

Central Challenges in Population-genetic Analysis of High-throughput Data

- Errors in small fragment alignment, with or without a “reference” genome, can lead to inflated estimates of variation and false allelism.
- Error rates as high as 0.01, resulting from a variety of sources (not just read quality), generate false heterozygosity and encourage inflated estimates of rare alleles.
- Most methods to estimate error rates are arbitrary, and “off-the shelf” estimates are not reliable – errors in the error rate lead to errors in population parameter estimates.
- Subjective methods for discarding potentially problematical sequences can lead to downward bias in variation estimates and can discard substantial amounts of data.
- At low to moderate coverages, there is a high probability that only one of the two alleles at the site within a diploid individual will have been sequenced, $2(1/2)^n$ – creating the false impression of homozygosity.

Estimation of Population-genetic Parameters With Single Diploid Individuals

- Estimates of average site-specific heterozygosity (π) can be achieved genome-wide, within chromosomal regions, or at classes of sites with particular functional significance.

```
AGCTTAAGTAGGTCACTATGT
GGTAAGCTTACGTAGGTCAGTATGTGAG
TTAAGTAGGTCACTATGTG
AGCTTACGTAGGTCAGTATGTGAGACCT
ACGTAGGTCACTATG
TAAGCTTAAGTAGGTCACTATGT
```

- Estimates of the correlation of heterozygosity across pairs of sites (Δ) separated by various physical distances can yield information on the degree of linkage disequilibrium.
- In randomly mating populations, these single-individual estimates are reasonably representative of the population average parameter estimates.
- Such analyses can yield insight into:
 - chromosomal regions subject to selective sweeps;
 - strength of selection operating on various classes of sites;
 - for neutral sites, the key parameters $4Nu$ and $4Nc$, respectively equal to the ratios of the power of mutation and the power of recombination to the power of drift.

A Maximum-Likelihood Approach

- **Data structure:** a site that has been sequenced n times within an individual will have a sequence profile (n_A, n_C, n_G, n_T) , with the sum of the four elements equaling n .
- Goal is to obtain estimates of the heterozygosity (π) and disequilibrium (Δ) that best explain the data, using the data itself to estimate and factor out the error rate (ε), and weighting each site by its information content.
- Assumes that all read fragments are properly aggregated, either by *de novo* assembly in the case of long reads or via a reference genome in the case of short reads, with complicating regions involving paralogs and mobile elements having been masked out.
- The raw sequence reads may be subject to trimming and quality control prior to analysis.
- The error structure of the data is assumed to be homogeneous, with each nucleotide having the same probability of misassignment to all others, but more complex scenarios are readily implemented.

Estimation of Average Heterozygosity, π

- For the full range of candidate values of π and ϵ , the likelihood of the data at each site is obtained by considering the probabilities of the observed data conditional on all possible genotypic states.
- Must condition on whether the site is truly homozygous or heterozygous.

Conditional on the site being homozygous, the likelihood of the observed data is obtained by summing over the likelihoods conditional on all four possible homozygous types (AA, CC, GG, and TT, with respective relative probabilities p_1 , p_2 , p_3 , and p_4),


$$\ell_1(n_1, n_2, n_3, n_4) = \sum_{i=1}^4 p_i \cdot b(n - n_i; n, \epsilon),$$

where $b(n-n_i; n, \epsilon)$ is the probability of $n-n_i$ errors, given error rate ϵ .

Conditional on the site being heterozygous, must incorporate:

- 1) the probabilities of alternative heterozygous genotypes,
- 2) the error probability distribution,
- 3) the binomial sampling distribution of the alternative alleles.

$$\begin{aligned}
 \ell_2(n_1, n_2, n_3, n_4) = & \underbrace{\sum_{i=1}^4 \sum_{j>i}^4 2p_i p_j}_{\text{genotype frequencies}} \cdot \underbrace{b(n - n_i - n_j; n, 2\epsilon/3)}_{\text{error distribution}} \\
 & \cdot \underbrace{p(n_i; n_i + n_j, 0.5)}_{\text{allele sampling}} / S,
 \end{aligned}$$


 summed heterozygote frequencies

The total likelihood of the observed data at the site, given the assumed values of π and ε :

