Introgression & Deriving Patterson's D

BIOL 1435

March 7, 2023

Overview

1. Motivation

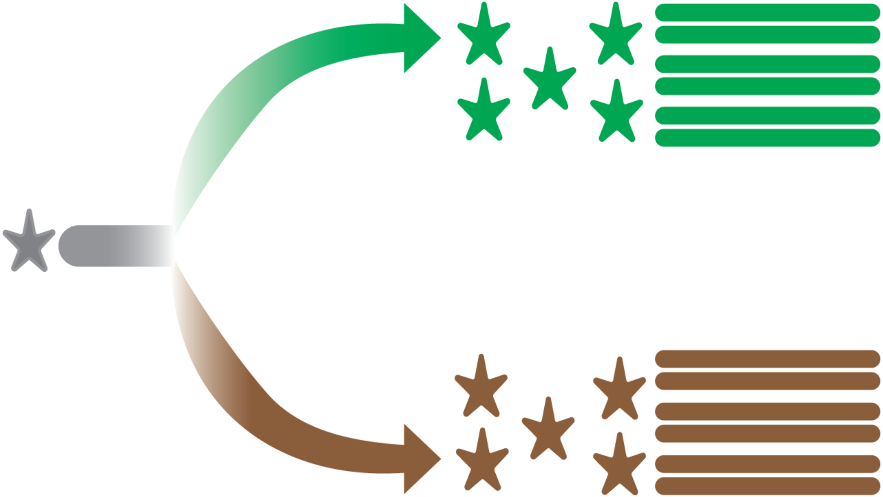
2. Preliminaries

3. Derivations

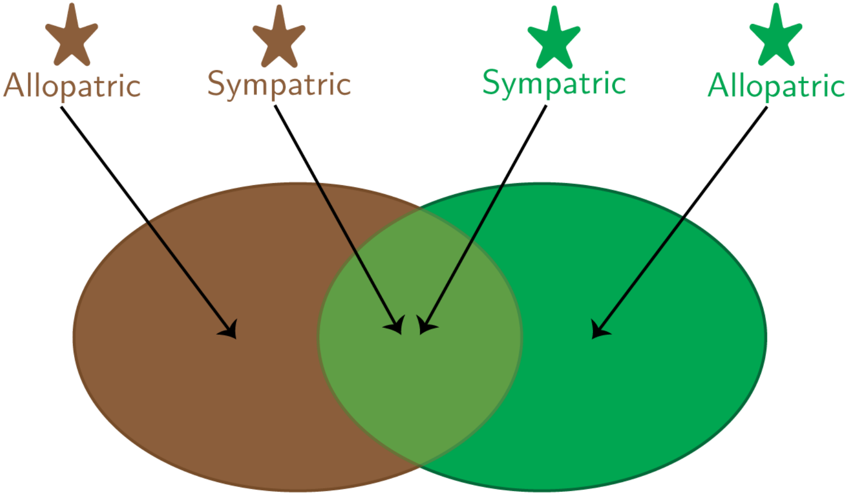
Overview

1. **Motivation**
2. Preliminaries
3. Derivations

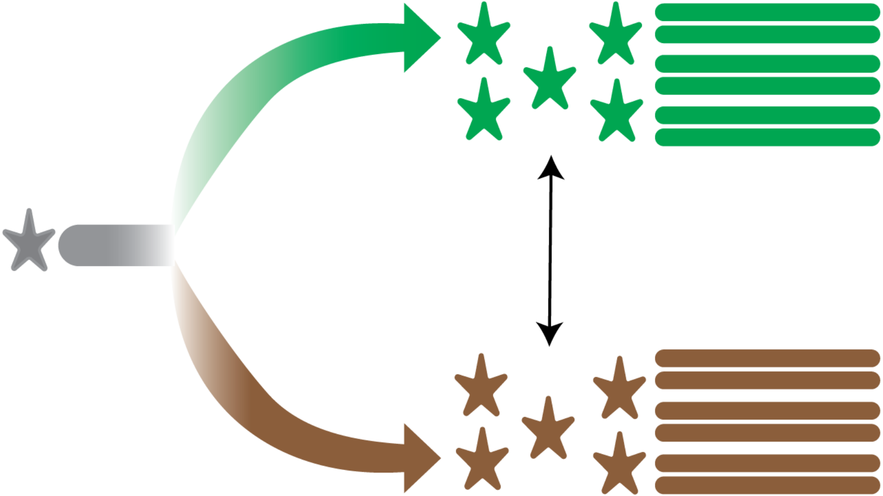
Evolution as a bifurcating process



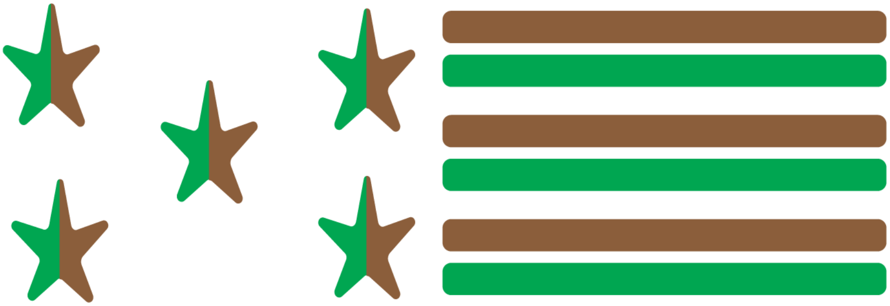
Sympatry is necessary for gene flow



