Closed-form sampling formulas for the coalescent with recombination

Yun S. Song

CS Division and Department of Statistics
University of California, Berkeley

September 7, 2009

Joint work with Paul Jenkins
Outline

1. Introduction
   - Motivation

2. Asymptotic Sampling Formula
   - Closed-form one-locus sampling formulas
   - Two-locus sample configuration
   - Universality
   - Details
   - Classifying the maximum likelihood estimate (MLE)

3. Accuracy Results
   - Detailed examples
   - Distribution of errors

4. Conclusion
   - Summary
   - Further work
The basic problem

What is the probability of observing a sample of DNA sequences for a given population genetics model?

This is behind many problems in population genetics, e.g.

- Estimating parameters: $L(\theta, \rho) = \mathbb{P}(D \mid \theta, \rho)$
- Ancestral inference
- Detecting departures from neutrality
The basic problem

What is the probability of observing a sample of DNA sequences for a given population genetics model?

- But, except in the one-locus case with a special model of mutation, closed-form sampling formulas are generally unknown.
- When recombination is involved, obtaining an analytic formula for the sampling distribution has so far remained an intractable problem.
Monte Carlo Approaches

- Importance Sampling
- MCMC
- Rejection algorithms
Monte Carlo Approaches

- Importance Sampling
- MCMC
- Rejection algorithms

The coalescent with recombination
Monte Carlo Approaches

- Importance Sampling
- MCMC
- Rejection algorithms

The coalescent with recombination

For large recombination rates, the genealogies sampled by Monte Carlo methods are typically very complicated, containing many recombination events.
Monte Carlo Approaches

- Importance Sampling
- MCMC
- Rejection algorithms

The coalescent with recombination

- For large recombination rates, the genealogies sampled by Monte Carlo methods are typically very complicated, containing many recombination events.
- However, we in fact expect the dynamics to be easier to study for large recombination rates, since the loci under consideration would then be less dependent.
Monte Carlo Approaches

- Importance Sampling
- MCMC
- Rejection algorithms

The coalescent with recombination

- For **large recombination rates**, the **genealogies** sampled by Monte Carlo methods are typically very **complicated**, containing many recombination events.

- However, we in fact expect the **dynamics** to be **easier** to study for large recombination rates, since the loci under consideration would then be **less dependent**.

- It seems plausible that there exists a stochastic process **simpler** than the standard coalescent with recombination, that describes the dynamics of the **relevant degrees of freedom** for large recombination rates.
Closed-form sampling formula for large recombination rates?
Closed-form sampling formula for large recombination rates?

A two-locus model

- The loci are labeled $A$ and $B$.
- $\rho$, the population-scaled recombination rate
Closed-form sampling formula for large recombination rates?

A two-locus model

- The loci are labeled $A$ and $B$.
- $\rho$, the population-scaled recombination rate
- $n$, sample configuration (defined later)
- $q(n)$, the sampling distribution of $n$
Closed-form sampling formula for large recombination rates?

A two-locus model

- The loci are labeled $A$ and $B$.
- $\rho$, the population-scaled recombination rate
- $n$, sample configuration (defined later)
- $q(n)$, the sampling distribution of $n$

Asymptotic Series

As $\rho \to \infty$, find

$$q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right),$$

where $q_0(n)$, $q_1(n)$, $q_2(n)$, ... are independent of $\rho$. 
Closed-form sampling formula for large recombination rates?

A two-locus model

- The loci are labeled $A$ and $B$.
- $\rho$, the population-scaled recombination rate
- $n$, sample configuration (defined later)
- $q(n)$, the sampling distribution of $n$
- $\theta_A, \theta_B$, the population-scaled mutation rates
- $P^A = (P^A_{ij}), P^B = (P^B_{ij})$, transition matrices for the mutation process in the case of a finite-alleles model

Asymptotic Series

As $\rho \to \infty$, find

$$q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right),$$

where $q_0(n), q_1(n), q_2(n), \ldots$ are independent of $\rho$. 
Outline

1. Introduction
   - Motivation

2. Asymptotic Sampling Formula
   - Closed-form one-locus sampling formulas
   - Two-locus sample configuration
   - Universality
   - Details
   - Classifying the maximum likelihood estimate (MLE)

