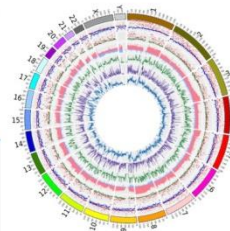
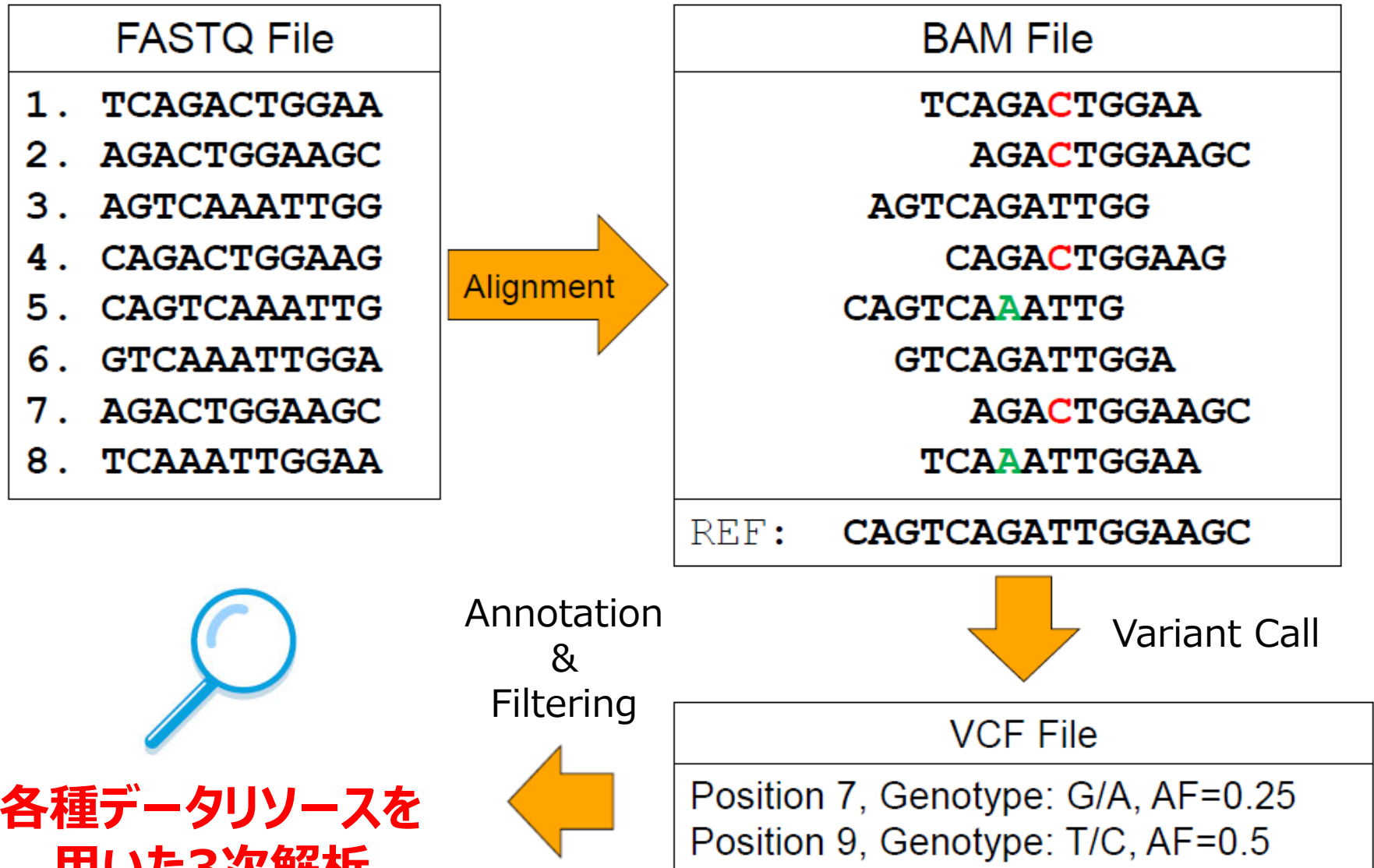


外部データリソースを利用した DNA-Seq疾患ゲノム解析

フィルジエン株式会社 バイオサイエンス部
(biosupport@filgen.jp)

- 次世代シーケンサーを用いたゲノム変異解析（DNA-Seq）では、大量の変異データが産出されるため、その中から重要な変異を見つけ出す必要がある。
- 公共の変異データベースなどでは、疾患との関連や、人種ごとの既知SNP、また生体への影響をスコア化したものなどがあり、これらの外部リソースを活用することで、効率的な解析を行うことが可能となる。



