多因子疾患における
レアバリアント関連解析

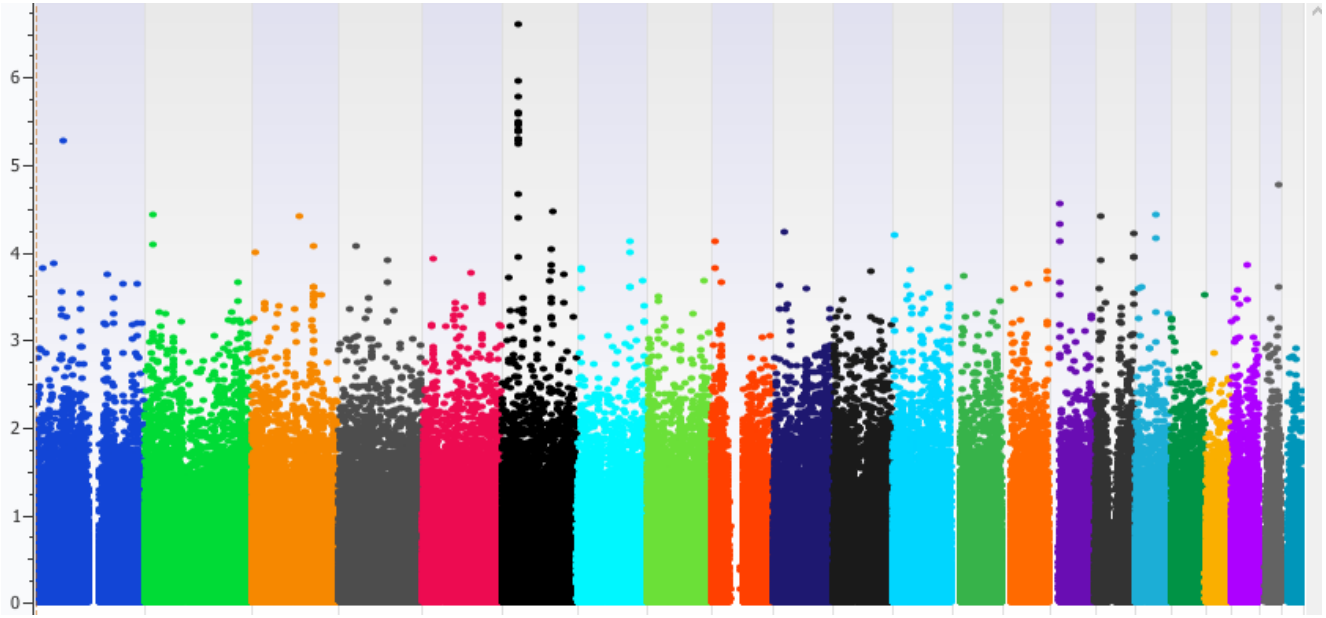
フィルジェン株式会社　バイオインフォマティクス部
(biosupport@filgen.jp)

はじめに

- 多因子疾患における疾患（表現型）と遺伝型の相関の研究においては、ゲノム上の一塩基多型であるSNPをマーカーとして用いる、マイクロアレイベースのゲノムワイド関連解析（GWAS）が広く利用されている

- 対して、DNAシークエンスデータを解析に用いる場合は、データ内に含まれる希少変異（レアバリアント）を扱うため、特別な解析手法が必要となる

- Golden Helix社SNP & Variation Suite (SVS)では、一般的なGWASに加え、レアバリアント解析用の解析ツールを搭載しており、SNPとレアバリアント両方の解析に対応できる



Golden Helix ソフトウェア製品ラインナップ

Filgen® biosciences & nanosciences

SNP & VARIATION SUITE

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

varSEQ®

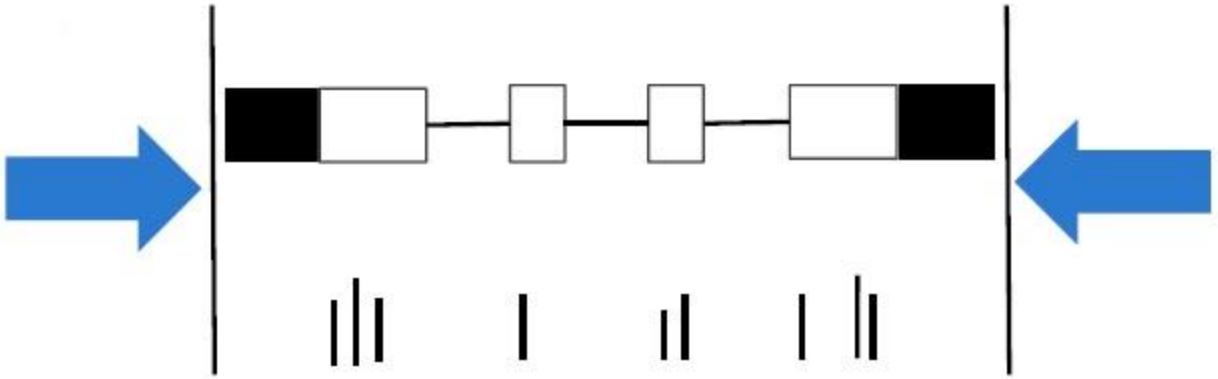
- Annotation and Filtering
- ACMG & AMP Guidelines
- CNV Calling
- Clinical Reporting
- Pipeline: Run Workflows

- Golden Helix社では、遺伝統計解析ソフトウェア「SNP & Variation Suite」と、クリニカルシークエンス解析ソフトウェア「VarSeq®」の2種類のソフトウェアパッケージを販売
- 医学・生物学研究や、家畜や作物の品種改良などの農学研究、さらに疾患の診断や最適な治療オプションの決定における医療分野、遺伝学的解析などの教育現場などで利用される

Collapsing Methods

■ レアバリアントに対する関連解析は、通常のGWASのアプローチでは、検出力が足りずに困難

バリアントを遺伝子ごとに集約し、単一ユニットとして解析

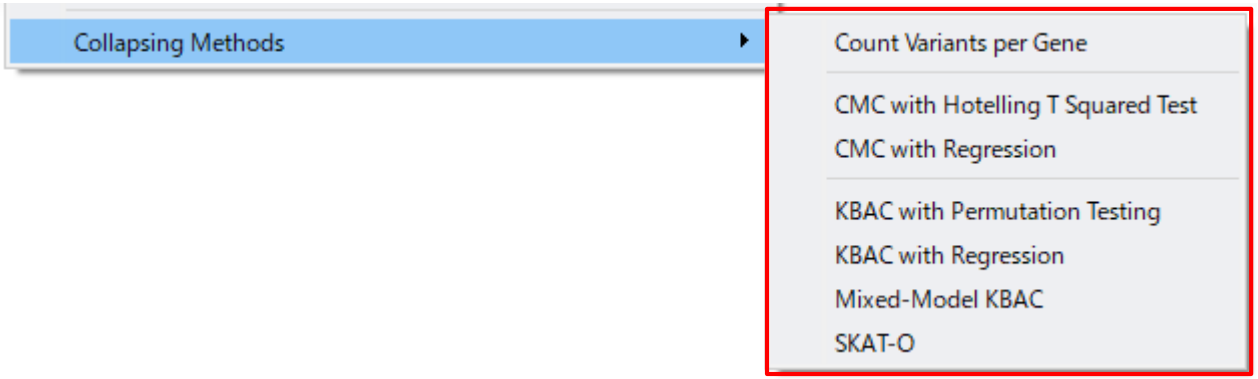


Collapsing Methods

解析手法の種類

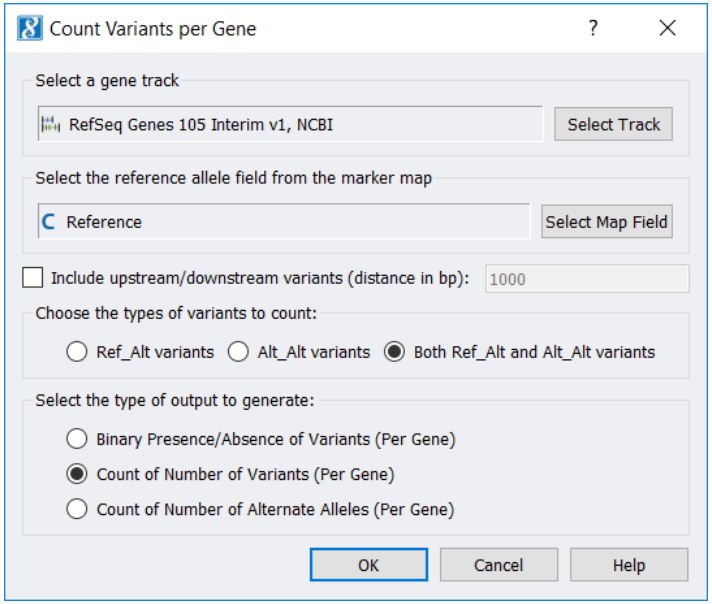
Burden Tests
- 遺伝子ごとにバリアントの効果を合計して検定を行う
- それぞれのバリアントの効果が、すべて同方向に作用することが前提となる
  - バリアントの重み付けなし：**CAST**, **CMC**
  - バリアントの重み付けあり：**KBAC**

