Processing of genetic data in Zaidi & White, et. al. 2018

Julie D. White

September 19, 2018

Contents

1 Preliminary data filtering

1.1 Extract individuals with 3D image data and genetic data and filter SNPs based on missingness

The Shriver lab has several thousands of 3D facial images, but we were only interested in those that also had genetic data. So, we matched the two ID lists to come up with a list of individuals that had both genetic data and 3D facial images for analysis: "ADAPT_InSymShape_WithGenotypes.txt" then used plink to create a new file of this intersection. In the meantime, we also wanted to filter out SNPs that were missing in more than 10% of our individuals.

```
plink --bfile ADAPT_2784ppl_567K_hg19 --keep ADAPT_InSymShape_WithGenotypes.txt --geno
    0.1 --make-bed --out ADAPT_2721ppl_SymShape_geno0.1
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2784ppl_567K_hg19
–geno 0.1
–keep ADAPT_InSymShape_WithGenotypes.txt
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\2_RemovePeopleNotInSymShape`
Start time: Thu Mar 22 10:03:05 2018

Random number seed: 1521727385
16262 MB RAM detected; reserving 8131 MB for main workspace.
1060536 variants loaded from .bim file.
2784 people (1030 males, 1754 females) loaded from .fam.
–keep: 2721 people remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 2641 founders and 80 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.545888.
492749 variants removed due to missing genotype data (–geno).
567787 variants and 2721 people pass filters and QC.
Note: No phenotypes present.
–make-bed to ADAPT_2721ppl_SymShape_geno0.1.bed +
ADAPT_2721ppl_SymShape_geno0.1.bim + ADAPT_2721ppl_SymShape_geno0.1.fam ...
done.

End time: Thu Mar 22 10:03:15 2018

1.2 Remove individuals with high missingness rates

Next, we removed individuals that were missing more than 10% of the typed SNPs in our sample.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1 --mind 0.1 --make-bed --out
    ADAPT_2721ppl_SymShape_geno0.1_mind0.1
```

```
PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1
–make-bed
–mind 0.1
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\2_
RemovePeopleNotInSymShape
Start time: Thu Mar 22 10:04:08 2018

Random number seed: 1521727448
16262 MB RAM detected; reserving 8131 MB for main workspace.
567787 variants loaded from .bim file.
2721 people (1005 males, 1716 females) loaded from .fam.
5 people removed due to missing genotype data (–mind).
IDs written to ADAPT_2721ppl_SymShape_geno0.1_mind0.1.irem .
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 2638 founders and 78 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.99521.
567787 variants and 2716 people pass filters and QC.
Note: No phenotypes present.
–make-bed to ADAPT_2721ppl_SymShape_geno0.1_mind0.1.bed +
ADAPT_2721ppl_SymShape_geno0.1_mind0.1.bim +
ADAPT_2721ppl_SymShape_geno0.1_mind0.1.fam ... done.

End time: Thu Mar 22 10:04:15 2018
```

1.3 Extract individuals 18 - 30 years old, with known height, weight, and gender

For our analysis, we also wanted to remove individuals with missing height and weight information, and those who were not between 18 and 30 years of age. We have age as a covariate in our analyses, but we wanted to minimize age variation regardless. Lastly, we removed individuals with a reported gender that was different from their chromosomal sex. We filtered our covariate files to create a list of people meeting these criteria and placed this in a text file: "ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_Keep.txt"

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1 --keep ADAPT_2721ppl_SymShape_geno0
    .1_mind0.1_AgeMissingData_Keep.txt --make-bed --out ADAPT_2721ppl_SymShape_geno0.1
    _mind0.1_AgeMissingData
```

```
PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
–keep ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_Keep.txt
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\3_Geno_
Mind_Relatives
```

```
Start time: Thu Mar 22 12:51:46 2018

Random number seed: 1521738848
16262 MB RAM detected; reserving 8131 MB for main workspace.
567787 variants loaded from .bim file.
2716 people (1003 males, 1713 females) loaded from .fam.
–keep: 1950 people remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1896 founders and 54 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.99509.
567787 variants and 1950 people pass filters and QC.
Note: No phenotypes present.
–make-bed to ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData.bed +
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData.bim +
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData.fam ... done.

End time: Thu Mar 22 12:51:56 2018
```

1.4 Relatedness

To avoid inflation of statistics, we removed all related individuals prior to analysis. A relatedness matrix was calculated using plink and 0.8 was chosen as the threshold for relatedness

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData --distance square0
    ibs --out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_Rel
```

```
PLINK v1.90b5.2 64-bit (9 Jan 2018)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData
–distance square0 ibs
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_Rel

Hostname: aci-lgn-001.acib.production.int.aci.ics.psu.edu
Working directory: /gpfs/scratch/jdw345/ADAPT
Start time: Thu Mar 22 13:14:08 2018

Random number seed: 1521737506
129063 MB RAM detected; reserving 64531 MB for main workspace.
567787 variants loaded from .bim file.
1950 people (731 males, 1219 females) loaded from .fam.
Using up to 9 threads (change this with –threads).
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.99509.
567787 variants and 1950 people pass filters and QC.
Note: No phenotypes present.
Excluding 18597 variants on non-autosomes from distance matrix calc.
Distance matrix calculation complete.
IDs written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_Rel.mibs.id .
IBS matrix written to
```

ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_Rel.mibs .
done.

End time: Thu Mar 22 13:14:14 2018

For each pair or group of individuals with relatedness above 0.8, a single individual was selected and placed in the list "KeepAfterRelativeRemoval.txt", along with all the individuals without any relative in the sample.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData --keep
    KeepAfterRelativeRemoval.txt --make-bed --out ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData
–keep KeepAfterRelativeRemoval.txt
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\3_Geno_Mind_Relatives`
Start time: Thu Mar 22 13:18:47 2018

Random number seed: 1521738887
16262 MB RAM detected; reserving 8131 MB for main workspace.
567787 variants loaded from .bim file.
1950 people (731 males, 1219 females) loaded from .fam.
–keep: 1921 people remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.995096.
567787 variants and 1921 people pass filters and QC.
Note: No phenotypes present.
–make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved.bed +
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved.bim +
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved.fam ...
done.

End time: Thu Mar 22 13:18:51 2018

2 HLA SNP heterozygosity

2.1 Extract HLA SNPs and calculate HLA heterozygosity in study sample

Using de Bakker, et. al. 2006, we created a list of SNPs that tag the MHC region. This list was used to create the "HLA_tagsnps.txt" file. We then filtered our genotype files and calculated heterozygosity for only those SNPs. We will later filter the resulting .het files so that we extract the information pertinent to our European sample.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved --
    extract HLA_tagsnps.txt --het --make-bed --out ADAPT_2721ppl_SymShape_geno0.1_mind0
    .1_AgeMissingData_RelativesRemoved_HLAHet
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved
–extract HLA_tagsnps.txt
–het
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_HLAHet

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\4_HLA_Heterozygosity`
Start time: Thu Mar 22 13:19:54 2018

Random number seed: 1521739194
16262 MB RAM detected; reserving 8131 MB for main workspace.
567787 variants loaded from .bim file.
1921 people (725 males, 1196 females) loaded from .fam.
–extract: 114 variants remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.993379.
114 variants and 1921 people pass filters and QC.
Note: No phenotypes present.
–make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_HLAHet.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_HLAHet.bim
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_HLAHet.fam
... done.
–het: 114 variants scanned, report written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_HLAHet.het

End time: Thu Mar 22 13:19:54 2018

2.2 Nucleotide diversity in the HLA region in our sample

The following exercise was to determine and illustrate nucleotide diversity in the HLA region in our sample, compared to the genomic background. Using the 1233 Europeans (identified in section 5 of this document),

we calculated nucleotide diversity within the HLA region in a genetic dataset that had not been pruned for linkage disequilibrium or palindromes.

```
#Create a vcf file of the HLA region using the 1233 Europeans in our sample.
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved --
    keep FinalSampleAfterEigenOutlierRemoval.txt --chr 6 --from-kb 28000 --to-kb 34000
    --recode vcf --out ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion
```

```
PLINK v1.90b5.2 64-bit (9 Jan 2018) www.cog-genomics.org/plink/1.9/
(C) 2005-2018 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion.log.
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved
–chr 6
–from-kb 28000
–keep FinalSampleAfterEigenOutlierRemoval.txt
–out ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion
–recode vcf
–to-kb 34000

32109 MB RAM detected; reserving 16054 MB for main workspace.
4809 out of 567787 variants loaded from .bim file.
1921 people (725 males, 1196 females) loaded from .fam.
–keep: 1233 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1193 founders and 40 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.990059.
4809 variants and 1233 people pass filters and QC.
Note: No phenotypes present.
–recode vcf to ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion.vcf ... done.
```

```
vcftools --vcf ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion.vcf --window-pi 50000 --
    window-pi-step 5000 --out ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion
```

```
VCFtools - 0.1.15
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:
–vcf ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion.vcf
–out ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion
–window-pi 50000
–window-pi-step 5000

After filtering, kept 1233 out of 1233 Individuals
Outputting Windowed Nucleotide Diversity Statistics...
After filtering, kept 4809 out of a possible 4809 Sites
Run Time = 1.00 seconds
```