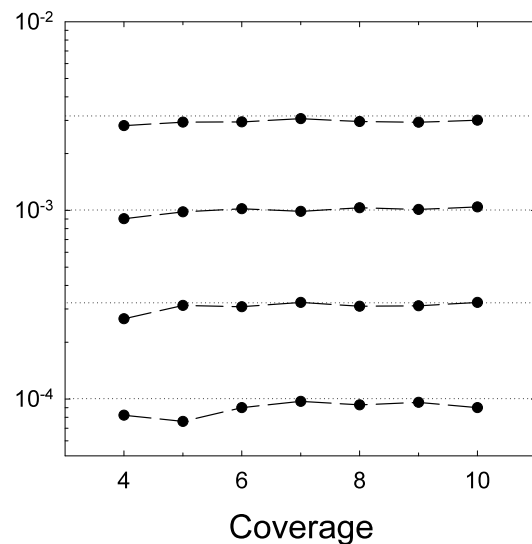
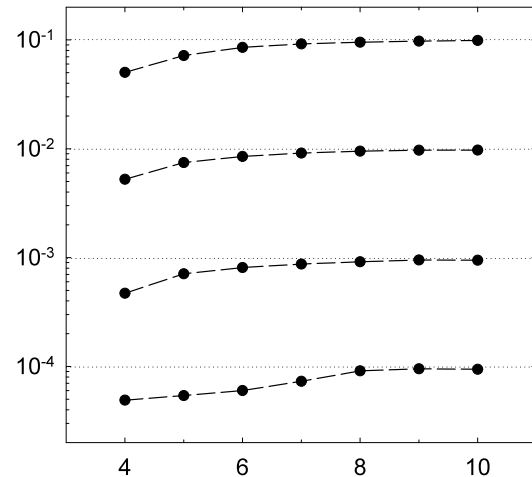
$$\ell(n_1, n_2, n_3, n_4) = (1 - \pi)\ell_1(n_1, n_2, n_3, n_4) + \pi\ell_2(n_1, n_2, n_3, n_4)$$

The total likelihood of all of the data is the product of the above over all sites, yielding the log likelihood

$$L = \sum N(n_1, n_2, n_3, n_4) \cdot \ln \left[\ell(n_1, n_2, n_3, n_4) \right]$$

Estimates of π and Their Standard Errors From Simulated Data

- $N = 10,000$ sites
- Error frequency = 0.001
- Even nucleotide frequencies



- Estimates are asymptotically unbiased with large numbers of sites.
- The downward bias can be removed entirely by dividing the estimate by $(1 - c)$, where $c = n(1/2)^{n-1}$ is the frequency of heterozygous sites with $(1, n - 1)$ allele-sampling configurations ($n = \text{coverage}$).

- The asymptotic sampling variance of the estimate of π at large N is

$$\pi(1-\pi)/N$$

which is the minimum possible sampling error.

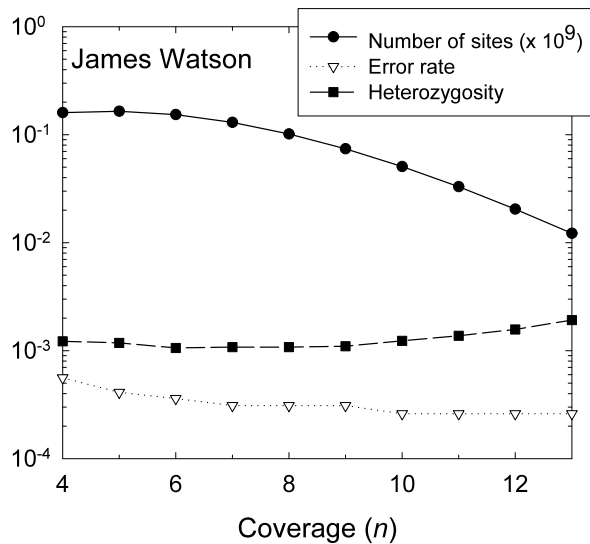
The *Daphnia* Genome:

Estimates of the fraction of heterozygous nucleotide sites in the *Daphnia pulex* genome derived by the ML and MM methods for pools of sites with various coverages. Raw Heterozygosity is the fraction of sites at which two or more nucleotides were observed in the sequence reads. The MM (method of moment) estimates were obtained by setting the error rate to 0.00120, the ML estimate of ϵ obtained by jointly analyzing all sites with $>3\times$ coverage.

