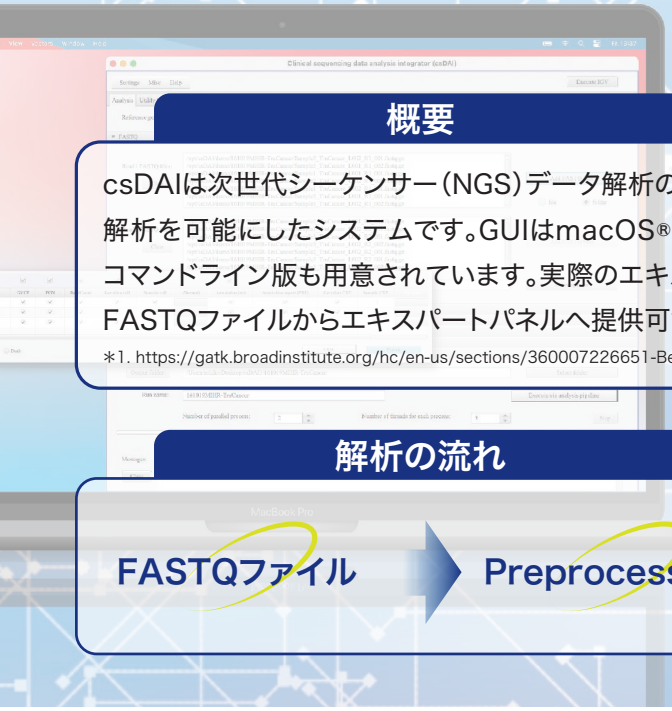
Clinical sequencing data analysis integrator (csDAI®)

macOS® 上のGUIからNGSデータを解析することが出来る統合解析システム



概要

csDAIは次世代シーケンサー(NGS)データ解析の統合パッケージであり、GUIからGATK Best Practices®*1に準拠した解析を可能にしたシステムです。GUIはmacOS®及びLinux®上で動作し、PCクラスター等のバッチシステムに対応したコマンドライン版も用意されています。実際のエキスパートパネルの現場と共に開発*2したアノテーションシステムにより、FASTQファイルからエキスパートパネルへ提供可能な資料作成までを簡易に実行することができます。

*1. https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows　*2. 国立がん研究センターとの共同研究

解析の流れ

FASTQファイル → Preprocess解析 → 変異解析 → アノテーション*

*個別データへ情報を付加すること

主な機能とソフトウエア

1 Quality control解析

- skewer
- bwa mem
- samtools
- verifyBamID
- picard
- GATK-4.2.6.1
- tabix
- 多数の独自ツール群

2 生殖細胞系列変異call

- HaplotypeCaller
- GermlineCNVCaller

3 体細胞変異call

- Mutect2
- cnv_somatic_pair_workflow

4 Call結果のアノテーション

- SnpEff
- HGMD®[1](ダウンロードライセンス契約が必要)
- COSMIC[2], ClinVar, gwasCatalog
- EnhancerAtlas, HACER
- HGVD, 38KJPN, GEM-J WGA
- ユーザー作成case/control

5 ゲノム構造変異解析

- Manta(Linux版のみ)

6 RNA-Seqデータ解析

- STAR
- STAR-Fusion, Arriba, FusionCatcher
- StringTie

7 補助機能

- Liftover
- 染色体read depth表示

アカデミックフリーソフトウエア/データ等は導入先にライセンスを用意して頂きます。

1. QIAGEN社のHGMD Downloadライセンスが必要
2. COSMICはWellcome Sanger Instituteのライセンスが必要

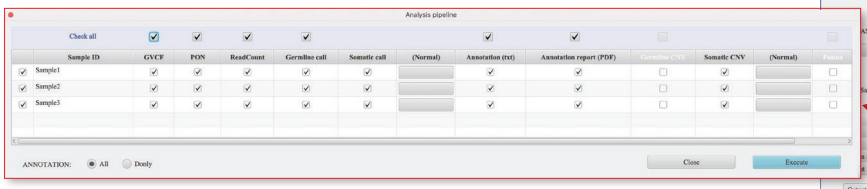
動作環境

項目	DNA-seqデータ	RNA-seqデータ
オペレーティングシステム	macOS® /Linux®	macOS®/Linux®
メインメモリー	16GB以上	32GB以上
ディスク容量	1TB以上	2TB以上
CPU	Intel® Core® i3/Apple M1®以上	Intel® Core® i5/Apple M1®以上
画面解像度	1920×1080以上	1920×1080以上