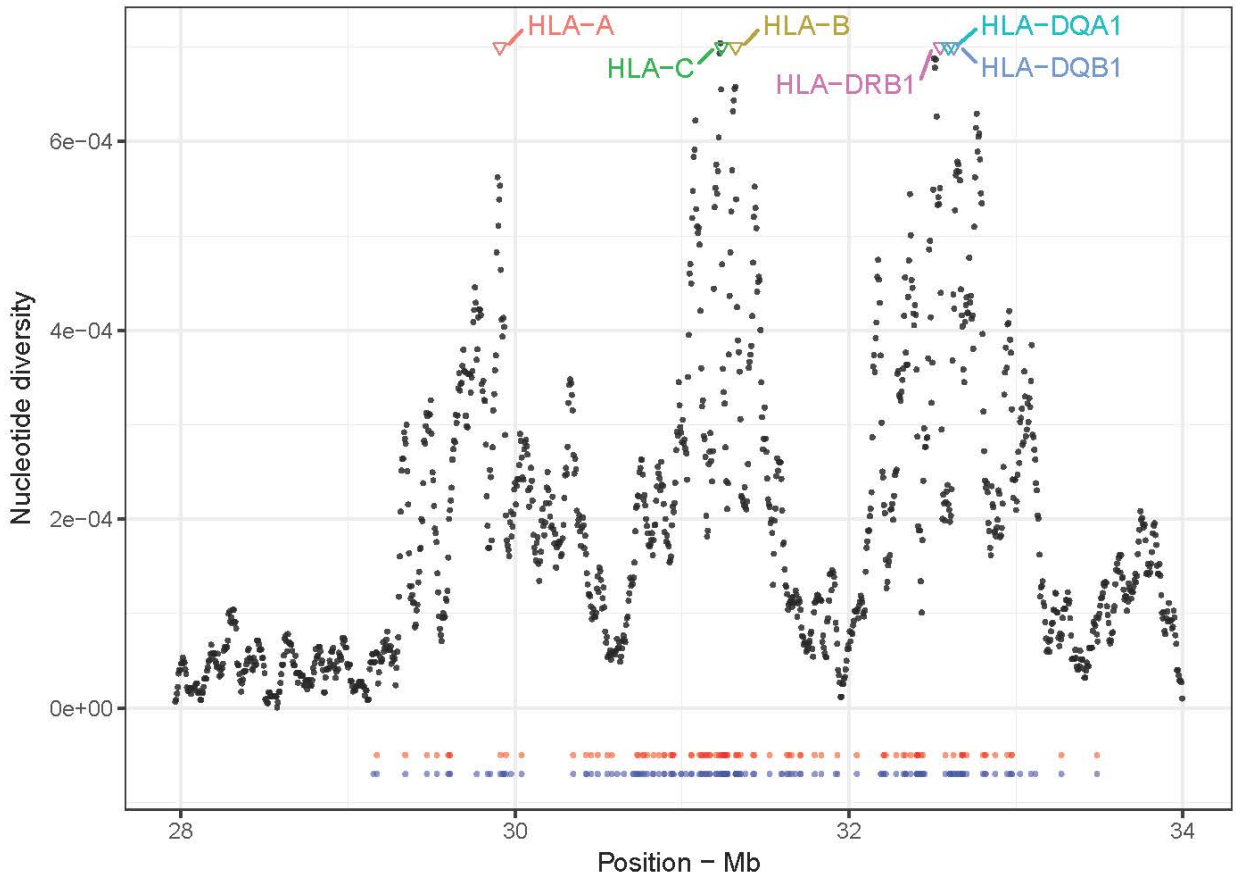


Figure S7: Nucleotide diversity in our sample at the MHC locus showing high diversity at the HLA genes compared to the genomic background. Pi was calculated in a sliding window of size 50kb with a step size of 5kb with vcftools. Location of HLA class-I and class-II genes is indicated on top. Blue points at the bottom indicate position of 195 SNPs tagging HLA alleles in Europeans while red points indicate position of the subset of 114 SNPs genotyped in our sample.

The following R code was used to create the above plot.

```
# Showing diversity in ADAPT samples at the MHC locus
setwd("C:/Users/jwhite/Box/MHC_paper/1_FinalDataCuration/10_NucleotideDiversity")

library(ggplot2)
library(ggrepel)

#Read in pi file created by vcftools
mhc.het<-read.table('ADAPT_AfterEigenOutlierRemoval_MHC_NarrowRegion.windowed.pi',header
    =T,sep="\t")

#Make the initial plot
piplot<-ggplot(mhc.het,aes(BIN_START/1000000,PI))+geom_point(alpha=0.7,size=0.5)+theme_
    bw()+labs(x="Position_-_Mb",y="Nucleotide_diversity")

#Create a dataframe with the HLA genes and their coordinates
hla.coordinates<-data.frame(hgnc=c("HLA-A","HLA-B","HLA-C","HLA-DRB1","HLA-DQA1","HLA-
    DQB1"),hg19start=c(29909037,31321649,31236526,32546546,32595956,32627244),hg19end=c
    (29913661,31324989,31239907,32557625,32614839,32636160))

#Make a new plot by layering the HLA genes and their coordinates with the HLA gene plot
hla.plot<-piplot+geom_text_repel(data=hla.coordinates,aes(y=0.0007,x=hg19start/1000000,
    label=hgnc,color=hgnc),min.segment.length = 0.3,point.padding = 0.5)+geom_point(data
    =hla.coordinates,aes(y=0.0007,x=hg19start/1000000,color=hgnc),shape=25)+theme(legend
```

```
      .position="none")

#load snp positions - 195 - full list of SNPs from paper
snp.full<-read.table('HLA_tagsnps195.txt',header=F)
colnames(snp.full)<-c("chr","start","stop","rsid")

#load snp positions - 97 subset overlapping with SNPs genotyped in ADAPT
snp.bim<-read.table('HLA_tagsnps.bim',header=F)
colnames(snp.bim)<-c("chr","rsid","cm","position","a1","a2")

#Add another layer of the full set of SNPs and the SNPs genotyped in our sample.
snp.plot<-hla.plot+geom_point(data=snp.bim,aes(y=-5e-05,x=position/1000000),color="red",
      size=0.5,alpha=0.5)+geom_point(data=snp.full,aes(y=-7e-05,x=start/1000000),alpha
      =0.5,size=0.5,color="blue")

#Save plot
ggsave("HLA_genomic_region_SNPs_pi.pdf",snp.plot,height=5,width=7)
```

2.3 Compare HLA heterozygosity in 1000G Europeans

We additionally used the European 1000G data to compare estimates of heterogyzosity calculated using only the SNPs for which our data were genotyped vs. that calculated from the full complement of SNPs identified in de Bakker, et al. 2006.
To do this, we calculated heterozygosity in the European 1000G data using the full and reduced set of SNPs

```
#For full set of SNPs
vcftools --gzvcf ALL.chr6.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.
      gz --keep HLA/1kg_Europeans.txt --snps HLA/HLA_tagsnps --het --out 1kg_hla_tag_all
```

VCFtools - v0.1.12
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:
–gzvcf ALL.chr6.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
–keep HLA/1kg_Europeans.txt
–het
–out 1kg_hla_tag_all
–snps HLA/HLA_tagsnps

Using zlib version: 1.2.8
Keeping individuals in 'keep' list
After filtering, kept 503 out of 2504 Individuals
Outputting Individual Heterozygosity
After filtering, kept 154 out of a possible 5024119 Sites
Run Time = 200.00 seconds

```
#For reduced set of SNPs
vcftools --gzvcf ../ALL.chr6.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.
      vcf.gz --keep 1kg_Europeans.txt --snps HLA_tagsnps_114.txt --het --out 1
      kg_hla_tag_114.sub.het
```

VCFtools - v0.1.12
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:

–gzvcf ../ALL.chr6.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
–keep 1kg_Europeans.txt
–het
–out 1kg_hla_tag_114.sub.het
–snps HLA_tagsnps_114.txt

Using zlib version: 1.2.8
Keeping individuals in 'keep' list
After filtering, kept 503 out of 2504 Individuals
Outputting Individual Heterozygosity
After filtering, kept 112 out of a possible 5024119 Sites
Run Time = 205.00 seconds

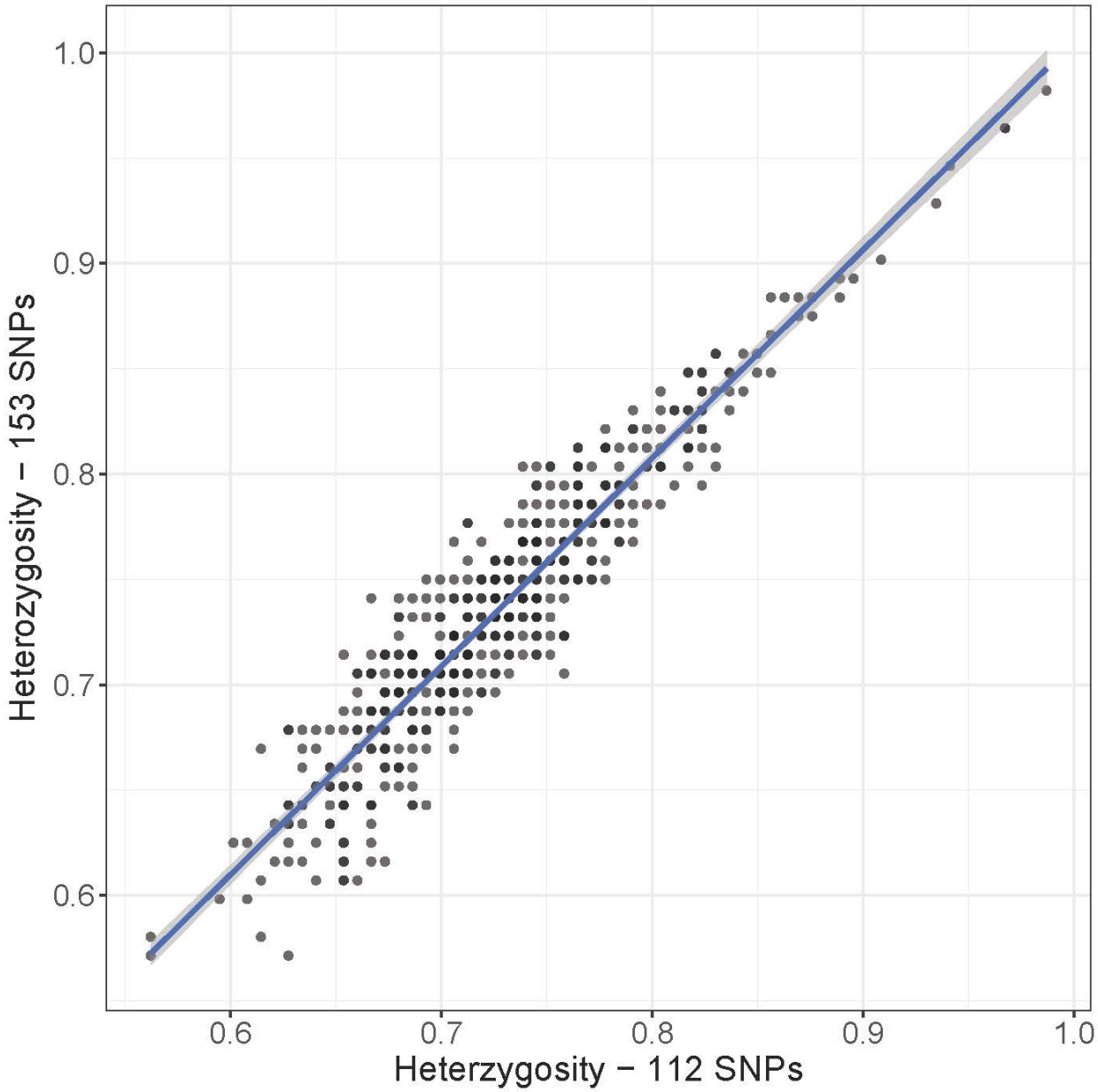


Figure S8: Comparison of Heterozygosity calculated using 154 SNPs tagging HLA haplotype variation in Europeans, the subset of SNPs reported in (de Bakker, et. al. 2006) for which there was 1000G information, and a reduced subset of 112 out of 114 SNPs genotyped in our sample. Heterozygosity is calculated in 503 individuals of European ancestry from the 1000 Genomes Project Data (1000 Genomes Project Consortium, et al. 2015).

