Using nuclear family WGS data to predict which child has pediatric cancer - scripts

Dustin Miller

03/04/2021

Python scripts executed in a Docker container to process VCF files and identify variants

The scripts listed below were executed with the docker container compound-het-vip. The scripts executed were adapted from CompoundHetVIP. The example here explains, in detail, how the scripts are used. The publication for CompoundHetVIP can be found here: https://doi.org/10.12688/f1000research.26848.2

Miller DB and Piccolo SR. CompoundHetVIP: Compound Heterozygous Variant Identification Pipeline [version 2; peer review: 2 approved]. F1000Research 2021, 9:1211 (https://doi.org/10.12688/f1000research.26848.2)

"keep_passed_variants.py" keeps variant positions that meet the filtering criteria of "PASS", a Quality Score >= 20 and removes multiallelic positions.

```
# Phased SNP files
docker run -d -v /Data:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/keep_passed_variants.py \
  PedFam/*/outs/phased_variants.vcf.gz \
  PedFam_analysis/phased_files/ \
  PedFam_analysis/passed_variants_summary.tsv \
  > PedFam_analysis/keep_passed_variants.out

# Phased deletion files
docker run -d -v /Data:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/keep_passed_variants.py \
  PedFam/*/outs/dels.vcf.gz \
  PedFam_analysis/dels_files/ \
  PedFam_analysis/passed_dels_summary.tsv \
  > PedFam_analysis/keep_passed_dels.out

# Phased structural variant files
docker run -d -v /Data:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/keep_passed_variants.py \
  PedFam/*/outs/large_svs.vcf.gz \
  PedFam_analysis/svs_files/ \
  PedFam_analysis/passed_svs_summary.tsv \
  > PedFam_analysis/keep_passed_svs.out
```

"concat_merge_phased_vcf.py" is used to merge all samples into a single file

```
# SNP file
docker run -d -v /Data:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/concat_merge_phased_vcf.py \
  PedFam_analysis/phased_files/ \
  PedFam_analysis/combined.vcf.gz \
  --output_fam_file PedFam_analysis/combined.fam \
  --concat_files n \
  --merge_files all \
  > PedFam_analysis/concat_merge_phased_vcf.out

# Deletion file
docker run -d -v /Data:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/concat_merge_phased_vcf.py \
  PedFam_analysis/dels_files/ \
  PedFam_analysis/dels_combined.vcf.gz \
  --output_fam_file PedFam_analysis/dels_combined.fam \
  --concat_files n \
  --merge_files none \
  > PedFam_analysis/concat_merge_dels.out

# Structural variant file
docker run -d -v /Data:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/concat_merge_phased_vcf.py \
  PedFam_analysis/svs_files/ \
  PedFam_analysis/svs_combined.vcf.gz \
  --output_fam_file PedFam_analysis/svs_combined.fam \
  --concat_files n \
  --merge_files none \
  > PedFam_analysis/concat_merge_svs.out
```

"vt_split_trim_left_align.py" is used to normalize and left-trim the variant calls. This only needed to be done on the SNP file, not the deletion or structural variant files.

```
# SNP file
docker run -d -v /Data:/proj -v /Data/PedFam_analysis/references:/references -w /proj \
  -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/vt_split_trim_left_align.py \
  PedFam_analysis/combined.vcf.gz \
  PedFam_analysis/combined_vt.vcf.gz \
  > PedFam_analysis/vt_split_trim_left_align.out
```

"annotate.py" uses snpEff to annotate the variants

```
# SNP file
docker run -d -v /Data:/proj \
  -v /Data/PedFam_analysis/references/snpEff_data:/snpEff/./data/GRCh38.86 -w /proj \
  -t dmill903/compound-het-vip:1.0 \
```

```
  python3 PedFam_analysis/PedFam/scripts/annotate.py \
  PedFam_analysis/combined_vt.vcf.gz \
  PedFam_analysis/combined_annotated.vcf \
  > PedFam_analysis/annotate.out

# Deletion file
docker run -d -v /Data:/proj \
  -v /Data/PedFam_analysis/references/snpEff_data:/snpEff/./data/GRCh38.86 -w /proj \
  -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/annotate.py \
  PedFam_analysis/dels_combined.vcf.gz \
  PedFam_analysis/dels_annotated.vcf \
  > PedFam_analysis/annotate_dels.out

# Structural Variant file
docker run -d -v /Data:/proj \
  -v /Data/PedFam_analysis/references/snpEff_data:/snpEff/./data/GRCh38.86 -w /proj \
  -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/annotate.py \
  PedFam_analysis/svs_combined.vcf.gz \
  PedFam_analysis/svs_annotated.vcf \
  > PedFam_analysis/annotate_svs.out
```

"gemini_load.py" uses vcf2db to create a GEMINI-compatible database

```
# SNP file
docker run -d -v /Data:/proj \
  -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/gemini_load.py \
  PedFam_analysis/combined_annotated.vcf \
  PedFam_analysis/combined.db \
  --fam_file PedFam_analysis/combined.fam \
  > PedFam_analysis/gemini_load.out

# Deletion file
docker run -d -v /Data:/proj \
  -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/gemini_load.py \
  PedFam_analysis/dels_annotated.vcf \
  PedFam_analysis/dels_combined.db \
  --fam_file PedFam_analysis/dels_combined.fam \
  > PedFam_analysis/gemini_load_dels.out

# Structural variant file
docker run -d -v /Data:/proj \
  -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/gemini_load.py \
  PedFam_analysis/svs_annotated.vcf \
  PedFam_analysis/svs_combined.db \
  --fam_file PedFam_analysis/svs_combined.fam \
  > PedFam_analysis/gemini_load_svs.out
```

"create_gnomAD_1K_cadd_file.py" was used to create a file that contained CADD, gnomAD MAF values, and 1000 Genomes MAF values for all congruent positions and is used during variant identification (subsequent steps) as a reference to obtain these values for each variant position.

```
docker run -d -v /Data/PedFam_analysis:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam/scripts/create_gnomAD_1K_cadd_file.py \
  > create_gnomAD_1K_cadd_file.out
```

"identify_CH_variants.py" is used to identify CH variants. When first executed, the GEMINI database that was created in the previous step is converted to a tsv. CADD, gnomAD MAF, 1000 genome MAF values are added to each variant position in the tsv and used to query for CH variants.

```
# SNP file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_CH_variants.py \
  PedFam_analysis/combined.db \
  PedFam_analysis/PedFam_CH_cadd20_maf01.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/combined.fam \
  --cadd 20 \
  --af 0.01 \
  > PedFam_analysis/identify_CH_variants.out

# Deletion file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_CH_variants.py \
  PedFam_analysis/dels_combined.db \
  PedFam_analysis/PedFam_CH_cadd20_maf01_dels.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/dels_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_CH_variants_dels.out

# Structural variant file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_CH_variants.py \
  PedFam_analysis/svs_combined.db \
  PedFam_analysis/PedFam_CH_cadd20_maf01_svs.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/svs_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_CH_variants_svs.out
```

"identify_deNovo_variants.py" is used to identify deNovo variants using the tsv that was created using the "identify_CH_variants.py" script

```
# SNP file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_deNovo_variants.py \
  PedFam_analysis/combined.db \
  PedFam_analysis/PedFam_deNovo_cadd20_maf01.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  PedFam_analysis/combined.fam \
  --cadd 20 \
  --af 0.01 \
  > PedFam_analysis/identify_deNovo_variants.out

# Deletion file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_deNovo_variants.py \
  PedFam_analysis/dels_combined.db \
  PedFam_analysis/PedFam_deNovo_cadd20_maf01_dels.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  PedFam_analysis/dels_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_deNovo_variants_dels.out

# Structural variant file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_deNovo_variants.py \
  PedFam_analysis/svs_combined.db \
  PedFam_analysis/PedFam_deNovo_cadd20_maf01_svs.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  PedFam_analysis/svs_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_deNovo_variants_svs.out
```

"identify_homAlt_variants.py" is used to identify homozygous alternate variants using the tsv that was created using the "identify_CH_variants.py" script

```
# SNP file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_homAlt_variants.py \
  PedFam_analysis/combined.db \
  PedFam_analysis/PedFam_homAlt_cadd20_maf01.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/combined.fam \
  --cadd 20 \
  --af 0.01 \
  > PedFam_analysis/identify_homAlt_variants.out
```

```
# Deletion file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_homAlt_variants.py \
  PedFam_analysis/dels_combined.db \
  PedFam_analysis/PedFam_homAlt_cadd20_maf01_dels.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/dels_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_homAlt_variants_dels.out

# Structural variant file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_homAlt_variants.py \
  PedFam_analysis/svs_combined.db \
  PedFam_analysis/PedFam_homAlt_cadd20_maf01_svs.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/svs_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_homAlt_variants_svs.out
```

"identify_het_variants.py" is used to identify heterozygous variants using the tsv that was created using the "identify_CH_variants.py" script

```
# SNP file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_het_variants.py \
  PedFam_analysis/combined.db \
  PedFam_analysis/PedFam_het_cadd20_maf01.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/combined.fam \
  --cadd 20 \
  --af 0.01 \
  > PedFam_analysis/identify_het_variants.out

# Deletion file
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_het_variants.py \
  PedFam_analysis/dels_combined.db \
  PedFam_analysis/PedFam_het_cadd20_maf01_dels.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/dels_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_het_variants_dels.out

# Structural variant file
```

```
docker run -d -v /Data/:/proj -w /proj -t dmill903/compound-het-vip:1.0 \
  python3 PedFam_analysis/PedFam/scripts/identify_het_variants.py \
  PedFam_analysis/svs_combined.db \
  PedFam_analysis/PedFam_het_cadd20_maf01_svs.tsv \
  PedFam_analysis/gnomAD_1K_cadd_GRCh38.tsv.gz \
  --fam_file PedFam_analysis/svs_combined.fam \
  --cadd 20 \
  --af 0.01 \
  --impact_filter_only y \
  > PedFam_analysis/identify_het_variants_svs.out
```