Definitions

Incomplete lineage sorting: The process where polymorphism at the time of lineage splitting sorts into two daughter lineages in a fashion that creates a gene tree vs. species tree conflict

Admixture: the mixing of differentiated populations (or species) the product of which is individuals with mixed ancestry

Introgression: the movement of alleles from one species to another

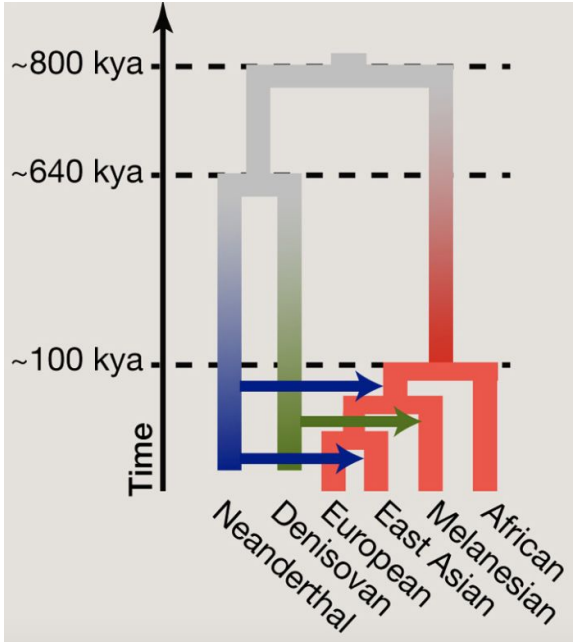
*introgression and admixture are often used interchangeably although perhaps

Admixture between Neanderthal and modern humans

Greene et al. (2010) reported introgression of Neanderthal alleles into modern humans

Reich et al. (2010) subsequently reported introgression of Denisovan alleles into modern humans

What tests were used to infer admixture?



Vattahil and Akey (2015) Cell

Incomplete Lineage Sorting (ILS)

Consider three populations and an outgroup that are related by a “species tree” (outer black lines)

The ancestral allele is blue

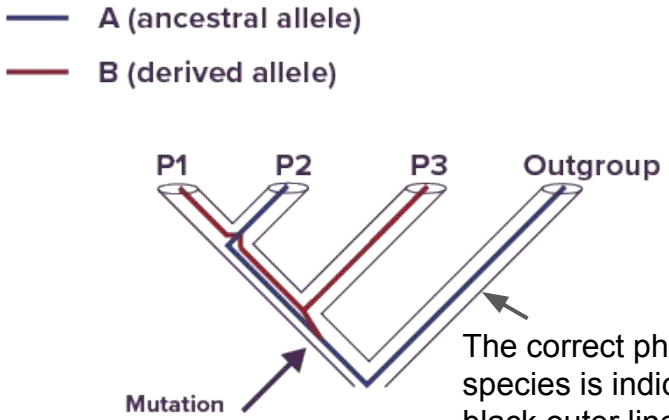
The blue line traces the presence of this allele over time

The mutation introduces a red allele in the common ancestor to P1,P2,P3

By chance, P1 inherits the red “B” allele, P2 inherits the blue “A” allele and P3 inherits the the red “B” allele

In this scenario, the “gene tree” shows a closer relationship between P1 and P3 and P2 and P4 (which does not match the true species tree)

Key point: the presence of polymorphism in a lineage at the time it splits into two daughter lineages creates the situation where incomplete lineage sorting can occur

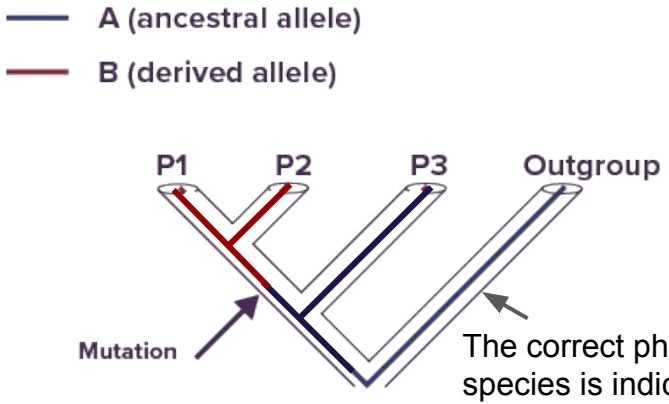


The correct phylogeny of the species is indicated by the black outer lines

Complete Lineage Sorting

In the absence of polymorphism at a locus in a population at the time it splits into two, there is no opportunity for ILS

Under this scenario, the gene tree (indicated by red and blue) does not conflict with the species tree (indicated in black)

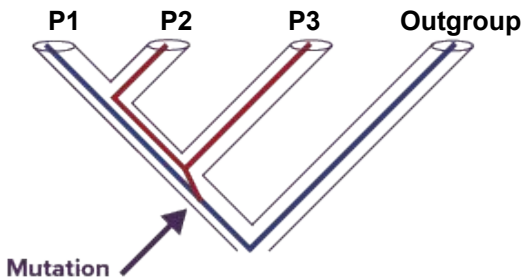


The correct phylogeny of the species is indicated by the outer black lines

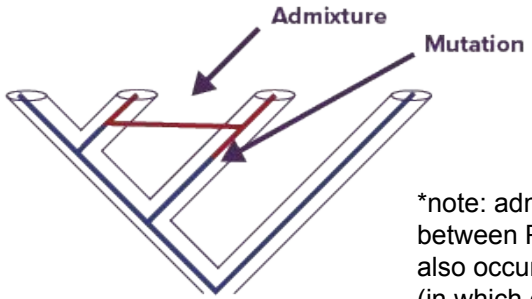
Why is ILS important in the context of testing for admixture?

Key point: Both ILS and admixture leave a similar signature of derived allele sharing that causes conflicts between a gene tree and species tree

Incomplete Lineage Sorting



Admixture/Introgression



*note: admixture shown here between P3 and P2, could also occur between P3 and P1 (in which case the pattern of sharing would be identical to the pattern created by lineage sorting on the left)

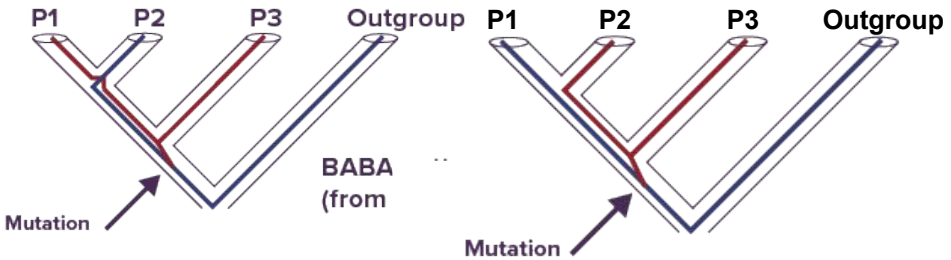
The principle of the D-test for admixture

Incomplete lineage sorting can result in P1 and P3 sharing an allele, or P2 and P3 sharing an allele

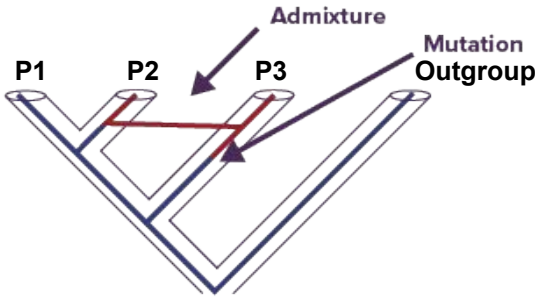
These patterns are expected at equal frequency

Introgression between P1 and P3 or P2 and P3 could produce either pattern, but an excess of one pattern is indicative of admixture

Incomplete Lineage Sorting



Admixture/Introgression



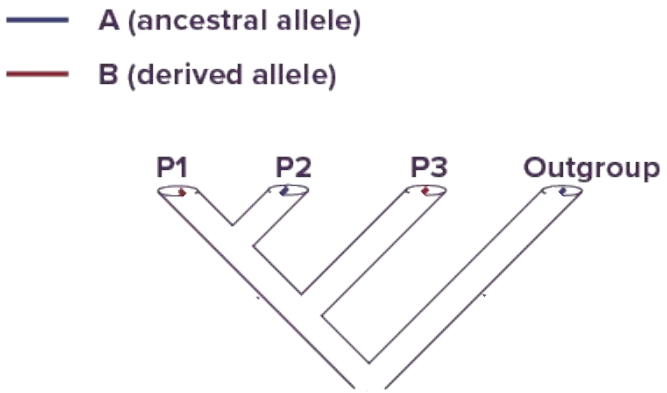
D-test (“ABBA-BABA” test)

Four population test

P1 and P2 are ‘ingroup’ populations

P3 is the “test” population

Outgroup is chosen which has not admixed with any of the other taxa



The test population (P3) is an outgroup to ingroup populations which we want to test for admixture with either ingroup population

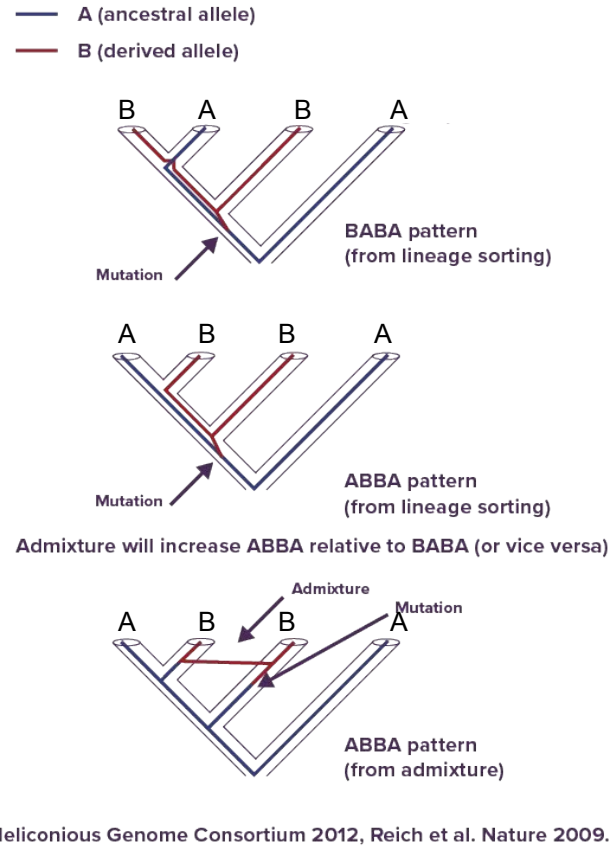
D-test (“ABBA-BABA” test)

“ABBA” and “BABA” correspond to ancestral (A) and derived (B) states in P1, P2, P3, and outgroup

Pattern	Explanation
ABBA	P2 and P3 share derived allele P1 and Outgroup share ancestral allele
BABA	P1 and P3 share derived allele P2 and Outgroup share ancestral allele

In example of admixture at the right, admixture between P3 and P2 will produce the ABBA pattern

Key point: in this example, the ABBA pattern can be produced by both ILS and admixture, whereas BABA pattern is produced only by ILS



Heliconious Genome Consortium 2012, Reich et al. Nature 2009.

D-test (“ABBA-BABA” test)

The ABBA-BABA test is based on the counting the number of ABBA and BABA sites genomewide and testing for a difference

D is the test statistic

E[D] = 0 if there is no admixture

E[D] > 0 if there is an excess of ABBA

E[D] < 0 if there is an excess of BABA

*E[D] is the “expectation of D”

D is compared against a distribution obtained by block jackknife to test statistical significance of the difference in ABBA and BABA site counts

$$D(P_1, P_2, P_3, O) \models \frac{\sum C_{ABBA}(i) - C_{BABA}(i)}{\sum C_{ABBA}(i) + C_{BABA}(i)}$$

i is the total number of ABBA and BABA sites

Each site only has an ABBA or BABA pattern

Example:

At an ABBA site we add 1 to the numerator (because C_{ABBA} - C_{BABA} = 1 - 0)

At a BABA site we subtract 1 from the numerator because (C_{ABBA} - C_{BABA} = 0 - 1)

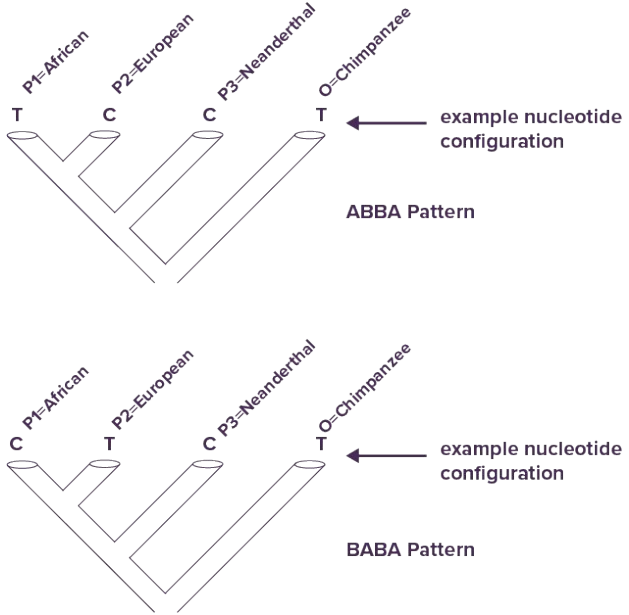
Example: Admixture between humans and Neanderthals

Whole genome sequencing of:

African (P1)
European (P2)
Neanderthal (P3)
Chimpanzee (O)

Calculate D across all biallelic SNPs with ABBA and BABA pattern genomewide

$$D(P_1, P_2, P_3, O) \models \frac{\sum C_{ABBA}(i) - C_{BABA}(i)}{\sum C_{ABBA}(i) + C_{BABA}(i)}$$



Green et al. Science (2010)

Example: Admixture between humans and Neanderthals

P1 = French (European), P2 = San (African), P3 = Neanderthal, Outgroup = chimpanzee

French-San-Neandertal-Chimpanzee	Counts
AAAA	818,322,920
AABA	5,827,247
ABAA	689,594
ABBA	95,347
ABCA	2,995
BAAA	756,324
BABA	103,612
BACA	3,544
BBAA	303,340
BBBA	8,156,936
BBCA	32,607
BCAA	972
BCBA	6,147
CBBA	6,264
BCDA	36

Green et al. (2010) Science

Example: Admixture between humans and Neanderthals

P1 = French (European), P2 = San (African), P3 = Neanderthal, Outgroup = chimpanzee

French-San-Neandertal-Chimpanzee	Counts
AAAA	818,322,920
AABA	5,827,247
ABAA	689,594
ABBA	95,347
ABCA	2,995
BAAA	756,324
BABA	103,612
BACA	3,544
BBAA	303,340
BBBA	8,156,936
BBCA	32,607
BCAA	972
BCBA	6,147
CBBA	6,264
BCDA	36

Example Site class	Interpretation	Informative about introgression
AAAA	all four alleles are identical	N
AABA	Derived allele in Neanderthal	N
ABAA	Derived allele in San	N
ABBA	Derived allele shared by San + Neanderthal	Y
BABA	Derived allele shared by French + Neanderthal	Y

Green et al. (2010) Science

Example: Admixture between humans and Neanderthals

P1 = French (European), P2 = San (African), P3 = Neanderthal, Outgroup = chimpanzee

French-San-Neandertal-Chimpanzee	Counts
AAAA	818,322,920
AABA	5,827,247
ABAA	689,594
ABBA	95,347
ABCA	2,995
BAAA	756,324
BABA	103,612
BACA	3,544
BBAA	303,340
BBBA	8,156,936
BBCA	32,607
BCAA	972
BCBA	6,147
CBBA	6,264
BCDA	36

Example Site class	Interpretation	Informative about introgression
AAAA	all four alleles are identical	N
AABA	Derived allele in Neanderthal	N
ABAA	Derived allele in San	N
ABBA	Derived allele shared by San + Neanderthal	Y
BABA	Derived allele shared by French + Neanderthal	Y

Green et al. (2010) Science

Example: Admixture between humans and Neanderthals

P1 = French (European), P2 = San (African), P3 = Neanderthal, Outgroup = chimpanzee

French-San-Neandertal-Chimpanzee	Counts
AAAA	818,322,920
AABA	5,827,247
ABAA	689,594
ABBA	95,347
ABCA	2,995
BAAA	756,324
BABA	103,612
BACA	3,544
BBAA	303,340
BBBA	8,156,936
BBCA	32,607
BCAA	972
BCBA	6,147
CBBA	6,264
BCDA	36

Example Site class	Interpretation	Informative about introgression
AAAA	all four alleles are identical	N
AABA	Derived allele in Neanderthal	N
ABAA	Derived allele in San	N
ABBA	Derived allele shared by San + Neanderthal	Y
BABA	Derived allele shared by French + Neanderthal	Y

The excess of 103,612 of BABA sites compared to 95,347 of ABBA sites suggests Neanderthal-French admixture, but is it statistically significant difference?

Green et al. (2010) Science

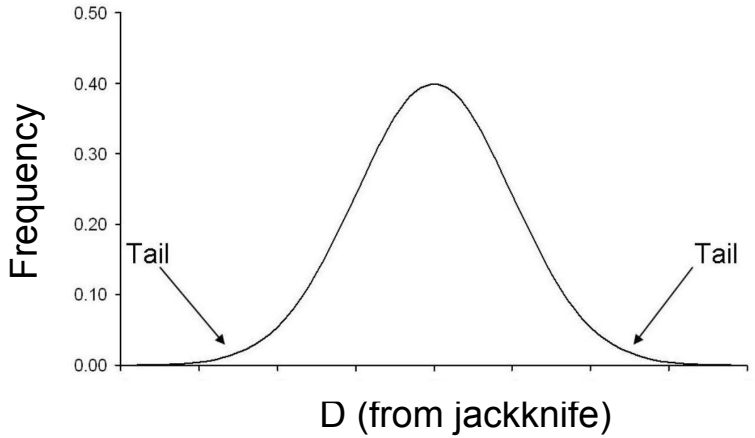
The block jackknife

Procedure:

Step 1. Divide genome into n intervals

Step 2. Calculate D statistic for each interval

Step 3. Compare observed D to jackknife distribution



If observed D < 0 and in left tail then this is evidence of an excess of BABA and the signature of admixture between P1 and P3

If observed D > 0 and in right tail then this is evidence of an excess of ABBA and the signature of admixture between P1 and P2

Green et al. (2010) Science

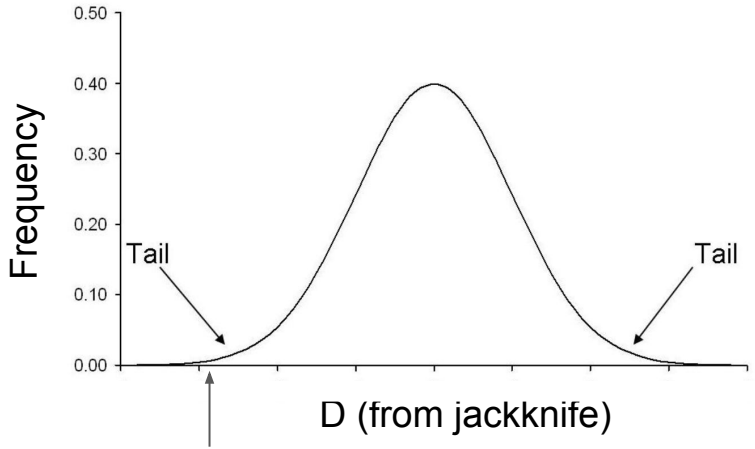
The block jackknife

Procedure:

Step 1. Divide genome into n intervals

Step 2. Calculate D statistic for each interval

Step 3. Compare observed D to jackknife distribution



D from Green et al. was significantly negative

If observed D < 0 and in left tail then this is evidence of an excess of BABA and the signature of admixture between P1 and P3

If observed D > 0 and in right tail then this is evidence of an excess of ABBA and the signature of admixture between P1 and P2

Green et al. (2010) Science

Example: human-Denisovan admixture

Ancient Phalanx (“pinky”) bone found in Siberia (Denisova)

Genome sequencing revealed divergent sequence from both humans and Neanderthals

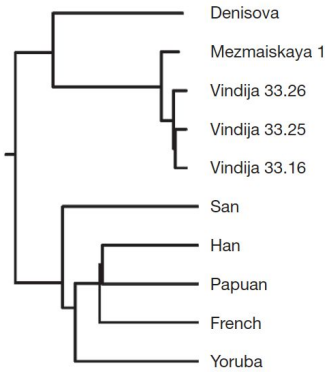


Figure 1 | A neighbour-joining tree based on pairwise autosomal DNA sequence divergences for five ancient and five present-day hominins. Vindija 33.16, Vindija 33.25 and Vindija 33.26 refer to the catalogue numbers of the Neanderthal bones.

Reich et al. (2010) Science

Example: human-Denisovan admixture

Note: in this example D is presented as nBABA-nABBA

P1 = X
P2 = Papuan
P3 = Neanderthal

P1 = X
P2 = Papuan
P3 = Denisova

Table 1 | Sharing of derived alleles between present-day and archaic hominins

Sample H_1	Sample H_2	Source of data for H_1 and H_2	$D(H_1, H_2$, Neanderthal, chimpanzee)					$D(H_1, H_2$, Denisova, chimpanzee)				
			n_{BABA}	n_{ABBA}	D (%)	s.e. (%)	Z-score	n_{BABA}	n_{ABBA}	D (%)	s.e. (%)	Z-score
Eurasian/Melanesian*												
French	Papuan1	Ref. 8	15,523	15,548	−0.1	0.8	−0.1	23,509	25,470	−4.0	0.7	−5.7†
Han	Papuan1	Ref. 8	15,059	14,677	1.3	0.9	1.5	22,262	24,198	−4.2	0.7	−5.8†

X=French

X=Han

We present the D statistic D(H1, H2, X, chimpanzee), the normalized difference between the number of sites at which the derived allele in an archaic read from X matches human sample H1 (nBABA) and human sample H2 (nBABA); thus, its value is D 5 (nBABA2nABBA)/(nBABA1nABBA).

Reich et al. (2010) Science