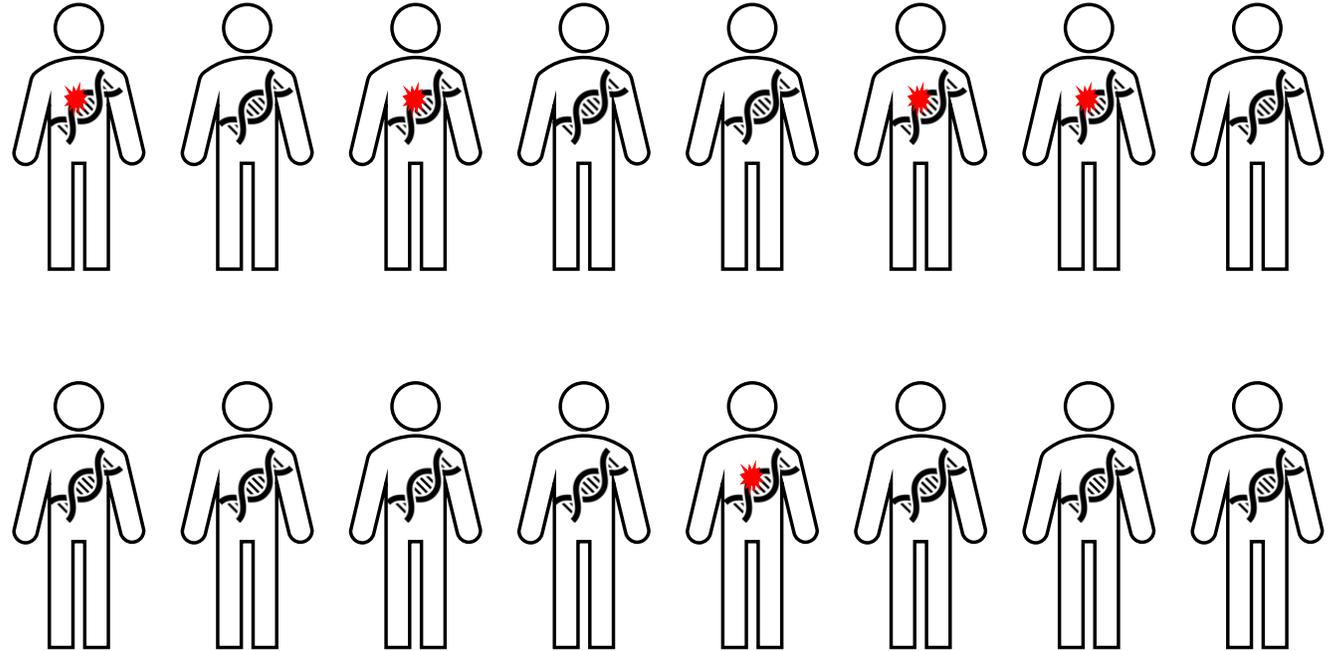
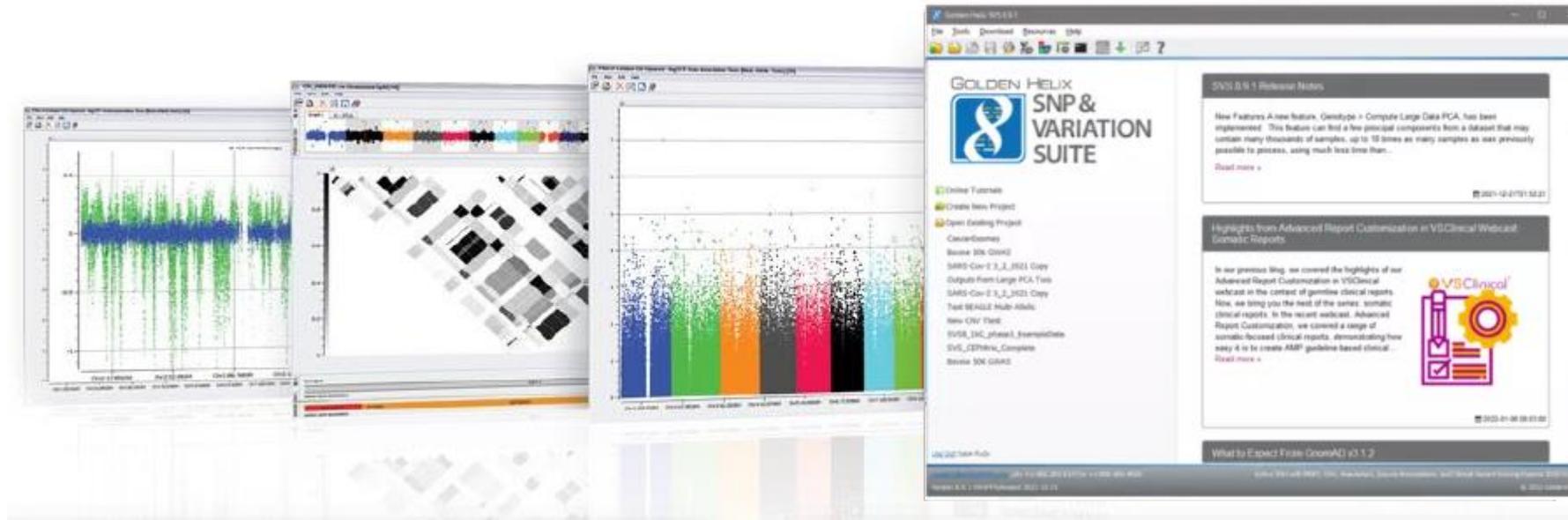


