

Algorithms for Association Mapping of Complex Diseases With Ancestral Recombination Graphs

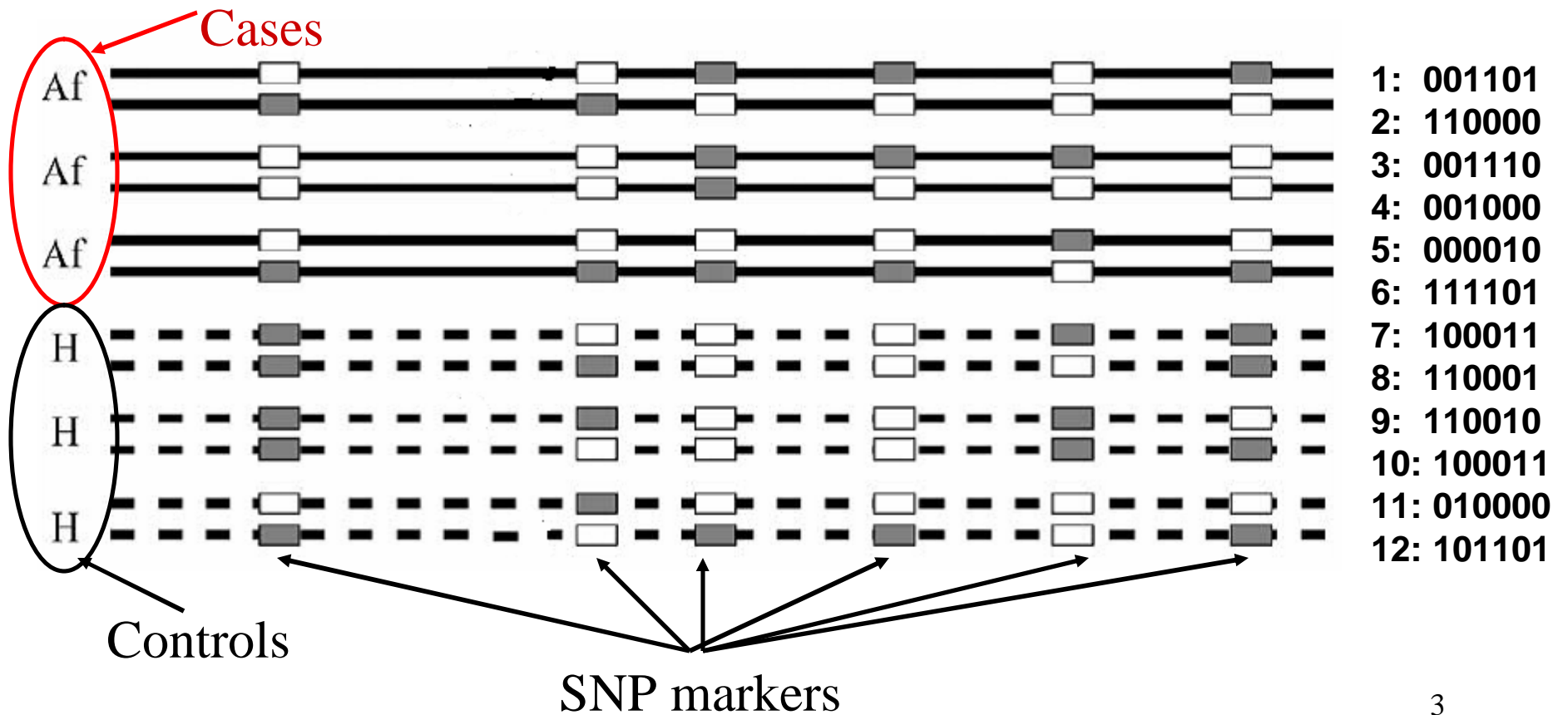
Yufeng Wu
UC Davis

RECOMB Satellite Workshop, 2007

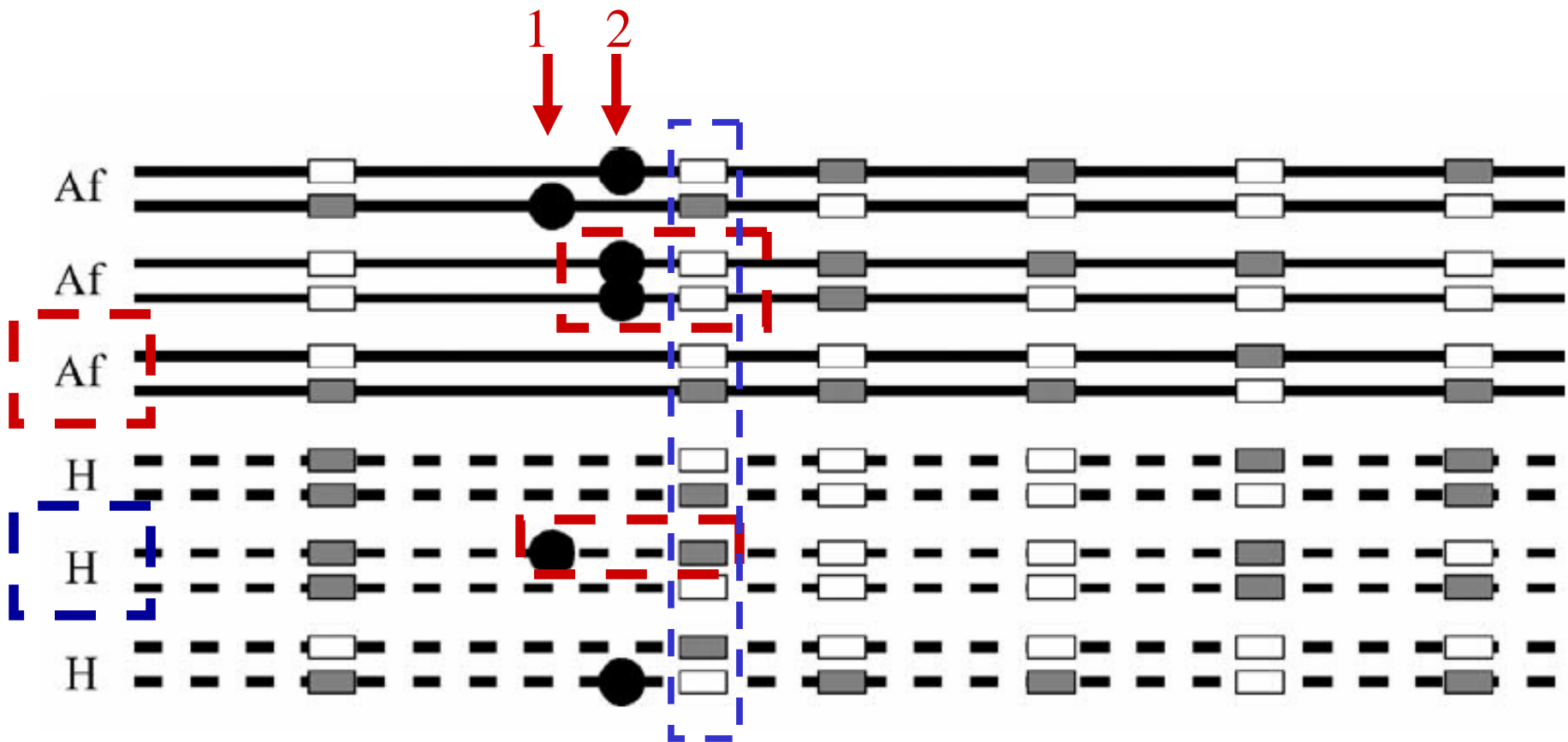
Association (or LD) Mapping

- Given a *subset* of SNPs from *unrelated* individuals, find **unobserved** genetic variations that strongly discriminate individuals with the trait (**cases**) and those without the trait (**controls**)
- Complex Diseases: difficult to map

Illustration (Zollner and Pritchard, Genetics, 2005)



