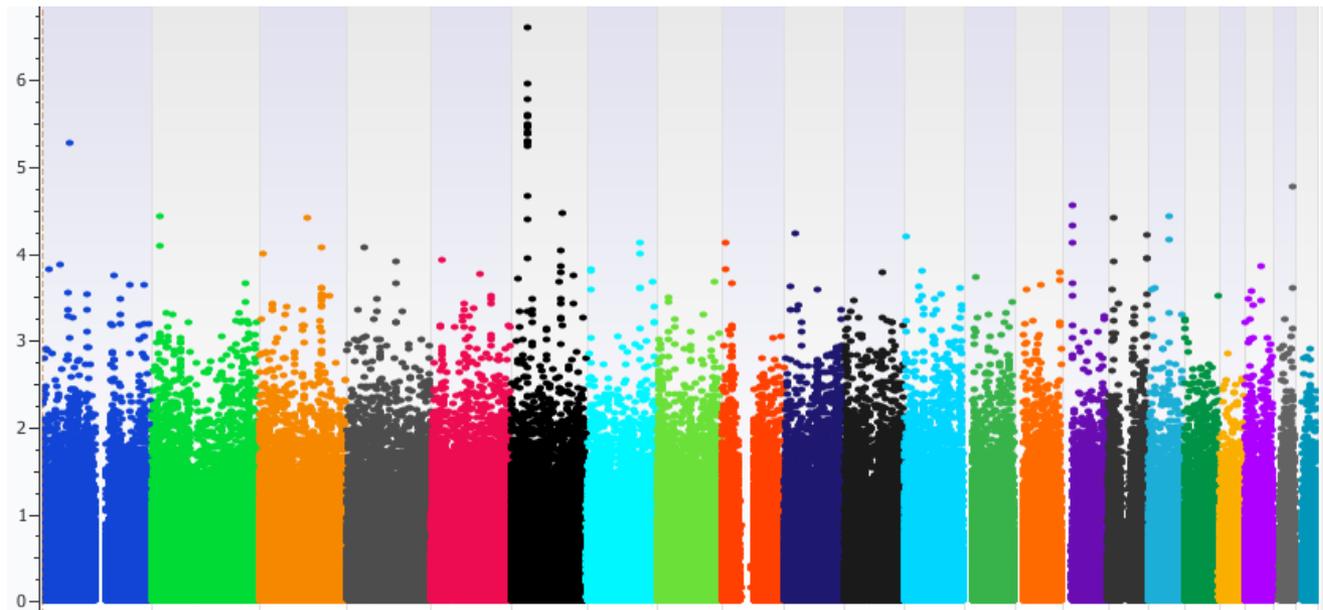


多因子疾患における レアバリエント関連解析

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)

- 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シーケ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（レアバリエント）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、レアバリエント解析用の解析ツールを搭載しており、SNPとレアバリエント両方の解析に対応できる





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



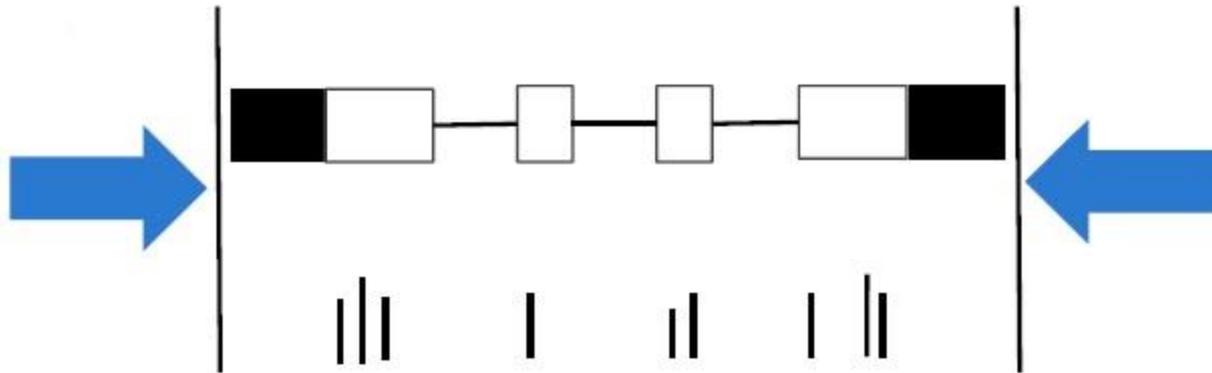
- Annotation and Filtering
- ACMG & AMP Guidelines
- CNV Calling
- Clinical Reporting
- Pipeline: Run Workflows

- Golden Helix社では、遺伝統計解析ソフトウェア「SNP & Variation Suite」と、クリニカルシーケ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

