Annotation of SNPs and indels from 1000 bulls project run 2.0

November 2012

Paul Stothard and Xiaoping Liao

Department of Agricultural, Food and Nutritional Science (AFNS)
1400 College Plaza
8215 - 112 Street
Edmonton, Alberta
Canada T6G 2C8

Input data

- 26,716,895 SNPs
- 1,555,106 indels

Annotation approach

- NGS-SNP (Grant et al., 2011)
  - annotate_SNPs.pl script for SNPs
  - annotate_INDELs.pl script for indels
- The following databases were used for annotation:
  - Ensembl release 68 (includes variants from dbSNP version 133).
  - Entrez Gene and UniProt used for some annotation fields (October 2012 versions).

Grant JR, Arantes AS, Liao X, Stothard P (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics 27:2300-2301.

Annotation approach

- Only “known” transcripts (have match to bovine transcript or protein in UniProtKB/Swiss-Prot or RefSeq) were considered when predicting SNP and indel consequences.
- Variants affecting pseudogenes are given appropriate functional classes (e.g. “nc_transcript_variant” for “non-coding transcript variant”).

Grant JR, Arantes AS, Liao X, Stothard P (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics 27:2300-2301.

Function class

- Each variant is assigned a function class using the Ensembl API.
- The function classes are defined relative to the reference genome sequence. For example:
  - stop_lost means that the non-reference allele in the input variant leads to the loss of a stop codon annotated on the reference genome.
  - stop_gained indicates that the alternative allele adds a stop codon to the coding region of a transcript annotated on the reference genome.

Function class

- A single variant can be assigned multiple function classes (also called consequences) due to the presence of multiple overlapping transcripts or genes, and due to overlap among the function classes. For example:
  - a SNP can be located in the 3'UTR of one transcript and translated region of another. Thus this SNP could have two consequence types: 3_prime_UTR_variant and missense_variant.
  - a SNP in a start codon can be assigned all of the following function classes: coding_sequence_variant, missense_variant, and initiator_codon_variant.

Function class

- One consequence is reported for each variant.
- When there are multiple consequences for a variant the one considered to be of the highest importance is reported (e.g. missense_variant will be reported over synonymous_variant).

Annotation fields for SNPs

- 32 annotation fields for SNPs are provided by annotate_SNPs.pl script.
- The “Comments” annotation field includes several (>20) key-value pairs providing more information about certain variant types such as SIFT score (for missense_variant SNPs), and length of protein sequence lost (for stop_gained SNPs).
- The “Model_Annotations” annotation field includes up to five key-value pairs providing information related to the human orthologue of the bovine gene containing the variant, if applicable. Examples of information provided in this field include KEGG pathway names, and phenotypes associated with the orthologue.
- The full list of fields available at:
  - http://www.ualberta.ca/~stothard/downloads/NGS-SNP/annotate_SNPs.html

Annotation fields for indels

- 22 annotation fields for indels are provided by annotate_INDELs.pl script.
- “Comments” and “Model_Annotations” fields provide additional information in the form of key-value pairs.
- Full list of fields available at:
  - http://www.ualberta.ca/~stothard/downloads/NGS-SNP/annotate_INDELs.html

Function class values

- For indels and SNPs the consequence type of the variant is given in the “Functional_Class” field.
- The values that appear in this field are defined by Ensembl and the Sequence Ontology (SO) project.

http://www.sequenceontology.org

Output files

- SNPs
  - Tab-delimited annotated SNPs.
  - Tab-delimited flanking sequence of each SNP for genotyping assay design or validation.
- Indels
  - Tab-delimited annotated indels.
  - Tab-delimited flanking sequence of each indel for genotyping assay design or validation.

Summary of results for SNPs

Known vs. novel SNPs

Indel type	Count
Known	5,751,736
Novel	20,965,159
All	26,716,895

“Known” is used here to describe input variants where the variant and all of its alleles exist already in the reference database.

Numbers of SNPs in each function class

SNP consequence type	Count
intergenic_variant	18,589,752
intron_variant	6,579,341
upstream_gene_variant	673,577
downstream_gene_variant	592,605
missense_variant	99,089
synonymous_variant	81,730
3_prime_UTR_variant	61,009
splice_region_variant	18,439
5_prime_UTR_variant	10,989
stop_gained	4,155
splice_donor_variant	2,268
non_coding_exon_variant	1,789
splice_acceptor_variant	1,628
initiator_codon_variant	184
stop_lost	118
coding_sequence_variant	101
stop_retained_variant	57
mature_miRNA_variant	52
nc_transcript_variant	12
Total	26,716,895

Sample annotated SNP

Field number	Field name	Value
1	CHROM	1
2	POS	145114963
3	ID	.
4	REF	T
5	ALT	C
6	QUAL	.
7	FILTER	.
8	INFO	.
9	Functional_Class	missense_variant
10	Chromosome	1

See http://www.ualberta.ca/~stothard/downloads/NGS-SNP/annotate_SNPs.html

Sample annotated SNP cont.