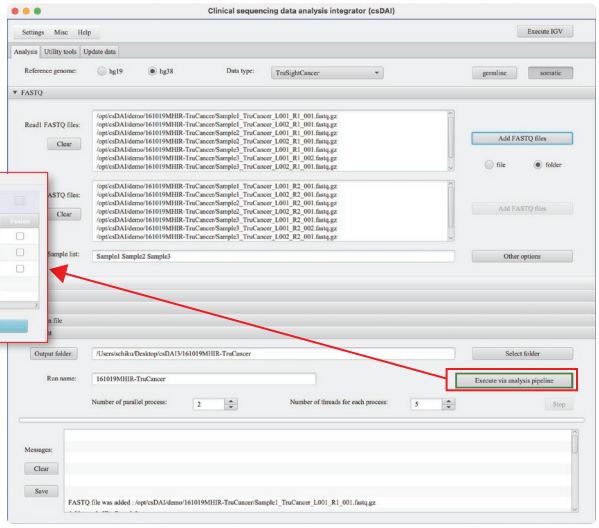
csDAIの実行方法

1. **FASTQファイルの指定（NGSランのフォルダーを選択）**
2. **Analysis pipelineで計算項目を確認**
3. **実行**

Analysis pipeline画面



csDAIのGUI画面からFASTQファイル群を選択した画面



解析結果

- QC report（text, PDF）
- Annotation report（text, PDF）
- CNV annotation（text）
- Fusion/Expression（text, PDF）

Preprocess解析

Quality control解析

- 統計値の一覧表（スプレッドシートで表示）
- 3つの表と検体毎の16個のグラフのPDF

1.1 161019MHIR-TruCancer

Table 1: 161019MHIR-TruCancer (TS): Number of exactly identical reads and adaptored reads

ID	lane#	#cluster	#PF	%PF	#identical	%ident	#adaptored	%adap.	#removed
Sample1	1	NA	167,756	NA	9,196	5.48	247	0.16	0
Sample1	2	NA	167,756	NA	9,190	5.48	253	0.16	2
Sample2	1	NA	121,853	NA	0	0.00	43	0.04	0
Sample2	2	NA	121,851	NA	0	0.00	54	0.04	0
Sample3	1	NA	207,800	NA	7	0.00	0	0.00	0
Sample3	2	NA	207,800	NA	2	0.00	0	0.00	0

Table 2: 161019MHIR-TruCancer: Base call informations

ID	#paired-read	read	%N	%A	%T	%G	%C	Q	Q_A	Q_T	Q_G	Q_C
Sample1	158,560	1	0.13	28.89	29.50	20.55	20.94	35.83	36.02	36.12	35.31	35.89
		2	0.13	28.96	29.34	20.71	20.86	35.69	35.88	35.97	35.17	35.77
Sample1	158,564	1	0.05	28.90	29.55	20.56	20.93	35.63	35.79	35.93	35.04	35.65
		2	0.10	29.02	29.38	20.68	20.82	35.44	35.61	35.73	34.90	35.51
Sample2	121,853	1	0.01	28.96	29.40	20.62	21.01	35.76	35.93	36.03	35.15	35.73
		2	0.00	29.13	29.25	20.85	20.76	35.72	35.88	35.97	35.14	35.75
Sample2	121,851	1	0.01	28.96	29.39	20.60	21.05	35.76	35.94	36.04	35.18	35.72
		2	0.00	29.12	29.28	20.84	20.76	35.71	35.86	35.96	35.13	35.73
Sample3	207,793	1	0.10	27.55	27.72	22.22	22.41	37.16	37.20	37.19	37.19	37.18
		2	0.10	27.50	27.68	22.31	22.42	33.22	33.30	33.31	33.19	33.19
Sample3	207,798	1	0.10	27.56	27.73	22.23	22.38	37.16	37.20	37.19	37.19	37.18
		2	0.10	27.50	27.68	22.29	22.43	33.22	33.31	33.30	33.18	33.18

