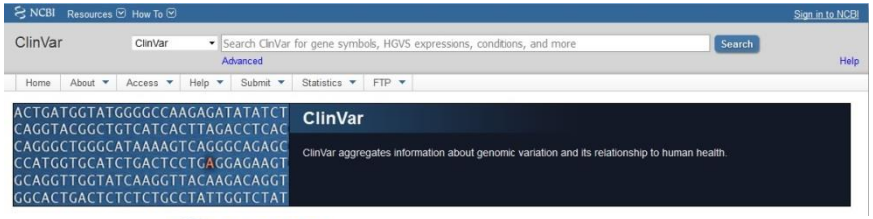
外部データリソースを利用した DNA-Seq疾患ゲノム解析

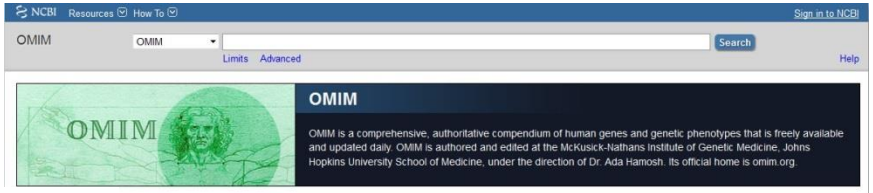
フィルジェン株式会社　バイオサイエンス部
(biosupport@filgen.jp)

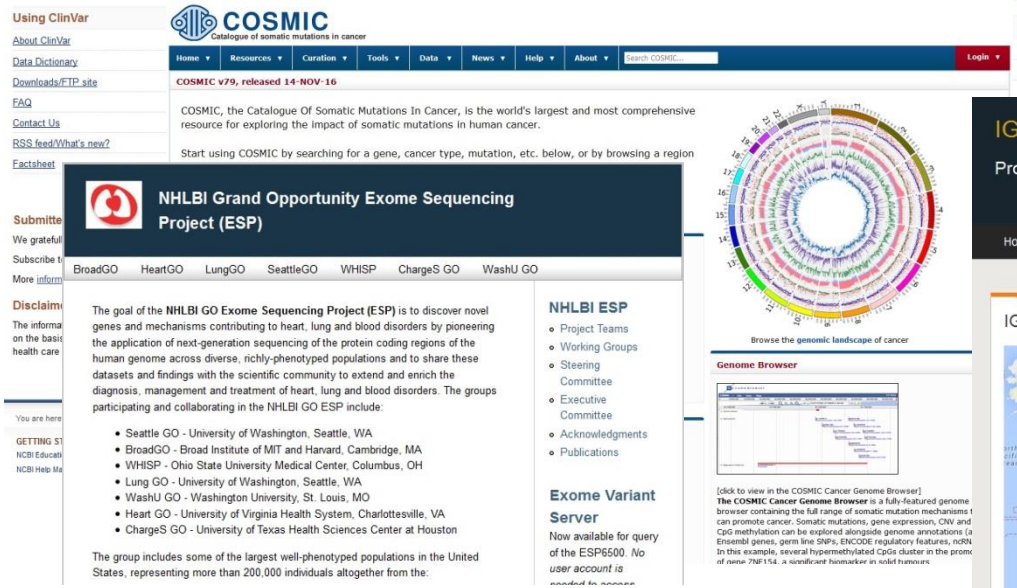
はじめに

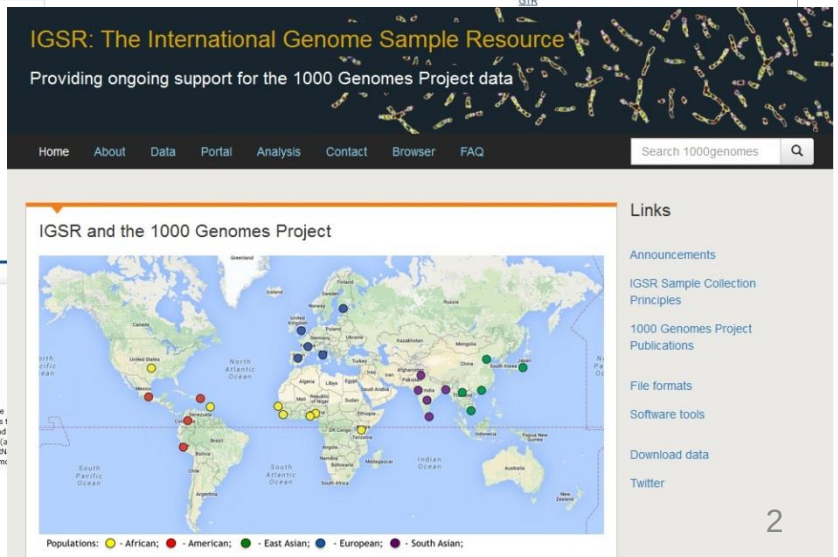
- 次世代シークエンサーを用いたゲノム変異解析（DNA-Seq）では、大量の変異データが産出されるため、その中から重要な変異を見つけ出す必要がある。

- 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。









DNA-Seq解析パイプライン

Filgen® biosciences & nanosciences

FASTQ File
1. TCAGACTGGAA
2. AGACTGGAAGC
3. AGTCAAATTGG
4. CAGACTGGAAG
5. CAGTCAAATTG
6. GTCAAATTGGA
7. AGACTGGAAGC
8. TCAAATTGGAA

Alignment

BAM File
TCAGACTGGAA
AGACTGGAAGC
AGTCAGATTGG
CAGACTGGAAG
CAGTCAAATTG
GTCAGATTGGA
AGACTGGAAGC
TCAAATTGGAA
REF: CAGTCAGATTGGAAGC

