Last updated: 07/17/2023

1. Overview of the "pedann" pipeline

Pedigree pipeline "pedann" is the NYGC's in-house developed pipeline that uses a family-level joint-genotyped VCF as input and that consists of the following steps:
1. De novo variant identification using FamSeq tool (Peng G *et al.* 2013 and 2014), as well as genotype calls generated by GATK's HaplotypeCaller.
2. Annotating variants with custom pedann annotations (trait and penetrance mode).
3. Annotating variants with functional annotations using the Variant Effect Predictor tool (VEP v104; McLaren W *et al.* 2016).

Pedann pipeline outputs an annotated VCF (`<prefix>`.annotated.vcf.gz) and tab delimited file (`<prefix>`.annotated.txt.gz).

2. Definitions of columns in the `<prefix>`.annotated.txt.gz output of pedann

1. **CHROM** chromosome
2. **POS** position
3. **REF** ref allele
4. **ALT** alt allele
5. **FILTER** VQSR filter
6. **Allele** the variant allele used to calculate the consequence
7. **Consequence** consequence of the variant on the protein sequence (e.g. frameshift, stop gained, missense etc.)
8. **IMPACT** impact level (HIGH, MODERATE, LOW, MODIFIER)
9. **SYMBOL** the gene symbol
10. **Gene** Ensembl stable ID of affected gene
11. **Feature_type** type of feature (Transcript, RegulatoryFeature, MotifFeature)
12. **Feature** Ensembl stable ID of feature
13. **BIOTYPE** biotype of transcript or regulatory feature
14. **EXON** the exon number (out of total number)
15. **INTRON** the intron number (out of total number)
16. **HGVSc** HGVS coding sequence name
17. **HGVSp** HGVS protein sequence name
18. **cDNA_position** relative position of base pair in cDNA sequence
19. **CDS_position** relative position of base pair in coding sequence
20. **Protein_position** relative position of amino acid in protein
21. **Amino_acids** reference and variant amino acids
22. **Codons** reference and variant codon sequence
23. **Existing_variation** identifier(s) of co-located known variants
24. **DISTANCE** shortest distance from variant to transcript
25. **STRAND** strand of the feature (1/-1)
26. **FLAGS** Transcript quality flags
27. **VARIANT_CLASS** sequence ontology variant class
28. **SYMBOL_SOURCE** the source of the gene symbol
29. **HGNC_ID** HGNC unique gene ID
30. **TSL** Transcript support level. NB: not available for GRCh37
31. **APPRIS** Annotates alternatively spliced transcripts as primary or alternate based on a range of computational methods. NB: not available for GRCh37
32. **GIVEN_REF** reference allele from input
33. **USED_REF** reference allele as used to get consequences
34. **SOURCE** NA
35. **GENE_PHENO** Indicates if overlapped gene is associated with a phenotype, disease or trait
36. **NEAREST** Identifier(s) of nearest transcription start site

37. **SIFT** SIFT prediction and/or score
38. **PolyPhen** PolyPhen prediction and/or score
39. **DOMAINS** source and identifier of any overlapping protein domains
40. **HGVS_OFFSET** Indicates by how many bases the HGVS notations for this variant have been shifted
41. **CLIN_SIG** ClinVar clinical significance of the dbSNP variant
42. **SOMATIC** somatic variation existing in the COSMIC database
43. **PHENO** Indicates if existing variant is associated with a phenotype, disease or trait; multiple values correspond to multiple values in the Existing_variation field
44. **PUBMED** Pubmed ID(s) of publications that cite existing variant
45. **MOTIF_NAME** the source and identifier of a transcription factor binding profile (TFBP) aligned at this position
46. **MOTIF_POS** the relative position of the variation in the aligned TFBP
47. **HIGH_INF_POS** flag indicating if the variant falls in a high information position of a TFBP
48. **MOTIF_SCORE_CHANGE** the difference in motif score of the reference and variant sequences for the TFBP
49. **TRANSCRIPTION_FACTORS** transcription factor binding site
50. **phyloP100** phyloP (phylogenetic p-value) conservation score based on the multiple alignments of 100 vertebrate species. Positive scores measure conservation, negative scores measure acceleration, i.e. faster than expected evolution. (Pollard *et al.* 2010)
51. **phastcons100** phastcons conservation scores based on the multiple alignments of 100 vertebrate species; score ranges from 0 to 1 and represents probability of negative selection (Siepel *et al.* 2005)
52. **gnomad_v3**
53. **gnomad_v3_AF**: alternate allele frequency in gnomas v3.1.1
54. **gnomad_v3_AN**: total number of called alleles in gnomad v3.1.1
55. **gnomad_v3_nhomalt**: Count of homozygous individuals in gnomad v3.1.1
56. **ClinVar_20210501**
57. **ClinVar_20210501_DBVARID**: nsv accessions from dbVar for the variant
58. **ClinVar_20210501_ALLELE_ID**: the ClinVar Allele ID
59. **ClinVar_20210501_CLNDN**: ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB
60. **ClinVar_20210501_CLNDISDB**: Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN
61. **ClinVar_20210501_MC**: comma separated list of molecular consequence in the form of Sequence Ontology ID|molecular_consequence
62. **ClinVar_20210501_CLNSIG**: Clinical significance for this single variant
63. **ClinVar_20210501_CLNSIGCONF**: Conflicting clinical significance for this single variant
64. **ClinVar_20210501_CLNREVSTAT**: ClinVar review status for the Variation ID
65. **ClinVar_20210501_ORIGIN**: Allele origin. One or more of the following values may be added: 0 - unknown; 1 - germline; 2 - somatic; 4 - inherited; 8 - paternal; 16 - maternal; 32 - de-novo; 64 - biparental; 128 - uniparental; 256 - not-tested; 512 - tested-inconclusive; 1073741824 - other
66. **AC** allele count in genotypes, for each ALT allele in the trio
67. **AN** total number of alleles in called genotypes
68. **AF** allele frequency in the trio
69. **SampleID_1.GT** Genotype called by GATK HaplotypeCaller
70. **SampleID_2.GT**
71. **SampleID_3.GT**
72. **SampleID_1.AD** Allelic depths for the ref and alt alleles in the order listed
73. **SampleID_2.AD**
74. **SampleID_3.AD**
75. **SampleID_1.DP** Approximate read depth; some reads may have been filtered
76. **SampleID_2.DP**
77. **SampleID_3.DP**
78. **SampleID_1.GQ** Genotype quality (Phred-scaled confidence that the genotype assignment (GT) is correct)

79. **SampleID__2.GQ**
80. **SampleID__3.GQ**
81. **SampleID__1.PL** Normalized Phred-scaled likelihoods of the possible genotypes
82. **SampleID__2.PL**
83. **SampleID__3.PL**
84. **SampleID__1.FGT** Genotype called by FamSeq
85. **SampleID__2.FGT**
86. **SampleID__3.FGT**
87. **SampleID__1.FPP** Posterior probability calculated by FamSeq
88. **SampleID__2.FPP**
89. **SampleID__3.FPP**
90. **SampleID__1.PEDANN__DESC** Description of SampleID's variant (e.g. DeNovo, maternal dominant etc.) generated by pedann pipeline (see below for more details; available for all affected and unaffected children in the pedigree)
91. **SampleID__2.PEDANN__DESC**
92. **SampleID__3.PEDANN__DESC**
93. **SampleID__1.PEDANN__TRAIT** Trait mode (Dom=dominant, Rec=recessive; available for all affected and unaffected children in the pedigree).
94. **SampleID__2.PEDANN__TRAIT**
95. **SampleID__3.PEDANN__TRAIT**
96. **SampleID__1.PEDANN__PNTR** Penetrance mode (cp=complete penetrance, vp=variable penetrance, lor="loss-of-resiliance" ("loss-of-resiliance" refers to loss of at least one copy of an alternate allele as compared to parental GTs); available for only for the affected children in the pedigree).
97. **SampleID__2.PEDANN__PNTR**
98. **SampleID__3.PEDANN__PNTR**

3. Definitions of variant annotations provided in columns "PEDANN__DESC"

Autosomes and PAR1/PAR2 regions of chromosome X:

- DeNovo_FS: de novo variant call supported only by the FamSeq tool (not supported by GATK's GT calls).
- DeNovo_GS: de novo variant call supported only by the GATK's GT calls (not supported by FamSeq tool). Note: FamSeq identifies de novo variants only among biallelic SNPs with non-missing GT calls, so all denovo variants idenfitied among biallelic indels and biallelic SNPs with some missing GT calls are based only on GATK's GT calls and thus have a "_GS" suffix.
- DeNovo_GFS: de novo variant call supported by both FamSeq and GATK.
- DeNovo_GS_HC/DeNovo_FS_HC/DeNovo_GFS_HC: high confidence de novo variant call; de novo call that meets the following criteria:
  - child's GT = "0/1", mother's GT = "0/0", father's GT = "0/0" (or FGT in case of DeNovo_FS_HC);
  - child's/mother's/father's DP > 9;
  - child's/mother's/father's GQ > 20;
  - child's AB = 0.25 < AB < 0.75 (AB computed from the AD field as (allele depth of ALT)/(allele depth of REF + allele depth of ALT));
  - father's and mother's AD of ALT allele = 0.
- MatHemi: maternal hemizygosity.
- PatHemi: paternal hemizygosity.
- UPD: uniparental disomy.
- Dom: dominant.
- Rec: recessive.
- MatDom: maternal dominant.
- PatDom: paternal dominant.
- RefHom: homozygous referent.

nonPAR regions of chromosome X:

- "PEDANN_DESC" annotation labels across nonPAR regions of chrX are based on GATK's GT calls only (FamSeq does not support mixed ploidies) and they start with "XLinked", e.g. "XLinkedDeNovo", "XLinkedMatDom", "XLinkedPatDom", etc.

Chromosome Y:

- Variants on chrY are not annotated with "PEDANN_DESC" labels.

4. Important notes.

- To ensure high sensitivity of de novo calling, we set the de novo prior probability (mRate) parameter within the FamSeq step to 1e-6 as default in the pedann pipeline (which is an order of magnitude higher than the default parameter within the FamSeq tool, i.e. 1e-7). The mRate parameter can be adjusted depending on the goals of the project. For example, if a more conservative set of de novos is desired, we recommend requesting the mRate to be lowered to 1e-7 or 1e-8.
- Pedann pipeline is designed to be used on complete families only (i.e. families that include at least one child (irrespective of phenotype status) and his/her mother and father).
- For optimal operation of the pipeline, it is strongly recommended that the input VCF and pedigree file contain only samples that are relevant to the pedigree analysis.
- "PEDANN_DESC", "PEDANN_TRAIT", and "PEDANN_PNTR" annotations are applied to all affected and unaffected children in the family (parents are not annotated unless they are a part of multi-generational family, see the note below).
- If multiple generations/extended family members are present in the input VCF, pedann will run on each subfamily (child + mother + father + optional sibling(s)) and output 1 annotated VCF file per each subfamily (each subfamily-VCF will contain GT calls for all members of the extended family, but pedann annotations will be included only for the children in the given subfamily).
- Only biallelic SNPs with no missing GT calls and no missing PL values are included in FamSeq de novo variant calling (note: e.g. in a family with 6 members, all 6 GT calls and PL values would have to be non-missing for a variant to be included in FamSeq de novo variant calling).
- Biallelic indels as well as some biallelic SNPs with missing GT calls and/or missing PL values that were not analyzed by FamSeq (e.g. affectedChild=0/1, mother=0/0, father=0/0, unaffectedChild =./.) can still be annotated as de novo variants (labeled "DeNovo_GS") based on their original GT calls in the input VCF file.
- Multiallelic variants are not annotated with pedann-specific annotations (i.e. "PEDANN_DESC", "PEDANN_TRAIT", and "PEDANN_PNTR" are all set to "." in case of multiallelic variants), they are only annotated by the VEP.
- Variants in the *annotated.vcf.gz as well as* annotated.txt.gz are normalized using "bcftools norm" (i.e. indels are left-aligned and normalized and multiallelic variants are split into separate rows). Original representation of variants in the *annotated.vcf.gz file is shown in the INFO field "PRE_NORM_VARIANT".
- Parents are always assumed to be unaffected in the pedann analysis (important to keep in mind when analyzing penetrance labels).
- Only affected children are annotated with penetrance type ("PEDANN_PNTR"). Penetrance is defined based on the comparison of GT call of a given affected child with the GT calls of the parents (who are always assumed to be unaffected) and all unaffected siblings (penetrance for parents and unaffected children is set to ".").
- Pedigree information for a family in the annotated VCF and txt file is included in the header.
- The PEDANN_DESC label definitions described in Section 3 above assume that the variants in the input VCF are called with sex-dependent ploidy settings on chromosomes X and Y. If variants in the input VCF were called using ploidy=2 setting across all chromosomes (including sex chromosomes) then sites on chrX and chrY will be annotated with the same "PEDANN_DESC" labels as those on autosomes.

5. Major updates.

07/17/2023. Changes introduced in pedann v0.15.0:

- Added FamSeq-specific fields (i.e. FGT (Genotype called by FamSeq), FPP (posterior probability calculated by FamSeq), and GPP (posterior probability calculated by single individual based method)) to the final outputs.
- GT and PL fields are no longer updated with FGT and FPP fields, respectively, in cases where variants are annotated with the "_FS" suffix in a "PEDANN_DESC" column. This is because we now provide FGT and FPP fields for all sites processed via FamSeq as mentioned above.
- Removed the AB (allele balance) field from the *pedann.annotated.txt.gz output. The AB field that we used to output there was coming from GATK HaplotypeCaller and was not the exact AB we use internally in the pipeline for filtering when defining high confidence de novos. The AB used for hard filtering is computed from the AD (allele depth) field provided by HaplotypeCaller (Haplotype Caller does not output AB for all sites which is why we're using AD to compute AB).

01/20/2022. Changes introduced in pedann v0.9.0:

- Mixed ploidy setting support on chrX and chrY.
- Renamed "PEDANN_INHRT" label to "PEDANN_TRAIT".

05/17/2021. Update to VEP annotations:

- Updated VEP from v93.2 to v104.
- Updated gnomad and ClinVar to latest versions.
- b37 now uses gnomad v2.1.1 genomes (no exomes) and b38 uses gnomad v3.1.1 genomes.
- Removed mtDNA annotations.

References

McLaren W *et al.* (2016). The Ensembl Variant Effect Predictor. Genome Biology Jun 6;17(1):122

Peng G *et al.* (2013). Rare variant detection using family-based sequencing analysis. Proceedings of the National Academy of Sciences, 110(10), 3985–3990. https://doi.org/10.1073/pnas.1222158110

Peng G *et al.* (2014). FamSeq: A Variant Calling Program for Family-Based Sequencing Data Using Graphics Processing Units. PLoS Computational Biology, 10(10), e1003880. https://doi.org/10.1371/journal.pcbi.1003880

Pollard KS *et al.* (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research, 20(1), 110–21. https://doi.org/10.1101/gr.097857.109

Siepel A *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8), 1034–1050. https://doi.org/10.1101/gr.3715005