Table 3: 161019MHIR-TruCancer: Mapping for Homo_sapiens_assembly38.fasta and depth on TruSightCancer.hg19to38

ID	Proper map(%) NotClip.	Clip.	Number of reads by samtools #mapped	%dup.	#proper	#on bait	%on bait	Cont. estim.	Depth mean
Sample1	94.09	5.88	316,921	3.05	307,190	260,545	84.82	0.00022	62.51
Sample1	94.46	5.50	317,028	2.97	307,504	260,606	84.75	0.00000	62.53
Sample2	94.30	5.30	243,637	2.63	236,404	198,972	84.17	0.00000	48.17
Sample2	94.40	5.20	243,626	2.53	236,610	199,181	84.18	0.00000	48.37
Sample3	99.19	0.81	415,586	0.18	414,838	301,747	72.74	0.00000	65.31
Sample3	99.17	0.83	415,596	0.17	414,890	301,178	72.59	0.00013	65.17

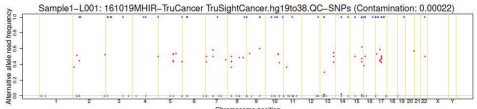
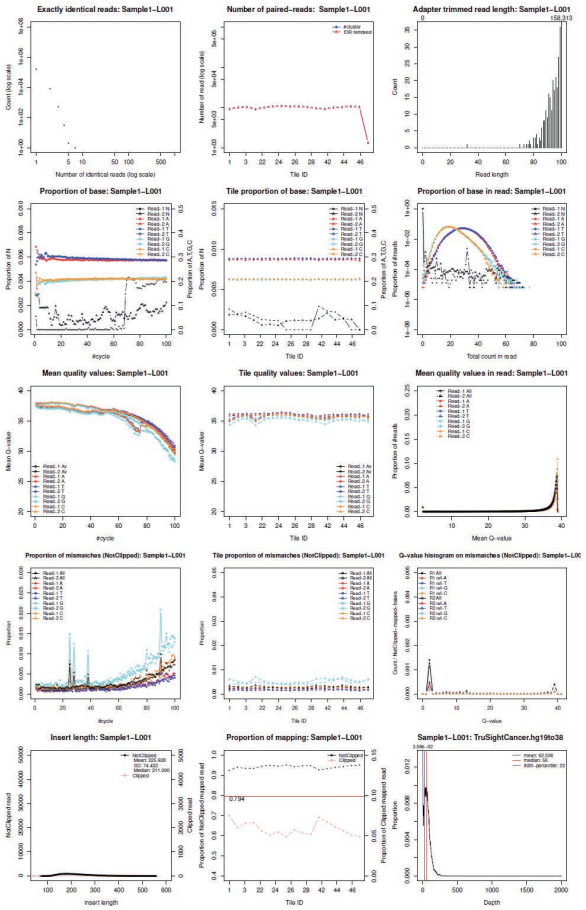


Figure 1: Sample1 (161019MHIR-TruCancer:Lane1) QC figures

- 他の確認可能項目
  - ✓ 性別の確認（性染色体上にbait領域が設定されている場合）
  - ✓ 検体間の血縁関係の推定



中間ファイル計算

- GVCF（Germline call）
- Read count（コピー数解析）
- Panel of normal（Mutect2）
- GTF（StringTie）

