clinvar_compare_df

Summer Elasady

Friday, April 24, 2015

Comparing impala and Brady's results by data frame

To compare every record of Brady's clinvar results with the results of the impala query, two matching data frames were created for comparison across all values.

```
library(readr)

##read in brady and impala query results
brady = read_tsv("~/GitHub/impala_scripts/queries/testing/clinvar/brady_clinvar.txt")
```

```
## Warning: 4043 problems parsing
## '~/GitHub/impala_scripts/queries/testing/clinvar/brady_clinvar.txt'. See
## problems(...) for more details.
```

```
impala = read_csv("~/GitHub/impala_scripts/queries/testing/clinvar/impala_query_results.csv")

##make data frames with matching variables
brady.df = data.frame(chr = gsub("chr", "", lapply(strsplit(as.character(brady$position), ":"), functio
                      pos = unlist(lapply(strsplit(as.character(brady$position), ":"), function(x) x[2]
                      ref = unlist(lapply(strsplit(as.character(brady$dna_change), "->"), function(x) x
                      alt = unlist(lapply(strsplit(as.character(brady$dna_change), "->"), function(x) x
                      zygosity = gsub("homozygous", "hom", brady$zygosity),
                      gene = unlist(lapply(strsplit(brady$gene_definition, ':'), function(x) x[1])),
                      sample_id = gsub(".*:","",brady$identifier_or_consent),
                      clin_sig = brady$clinvar_pathogenicity)

impala.df = data.frame(chr = as.character(impala$chr),
                       pos = as.character(impala$start),
                       ref = impala$ref,
                       alt = impala$allele1seq,
                       zygosity = impala$zygosity,
                       gene = unlist(lapply(strsplit(impala$clin_geneinfo, ':'), function(x) x[1])),
                       sample_id = impala$sample_id,
                       clin_sig = as.character(impala$clin_sigid))

##coerce all values to character for matching
#coercing columns to same class for comparison
i = sapply(brady.df, is.factor)
brady.df[i] = lapply(brady.df[i], as.character)

j = sapply(impala.df, is.factor)
impala.df[j] = lapply(impala.df[j], as.character)

##order data frames by sample id, chr, pos for matching
#order both data frames for matching
brady.df = brady.df[with(brady.df, order(sample_id, chr, pos)),]
```

```
impala.df = impala.df[with(impala.df, order(sample_id, chr, pos)),]

##subest brady's results to match with impala query
pathogenic = c("4", "5")
not_pathogenic = c("2", "3")
brady_filter = brady.df[which(brady.df$chr == "8" & brady.df$zygosity == "hom"
                          &  grep(paste(pathogenic,collapse="|"), brady.df$clin_sig) &
                            grep(paste(not_pathogenic,collapse="|"), brady.df$clin_sig, invert=TRUE)),]
```

```
## Warning in brady.df$chr == "8" & brady.df$zygosity == "hom" &
## grep(paste(pathogenic, : longer object length is not a multiple of shorter
## object length
```

```
##clear rownames for matching
rownames(impala.df) = NULL
rownames(brady_filter) = NULL

##since the impala set does not include piping, and all the pipes from Brady's results are "5|5" gsub w
##yes I realize this is a bit of cheating
brady_filter$clin_sig = as.character("5")
```

Once data frames were created from each result with matching column names and class type structures, dplyr was be used to search for any differences in the data frames. All rows not returned in the final output mean that the entries are identical.

```
require(dplyr)
```

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#records in impala not in brady's results
not_in_brady = unique(anti_join(impala.df,brady_filter))
```

```
## Joining by: c("chr", "pos", "ref", "alt", "zygosity", "gene", "sample_id", "clin_sig")
```

```
dim(not_in_brady)
```

```
## [1] 7 8
```

```
not_in_brady
```

```
##   chr       pos ref alt zygosity    gene sample_id clin_sig
## 1   8  67091182   G   T      hom     CRH 101-180-M        5
## 2   8  67091182   G   T      hom     CRH 101-445-M        5
## 3   8  67091182   G   T      hom     CRH 101-728-F        5
## 4   8  67091182   G   T      hom     CRH 101-270-F        5
## 5   8 143994266   A   G      hom CYP11B2 101-506-M        5
## 7   8  67091182   G   T      hom     CRH 101-525-M        5
## 8   8 143994266   A   G      hom CYP11B2 101-878-M        5
```

```
##records in brady's results but not in impala
not_in_impala = unique(anti_join(brady_filter,impala.df))
```

```
## Joining by: c("chr", "pos", "ref", "alt", "zygosity", "gene", "sample_id", "clin_sig")
```