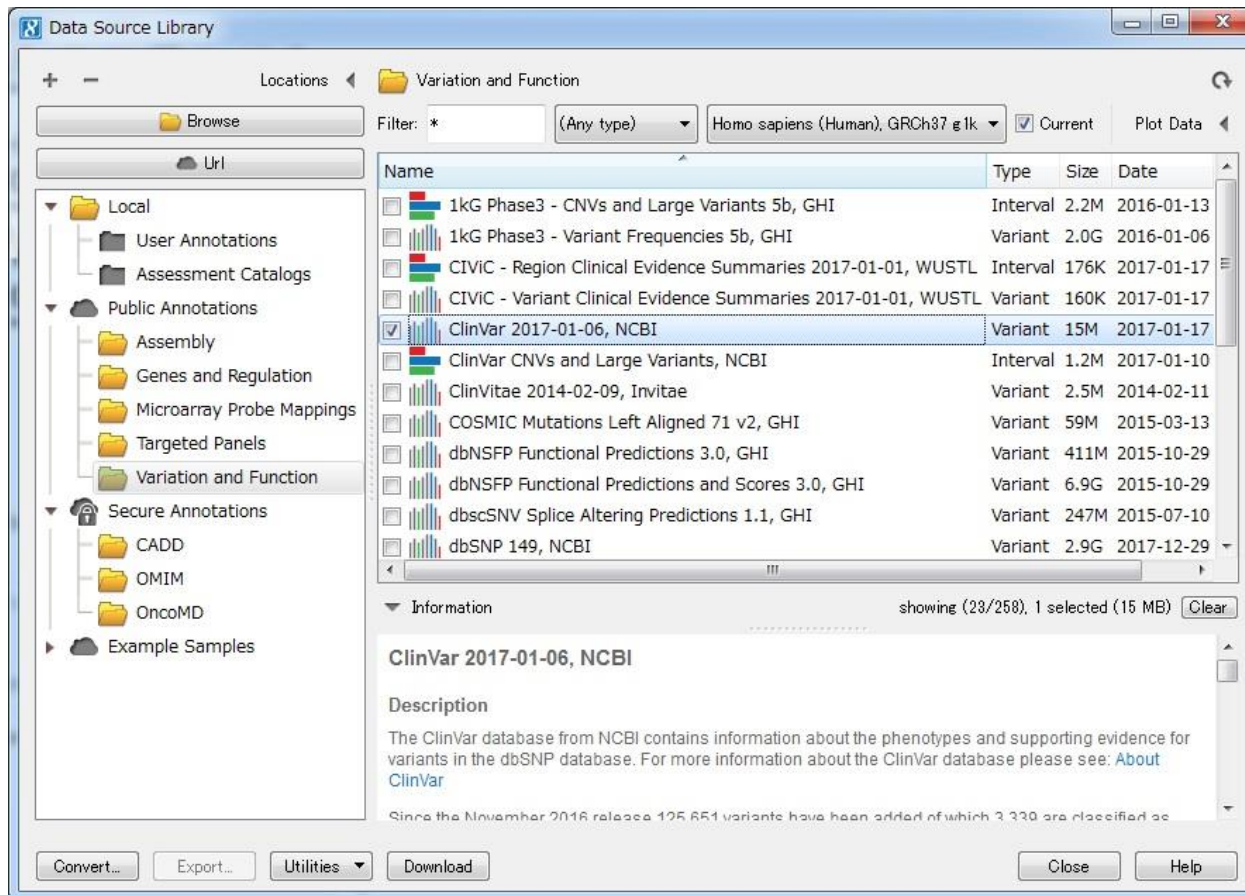




- **GWAS & SNP Analysis**
- **Large-N DNA-Seq Analysis**
- **Genomic Prediction**
- **Copy Number Analysis**
- **RNA-Seq Analysis**



- **Variant Interpretation**
- **Cancer Diagnostics**
- **CNV Calling**
- **Clinical Reporting**
- **High-throughput NGS Testing**



- どちらもデータベース管理ツールを搭載し、簡単な操作で各種データソースのデータを専用サーバーよりダウンロードが可能。
- 各データソースは、Golden Helix社によってメンテナンスされており、高品質なアノテーションデータを使用できる。

生体への影響などを予測してスコア化したもの

- 例：dbNSFP, CADDなど。
- 特長：各変異を様々な条件で数値化し、検索時の優先順位などを付けることができる。

疾患との関連が明らかになっているもの

- 例：ClinVar, OMIM, COSMIC, OncoMDなど。
- 特長：論文などで、すでに疾患との関連が報告されている変異を、容易にピックアップできる。

人種ごとのアレル頻度をまとめたもの

- 例：1000 Genome, HapMap, HGVDなど。
- 特長：特定の人種や民族ごとに、集団内に高頻度で存在するコモンSNPなどを排除して検索を行うことができる。

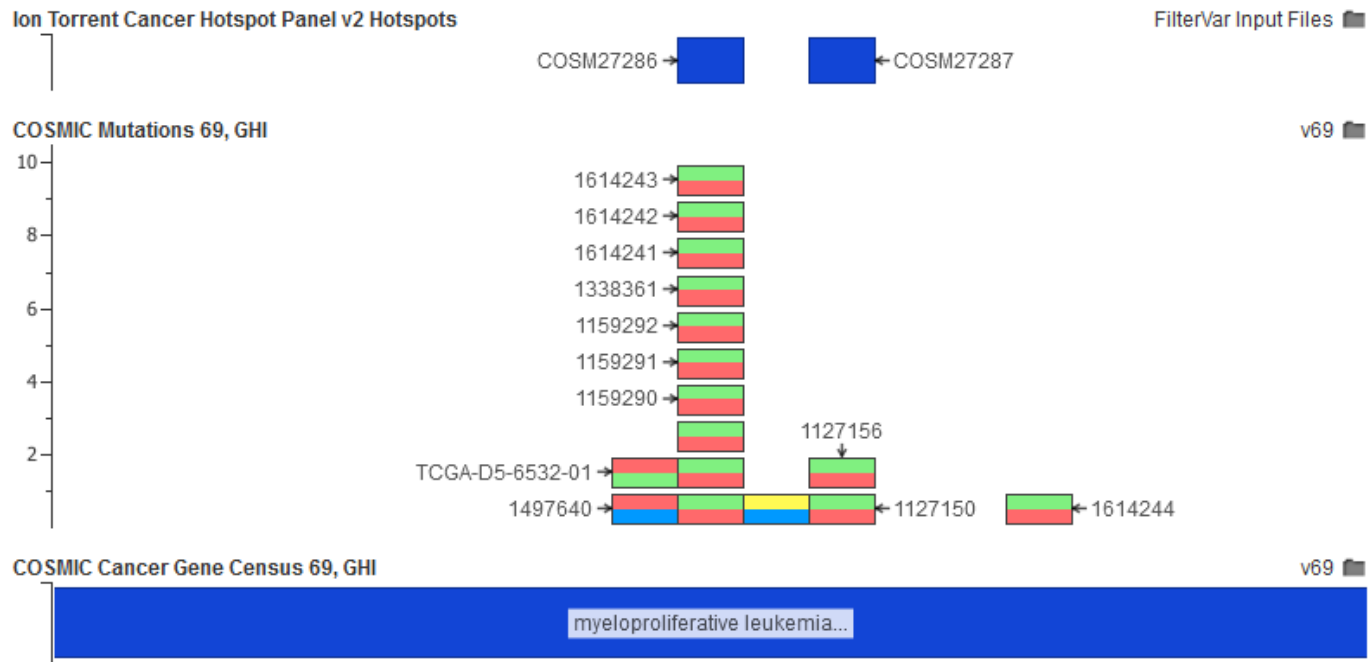
- **SIFT, Polyphen2, Mutation Taster, MutationAssessor, FATHMM, FATHMM-MKL** の6種類のスコア化アルゴリズムを同時に使用し、変異の評価が可能。
- 各アルゴリズムは、生物種間の配列保存性や、タンパク質立体構造への影響などから、各変異の生体への影響を評価している。
- データソースには、各アルゴリズムごとの評価結果スコアの生データのもの、スコアから Tolerated, Damaging などのカテゴリー名に変換したものの2種類が使用可能。

N of 6 Predicted Damaging	SIFT Pred (C)	Polyphen2 HVAR Pred (C)	MutationTaster Pred (C)
3 of 6 Predicted as Damaging	Tolerated	Benign	Damaging
0 of 6 Predicted as Damaging	?	?	Tolerated
3 of 6 Predicted as Damaging	Damaging	Possibly damaging	Tolerated
1 of 6 Predicted as Damaging	Tolerated	Benign	Tolerated
2 of 6 Predicted as Damaging	Tolerated	Benign	Damaging
0 of 6 Predicted as Damaging	?	Benign	Tolerated
0 of 6 Predicted as Damaging	Tolerated	Benign	Tolerated
4 of 6 Predicted as Damaging	Damaging	Probably damaging	Damaging
1 of 6 Predicted as Damaging	Tolerated	Benign	Tolerated
2 of 6 Predicted as Damaging	Damaging	Benign	Tolerated
0 of 6 Predicted as Damaging	Tolerated	Benign	Tolerated
0 of 6 Predicted as Damaging	Tolerated	Benign	Tolerated
3 of 6 Predicted as Damaging	Tolerated	Probably damaging	Damaging
2 of 6 Predicted as Damaging	Tolerated	Benign	Tolerated
6 of 6 Predicted as Damaging	Damaging	Probably damaging	Damaging

- OMIMやHGMDなど、ヒトの変異と疾患との関連情報をまとめたデータベース。
- 各変異について、関連する疾患名の他に、「Pathogenic」「Benign」といった Clinical Significance データも含み、病原性の高い変異のみを容易に検索が可能になる。
- Review Statusも検索条件に入れることができ、疾患との関連情報の投稿者が多い変異の検索も可能。

