CITATION

1) For any resultant publications using single samples 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

2) For any resultant publications using paired cancer samples please cite:

Wilmott, J. S., M. A. Field, P. A. Johansson, H. Kakavand, P. Shang, R. De Paoli-Iseppi, R. E. Vilain, G. M. Pupo, V. Tembe, V. Jakrot, C. A. Shang, J. Cebon, M. Shackleton, A. Fitzgerald, J. F. Thompson, N. K. Hayward, G. J. Mann and R. A. Scolyer (2015). "Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes." Pathology. doi:10.1097/PAT.0000000000000324

FILE CONTENT / FORMAT

The file contains annotated snvs that overlap either ensembl canonical coding exons or splice site regions (defined as 10bp either side of an ensembl coding exon).
The lines are tab delimited and designed for loading into excel (set delimiter as tab when loading).

SNV GROUPINGS

The SNVs are divided in three major groups and divided by lines containing '---' in all cells:

a) SNV NOVEL entries
   -Novel SNVs (i.e. no matching dbsnp entry) 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) SNV RARE entries
   -Known SNVs (i.e. matches dbsnp entry) where variant allele frequency is <2% according to dbsnp. This filter is only applied to known SNVs where frequency is reported by dbsnp.
c) LOW_PRIORITY entries
   -Remaining SNVs that don't fall into a) or b). These entries are sorted by chromosome and then coordinate

COLUMN DEFINITIONS

Report columns are dependent on the type of sample; currently we are able to analyse human and mouse samples with different reports generated for cancer samples versus non cancer samples.

HUMAN SAMPLES (NON-CANCER) REPORTS

A) **chr**: chromosome

B) **coord**: genomic coordinate of snv

C) **ref_allele**: reference base at that coordinate

D) **ref_allele_count:** number of times reference allele is observed from the pileup string

E) **var_allele**: variant base call at that coordinate

F) **var_allele_count:** number of times variant allele is observed from the pileup string

G) **gene(hgnc)**: hugo gene name

H) **snv_score**: quality SNV score from samtools mpileup (higher is better; max 225)

I) **final_status**: contains final overall PASS/FAIL status. If failed it will list all the reasons for failing as there may be multiple reasons for failure. Generally variants are passed when they are novel or represent a rare variant, have high quality bases, and high read depth

J) **other_alleles:** string that show any additional alleles and their respective counts at that coordinate for the sample. Data is obtained from the pileup string
Example: C:+2AG_3:4 → shows there are 3 C's and 4 AG insertions at that coordinate in addition to the reference and called variant counts

K) **read_allele_freq**: reports the reference_base:variant_base ratio from the read data
Example: reads:R0.31:V0.66 → shows 31% of our reads match the reference base and 66% of the reads match the variant base

L) **known_variation:** reports all known variation at that coordinate from ENSEMBL's variant effect predictor tool. This report contains entries from dbsnp and cosmic in addition to other population study data

M) **gmaf_1000_genomes:** reports the minor allele frequency and corresponding base from 1000 genomes data. When the minor allele is the reference genome a 'REF' string is present.
Example: 0.069 (G) → 6.9% MAF frequency for variant base G
Example: 0.2143(C:REF) → 21.43% MAF frequency for reference base C

N) **dbsnp_match**: for dbsnp overlapping SNVs reports the rs id as well as the reported frequencies for the reference base and the variant allele. Possible entries:
Example: dbsnp:rs139422176:R0.999:V0.001 → rs id with ref freq 0.999 and var freq 0.001
Example: dbsnp:rs139422176:No_freq → rs id with no reported frequencies
Note that this information is obtained from the latest version of dbsnp which may differ from the data reported in the known_variation column which comes from the external tool (variant effect predictor) that may not be up to date with regard to dbsnp

O) **dbsnp_var_allele_freq:** variant allele frequency from the dbsnp match above; separate

column added for sorting. Possible values are a number from 0-1 and 'No freq' for entries with no frequency information available from dbsnp