The following R code was used to create the above plot.

```
#comparing heterozygosity in 503 Europeans from the 1000G Project data across 153 and
    112 genes

setwd("~/Box Sync/MHC_paper/Julie_1000G_Analysis/MHC_tagsnps/Analysis_04292018/
    Comparison_of_allvreduced_hla_tagsnps_on_heterozygosity")
library(ggplot2)
library(reshape2)

#read heterozygosity data
het.dat<-read.table("hla_154_112_combined",header=T,sep="\t")

#calculate observed_homozygosity
het.dat$Ohom_p<-with(het.dat,Ohom/N_sites)

#dcast to create two columns
dhet.dat<-dcast(het.dat,INDV~Set,value.var='Ohom_p')

with(dhet.dat,cor(All_154,All_112))

comp.plt<-ggplot(dhet.dat,aes(All_154,All_112))+
  geom_point(alpha=0.5)+
  stat_smooth(method="lm")+
  labs(x="Heterzygosity - 112 SNPs",y="Heterozygosity - 153 SNPs")+
  theme_bw()+
  theme(axis.text=element_text(size=14),axis.title=element_text(size=16))

ggsave("1KG_het_112_154.pdf",comp.plt,width=7,height=7)
```

3 Merge with 1000G for ancestry assessment

3.1 SNP intersection with 1000G

We completed this step by first finding the intersection between 1000 Genomes Phase 3 and our sample, then by merging the two files at that intersection.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved --
    write-snplist --out ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved
```

```
PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved
–write-snplist

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_
MergeWith1000G
Start time: Thu Mar 22 14:20:10 2018

Random number seed: 1521742810
16262 MB RAM detected; reserving 8131 MB for main workspace.
567787 variants loaded from .bim file.
1921 people (725 males, 1196 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.995096.
567787 variants and 1921 people pass filters and QC.
Note: No phenotypes present.
List of variant IDs written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved.snplist

End time: Thu Mar 22 14:20:11 2018
```

Create the SNP lists for 1000 Genomes

```
for i in {1..22}; do
plink --bfile ALL.chr$[i].phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes --
    write-snplist --out ALL.chr$[i].phase3_shapeit2_mvncall_integrated_v5a.20130502.
    genotypes
done

plink --bfile ALL.chrX.phase3_shapeit2_mvncall_integrated_v1b.20130502.genotypes --write
    -snplist --out ALL.chrX.phase3_shapeit2_mvncall_integrated_v1b.20130502.genotypes

#Not going to give the plink output files since there is one for each chromosome.

#Concatenate these files into a single list
for i in {1..22}; do
cat ALL.chr$[i].phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.snplist >>
    1000G_Phase3_Snplist;
done

cat ALL.chrX.phase3_shapeit2_mvncall_integrated_v1b.20130502.genotypes >> 1000
    G_Phase3_Snplist
```

Use grep to find the intersection between the two snplists

```
grep -Fx -f ADAPT_2721ppl_SymShape_geno0.1_mnid0.1_AgeMissingData_RelativesRemoved.snplist 1000G_Phase3_Snplist > ADAPT_1000G_Intersection.txt
```

3.2 Filter genotype files to keep only intersecting SNPs

This was used to filter both the study samples and 1000 Genomes files to keep only the intersection
For 1000 Genomes:

```
for i in {1..22}; do
plink --bfile ALL.chr$[i].phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes --extract ADAPT_1000G_Intersection.txt --make-bed --out 1000G_ADAPT_Intersection_chr$[i]
done

plink --bfile ALL.chrX.phase3_shapeit2_mvncall_integrated_v1b.20130502.genotypes --extract ADAPT_1000G_Intersection.txt --make-bed --out 1000G_ADAPT_Intersection_chr23

#Not going to give the plink output files since there is one for each chromosome.

#Merge the per-chromosome files into a single file
for i in {2..23}; do
echo "1000G_ADAPT_Intersection_chr$[i]" >> 1000G_ADAPT_MergeList.txt
done

plink --bfile 1000G_ADAPT_Intersection_chr1 --merge-list 1000G_ADAPT_MergeList --make-bed --out 1000G_ADAPT_Intersection
```

PLINK v1.90b5.2 64-bit (9 Jan 2018)
Options in effect:
–bfile 1000G_ADAPT_Intersection_chr1
–make-bed
–merge-list 1000G_ADAPT_MergeList.txt
–out 1000G_ADAPT_Intersection

Hostname: aci-lgn-001.acib.production.int.aci.ics.psu.edu
Working directory: /storage/work/jdw345/ReferenceDatasets/1000G_Phase3_Plink_gp0.9_BiallelicOnly
Start time: Thu Mar 22 14:45:39 2018

Random number seed: 1521744339
129063 MB RAM detected; reserving 64531 MB for main workspace.
Performing single-pass merge (2504 people, 526548 variants).
Merged fileset written to 1000G_ADAPT_Intersection-merge.bed + 1000G_ADAPT_Intersection-merge.bim + 1000G_ADAPT_Intersection-merge.fam .
526548 variants loaded from .bim file.
2504 people (0 males, 0 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to 1000G_ADAPT_Intersection.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 2504 founders and 0 nonfounders present.
Calculating allele frequencies... done.
526548 variants and 2504 people pass filters and QC.
Note: No phenotypes present.
–make-bed to 1000G_ADAPT_Intersection.bed + 1000G_ADAPT_Intersection.bim + 1000G_ADAPT_Intersection.fam ... done.

End time: Thu Mar 22 14:45:48 2018

For the study samples

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved --
    extract ADAPT_1000G_Intersection.txt --make-bed --out ADAPT_2721ppl_SymShape_geno0.1
    _mind0.1_AgeMissingData_RelativesRemoved_Intersection
```

```
PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
--bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved
--extract ADAPT_1000G_Intersection.txt
--make-bed
--out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_
MergeWith1000G
Start time: Thu Mar 22 15:04:07 2018

Random number seed: 1521745447
16262 MB RAM detected; reserving 8131 MB for main workspace.
567787 variants loaded from .bim file.
1921 people (725 males, 1196 females) loaded from .fam.
--extract: 526548 variants remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.995225.
526548 variants and 1921 people pass filters and QC.
Note: No phenotypes present.
--make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection.bim
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection.fam
... done.

End time: Thu Mar 22 15:04:09 2018
```

3.3 Merge with 1000G

Try a first round of merging

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection --bmerge 1000G_ADAPT_Intersection --
    make-bed --out ADAPT_2721_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_1000G
```

```
PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
--bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection
```

–bmerge 1000G_ADAPT_Intersection
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-
1000G

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G
Start time: Thu Mar 22 15:05:35 2018

Random number seed: 1521745535
16262 MB RAM detected; reserving 8131 MB for main workspace.
1921 people loaded from
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection.fam.
2504 people to be merged from 1000G_ADAPT_Intersection.fam.
Of these, 2504 are new, while 0 are present in the base dataset.
Warning: Multiple chromosomes seen for variant 'rs3021087'.
Warning: Multiple chromosomes seen for variant 'rs61774271'.
Warning: Multiple positions seen for variant 'rs2155163'.
526548 markers loaded from
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection.bim.
526548 markers to be merged from 1000G_ADAPT_Intersection.bim.
Of these, 0 are new, while 526548 are present in the base dataset.
Error: 244 variants with 3+ alleles present.
* If you believe this is due to strand inconsistency, try –flip with
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-
1000G-merge.missnp.
(Warning: if this seems to work, strand errors involving SNPs with A/T or C/G
alleles probably remain in your data. If LD between nearby SNPs is high,
–flip-scan should detect them.)
* If you are dealing with genuine multiallelic variants, we recommend exporting
that subset of the data to VCF (via e.g. '–recode vcf'), merging with
another tool/script, and then importing the result; PLINK is not yet suited
to handling them.

End time: Thu Mar 22 15:05:36 2018

Since there were 244 variants with 3+ alleles present, we will first try to flip those variants to see if they are genuinely a result of strand inconsistencies.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection --flip ADAPT_2721ppl_SymShape_geno0.1
    _mind0.1_AgeMissingData_RelativesRemoved_Intersection_1000G-merge.missnp --make-bed
    --out ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection
–flip ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-
1000G-merge.missnp
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-

Flip

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G
Start time: Thu Mar 22 15:07:16 2018

Random number seed: 1521745636
16262 MB RAM detected; reserving 8131 MB for main workspace.
526548 variants loaded from .bim file.
1921 people (725 males, 1196 females) loaded from .fam.
–flip: 244 SNPs flipped.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.995225.
526548 variants and 1921 people pass filters and QC.
Note: No phenotypes present.
–make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip.bim
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip.fam
... done.

End time: Thu Mar 22 15:07:17 2018

Now retry the merge:

```
plink --bfile 1000G_ADAPT_Intersection --bmerge ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip --make-bed --out
    ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_1000G
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile 1000G_ADAPT_Intersection
–bmerge ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_1000G

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G Start time: Thu Mar 22 15:08:37 2018

Random number seed: 1521745717
16262 MB RAM detected; reserving 8131 MB for main workspace.