- レアバリエ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



バリエントを遺伝子ごとに集約し、単一ユニットとして解析



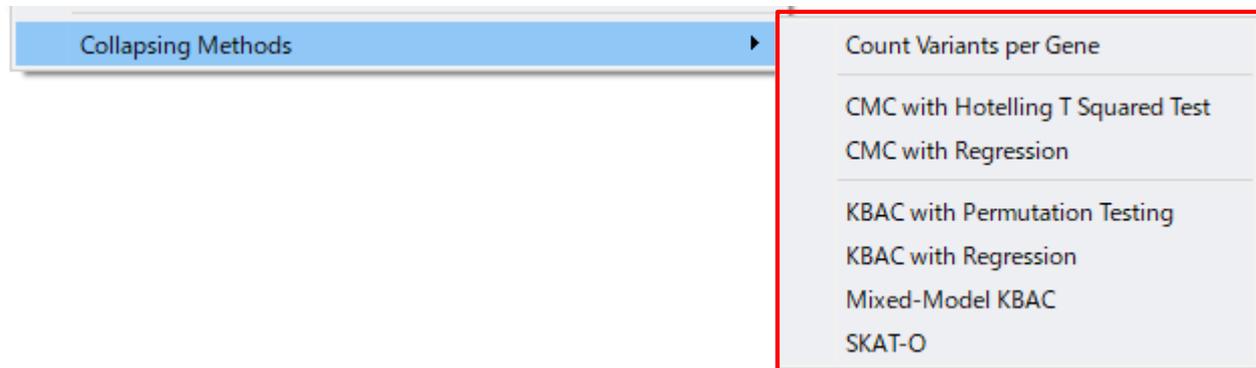
Collapsing Methods

Burden Tests

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

Kernel Tests

- 遺伝子ごとに、各バリエーションの統計スコアの2乗を合計して検定を行う
- 各バリエーションの効果が、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エーション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、バリエーション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

The 'Count Variants per Gene' dialog box is shown with the following settings:

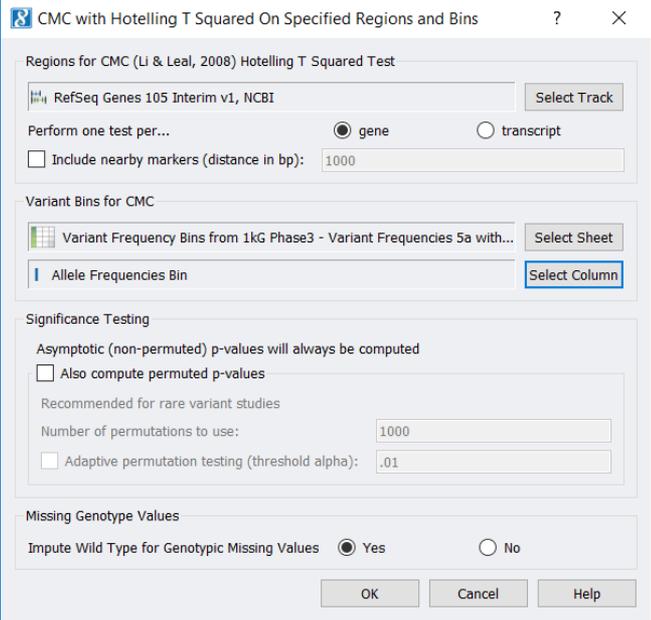
- Select a gene track: RefSeq Genes 105 Interim v1, NCBI
- Select the reference allele field from the marker map: Reference
- Include upstream/downstream variants (distance in bp): 1000
- Choose the types of variants to count: Both Ref_Alt and Alt_Alt variants
- Select the type of output to generate: Count of Number of Variants (Per Gene)

The 'Numeric Association Tests' dialog box is shown with the following settings:

- Case/Control dependent variable: Case? (47 cases and 650 controls), 1744 markers from chromosomes: 5, 6
- Association Test Parameters: Correlation/Trend test (checked), T-test (unchecked), Logistic regression (unchecked)
- Multiple Testing Correction: Bonferroni adjustment (on N covariates) (checked), False Discovery Rate (FDR) (checked), Single value permutations (unchecked), Full scan permutations (unchecked)
- Number of permutations: (empty field)
- Additional Outputs: Output data for P-P/Q-Q plots (unchecked), Output $-\log_{10}(P)$ (checked)
- Principal Components Analysis (PCA): Correct for batch effects/stratification with PCA (unchecked), Use corrected dependent (unchecked)

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、バリエ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 T² 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）



CMC with Hotelling T Squared On Specified Regions and Bins

Regions for CMC (Li & Leal, 2008) Hotelling T Squared Test

RefSeq Genes 105 Interim v1, NCBI Select Track

Perform one test per... gene transcript

Include nearby markers (distance in bp): 1000

Variant Bins for CMC

Variant Frequency Bins from 1kG Phase3 - Variant Frequencies 5a with... Select Sheet

Allele Frequencies Bin Select Column

Significance Testing

Asymptotic (non-permuted) p-values will always be computed

Also compute permuted p-values

Recommended for rare variant studies

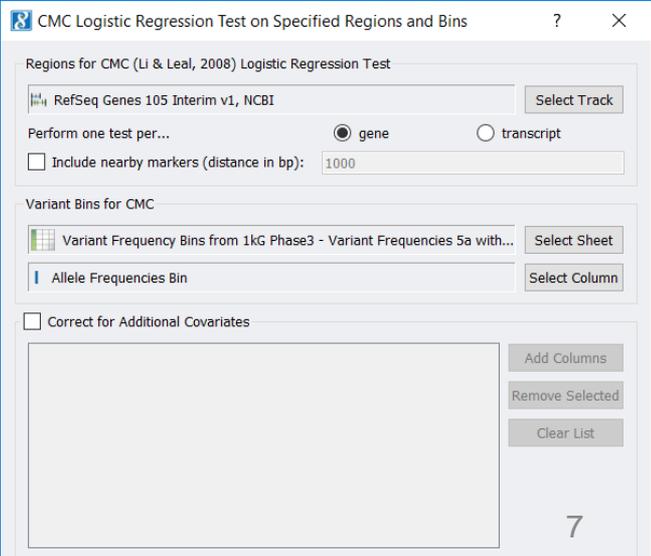
Number of permutations to use: 1000

Adaptive permutation testing (threshold alpha): .01

Missing Genotype Values

Impute Wild Type for Genotypic Missing Values Yes No

OK Cancel Help



CMC Logistic Regression Test on Specified Regions and Bins

Regions for CMC (Li & Leal, 2008) Logistic Regression Test

RefSeq Genes 105 Interim v1, NCBI Select Track

Perform one test per... gene transcript

Include nearby markers (distance in bp): 1000

Variant Bins for CMC

Variant Frequency Bins from 1kG Phase3 - Variant Frequencies 5a with... Select Sheet

