using the pipeline of ANGSD and ngsTools to generate PCA plot of ngs data

1. Using angsd to generate the geno file

```
angsd -bam allSample.bamlist -doMaf 2 -doMajorMinor 1 -uniqueOnly 1 -minMapQ
30 -minQ 20 -SNP_pval 0.000001 -minInd 20 -doCounts 1 -setMinDepth 300 -
setMaxDepth 2400 -doGeno 32 -doPost 1 -postCutoff 0.95 -geno_minDepth 10 -
doSaf 1 -anc /home/lwang/ref/Zea_mays.AGPv3.22.dna.genome.fa -GL 2 -P 24 -
out allSample -fold 1
```

setting -doGeno 32 is important, as ngsCovar in ngsTools requires the binary genotype data with posterior probability of genotypes.

2. ngsCovar to compute the covariance matrix

```
ngsCovar -probfile allSample.geno -outfile allSample.covar -nind 30 -nsites
94795215 -block_size 20000 -call 0 -norm 0 -isfold TRUE
```

nsites here means the total number of sites filtering out of the cutoff, not only variable sites. This number can be obtained by counting the lines of the saf.pos file generated via angsd.

3. plotting it out

```
Rscript --vanilla --slave plotPCA.R -i allSample.covar -c 1-2 -a
allSample.clst.txt -o allSamplePCA.eps
```

```
# Usage: Rscript -i infile.covar -c component1-component2 -a annotation.file
-o outfile.eps

library(optparse)
library(ggplot2)

option_list <- list(make_option(c('-i','--in_file'), action='store',
type='character', default=NULL, help='Input file (output from ngsCovar)'),
                    make_option(c('-c','--comp'), action='store',
type='character', default=1-2, help='Components to plot'),
                    make_option(c('-a','--annot_file'), action='store',
type='character', default=NULL, help='Annotation file with individual
classification (2 column TSV with ID and ANNOTATION)'),
                    make_option(c('-o','--out_file'), action='store',
type='character', default=NULL, help='Output file')
                    )
opt <- parse_args(OptionParser(option_list = option_list))
```

```
# Annotation file is in plink cluster format

################################################################################

# Read input file
covar <- read.table(opt$in_file, stringsAsFact=F);

# Read annot file
annot <- read.table(opt$annot_file, sep=" ", header=T); # note that plink cluster files are usually tab-separated instead

# Parse components to analyze
comp <- as.numeric(strsplit(opt$comp, "-", fixed=TRUE)[[1]])

# Eigenvalues
eig <- eigen(covar, symm=TRUE);
eig$val <- eig$val/sum(eig$val);
cat(signif(eig$val, digits=3)*100,"\n");

# Plot
PC <- as.data.frame(eig$vectors)
colnames(PC) <- gsub("V", "PC", colnames(PC))
PC$Pop <- factor(annot$CLUSTER)

title <- paste("PC",comp[1]," (",signif(eig$val[comp[1]], digits=3)*100,"%)"," / PC",comp[2]," (",signif(eig$val[comp[2]], digits=3)*100,"%)",sep="",collapse="")

x_axis = paste("PC",comp[1],sep="")
y_axis = paste("PC",comp[2],sep="")

ggplot() + geom_point(data=PC, aes_string(x=x_axis, y=y_axis, color="Pop")) + ggtitle(title)
ggsave(opt$out_file)
unlink("Rplots.pdf", force=TRUE)
```

The plotPCA.R is part of the scripts of ngsPopGen folder. Here, the annotation file is given by “-a”. An example of it is as follows:

FID	IID	CLUSTER
RIMMA0438	1	Andean
RIMMA0466	1	Andean
RIMMA0468	1	Andean
RIMMA0662	1	Andean
RIMMA0665	1	Andean
RIMMA0421	1	MexHigh
RIMMA0625	1	MexHigh

FID	IID	CLUSTER
RIMMA0626	1	MexHigh
RIMMA0672	1	MexHigh
RIMMA0677	1	MexHigh
RIMMA0670	1	GuaHigh
RIMMA1007	1	GuaHigh
RIMMA1008	1	GuaHigh
RIMMA1011	1	GuaHigh
RIMMA0383	1	SW_US
RIMMA0384	1	SW_US
RIMMA0385	1	SW_US
RIMMA0387	1	SW_US
RIMMA0415	1	SW_US
RIMMA0409	1	MexLow
RIMMA0703	1	MexLow
RIMMA0720	1	MexLow
RIMMA0733	1	MexLow
RIMMA1010	1	MexLow
RIMMA0390	1	SA_Low
RIMMA0392	1	SA_Low
RIMMA0393	1	SA_Low
RIMMA0395	1	SA_Low
RIMMA0399	1	SA_Low