Variant Call

VCF File
Position 7, Genotype: G/A, AF=0.25 Position 9, Genotype: T/C, AF=0.5

Annotation
&
Filtering

各種データリソースを用いた3次解析

Golden Helix 3次解析ソフトウェア

Filgen® biosciences & nanosciences

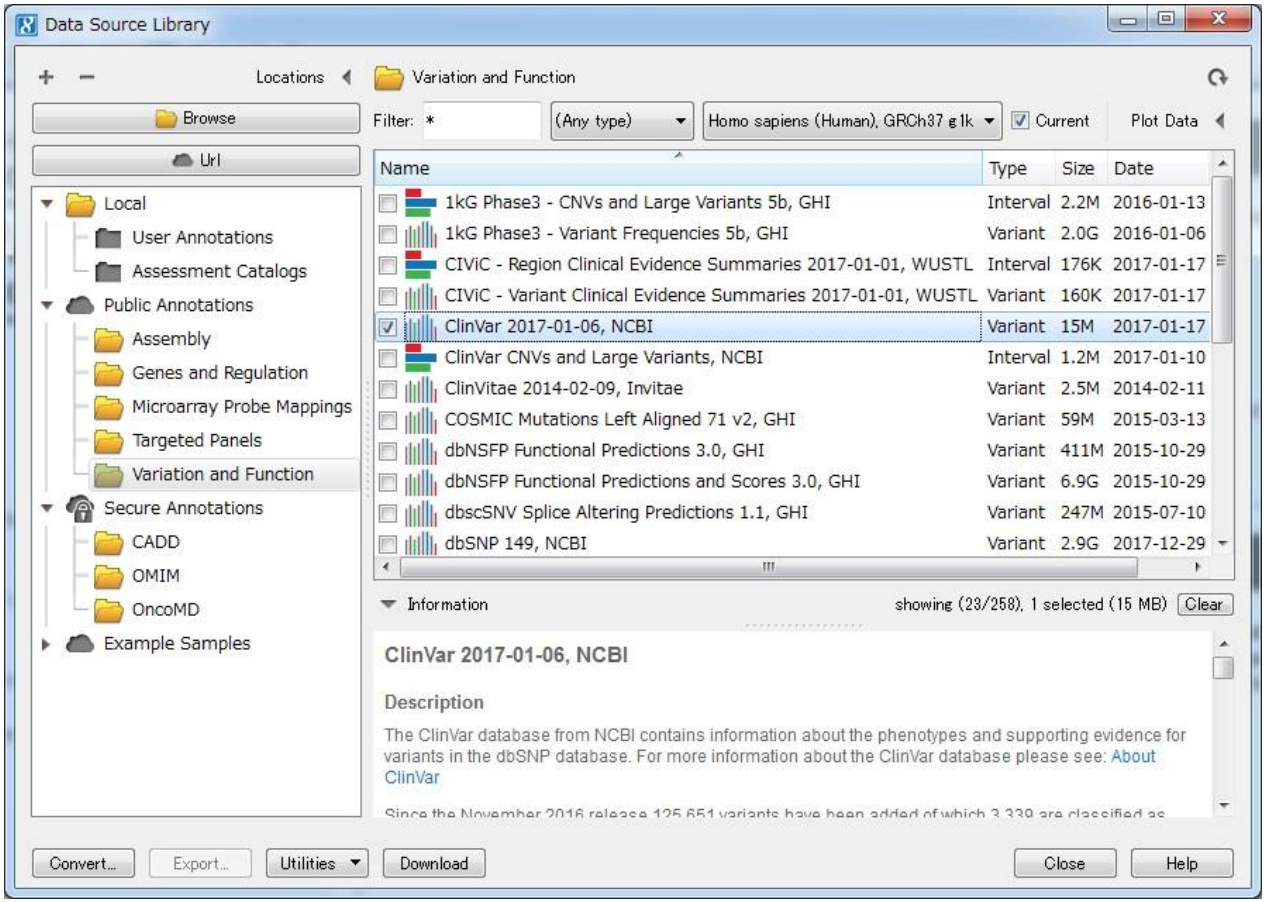
SNP & VARIATION SUITE

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

varSEQ®

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

データベース管理ツール



- 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。

データリソースの種類

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。

dbNSFP

- **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。

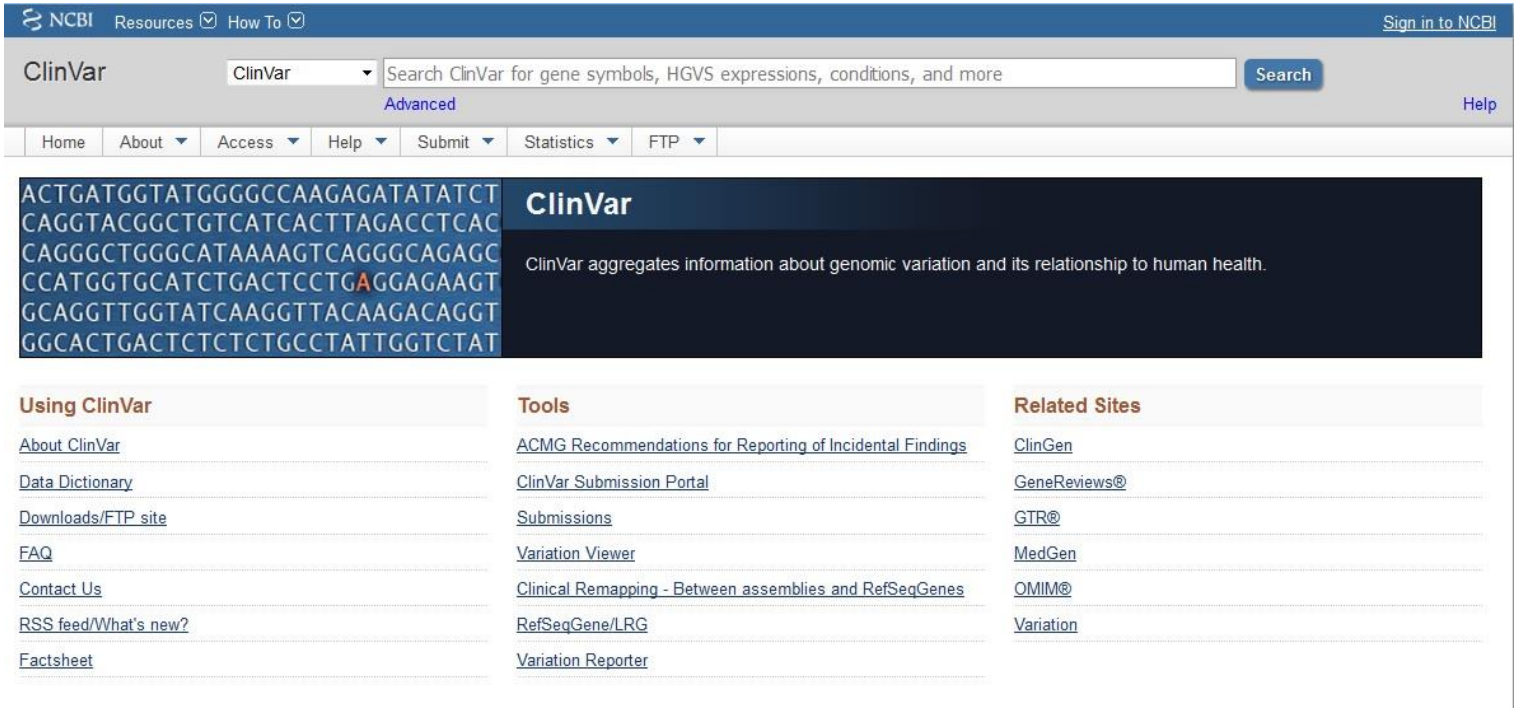
- 各アルゴリズムは、生物種間の配列保存性や、タンパク質立体構造への影響などから、各変異の生体への影響を評価している。

- 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と、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

