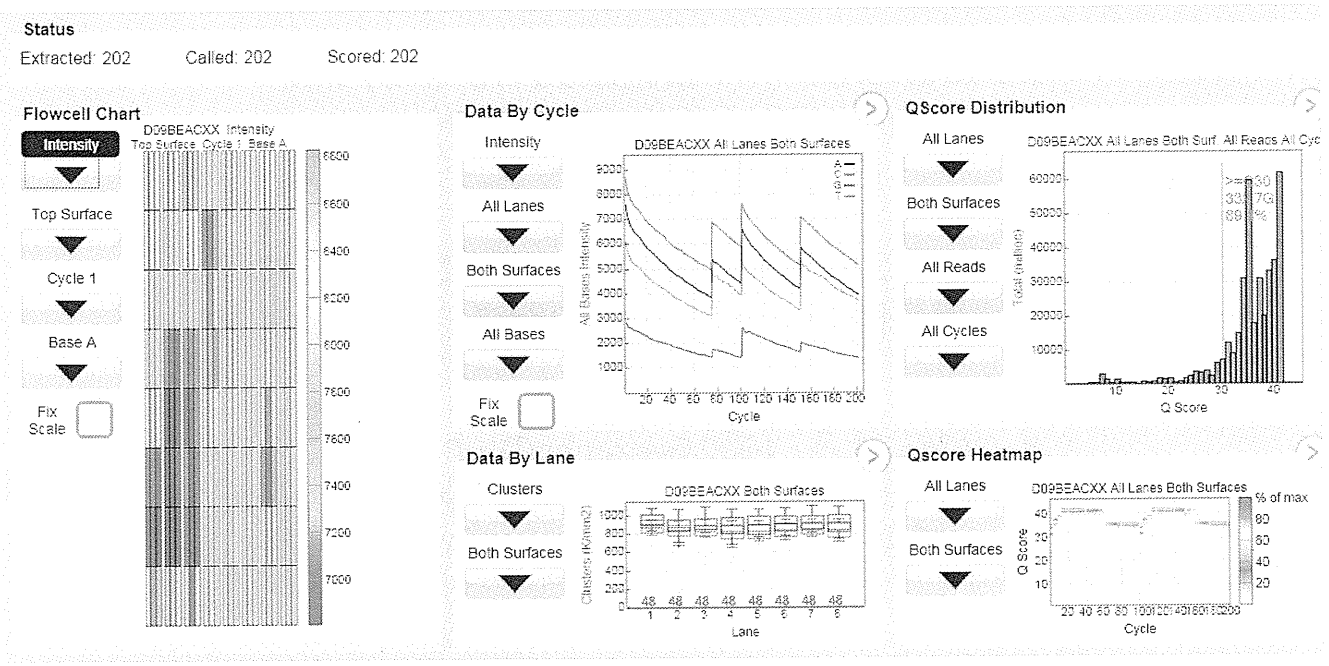


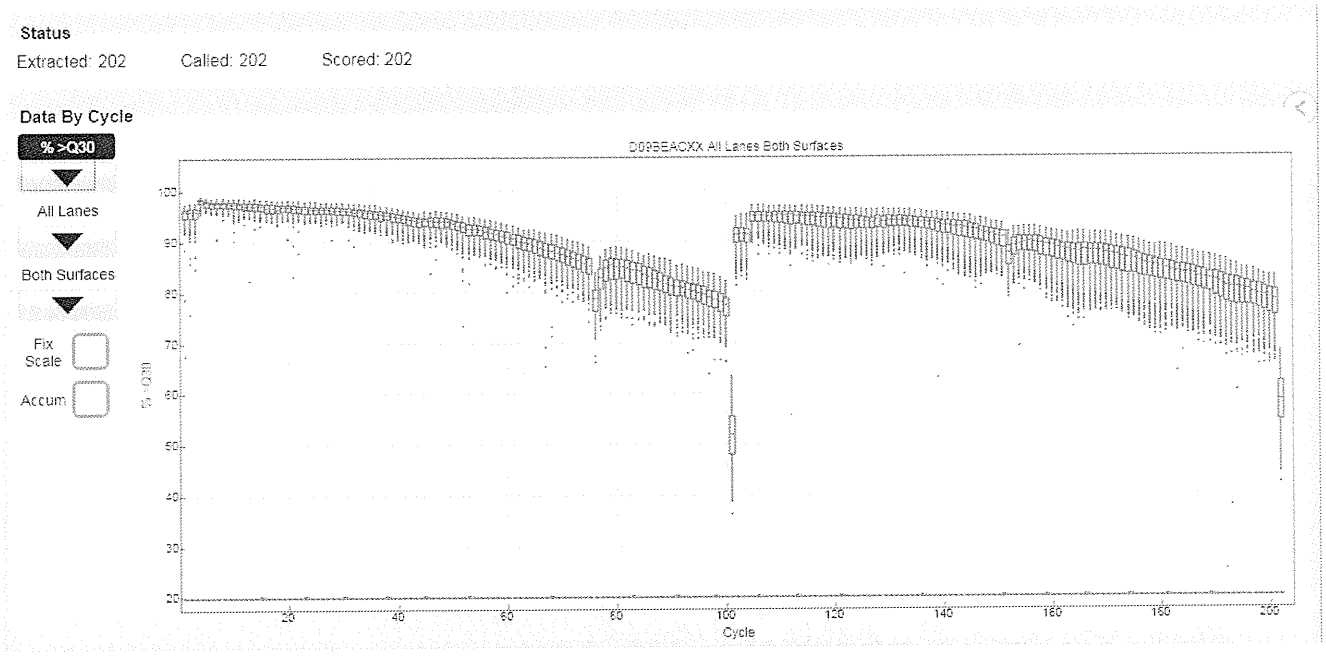


図10: HiSeq1000による大量並列シーケンシング

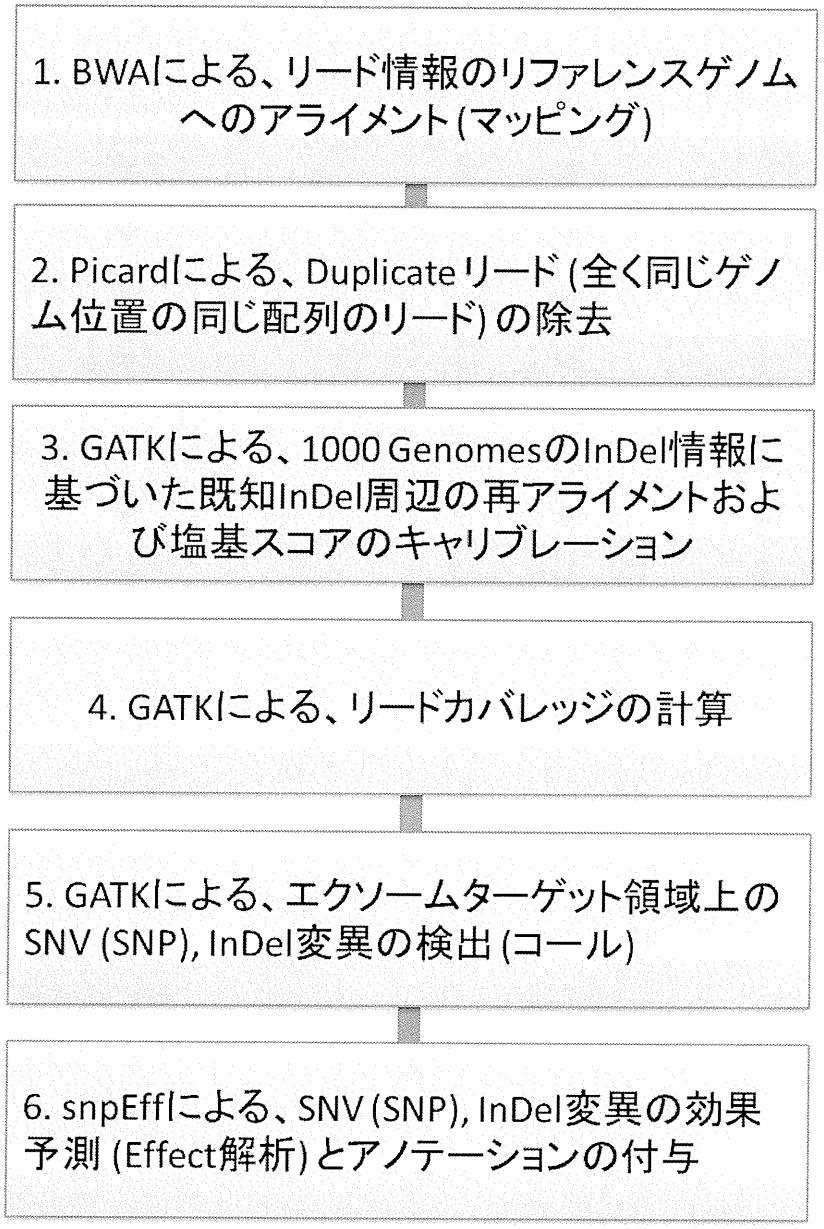


図11: エクソームデータ解析パイプライン
試験的に構築した幾つかの解析パイプラインのうち、上記に示すパイプラインを主に用いた。

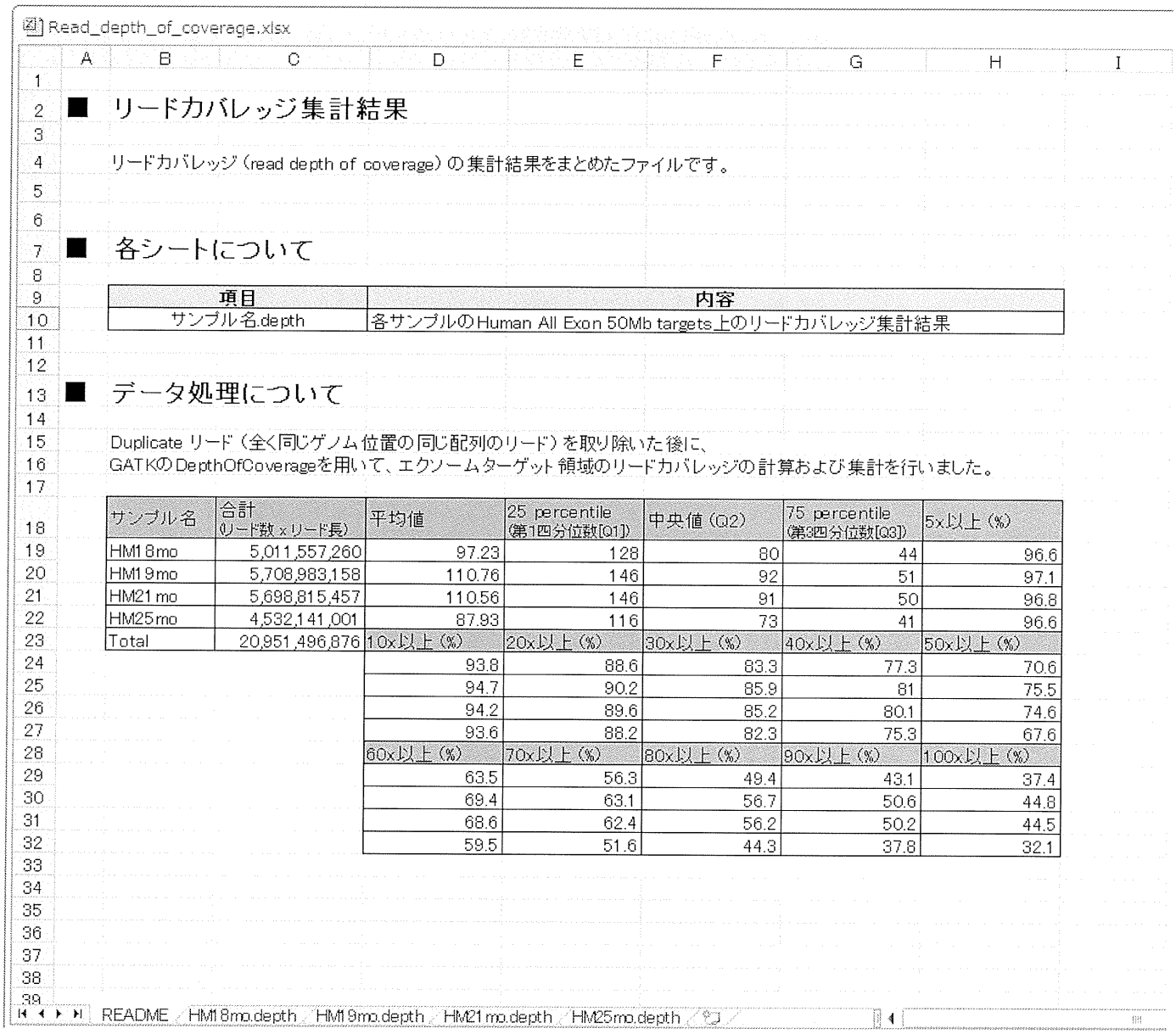
BWA によるアライメントの実行

BWA の実行コマンドは、ラッパーとして Perl スクリプト中に組込み実行した。BWA を実行するための Perl スクリプトを以下に示す。

```
  $command = "$bwa aln -t $threads $hg19 $sample¥_R1.fastq >
$sample¥_R1.sai";
  print STDERR "¥n$command¥n";
  `$command 2>> $log_dir/$sample.bwa.log`;

  $command = "$bwa aln -t $threads $hg19 $sample¥_R2.fastq >
$sample¥_R2.sai";
  print STDERR "¥n$command¥n";
  `$command 2>> $log_dir/$sample.bwa.log`;

  $id++;
  $command = "$bwa sampe -P $hg19 $sample¥_R1.sai $sample¥_R2.sai
$sample¥_R1.fastq $sample¥_R2.fastq ".
             "-r '¥@RG¥¥tID:$id¥¥tSM:$sample¥¥tPL:Illumina' >
$sample.sam";
  print STDERR "¥n$command¥n";
  `$command 2>> $log_dir/$sample.bwa.log`;

  $command = "$samtools view -bS $sample.sam > $sample.bam";
  print STDERR "¥n$command¥n";
  `$command 2>> $log_dir/$sample.bwa.log`;

  $command = "$samtools sort $sample.bam $sample.sort";
  print STDERR "¥n$command¥n";
  `$command 2>> $log_dir/$sample.bwa.log`;

  $command = "samtools index $sample.sort.bam";
  print STDERR "¥n$command¥n";
  `$command`;

  $command = "rm -f $sample¥_R1.sai $sample¥_R2.sai $sample.sam
$sample.bam";
  print STDERR "¥n$command¥n";
  `$command`;
```