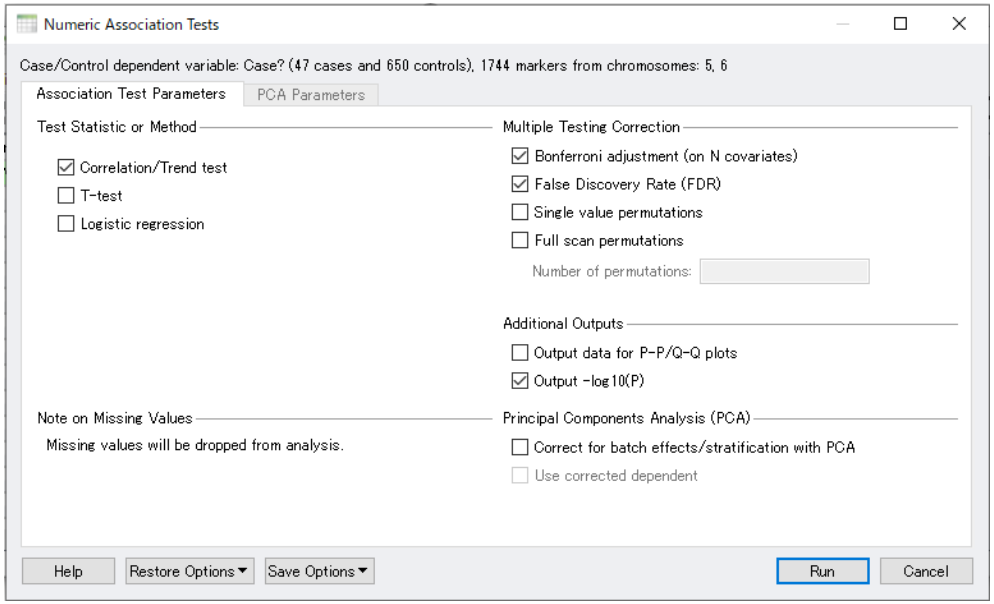
Kernel Tests
- 遺伝子ごとに、各バリアントの統計スコアの2乗を合計して検定を行う
- 各バリアントの効果が、様々な方向に作用する場合に用いられる
  - **SKAT**
  - **SKAT-O** (Burden Testsとの組み合わせ)



CAST (Cohort Allelic Sums Test)

- ✓ 各遺伝子ごとにバリアント数をカウントしたテーブルを作成し、関連解析を実行
- ✓ カウントテーブル作成前に、バリアントの機能予測プログラム（SIFT, Polyphen-2など）を用いて、非同義変異や機能喪失型変異のみを使用することが多い



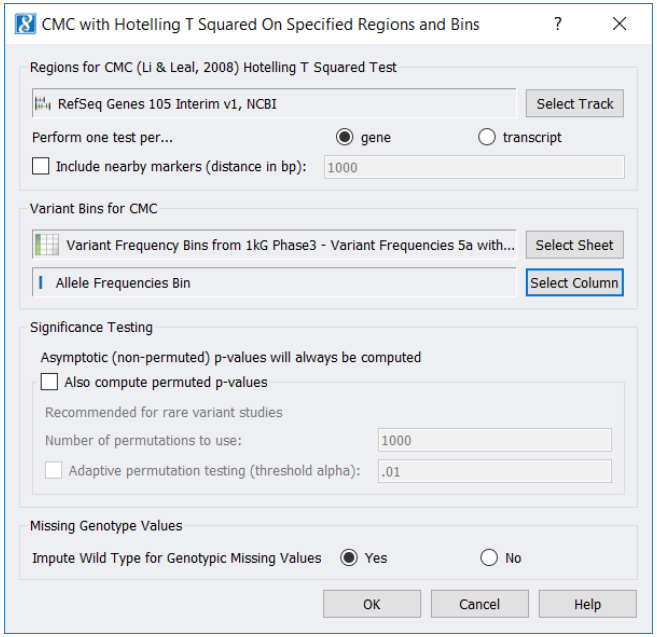


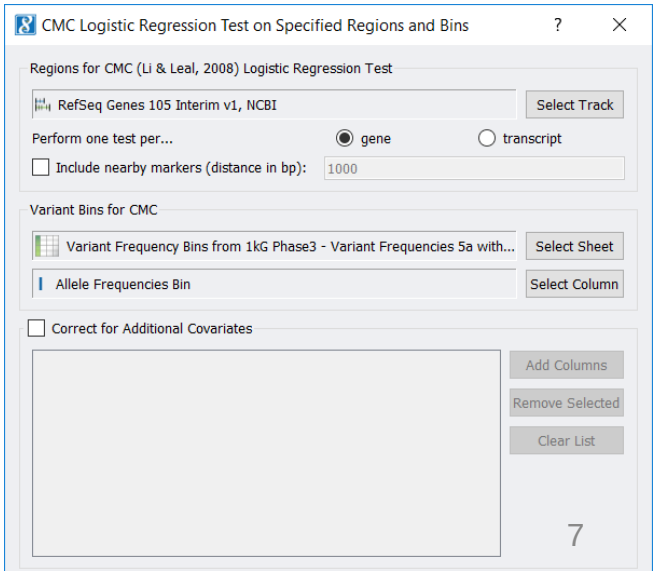
CMC

CMC (Combined Multivariate and Collapsing)

- ✓ 任意のアリル頻度データベース（1000 Genomes projectデータなど）を用いて、バリアントをアリル頻度ごとにグループ分けし、関連解析を実行

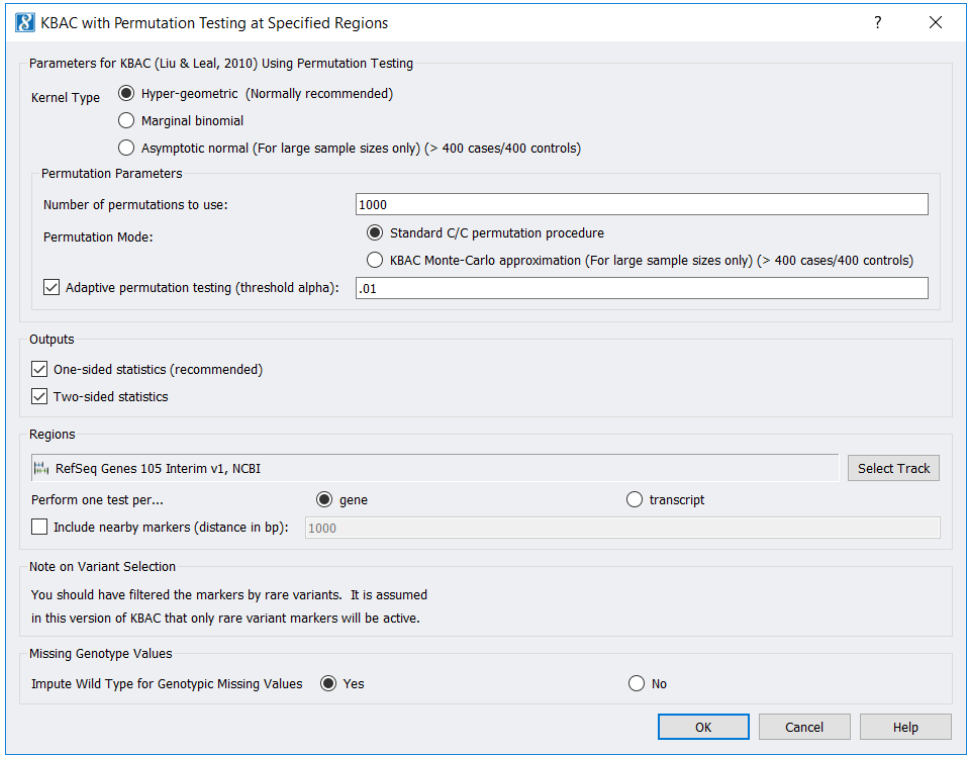
- ✓ 2種類の統計手法を利用可能
  - Hotelling T^2 Test（ケース・コントロール形質）
  - Regression Test（量的形質、共変量による調整）





KBAC (Kernel-Based Adaptive Cluster)

