SNPs/InDels

We used GATK to call single nucleotide polymorphisms (SNPs) and small insertions and deletions (small InDels) from BAM files and used ANNOVAR to annotate variants.

1 Summary of result files

1.1 Files in the folder 'Annotation'

- ***.annovar.hg19_multianno.xls.gz**

  This file contains annotation information for SNVs or InDels in gzip-compressed format. After decompressed, the .xls file can be viewed with a plain text editor or Excel.

1.2 Files in the folder 'Vcf'

- ***.snp.vcf.gz**

  This file contains SNV or InDels calls in bgzip-compressed VCF (Variant Call Format) format.

  VCF is a flexible and extendable line-oriented text format developed by the 1000 Genomes Project for releases of SNVs, indels, copy number variants and structural variants discovered by the project. VCF format is described in VCFv4.2.pdf (Download link: https://samtools.github.io/hts-specs/VCFv4.2.pdf)

1.3 Files in the current folder 'SNP'

- ***_function.stat.xls (* denotes snp or indel)**

  This file shows the numbers of SNPs/InDels localized in various types of genomic elements in each sample.

- ***_features.xls (* denotes snp or indel)**

This files contains statistical results of SNPs/InDels.

Note: As many databases in ANNOVAR are processed by the allele-splitting and left-normalization pipeline described in http://annovar.openbioinformatics.org/en/latest/articles/VCF/ (including dbSNP, 1000Genomes, ExAC, ESP6500, ClinVar database), so the annotation workflow at Novogene is: (1) split VCF lines so that each line contains one and only one variant; (2) left-normalize all VCF lines; (3) annotate by ANNOVAR.

2 Explanation on Headers of Files

There are multiple fields in each file. Detailed explanations for each filed are listed as follows:

- ***.annovar.hg19_multianno.xls.gz**

Note: Annotation information includes fiveparts: Chrosomal regions and gene structures related to this variation (1-20), Database annotation (21-39), Functional prediction (40-51), Basic information on the variation (52-56), and Gene function and pathway annotation (57-67).

The first part shows information of chromosomal regions and gene structures related to the variant.

(1) CHROM: Chromosome ID.

(2) POS: The position of the variant on chromosomes. The value refers to the position of the first base in the REF sting.

(3) ID: The rs number of the variant in dbSNP.

(4) REF: Reference base(s).

(5) ALT: Alternate base(s). Comma separated list of alternate non-reference alleles called on at least one of the samples.

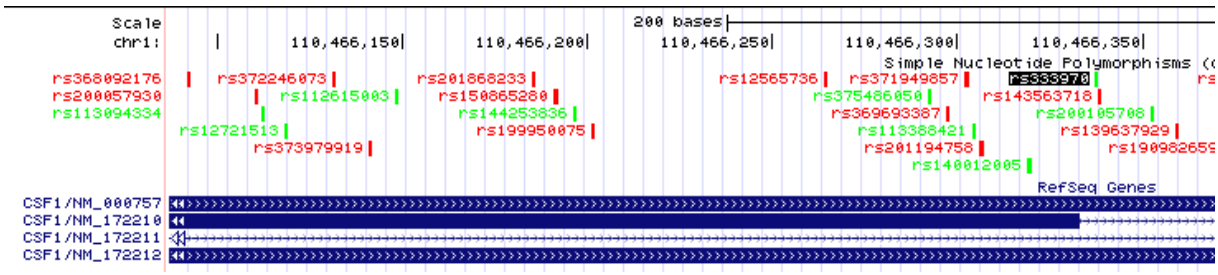
(6) QUAL: Quality value for the variant. Phred-scaled quality score for the assertion made in ALT. i.e. -10log10 prob (call in ALT is wrong).

(7) FILTER: Filter status, PASS if the position has passed all filters.

(8) GeneName: Names of genes in which this variant is located according to the refGene annotations.