変異解析

使用ツール・パイプライン

項目	DNA-seqデータ	RNA-seqデータ
Germline call	HaplotypeCaller	HaplotypeCaller
Somatic call	Mutect2	Mutect2
Germline copy number	GermlineCNVCaller	-
Somatic copy number	cnv_somatic_pair_workflow	-
Structural variation	Manta	-
Expression	-	stringtie
Fusion	-	STAR-Fusion, Arriba, FusionCatcher

Annotation

csDAIのアノテーションシステムではスプレッドシートで絞り込みを行うことができるテキストファイルの一覧表と、指定した頻度以下の変異毎にアノテーション項目の表とIGV*によるVCF及びBAMファイルのスナップショットを取り込んだPDFファイルを提供します。

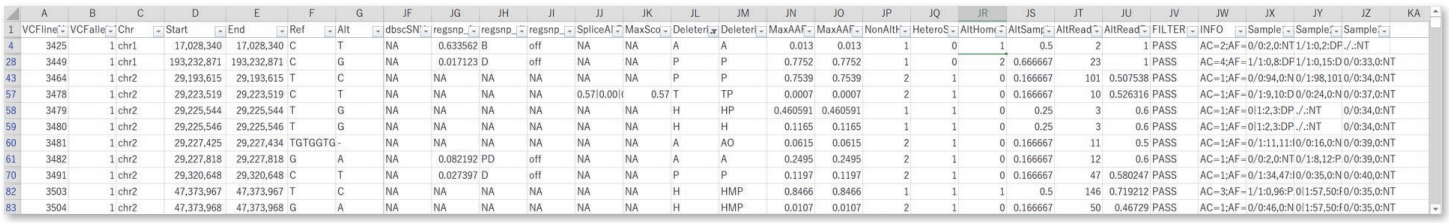
*. https://software.broadinstitute.org/software/igv/

アノテーション項目

カテゴリ	概要
遺伝子情報	SnpEff[1]によるEnsembl[2](GRCh38.105)とRefSeq[3](GRCh38.p14)
ゲノム特徴量	UCSC[4]ゲノム領域情報(cytoBand、genomicSuperDups、rmsk)
集団頻度	NHLBI-ESP[5]、ExAC、1000genome[6]、gnomAD[7]、Kaviar[8]、HRC[9]、ABraOM[10]、京都大学のHGVD[11]、東北メディカル・メガバンク(ToMMo)[12]、TogoVarのGEM-J WGA[13]
変異・疾患情報	dbSNP[14]、GWAS catalog[15]、ICGC[16]、COSMIC[17]、HGMD[18]情報、ClinVar[19]の情報
エンハンサー領域	EnhancerAtlas2.0[20]及びHACER[21]の情報
In silico 予測	dbNSFP[22]、SpliceAI[23](SNV, 1 bp insertions and 1-4 bp deletions[24])のスコア0.2以上の情報
統合情報	Deleterious フラグ(独自の変異重要度指標)、集団頻度中の最大allele頻度(MaxAAF)
遺伝子型call情報	遺伝子型とallele毎のdepth情報及び独自のハードフィルター情報

[1] http://pcingola.github.io/SnpEff/
[2] https://asia.ensembl.org/index.html
[3] https://www.ncbi.nlm.nih.gov/refseq/
[4] https://genome.ucsc.edu/
[5] https://esp.gs.washington.edu/drupal/
[6] https://www.internationalgenome.org/
[7] https://gnomad.broadinstitute.org/
[8] https://db.systemsbiology.net/kaviar/
[9] http://www.haplotype-reference-consortium.org/
[10] https://abraom.ib.usp.br/
[11] https://www.hgvd.genome.med.kyoto-u.ac.jp/
[12] https://jmorp.megabank.tohoku.ac.jp/202206/
[13] https://togovar.biosciencedbc.jp/doc/ja/datasets/gem_j_wga
[14] https://www.ncbi.nlm.nih.gov/snp/
[15] https://www.ebi.ac.uk/gwas/
[16] https://dcc.icgc.org/
[17] https://cancer.sanger.ac.uk/cosmic
[18] https://www.hgmd.cf.ac.uk/ac/index.php
[19] https://www.ncbi.nlm.nih.gov/clinvar/
[20] http://www.enhanceratlas.org/
[21] http://bioinfo.vanderbilt.edu/AE/HACER/
[22] https://sites.google.com/site/jpopgen/dbNSFP
[23] https://github.com/Illumina/SpliceAI
[24] https://basespace.illumina.com/s/otSPW8hnhaZR