Field number	Field name	Value
11	Chromosome_Position	145114963
12	Chromosome_Strand	forward
13	Chromosome_Reference	T
14	Chromosome_Reads	C
15	Gene_Description	Bos taurus integrin, beta 2 (complement component 3 receptor 3 and 4 subunit) (ITGB2), mRNA. [Source:RefSeq mRNA;Acc:NM_175781]
16	Ensembl_Gene_ID	ENSBTAG00000017060
17	Entrez_Gene_Name	ITGB2
18	Entrez_Gene_ID	281877
19	Ensembl_Transcript_ID	ENSBTAT00000022687
20	Transcript_SNP_Position	488

Sample annotated SNP cont.

Field number	Field name	Value
21	Transcript_SNP_Reference	A
22	Transcript_SNP_Reads	G
23	Transcript_To_Chromosome_Strand	reverse
24	Ensembl_Protein_ID	ENSBTAP00000022687
25	UniProt_ID	ITB2_BOVIN
26	Amino_Acid_Position	128
27	Overlapping_Protein_Domains	superfamily;pfam IPR002369 Integrin_bsu_N;smart IPR002369 Integrin_bsu_N;pirsf IPR015812 Integrin_bsu;prints IPR015812 Integrin_bsu
28	Overlapping_Protein_Features	CHAIN:23:769:Integrin beta-2.;TOPO_DOM:23:700:Extracellular (Potential).;DOMAIN:124:363:VWFA.;DISULFID:33:447:By similarity.;VARIANT:128:128:D -> G (in LAD).

Sample annotated SNP cont.

Field number	Field name	Value
29	Amino_Acid_Reference	D
30	Amino_Acid_Reads	G
31	Amino_Acids_In_Orthologues	DDDDDDDDDDDDXDDDDDDDNDDDDDXDDDDDDDD DDDDDDDDDDDDDDDDDDDDDDDDDDDDX
32	Alignment_Score_Change	-0.438
33	C_blosum	0.9
34	Context_Conservation	84.6
35	Orthologue_Species	Ovis_aries;Tursiops_truncatus;Sus_scrofa;Myotis_lucifugus;Pteropus_vampyrus;Ailuropoda_melanoleuca;Mustela_putorius_furo;Equus_caballus;Canis_lupus_familiaris;Felis_catus;Callithrix_jacchus;Pan_troglodytes;Tarsius_syrichta;Otolemur_garnettii;Nomascus_leucogenys;Pongo_abelii;Ictidomys_tridecemlineatus;Microcebus_murinus;Mus_musculus;Dipodomys_ordii;Rattus_norvegicus;Mus_musculus;Homo_sapiens;Cavia_porcellus;Macaca_mulatta;Rattus_norvegicus;Tupaia_belangeri;Gorilla_gorilla_gorilla;Dasypus_novemcinctus;Loxodonta_africana;Echinops_telfairi;Procavia_capensis;Macropus_eugenii;Monodelphis_domestica;Pelodiscus_sinensis;Ficedula_albicollis;Gallus_gallus;Anas_platyrhynchos;Meleagris_gallopavo;Anolis_carolinensis;Taeniopygia_guttata;Petromyzon_marinus;Ciona_savignyi;Ciona_intestinalis;Drosophila_melanogaster;Caenorhabditis_elegans;Oryctolagus_cuniculus;Ochotona_princeps;Sorex_araneus;Xenopus_tropicalis;Ornithorhynchus_anatinus;Lepisosteus_oculatus;Latimeria_chalumnae;Tetraodon_nigroviridis;Gadus_morhua;Astyanax_mexicanus;Takifugu_rubripes;Oryzias_latipes;Gasterosteus_aculeatus;Xiphophorus_maculatus;Danio_rerio;Oreochromis_niloticus;Oreochromis_niloticus;Vicugna_pacos

Sample annotated SNP cont.