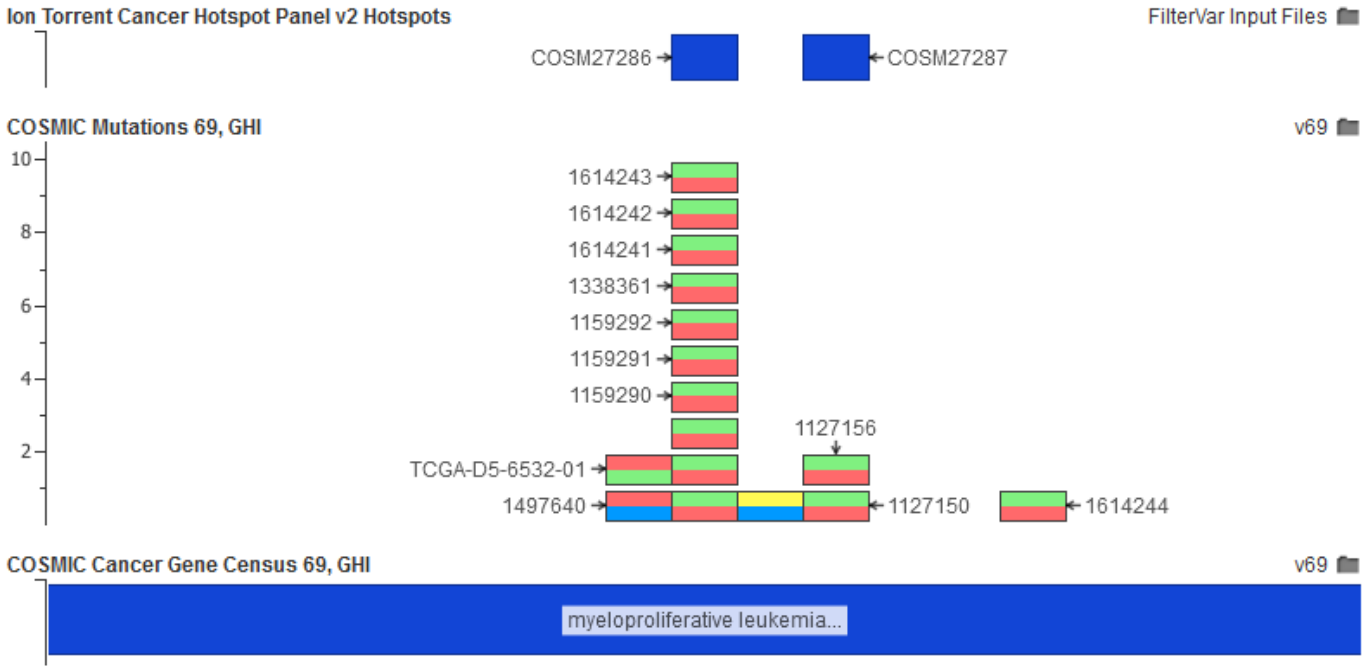
ClinVar

- 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。



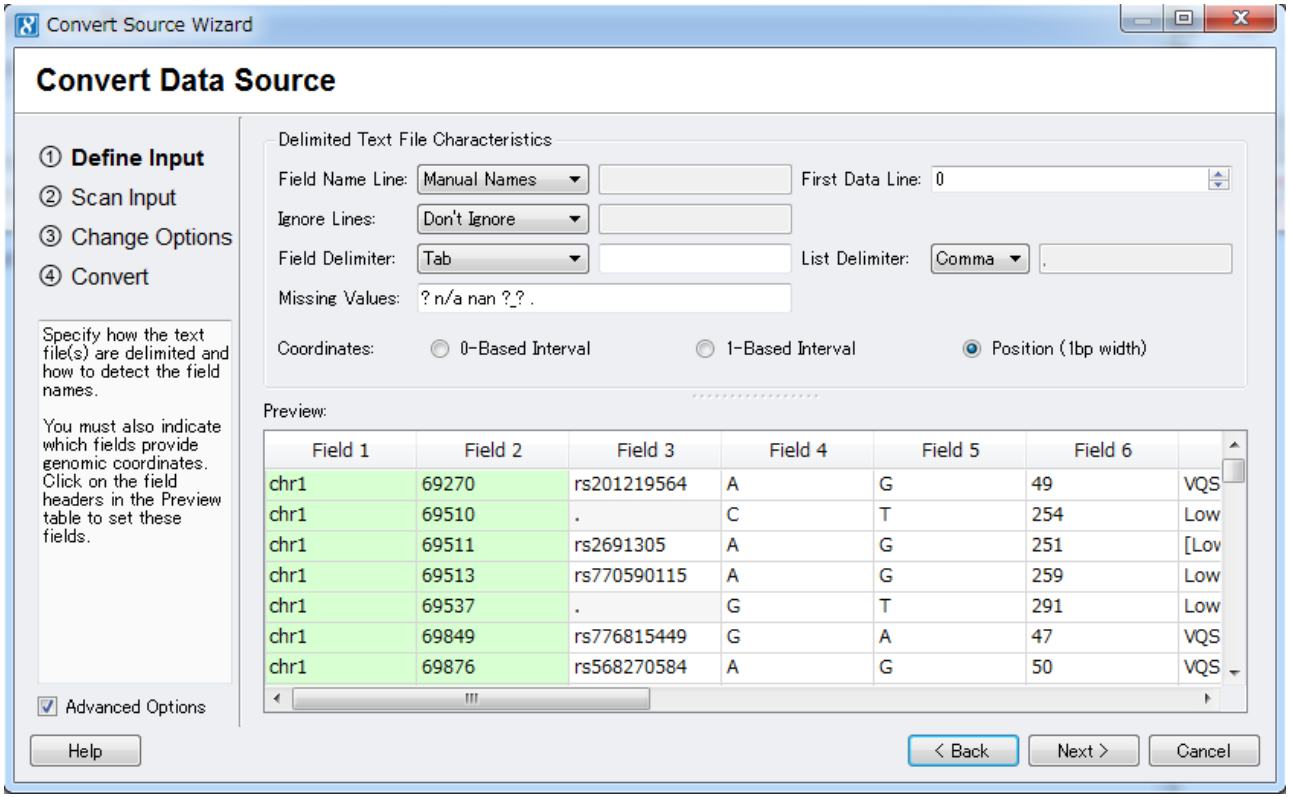
COSMIC

- 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。



HGVD

- 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。



有償データリソース

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。

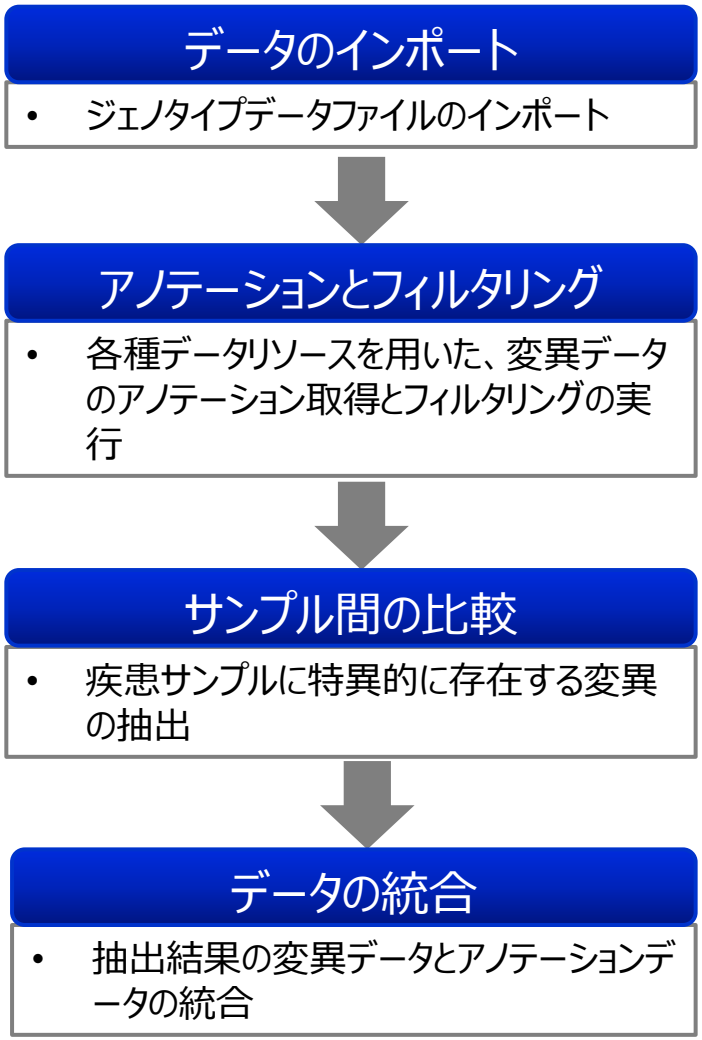
OncoMD