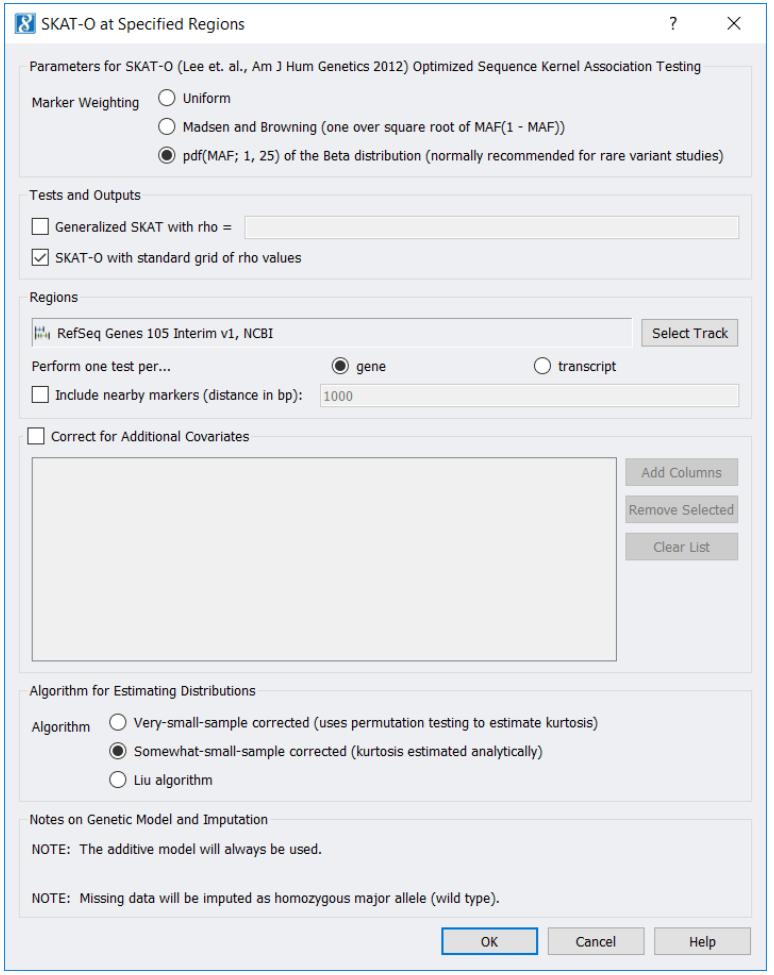
- ✓ 各バリアントに対して、サンプルをケース・コントロールに分けた際にどちらのグループに高頻度に存在するかに応じて、バリアントを重み付けし、関連解析を実行
- ✓ レアバリアントのみを解析に用いるように、解析前にバリアントのフィルタリングを行う必要がある
- ✓ 3種類の統計手法を利用可能
  - Permutation Test
  - Regression Test（共変量による調整）
  - Mixed-Model Test（血縁関係の補正）
- ✓ 片側検定の結果を採用することが推奨



SKAT, SKAT-O

SKAT (Sequence Kernel Association Test)

- ✓ 各バリアントの集団内頻度に応じて、バリアントを重み付けすることが可能
- ✓ 各バリアントの効果が、すべて同じ方向に作用するデータの場合は、Burden Testに比べて検出力が弱い
- ✓ Generalized SKATでは、rhoパラメータを0～1にバランス調整することで、Burden Testと組み合わせての計算が可能



SKAT-O (Optimized SKAT)

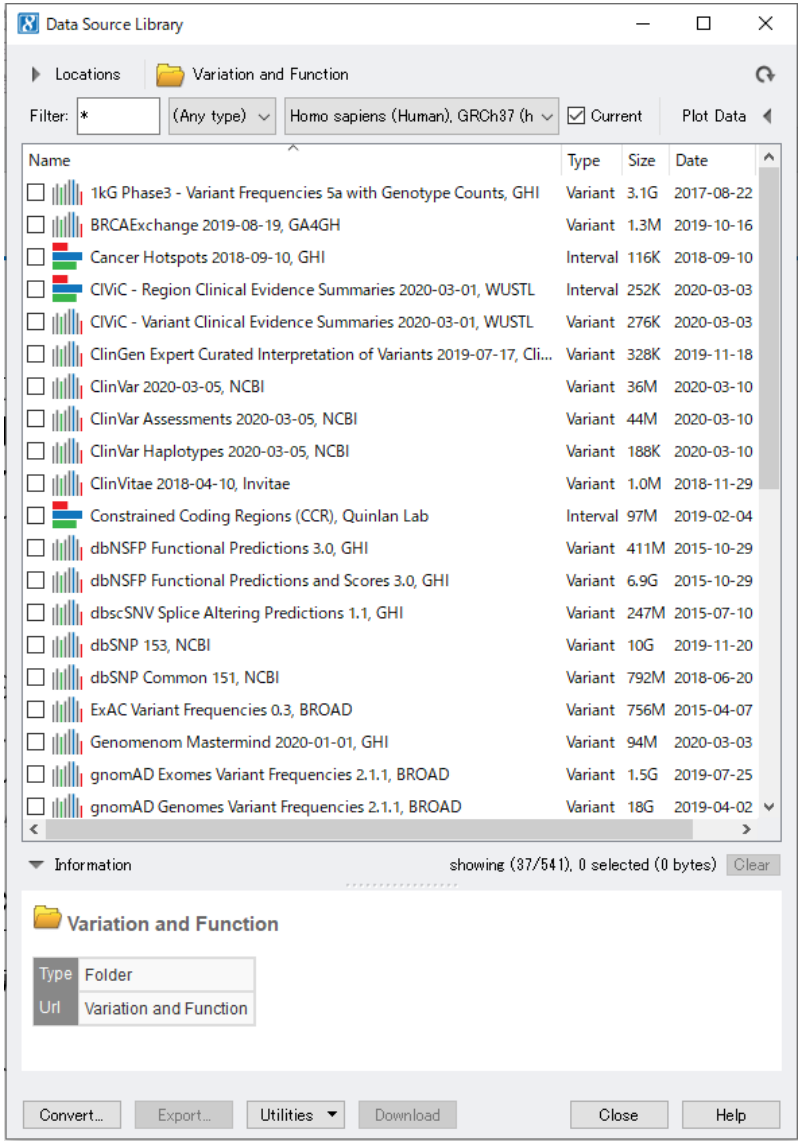
- ✓ Generalized SKATにおけるBurden Testとのバランスを、最適な値に調整して計算が可能
- ✓ 各バリアントの効果が、様々な方向またはすべて同じ方向に作用する場合のどちらでも、高い検出力を持つ

フィルタリングによるバリアントの選択

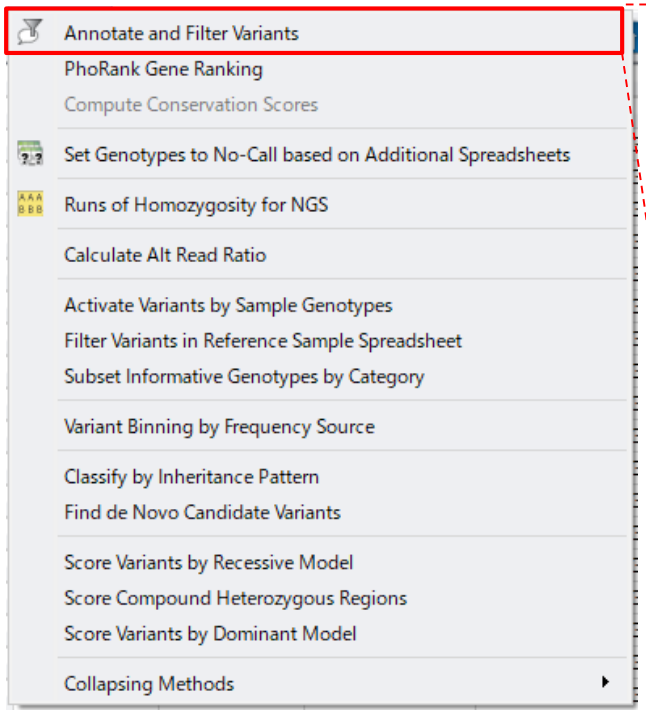
- ✓ Collaping Methods実行前に、バリアントデータのフィルタリングを実行
  - タンパク質機能にダメージを与えるバリアントのみを抽出
  - 特定集団におけるSNPの除去

- ✓ フィルタリングに用いるバリアントデータベースは、ソフトウェア搭載のデータベース管理ツールよりダウンロードが可能
  - RefSeq Gene
  - dbNSFP (SIFT, Polyphen…)
  - アリル頻度データベース（dbSNP, 1000 Genomes…）

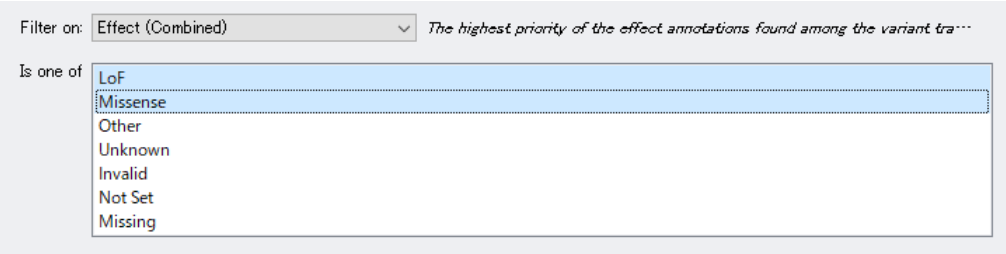
- ✓ フィルタリング実行時には、ソフトウェア搭載のフィルタリング用ツールを使用