Field number	Field name	Value
36	Gene_Ontology	[GO:0002376]:immune system process;[GO:0003674]:molecular_function;[GO:0005575]:cellular_component;[GO:0005623]:cell;[GO:0005886]:plasma membrane;[GO:0007155]:cell adhesion;[GO:0007165]:signal transduction;[GO:0008150]:biological_process;[GO:0019899]:enzyme binding;[GO:0040011]:locomotion;[GO:0043167]:ion binding;[GO:0043234]:protein complex;[GO:0048870]:cell motility
37	Model_Annotations	Phenotypes_Position=Source: OMIM Description: LEUKOCYTE ADHESION DEFICIENCY Variation_names: rs137852615|Source: Uniprot Description: Leukocyte adhesion deficiency type 1 Variation_names: rs137852615|Source: Uniprot Phenotype_name: LAD1 Description: Leukocyte adhesion deficiency 1 Variation_names: rs137852615|Source: dbSNP_ClinVar Description: LEUKOCYTE_ADHESION_DEFICIENCY|Source: HGMD-PUBLIC Phenotype_name: HGMD_MUTATION Description: Annotated by HGMD but no phenotype description is publicly available;Phenotypes_Gene=Leukocyte adhesion deficiency type 1|GTR|MedGen|OMIM;Interactions_Count=11;Overlapping_Protein_Features=TOPO_DOM:23:700:Extracellular (Potential).|DOMAIN:124:363:VWFA.|VARIANT:128:128:D -> N (in LAD1 dbSNP:rs137852615).|VARIANT:128:128:D -> Y (in LAD1 dbSNP:rs137852615).;

Sample annotated SNP cont.

Field number	Field name	Value
38	Comments	Gene_Status=KNOWN;Transcript_Status=KNOWN;Gene_Biotype=protein_coding;Transcript_Biotype=protein_coding;SIFT_Prediction_Ensembl=deleterious(0);
39	Ref_SNPs	.
40	Is_Fully_Known	no

Summary of results for indels

Known vs. novel indels

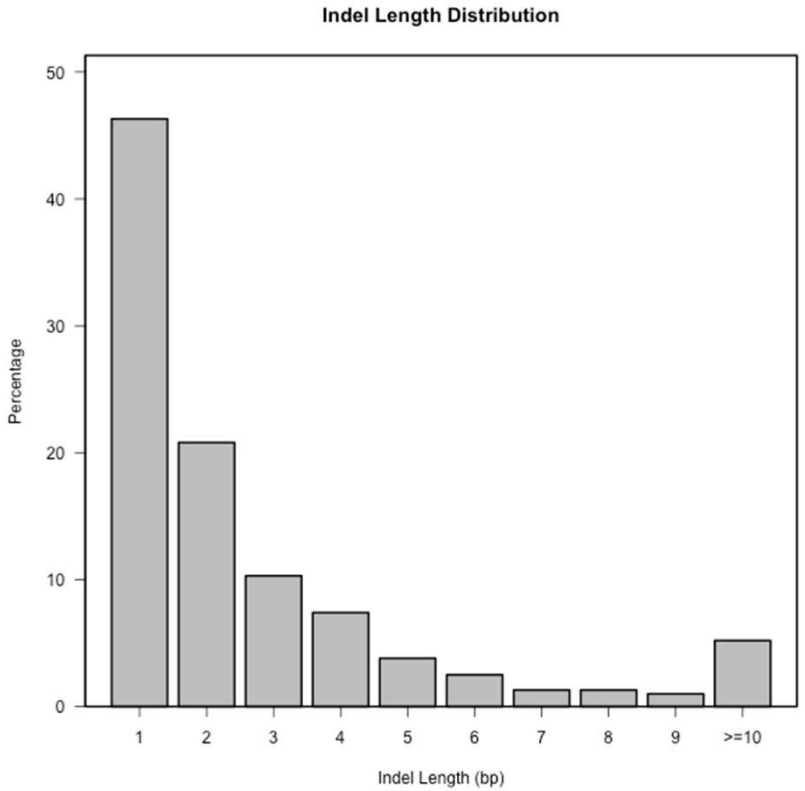
Indel type	Count
Known	15,222
Novel	1,539,884
All	1,555,106

“Known” is used here to describe input variants where the variant and all of its alleles exist already in the reference database.

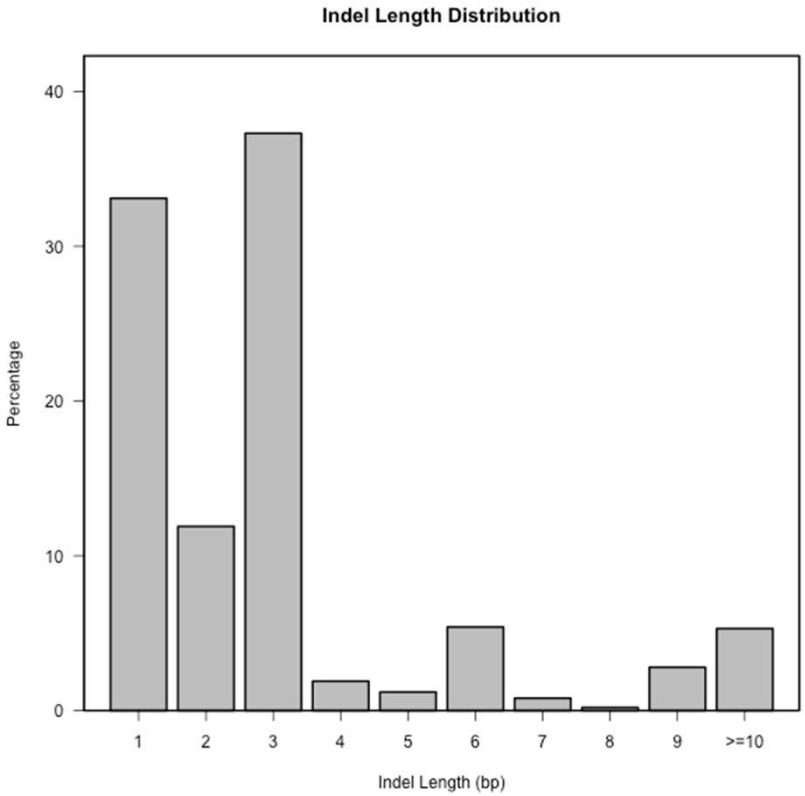
Numbers of indels in each function class

Indel consequence type	Count
INTERGENIC	1,074,284
intron_variant	391,227
upstream_gene_variant	42,693
downstream_gene_variant	38,781
3_prime_UTR_variant	4,147
frameshift_variant	1,028
splice_region_variant	974
inframe_deletion	717
5_prime_UTR_variant	675
inframe_insertion	254
splice_acceptor_variant	84
splice_donor_variant	81
missense_variant	55
non_coding_exon_variant	49
coding_sequence_variant	47
nc_transcript_variant	5
stop_gained	4
mature_miRNA_variant	1
Total	1,555,106

Length distribution of all indels and indels in coding regions



All indels



Coding-region indels

Sample annotated indel

Field number	Field name	Value
1	CHROM	4
2	POS	51126779
3	ID	.
4	REF	ACCC
5	ALT	ACC
6	QUAL	.
7	FILTER	.
8	INFO	.
9	Functional_Class	frameshift_variant
10	Chromosome_Reference	C

See http://www.ualberta.ca/~stothard/downloads/NGS-SNP/annotate_INDELs.html

Sample annotated indel cont.

Field number	Field name	Value
11	Chromosome_Reads	-
12	Gene_Description	cystic fibrosis transmembrane conductance regulator [Source:RefSeq peptide;Acc:NP_776443]
13	Ensembl_Gene_ID	ENSBTAG00000006589
14	Entrez_Gene_Name	CFTR
15	Entrez_Gene_ID	281067
16	Ensembl_Transcript_ID	ENSBTAT00000053450
17	Transcript_INDEL_Position	2540
18	Transcript_INDEL_Reference	G
19	Transcript_INDEL_Reads	-
20	Transcript_To_Chromosome_Strand	reverse

Sample annotated indel cont.

Field number	Field name	Value
21	Ensembl_Protein_ID	ENSBTAP00000049907
22	UniProt_ID	.
23	Amino_Acid_Position	830
24	Overlapping_Protein_Domains	superfamily;tigrfam IPR005291 cAMP_cl_channel
25	Overlapping_Protein_Features	.
26	Gene_Ontology	[GO:0005575]:cellular_component;[GO:0005622]:intracellular; [GO:0005623]:cell; [GO:0005737]:cytoplasm; [GO:0005768]:endosome; [GO:0005783]:endoplasmic reticulum; [GO:0005886]:plasma membrane;

Sample annotated indel cont.

Field number	Field name	Value
27	Model_Annotations	Phenotypes_Position=Source: HGMD-PUBLIC Phenotype_name: HGMD_MUTATION Description: Annotated by HGMD but no phenotype description is publicly available Variation_names: CM930120;Phenotypes_Gene=Bronchiectasis|Gene Reviews|GTR|OMIM|Congenital bilateral absence of the vas deferens|Cystic fibrosis
28	Comments	Gene_Status=KNOWN;Transcript_Status=KNOWN;Length_Downstream_Protein=651(43.96);
29	Ref_INDELs	.
30	Is_Fully_Known	no