3. Accuracy Results
   - Detailed examples
   - Distribution of errors

4. Conclusion
   - Summary
   - Further work
Closed-form one-locus sampling formulas

- $n = (n_1, \ldots, n_K)$, where $n_i$ denotes the number of gametes with allele $i$ at the locus.
- $q(n)$, probability of an ordered sample with configuration $n$.
- $(x)_n := x(x + 1) \cdots (x + n - 1)$

Infinite-alleles model

Ewens’ sampling formula (1972): $q_{ESF}(n) = \frac{\theta^K}{(\theta)_n} \prod_{i=1}^{K} (n_i - 1)!$

Finite-alleles Parent-Independent Mutation (PIM) model

- Mutation transition matrix satisfies $P_{ij} = P_j$.
- Wright’s sampling formula (1949): $q_{WSF}(n) = \frac{1}{(\theta)_n} \prod_{i=1}^{K} (\theta P_i)^{n_i}$
- Any diallelic recurrent mutation model can be transformed into a PIM model.
Two-locus sample configuration $n = (a, b, c)$

- $a = (a_i)$, $K$-dim vector
- $b = (b_j)$, $L$-dim vector
- $c = (c_{ij})$, $K$-by-$L$ matrix

A sample of 10 gametes

\[
\begin{array}{cccc}
\text{A sample of 10 gamet} & & & \\
\hline
\text{Red} & \text{Blue} & \text{Red} & \text{Blue} \\
\text{Red} & \text{Blue} & \text{Red} & \text{Green} \\
\text{Red} & \text{Green} & \text{Red} & \text{Green} \\
\text{Red} & \text{Blue} & \text{Red} & \text{Green} \\
\text{Yellow} & \text{Blue} & \text{Yellow} & \text{Blue} \\
\text{Yellow} & \text{Red} & \text{Yellow} & \text{Green} \\
\text{Yellow} & \text{Green} & \text{Yellow} & \text{Green} \\
\text{Orange} & \text{Blue} & \text{Orange} & \text{Blue} \\
\end{array}
\]

$c = (c_{ij})$, a $K$-by-$L$ matrix

\[
\begin{array}{c|cc|}
& 1, \ldots, L \\
\hline
j \in & 1 & 2 \\
\text{Red} & 3 & 2 \\
\text{Green} & 3 & 1 \\
\text{Yellow} & 1 & 0 \\
\end{array}
\]

Sample size is $c = |c| = 10$. 

---

**Introduction**

**Accuracy Results**

**Conclusion**
Two-locus sample configuration

\[ a = (a_i)_{i \in \{1, \ldots, K\}}, \text{ multiplicity of left half-fragments.} \]

\[ b = (b_j)_{j \in \{1, \ldots, L\}}, \text{ multiplicity of right half-fragments.} \]

The complete sampling distribution is then \( q(a, b, c) \).
Two-locus sample configuration

\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]
Two-locus sample configuration

\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

\( q_0(a, b, c) \) is the exact sampling distribution when the two loci are unlinked \((\rho = \infty)\).
Two-locus sample configuration

\[ q(a, b, c | \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

\[ q_0(a, b, c) \]

\( q_0(a, b, c) \) is the exact sampling distribution when the two loci are unlinked (\( \rho = \infty \)).

\[ c_A = (c_i) \quad \text{and} \quad c_B = (c_j) \]
Two-locus sample configuration

\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

\[ q_0(a, b, c) \]

\( q_0(a, b, c) \) is the exact sampling distribution when the two loci are unlinked (\( \rho = \infty \)).

\( c_A = (c_i) \)

add to \( a \) to get

\( q^A(a + c_A) \)

\( c_B = (c_j) \)

add to \( b \) to get

\( q^B(b + c_B) \).
Two-locus sample configuration

\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

\[ q_0(a, b, c) \]

\( q_0(a, b, c) \) is the exact sampling distribution when the two loci are unlinked \((\rho = \infty)\).

- Infinite alleles:

  \[ q_0(a, b, c) = q_{ESF}^A(a + c_A)q_{ESF}^B(b + c_B) \]
Two-locus sample configuration

\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

\( q_0(a, b, c) \)

\( q_0(a, b, c) \) is the exact sampling distribution when the two loci are unlinked (\( \rho = \infty \)).