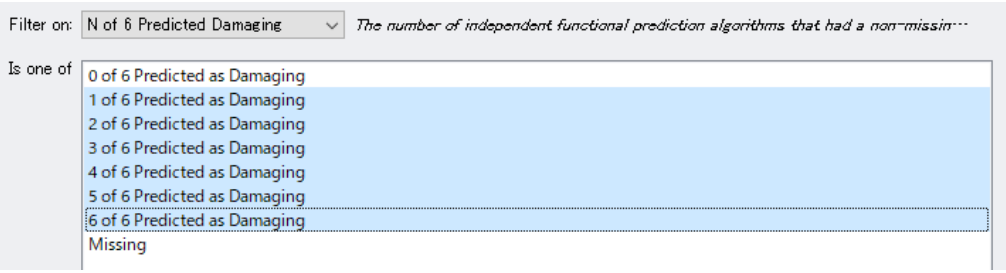
フィルタリング例



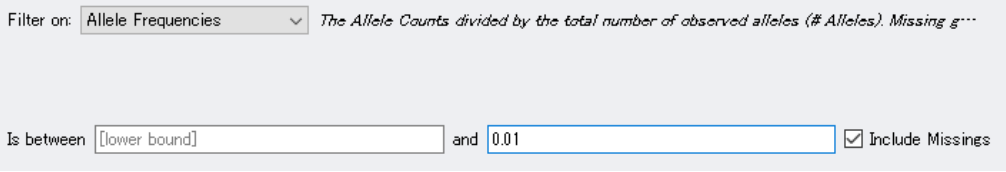
- RefSeq Gene（非同義変異・機能喪失変異などの抽出）



- dbNSFP（タンパク質機能にダメージを与えるバリアントの抽出）

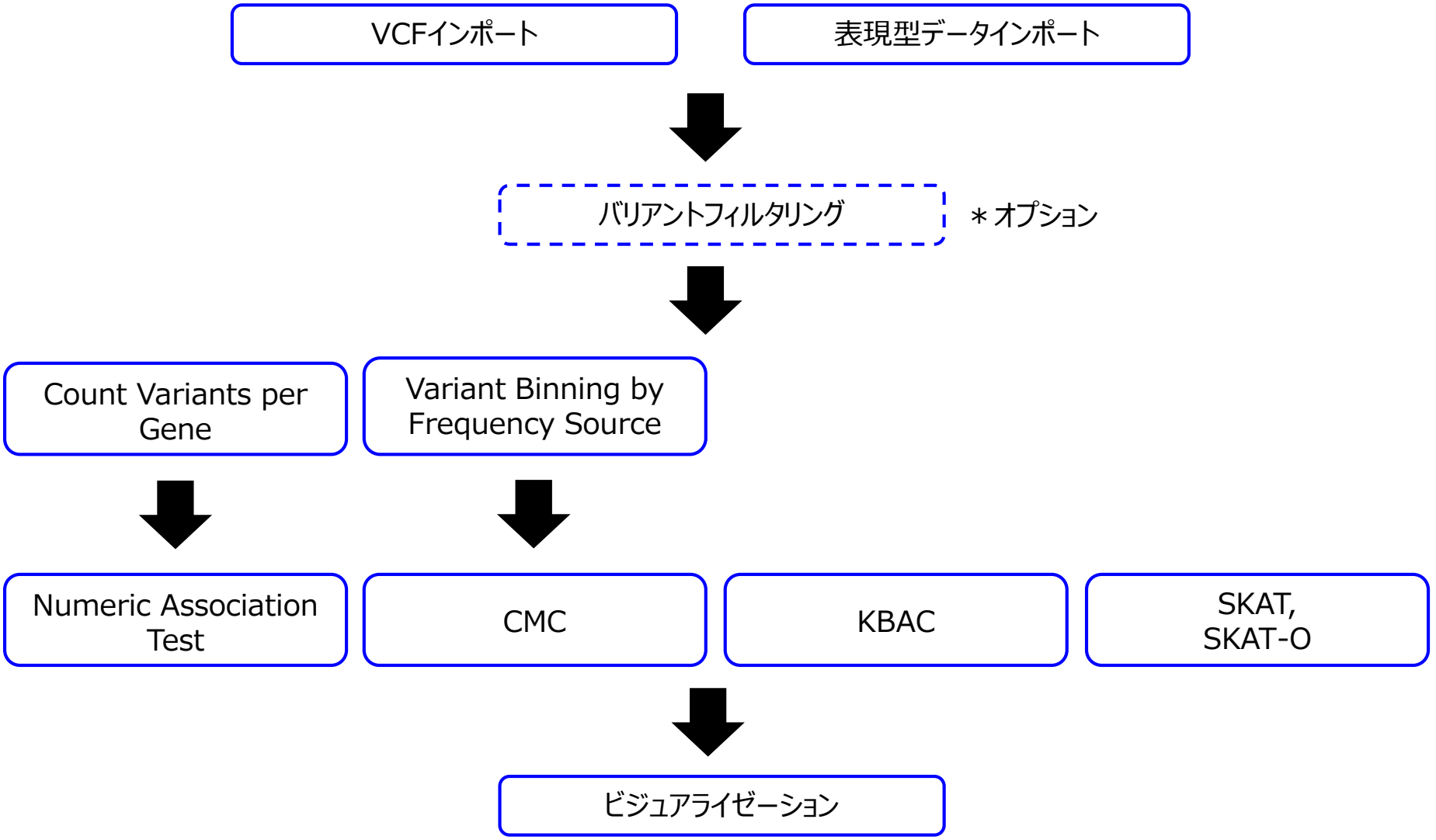


- 1000 Genomes（特定集団内SNPなどの除去）

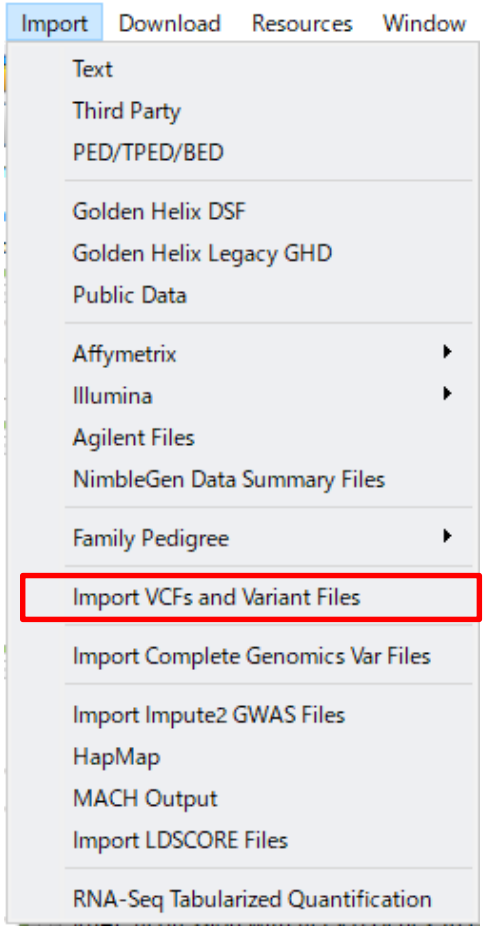


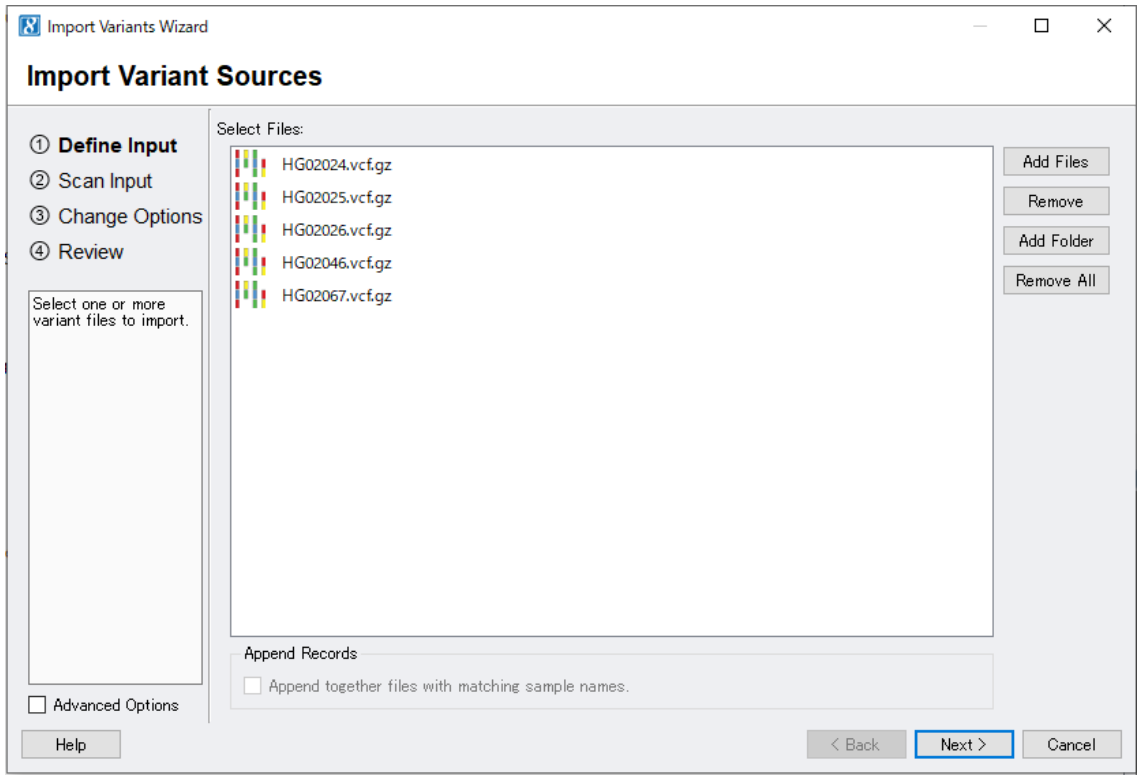
✓ フィルタリング実行時には、ソフトウェア搭載のフィルタリング用ツールを使用

解析ワークフロー


VCFインポート
表現型データインポート
バリアントフィルタリング
＊オプション
Count Variants per Gene
Variant Binning by Frequency Source
Numeric Association Test
CMC
KBAC
SKAT, SKAT-O
ビジュアライゼーション