- 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。

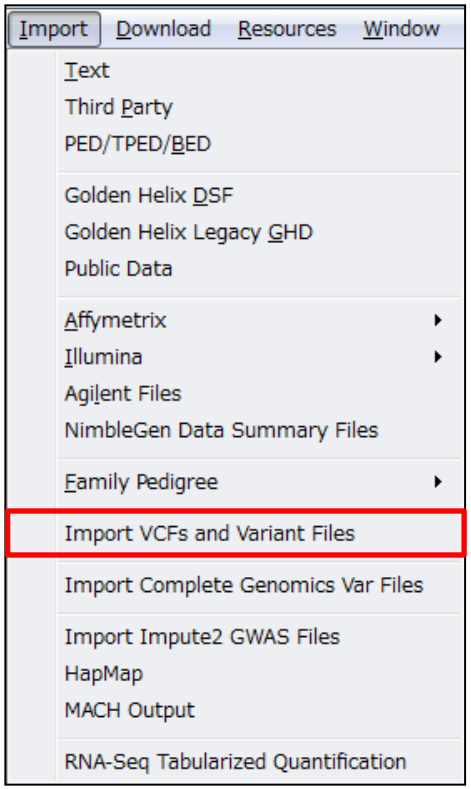
解析の流れ

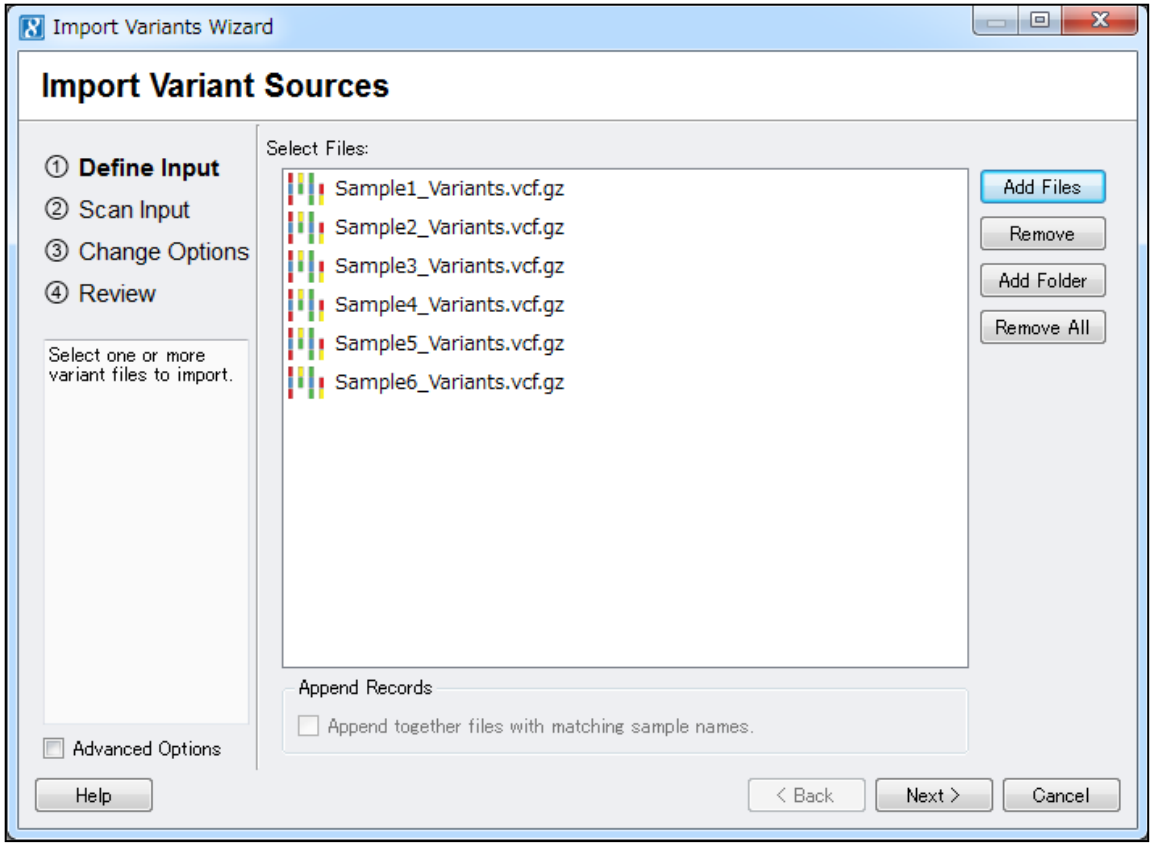
使用するジェノタイプデータ:

- 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



データのインポート

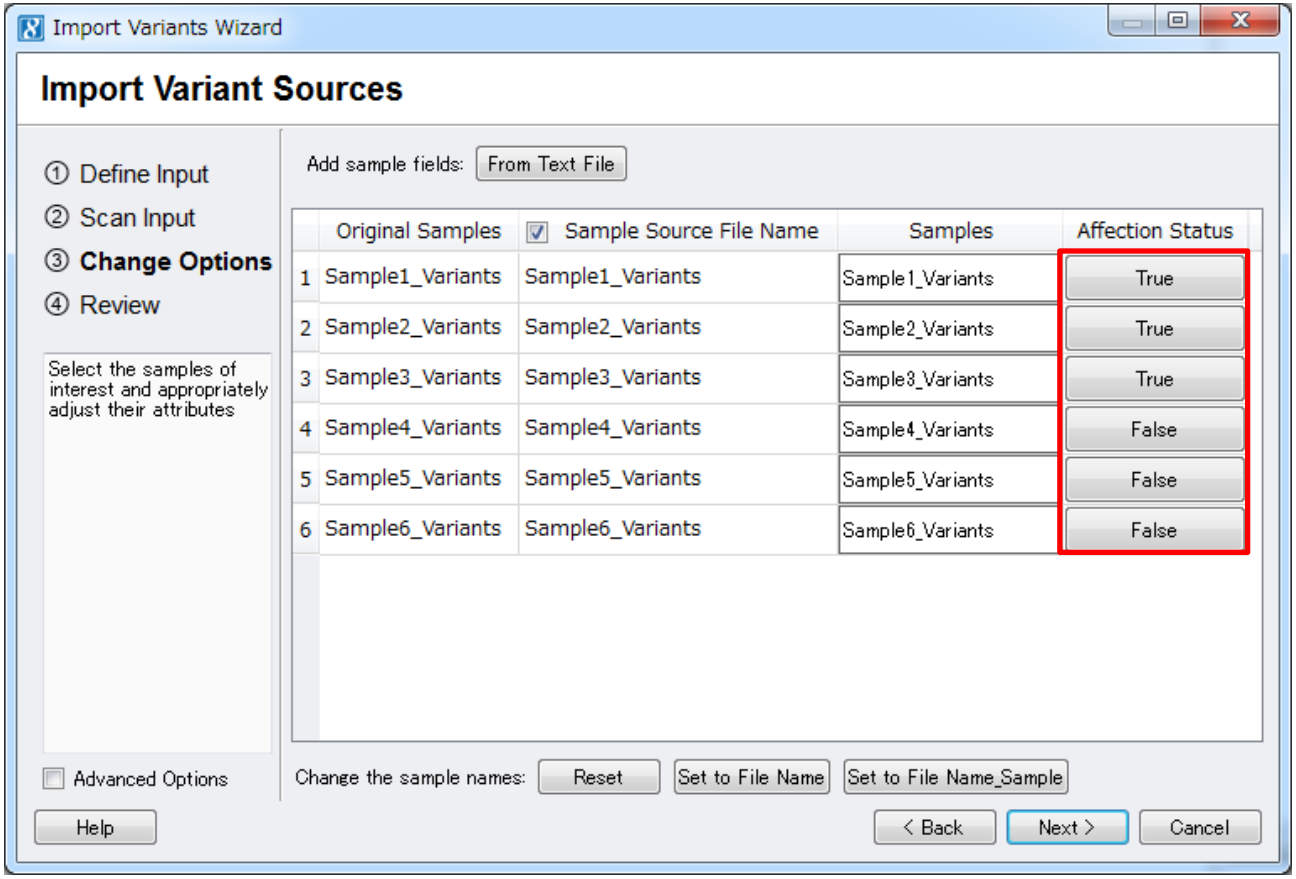




ジェノタイプデータファイル（VCFファイル）のインポート

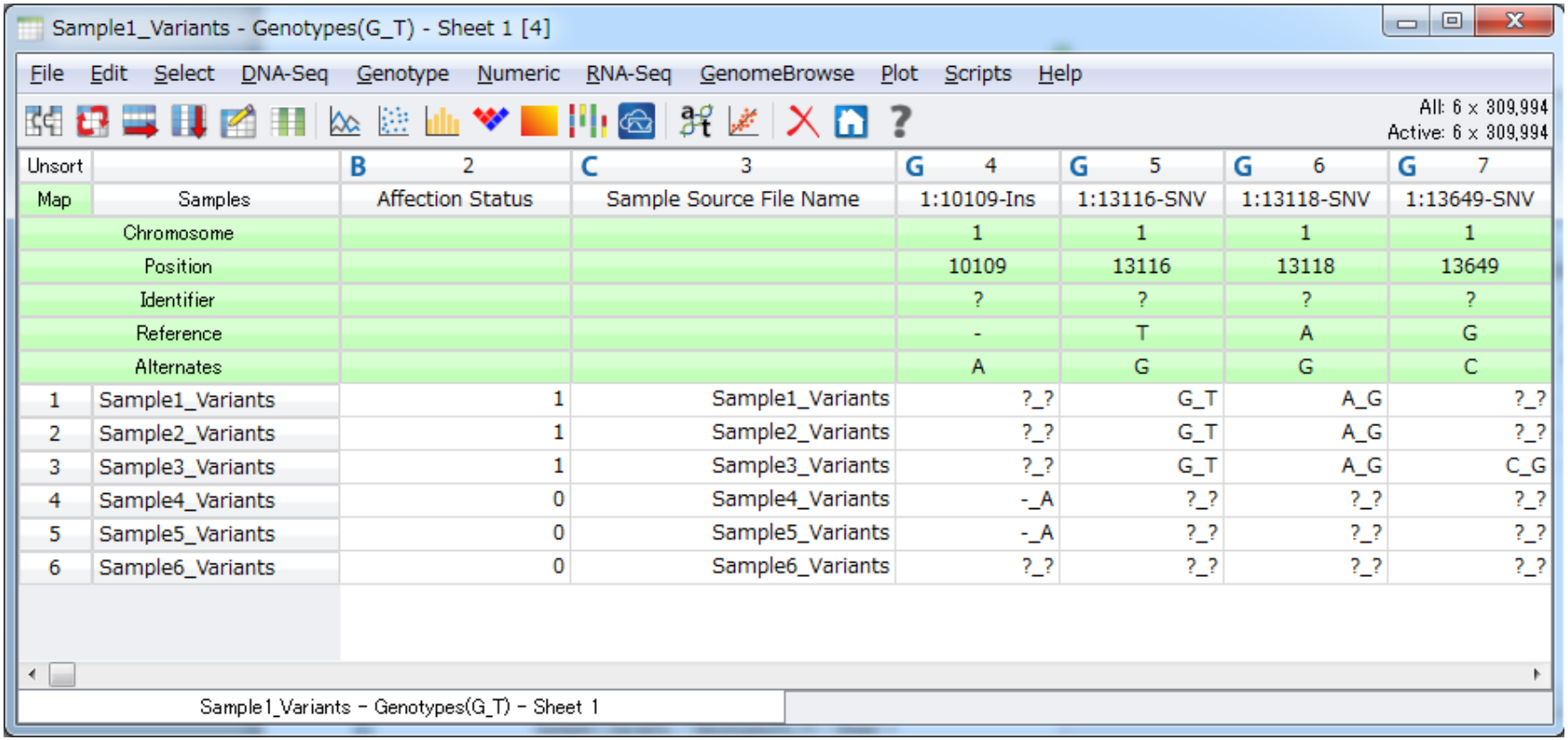
- ジェノタイプデータファイルは、バリアントコール用ツールなどで作成した、VCFファイルを使用する。