Allele Frequencies Bin Select Column

Correct for Additional Covariates

Add Columns

Remove Selected

Clear List

7

KBAC (Kernel-Based Adaptive Cluster)

- ✓ 各バリエントに対して、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、バリエントを重み付けし、関連解析を実行
- ✓ レアバリエントのみを解析に用いるように、解析前にバリエ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

The screenshot shows the 'KBAC with Permutation Testing at Specified Regions' dialog box. The 'Parameters for KBAC (Liu & Leal, 2010) Using Permutation Testing' section includes the following settings:

- Kernel Type:** Hyper-geometric (Normally recommended), Marginal binomial, Asymptotic normal (For large sample sizes only) (> 400 cases/400 controls)
- Permutation Parameters:**
 - Number of permutations to use: 1000
 - Permutation Mode: Standard C/C permutation procedure, KBAC Monte-Carlo approximation (For large sample sizes only) (> 400 cases/400 controls)
 - Adaptive permutation testing (threshold alpha): .01
- Outputs:** One-sided statistics (recommended), Two-sided statistics
- Regions:** RefSeq Genes 105 Interim v1, NCBI (with a 'Select Track' button)
- Perform one test per...:** gene, transcript
- Include nearby markers (distance in bp):** 1000
- Note on Variant Selection:** You should have filtered the markers by rare variants. It is assumed in this version of KBAC that only rare variant markers will be active.
- Missing Genotype Values:** Impute Wild Type for Genotypic Missing Values, No

Buttons at the bottom: OK, Cancel, Help.

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と組み合わせた計算が可能

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

SKAT-O at Specified Regions

Parameters for SKAT-O (Lee et. al., Am J Hum Genetics 2012) Optimized Sequence Kernel Association Testing

Marker Weighting Uniform
 Madsen and Browning (one over square root of $MAF(1 - MAF)$)
 pdf(MAF; 1, 25) of the Beta distribution (normally recommended for rare variant studies)

Tests and Outputs

Generalized SKAT with rho =
 SKAT-O with standard grid of rho values

Regions

RefSeq Genes 105 Interim v1, NCBI

Perform one test per... gene transcript

Include nearby markers (distance in bp):

Correct for Additional Covariates

Algorithm for Estimating Distributions

Algorithm Very-small-sample corrected (uses permutation testing to estimate kurtosis)
 Somewhat-small-sample corrected (kurtosis estimated analytically)
 Liu algorithm

Notes on Genetic Model and Imputation

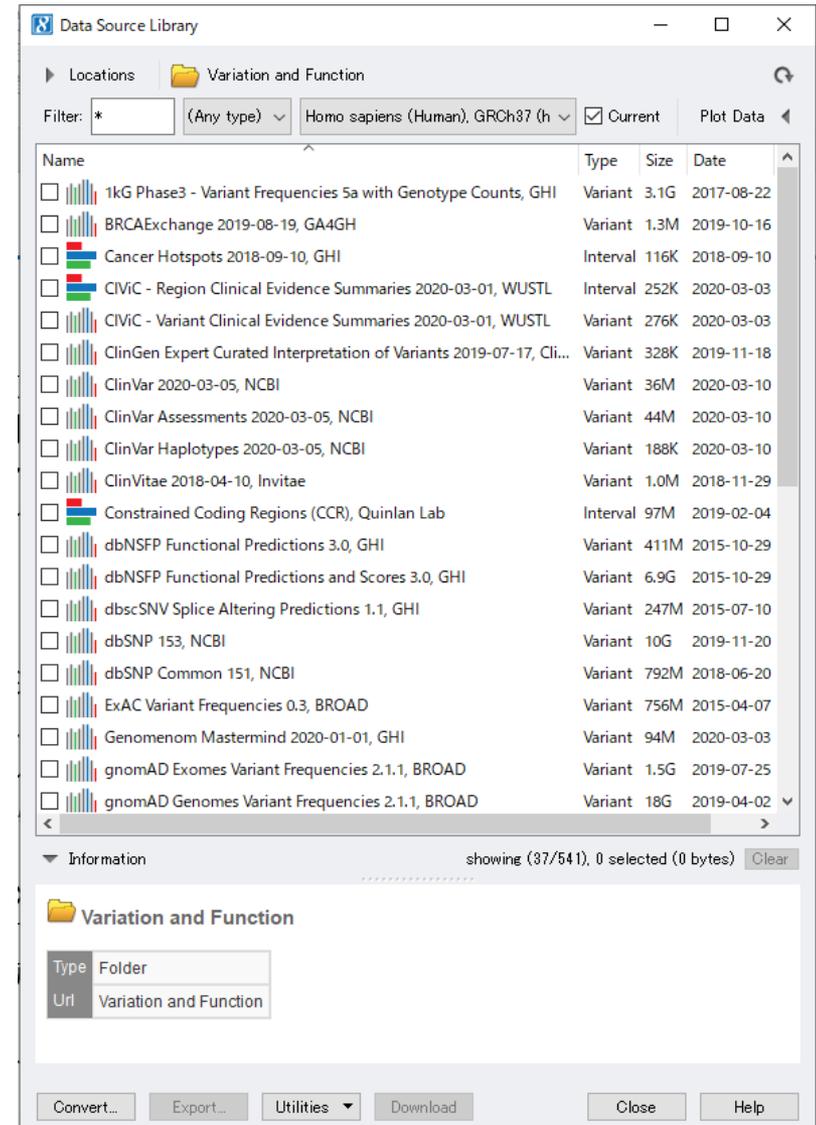
NOTE: The additive model will always be used.