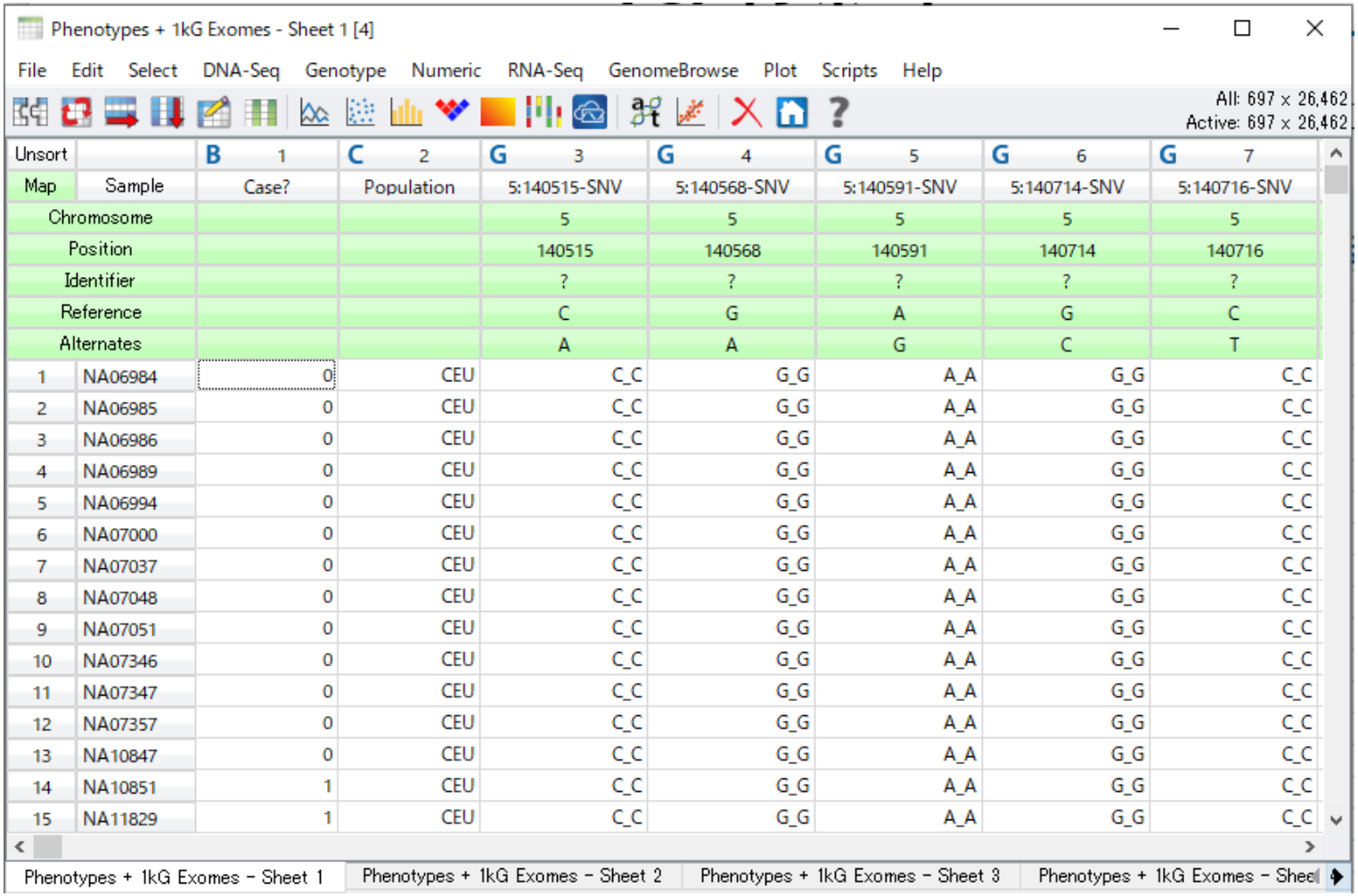
VCFインポート





- ✓ バリアントコール用ツールなどで作成したVCFファイルをインポート
- ✓ 1ファイル内に複数サンプルのデータを含んだMulti VCFファイルのインポートにも対応

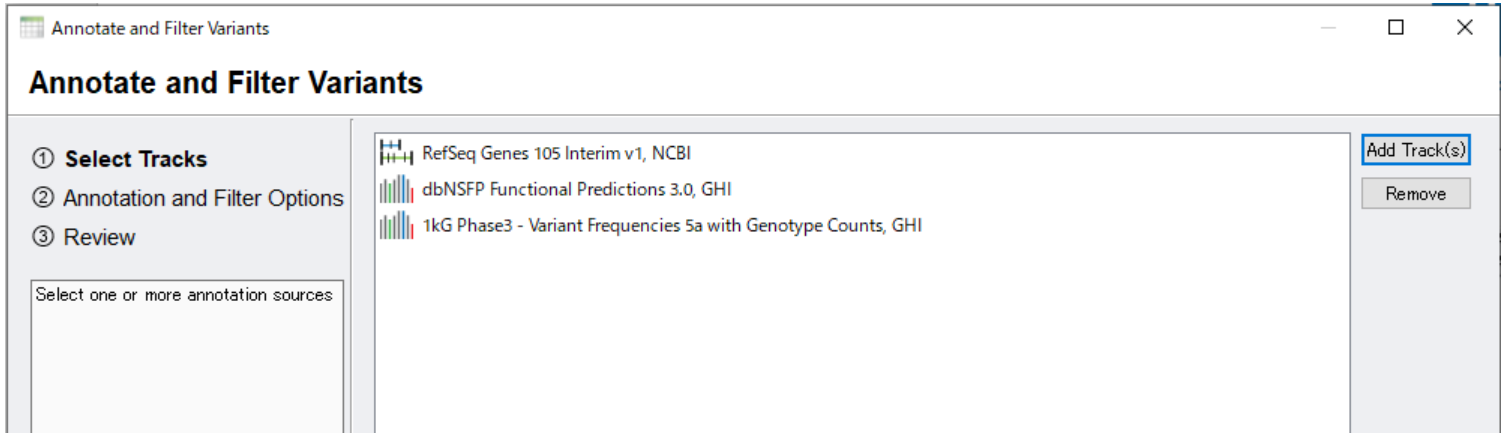
VCFインポート

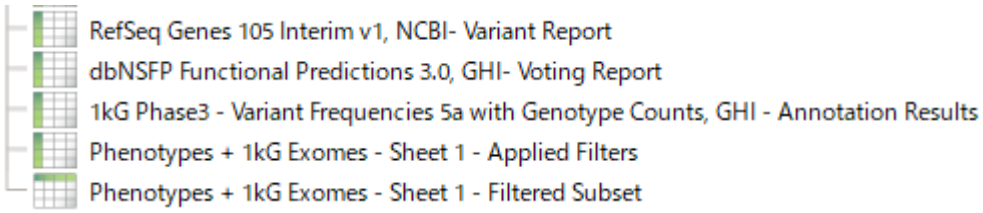


Unsort		B 1	C 2	G 3	G 4	G 5	G 6	G 7
Map	Sample	Case?	Population	5:140515-SNV	5:140568-SNV	5:140591-SNV	5:140714-SNV	5:140716-SNV
Chromosome				5	5	5	5	5
Position				140515	140568	140591	140714	140716
Identifier				?	?	?	?	?
Reference				C	G	A	G	C
Alternates				A	A	G	C	T
1	NA06984	0	CEU	C_C	G_G	A_A	G_G	C_C
2	NA06985	0	CEU	C_C	G_G	A_A	G_G	C_C
3	NA06986	0	CEU	C_C	G_G	A_A	G_G	C_C
4	NA06989	0	CEU	C_C	G_G	A_A	G_G	C_C
5	NA06994	0	CEU	C_C	G_G	A_A	G_G	C_C
6	NA07000	0	CEU	C_C	G_G	A_A	G_G	C_C
7	NA07037	0	CEU	C_C	G_G	A_A	G_G	C_C
8	NA07048	0	CEU	C_C	G_G	A_A	G_G	C_C
9	NA07051	0	CEU	C_C	G_G	A_A	G_G	C_C
10	NA07346	0	CEU	C_C	G_G	A_A	G_G	C_C
11	NA07347	0	CEU	C_C	G_G	A_A	G_G	C_C
12	NA07357	0	CEU	C_C	G_G	A_A	G_G	C_C
13	NA10847	0	CEU	C_C	G_G	A_A	G_G	C_C
14	NA10851	1	CEU	C_C	G_G	A_A	G_G	C_C
15	NA11829	1	CEU	C_C	G_G	A_A	G_G	C_C