コモンバリアントとレアバリアントの疾患ゲノム解析

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

- 遺伝子バリエーションの中には、サンプル集団内に比較的高頻度に存在するコモンバリエーションと、ごくまれに存在するレアバリエーションがあり、それぞれ異なる解析アプローチが必要
- Golden Helix社SNP & Variation Suite (SVS) では、コモンバリエーションとレアバリエーションそれぞれに適した解析ツールが搭載されていて、データに合わせて使い分けることが可能





Core Features

- Powerful Data Management
- Rich Visualizations (GenomeBrowse)
- Robust Static
- Flexible

Applications

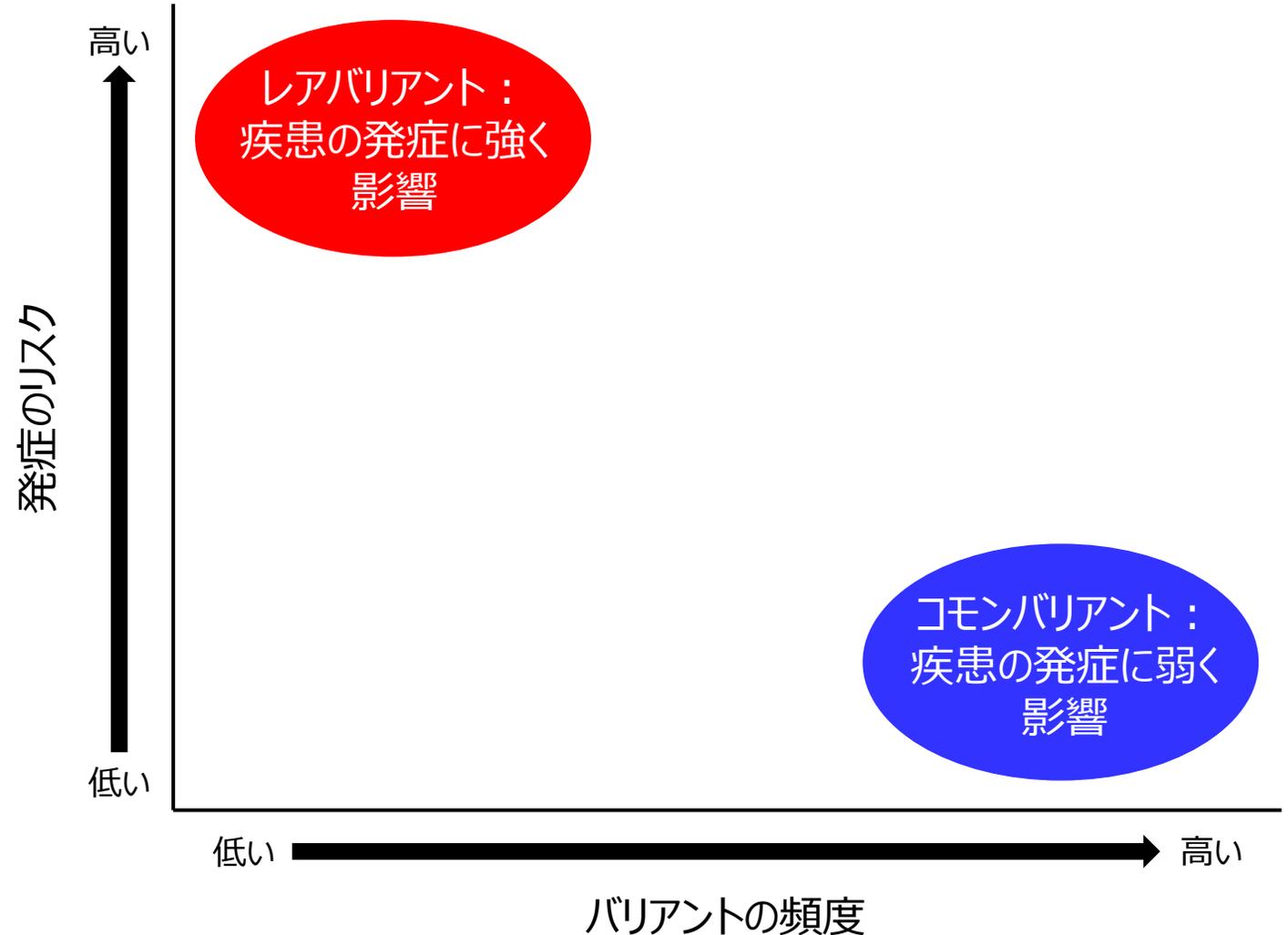
- Genotype Analysis
- Agrigenomics Analysis
- DNA/RNA Sequence Analysis
- CNV Analysis

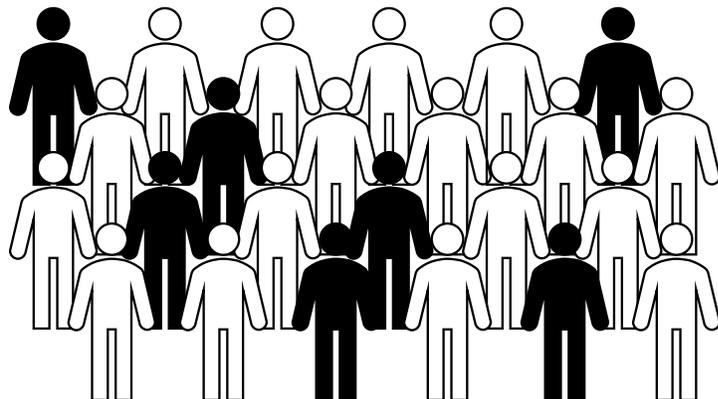
■ コモンバリエント

- ✓ 一般集団内に比較的高い頻度（1～5%以上）で存在し、多因子疾患などのありふれた疾患の発症に影響
- ✓ バリエント一つ一つは疾患発症に対する効果は弱く、複数のバリエントの効果が積み重なって発症すると考えられている
- ✓ SNPマイクロアレイによるジェノタイピングで同定することが多い

