Clinvar_Testing

Impala Query to locate clinvar variants

The following query was run on impala, returning 57 rows:

```
WITH clinvar AS
(SELECT clin.chromosome as clin_chrom, clin.pos as clin_pos, clin.ref as clin_ref, clin.alt AS clin_alt
 clin.rs AS rsID, clin.geneinfo as clin_geneinfo, clin.clnsig AS clin_sigid, clin.clnhgvs as clin_hgvs,
 clnsum.type as clin_vartype, clnsum.clinicalsignificance as clin_sig, clin.clndbn as clin_clindbn,
 clnsum.cytogenetic as clin_cytogenetic, clnsum.guidelines as clin_guidelines
 FROM public_hg19.clinvar_summary as clnsum
 JOIN public_hg19.clinvar AS clin
 ON clin.rs = clnsum.rsnum_dbsnp
 WHERE ((FIND_IN_SET('4', clin.clnsig) > 0 OR FIND_IN_SET('5', clin.clnsig)>0)
       AND
       (FIND_IN_SET('3', clin.clnsig)=0 AND FIND_IN_SET('2', clin.clnsig)=0)
       )
 AND clnsum.assembly = "GRCh37"
 )

SELECT *
FROM p7_ptb.comgen_variant AS comgen, clinvar
WHERE comgen.chr = "8"
AND comgen.zygosity = "hom"
AND (comgen.sample_id LIKE "%F" OR comgen.sample_id LIKE "%M")
AND clinvar.clin_chrom = comgen.chr
AND comgen.ref = clinvar.clin_ref
AND comgen.allele1seq = clinvar.clin_alt
AND clinvar.clin_pos = comgen.start
```

Comparing Results with Brady's Pipeline

The results were read into R:

```
library(readr)
impala = read_csv("~/impala_scripts/queries/testing/clinvar/query_result.csv")
head(impala)
```

```
##       start      stop zygosity vartype ref allele1seq allele2seq
## 1 143994266 143994267      hom     snp   A          G          G
## 2 143994266 143994267      hom     snp   A          G          G
## 3  19813529  19813530      hom     snp   A          G          G
## 4  21976710  21976711      hom     snp   T          C          C
## 5  67091182  67091183      hom     snp   G          T          T
## 6  21976710  21976711      hom     snp   T          C          C
##   allele1varquality allele2varquality totalreadcount sample_id chr
## 1            VQHIGH            VQHIGH             29 101-589-M   8
## 2            VQHIGH            VQHIGH             29 101-589-M   8
## 3            VQHIGH            VQHIGH             68 101-264-M   8
## 4            VQHIGH            VQHIGH             22 101-013-F   8
```

```
## 5            VQHIGH            VQHIGH             54 101-180-M   8
## 6            VQHIGH            VQHIGH             20 101-190-M   8
##   clin_chrom  clin_pos clin_ref clin_alt     rsid clin_geneinfo clin_sigid
## 1          8 143994266        A        G 61757294  CYP11B2:1585          5
## 2          8 143994266        A        G 61757294  CYP11B2:1585          5
## 3          8  19813529        A        G      268      LPL:4023          5
## 4          8  21976710        T        C  7014851      HR:55806          5
## 5          8  67091182        G        T 12721510      CRH:1392          5
## 6          8  21976710        T        C  7014851      HR:55806          5
##                      clin_hgvs              clin_vartype   clin_sig
## 1 NC_000008.10:g.143994266A>G single nucleotide variant Pathogenic
## 2 NC_000008.10:g.143994266A>G single nucleotide variant Pathogenic
## 3  NC_000008.10:g.19813529A>G single nucleotide variant Pathogenic
## 4  NC_000008.10:g.21976710T>C single nucleotide variant Pathogenic
## 5  NC_000008.10:g.67091182G>T single nucleotide variant Pathogenic
## 6  NC_000008.10:g.21976710T>C single nucleotide variant Pathogenic
##                                                                                                clin_clindbn
## 1 Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
## 2 Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
## 3                                                                  Hyperlipidemia\\x2c_familial_combined
## 4                                                                         Alopecia_universalis_congenita
## 5                                                Autosomal_dominant_nocturnal_frontal_lobe_epilepsy
## 6                                                                         Alopecia_universalis_congenita
##   clin_cytogenetic clin_guidelines
## 1           8q24.3            NULL
## 2           8q24.3            NULL
## 3           8p21.3            NULL
## 4           8p21.3            NULL
## 5             8q13            NULL
## 6           8p21.3            NULL
```

```
dim(impala)
```

```
## [1] 57 25
```

```
impala[grep("3", impala$clin_sigid),]
```

```
##  [1] start             stop              zygosity
##  [4] vartype           ref               allele1seq
##  [7] allele2seq        allele1varquality allele2varquality
## [10] totalreadcount    sample_id         chr
## [13] clin_chrom        clin_pos          clin_ref
## [16] clin_alt          rsid              clin_geneinfo
## [19] clin_sigid        clin_hgvs         clin_vartype
## [22] clin_sig          clin_clindbn      clin_cytogenetic
## [25] clin_guidelines
## <0 rows> (or 0-length row.names)
```