- ✓ タブ区切りテキストファイルなどから別にインポートした表現型データと統合し、スプレッドシート形式でデータを管理

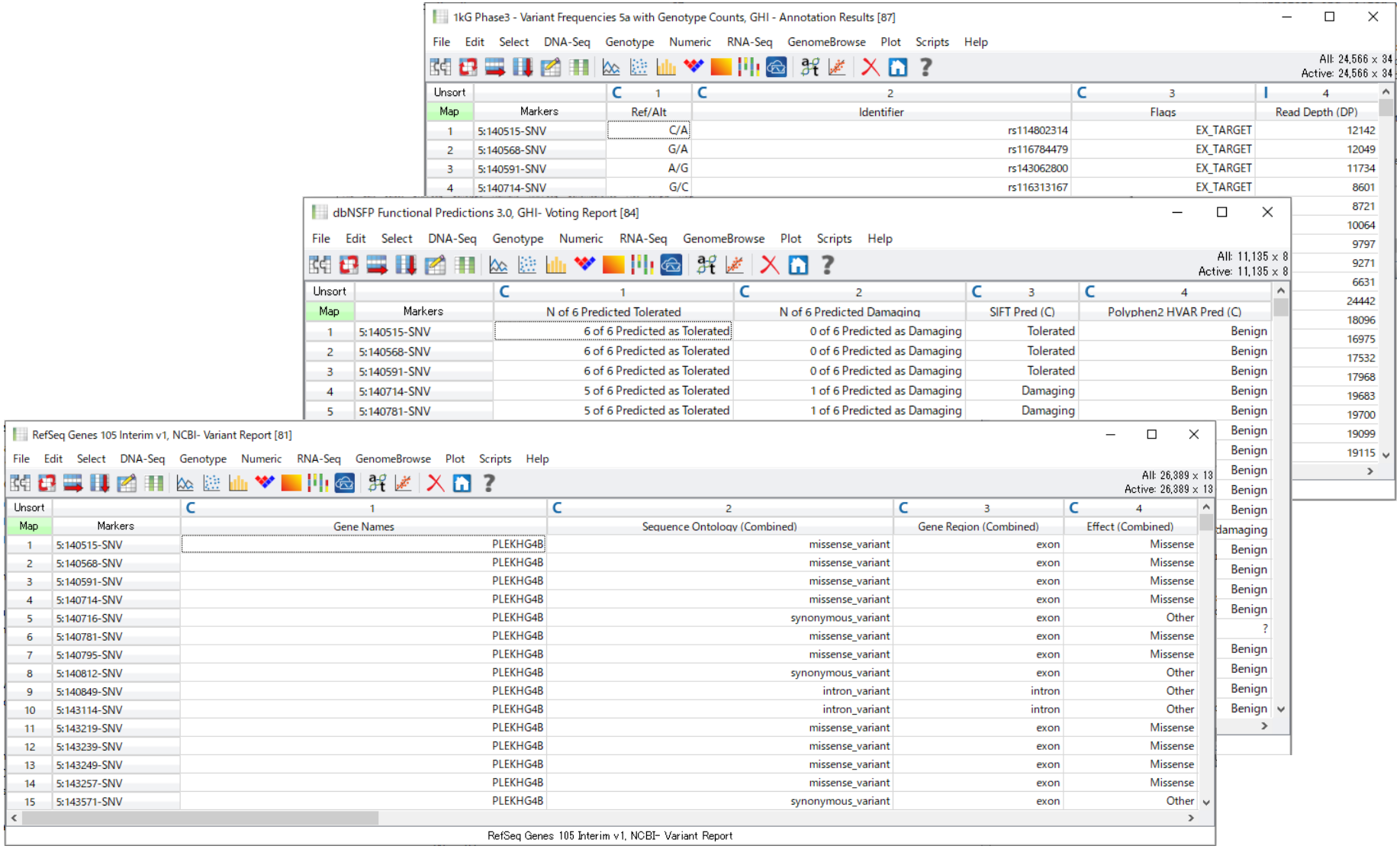
バリアントフィルタリング





- ✓ フィルタリングに用いるデータベースを選択し、検索条件を設定
- ✓ 検索終了後、各データベースごとにアノテーション付けされたバリアントのリストと、設定した検索条件でバリアントが絞り込まれたシートを出力