この操作により、リード情報をリファレンスゲノムへアライメントし、圧縮、ソートしたアライメントファイル（サンプル名.sort.bam）が作成される。

図12：エクソームデータ解析パイプラインで用いたPerlスクリプト例



Read_depth_of_coverage.xlsx

■ リードカバレッジ集計結果

リードカバレッジ（read depth of coverage）の集計結果をまとめたファイルです。

■ 各シートについて

項目	内容
サンプル名.depth	各サンプルのHuman All Exon 50Mb targets上のリードカバレッジ集計結果

■ データ処理について

Duplicate リード（全く同じゲノム位置の同じ配列のリード）を取り除いた後に、
GATKのDepthOfCoverageを用いて、エクソームターゲット領域のリードカバレッジの計算および集計を行いました。

サンプル名	合計 (リード数 x リード長)	平均値	25 percentile (第1四分位数[Q1])	中央値 (Q2)	75 percentile (第3四分位数[Q3])	5x以上 (%)
HM18mo	5,011,557,260	97.23	128	80	44	96.6
HM19mo	5,708,983,158	110.76	146	92	51	97.1
HM21mo	5,698,815,457	110.56	146	91	50	96.8
HM25mo	4,532,141,001	87.93	116	73	41	96.6
Total	20,951,496,876	10x以上 (%)	20x以上 (%)	30x以上 (%)	40x以上 (%)	50x以上 (%)
		93.8	88.6	83.3	77.3	70.6
		94.7	90.2	85.9	81	75.5
		94.2	89.6	85.2	80.1	74.6
		93.6	88.2	82.3	75.3	67.6
		60x以上 (%)	70x以上 (%)	80x以上 (%)	90x以上 (%)	100x以上 (%)
		63.5	56.3	49.4	43.1	37.4
		69.4	63.1	56.7	50.6	44.8
		68.6	62.4	56.2	50.2	44.5
		59.5	51.6	44.3	37.8	32.1

README | HM18mo.depth | HM19mo.depth | HM21mo.depth | HM25mo.depth

図13:
GATKのDepthOfCoverage機能を用いたエクソームターゲット領域のリードカバレッジ (read depth of coverage) 集計

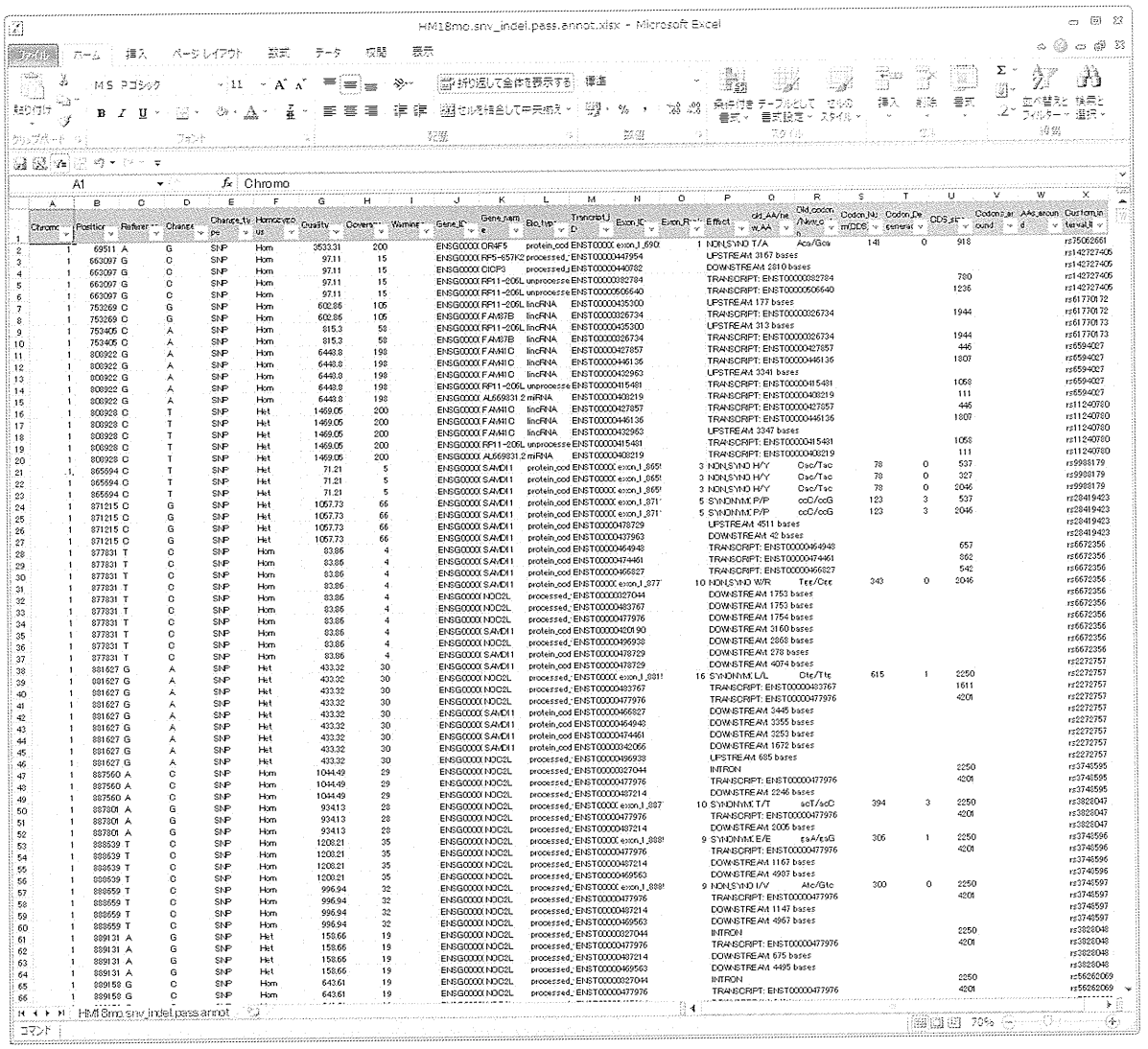


図 14： (sample name).snv_indel.pass.annot.xlsx ファイルの例

フィルタをパスした SNV, InDel の効果予測とアノテーションを transcript (転写産物) 単位でまとめたファイルである。transcript 単位のため、同じ SNV, InDel が異なる transcript 毎に出力される。

Column	Notes
Chromosome	Chromosome name (usually without any leading 'chr' string)
Position	One based position
Reference	Reference
Change	Sequence change
Change type	Type of change {SNP, MNP, INS, DEL}
Homozygous	Is this homozygous or heterozygous {Hom, Het}
Quality	Quality score (from input file)
Coverage	Coverage (from input file)
Warnings	Any warnings or errors.
Gene_ID	Gene ID (usually ENSEMBL)
Gene_name	Gene name
Bio_type	BioType, as reported by ENSEMBL.
Trancript_ID	Transcript ID (usually ENSEMBL)
Exon_ID	Exon ID (usually ENSEMBL)
Exon_Rank	Exon number on a transcript
Effect	Effect of this variant. See details below.
old_AA/new_AA	Amino acid change
old_codon/new_codon	Codon change
Codon_Num(CDS)	Codon number in CDS
Codon_degenaracy	Codon degenaracy (see below).
CDS_size	CDS size in bases
Custom_interval_ID	If any custom interval was used, add the IDs here (may be more than one).