Coverage (n)	Number of Sites	Raw Hetero. (H)	ML Error Estimate (ϵ)	Nucleotide Heterozygosity	
				ML Estimate (SE)	MM Estimate (SE)
2	2293860	0.00569	0.0026	0.00150 (0.00003)	0.00662 (0.00010)
3	3172155	0.00693	0.0021	0.00104 (0.00001)	0.00446 (0.00006)
4	4632345	0.00766	0.0016	0.00150 (0.00002)	0.00330 (0.00005)
5	6447634	0.00832	0.0014	0.00131 (0.00001)	0.00251 (0.00004)
6	8329666	0.00913	0.0013	0.00130 (0.00001)	0.00203 (0.00003)
7	9844796	0.00993	0.0013	0.00124 (0.00001)	0.00161 (0.00003)
8	10654782	0.01086	0.0012	0.00122 (0.00001)	0.00133 (0.00003)
9	10622205	0.01173	0.0012	0.00118 (0.00001)	0.00100 (0.00003)
10	9828605	0.01246	0.0012	0.00113 (0.00001)	0.00054 (0.00004)
11	8459954	0.01341	0.0011	0.00117 (0.00001)	0.00030 (0.00004)
12	6863918	0.01429	0.0011	0.00117 (0.00001)	-0.00001 (0.00005)
13	5216553	0.01531	0.0011	0.00123 (0.00002)	-0.00017 (0.00005)

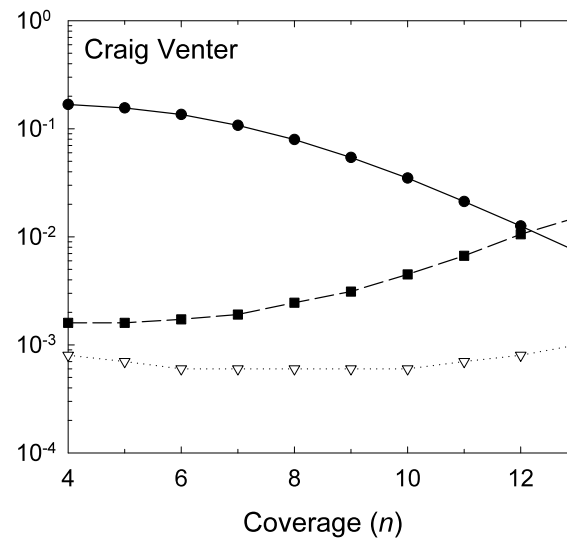
Nucleotide Heterozygosity Across the Human Genome

- Known repeat regions and paralogous genes masked out.
- Restricted to sequences uniquely mapping to single positions, with >90% sequence identity.



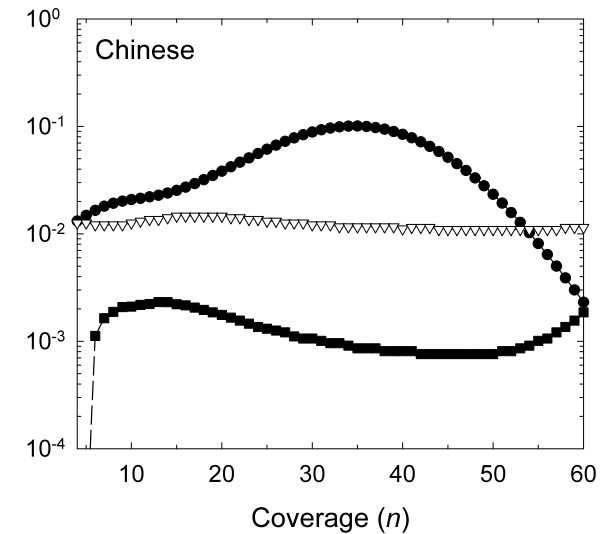
454 sequence
Wheeler et al. (2008)