The screenshot shows the ClinVar website interface. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' links. Below this is a search bar with the text 'Search ClinVar for gene symbols, HGVS expressions, conditions, and more' and a 'Search' button. A 'Help' link is also visible. Below the search bar is a navigation menu with links for 'Home', 'About', 'Access', 'Help', 'Submit', 'Statistics', and 'FTP'. The main content area features a dark blue header with the ClinVar logo and the text 'ClinVar aggregates information about genomic variation and its relationship to human health.' Below this header, there are three columns of links: 'Using ClinVar' (including About ClinVar, Data Dictionary, Downloads/FTP site, FAQ, Contact Us, RSS feed/What's new?, and Factsheet), 'Tools' (including ACMG Recommendations for Reporting of Incidental Findings, ClinVar Submission Portal, Submissions, Variation Viewer, Clinical Remapping - Between assemblies and RefSeqGenes, RefSeqGene/LRG, and Variation Reporter), and 'Related Sites' (including ClinGen, GeneReviews®, GTR®, MedGen, OMIM®, and Variation).

- サンガー研究所が作成している、がんの体細胞変異の情報を集めたデータベース。
- 論文報告されている、変異と関連するがんのタイプや、原発部位情報などが登録されている。
- The Cancer Genome Atlasの情報を統合した、COSMIC v71のデータを使用が可能。



- 1208名の日本人エクソーム解析によって検出された、遺伝型データのカタログ。
- アリルごとの検出サンプル数の情報を含み、変異データのフィルタリングに使用することで、日本人に高頻度で存在するアリルを除去することができる。
- SNP & Variation SuiteとVarSeqソフトウェアには、カスタムフォーマットファイルとしてインポートが可能。

Convert Source Wizard

Convert Data Source

① Define Input
② Scan Input
③ Change Options
④ Convert

Specify how the text file(s) are delimited and how to detect the field names.
You must also indicate which fields provide genomic coordinates. Click on the field headers in the Preview table to set these fields.

Advanced Options

Help

Delimited Text File Characteristics

Field Name Line: Manual Names [] First Data Line: 0 []

Ignore Lines: Don't Ignore []

Field Delimiter: Tab [] List Delimiter: Comma []

Missing Values: ? n/a nan ?_? []

Coordinates: 0-Based Interval 1-Based Interval Position (1bp width)

Preview:

Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	
chr1	69270	rs201219564	A	G	49	VQS
chr1	69510	.	C	T	254	Low
chr1	69511	rs2691305	A	G	251	[Low
chr1	69513	rs770590115	A	G	259	Low
chr1	69537	.	G	T	291	Low
chr1	69849	rs776815449	G	A	47	VQS
chr1	69876	rs568270584	A	G	50	VQS

< Back Next > Cancel

CADD

- 変異の有害性をスコア化したデータベース。
- データベースに含まれないInsertion/Deletionについては、シミュレーションによりスコアを自動的に計算する。

OMIM

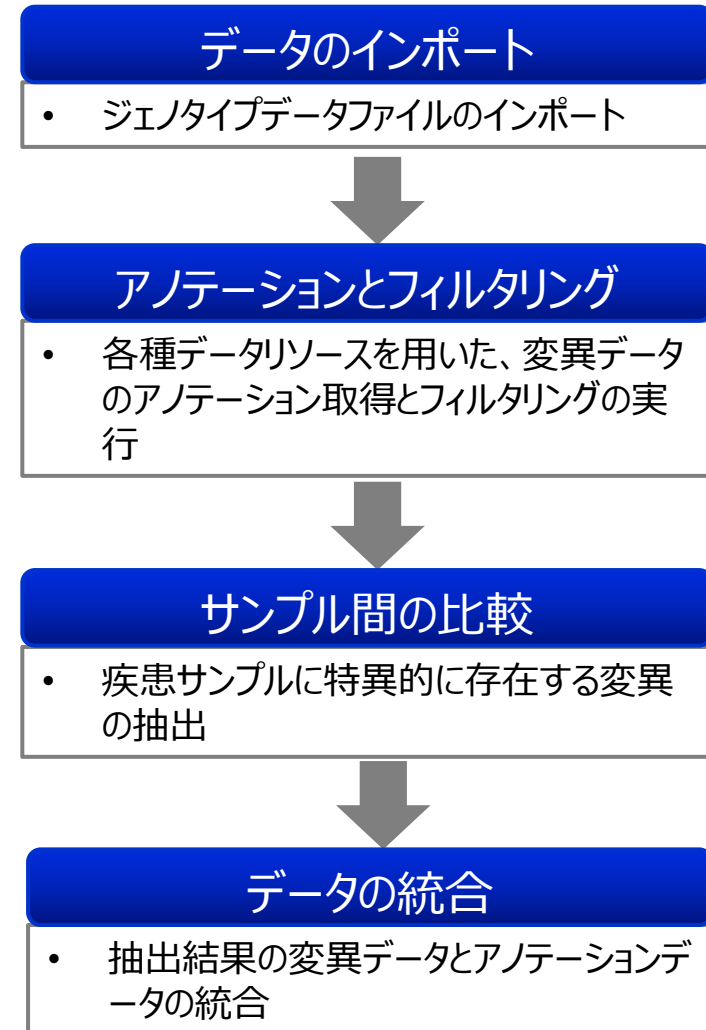
- ヒトの遺伝子と遺伝病の情報をまとめたデータベース。
- 遺伝病と関連する遺伝子名や遺伝形式、変異情報などを含み、データベースのIDや出典論文の情報も取得できる。