- Infinite alleles:  \( q_0(a, b, c) = q_{\text{ESF}}^A(a + c_A)q_{\text{ESF}}^B(b + c_B) \)
- Finite-alleles PIM:  \( q_0(a, b, c) = q_{\text{WSF}}^A(a + c_A)q_{\text{WSF}}^B(b + c_B) \)
Two-locus sample configuration

\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

\( q_0(a, b, c) \) is the exact sampling distribution when the two loci are unlinked (\( \rho = \infty \)).

- Infinite alleles: \( q_0(a, b, c) = q^A_{ESF}(a + c_A)q^B_{ESF}(b + c_B) \)
- Finite-alleles PIM: \( q_0(a, b, c) = q^A_{WSF}(a + c_A)q^B_{WSF}(b + c_B) \)
- In fact, for any model of mutation,

\[ q_0(a, b, c) = q^A(a + c_A)q^B(b + c_B) \]

where \( q^A, q^B \) are (possibly unknown) one-locus sampling distributions.
Universality

For all mutation models, the functional form of $q_0(a, b, c)$ is

$$q_0(a, b, c) = F(q^A, q^B; a, b, c) := q^A(a + c_A)q^B(b + c_B)$$
Universality

For all mutation models, the functional form of $q_0(a, b, c)$ is

$$q_0(a, b, c) = F(q^A, q^B; a, b, c) := q^A(a + c_A)q^B(b + c_B)$$

Theorem 1

$q_1(n)$ also exhibits universality:

$$q_1(a, b, c) = G(q^A, q^B; a, b, c)$$
Universality

For all mutation models, the functional form of $q_0(a, b, c)$ is

$$q_0(a, b, c) = F(q^A, q^B; a, b, c) := q^A(a + c_A)q^B(b + c_B)$$

Theorem 1

$q_1(n)$ also exhibits universality:

$$q_1(a, b, c) = G(q^A, q^B; a, b, c)$$

$$q_1(a, b, c) = \binom{c}{2} q^A(a + c_A)q^B(b + c_B)$$

$$- q^B(b + c_B) \sum_{i=1}^{K} \binom{c_i}{2} q^A(a + c_A - e_i)$$

$$- q^A(a + c_A) \sum_{j=1}^{L} \binom{c_j}{2} q^B(b + c_B - e_j)$$

$$+ \sum_{i=1}^{K} \sum_{j=1}^{L} (c_{ij}) q^A(a + c_A - e_i)q^B(b + c_B - e_j),$$

where $e_i$ is a unit vector with a 1 at entry $i$ and 0 otherwise.
The full sampling distribution $q(a, b, c | \rho)$ satisfies a recursion relation.

\[ q(a, b, c | \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]
Derivation of $q_1(a, b, c)$

The full sampling distribution $q(a, b, c)$ satisfies a recursion relation.

**Infinite-alleles case:**

$$[n(n - 1) + \theta_A(a + c) + \theta_B(b + c) + \rho c]q(a, b, c) =$$

$$\sum_{i=1}^{K} a_i(a_i - 1 + 2c_i.)q(a - e_i, b, c) + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_j)q(a, b - e_j, c)$$

$$+ \sum_{i=1}^{K} \sum_{j=1}^{L} [c_{ij}(c_{ij} - 1)q(a, b, c - e_{ij}) + 2a_ib_jq(a - e_i, b - e_j, c + e_{ij})]$$

$$+ \theta_A \sum_{i=1}^{K} \left[ \sum_{j=1}^{L} \delta_{a_i + c_i, 1} \delta_{c_{ij}, 1} q(a, b + e_j, c - e_{ij}) + \delta_{a_i, 1} \delta_{c_{ij}, 0} q(a - e_i, b, c) \right]$$

$$+ \theta_B \sum_{j=1}^{L} \left[ \sum_{i=1}^{K} \delta_{b_j + c_j, 1} \delta_{c_{ij}, 1} q(a + e_i, b, c - e_{ij}) + \delta_{b_j, 1} \delta_{c_{ij}, 0} q(a, b - e_j, c) \right]$$

$$+ \rho \sum_{i=1}^{K} \sum_{j=1}^{L} c_{ij} q(a + e_i, b + e_j, c - e_{ij}),$$

with boundary conditions are $q(e_i, 0, 0) = q(0, e_j, 0) = 1$ for all $i \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, L\}$. 