4-8x, $\pi = 0.00086$
 4-10x, $\pi = 0.00091$
 4-12x, $\pi = 0.00096$
 4-14x, $\pi = 0.00096$



Sanger sequence
Levy et al. (2007)

4-8x, $\pi = 0.0013$
 4-10x, $\pi = 0.0017$
 4-12x, $\pi = 0.0019$
 4-14x, $\pi = 0.0019$



Illumina sequence
Wang et al. (2008)

Estimating the Correlation of Heterozygosity for All Pairs of Sites Separated by d Nucleotides, Δ_d .

$n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}$ = octet of observed nucleotide counts at loci a and b

Likelihood of the data observed at the pair of sites:

$$\ell(n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}) = \underbrace{\left[(1 - \pi)^2 + \Delta\pi(1 - \pi) \right]}_{\text{joint homozygosity}} \ell_{1a}\ell_{1b} + \underbrace{\left[\pi^2 + \Delta\pi(1 - \pi) \right]}_{\text{joint heterozygosity}} \ell_{2a}\ell_{2b} + \underbrace{\left[\pi(1 - \pi)(1 - \Delta) \right]}_{\text{heterozygosity / homozygosity}} (\ell_{1a}\ell_{2b} + \ell_{1b}\ell_{2a}),$$

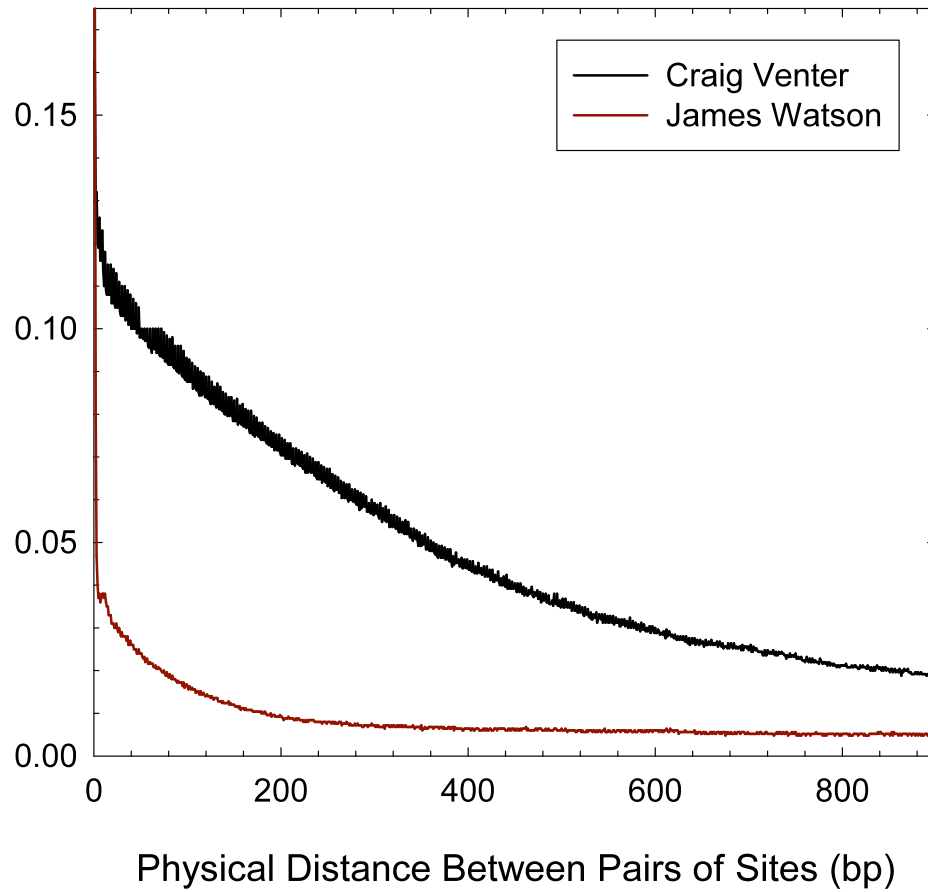
ℓ_{1a}, ℓ_{1b} = likelihoods of homozygosity at loci a and b , given the observed quartets at the sites and ε .

