Lesson 7: Outgroup f3-statistics

Friday July 21, 2018: 9:00 – 11:30 am

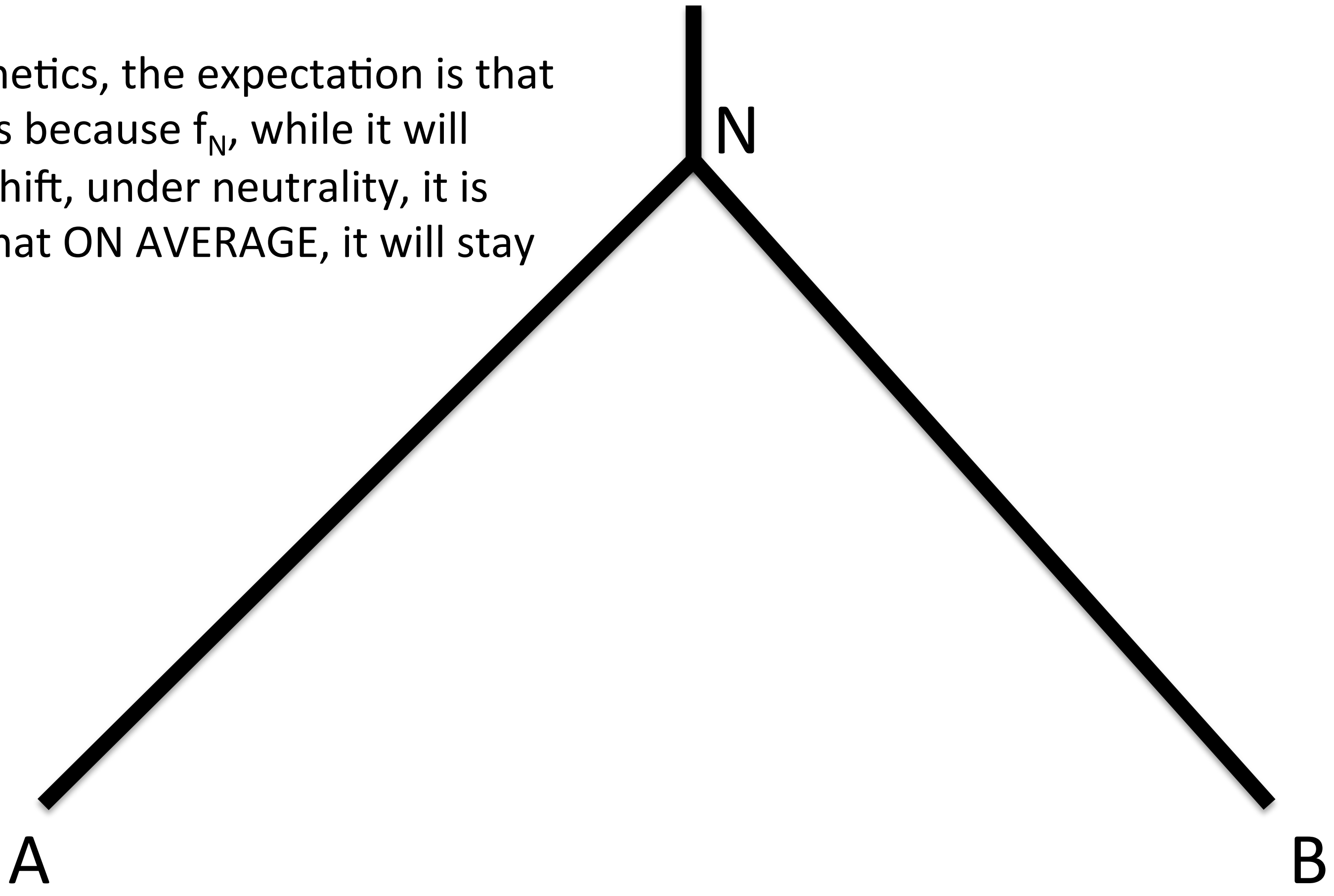
Up to Now

- Specific hypothesis testing – D-statistics
- Survey tool – PCA
- An f-statistic based survey tool – outgroup f3-statistics
  - Originated as a test of admixture, which you will learn next lesson.
  - Here, we are considering specifically:

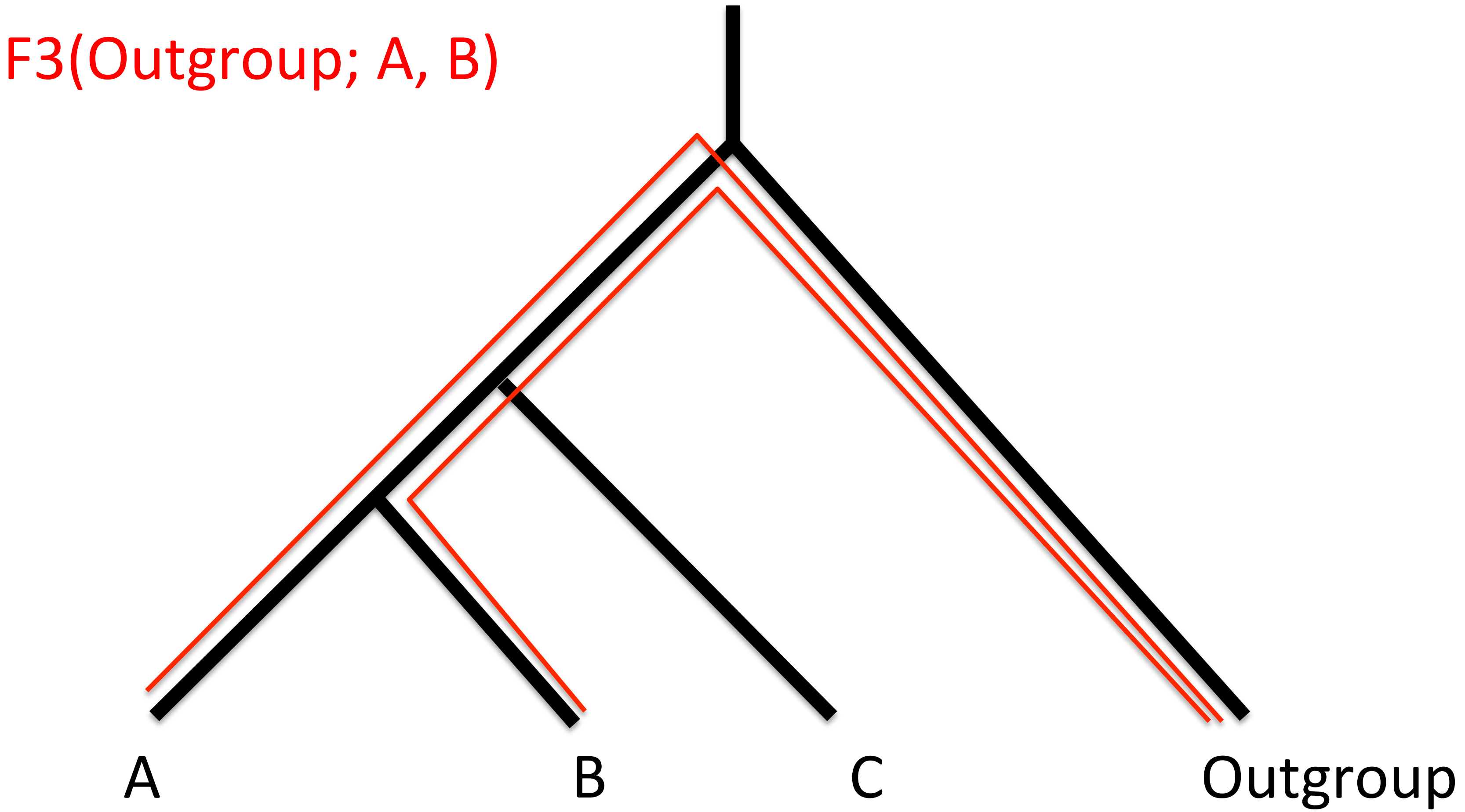
f3(X, Y; Outgroup)