P) **read_depth**: total number of reads aligned to that coordinate

Q) **median_quality_score**: median quality score for all bases aligned at that coordinate (max score is 40 and any score below 30 is low)

R) **aa_change**: show amino acid change for non-synonomous snvs. Possible entries:
Example: P->L → missense mutation
Example: P->STOP → nonsense mutation
Example: G->N (COMBINED:38122470-38122471) → missense mutation where multiple bases in the codon were mutated. The amino acid change reflects the combined effect of both mutations

S) **aa_position:** position of mutated amino acid within the ENSEMBL transcript

T) **aa_length:** total length in amino acids of the ENSEMBL transcript

U) **exon_intron_count:** relative position within the transcript from the variant effect predictor
Example: EXON->10/23 → Variant located in the 10th of 23 total exons
Example: INTRON->1/2 → Variant located in the 1st of 2 total introns

V) **snv_exon_type**: indicates whether the SNV overlaps the splice site or the exon and whether amino acid change is synonomous, missense, or nonsense (SYN, MISSENSE, NONSENSE or SPLICE). Currently to qualify for splice category the coordinate must be within 10bp of a coding exon

W) **sift_prediction**: the text classification returned from sift which estimates how damaging the amino acid change is to the protein. Possible values are deleterious or tolerated. 'N/A' entries indicate either synonymous variants, nonsense mutations, splice site variants, or cases where data is not available

X) **sift_score**: number from 0-1 with 0 being most likely damaging

Y) **polyphen_prediction**: the text classification returned from polyphen which estimates how damaging the amino acid change is to the protein. Possible values are benign, unknown, possiblydamaging, or probablydamaging. 'N/A' entries indicate either synonymous variants, nonsense mutations, or splice site variants

Z) **polyphen_score**: number from 0-1 with 1 being most likely damaging

AA) **cadd_phred:** phred quality score from CADD with higher being more damaging

AB) **filter_dbsnp_snv**: indicates whether coordinate matches known dbsnp entry and if it matches what type of match it is. Possible values are:
NOVEL -> no dbsnp match
RARE_ALLELE -> cases where the variant allele frequency is reported to be <2%
RARE_REF -> cases where the reference allele frequency is reported to be <2%
NO_FREQ -> cases where no dbsnp frequency information is given

FAIL -> all other dbsnp matching cases

AC) **filter_exac_snv:** exac snv information
reports variant frequency information from the ExAC resource. For each matching coordinates all variants are reports, not just for the variant being reported
Example: A->G:0.000119 → Reports snv of A->G with frequency of 0.000119
Example: C->T:0.034234,C->G:0.000434 → Reports snvs of C->T and C->G

AD) **filter_clinvar_snv:** Clinical significance reported for this variant in NCBI's ClinVar, followed by HGVS expression and corresponding condition(s) when provided.
Note some may have conflicting CLINSIG terms

AE) **ensembl**: ENSEMBL gene link

AF) **ens_canonical_trans:** canonical transcript from ENSEMBL. All variant information is relative to this transcript

AG) **protein_domains:** the source and unique identifier of any protein domains at that coordinate from the variant effect predictor.

AH) **uniprot**: uniprot gene name(s). This is a comma delimited list of all uniprot genes that map to this coordinate.

AI) **ccds**: ccds name(s). This is a comma delimited list of all ccds names that map to this coordinate

AJ) **refseq**: refseq gene name(s). This is a comma delimited list of all refseq names that map to this coordinate

AK) **description:** description of gene function from hugo

AL) **omim**: link to omim entry if it exists. Note that these entries are linked by gene name and not coordinates so if there are discrepancies in the gene name this link will not be reported

AM) **pubmed:** link to pubmed ID(s) of publications that cite existing variant

AN) **phenotype:** phenotype information from mouse data from MGI if available

AO) **cosmic**: link to gene in cosmic (catalogue of somatic mutations in cancer)

AP) **vogelstein**: indicates whether the gene is cited in vogelstein's landmark paper "Cancer genes and the pathways they control". If it is then the column contains information from the original vogelstein paper