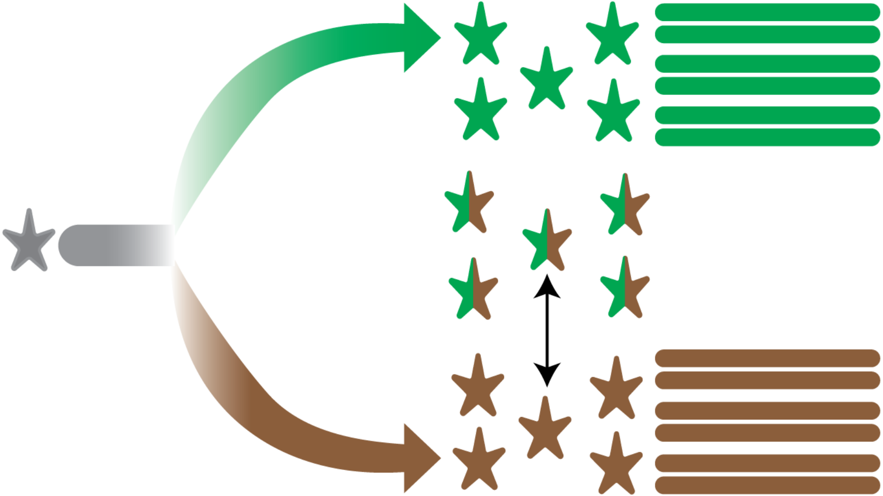
Incomplete RI leads to hybridization



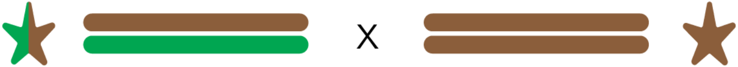
F1 hybrids = uniform mixture of parental ancestry



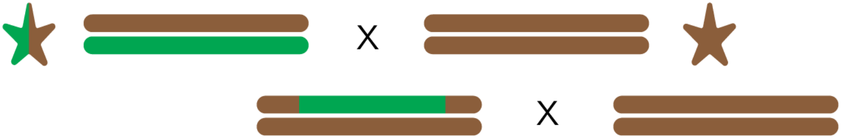
Hybridization is NOT sufficient for introgression



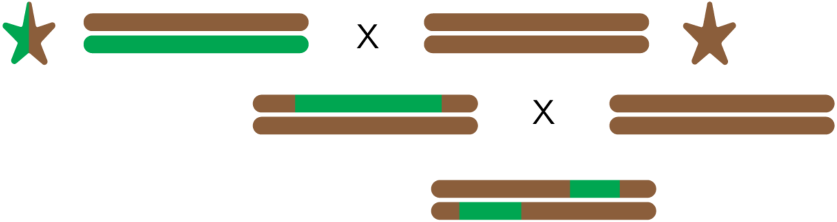
Incorporation of heterospecific loci via backcrossing



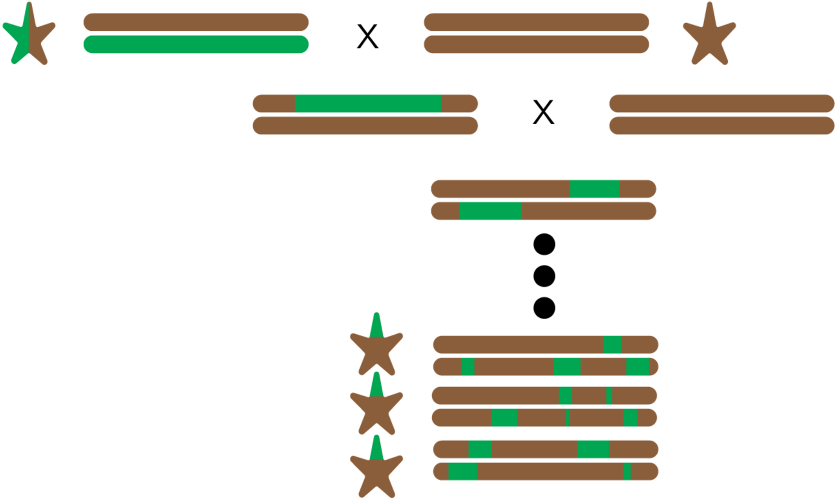
Incorporation of heterospecific loci via backcrossing