The results of Brady's pipeline were read into R for comparison:

```
library(readxl)
brady = read_excel("/Users/selasady/impala_scripts/queries/testing/clinvar/clinvar_all_itmi_adult_hits_
head(brady)
```

```
##   gene_name clinvar_pathogenicity                            phenotype
## 1    FCGR3B                    5 Neutrophil-specific_antigens_na1/na2
## 2    FCGR3B                    5 Neutrophil-specific_antigens_na1/na2
## 3    FCGR3B                    5 Neutrophil-specific_antigens_na1/na2
## 4    FCGR3B                    5 Neutrophil-specific_antigens_na1/na2
## 5    FCGR3B                    5 Neutrophil-specific_antigens_na1/na2
## 6    FCGR3B                    5 Neutrophil-specific_antigens_na1/na2
##       gene_definition  chr       pos identifier_or_consent coding_change
## 1 FCGR3B:NM_001244753 chr1 161599643             101-009-F  SUBSTITUTION
## 2    FCGR3B:NM_000570 chr1 161599643             101-009-F  SUBSTITUTION
## 3 FCGR3B:NM_001244753 chr1 161599693             101-009-F  SUBSTITUTION
## 4    FCGR3B:NM_000570 chr1 161599693             101-009-F  SUBSTITUTION
## 5 FCGR3B:NM_001244753 chr1 161599779             101-009-F  SUBSTITUTION
## 6    FCGR3B:NM_000570 chr1 161599779             101-009-F  SUBSTITUTION
##   protein_definition amino_acid_position amino_acid_change   zygosity
## 1       NP_001231682                118              N->D homozygous
## 2          NP_000561                 82              N->D homozygous
## 3       NP_001231682                101              N->S homozygous
## 4          NP_000561                 65              N->S homozygous
## 5       NP_001231682                 72              S->R homozygous
## 6          NP_000561                 36              S->R homozygous
##   variant_call reference_major_or_minor data_set_minor_allele_fraction
## 1          1/1                ref_major                         0.38239
## 2          1/1                ref_major                         0.38239
## 3          0/0                ref_minor                         0.39696
## 4          0/0                ref_minor                         0.39696
## 5          1/1                ref_major                         0.38602
## 6          1/1                ref_major                         0.38602
##   kaviar_minor_allele_fraction
## 1                       0.0117
## 2                       0.0117
## 3                       0.0082
## 4                       0.0082
## 5                       0.0031
## 6                       0.0031
```