Using that recursion, one can show that this equation admits a probabilistic interpretation as a random variable, and $f$ is a function of the zeroth-order term.
Using that recursion, one can show that the full sampling distribution $q(a, b, c | \rho)$ satisfies
\[
q(a, b, c | \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right)
\]

### Derivation of $q_1(a, b, c)$

1. The full sampling distribution $q(a, b, c)$ satisfies a recursion relation.
2. Using that recursion, one can show that $q_1(a, b, c)$ satisfies
\[
q_1(a, b, c) = f_1(a, b, c) + \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{c_{ij}}{c} q_1(a + e_i, b + e_j, c - e_{ij}),
\]
where $f_1(a, b, c)$ is a function of the zeroth-order term $q_0$. 
Using that recursion, one can show that the equation admits a probabilistic interpretation:

\[ q(a, b, c | \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

**Derivation of** \( q_1(a, b, c) \)

1. The full sampling distribution \( q(a, b, c) \) satisfies a recursion relation.
2. Using that recursion, one can show that \( q_1(a, b, c) \) satisfies

\[ q_1(a, b, c) = f_1(a, b, c) + \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{c_{ij}}{c} q_1(a + e_i, b + e_j, c - e_{ij}), \]

where \( f_1(a, b, c) \) is a function of the zeroth-order term \( q_0 \).
3. This equation admits a probabilistic interpretation:

\[ q_1(a, b, c) = q_1(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_1(A^{(m)}, B^{(m)}, C^{(m)})], \]

where \( C^{(m)} = (C_{ij}^{(m)}) \) is a multivariate hypergeometric \((c, c, m)\) random variable, and \( A^{(m)}, B^{(m)} \) depend on \( C^{(m)} \).
\[ q_1(a, b, c) = q_1(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_1(A^{(m)}, B^{(m)}, C^{(m)})] \]

- Random matrix \( C^{(m)} = (C^{(m)}_{ij}) \) corresponds to a random subsample obtained by sampling without replacement \( m \) gametes from \( c \)

- \( A^{(m)} = a + c_A - C^{(m)}_A \) and \( B^{(m)} = b + c_B - C^{(m)}_B \), where \( C^{(m)}_A = (C^{(m)}_i) \) and \( C^{(m)}_B = (C^{(m)}_j) \)

\[
P \left( \bigcap_{(i,j)} \left[ C^{(m)}_{ij} = c^{(m)}_{ij} \right] \right) = \frac{1}{c \choose m} \prod_{(i,j)} \left( \begin{array}{c} c_{ij} \\ c^{(m)}_{ij} \end{array} \right).\]
\[ q_1(a, b, c) = q_1(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_1(A^{(m)}, B^{(m)}, C^{(m)})] \]

- Random matrix \( C^{(m)} = (C_{ij}^{(m)}) \) corresponds to a random subsample obtained by sampling without replacement \( m \) gametes from \( c \)
- \( A^{(m)} = a + c_A - C_A^{(m)} \) and \( B^{(m)} = b + c_B - C_B^{(m)} \), where \( C_A^{(m)} = (C_{i.}^{(m)}) \) and \( C_B^{(m)} = (C_{.j}^{(m)}) \)

\( f_1 \) is a deg-2 polynomial in the entries \( C_{ij} \) of \( C^{(m)} \), so the above expectation can be easily computed.
\[ q_1(a, b, c) = q_1(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_1(A^{(m)}, B^{(m)}, C^{(m)})] \]

- Random matrix \( C^{(m)} = (C_{ij}^{(m)}) \) corresponds to a \textit{random subsample} obtained by sampling without replacement \( m \) gametes from \( c \)
- \( A^{(m)} = a + c_A - C_A^{(m)} \) and \( B^{(m)} = b + c_B - C_B^{(m)} \), where \( C_A^{(m)} = (C_{i.}^{(m)}) \) and \( C_B^{(m)} = (C_.j^{(m)}) \)

\( f_1 \) is a \textit{deg-2 polynomial} in the entries \( C_{ij} \) of \( C^{(m)} \), so the above expectation can be easily computed.