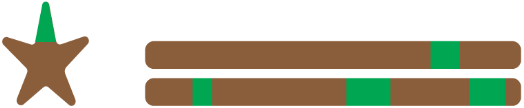
Incorporation of heterospecific loci via backcrossing



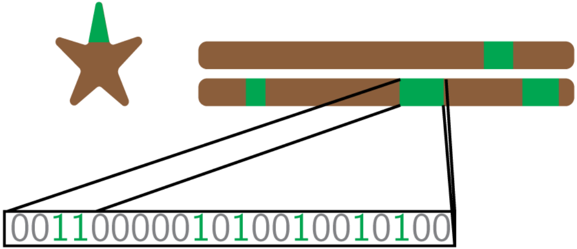
Incorporation of heterospecific loci via backcrossing



Genetic composition of an introgressed segment



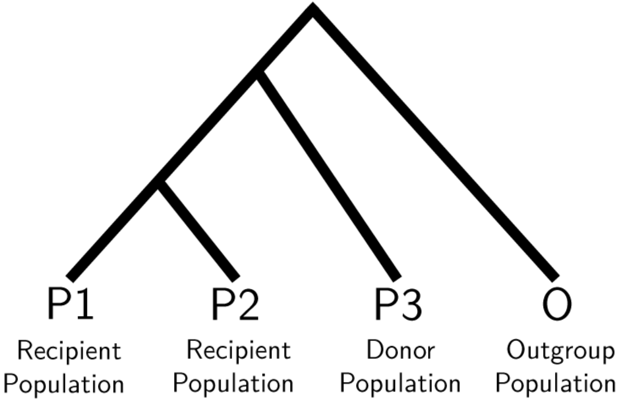
Introgression segments leaves a genomic footprint



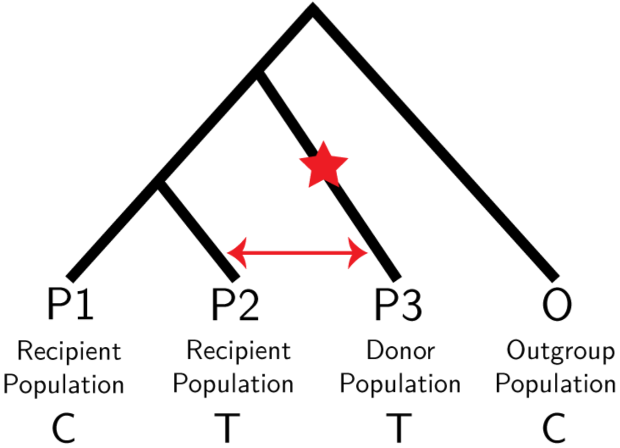
0 = reintroduced ancestral allele

1 = newly introduced derived allele

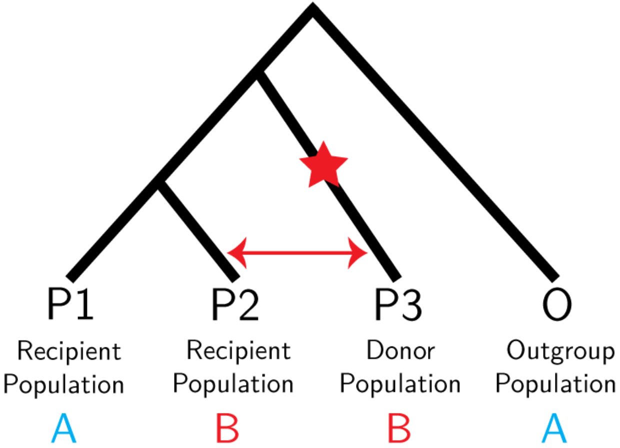
Site pattern tests of introgression



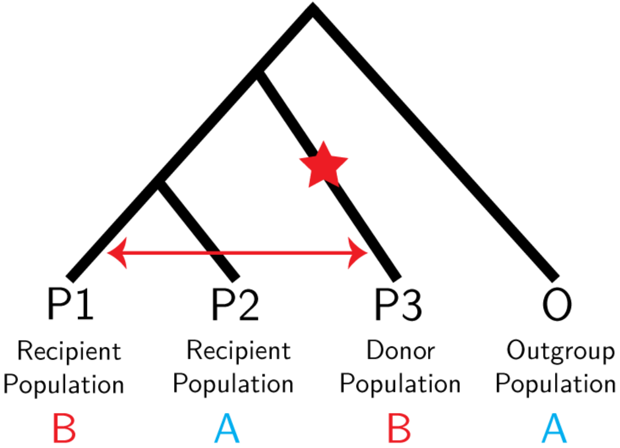
Derived allele sharing (P2 $\longleftrightarrow$ P3)