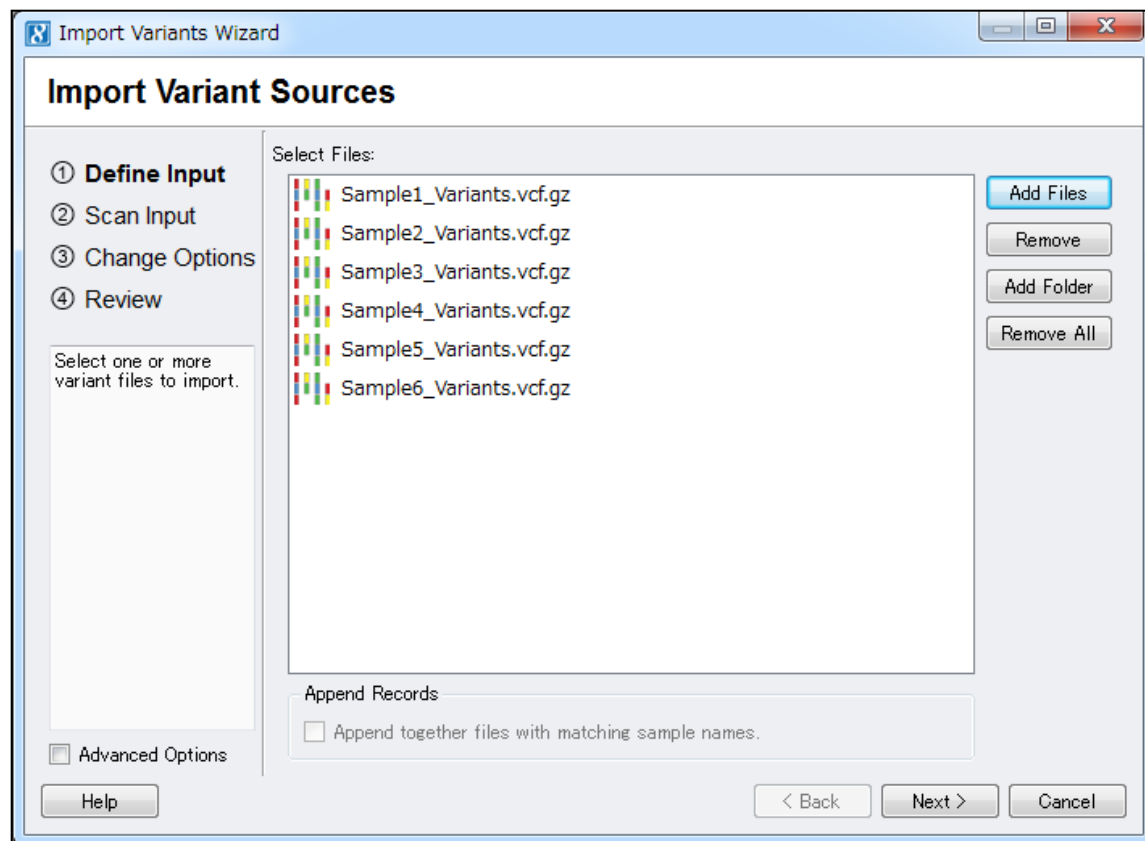
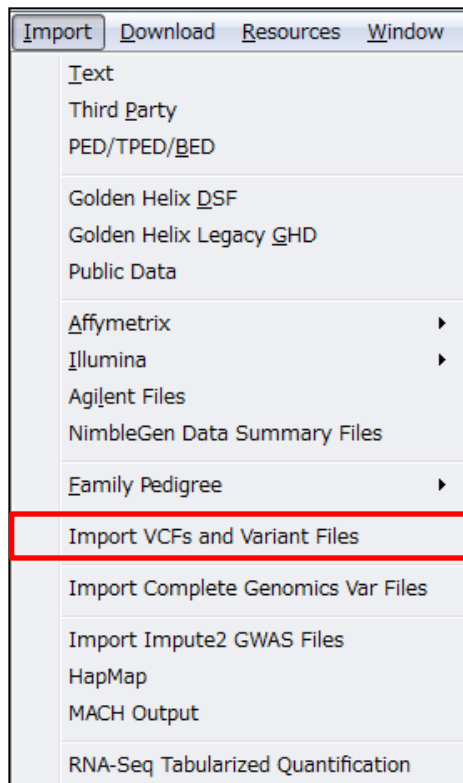
OncoMD

- がんと関連する遺伝子や変異などの情報をまとめたデータベース。
- 臨床試験情報や承認薬に対する反応率、論文情報なども得ることができる。

使用するジェノタイプデータ:

- 疾患サンプル数: 3例
- 健常者サンプル数: 3例
- ファイルフォーマット: VCFファイル
- 変異数: 約300,000個





- ジェノタイプデータファイルは、バリアントコール用ツールなどで作成した、VCFファイルを使用する。

ジェノタイプデータファイル (VCFファイル) のインポート

Import Variants Wizard

Import Variant Sources

① Define Input
② Scan Input
③ **Change Options**
④ Review

Select the samples of interest and appropriately adjust their attributes

Add sample fields:

	Original Samples	<input checked="" type="checkbox"/> Sample Source File Name	Samples	Affection Status
1	Sample1_Variants	Sample1_Variants	Sample1_Variants	<input type="button" value="True"/>
2	Sample2_Variants	Sample2_Variants	Sample2_Variants	<input type="button" value="True"/>
3	Sample3_Variants	Sample3_Variants	Sample3_Variants	<input type="button" value="True"/>
4	Sample4_Variants	Sample4_Variants	Sample4_Variants	<input type="button" value="False"/>
5	Sample5_Variants	Sample5_Variants	Sample5_Variants	<input type="button" value="False"/>
6	Sample6_Variants	Sample6_Variants	Sample6_Variants	<input type="button" value="False"/>

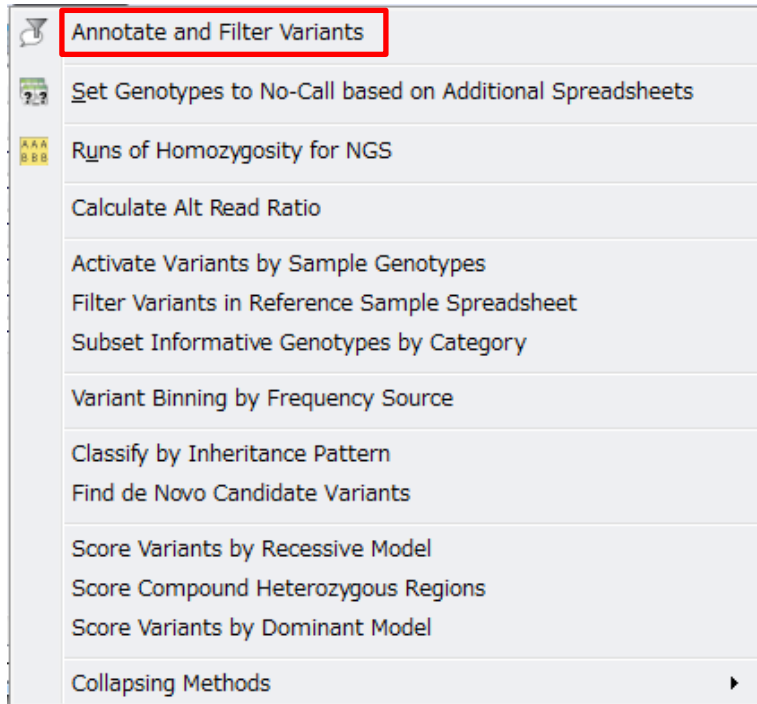
Advanced Options

Change the sample names:

- 各データの疾患／正常サンプルの設定を行う。
- サンプル数が多い場合は、サンプル情報を記載してあるテキストファイルを使用して、設定を行うこともできる。

Unsort		B	2	C	3	G	4	G	5	G	6	G	7
Map		Samples	Affection Status	Sample Source File Name		1:10109-Ins	1:13116-SNV	1:13118-SNV	1:13649-SNV				
Chromosome						1	1	1	1				
Position						10109	13116	13118	13649				
Identifier						?	?	?	?				
Reference						-	T	A	G				
Alternates						A	G	G	C				
1	Sample1_Variants		1	Sample1_Variants		??	G_T	A_G	??				
2	Sample2_Variants		1	Sample2_Variants		??	G_T	A_G	??				
3	Sample3_Variants		1	Sample3_Variants		??	G_T	A_G	C_G				
4	Sample4_Variants		0	Sample4_Variants		-_A	??	??	??				
5	Sample5_Variants		0	Sample5_Variants		-_A	??	??	??				
6	Sample6_Variants		0	Sample6_Variants		??	??	??	??				