**Lemma**

For all \( a \) and \( b \), \( q_1(a, b, 0) = 0. \)

That \( q_1(a, b, 0) \) vanishes is not expected \textit{a priori}.
\[
q_1(a, b, c) = q_1(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_1(A^{(m)}, B^{(m)}, C^{(m)})],
\]

**Theorem 1**

For either an infinite-alleles or an arbitrary finite-alleles model,

\[
q_1(a, b, c) = \binom{c}{2} q_A(a + c_A) q_B(b + c_B)
\]

\[
- q_B(b + c_B) \sum_{i=1}^{K} \binom{c_i}{2} q_A(a + c_A - e_i)
\]

\[
- q_A(a + c_A) \sum_{j=1}^{L} \binom{c_j}{2} q_B(b + c_B - e_j)
\]

\[
+ \sum_{i=1}^{K} \sum_{j=1}^{L} \binom{c_{ij}}{2} q_A(a + c_A - e_i) q_B(b + c_B - e_j),
\]

where \(e_i\) is a unit vector with a 1 at entry \(i\) and 0 otherwise.
\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

**Theorem 2**

For either an infinite-alleles or an arbitrary finite-alleles model,
\[
q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})],
\]

where \( f_2 \) is a deg-4 polynomial in the entries \( C_{ij} \) of \( C^{(m)} \).
\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

**Theorem 2**

For either an infinite-alleles or an arbitrary finite-alleles model,

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})], \]

where \( f_2 \) is a deg-4 polynomial in the entries \( C_{ij} \) of \( C^{(m)} \).

**Remarks**

- Found a closed-form formula for \( \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})] \).
\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O \left( \frac{1}{\rho^3} \right) \]

**Theorem 2**

For either an infinite-alleles or an arbitrary finite-alleles model,

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})], \]

where \( f_2 \) is a deg-4 polynomial in the entries \( C_{ij} \) of \( C^{(m)} \).

**Remarks**

- Found a closed-form formula for \( \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})] \).
- Unfortunately, \( q_2(a, b, 0) \neq 0 \) in general and we don't have a closed-form expression for it.
\[ q(a, b, c \mid \rho) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

**Theorem 2**

For either an infinite-alleles or an arbitrary finite-alleles model,

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})], \]

where \( f_2 \) is a deg-4 polynomial in the entries \( C_{ij} \) of \( C^{(m)} \).

**Remarks**

- Found a closed-form formula for \( \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})] \).
- Unfortunately, \( q_2(a, b, 0) \neq 0 \) in general and we don’t have a closed-form expression for it.
- However, it satisfies a simple recursion relation that can be easily solved numerically using dynamic programming.
Theorem 2

For either an infinite-alleles or an arbitrary finite-alleles model,

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})], \]

where \( f_2 \) is a deg-4 polynomial in the entries \( C_{ij} \) of \( C^{(m)} \).

Remarks

- Found a closed-form formula for \( \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})] \).
- Unfortunately, \( q_2(a, b, 0) \neq 0 \) in general and we don't have a closed-form expression for it.
- However, it satisfies a simple recursion relation that can be easily solved numerically using dynamic programming.
- The contribution of \( q_2(a, b, 0) \neq 0 \) is negligibly small compared to that of the other terms.
Classifying the maximum likelihood estimate (MLE)

Application: Classification of the MLE of $\rho$

**Theorem 3 (A sufficient condition for finite MLE)**

For a given sample configuration $(a, b, c)$, if

$$q_1(a, b, c) > 0,$$

then the MLE of $\rho$ is finite.
Classifying the maximum likelihood estimate (MLE)

Application: Classification of the MLE of $\rho$

**Theorem 3 (A sufficient condition for finite MLE)**

For a given sample configuration $(a, b, c)$, if

$$q_1(a, b, c) > 0,$$

then the MLE of $\rho$ is finite.

**Is the converse true?**

If $q_1(a, b, c) < 0$, then is MLE infinite?
Theorem 3 (A sufficient condition for finite MLE)

For a given sample configuration \((a, b, c)\), if

\[ q_1(a, b, c) > 0, \]

then the MLE of \(\rho\) is finite.

Counter-example to the converse

- Finite alleles PIM model, \(\theta = 0.01\).
- \(a = b = 0\),
- \(c = \begin{pmatrix} 6 & 3 \\ 1 & 0 \end{pmatrix}\).
- \(q_1(a, b, c) < 0\).
Classifying the maximum likelihood estimate (MLE)

Application: Classification of the MLE of $\rho$

**Theorem 3 (A sufficient condition for finite MLE)**

For a given sample configuration $(a, b, c)$, if $q_1(a, b, c) > 0$, then the MLE of $\rho$ is finite.

**Counter-example to the converse**

- Finite alleles PIM model, $\theta = 0.01$.
- $a = b = 0$,
  $c = \binom{6}{1} \binom{3}{0}$.
- $q_1(a, b, c) < 0$.

![Graph showing log-likelihood vs. $\rho$]

$q(n)$
Theorem 3 (A sufficient condition for finite MLE)

For a given sample configuration \((a, b, c)\), if \(q_1(a, b, c) > 0\), then the MLE of \(\rho\) is finite.

Example

\[-19.73272406\]
\[-19.73272408\]
\[-19.73272406\]

Counter-example to the converse

\(-19.7326\)
\(-19.7327\)
\(-19.7327\)

\[q_1(a, b, c) < 0.\]
Outline

1 Introduction
   - Motivation

2 Asymptotic Sampling Formula
   - Closed-form one-locus sampling formulas
   - Two-locus sample configuration
   - Universality
   - Details
   - Classifying the maximum likelihood estimate (MLE)

3 Accuracy Results
   - Detailed examples
   - Distribution of errors

4 Conclusion
   - Summary
   - Further work
Truncate at order 2

\[ q_{\text{ASF}}(a, b, c) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} \]

should approximate \( q(a, b, c) \) accurately for sufficiently large \( \rho \).
**Truncate at order 2**

\[ q_{\text{ASF}}(a, b, c) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} \]

should approximate \( q(a, b, c) \) accurately for sufficiently large \( \rho \).

**Example 1**

- \( a = b = 0, \ c = \begin{pmatrix} 10 & 7 \\ 2 & 1 \end{pmatrix} \)
- Finite alleles PIM model, \( \theta = 0.01 \) per locus.
Truncate at order 2

\[ q_{\text{ASF}}(a, b, c) = q_0(a, b, c) + \frac{q_1(a, b, c)}{\rho} + \frac{q_2(a, b, c)}{\rho^2} \]

should approximate \( q(a, b, c) \) accurately for sufficiently large \( \rho \).

Example 2

\( a = b = 0, \ c = \begin{pmatrix} 10 & 1 \\ 2 & 7 \end{pmatrix} \)

Finite alleles PIM model, \( \theta = 0.01 \) per locus.
A measure for accuracy: unsigned relative error

$$\left| \frac{q_{\text{ASF}}(n) - q(n)}{q(n)} \right| \times 100\%$$

Accuracy across all dimorphic \((0,0,c)\) samples of size 20, when \(\rho = 50\) in the symmetric PIM model.

\[
\begin{align*}
q_0(n) \\
q_0(n) + \frac{q_1(n)}{\rho} \\
q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2}
\end{align*}
\]
Accuracy: effect of recombination rate

Accuracy of $q_0(0, 0, c) + \frac{q_1(0, 0, c)}{\rho} + \frac{q_2(0, 0, c)}{\rho^2}$: Fix $c = 20$, vary $\rho$.

Accuracy across all dimorphic samples.

$\theta = 0.01$

$\theta = 1$
For which samples is the ASF most accurate?

\[ \theta = 0.01 \]
For which samples is the ASF most accurate?

\[ \theta = 0.01 \]

- Monomorphomic at both loci.
- At least one allele has multiplicity one.
- Very low LD: \( r^2 \leq 0.02 \).
- Perfect LD.
- Perfect LD except for one haplotype.
- None of the above.
For which samples is the ASF most accurate?

- △ Monomorphic at both loci.
- ▲ Monomorphic at one locus.

\[ \theta = 0.01 \]

- None of the above.
For which samples is the ASF most accurate?

\[ \theta = 0.01 \]

- ▽ Monomorphic at both loci.
- △ Monomorphic at one locus.
- ○ At least one allele has multiplicity one.
- • None of the above.
Distribution of errors

For which samples is the ASF most accurate?

\[ \theta = 0.01 \]

- ▽ Monomorphic at both loci.
- △ Monomorphic at one locus.
- ○ At least one allele has multiplicity one.
- + Very low LD: \( r^2 \leq 0.02 \).

- None of the above.
For which samples is the ASF most accurate?

- ▽ Monomorphic at both loci.
- △ Monomorphic at one locus.
- ○ At least one allele has multiplicity one.
- + Very low LD: $r^2 \leq 0.02$.
- ◊ Perfect LD.