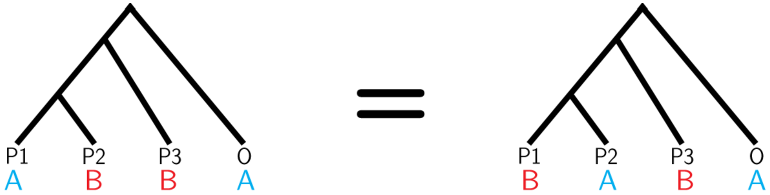
Derived allele sharing (P2 ⟷ P3)



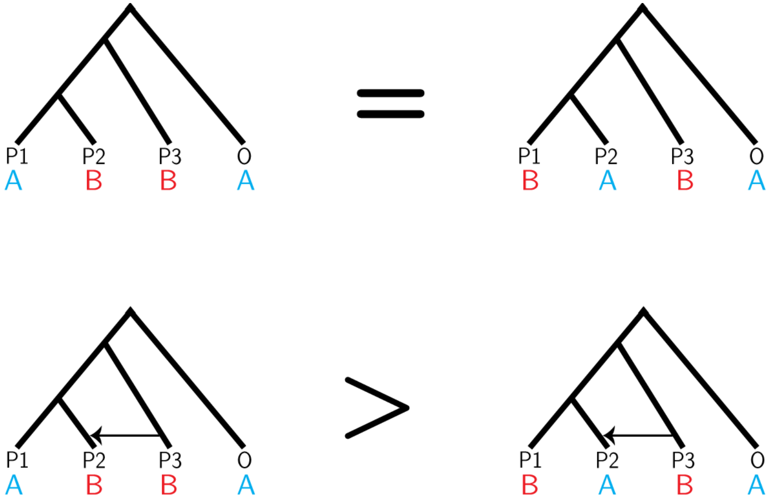
Derived allele sharing (P1 ⟷ P3)



Patterson's D



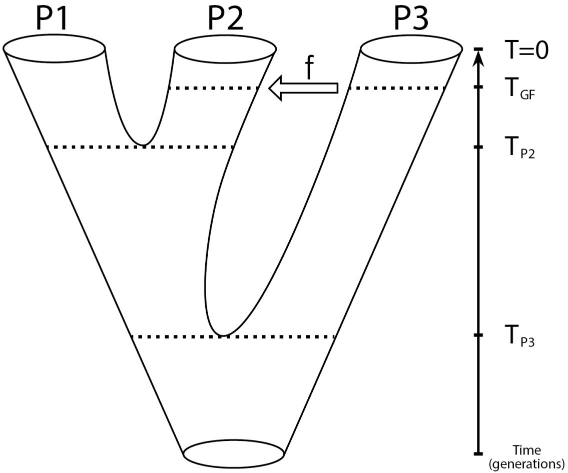
Patterson's D



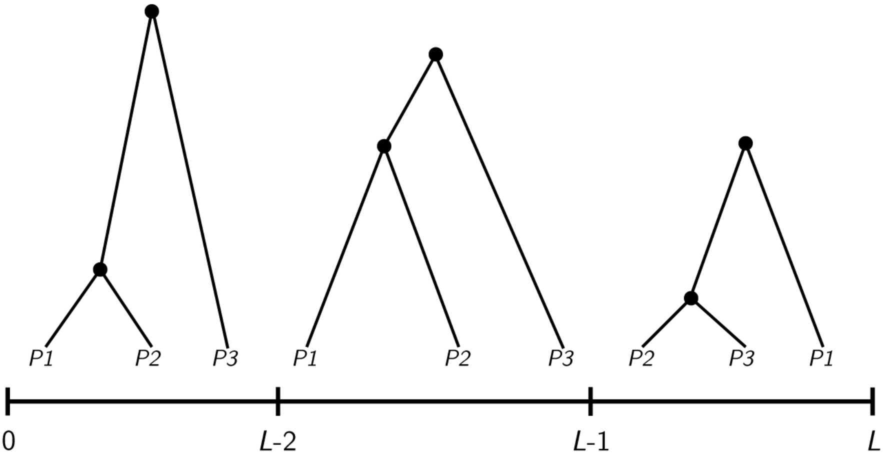
Overview

1. Motivation
2. **Preliminaries**
3. Derivations

Species tree model



Each locus has its own coalescent history



Probability of gene flow

Equation

$$Pr\left(gene\ flow\right) = f \tag{1}$$

$$Pr\left(no\ gene\ flow\right) = \left(1 - f\right) \tag{2}$$

Where f represents the admixture proportion—the probability that any lineage from *P3* migrates to *P2*.

Probability of no coalescence during time interval t

Probability of no coalescence during time interval t

Equation

$$Pr\left(no\ coalescences\right) = \left(1 - \frac{1}{2N}\right)^{t} \tag{3}$$

Where t denotes the time interval where coalescence can occur.

Probability of coalescence during time interval t

Probability of coalescence during time interval t

Equation

$$Pr\left(coalescences\right) = 1 - \left(1 - \frac{1}{2N}\right)^{t} \tag{4}$$

Where t denotes the time interval where coalescence can occur.

Expected time of coalescence T_2 & T_3

Expected time of coalescence T_2 & T_3

Equation

$$T_2 = 2N \tag{5}$$

$$T_3 = \frac{2N}{3} \tag{6}$$

Expected time of coalescence | coalescence during t

Expected time of coalescence | coalescence during t

Equation

$$\bar{t} \sim Tgeo\left(t|p, c\right) \tag{7}$$

Expected time of coalescence | coalescence during t

Equation

$$\bar{t} \sim Tgeo\left(t \middle| p, c\right) \tag{7}$$

$$f\left(\bar{t}\right) = \frac{\frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{t-1}}{1 - \left(1 - \frac{1}{2N}\right)^{c}} \tag{8}$$

Expected time of coalescence | coalescence during t

Equation

$$\bar{t} \sim Tgeo\left(t|p, c\right) \tag{7}$$

$$f\left(\bar{t}\right) = \frac{\frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{t-1}}{1 - \left(1 - \frac{1}{2N}\right)^{c}} \tag{8}$$

$$\mathbb{E}\left(\bar{t}\right) = \sum_{t=1}^{c} t \frac{\frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{t-1}}{1 - \left(1 - \frac{1}{2N}\right)^{c}} = \frac{2N - \left(\left(1 - \frac{1}{2N}\right)^{c}\left(c + 2N\right)\right)}{1 - \left(1 - \frac{1}{2N}\right)^{c}} \tag{9}$$

Where c denotes the time interval where coalescence must occur.

Overview

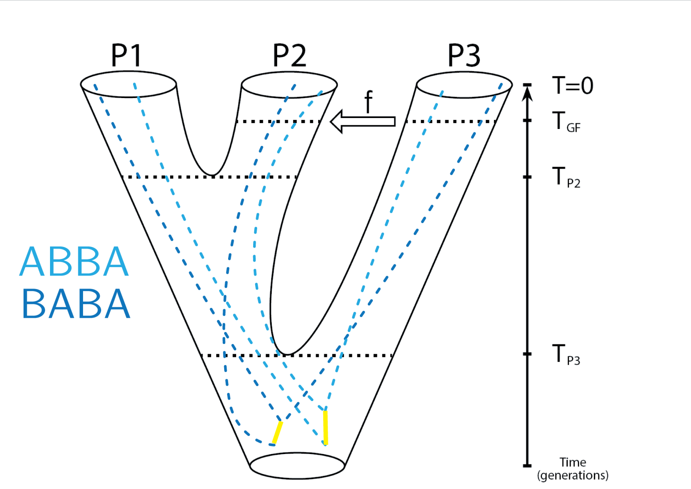
1. Motivation

2. Preliminaries

3. **Derivations**

Coalescent history 1

No gene flow from *P3* → *P2*, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P2* don't coalesce.



Coalescent history 1

No gene flow from $P3 \rightarrow P2$, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P2* don't coalesce.

Coalescent history 1

No gene flow from $P3 \rightarrow P2$, $P1$ & $P2$ don't coalesce between T_{P2} & T_{P3}, and $P1$ & $P2$ don't coalesce.

Derivation

$$Pr\left(no\ gene\ flow\right) = (1 - f) \tag{10}$$

Coalescent history 1

No gene flow from *P3* → *P2*, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$Pr\left(no\ gene\ flow\right) = (1 - f) \tag{10}$$

$$Pr\left(P1\ \&\ P2\ don't\ coalesce\ between\ T_{P2}\ \&\ T_{P3}\right) = \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{P2}} \tag{11}$$

Coalescent history 1

No gene flow from $P3 \rightarrow P2$, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$Pr\left(\textit{no gene flow}\right) = (1 - f) \tag{10}$$

$$Pr\left(\textit{P1 \& P2 don't coalesce between } T_{P2} \textit{ \& } T_{P3}\right) = \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{P2}} \tag{11}$$

$$Pr\left(\textit{P1 \& P2 don't coalesce}\right) = \frac{1}{3} \tag{12}$$

Coalescent history 1

No gene flow from *P3* → *P2*, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$Pr\left(no\ gene\ flow\right) = (1 - f) \tag{10}$$

$$Pr\left(P1\ \&\ P2\ don't\ coalesce\ between\ T_{P2}\ \&\ T_{P3}\right) = \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{P2}} \tag{11}$$

$$Pr\left(P1\ \&\ P2\ don't\ coalesce\right) = \frac{1}{3} \tag{12}$$

$$\mathbb{E}\left(Branch\ length\ between\ the\ 1^{st}\ \&\ 2^{nd} coalescent\ event\right) = 2N \tag{13}$$