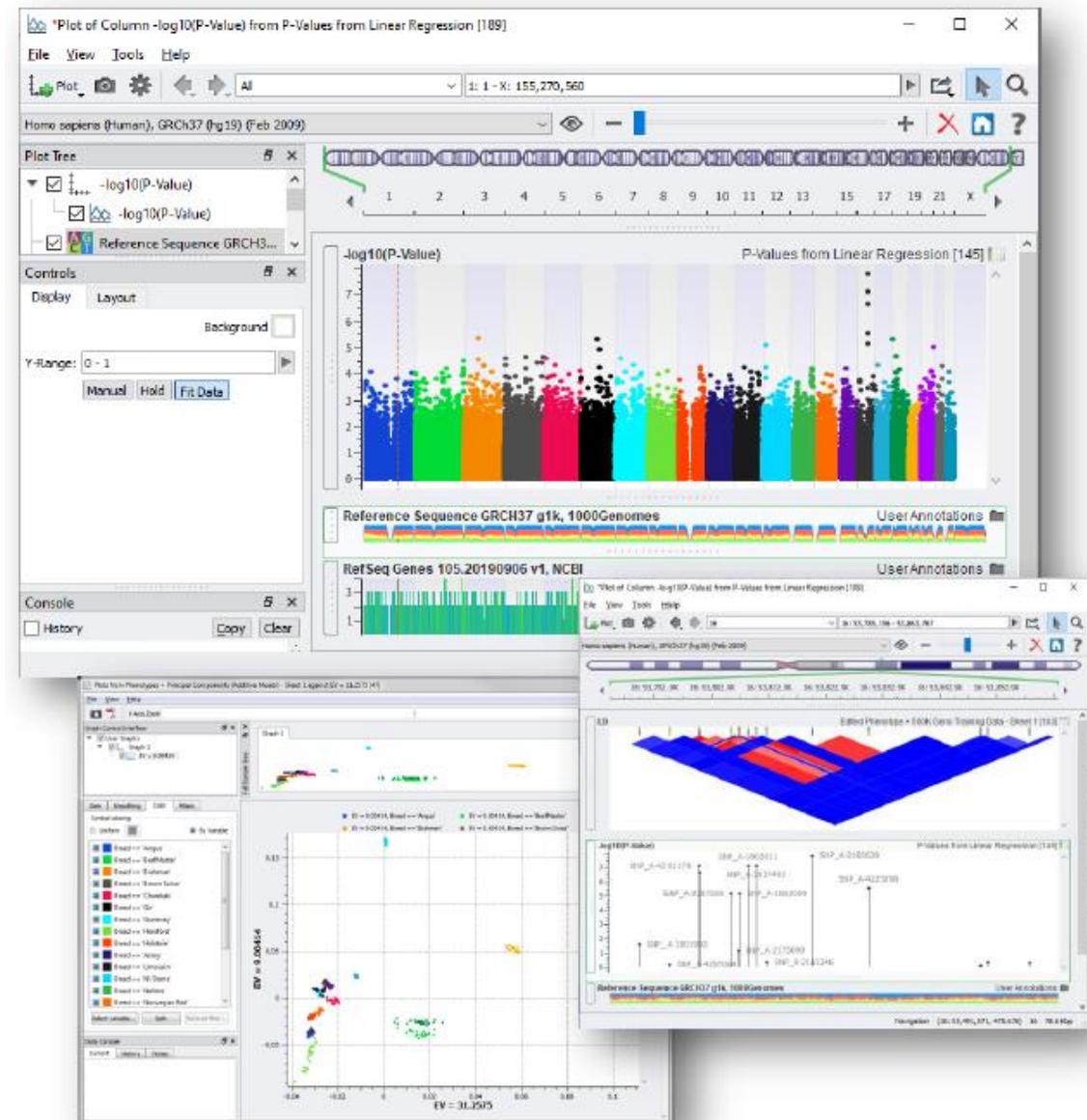
■ レアバリエント

- ✓ 一般集団内にまれに存在するバリエントで、単一遺伝子病などの遺伝性疾患の発症に影響
- ✓ 発症リスクが大きく、単独のバリエントのみで発症
- ✓ 同定には、次世代シーケンサーなどによるシーケンスが必要



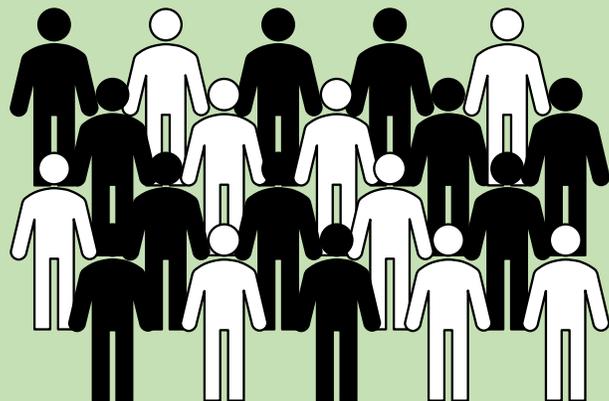


統計解析

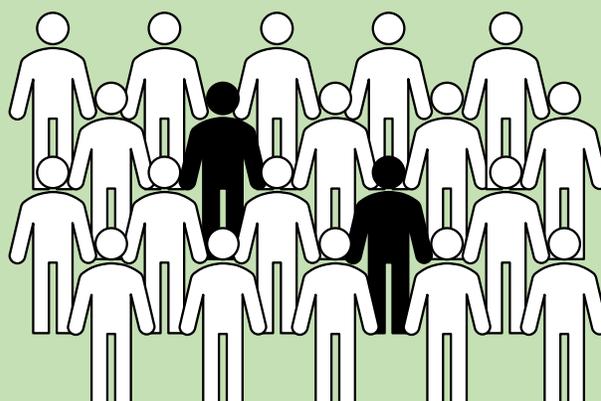


- ✓ 発症に関する影響が弱いバリエーションを検出するため、多数のサンプルデータを用いて統計処理を行う手法（ゲノムワイド関連解析；GWAS）が用いられている
- ✓ SNPごとに疾患リスクを評価したP値が算出されるため、疾患の発症と関連するSNPを網羅的に測定が可能

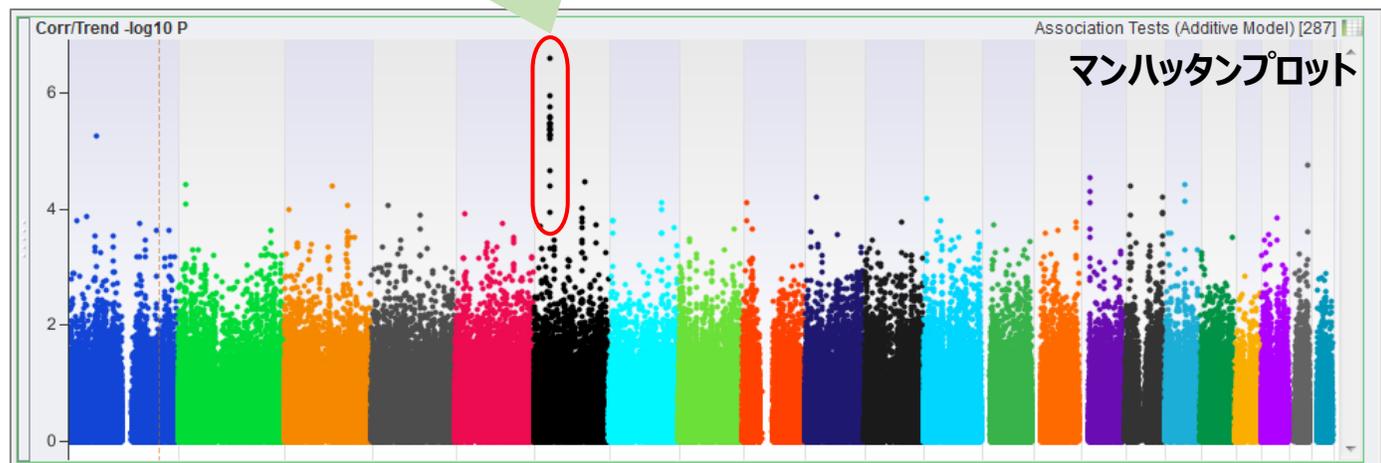
Case群



Control群



- ✓ サンプル集団を疾患ごとのCaseとControlに分け、それぞれのグループにおけるバリエーションの頻度の違いから、グループ特徴的なバリエーションを同定する手法
- ✓ 解析結果はバリエーションごとに疾患との関連を表したP値で出力され、そのP値をゲノム上にプロットしたマンハッタンプロットが可視化に用いられる
- ✓ 解析に使えるのはSNPなどのコモンバリエーションだけで、レアバリエーションは検出力が足りず使用不可能