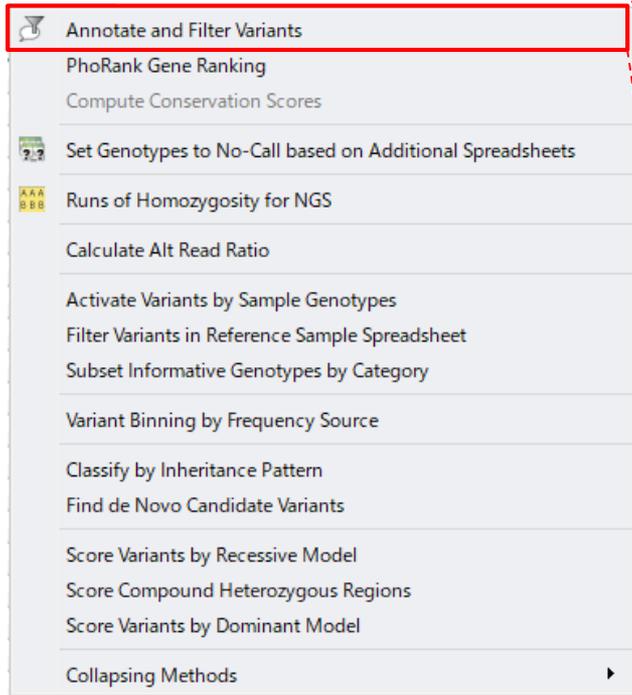
NOTE: Missing data will be imputed as homozygous major allele (wild type).

- ✓ Collaping Methods実行前に、バリエーション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エーションのみを抽出
 - 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エーションデータベースは、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





- RefSeq Gene (非同義変異・機能喪失変異などの抽出)

Filter on: Effect (Combined) The highest priority of the effect annotations found among the variant tra...

Is one of

- LoF
- Missense
- Other
- Unknown
- Invalid
- Not Set
- Missing

- dbNSFP (タンパク質機能にダメージを与えるバリエーションの抽出)

Filter on: N of 6 Predicted Damaging The number of independent functional prediction algorithms that had a non-missin...

Is one of

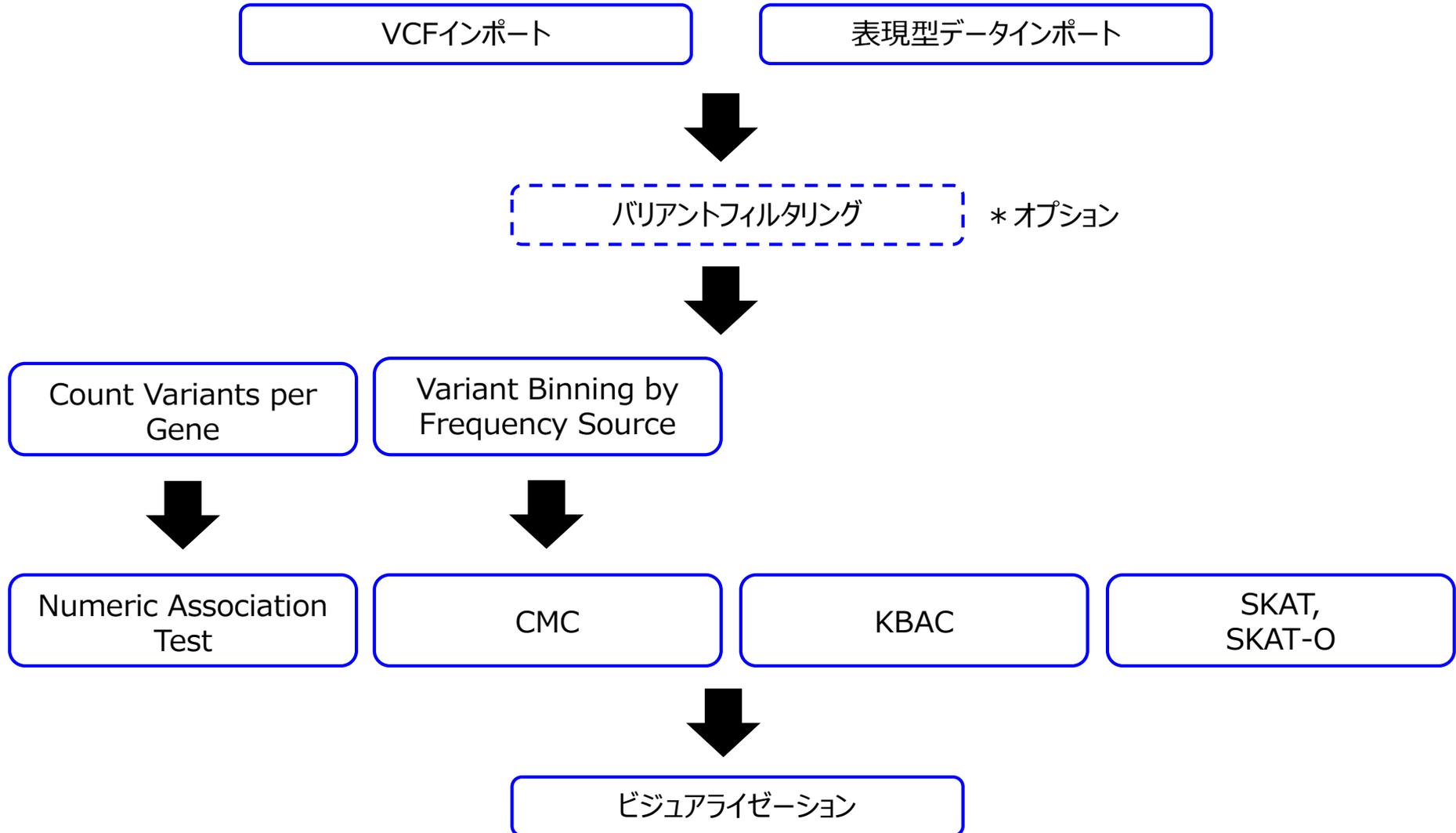
- 0 of 6 Predicted as Damaging
- 1 of 6 Predicted as Damaging
- 2 of 6 Predicted as Damaging
- 3 of 6 Predicted as Damaging
- 4 of 6 Predicted as Damaging
- 5 of 6 Predicted as Damaging
- 6 of 6 Predicted as Damaging
- Missing