F-statistics

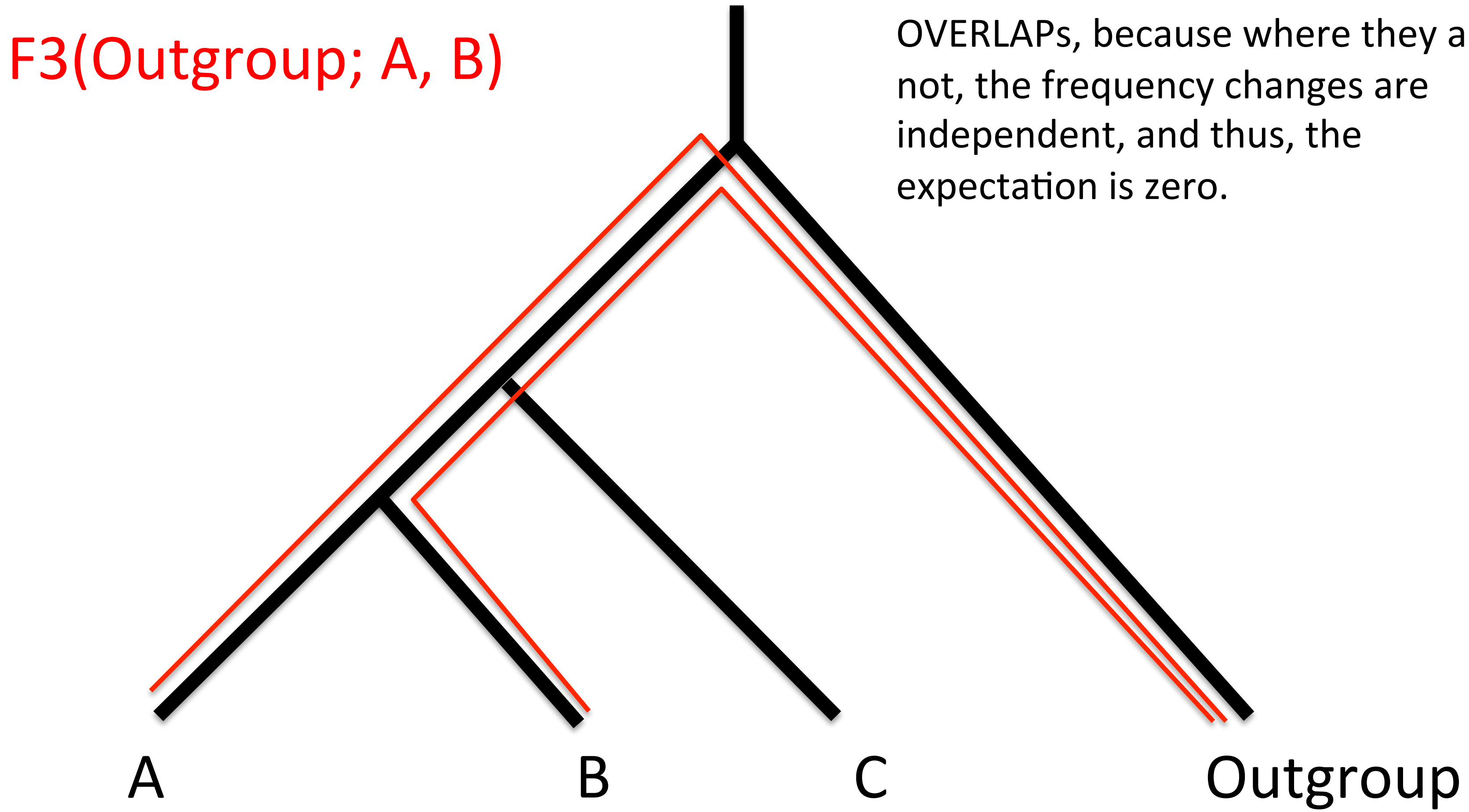
F2(A,B) – Thinking of allele frequency change.
In phylogenetics, the expectation is that f_A=f_B. This is because f_N, while it will randomly shift, under neutrality, it is expected that ON AVERAGE, it will stay the same.



F3(Outgroup; X, Y)



F3(Outgroup; X, Y)



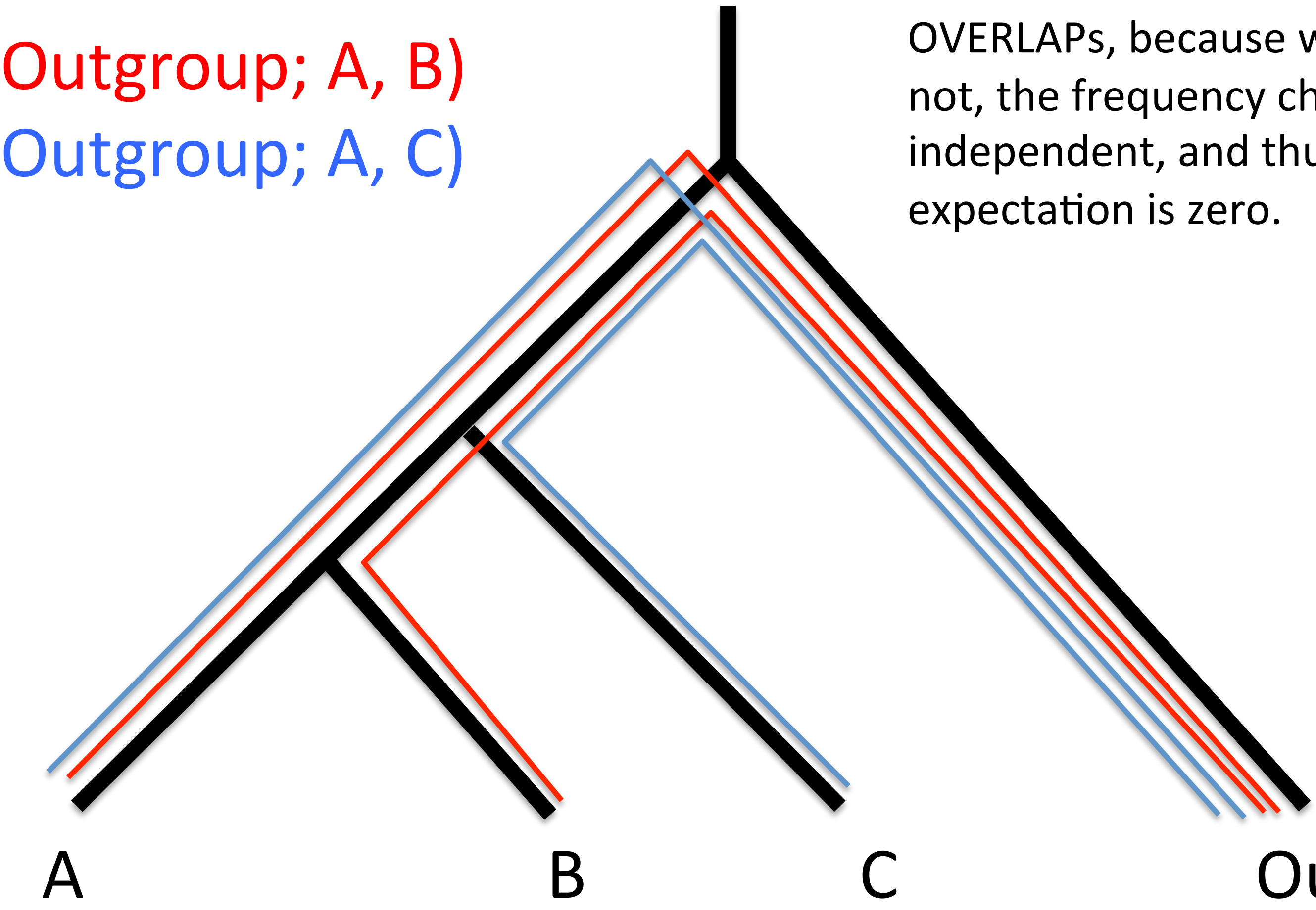
What we are interested in are OVERLAPs, because where they are not, the frequency changes are independent, and thus, the expectation is zero.