θ = 0.01

- None of the above.
For which samples is the ASF most accurate?

- ▼ Monomorphic at both loci.
- △ Monomorphic at one locus.
- ○ At least one allele has multiplicity one.
- + Very low LD: $r^2 \leq 0.02$.
- ◊ Perfect LD.
- □ Perfect LD except for one haplotype.
- • None of the above.

$\theta = 0.01$
For which samples is the ASF most accurate?

\[ \theta = 0.01 \]  
\[ \theta = 1 \]
Asymptotic series: \( q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \)

Summary of results

1. Found a **closed-form** formula for the first-order term \( q_1(n) \).

2. This formula is **universal** to all mutation models.

3. Found a **closed-form** formula + (**extra bit**) for \( q_2(n) \).

4. The "**extra bit**" is negligibly small and can be easily evaluated using dynamic programming if one wishes.

5. Found a simple **sufficient condition** for a given two-locus sample configuration to have a **finite MLE** of \( \rho \).
Asymptotic series: $q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right)$

Summary of results

1. Found a **closed-form** formula for the first-order term $q_1(n)$.
2. This formula is **universal** to all mutation models.
Summary

Asymptotic series: \( q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \)

Summary of results

1. Found a **closed-form** formula for the first-order term \( q_1(n) \).
2. This formula is **universal** to all mutation models.
3. Found a **closed-form** formula + (extra bit) for \( q_2(n) \).

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})]. \]
Summary

Asymptotic series: \( q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \)

Summary of results

1. Found a **closed-form** formula for the first-order term \( q_1(n) \).
2. This formula is **universal** to all mutation models.
3. Found a **closed-form** formula + (extra bit) for \( q_2(n) \).

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})] \]

4. The “extra bit” is negligibly small and can be easily evaluated using dynamic programming if one wishes.
Asymptotic series:

\[ q(n|\rho) = q_0(n) + \frac{q_1(n)}{\rho} + \frac{q_2(n)}{\rho^2} + O\left(\frac{1}{\rho^3}\right) \]

Summary of results

1. Found a **closed-form** formula for the first-order term \( q_1(n) \).
2. This formula is **universal** to all mutation models.
3. Found a **closed-form** formula + (extra bit) for \( q_2(n) \).

\[ q_2(a, b, c) = q_2(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_2(A^{(m)}, B^{(m)}, C^{(m)})]. \]

4. The “extra bit” is negligibly small and can be easily evaluated using dynamic programming if one wishes.
5. Found a simple **sufficient condition** for a given two-locus sample configuration to have a finite MLE of \( \rho \).
Applications

- Composite-likelihood method such as LDhat. (Monte Carlo methods become substantially less efficient as $\rho$ increases, so our analytic work will be of practical utility.)
- Test for epistasis.
- Sample-wise LD: $P[r^2]$ and $E[r^2]$ as $\rho \to \infty$.

Papers

- Infinite-alleles case:

- Arbitrary finite-alleles recurrent mutation case:
Future work

1. Does there exist a stochastic process “dual” to the coalescent with recombination?

2. Is there a combinatorial interpretation for $q_1(a, b, c)$?
   - cf. Interpretation of Ewens’ sampling formula as the joint distribution of cycle counts in a random permutation

3. Does the universality property extend to $q_k(c)$ for $k > 1$?

4. For all $k > 0$, we think the following is true:

   $$ q_k(a, b, c) = q_k(a + c_A, b + c_B, 0) + \sum_{m=1}^{c} \mathbb{E}[f_k(A^{(m)}, B^{(m)}, C^{(m)})] $$

   where $f_k$ is a deg-2$k$ polynomial in the entries $C_{ij}$ of $C^{(m)}$. Can we automate the computation of the expectation?

5. Does $\sum_{k=0}^{\infty} \frac{q_k(a, b, c)}{\rho^k}$ converge?

6. Extend to multiple loci?
Acknowledgments

Thank you for your attention!

Useful discussions
- Bob Griffiths
- Chuck Langley
- Rasmus Nielsen
- Josh Paul
- Monty Slatkin

Programming
- Fulton Wang

Research supported in part by NIH, Alfred P. Sloan Research Fellowship, and Packard Fellowship for Science and Engineering