[toc]

# Pipeline

Selve pipeline run dvs. fra FastQC til GT

# T.o.m haplotypeCaller, chrashede derefter, så blev startet op igen med vcfonly  
export NXF\_LAUNCHER=$SNIC\_TMP # For Rackham  
export NXF\_TEMP=$SNIC\_TMP # For Rachham  
export NXF\_SINGULARITY\_CACHEDIR=/proj/uppstore2017228/KLT.05.GRUS/sophieflatr/K9-WGS-Pipeline/singularity  
  
nextflow -log runallagain.log run -resume /proj/uppstore2017228/KLT.05.GRUS/sophieflatr/K9-WGS-Pipeline/main.nf -profile rackham \  
 --fastqDir /home/sesagger/FCRfastqs/ \  
 --reference /proj/uppstore2017228/KLT.05.GRUS/DATA\_merge/canFam3.1.fa \  
 --known /proj/uppstore2017228/KLT.05.GRUS/DATA\_merge/dog\_dbSNP.recode.vcf \  
 --outdir /proj/uppstore2017228/KLT.05.GRUS/sophieflatr/newrun \  
 --project snic2017-7-384  
   
# Run som stod for genotypning og derfra   
export NXF\_LAUNCHER=$SNIC\_TMP # For Rackham  
export NXF\_TEMP=$SNIC\_TMP # For Rachham  
export NXF\_SINGULARITY\_CACHEDIR=/proj/uppstore2017228/KLT.05.GRUS/sophieflatr/K9-WGS-Pipeline/singularity  
  
nextflow -log runallagain.log run /proj/uppstore2017228/KLT.05.GRUS/sophieflatr/K9-WGS-Pipeline/vcf-new.nf -profile rackham1 \  
 --vcfDir /proj/uppstore2017228/KLT.05.GRUS/sophieflatr/newrun/haplotypeCaller \  
 --reference /proj/uppstore2017228/KLT.05.GRUS/DATA\_merge/canFam3.1.fa \  
 --known /proj/uppstore2017228/KLT.05.GRUS/DATA\_merge/dog\_dbSNP.recode.vcf \  
 --outdir /proj/uppstore2017228/KLT.05.GRUS/sophieflatr/newrun/onlyvcfs \  
 --project snic2017-7-384

## Make files

### Annotate files

eriks  
---------------------------------  
module load bioinfo-tools htslib/1.10 bcftools/1.10  
bcftools annotate 160DG.99.9.recalibrated.variants.eff.vcf.gz --set-id +'%CHROM\\_%POS' --output erik\_renamed.vcf  
  
bgzip -c erik\_renamed.vcf > erik\_renamed.vcf.gz  
  
tabix erik\_renamed.vcf.gz  
  
iDog  
---------------------------------  
bcftools annotate Filtred\_Published.vcf.gz --set-id +'%CHROM\\_%POS' --output iDog-filtered\_renamed.vcf  
  
bgzip -c iDog-filtered\_renamed.vcf > iDog-filtered\_renamed.vcf.gz  
  
tabix iDog-filtered\_renamed.vcf.gz

### Merging

module load bioinfo-tools bcftools/1.10 vcftools/0.1.15 htslib/1.10 plink/1.90b4.9 plink2/2.00-alpha-2-20190429  
  
bcftools merge mine\_hunde\_renamed.vcf.gz ../../../eriks/erik\_renamed.vcf.gz ../iDog/iDog-filtered\_renamed.vcf.gz --output all\_by\_bcfttools.vcf  
  
bgzip -c all\_by\_bcfttools.vcf > all\_by\_bcfttools.vcf.gz  
  
tabix all\_by\_bcfttools.vcf.gz

## Analyses

### Calculation of pooled heterogozity

module load bioinfo-tools bcftools/1.10 vcftools/0.1.15 htslib/1.10 plink/1.90b4.9 plink2/2.00-alpha-2-20190429  
plink \  
 --allow-extra-chr \  
 --allow-no-sex \  
 --bfile $fcr/all\_by\_bcf\_use\_this \  
 --dog \  
 --double-id \  
 --freq counts \  
 --keep phenos/pheno\_hsfcr\_vs\_fcr.txt \  
 --out fcr\_frq\_count\_all \  
 --pheno phenos/pheno\_hsfcr\_vs\_fcr.txt  
   
