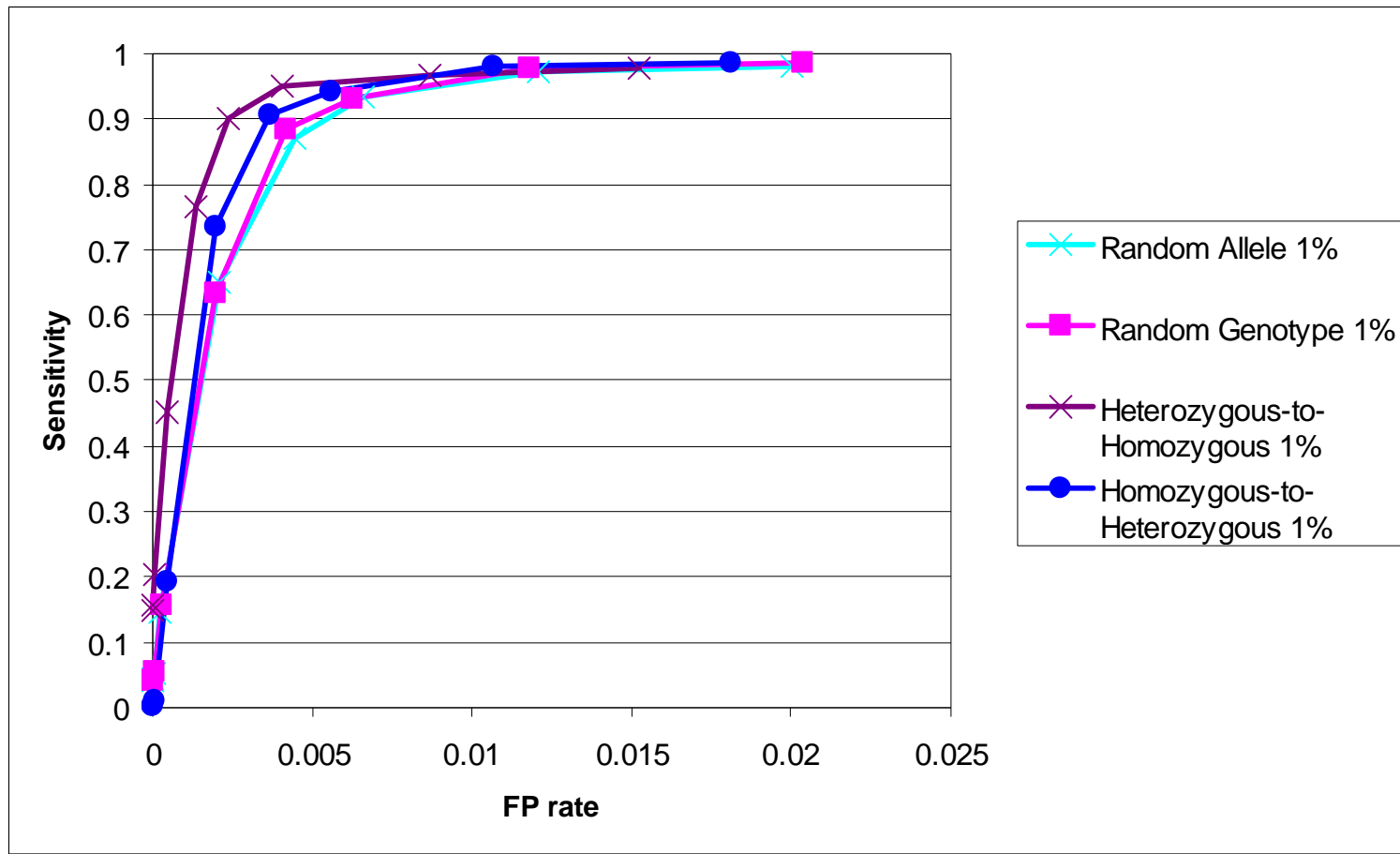
Comparison with FAMHAP (Random Allele Errors)