- データがインポートされ、スプレッドシートにまとめて表示される。
- VCFファイルに各アレルのリード深度やカバレッジ情報などが含まれている場合は、それらのデータは別シートにインポートされ、フィルタリングなどに使用できる。



以下データリソースを使用し、サンプルの変異データの
アノテーション取得とフィルタリングを実行

1. HGVD

Alternate alleleを100サンプル以上もつ変異を除外

2. RefSeq Genes

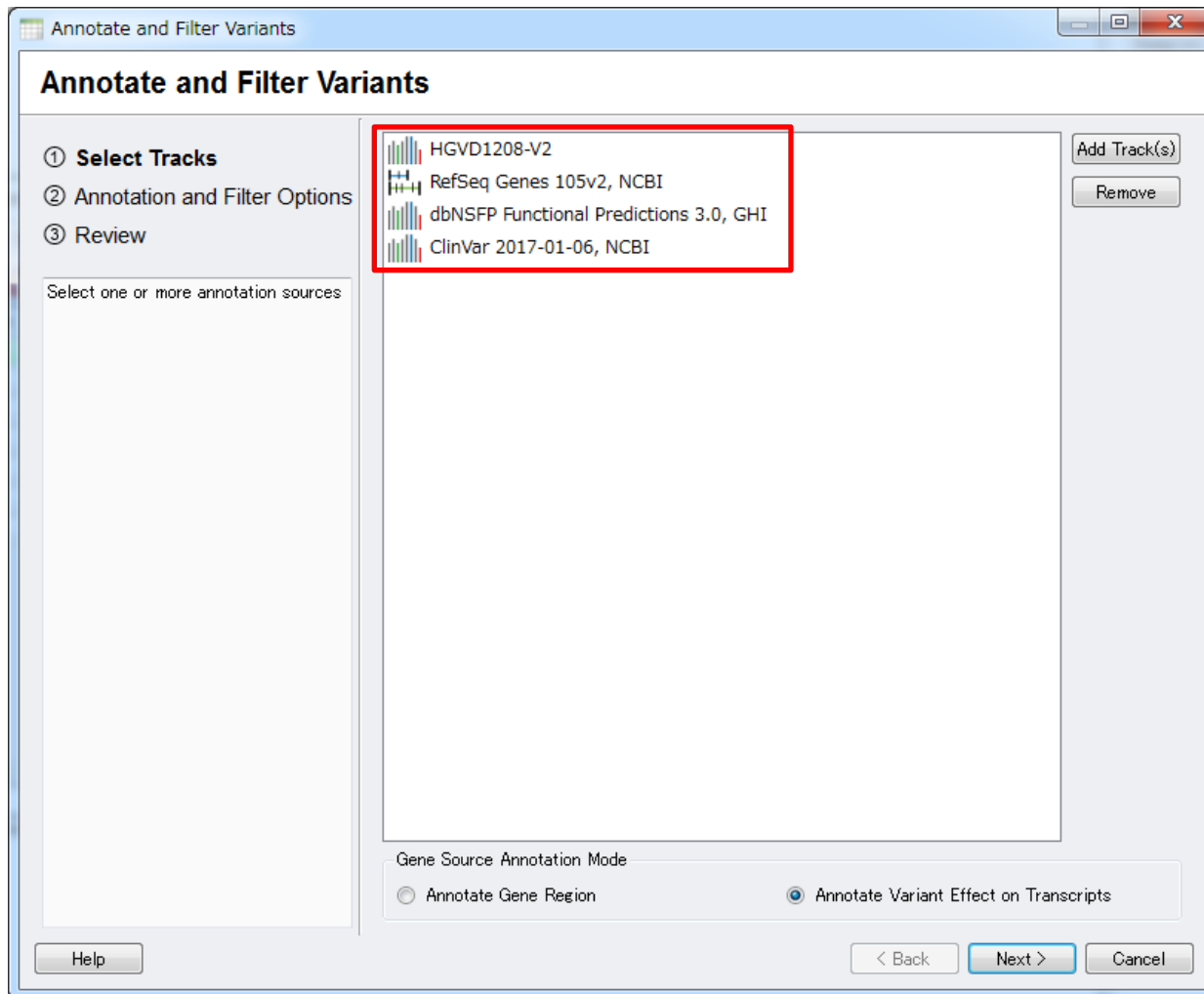
エクソン領域内の変異のみを抽出

3. dbNSFP

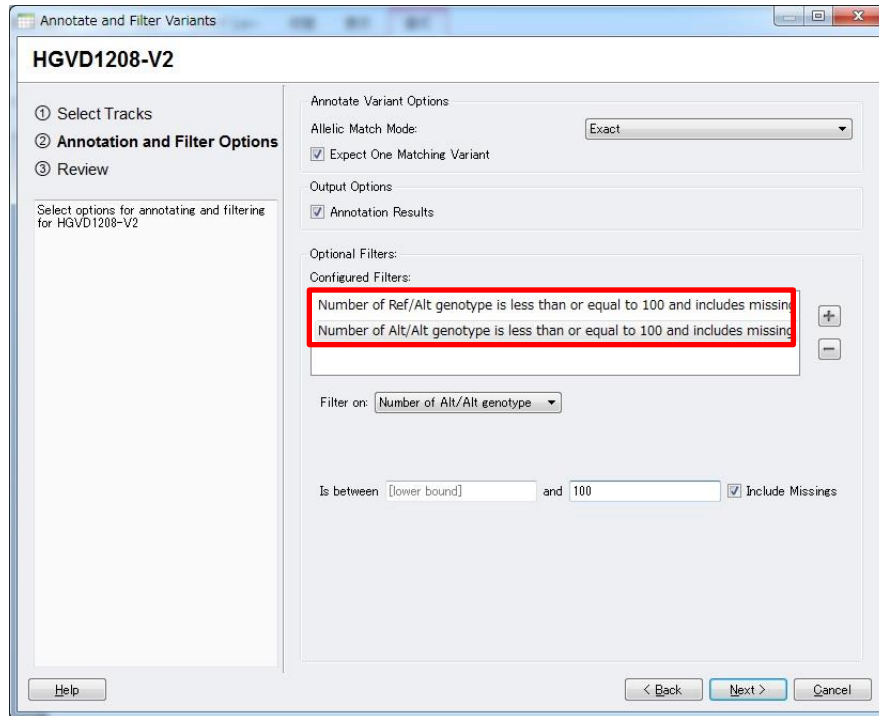