データのインポート



- 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。

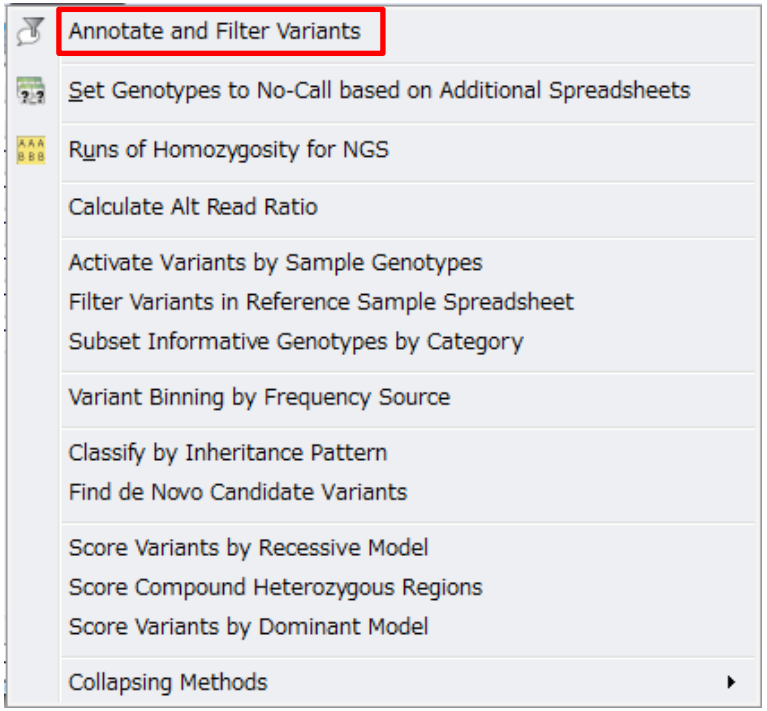
データのインポート



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ファイルに各アリ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。

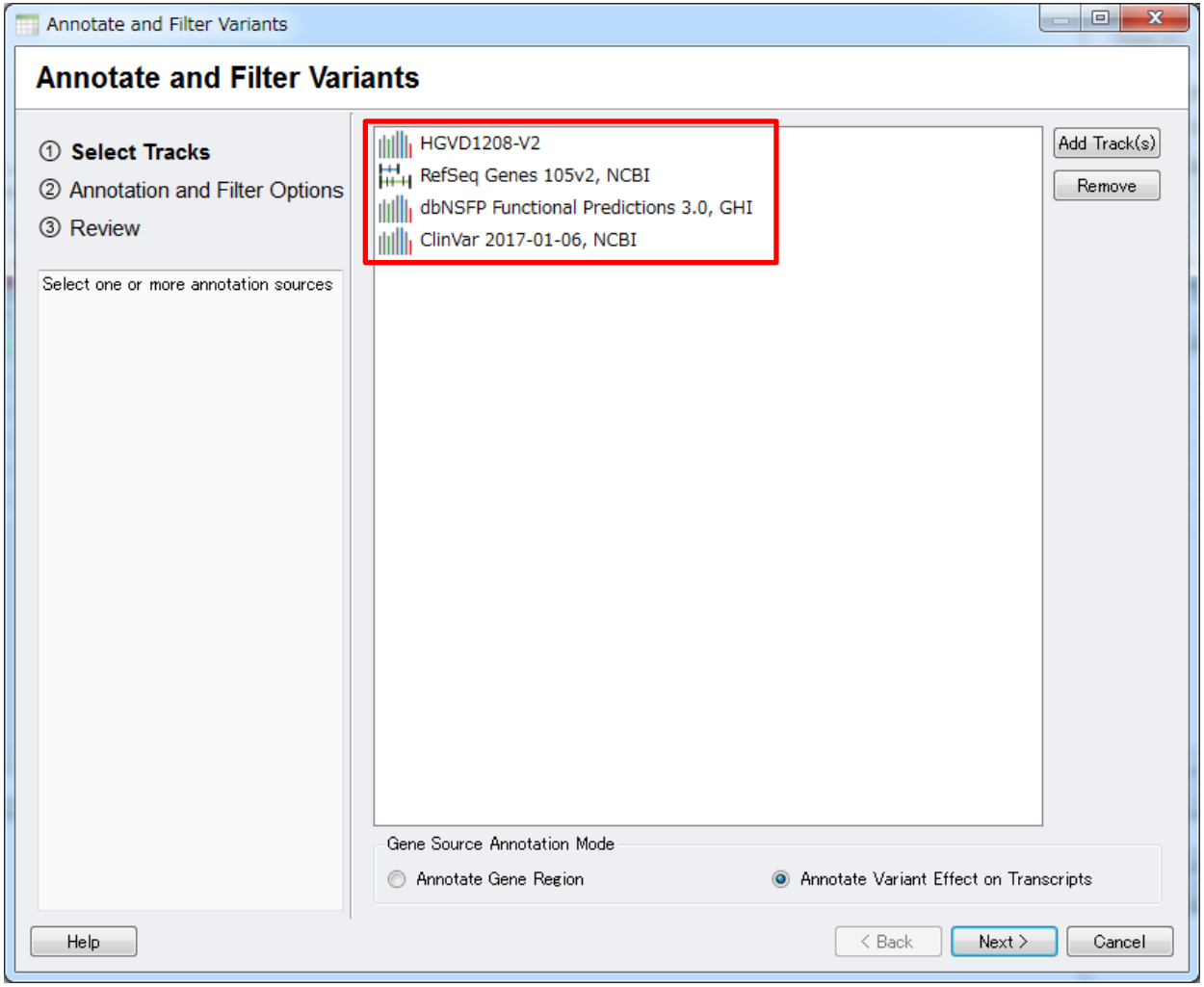
アノテーションとフィルタリング



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

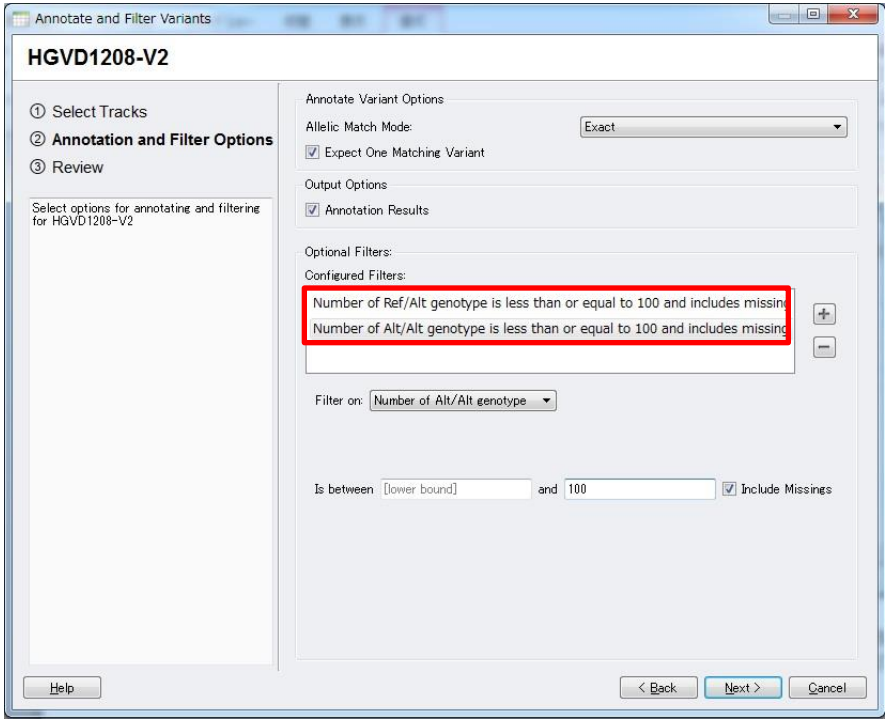
1. **HGVD**
   Alternate alleleを100サンプル以上もつ変異を除外
2. **RefSeq Genes**
   エクソン領域内の変異のみを抽出
3. **dbNSFP**
   6個のうち4個以上の予測アルゴリズムで、生体に有害と判定された変異のみを抽出
4. **ClinVar**
   疾患関連情報を取得

