CITATION

For any resultant publications using please cite:
Matthew A. Field, Vicky Cho, T. Daniel Andrews, and Chris C. Goodnow (2015). "Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies" PlosOne. doi: 10.1371/journal.pone.0143199

AND

Field, M. A., V. Cho, M. C. Cook, A. Enders, C. Vinuesa, B. Whittle, T. D. Andrews and C. C. Goodnow (2015). "Reducing the search space for causal genetic variants with VASP: Variant Analysis of Sequenced Pedigrees." Bioinformatics.

FILE CONTENT/FORMAT

The file contains a summary of all snvs or indels called in at least one member of the cohort. This consist of all variants that either overlap an ensembl canonical coding exons or splice site regions (defined as 10bp either side of an ensembl coding exon). Each line in this file represents everything we know about a particular variant for the entire cohort. For information specific to an individual in the cohort see the individual summary files. The lines are tab delimited and designed for loading into excel (set delimiter as tab).

CUTOFFS USED

For this report there are several important cutoffs utilised:

1) **Variant classification**: For a coordinate to be called 'variant' at a specific coordinate 10% or more of the total bases must be different from the reference base. Non reference bases include a single base other than the reference, an insertion of any size, or a deletion of any size. Note that this definition of variant is different from the definition applied to an individual which uses numerous cutoff such as samtools snv score, min read depth, min average base quality, etc
2) **Zygosity classification**: There are three categories for zygosity in this file
    1. **ref**: >90% of the bases match the reference
    2. **hom**: >90% of the bases are the same variant base
    3. **het**: everything not in the above categories is called het (i.e. variant base frequency is between 11%-89%)

VARIANT GROUPINGS

The variants are divided in two major groups and divided by lines containing '---' in all cells:

a) NOVEL_OR_RARE entries

-Novel variants (i.e. no matching dbsnp entry) and rare variants (variant allele frequency is <2%) that either cause a non-synonomous change in a ensembl entry or overlap a splice site. Each group is ordered by descending read depth.

b) LOW_PRIORITY entries

-Remaining variants that don't fall into a).

DEFAULT SORTING

The files is divided into two main groups; first passed variants (either novel or rare population variants i.e. <2% MAF) followed by all other lower priority variants. Within each group the lines are sorted first on the maximum number of affected samples with a variant and next on the minimum number of unaffected samples with a variant. Finally within these groups variants are sorted by increasing MAF. For an example suppose we have a cohort consisting of two affected and one unaffected samples:

passed	2_affected_samples_variant	0_unaffected_samples_variant
passed	2_affected_samples_variant	1_unaffected_samples_variant
passed	1_affected_samples_variant	0_unaffected_samples_variant
passed	1_affected_samples_variant	1_unaffected_samples_variant
passed	0_affected_samples_variant	1_unaffected_samples_vairant
low-priority	2_affected_samples_variant	0_unaffected_samples_variant
….		
low-priority	0_affected_samples_variant	1_unaffected_samples_variant

PRIORITISING VARIANTS

This is a big file with lots of information but there are ways to prioritise for the variants most likely to be causal. Please see the column definitions in the next section for descriptions of the columns mentioned below. Based on previously analysed cohorts where we've found the causal variant(s), they usually fall into the following categories.

Cases likely to contain variants of interest

1) variants flagged as de novo in **mendelian_inheritance** column → beware cases with low coverage as this often results in unreliable zygosity calls
2) variants common to affected individuals and absent in unaffected individuals; use columns **number_affected_variant** and **number_unaffected_variant** for this
3) variants flagged as auto-dominant, auto-recessive, X-linked-dominant, or X-linked-recessive in the **disease_inheritance** column
4) Variants where the mutant allele for all affected samples comes from a single parent while the unaffected samples get the reference allele. Look for 'Father gives affected allele' or 'Mother gives affected allele' in the **parent_allele_common_to_affected**
5) Compound hets with low gmaf frequencies found in the **definite_compound_het** and **possible_compound_het** columns
6) Variants either not in dbsnp or flagged as rare_alleles (defined as <2% gmaf); use the **filter_dbsnp_snv** or **filter_dbsnp_indel** column to find entries containing NOVEL or RARE_ALLELE
7) Nonsense mutations; use the **snv_exon_type** column and search for NONSENSE
8) Missense mutations determined to be damaging by polyphen and/or SIFT. Both pieces of software are know to have false positive rates of ~20% but perform better when both programs determine a variant is damaging; use the columns **polyphen_prediction** (look for probably_damaging) and **sift_prediction** (look for deleterious)
9) For splice site variants (defined as 10bp either side of the exon) the closer to the exon