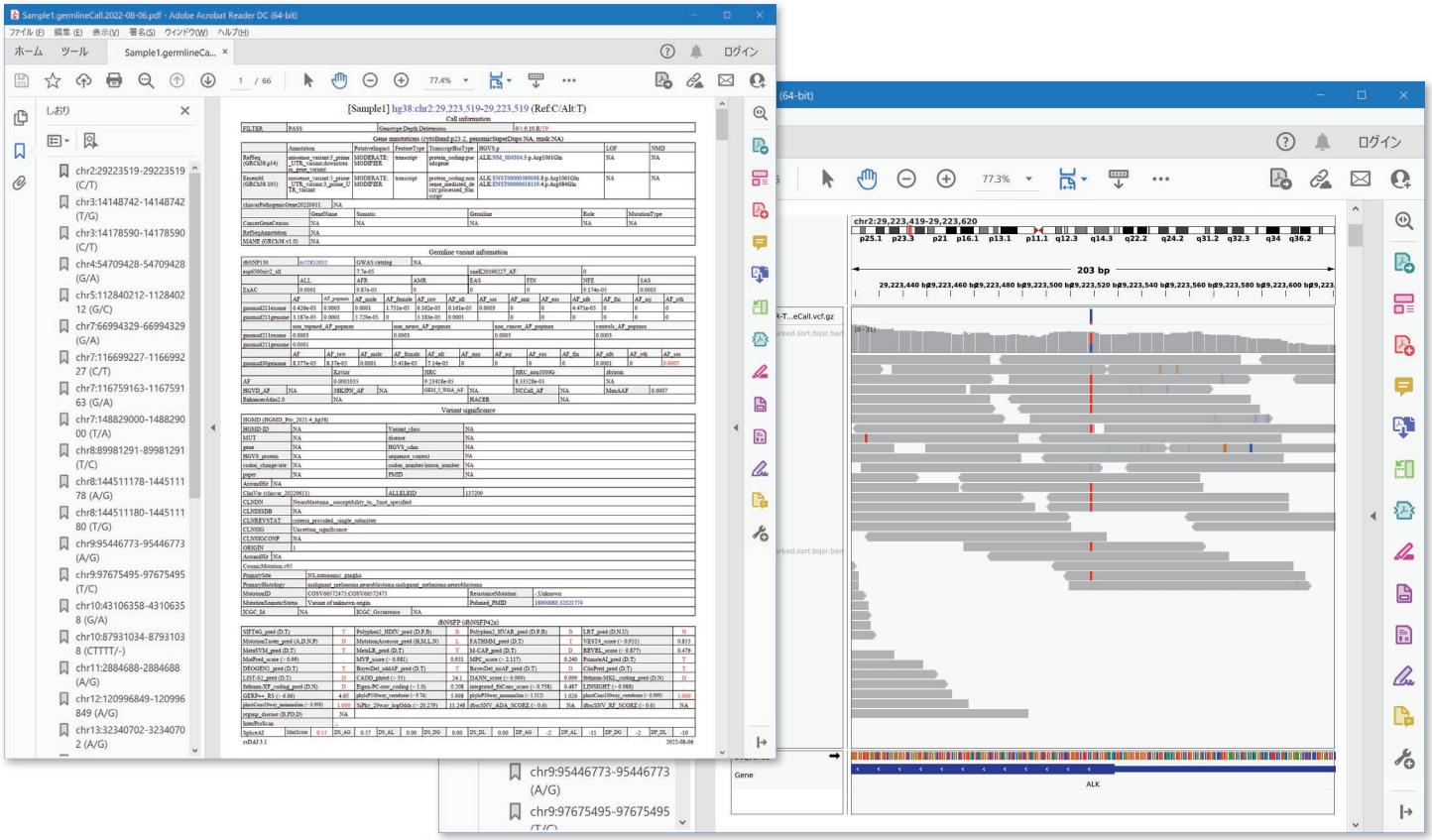
アノテーションの一覧表示例



	A	B	C	D	E	F	G	JF	JG	JH	JI	JJ	JK	JL	JM	JN	JO	JP	JQ	JR	JS	JT	JU	JV	JW	JX	JY	JZ	KA
1	VCFline	VCFalle	Chr	Start	End	Ref	Alt	dbscSN	regsnp_	regsnp_	regsnp_	SpliceAI	MaxSco	Deleteri	Deleteri	MaxAAF	MaxAAF	NonAltF	HeteroS	AltHom	AltSamp	AltRead	AltRead	FILTER	INFO	Sample	Sample	Sample	
4	3425		1	chr1	17,028,340	17,028,340	C	T		NA	0.633562	B		off	NA	NA	A	A	0.013	0.013	1	0	1	0.5	2	1	PASS	AC=2;AF=0/0:2,0:NT	1/1:0,2:DP./.:NT		
28	3449		1	chr1	193,232,871	193,232,871	C	G		NA	0.017123	D		off	NA	NA	P	P	0.7752	0.7752	1	0	2	0.666667	23	1	PASS	AC=4;AF=1/1:0,8:DP	1/1:0,15:D	0/0:33,0:NT	
43	3464		1	chr2	29,193,615	29,193,615	T	C		NA	NA	NA		NA	NA	NA	P	P	0.7539	0.7539	2	1	0	0.166667	101	0.507538	PASS	AC=1;AF=0/0:94,0:N	0/1:98,101	0/0:34,0:NT	
57	3478		1	chr2	29,223,519	29,223,519	C	T		NA	NA	NA		NA	0.57(0.00(	0.57	T	TP	0.0007	0.0007	2	1	0	0.166667	10	0.526316	PASS	AC=1;AF=0/1:9,10:D	0/0:24,0:N	0/0:37,0:NT	