6個のうち4個以上の予測アルゴリズムで、生体に有害と
判定された変異のみを抽出

4. ClinVar

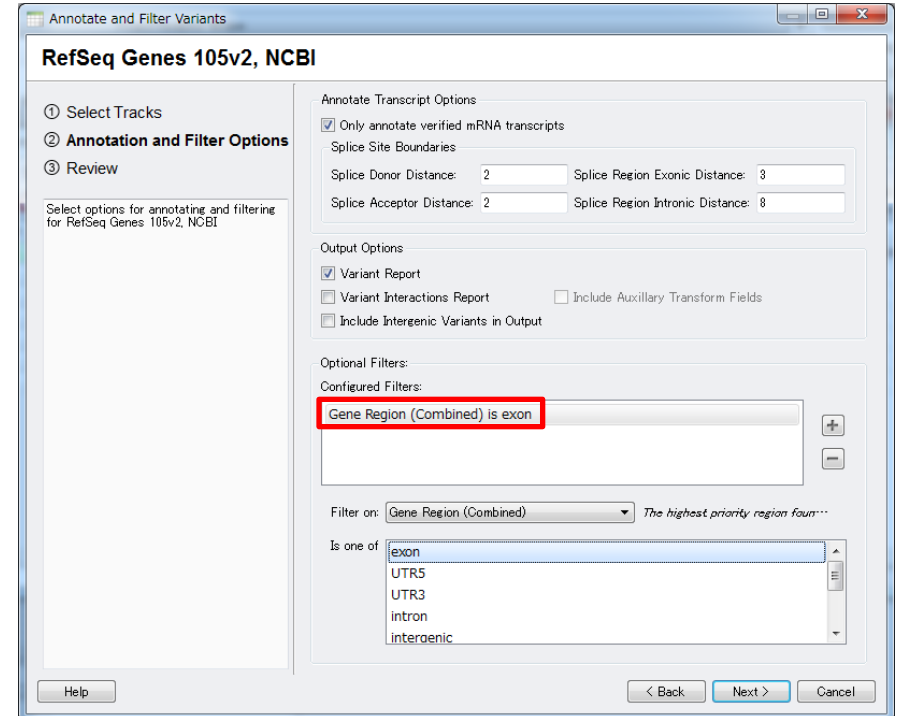
疾患関連情報を取得



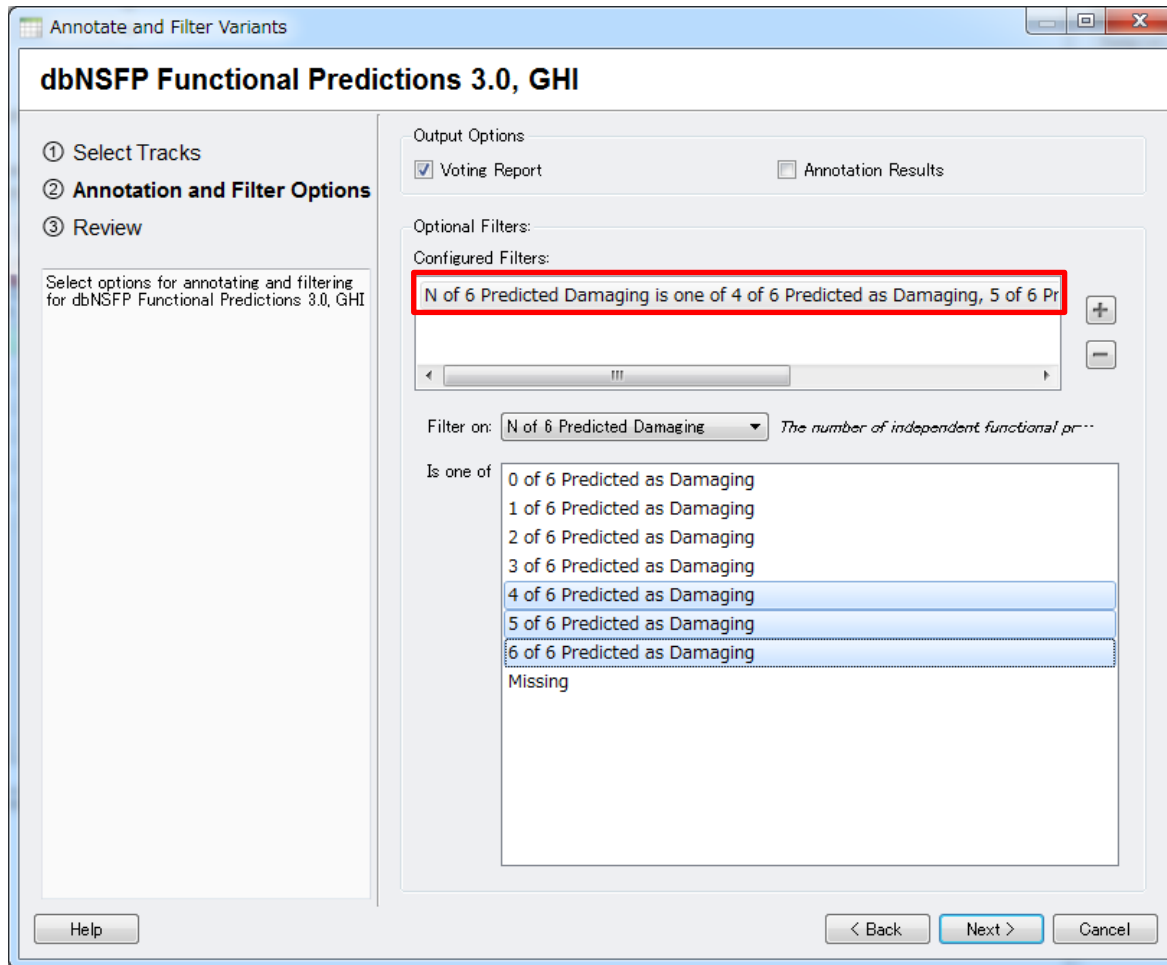
- あらかじめデータベース管理ツールでダウンロード／インポートしておいた各種データリソースを選択する。



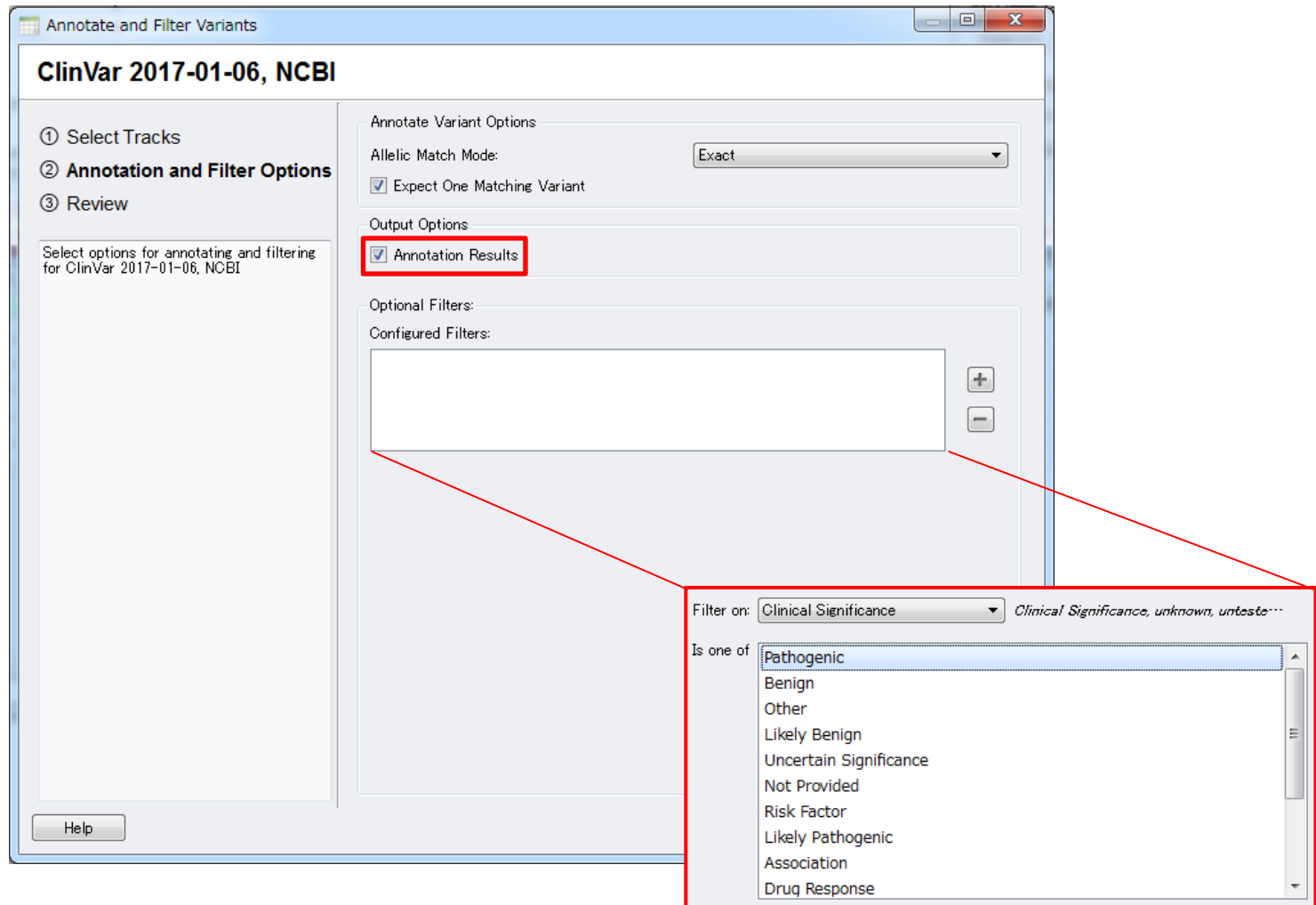
- HGVDデータセット内に、Alternate Allele (Ref/AltとAlt/Alt) として100サンプル以上登録されている変異を除外する。