Coalescent history 1

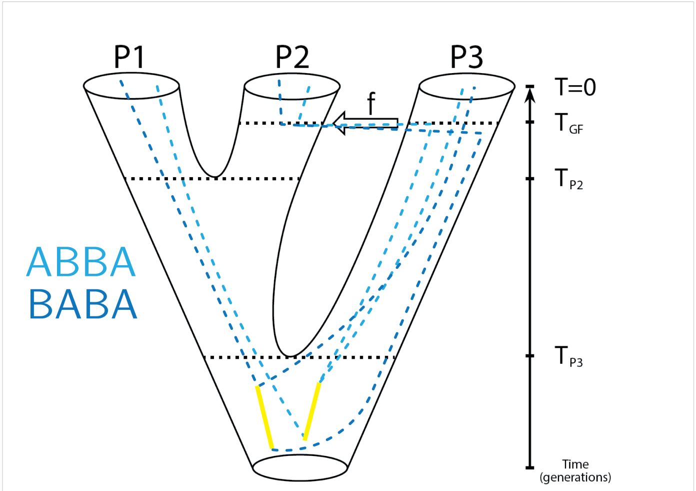
No gene flow from $P3 \rightarrow P2$, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$\mathbb{E}\left(C_{ABBA_1}\right) = \mathbb{E}\left(C_{BABA_1}\right) = (1 - f)\left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{P2}} \frac{2N}{3} \tag{14}$$

Coalescent history 2

Gene flow from *P3* → *P2*, *P2* & *P3* don't coalesce between T_{GF} & T_{P3}, and *P1* & *P2* don't coalesce.



Coalescent history 2

Gene flow from $P3 \rightarrow P2$, *P2* & *P3* don't coalesce between T_{GF} & T_{P3}, and *P1* & *P2* don't coalesce.

Coalescent history 2

Gene flow from $P3 \rightarrow P2$, *P2* & *P3* don't coalesce between T_{GF} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$Pr\left(gene\ flow\right) = f \tag{15}$$

Coalescent history 2

Gene flow from $P3 \rightarrow P2$, *P2* & *P3* don't coalesce between T_{GF} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$Pr\left(gene\ flow\right) = f \tag{15}$$

$$Pr\left(P2\ \&\ P3\ don't\ coalesce\ between\ T_{GF}\ \&\ T_{P3}\right) = \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{GF}} \tag{16}$$

Coalescent history 2

Gene flow from $P3 \rightarrow P2$, $P2$ & $P3$ don't coalesce between T_{GF} & T_{P3}, and $P1$ & $P2$ don't coalesce.

Derivation

$$Pr\left(gene\ flow\right) = f \tag{15}$$

$$Pr\left(P2\ \&\ P3\ don't\ coalesce\ between\ T_{GF}\ \&\ T_{P3}\right) = \left(1 - \frac{1}{2N}\right)^{T_{P3}-T_{GF}} \tag{16}$$

$$Pr\left(P1\ \&\ P2\ don't\ coalesce\right) = \frac{1}{3} \tag{17}$$

Coalescent history 2

Gene flow from $P3 \rightarrow P2$, $P2$ & $P3$ don't coalesce between T_{GF} & T_{P3}, and $P1$ & $P2$ don't coalesce.

Derivation

$$Pr\left(gene\ flow\right) = f \tag{15}$$

$$Pr\left(P2\ \&\ P3\ don't\ coalesce\ between\ T_{GF}\ \&\ T_{P3}\right) = \left(1 - \frac{1}{2N}\right)^{T_{P3}-T_{GF}} \tag{16}$$

$$Pr\left(P1\ \&\ P2\ don't\ coalesce\right) = \frac{1}{3} \tag{17}$$

$$\mathbb{E}\left(Branch\ length\ between\ the\ 1^{st}\ \&\ 2^{nd} coalescent\ event\right) = 2N \tag{18}$$

Coalescent history 2

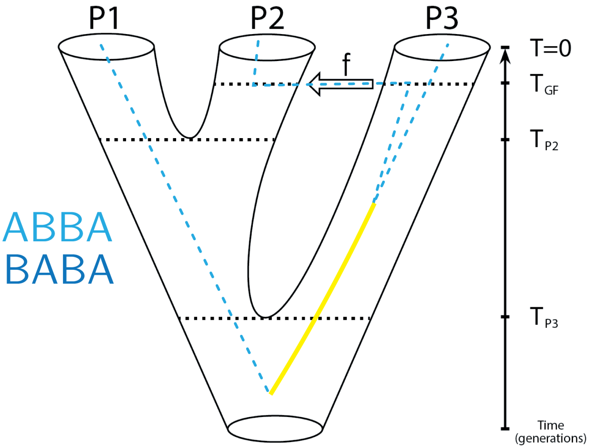
Gene flow from $P3 \rightarrow P2$, *P2* & *P3* don't coalesce between T_{GF} & T_{P3}, and *P1* & *P2* don't coalesce.