Brady's result set was subset to match the parameters of the impala query:
- chromsome 8
- Mothers or Fathers
- clinical significance of 4 or 5
- homozygous

```
##first we'll add a chromosome column for subsetting by chromosome 8
brady$chrom = gsub("chr", "", brady$chr)

##changing "-" to "_" to avoid breaking R
brady$identifier_or_consent = gsub("-", "_", brady$identifier_or_consent)
impala$sample_id = gsub("-", "_", impala$sample_id)

##next we'll add a subject type for subsetting for Mother and Father
brady$subj_type = lapply(strsplit(as.character(brady$identifier_or_consent), "_"), function(x) x[3])
##checkibrng if entries are already subset for M and F
levels(as.factor(unlist(brady$subj_type)))
```

```
## [1] "F" "M"
```

```
##now we'll subset by choromosome and zygosity
chr8_hom = brady[which(brady$chrom == "8" & brady$zygosity == "homozygous"),]

#subset for clnsig that was rated 4 or 5 but not 2 or 3
pathogenic = c("4", "5")
not_pathogenic = c("2", "3")

patho = chr8_hom[(
  grep(paste(pathogenic,collapse="|"), chr8_hom$clinvar_pathogenicity) &
      grep(paste(not_pathogenic,collapse="|"), chr8_hom$clinvar_pathogenicity, invert=TRUE)),]

#lets make sure data looks normal
head(patho)
```

```
##      gene_name clinvar_pathogenicity
## 76     CYP11B2                   5|5
## 89          HR                     5
## 90          HR                     5
## 482    CYP11B2                   5|5
## 2464        HR                     5
## 2465        HR                     5
##                                                                                                  phenotype
## 76   Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
## 89                                                                                Alopecia_universalis_congenita
## 90                                                                                Alopecia_universalis_congenita
## 482  Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
## 2464                                                                              Alopecia_universalis_congenita
## 2465                                                                              Alopecia_universalis_congenita
##        gene_definition  chr       pos identifier_or_consent coding_change
## 76   CYP11B2:NM_000498 chr8 143994266             101_012_M  SUBSTITUTION
## 89        HR:NM_005144 chr8  21976710             101_013_F  SUBSTITUTION
## 90        HR:NM_018411 chr8  21976710             101_013_F  SUBSTITUTION
## 482  CYP11B2:NM_000498 chr8 143994266             101_049_F  SUBSTITUTION
## 2464      HR:NM_005144 chr8  21976710             101_190_M  SUBSTITUTION
## 2465      HR:NM_018411 chr8  21976710             101_190_M  SUBSTITUTION
##      protein_definition amino_acid_position amino_acid_change   zygosity
## 76              P19099                 386              V->A homozygous
## 89              O43593                1022              T->A homozygous
## 90            O43593-2                1022              T->A homozygous
## 482             P19099                 386              V->A homozygous
## 2464            O43593                1022              T->A homozygous
## 2465          O43593-2                1022              T->A homozygous
##      variant_call reference_major_or_minor data_set_minor_allele_fraction
## 76            1/1               ref_major                         0.04937
## 89            1/1               ref_major                         0.04326
## 90            1/1               ref_major                         0.04326
## 482           1/1               ref_major                         0.04937
## 2464          1/1               ref_major                         0.04326
## 2465          1/1               ref_major                         0.04326
##      kaviar_minor_allele_fraction chrom subj_type
## 76                         0.0678     8         M
## 89                          0.092     8         F
```

```
## 90                          0.092     8         F
## 482                        0.0678     8         F
## 2464                        0.092     8         M
## 2465                        0.092     8         M
```

```
#how many results did we get back?
dim(patho)
```

```
## [1] 58 18
```

Analyzing differences in results

The impala results give us back 57 rows, Brady's pipeline gives us back 58. Let's examine the results:

```
##brady's clinically significant results
unique(patho$gene_name)
```

Unique gene names returned

```
## [1] "CYP11B2" "HR"      "LPL"     "GDF6"    "ZFPM2"   "RP1"     "NAT1"
```

```
##impala results
unique(impala$clin_geneinfo)
```

```
## [1] "CYP11B2:1585" "LPL:4023"     "HR:55806"     "CRH:1392"
## [5] "RP1:6101"     "GDF6:392255"  "ZFPM2:23414"
```

The genes returned are identical except that Brady's set contains NAT1 and the impala set contains CRH.

```
##brady's clinically significant results
brady_samples = as.character(gsub(".*:", "", patho$identifier_or_consent))
brady_samples = data.frame(sample_id = as.character(brady_samples[order(brady_samples)]))

##impala results
impala_samples = impala$sample_id
impala_samples = data.frame(sample_id = as.character(impala_samples[order(impala_samples)]))

#compare results unique to each set
data.frame(brady_only=c(unique(as.character(brady_samples[!(brady_samples$sample_id %in% impala_samples$
           impala_only=unique(as.character(impala_samples[!(impala_samples$sample_id %in% brady_samples$
```

Sample ID's returned

```
##   brady_only impala_only
## 1  101_803_M   101_180_M
## 2  101_875_F   101_270_F
## 3  101_876_M   101_445_M
## 4  101_927_F   101_506_M
## 5         NA   101_525_M
## 6         NA   101_728_F
## 7         NA   101_878_M
```

Results in Brady's set not in imapala

All sample ID's that were found in Brady's study but not impala were determined to be missing from the comgen_variant table. Joe is investigating the reason for missing sample ID's.

Examine impala results not included in Brady's analysis

Now let's look at results from impala that were not included in Brady's analysis.

Sample 101-180-M Viewing impala annotation:

```
impala[grep("101_180_M", impala$sample_id),]
```

```
##      start     stop zygosity vartype ref allele1seq allele2seq
## 5 67091182 67091183      hom     snp   G          T          T
##   allele1varquality allele2varquality totalreadcount sample_id chr
## 5            VQHIGH            VQHIGH             54 101_180_M   8
##   clin_chrom clin_pos clin_ref clin_alt     rsid clin_geneinfo clin_sigid
## 5          8 67091182        G        T 12721510      CRH:1392          5
##                     clin_hgvs              clin_vartype   clin_sig
## 5 NC_000008.10:g.67091182G>T single nucleotide variant Pathogenic
##                                        clin_clindbn clin_cytogenetic
## 5 Autosomal_dominant_nocturnal_frontal_lobe_epilepsy             8q13
##   clin_guidelines
## 5            NULL
```

This query was run on impala to make sure the comgen variant info is correct:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.chr = "8"
AND comgen.sample_id = "101-180-M"
AND comgen.start = 67091182
```

And the results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQHIGH
- totalreadcount = 54

The information from impala matches with ClinVar:
http://www.ncbi.nlm.nih.gov/clinvar/variation/38800/

This variant is marked as pathogenic, with only one entry and a rank of 5.

Searching for this gene id in impala gives consistent results:

```
SELECT *
FROM public_hg19.clinvar as clin
WHERE clin.geneinfo = "CRH:1392"
```

This sample ID appears only once in Brady's results with a hit on chromosome 9.

Sample ID 101-270-F Viewing impala annotaiton

```
impala[grep("101_270_F", impala$sample_id),]
```

```
##       start     stop zygosity vartype ref allele1seq allele2seq
## 46 67091182 67091183      hom     snp   G          T          T
##    allele1varquality allele2varquality totalreadcount sample_id chr
## 46            VQHIGH            VQHIGH             67 101_270_F   8
##    clin_chrom clin_pos clin_ref clin_alt     rsid clin_geneinfo clin_sigid
## 46          8 67091182        G        T 12721510      CRH:1392          5
##                     clin_hgvs              clin_vartype   clin_sig
## 46 NC_000008.10:g.67091182G>T single nucleotide variant Pathogenic
##                                        clin_clindbn clin_cytogenetic
## 46 Autosomal_dominant_nocturnal_frontal_lobe_epilepsy             8q13
##    clin_guidelines
## 46            NULL
```

Infomation shown is consistent with info found on clinvar: http://www.ncbi.nlm.nih.gov/clinvar/variation/38800/

Pulling up this info in impala:

```
SELECT *
FROM public_hg19.clinvar as clin
WHERE clin.id = 'rs12721510'
```

The information matches clinvar info on impala and is marked as pathogenic by one source with a rating of 5.

This sample ID does not appear in Brady's results for chromosome 8, but does appear on other chromosomes. Verifying that the entry does appear in comgen_variant:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.sample_id = "101-270-F"
AND comgen.chr = "8"
AND comgen.start = 67091182
```

Results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQHIGH
- totalreadcount = 67

Sample ID 101-445-M Viewing impala notation:

```
impala[grep("101_445_M", impala$sample_id),]
```

```
##       start     stop zygosity vartype ref allele1seq allele2seq
## 55 67091182 67091183      hom     snp   G          T          T
##    allele1varquality allele2varquality totalreadcount sample_id chr
## 55            VQHIGH            VQHIGH             40 101_445_M   8
##    clin_chrom clin_pos clin_ref clin_alt     rsid clin_geneinfo clin_sigid
## 55          8 67091182        G        T 12721510      CRH:1392          5
##                     clin_hgvs              clin_vartype   clin_sig
## 55 NC_000008.10:g.67091182G>T single nucleotide variant Pathogenic
##                                        clin_clindbn clin_cytogenetic
## 55 Autosomal_dominant_nocturnal_frontal_lobe_epilepsy             8q13
##    clin_guidelines
## 55            NULL
```

Results are consistent with clinvar:
http://www.ncbi.nlm.nih.gov/clinvar/variation/38800/

Entry is listed as pathogenic by 1 submission with a rating of 5.

This sample ID appears in Brady's results on chromosome 5, but not on chromsome 8.

Verifying this entry in impala:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.sample_id = "101-445-M"
AND comgen.chr = "8"
AND comgen.start = 67091182
```

Results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQHIGH
- totalreadcount = 40

Sample ID 101-506-M Viewing impala notation:

```
impala[grep("101_506_M", impala$sample_id),]
```

```
##        start      stop zygosity vartype ref allele1seq allele2seq
## 23 143994266 143994267      hom     snp   A          G          G
## 24 143994266 143994267      hom     snp   A          G          G
##    allele1varquality allele2varquality totalreadcount sample_id chr
## 23             VQLOW            VQHIGH             20 101_506_M   8
## 24             VQLOW            VQHIGH             20 101_506_M   8
##    clin_chrom  clin_pos clin_ref clin_alt     rsid clin_geneinfo
## 23          8 143994266        A        G 61757294  CYP11B2:1585
## 24          8 143994266        A        G 61757294  CYP11B2:1585
##    clin_sigid                   clin_hgvs              clin_vartype
## 23          5 NC_000008.10:g.143994266A>G single nucleotide variant
## 24          5 NC_000008.10:g.143994266A>G single nucleotide variant
##      clin_sig
## 23 Pathogenic
## 24 Pathogenic
##                                                                                  clin_clindbn
## 23 Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
## 24 Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
```

```
##    clin_cytogenetic clin_guidelines
## 23           8q24.3            NULL
## 24           8q24.3            NULL
```

Results are consistent with clinvar:
http://www.ncbi.nlm.nih.gov/clinvar/variation/16876/

Entry is listed as pathogenic by 1 submission with a rating of 5.

This sample ID does not appear in Brady's results.

Verifying this entry in impala:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.sample_id = "101-506-M"
AND comgen.chr = "8"
AND comgen.start = 143994266
```

Results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQLOW
- totalreadcount = 40

Sample ID 101-525-M Viewing impala notation:

```
impala[grep("101_525_M", impala$sample_id),]
```

```
##       start     stop zygosity vartype ref allele1seq allele2seq
## 30 67091182 67091183      hom     snp   G          T          T
##    allele1varquality allele2varquality totalreadcount sample_id chr
## 30            VQHIGH            VQHIGH             37 101_525_M   8
##    clin_chrom clin_pos clin_ref clin_alt     rsid clin_geneinfo clin_sigid
## 30          8 67091182        G        T 12721510      CRH:1392          5
##                      clin_hgvs              clin_vartype   clin_sig
## 30 NC_000008.10:g.67091182G>T single nucleotide variant Pathogenic
##                                        clin_clindbn clin_cytogenetic
## 30 Autosomal_dominant_nocturnal_frontal_lobe_epilepsy             8q13
##    clin_guidelines
## 30            NULL
```

Results are consistent with clinvar:
http://www.ncbi.nlm.nih.gov/clinvar/variation/38800/

Entry is listed as pathogenic by 1 submission with a rating of 5.

This sample ID does not appear in Brady's results.