58	3479		1	chr2	29,225,544	29,225,544	T	G		NA	NA	NA		NA	NA	NA	H	HP	0.460591	0.460591	1	1	0	0.25	3	0.6	PASS	AC=1;AF=0/1:2,3:DP	/.::NT	0/0:34,0:NT	
59	3480		1	chr2	29,225,546	29,225,546	T	G		NA	NA	NA		NA	NA	NA	H	H	0.1165	0.1165	1	1	0	0.25	3	0.6	PASS	AC=1;AF=0/1:2,3:DP	/.::NT	0/0:34,0:NT	
60	3481		1	chr2	29,227,425	29,227,434	TGTGGTG	-		NA	NA	NA		NA	NA	NA	A	AO	0.0615	0.0615	2	1	0	0.166667	11	0.5	PASS	AC=1;AF=0/1:11,11	I0/0:16,0:N	0/0:39,0:NT	
61	3482		1	chr2	29,227,818	29,227,818	G	A		NA	0.082192	PD		off	NA	NA	A	A	0.2495	0.2495	2	1	0	0.166667	12	0.6	PASS	AC=1;AF=0/0:2,0:NT	0/1:8,12:P	0/0:39,0:NT	
70	3491		1	chr2	29,320,648	29,320,648	C	T		NA	0.027397	D		off	NA	NA	P	P	0.1197	0.1197	2	1	0	0.166667	47	0.580247	PASS	AC=1;AF=0/1:34,47	I0/0:35,0:N	0/0:40,0:NT	
82	3503		1	chr2	47,373,967	47,373,967	T	C		NA	NA	NA		NA	NA	NA	H	HMP	0.8466	0.8466	1	1	1	0.5	146	0.719212	PASS	AC=3;AF=1/1:0,96:P	0|1:57,50:F	0/0:35,0:NT	
83	3504		1	chr2	47,373,968	47,373,968	G	A		NA	NA	NA		NA	NA	NA	H	HMP	0.0107	0.0107	2	1	0	0.166667	50	0.46729	PASS	AC=1;AF=0/0:46,0:N	0|1:57,50:F	0/0:35,0:NT	

アノテーションレポート

Deleterious　flagがNA以外で且つMaxAAFが0.05(default値の場合)以下の変異について検体毎にPDF形式のアノテーションレポートを作成します。



RNA-seqデータ解析の出力

RNA-seqデータの解析ではtranscripts per million（TPM）の発現量にGENCODE*1の遺伝子シンボルやGene_type、Transcript_type情報を加えたスプレッドシートで閲覧できる一覧表を作成します。また融合遺伝子解析では各ソフトウエアの出力結果を統合し、Arriba*2からは融合遺伝子の構造を示したPDFファイルを得ることが出来ます。

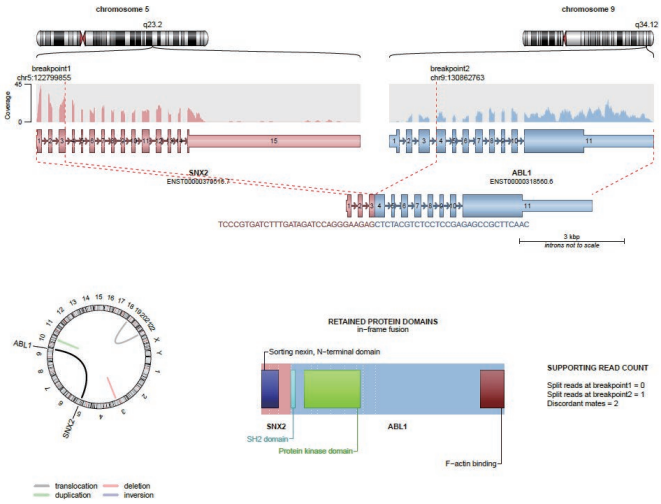
遺伝子毎のTPM

	A	B	C	D	E	F	G	H	I	J	K
1	ENSGs	Symbol	Chr	Strand	Start	End	NumRe	Gene_type	Mean	DRR01411	DRR01411
2	ENSG00000223972.5	DDX11L1	chr1	+	11,869	14,409	1	transcribe	0.28776	0	0.575519
3	ENSG00000227232.5	WASH7P	chr1	−	14,404	29,570	1	unprocess	20.55396	15.31022	25.79771
4	ENSG00000278267.1	MIR6859−1	chr1	−	17,369	17,436	1	miRNA	0	0	0
5	ENSG00000243485.5	MIR1302−2H	chr1	+	29,554	31,109	1	lncRNA	0	0	0
6	ENSG00000284332.1	MIR1302−2	chr1	+	30,366	30,503	1	miRNA	0	0	0
7	ENSG00000237613.2	FAM138A	chr1	−	34,554	36,081	1	lncRNA	0	0	0
8	ENSG00000268020.3	OR4G4P	chr1	+	52,473	53,312	1	unprocess	0	0	0
9	ENSG00000240361.2	OR4G11P	chr1	+	57,598	64,116	1	transcribe	0	0	0
10	ENSG00000186092.6	OR4F5	chr1	+	65,419	71,585	1	protein_co	0	0	0

転写物毎のTPM

	A	B	C	D	E	F	G	H	I	J	K
1	Transcript_id	Transcript_type	Chr	Strand	Start	End	Gene_id	Gene_name	Mean	DRR01411	DRR01411
2	ENST00000456328.2	processed_transcript	chr1	+	11,869	14,409	ENSG0000022	DDX11L1	0.28776	0	0.575519
3	ENST00000450305.2	transcribed_unprocess	chr1	+	12,010	13,670	ENSG0000022	DDX11L1	0	0	0
4	ENST00000488147.1	unprocessed_pseudog	chr1	−	14,404	29,570	ENSG0000022	WASH7P	20.55396	15.31022	25.79771
5	ENST00000619216.1	miRNA	chr1	−	17,369	17,436	ENSG0000027	MIR6859−1	0	0	0
6	ENST00000473358.1	lncRNA	chr1	+	29,554	31,097	ENSG0000024	MIR1302−2HG	0	0	0
7	ENST00000469289.1	lncRNA	chr1	+	30,267	31,109	ENSG0000024	MIR1302−2HG	0	0	0
8	ENST00000607096.1	miRNA	chr1	+	30,366	30,503	ENSG0000028	MIR1302−2	0	0	0
9	ENST00000417324.1	lncRNA	chr1	−	34,554	36,081	ENSG0000023	NA	0	0	0
10	ENST00000461467.1	lncRNA	chr1	−	35,245	36,073	ENSG0000023	NA	0	0	0

Arribaよる融合遺伝子構造図の出力例



*1. https://www.gencodegenes.org/
*2. https://github.com/suhrig/arriba

Utility tools

Utility toolsでは様々な補助機能を提供しています。外部から持ち込まれたVCFに含まれる検体の確認や、検体数の増加に伴うコピー数解析などのまとめ解析の場合に有用です。

Utility tools機能一覧

機能名	内容
Annotation (chromosomal location)	ゲノム位置情報と変異情報を入力し、csDAIのアノテーション情報を付与する。 VCFとBAMを指定するとIGVのスナップショットも付与される。
Merge BAM	複数のBAMファイルをマージする。Readを追加した場合などに使用する。
Merge Mutect2 PON	Preprocessで計算したpon.vcf.gzファイルを統合してMutect2用のpanel of nomarlファイルを作成する。
Somatic CNV PON	Preprocessで計算したread countファイル（tsv）からGATKのcnv_somatic_pair_workflowで使用するsomatic CNV用のpanel of normalファイルを作成する。
Chromosomal AARF view	VCFファイルからalternative allele read数割合（variant allele frequency; VAF）の染色体プロットを作成する。 がん部と非がん部の確認などに用いることが出来る。
Genotype concordance	VCFファイル（複数指定可）からサンプル間の遺伝子型一致割合を計算する。 サンプル間の血縁関係の確認に用いることが出来る。
Liftover	ゲノム位置座標をhg19からhg38もしくはhg38からhg19に変換する。
From Takara Bio FASTQ	タカラバイオから送付されたHDDにコピーされているFASTQファイルをcsDAI形式に変換する。

ACMGガイドライン*1バリアント評価データベースシステム*2

アノテーション項目はcsDAIのバージョンアップと共に増えています。この様な項目が変わっていくデータの場合リレーショナルデータベースの構築は不向きです。そこでXML DBを利用したデータベース機能を提供することで、csDAIがバージョンアップしても過去のアノテーションデータに対してユーザーが付与したACMGガイドラインのバリアント評価結果DBを再構築無しに検索することが出来ます。

*1. Richards et al., Genetics in Medicine volume 17, pages405–423(2015), doi: 10.1038/gim.2015.30
*2. 本機能はオプションであり、実際の構築時はご相談の上で仕様を調整させて頂きます。

問い合わせ先
みずほリサーチ&テクノロジーズ 情報通信研究部社会基盤技術チーム
〒101-8443 東京都千代田区神田錦町 2-3
TEL：03-5281-5289　FAX：03-5281-3457
e-mail：icd-joho@mizuho-rt.co.jp
URL：https://www.mizuho-rt.co.jp

ご相談やお見積もり等の詳細については、
左記までお気軽にお問い合わせください。