F3(Outgroup; X, Y)

F3(Outgroup; A, B)

F3(Outgroup; A, C)

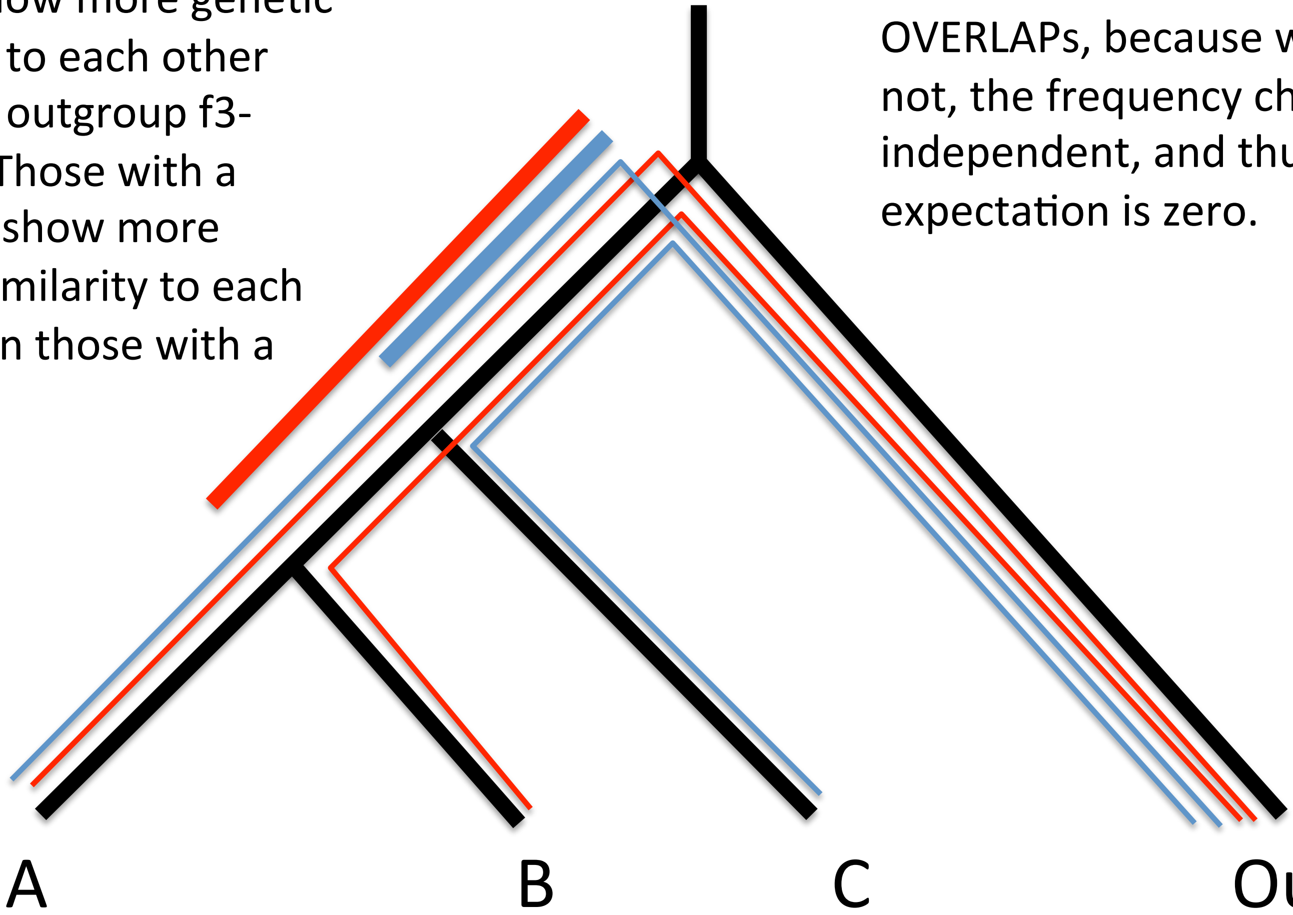
What we are interested in are OVERLAPs, because where they are not, the frequency changes are independent, and thus, the expectation is zero.



F3(Outgroup; X, Y)

Thus, we can test which groups show more genetic similarity to each other using the outgroup f3-statistic. Those with a higher f3 show more genetic similarity to each other than those with a lower f3.

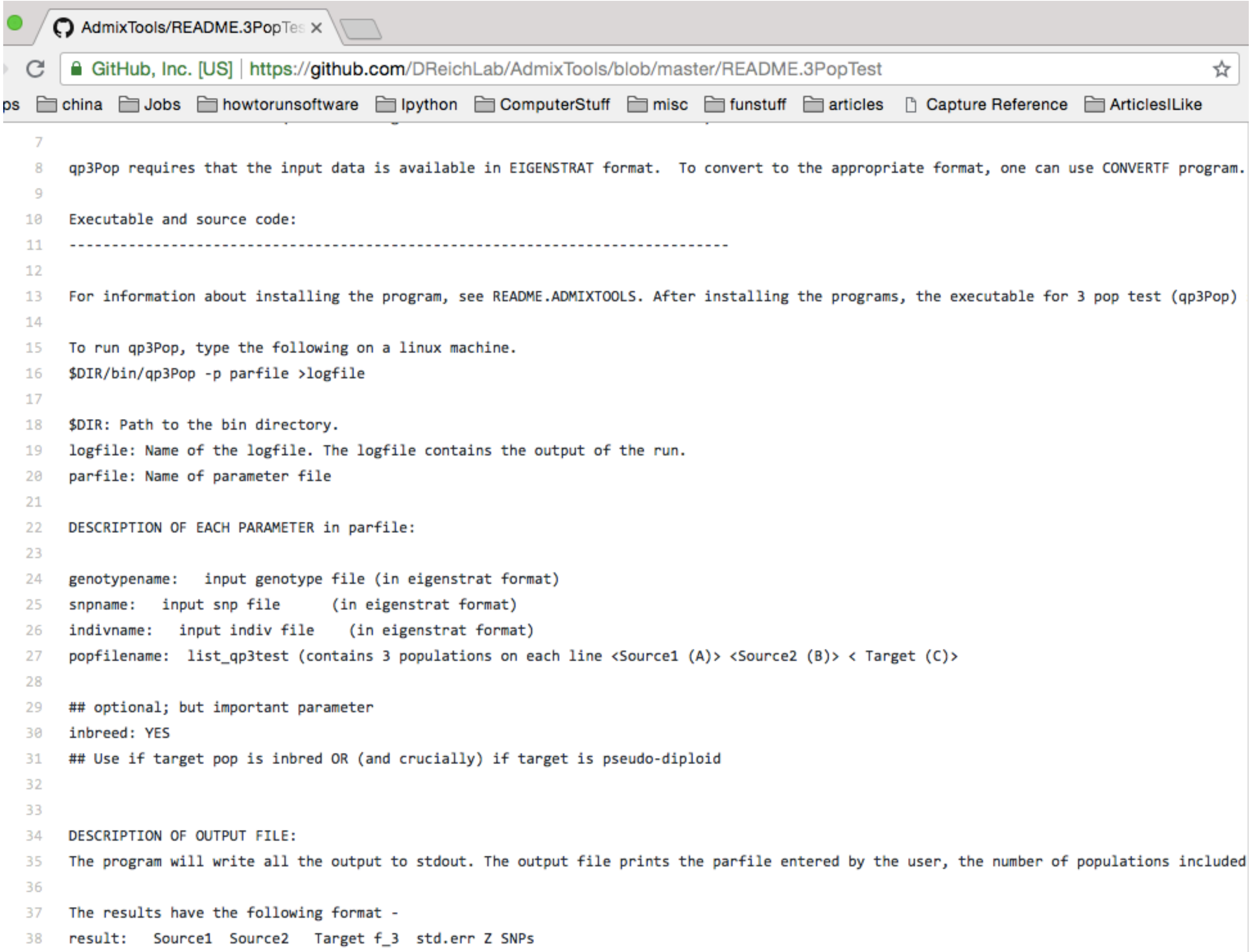
What we are interested in are OVERLAPs, because where they are not, the frequency changes are independent, and thus, the expectation is zero.



