InbreedingCoeff - Multi-Threaded GenotypeGVCFs

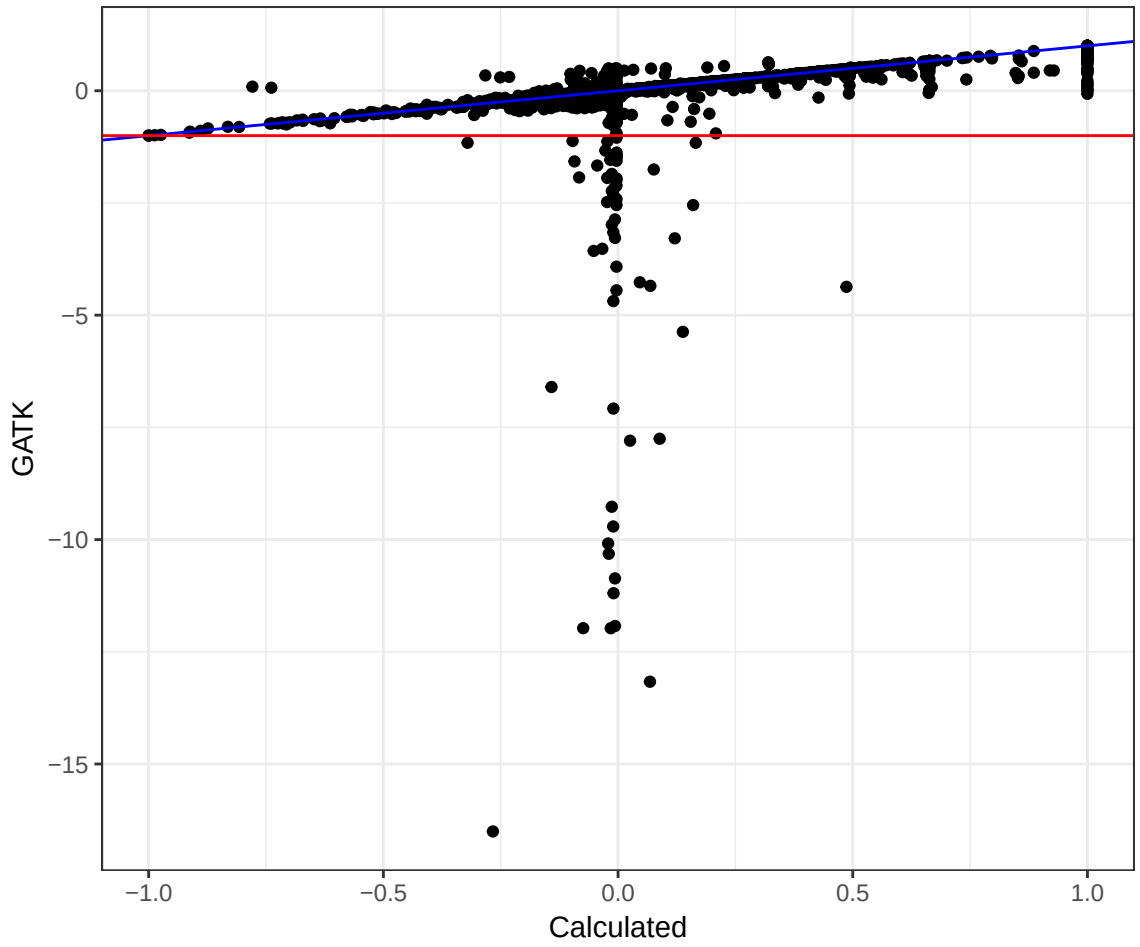
Camden Lopez

3/27/2018

```
library(dplyr)
library(ggplot2)
library(vcfR)
vcf <- read.vcfR("data/exome.bwa.gatk.multithread.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, GATK) %>% head(5)
```

```
##   VCF_IDX           ID REF HET ALT   HET_EXP     GATK   Calculated
## 1  227308  6:167591954  88  64   0 50.526316 -16.5028 -0.266666667
## 2  573050  22:19773358  98  46   8 49.355263 -13.1667  0.067981872
## 3  343310  11:32852167 127  25   1 24.617647 -11.9750 -0.015531661
## 4  216154   6:90437541  96  56   5 52.127389 -11.9741 -0.074291300
## 5  393486 12:107366663 152   2   0  1.987013 -11.9252 -0.006535948
```

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