the better. Use the **snv_exon_type** column to get this information (e.g. SPLICE (1) or SPLICE(2), etc)

Cases likely to not contain variants of interest.

1) Low coverage and/or missing allele information; use the different **sample alleles** columns to check for either low coverage levels or no data entries. These low coverage levels often result in incorrect zygosity calls.
2) Variants in genes with a large number of mutations; use the **gene_total_variants_in_gene** column to check for high number
3) Variants where the reference allele is the rare allele in the population. Use the **filter_dbsnp** columns to check for RARE_REF cases

COLUMN DEFINITIONS

Report columns are slightly dependent on the makeup of the cohort. For example if we have no parent sequence data you will not see columns 'mother_allele' and 'father_allele'

A) **samples**: List of all samples containing the variant according to the pileup files generated

B) **number_affected_variant/total_affected:** Number of **affected** samples defined as variant (see definition above) / Total number of affected samples in the cohort

e.g. 2 out of 2 → both affected samples in the cohort contain this variant
e.g. 1 out of 3 → 1 out of a total 3 affected samples contain this variant

C) **number_unaffected_variant/total_unaffected:** Number of **unaffected** samples defined as variant (see definition above) / Total number of unaffected samples in the cohort

e.g. 1 out of 3 → 1 out of a 3 unaffected samples contain this variant
e.g. 0 out of 1 → the single unaffected sample in the cohort is not variant at this coordinate

D) **mendelian_inheritance:** (cases with at least one parent)
Columns indicates whether Mendelian inheritance rules are observed within the cohort. If Mendelian inheritance rules are not followed one of the exception(s) will be reported. For example 'hom_parent ref_child' would indicate there is a parent who is homozygous at a particular allele while their child is reference which shouldn't happen with normal Mendelian inheritance. Cases following a de novo pattern are also flagged (defined as both parents being reference and the child being heterozygous) and represent potential variants of interest.

e.g. mendelian_rules_followed → allele distribution within the cohort follow Mendelian rules
e.g. de novo (ref_parents het_child) → de novo case where affected child has de novo mutant
e.g. hom_mother ref_child → broken Mendelian rule; here it's a hom_mother and a ref_child
e.g. missing allele_info → some samples contain no sequence information

E) **disease_inheritance:** disease inheritance pattern observedfor each variant. There are five distinct possible values for this field any of which may also be additionally annotated as

def_com_het or pos_com_het (see column definitions below). The five options are auto-recessive, auto-dominant, x-linked-recessive, x-linked-dominant, or none. Auto-recessive is defined as all affected being homs and all unaffected being non homs while auto-dominant is defined as all affected being heterozygous and all unaffected reference. The x-linked categories are the same except the variant falls on the X chromosome.

e.g. auto-dominant
e.g x-linked-recessive
e.g. auto-recessive,pos_com_het
e.g. x-linked-dominant,def_com_het

F) **gene:** The mutated gene name from hgnc

G) **gene_total_variants_in_gene:** Total number of mutations in the gene for the entire cohort. For example if two individuals share three different mutations in the same gene the value will be 6.

H) **unique_coord_gene_variants:** Total number of distinct coordinates mutated within the cohort for a single gene. For example if two individuals share three different mutations in the same gene the value will be 3.

I) **%_coding_bases_variant:** The percent of total coding bases in the gene found to be variant. Splice site variants will have N/A for this column as we are looking at coding bases only.

J) **mother_allele:** (cases with data from the mother)
Contains the allele inherited for each child from the mother when this can be determined.

e.g. affected1(A),affected2(ref),unaffected1(?) → affected1 get a mutant A allele from the mother, affected2 gets the ref allele from the mother, and we can't determine what allele unaffected1 gets from the mother
e.g. affected1(+1A),affected2(+1A),unaffected1(-2AA) → affected1 and affected2 get an insertion of an A from the father, and unaffected1 gets deletion of AA from the mother

K) **father_allele:** (cases with data from the father)
Same as above column but for the father

L) **parent_allele_common_to_affected:** (cases with at least one parent)
Flags cases where the following conditions are met:
-All affected children inherit the same mutant allele from the same parent
-All unaffected inherit the reference allele from this parent
-All samples inherit only the reference allele from the other parent
These cases often cluster and are likely to be regions of interest for causal variants. Possible values are 'no', '?','Mother gives affected allele', and 'Father gives affected allele'

e.g. If we have mother_allele = affected1(A),affected2(A),unaffected1(ref) and father_allele = affected1(ref),affected2(ref),unaffected1(ref) then the value will be 'Mother gives affected

allele'

M) **affected_allele_block:** Contains information on genomic blocks of cases described in L) i.e. cases where variant allele goes to affected children only from the same parent. These blocks may contain inconclusive variants but not cases where we can definitely state conditions described in L) are not met.

e.g. 4 variants; Mother; 22:21988602-22049369; 60767bp → means 4 consecutive variants in the ~60kb region from 22:21988602-22049369 are from the Mother

N) **definite_compound_het:** (cases with at least one parent)
To a gene to be defined as definitely compound het the following conditions must be met:

I) Gene contains at least one heterozygous variant inherited from each parent
II) The variant must not be homozygous in any unaffected individuals
III) The variant must be heterozygous in all affected individuals
IV) Unaffected and affected siblings must not share the exact same heterozygous variants

The variants where the mother gives the mutant allele begin with 'M=' while 'F=' represents cases where the father gives the mutant allele. This calculation is independent of variant type (i.e. a child can inherit an indel allele from their mother and a snv allele from their father). The strings contain lots of information and can get very long for genes with lots of mutations.

e.g. affected1(M=13:40261945:snv:gmaf=0.2766(G),F=13:40229957:snv:gmaf=0.4638(A)) → this means child affected1 gets a mutant allele from their mother (M=) for a snv at chr13:40261945 with gmaf of 0.2766 AND gets the mutant allele from their father (F=) for another snv at chr13:40229957 with gmaf of 0.4638.

O) **possible_compound_het:** (cases with at least one parent)
Cases similar to the above except that condition I) cannot be guaranteed due to parent and child both being hets meaning we cannot definitively know which allele the child got from each parent. These variants are flagged as AMB in the string below.

e.g. affected2(F=13:25876011:snv:gmaf=0.4075(A),AMB=13:25882049:DEL:gmaf=N/A) → this means child affected gets a mutant allele from their mother (F=) for a snv at chr13:25876011 with gmaf of 0.4075 AND **possibly** gets the mutant allele from their father (AMB=) for a deletion at chr13:25882049 with no known gmaf score

P...) **Sample_name (relation):** These columns report the sequence content for each sample at the genomic coordinate as well as the zygosity call (see definitions above). There will be one column for each sample.

The format of these columns is a string that show the sequence content and their respective counts at that coordinate for the sample followed by a zygosity call. Data is obtained from the pileup string
Example: ref:10_3:C_4:+2AG (het) → shows there are 10 reference base, 3 C's, and 4 AG insertions at that coordinate and that the zygosity is heterozygous

The remaining columns are either the snv entries or the indel entries depending on the file type. When samples have different values for a column (such as read_depth) all the values are listed in order separated by semicolons. When samples have the same value for a column (such as gene_name) then only one value is listed. See the individual report documentation for information on these columns