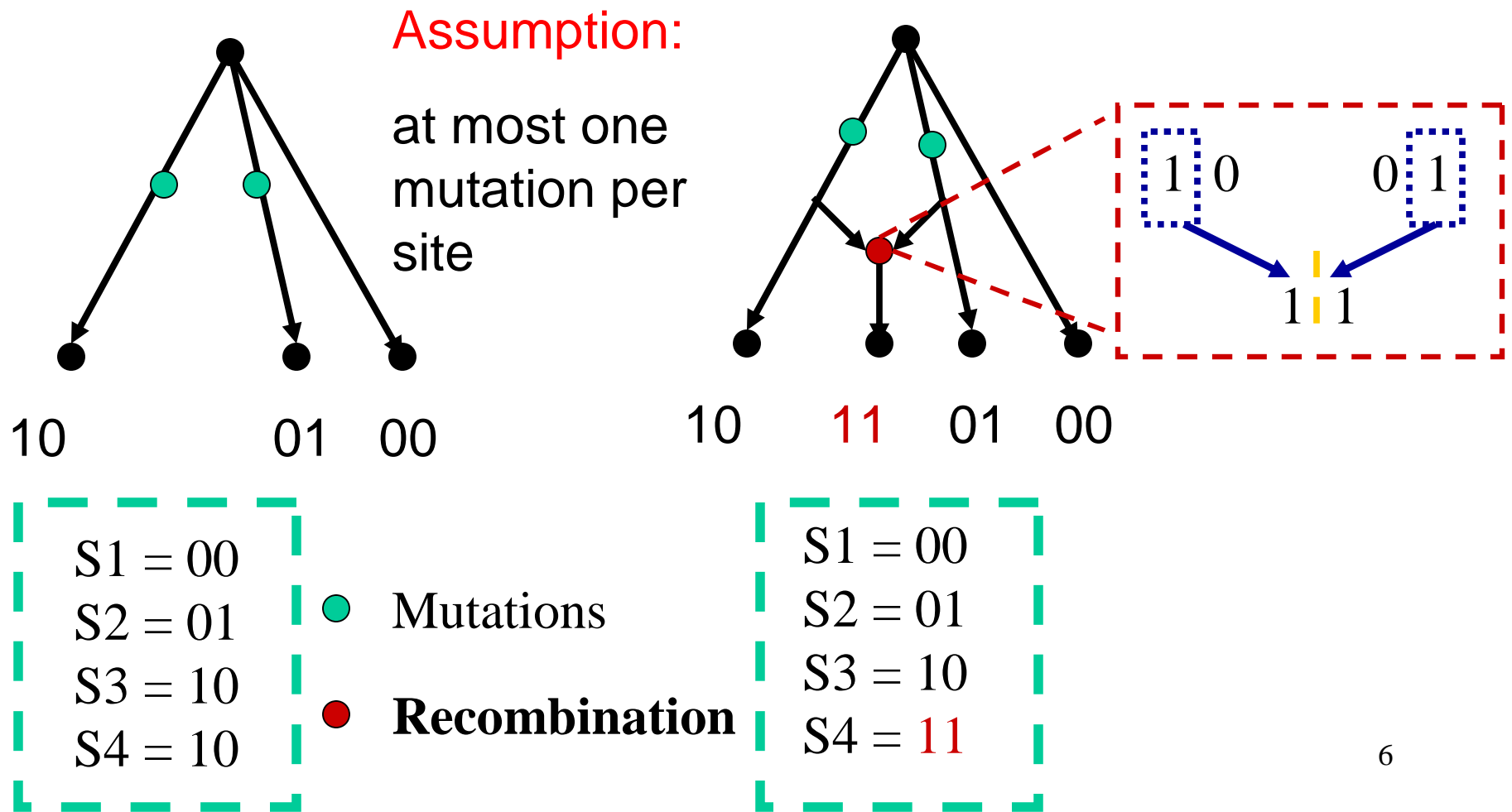
Some Challenges in Association Mapping



The Genealogy Approach

- “..the *best* information that we could possibly get about association is to know the full coalescent **genealogy**...”
 - Zollner and Pritchard
- Goal: **infer** genealogy from marker data with **recombination**
 - Approximation (e.g. in Zollner and Pritchard)

Ancestral Recombination Graph (ARG)



Full-ARG Approaches

- First full ARG mapping method (Minichiello and Durbin)
 - Use full *plausible* ARG, but heuristic
 - Disease model is somewhat less complex than Zollner and Pritchard's (and also of this work)
- Our results
 - Sampling **full** ARGs with provable property, and work on more complex disease model

What Type of ARGs?

