


*Exact and algorithmic
methods for haplotype
frequency inference: what
do they tell us?*

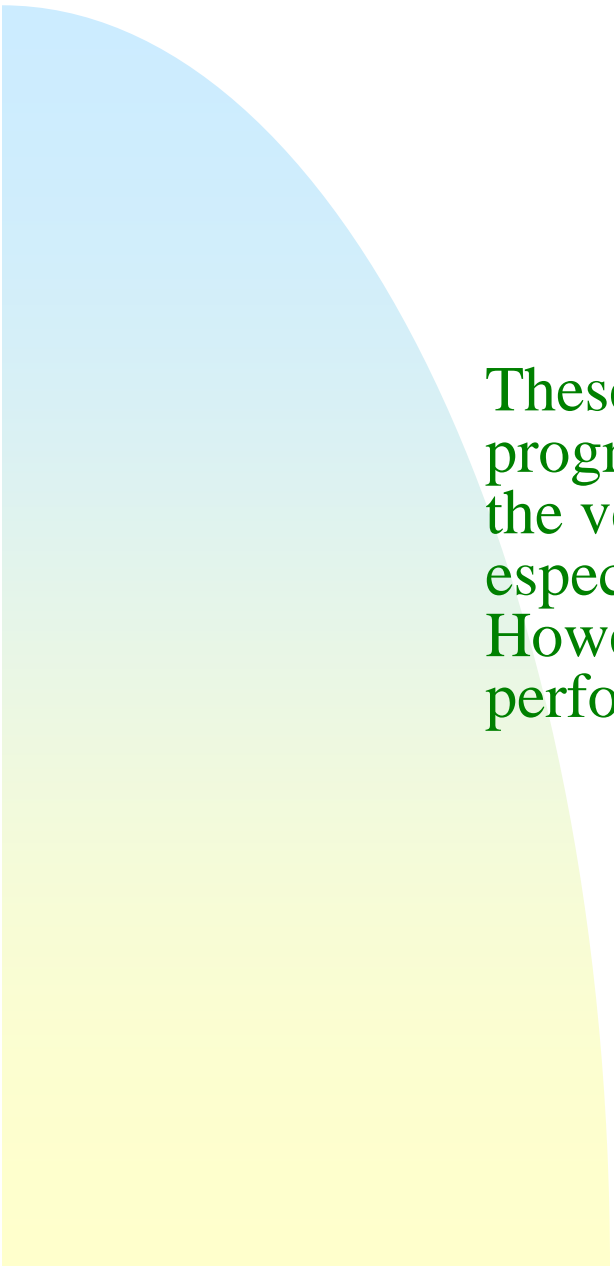
Steven Orzack, Dan Gusfield, Lakshman
Subrahmanyam, Sebastien Lissarrague,
and Laurent Essioux

Fresh Pond Research Institute

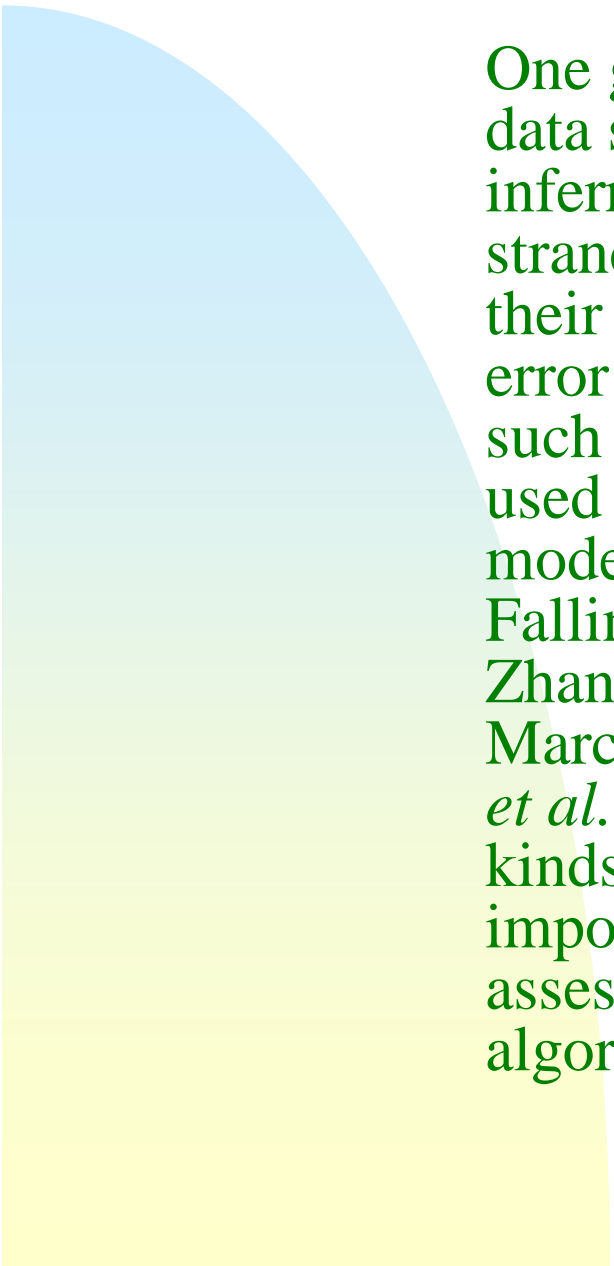
*RECOMB University of Southern
California January 27-28 2007*



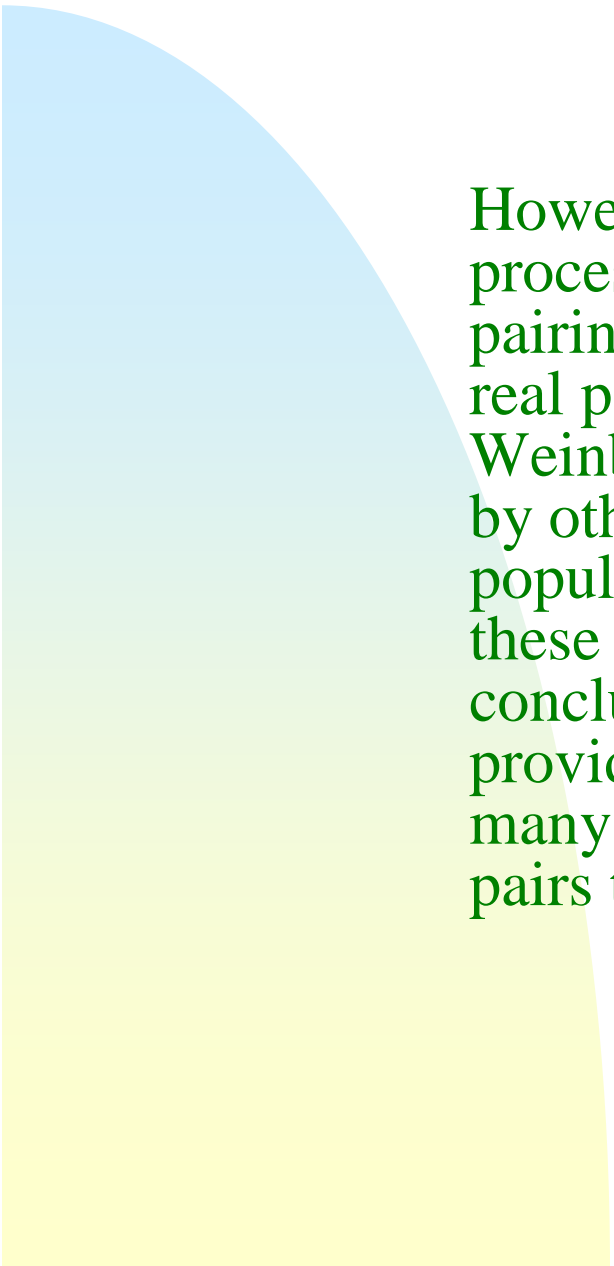
Both molecular and algorithmic methods have been developed for inferring haplotypes from samples of unrelated individuals. The algorithms include expectation-maximization (EM) (Templeton *et al.* 1988, Excoffier and Slatkin 1995, Long *et al.* 1995, Hawley and Kidd 1995, and Fallin and Schork 2000), partial ligation using the Gibbs sampler (Niu *et al.* 2002) or EM (Qin *et al.* 2002), the coalescent-based approach of Stephens *et al.* (2001a) (see also Lin *et al.* 2002 and Stephens and Donnelly 2003), the rule-based approach of Clark (1990), Gusfield (2001), and Orzack *et al.* (2003), and the perfect-phylogeny-based approach of Gusfield (2002) and Eskin *et al.* (2003). These algorithms generate estimates of haplotype frequencies as well as infer pairs of haplotypes for each ambiguous genotype.



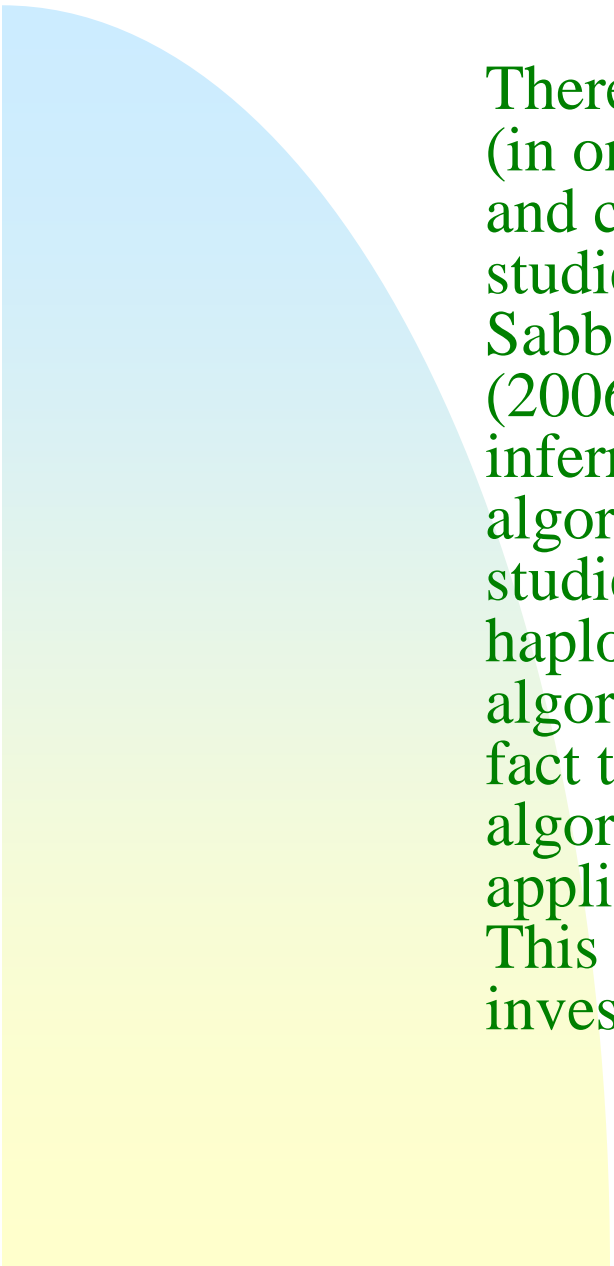
These algorithms and others reflect significant progress in the development of tools for analyzing the veritable flood of data on genetic variation, especially data on DNA sequence variation. However, there are gaps in the assessment of the performance of these tools.



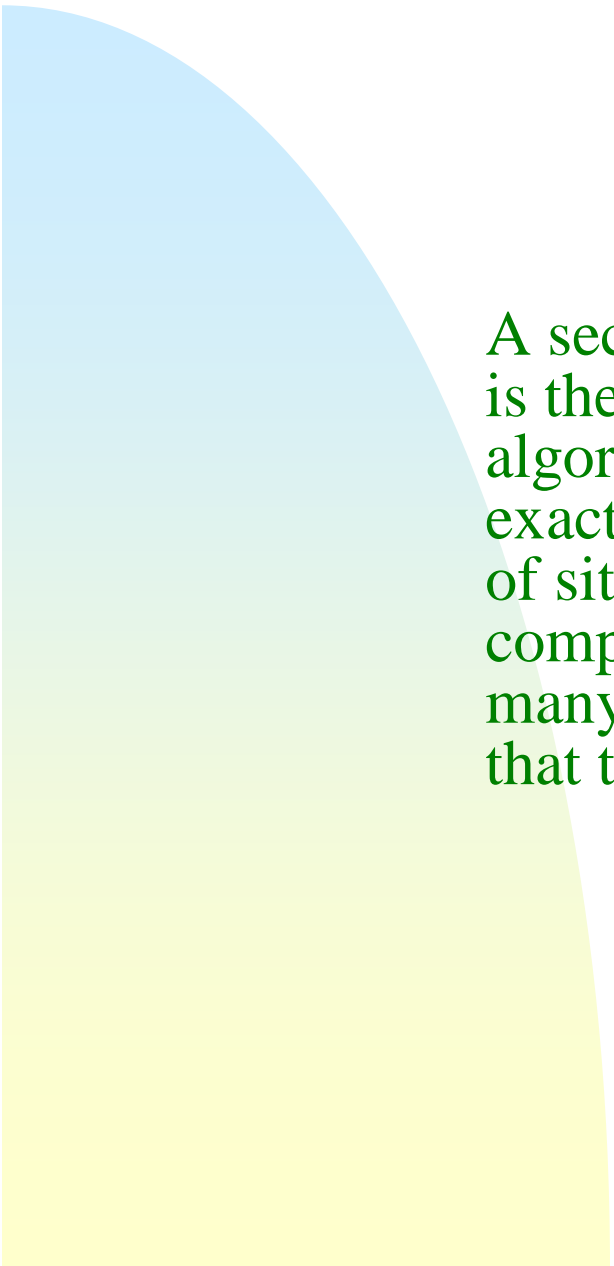
One gap is the scarcity of comparisons involving data sets containing “real” haplotype pairs (those inferred from molecular analyses, e.g., cloning, strand-specific PCR, or somatic cell hybridization; their error rate is likely very low as compared to the error rate for algorithmic inferral.) Instead of using such data, most studies of inferral accuracy have used either simulated data (generated by a neutral model) or randomly-paired real haplotypes (e.g., Fallin and Schork 2000, Niu *et al.* 2002, Niu 2004, Zhang *et al.* 2005, Brinza and Zelikovsky 2006, Marchini *et al.* 2006, Zhang and Zhao 2006, Zhang *et al.* 2006; several of these studies contain other kinds of comparisons as well). These studies are important; they generally have favorable assessments of the performance of inferral algorithms.



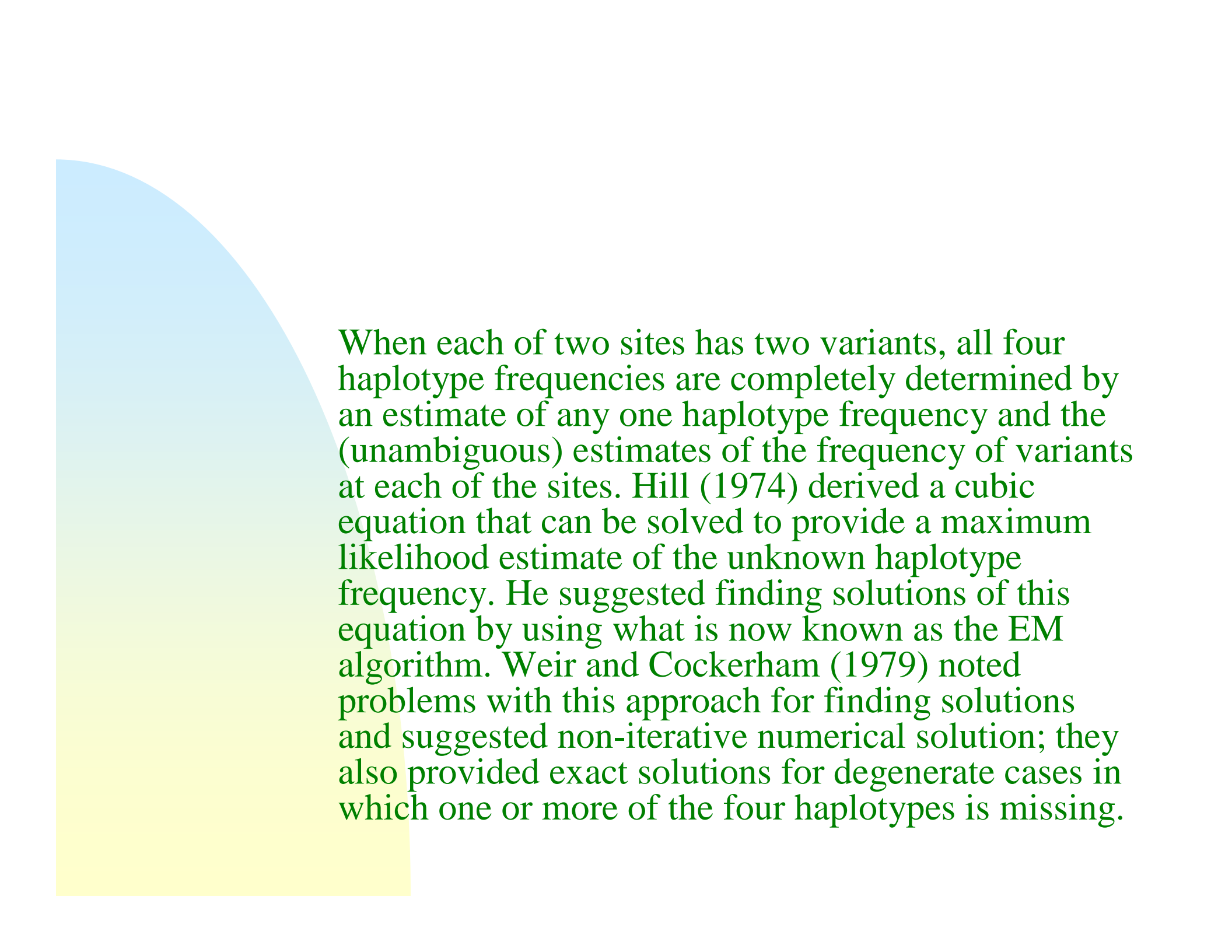
However, such samples are affected by some processes (e.g., random mating or random pairing of haplotypes) that may not influence real populations (even if they are in Hardy-Weinberg equilibrium) and they are not affected by other processes that may influence real populations (natural selection). To this extent, these studies do not by themselves allow us to conclude that these algorithms are generally provide accurate results, which we define as many others do as the proportion of haplotype pairs that is correctly inferred.



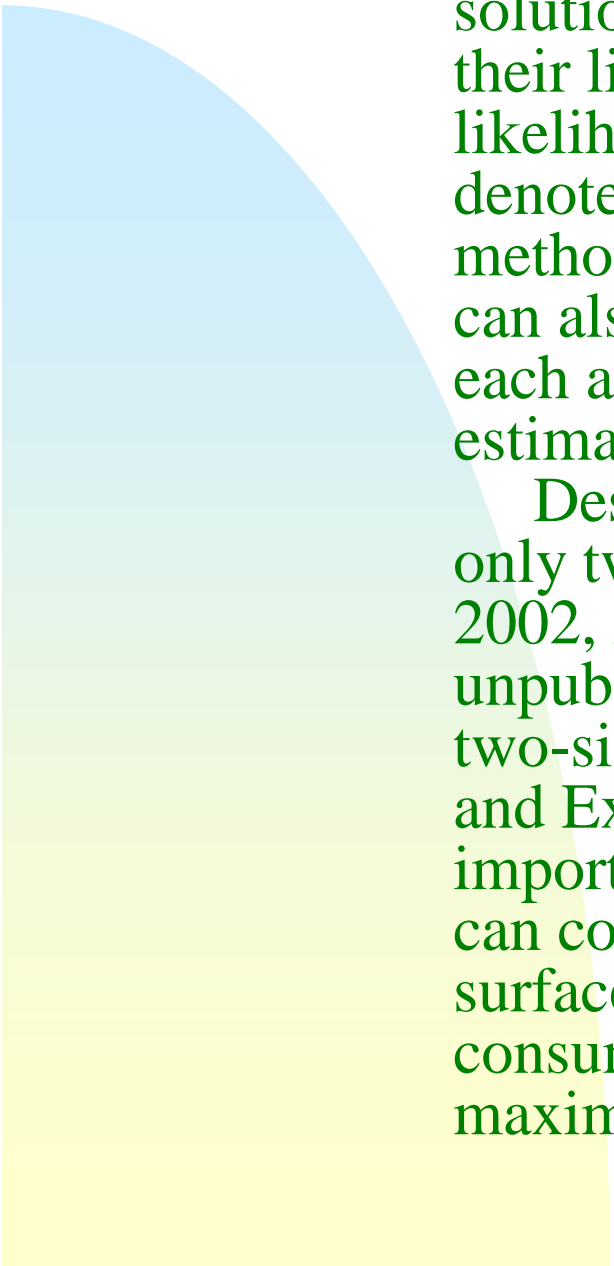
There are only a few studies that involve many sites (in one or more loci), a reasonably large sample size, and completely implement the algorithms being studied. Of these studies, Orzack *et al.* 2003, Sabbagh and Darlu (2005), and Proudnikov *et al.* (2006) assessed algorithmic and experimental inferences and Halperin and Eskin (2004) compared algorithmic and pedigree inferences. While these studies are encouraging in that high proportions of haplotype pairs are shown to be inferred by some algorithms, the paucity of such studies speaks to the fact that it is premature to conclude that present algorithms generally provide accurate results when applied to data sets containing real haplotype pairs. This is the canonical type of application most investigators have in mind.



A second gap in our assessment of inferal methods is the lack of comparisons between exact and algorithmic results, a likely reason being the lack of exact results for genotypes with an arbitrary number of sites. Such comparisons are essential given the complexity of the inference problem and the fact that many of the algorithms are stochastic, which implies that the results may be sample-path-dependent.



When each of two sites has two variants, all four haplotype frequencies are completely determined by an estimate of any one haplotype frequency and the (unambiguous) estimates of the frequency of variants at each of the sites. Hill (1974) derived a cubic equation that can be solved to provide a maximum likelihood estimate of the unknown haplotype frequency. He suggested finding solutions of this equation by using what is now known as the EM algorithm. Weir and Cockerham (1979) noted problems with this approach for finding solutions and suggested non-iterative numerical solution; they also provided exact solutions for degenerate cases in which one or more of the four haplotypes is missing.




One can use standard formulae to generate exact solutions of the cubic equation; one can then compare their likelihoods in order to find the maximum likelihood estimate of the haplotype frequency. We denote this sequence of two steps as the “exact” method. Given the maximum likelihood estimate, one can also generate the most probable haplotype pair for each ambiguous genotype. Here we focus only on the estimation of haplotype frequency.

Despite the existence of these results, we know of only two uses of this exact approach (De Vivo *et al.* 2002, Zee *et al.* 2004); both stem from our unpublished analyses. Instead, most analyses of the two-site case employ the EM algorithm (e.g., Slatkin and Excoffier 1996.) While this algorithm is an important inferential tool, its use can be problematic. It can converge to a local maximum of the likelihood surface, convergence to the maximum can be time consuming as compared to the exact method, and two maxima can have identical likelihoods.

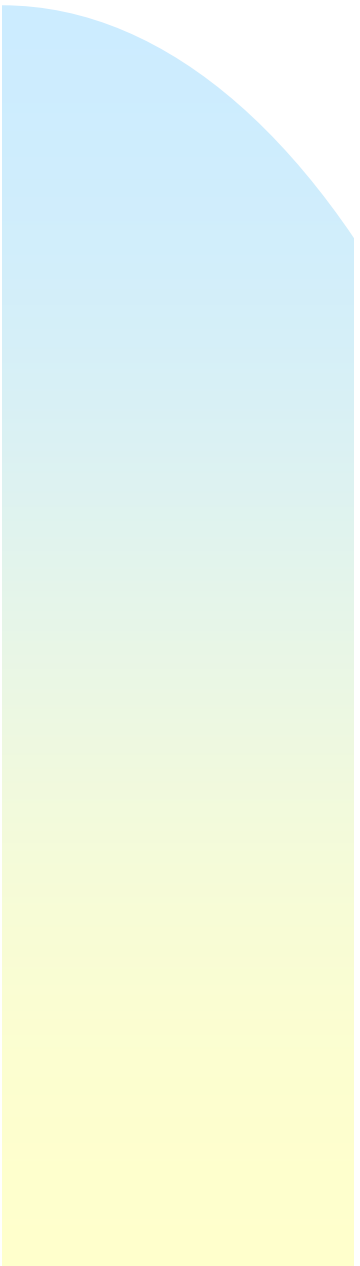


A natural question arises: how do the results of algorithmic methods compare with the exact results?

Genotypes				Haplotypes					
	BB	Bb	bb	AB	Ab	aB	ab	AB frequency	Ln(Likelihood)
AA	7	13	7	27	29	2	0	0.46551725 GM	-47.480226
Aa	0	2	0	28	28	1	1	0.48275861 I	-47.515948
aa	0	0	0	29	27	0	2	0.50000000 GM	-47.480226
AA	31	1	0	101	63	227	39	0.23544239 GM	-464.800460
Aa	21	79	0	141	23	186	80	0.32807544 I	-467.402927
aa	72	0	11	158	6	170	96	0.36787754 LM	-466.885080
AA	20	5	61	105	215	116	166	0.17484833 LM	-714.089840
Aa	2	141	5	115	205	106	176	0.19149581 I	-714.095945
aa	15	1	51	137	183	84	198	0.22750969 GM	-714.046978
AA	2	2	2	62	21	386	151	0.09993113 GM	-601.073664
Aa	49	13	9						
aa	106	119	8						
AA	10	0	30	38	128	80	70	0.11870803 LM	-369.716729
Aa	1	82	3	62	104	56	94	0.19580439 I	-370.588124
aa	7	0	25	89	77	29	121	0.28042430 GM	-369.431641



Case	Site	χ^2 test value	χ^2 P value	Exact Test P value
1	1	0.047	0.828	0.806
	2	1.895	0.169	0.203
2	1	0.044	0.834	0.885
	2	0.171	0.679	0.850
3	1	0.048	0.826	0.817
	2	0.782	0.376	0.456
4	1	0.037	0.847	1.0
	2	0.034	0.853	1.0
5	1	1.320	0.251	0.338
	2	1.879	0.170	0.233



QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Direct Estimates

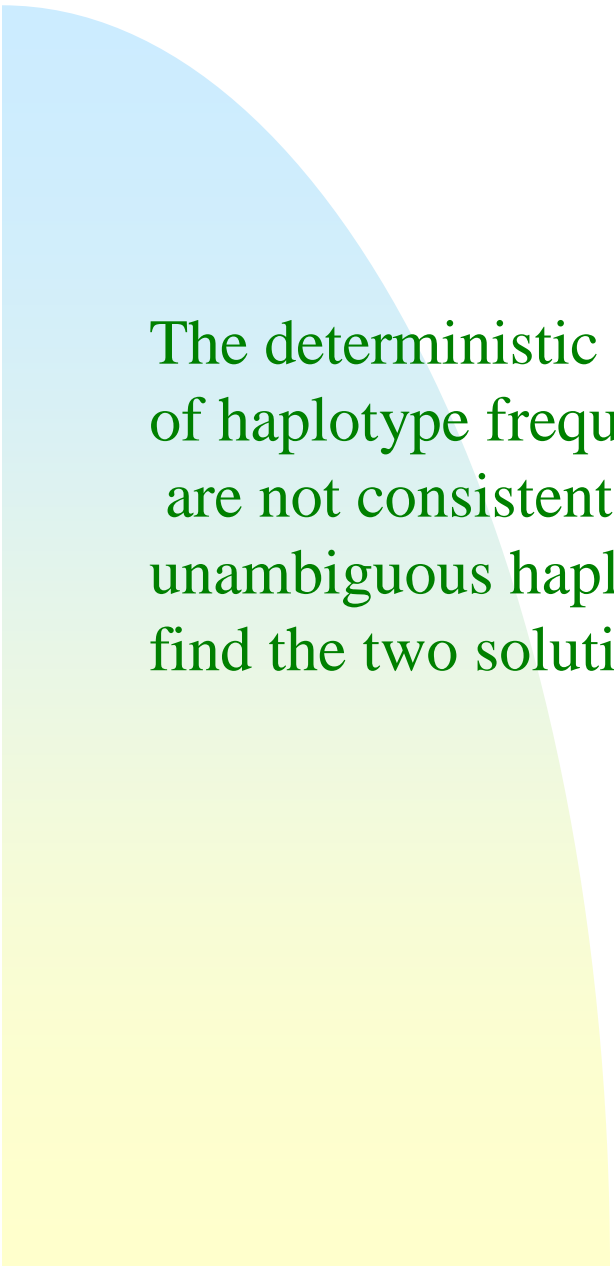
		Number of Sample Paths									
		Method									
Number of AB haplotypes	EM		PL-EM		RB		Phase MR		Phase MS		
	Simulation										
	1	2	1	2	1	2	1	2	1	2	
27 GM	4917	5044	4358	4474	2474	2550	0	21	0	0	
28 I	0	1	589	530	4971	4835	2000	1973	2000	2000	
29 GM	5083	4955	5053	4996	2555	2615	0	6	0	0	
84 - 100	0	0	0	0	16	10	0	0	0	0	
101 GM	7197	7312	7180	7528	15	18	0	0	0	0	
102 - 140	0	0	0	0	9969	9972	1998	2000	1991	1988	
141 I	0	0	0	0	0	0	2	0	1	7	
142 - 157	0	0	0	0	0	0	0	0	8	5	
158 LM	2803	2688	2820	2472	0	0	0	0	0	0	
159 - 163	0	0	0	0	0	0	0	0	0	0	
47 - 104	0	0	0	0	624	633	0	25	0	0	
105 LM	4791	4892	4834	5053	189	204	0	7	0	0	
106 - 114	0	0	0	0	3722	3708	74	269	195	181	
115 I	2	1	168	86	540	529	46	70	61	63	
116 - 136	0	0	0	0	4909	4908	1880	1619	1744	1756	
137 GM	5207	5107	4998	4861	7	7	0	2	0	0	
138 - 188	0	0	0	0	9	11	0	8	0	0	
55	0	0	0	0	49	62	0	0	0	0	
56 - 61	0	0	0	0	8855	8899	1552	1834	875	878	
62 GM	10000	10000	10000	10000	709	668	448	165	1125	1122	
63 - 67	0	0	0	0	387	371	0	1	0	0	
68	0	0	0	0	0	0	0	0	0	0	
21	0	0	0	0	0	0	0	0	0	0	
22 - 37	0	0	0	0	0	0	0	3	0	0	
38 LM	4906	5025	5276	4783	0	0	0	0	0	0	
39 - 61	0	0	0	0	5414	5417	459	788	439	420	
62 I	0	0	0	0	708	690	200	251	96	109	
63 - 88	0	0	0	0	3878	3893	1341	957	1465	1471	
89 GM	5094	4975	4724	5217	0	0	0	1	0	0	
90 - 102	0	0	0	0	0	0	0	0	0	0	
103	0	0	0	0	0	0	0	0	0	0	

denotes $P > 0.05$
 denotes $P < 0.05$

Indirect Estimates

Number of AB haplotypes		Number of Sample Paths									
		Method									
		EM		PL-EM		Haplotyper		Phase MR		Phase MS	
		Simulation									
		1	2	1	2	1	2	1	2	1	2
27		4917	5044	4458	4739	5065	4832	942	914	904	894
28		0	1	0	0	0	0	168	164	176	173
29		5083	4955	5542	5261	4935	5168	890	922	920	933
84		7197	7312	7180	7528	7729	7783	1737	1711	1663	1638
85 - 162		0	0	0	0	0	0	14	16	12	17
163		2803	2688	2820	2472	2271	2217	249	273	325	345
47		4793	4893	5002	5139	4432	4599	632	817	459	427
48 - 187		0	0	0	0	0	0	90	253	101	92
188		5207	5107	4998	4861	5568	5401	1278	930	1440	1481
55		0	0	0	0	1666	1846	0	1	0	0
56 - 67		0	0	0	0	0	0	0	3	0	0
68		10000	10000	10000	10000	8334	8154	2000	1996	2000	2000
21		4906	5025	5276	4783	4333	4501	509	830	455	438
22 Š 102		0	0	0	0	0	0	26	116	29	38
103		5094	4975	4724	5217	5667	5499	1465	1054	1516	1524

denotes $P > 0.05$
 denotes $P < 0.05$



The deterministic algorithm, DPPH, produced a direct estimate of haplotype frequency only for case 1, since the other four cases are not consistent with a perfect phylogeny (each has four unambiguous haplotypes present). For case 1, the program did find the two solutions that have identical likelihoods.

The second deterministic algorithm, HAP, produced results in which all of the ambiguous genotypes were identically resolved.

The solutions were

case 1: $AB = 29$, $Ab = 27$, $aB = 0$, and $ab = 2$; LM

case 2: $AB = 84$, $Ab = 80$, $aB = 244$, and $ab = 22$;

case 3: $AB = 47$, $Ab = 273$, $aB = 174$, and $ab = 108$;

case 4: $AB = 55$, $Ab = 28$, $aB = 393$, and $ab = 144$;

case 5: $AB = 21$, $Ab = 145$, $aB = 97$, and $ab = 53$.

Ambiguous genotypes were resolved as AB/ab in Case 1 and as Ab/aB for the four other cases.

The third deterministic algorithm, 2SNP, also produced results in which all of the ambiguous genotypes were identically resolved.

The solutions were

case 1: $AB = 27$, $Ab = 29$, $aB = 2$, and $ab = 0$; LM

case 2: $AB = 163$, $Ab = 1$, $aB = 165$, and $ab = 101$;

case 3: $AB = 188$, $Ab = 132$, $aB = 33$, and $ab = 249$;

case 4: $AB = 68$, $Ab = 15$, $aB = 380$, and $ab = 157$;

case 5: $AB = 103$, $Ab = 15$, $aB = 63$, and $ab = 135$.

The resolutions for the five cases were opposite to those for HAP in that ambiguous genotypes were resolved as Ab/aB in Case 1 and as AB/ab for the four other cases. The reasons for this difference are unknown.

The fourth deterministic algorithm, HaploFreq, does not produce results in which all of the ambiguous genotypes were identically resolved. The (rounded) solutions were

case 1: AB = 27, Ab = 29, aB = 2, and ab = 0; “LM”

case 2: AB = 101, Ab = 63, aB = 227, and ab = 79; GM

case 3: AB = 105, Ab = 215, aB = 116, and ab = 166; LM

case 4: AB = 62, Ab = 21, aB = 386, and ab = 151; GM

case 5: AB = 38, Ab = 128, aB = 80, and ab = 70. LM

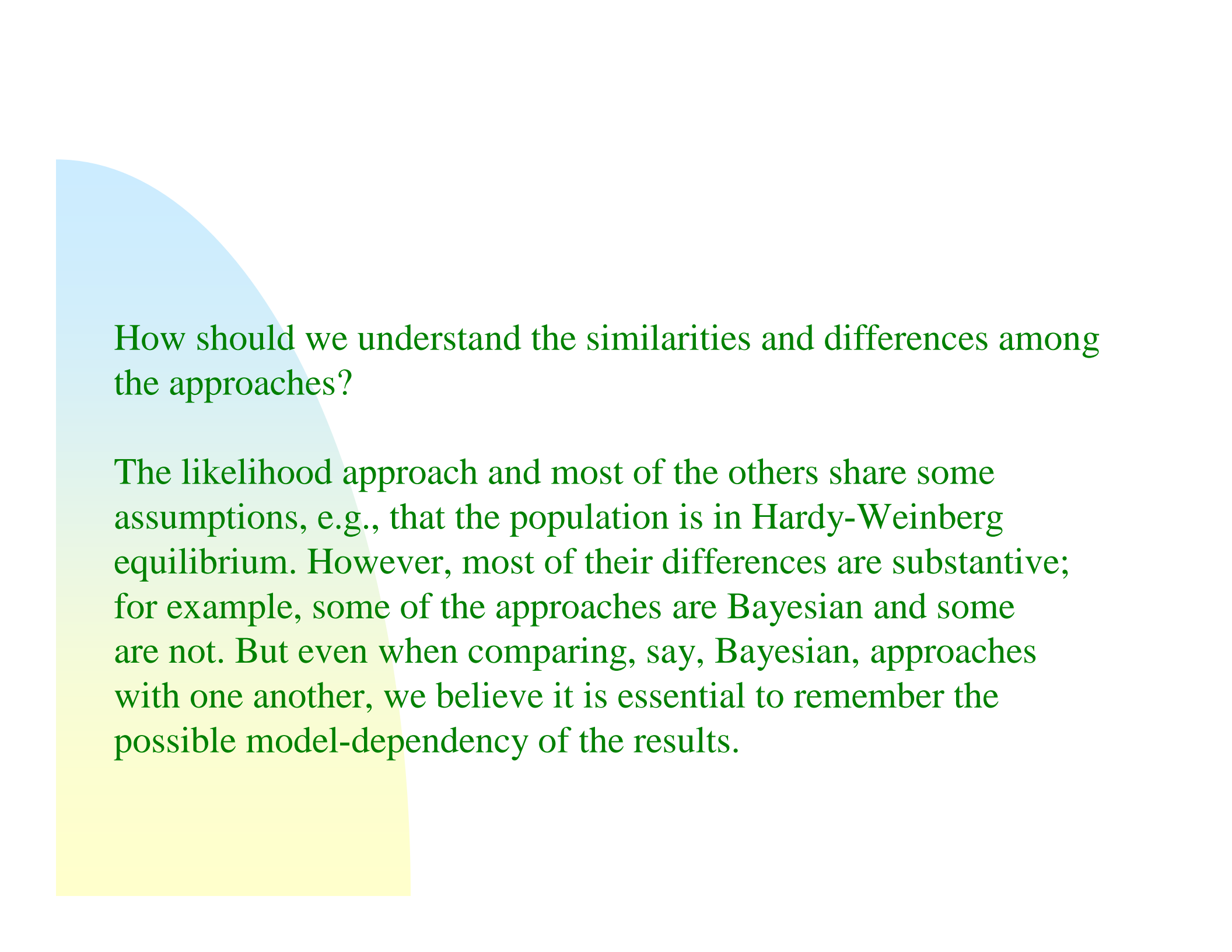
Case 3: $f(AB)$ 0.175 vs 0.228. Case 5: $f(AB)$ 0.119 vs. 0.280

Halperin and Hazan (2006): “the maximum relaxed likelihood can be found in polynomial time and that the optimal solution of the relaxed likelihood approaches asymptotically to the haplotype frequencies in the population.”