(9) Func: This field tells whether the variant hits exons or hits intergenic regions, or hits introns, or hits a non-coding RNA genes. The value of this field takes the following precedence: exonic = splicing > ncRNA > UTR5/UTR3 > intronic > upstream/downstream > intergenic. Notes: 1. When a variant hits different genes or transcripts, the variant may fit multiple functional categories, and then the precedence mentioned above is used to decide what function to print out; 2. The 'exonic' here refers only to coding exonic portion, but not UTR portion, as there are two keywords (UTR5, UTR3) that are specifically reserved for UTR annotations; 3. If a variant is located in both 5'UTR and 3'UTR region (possibly for two different genes), then the 'UTR5,UTR3' will be printed as the output; 4. 'splicing' in ANNOVAR is defined as variant that is within 2bp away from an exon/intron boundary by default; 5. 'splicing' in ANNOVAR only refers to the 2bp in the intron that is close to an exon; 6. The term 'upstream' and 'downstream' is defined as 1kb away from transcription start site or transcription end site, respectively, taking in account of the strand of the mRNA. If a variant is located in both downstream and upstream region (possibly for 2 different genes), then the 'upstream, downstream' will be printed as the output.

(10) Gene: The transcript name(s). If a variant has 'intergenic' in 'Func' field, this field will give the two neighboring transcripts. If a variant hits multiple transcripts with different functional categories, only transcript names in accordance with the value of 'Func' field will be output. For example, rs333970 hits the exonic, splicing, intronic, exonic of the four transcripts of gene CSF1, the 'Func' value will be 'exonic; splicing' and the 'Gene' value will be 'NM_000757, NM_172210, NM_172212' (NM_172211 will be ignored).



(11) GeneDetail: Description of the sequence change in UTR, splicing, ncRNA_splicing or intergenic region. If 'Func' is 'exonic; splicing' or 'splicing', this field gives the sequence change in splicing region(s); for example, NM_172210:exon6:c.1090+5C>A, NM_172210 is the transcript identifier; exon6:c.1090+5C>A is the sequence change and means that this C>A substitution is at the fifth base downstream from the 6th exon (1090 is the end position of the 6th exon of the cDNA). If 'Func' is 'intergenic', this field gives the distance to the neighboring transcripts, such as 'dist=1366; dist=22344'. If 'Func' is 'UTR*', this field gives the sequence change in UTR; for example, NM_198576:c.*19C>T means that this C>T substitution is at the 19th base downstream from stop codon on NM_198576.

(12) ExonicFunc: This field tells the functional consequences of the variant (possible values include: missense SNV, synonymous SNV, frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, frameshift block substitution, nonframshift block substitution, stopgain, stoploss, unknown).

(13) AAChange: This field tells the amino acid changes as a result of the exonic variant. Only exonic variants have information in this field, i.e. when 'Func' is 'exonic' or 'exonic; splicing', this field gives the amino acid change in each related transcript. For example, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L is gene name; NM_001039775 is the transcript identifier; exon2 means this variant is on the second exon of NM_001039775; c.C2768T is the sequence change and means that this C>T substitution is at the 2,768 position on the cDNA; p.P923L is the amino acid change and means that the 923 amino acid on protein is changed from Pro to Leu due to this variant. Another example, NADK:NM_001198995:exon10:c.1240_1241insAGG:p.G414delinsEG, c.1240_1241insAGG is the sequence change and means that there is a 3bp insertion between position 1,240 and 1,241 on the cDNA; p.G414delinsEG is the amono acid change and means that Gly at the 414th amino acid on protein is changed to Glu-Gly.

(14) Gencode: The transcript name(s) in which this variant is located according to Gencode gene definitions.

(15) cytoband: This field gives the Giemsa-stained chromosomes bands. When a variant spans multiple bands, they will be connected by a dash (for example, 1q21.1-q23.3).

(16) wgRna: Gene names of snoRNAs and microRNAs based on the miRBase Release and snoRNABase.

(17) targetScanS: The targetScanS annotation database offered by UCSC gives conserved mammalian microRNA regulatory target sites for conserved microRNA families in the 3' UTR regions of Refseq Genes, as predicted by TargetScanHuman 5.1. This field tells whether the variant disrupts predicted microRNA binding sites. The output consists of a score and a name. The score of target site ranges from from 0-1000; the smaller the score, the target site is more confident. The name shows the name of microRNA acting on the target. For instance, “Score=62;Name=KRAS:miR-181:1” means that the predicted target site is within the UTR3 region of gene KRAS and that the microRNA named

miR-181:1 acts on this target site.

(18) tfbsConsSites: This field tells whether the variant disrupts transcription factor binding sites conserved in the human/mouse/rat alignment and gives the Score and Name annotation for the transcription factor binding sites. The score represents the normalized score. The name represents binding site motif name. For example, Score=765;Name=V$PAX5_02. Users can investigate what transcription factors may recognize this motif using many online resources, for example, MSigDB provides gene list that recognize these motifs, see for example http://www.broadinstitute.org/gsea/msigdb/cards/V$PAX5_02.

(19) genomicSuperDups: This field tells whether the variant hits segmental duplications in reference genome. Variants that are mapped to segmental duplications are most likely sequence alignment errors and should be treated with extreme caution. The 'Score' field in output is the sequence identity ranging from 0 to 1 between two genomic segments. The 'Name' field represents the other "matching" segments in genome. For example, 'Score=0.994828; Name=chr19:60000' means that the fragment at the position of chr19:60000 is homologous to the fragment containing this variant, and the sequence identity is 0.994828. Note, for a region to be included in the segmental duplications, at least 1 Kb of the total sequence (containing at least 500bp of non-RepeatMasked sequence) had to align and a sequence identity of at least 90% was required.

(20) Repeat: This field tells whether the variant hits interspersed repeats and low complexity DNA sequences output by RepeatMasker program, such as SINE, LINE and Simple repeats. For example, 'Score=180;Name=1385:(CACCC) n(Simple_repeat)', 180 is the score of the repeat, (CACCC) n is the name of the repeat, 'Simple_repeat' is type of repeat. Note, variants mapped to repeats are likely to be false and should be treated with extreme caution.

The second item is database annotataion—There are a great number of common polymorphism sites in human population, while many deleterious variants are rare or low-frequency. This part gives the allele frequency and clinical information for each variant.

(21) avsnp147: The RS number of the variant in dbSNP (build 147).

(22) cosmic70: The Cosmic id in Catalogue Of Somatic Mutations In Cancer (COSMIC) database.

(23) clinvar_20150330: The ClinVar database archives and aggregates information about relationships among variations and human health. An example is 'CLINSIG=probable-non-pathogenic;CLNDBN=not_specified;CLNREVSTAT=single;CLNACC=RCV000116272.2;CLNDSDB=MedGen;CLNDSDBID=CN169374'. CLINSIG refers to Variant Clinical Significance, including unknown, untested, non-pathogenic, probable-non-pathogenic, probable-pathogenic, pathogenic, drug-response, histocompatibility, other; CLINDBN refers to variant disease name; CLNREVSTAT refers to ClinVar Review Status, mult-Classified by multiple submitters, single-Classified by single submitter, not-Classified by submitter, exp-Reviewed by expert panel, prof-Review by professional society; CLINACC refers to Variant Accession and Versions; CLNDSDB refers to variant disease database name; CLNDSDBID refers to variant disease database ID.

(24) gwasCatalog: This field tells whether this variant was previously reported to be associated with diseases or traits in genome-wide association studies. It lists the disease names related to this variation. "." means this variation has not been reported by published GWAS study.

(25) 1000g2015aug_eas: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in East Asian population.

(26) 1000g2015aug_sas: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in South Asian population.

(27) 1000g2015aug_eur: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in European population.

(28) 1000g2015aug_afr: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in African population.

(29) 1000g2015aug_amr: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in Admixed American population.

(30) 1000g2015aug_all: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in ALL population.

(31) esp6500siv2_all: The ESP is a NHLBI funded exome sequencing project aiming to identify genetic variants in exonic regions from over 6000 individuals, including healthy ones as well as subjects with different diseases. This field gives alternative allele frequency for the variant in ESP.

(32) ExAC_ALL: ExAC is short for Exome Aggregation Consortium. The data set spans 60,706 unrelated individuals and should serve as a useful reference set of allele frequencies for severe disease studies. Currently supported population groups include ALL, AFR (African), AMR (Admixed American), EAS (East Asian), FIN (Finnish), NFE (Non-finnish European), OTH (other) and SAS (South Asian). ExAC_ALL gives alternative allele frequency for the variation in ALL ExAC samples.

(33) ExAC_AFR: The alternative allele frequency for the variation in ExAC for African population.

(34) ExAC_AMR: The alternative allele frequency for the variation in ExAC for Admixed American population.

(35) ExAC_EAS: The alternative allele frequency for the variation in ExAC for East Asian population.

(36) ExAC_FIN: The alternative allele frequency for the variation in ExAC for Finnish population.

(37) ExAC_NFE: The alternative allele frequency for the variation in ExAC for Non-Finnish European population.

(38) ExAC_OTH: The alternative allele frequency for the variation in ExAC for other population.

(39) ExAC_SAS: The alternative allele frequency for the variation in ExAC for South Asian population.

The third item is functional prediction—These annotations can help to evaluate deleteriousness of a variation. Note 1: SIFT, Polyphen2, MutationTaster, LRT, MutationAssessor and FATHMM are similar and all predict whether an amino acid substitution affects protein function; only coding variants have these annotations. Note 2: phyloP, SiPhy, gerp++ and CADD are similar and predict the conservation level of the site; these types of 'conservation scores' only consider conservation level at the current base, and they do not care about the actual nucleotide identity, so synonymous and non-synonymous variants at the same site will be scored as the same; these scores are used for finding functionally important sites, so variants that confer increased susceptibility may be scored well.

(40) SIFT: SIFT annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction; the scores and predictions are separated by comma. There are two possible predictions: D (Deleterious, score<=0.05)); T (Tolerated, score>0.05).

(41) Polyphen2_HVAR: PolyPhen 2 (dbNSFP version 3.0a) annotation based on HumanVar database. This annotation should be used for diagnostics of Mendelian diseases. The annotation consists of score and categorical prediction. There are three possible predictions: D (Porobably damaging, score>=0.909), P (possibly damaging, 0.447<=score<=0.909), B (benign, score<=0.446).

(42) Polyphen2_HDIV: PolyPhen 2 (dbNSFP version 3.0a) annotation based on HumanDiv database. This annotation should be used when evaluating rare alleles at loci potentially involved in complex phenotypes, dense mapping of regions identified by genome-wide association studies, and analysis of natural selection from sequence data. The annotation consists of score and categorical prediction. There are three possible predictions: D (Porobably damaging, score>=0.957), P (possibly damaging, 0.453<=score<=0.956), B (benign, score<=0.452).

(43) MutationTaster: MutationTaster annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are four possible predictions: 'A' (Disease_causing_automatic), 'D (Disease_causing), 'N' (Polymorphism), 'P' (Polymorphism_automatic'). D and N are categorized by only score, while A and P are categorized by score and other information (if nonsynonymous SNV leads to stop-gain, the variation will be predicted an 'A'; if all three genotypes of nonsynonymous SNV has frequency information in HapMap, the variation will be predicted a 'P'). So, both A and D should be considered deleterious.

(44) LRT: LRT annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are three possible predictions: D (Deleterious), N (Neutral), U (Unknown).

(45) MutationAssessor: MutationAssessor annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are four possible predictions: H (high), M (medium), L (low), N (neutral). H/M means functional and L/N means non-functional.

(46) FATHMM: FATHMM annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are two possible predictions: D (Deleterious, score<=-1.5)); T (Tolerated, score>-1.5).