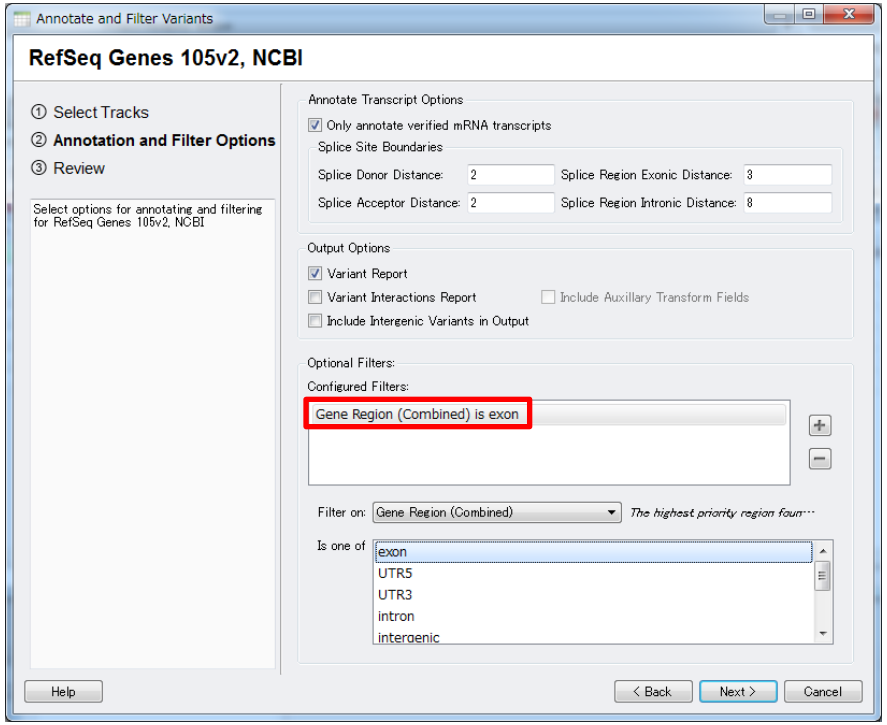
アノテーションとフィルタリング



- 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。

アノテーションとフィルタリング

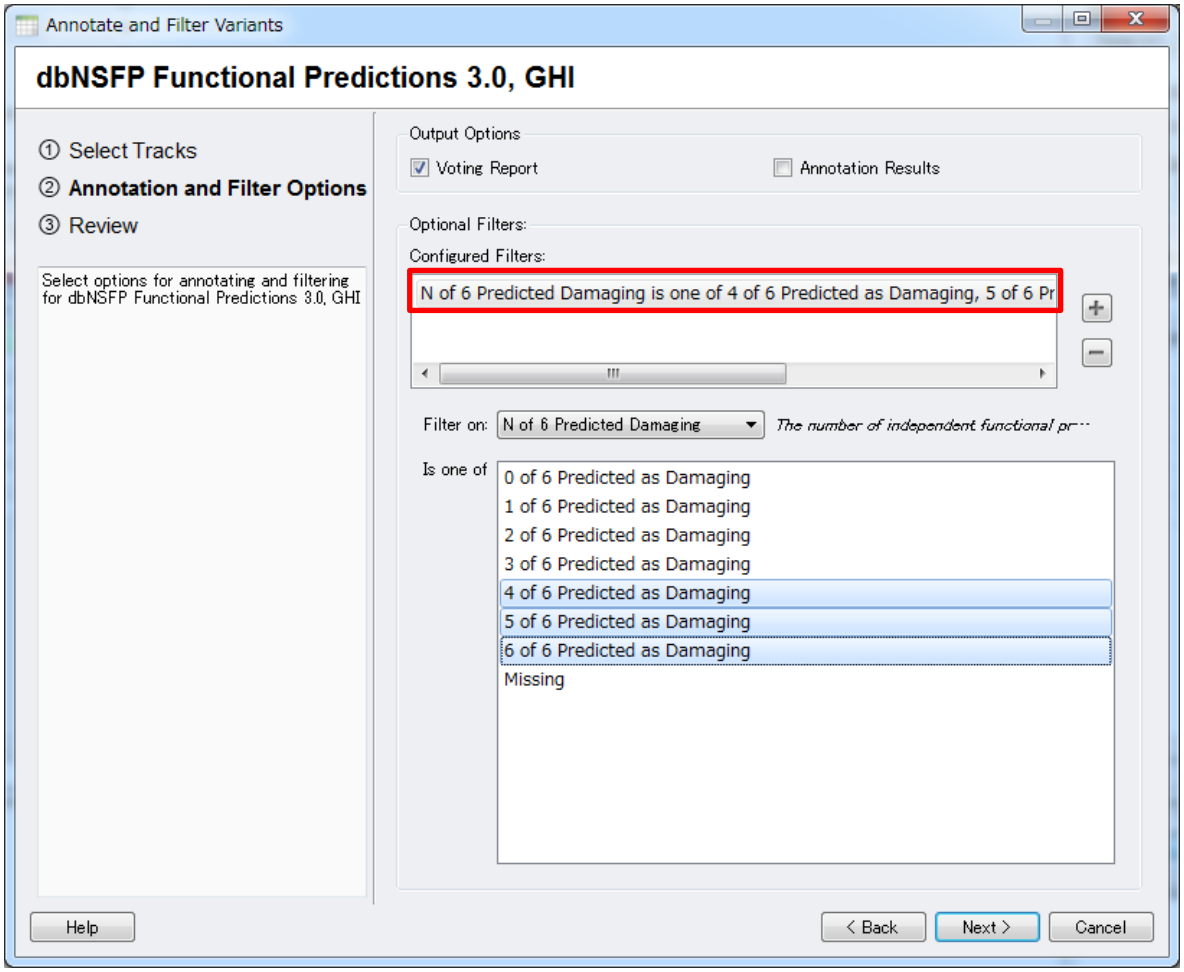




- HGVDデータセット内に、Alternate Allele (Ref/AltとAlt/Alt) として100サンプル以上登録されている変異を除外する。

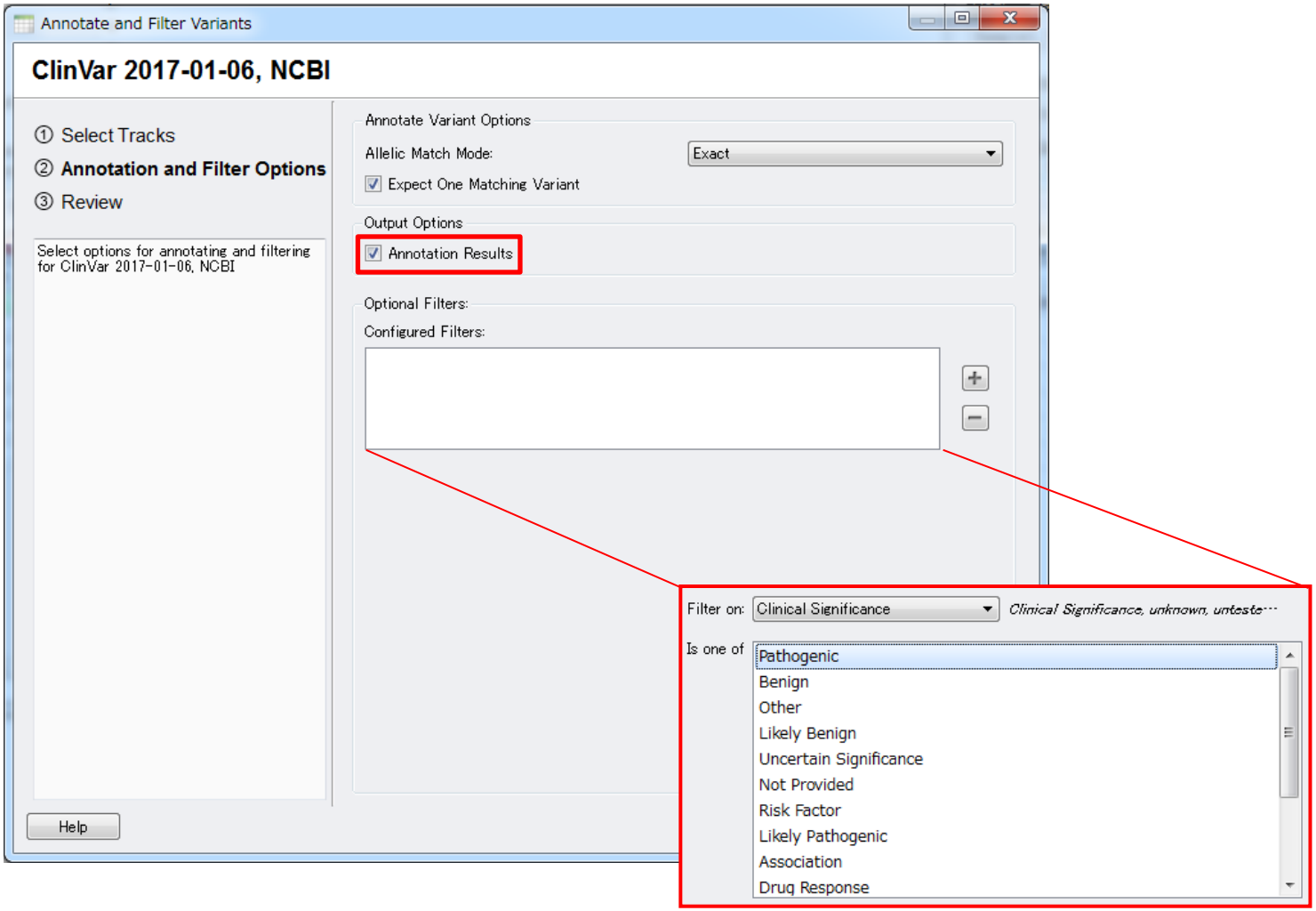
- RefSeq遺伝子アノテーションにおいて、エクソン領域内に存在する変異のみを抽出する。

アノテーションとフィルタリング



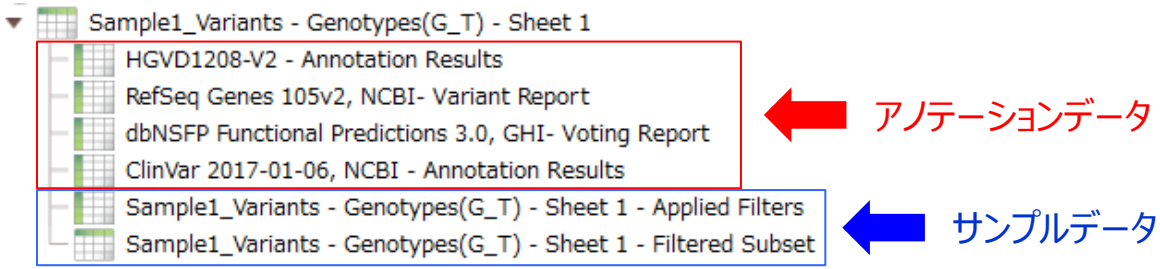
- dbNSFPにおいて、6個のうち4個以上の予測アルゴリズムで、生体に有害と判定された変異のみを抽出する。

アノテーションとフィルタリング



- 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。

アノテーションとフィルタリング