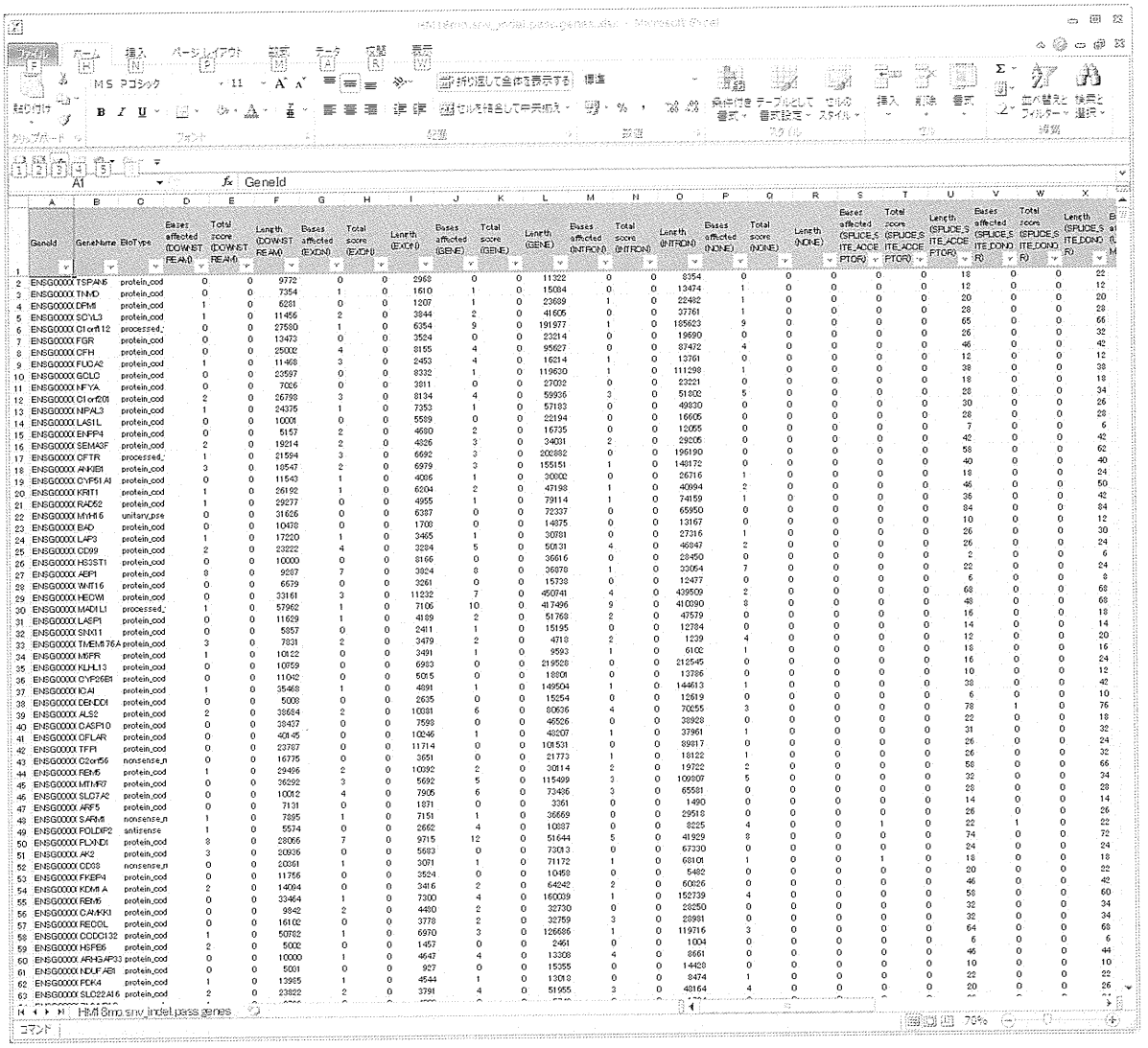
図 15： **(sample name).snv_indel.pass.annot.xlsx** に含まれるアノテーション情報

Effect	Note	Example
INTERGENIC	The variant is in an intergenic region	
UPSTREAM	Upstream of a gene (default length: 5K bases)	
UTR_5_PRIME	Variant hits 5'UTR region	
UTR_5_DELETED	The variant deletes and exon which is in the 5'UTR of the transcript	
START_GAINED	A variant in 5'UTR region produces a three base sequence that can be a START codon.	
SPLICE_SITE_ACCEPTOR	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).	
SPLICE_SITE_DONOR	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).	
START_LOST	Variant causes start codon to be mutated into a non-start codon.	aTg/aGg, M/R
SYNONYMOUS_START	Variant causes start codon to be mutated into another start codon.	Ttg/Ctg, L/L (TTG and CTG can be START codons)
CDS	The variant hits a CDS.	
GENE	The variant hits a gene.	
TRANSCRIPT	The variant hits a transcript.	
EXON	The vairant hist an exon.	
EXON_DELETED	A deletion removes the whole exon.	
NON_SYNONYMOUS_CODING	Variant causes a codon that produces a different amino acid	Tgg/Cgg, W/R
SYNONYMOUS_CODING	Variant causes a codon that produces the same amino acid	Ttg/Ctg, L/L
FRAME_SHIFT	Insertion or deletion causes a frame shift	An indel size is not multple of 3
CODON_CHANGE	One or many codons are changed	An MNP of size multiple of 3

図 16-1： (sample name).snv_indel.pass.annot.xlsx に含まれる変異効果予測アノテーション全 32 種類の定義

Effect	Note	Example
CODON_INSERTION	One or many codons are inserted	An insert multiple of three in a codon boundary
CODON_CHANGE_ PLUS_CODON_INSERTION	One codon is changed and one or many codons are inserted	An insert of size multiple of three, not at codon boundary
CODON_DELETION	One or many codons are deleted	A deletion multiple of three at codon boundary
CODON_CHANGE_ PLUS_CODON_DELETION	One codon is changed and one or more codons are deleted	A deletion of size multiple of three, not at codon boundary
STOP_GAINED	Variant causes a STOP codon	Cag/Tag, Q/*
SYNONYMOUS_STOP	Variant causes stop codon to be mutated into another stop codon.	taA/taG, */*
STOP_LOST	Variant causes stop codon to be mutated into a non-stop codon	Tga/Cga, */R
INTRON	Variant hist and intron. Technically, hits no exon in the transcript.	
UTR_3_PRIME	Variant hits 3'UTR region	
UTR_3_DELETED	The variant deletes and exon which is in the 3'UTR of the transcript	
DOWNSTREAM	Downstream of a gene (default length: 5K bases)	
INTRON_CONSERVED	The variant is in a highly conserved intronic region	
INTERGENIC_CONSERVED	The variant is in a highly conserved intergenic region	
INTRAGENIC	The variant hits a gene, but no transcripts within the gene	

図 16-2： (sample name).snv_indel.pass.annot.xlsx に含まれる変異効果予測アノテーション全 32 種類の定義



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Geneid	GeneName	EtoType	Bases affected (DOWNSTREAM)	Total score (DOWNSTREAM)	Length (DOWNSTREAM)	Bases affected (EXON)	Total score (EXON)	Length (EXON)	Bases affected (GENE)	Total score (GENE)	Length (GENE)	Bases affected (INTRON)	Total score (INTRON)	Length (INTRON)	Bases affected (NONE)	Total score (NONE)	Length (NONE)	Bases affected (SPLICE_SITE_ACCEPTOR)	Total zscore (SPLICE_SITE_ACCEPTOR)	Length (SPLICE_SITE_ACCEPTOR)	Bases affected (SPLICE_SITE_DONOR)	Total zscore (SPLICE_SITE_DONOR)	Length (SPLICE_SITE_DONOR)	B al (L M
2	ENSG00000	TSPAN6	protein_cod	0	0	9772	0	0	2968	0	0	11222	0	0	8354	0	0	0	0	0	0	18	0	0	22
3	ENSG00000	TNMD	protein_cod	0	0	7354	1	0	1610	1	0	15084	0	0	13474	1	0	0	0	0	0	12	0	0	12
4	ENSG00000	DPM1	protein_cod	1	0	6281	0	0	1207	1	0	23689	1	0	22482	1	0	0	0	0	0	20	0	0	20
5	ENSG00000	SCYL3	protein_cod	1	0	11456	2	0	3844	2	0	41605	0	0	37761	1	0	0	0	0	0	28	0	0	28
6	ENSG00000	C1orf112	processed_	0	0	27580	1	0	6354	9	0	191977	1	0	185623	9	0	0	0	0	0	65	0	0	66
7	ENSG00000	FGR	protein_cod	0	0	13473	0	0	3524	0	0	23214	0	0	19690	0	0	0	0	0	0	26	0	0	32
8	ENSG00000	CFH	protein_cod	0	0	25002	4	0	8155	4	0	95627	0	0	87472	4	0	0	0	0	0	46	0	0	42
9	ENSG00000	FUCA2	protein_cod	1	0	11468	3	0	2453	4	0	16214	1	0	13761	0	0	0	0	0	0	12	0	0	12
10	ENSG00000	GCLC	protein_cod	0	0	33597	0	0	8332	1	0	119620	1	0	111298	1	0	0	0	0	0	38	0	0	38
11	ENSG00000	NFYA	protein_cod	0	0	7026	0	0	3911	0	0	27002	0	0| 23221	0	0	0	0	0	0	18	0	0	18
12	ENSG00000	C1orf201	protein_cod	2	0	26798	3	0	8134	4	0	59935	3	0	51809	5	0	0	0	0	0	28	0	0	34
13	ENSG00000	NPAL3	protein_cod	1	0	24375	1	0	7353	1	0	57183	0	0	49830	0	0	0	0	0	0	30	0	0	26
14	ENSG00000	LAS1L	protein_cod	0	0	10001	0	0	5589	0	0	22194	0	0	16605	0	0	0	0	0	0	28	0	0	28
15	ENSG00000	ENPP4	protein_cod	0	0	5157	2	0	4680	2	0	16735	0	0	12055	0	0	0	0	0	0	7	0	0	6
16	ENSG00000	SEMA3F	protein_cod	2	0	19214	2	0	4826	3	0	34031	2	0	29205	0	0	0	0	0	0	42	0	0	42
17	ENSG00000	CFTR	processed_	1	0	31594	3	0	6692	3	0	202882	0	0	196190	0	0	0	0	0	0	58	0	0	62
18	ENSG00000	ANKIB1	protein_cod	3	0	10547	2	0	6979	3	0	155151	1	0	149172	0	0	0	0	0	0	40	0	0	40
19	ENSG00000	CYP51A1	protein_cod	0	0	11543	1	0	4006	1	0	30002	0	0	26716	1	0	0	0	0	0	18	0	0	24
20	ENSG00000	KRIT1	protein_cod	1	0	56192	1	0	6004	2	0	47198	1	0	40994	2	0	0	0	0	0	46	0	0	50
21	ENSG00000	RAD52	protein_cod	1	0	29277	0	0	4955	1	0	79114	1	0	74159	1	0	0	0	0	0	36	0	0	42
22	ENSG00000	MYH6	unitary,pre	0	0	31626	0	0	6387	0	0	72337	0	0	65950	0	0	0	0	0	0	84	0	0	84
23	ENSG00000	BAD	protein_cod	0	0	10478	0	0	1708	0	0	14875	0	0	13167	0	0	0	0	0	0	10	0	0	12
24	ENSG00000	LAP3	protein_cod	1	0	17220	1	0	3465	1	0	30781	0	0	27316	1	0	0	0	0	0	26	0	0	30
25	ENSG00000	CD99	protein_cod	2	0	23222	4	0	3284	5	0	50131	4	0	46947	2	0	0	0	0	0	26	0	0	24
26	ENSG00000	HS3ST1	protein_cod	0	0	10000	0	0	3166	0	0	36616	0	0	33450	0	0	0	0	0	0	2	0	0	6
27	ENSG00000	ABP1	protein_cod	3	0	9207	7	0	3924	8	0	36978	1	0	33054	7	0	0	0	0	0	22	0	0	24
28	ENSG00000	WIF1	protein_cod	0	0	6539	0	0	3261	0	0	15738	0	0	12477	0	0	0	0	0	0	6	0	0	8
29	ENSG00000	HEXIM	protein_cod	0	0	33161	3	0	11232	7	0	450741	4	0	439509	2	0	0	0	0	0	68	0	0	68
30	ENSG00000	MADILI	processed_	1	0	57962	1	0	7106	10	0	417496	9	0	410090	8	0	0	0	0	0	48	0	0	68
31	ENSG00000	LASP1	protein_cod	0	0	11629	1	0	4189	2	0	51768	2	0	47579	0	0	0	0	0	0	16	0	0	18
32	ENSG00000	SNX11	protein_cod	0	0	5857	0	0	2411	1	0	15195	0	0	12784	0	0	0	0	0	0	14	0	0	14
33	ENSG00000	TMEM176A	protein_cod	3	0	7831	2	0	3479	2	0	4718	2	0	1239	4	0	0	0	0	0	18	0	0	16
34	ENSG00000	M6PR	protein_cod	1	0	10122	0	0	3491	1	0	9593	1	0	6102	1	0	0	0	0	0	16	0	0	24
35	ENSG00000	KLHL13	protein_cod	0	0	10759	0	0	6983	0	0	219528	0	0	212545	0	0	0	0	0	0	10	0	0	24
36	ENSG00000	CYP26B1	protein_cod	0	0	11042	0	0	5015	0	0	18801	0	0	13786	0	0	0	0	0	0	10	0	0	12
37	ENSG00000	ICAI	protein_cod	1	0	35468	1	0	4891	1	0	149504	1	0	144613	1	0	0	0	0	0	38	0	0	42
38	ENSG00000	DENND1	protein_cod	0	0	5008	0	0	2635	0	0	15254	0	0	12619	0	0	0	0	0	0	6	0	0	10
39	ENSG00000	ALS2	protein_cod	2	0	38684	2	0	10091	6	0	80636	4	0	70255	3	0	0	0	0	0	78	1	0	76
40	ENSG00000	CASP10	protein_cod	0	0	38437	0	0	7599	0	0	46526	0	0	38928	0	0	0	0	0	0	22	0	0	18
41	ENSG00000	CFLAR	protein_cod	0	0	40145	0	0	10246	1	0	48207	1	0	37961	1	0	0	0	0	0	31	0	0	32
42	ENSG00000	TFP1	protein_cod	0	0	23767	0	0	11714	0	0	101501	0	0	89917	0	0	0	0	0	0	26	0	0	24
43	ENSG00000	C2orf56	nonsense_n	0	0	16775	0	0	3651	0	0	21773	1	0	18122	1	0	0	0	0	0	26	0	0	32
44	ENSG00000	REM5	protein_cod	1	0	29466	2	0	10092	2	0	30114	2	0	19722	2	0	0	0	0	0	58	0	0	56
45	ENSG00000	MTMR7	protein_cod	0	0	36292	3	0	5692	5	0	115499	3	0	109907	5	0	0	0	0	0	32	0	0	34
46	ENSG00000	SLC7A2	protein_cod	0	0	10012	4	0	7905	6	0	73486	3	0	65581	0	0	0	0	0	0	28	0	0	28
47	ENSG00000	ARF5	protein_cod	0	0	7131	0	0	1871	0	0	3361	0	0	1490	0	0	0	0	0	0	14	0	0	14
48	ENSG00000	SARM1	nonsense_n	1	0	7895	1	0	7151	1	0	36669	0	0	29518	0	0	0	0	0	0	26	0	0	26
49	ENSG00000	POLDIP2	antisense	1	0	5574	0	0	2662	4	0	10887	0	0	8225	4	0	0	0	1	0	22	1	0	22
50	ENSG00000	PLXND1	protein_cod	8	0	28066	7	0	9715	12	0	51644	5	0	41929	8	0	0	0	0	0	74	0	0	72
51	ENSG00000	AK2	protein_cod	3	0	20926	0	0	5693	0	0	73013	0	0	67330	0	0	0	0	0	0	24	0	0	24
52	ENSG00000	CD93	nonsense_n	0	0	20391	1	0	3071	1	0	71172	1	0	68101	1	0	0	0	1	0	18	0	0	18
53	ENSG00000	FKBP4	protein_cod	0	0	11766	0	0	3524	0	0	10458	0	0	5482	0	0	0	0	0	0	20	0	0	22
54	ENSG00000	KDM1A	protein_cod	2	0	14094	0	0	3416	2	0	64242	2	0	60026	0	0	0	0	0	0	46	0	0	42
55	ENSG00000	REM6	protein_cod	0	0	33464	1	0	7300	4	0	160039	1	0	152739	4	0	0	0	0	0	58	0	0	60
56	ENSG00000	CAMKK1	protein_cod	0	0	9842	2	0	4480	2	0	32730	0	0	28250	0	0	0	0	0	0	32	0	0	34
57	ENSG00000	RECQL	protein_cod	0	0	16102	0	0	3778	2	0	32759	3	0	28981	0	0	0	0	0	0	32	0	0	34
58	ENSG00000	CCDC132	protein_cod	1	0	50762	1	0	6970	3	0	126686	1	0	119716	3	0	0	0	0	0	64	0	0	68
59	ENSG00000	HSPB6	protein_cod	2	0	5002	0	0	1457	0	0	2461	0	0	1004	0	0	0	0	0	0	6	0	0	6
60	ENSG00000	ARHGAP33	protein_cod	0	0	10000	1	0	4647	4	0	13308	4	0	8661	0	0	0	0	0	0	46	0	0	44
61	ENSG00000	NDUFAB1	protein_cod	0	0	5091	0	0	927	0	0	15355	0	0	14428	0	0	0	0	0	0	10	0	0	10
62	ENSG00000	PDK4	protein_cod	1	0	13965	1	0	4544	1	0	13018	0	0	8474	1	0	0	0	0	0	22	0	0	22
63	ENSG00000	SLC22A16	protein_cod	2	0	23822	2	0	3791	4	0	51955	3	0	48164	4	0	0	0	0	0	20	0	0	26

図 17： (sample name).snv_indel.genes.xlsx ファイルの例

SNV, InDel の効果予測とアノテーションを遺伝子単位 (またはゲノム領域) でまとめた。(sample name).snv_indel.pass.genes.xlsx はフィルタをパスした SNV, InDel をまとめたファイル、(sample name).サンプル名.snv_indel.genes.xlsx は全ての SNV, InDel をまとめたファイルとした。変異の効果予測については図１５を参照。

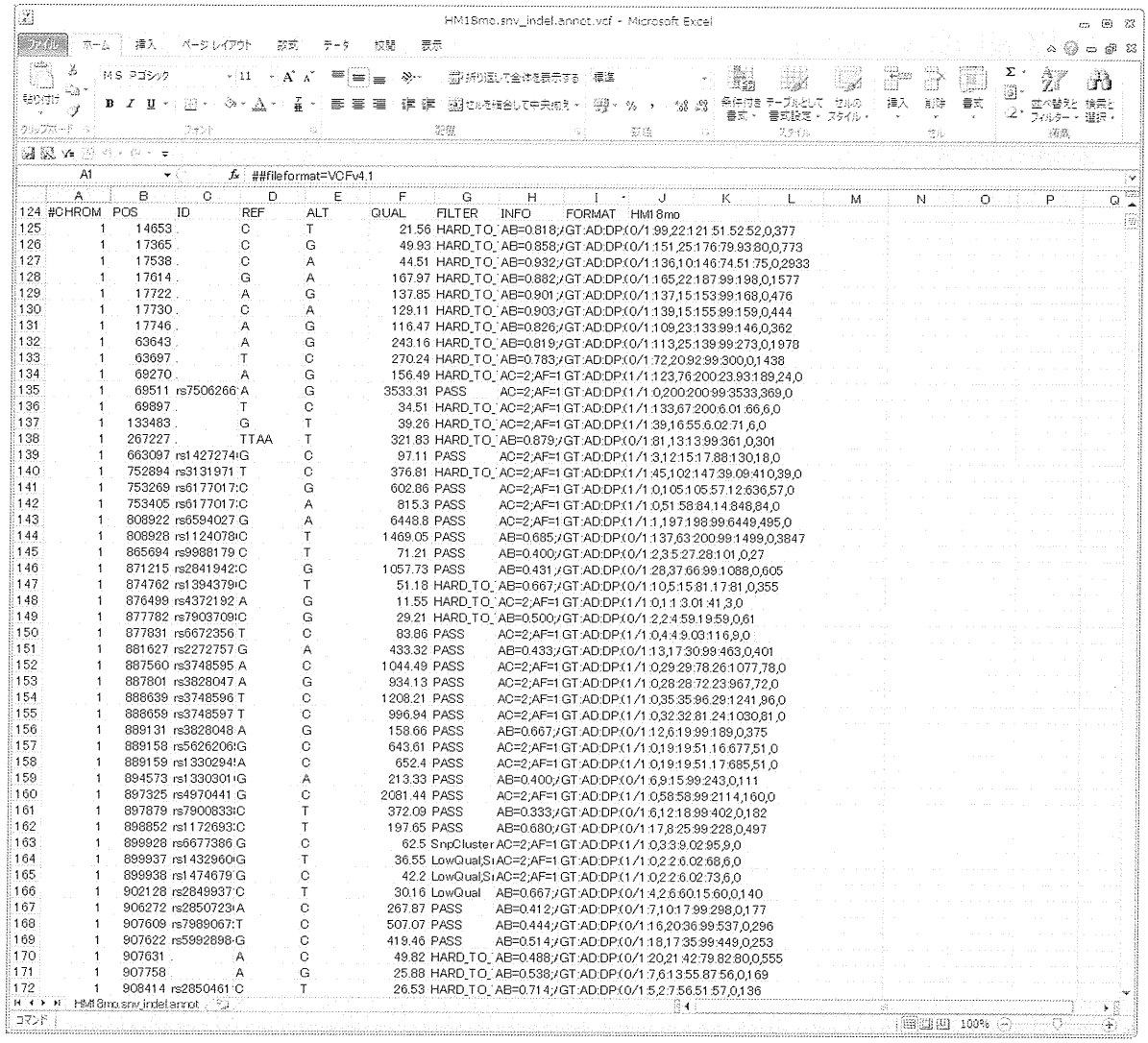


図 18： (sample name).snv_indel.annot.vcf ファイルの例 (エクセル使用)

VCF は、変異を記述するフォーマットとしてスタンダードなものとなっており、変異のフィルタリングや他ツールへの取り込み等に用いることができる。VCF の詳細については、1000 Genomes ウェブサイト http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41 を参照されたい。

Sub-field	Notes
Effect	Effect of this variant. See details below.
Effect impact	Effect impact {High, Moderate, Low, Modifier}. See details below.
Codon_Change	Codon change: old_codon/new_codon
Amino_Acid_change	Amino acid change: old_AA/new_AA
Warnings	Any warnings or errors.
Gene_name	Gene name
Gene_BioType	BioType, as reported by ENSEMBL.
Coding	[CODING | NON_CODING]. If information reported by ENSEMBL (e.g. has 'protein_id' information in GTF file).
Trancript	Transcript ID (usually ENSEMBL)
Exon	Exon ID (usually ENSEMBL)
Warnings	Any warnings or errors (not shown if empty).

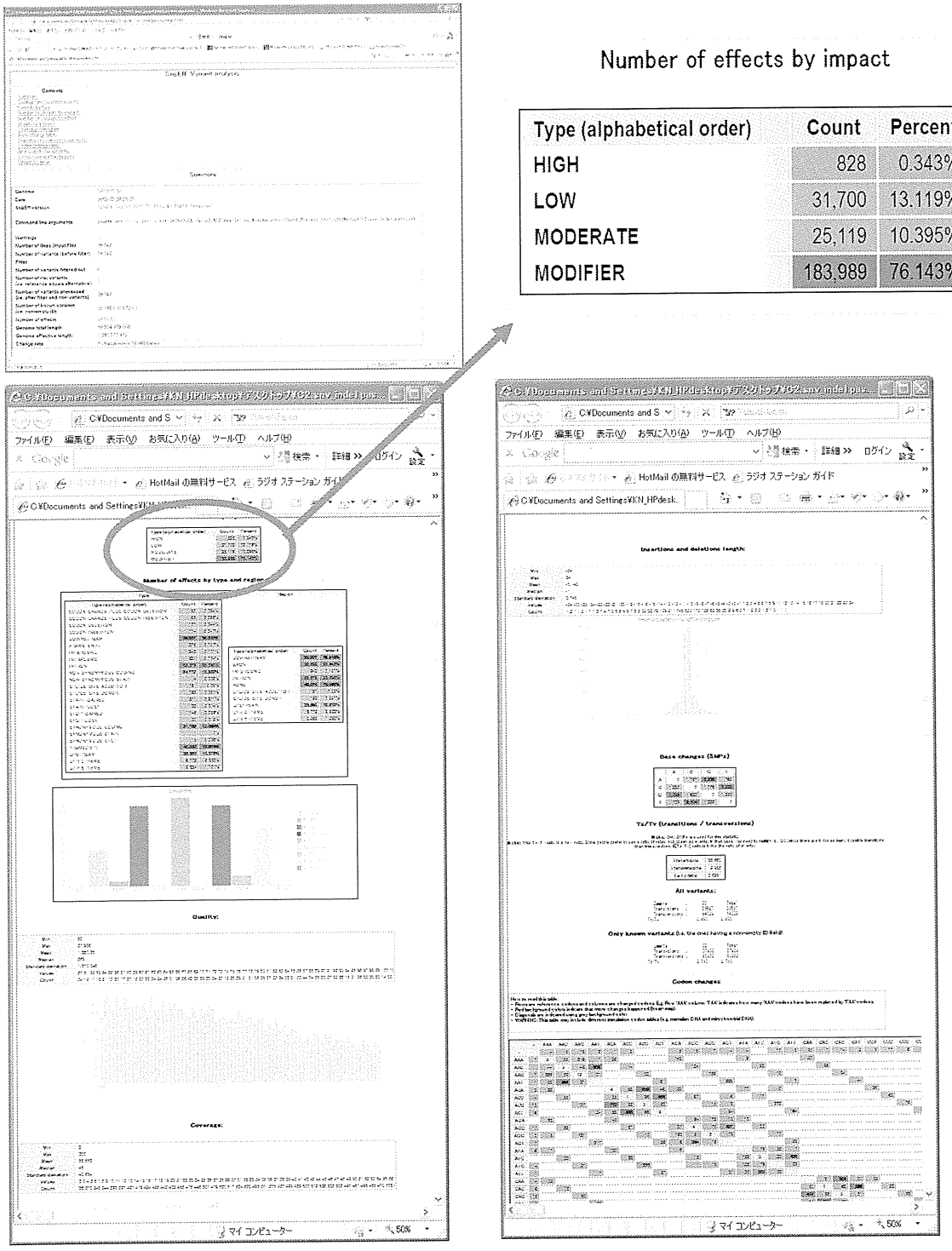
図 19：snpEff による変異の効果予測のサブフィールド定義

snpEff による変異の効果予測は、VCF の INFO フィールドに記述され、以下のようなフォーマットになる。複数の効果予測がある場合は、コンマ (,) で区切られる。