Verifying this entry in impala:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.sample_id = "101-525-M"
AND comgen.chr = "8"
AND comgen.start = 67091182
```

Results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQHIGH
- totalreadcount = 37

Sample ID 101-728-F Viewing impala notation:

```
impala[grep("101_728_F", impala$sample_id),]
```

```
##       start     stop zygosity vartype ref allele1seq allele2seq
## 16 67091182 67091183      hom     snp   G          T          T
##    allele1varquality allele2varquality totalreadcount sample_id chr
## 16            VQHIGH            VQHIGH             46 101_728_F   8
##    clin_chrom clin_pos clin_ref clin_alt     rsid clin_geneinfo clin_sigid
## 16          8 67091182        G        T 12721510      CRH:1392          5
##                     clin_hgvs              clin_vartype   clin_sig
## 16 NC_000008.10:g.67091182G>T single nucleotide variant Pathogenic
##                                        clin_clindbn clin_cytogenetic
## 16 Autosomal_dominant_nocturnal_frontal_lobe_epilepsy             8q13
##    clin_guidelines
## 16            NULL
```

Results are consistent with clinvar:
http://www.ncbi.nlm.nih.gov/clinvar/variation/38800/

Entry is listed as pathogenic by 1 submission with a rating of 5.

This sample ID appears in Brady's results on chromosome 5 and 7, but not 8.

Verifying this entry in impala:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.sample_id = "101-728-F"
AND comgen.chr = "8"
AND comgen.start = 67091182
```

Results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQHIGH
- totalreadcount = 46

Sample ID 101-878-M Viewing impala notation:

```
impala[grep("101_878_M", impala$sample_id),]
```

```
##        start      stop zygosity vartype ref allele1seq allele2seq
## 47 143994266 143994267      hom     snp   A          G          G
## 48 143994266 143994267      hom     snp   A          G          G
##    allele1varquality allele2varquality totalreadcount sample_id chr
## 47             VQLOW            VQHIGH             18 101_878_M   8
## 48             VQLOW            VQHIGH             18 101_878_M   8
##    clin_chrom  clin_pos clin_ref clin_alt     rsid clin_geneinfo
## 47          8 143994266        A        G 61757294  CYP11B2:1585
## 48          8 143994266        A        G 61757294  CYP11B2:1585
##    clin_sigid                   clin_hgvs              clin_vartype
## 47          5 NC_000008.10:g.143994266A>G single nucleotide variant
## 48          5 NC_000008.10:g.143994266A>G single nucleotide variant
##      clin_sig
```

```
## 47 Pathogenic
## 48 Pathogenic
##                                                                                                clin_clindbn
## 47 Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
## 48 Corticosterone_methyloxidase_type_2_deficiency|Corticosterone_methyloxidase_type_1_deficiency
##    clin_cytogenetic clin_guidelines
## 47           8q24.3            NULL
## 48           8q24.3            NULL
```

Results are consistent with clinvar:
http://www.ncbi.nlm.nih.gov/clinvar/variation/16876/

Entry is listed as pathogenic by 1 submission with a rating of 5.

This sample ID appears in Brady's results on chromosome 7, but not 8.

Verifying this entry in impala:

```
SELECT *
FROM p7_ptb.comgen_variant as comgen
WHERE comgen.sample_id = "101-878-M"
AND comgen.chr = "8"
AND comgen.start = 143994266
```

Results were consistent with the annotation. The snp begins on the record's start site.
- allele1varquality = VQLOW
- totalreadcount = 18

Conclusions

For sample ID's found in Brady's result set and not included in imapala, we need to determine why these sample ID's are not found in the comgen_variant table. Once the sample ID's are located and added to impala, this analysis can be re-run to ensure results are consistent.

For the sample ID's found in impala, but not in Brady's set, one possible source of difference may be due to filtering whereas records with quality scores lower than VQHIGH or readcount below a certain total are filtered out in Prachi and Varsha's pipeline. However, for sample ID 101-868-M from Brady's result set, the read count was 14, so this does not seem to be the case.

I'll need to meet with Brady to determine the source of these differences.