Issue with CADD

Description

Only a small subset of markers in a single VCF return annotations. The rest return only RawScore and Phred in the same output file as the others. As an example, here is an illustration from running on Chr 12.

```
chrom=12
datadir='/scratch/groups/abattle4/victor/WatershedAFR/data/'
envloc='~/.conda/envs/cadd'
cadd_loc='/scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts'
vcf_loc=paste0(datadir,'/rare_variants_gnomad/gene-AFR-rv.CADD.vcf')
chr_vcf_loc=paste0(datadir,'/rare_variants_gnomad/gene-AFR-rv.CADD.chr',chrom,'.vcf')
outprefix=paste0(datadir,'/annotation/debug/gene-AFR-rv.CADD.chr',chrom)

Sys.setenv(datadir=datadir,
envloc=envloc,
cadd_loc=cadd_loc,
chrom=chrom,
vcf_loc=vcf_loc,
chr_vcf_loc=chr_vcf_loc,
outprefix=outprefix
)
```

```
# only looking at chromosome 12 for example

## store path variables
# datadir=/scratch/groups/abattle4/victor/WatershedAFR/data/
# envloc=/home-2/jbonnie1@jhu.edu/.conda/envs/cadd
# cadd_loc=/scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts
# ml anaconda
# conda activate ${envloc}

# vcf_loc=${datadir}/rare_variants_gnomad/gene-AFR-rv.CADD.vcf
# chr_vcf_loc=$datadir/rare_variants_gnomad/gene-AFR-rv.CADD.chr${chrom}.vcf
# outprefix=$datadir/annotation/debug/gene-AFR-rv.CADD.chr${chrom}

# bash $cadd_loc/CADD.sh -a -g GRCh38 -o ${outprefix}.tsv.gz \
#   ${chr_vcf_loc} 2>&1 | tee ${outprefix}.log
cat ${outprefix}.log
```

```
## CADD-v1.6 (c) University of Washington, Hudson-Alpha Institute for Biotechnology a
nd Berlin Institute of Health 2013-2020. All rights reserved.
## Running snakemake pipeline:
## snakemake /tmp/tmp.9mvTSkIqhg/gene-AFR-rv.CADD.chr12.tsv.gz --use-conda --conda-pr
efix /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/envs --cores 1
## --configfile /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/config
/config_GRCh38_v1.6.yml --snakefile /scratch/groups/abattle4/jessica/RareVar_AFR/cadd
/CADD-scripts/Snakefile -q
## Job stats:
## job           count    min threads    max threads
## ----------  -------  -------------  -------------
## annotation        1              1              1
## imputation        1              1              1
## join              1              1              1
## prepare           1              1              1
## prescore          1              1              1
## score             1              1              1
## total             6              1              1
##
## Opening /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/data/presco
red/GRCh38_v1.6/incl_anno/gnomad.genomes.r3.0.indel_inclAnno.tsv.gz...
## Opening /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/data/presco
red/GRCh38_v1.6/incl_anno/gnomad.genomes.r3.0.snv.tsv.gz...
## Opening /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/data/presco
red/GRCh38_v1.6/incl_anno/whole_genome_SNVs_inclAnno.tsv.gz...
## Possible precedence issue with control flow operator at /scratch/groups/abattle4/j
essica/RareVar_AFR/cadd/CADD-scripts/envs/2d842d5bbeba7229d999795b70331783/lib/site_p
erl/5.26.2/Bio/DB/IndexedBase.pm line 805.
## Input file /tmp/tmp.9mvTSkIqhg/gene-AFR-rv.CADD.chr12.csv.gz is empty.
##
## CADD scored variants written to file: /scratch/groups/abattle4/victor/WatershedAFR
/data//annotation/debug/gene-AFR-rv.CADD.chr12.tsv.gz
```

Here are the number of lines in the output file that have more than 8 fields:

```
zcat ${outprefix}.tsv.gz | awk 'NF > 8' | wc -l
```

```
## 2679
```

Here are the number of lines in the output file that have less than 8 fields:

```
zcat ${outprefix}.tsv.gz | awk 'NF < 8' | wc -l
```

```
## 252816
```

Here are the first 5 output lines with too few fields:

```
zcat ${outprefix}.tsv.gz | awk 'NF < 8' | awk 'NR < 6' > ${outprefix}_n5.tsv
cat ${outprefix}_n5.tsv
```

```
## 12   39075   A   T   0.397017    5.394
## 12   39378   T   C   0.504847    6.555
## 12   39383   G   A   0.078593    1.906
## 12   39933   G   T   0.294273    4.222
## 12   43601   T   C   0.177167    2.867
```

What if we only run those lines?

```
# awk '{print $2}' ${outprefix}_n5.tsv > ${outprefix}_n5.txt
# awk 'NR==FNR{a[$1]++; next} $2 in a' ${outprefix}_n5.txt ${chr_vcf_loc} > ${outpref
ix}_n5.vcf
# bash $cadd_loc/CADD.sh -a -g GRCh38 -o ${outprefix}_n5_out.tsv.gz ${outprefix}_n5.v
cf 2>&1 | tee ${outprefix}_n5_out.log
cat ${outprefix}_n5_out.log
zcat ${outprefix}_n5_out.tsv.gz
```

```
## CADD-v1.6 (c) University of Washington, Hudson-Alpha Institute for Biotechnology a
nd Berlin Institute of Health 2013-2020. All rights reserved.
## Running snakemake pipeline:
## snakemake /tmp/tmp.cpbqzbZcp6/gene-AFR-rv.CADD.chr12_n5.tsv.gz --use-conda --conda
-prefix /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/envs --cores 1
## --configfile /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/config
/config_GRCh38_v1.6.yml --snakefile /scratch/groups/abattle4/jessica/RareVar_AFR/cadd
/CADD-scripts/Snakefile -q
## Job stats:
## job           count    min threads    max threads
## ----------  -------  -------------  -------------
## annotation        1              1              1
## imputation        1              1              1
## join              1              1              1
## prepare           1              1              1
## prescore          1              1              1
## score             1              1              1
## total             6              1              1
##
## Opening /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/data/presco
red/GRCh38_v1.6/incl_anno/gnomad.genomes.r3.0.indel_inclAnno.tsv.gz...
## Opening /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/data/presco
red/GRCh38_v1.6/incl_anno/gnomad.genomes.r3.0.snv.tsv.gz...
## Opening /scratch/groups/abattle4/jessica/RareVar_AFR/cadd/CADD-scripts/data/presco
red/GRCh38_v1.6/incl_anno/whole_genome_SNVs_inclAnno.tsv.gz...
## Possible precedence issue with control flow operator at /scratch/groups/abattle4/j
essica/RareVar_AFR/cadd/CADD-scripts/envs/2d842d5bbeba7229d999795b70331783/lib/site_p
erl/5.26.2/Bio/DB/IndexedBase.pm line 805.
## Input file /tmp/tmp.cpbqzbZcp6/gene-AFR-rv.CADD.chr12_n5.csv.gz is empty.
##
## CADD scored variants written to file: /scratch/groups/abattle4/victor/WatershedAFR
/data//annotation/debug/gene-AFR-rv.CADD.chr12_n5_out.tsv.gz
## ##CADD GRCh38-v1.6 (c) University of Washington, Hudson-Alpha Institute for Biotec
hnology and Berlin Institute of Health 2013-2020. All rights reserved.
## #Chrom   Pos Ref Alt Type    Length  AnnoType    Consequence ConsScore   ConsDetai
l  GC  CpG motifECount motifEName  motifEHIPos motifEScoreChng oAA nAA GeneID  Featur
eID   GeneName    CCDS    Intron  Exon    cDNApos relcDNApos  CDSpos  relCDSpos   pro
tPos relProtPos  Domain  Dst2Splice  Dst2SplType minDistTSS  minDistTSE  SIFTcat SIFT
val PolyPhenCat PolyPhenVal priPhCons   mamPhCons   verPhCons   priPhyloP   mamPhyloP
verPhyloP   bStatistic  targetScan  mirSVR-Score    mirSVR-E    mirSVR-Aln  cHmm_E1 c
Hmm_E2 cHmm_E3 cHmm_E4 cHmm_E5 cHmm_E6 cHmm_E7 cHmm_E8 cHmm_E9 cHmm_E10    cHmm_E11
cHmm_E12    cHmm_E13    cHmm_E14    cHmm_E15    cHmm_E16    cHmm_E17    cHmm_E18    c
Hmm_E19    cHmm_E20    cHmm_E21    cHmm_E22    cHmm_E23    cHmm_E24    cHmm_E25    Ge
rpRS  GerpRSpval  GerpN   GerpS   tOverlapMotifs  motifDist   EncodeH3K4me1-sum   Enc
odeH3K4me1-max   EncodeH3K4me2-sum   EncodeH3K4me2-max   EncodeH3K4me3-sum   EncodeH3
K4me3-max   EncodeH3K9ac-sum    EncodeH3K9ac-max    EncodeH3K9me3-sum   EncodeH3K9me3
-max   EncodeH3K27ac-sum   EncodeH3K27ac-max   EncodeH3K27me3-sum  EncodeH3K27me3-max
EncodeH3K36me3-sum  EncodeH3K36me3-max  EncodeH3K79me2-sum  EncodeH3K79me2-max  Encod
eH4K20me1-sum  EncodeH4K20me1-max  EncodeH2AFZ-sum EncodeH2AFZ-max EncodeDNase-sum En
codeDNase-max EncodetotalRNA-sum  EncodetotalRNA-max  Grantham    SpliceAI-acc-gain
SpliceAI-acc-loss   SpliceAI-don-gain   SpliceAI-don-loss   MMSp_acceptorIntron MMSp_
```

```
acceptor   MMSp_exon   MMSp_donor  MMSp_donorIntron    Dist2Mutation   Freq100bp   Ra
re100bp   Sngl100bp   Freq1000bp  Rare1000bp  Sngl1000bp  Freq10000bp Rare10000bp Sng
l10000bp EnsembleRegulatoryFeature   dbscSNV-ada_score   dbscSNV-rf_score    RemapOve
rlapTF  RemapOverlapCL  RawScore    PHRED
## 12   39075   A   T   0.397017    5.394
## 12   39378   T   C   0.504847    6.555
## 12   39383   G   A   0.078593    1.906
## 12   39933   G   T   0.294273    4.222
## 12   43601   T   C   0.177167    2.867
```

If we upload this same vcf to the online GUI, however, we get that there are no lines with less than 8 columns and we can see that at least one of them has all the expected fields.

```
zcat ${datadir}/annotation/debug/GRCh38-v1.6_anno_e5923a31b059563633853772767 7e5d8.ts
v.gz | awk 'NF < 8' | wc -l
zcat ${datadir}/annotation/debug/GRCh38-v1.6_anno_e5923a31b059563633853772767 7e5d8.ts
v.gz | head -n4 | tail -n1
```

```
## 0
## 12   39378   T   C   SNV 0   Intergenic  UPSTREAM    1   upstream    0.245   0   N
A  NA  NA  NA  NA  NA  ENSG00000249054 ENST00000546223 FAM138D NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  1245    2717    NA  NA  NA  NA  0.058   0.004   0.004   0.321
0.365   0.383   NA  NA  NA  NA  NA  0   0   2   0   0   0   1   0   0   0   0   1   1
2   8   0   0   0   1   0   10  19  3   0   0   NA  NA  3.69    1.27    NA  NA  0.18
0.18    1.03    1.03    0.89    0.89    NA  NA  NA  NA  NA  NA  6.16    1.21    NA  N
A  1.93    1   0.97    0.97    1.83    1.17    0.14    0.04    NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  118 0   0   1   0   0   28  0   0   75  NA  NA  NA  NA  NA
0.504847    6.555
```

Let's make sure that at least one of these is in those source files we got through CADD-scripts:

```
# wg_ia=${cadd_loc}/data/prescored/GRCh38_v1.6/incl_anno/whole_genome_SNVs_inclAnno.t
sv.gz
# ml htslib
# tabix ${wg_ia} 12:39378-39378 | head -n2 | tail -n1 > ${outprefix}_wgia_chr12_3937
8.tsv
cat ${outprefix}_wgia_chr12_39378.tsv
```

```
## 12   39378   T   C   SNV 0   Intergenic  UPSTREAM    1   upstream    0.245   0   N
A  NA  NA  NA  NA  NA  ENSG00000249054 ENST00000546223 FAM138D NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  1245    2717    NA  NA  NA  NA  0.058   0.004   0.004   0.321
0.365   0.383   NA  NA  NA  NA  NA  0   0   2   0   0   0   1   0   0   0   0   1   1
2   8   0   0   0   1   0   10  19  3   0   0   NA  NA  3.69    1.27    NA  NA  0.18
0.18    1.03    1.03    0.89    0.89    NA  NA  NA  NA  NA  NA  6.16    1.21    NA  N
A  1.93    1   0.97    0.97    1.83    1.17    0.14    0.04    NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  118 0   0   1   0   0   28  0   0   75  NA  NA  NA  NA  NA
0.504847    6.555
```

So, it's in that one.

```
gnmd=${cadd_loc}/data/prescored/GRCh38_v1.6/incl_anno/gnomad.genomes.r3.0.snv.tsv.gz
# tabix ${gnmd} 12:39378-39378 > ${outprefix}_gnmd_chr12_39378.tsv
cat ${outprefix}_gnmd_chr12_39378.tsv
```

```
## 12   39378   T   C   0.504847    6.555
```

And also that one.

```
gnmd_indel=${cadd_loc}/data/prescored/GRCh38_v1.6/incl_anno/gnomad.genomes.r3.0.indel
_inclAnno.tsv.gz
# tabix ${gnmd_indel} 12:39378-39378 > ${outprefix}_gnmd_indel_chr12_39378.tsv
cat ${outprefix}_gnmd_indel_chr12_39378.tsv
```

It’s not in the last one, but I wouldn’t expect it to be.

SO. I don’t know what is wrong.