ℓ_{2a}, ℓ_{2b} = likelihoods of heterozygosity at loci a and b , given the observed quartets at the sites and ε .

Total log likelihood summed over all pairs of sites:

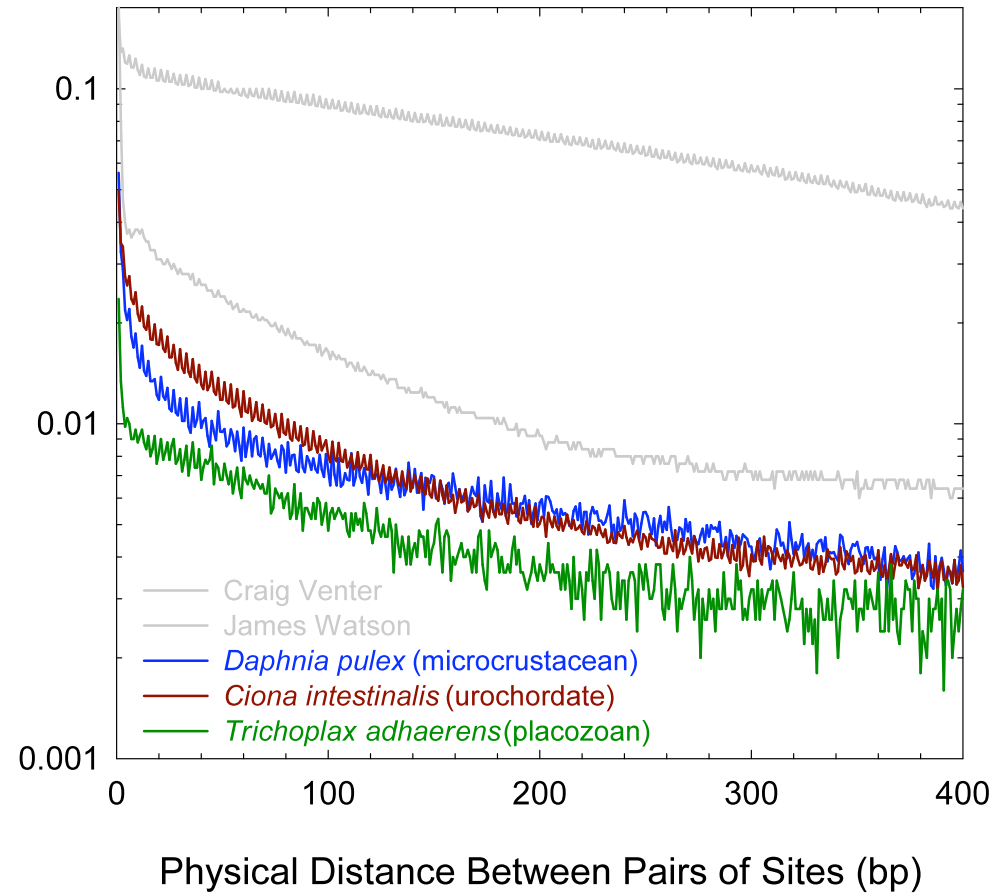
$$L = \sum N(n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}) \cdot \ln \left[\ell(n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}) \right]$$

