

InbreedingCoeff - Single-Threaded GenotypeGVCFs

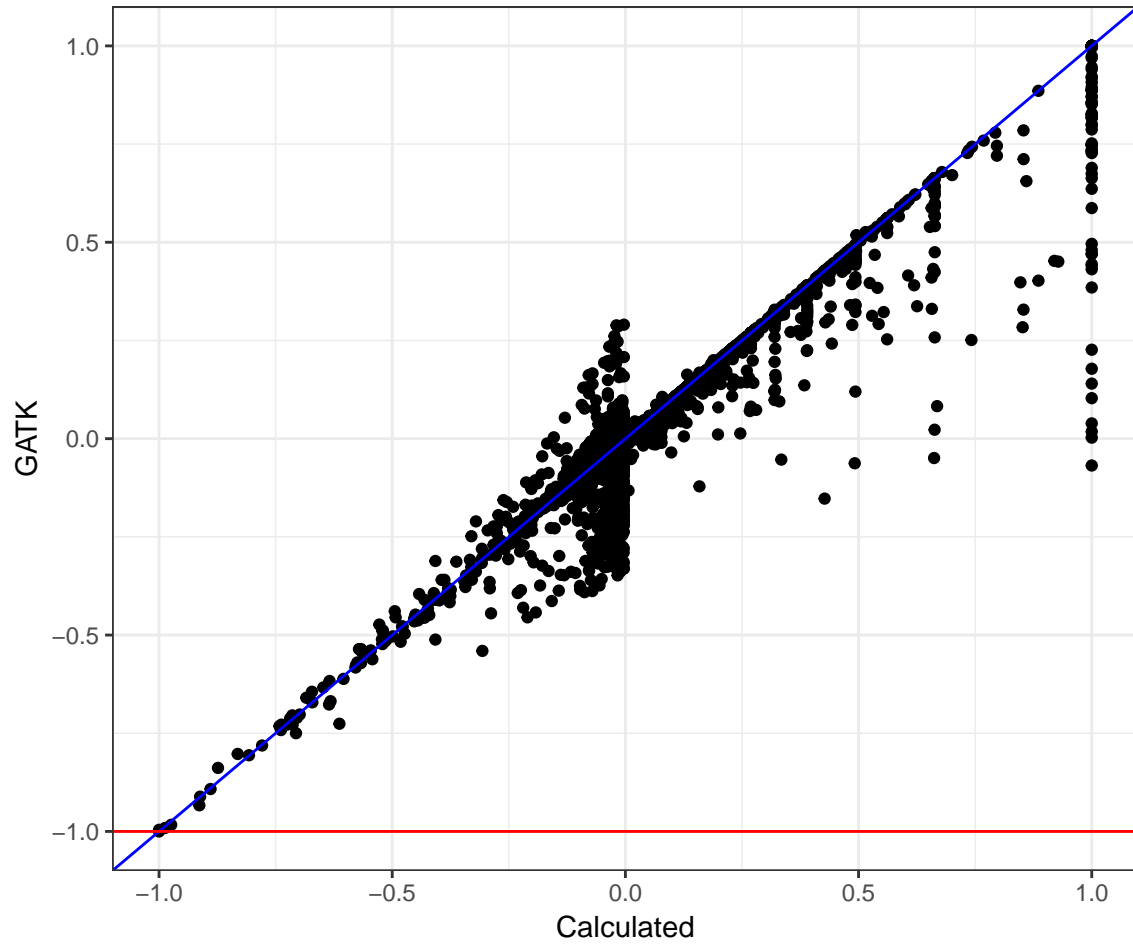
Camden Lopez

3/27/2018

```
library(dplyr)
library(ggplot2)
library(vcfR)
vcf <- read.vcfR("data/exome.bwa.gatk.singlethread.vcf.gz",
                verbose = FALSE)
inbreed <- as.double(extract.info(vcf, "InbreedingCoeff"))
check <- sample(1:length(inbreed), 10000)
gt <- t(apply(vcf@gt[check, -1], 1, function(x) substr(x, 1, 3)))
table(gt[, 1], useNA = "ifany")

##
## ./ 0/0 0/1 0/2 0/3 1/1 1/2 1/3 2/2 2/3 3/3
## 51 8553 827 52 10 483 10 2 6 5 1

IC <- data.frame(VCF_IDX = check,
                 ID = paste(vcf@fix[check, "CHROM"],
                             vcf@fix[check, "POS"],
                             sep = ":"))
IC$REF <- apply(gt, 1, function(x)
               sum(x == "0/0", na.rm = TRUE))
IC$HET <- apply(gt, 1, function(x)
               sum(x %in% paste(0, 1:3, sep = "/"), na.rm = TRUE))
IC$ALT <- apply(gt, 1, function(x)
               sum(x %in% c(paste(1, 1:3, sep = "/"),
                             paste(2, 2:3, sep = "/"),
                             paste(3, 3, sep = "/")), na.rm = TRUE))
IC$HET_EXP <- with(IC, {
  p <- (2 * REF + HET) / (2 * (REF + HET + ALT))
  q <- 1 - p
  2 * p * q * (REF + HET + ALT)
})
IC$GATK = inbreed[check]
IC$Calculated <- with(IC, 1 - HET / HET_EXP)
ggplot(IC) +
  geom_point(aes(x = Calculated, GATK)) +
  geom_abline(intercept = 0, slope = 1, color = "blue") +
  geom_hline(yintercept = -1, color = "red") +
  theme_bw()
```



```
arrange(IC, desc(abs(GATK - Calculated))) %>% head(5)
```

##	VCF_IDX	ID	REF	HET	ALT	HET_EXP	GATK	Calculated
## 1	282140	8:145315963	54	0	1	1.963636	-0.0683	1
## 2	238055	7:44365158	102	0	1	1.980583	0.0026	1
## 3	527559	19:14200674	123	0	1	1.983871	0.0180	1
## 4	278207	8:125834192	140	0	1	1.985816	0.0389	1
## 5	475414	16:81069449	1	0	11	1.833333	0.1031	1

```
write.csv(IC, "data/exome_inbreedcoeff_singlethread.csv",
          row.names = FALSE)
```