grep -v "NaN" fcr\_frq\_count\_all.frq.counts

#Libraries  
# Rcpp\_1.0.3 compiler\_3.6.0 pillar\_1.4.3   
# prettyunits\_1.1.1 progress\_1.2.2 bitops\_1.0-6   
# tools\_3.6.0 digest\_0.6.23 zeallot\_0.1.0   
# bit\_1.1-15.2 RSQLite\_2.2.0 memoise\_1.1.0   
# lifecycle\_0.1.0 tibble\_2.1.3 gtable\_0.3.0   
# pkgconfig\_2.0.3 rlang\_0.4.0 DBI\_1.1.0   
# curl\_4.3 parallel\_3.6.0 httr\_1.4.1   
# hms\_0.5.1 IRanges\_2.18.3 S4Vectors\_0.22  
# vctrs\_0.2.0 gtools\_3.8.1 stats4\_3.6.0   
# bit64\_0.9-7 grid\_3.6.0 tidyselect\_0.2  
# Biobase\_2.44.0 glue\_1.3.1 calibrate\_1.7.  
# R6\_2.4.1 AnnotationDbi\_1.46.1 XML\_3.99-0.3   
# purrr\_0.3.2 blob\_1.2.0 magrittr\_1.5   
# BiocGenerics\_0.30.0 backports\_1.1.5 MASS\_7.3-51.4   
# assertthat\_0.2.1 colorspace\_1.4-1 stringi\_1.4.5   
# RCurl\_1.98-1.1 crayon\_1.3.4 tidyr\_1.0.2   
# biomaRt\_2.40.5 forcats\_0.4.0 gdata\_2.18.0   
# stringr\_1.4.0 dplyr\_0.8.4 qqman\_0.1.4   
  
hentfil("fcr\_freq\_counts\_no\_allmiss.tsv",df)  
df$Location<-str\_split\_fixed(df$SNP,"\_",2)[,2]  
hp\_calc<-function (df,bin\_size,bin\_step)  
{   
 df3<-c()  
 i<-0  
 j<-1  
 while (j<39){  
 a<-subset(df,CHR==j)  
 if(dim(a)[1]!=0){  
 i<-0  
 while (i < max(a$Location)){  
 bin\_start<-i  
 bin\_end<-i+bin\_size  
 b<-subset(a,a$Location>=bin\_start&a$Location<bin\_end)  
 C1\_sum<-sum(b$C1,na.rm=TRUE)  
 C2\_sum<-sum(b$C2,na.rm=TRUE)  
 G0\_sum<-sum(b$G0,na.rm=TRUE)  
 df2<-cbind(j,bin\_start,bin\_end,C1\_sum,C2\_sum,G0\_sum)  
 i<-i+bin\_step  
 if(C1\_sum!=0|C2\_sum!=0|G0\_sum!=0){  
 df3<-bind\_rows(df3,as.data.frame(df2))  
 }  
 }  
 }   
 j<-j+1  
 }  
 df3$hp<-2\*df3$C1\_sum\*df3$C2\_sum/(df3$C1\_sum+df3$C2\_sum)^2  
 df3$Z\_hp<-scale(df3$hp,center=TRUE,scale=TRUE)  
 assign(paste(deparse(substitute(df)), "summed", sep = "\_"),   
 df3, envir = .GlobalEnv)  
}

### Prune for $\bold{F\_{ST}}$

#PLINK v1.90b4.9 64-bit (13 Oct 2017)  
plink \  
 --allow-extra-chr \  
 --allow-no-sex \  
 --bfile $fcr/SNP/all\_by\_bcf\_use\_this \  
 --dog \  
 --double-id \  
 --geno 0.05 \  
 --indep-pairwise 50 10 0.5 \  
 --out fcrpruned\_pendleton\_r2\_0.05\_geno\_0.05 \  
 --pheno $fcr/SNP/phenos/pheno\_fcr\_vs\_eriks.txt \  
 --remove $fcr/SNP/phenos/3dogs.txt

### Calculate

#### Windowed

##### Make the file

module load bioinfo-tools bcftools/1.10 vcftools/0.1.15 htslib/1.10 plink/1.90b4.9 plink2/2.00-alpha-2-20190429  
  
vcftools \  
 --gzvcf all\_by\_bcfttools.vcf.gz \  
 --fst-window-size {100000/200000} \  
 --fst-window-step {50000/100000} \  
 --weir-fst-pop fcr\_for\_fst.txt \  
 --weir-fst-pop {all,iDog,eriks}\_for\_fst.txt \  
 --keep fcr\_for\_fst.txt \  
 --keep {all,iDog,eriks}\_for\_fst.txt \  
 --out fcr\_{all,iDog,eriks}\_fst\_w\_{100000,200000}\_s\_{50000,100000}kb

##### Databehandling

is used, unless otherwise specified.