Effect (Effect_Impact | Codon_Change | Amino_Acid_change | Gene_Name | Gene_BioType | Coding | Transcript | Exon [| ERRORS | WARNINGS])

Impact	Effects
High	SPLICE_SITE_ACCEPTOR SPLICE_SITE_DONOR START_LOST EXON_DELETED FRAME_SHIFT STOP_GAINED STOP_LOST
Moderate	NON_SYNONYMOUS_CODING CODON_CHANGE CODON_INSERTION CODON_CHANGE_PLUS_CODON_INSERTION CODON_DELETION CODON_CHANGE_PLUS_CODON_DELETION UTR_5_DELETED UTR_3_DELETED
Low	SYNONYMOUS_START NON_SYNONYMOUS_START START_GAINED SYNONYMOUS_CODING SYNONYMOUS_STOP NON_SYNONYMOUS_STOP
Modifier	UTR_5_PRIME UTR_3_PRIME REGULATION UPSTREAM DOWNSTREAM GENE TRANSCRIPT EXON INTRON_CONSERVED INTRON INTRAGENIC INTERGENIC INTERGENIC_CONSERVED NONE

図 20：snpEff による変異の効果予測のサブフィールド定義：
“Effect_Impact”の分類



Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	828	0.343%
LOW	31,700	13.119%
MODERATE	25,119	10.395%
MODIFIER	183,989	76.143%

図 21：(sample name).snv_indel.summary.html の例

検出された SNV, InDel の様々な統計情報 (効果予測、カバレッジ、アミノ酸置換等) を記述した HTML ファイル（ウェブブラウザーで閲覧可能）。

III. 研究成果の刊行に関する一覧表

研究成果の刊行に関する一覧表

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年	ページ

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Fukami M Tsuchiya T, Takada S, Kanbara A, Asahara A, Igarashi A, Kamiyama Y, Nishimura G, Ogata T	Complex Genomic Rearrangement in the *SOX9* 5′ Region in a Patient with Pierre Robin Sequence and Hypoplastic Left Scapula.	Am J Med Genet A	In press		2012

IV. 研究成果の刊行物・別刷

Complex Genomic Rearrangement in the *SOX9* 5′ Region in a Patient with Pierre Robin Sequence and Hypoplastic Left Scapula

Maki Fukami[1,*,#], Takayosi Tsuchiya[1,3,#], Shuji Takada[2,#], Akiko Kanbara[4], Hiroshi Asahara[2], Arisa Igarashi[2], Yasunori Kamiyama[4], Gen Nishimura[5] and Tsutomu Ogata[1,6]

Departments of [1]Molecular Endocrinology and [2]Systems Biomedicine, National Research Institute for Child Health and Development, Tokyo, Japan

[3]Department of Pediatrics, Dokkyo University Koshigaya Hospital, Koshigaya, Japan

[4]Department of Pediatrics, Saiseikai Utsunomiya Hospital, Utsunomiya, Japan

[5]Department of Pediatric Imaging, Tokyo Metropolitan Children's Medical Center, Tokyo, Japan

[6] Department of Pediatrics, Hamamatsu University School of Medicine, Hamamatsu, Japan

Running title: Fukami et al., Genomic abnormality in Pierre Robin sequence

*Correspondence to:

Maki Fukami, M.D., Department of Molecular Endocrinology, National Research Institute for Child Health and Development, 2-10-1 Ohkura, Setagaya, Tokyo, 157-8535, Japan, E-mail: mfukami@nch.go.jp, Telephone number: +81-3-5494-7025, Fax number: +81-3-5494-7026

[#]These authors are equally contributed to this manuscript.

Pierre Robin sequence (PRS) can occur as a component of campomelic dysplasia (CD) and acampomelic campomelic dysplasia (ACD) caused by dysfunction or dysregulation of SOX9, although it can also take place as an isolated form. Recently, genomic alterations in the far upstream and the far downstream region of *SOX9* have been identified in patients with isolated PRS. Here, we report a male patient with PRS and a heterozygous genomic rearrangement in the 5′ region of *SOX9*. Clinical analysis revealed PRS-compatible craniofacial anomalies, mild hypoplasia of the left scapula and normal male external genitalia. Molecular analysis identified a paracentric inversion on the long arm of chromosome 17 with breakpoints at 17q21.31 and 17q24.3, and a microdeletion spanning from -4.15 to -1.16 Mb relative to *SOX9*. These findings indicate that the chromosomal region more than 1.16 Mb apart from *SOX9* contains at least one developmental enhancer(s) for *SOX9* that plays a critical role in the development of the mandible and a relatively small role in the development of the scapula. Moreover, the concept of exclusion mapping argues that putative CD/ACD loci are located within the 1.16 Mb region closest to *SOX9* coding exons, which remain intact in this non-CD/ACD patient. This study provides a novel example for long-range cis-regulatory mutations of *SOX9*.

Key words:
campomelic dysplasia, deletion, inversion, enhancer, non-coding element

INTRODUCTION

Pierre Robin sequence (PRS) (OMIM 261800) is a congenital malformation sequence characterized by micrognathia, glossoptosis and posterior U-shaped cleft palate [Robin et al., 1923]. The primary defect of PRS is assumed to be mandibular hypoplasia caused by impaired chondrogenesis or aberrant proliferation of neural crest cells [Gordon et al., 2009]. PRS frequently occurs as a component of known syndromes such as campomelic dysplasia (CD) (OMIM 114290), acampomelic campomelic dysplasia (ACD) and Stickler syndrome (OMIM 108300), although PRS can also take place as an isolated (non-syndromic) form [Holder-Espinasse et al., 2001].

CD and ACD are caused by dysfunction or dysregulation of SOX9; multiple intragenic mutations of *SOX9* as well as various types of chromosomal rearrangements around the coding exons have been identified in patients with CD and ACD [Meyer J et al., 1997; Gordon et al., 2009]. In addition to PRS, patients with CD manifest bowing of the long bones (campomelia), hypoplastic scapulae, pelvic malformations, a missing pair of ribs, clubfeet and 46, XY gonadal dysgenesis. ACD represents a mild variant of CD lacking campomelia. Since PRS is present in most patients with CD and ACD [Gordon et al., 2009], SOX9 likely plays a particularly important role in the development of the mandible.

Recently, molecular defects in the far upstream and the far downstream region of *SOX9* have been identified in patients with isolated PRS. Jamshidi et al. [2004] and Jakobsen et al. [2007] identified balanced translocations of t(2;17) in familial and sporadic PRS cases respectively, and found that the 17q breakpoints are located more than 1.0 Mb upstream of *SOX9*. Subsequently, Benko et al. identified variable genomic abnormalities (translocations, deletions and a nucleotide substitution) at a position more than 1.0 Mb apart from *SOX9* in two sporadic and five familial cases with PRS [2009]. Furthermore, Benko et al. showed that the deletions and translocations included several highly conserved non-coding elements (HCNE) and the nucleotide substitution abolished the tissue-specific enhancer activity of one of these HCNEs (HCNE-F2) [2009]. These data provide the first evidence that dysfunction of the