MAGERI benchmark using reference standard DNA library

Mikhail Shugay

December 1, 2016

Load metadata

```
df.vmeta <- read.table("hd734_variant_metadata.txt", sep="\t", header=T) # variants observed in HD734 a
df.smeta <- read.table("sample_metadata.txt", sep="\t", header=T) # metadata for amplicon sequencing sa
```

VCF parsing function

```
library(stringr)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(ggplot2)
library(ggbeeswarm)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
read_vcf <- function(file_name) {
  .vcf <- read.table(file_name, header = F, sep = "\t", stringsAsFactors = F)
  colnames(.vcf) <- c("chromosome", "position", "skip1", "from", "to", "qual", "skip2", "info", "skip3"
  .vcf$skip1 <- NULL
  .vcf$skip2 <- NULL
  .vcf$skip3 <- NULL
  .vcf$skip4 <- NULL
```

```
  .vcf <- subset(.vcf, nchar(from) != nchar(to)) # indels only

  if (nrow(.vcf) == 0) {
    return(data.frame())
  }

  .infosplit <- str_split_fixed(.vcf$info, regex("[=;]"), 5)[,c(2, 4)]

  if(is.null(dim(.infosplit))) {
    return(data.frame())
  }

  .vcf$coverage <- as.numeric(.infosplit[,1])
  .vcf$frequency <- as.numeric(.infosplit[,2])
  .vcf$info <- NULL

  .vcf$count <- as.integer(round(.vcf$coverage * .vcf$frequency))

  # the variant below is read boundary/alignment artefact
  # always inspect your indels manually using SAM files when running amplicon-seq datasets:
  # - UMIs don't guarantee indel-proof data
  # - and there is no scoring/filtering of indels in MAGERI
  subset(.vcf, !(position == 212578380 & chromosome == "chr2"))
}
```

Read samples with HD734 standard DNA and control human DNA, append metadata

```
df <- data.frame()

read_vcf_with_metadata <- function(file_name, primer_set, replica, ratio, type) {
  .vcf <- read_vcf(file_name)

  .vcf <- merge(.vcf, df.vmeta, all.x = type != "standard", all.y = F)

  if (nrow(.vcf) == 0) {
    return(data.frame())
  }

  .vcf$known.frequency <- .vcf$known.frequency * ratio
  .vcf$known.frequency[is.na(.vcf$known.frequency)] <- 0
  .vcf$primer_set <- primer_set
  .vcf$replica <- primer_set
  .vcf$type <- type

  .vcf
}

for (i in 1:nrow(df.smeta)) {
  .df <- with(df.smeta, read_vcf_with_metadata(paste(prefix[i], "vcf", sep="."),
                                               primer_set[i],
                                               replica[i],
                                               ratio[i],
                                               type[i]))
```

```
  df <- rbind(df, .df)
}
```

Some grooming

```
df$bases.diff <- abs(nchar(df$from) - nchar(df$to))
df$type <- with(df, ifelse(nchar(from) > nchar(to), "deletion", "insertion"))
df$text <- ifelse(is.na(df$id), NA, paste(df$id, "@", df$known.frequency*100, "%", sep=""))
```

Indel rate and known EGFR variant

```
ggplot(df, aes(x=bases.diff, y=frequency, fill=text)) +
  geom_point(shape=21,size=3) +
  scale_y_log10("Variant frequency") +
  xlab("Indel size") +
  scale_fill_brewer("Variant", palette = "Set1") +
  theme_bw()
```