F3(Outgroup; A, B)>F3(Outgroup; A, C)

https://github.com/DReichLab/AdmixTools/blob/master/README.3PopTest

qp3Pop


AdmixTools/README.3PopTes ×
GitHub, Inc. [US] | https://github.com/DReichLab/AdmixTools/blob/master/README.3PopTest
ps
china
Jobs
howtorunsoftware
Ipython
ComputerStuff
misc
funstuff
articles
Capture Reference
ArticlesILike

```
qp3Pop requires that the input data is available in EIGENSTRAT format.  To convert to the appropriate format, one can use CONVERTF program.

Executable and source code:
------------------------------------------------------------------------------


For information about installing the program, see README.ADMIXTOOLS. After installing the programs, the executable for 3 pop test (qp3Pop)

To run qp3Pop, type the following on a linux machine.
$DIR/bin/qp3Pop -p parfile >logfile

$DIR: Path to the bin directory.
logfile: Name of the logfile. The logfile contains the output of the run.
parfile: Name of parameter file

DESCRIPTION OF EACH PARAMETER in parfile:

genotypename:   input genotype file (in eigenstrat format)
snpname:   input snp file      (in eigenstrat format)
indivname:   input indiv file    (in eigenstrat format)
popfilename:  list_qp3test (contains 3 populations on each line <Source1 (A)> <Source2 (B)> < Target (C)>

## optional; but important parameter
inbreed: YES
## Use if target pop is inbred OR (and crucially) if target is pseudo-diploid



DESCRIPTION OF OUTPUT FILE:
The program will write all the output to stdout. The output file prints the parfile entered by the user, the number of populations included

The results have the following format -
result:   Source1  Source2   Target f_3  std.err Z SNPs
```

qp3Pop

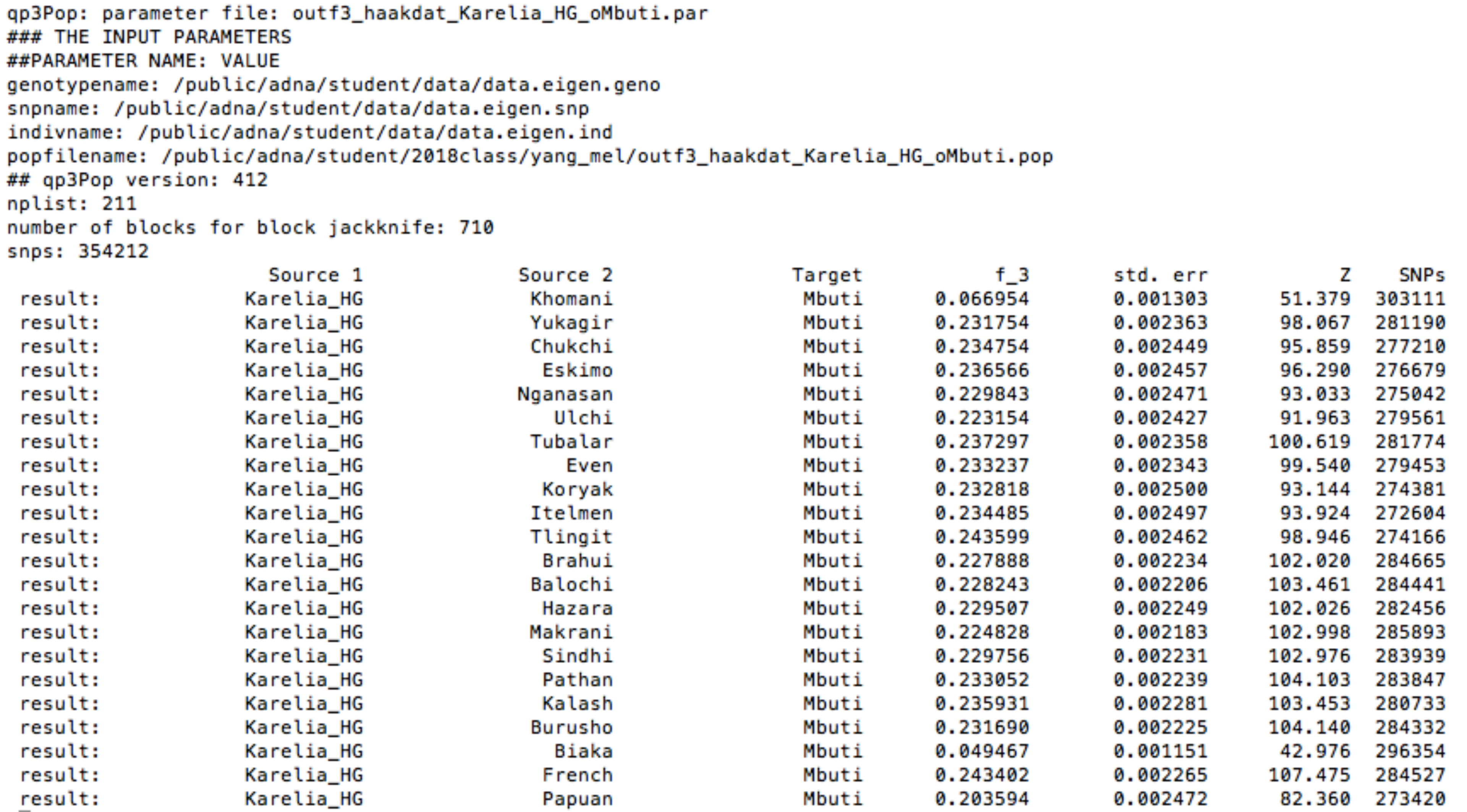
- Make a pop file.
- Make a par file.
- Run qp3Pop –p [parfilename] > [logfilename]

F3(Karelia, X; Mbuti)

- How should we find those populations that share the most genetic similarity to Karelia?

F3(Karelia, X; Mbuti)

- How should we find those populations that share the most genetic similarity to Karelia?



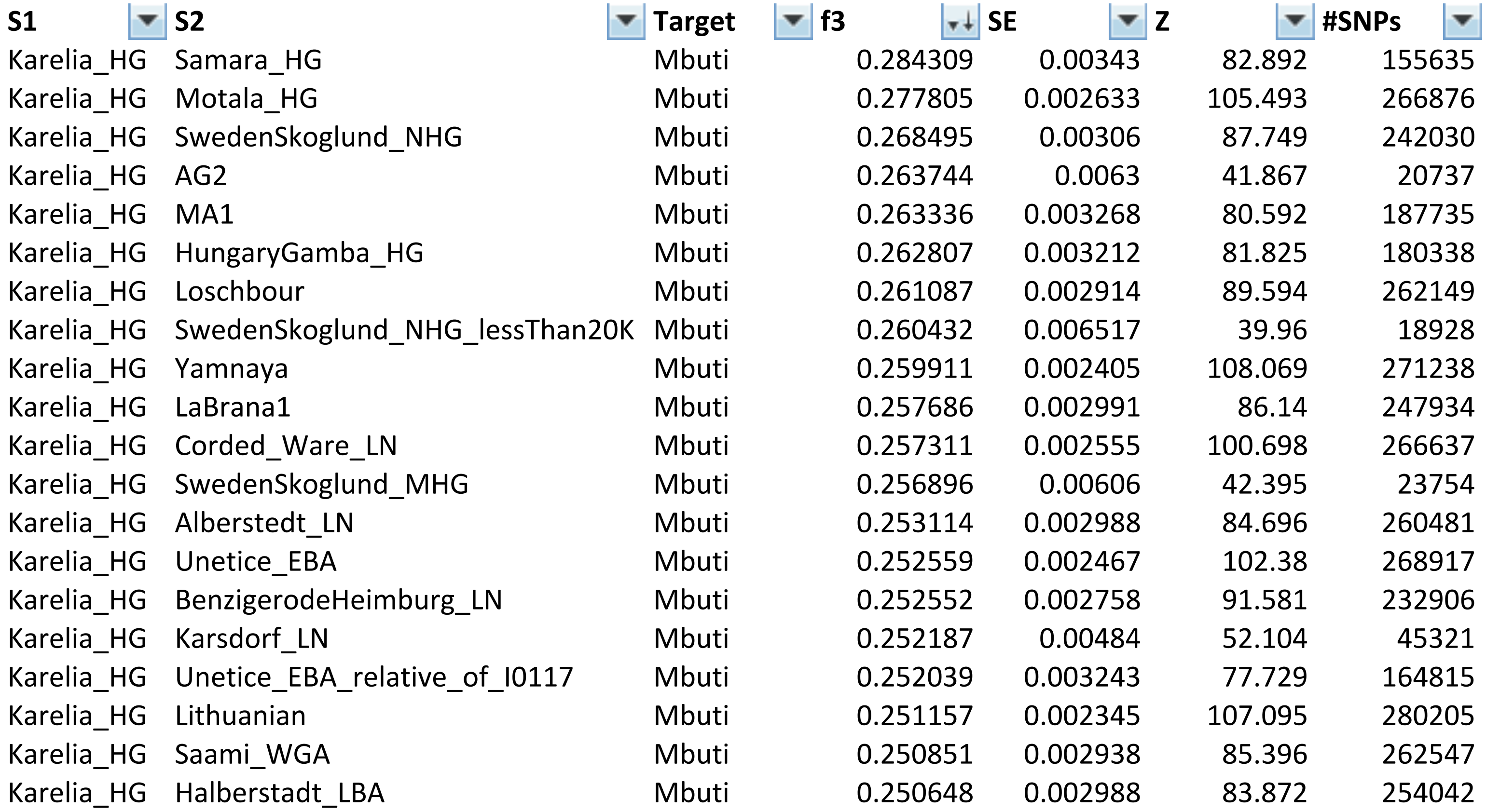
```
qp3Pop: parameter file: outf3_haakdat_Karelia_HG_oMbuti.par
### THE INPUT PARAMETERS
##PARAMETER NAME: VALUE
genotypename: /public/adna/student/data/data.eigen.geno
snpname: /public/adna/student/data/data.eigen.snp
indivname: /public/adna/student/data/data.eigen.ind
popfilename: /public/adna/student/2018class/yang_mel/outf3_haakdat_Karelia_HG_oMbuti.pop
## qp3Pop version: 412
nplist: 211
number of blocks for block jackknife: 710
snps: 354212
                      Source 1            Source 2             Target        f_3        std. err          Z     SNPs
 result:            Karelia_HG             Khomani              Mbuti   0.066954        0.001303     51.379   303111
 result:            Karelia_HG             Yukagir              Mbuti   0.231754        0.002363     98.067   281190
 result:            Karelia_HG             Chukchi              Mbuti   0.234754        0.002449     95.859   277210
 result:            Karelia_HG              Eskimo              Mbuti   0.236566        0.002457     96.290   276679
 result:            Karelia_HG            Nganasan              Mbuti   0.229843        0.002471     93.033   275042
 result:            Karelia_HG               Ulchi              Mbuti   0.223154        0.002427     91.963   279561
 result:            Karelia_HG             Tubalar              Mbuti   0.237297        0.002358    100.619   281774
 result:            Karelia_HG                Even              Mbuti   0.233237        0.002343     99.540   279453
 result:            Karelia_HG              Koryak              Mbuti   0.232818        0.002500     93.144   274381
 result:            Karelia_HG             Itelmen              Mbuti   0.234485        0.002497     93.924   272604
 result:            Karelia_HG             Tlingit              Mbuti   0.243599        0.002462     98.946   274166
 result:            Karelia_HG              Brahui              Mbuti   0.227888        0.002234    102.020   284665
 result:            Karelia_HG             Balochi              Mbuti   0.228243        0.002206    103.461   284441
 result:            Karelia_HG              Hazara              Mbuti   0.229507        0.002249    102.026   282456
 result:            Karelia_HG             Makrani              Mbuti   0.224828        0.002183    102.998   285893
 result:            Karelia_HG              Sindhi              Mbuti   0.229756        0.002231    102.976   283939
 result:            Karelia_HG              Pathan              Mbuti   0.233052        0.002239    104.103   283847
 result:            Karelia_HG              Kalash              Mbuti   0.235931        0.002281    103.453   280733
 result:            Karelia_HG             Burusho              Mbuti   0.231690        0.002225    104.140   284332
 result:            Karelia_HG               Biaka              Mbuti   0.049467        0.001151     42.976   296354
 result:            Karelia_HG              French              Mbuti   0.243402        0.002265    107.475   284527
 result:            Karelia_HG              Papuan              Mbuti   0.203594        0.002472     82.360   273420
```

F3(Karelia, X; Mbuti)

- How should we find those populations that share the most genetic similarity to Karelia?
  - Transfer file to computer
  - Convert to excel
  - Sort
  - Examine

F3(Karelia, X; Mbuti)

- How should we find those populations that share the most genetic similarity to Karelia?



S1	S2	Target	f3	SE	Z	#SNPs
Karelia_HG	Samara_HG	Mbuti	0.284309	0.00343	82.892	155635
Karelia_HG	Motala_HG	Mbuti	0.277805	0.002633	105.493	266876
Karelia_HG	SwedenSkoglund_NHG	Mbuti	0.268495	0.00306	87.749	242030
Karelia_HG	AG2	Mbuti	0.263744	0.0063	41.867	20737
Karelia_HG	MA1	Mbuti	0.263336	0.003268	80.592	187735
Karelia_HG	HungaryGamba_HG	Mbuti	0.262807	0.003212	81.825	180338
Karelia_HG	Loschbour	Mbuti	0.261087	0.002914	89.594	262149
Karelia_HG	SwedenSkoglund_NHG_lessThan20K	Mbuti	0.260432	0.006517	39.96	18928
Karelia_HG	Yamnaya	Mbuti	0.259911	0.002405	108.069	271238
Karelia_HG	LaBrana1	Mbuti	0.257686	0.002991	86.14	247934
Karelia_HG	Corded_Ware_LN	Mbuti	0.257311	0.002555	100.698	266637
Karelia_HG	SwedenSkoglund_MHG	Mbuti	0.256896	0.00606	42.395	23754
Karelia_HG	Alberstedt_LN	Mbuti	0.253114	0.002988	84.696	260481
Karelia_HG	Unetice_EBA	Mbuti	0.252559	0.002467	102.38	268917
Karelia_HG	BenzigerodeHeimburg_LN	Mbuti	0.252552	0.002758	91.581	232906
Karelia_HG	Karsdorf_LN	Mbuti	0.252187	0.00484	52.104	45321
Karelia_HG	Unetice_EBA_relative_of_I0117	Mbuti	0.252039	0.003243	77.729	164815
Karelia_HG	Lithuanian	Mbuti	0.251157	0.002345	107.095	280205
Karelia_HG	Saami_WGA	Mbuti	0.250851	0.002938	85.396	262547
Karelia_HG	Halberstadt_LBA	Mbuti	0.250648	0.002988	83.872	254042

Comparing f3 and D

- D (or f4) is essentially equivalent to comparing two f3-statistics.
  - If F3(Karelia, X1; Mbuti)>F3(Karelia, X2; Mbuti), how do we determine is this difference is meaningful?
  - Use D-statistics!
  - Outgroup f3 is a good survey tool, but D-statistic does the actual hypothesis testing.

Making the pop file

- I’ve provided a python script **make_out f3pop.py**, that takes as arguments the working directory, population to put in S1, and outgroup, and this outputs a POP file.
- For each person, take one of the Haak et al. sets and get the outgroup f3-statistics for that set.
- The run should take about seven minutes.

Ancient sets

- EHG: Karelia, Samara
- SHG: Motala
- EN: **LBK,** LBKT, **Starcevo, Spain_EN**
- MN/CA: **Yamnaya, Esperstedt, Baalberge, Spain_MN**
- LN: **Alberstedt, BenzigerodeHeimburg, Bell Beaker,** Karsdorf, **Corded Ware**
- BA/IA: **Halberstadt, Unetice**

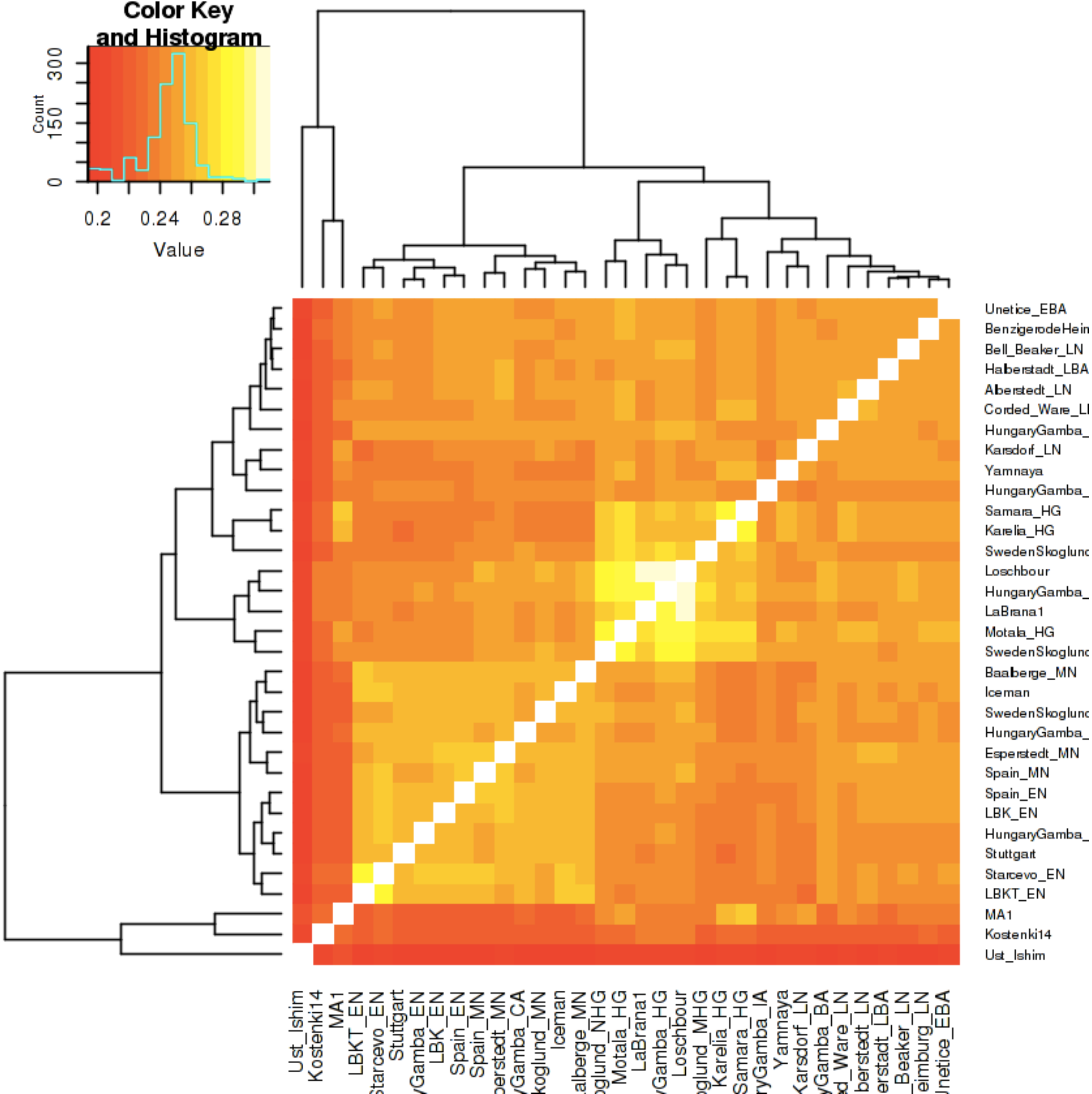
Conclusions from outgroup f3 analysis

Groups – Exercise 7.3

1. HungaryGamba_EN (8)
2. Unetice_EBA (7)
3. BenzigerodeHeimburg_LN (3)
4. Bell_Beaker_LN (6)
5. Corded_Ware_LN (4)
6. Yamnaya (9)
7. Baalberge_MN (3)
8. LBK_EN (12)
9. Motala_HG (7)
10. Spain_MN (4)
11. Spain_EN (4)
12. Sweden_Skoglund_NHG (3)

Summarizing all the information

- How does the Haak et al. (2015) paper summarize the outgroup f3 results?

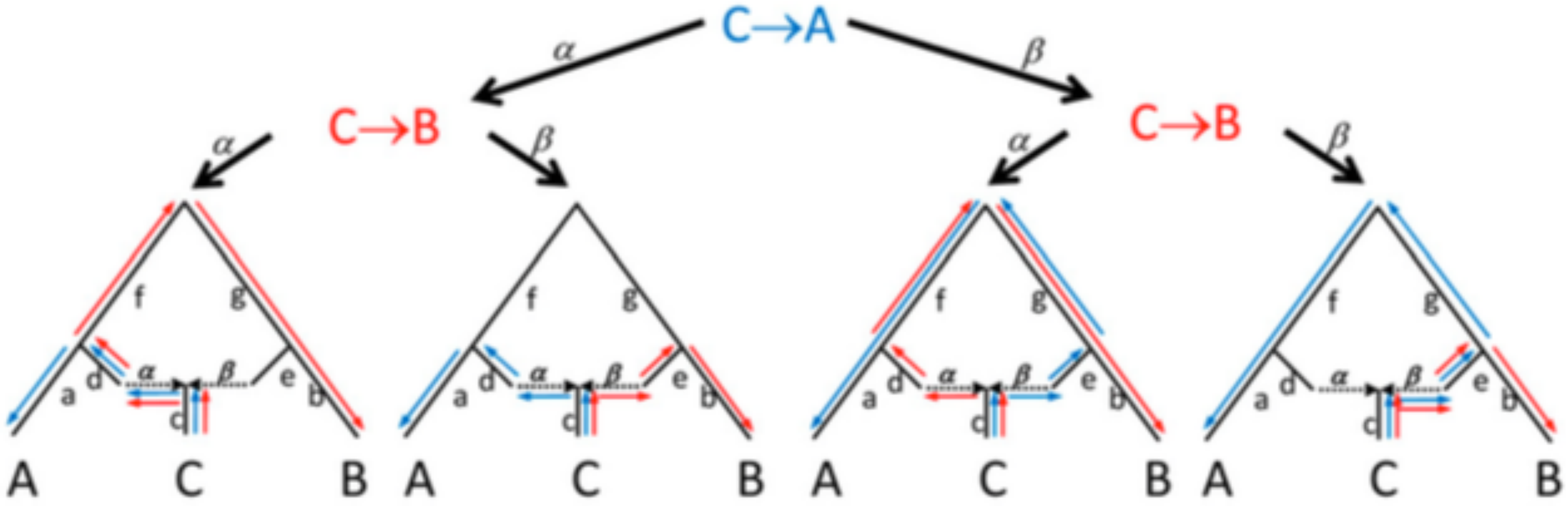

Color Key
and Histogram
Count
0
150
300
0.2
0.24
0.28
Value
Unetice_EBA
BenzigerodeHein
Bell_Beaker_LN
Halberstadt_LBA
Alberstedt_LN
Corded_Ware_LI
HungaryGamba_
Karsdorf_LN
Yamnaya
HungaryGamba_
Samara_HG
Karelia_HG
SwedenSkoglunc
Loschbour
HungaryGamba_
LaBrana1
Motala_HG
SwedenSkoglunc
Baalberge_MN
Iceman
SwedenSkoglunc
HungaryGamba_
Esperstedt_MN
Spain_MN
Spain_EN
LBK_EN
HungaryGamba_
Stuttgart
Starcevo_EN
LBKT_EN
MA1
Kostenki14
Ust_Ishim
Ust_Ishim
Kostenki14
MA1
LBKT_EN
tarcevo_EN
Stuttgart
yGamba_EN
LBK_EN
Spain_EN
Spain_MN
perstedt_MN
yGamba_CA
koglund_MN
Iceman
alberge_MN
oglund_NHG
Motala_HG
LaBrana1
yGamba_HG
Loschbour
oglund_MHG
Karelia_HG
Samara_HG
ryGamba_IA
Yamnaya
Karsdorf_LN
yGamba_BA
d_Ware_LN
berstedt_LN
erstadt_LBA
Beaker_LN
eimburg_LN
Jnetice_EBA

F3-statistics

- Outgroup f3-statistics were not original purpose of an f3-statistic – testing for admixture was.
- The admixture test explored admixture into the the test population, which holds if F3(S1, S2; Test)<0.
  - No information on whether there was admixture if positive.
    - High population specific drift in Test can mask signal (p. 1068)
  - Negative result MUST mean Test is admixed, though whether the best sources are S1 and S2 is unknown.

Negative f3-statistics – the admixture case, some theory

F3(A, B; C) - We need to consider all paths for admixture.



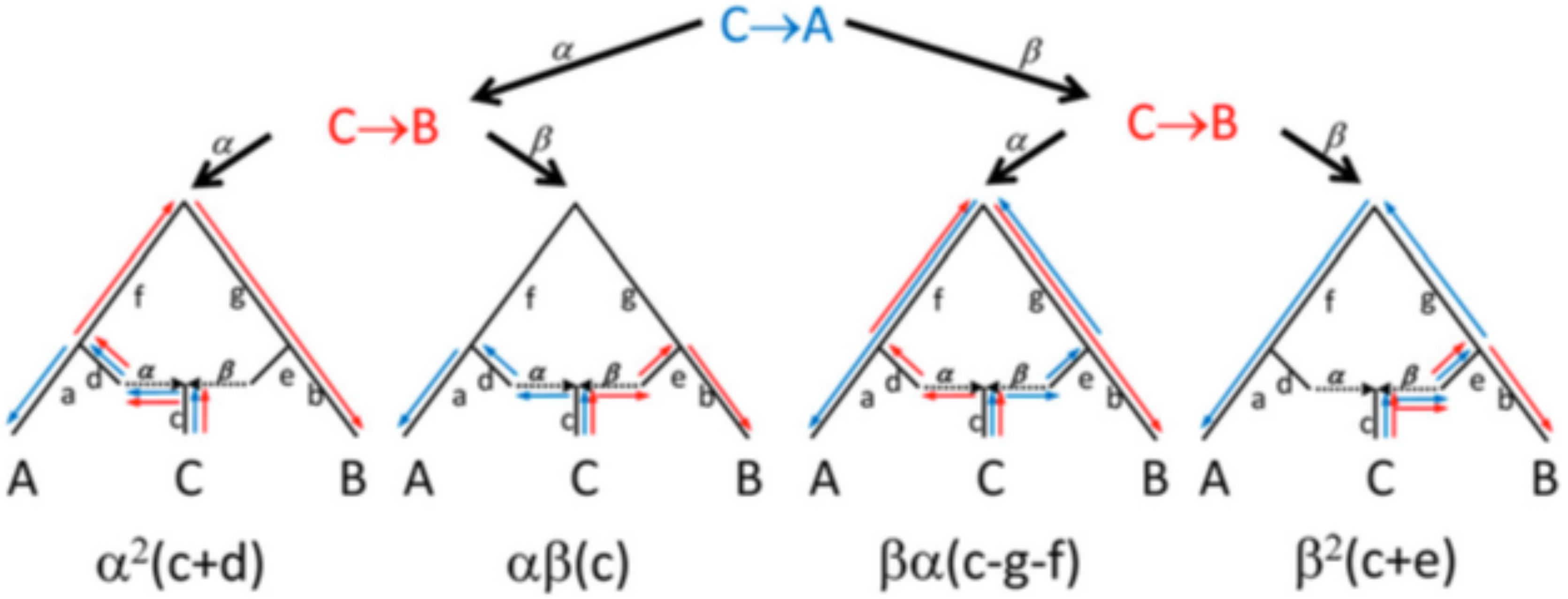
$\alpha^2(c+d)$ $\alpha\beta(c)$ $\beta\alpha(c\text{-}g\text{-}f)$ $\beta^2(c+e)$

$$F_3(C;A,B) = c + \alpha^2 d + \beta^2 e - \alpha\beta(g+f)$$

Negative f3-statistics – the admixture case, some theory

F3(A, B; C) - We need to consider all paths for admixture.