Linkage Disequilibrium (Δ) Within Human Genomes



Note: $E(D^2) = \Delta\pi(1-\pi)/4$

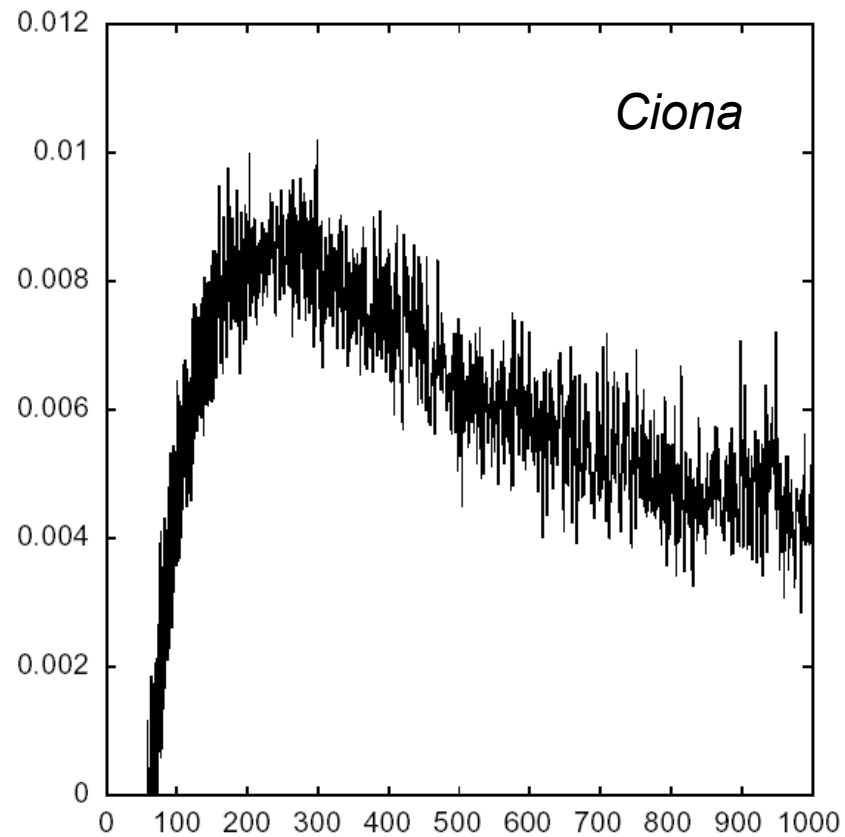
LD is Relatively High in Humans



Estimating the Population Recombination Rate ($4N_e c$) from Δ

- Strobeck and Morgan (Genetics, 1978) – at drift-mutation-recombination equilibrium for neutral sites, Δ is a function of π (which estimates $4N_e u$) and $4N_e c$.

$4N_e c / d$
(ratio of the power of
recombination to the
power of drift at
adjacent nucleotide sites)



Physical Distance Between Sites, d (bp)

Estimation of Population-level Allele Frequencies by ML

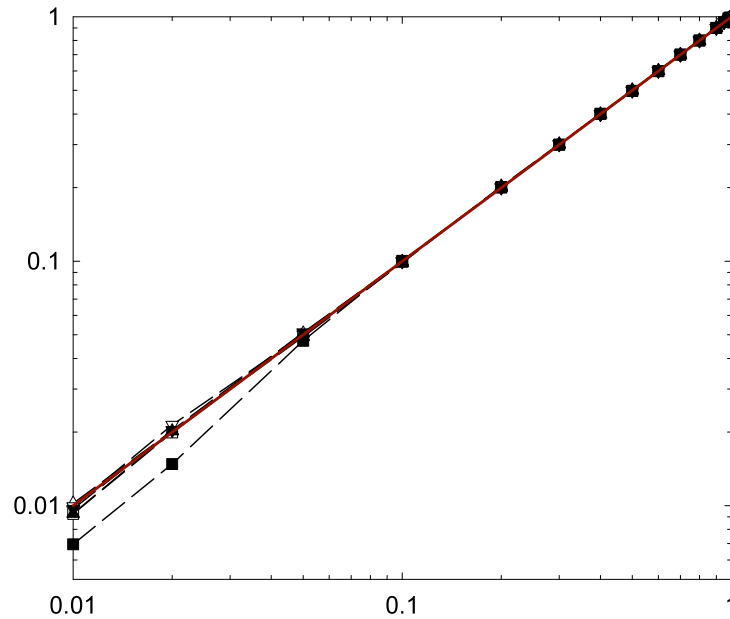
- Eliminates the necessity of an external error rate.
- Yields unbiased estimates of frequencies of rare alleles (rarer than the error rate).
- Sampling variance of p is very close to the theoretical minimum, $p(1-p)/(2N)$, for coverages $>4x$.