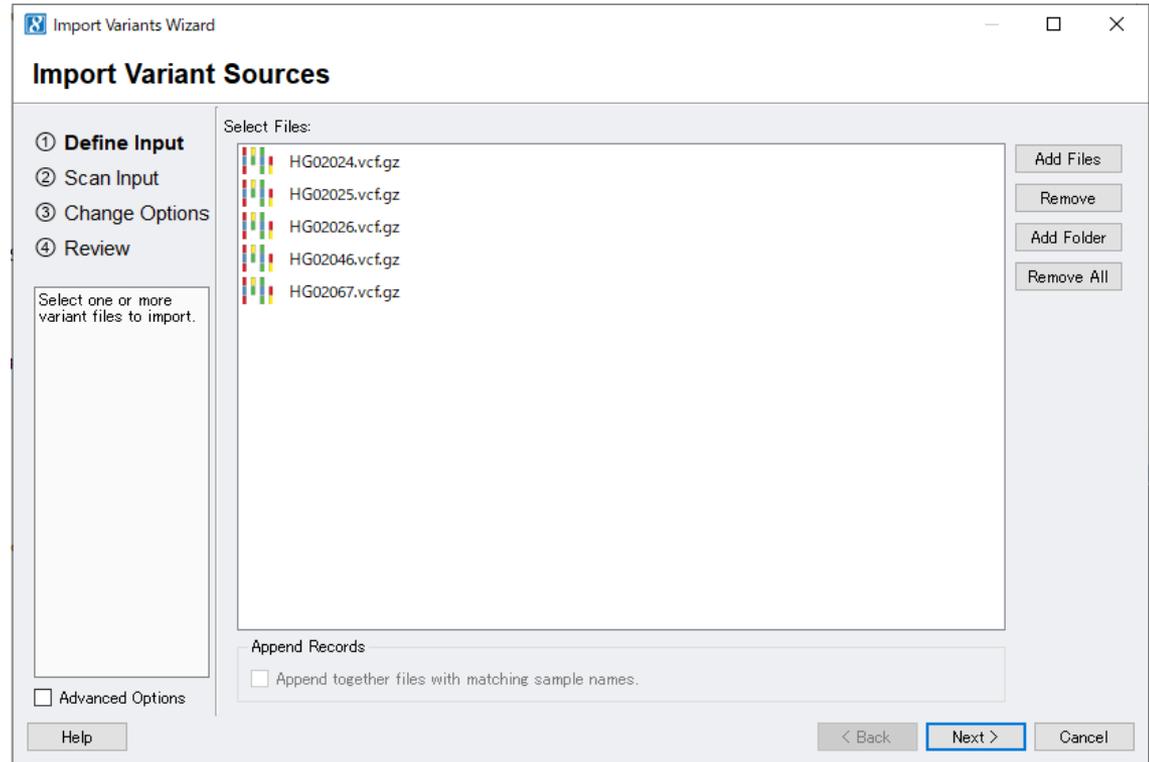
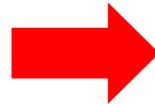
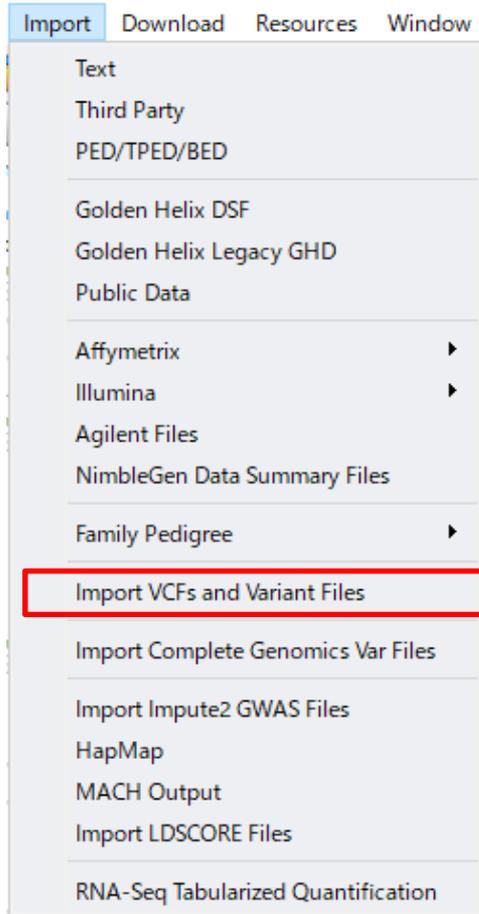
- 1000 Genomes (特定集団内SNPなどの除去)

Filter on: Allele Frequencies The Allele Counts divided by the total number of observed alleles (# Alleles). Missing g...