Derivation

$$\mathbb{E}\left(C_{ABBA_2}\right) = \mathbb{E}\left(C_{BABA_2}\right) = f\left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{GF}} \frac{2N}{3} \tag{19}$$

Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and $P2$ & $P3$ coalesce between T_{GF} & T_{P3}.



Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and *P2* & *P3* coalesce between T_{GF} & T_{P3}.

Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and $P2$ & $P3$ coalesce between T_{GF} & T_{P3}.

Derivation

$$Pr\left(gene\ flow\right) = f \tag{20}$$

Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and $P2$ & $P3$ coalesce between T_{GF} & T_{P3}.

Derivation

$$Pr\left(gene\ flow\right) = f \tag{20}$$

$$Pr\left(P2\ \&\ P3\ coalesce\ between\ T_{GF}\ \&\ T_{P3}\right) = 1 - \left(1 - \frac{1}{2N}\right)^{T_{P3}-T_{GF}} \tag{21}$$

Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and $P2$ & $P3$ coalesce between T_{GF} & T_{P3}.

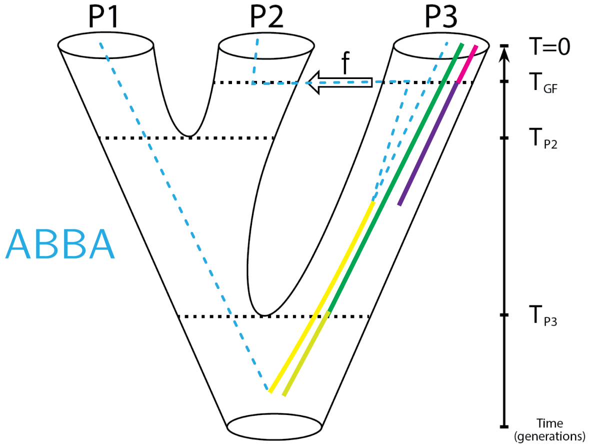
Derivation

$$Pr\left(gene\ flow\right) = f \tag{20}$$

$$Pr\left(P2\ \&\ P3\ coalesce\ between\ T_{GF}\ \&\ T_{P3}\right) = 1 - \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{GF}} \tag{21}$$

$$\mathbb{E}\left(Branch\ length\ between\ the\ 1^{st}\ \&\ 2^{nd} coalescent\ event\right) = \left(T_{P3} + 2N\right) - \left(T_{GF} + \overline{t}\right) \tag{22}$$

Coalescent history 3 ($(\bar{t})$)



$$\text{IBL} = T_{P3} + 2N - (T_{GF} + \bar{t})$$

Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and $P2$ & $P3$ coalesce between T_{GF} & T_{P3}.

Derivation

$$\mathbb{E}\left(C_{ABBA_3}\right) = f\left(1 - \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{GF}}\right)\left(\left(T_{P3} + 2N\right) - \left(T_{GF} + \overline{t}\right)\right) \tag{23}$$

Coalescent history 3

Gene flow from $P3 \rightarrow P2$ and $P2$ & $P3$ coalesce between T_{GF} & T_{P3}.

Derivation

$$\mathbb{E}\left(C_{ABBA_3}\right) = f\left(1 - \left(1 - \frac{1}{2N}\right)^{T_{P3} - T_{GF}}\right)\left(\left(T_{P3} + 2N\right) - \left(T_{GF} + \bar{t}\right)\right) \tag{23}$$

$$\mathbb{E}\left(C_{BABA_3}\right) = 0 \tag{24}$$

$\mathbb{E}\left(ABBA\right)$

Derivation

$$\mathbb{E}\left(\tau_{ABBA}\right) = C_{ABBA_1} + C_{ABBA_2} + C_{ABBA_3} \tag{25}$$

$\mathbb{E}\left(ABBA\right)$

Derivation

$$\mathbb{E}\left(\tau_{ABBA}\right) = C_{ABBA_1} + C_{ABBA_2} + C_{ABBA_3} \tag{25}$$

$$\begin{aligned}\mathbb{E}\left(\tau_{ABBA}\right) = {} & (1 - f)\left({}^{2N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3} - T_{P2}}\right) \\ & + (f)\left(\left({}^{2N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3} - T_{GF}}\right) + \left(T_{P3} - T_{GF}\right)\right)\end{aligned} \tag{26}$$

$\mathbb{E}\left(ABBA\right)$

Derivation

$$\mathbb{E}\left(\tau_{ABBA}\right) = C_{ABBA_1} + C_{ABBA_2} + C_{ABBA_3} \tag{25}$$

$$\begin{aligned}\mathbb{E}\left(\tau_{ABBA}\right) = {} & (1 - f)\left({}^{2N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3} - T_{P2}}\right) \\ & + (f)\left(\left({}^{2N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3} - T_{GF}}\right) + \left(T_{P3} - T_{GF}\right)\right)\end{aligned} \tag{26}$$

$$\mathbb{E}\left(ABBA_{sites}\right) = \mathbb{E}\left(\tau_{ABBA}\right) \times \mu \times L \tag{27}$$

Where μ represents the mutation rate and L represents the sequence length.

$\mathbb{E}\left(BABA\right)$

Derivation

$$\mathbb{E}\left(\tau_{BABA}\right) = C_{BABA_1} + C_{BABA_2} \tag{28}$$

$\mathbb{E}\left(BABA\right)$

Derivation

$$\mathbb{E}\left(\tau_{BABA}\right) = C_{BABA_1} + C_{BABA_2} \tag{28}$$

$$\begin{aligned}\mathbb{E}\left(\tau_{BABA}\right) = (1-f)&\left({}^{2N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{P2}}\right)\\ +(f)&\left({}^{2N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{GF}}\right)\end{aligned} \tag{29}$$

$\mathbb{E}\left(BABA\right)$

Derivation

$$\mathbb{E}\left(\tau_{BABA}\right) = C_{BABA_1} + C_{BABA_2} \tag{28}$$

$$\begin{aligned}\mathbb{E}\left(\tau_{BABA}\right) = (1 - f)\left({}^{2N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3} - T_{P2}}\right) \\ + (f)\left({}^{2N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3} - T_{GF}}\right)\end{aligned} \tag{29}$$

$$\mathbb{E}\left(BABA_{sites}\right) = \mathbb{E}\left(\tau_{BABA}\right) \times \mu \times L \tag{30}$$

Where μ represents the mutation rate and L represents the sequence length.

Patterson's D

Derivation

$$\mathbb{E}\left(D\right) = \frac{\mathbb{E}\left(\tau_{ABBA}\right) - \mathbb{E}\left(\tau_{BABA}\right)}{\mathbb{E}\left(\tau_{ABBA}\right) + \mathbb{E}\left(\tau_{BABA}\right)} \tag{31}$$

Patterson's D

Derivation

$$\mathbb{E}\left(D\right) = \frac{\mathbb{E}\left(\tau_{ABBA}\right) - \mathbb{E}\left(\tau_{BABA}\right)}{\mathbb{E}\left(\tau_{ABBA}\right) + \mathbb{E}\left(\tau_{BABA}\right)} \tag{31}$$

$$\mathbb{E}\left(D\right) = \frac{\left(f\right)\left(T_{P3} - T_{GF}\right)}{\left(1-f\right)\left[{}^{4N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3}-T_{P2}}\right] + \left(f\right)\left[\left({}^{4N}\!/_{3}\left(1 - {}^{1}\!/_{2N}\right)^{T_{P3}-T_{GF}}\right) + \left(T_{P3} - T_{GF}\right)\right]} \tag{32}$$

Patterson's D

Derivation

$$\mathbb{E}\left(D\right) = \frac{\mathbb{E}\left(\tau_{ABBA}\right) - \mathbb{E}\left(\tau_{BABA}\right)}{\mathbb{E}\left(\tau_{ABBA}\right) + \mathbb{E}\left(\tau_{BABA}\right)} \tag{31}$$

$$\mathbb{E}\left(D\right) = \frac{\left(f\right)\left(T_{P3} - T_{GF}\right)}{\left(1-f\right)\left[{}^{4N}\!/\!_{3}\left(1 - {}^{1}\!/\!_{2N}\right)^{T_{P3}-T_{P2}}\right] + \left(f\right)\left[\left({}^{4N}\!/\!_{3}\left(1 - {}^{1}\!/\!_{2N}\right)^{T_{P3}-T_{GF}}\right) + \left(T_{P3} - T_{GF}\right)\right]} \tag{32}$$

$$\mathbb{E}\left(D\right) = \frac{\sum_{i=1}^{L}(1-p_{i1})p_{i2}p_{i3}(1-p_{iO}) - p_{i1}(1-p_{i2})p_{i3}(1-p_{iO})}{\sum_{i=1}^{L}(1-p_{i1})p_{i2}p_{i3}(1-p_{iO}) + p_{i1}(1-p_{i2})p_{i3}(1-p_{iO})} \tag{33}$$

Where $p_{i\#}$ represents the derived allele frequency at site i and L represents the sequence length.