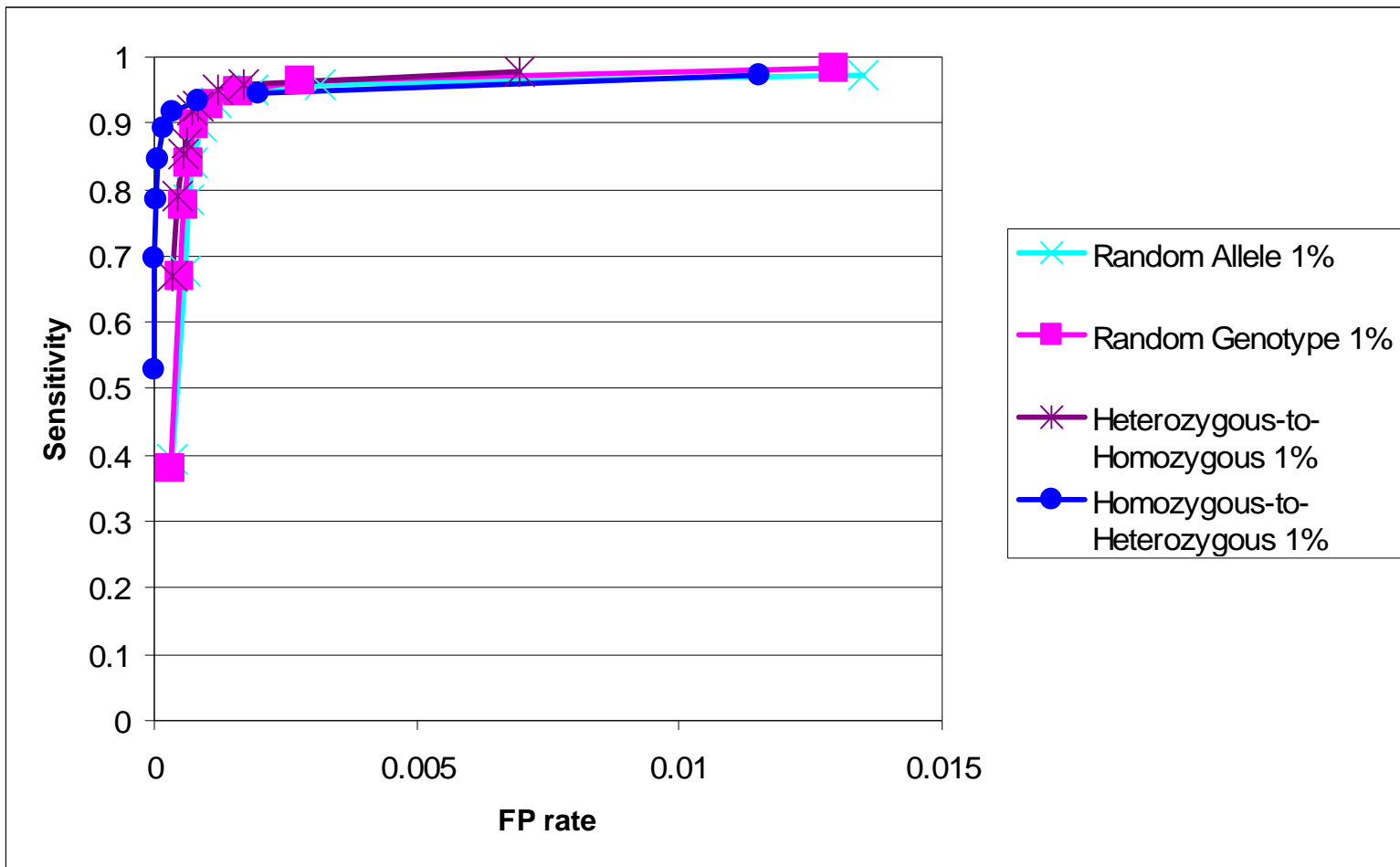
Children vs. Parents (Random Allele Errors)