(47) phyloP7way_vertebrate: PhyloP score (dbNSFP version 3.0a) based on the whole genome alignment of 7 vertebrates. Generally the higher the score, the more conserved the site.

(48) phyloP20way_mammalian: PhyloP score (dbNSFP version 3.0a) based on the whole genome alignment of 20 mammals.

(49) SiPhy_29way_logOdds: SiPhy score (dbNSFP version 3.0a) based on the whole genome alignment of 29 mammals genomes. The larger the score is, the more conserved the site.

(50) gerp++gt2: GERP++ scores for all mutations with GERP++>2 in human genome, as this threshold is typically regarded as evolutionarily conserved and potentially functional. Variants with '.' in this field should be considered not conserved. The larger the score is, the more conserved the site.

(51) CADD: CADD (Combined Annotation Dependent Depletion) is a score that is based on SVM on multiple other scores. It assigns a score to each possible mutation in the human genome including non-coding and coding variants. In the output, the comma-delimited values are raw scores and phred-scaled scores. "." stands for CADD score<10. For phred-scaled scores, 10 means 10% percentile highest scores, 20 means 1% percentile highest scores, and 30% means 0.1% percentile highest scores. CADD official website suggests 15 as a cutoff; in published studies, 10 or 15 is used as a cutoff.

The fourth item is basic information on the variation —This part shows the detail information of variation sites, including allelic depths , types of DNA base before and after mutation, and genotype information, et al. those information play a critical role in pedigree analysis.

(52) INFO: Information about this variation from variant calling software. INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data].

(53) FORMAT: The FORMAT field specifies the data types and order (colon-separated alphanumeric String). This is followed by one field per sample, with the colon-separated data

corresponding to the types specified in the FORMAT.

GT: genotype, encoded as allele values separated by either of / or |. The allele values are 0 for the reference allele (what is in the Ori_REF field), 1 for the first allele listed in Ori_ALT, 2 for the second allele list in Ori_ALT and so on. 0/0 and 1/1 represent homozygous. 0/1 represents heterozygous. '.' means that a call cannot be made for a sample at a given locus.

AD: Allelic depths for the ref and alt alleles in the order listed (Allelic depths).

DP: Approximate read depth (reads with MQ=255 or with bad mates are filtered).

GQ: Genotype Quality.

PL: Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification.

(54) SampleID: The colon-separated data in this sample corresponding to the types specified in the FORMAT.

(55) Ori_REF: The reference allele (what is in the REF field) in VCF file. According to the annotation workflow at Novogene (mentioned above), for InDel, the allele in "REF" field in this file may be different from (usually shorter than) the "REF" in VCF file.

(56) Ori_ALT: The alternative allele(s) (what is in the ALT field) in VCF file. In this file, the allele in "ALT" field corresponds to one alelle in "Ori_ALT" field; according to the annotation workflow at Novogene (mentioned above), for InDel, the allele in "ALT" field may be different from (usually shorter than) the corresponding allele in the "Ori_ALT" field.

The fifth item is gene function and pathway annotation: These annotations are for genes containing this variation.

(57) OMIM: Annotation from Online Mendelian Inheritance in Man (OMIM).

(58) GWAS_Pubmed_pValue: Annotation from the NHGRI-EBI GWAS Catalog. The value is like 'pubmedID(p-value);pubmedID(p-value)'. 'pubmedID' is PubMed ID of publication of the study which reported the association between the variation and disease. 'p-value' is the corresponding p-value in the publication.

(59) HGMD_ID_Diseasename: Annotation from the Human Gene Mutation Database (HGMD®). The value is like 'ID_HGMD(Disease_name);ID_HGMD(Disease_name)'. ID_HGMD is HGMD internal identifier. Disease_name is the name for the disease or condition associated with the mutation.

(60) HGMD_mutation: Annotation from the Human Gene Mutation Database (HGMD®). The value is infomation about this variant.

(61-63) GO_BP, GO_CC, GO_MF: Annotation from Gene Ontology. BP is Biological Process; CC is cellular component; MF is molecular function.

(64) KEGG_PATHWAY: Annotation from KEGG PATHWAY Database.

(65) PID_PATHWAY: Annotation from PID (Pathway Interaction Database).

(66) BIOCARTA_PATHWAY: Annotation from BioCarta.

(67) REACTOME_PATHWAY: Annotation from Reactome Pathway Database.

- **snp_function.stat.xls**

(1) Sample: Sample name

(2) CDS: the number of SNPs in exonic region

(3) synonymous_SNP: a single nucleotide change that does not cause an amino acid change

(4) missense_SNP: a single nucleotide change that cause an amino acid change

(5) stopgain: a nonsynonymous SNPs that lead to the immediate creation of stop codon

(6) stoploss: a nonsynonymous SNPs that lead to the immediate elimination of stop codon

(7) unknown: unknown function (due to various errors in gene structure annotations)

(8) intronic: the number of SNPs in intronic region

(9) UTR3: the number of SNPs in 3'UTR region

(10) UTR5: the number of SNPs in 5'UTR region

(11) splicing: the number of SNPs in 2bp splicing junction region

(12) ncRNA_exonic: the number of SNPs in non-coding RNA exonic region

(13) ncRNA_intronic: the number of SNPs in non-coding RNA intronic region

(14) ncRNA_splicing: the number of SNPs in 2bp splicing junction of non-coding RNA

(15) upstream: the number of SNPs in the 1kb upstream region of transcription start site

(16) downsteam: the number of SNPs in the 1kb downstream region of transcription termination site

(17) intergenic: the number of SNPs in intergenic region

(18) Total: the total number of SNPs

- **snp_features.xls**

(1) Sample: sample name

(2) Total: the total number of variants

(3) Het: the number of heterozygotes

(4) Hom: the number of homozygotes

(5) transition(ts): the number of transitions

(6) transversion(tv): the number of transversions

(7) ts/tv: the number of transitions divided by the number of transversions

(8) dbSNP percentage: the number of SNPs that have been reported in dbSNP database divided by the total number of called SNPs

(9) novel: the number of SNPs that have not been reported in dbSNP

(10) novel ts: the number of ts SNPs that have not been reported in dbSNP

(11) novel tv: the number of tv SNPs that have not been reported in dbSNP

(12) novel ts/tv: novel ts divided by novel tv

- **indel_function.stat.xls**

(1) Sample: Sample name

(2) CDS: the number of InDels in coding region

(3) frameshift_deletion: a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence

(4) frameshift_insertion: an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence

(5) nonframeshift_deletion: a deletion that does not cause frameshift changes

(6) nonframeshift_insertion: an insertion that does not cause frameshift changes

(7) stopgain: an insertion or a deletion that leads to the immediate creation of stop codon at the variant site

(8) stoploss: an insertion or a deletion that leads to the immediate elimination of stop codon at the variant site

(9) unknown: unknown function (due to various errors in the gene structure definition in the database file)

(10) intronic: the number of InDels in intronic region

(11) UTR3: the number of InDels in 3'UTR region

(12) UTR5: the number of InDels in 5'UTR region

(13) splicing: the number of InDels within 2bp away from an exon/intron boundary

(14) ncRNA_exonic: the number of InDels in exonic region of non-coding RNAs

(15) ncRNA_intronic: the number of InDels in intronic region of non-coding RNAs

(16) ncRNA_splicing: the number of InDels within 2bp away from an exon/intron boundary of non-coding RNAs

(17) upstream: the number of InDels within 1kb away from transcription start site

(18) downstream: the number of InDels iwithin 1kb away from transcription ending site

(19) intergenic: the number of InDels in intergenic region

(20) Total: the total number of InDels

- **indel_features.xls**

(1) Sample: sample name

(2) Total: the total number of variants

(3) Het: the number of heterozygotes

(4) Hom: the number of homozygotes

(5) dbSNP percentage: the number of INDELs that have been reported in dbSNP database divided by the total number of called INDELs

(6) novel: the number of INDELs that have not been reported in dbSNP