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

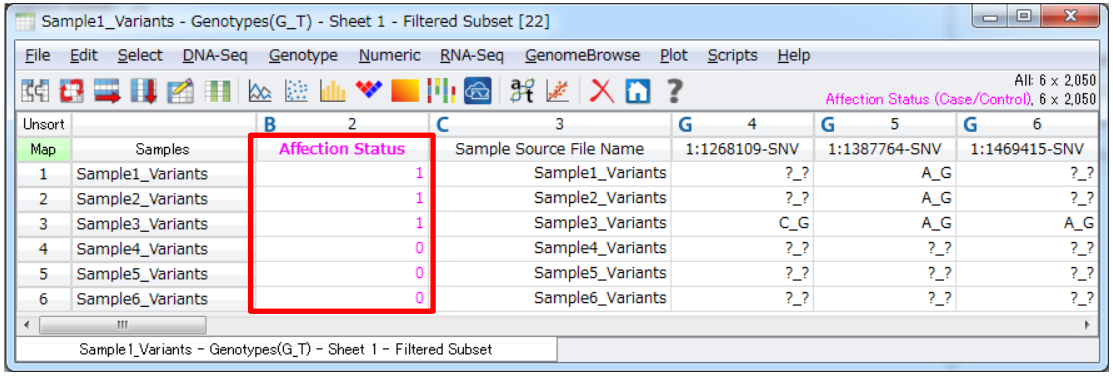
All: 6 x 2,050
Active: 6 x 2,050

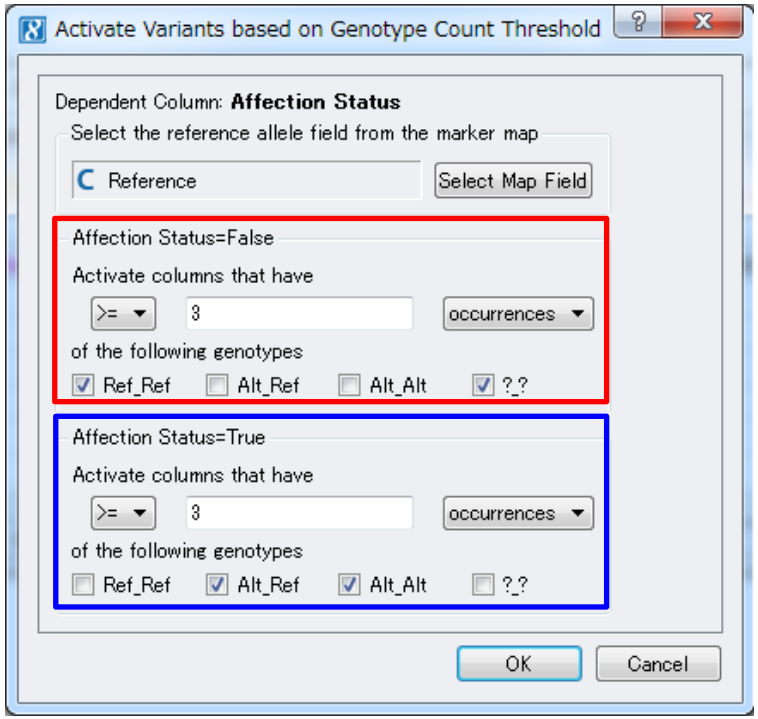
Unsort		B 2	C 3	G 4	G 5	G 6
Map	Samples	Affection Status	Sample Source File Name	1:1268109-SNV	1:1387764-SNV	1:1469415-SNV
	Chromosome			1	1	1
	Position			1268109	1387764	1469415
	Identifier			?	?	?
	Reference			G	G	A
	Alternates			C	A	G
1	Sample1_Variants	1	Sample1_Variants	?_?	A_G	?_?
2	Sample2_Variants	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、各データリ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。