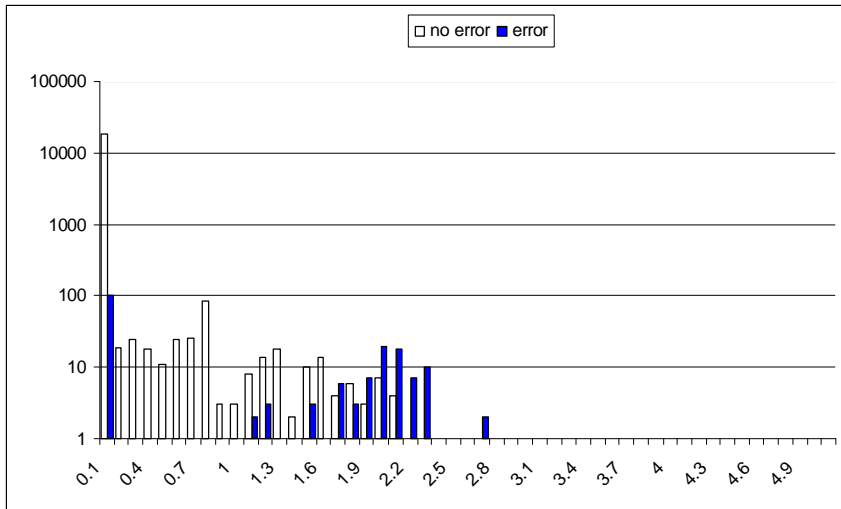
Error Model Comparison (TrioProb-1 Parents)



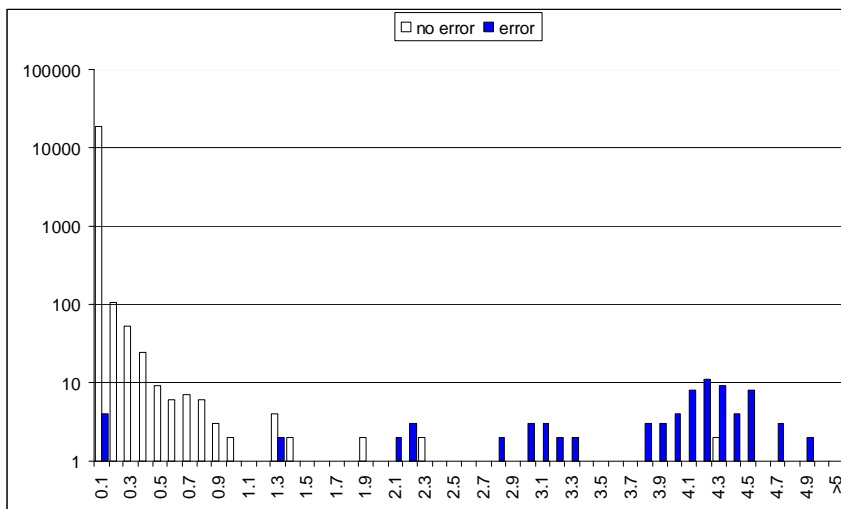
Error Model Comparison (TrioProb-1 Children)



Unrelated vs. Trio Likelihood Sensitivity

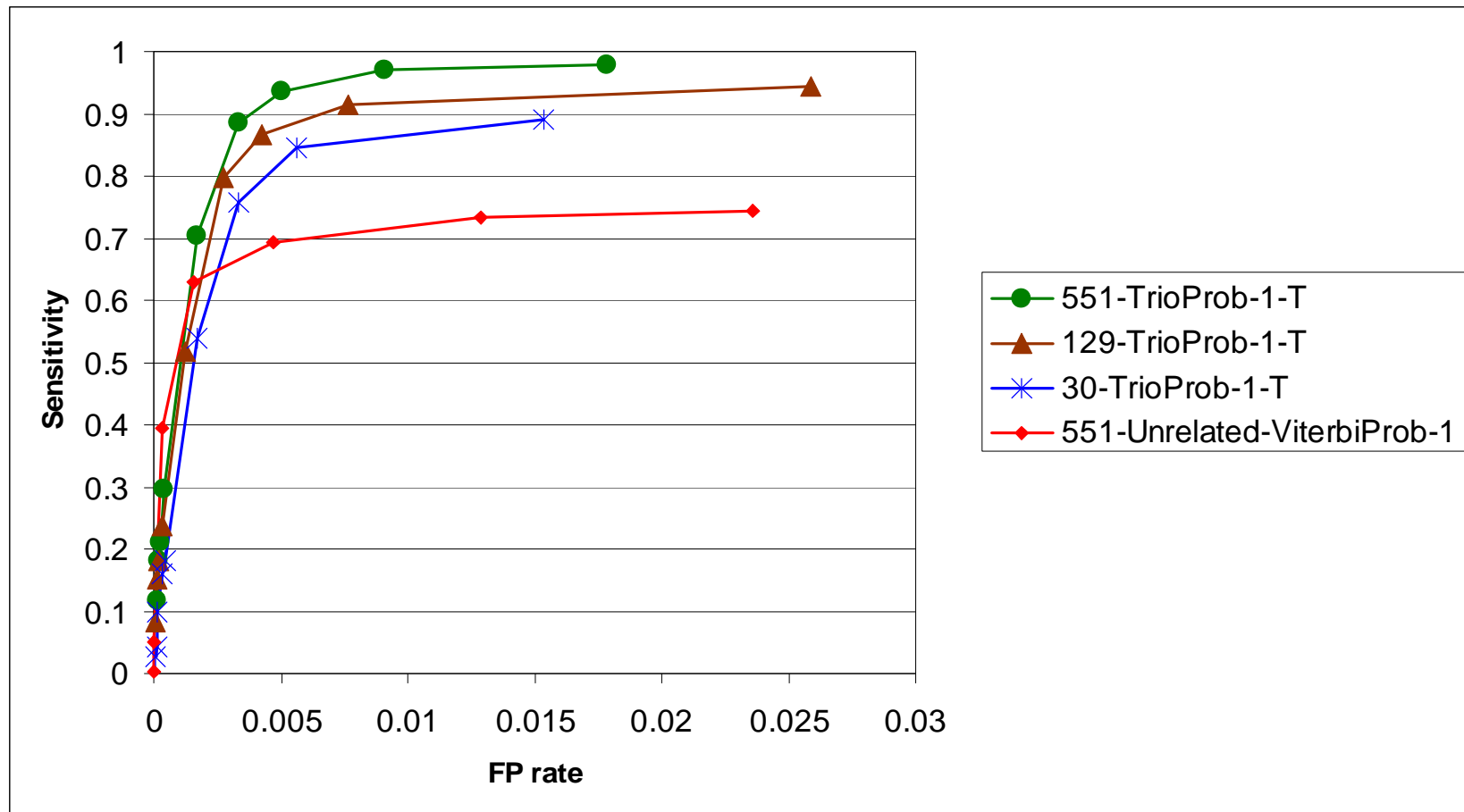


Unrelated ViterbiProb-1
Likelihood ratios (children)



Trio ViterbiProb-1
Likelihood ratios (children)

Pedigree Info vs. Sample Size Effect



TrioProb-1 Results on Real Dataset

Threshold	Total Signals			True Positives			False Positives			Unknown		
	2	3	4	2	3	4	2	3	4	2	3	4
Parents	80	15	9	9	9	8	2	1	1	69	5	0
Children	27	21	17	11	10	10	3	3	1	13	8	6
Total	107	36	26	20	19	18	5	4	2	82	13	6

- n [Becker et al. 06] resequenced all trio members at 41 loci flagged by FAMHAP-3
 - n 23 SNP genotypes were identified as true errors
 - n $41 \times 3 - 23 = 100$ resequenced SNP genotypes agree with original calls
 - n Predictive value for $R=10^4$ is between $18/26=69\%$ and $24/26=92\%$, compared to $23/41=56\%$ for FAMHAP-3

Outline



n Introduction

n Likelihood Sensitivity Approach to Error Detection

n HMM-Based Algorithms

n Experimental Results

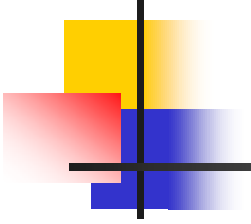
n Conclusion

Conclusion



- n We have proposed efficient methods for error detection in trio genotype data based on a HMM model of haplotype diversity
 - n Exploiting pedigree info is very useful
 - n Improved detection accuracy compared to FAMHAP
 - n Runtime linear in #SNPs and #trios
- n Ongoing work
 - n Improve detection accuracy via iteration
 - n Fix MIs using likelihood before error detection
 - n Correct errors with high likelihood ratio, then recompute likelihood ratios (possibly after re-phasing and HMM re-training)
 - n Integration with genotype calling algorithms
 - n Combine low level intensity data with haplotype-based likelihoods

Questions?



Accuracy Measures



- n Sensitivity
 - n $TP/(TP+TF)$

- n Predictive value
 - n $TP/(TP+FP)$

- n Specificity
 - n $TN/(FP+TN)$

- n False Positive rate
 - n $1-\text{Specificity}$