- Focus on parsimonious history
 - minARGs: ARGs that use the **minimum** number of recombinations (NP-hard)
 - Near minimum ARGs
- **Uniform** sampling of minARGs
 - Practical for certain range of data

Special Case: ARG with Only Input Sequences

- Self-derivability (SD) Problem: construct an ARG with **only** the input sequences
- In fact, such ARG, if exists, must be a minARG
- Runs in $O(2^n)$ time
- Heuristics to extend to non-self-derivable data

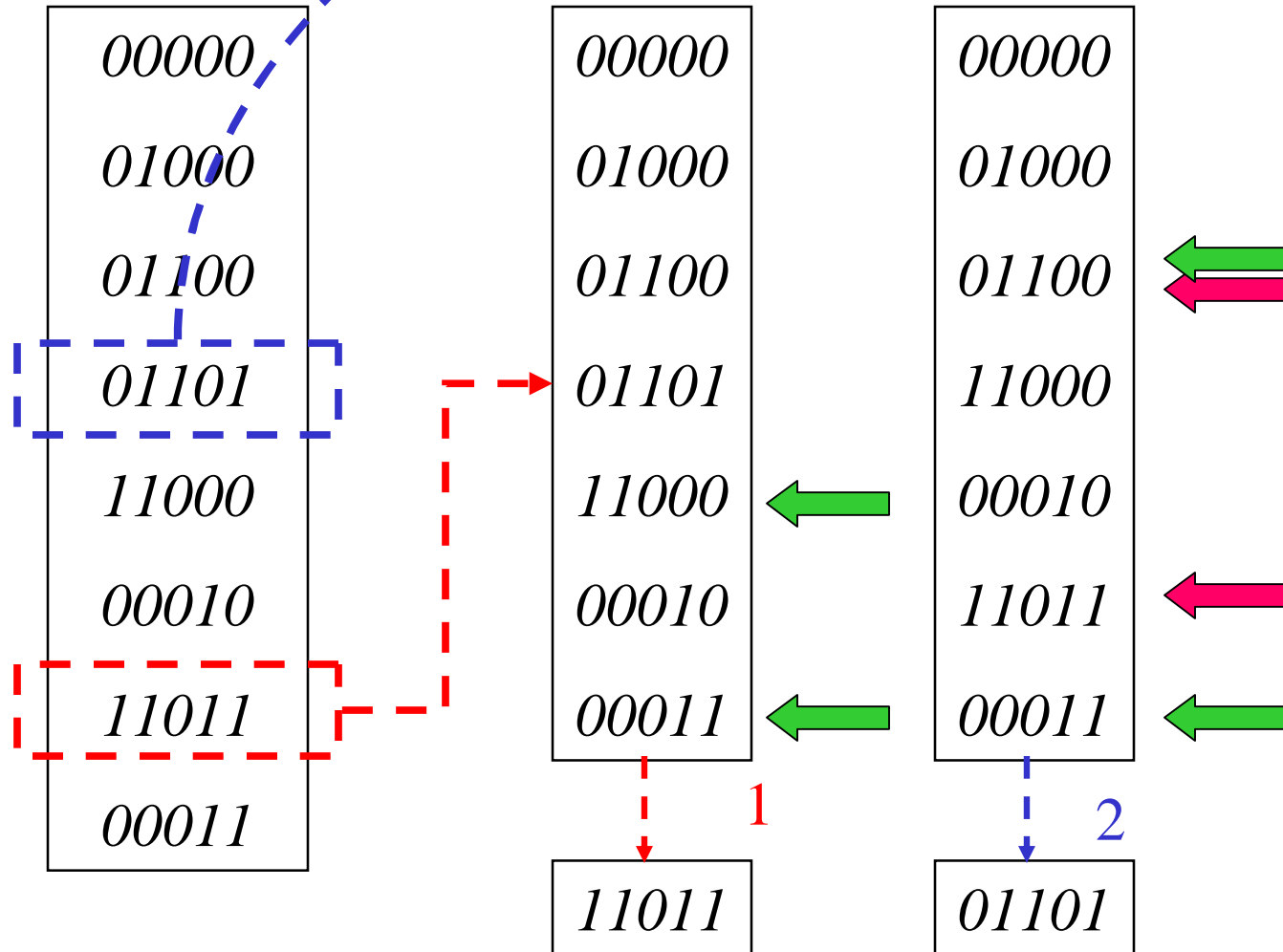
Counting Self-derived ARGs

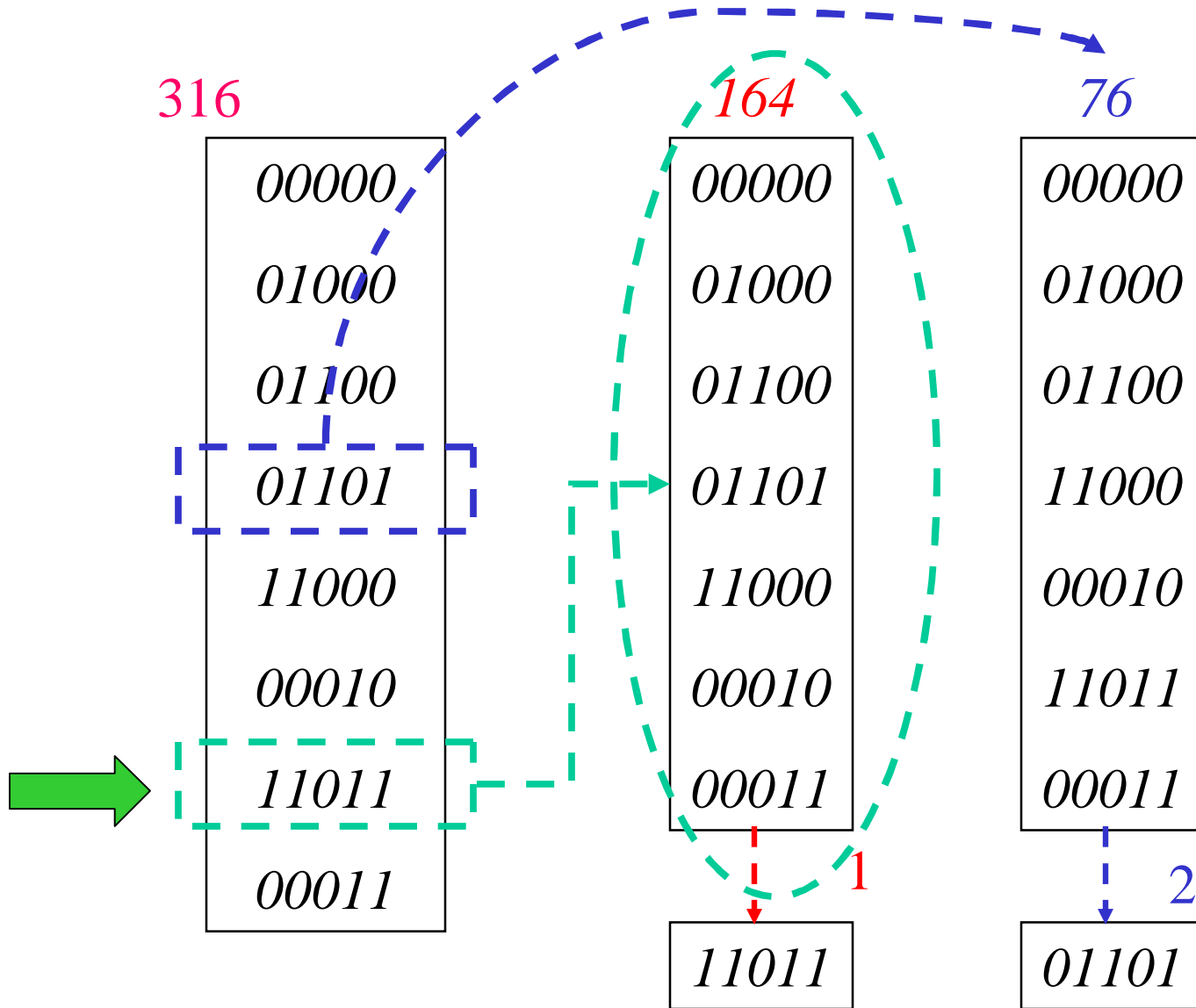
$$N = 164 * 1 + 76 * 2$$

$$= 316$$

$$N1 = 164$$

$$N2 = 76$$

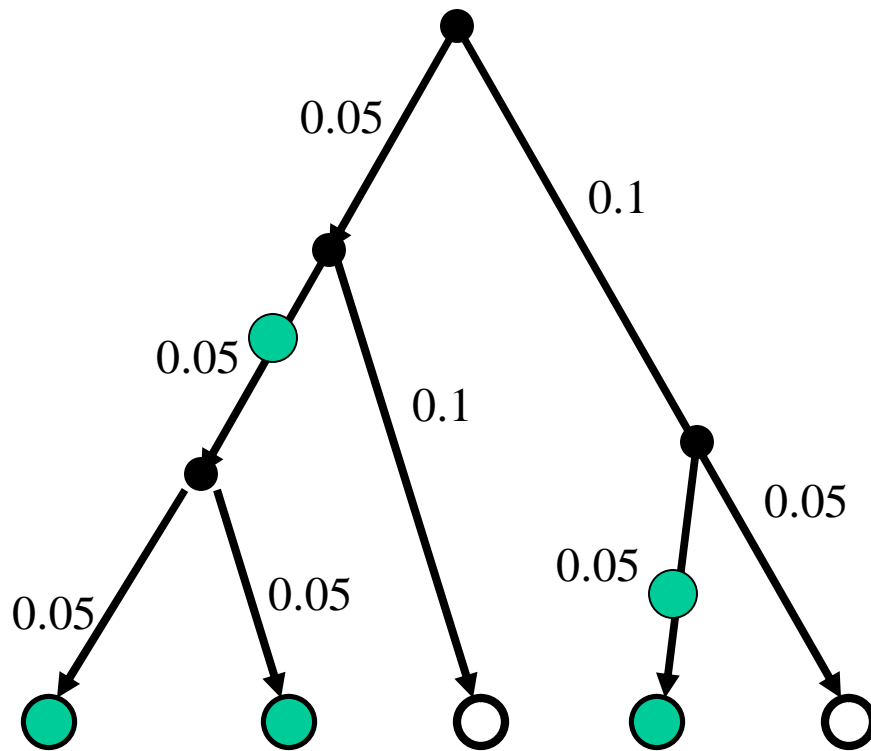




1. Random value $Rnd = 0.3 < 0.52$
2. Pick seq = 11011 as last row to derive
3. Move to reduced matrix

Select 11011 with prob = $164/316 = \mathbf{0.52}$, and 01101 with prob = $76 \cdot 2/316 = \mathbf{0.48}$

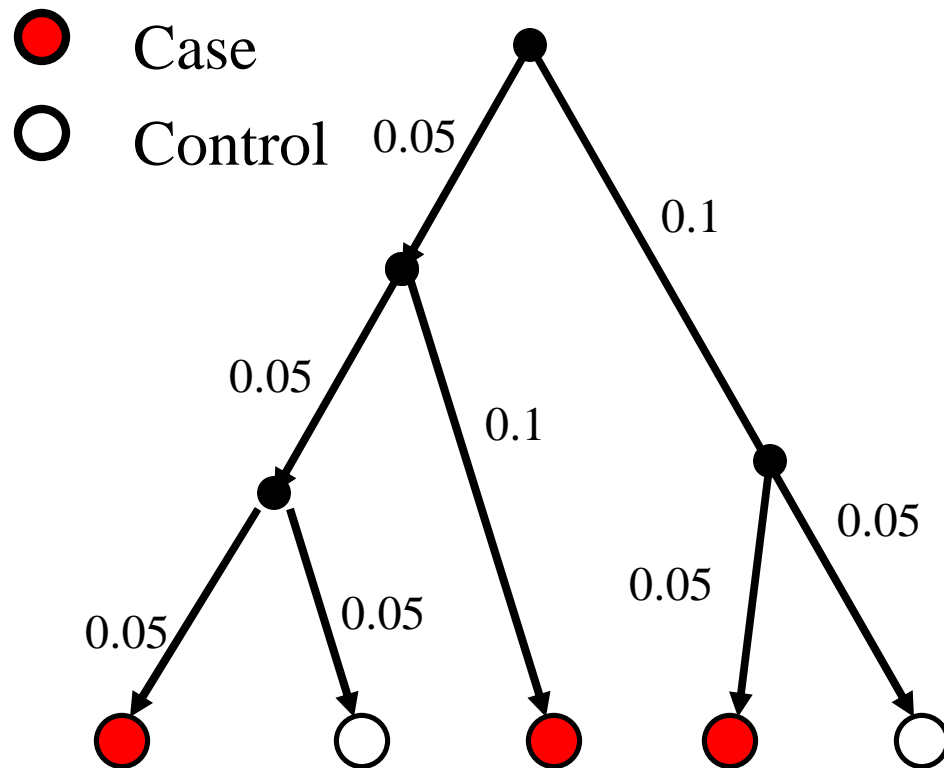
Disease Model (Zollner & Pritchard)



Disease mutations:
Poisson Process

Two alleles: wild-type and **mutant**

Disease Penetrance (Zollner & Pritchard)



$P_{A,1}$: probability of a mutant sequence becomes a **case**

$$P_{C,1} = 1.0 - P_{A,1}$$

$P_{A,0}$: probability of a wild-type sequence becomes a **case**

$$P_{C,0} = 1.0 - P_{A,0}$$

Phenotype Likelihood

(Zollner and Pritchard)

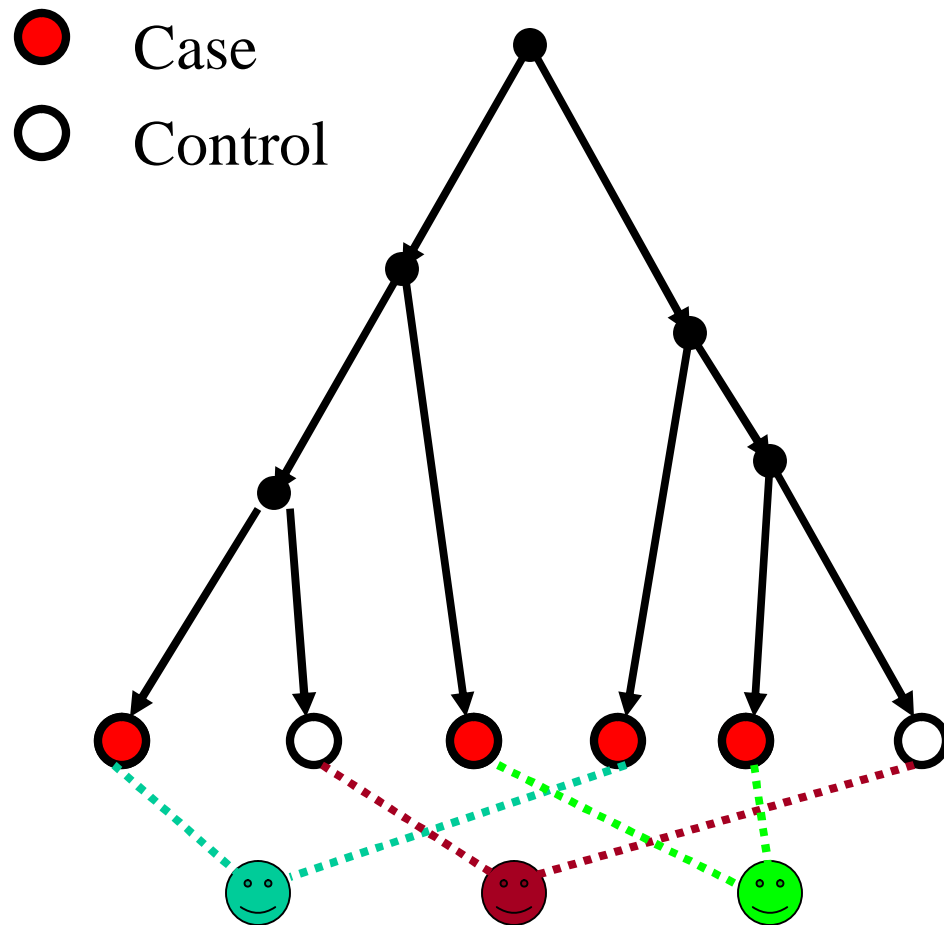
- Given a tree T_x at position x and case/control phenotype F of its leaves, what is the *probability* $\Pr(F | T_x)$ of observing F on T_x ? (Zollner & Pritchard)
 - Sum over all subset of mutated edges
- Adopted in this work

Expected Phenotype Likelihood

- Need for assessing statistical significance
- **Null model:** randomly permute case/control labels
- Our result: $O(n^3)$ algorithm for computing *expected* value of phenotype likelihood
 - For randomly permuted phenotypes

Diploid Penetrance

(Zollner and Pritchard)



Diploid: **two** sequences per individual

Diploid enetrance:

$P_{A,00}$: prob. Individual with two **wild-type** sequences becomes a **case**

$P_{A,01} : \dots$

$P_{A,11} : \dots$

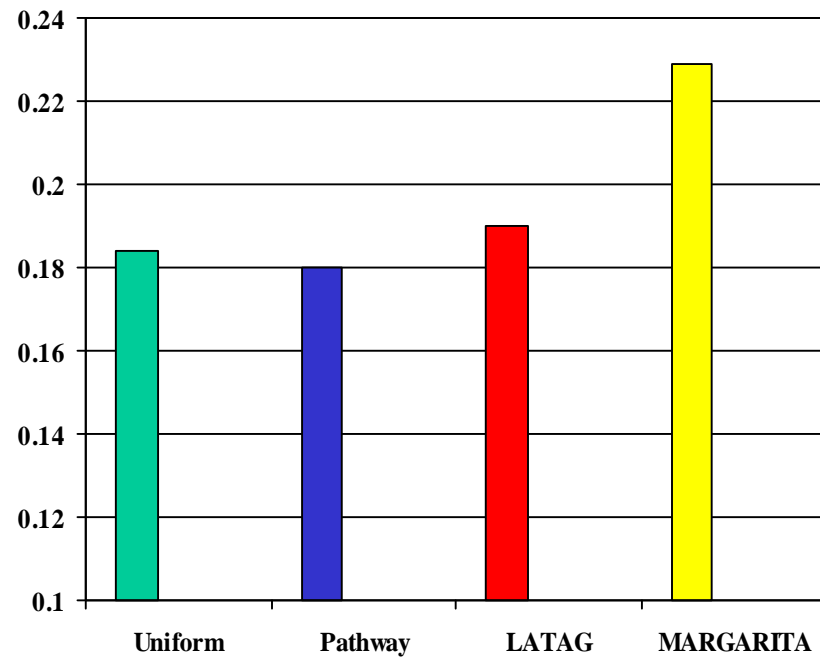
Diploid Penetrance Is Hard

(Wu, 2007)

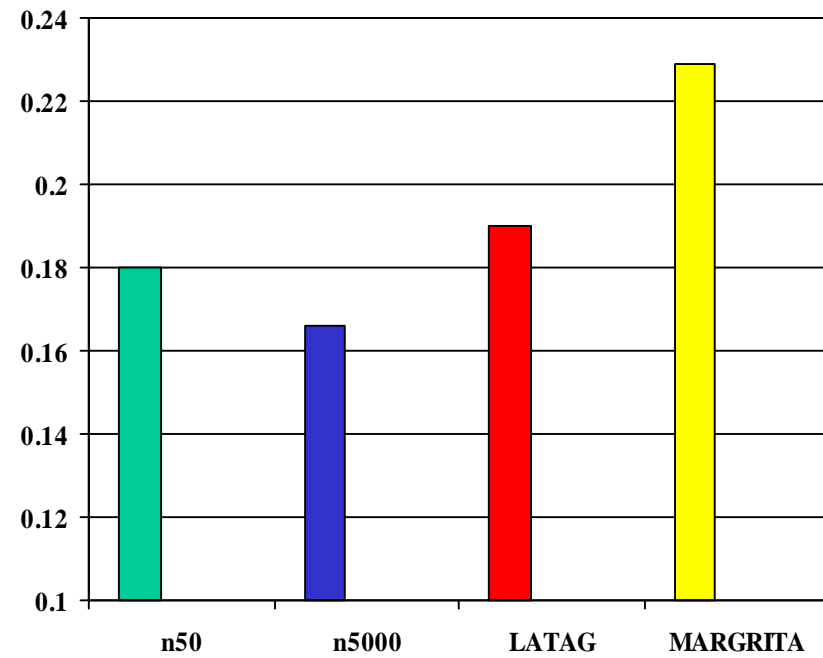
- Efficient computation of phenotype likelihood is desired
 - Stated but unresolved in Zollner and Pritchard
- Our result: computing phenotype likelihood with diploid penetrance is **NP-hard**

Simulation Results

50 ARGs per data



50/5000 ARGs per data



Comparison: TMARG (uniform), TMARG (pathway),
LATAG, MARGARITA

Acknowledgement

- Software available at:
<http://wwwcsif.cs.ucdavis.edu/~wuyu>
- I want to thank
 - Dan Gusfield
 - Dan Brown
 - Chuck Langley
 - Yun S. Song