Is between [lower bound] and 0.01 Include Missings

- ✓ 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

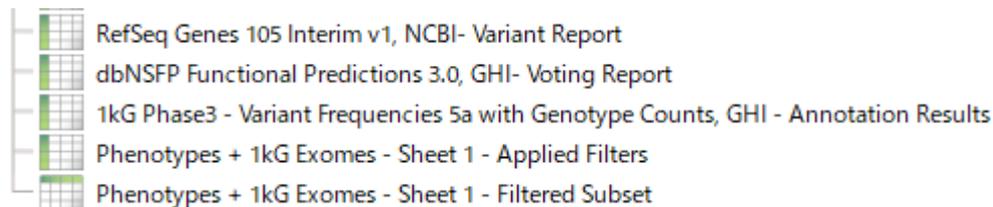
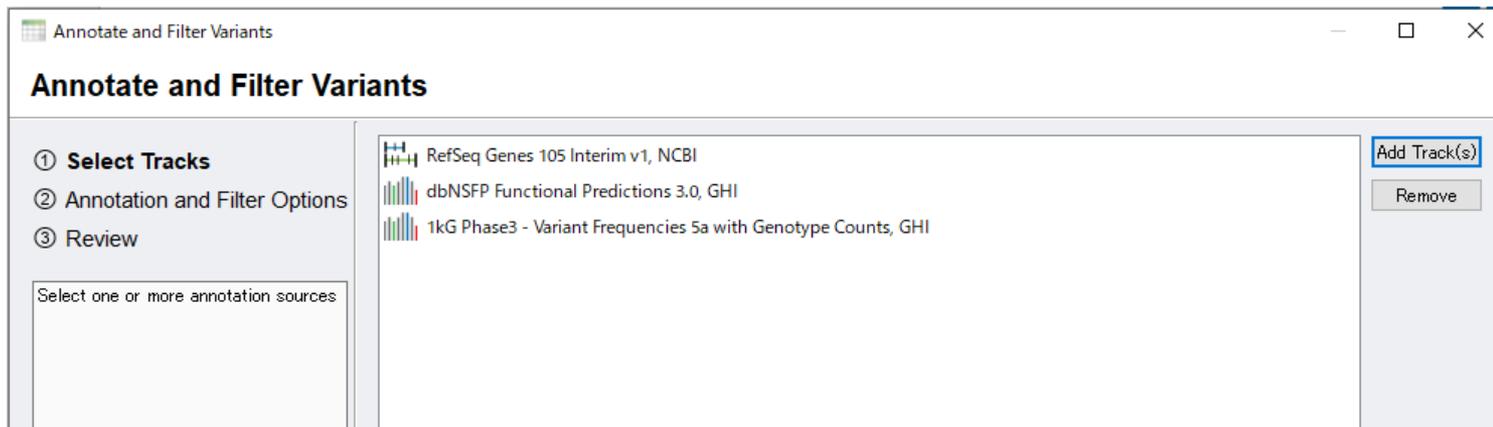




- ✓ バリエントコール用ツールなどで作成したVCFファイルをインポート
- ✓ 1ファイル内に複数サンプルのデータを含んだMulti VCFファイルのインポートにも対応

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



- ✓ 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エントのリストと、設定した検索条件でバリエントが絞り込まれたシートを出力

1kg Phase3 - Variant Frequencies 5a with Genotype Counts, GHI - Annotation Results [87]

Unsort	Map	Markers	1 Ref/Alt	2 Identifier	3 Flags	4 Read Depth (DP)
1	5:	140515-SNV	C/A	rs114802314	EX_TARGET	12142
2	5:	140568-SNV	G/A	rs116784479	EX_TARGET	12049
3	5:	140591-SNV	A/G	rs143062800	EX_TARGET	11734
4	5:	140714-SNV	G/C	rs116313167	EX_TARGET	8601

dbNSFP Functional Predictions 3.0, GHI- Voting Report [84]

Unsort	Map	Markers	1 N of 6 Predicted Tolerated	2 N of 6 Predicted Damaging	3 SIFT Pred (C)	4 Polyphen2 HVAR Pred (C)
1	5:	140515-SNV	6 of 6 Predicted as Tolerated	0 of 6 Predicted as Damaging	Tolerated	Benign
2	5:	140568-SNV	6 of 6 Predicted as Tolerated	0 of 6 Predicted as Damaging	Tolerated	Benign
3	5:	140591-SNV	6 of 6 Predicted as Tolerated	0 of 6 Predicted as Damaging	Tolerated	Benign
4	5:	140714-SNV	5 of 6 Predicted as Tolerated	1 of 6 Predicted as Damaging	Damaging	Benign
5	5:	140781-SNV	5 of 6 Predicted as Tolerated	1 of 6 Predicted as Damaging	Damaging	Benign

RefSeq Genes 105 Interim v1, NCBI- Variant Report [81]

Unsort	Map	Markers	1 Gene Names	2 Sequence Ontology (Combined)	3 Gene Region (Combined)	4 Effect (Combined)
1	5:	140515-SNV	PLEKHG4B	missense_variant	exon	Missense
2	5:	140568-SNV	PLEKHG4B	missense_variant	exon	Missense
3	5:	140591-SNV	PLEKHG4B	missense_variant	exon	Missense
4	5:	140714-SNV	PLEKHG4B	missense_variant	exon	Missense
5	5:	140716-SNV	PLEKHG4B	synonymous_variant	exon	Other
6	5:	140781-SNV	PLEKHG4B	missense_variant	exon	Missense
7	5:	140795-SNV	PLEKHG4B	missense_variant	exon	Missense
8	5:	140812-SNV	PLEKHG4B	synonymous_variant	exon	Other
9	5:	140849-SNV	PLEKHG4B	intron_variant	intron	Other
10	5:	143114-SNV	PLEKHG4B	intron_variant	intron	Other
11	5:	143219-SNV	PLEKHG4B	missense_variant	exon	Missense
12	5:	143239-SNV	PLEKHG4B	missense_variant	exon	Missense
13	5:	143249-SNV	PLEKHG4B	missense_variant	exon	Missense
14	5:	143257-SNV	PLEKHG4B	missense_variant	exon	Missense
15	5:	143571-SNV	PLEKHG4B	synonymous_variant	exon	Other

✓ バリエントリストでは、各バリエントのアノテーション情報を確認が可能

Count Variants per Gene

Variant counts per gene - Both Alt_Alt and Ref_Alt variants [20]

Unsort	1	2	3	4	5	
Map	Sample	PLEKHG4B	LRRC14B	CCDC127	SDHA	PDCD6
1	NA06984	0	0	0	4	0
2	NA06985	0	0	0	0	0
3	NA06986	0	0	0	0	0
4	NA06989	0	0	0	0	0
5	NA06994	0	0	0	3	0
6	NA07000	0	0	0	4	0
7	NA07037	0	0	0	4	0
8	NA07048	0	0	0	1	0
9	NA07051	0	0	0	1	0
10	NA07346	0	0	0	0	0
11	NA07347	0	0	0	1	0
12	NA07357	0	0	0	0	0
13	NA10847	0	0	0	0	0
14	NA10851	0	0	0	4	0
15	NA11829	0	0	0	4	0
16	NA11830	0	0	0	4	0

Variant counts per gene - Both Alt_Alt and Ref_Alt variants

Variant Binning by Frequency Source

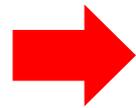
Variant Frequency Bins from 1kG Phase3 - Variant Frequencies 5a with Genotype Counts, GHI [30]

Unsort	1	2	3	
Map	Probe	Allele Frequencies Bin	Allele Frequencies	Alleles Present
1	5:140515-SNV	0	0.0027955272235	A/C
2	5:140568-SNV	0	0.000399361015297	A/G
3	5:140591-SNV	0	0.00299520767294	A/G
4	5:140714-SNV	0	0.00319488812238	C/G
5	5:140716-SNV	0	0.000998402596451	C/T
6	5:140781-SNV	0	?	G
7	5:140795-SNV	0	0.000998402596451	A/G
8	5:140812-SNV	0	0.000199680507649	C/T
9	5:140849-SNV	0	0.0027955272235	C/T
10	5:143114-SNV	0	0.000399361015297	C/T
11	5:143219-SNV	0	0.000199680507649	A/G
12	5:143239-SNV	0	0.000798722030595	A/G
13	5:143249-SNV	0	0.00299520767294	C/G
14	5:143257-SNV	0	?	C
15	5:143571-SNV	0	0.000399361015297	C/T
16	5:143626-SNV	0	?	G

Variant Frequency Bins from 1kG Phase3 - Variant Frequencies 5a with Genotype Counts, GHI

- ✓ CASTでは、遺伝子ごとバリエーション数のカウントテーブルシート、CMCでは、バリエーション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

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



SKAT-O at Specified Regions

Parameters for SKAT-O (Lee et. al, Am J Hum Genetics 2012) Optimized Sequence Kernel Association Testine

Uniform
 Madsen and Browning (one over square root of $MAF(1 - MAF)$)
 pdf(MAF: 1, 25) of the Beta distribution (normally recommended for rare variant studies)

Tests and Outputs

Generalized SKAT with rho =
 SKAT-O with standard grid of rho values

Regions

RefSeq Genes 105 Interim v1, NCBI Select Track

Perform one test per... gene transcript

Include nearby markers (distance in bp):

Correct for Additional Covariates
 Add Columns
Remove Selected
Clear List

Algorithm for Estimating Distributions

Algorithm Very-small-sample corrected (uses permutation testing to estimate kurtosis)
 Somewhat-small-sample corrected (kurtosis estimated analytically)
 Liu algorithm

Notes on Genetic Model and Imputation

NOTE: The additive model will always be used.

NOTE: Missing data will be imputed as homozygous major allele (wild type).

- ✓ 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]

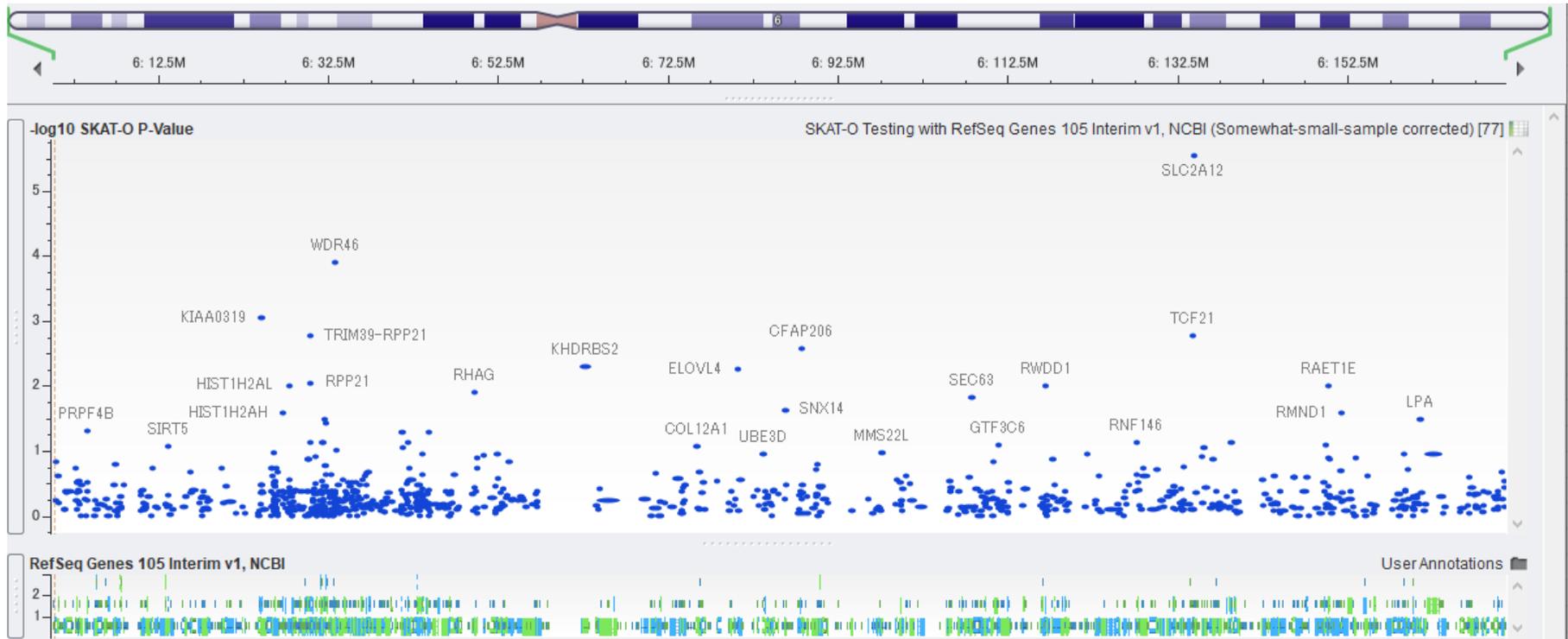
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