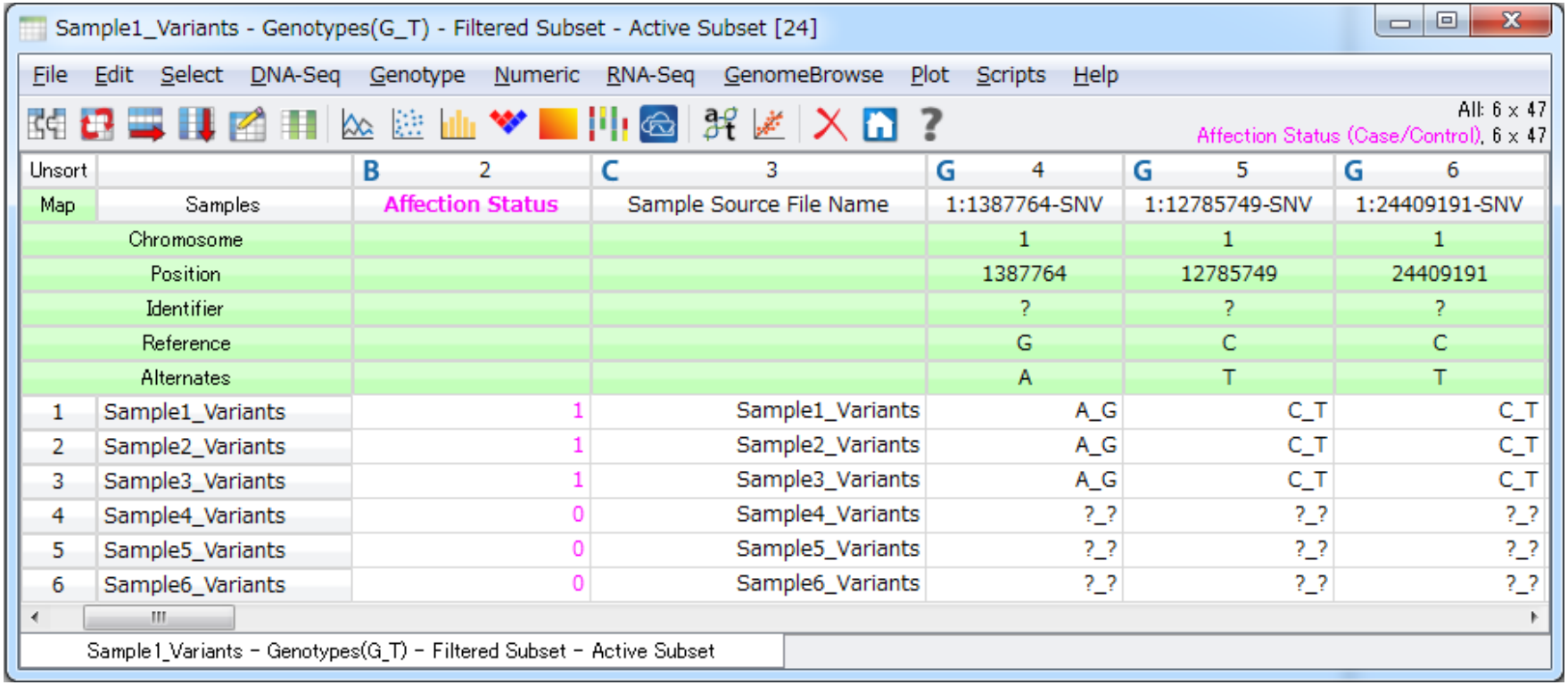
サンプル間の比較





- 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、アリルの検索条件を指定する。

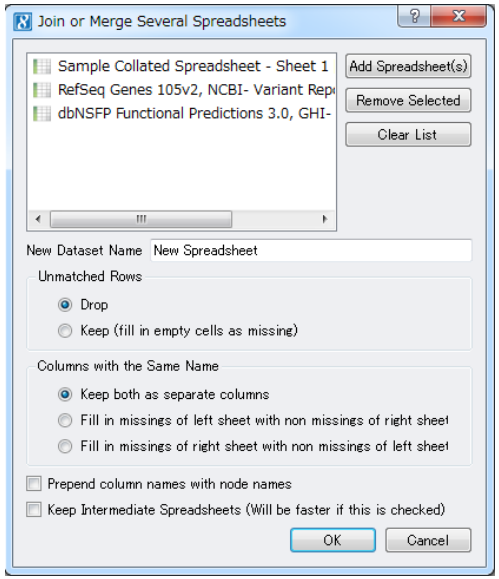
サンプル間の比較



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	?_?	?_?	?_?

- 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。

データの統合



- 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。

New Spreadsheet - Sheet 1 [39]

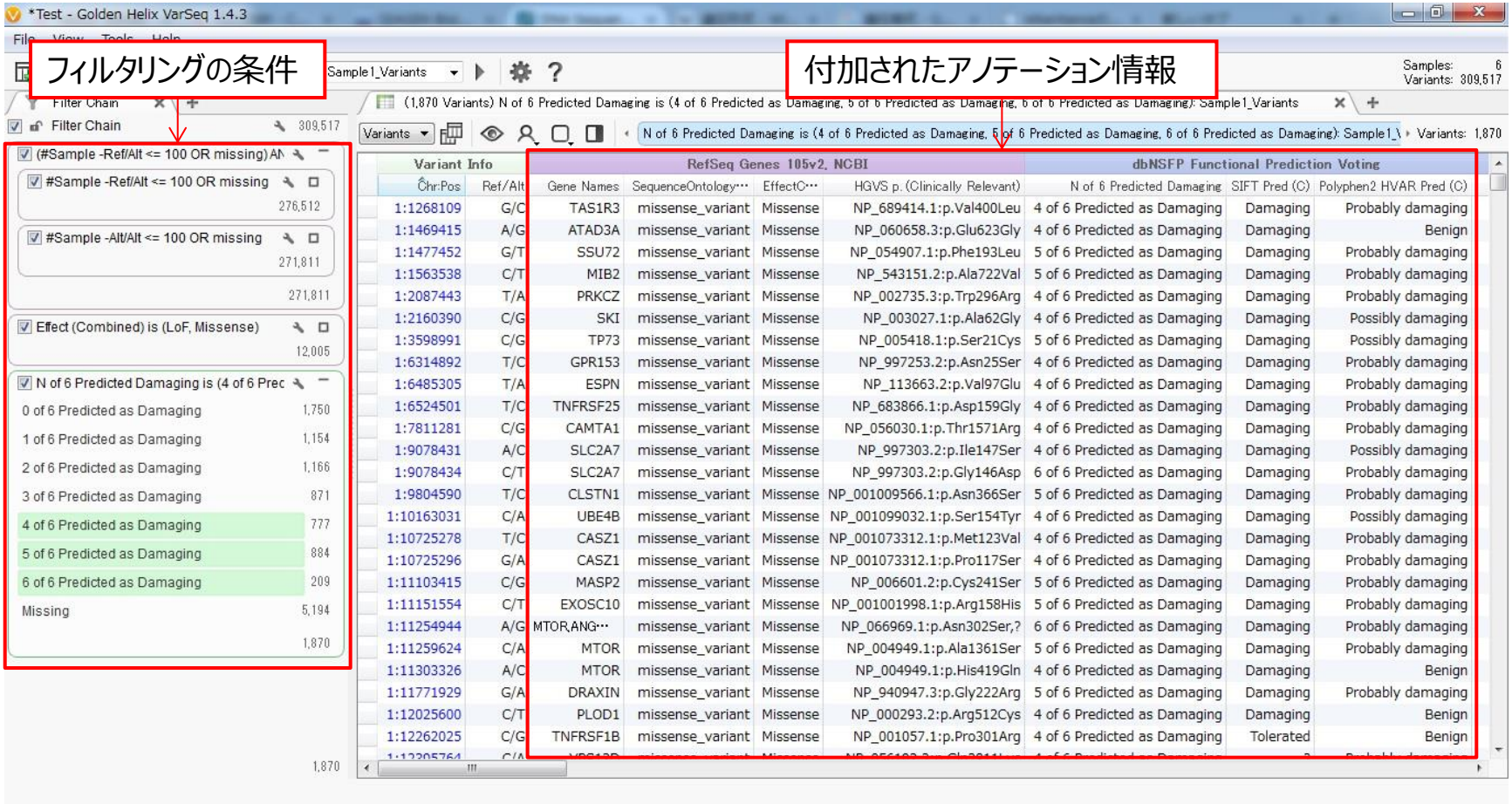
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

All: 44 x 27
Active: 44 x 27

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

New Spreadsheet - Sheet 1

VarSeqでは…



- VarSeqでも同じデータリソースを使用して、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：フィルジェン株式会社

TEL: 052-624-4388 (9:00～17：00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp