Patterns of
D(archaic1, archaic2, modern human, ape)
stratified by B-allele frequency
in modern humans

- Disclaimer:
  - this is based on observations from real data, some simulations and discussions in meetings
  - to my knowledge it has not been coherently
    - formally written down
    - explored with simulations
      (but see Supplement S9b, Figures S47-S66 from Prüfer et al. (2017).
      A high-coverage Neandertal genome from Vindija Cave in Croatia.
      which covers most of it)

➔ great future project! ;)

D(archaic1, archaic2, modern humans, ape)

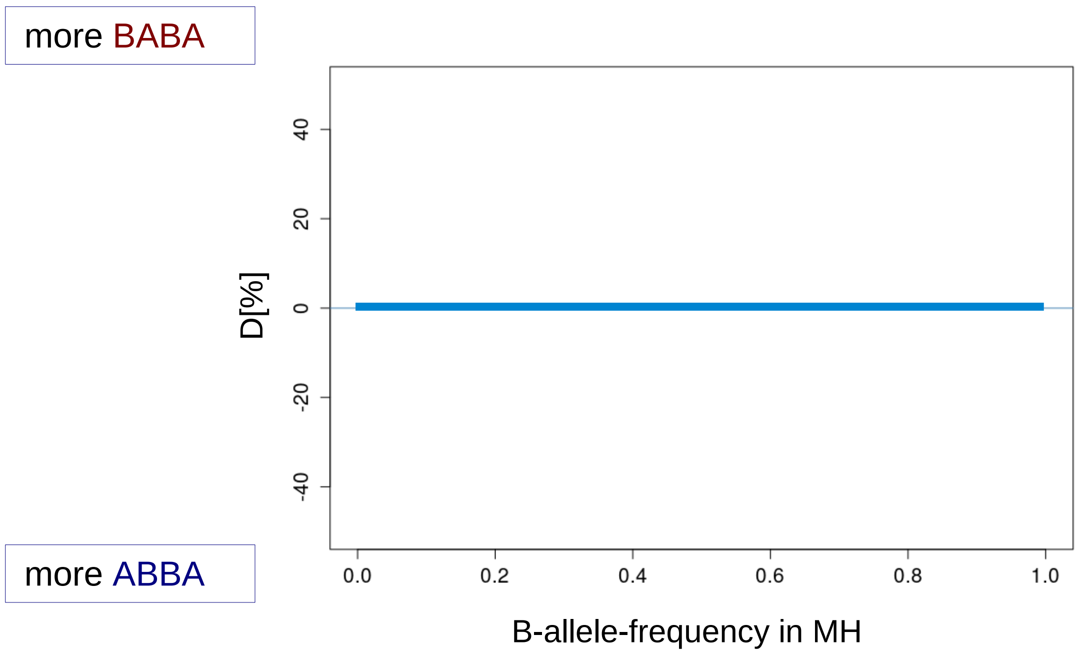
Null-hypothesis:

- modern humans are an outgroup to *archaic1* and *archaic2*
- no introgression

Arch1	Arch2	MH	ape
A	B	B	A
Arch1	Arch2	MH	ape
B	A	B	A

then:

- **numbers of ABBA and BABA sites are equal**
- independent of modern human allele frequencies (?)
(effect of different *Ne* etc in archaics to be checked)
- D = (BABA-ABBA) / (BABA+ABBA)
- E(D) = 0

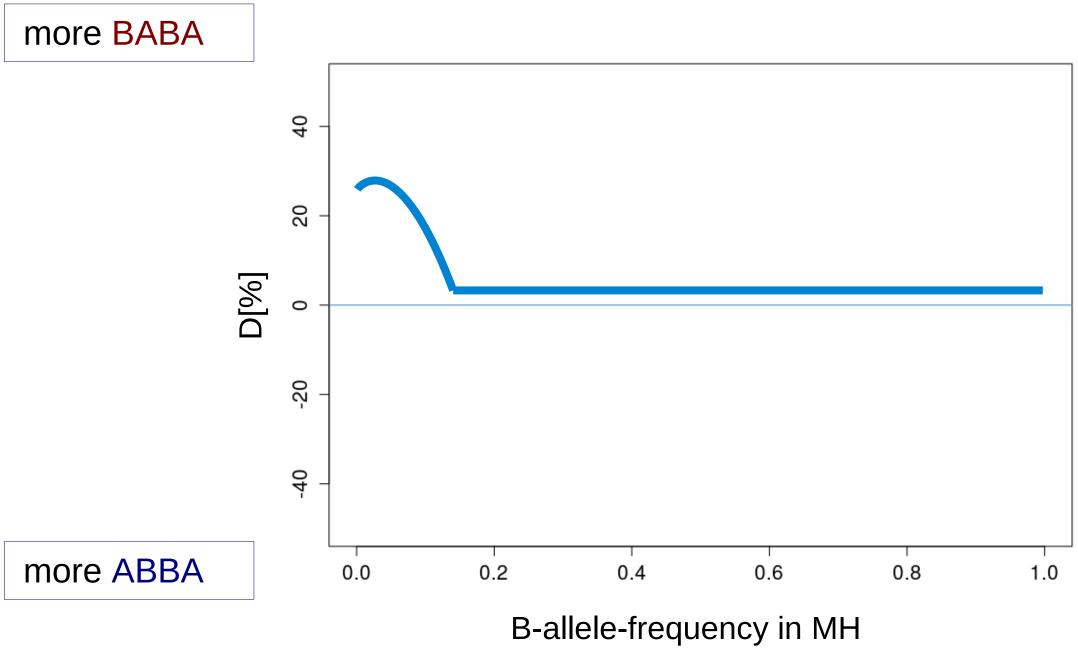


(1) effect of introgression

archaic → modern humans

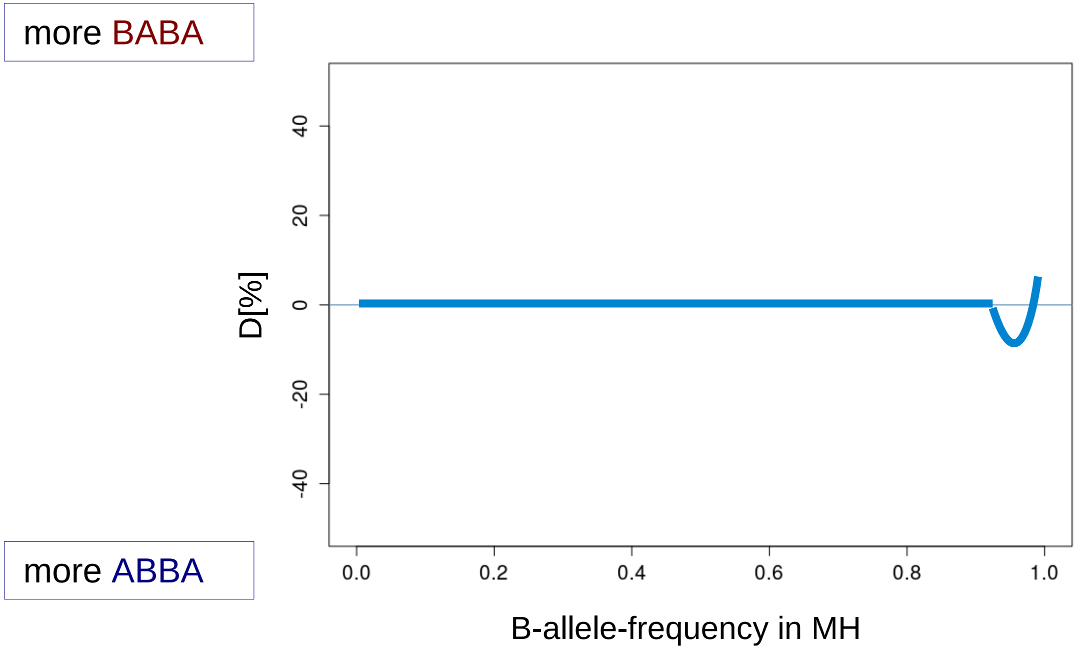
- example for introgression from *archaic1*
    - **effect 1: BAAA → BABA**
    - introgression of *B* allele

- strongest at introgressed allele frequency
- stronger the more diverged *archaic1* and *archaic2* are



- example for introgression from *archaic1*
    - **effect 2: ABBA → ABAA**
    - introgression of *A* allele

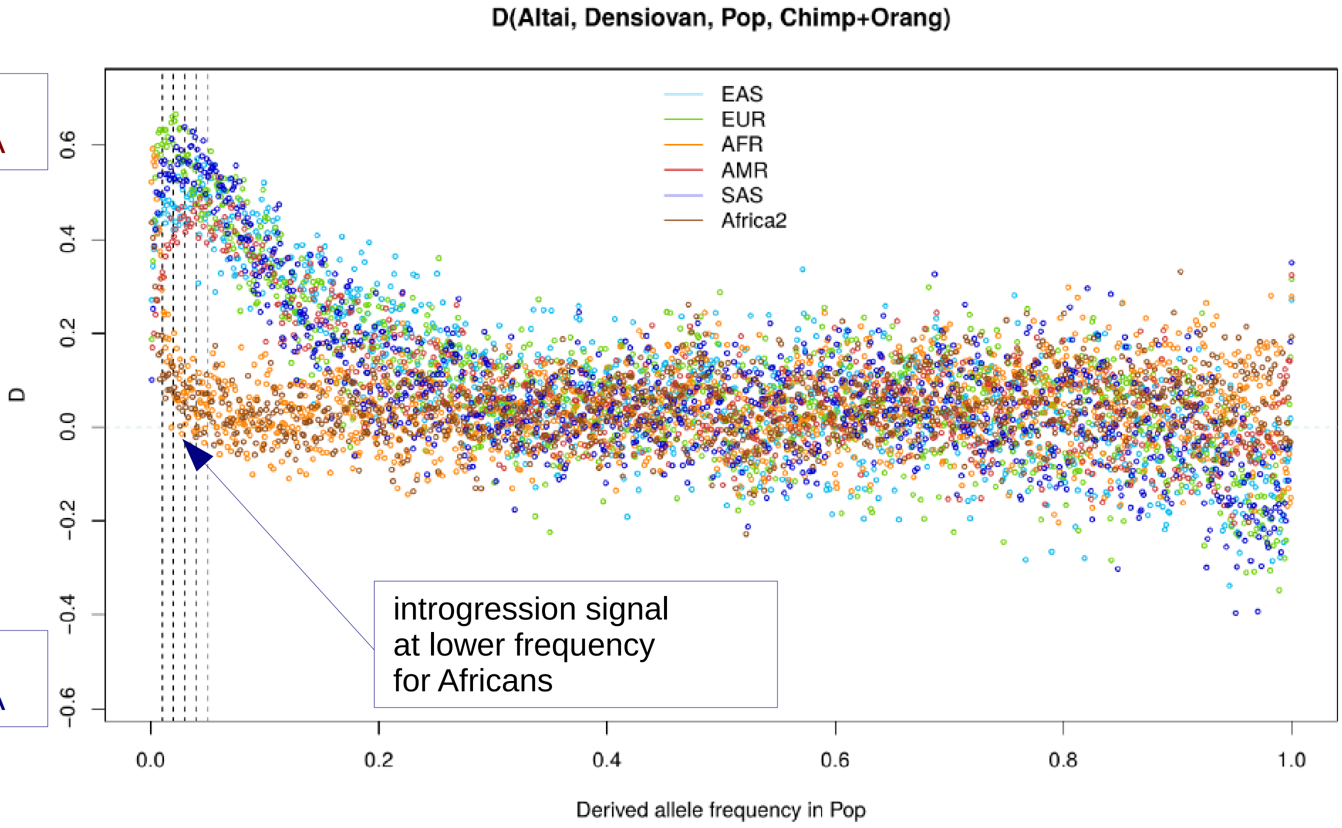
- strongest at introgressed allele frequency
    - *A*-allele-frequency = 1 - *B*-allele-frequency
    - **ABBA** sites get removed from fixed to high-frequency