関連解析前処理

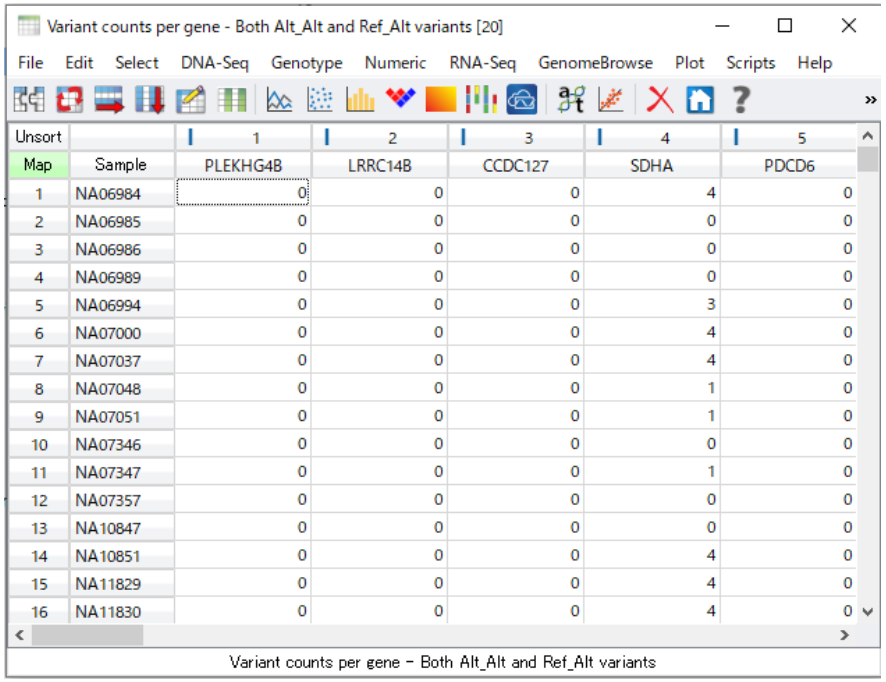


- ✓ バリアントリストでは、各バリアントのアノテーション情報を確認が可能

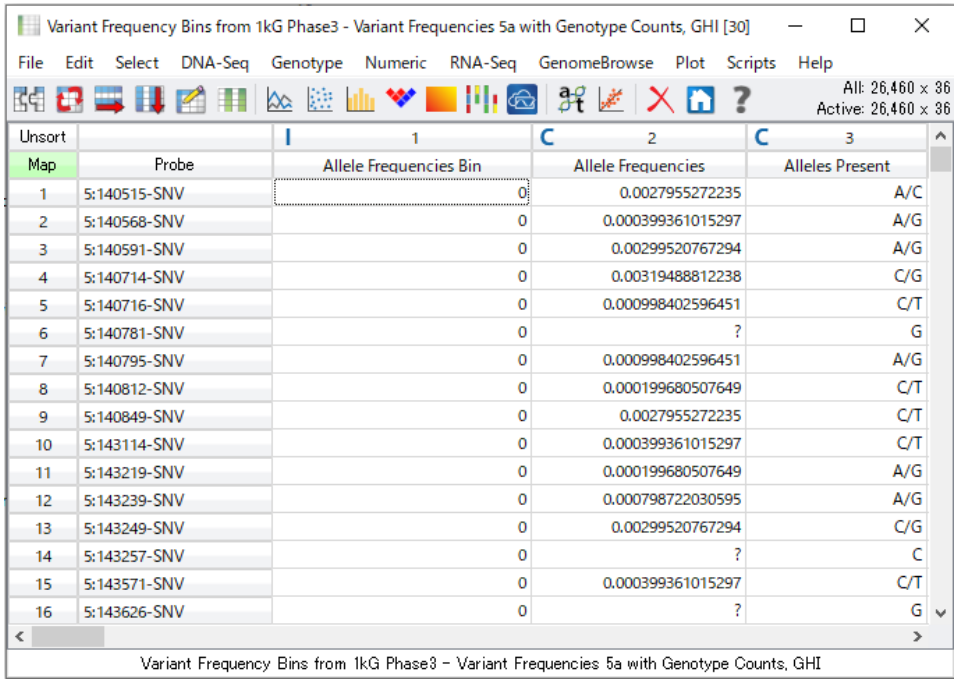
関連解析前処理

Filgen® biosciences & nanosciences

Count Variants per Gene



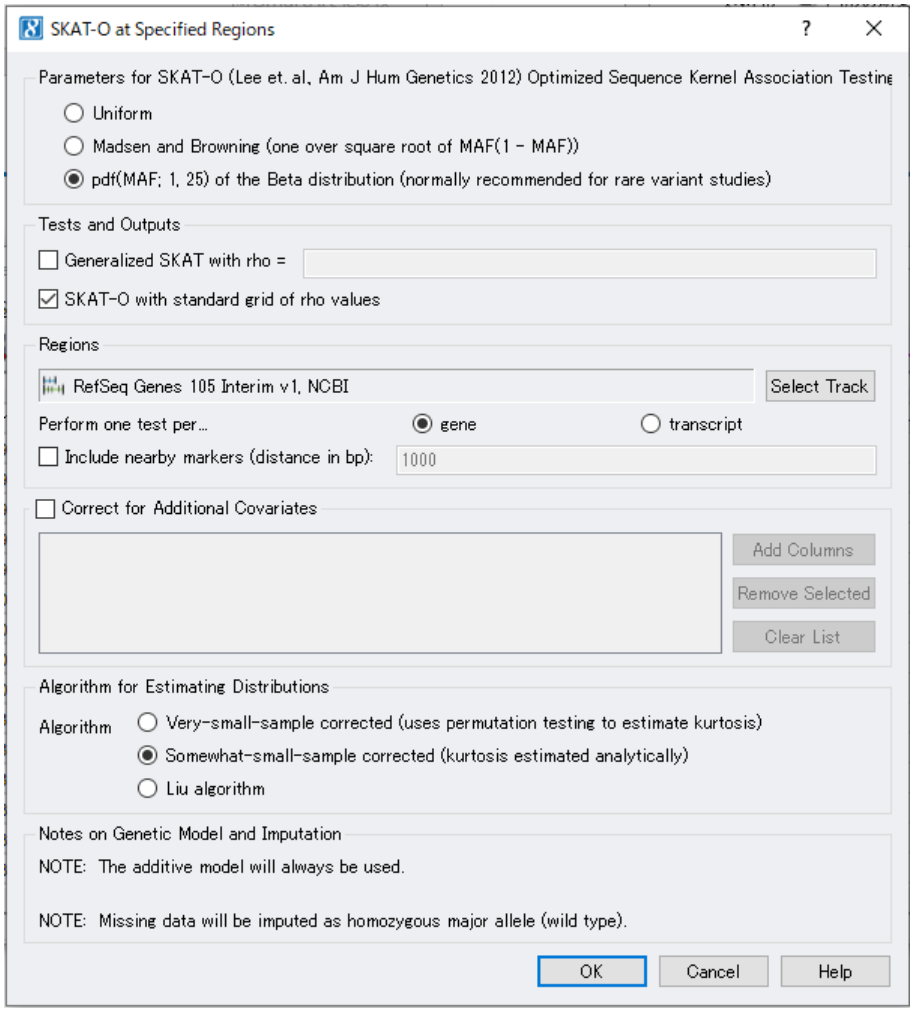
Variant Binning by Frequency Source



- ✓ CASTでは、遺伝子ごとバリアント数のカウントテーブルシート、CMCでは、バリアントをアリル頻度ごとに分類・集約したシートを、関連解析前に作成
- ✓ CASTでは、カウントテーブルシートに、表現型データを結合したのちに、関連解析を実行

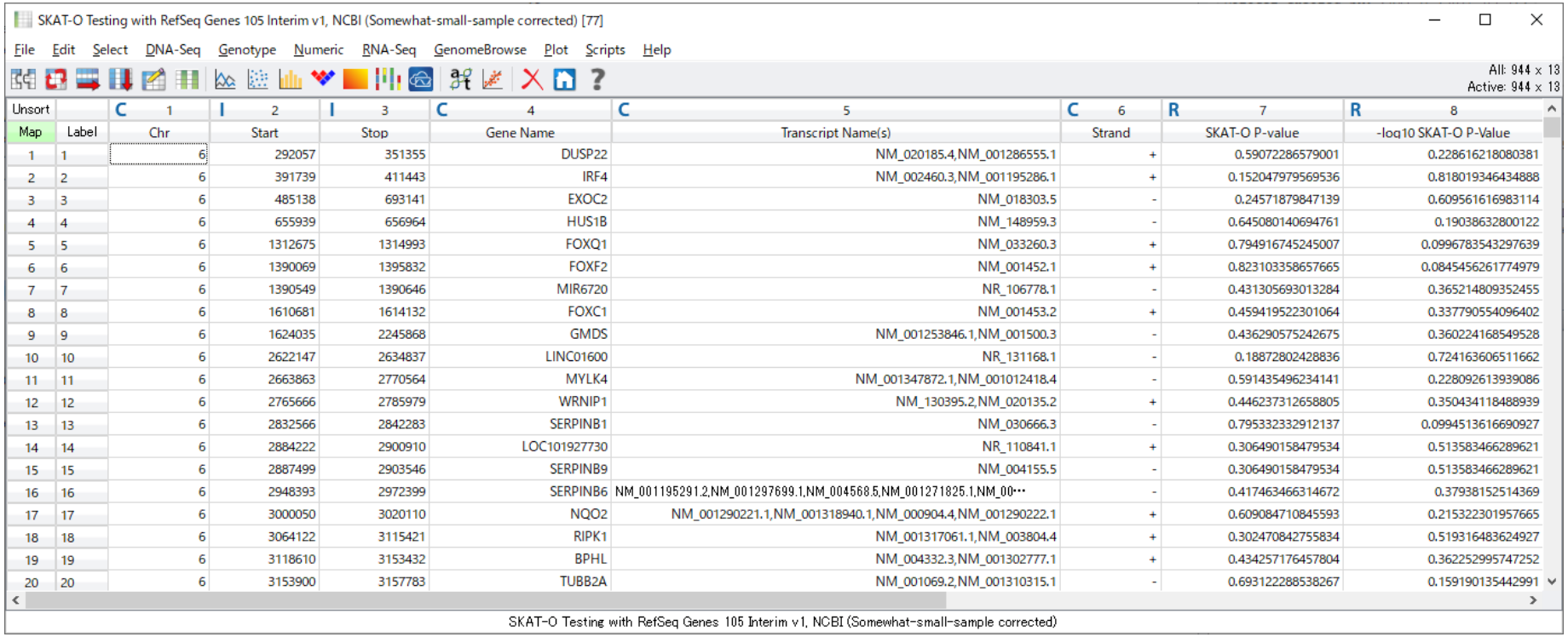
関連解析の実行

Unsort		B 1	C 2	G 3	G 4
Map	Sample	Case?	Population	5:140515-SNV	5:140568-SNV
1	NA06984	0	CEU	C_C	G_G
2	NA06985	0	CEU	C_C	G_G
3	NA06986	0	CEU	C_C	G_G
4	NA06989	0	CEU	C_C	G_G
5	NA06994	0	CEU	C_C	G_G
6	NA07000	0	CEU	C_C	G_G
7	NA07037	0	CEU	C_C	G_G
8	NA07048	0	CEU	C_C	G_G
9	NA07051	0	CEU	C_C	G_G
10	NA07346	0	CEU	C_C	G_G
11	NA07347	0	CEU	C_C	G_G
12	NA07357	0	CEU	C_C	G_G
13	NA10847	0	CEU	C_C	G_G
14	NA10851	1	CEU	C_C	G_G
15	NA11829	1	CEU	C_C	G_G



- ✓ スプレッドシート上で、検定に用いる表現型データのカラムを選択し、関連解析用ツールを起動
- ✓ 各種パラメータや、必要となる外部のデータリソース（遺伝子アノテーションデータベースなど）を設定し、解析を実行

解析結果


SKAT-O Testing with RefSeq Genes 105 Interim v1, NCBI (Somewhat-small-sample corrected) [77]