AQ) **go**: gene GO terms

MOUSE SAMPLES (NON-CANCER) REPORTS

A) **chr**: chromosome

B) **coord**: genomic coordinate of snv

C) **ref_allele**: reference base at that coordinate

D) **ref_allele_count:** number of times reference allele is observed from the pileup string

E) **var_allele**: variant base call at that coordinate

F) **var_allele_count:** number of times variant allele is observed from the pileup string

G) **gene(mgi)**: mgi gene name

H) **snv_score**: quality SNV score from samtools mpileup (higher is better; max 225)

I) **final_status**: contains final overall PASS/FAIL status. If failed it will list all the reasons for failing as there may be multiple reasons for failure. Generally variants are passed when they are novel or represent a rare variant, have high quality bases, and high read depth

J) **other_alleles:** string that show any additional alleles and their respective counts at that coordinate for the sample. Data is obtained from the pileup string
Example: C:+2AG_3:4 → shows there are 3 C's and 4 AG insertions at that coordinate in addition to the reference and called variant counts

K) **read_allele_freq**: reports the reference_base:variant_base ratio from the read data
Example: reads:R0.31:V0.66 → shows 31% of our reads match the reference base and 66% of the reads match the variant base

L) **known_variation:** reports all known variation at that coordinate from ENSEMBL's variant effect predictor tool. This report contains entries from dbsnp

M) **dbsnp_match**: for dbsnp overlapping SNVs reports the rs id as well as the reported frequencies for the reference base and the variant allele. Possible entries:
Example: dbsnp:rs139422176:R0.999:V0.001 → rs id with ref freq 0.999 and var freq 0.001
Example: dbsnp:rs139422176:No_freq → rs id with no reported frequencies
Note that this information is obtained from the latest version of dbsnp which may differ from the data reported in the known_variation column which comes from the external tool (variant effect predictor) that may not be up to date with regard to dbsnp

N) **dbsnp_var_allele_freq:** variant allele frequency from the dbsnp match above; separate column added for sorting. Possible values are a number from 0-1 and 'No freq' for entries with no frequency information available from dbsnp

O) **read_depth**: total number of reads aligned to that coordinate

P) **median_quality_score**: median quality score for all bases aligned at that coordinate (max

score is 40 and any score below 30 is low)

Q) **aa_change**: show amino acid change for non-synonomous snvs. Possible entries:
Example: P->L → missense mutation
Example: P->STOP → nonsense mutation
Example: G->N (COMBINED:38122470-38122471) → missense mutation where multiple bases in the codon were mutated. The change reflects the combined effect of both mutations

R) **aa_position:** position of mutated amino acid within the ENSEMBL transcript

S) **aa_length:** total length in amino acids of the ENSEMBL transcript

T) **exon_intron_count:** relative position within the transcript from the variant effect predictor
Example: EXON->10/23 → Variant located in the 10th of 23 total exons
Example: INTRON->1/2 → Variant located in the 1st of 2 total introns

U) **snv_exon_type**: indicates whether the SNV overlaps the splice site or the exon and whether amino acid change is synonomous, missense, or nonsense (SYN, MISSENSE, NONSENSE or SPLICE). Currently splice category means within 10bp of an exon

V) **sift_prediction**: the text classification returned from sift which estimates how damaging the amino acid change is to the protein. Possible values are deleterious or tolerated. 'N/A' entries indicate either synonymous variants, nonsense mutations, splice site variants, or cases where data is not available

W) **sift_score**: number from 0-1 with 0 being most likely damaging

X) **polyphen_prediction**: the text classification returned from polyphen which estimates how damaging the amino acid change is to the protein. Possible values are benign, unknown, possiblydamaging, or probablydamaging. 'N/A' entries indicate either synonymous variants, nonsense mutations, or splice site variants

Y) **polyphen_score**: number from 0-1 with 1 being most likely damaging

Z) **polyphen_info: c**ontains the info needed to run polyphen on the website as we are only running polyphen for the passed snps.
There are two types of entries in this column, mapped and unmapped. This difference is whether we can match our ensembl transcript to a uniprot protein (polyphens input).
i) Most entries are mapped and look like: O35199:258:K:AAA:A:1:+:ENSMUST00000159611
The first field is the uniprot protein id and the second field is the position of the change in the amino acid. Between this and the aa_change column you can run polyphen on the web site.
ii) When an entry is not mapped the data looks like the following:
Example2: NO_SP:346:T:ACA:C:1:+:CCDS20787.1_NO_MAP:MFIFME...REKSYF
The only difference is we input the protein sequence (the last field) instead of the uniprot id

AA) **filter_dbsnp_snv**: indicates whether coordinate matches known dbsnp entry and if it matches what type of match it is. Possible values are:
NOVEL -> no dbsnp match
RARE_ALLELE -> cases where the variant allele frequency is reported to be <2%

RARE_REF -> cases where the reference allele frequency is reported to be <2%
NO_FREQ -> cases where no dbsnp frequency information is given
FAIL -> all other dbsnp matching cases

AB) **filter_common:** indicates whether we've observed the variant previously in unrelated mice. This list is generated from the roughly 1000 mice we've analysed to date. Possible values are NOT_SEEN or PREVIOUSLY_SEEN.

AC) **ensembl**: ENSEMBL gene link

AD) **ens_canonical_trans:** canonical transcript from ENSEMBL. All variant information is relative to this transcript

AE) **protein_domains:** the source and unique identifier of any protein domains at that coordinate from the variant effect predictor.

AF) **uniprot**: uniprot gene name(s). This is a comma delimited list of all uniprot genes that map to this coordinate.

AG) **ccds_name**: ccds name(s). This is a comma delimited list of all ccds names that map to this coordinate

AH) **refseq**: refseq gene name(s). This is a comma delimited list of all refseq names that map to this coordinate

AI) **description:** description of gene function from mgi

AJ) **omim**: link to omim entry if it exists for the human homolog. These entries are linked by gene name to the human homolog so discrepancies in the gene name cause problems

AK) **pubmed:** link to pubmed ID(s) of publications that cite existing variant

AL) **phenotype:** phenotype information from mouse data from MGI if available

AM) **homolog:** human homolog gene

AN) **immgen: i**mmGen expression levels from Clustering Analysis of Tissue Specific Pattern of mRNA Expression in ImmGen Key Populations Dataset

AO) **gnf:** GNF expression levels from Clustering Analysis of Tissue Specific Pattern of mRNA Expression in Mouse Gene Atlas Data

AP) **go**: gene GO term

HUMAN CANCER SAMPLES REPORTS

A) **chr**: chromosome

B) **coord**: genomic coordinate of snv

C) **ref_allele**: reference base at that coordinate

D) **ref_allele_count:** number of times reference allele is observed from the pileup string

E) **var_allele**: variant base call at that coordinate

F) **var_allele_count:** number of times variant allele is observed from the pileup string

G) **gene(hgnc)**: hugo gene name

H) **snv_score**: quality SNV score from samtools mpileup (higher is better; max 225)

I) **clr_score:** Phred like probability score that indicates when tumour is mutant and control matches the reference (higher is better; max 255)

J) **final_status**: contains final overall PASS/FAIL status. If failed it will list all the reasons for failing as there may be multiple reasons for failure. Generally variants are passed when they are novel or represent a rare variant, have high quality bases, and high read depth

K) **tumour_other_alleles:** string that show any additional alleles and their respective counts at that coordinate for the tumour sample. Data is obtained from the pileup string
Example: C:+2AG_3:4 → shows there are 3 C's and 4 AG insertions at that coordinate in addition to the reference and called variant counts

L) **normal_alleles**: string that show all alleles and respective counts at that coordinate for the normal sample in the same format as the above column. Ideally most these alleles will represent the reference allele

M) **read_allele_freq**: reports the reference_base:variant_base ratio from the read data
Example: reads:R0.31:V0.66 → shows 31% of our reads match the reference base and 66% of the reads match the variant base

N) **known_variation:** reports all known variation at that coordinate from ENSEMBL's variant effect predictor tool. Contains entries from dbsnp, cosmic and other population study data

O) **gmaf_1000_genomes:** reports the minor allele frequency and corresponding base from 1000 genomes data. When the minor allele is the reference genome a 'REF' string is present.
Example: 0.069 (G) → 6.9% MAF frequency for variant base G
Example: 0.2143(C:REF) → 21.43% MAF frequency for reference base C

P) **dbsnp_match**: for dbsnp overlapping SNVs reports the rs id as well as the reported frequencies for the reference base and the variant allele. Possible entries:
Example: dbsnp:rs139422176:R0.999:V0.001 → rs id with ref freq 0.999 and var freq 0.001
Example: dbsnp:rs139422176:No_freq → rs id with no reported frequencies

Note that this information is obtained from the latest version of dbsnp which may differ from the data reported in the known_variation column which comes from the external tool

Q) **dbsnp_var_allele_freq:** variant allele frequency from the dbsnp match above; separate column added for sorting. Possible values are a number from 0-1 and 'No freq' for entries with no frequency information available from dbsnp

R) **read_depth**: total number of reads aligned to that coordinate

S) **median_quality_score**: median quality score for all bases aligned at that coordinate (max score is 40 and any score below 30 is low)

T) **aa_change**: show amino acid change for non-synonomous snvs. Possible entries:
Example: P->L → missense mutation
Example: P->STOP → nonsense mutation
Example: G->N (COMBINED:38122470-38122471) → missense mutation where multiple bases in the codon were mutated. The change reflects the combined effect of both mutations

U) **aa_position:** position of mutated amino acid within the ENSEMBL transcript

V) **aa_length:** total length in amino acids of the ENSEMBL transcript

W) **exon_intron_count:** relative position within the transcript from the variant effect predictor
Example: EXON->10/23 → Variant located in the 10th of 23 total exons
Example: INTRON->1/2 → Variant located in the 1st of 2 total introns

X) **snv_exon_type**: indicates whether the SNV overlaps the splice site or the exon and whether amino acid change is synonomous, missense, or nonsense (SYN, MISSENSE, NONSENSE or SPLICE). 'SPLICE' is defined as within 10bp of a coding exon

Y) **snv_class:** snv classification based on tumour and normal alleles. Currently only reporting somatic mutants although loh snps are currently flagged, just not reported.

Z) **sift_prediction**: the text classification returned from sift which estimates how damaging the amino acid change is to the protein. Possible values are deleterious or tolerated. 'N/A' entries indicate either synonymous variants, nonsense mutations, splice site variants, or cases where data is not available

AA) **sift_score**: number from 0-1 with 0 being most likely damaging

AB) **polyphen_prediction**: the text classification returned from polyphen which estimates how damaging the amino acid change is to the protein. Possible values are benign, unknown, possiblydamaging, or probablydamaging. 'N/A' entries indicate either synonymous variants, nonsense mutations, or splice site variants

AC) **polyphen_score**: number from 0-1 with 1 being most likely damaging

AD) **cadd_phred:** phred quality score from CADD with higher being more damaging

AE) **filter_dbsnp_snv**: indicates whether coordinate matches known dbsnp entry and if it matches what type of match it is. Possible values are:
NOVEL -> no dbsnp match
RARE_ALLELE -> cases where the variant allele frequency is reported to be <2%
RARE_REF -> cases where the reference allele frequency is reported to be <2%
NO_FREQ -> cases where no dbsnp frequency information is given
FAIL -> all other dbsnp matching cases

AF) **filter_exac_snv:** exac snv information
reports variant frequency information from the ExAC resource. For each matching coordinates all variants are reports, not just for the variant being reported
Example: A->G:0.000119 → Reports snv of A->G with frequency of 0.000119
Example: C->T:0.034234,C->G:0.000434 → Reports snvs of C->T and C->G

AG) **filter_clinvar_snv:** Clinical significance reported for this variant in NCBI's ClinVar, followed by HGVS expression and corresponding condition(s) when provided.
Note some may have conflicting CLINSIG terms

AH) **ensembl**: ENSEMBL gene link

AI) **cosmic_coord**: entries whose coordinate overlaps existing cosmic mutation. Overlapping entries contain a string containing information from cosmic.
cancer_type:mutant_type^^^base_change^^^strand^^^amino_acid_change^^^genename^^^ histology^^^tumour_type^^^somatic_status

AJ) **ens_canonical_trans:** canonical transcript from ENSEMBL. All variant information is relative to this transcript

AK) **protein_domains:** the source and unique identifier of any protein domains at that coordinate from the variant effect predictor.

AL) **uniprot**: uniprot gene name(s). This is a comma delimited list of all uniprot genes that map to this coordinate.

AM) **ccds**: ccds name(s). This is a comma delimited list of all ccds names that map to this coordinate

AN) **refseq**: refseq gene name(s). This is a comma delimited list of all refseq names that map to this coordinate

AO) **description:** description of gene function from hugo

AP) **omim**: link to omim entry if it exists. Note that these entries are linked by gene name and not coordinates so if there are discrepancies in the gene name this link won't be reported

AQ) **pubmed:** link to pubmed ID(s) of publications that cite existing variant

AR) **phenotype:** phenotype information from mouse data from MGI if available

AS) **cosmic**: link to gene in cosmic (catalogue of somatic mutations in cancer)

AT) **vogelstein**: indicates whether the gene is cited in vogelstein's landmark paper “Cancer genes and the pathways they control”. Matches contains information from the original paper.

AU) **go**: gene GO terms

MOUSE CANCER SAMPLES REPORTS

A) **chr**: chromosome

B) **coord**: genomic coordinate of snv

C) **ref_allele**: reference base at that coordinate

D) **ref_allele_count:** number of times reference allele is observed from the pileup string

E) **var_allele**: variant base call at that coordinate

F) **var_allele_count:** number of times variant allele is observed from the pileup string

G) **gene(mgs)**: mgi gene name

H) **snv_score**: quality SNV score from samtools mpileup (higher is better; max 225)

I) **clr_score:** Phred like probability score that indicates when tumour is mutant and control matches the reference (higher is better; max 255)

J) **final_status**: contains final overall PASS/FAIL status. If failed it will list all the reasons for failing as there may be multiple reasons for failure. Generally variants are passed when they are novel or represent a rare variant, have high quality bases, and high read depth

K) **tumour_other_alleles:** string that show any additional alleles and their respective counts at that coordinate for the tumour sample. Data is obtained from the pileup string
Example: C:+2AG_3:4 → shows there are 3 C's and 4 AG insertions at that coordinate in addition to the reference and called variant counts

L) **normal_alleles**: string that show all alleles and respective counts at that coordinate for the normal sample in the same format as the above column. Ideally alleles will be reference allele

M) **read_allele_freq**: reports the reference_base:variant_base ratio from the read data
Example: reads:R0.31:V0.66 → shows 31% of our reads match the reference base and 66% of the reads match the variant base

N) **known_variation:** reports all known variation at that coordinate from ENSEMBL's variant effect predictor tool. Contains entries from dbsnp, cosmic and other population study data

O) **dbsnp_match**: for dbsnp overlapping SNVs reports the rs id as well as the reported frequencies for the reference base and the variant allele. Possible entries:
Example: dbsnp:rs139422176:R0.999:V0.001 → rs id with ref freq 0.999 and var freq 0.001
Example: dbsnp:rs139422176:No_freq → rs id with no reported frequencies
Note that this information is obtained from the latest version of dbsnp which may differ from the data reported in the known_variation column which comes from the external tool

P) **dbsnp_var_allele_freq:** variant allele frequency from the dbsnp match above; separate column added for sorting. Possible values are a number from 0-1 and 'No freq' for entries with no frequency information available from dbsnp

Q) **read_depth**: total number of reads aligned to that coordinate

R) **median_quality_score**: median quality score for all bases aligned at that coordinate (max score is 40 and any score below 30 is low)

S) **aa_change**: show amino acid change for non-synonomous snvs. Possible entries:
Example: P->L → missense mutation
Example: P->STOP → nonsense mutation
Example: G->N (COMBINED:38122470-38122471) → missense mutation where multiple bases in the codon were mutated. The change reflects the combined effect of both mutations

T) **aa_position:** position of mutated amino acid within the ENSEMBL transcript

U) **aa_length:** total length in amino acids of the ENSEMBL transcript

V) **exon_intron_count:** relative position within the transcript from the variant effect predictor
Example: EXON->10/23 → Variant located in the 10th of 23 total exons
Example: INTRON->1/2 → Variant located in the 1st of 2 total introns

W) **snv_exon_type**: indicates whether the SNV overlaps the splice site or the exon and whether amino acid change is synonomous, missense, or nonsense (SYN, MISSENSE, NONSENSE or SPLICE). 'SPLICE' is defined as within 10bp of a coding exon

X) **snv_class:** snv classification based on tumour and normal alleles. Currently only reporting somatic mutants although loh snps are currently flagged, just not reported.

Y) **sift_prediction**: the text classification returned from sift which estimates how damaging the amino acid change is to the protein. Possible values are deleterious or tolerated. 'N/A' entries indicate either synonymous variants, nonsense mutations, splice site variants, or cases where data is not available

Z) **sift_score**: number from 0-1 with 0 being most likely damaging

AA) **polyphen_prediction**: the text classification returned from polyphen which estimates how damaging the amino acid change is to the protein. Possible values are benign, unknown, possiblydamaging, or probablydamaging. 'N/A' entries indicate either synonymous variants, nonsense mutations, or splice site variants

AB) **polyphen_score**: number from 0-1 with 1 being most likely damaging

AC) **polyphen_info:** the info needed to run polyphen on the website if local version fails:
There are two types of entries in this column, mapped and unmapped. This difference is whether we can match our ensembl transcript to a uniprot protein (polyphens input).
i) Example: O35199:258:K:AAA:A:1:+:ENSMUST0159611 → uniprot id O35199 and pos 258
Between this and the aa_change column you can run polyphen on the web site.
ii) Example2: NO_SP:346:T:ACA:C:1:+:CCDS20787.1_NO_MAP:MFIFME...REKSYF
The only difference is we input the protein sequence (the last field) instead of the uniprot id

AD) **filter_dbsnp_snv**: indicates whether coordinate matches known dbsnp entry and if it matches what type of match it is. Possible values are:
NOVEL -> no dbsnp match
RARE_ALLELE -> cases where the variant allele frequency is reported to be <2%
RARE_REF -> cases where the reference allele frequency is reported to be <2%
NO_FREQ -> cases where no dbsnp frequency information is given
FAIL -> all other dbsnp matching cases

AE) **ensembl**: ENSEMBL gene link

AF) **ens_canonical_trans:** canonical transcript from ENSEMBL. All variant information is relative to this transcript

AG) **protein_domains:** the source and unique identifier of any protein domains at that coordinate from the variant effect predictor.

AH) **uniprot**: uniprot gene name(s). This is a comma delimited list of all uniprot genes that map to this coordinate.

AI) **ccds**: ccds name(s). This is a comma delimited list of all ccds names that map to this coordinate

AJ) **refseq**: refseq gene name(s). This is a comma delimited list of all refseq names that map to this coordinate

AK) **description:** description of gene function from hugo

AL) **omim**: link to omim entry if it exists. Note that these entries are linked by gene name and not coordinates so if there are discrepancies in the gene name this link won't be reported

AM) **pubmed:** link to pubmed ID(s) of publications that cite existing variant

AN) **phenotype:** phenotype information from mouse data from MGI if available

AO) **homolog:** human homolog gene

AP) **immgen: i**mmGen expression levels from Clustering Analysis of Tissue Specific Pattern of mRNA Expression in ImmGen Key Populations Dataset

AQ) **gnf:** GNF expression levels from Clustering Analysis of Tissue Specific Pattern of mRNA Expression in Mouse Gene Atlas Data

AR) **go:** gene GO terms