Likelihood function for the read profile of an individual:

$$\begin{aligned}
 P(n_1, n_2, n_3 | n, p, \epsilon) = & \overbrace{p^2} \phi_e(n_2, n_3; n, \epsilon/3, 2\epsilon/3) + \overbrace{(1-p)^2} \phi_e(n_1, n_3; n, \epsilon/3, 2\epsilon/3) \\
 & + \underbrace{2p(1-p)}_{\text{genotype}} \underbrace{\phi_e(n_3; n, 2\epsilon/3)}_{\text{error distribution}} \underbrace{p(n_1; n_1 + n_2, 0.5)}_{\text{allele sampling}}.
 \end{aligned}$$

n_1 and n_2 = number of reads for the major and minor alleles.

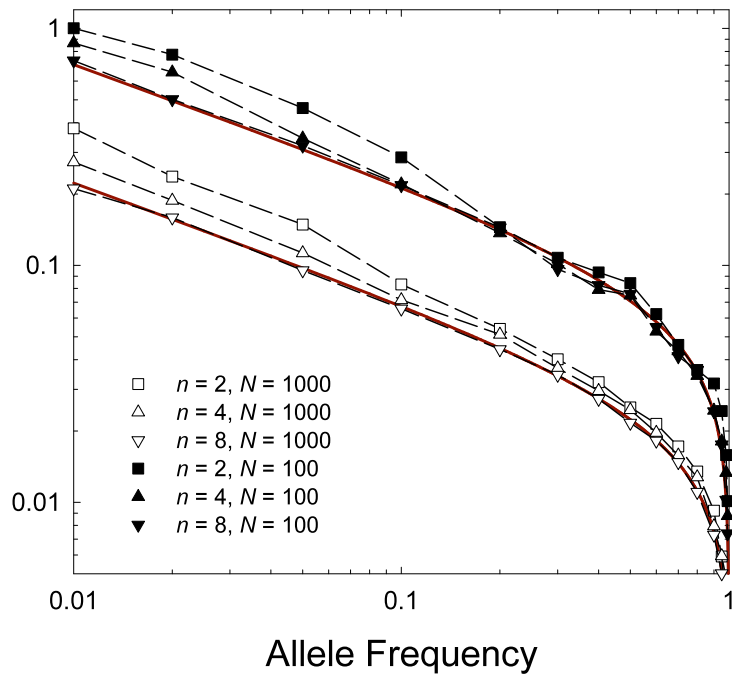
n_3 = number of erroneous reads.



Simulation Results:

- Error rate, $\varepsilon = 0.01$.
- n = coverage per site.
- N = number of diploid individuals.

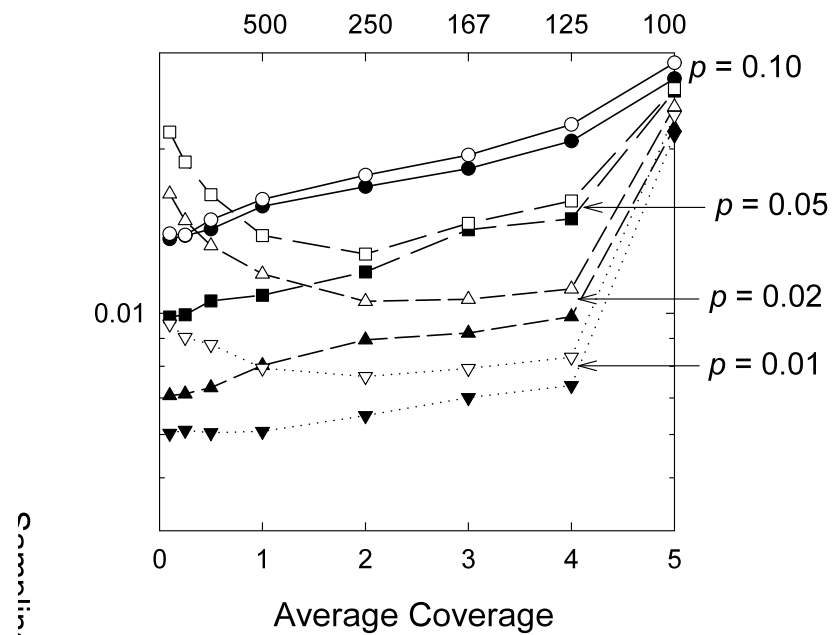
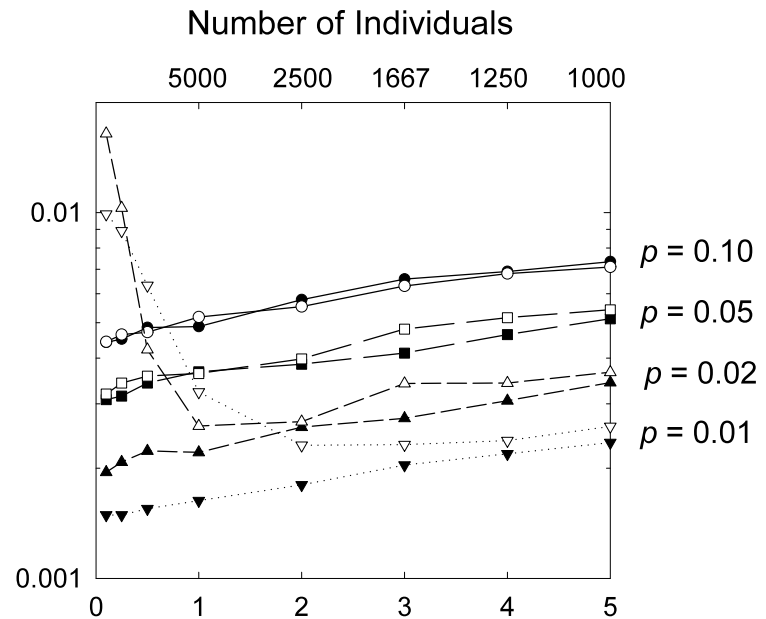
← Estimates are essentially unbiased.



← Sampling variance is close to the asymptotic minimum for binomial sampling.

The ML Approach Provides
 a Logical Basis for Identifying
 Optimal Survey Designs, Given
 a Fixed Amount of Resources

Open symbols, error rate = 0.01
 Closed symbols, error rate = 0.001



General Conclusions:

- Despite uneven coverage and the presence of sequence errors, accurate information can be extracted from whole-genome analyses of single diploid individuals.
 - Neither arbitrary coverage cutoffs nor external measures of the base-call error rate are necessary, or even desirable, to obtain meaningful estimates.
 - This is a preferred situation, as the former can discard substantial amounts of data, and the latter can involve extrapolations from extrinsic studies with uncertain justification.
- The ML approach accounts for all sources of error (not just machine-read errors), including true sequences of somatic mutations, errors incurred during sample storage or preparation (ancient DNA), and perhaps some misalignment errors.
- There are, however, limitations to what can be accomplished – completely unbiased estimates of population-genetic parameters may not be possible at very low coverages.

Future improvements:

- The assumption of homogeneous error rates is readily relaxed by incorporating into the likelihood functions multiple terms for alternative nucleotide changes.
- Additional complexity may also be incorporated by distinguishing alternative types of heterozygotes (e.g., transitions vs. transversions).
- The utility of these kinds of modifications can be easily evaluated by testing for the significance of the improved model fit by using conventional likelihood-ratio test statistics.

Key Acknowledgments:

- Xiang Gao, NIH Ruth Kirschstein Fellow,
Indiana University



- Bernhard Haubold, Group Leader, Max Planck Institute
for Evolutionary Biology



Programs available at:

<http://guanine.evolbio.mpg.de/mlRho/>

Trichoplax adhaerens, sole known member of the phylum Placozoa.

- Sanger sequence of ~100 Mb diploid genome to ~8x total coverage.
- Recombination and sexual stages are unknown.



$$\pi = 0.0061$$
$$\varepsilon = 0.0050$$