2504 people loaded from 1000G_ADAPT_Intersection.fam.
1921 people to be merged from
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-Flip.fam.
Of these, 1921 are new, while 0 are present in the base dataset.
Warning: Multiple chromosomes seen for variant 'rs61774271'.
Warning: Multiple positions seen for variant 'rs2155163'.
Warning: Multiple chromosomes seen for variant 'rs3021087'.
526548 markers loaded from 1000G_ADAPT_Intersection.bim.
526548 markers to be merged from
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-Flip.bim.
Of these, 0 are new, while 526548 are present in the base dataset.
Error: 46 variants with 3+ alleles present.
* If you believe this is due to strand inconsistency, try –flip with
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-1000G-merge.missnp.
(Warning: if this seems to work, strand errors involving SNPs with A/T or C/G
alleles probably remain in your data. If LD between nearby SNPs is high,
–flip-scan should detect them.)
* If you are dealing with genuine multiallelic variants, we recommend exporting
that subset of the data to VCF (via e.g. '–recode vcf'), merging with
another tool/script, and then importing the result; PLINK is not yet suited
to handling them.

End time: Thu Mar 22 15:08:38 2018

There are still 46 variants that aren't merging well, so we are going to remove these from both the study sample and the 1000G data.

```
#For 1000G
plink --bfile 1000G_ADAPT_Intersection --exclude ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_1000G-merge.missnp --make-bed --
    out 1000G_ADAPT_Intersection_RemoveMultiallelic
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile 1000G_ADAPT_Intersection
–exclude ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_1000G-merge.missnp
–make-bed
–out 1000G_ADAPT_Intersection_RemoveMultiallelic

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G`
Start time: Thu Mar 22 15:10:28 2018

Random number seed: 1521745828
16262 MB RAM detected; reserving 8131 MB for main workspace.
526548 variants loaded from .bim file.
2504 people (0 males, 0 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to 1000G_ADAPT_Intersection_RemoveMultiallelic.nosex

. –exclude: 526502 variants remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 2504 founders and 0 nonfounders present.
Calculating allele frequencies... done.
526502 variants and 2504 people pass filters and QC.
Note: No phenotypes present.
–make-bed to 1000G_ADAPT_Intersection_RemoveMultiallelic.bed +
1000G_ADAPT_Intersection_RemoveMultiallelic.bim +
1000G_ADAPT_Intersection_RemoveMultiallelic.fam ... done.

End time: Thu Mar 22 15:10:30 2018

```
#Now for the study samples
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip --exclude
    ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_1000G-merge.missnp --make-bed --
    out ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip
–exclude ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_1000G-merge.missnp
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G`
Start time: Thu Mar 22 15:11:53 2018

Random number seed: 1521745913
16262 MB RAM detected; reserving 8131 MB for main workspace.
526548 variants loaded from .bim file.
1921 people (725 males, 1196 females) loaded from .fam.
–exclude: 526502 variants remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1874 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.995225.
526502 variants and 1921 people pass filters and QC.
Note: No phenotypes present.
–make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic.bim

+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic.fam
... done.

End time: Thu Mar 22 15:11:55 2018

Try the merge a third time:

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic --bmerge 1000
    G_ADAPT_Intersection_RemoveMultiallelic --make-bed --out
    ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic
–bmerge 1000G_ADAPT_Intersection_RemoveMultiallelic
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G`
Start time: Thu Mar 22 15:12:36 2018

Random number seed: 1521746016
16262 MB RAM detected; reserving 8131 MB for main workspace.
1921 people loaded from
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic.fam.
2504 people to be merged from 1000G_ADAPT_Intersection_RemoveMultiallelic.fam.
Of these, 2504 are new, while 0 are present in the base dataset.
Warning: Multiple chromosomes seen for variant 'rs3021087'.
Warning: Multiple chromosomes seen for variant 'rs61774271'.
Warning: Multiple positions seen for variant 'rs2155163'.
526502 markers loaded from
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic.bim.
526502 markers to be merged from
1000G_ADAPT_Intersection_RemoveMultiallelic.bim.
Of these, 0 are new, while 526502 are present in the base dataset.
Performing single-pass merge (4425 people, 526502 variants).
Merged fileset written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G-merge.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G-merge.bim
+

ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G-merge.fam
.
526502 variants loaded from .bim file.
4425 people (725 males, 1196 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G.nosex
.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 4378 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997927.
526502 variants and 4425 people pass filters and QC.
Note: No phenotypes present.
–make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G.bim
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G.fam
... done.

End time: Thu Mar 22 15:12:51 2018

3.4 Remove palindromic SNPs

We also wanted to additionally remove palindromes (A/T, G/C) SNPs, since we are not dealing with phased data and thus don't know which strand of the chromosome the SNPs are on.
The ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G.bim file was read into Excel and filters were used to create a list of all non-palindromic snps: "ADAPT_1000G_NoPalindromes.snplist.txt" and this list was used to create a new file without palindromic SNPs.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G --
    extract ADAPT_1000G_NoPalindromes.snplist.txt --make-bed
    ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin
    --out
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000
–exclude ADAPT_1000G_NoPalindromes.snplist.txt
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-Flip_RemoveMultiallelic_1000G_NoPalin

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\5_MergeWith1000G`
Start time: Thu Mar 22 15:13:28 2018

Random number seed: 1536869257
16262 MB RAM detected; reserving 8131 MB for main workspace.
526502 variants loaded from .bim file.
4425 people (725 males, 1196 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.nosex
–exclude: 522918 variants remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 4378 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997932
522918 variants and 4425 people pass filters and QC.
Note: No phenotypes present.
–make-bed to ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.bed + ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.bim + ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.fam ... done.

End time: Thu Mar 22 15:13:30 2018

4 ADMIXTURE

4.1 Prune for LD

Prior to running admixture, we pruned for linkage disequilibrium by first calculating SNPs in linkage

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
      _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin
      --indep 50 5 2 --out ADAPT_2721ppl_SymShape_geno0.1_mind0.1
      _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin
–indep 50 5 2
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\Box\MHC_paper\1_FinalDataCuration\5_MergeWith1000G\WithoutPalindromes`
Start time: Wed Mar 28 23:14:24 2018

Random number seed: 1522293264
16262 MB RAM detected; reserving 8131 MB for main workspace.
522918 variants loaded from .bim file.
4425 people (725 males, 1196 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.nosex
.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 4378 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997932.
522918 variants and 4425 people pass filters and QC.
Note: No phenotypes present.
Pruned 24855 variants from chromosome 1, leaving 15736.
Pruned 26711 variants from chromosome 2, leaving 15143.
Pruned 21939 variants from chromosome 3, leaving 12857.
Pruned 19410 variants from chromosome 4, leaving 11952.
Pruned 19487 variants from chromosome 5, leaving 11684.
Pruned 22602 variants from chromosome 6, leaving 12157.
Pruned 17581 variants from chromosome 7, leaving 10788.
Pruned 18246 variants from chromosome 8, leaving 9992.
Pruned 15461 variants from chromosome 9, leaving 9055.
Pruned 17003 variants from chromosome 10, leaving 10059.
Pruned 15691 variants from chromosome 11, leaving 9338.
Pruned 15241 variants from chromosome 12, leaving 9452.
Pruned 12017 variants from chromosome 13, leaving 7351.
Pruned 10387 variants from chromosome 14, leaving 6559.
Pruned 9510 variants from chromosome 15, leaving 6366.
Pruned 9350 variants from chromosome 16, leaving 6958.

Pruned 7904 variants from chromosome 17, leaving 6417.
Pruned 9285 variants from chromosome 18, leaving 6224.
Pruned 4712 variants from chromosome 19, leaving 4895.
Pruned 7863 variants from chromosome 20, leaving 5484.
Pruned 4520 variants from chromosome 21, leaving 3182.
Pruned 4391 variants from chromosome 22, leaving 3343.
Pruned 7710 variants from chromosome 23, leaving 6049.
Pruned 0 variants from chromosome 26, leaving 1.
Pruning complete. 321876 of 522918 variants removed.
Marker lists written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G_NoPalin.prune.in
and
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G_NoPalin.prune.out
.

End time: Wed Mar 28 23:14:59 2018

Then by removing those SNPs from the dataset

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin
    --exclude ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.
    prune.out --make-bed --out ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin
–exclude ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin.prune.out
–make-bed
–out ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_-Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\Box\MHC_paper\1_FinalDataCuration\5_MergeWith1000G\WithoutPalindromes
Start time: Wed Mar 28 23:15:41 2018

Random number seed: 1522293341
16262 MB RAM detected; reserving 8131 MB for main workspace.
522918 variants loaded from .bim file.
4425 people (725 males, 1196 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-RemoveMultiallelic_1000G_NoPalin_LDPrune.nosex
.
–exclude: 201042 variants remaining.

```
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 4378 founders and 47 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997761.
201042 variants and 4425 people pass filters and QC.
Note: No phenotypes present.
–make-bed to
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-
RemoveMultiallelic_1000G_NoPalin_LDPrune.bed
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-
RemoveMultiallelic_1000G_NoPalin_LDPrune.bim
+
ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_-
RemoveMultiallelic_1000G_NoPalin_LDPrune.fam
... done.

End time: Wed Mar 28 23:16:02 2018
```

4.2 Create and submit files for ADMIXTURE run

We ran ADMIXTURE from K levels of 2 to 16. This first required setting up all of those files, which was done in the following manner:

```
for i in {2..16}; do
echo '#PBS -l walltime=150:00:00' >> ADAPT_1000G_Admixture_K$[i].pbs
echo '#PBS -l nodes=1:ppn=10' >> ADAPT_1000G_Admixture_K$[i].pbs
echo '#PBS -l pmem=8gb' >> ADAPT_1000G_Admixture_K$[i].pbs
echo '#PBS -A jlt22_b_g_sc_default' >> ADAPT_1000G_Admixture_K$[i].pbs
echo '#PBS -j oe' >> ADAPT_1000G_Admixture_K$[i].pbs
echo 'cd $PBS_O_WORKDIR' >> ADAPT_1000G_Admixture_K$[i].pbs
echo "admixture -j10 --cv ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune
    .bed $[i] | tee ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune_K$
    [i].log" >> ADAPT_1000G_Admixture_K$[i].pbs
done
```

The files were then submitted to the Penn State ACI-B cluster using:

```
for i in {2..16}; do qsub ADAPT_1000G_Admixture_K$[i].pbs; done
```

4.3 Analyze ADMIXTURE results and determine appropriate K value

As a first guess of which K value to choose for our ancestry filtering, we looked at all the CV error values produced by the ADMIXTURE program, with the understanding that low values are better.

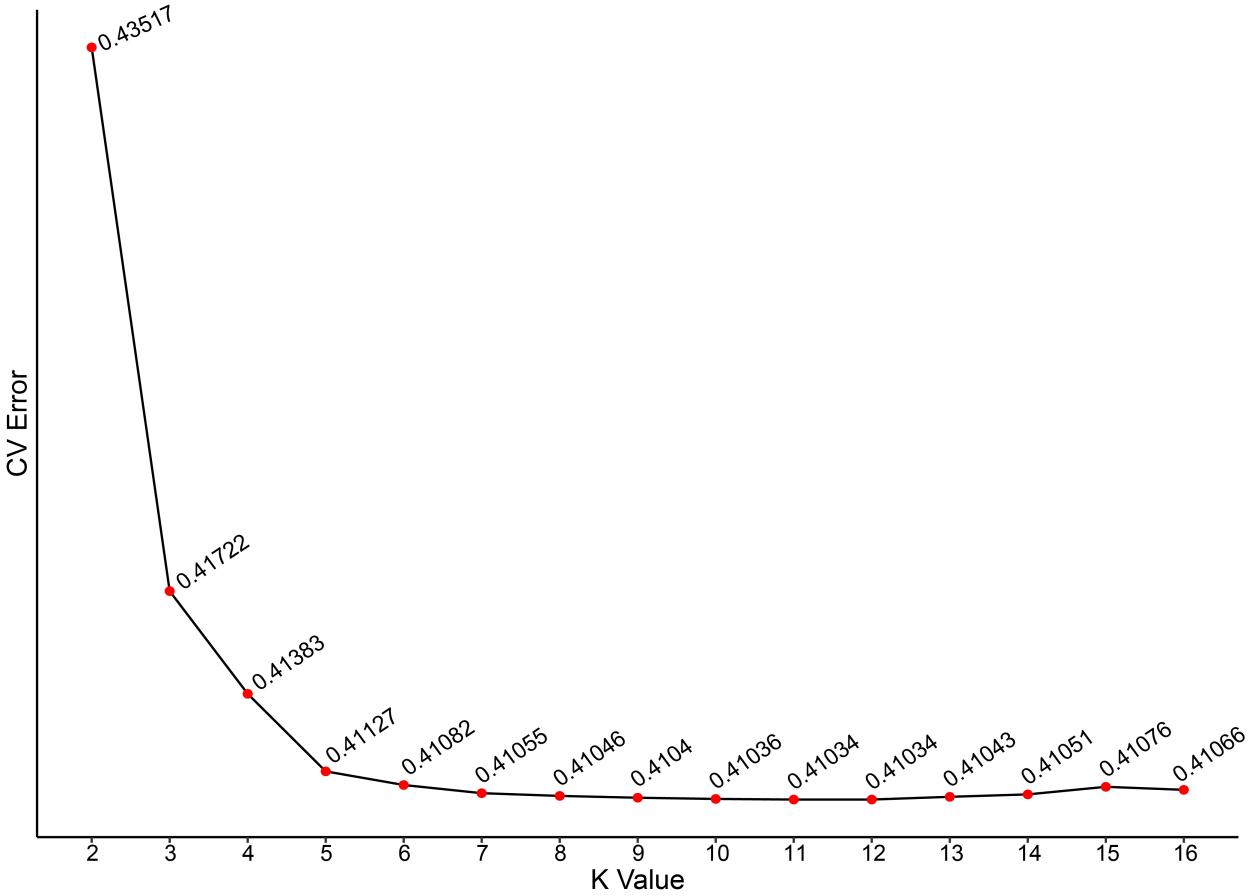


Figure S3: Cross validation error for ADMIXTURE run using K from 2 to 16. A K value of 11 and 12 gave the absolute lowest CV error. However, an elbow in the curve can be seen at K = 6. Therefore, K=6 was chosen to preliminary select individuals with European-derived ancestry, as it facilitated easy separation of the five 1000 Genomes (1) super-populations (Fig. S1) and had a low CV error.

We also created some ancestry plots with the 1000G individuals as reference, to see how the continental groups were differentiated. This was first performed by merging the population labels from 1000G with the .fam files, using VLOOKUP in Excel. Then, each ADMIXTURE component was sorted from largest to smallest and a note was made of which 1000G population had the most ancestry deriving from that component. This was repeated until we had a rough idea of which component was best associated with which 1000G population.

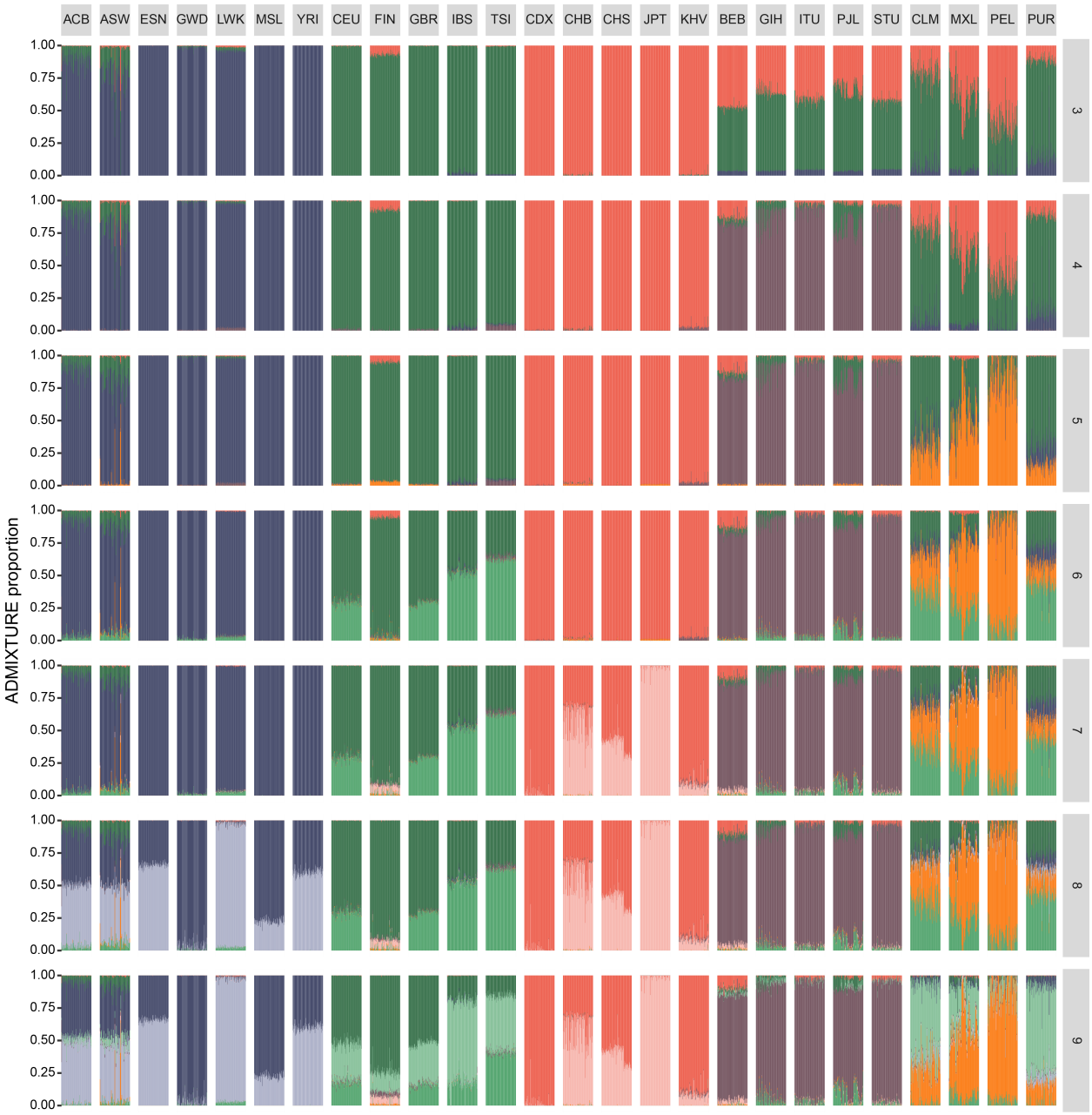


Figure S2: ADMIXTURE plot for 1000 Genomes samples for K from 3 to 9. Scale on the left Y-axis shows the ADMIXTURE fraction from each cluster (represented by different colors). Each vertical bar represents an individuals ancestry fraction and the relative heights of a color in each bar represents the ancestry fraction from that cluster. Individuals are grouped column-wise into known population and continental groupings as shown on top (see Table S1 for description of abbreviations). Results are further grouped row-wise for different values of K indicated on the right.

The R code for creating a single-page ADMIXTURE graph similar to the one above is as follows:

```
1 library(ggplot2)
2 library(reshape2)
3 library(readr)
4
5 setwd("C:/Users/Julie_White/Box/MHC_paper/1_FinalDataCuration/6_ADMIXTURE/
      WithoutPalindromes")
6
```

```
q3<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.3.Q',header=F,sep=" ")
q4<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.4.Q',header=F,sep=" ")
q5<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.5.Q',header=F,sep=" ")
q6<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.6.Q',header=F,sep=" ")
q7<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.7.Q',header=F,sep=" ")
q8<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.8.Q',header=F,sep=" ")
q9<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.9.Q',header=F,sep=" ")
q10<-read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDprune.10.Q',header=F,sep=" ")

fam <- read.table('ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune.fam',header=F,sep=" ")
groups <- read.table('AdmixtureGroupLabels.txt', header = T, sep = "\t")

q3 <- cbind(fam[2], q3)
names(q3) <- c("IID", "C1", "C2", "C3")
q3 <- merge(groups, q3, by = "IID")

q4 <- cbind(fam[2], q4)
names(q4) <- c("IID", "C1", "C2", "C3", "C4")
q4 <- merge(groups, q4, by = "IID")

q5 <- cbind(fam[2], q5)
names(q5) <- c("IID", "C1", "C2", "C3", "C4", "C5")
q5 <- merge(groups, q5, by = "IID")

q6 <- cbind(fam[2], q6)
names(q6) <- c("IID", "C1", "C2", "C3", "C4", "C5", "C6")
q6 <- merge(groups, q6, by = "IID")

q7 <- cbind(fam[2], q7)
names(q7) <- c("IID", "C1", "C2", "C3", "C4", "C5", "C6", "C7")
q7 <- merge(groups, q7, by = "IID")

q8 <- cbind(fam[2], q8)
names(q8) <- c("IID", "C1", "C2", "C3", "C4", "C5", "C6", "C7", "C8")
q8 <- merge(groups, q8, by = "IID")

q9 <- cbind(fam[2], q9)
names(q9) <- c("IID", "C1", "C2", "C3", "C4", "C5", "C6", "C7", "C8", "C9")
q9 <- merge(groups, q9, by = "IID")

q10 <- cbind(fam[2], q10)
names(q10) <- c("IID", "C1", "C2", "C3", "C4", "C5", "C6", "C7", "C8", "C9", "C10")
q10 <- merge(groups, q10, by = "IID")

mq3<-melt(q3[which(!is.na(q3$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq4<-melt(q4[which(!is.na(q4$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq5<-melt(q5[which(!is.na(q5$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq6<-melt(q6[which(!is.na(q6$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq7<-melt(q7[which(!is.na(q7$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq8<-melt(q8[which(!is.na(q8$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq9<-melt(q9[which(!is.na(q9$Pop)),], id.vars = c("FID", "IID", "Pop"))
mq10<-melt(q10[which(!is.na(q10$Pop)),], id.vars = c("FID", "IID", "Pop"))

mq3$components<-3
mq4$components<-4
mq5$components<-5
mq6$components<-6
mq7$components<-7
```

```
mq8$components<-8
mq9$components<-9
mq10$components<-10

mq3$color<-3
mq4$color<-4
mq5$color<-5
mq6$color<-6
mq7$color<-7
mq8$color<-8
mq9$color<-9
mq10$color<-10

mq3 <- within(mq3, color[variable == 'C1'] <- 'GWD')
mq3 <- within(mq3, color[variable == 'C2'] <- 'CDX')
mq3 <- within(mq3, color[variable == 'C3'] <- 'FIN')

mq4 <- within(mq4, color[variable == 'C1'] <- 'FIN')
mq4 <- within(mq4, color[variable == 'C2'] <- 'ITU')
mq4 <- within(mq4, color[variable == 'C3'] <- 'GWD')
mq4 <- within(mq4, color[variable == 'C4'] <- 'CDX')

mq5 <- within(mq5, color[variable == 'C1'] <- 'CDX')
mq5 <- within(mq5, color[variable == 'C2'] <- 'PEL')
mq5 <- within(mq5, color[variable == 'C3'] <- 'GWD')
mq5 <- within(mq5, color[variable == 'C4'] <- 'ITU')
mq5 <- within(mq5, color[variable == 'C5'] <- 'FIN')

mq6 <- within(mq6, color[variable == 'C1'] <- 'PEL')
mq6 <- within(mq6, color[variable == 'C2'] <- 'CDX')
mq6 <- within(mq6, color[variable == 'C3'] <- 'GWD')
mq6 <- within(mq6, color[variable == 'C4'] <- 'FIN')
mq6 <- within(mq6, color[variable == 'C5'] <- 'TSI')
mq6 <- within(mq6, color[variable == 'C6'] <- 'ITU')

mq7 <- within(mq7, color[variable == 'C1'] <- 'JPT')
mq7 <- within(mq7, color[variable == 'C2'] <- 'GWD')
mq7 <- within(mq7, color[variable == 'C3'] <- 'PEL')
mq7 <- within(mq7, color[variable == 'C4'] <- 'CDX')
mq7 <- within(mq7, color[variable == 'C5'] <- 'ITU')
mq7 <- within(mq7, color[variable == 'C6'] <- 'TSI')
mq7 <- within(mq7, color[variable == 'C7'] <- 'FIN')

mq8 <- within(mq8, color[variable == 'C1'] <- 'PEL')
mq8 <- within(mq8, color[variable == 'C2'] <- 'JPT')
mq8 <- within(mq8, color[variable == 'C3'] <- 'TSI')
mq8 <- within(mq8, color[variable == 'C4'] <- 'FIN')
mq8 <- within(mq8, color[variable == 'C5'] <- 'GWD')
mq8 <- within(mq8, color[variable == 'C6'] <- 'LWK')
mq8 <- within(mq8, color[variable == 'C7'] <- 'ITU')
mq8 <- within(mq8, color[variable == 'C8'] <- 'CDX')

mq9 <- within(mq9, color[variable == 'C1'] <- 'LWK')
mq9 <- within(mq9, color[variable == 'C2'] <- 'ITU')
mq9 <- within(mq9, color[variable == 'C3'] <- 'CDX')
mq9 <- within(mq9, color[variable == 'C4'] <- 'GWD')
mq9 <- within(mq9, color[variable == 'C5'] <- 'PEL')
mq9 <- within(mq9, color[variable == 'C6'] <- 'FIN')
mq9 <- within(mq9, color[variable == 'C7'] <- 'TSI')
mq9 <- within(mq9, color[variable == 'C8'] <- 'IBS')
mq9 <- within(mq9, color[variable == 'C9'] <- 'JPT')

mq10 <- within(mq10, color[variable == 'C1'] <- 'CDX')
mq10 <- within(mq10, color[variable == 'C2'] <- 'FIN')
mq10 <- within(mq10, color[variable == 'C3'] <- 'ITU')
mq10 <- within(mq10, color[variable == 'C4'] <- 'JPT')
mq10 <- within(mq10, color[variable == 'C5'] <- 'GBR')
mq10 <- within(mq10, color[variable == 'C6'] <- 'GWD')
```

```
mq10 <- within(mq10, color[variable == 'C7'] <- 'LWK')
mq10 <- within(mq10, color[variable == 'C8'] <- 'PEL')
mq10 <- within(mq10, color[variable == 'C9'] <- 'CLM')
mq10 <- within(mq10, color[variable == 'C10'] <- 'TSI')

mq3_9<-rbind(mq3,mq4,mq5,mq6,mq7, mq8, mq9)

mq3_9$POP2<-factor(mq3_9$Pop,levels=c("ACB","ASW","ESN","GWD","LWK","MSL","YRI","CEU","
    FIN","GBR","IBS","TSI","CDX","CHB","CHS","JPT","KHV","BEB","GIH","ITU","PJL","STU","
    CLM","MXL","PEL","PUR"))

ancestrycolor <- c("GWD" = "#484D6D", "LWK" = "#AFB3CA", #blue
                   "FIN" = "#3d7550", "TSI" = "#57a772", "IBS" = "#89c29c", #green
                   "CDX" = "#EE6352", "JPT" = "#F7B8B0", #red
                   "ITU" = "#785964", #purple
                   "PEL" = "#fb801c", "CLM" = "#fcaa67") #orange

anc.plt <- ggplot(mq3_9,aes(IID,value,fill=color))+geom_bar(stat="identity",width=1)+
  theme(axis.text.x=element_blank(),axis.ticks.x = element_blank(), legend.position = '
    none')+
  scale_fill_manual(values=ancestrycolor)+
  facet_grid(components~POP2,scales="free_x")+labs(y="ADMIXTURE proportion",x="
    Populations",fill="Components")

ggsave('ADAPT_1000G_AdmixPlot.pdf',anc.plt,width=10,height=7.5)
```

4.4 Select European Individuals

Based on the low CV error, as well as a good differentiation between continental ancestry groups, we chose K = 6 as the appropriate K value for our analysis. For K = 6,
Component 1 is represented most in the PEL 1000 Genomes population
Component 2 is represented most in the CHS 1000 Genomes population
Component 3 is represented most in the GWD 1000 Genomes population
Component 4 is represented most in the FIN 1000 Genomes population
Component 5 is represented most in the TSI 1000 Genomes population
Component 6 is represented most in the ITU 1000 Genomes population.

To determine the European individuals in our dataset, we summed components 4 and 5 to create a broadly European ancestry component. Then we instituted the following filters:

Less than 10% ancestry from C1, C2, C3, and C6 and more than 90% ancestry from C4 + C5. Individuals left after these filters were designated as European and placed in the list "ADAPT_1000G_EuroList.txt".

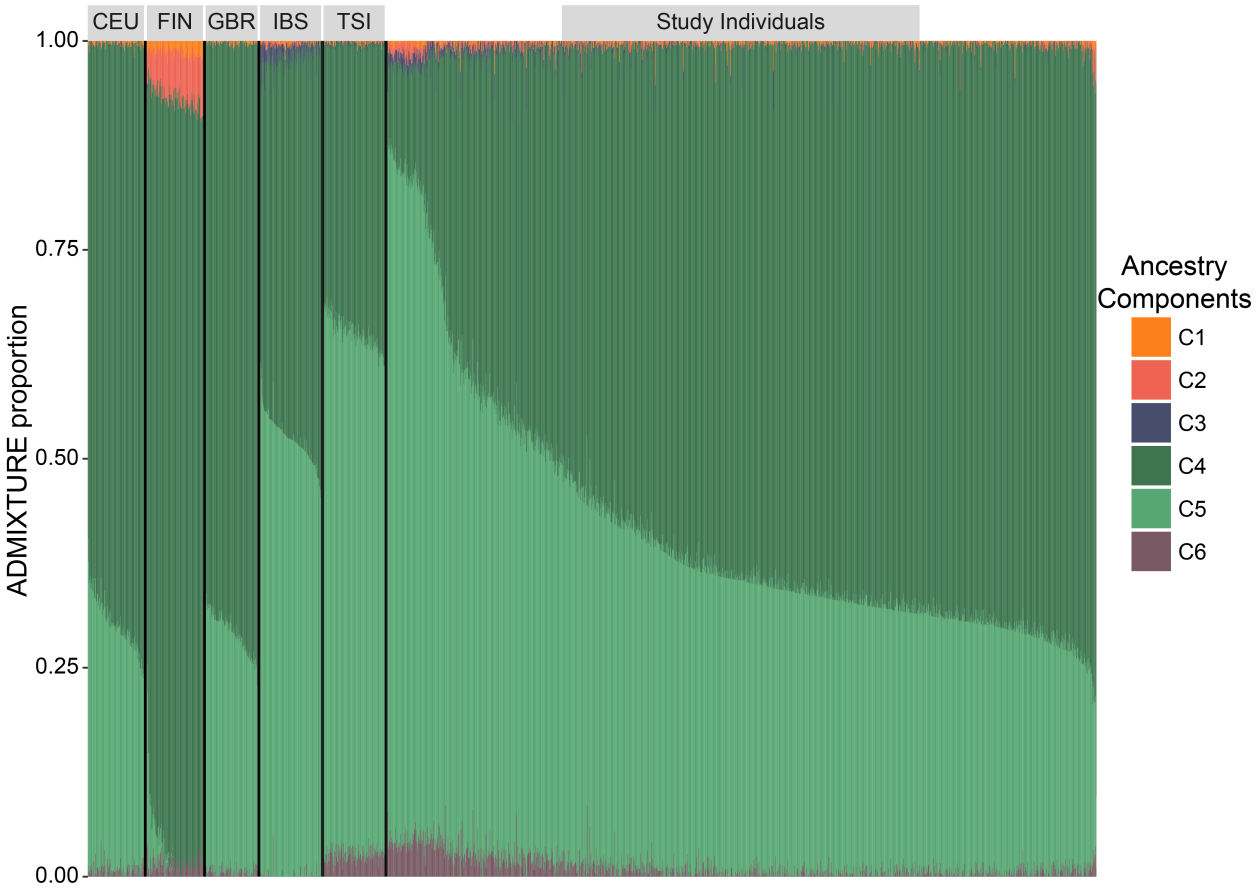


Figure S4: ADMIXTURE (K=6) results of individuals of European ancestry. The 1000 Genomes European individuals (CEU, FIN, GBR, IBS, and TSI) are used here as reference to illustrate that the individuals used for analyses in this paper are primarily of European ancestry.