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	GMD5	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



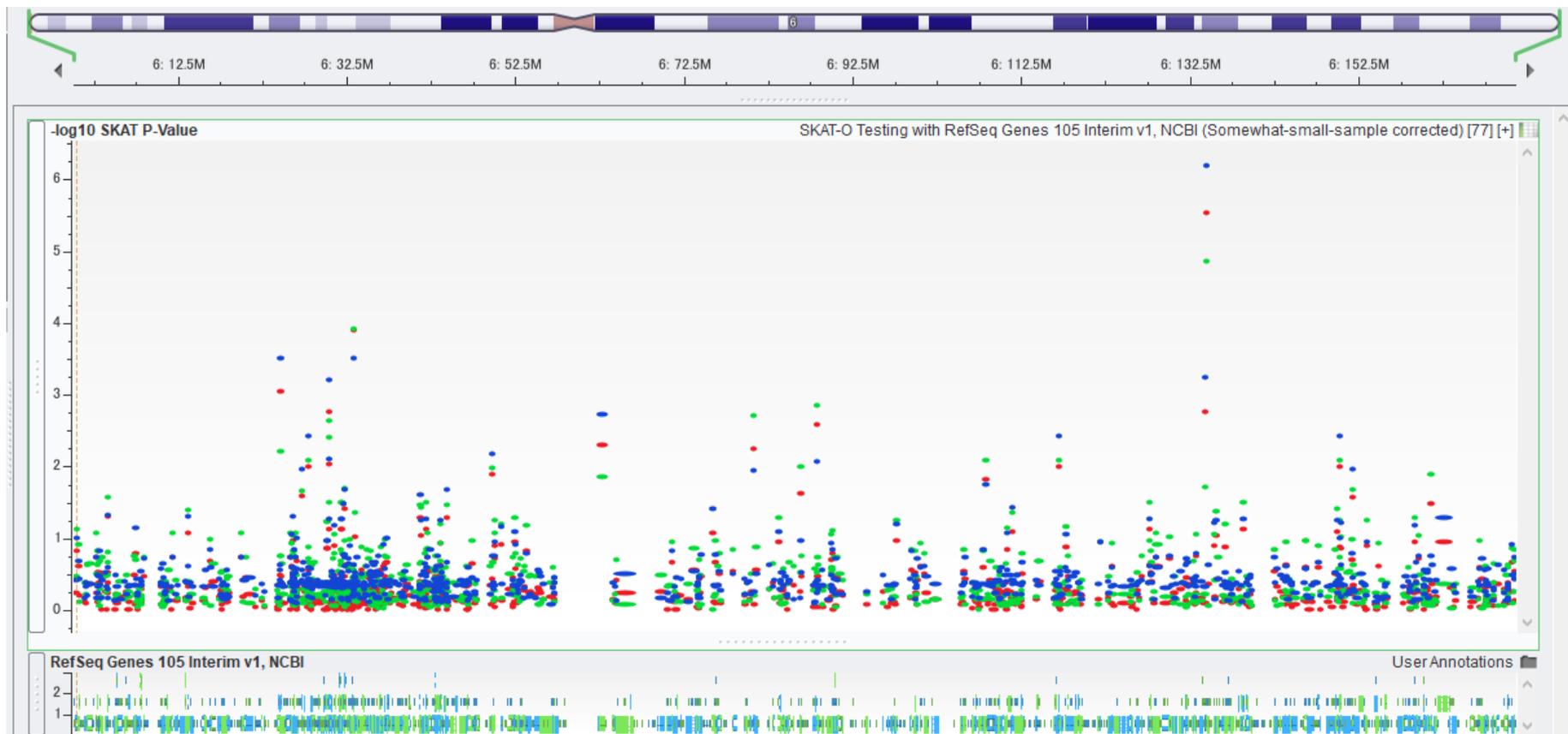
- ✓ ゲノムブラウザーには、各種アノテーションデータベースに加え、サンプルのバリエーションデータや、他の統計手法で計算した値などもプロットが可能
- ✓ プロットの表示色や属性によるラベル付けなどを行うことで、データの識別や解釈などを行いやすくなる

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：フィルジエン株式会社

TEL: 052-624-4388 (9:00～18 : 00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp