Cytoplasmic Tree (Maternal)
Cytoplasmic Tree (Maternal)

Zygotic Pedigree (bi-parental)
Cytoplasmic Tree (Maternal) \[ \tau = 0 \]

Zygotic Pedigree (bi-parental) \[ \tau = 1 \]

Sub-Karyotic Pedigree \[ 0 < \tau < 1 \]

Common Prob. Space from a Markov Chain on Multi-set Partitions.
Hermaphrodite Wright-Fisher (Chang's Model) \[ \tau_n, \mu_n \]

Haploid Wright-Fisher

\[ \tau_n = \mu_n = 4 \]

\[ n \to \infty \quad r = 0 \]

Kingman's Coalescent

\[ \text{what if } 0 < r < 1 \]

Recombining Wright-Fisher

\[ n \to \infty \]

[4] = \{1, 2, 3, 4\} is C.A.

48 gens. Pedigree (10 gens DNA seqn.) of FL Scrub Jay
Consider a population of constant size $n$, $3n + 1$.
Let's ask (not done before)
What is the discrete combinatorial structure underpinning the recombinating Wright-Fisher Model?

Let pop. size be $2n$ at every generation.
Let Label Set be $[2n] := \{1, 2, \ldots, 2n\}$ at gen. $k$

$$V_i = \# \text{ of non-recomb. offspring of indiv. } i \text{ (of gen } k\text{)}$$

$$U_{i,j} = \# \text{ of recombinant } \text{pair of indiv. } i,j \text{ (of gen } k\text{)}$$

$$V_0 = \sum_{i=1}^{2n} V_i$$

$$U_{0,0} = \sum_{i,j \in [2n] : i \neq j} U_{i,j}$$

Constant pop. size $\Rightarrow V_0 + U_{0,0} = 2n$

Then $(V, U) = (V_1, V_2, \ldots, V_{2n}, U_{1,2}, U_{1,3}, \ldots, U_{2n-1,2n})$

is Multinomial:

$$\Pr((V, U) = (v, u)) = \frac{(2n)!}{v_1! \cdots v_{2n}! \cdots u_{1,2}! \cdots u_{2n-1,2n}!} (1-r)^V_0 r^{U_{0,0}}$$

This reproduction scheme is enforced in each gen.

Special case ($r=0$) (non-recomb. W-F Model):

$V_0=2n$, $U_{0,0}=0$ and $\Pr((V, U) = (v, u)) = \frac{\binom{2n}{v} v_1! \cdots v_{2n}!}{\binom{2n}{v_0} V_0!}$

Special case ($r=1$) (Chains (Zygote) pedigree):

$V_0=0$, $U_{0,0}=2n$ and $\Pr((V, U) = (v, u)) = \frac{\binom{2n}{v} (1-r)^V_0}{\binom{2n}{v_0} V_0!}$
Now \(\text{Ancestral size Markov Chain.} (\text{backwards in time})\) 
\(\text{never done before exactly}\)

\[ \begin{align*}
\text{# of diploid indiv.} \\
\text{per gen prob of recomb} \\
\text{# gens ago} \\
\{2n, r\} \\
X(t) \quad t \in \mathbb{Z} \\
\overline{X} := \{1, 2, \ldots, 2n\}
\end{align*} \]

Thm 1: The one-step transition prob matrix of \(\{2n, r\} X(t)\) is:
\[ P_{i,j} = \binom{2n}{j} \sum_{k=0}^{i} \binom{i}{k} r^k (1-r)^{i-k} \frac{\sum_{m=0}^{i} (-1)^m \binom{i}{m} (j-m-r^k) (j-m-1-k)}{(2n)^i (2n)^k} \]

Proof:
\[ \text{Fix sets } I, J, K, \text{ at: } |I| = j, |J| = i, |K| = k, K \subseteq I \backslash K \text{ non-recomb.} \]

\[ P(j \mid I) = \sum_{J: |J| = j} P(J \mid I) = \binom{2n}{j} P(J \mid I) \]

\[ P(J \mid I) = \sum_{K \subseteq I} r^{\mid K \mid} (1-r)^{|I|-|K|} \]

\[ P(J \mid I, K) = \frac{\binom{b(J \mid I, K)}{1}}{\binom{|I|-|K|}{2}} \]

Set of bipartite graphs with vertices \(I \cup J, K\), with bipartition \(J \mid I, K\), s.t.

\[ \text{vertices in } K \text{ are of degree 2,} \]
\[ \text{# of ways in which indiv in } I \]
\[ \text{I} \backslash K \text{ are not isolated (and } \text{K are not} \]

all together: we have proved Thm 1.

\[ P_{i,j} = P(j \mid i) = \binom{2n}{i} \sum_{k=0}^{i} \binom{i}{k} r^k (1-r)^{i-k} \frac{\binom{b(i \mid i, k)}{1}}{(2n)^{i-k} (2n)^2} \]
Thm 2. Let \( T_n \) = \# of gens. to MRCA of all present-day indivs. at a recombining locus (per gen. rebomb. prob. \( r \)) in diploid popn. of size \( n \).

Then \( \forall \, \varepsilon > 0, \)

\[
\lim_{n \to \infty} \left( 1 - \varepsilon \right) \frac{C(r)}{\ln n} \leq T_n \leq \left( 1 + \varepsilon \right) \frac{C(r)}{\ln n} \right\} = 1
\]

\[ C(r) = \frac{1}{\ln(1+r)} - \frac{1}{\ln(1-r)} \]

\( \Rightarrow \) because it takes more time to reach \( n \) (unlike Chang's case with \( r=1 \))

this term is missing in Wiuf-Hien conjecture.

Thm 3. \( U_n \) = \# of gens. ago when each indiv. is either a CA or not a CA of all present-day indivs.

\[ P = p(x) \] be unique soln. in \( (0,1) \) to \( x = e^{-C(r)(1-x)} \)

i.e. Prob \{ extinction of GWW(Pois(\( C(r) \)) \}

Then \( \forall \, \varepsilon > 0, \)

\[ \lim_{n \to \infty} \left( 1 - \varepsilon \right) \frac{C(r)}{\ln n} \leq U_n \leq \left( 1 + \varepsilon \right) \frac{C(r)}{\ln n} \right\} = 1
\]

Extra time is due to waiting for all families to have 0 or \( (\ln n)^2 \) descendants.

Covr. 4. The fraction of popn. at \( U_n \) that are CAs is \( 1 - p(x) \).

Thm 5. Griffith's ARG is the asymptotic approx. of \( \{ X_t \} \)

(Hudson)

based on a random sample as \( n \to \infty \) but \( n \) is held constant.
\[ n = 10^4, \ r = 0.1 \]

\[
\begin{array}{ccc}
T_n, U_n, \text{ CA}_n/n \\
117 & 223 & 0.17 \\
120 & 232 & 0.18 \\
129 & 228 & 0.18 \\
123 & 215 & 0.18 \\
121 & 210 & 0.17 \\
\hline
184 & [2764, & 0.02 \\
\end{array}
\]

\[ n = 10^5, \ r = 0.5 \]

\[
\begin{array}{ccc}
T_n, U_n, \text{ CA}_n/n \\
29 & 60 & 0.58 \\
30 & 53 & 0.58 \\
29 & 61 & 0.58 \\
29 & 51 & 0.58 \\
29 & 52 & 0.58 \\
\hline
36 & [55, & 0.58 \\
\end{array}
\]

\[ n = 10^6, \ r = 0.9 \]

\[
\begin{array}{ccc}
T_n, U_n, \text{ CA}_n/n \\
16 & 30 & 0.76 \\
16 & 32 & 0.76 \\
16 & 30 & 0.77 \\
16 & 31 & 0.77 \\
16 & 30 & 0.77 \\
\hline
18 & [30, & 0.77 \\
\end{array}
\]

Note: if \( r = 0.01 \) then \[ \frac{820}{930} \] \( \approx \frac{1560}{1800} \) \( \approx 0.02 \) so \( T_n \) increases as \( r \) decreases.

Also:

\[
\prod_{i=1}^{n} a_i = \lim_{n \to \infty} \frac{a_1 \cdot a_2 \cdots a_n}{n} \to 1 - p(r)
\]

Nowhere near asymptotic regime .... (proof needed).

Now consider a Natural Embedding (a thinning scheme).

**Ancestors levels**

- **Level 0**: \( \frac{r = 1}{1} \) (all genomic content)
  - Let \( w \) be a sample pedigree \( (r = 1) \).
  - Consider a fragment of recombinant DNA, with \( r = 0.9 \).
  - Get \( w' \) from \( w \) by thinning.

- **Level 1**: Consider a sub-fragment, with \( r = 0.5 \).
  - Get \( w'' \) from \( w' \) by further thinning.

- **Level 2**: Consider a different large fragment, say chr. 2.
  - Consider its sub-fragment ...
\(G_t^i = \text{set of descendants of } I_0, i \text{ after } t \text{ generations.}\)

\[|G_t^i| = \# \text{ of descendants of } I_0, i\]

\[\{G_t^i\}_{t \in \mathbb{Z}^+}\]

is a M.C. with tr. pr.

\[\text{super-critical}\]

\[\left( G_{t+1}^i \mid G_t^i \right) \sim \text{Bin} \left( n, \frac{g_t^i}{n} \right) - r \left( \frac{(g_t^i)^2}{n^2} \right)\]

\[\left( 1 - r \right) \frac{g_t^i}{n} + r \left( 1 - \left( \frac{g_t^i}{n} \right)^2 \right)\]

1. Stage G1

\[\text{after } T_n \approx 2 \ln \ln n / \ln(1+r) \text{ gens. with high prob.}\]

2. Stage G2

\[\text{\# reaches at least } (\ln n)^2.\]

3. Stage G3

\[\text{\# increases from } (\ln n)^2 \text{ to } g_2 n \text{ in } T_n \approx \frac{\ln n}{\ln(1+r)} \]

4. Stage B1

\[\text{\# decreases from } a/2 \text{ to } b/n \text{ in } T_n \approx \ln \ln n \text{ gens.}\]

\[P \left( T_n \geq \ln \ln n \right) = 1 - O(1/n)\]

5. Stage B2

\[\text{further decreases from } b/n \text{ to } (\ln n)^2 \text{ in } T_n \approx -\ln n / \ln(1-r)\]

6. Stage B3

\[\text{go extinct from } (\ln n)^2 \text{ to } 0 \text{ in } T_n \approx -2 \ln \ln n / \ln(1-r) \text{ gens.}\]
All 6 stages together gives, with prob → 1 as \( n \to \infty \),

The first time when an indiv. becomes a C.A.

\[ \text{Proof: stage G1} \]

Let \( \{ X_t \}_{t \geq 0} \) be some \( \mathbb{Z}^+ \)-valued process,

where \( \tau^X_s = \inf \{ t : X_t = s \} \).

Show with high prob. within \( o(n) \), gens. \( (ln(n))^2 \) descendants of an indiv. from \( t = 0 \).

Idea: Approx. "# of descendants of G process" \( \{ G_t \}_{t} \) using \( \{ Y_t \}_{t} \)

provided we look early enough & with popn size small enough.

Lemma 17(18). If \( k_n \to \infty, b_n \to \infty \) as \( n \to \infty \) s.t. \( k_nb_n = o(n) \)

Then \( P_{\tau_b^G} > k_n \) = \( P_{\tau_b^Y} > k_n \) \((1 + o(1)) \) \( \ast \)

Starting at \( Y_0 = G_0 = 1 \)

Proof: Bounding the ratio of transition probs. of \( G_n \) & \( Y^+ \) we can get

\[ P \{ \tau_b^Y > k_n \}e^{-c_k b_n^2/n} \leq P \{ \tau_b^G > k_n \} \]

so need \( k_nb_n = o(n) \) for \( \ast \) to hold.

Take \( b_n = (ln(n))^2 \)

\[ \tilde{b_n} = \frac{3 \ln \ln n}{\ln(1 + r)} \]

\[ m_0 = \ln \ln n \] many geometric trials

By \( \ast \) & a non-neg. Martingale argument:

\[ \lim \inf P \{ \tau(b_n)^2 \leq \frac{3 \ln(\ln n)}{\ln(1 + r)} \} = c > 0 \]

\( \{ M_{t} = \{ Y^+_{(t+1)} \} \} \) is a non-neg. mart \( a.s. \) to \( M_\infty \), \( P \{ M_\infty = \emptyset \} = p < 1 \)

\[ \lim \sup P \{ \tau(b_n)^2 \leq \frac{3 \ln(\ln n)}{\ln(1 + r)} \} \leq \lim \sup P \{ \tau(b_n)^2 \leq \frac{3 \ln \ln n}{\ln(1 + r)} \} \]

\[ \leq \lim \sup P \{ M_{\frac{3 \ln \ln n}{\ln(1 + r)}} < \ln(n) \} \]

\[ \leq P \{ M_{\frac{3 \ln \ln n}{\ln(1 + r)}} < \ln(n) \} = P \{ M_{\infty} = \emptyset \} = p < 1 \]
Thus, \( R \{ \text{no indv. at time } 0 \text{ has at least } (\ln n)^2 \} \cap \left( \frac{\ln n}{\ln(1+r)} \right)^{1-c} \to 0 \text{ as } n \to \infty \).

Next we study the family 0 of such a thriving indv from generation 0 with \( \geq (\ln n)^2 \) descendants. By time 0(\ln n),

**Stage G2.** \( G_t \geq (\ln n)^2 \to G_t \geq g_2 n \), \( g_2 \in (0, \frac{1}{2}) \)

**Idea.** \( G_t \) is large enough to behave like its expectation.

**Use Bernstein's Inequality.** If \( X \sim \text{Bin}(m, p) \), \( x > 0 \), then

\[
P \{ X \geq np + x \} \leq \exp \left( \frac{-x^2}{2np(1-p) + x/3} \right)
\]

and

\[
P \{ X \leq np - x \} \leq \exp \left( \frac{-x^2}{2np(1-p) + x/3} \right)
\]

With \( x = \eta (G_{t-1} - G_t^2/n) > 0 \), \( \eta > 0 \) s.t.

\[
\ln (1+r-\eta) > \frac{\ln(1+r)}{1+\varepsilon/2}
\]

in

\[
P \{ G_{t+1} \leq (1+r-\eta) G_t, (\ln n)^2 \leq G_t \leq g_2 n \} \leq \exp \left( -\frac{(\eta - g_2^2) (\ln n)^2}{2(1+r) + \eta/3} \right) \to 0
\]

If \( G_{t+1} \geq (1+r-\eta) G_t \) \( \forall t \leq m_n = \left[ \frac{\ln n - 2 \ln n + 1}{\ln(1+r-\eta)} \right] \),

Then \( G_{m_n} \geq (1+r-\eta)^{m_n} G_0 \geq g_2 n \) is satisfied.

Thus,

\[
P \{ G_t < g_2 n, \forall t \leq m_n | G_0 \geq (\ln n)^2 \} \leq \sum_{t=0}^{m_n} P \{ G_t \geq g_2 n \} = o(\frac{1}{n})
\]

Let

\[
\zeta_n = \inf \{ t : G_t \geq g_2 n \} \quad \text{and} \quad \zeta_0 \geq (\ln n)^2
\]

Then,

\[
P \{ \zeta_n < (1-\varepsilon) \ln n \ln(1+r) \} = o(\frac{1}{n})
\]

as \( n \to \infty \)

\( \text{(Stage G2 done)} \)
Stage G3 from $G_t \geq g_2 n$ to $G_t \geq n/2$.

Same idea in place of stage G2 gives:

Let $G_0 \geq g_2 n$, where $0 < g_2 < \frac{\varphi}{r}$, $\ln(1+r-\varphi) > \frac{\ln(1+r)}{1 + \epsilon/2}$.

$\mathcal{I}_n = \inf \{t : G_t \geq n/2\}$

Then, as $n \to \infty$,

$P\{\mathcal{I}_n \geq \ln \ln n\} = o\left(\frac{1}{n}\right)$

Amandine's lemma

Detailed Version

Consider $g_t = \frac{G_t}{n}$, $E(g_{t+1} | g_t) = \frac{[n(1+r)G_t - rG_t^2]}{n^2} = (1+r)g_t - \frac{rG_t^2}{n} = \frac{g_t(1+r-rg_t)}{1 + \frac{\epsilon}{2}g_t}$

So, need $g_2 n \left[1 + \frac{r}{3}\right] = \frac{n}{2}$.

Thus, $m_n = \log_{1 + \frac{r}{2}} \left(\frac{n}{2} - \frac{1}{g_2 n}\right) = \log_{1 + \frac{r}{2}} \left(\frac{1}{2g_2}\right)$

Let $G_0 \geq g_2 n$, $\mathcal{I}_n = \inf \{t : G_t \geq \frac{n}{2}\}$

Then $P\{\mathcal{I}_n \geq \log_{1 + \frac{r}{2}} \left(\frac{1}{2g_2}\right)\} = o\left(\frac{1}{n}\right)$ as $n \to \infty$.

Proof:

For $g_2 n \leq G_t \leq \frac{n}{2}$, by Bernoulli,

$P\{G_{t+1} \leq (1 + \frac{r}{3})G_t | G_t\} \leq \exp\left(-C(\Delta)g_2 n\right)$

So, if $\mathcal{I}_n \geq \log_{1 + \frac{r}{2}} \left(\frac{1}{2g_2}\right)$ then there was some $t \leq \log_{1 + \frac{r}{3}} \left(\frac{1}{2g_2}\right)$ such that $g_2 n \leq G_t \leq \frac{n}{2}$. Thus, $P\{\mathcal{I}_n \geq \log_{1 + \frac{r}{3}} \left(\frac{1}{2g_2}\right)\} \leq \log_{1 + \frac{r}{3}} \left(\frac{1}{2g_2}\right) e^{-C(\Delta)g_2 n} = o(1)$ as $n \to \infty$.
Stage B.1. From $B_t \leq n/2$ to $B_t < b, n$, $b, \epsilon (0, 1/2)$

$B_t = n - G_t = \# of non-desc. of I$

$- (B_{t+1} | B_t) \sim \text{Bin} \left( n, (1-r) \frac{B_t}{n} + r \frac{B_t^2}{n} \right)$

Here, $B_t$ decreases nearly deterministically (being of order $n$) at a rate $\geq 1 - \frac{r + \sqrt{r/2}}{1 - r/2} = 1 - \frac{r}{2}$.

(Exactly same as in Cn3 phase) we get: $\tau^{(B)}_n = \inf \{ t : B_t \leq b, n \}$

If $B_0 \leq n/2$ and $b, \epsilon (0, 1/2)$

then as $n \to \infty$ $P \{ \tau^{(B)}_n > \ln \ln n \} = o \left( \frac{1}{n} \right)$.

Stage B.2. From $B_t \leq b, n$ to $B_t \leq (\ln n)^2$

Proceeding as in Cn2.

Let $\eta \in (0, 1-r)$ and fix $b_1 > 0$ s.t. $\eta > rb_1$, By Bernst,

$P \{ B_{t+1} > (1-r+\eta) B_t, (\ln n)^2 \leq B_t \leq b_1, n \} \leq \exp \left\{ - \frac{(\eta - rb_1)^2 (\ln n)^2}{2 + 2 \eta / 3} \right\}$

Let $m_n = \left[ \frac{\ln n - 2 \ln \ln n + \ln b_1}{-\ln (1-r+\eta)} \right]$. 

If $B_{t+1} \leq (1-r+\eta) B_t$, $\forall t \leq m_n$,

Then $G_{m_n} \leq (1-r+\eta)^{m_n} B_0 \leq (\ln n)^2$

so, $P \{ B_t > (\ln n)^2, \forall t \leq m_n \mid B_0 \leq b_1, n \} \leq m_n e^{-C'(\tau, \eta, b)(\ln n)^2} = o \left( \frac{1}{n} \right)$

and if we choose $\eta > 0$ and $b, \epsilon (0, 1/2)$ s.t. $\ln (1-r+\eta) \leq \frac{\ln (1-r)}{1+\epsilon/2}$ and $b_1 < \frac{1}{r}$, then we get:
Lemma 11. If \( B_0 \leq b, n \), \( T_n^{(B_2)} = \inf \{ t : B_t \leq (\ln n)^2 \} \). Then as \( n \to \infty \),

\[
P \left\{ T_n^{(B_2)} > \left( 1 + \frac{\epsilon}{2} \right) \frac{\ln n}{-\ln (1-r)} \right\} = o \left( \frac{1}{n^2} \right)
\]

Lemma 12. Likewise, (up to taking smaller \( r \) \& \( b \)).

If \( n / \log n \leq B_0 \leq b, n \). Then as \( n \to \infty \),

\[
P \left\{ \frac{\ln n}{-\ln (1-r)} \leq T_n^{(B_2)} \right\} = o \left( \frac{1}{n^2} \right)
\]

Stage B3. Extinction of \( \{ B_t \}_{t \in \mathbb{Z}_+} \).

Let \( B_0 = (\ln n)^2 \), \( T_n^{(B_3)} = \inf \{ t : B_t = 0 \} \).

Lemma 17 (iii) : If for some \( \alpha \in (0, \frac{1}{2}) \), \( Y_t \sim (2\alpha, \frac{1}{2}) \) we have \( i = O(n^\alpha) \) and \( k = o(n^{1-2\alpha}) \), then as \( n \to \infty \),

\[
P_i \left\{ T_0 > k \right\} = P_i \left\{ T_0 > k \right\} (1 + o(1))
\]

\[
\inf \{ t : B_t = 0 \}. \quad Y^* \sim GW(Pois (1-r))
\]

By Lemma 15, \& branching. Prep : \( \mathcal{B}_0 \)

\[ T_0 = \inf \{ t : Y_t = 0 \} \]

\[ Y_t \sim GW(\text{mean } m) \quad m < 1 \]

Then, \( k \in \mathbb{Z}_+ \),

\[ P_i \left\{ T_0 > k \right\} < m^k \]

By Taylor exp.

\[ e^{(\ln n)^2} (1 - (1-r)) (\ln n)^2 \]

\[ \rightarrow 1 \quad \text{as } n \to \infty \] if \( C > \frac{2}{\ln (1-r)} \).