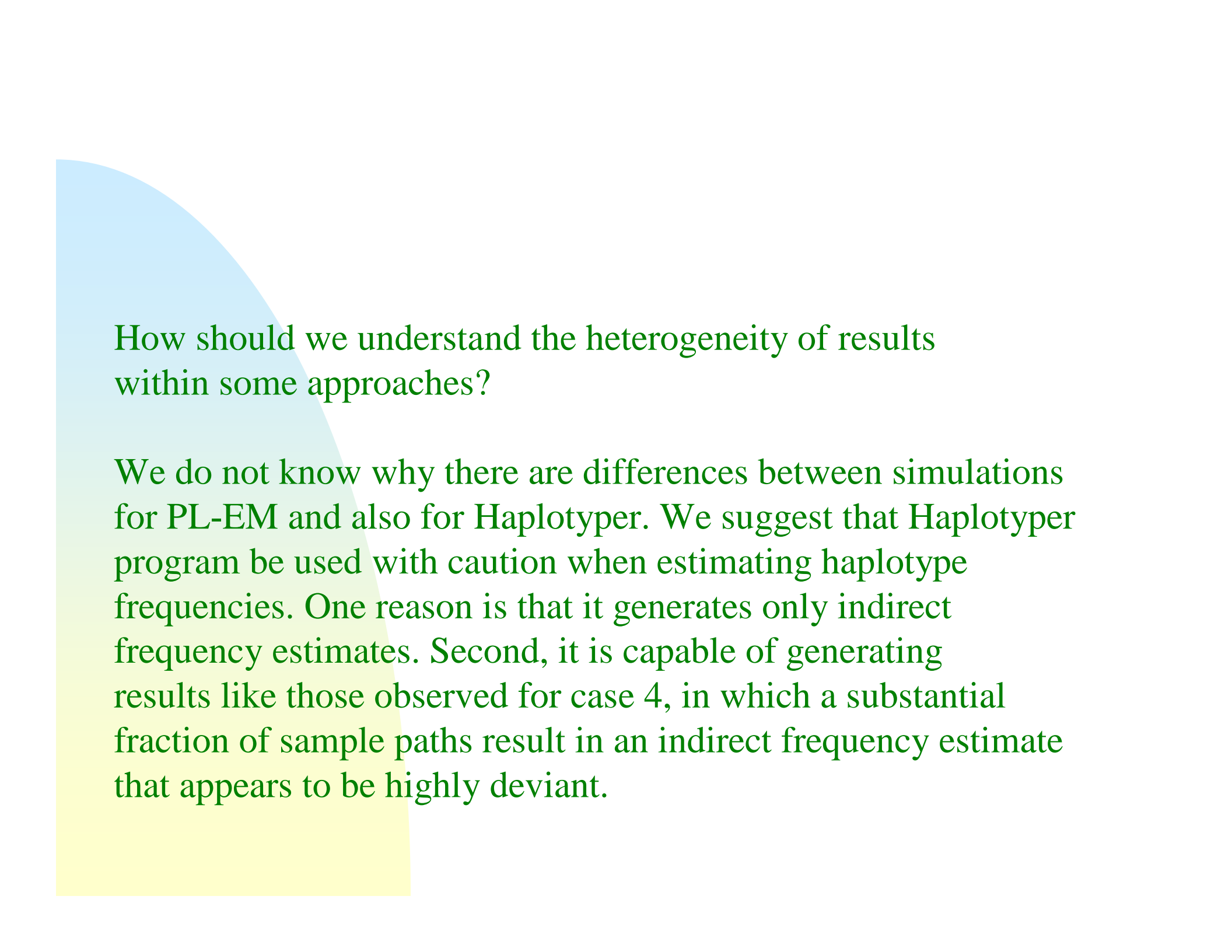
How to relate the exact likelihood approach to the EM approach?

The exact approach is clearly preferable as compared to the EM algorithm if one is using the likelihood approach to analyze two-site data. Results are obtained quickly and one can readily determine whether there is a complexity to the inference problem that a partial or even comprehensive implementation of the EM algorithm might not reveal.



How should we understand the similarities and differences among the approaches?

The likelihood approach and most of the others share some assumptions, e.g., that the population is in Hardy-Weinberg equilibrium. However, most of their differences are substantive; for example, some of the approaches are Bayesian and some are not. But even when comparing, say, Bayesian, approaches with one another, we believe it is essential to remember the possible model-dependency of the results.



How should we understand the heterogeneity of results within some approaches?

We do not know why there are differences between simulations for PL-EM and also for Haplotyper. We suggest that Haplotyper program be used with caution when estimating haplotype frequencies. One reason is that it generates only indirect frequency estimates. Second, it is capable of generating results like those observed for case 4, in which a substantial fraction of sample paths result in an indirect frequency estimate that appears to be highly deviant.

How should we understand the relevance of our results to the general problem of haplotype inference?

It is possible to argue that it is *more* difficult to infer haplotype frequencies or haplotype pairs in the two-site case than in the multi-site case. In the former case, each ambiguous individual is heterozygous at every site, while in the latter case, this is possible but unlikely. To this extent, there is no “partial” information in the two-site case about the haplotypes in ambiguous individuals. In the multi-site case, one ambiguous individual may be, say, homozygous for several sites that are multiply-heterozygous in other ambiguous individuals.

How should we understand the relevance of our results to the general problem of haplotype inference?

On the other hand, it is possible to argue that it is *less* difficult to infer haplotype frequencies or haplotype pairs in the two-site case than in the multi-site case. The fact that all ambiguous individuals in the two-variant, two-site case have identical genotypes would seem to simplify the inference problem.

At present, we are unaware of a satisfactory means of reconciling these conflicting inferences.

Our overall conclusion is that meaningful algorithmic inferral of even common haplotype frequencies requires careful interpretation of the results generated by multiple sample paths. It is also important to compare the results generated by several programs. Of course, knowledge of the evolutionary processes underlying the evolution of the sequences might allow one to prefer one algorithm. Even so, the investigator potentially faces the challenge of reconciling sets of inferrals. In such a circumstance, one possibility is a consensus method, as suggested by Orzack *et al.* (2003) (see also Fullerton *et al.* 2004), in which, say, the most common set of inferrals is used. As discussed by Orzack *et al.*, additional genetic criteria can be used to determine which inferral sets are queried in order to determine the consensus inferrals. This general approach is promising, especially in as much as it provides a clear way to determine an answer and they show that it performs well for the locus they studied. Nonetheless, additional research is needed to assess how generally useful it will be.



Thanks!

This work has been partially supported by NSF, NIA, NIH,
and by Variagenics, Inc.