- RefSeq遺伝子アノテーションにおいて、エクソン領域内に存在する変異のみを抽出する。



- dbNSFPにおいて、6個のうち4個以上の予測アルゴリズムで、生体に有害と判定された変異のみを抽出する。



- ClinVarのデータでアノテーション付けを行う。
- フィルタリングを行う場合は、「Clinical Significance」を「Pathogenic」などと指定することで、病原性をもつ変異のみを抽出できる。

▼ Sample1_Variants - Genotypes(G_T) - Sheet 1

- HGVD1208-V2 - Annotation Results
- RefSeq Genes 105v2, NCBI- Variant Report
- dbNSFP Functional Predictions 3.0, GHI- Voting Report
- ClinVar 2017-01-06, NCBI - Annotation Results
- Sample1_Variants - Genotypes(G_T) - Sheet 1 - Applied Filters
- Sample1_Variants - Genotypes(G_T) - Sheet 1 - Filtered Subset

← アノテーションデータ

← サンプルデータ

フィルタリング結果の変異データ数

Sample1_Variants - Genotypes(G_T) - Sheet 1 - Filtered Subset [22]

File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

All: 6 x 2,050
Active: 6 x 2,050

Unsort	B	C	G	G	G	
Map	Samples	Affection Status	Sample Source File Name	1:1268109-SNV	1:1387764-SNV	1:1469415-SNV
Chromosome				1	1	1
Position				1268109	1387764	1469415
Identifier				?	?	?
Reference				G	G	A
Alternates				C	A	G
1	Sample1_Variants	1	Sample1_Variants	?_?	A_G	?_?
2	Sample2_Variants	1	Sample2_Variants	?_?	A_G	?_?
3	Sample3_Variants	1	Sample3_Variants	C_G	A_G	A_G
4	Sample4_Variants	0	Sample4_Variants	?_?	?_?	?_?
5	Sample5_Variants	0	Sample5_Variants	?_?	?_?	?_?
6	Sample6_Variants	0	Sample6_Variants	?_?	?_?	?_?

Sample1_Variants - Genotypes(G_T) - Sheet 1 - Filtered Subset

- 計算が終了すると、各データソースごとのアノテーションデータのシート、およびフィルタリング結果をまとめたサンプルデータのシートが作成される。
- フィルタリングによって残った変異データは、「Filtered Subset」シートにまとめられている。

Map	Samples	Affection Status	Sample Source File Name	1:1268109-SNV	1:1387764-SNV	1:1469415-SNV
1	Sample1_Variants	1	Sample1_Variants	?_?	A_G	?_?
2	Sample2_Variants	1	Sample2_Variants	?_?	A_G	?_?
3	Sample3_Variants	1	Sample3_Variants	C_G	A_G	A_G
4	Sample4_Variants	0	Sample4_Variants	?_?	?_?	?_?
5	Sample5_Variants	0	Sample5_Variants	?_?	?_?	?_?
6	Sample6_Variants	0	Sample6_Variants	?_?	?_?	?_?

- 最初に、疾患／正常グループ情報のカラムを指定する。
- それぞれのグループにおける、アレルの検索条件を指定する。

Activate Variants based on Genotype Count Threshold

Dependent Column: **Affection Status**

Select the reference allele field from the marker map

C Reference

Affection Status=False

Activate columns that have

>= 3 occurrences

of the following genotypes

Ref_Ref Alt_Ref Alt_Alt ?_?

Affection Status=True

Activate columns that have

>= 3 occurrences

of the following genotypes

Ref_Ref Alt_Ref Alt_Alt ?_?

Sample1_Variants - Genotypes(G_T) - Filtered Subset - Active Subset [24]

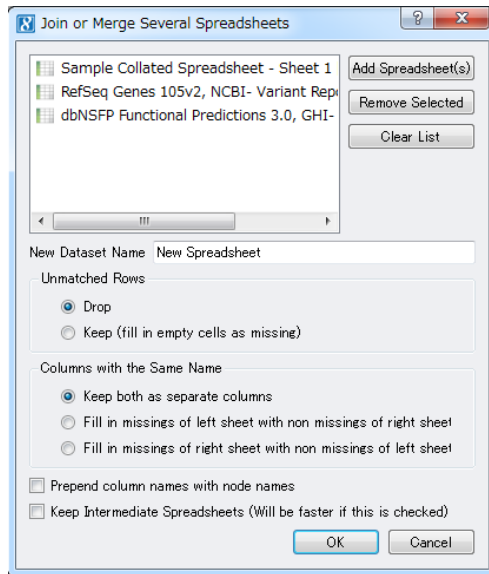
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

All: 6 x 47
Affection Status (Case/Control), 6 x 47

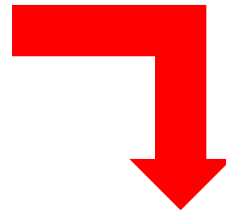
Unsort		B	2	C	3	G	4	G	5	G	6
Map	Samples	Affection Status		Sample Source File Name		1:1387764-SNV		1:12785749-SNV		1:24409191-SNV	
	Chromosome					1		1		1	
	Position					1387764		12785749		24409191	
	Identifier					?		?		?	
	Reference					G		C		C	
	Alternates					A		T		T	
1	Sample1_Variants		1	Sample1_Variants		A_G		C_T		C_T	
2	Sample2_Variants		1	Sample2_Variants		A_G		C_T		C_T	
3	Sample3_Variants		1	Sample3_Variants		A_G		C_T		C_T	
4	Sample4_Variants		0	Sample4_Variants		?_?		?_?		?_?	
5	Sample5_Variants		0	Sample5_Variants		?_?		?_?		?_?	
6	Sample6_Variants		0	Sample6_Variants		?_?		?_?		?_?	

Sample1_Variants - Genotypes(G_T) - Filtered Subset - Active Subset

- 計算が終了すると、指定の検索条件で抽出された変異データのシートが作成される。



- ジェノタイプデータのシートとアノテーションデータのシートを統合し、1シートでまとめてデータを確認することが可能。



Unsort	Map	Variant	G 3 Sample3_Variants_GT	G 4 Sample4_Variants_GT	G 5 Sample5_Variants_GT	G 6 Sample6_Variants_GT	C 7 Gene Names	C 8 Sequence Ontology (Combined)
1	1	1:1387764-SNV	A_G	??	??	??	ATAD3C	missense_variant
2	1	1:12785749-SNV	C_T	??	??	??	AADACL3	missense_variant
3	1	1:24409191-SNV	C_T	??	??	??	MYOM3	missense_variant
4	1	1:38329999-SNV	G_G	??	??	??	INPP5B	missense_variant
5	1	1:53712727-SNV	C_T	??	??	??	LRP8	missense_variant
6	1	1:172410967-SNV	A_G	??	??	??	C1orf105,PIGC	missense_variant
7	1	1:183184690-SNV	C_T	??	??	??	LAMC2	missense_variant
8	1	1:216424275-SNV	C_G	??	??	??	USH2A	missense_variant
9	1	1:246930564-SNV	C_G	??	??	??	SCCPDH	missense_variant
10	2	2:99279525-SNV	A_G	??	??	??	MGAT4A	missense_variant
11	2	2:152436012-SNV	G_T	??	??	??	NEB	missense_variant

フィルタリングの条件

- (#Sample -Ref/Alt <= 100 OR missing) AN 309,517
- #Sample -Ref/Alt <= 100 OR missing 276,512
- #Sample -Alt/Alt <= 100 OR missing 271,811
- Effect (Combined) is (LoF, Missense) 12,005
- N of 6 Predicted Damaging is (4 of 6 Prec 777
- 0 of 6 Predicted as Damaging 1,750
- 1 of 6 Predicted as Damaging 1,154
- 2 of 6 Predicted as Damaging 1,166
- 3 of 6 Predicted as Damaging 871
- 4 of 6 Predicted as Damaging 777
- 5 of 6 Predicted as Damaging 884
- 6 of 6 Predicted as Damaging 209
- Missing 5,194
- 1,870

付加されたアノテーション情報

Variant Info		RefSeq Genes 105v2, NCBI			dbNSFP Functional Prediction Voting			
Chr:Pos	Ref/Alt	Gene Names	SequenceOntology...	EffectC...	HGVS p. (Clinically Relevant)	N of 6 Predicted Damaging	SIFT Pred (C)	Polyphen2 HVAR Pred (C)
1:1268109	G/C	TAS1R3	missense_variant	Missense	NP_689414.1:p.Val400Leu	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:1469415	A/G	ATAD3A	missense_variant	Missense	NP_060658.3:p.Glu623Gly	4 of 6 Predicted as Damaging	Damaging	Benign
1:1477452	G/T	SSU72	missense_variant	Missense	NP_054907.1:p.Phe193Leu	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:1563538	C/T	MIB2	missense_variant	Missense	NP_543151.2:p.Ala722Val	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:2087443	T/A	PRKCZ	missense_variant	Missense	NP_002735.3:p.Trp296Arg	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:2160390	C/G	SKI	missense_variant	Missense	NP_003027.1:p.Ala62Gly	4 of 6 Predicted as Damaging	Damaging	Possibly damaging
1:3598991	C/G	TP73	missense_variant	Missense	NP_005418.1:p.Ser21Cys	5 of 6 Predicted as Damaging	Damaging	Possibly damaging
1:6314892	T/C	GPR153	missense_variant	Missense	NP_997253.2:p.Asn25Ser	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:6485305	T/A	ESPN	missense_variant	Missense	NP_113663.2:p.Val97Glu	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:6524501	T/C	TNFRSF25	missense_variant	Missense	NP_683866.1:p.Asp159Gly	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:7811281	C/G	CAMTA1	missense_variant	Missense	NP_056030.1:p.Thr1571Arg	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:9078431	A/C	SLC2A7	missense_variant	Missense	NP_997303.2:p.Ile147Ser	4 of 6 Predicted as Damaging	Damaging	Possibly damaging
1:9078434	C/T	SLC2A7	missense_variant	Missense	NP_997303.2:p.Gly146Asp	6 of 6 Predicted as Damaging	Damaging	Probably damaging
1:9804590	T/C	CLSTN1	missense_variant	Missense	NP_001009566.1:p.Asn366Ser	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:10163031	C/A	UBE4B	missense_variant	Missense	NP_001099032.1:p.Ser154Tyr	4 of 6 Predicted as Damaging	Damaging	Possibly damaging
1:10725278	T/C	CASZ1	missense_variant	Missense	NP_001073312.1:p.Met123Val	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:10725296	G/A	CASZ1	missense_variant	Missense	NP_001073312.1:p.Pro117Ser	4 of 6 Predicted as Damaging	Damaging	Probably damaging
1:11103415	C/G	MASP2	missense_variant	Missense	NP_006601.2:p.Cys241Ser	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:11151554	C/T	EXOSC10	missense_variant	Missense	NP_001001998.1:p.Arg158His	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:11254944	A/G	MTORANG...	missense_variant	Missense	NP_066969.1:p.Asn302Ser,?	6 of 6 Predicted as Damaging	Damaging	Probably damaging
1:11259624	C/A	MTOR	missense_variant	Missense	NP_004949.1:p.Ala1361Ser	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:11303326	A/C	MTOR	missense_variant	Missense	NP_004949.1:p.His419Gln	4 of 6 Predicted as Damaging	Damaging	Benign
1:11771929	G/A	DRAXIN	missense_variant	Missense	NP_940947.3:p.Gly222Arg	5 of 6 Predicted as Damaging	Damaging	Probably damaging
1:12025600	C/T	PLOD1	missense_variant	Missense	NP_000293.2:p.Arg512Cys	4 of 6 Predicted as Damaging	Damaging	Benign
1:12262025	C/G	TNFRSF1B	missense_variant	Missense	NP_001057.1:p.Pro301Arg	4 of 6 Predicted as Damaging	Damaging	Tolerated
1:12265764	C/A	...	missense_variant	Missense	NP_006601.2:p.Cys241Ser	5 of 6 Predicted as Damaging	Damaging	Probably damaging

- VarSeqでも同じデータソースを使用して、変異のアノテーション付けとフィルタリングが可能。
- SVSと違い、グラフィカルな操作でフィルタリング設定を行い、設定の変更を行った場合は、結果がリアルタイムに表示される。

お問い合わせ先：フィルジエン株式会社

TEL: 052-624-4388 (9:00～17:00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp