Coalescent models with linked selection

Graham Coop and Peter Ralph
Dept. of Evolution and Ecology
UC Davis
Outline

• Genome-wide evidence for hitchhiking
• Multiple merger coalescent of full sweeps
• A multiple merger model of recurrent partial sweeps
• A simultaneous multiple merger model of recurrent soft sweeps
The effect of selective sweeps on linked neutral variants

Maynard Smith and Haigh, Kaplan et al ‘89, etc

Selective sweep results in a characteristic reduction in coalescent time at linked neutral sites. Also a distortion in the genealogical tree towards external branches and away from internal branches.

Background selection can also lead to a reduction in diversity, but lead to only a weak skew towards rare alleles.
Evidence for linked selection in *Drosophila melanogaster*

e.g. Shapiro et al 2007
Evidence for variation-reducing selection in humans

But not clear what mode of linked selection acts.

Cai et al. 2009

Lohmueller et al., 2011
Time-scale of selective sweep = $t = \frac{2\log(2N_e)}{s}$

Probability of failing to recombine off $q = \exp(-rt/2)$

Probability that $i$ out of $k$ lineages are forced to coalesce $\sim \text{Binom}(k,q)$

Maynard Smith and Haigh, Barton, 1998; Durrett and Schweinsberg, 2004, etc

Barton, 1998; Durrett and Schweinsberg, 2004; Etheridge et al., 2006; Pfaffelhuber et al., 2006, …
Sweeps occur at rate $\nu$ with $q \sim f(q)$ a iid r.v. across sweeps. $i$ lineages out of $k$ lineages forced to coalesce at rate:

$$\lambda_{k,i} = \binom{k}{2} \frac{1}{2N} \delta_{i,2} + \nu I_{k,i} \quad \text{for } 2 \leq i \leq k,$$

$$I_{k,i} = \binom{k}{i} \int_0^1 q^i (1 - q)^{k-i} f(q) dq.$$

Gillespie ‘00, Durrett & Schweinsberg 05

Lambda coalescent:

$$\Lambda(dq) = q^2 \nu f(q) dq + \delta_0(dq)/2N$$
Homogeneous sweeps at rate $\nu_{BP}$, recombination at rate $r_{BP}$.
Then $i$ out of $k$ lineages coalesce at rate:

$$\lambda_{k,i} = \frac{1}{2N} \binom{k}{2} \delta_{i,2} + \frac{\nu_{BP}}{r_{BP}} J_{k,i} \quad \text{for } 2 \leq i \leq k,$$

Kaplan et al. 1989,

$$J_{k,i} = \binom{k}{i} \int_0^\infty q(r)^i (1 - q(r))^{k-i} dr$$

Durrett & Schweinsberg 05

$$\mathbb{E}(\pi) = 2u\mathbb{E}(T_2) = \frac{4Nu}{1 + 2N\nu_{BP} J_{2,2}/r_{BP}}$$

Kaplan et al. (1989) and Stephan et al. (1992)
\[ r_{BP} = \text{Recombination rate cM/kb} \]

\[ \mathbb{E}(\pi) = 2u\mathbb{E}(T_2) = \frac{4Nu}{1 + 2Nu_{BP}J_{2,2}/r_{BP}} \]

Kaplan et al. (1989) and Stephan et al. (1992)
What if most newly arisen selected alleles do not sweep to rapidly fixation?

E.g. due to changing environment or genomic background (Due to parallel mutation, other standing variation etc)

Pennings and Hermisson, 2006a,b; Chevin and Hospital, 2008; Ralph and Coop, 2010, Innan and Kim, 2004; Hermisson and Pennings, 2005; Przeworski et al., 2005
What if most newly arisen selected alleles do not sweep to rapidly fixation?
E.g. due to changing environment or genomic background (Due to parallel mutation, other variation etc)
The derived allele arose $\tau$ Generations ago

Conditions on trajectory: Selected allele initially quickly increases in frequency. If it approaches 0 or 1 it does not reenter the Population.

$X(t)$ is the frequency of the Derived allele at time $t$
Imagine a neutral site a genetic distance $r$ away from the selected locus

Probability that the lineage is of the derived type at time $0 = q(r, X) = r \int_0^\tau e^{-r t} X(t) \, dt$

For $r \tau >> 1$
\[ q(r, X) = r \int_0^\tau e^{-rt} X(t) \, dt \]

Probability that \( i \) out of \( k \) lineages are force to coalesce is binomial:

\[
\binom{k}{i} q^i (1 - q)^{k-i},
\]

for \( 2 \leq i \leq k \),

Assuming that the all coalescence happens close to time 0, \( rN >> 1 \)
Simple trajectories

Selected allele moves quickly from $1/2N$ to $x$ in time $t_x$

Then stays at $x$, or goes to fixation, or loss on a slower time-scale (e.g. with selection coefficient $s_2$, $-s_2$, or 0 respectively)

$$q \approx xe^{-rt_x}$$

Also holds for other trajectories when $r >> s_2$

$$\mathcal{E}(T_2) = 2N(1 - q_x^2e^{\tau/(2N)})$$
A. Multiple mergers coalescent

B. Coalescent with trajectory
Recurrent sweep process

- Assume Neutral pairwise rate of coalescence: $1/(2N)$
- Sweeps happen at rate $\nu$
- At a fixed position, with constant $q$
- Total rate of coalescence of $i$ out of $k$:

$$\lambda_{k,i} = \binom{k}{2} \frac{1}{2N} \delta_{i,2} + \nu I_{k,i} \quad \text{for} \ 2 \leq i \leq k,$$

$$I_{k,i} = \binom{k}{i} q^i (1 - q)^{k-i}.$$
\[ E(T^2) = \frac{2N}{1 + 2N \nu q^2} \]

- For our simple approximation \( q \approx xe^{-rt_x} \)

\[
\begin{align*}
\text{Position, } 4N\nu & \quad 0 \quad 200 \quad 400 \quad 600 \quad 800 \quad 1000 \\
\text{0.2} & \quad 0.4 & \quad 0.6 & \quad 0.8 & \quad 1.0 \\
\text{x = 0.8} & \quad \text{--- Recurrent top-hat traj.} & \quad \text{--- Recurrent step traj.} \\
\text{x = 0.8} & \quad 4N\nu = 1, & \quad 4N\nu = 2, & \quad 4N\nu = 4, & \quad \text{Approx.}
\end{align*}
\]
Run mssel for recurrent top-hat trajectories for 20 sequences

$$2 \frac{r}{s} \log(2N) = y = 0.61$$

Calculate for partial sweep coalescent

$$q = x e^{-y}$$

Reduction in diversity

$$\frac{\mathbb{E}(T_2)}{2N}$$
Homogeneous sweeps at rate $\nu_{BP}$, recombination at rate $r_{BP}$. Then $i$ out of $k$ lineages coalesce at rate:

$$
= \frac{\nu_{BP}}{r_{BP}} J_{k,i} \quad \text{for } 2 \leq i \leq k,
$$

$$
J_{k,i} = \binom{k}{i} \mathbb{E}_X \left[ \int_0^\infty q(r, X)^i (1 - q(r, X))^{k-i} \, dr \right]
$$

Where $J_{k,i}$ depend only on the form taken by trajectories
So rate of coalescence controlled by $\frac{\nu_{BP}}{r_{BP}}$

E.g. for our simple trajectory $J_{k,i}$ is a function of $x$ (freq. sweeps achieve) and so number of lineages forced to coalesce by $x$ (or distribution on $x$).
Under our simple partial sweep model: $J_{2,2} = \frac{x^2}{t_x}$

$t_x = 1000$ gens \hspace{1cm} \(s \sim 0.1\%\), \(N = 10^6\), $v_{BP}$ $x^2 = 3 \times 10^{-13}$

$x = \begin{array}{ccc} 100\% & 20\% & 5\% \\ \end{array}$

$v_{BP} = \begin{array}{ccc} 3 \times 10^{-13} & 8 \times 10^{-12} & 1 \times 10^{-10} \text{ per generation} \\ \end{array}$

$\Sigma = \text{Syn. Diversity (\%)}$

$r_{BP} = \text{Recombination rate cM/kb}$

$E(\pi) = 2uE(T_2) = \frac{4 Nu}{1 + 2Nv_{BP}J_{2,2}/r_{BP}}$

Data from *Drosophila melanogaster* (Shapiro et al 2007)

$2Nv_{BP}J_{2,2} = 7 \times 10^{-9}$

Assuming none of the reduction is due to Background Selection
For same reduction in diversity we can get very different distortions to frequency spectrum

\[ \frac{\pi}{(4Nu)} = 0.1 \]

\[
F_{n,k}^N = \mathbb{E}(\text{Fraction of sites seen in } k \text{ out of } n)
\]

Under Kingman coalescent

\[
F_{n,k}^N = \frac{1}{k} / \sum_{j=1}^{n-1} \left( \frac{1}{j} \right)
\]
For same reduction in diversity we can get very different distortions to frequency spectrum

\[
\pi / (4Nu) = 0.1
\]
Soft Sweeps

Selection on multiple mutations either standing or new

Hermisson and Pennings 05, Pennings and Hermisson 06

Selection on standing variation

Przeworski, Coop and Wall 2005
Kim and Innan 05
Soft Sweeps

Pennings and Hermisson showed:
Mutation rate at selected site = $\rho$

At selected site: Lineages assigned to coalescent families (tables) following infinite alleles model with param. $4N\rho$

$q = e^{-rt}$

Where $t$ = time of sweep

Remaining lineages assigned to coalescent families

Hermisson and Pennings 05,
Pennings and Hermisson 06
Recurrent Soft Sweeps

Neutral coalescence at rate $1/(2N)$

Sweeps occur at rate $\nu_{BP}$ homogeneously along sequence recombining at rate $r_{BP}$
i out of k lineages caught in sweep at rate:

$$\binom{k}{i} \frac{\nu_{BP}}{t r_{BP}} \int_{0}^{\infty} (e^{-r})^i (1 - e^{-r})^{k-i} \, dr$$

The i lineages are then forced into coalescence families according to infinite alleles model with parameter $4N\rho$

$$\mathbb{E}[\pi] = \frac{\theta}{1 + 2N\nu_{BP} J_{2,2} / (r_{BP}(1 + 4N\rho))}$$
Recurrent Soft Sweeps

\[ \frac{\pi}{(4Nu)} = 0.1 \]
Conclusions

• A broad range of linked selection models can be approximated by coalescent models with multiple mergers

• Range of biological models of linked selection depressingly large and predictions overlap.

• Idea: Rather than estimating one model why not estimate rates of different types of coalescence across genome.
What we need

• Given that the rate of sweeps differs across the genome, what can we hope to learn about the multiple merger process?
• We need theory to predict frequency spectra and haplotype patterns under these models.
• What set of statistics are most informative?
• What set of coalescent processes can we hope to distinguish?
Thanks

Peter Ralph

Thanks to Yaniv Brandvain, Chuck Langley, Molly Przeworski, Alisa Sedghifar, and Guy Sella for helpful conversations
• For our simple approximation \( q \approx xe^{-rt_x} \)

\[
E(T) = \frac{1}{1 + 2Nvq^2}
\]

Simulate ms sel with either

Loss trajectory motif

fixed trajectory motif

Repeat motif with waiting time \( \sim \text{Exp}(v) \) between them
Evidence for variation-reducing selection in humans
But not clear what mode of linked selection acts.

Polymorphism, $\pi$

Average minor allele frequency

Cai et al. 2009

Lohmueller et al., 2011
Matching the reduction in $\pi$ the distortion to the site frequency spectrum

**Graph**

- **Y-axis**: Fraction of singleton sites
- **X-axis**: Recombination
- **Legend**
  - Frequency selected alleles sweep to
  - 100%
  - 50%
  - 20%
  - 5%

**Sample size = 20**
Soft sweep model due to Parallel mutation during sweep

Fail to recombine off derived background, forced to coalesce
Conclusions

• P
Hellmann et al using similar data

\[ \pi \approx \frac{r_{BP} \pi_0}{r_{BP} + \alpha} \]

Estimated \( \pi_0 = 1.6 \times 10^{-3} \), \( \alpha = 6 \times 10^{-11} \)

Assuming none of the reduction is due to BS

### Data from Humans

- **e.g. Cai et al 2009**

**Rate of recombination (cM/Mb)**

\[ \alpha = 2Nv_{BP} \left( \frac{x^2}{t_x} \right) \]

\[ t_x = 1000 \quad (s\sim1\%) \]

\[ N = 10000 \]

\[ v_{BP} x^2 = 3 \times 10^{-12} \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>100%</th>
<th>50%</th>
<th>20%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v_{BP} )</td>
<td>3e-12</td>
<td>1e-11</td>
<td>8e-11</td>
<td>1e-09 !!!</td>
</tr>
</tbody>
</table>

Note humans need a high sweep rate despite smaller effect of HH.
Solid coloured line recurrent loss trajectory.
Dashed coloured line recurrent fix trajectory

\[ t_x / 2N = 0.0015 \]
Pauses for 0.02 (2N generations)