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

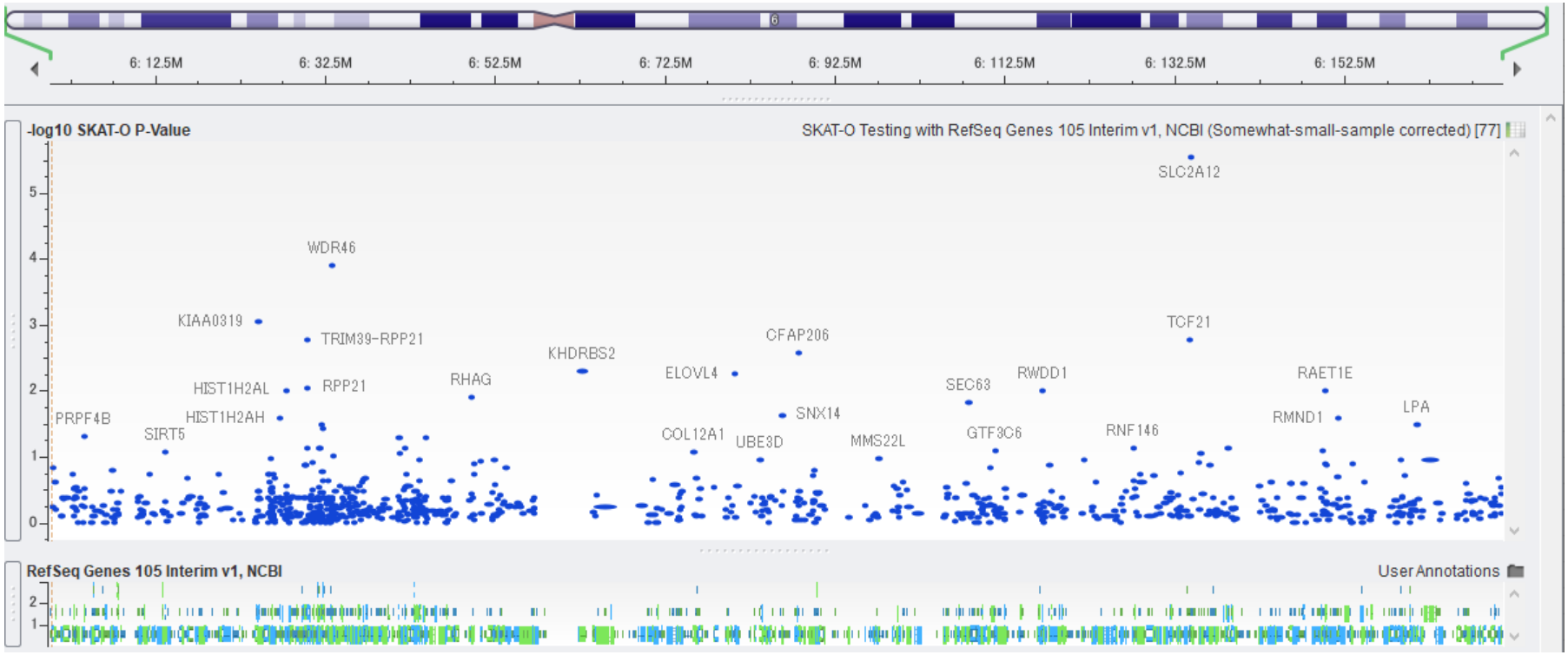
All: 944 x 13
Active: 944 x 13

Unsort		C 1	I 2	I 3	C 4	C 5	C 6	R 7	R 8
Map	Label	Chr	Start	Stop	Gene Name	Transcript Name(s)	Strand	SKAT-O P-value	-log10 SKAT-O P-Value
1	1	6	292057	351355	DUSP22	NM_020185.4,NM_001286555.1	+	0.59072286579001	0.228616218080381
2	2	6	391739	411443	IRF4	NM_002460.3,NM_001195286.1	+	0.152047979569536	0.818019346434888
3	3	6	485138	693141	EXOC2	NM_018303.5	-	0.24571879847139	0.609561616983114
4	4	6	655939	656964	HUS1B	NM_148959.3	-	0.645080140694761	0.19038632800122
5	5	6	1312675	1314993	FOXQ1	NM_033260.3	+	0.794916745245007	0.0996783543297639
6	6	6	1390069	1395832	FOXF2	NM_001452.1	+	0.823103358657665	0.0845456261774979
7	7	6	1390549	1390646	MIR6720	NR_106778.1	-	0.431305693013284	0.365214809352455
8	8	6	1610681	1614132	FOXC1	NM_001453.2	+	0.459419522301064	0.337790554096402
9	9	6	1624035	2245868	GMDS	NM_001253846.1,NM_001500.3	-	0.436290575242675	0.360224168549528
10	10	6	2622147	2634837	LINC01600	NR_131168.1	-	0.18872802428836	0.724163606511662
11	11	6	2663863	2770564	MYLK4	NM_001347872.1,NM_001012418.4	-	0.591435496234141	0.228092613939086
12	12	6	2765666	2785979	WRNIP1	NM_130395.2,NM_020135.2	+	0.446237312658805	0.350434118488939
13	13	6	2832566	2842283	SERPINB1	NM_030666.3	-	0.795332332912137	0.0994513616690927
14	14	6	2884222	2900910	LOC101927730	NR_110841.1	+	0.306490158479534	0.513583466289621
15	15	6	2887499	2903546	SERPINB9	NM_004155.5	-	0.306490158479534	0.513583466289621
16	16	6	2948393	2972399	SERPINB6	NM_001195291.2,NM_001297699.1,NM_004568.5,NM_001271825.1,NM_00…	-	0.417463466314672	0.37938152514369
17	17	6	3000050	3020110	NQO2	NM_001290221.1,NM_001318940.1,NM_000904.4,NM_001290222.1	+	0.609084710845593	0.215322301957665
18	18	6	3064122	3115421	RIPK1	NM_001317061.1,NM_003804.4	+	0.302470842755834	0.519316483624927
19	19	6	3118610	3153432	BPHL	NM_004332.3,NM_001302777.1	+	0.434257176457804	0.362252995747252
20	20	6	3153900	3157783	TUBB2A	NM_001069.2,NM_001310315.1	-	0.693122288538267	0.159190135442991

SKAT-O Testing with RefSeq Genes 105 Interim v1, NCBI (Somewhat-small-sample corrected)

- ✓ 計算が終了すると、遺伝子ごとのP-valueや各種統計値をまとめたシートが出力される
- ✓ シート上より、各種のデータ（-log10 P-valueなど）をゲノムブラウザーにプロットが可能

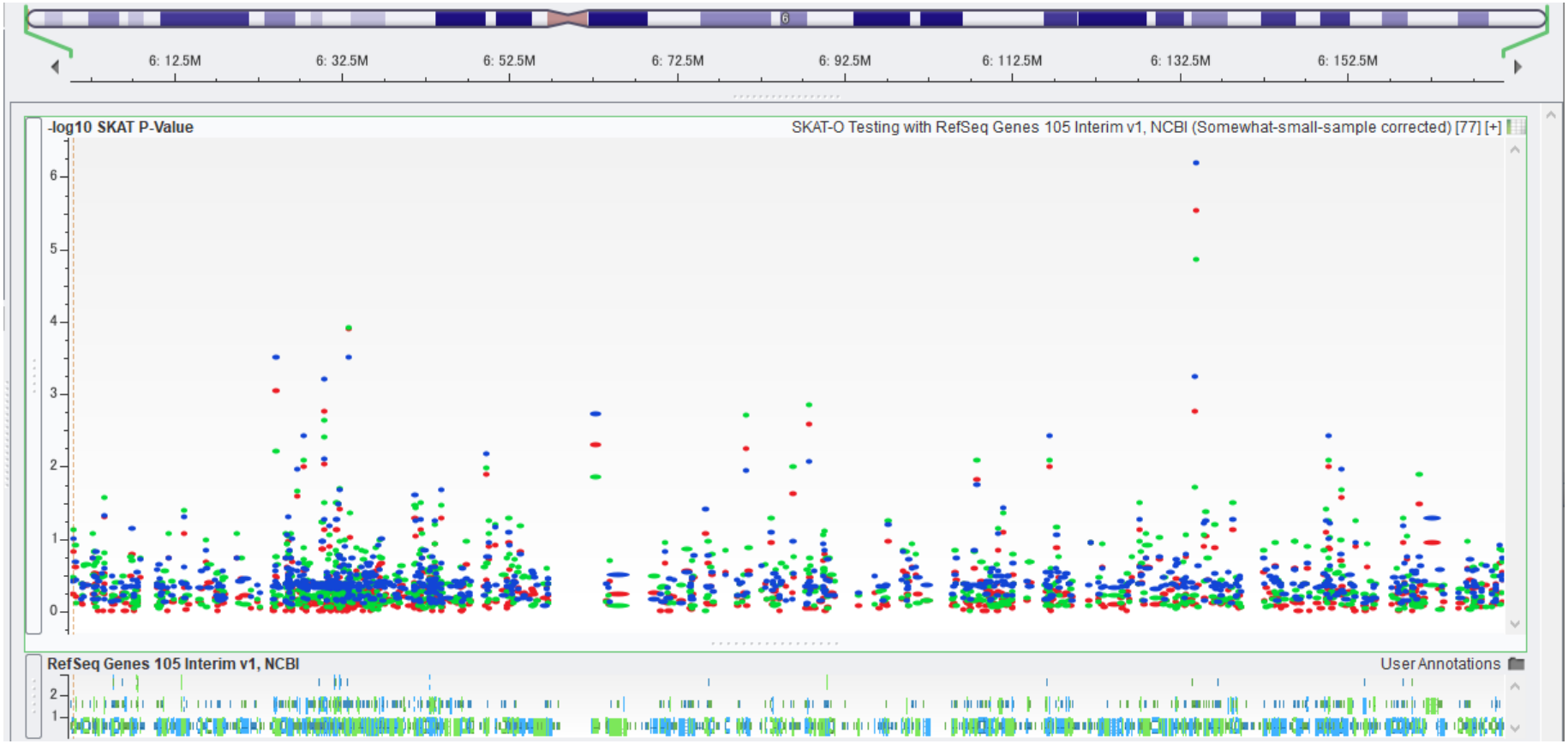
ビジュアライゼーション



- ✓ ゲノムブラウザーには、各種アノテーションデータベースに加え、サンプルのバリアントデータや、他の統計手法で計算した値などもプロットが可能
- ✓ プロットの表示色や属性によるラベル付けなどを行うことで、データの識別や解釈などを行いやすくなる

複数の解析結果の比較

赤：SKAT-O　　青：Burden Test (rho = 1)　　緑：SKAT (rho = 0)



お問い合わせ先：フィルジェン株式会社

TEL: 052-624-4388 (9:00～18：00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp