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 indels 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).

INDEL GROUPINGS

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

a) DEL NOVEL entries
   -Novel deletions (i.e. no matching dbsnp entry) that either overlap an ensembl exon or a splice site. Ordered by descending read depth.
b) INS NOVEL entries
   - Novel insertions (i.e. no matching dbsnp entry) that either overlap an ensembl exon or a splice site. Ordered by descending read depth.
c) DEL RARE entries
   -Known deletions (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.
d) INS RARE entries
   -Known insertions (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.
e) LOW_PRIORITY entries
   -Remaining indels that don't fall into a) or b) or c). These entries are sorted by chromosome and then coordinate

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 Sample (non cancer) reports are described on pages 2-4
Mouse Sample (non cancer) reports are described on pages 5-7
Human Cancer Sample reports are described on pages 8-10
Mouse Cancer Sample reports are described on pages 11-13

HUMAN SAMPLES (NON-CANCER) REPORTS

A) **chr**: chromosome

B) **start_coord**: genomic start coordinate of indel (for insertions this is the reference base coordinate before the insertion while for deletions this is the first base deleted)

C) **end_coord:** genomic end coordinate of indel (same as start_coord for insertion and the last base deleted for deletions)

D) **ref_allele:** reference allele (deleted bases for deletions and N/A for insertions)

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

F) **var_type**: DEL for deletion and INS for insertion

G) **var_length**: length of variant (# bases deleted for deletions and # bases inserted for insertions)

H) **var_allele**: variant allele (inserted bases for insertion and deleted base for deletions)
Example: +AG → AG insertion
Example: -GTG → GTG deletion

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

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

K) **var_score**: indel quality score from samtools mpileup (higher is better; max 225)

L) **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

M) **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

N) **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

O) **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

P) **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

Q) **dbsnp_match**: for dbsnp overlapping indels 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

R) **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

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

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

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) **exon_overlap**: Indicates whether the indel overlaps the splice site or the exon (EXON or SPLICE). To qualify for splice category the coordinate must be within 10bp of a coding exon

Y) **filter_dbsnp_indel**: 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

Z) **filter_exac_indel:** exac indel 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: -G:0.000119 → Reports deletion of G with frequency of 0.000119
Example: -GA:0.034234,+A:0.000434 → Reports deletion and insertion

AA) **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

AB) **ensembl**: ENSEMBL gene link

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

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

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

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

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

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

AI) **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

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

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

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

AM) **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

AN) **go**: gene GO terms

MOUSE SAMPLES (NON-CANCER) REPORTS

A) **chr**: chromosome

B) **start_coord**: genomic start coordinate of indel (for insertions this is the reference base coordinate before the insertion while for deletions this is the first base deleted)

C) **end_coord:** genomic end coordinate of indel (same as start_coord for insertion and the last base deleted for deletions)

D) **ref_allele:** reference allele (deleted bases for deletions and N/A for insertions)

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

F) **var_type**: DEL for deletion and INS for insertion

G) **var_length**: length of variant (# bases deleted for deletions and # bases inserted for insertions)

H) **var_allele**: variant allele (inserted bases for insertion and deleted base for deletions)
Example: +AG → AG insertion
Example: -GTG → GTG deletion

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

J) **gene(mgi)**: mgi gene name

K) **var_score**: indel score from samtools mpileup (higher is better; max 225)

L) **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

M) **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

N) **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

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

P) **dbsnp_match**: for dbsnp overlapping indels 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

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_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) **exon_overlap**: Indicates whether the indel overlaps the splice site or the exon (EXON or SPLICE). To qualify for splice category the coordinate must be within 10bp of a coding exon

X) **filter_dbsnp_indel**: 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

Y) **filter_common_indel:** 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.

Z) **ensembl**: ENSEMBL gene link

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

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

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

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

this coordinate

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

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

AG) **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

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

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

AJ) **homolog:** human homolog gene

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

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

AM) **go**: gene GO term

HUMAN CANCER SAMPLES REPORTS

A) **chr**: chromosome

B) **start_coord**: genomic start coordinate of indel (for insertions this is the reference base coordinate before the insertion while for deletions this is the first base deleted)

C) **end_coord:** genomic end coordinate of indel (same as start_coord for insertion and the last base deleted for deletions)

D) **ref_allele:** reference allele (deleted bases for deletions and N/A for insertions)

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

F) **var_type**: DEL for deletion and INS for insertion

G) **var_length**: length of variant (# bases deleted for deletions and # bases inserted for insertions)

H) **var_allele**: variant allele (inserted bases for insertion and deleted base for deletions)
Example: +AG → AG insertion
Example: -GTG → GTG deletion

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

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

K) **var_score**: indel quality score from samtools mpileup (higher is better; max 225)

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

M) **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

N) **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

O) **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

P) **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

Q) **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

R) **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

S) **dbsnp_match**: for dbsnp overlapping indels 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

T) **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

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

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

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

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

Y) **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

Z) **exon_overlap**: Indicates whether the indel overlaps the splice site or the exon (EXON or SPLICE). To qualify for splice category the coordinate must be within 10bp of a coding exon

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

AB) **filter_dbsnp_indel**: 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_indel:** exac indel 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: -G:0.000119 → Reports deletion of G with frequency of 0.000119
Example: -GA:0.034234,+A:0.000434 → Reports deletion and insertion

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) **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

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

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

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

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

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

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

AM) **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

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

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

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

AQ) **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.

AR) **go**: gene GO terms

MOUSE CANCER SAMPLES REPORTS

A) **chr**: chromosome

B) **start_coord**: genomic start coordinate of indel (for insertions this is the reference base coordinate before the insertion while for deletions this is the first base deleted)

C) **end_coord:** genomic end coordinate of indel (same as start_coord for insertion and the last base deleted for deletions)

D) **ref_allele:** reference allele (deleted bases for deletions and N/A for insertions)

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

F) **var_type**: DEL for deletion and INS for insertion

G) **var_length**: length of variant (# bases deleted for deletions and # bases inserted for insertions)

H) **var_allele**: variant allele (inserted bases for insertion and deleted base for deletions)
Example: +AG → AG insertion
Example: -GTG → GTG deletion

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

J) **gene(mgi)**: mgi gene name

K) **var_score**: indel quality score from samtools mpileup (higher is better; max 225)

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

M) **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

N) **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

O) **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

P) **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

Q) **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

R) **dbsnp_match**: for dbsnp overlapping indels 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

S) **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

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

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

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

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

X) **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

Y) **exon_overlap**: Indicates whether the indel overlaps the splice site or the exon (EXON or SPLICE). To qualify for splice category the coordinate must be within 10bp of a coding exon

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

AA) **filter_dbsnp_indel**: 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) **ensembl**: ENSEMBL gene link

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

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

AE) **uniprot**: uniprot gene name(s). This is a comma delimited list of all uniprot genes that

map to this coordinate.

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

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

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

AI) **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

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

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

AL) **homolog:** human homolog gene

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

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

AO) **go:** gene GO terms