```
dim(not_in_impala)
```

```
## [1] 4 8
```

```
not_in_impala
```

```
##    chr       pos ref alt zygosity    gene sample_id clin_sig
## 1    8 143994266   A   G      hom CYP11B2 101-876-M        5
## 2    8  18080001   G   A      hom    NAT1 101-875-F        5
## 11   8 143994266   A   G      hom CYP11B2 101-927-F        5
## 12   8 143994266   A   G      hom CYP11B2 101-803-M        5
```

These results are identical with the previous analysis.

How do these results compare to similar results between data sets?

To answer the question of how the results compare across data sets, results similar to each of the differences were extracted from each data set.

Results not found in Brady's set

```
##Sample ID
brady_filter[which(brady_filter$sample_id == "101-180-M"),]
```

Sample 101-180-M and all results for CRH:

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

```
impala.df[which(impala.df$sample_id == "101-180-M"),]
```

```
##   chr      pos ref alt zygosity gene sample_id clin_sig
## 6   8 67091182   G   T      hom  CRH 101-180-M        5
```

```
brady.df[which(brady.df$sample_id == "101-180-M"),]
```

```
##       chr       pos ref alt zygosity  gene sample_id clin_sig
## 10137   9 132580901   C   G      hom TOR1A 101-180-M    255|5
```

This sample ID only has a result on chromosome 9 in Brady's set. Let's see if there are any results for the CRH gene:

```
##Gene
brady_filter[which(brady_filter$gene == "CRH"),] #not in filtered set
```

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

```
brady.df[which(brady.df$gene == "CRH"),]
```

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

The CRH gene does not appear in Brady's results. Let's look for anything that might be in that gene region.

```
##Region
brady_filter[which(brady_filter$chr == "8" & brady_filter$pos =="67091182" ),] #not in filtered set
```

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

```
brady.df[which(brady.df$chr == "8" & (brady.df$pos >= "67091000" & brady.df$pos <= "67091300")),] #noth
```

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

Nothing from Brady's results set falls in this region.

Sample 101-506-M and all results with CYP11B2 Let's look at hits for sample ID 101-506-M

```
##Sample ID
brady_filter[which(brady_filter$sample_id == "101-506-M"),] #not in filtered set
```

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

```
impala.df[which(impala.df$sample_id == "101-506-M"),]
```

```
##    chr       pos ref alt zygosity    gene sample_id clin_sig
## 37   8 143994266   A   G      hom CYP11B2 101-506-M        5
## 38   8 143994266   A   G      hom CYP11B2 101-506-M        5
```

```
brady.df[which(brady.df$sample_id == "101-506-M"),]
```

```
## [1] chr       pos       ref       alt       zygosity  gene      sample_id
## [8] clin_sig
## <0 rows> (or 0-length row.names)
```

```
##sample ID not in Brady's results
```

This sample ID does not appear in Brady's results. Let's look for the CYP11B2 gene:

```
##Gene
brady_filter[which(brady_filter$gene == "CYP11B2"),]
```

```
##    chr       pos ref alt zygosity    gene sample_id clin_sig
## 1    8 143994266   A   G      hom CYP11B2 101-012-M        5
## 4    8 143994266   A   G      hom CYP11B2 101-049-F        5
## 7    8 143994266   A   G      hom CYP11B2 101-191-F        5
## 8    8 143994266   A   G      hom CYP11B2 101-253-F        5
## 9    8 143994266   A   G      hom CYP11B2 101-259-F        5
## 11   8 143994266   A   G      hom CYP11B2 101-267-F        5
## 19   8 143994266   A   G      hom CYP11B2 101-354-F        5
## 22   8 143994266   A   G      hom CYP11B2 101-367-F        5
## 23   8 143994266   A   G      hom CYP11B2 101-408-F        5
## 24   8 143994266   A   G      hom CYP11B2 101-432-F        5
## 25   8 143994266   A   G      hom CYP11B2 101-436-F        5
## 28   8 143994266   A   G      hom CYP11B2 101-492-F        5
## 31   8 143994266   A   G      hom CYP11B2 101-585-M        5
## 34   8 143994266   A   G      hom CYP11B2 101-589-M        5
## 37   8 143994266   A   G      hom CYP11B2 101-627-M        5
## 44   8 143994266   A   G      hom CYP11B2 101-793-F        5
## 45   8 143994266   A   G      hom CYP11B2 101-803-M        5
## 57   8 143994266   A   G      hom CYP11B2 101-876-M        5
## 58   8 143994266   A   G      hom CYP11B2 101-927-F        5
```

```
brady.df[which(brady.df$gene == "CYP11B2"),]
```

```
##       chr       pos ref alt zygosity    gene sample_id clin_sig
## 10061   8 143994266   A   G      hom CYP11B2 101-012-M      5|5
## 10048   8 143994266   A   G      hom CYP11B2 101-049-F      5|5
## 10049   8 143994266   A   G      hom CYP11B2 101-191-F      5|5
## 10050   8 143994266   A   G      hom CYP11B2 101-253-F      5|5
## 10051   8 143994266   A   G      hom CYP11B2 101-259-F      5|5
## 10052   8 143994266   A   G      hom CYP11B2 101-267-F      5|5
## 10053   8 143994266   A   G      hom CYP11B2 101-354-F      5|5
```

```
## 10054   8 143994266   A   G      hom CYP11B2 101-367-F      5|5
## 10055   8 143994266   A   G      hom CYP11B2 101-408-F      5|5
## 10056   8 143994266   A   G      hom CYP11B2 101-432-F      5|5
## 10057   8 143994266   A   G      hom CYP11B2 101-436-F      5|5
## 10058   8 143994266   A   G      hom CYP11B2 101-492-F      5|5
## 10062   8 143994266   A   G      hom CYP11B2 101-585-M      5|5
## 10063   8 143994266   A   G      hom CYP11B2 101-589-M      5|5
## 10064   8 143994266   A   G      hom CYP11B2 101-627-M      5|5
## 10059   8 143994266   A   G      hom CYP11B2 101-793-F      5|5
## 10065   8 143994266   A   G      hom CYP11B2 101-803-M      5|5
## 10066   8 143994266   A   G      hom CYP11B2 101-876-M      5|5
## 10060   8 143994266   A   G      hom CYP11B2 101-927-F      5|5
```

```
##this gene is in Brady's results with matching coords, but that sample is not included
```

This gene appears in Brady's results, but since that sample ID is missing, this hit is not a match. Let's check for anything in that region:

```
##Region
brady_filter[which(brady_filter$chr == "8" & brady_filter$pos =="143994266" ),]
```

```
##    chr       pos ref alt zygosity    gene sample_id clin_sig
## 1    8 143994266   A   G      hom CYP11B2 101-012-M        5
## 4    8 143994266   A   G      hom CYP11B2 101-049-F        5
## 7    8 143994266   A   G      hom CYP11B2 101-191-F        5
## 8    8 143994266   A   G      hom CYP11B2 101-253-F        5
## 9    8 143994266   A   G      hom CYP11B2 101-259-F        5
## 11   8 143994266   A   G      hom CYP11B2 101-267-F        5
## 19   8 143994266   A   G      hom CYP11B2 101-354-F        5
## 22   8 143994266   A   G      hom CYP11B2 101-367-F        5
## 23   8 143994266   A   G      hom CYP11B2 101-408-F        5
## 24   8 143994266   A   G      hom CYP11B2 101-432-F        5
## 25   8 143994266   A   G      hom CYP11B2 101-436-F        5
## 28   8 143994266   A   G      hom CYP11B2 101-492-F        5
## 31   8 143994266   A   G      hom CYP11B2 101-585-M        5
## 34   8 143994266   A   G      hom CYP11B2 101-589-M        5
## 37   8 143994266   A   G      hom CYP11B2 101-627-M        5
## 44   8 143994266   A   G      hom CYP11B2 101-793-F        5
## 45   8 143994266   A   G      hom CYP11B2 101-803-M        5
## 57   8 143994266   A   G      hom CYP11B2 101-876-M        5
## 58   8 143994266   A   G      hom CYP11B2 101-927-F        5
```

```
brady.df[which(brady.df$chr == "8" & (brady.df$pos >= "143994100" & brady.df$pos <= "143994300")),] #no
```

```
##       chr       pos ref alt zygosity    gene sample_id clin_sig
## 10061   8 143994266   A   G      hom CYP11B2 101-012-M      5|5
## 10048   8 143994266   A   G      hom CYP11B2 101-049-F      5|5
## 10049   8 143994266   A   G      hom CYP11B2 101-191-F      5|5
## 10050   8 143994266   A   G      hom CYP11B2 101-253-F      5|5
## 10051   8 143994266   A   G      hom CYP11B2 101-259-F      5|5
## 10052   8 143994266   A   G      hom CYP11B2 101-267-F      5|5
## 10053   8 143994266   A   G      hom CYP11B2 101-354-F      5|5
```

```
## 10054   8 143994266   A   G      hom CYP11B2 101-367-F      5|5
## 10055   8 143994266   A   G      hom CYP11B2 101-408-F      5|5
## 10056   8 143994266   A   G      hom CYP11B2 101-432-F      5|5
## 10057   8 143994266   A   G      hom CYP11B2 101-436-F      5|5
## 10058   8 143994266   A   G      hom CYP11B2 101-492-F      5|5
## 10062   8 143994266   A   G      hom CYP11B2 101-585-M      5|5
## 10063   8 143994266   A   G      hom CYP11B2 101-589-M      5|5
## 10064   8 143994266   A   G      hom CYP11B2 101-627-M      5|5
## 10059   8 143994266   A   G      hom CYP11B2 101-793-F      5|5
## 10065   8 143994266   A   G      hom CYP11B2 101-803-M      5|5
## 10066   8 143994266   A   G      hom CYP11B2 101-876-M      5|5
## 10060   8 143994266   A   G      hom CYP11B2 101-927-F      5|5
```

This gene is the only one that hits in this region.