1. Only overlap matters!

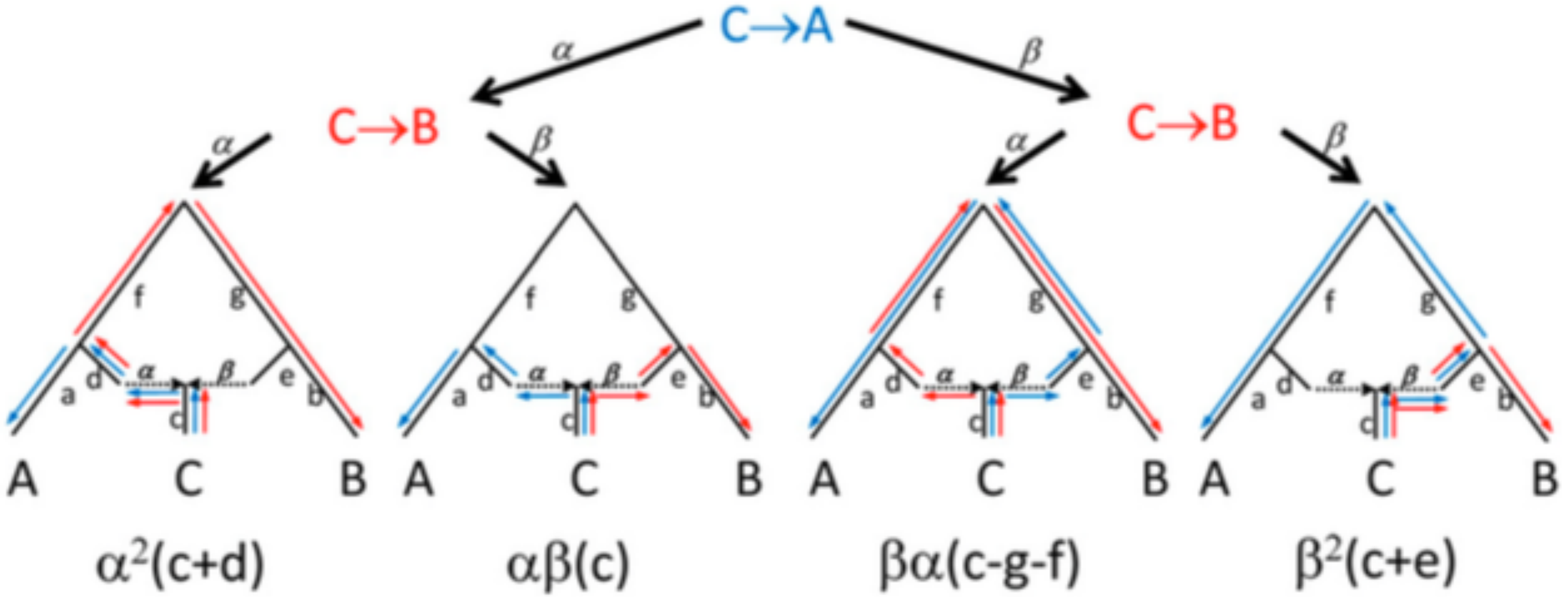


$$F_3(C;A,B) = c + \alpha^2 d + \beta^2 e - \alpha\beta(g+f)$$

2. Some algebra…

Negative f3-statistics – the admixture case, some theory

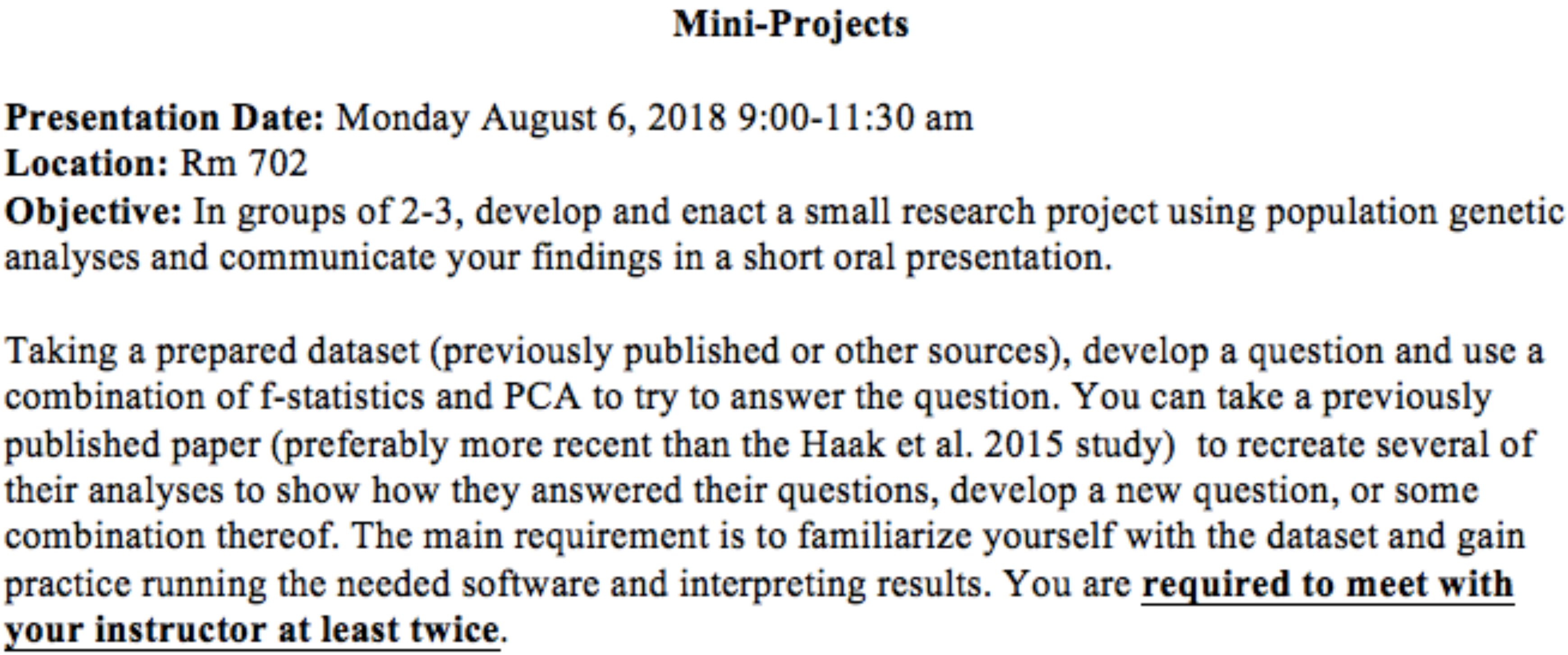
F3(A, B; C) - We need to consider all paths for admixture.



$$F_3(C;A,B) = c + \alpha^2 d + \beta^2 e - \boxed{\alpha\beta(g+f)}$$

This term decides if F3 is negative!
If α or β is 0, then F3>0, while otherwise, F3 MIGHT be negative.

Mini-Project

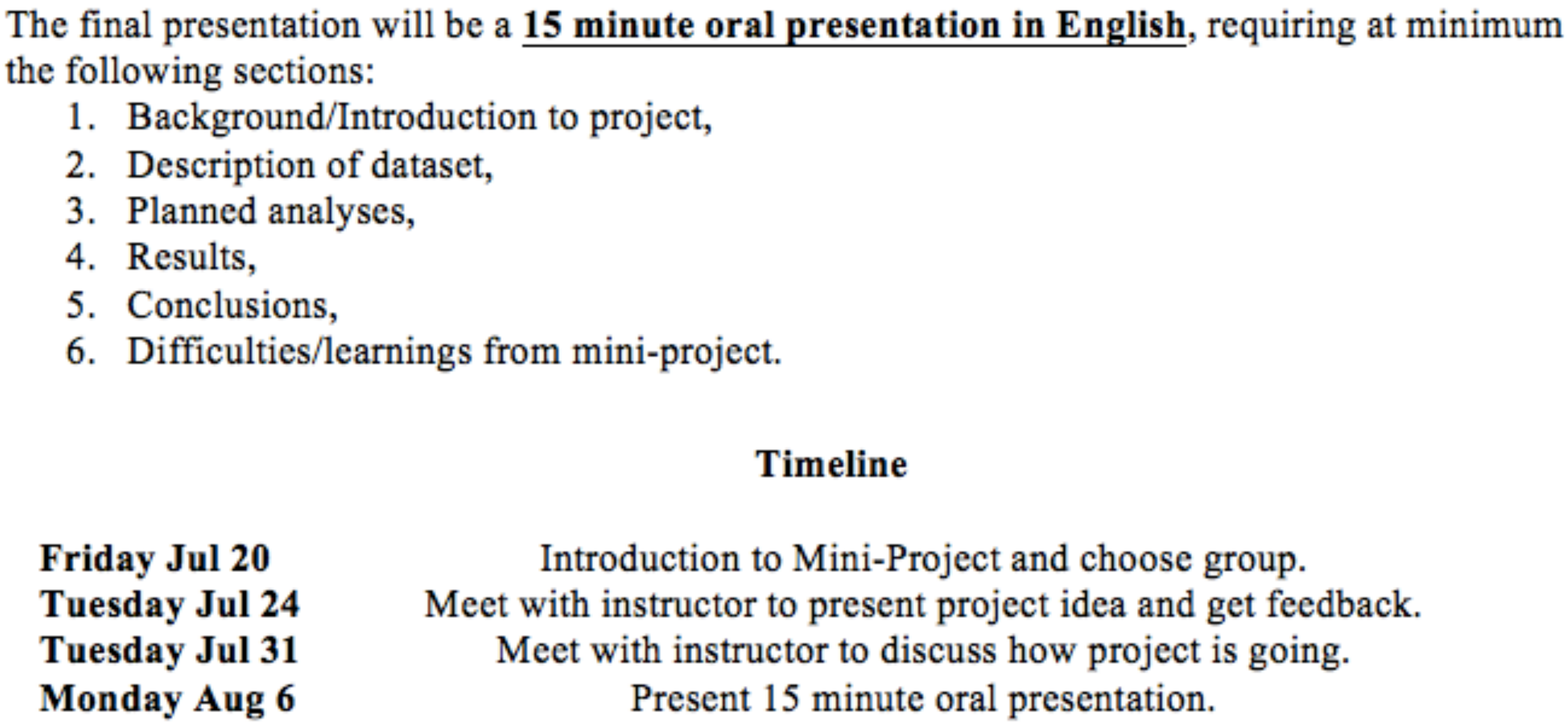


Mini-Projects

Presentation Date: Monday August 6, 2018 9:00-11:30 am
Location: Rm 702
Objective: In groups of 2-3, develop and enact a small research project using population genetic analyses and communicate your findings in a short oral presentation.

Taking a prepared dataset (previously published or other sources), develop a question and use a combination of f-statistics and PCA to try to answer the question. You can take a previously published paper (preferably more recent than the Haak et al. 2015 study) to recreate several of their analyses to show how they answered their questions, develop a new question, or some combination thereof. The main requirement is to familiarize yourself with the dataset and gain practice running the needed software and interpreting results. You are **required to meet with your instructor at least twice**.

Mini-Project



The final presentation will be a **15 minute oral presentation in English**, requiring at minimum the following sections:

1. Background/Introduction to project,
2. Description of dataset,
3. Planned analyses,
4. Results,
5. Conclusions,
6. Difficulties/learnings from mini-project.

Timeline

Friday Jul 20	Introduction to Mini-Project and choose group.
Tuesday Jul 24	Meet with instructor to present project idea and get feedback.
Tuesday Jul 31	Meet with instructor to discuss how project is going.
Monday Aug 6	Present 15 minute oral presentation.

Mini-Project

- 2:00 pm – Group
- 2:30 pm – Group
- 3:00 pm – Group
- 3:30 pm – Group
- 4:00 pm – Group

Mini-Project

- Before Tuesday, explore some more recent papers, and see if there’s a region or dataset you’re interested in learning.
- Highly recommend you use a dataset similar to the Haak dataset, as that way you don’t have to worry about file formats!
  - https://reich.hms.harvard.edu/datasets
- Come with ideas, so I can help you develop them further!

Questions?