Deriving *Patterson's D* & f_{HOM}

1 Preliminaries

1.1 Instantaneous Unidirectional Admixture Model of Introgression

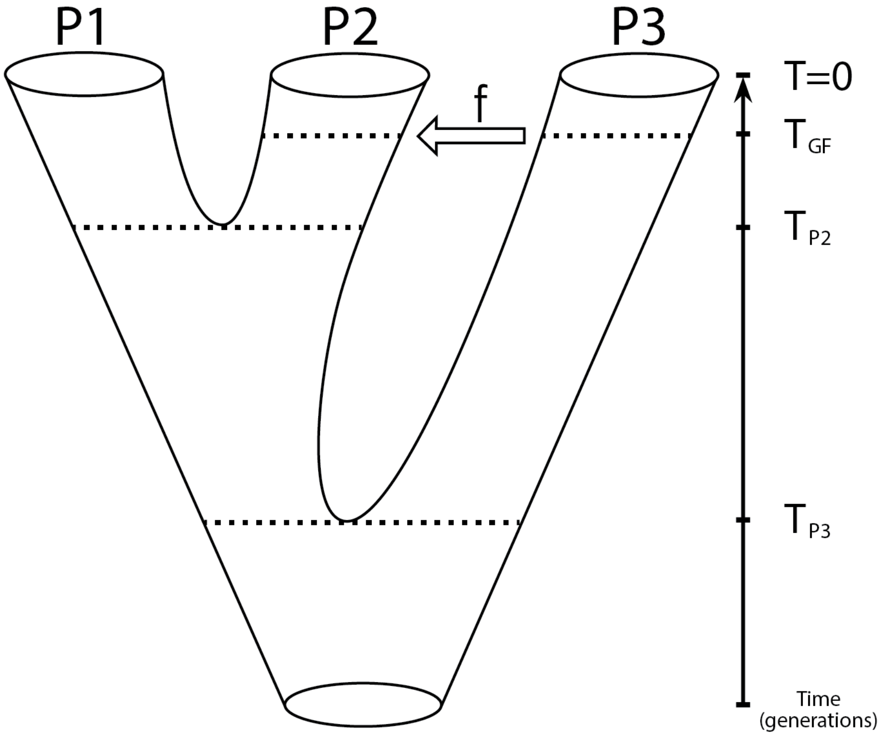

Figure 1: Instantaneous unidirectional admixture model (IUA) of introgression.

Figure 1 graphically depicts the instantaneous unidirectional admixture (IUA) model of introgression, which assumes a three-taxon species tree—in Newick format: ((*P1*, *P2*), *P3*)—where the *P1* and *P2* populations diverged from an ancestral population *P12* at time T_{P2} and where the *P3* and *P12* populations diverged from an ancestral population *P123* at time T_{P3}. Additionally, the IUA model assumes that a single introgression event took place at time T_{GF}, which can occur at any generation between T_{P2} and generation zero, with a probability f that a *P2* lineage traces its ancestry from the *P3* side of the phylogeny. Additionally, we assume that all populations have an equal and constant effective population size throughout the entire demographic history, an infinite sites mutation model, and that ancestral states are known. It should be noted that in practice one would also consider an additional taxon—in Newick format: (((*P1*, *P2*), *P3*), *O*)—where the *O* population is ancestral to the *P123* population, and is used as a putative outgroup to polarize ancestral states, since true ancestral states are rarely known. We will use the IUA model to derive the expected branch lengths that would generate an *ABBA* and *BABA* site patterns.

1.2 Expected Time of Coalescence Between Two Lineages Given that Coalescence Occurs During the Time Interval *c*

For some coalescent histories we will need to find the expected coalescence time between two lineages conditioned on the coalescent event occurring within the time interval c which we denote as $(\bar{t})$. To do so we utilize the fact that the expected time to coalescence between two lineages is geometrically distributed. Thus, we can derive $\bar{t}$ by defining the probability mass function for the random variable i where $i \in \{1, ..., c\}$ which is conveniently known as the truncated geometric distribution.

$$f\left(i\right) = \frac{\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}}{1-\left(1-\frac{1}{2N}\right)^{c}} \tag{1a}$$

$$\therefore \bar{t} = \sum_{i=1}^{c} i \frac{\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}}{1-\left(1-\frac{1}{2N}\right)^{c}} \tag{1b}$$

$$= \frac{2N-\left(\left(1-\frac{1}{2N}\right)^{c}\left(c+2N\right)\right)}{1-\left(1-\frac{1}{2N}\right)^{c}} \tag{1c}$$

2 Site Pattern Derivations

2.1 *ABBA*

There are three unique coalescent histories that result in a branch where a mutation would generate an *ABBA* site pattern. For each scenario we derive the expected contribution C_n of the scenario n. We define C_n as the probability of obtaining each scenario n multiplied by the expected branch length from the parent node u of the leaves $(P2, P3)$ to its parent node, conditioned on the coalescent history in the scenario n.

2.1.1 Coalescent History 1. No gene flow from ***P3*** → ***P2***, ***P1*** & ***P2*** don't coalesce between $\boldsymbol{T_{P2}}$ & $\boldsymbol{T_{P3}}$, and ***P2*** & ***P3*** coalesce in the first coalescent event after $\boldsymbol{T_{P3}}$:

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

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

$$Pr\left(\textit{P2 \& P3 coalesce in the first coalescent event}\right) = \frac{1}{3} \tag{2c}$$

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

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

2.1.2 Coalescent History 2. Gene flow from ***P3*** → ***P2***, ***P2*** & ***P3*** don't coalesce between $\boldsymbol{T_{GF}}$ & $\boldsymbol{T_{P3}}$, and ***P2*** & ***P3*** coalesce in the first coalescent event after $\boldsymbol{T_{P3}}$:

$$Pr\left(\textit{gene flow}\right) = f \tag{3a}$$

$$Pr\left(\textit{P2 \& P3 don't coalesce between } T_{GF} \textit{ \& } T_{P3}\right) = \left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}} \tag{3b}$$

$$Pr\left(\textit{P2 \& P3 coalesce in the first coalescent event}\right) = \frac{1}{3} \tag{3c}$$

$$\mathbb{E}\left(\textit{branch length between the } 1^{st} \textit{ \& } 2^{nd} \textit{coalescent event}\right) = 2N \tag{3d}$$

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

2.1.3 Coalescent History 3. Gene flow from ***P3*** → ***P2*** and ***P2*** & ***P3*** coalesce between $\boldsymbol{T_{GF}}$ & $\boldsymbol{T_{P3}}$:

$$Pr\left(\textit{gene flow}\right) = f \tag{4a}$$

$$Pr\left(\textit{P2 \& P3 coalesce between } T_{GF} \textit{ \& } T_{P3}\right) = 1-\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}} \tag{4b}$$

$$\mathbb{E}\left(\textit{branch length between the } 1^{st} \textit{ \& } 2^{nd} \textit{coalescent event}\right) = \left(T_{P3}+2N\right)-\left(T_{GF}+\bar{t}\right) \tag{4c}$$

$$\mathbb{E}\left(C_{ABBA_3}\right) = f\left(1-\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}}\right)\left(\left(T_{P3}+2N\right)-\left(T_{GF}+\bar{t}\right)\right) \tag{4d}$$

2.1.4 Expected Branch Length of the *ABBA* Site Pattern

Using the the law of total probability, we can write $\mathbb{E}\left(\tau_{ABBA}\right)$ as the sum of the three contributions C_{ABBA_1}, C_{ABBA_2} and C_{ABBA_3}:

$$\begin{aligned}\mathbb{E}\left(\tau_{ABBA}\right) = (1-f)&\left({}^{2N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{P2}}\right)\\ &+ (f)\left(\left({}^{2N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{GF}}\right)+\left(T_{P3}-T_{GF}\right)\right)\end{aligned} \tag{5}$$

2.2 *BABA*

There are two unique coalescent histories that result in a branch where a mutation would generate a *BABA* site pattern. For each scenario we derive the expected contribution C_n of the scenario n. We define C_n as the probability of obtaining each scenario n multiplied by the expected branch length from the parent node u of the leaves $(P1, P3)$ to its parent node, conditioned on the coalescent history in the scenario n.

2.2.1 Coalescent History 1. No gene flow from *P3* → *P2*, *P1* & *P2* don't coalesce between T_{P2} & T_{P3}, and *P1* & *P3* coalesce in the first coalescent event after T_{P3}:

$$Pr\left(\textit{no gene flow}\right) = (1-f) \tag{6a}$$

$$Pr\left(\textit{P1 \& P2 don't coalesce between } T_{P2} \textit{ \& } T_{P3}\right) = \left(1-\frac{1}{2N}\right)^{T_{P3}-T_{P2}} \tag{6b}$$

$$Pr\left(\textit{P1 \& P3 coalesce in the first coalescent event}\right) = \frac{1}{3} \tag{6c}$$

$$\mathbb{E}\left(\textit{branch length between the } 1^{st} \textit{ \& } 2^{nd} \textit{coalescent event}\right) = 2N \tag{6d}$$

$$\mathbb{E}\left(C_{BABA_1}\right) = (1-f)\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{P2}}\frac{2N}{3} \tag{6e}$$

2.2.2 Coalescent History 2. Gene flow from *P3* → *P2*, *P2* & *P3* don't coalesce between T_{GF} & T_{P3}, and *P1* & *P3* coalesce in the first coalescent event after T_{P3}:

$$Pr\left(\textit{gene flow}\right) = f \tag{7a}$$

$$Pr\left(\textit{P2 \& P3 don't coalesce between } T_{GF} \textit{ \& } T_{P3}\right) = \left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}} \tag{7b}$$

$$Pr\left(\textit{P1 \& P3 coalesce in the first coalescent event}\right) = \frac{1}{3} \tag{7c}$$

$$\mathbb{E}\left(\textit{branch length between the } 1^{st} \textit{ \& } 2^{nd} \textit{coalescent event}\right) = 2N \tag{7d}$$

$$\mathbb{E}\left(C_{BABA_2}\right) = f\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}}\frac{2N}{3} \tag{7e}$$

2.2.3 Expected Branch Length of the *BABA* Site Pattern

Using the the law of total probability, we can write $\mathbb{E}\left(\tau_{BABA}\right)$ as the sum of the two contributions C_{BABA_1} and C_{BABA_2}:

$$\begin{aligned}\mathbb{E}\left(\tau_{BABA}\right) = & \left(1-f\right)\left({}^{2N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{P2}}\right) \\ & + \left(f\right)\left({}^{2N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{GF}}\right)\end{aligned} \tag{8}$$

3 Introgression Metrics Derivations

3.1 *Patterson's D*

Patterson's D tests for the presence of introgression by assessing if there is a significant excess of derived allele sharing between the donor and recipient populations. This is accomplished by using observed site patterns as a proxy for genealogical relationships between populations. It should be noted that a statistically significant non-zero value of *Patterson's D* should be interpreted as evidence for rejecting an incomplete lineage sorting (ILS)—the process where two lineages fail to coalesce in their most recent common ancestral population—only model. This is due to the fact that in the absence of introgression it is only possible to generate *ABBA* and *BABA* site patterns via genealogical histories of ILS.

$$\mathbb{E}\left(D\right) = \frac{\mathbb{E}\left(\tau_{ABBA}\right)-\mathbb{E}\left(\tau_{BABA}\right)}{\mathbb{E}\left(\tau_{ABBA}\right)+\mathbb{E}\left(\tau_{BABA}\right)} \tag{9a}$$

$$\mathbb{E}\left(D\right) = \frac{\left(f\right)\left(T_{P3}-T_{GF}\right)}{\left(1-f\right)\left({}^{4N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{P2}}\right)+\left(f\right)\left(\left({}^{4N}\!/_{3}\left(1-{}^{1}\!/_{2N}\right)^{T_{P3}-T_{GF}}\right)+\left(T_{P3}-T_{GF}\right)\right)} \tag{9b}$$

3.2 f_{HOM}

f_{HOM} quantifies the fraction of the genome shared through introgression—hereafter referred to as the admixture proportion—by assessing the ratio of observed difference in allele sharing between *P3* and *P2* in the numerator and the expected differences assuming the entire genome was introgressed—i.e., complete homogenization of allele sharing—from *P3* to *P2* in the denominator. This is accomplished by replacing *P2* with *P3* in the assumed underlying species tree—i.e., in newick format: *(((P1, P3), P3,), O)*—when computing the denominator. Consequently, when we make the assumption that the entire genome was introgressed, in computing the denominator we assume that $f = 1$ and $T_{GF} = 0$.

$$S\left(P1,P2,P3,O\right) = \mathbb{E}\left(\tau_{ABBA}\right)-\mathbb{E}\left(\tau_{BABA}\right) = \left(f\right)\left(T_{P3}-T_{GF}\right) \tag{10a}$$

$$S\left(P1,P3,P3,O\right) = \mathbb{E}\left(\tau_{ABBA}\right)-\mathbb{E}\left(\tau_{BABA}\right) = T_{P3}+2N \tag{10b}$$

$$f_{hom} = \frac{S\left(P1,P2,P3,O\right)}{S\left(P1,P3,P3,O\right)} = \frac{\left(f\right)\bullet\left(T_{P3}-T_{GF}\right)}{T_{P3}+2N} \tag{10c}$$