- example for introgression with high-coverage genomes:
  - *Altai vs. Denisova* → introgression from Neandertals

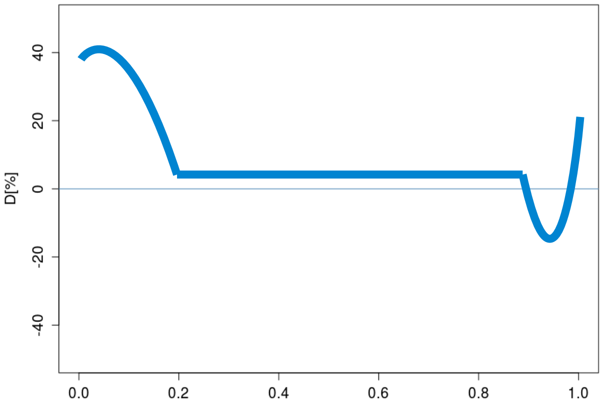
more
BABA

more
ABBA



introgression signal
at lower frequency
for Africans

- introgression:



introgression
BAAA → BABA

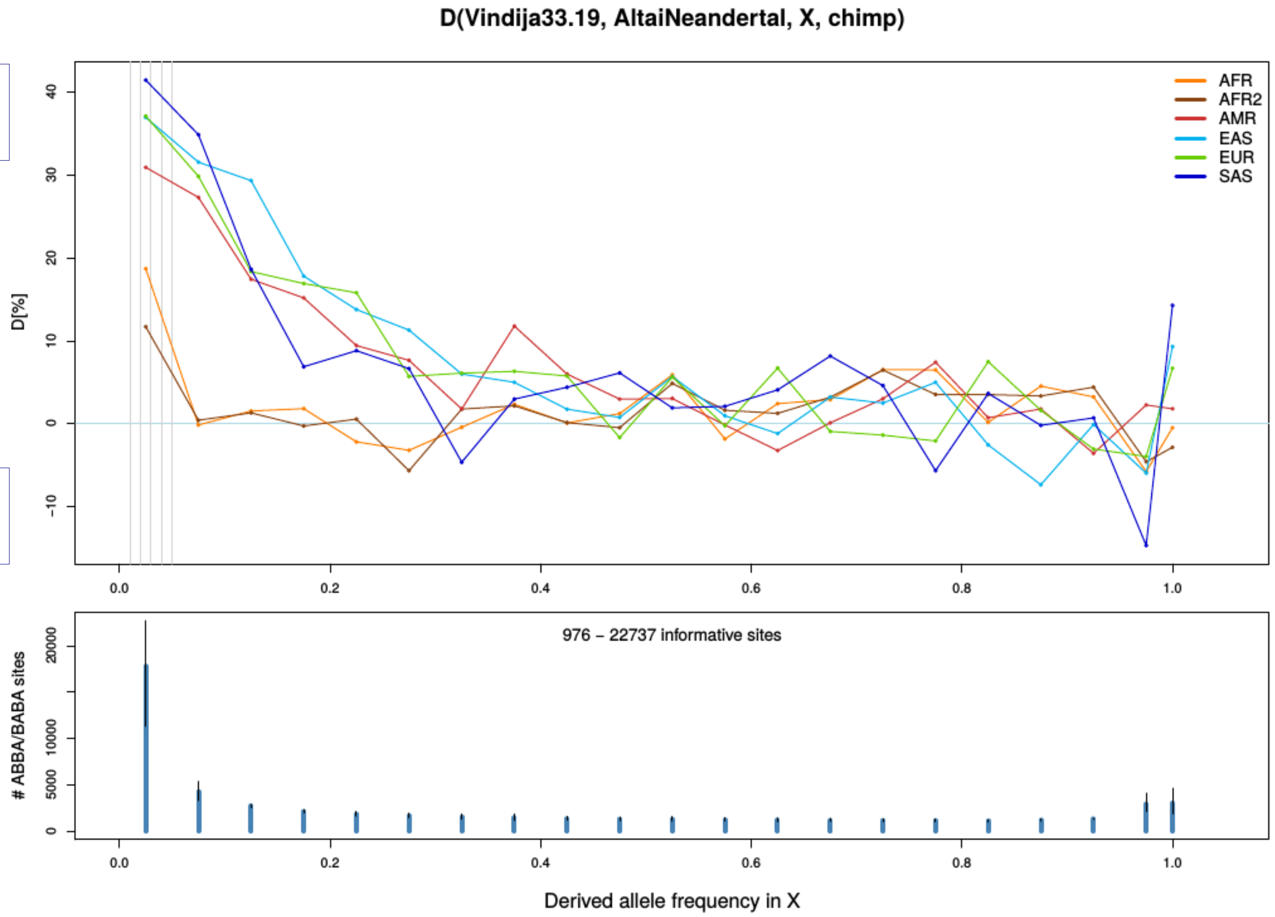
introgression
of ancestral
allele
ABBA →
ABAA:

fixed ABBA
gets
converted to
high-freq
ABBA

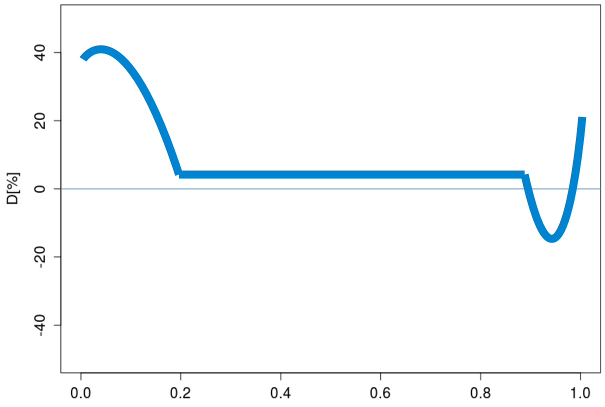
- example for introgression with high-coverage genomes:
  - *Vindija* vs. *Altai* → *Vindija* is closer to introgressing Neandertal

more BABA

more ABBA



- introgression:

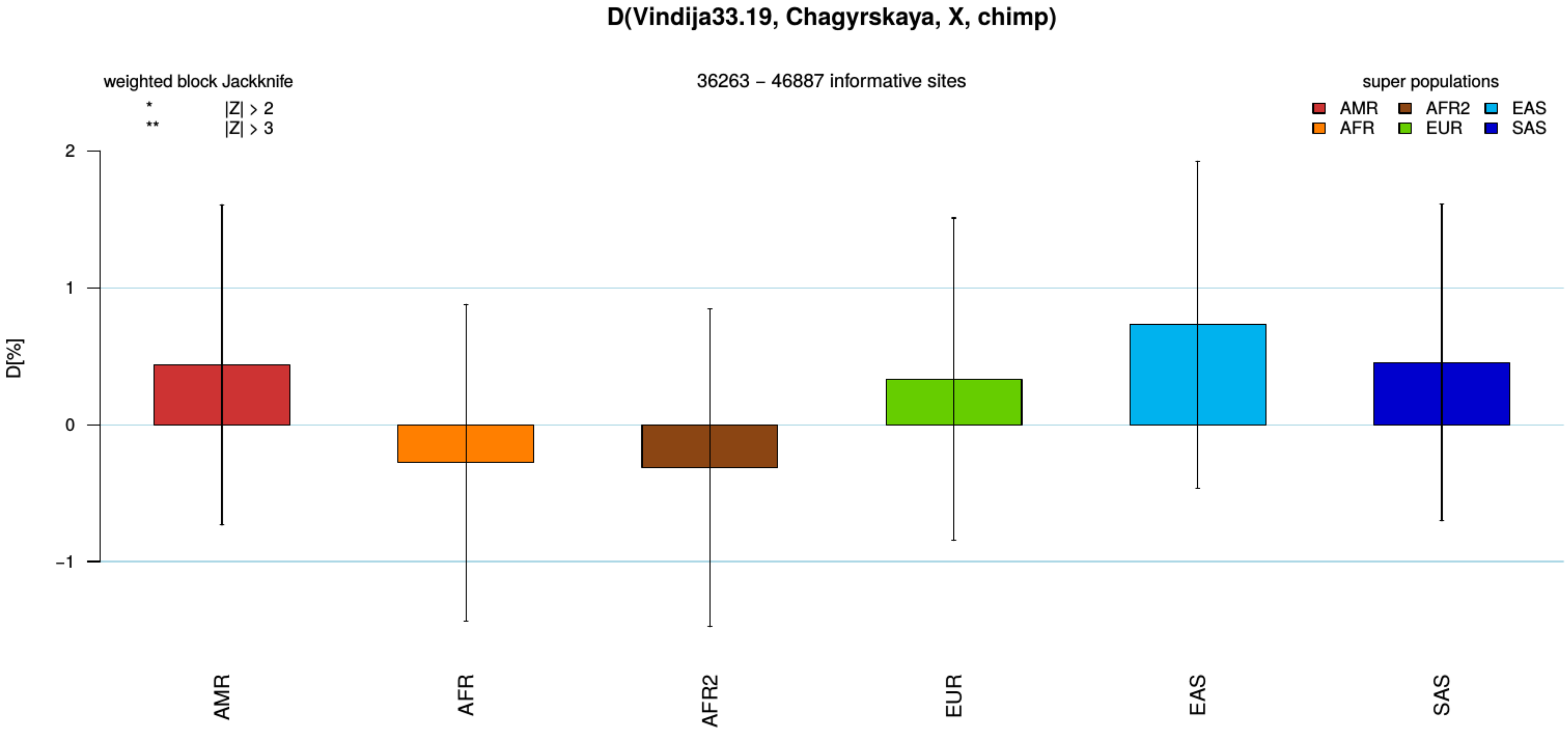


introgression
BAAA → BABA

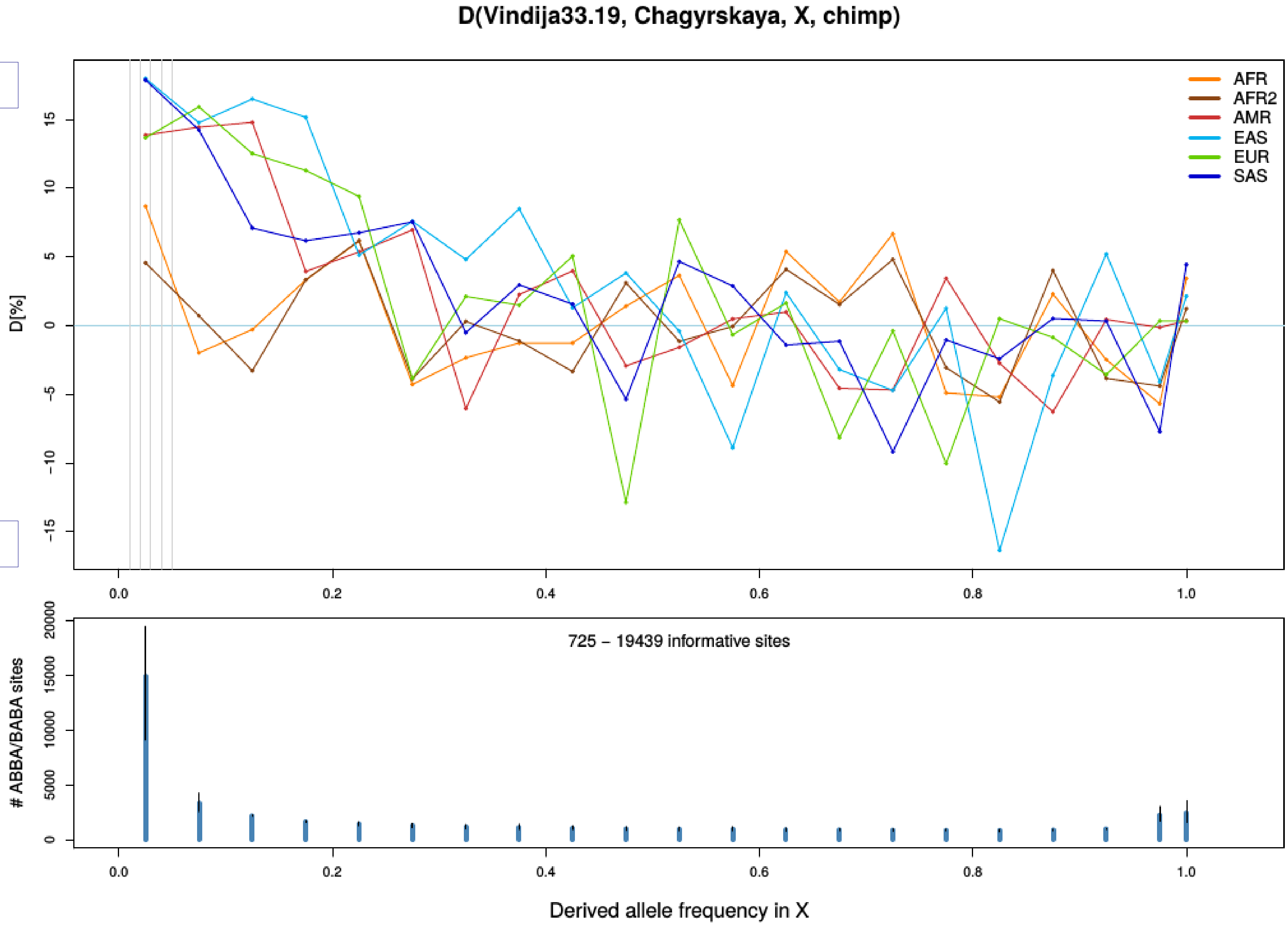
introgression of ancestral allele ABBA → ABAA:

fixed ABBA gets converted to high-freq ABBA

- example for how stratified D-statistics can increase the power:
    - regular D-statistics show no significant difference between *Vindija* and *Chagyrskaya*



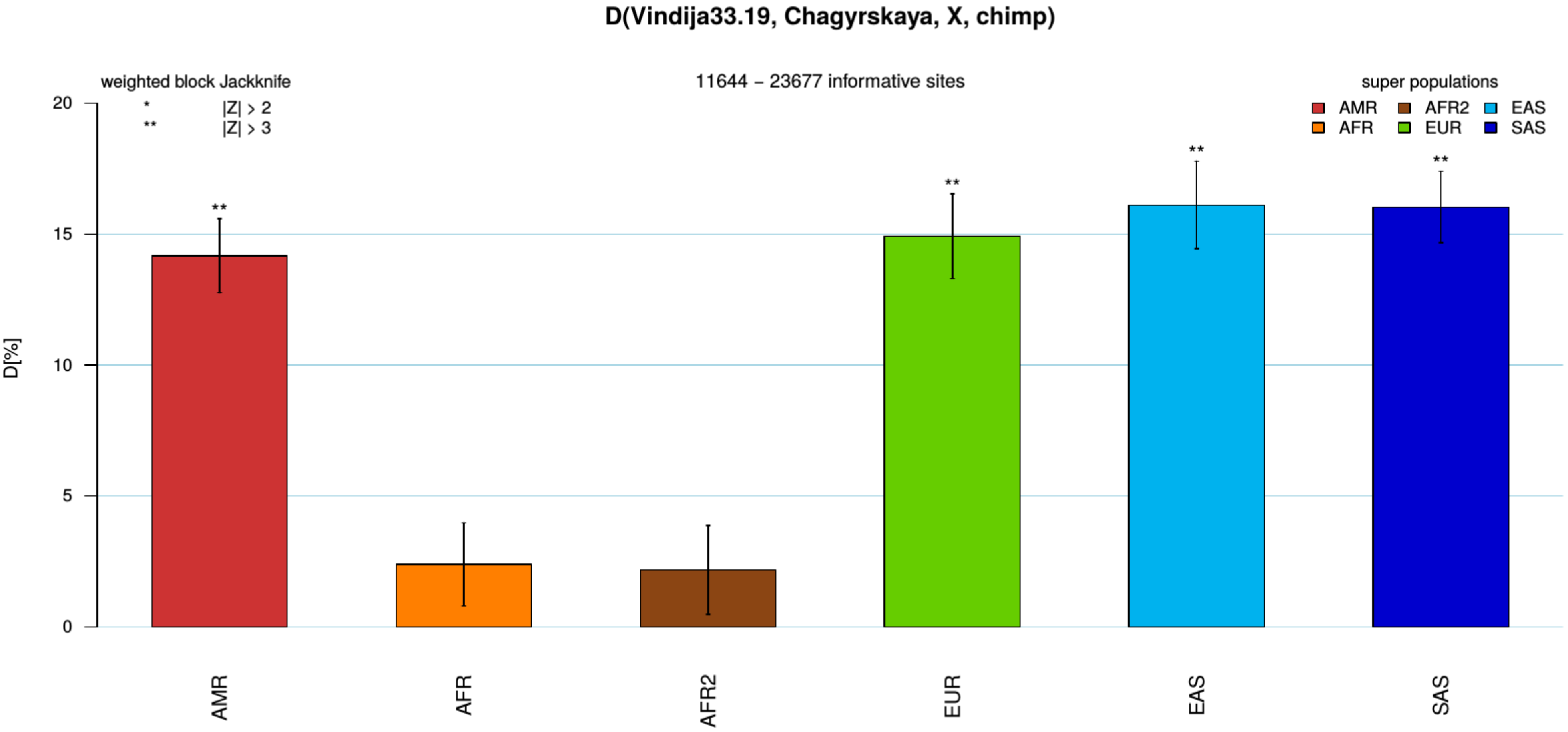
- at low *B*-frequencies in modern humans *Vindija* shares more derived alleles with modern humans than *Chagyrskaya*
- → *Vindija* is closer to the introgressing Neandertal than *Chayrskaya*



more BABA

more ABBA

- filter for **B-frequency <= 10%** in modern humans:
  - signal for *Vindija* being closer to modern humans gets significant

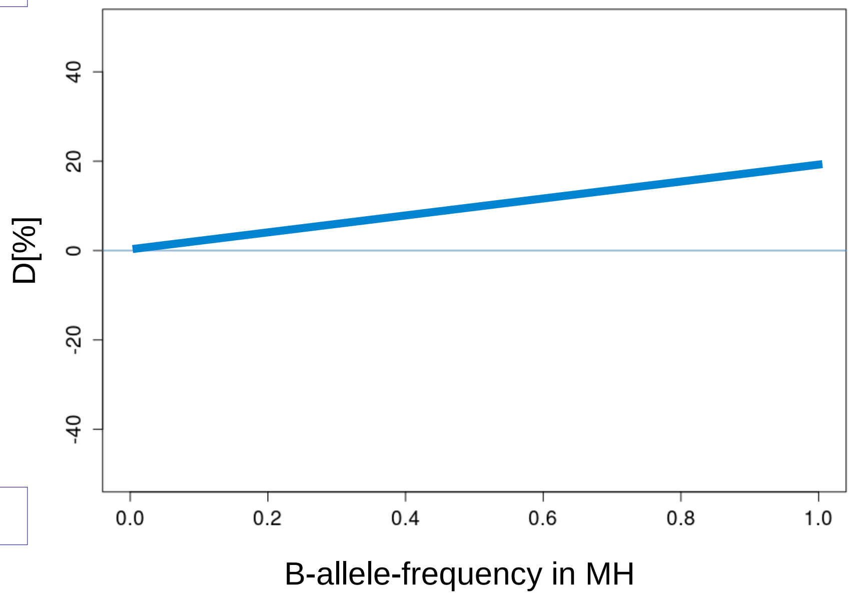


(2) effect of contamination

(or MH → archaic introgression)

- example for modern human contamination into archaic1
  - **effect 1: AABA → BABA**

- correlated with MH allele frequency
  - contaminant more likely to share the B-allele with rising frequency of B



more BABA

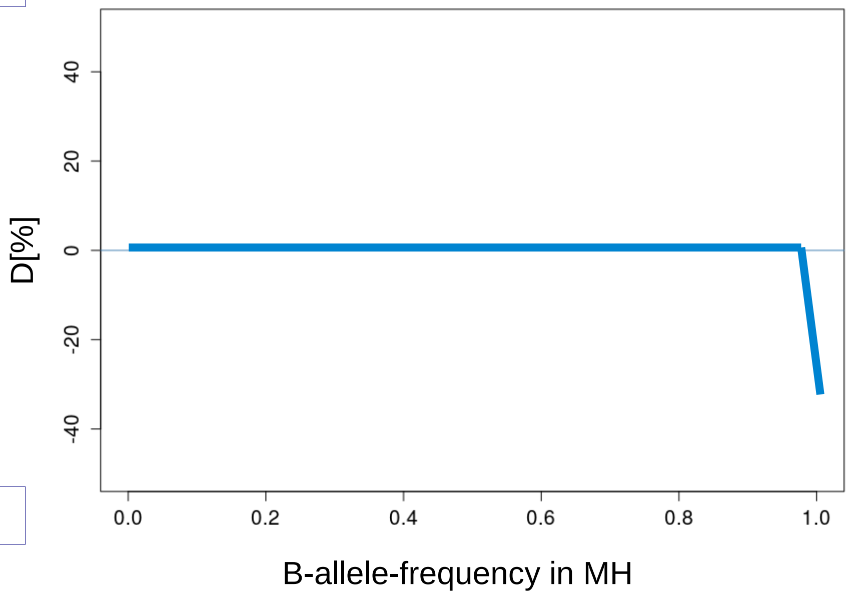
more ABBA

(3) effect of error

- example for more errors in archaic1
    - **effect 1: BBBA → ABBA**

- most visible at fixed B in MH (most BBBA sites are fixed)
- gets stronger with outgroup branch length (more BBBA sites)

more BABA

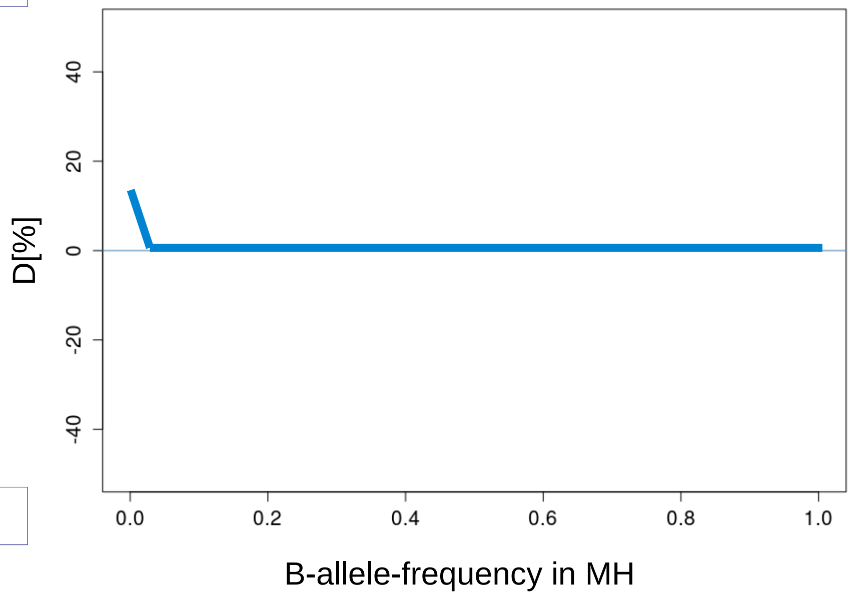


more ABBA

- example for more errors in archaic1
    - **effect 2: AABA → BABA**

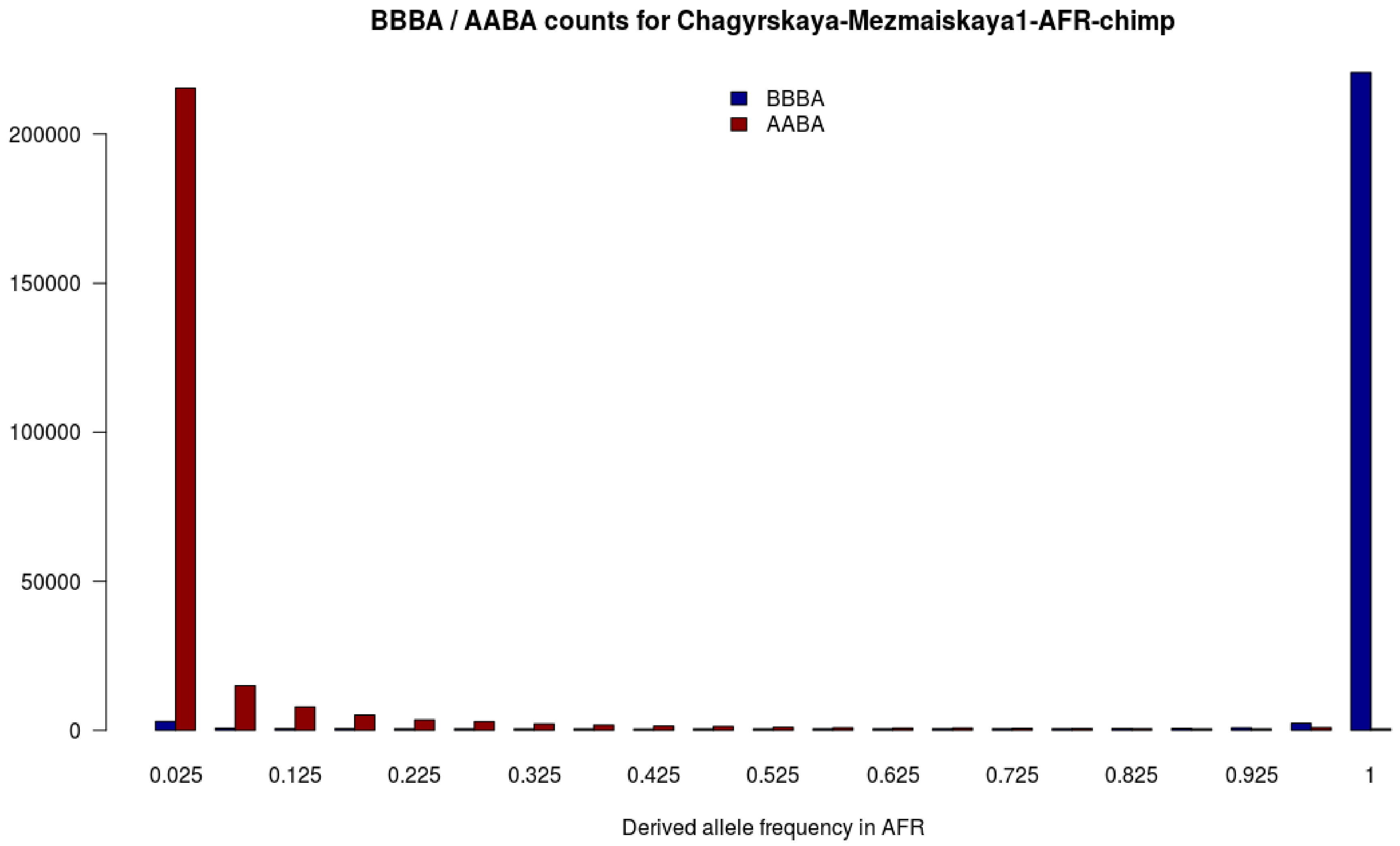
- most visible at low B-frequency in MH (most AABA sites are low frequency)
- more effect in Africans (more AABA sites)

more BABA

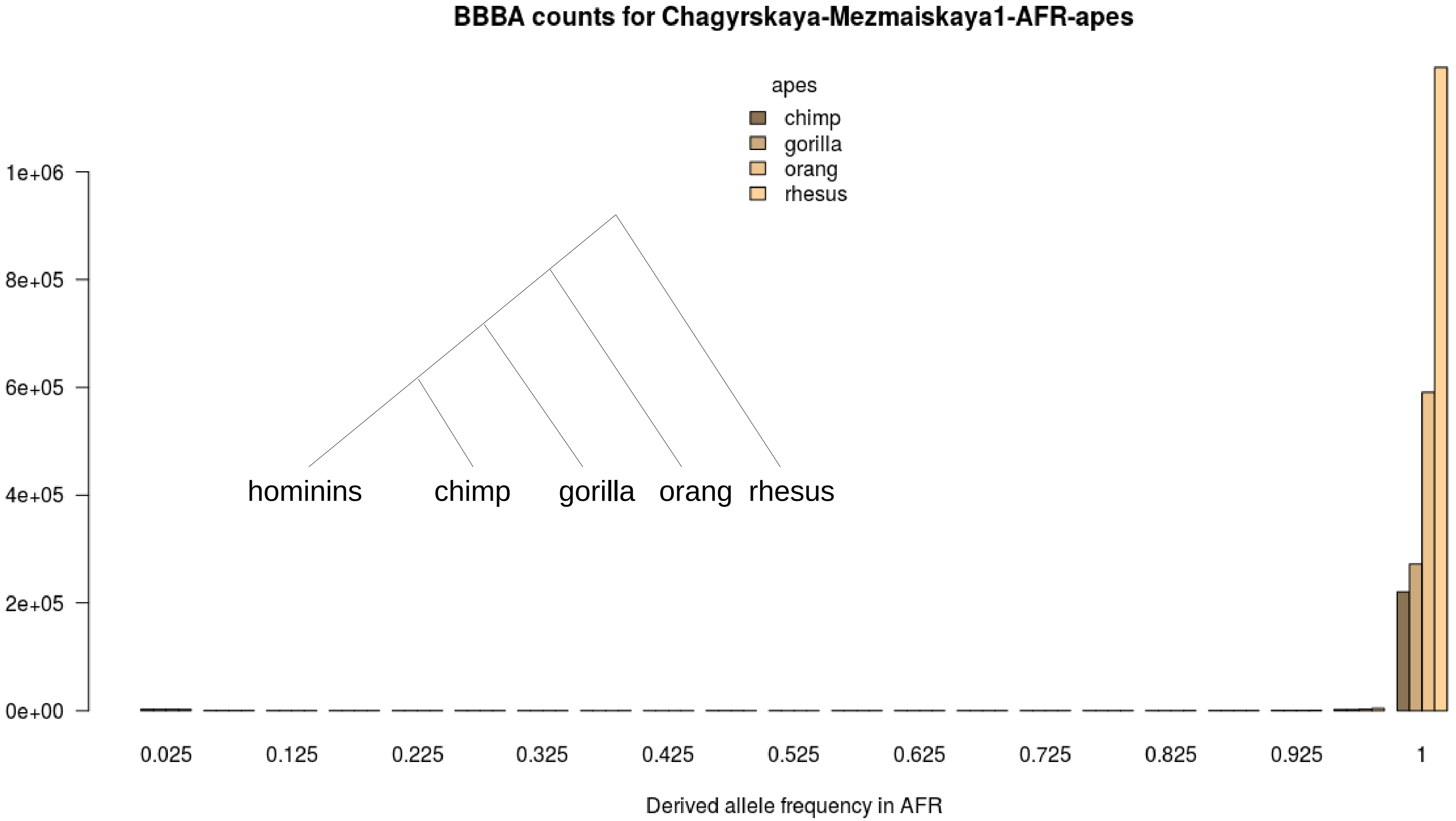


more ABBA

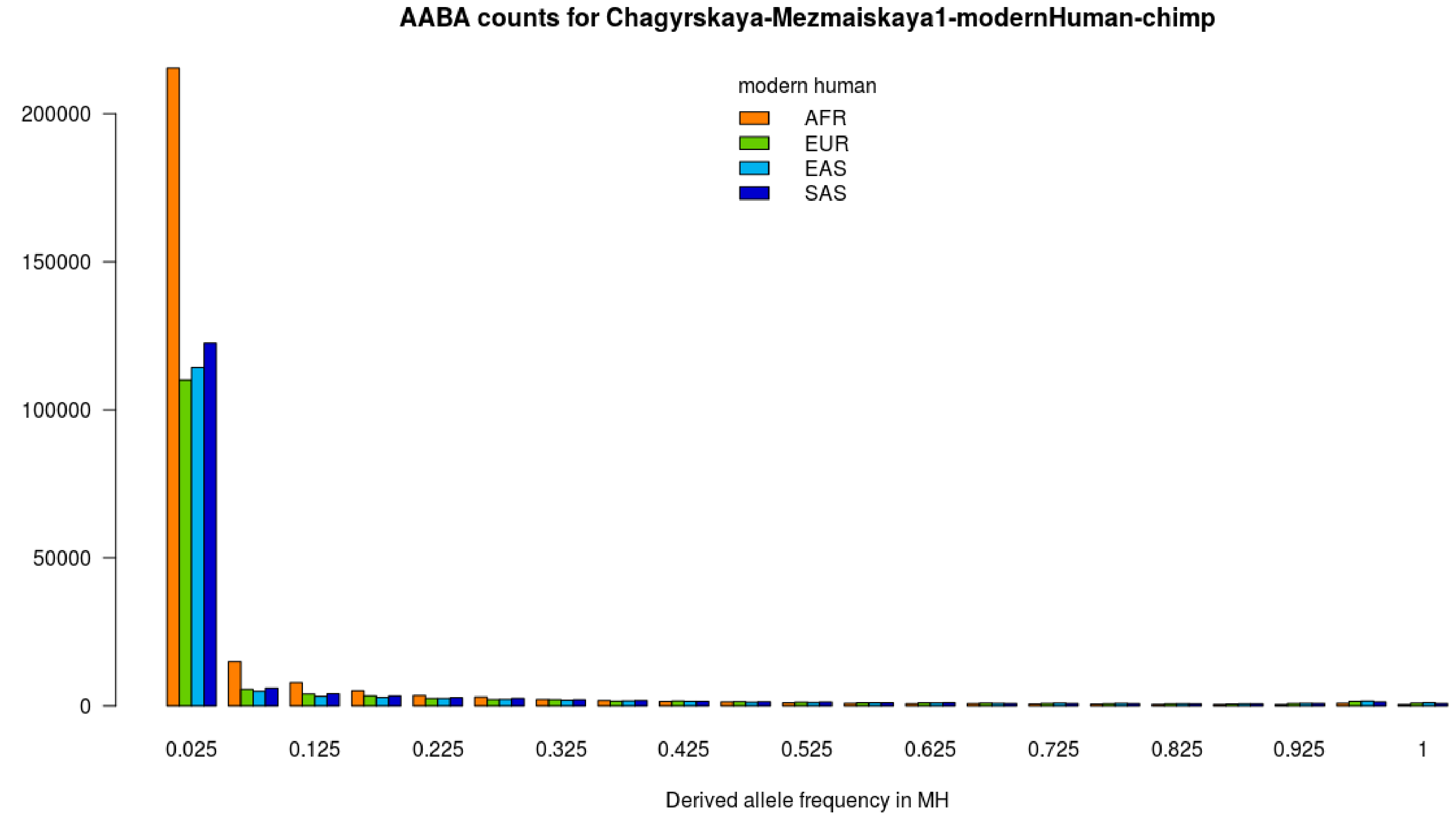
- example of BBBA counts and ABAA counts per MH B-allele-frequency
- starting point for effect 1 and effect 2 from errors in *archaic1*



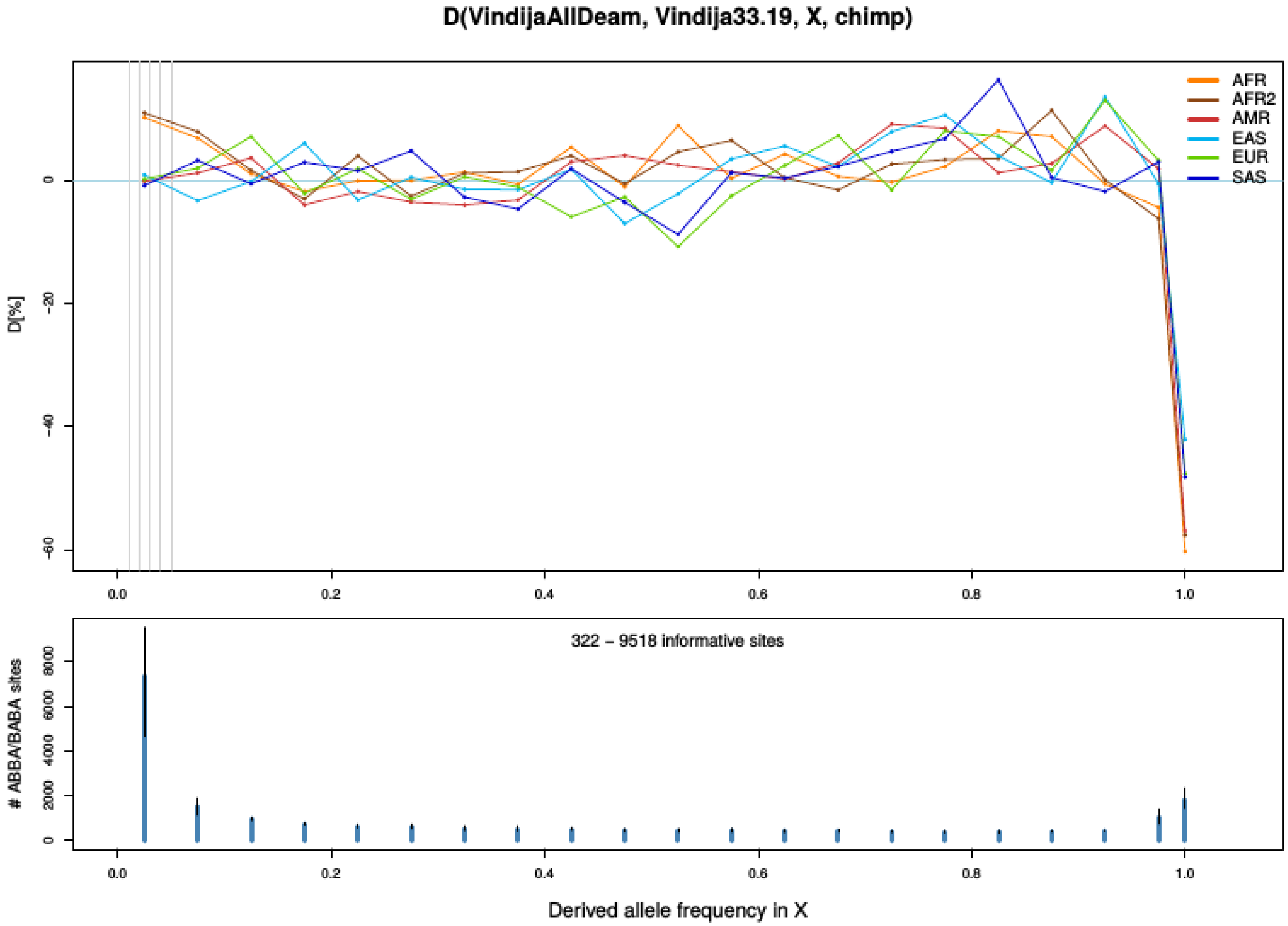
- example of BBBA counts per MH B-allele frequency for different outgroup branch lengths
- starting point for effect 1 from errors in *archaic1*



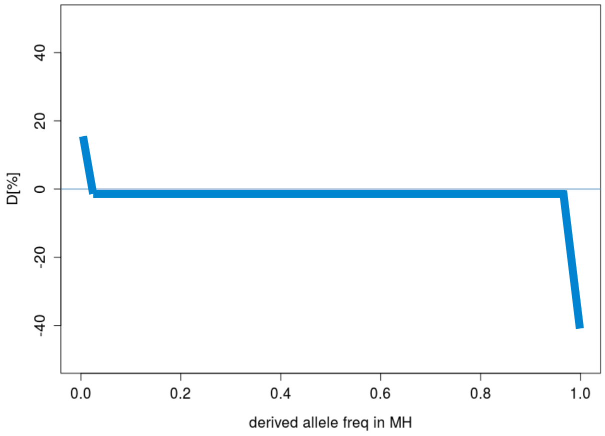
- example of AABA counts per MH B-allele frequency for different modern human populations
- starting point for effect 2 from errors in *archaic1*



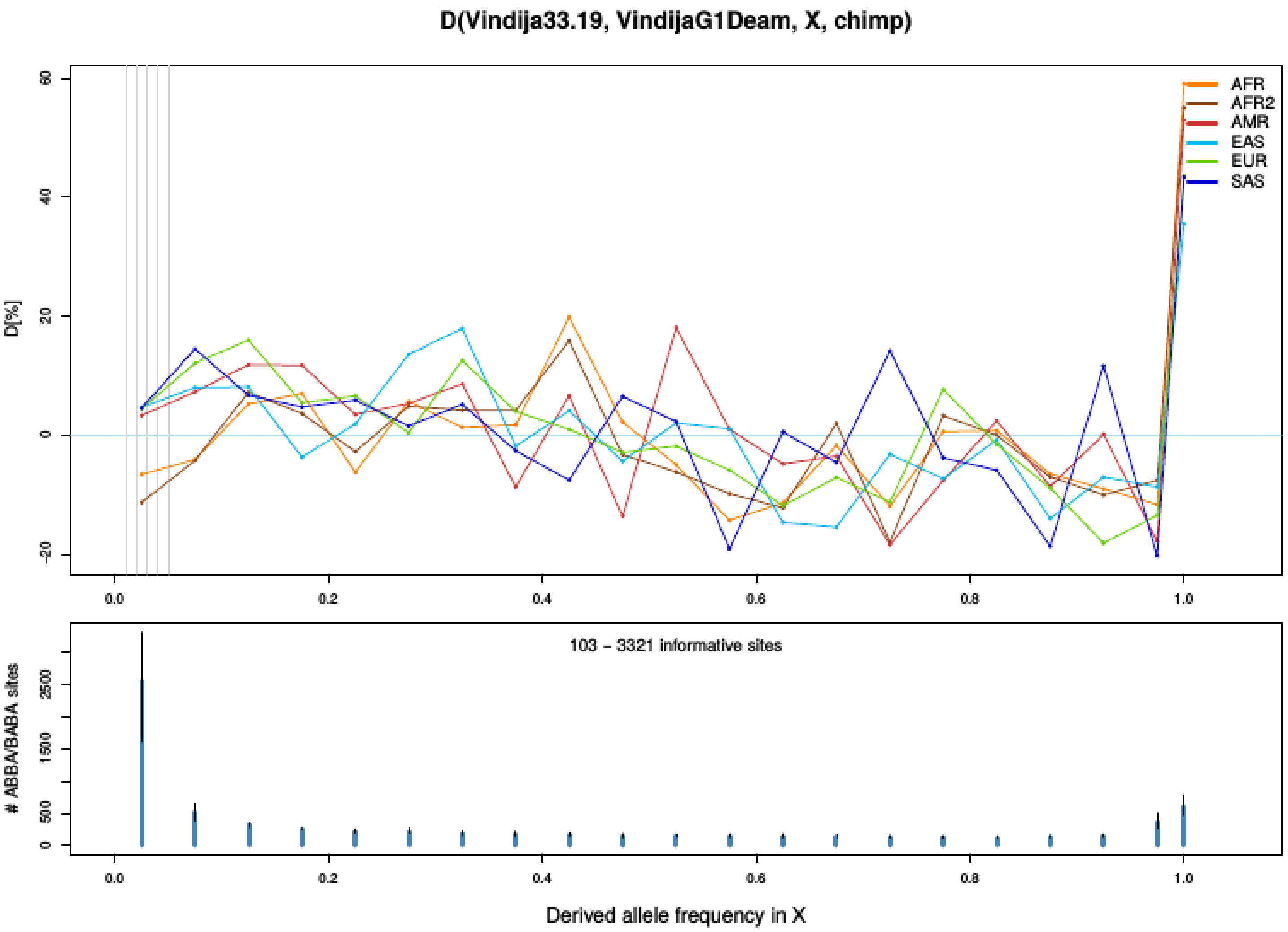
- example for effect of errors:
  - high-coverage *Vindija33.19* genotypes vs. *Vindija33.19* random reads
  - same individual → 0-hypothesis of equidistance to introgressing Neandertal is true



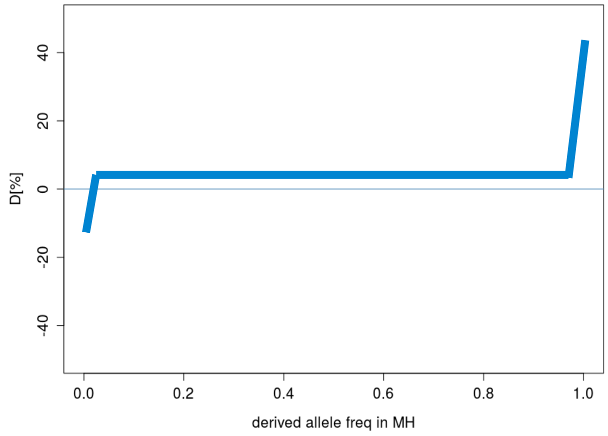
- more errors:



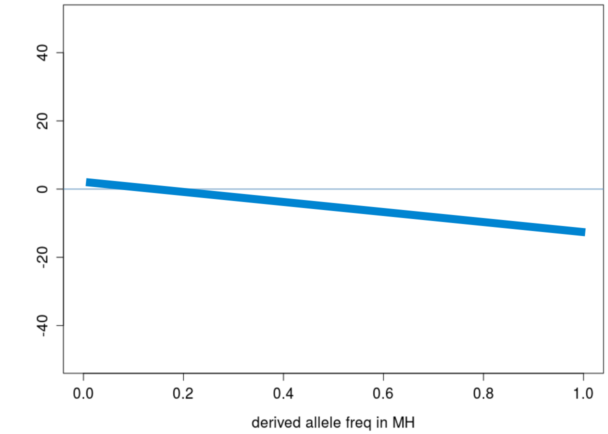
- example for effect of errors + contamination:
  - high-coverage *Vindija33.19* genotypes vs. *VindijaG1* random reads
  - same individual → 0-hypothesis of equidistance to introgressing Neandertal is true



- more errors:



- contamination:

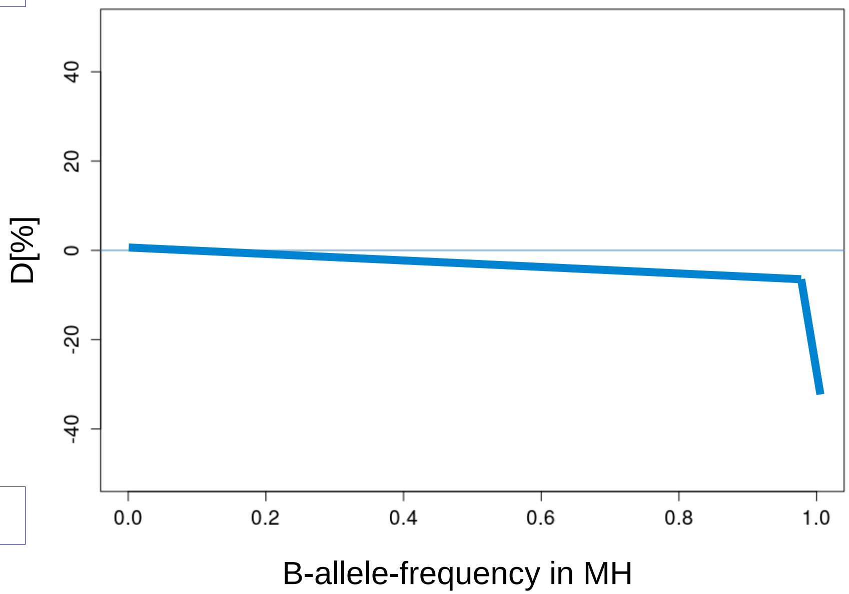


(4) effect of super-archaic introgression

- example for super-archaic introgression into archaic1
    - **effect 1: BBBA → ABBA**

- most visible at fixed B in MH (most BBBA sites are fixed)
- should not increase with outgroup branch length (unlike BBBA → ABBA error)

more BABA



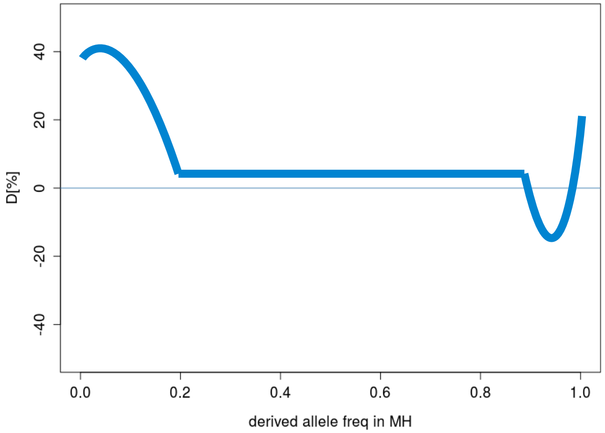
more ABBA

- summary of patterns for B-allele-frequency-stratified D-statistics

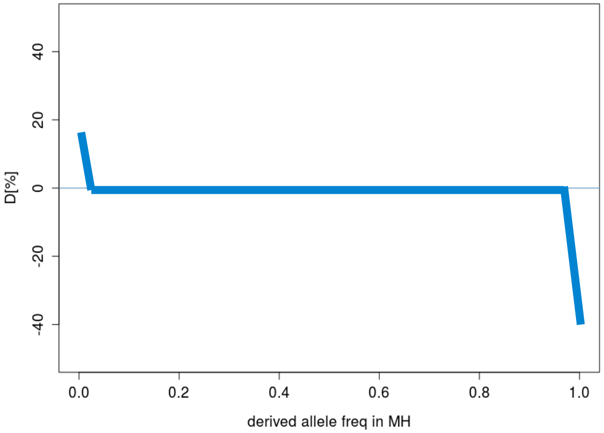
more BABA

more ABBA

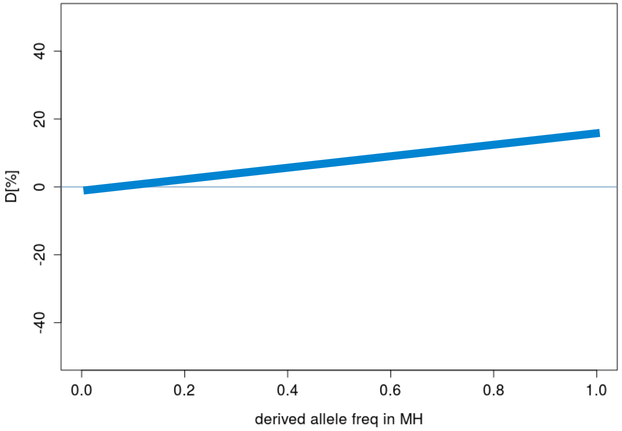
- introgression from archaic1:
  (1) BAAA → BABA
  (2) ABBA → ABAB



- more errors in archaic1:
  (1) BBBA → ABBA
  (2) AABA → BABA



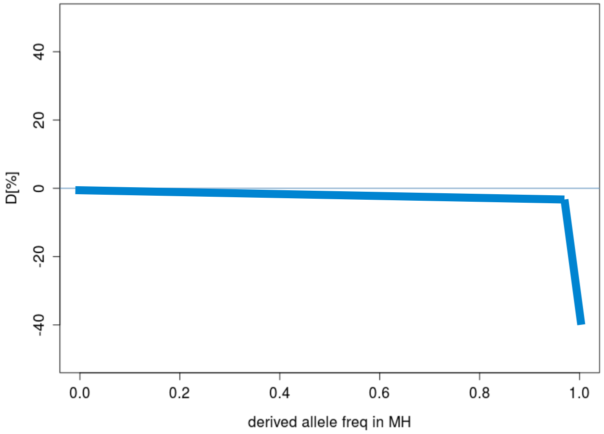
- contamination in archaic1:
  (1) AABA → BABA



more BABA

more ABBA

- superarchaic into archaic1:
  (1) BBBA → ABBA



- summary of patterns for B-allele-frequency-stratified D-statistics

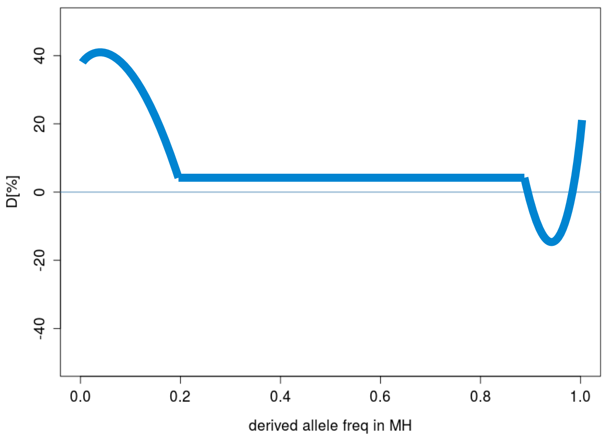
- introgression from archaic1:
  (1) BAAA → BABA
  (2) ABBA → ABAB

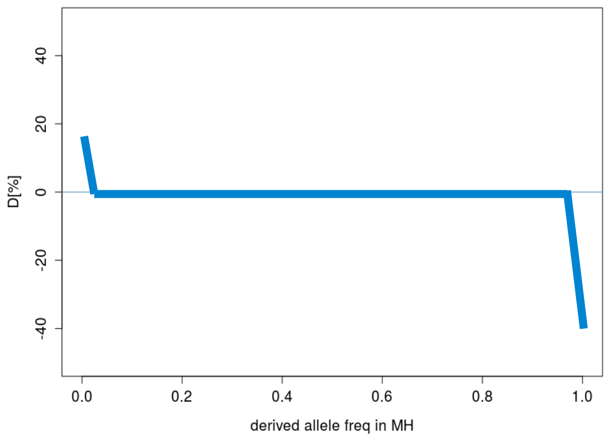
- more errors in archaic1:
  (1) BBBA → ABBA
  (2) AABA → BABA

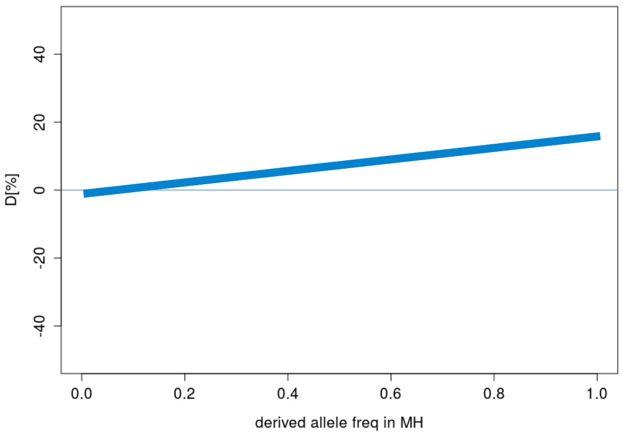
- contamination in archaic1:
  (1) AABA → BABA

more BABA

more ABBA



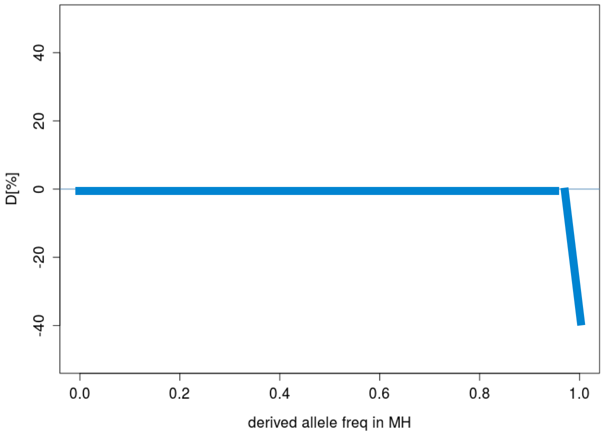




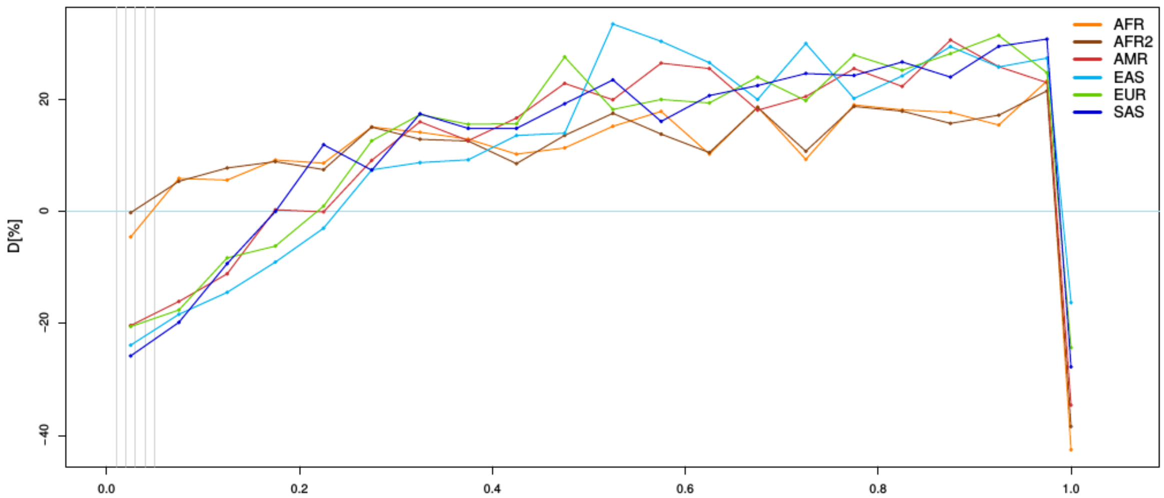
- superarchaic into archaic1:
  (1) BBBA → ABBA

more BABA

more ABBA



D(Mezmaiskaya1, Vindija33.19, X, chimp)



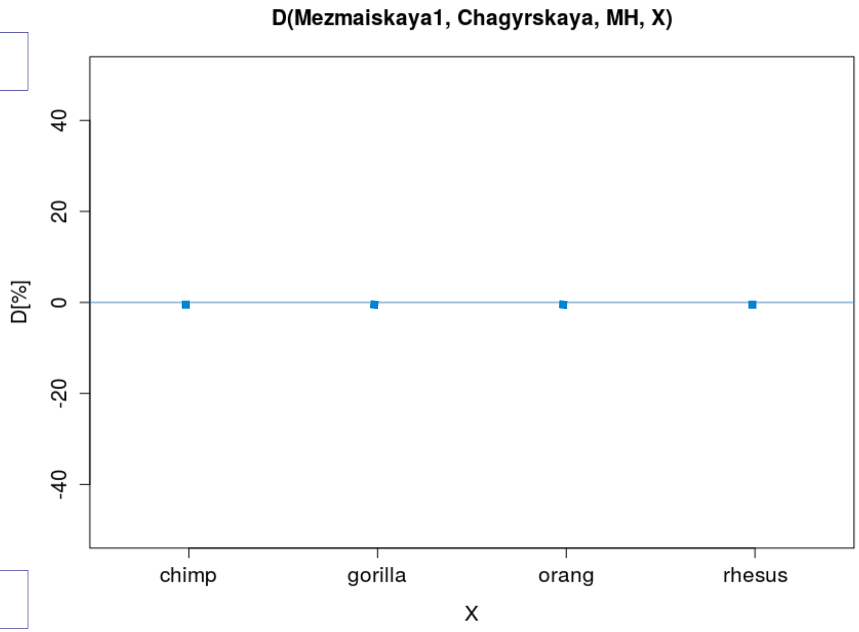
(4) effect of errors and outgroup branch length

- example for error in *archaic1*
    - → bias genomewide D towards Chagyrskaya
    - **effect 1: BBBA → ABBA**

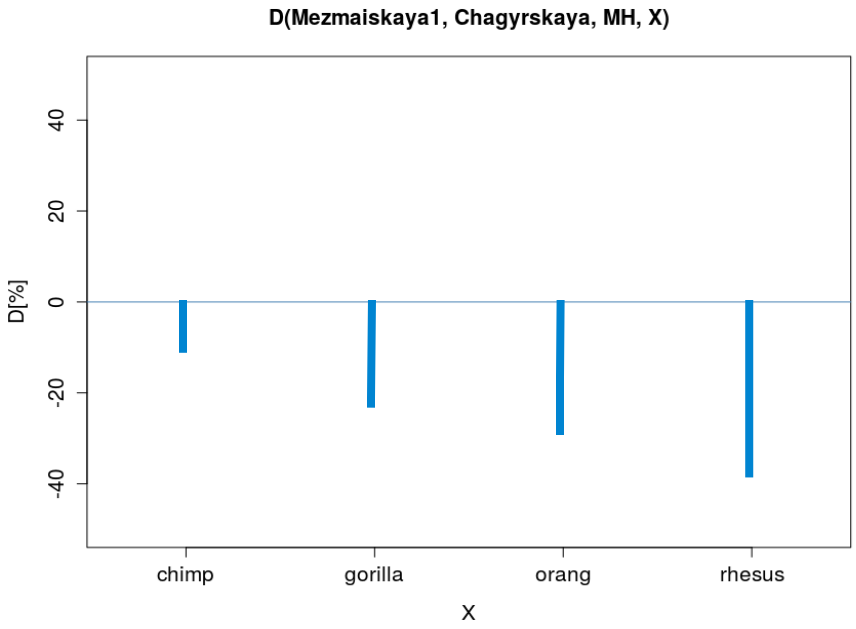
- expectation if *Mezmaiskaya1* and *Chagyrskaya* are equally close to introgressing Neandertal:
    - with same quality:

more BABA

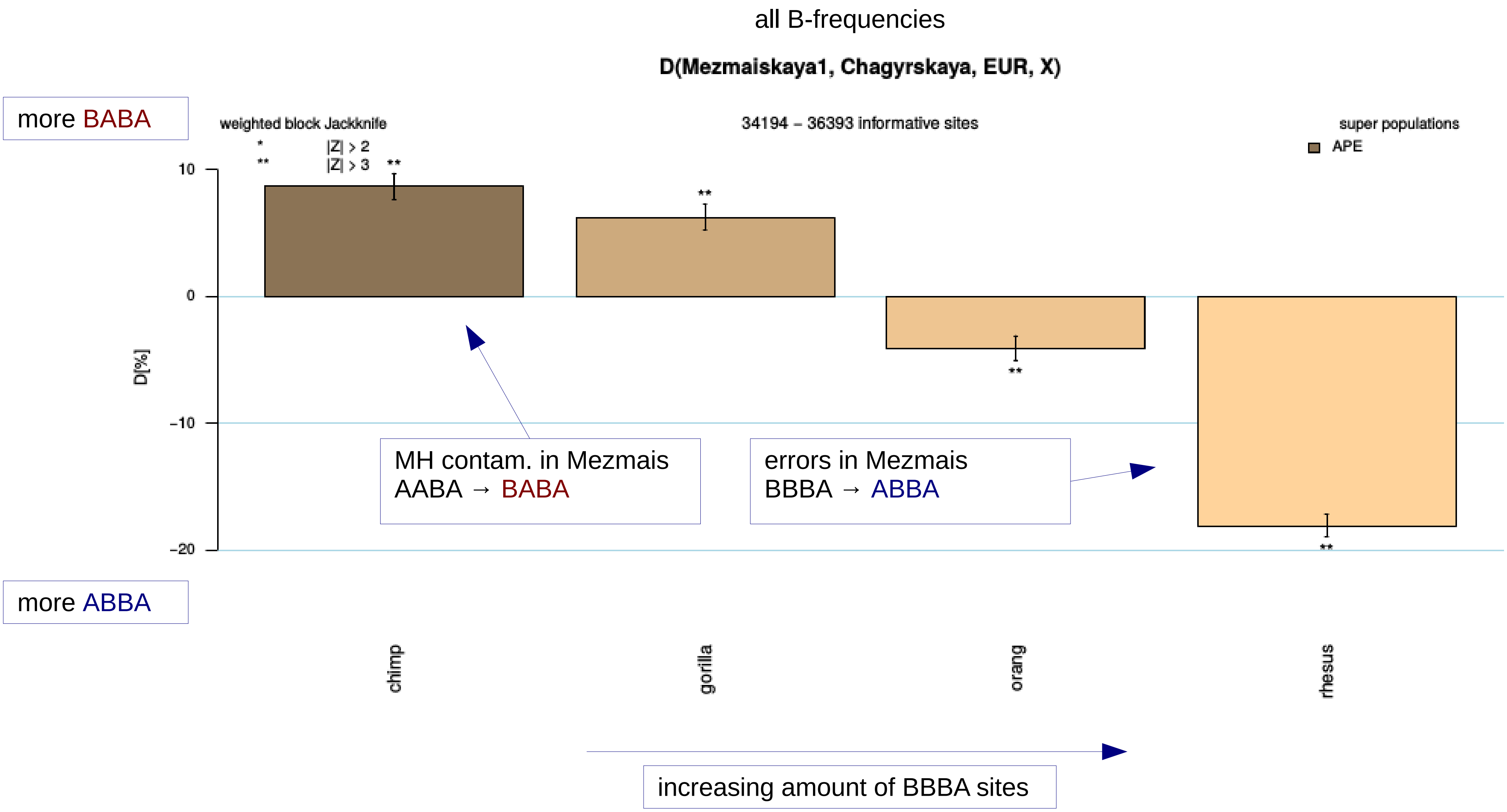
more ABBA



- with more errors in Mez1:



- check for the effect of long branch attraction using different outgroups
- 1) for D using all sites the expected effect is very strong



all B-frequencies

more BABA

more ABBA

MH contam. in Mezmais
AABA → BABA

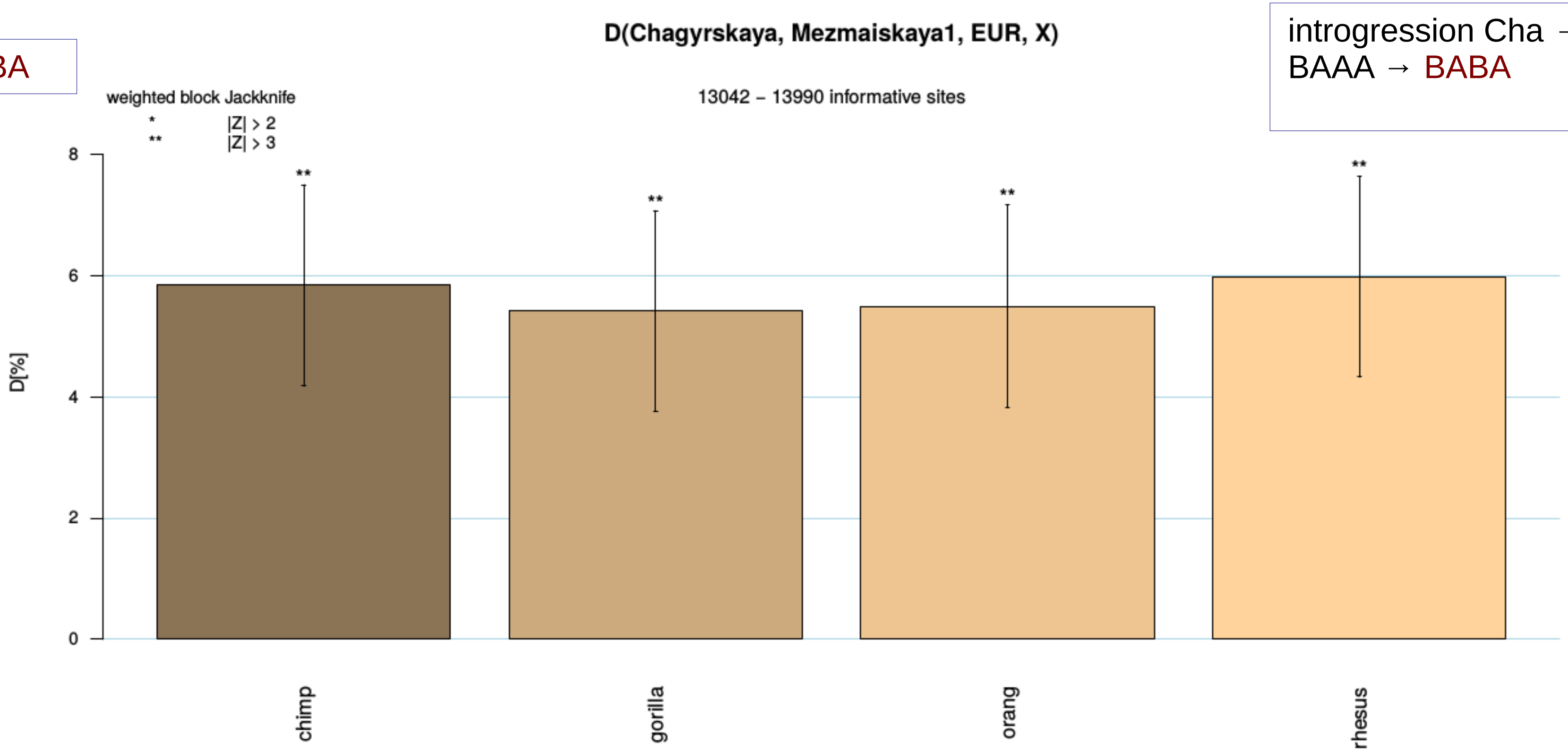
errors in Mezmais
BBBA → ABBA

increasing amount of BBBA sites

- 2) for D using only sites with B-freq <= 10% however long branch attraction is not observed at all
    - because most BBBA sites are fixed B in modern humans
        - ➔ errors affect mostly fixed B (also see freq-stratified D above)
        - ➔ for low frequency B in modern humans there are few BBBA sites that can be converted to ABBA sites

B <= 10% in EUR

more BABA



introgression Cha → EUR
BAAA → BABA

more ABBA

also note:
stratifying by B-allele-frequency in
pop1 or *pop2*
makes no sense

- high B-freq in *pop2*:
  - low A-freq in *pop2*
  - more ABBA

- low B-freq in *pop2*:
  - high A-freq in *pop2*
  - more BABA

- this was also confusing Bill Amos, who claimed D(Afr, Eur, Nean, out) should have a stronger introgression signal at low B-frequencies in Europeans. But in fact he observed the opposite and interpreted that as evidence against introgression theory.

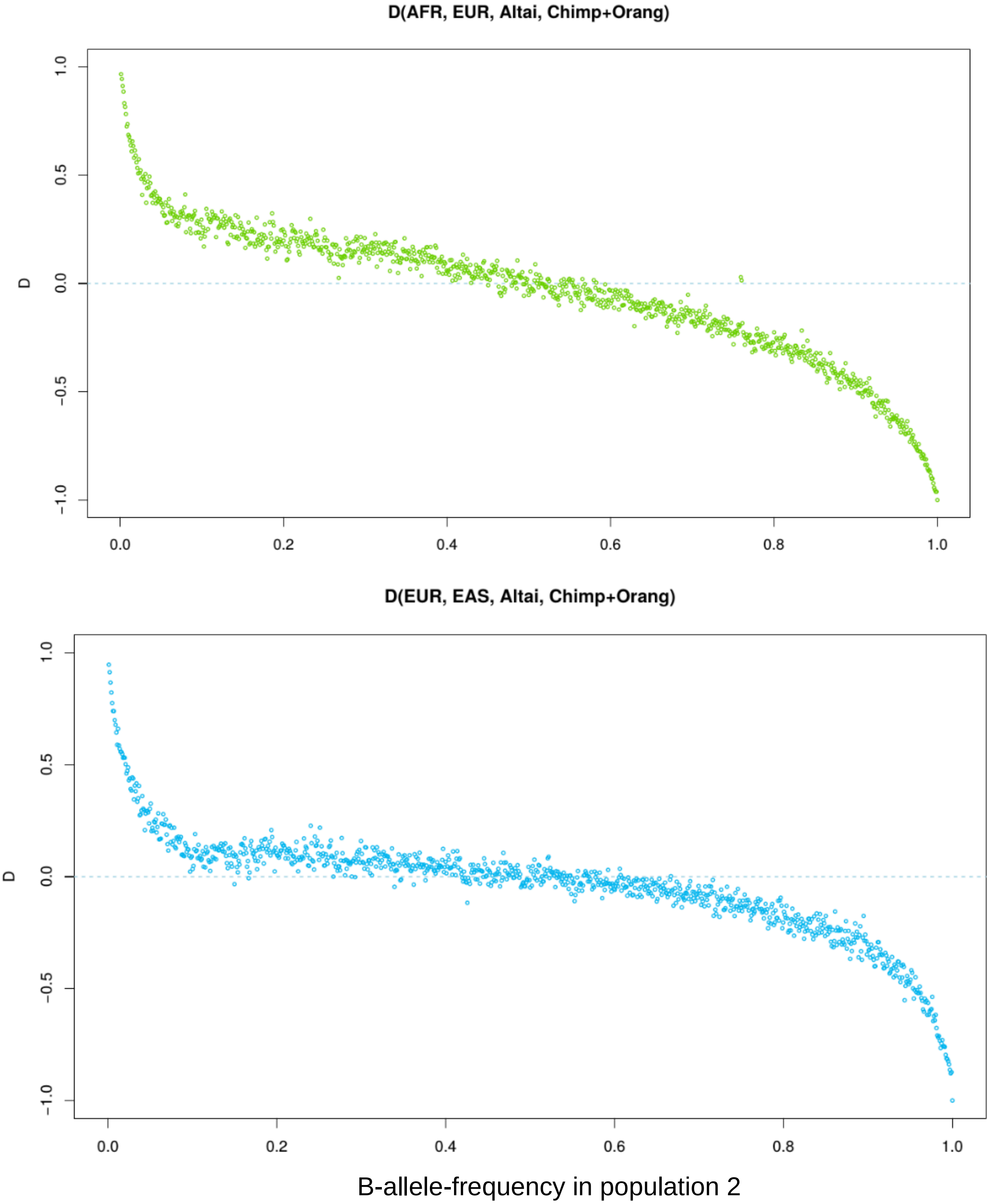
  In fact one will always observe a pattern like on the right, independent of which MH pops are used

more
BABA

more
ABBA

more
BABA

more
ABBA


D(AFR, EUR, Altai, Chimp+Orang)
D
D(EUR, EAS, Altai, Chimp+Orang)
D
B-allele-frequency in population 2