Fejltjek:

str(df)  
summary(df)  
#new\_df <- something(df)  
dim(newdf)==dim(df)  
# Summarize  
 df %>% dplyr::select(-Feature) %>% subset(filter options) %>% distinct(Location,.keep\_all=TRUE) #Check length=length(count(Location)) - to see if any locations have been dropped  
MAF=ifelse(as.character(fcr\_e\_g5$Allele)==as.character(fcr\_e\_g5$A1),fcr\_e\_g5$MAF\_A,1-fcr\_e\_g5$MAF\_A)#To take into accountt thatt plink somettimes switches the min og maj  
a<-c("CHR","SNP","Gene","Location","POS","NMISS","FST","Z\_FST","#Uploaded\_variation","Allele","Consequence","Amino\_acids","Codons","Extra","external\_gene\_name","A1","MAF\_A","MAF\_U","r\_MAF\_A")#Relevante kolonner

#Make tables:  
df %>% dplyr::select(-Feature) %>% subset(filter options) %>% distinct(Location,.keep\_all=TRUE)%>% dplyr::select(Consequence ,external\_gene\_name,Gene,Location,FST , Z\_FST,MAF\_A,MAF\_U,Codons,Extra,NCHROBS\_A,NCHROBS\_U)  
  
dplyr::select(Consequence ,external\_gene\_name,Gene,Location,A1,A2,NCHROBS\_A,FST , Z\_FST,MAF\_A,Codons,Extra)

[]

##### Figures

gg.manhattan\_chr  
function (df1, threshold, hlight, col, ylims, title)   
{  
 if (is.data.frame(df1) == FALSE) {  
 print("Input must be a data.frame")  
 }  
 else if (is.data.table(df1) == TRUE) {  
 print("Input must be data.frame, not data.table")  
 }  
 else {  
 if ((missing(threshold))) {  
 threshold <- 5  
 }  
 if ((missing(col))) {  
 col <- c("#5D82BB", "#3B64A5", "#1E4F9E", "#103B7E",   
 "#082B64")  
 }  
 if ((missing(title))) {  
 title <- "Unknown"  
 }  
 df <- subset(df1, Z\_FST > 0)  
 df.tmp <- df %>% group\_by(CHR) %>% summarise(chr\_len = max(POS)) %>%   
 mutate(tot = cumsum(as.numeric(chr\_len)) - chr\_len) %>%   
 dplyr::select(-chr\_len) %>% left\_join(df, ., by = c(CHR = "CHR")) %>%   
 arrange(CHR, POS) %>% mutate(BPcum = POS + tot) %>%   
 mutate(is\_annotate = ifelse(Z\_FST > threshold, "yes",   
 "no")) %>% mutate(is\_annotate1 = ifelse(Z\_FST >   
 5.5, "yes", "no"))  
 axisdf <- df.tmp %>% group\_by(CHR) %>% summarize(center = (max(BPcum) +   
 min(BPcum))/2)  
 ggplot(df.tmp, aes(x = BPcum, y = Z\_FST)) + geom\_point(aes(color = as.factor(CHR)),   
 alpha = 0.8, size = 2) + scale\_color\_manual(values = rep(col,   
 38)) + scale\_x\_continuous(label = axisdf$CHR, breaks = axisdf$center) +   
 scale\_y\_continuous(expand = c(0, 0), limits = ylims) +   
 ggtitle(paste0(title)) + labs(x = "Chromosome", y = expression(ZF["ST"])) +   
 geom\_hline(yintercept = 5, color = "orange", linetype = "dashed") +   
 geom\_hline(yintercept = 5.5, color = "orange") +   
 geom\_point(data = subset(df.tmp, is\_annotate == "yes"),   
 color = "orange", size = 2) + theme\_bw(base\_size = 22) +   
 theme(plot.title = element\_text(hjust = 0.5), legend.position = "none",   
 panel.border = element\_blank(), panel.grid.major.x = element\_blank(),   
 panel.grid.minor.x = element\_blank())  
 }  
}

#### All

module load bioinfo-tools bcftools/1.10 vcftools/0.1.15 htslib/1.10 plink/1.90b4.9 plink2/2.00-alpha-2-20190429  
plink \  
 --allow-extra-chr \  
 --allow-no-sex \  
 --bfile all\_by\_bcf\_use\_this \  
 --dog \  
 --double-id \  
 --fst case-control \  
 --geno 0.05 \  
 --keep phenos/pheno\_fcr\_vs\_{both,iDog,eriks}.txt \  
 --out fcr\_vs\_{both,iDog,eriks}\_fst\_g5percent.fst \  
 --pheno phenos/pheno\_fcr\_vs\_{both,iDog,eriks}.txt \  
 --within phenos/pheno\_fcr\_vs\_{both,iDog,eriks}.txt

### Annotering af SNPs

## Ostranders

Dogs that in file that could be found under bioproject, was included |[#paechDifferentialLigandActivation1997][][#qanbariMappingSignaturesPositive2014][]

# Bibliography