The above plot was created using the following R code:

```
PlotEuroADAPT_K6 <- read.table("ADAPT_1000G_PlotEuro_K6.txt", sep = "\t", header = TRUE)
mk6 <- melt(PlotEuroADAPT_K6, id.vars = c("FID", "IID", "Pop", "Order"))

ancestrycolor <- c("C1" = "#fb801c", #orange
                   "C2" = "#EE6352",#red
                   "C3" = "#484D6D", #blue
                   "C4" = "#3d7550", #green
                   "C5" = "#57a772",
                   "C6" = "#785964", #purple
                   "Blank" = "black")

anc.plt.euro <- ggplot(mk6, aes(x=Order, y = value, fill = variable))+geom_bar(stat = "
    identity")+
  theme(axis.text.x=element_blank(),axis.ticks.x = element_blank(), panel.background =
    element_blank())+
  scale_fill_manual(values=ancestrycolor)+
  labs(x="Individuals", y="ADMIXTURE proportion", fill="Components")

ggsave('ADAPT_1000G_EuroPlot_K6.pdf',anc.plt.euro,width=10,height=7.5)
```

4.5 Filter genetic data to keep only European individuals

Individuals selected as being European were filtered to create a new European genetic dataset.

```
plink --bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1
    _AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune
     --keep ADAPT_1000G_EuroList.txt --make-bed --out ADAPT_1000G_NoPalin_LDPrune_Euro
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_RelativesRemoved_Intersection_Flip_RemoveMultiallelic_1000G_NoPalin_LDPrune
–keep ADAPT_1000G_EuroList.txt
–make-bed
–out ADAPT_1000G_NoPalin_LDPrune_Euro

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\Box\MHC_paper\1_FinalDataCuration\7_ExtractNewEuroSet\WithoutPalindromes
Start time: Thu Mar 29 11:44:23 2018

Random number seed: 1522338263
16262 MB RAM detected; reserving 8131 MB for main workspace.
201042 variants loaded from .bim file.
4425 people (725 males, 1196 females, 2504 ambiguous) loaded from .fam.
Ambiguous sex IDs written to ADAPT_1000G_NoPalin_LDPrune_Euro.nosex.
–keep: 1752 people remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1710 founders and 42 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.996223.
201042 variants and 1752 people pass filters and QC.
Note: No phenotypes present.
–make-bed to ADAPT_1000G_NoPalin_LDPrune_Euro.bed + ADAPT_1000G_NoPalin_LDPrune_Euro.bim + ADAPT_1000G_NoPalin_LDPrune_Euro.fam ... done.

End time: Thu Mar 29 11:44:30 2018

5 Eigensoft

5.1 SmartPCA run

Eigensoft (https://www.hsph.harvard.edu/alkes-price/software/) is a tool by which you can remove outliers based on PC scores and was utilized here to check for outliers in our European sample of individuals.

In this run, we specify that the study samples are the group by which the PCA space should be built on, so while the 1000G samples are included in the run, they don't play a role in the outlier statistics. The PCA scores for the study sample are the same by doing this vs. a run with the genetic data only including study samples.

```
1 smartpca -p ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.parfile
```

```
parameter file: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.parfile
THE INPUT PARAMETERS
PARAMETER NAME: VALUE
genotypename: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins.bed
snpname: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins.bim
indivname: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins.fam
evecoutname: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.pca.evec
evaloutname: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.pca.eval
altnormstyle: NO
numoutevec: 10
numoutlieriter: 5
numoutlierevec: 10
outliersigmathresh: 6
qtmode: 0
snpweightoutname: ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.snpwt
poplistname: ADAPT_Reference.txt
smartpca version: 16000
norm used

genetic distance set from physical distance
genotype file processed
number of samples used: 1249 number of snps used: 194992
Using 9 threads, and partial sum lookup algorithm.
snp rs200498230 ignored . allelecnt: 0 missing: 1
snp rs201136045 ignored . allelecnt: 0 missing: 21
snp rs28569008 ignored . allelecnt: 0 missing: 1
snp rs187819489 ignored . allelecnt: 0 missing: 55
snp rs35687686 ignored . allelecnt: 0 missing: 0
snp rs35072750 ignored . allelecnt: 0 missing: 0
snp rs12068394 ignored . allelecnt: 0 missing: 41
snp rs2235733 ignored . allelecnt: 0 missing: 0
snp rs75355390 ignored . allelecnt: 0 missing: 2
snp rs148831269 ignored . allelecnt: 0 missing: 25
total number of snps killed in pass: 2512 used: 192480
REMOVED outlier ADAPT:140521 iter 1 evec 4 sigmage 19.822 pop: ADAPT
REMOVED outlier ADAPT:141012 iter 1 evec 7 sigmage -6.045 pop: ADAPT
REMOVED outlier ADAPT:141642 iter 1 evec 5 sigmage -10.520 pop: ADAPT
REMOVED outlier ADAPT:141751 iter 1 evec 3 sigmage -12.441 pop: ADAPT
REMOVED outlier ADAPT:141752 iter 1 evec 3 sigmage -29.271 pop: ADAPT
REMOVED outlier ADAPT:143507 iter 1 evec 7 sigmage -17.080 pop: ADAPT
```

```
REMOVED outlier ADAPT:143518 iter 1 evec 6 sigmage -7.506 pop: ADAPT
REMOVED outlier ADAPT:143612 iter 1 evec 2 sigmage 6.150 pop: ADAPT
number of samples after outlier removal: 1241
total number of snps killed in pass: 2547 used: 192445
REMOVED outlier ADAPT:140077 iter 2 evec 6 sigmage -6.722 pop: ADAPT
REMOVED outlier ADAPT:140211 iter 2 evec 5 sigmage -16.818 pop: ADAPT
REMOVED outlier ADAPT:141500 iter 2 evec 5 sigmage 13.257 pop: ADAPT
REMOVED outlier ADAPT:141673 iter 2 evec 6 sigmage -9.771 pop: ADAPT
number of samples after outlier removal: 1237
total number of snps killed in pass: 2571 used: 192421
REMOVED outlier 42:140522 iter 3 evec 7 sigmage 8.143 pop: ADAPT
REMOVED outlier ADAPT:141534 iter 3 evec 7 sigmage 8.963 pop: ADAPT
number of samples after outlier removal: 1235
total number of snps killed in pass: 2571 used: 192421
REMOVED outlier 70:141460 iter 4 evec 6 sigmage -6.092 pop: ADAPT
REMOVED outlier ADAPT:140049 iter 4 evec 9 sigmage 6.064 pop: ADAPT
number of samples after outlier removal: 1233
total number of snps killed in pass: 2575 used: 192417
end of smartpca run
```

When it came to plotting the PCs, we found that we had better resolution when plotting self reported ancestry vs. 1000 Genomes population codes.

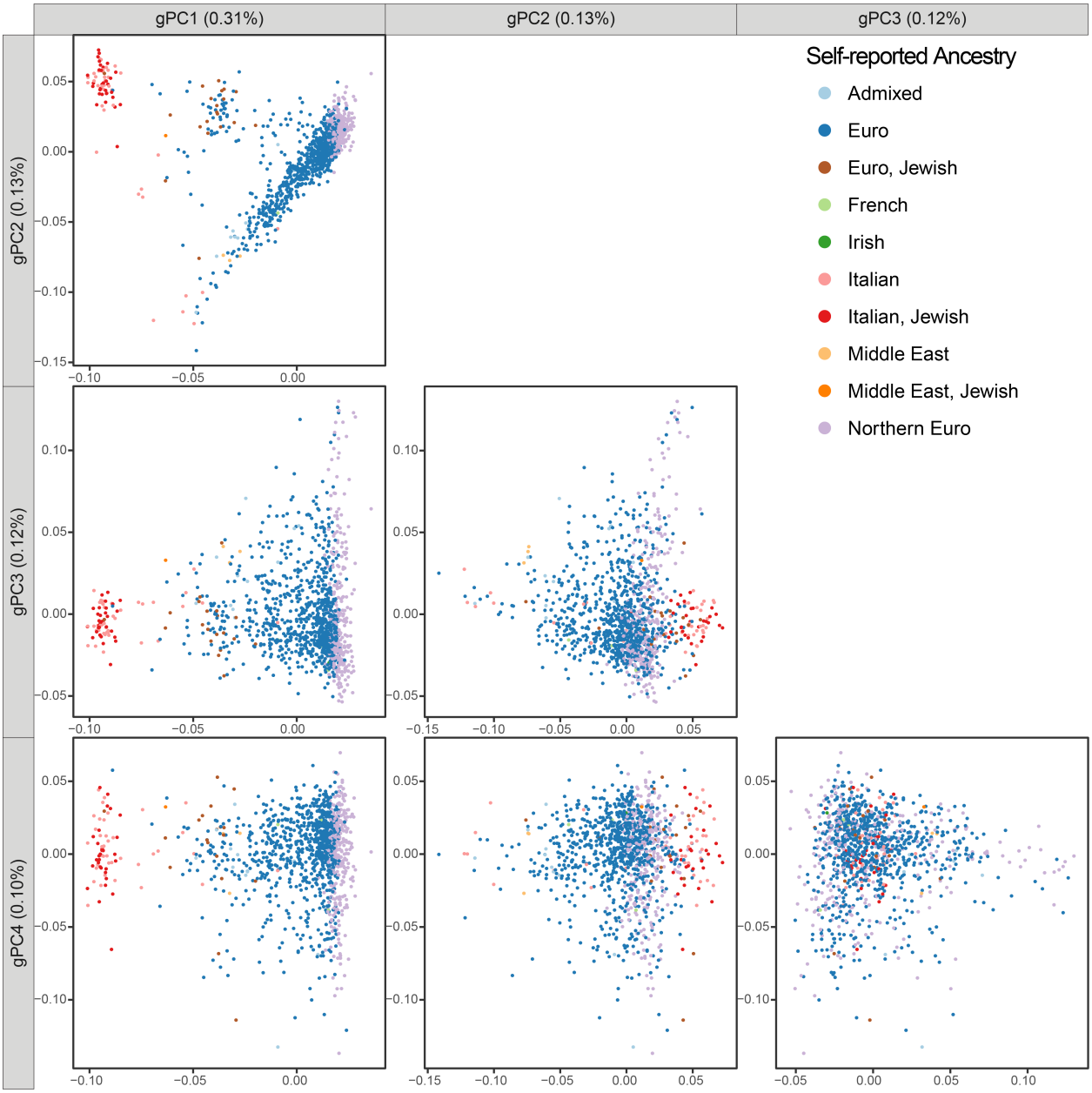


Figure S5: PCA of individuals of European ancestry. PCA on the study individuals was performed using Eigensoft and outliers were removed. Points are colored based on self-reported ancestry.

The following R code was used to make the above plot:

```
#Plot Eigensoft results
setwd("C:/Users/Julie_White/Box/MHC_paper/1_FinalDataCuration/8_Eigensoft/FullSample")

#Read in eigensoft data
dat <- read_delim("ADAPT_1000G_Merge3_geno0.01_LDPrune_1000GRef_Euros.pca.evec", "\t",
    escape_double = FALSE, trim_ws = TRUE)

for (i in 2:7){
  for (j in 3:8){
    if (i < j){
      p <- ggplot()+
        geom_point(data=dat[which(dat$Group=="ADAPT"),], aes_(x = as.name(colnames(dat)[
    i]), y = as.name(colnames(dat)[j])), alpha=0.5, size = 0.5) +
```

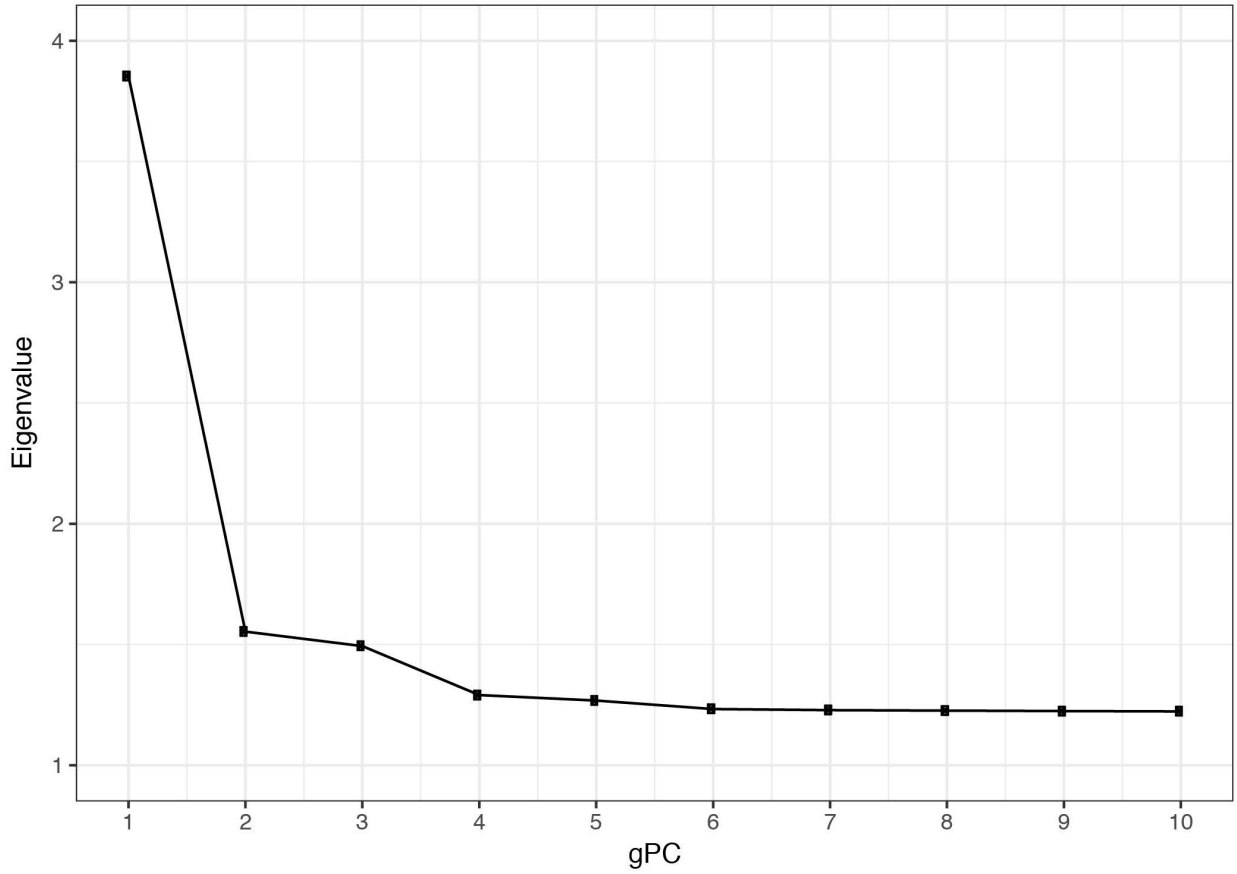
```
        theme_bw()+
        theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    legend.position = "none")

      ggsave(p, file = paste0("ADAPT_Merge3_geno0.01_LDPrune_Euros_gPC", i-1, "vsgPC", j
    -1, ".pdf"), device = "pdf", width = 3, height = 3, units = "in")
    }
  }
}

#Plot using self-reported data
for (i in 4:8){
  for (j in 5:7){
    if (i < j){
      p <- ggplot(dat, aes_(x = as.name(colnames(dat)[i]), y=as.name(colnames(dat)[j]),
    color=as.name(colnames(dat)[3])))+geom_point()+scale_color_manual(values=c("#a6cee3"
    ,"#1f78b4","#b15928","#b2df8a","#33a02c","#fb9a99","#e31a1c","#fdbf6f","#ff7f00","#
    cab2d6","#6a3d9a"))+theme_bw()+theme(panel.grid.major = element_blank(), panel.grid.
    minor = element_blank(), legend.position="none")

      ggsave(p, file = paste0("ADAPT_SelfReport_gPC", i-1, "vsgPC", j-1, ".pdf"), device
    = "pdf", width = 3, height = 3, units = "in")
    }
  }
}
```

The below plot is a scree plot of the genetic PCs. For our analysis we chose the first 3 PCs because the scree plot is quite flat after the third PC, and adding additional PCs does not explain much more variation.



The following R code was used to create the above plot:

```
library(ggplot2)
```

```
library(data.table)

evec<-fread("../../Results/Genetic_pca/ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.pca.evec",header=T)
eval<-fread("../../Results/Genetic_pca/ADAPT_1000G_NoPalin_LDPrune_Euro_NoFins_ADAPTRef.pca.eval")
colnames(eval)<-c("Eigenvalue")
eval$gPC<-seq(1,nrow(eval),1)

scree<-ggplot(data=eval[which(eval$gPC<11),],aes(gPC,Eigenvalue))+
  geom_point()+
  geom_line()+
  theme_bw()+
  scale_x_continuous(breaks=seq(1,10))+
  ylim(c(1,4))

ggsave("../../Results/Genetic_pca/gPC_screeplot_1_10_09102018.pdf",scree,height=5,width=7)
```

5.2 Remove 16 Eigensoft outliers

The 16 individuals identified as outliers in the above Eigensoft run were removed from the subsequent analyses.

```
plink --bfile ADAPT_1000G_NoPalin_LDPrune_Euro --remove IndividualsRemovedByEigensoft_And1000GToRemove.txt --make-bed --out ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft
```

```
PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
--bfile ADAPT_1000G_NoPalin_LDPrune_Euro
--make-bed
--out ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft
--remove IndividualsRemovedByEigensoft_And1000GToRemove.txt

Hostname: JWHITEPC
Working directory: C:\Users\JulieWhite\Box\MHC_paper\1_FinalDataCuration\8_Eigensoft\WithoutPalindromes
Start time: Thu Mar 29 12:40:44 2018

Random number seed: 1524170041
16262 MB RAM detected; reserving 8131 MB for main workspace.
201042 variants loaded from .bim file.
1752 people (478 males, 771 females, 503 ambiguous) loaded from .fam.
Ambiguous sex IDs written to
ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft.nosex .
--remove: 1233 people remaining.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1193 founders and 40 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.994686.
201042 variants and 1233 people pass filters and QC.
Note: No phenotypes present.
--make-bed to ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft.bed +
ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft.bim +
ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft.fam ... done.
```

```
End time: Thu Mar 29 12:40:46 2018
```

6 Calculate Genome and HLA heterozygosity

The following command was used to compute the necessary statistics to calculate genome-wide heterozygosity in the 1233 European samples

```
plink --bfile ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft --het --out
    ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft
```

PLINK v1.90b3.29 64-bit (24 Dec 2015)
Options in effect:
–bfile ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft
–het
–out ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft

Hostname: JWHITEPC
Working directory: `C:\Users\JulieWhite\BoxSync\MHC_paper\1_FinalDataCuration\9_AfterEigenOutlierRemoval`
Start time: Thu Mar 29 12:45:25 2018

Random number seed: 1522341925
16262 MB RAM detected; reserving 8131 MB for main workspace.
201042 variants loaded from .bim file.
1233 people (471 males, 762 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 1193 founders and 40 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.994686.
201042 variants and 1233 people pass filters and QC
–het: 192385 variants scanned, report written to
ADAPT_NoPalin_LDPrune_IndividualsKeptByEigensoft.het

End time: Thu Mar 29 12:45:26 2018

From the .het files produced, we calculated heterozygosity using the following equation: (N(NM)-O(HOM)) / N(NM)

To pull out the HLA heterozygosity information, we calculated heterozygosity using the above equation and the information in the following file: ADAPT_2721ppl_SymShape_geno0.1_mind0.1_AgeMissingData_-RelativesRemoved_HLAHet.het. Then, we used Excel VLOOKUP to get HLA heterozygosity information from only the 1233 European individuals.