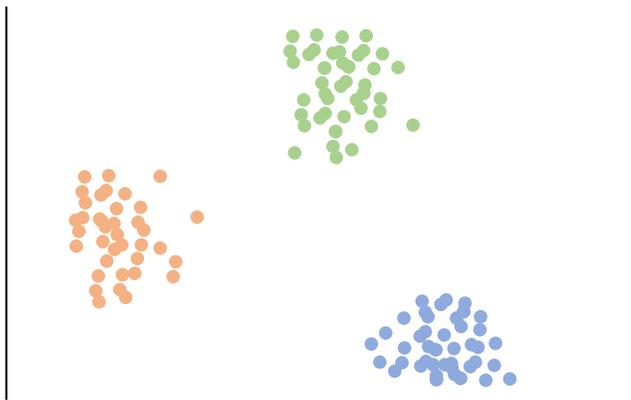


GWAS ワークフロー

✓ SNPデータ

	SNP1	SNP2	SNP3
Sample1	A/G	C/T	G/G
Sample2	A/G	C/T	G/G
Sample3	A/A	C/T	G/G
Sample4	A/G	C/T	G/G
Sample5	A/G	T/T	G/G

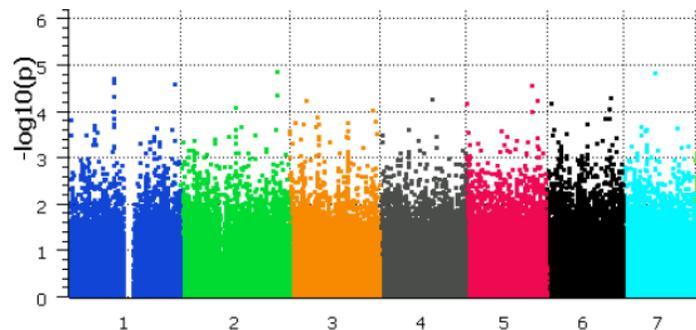
✓ クオリティコントロール



✓ GWASのワークフローは、解析用データの読み込みに始まり、データのクオリティコントロール、統計解析、解析結果の可視化、の4ステップに分かれている

✓ SNP & Variation Suiteでは、これらすべての処理をグラフィカルな画面で簡単に実行が可能

✓ 可視化



✓ 統計解析

	A/A	A/G	G/G
Case	96	165	217
Control	236	110	89

✓ SNPデータのテキストファイル

	A	B	C	D	E	F
1	Sample	SNP_A-1909444	SNP_A-4303947	SNP_A-1886933	SNP_A-2236359	SNP_A-2205441
2	GSM233256	T_T	A_A	G_G	T_T	C_C
3	GSM233258	C_T	A_A	G_G	T_T	C_C
4	GSM233260	C_C	?_?	A_A	T_T	C_C
5	GSM233262	C_T	A_G	A_G	T_T	C_C
6	GSM233264	T_T	A_A	G_G	T_T	C_C
7	GSM233266	T_T	A_A	G_G	T_T	C_C
8	GSM233268	C_T	A_G	A_G	T_T	C_C
9	GSM233270	T_T	A_A	G_G	T_T	C_C
10	GSM233272	T_T	A_A	G_G	T_T	C_C
11	GSM233274	T_T				
12	GSM233276	T_T				
13	GSM233278	T_T				
14	GSM233280	T_T				
15	GSM233282	T_T				
16	GSM233284	C_T				
17	GSM233286	T_T				
18	GSM233288	C_T				
19	GSM233290	C_C				
20	GSM233292	T_T				

✓ 多くの場合は、SNPと表現型それぞれのタブ区切りテキストファイルを作成し、SVSにインポートして使用

✓ 表現型データには、バイアス補正などに利用する共変量も含めることが可能

✓ 表現型データのテキストファイル

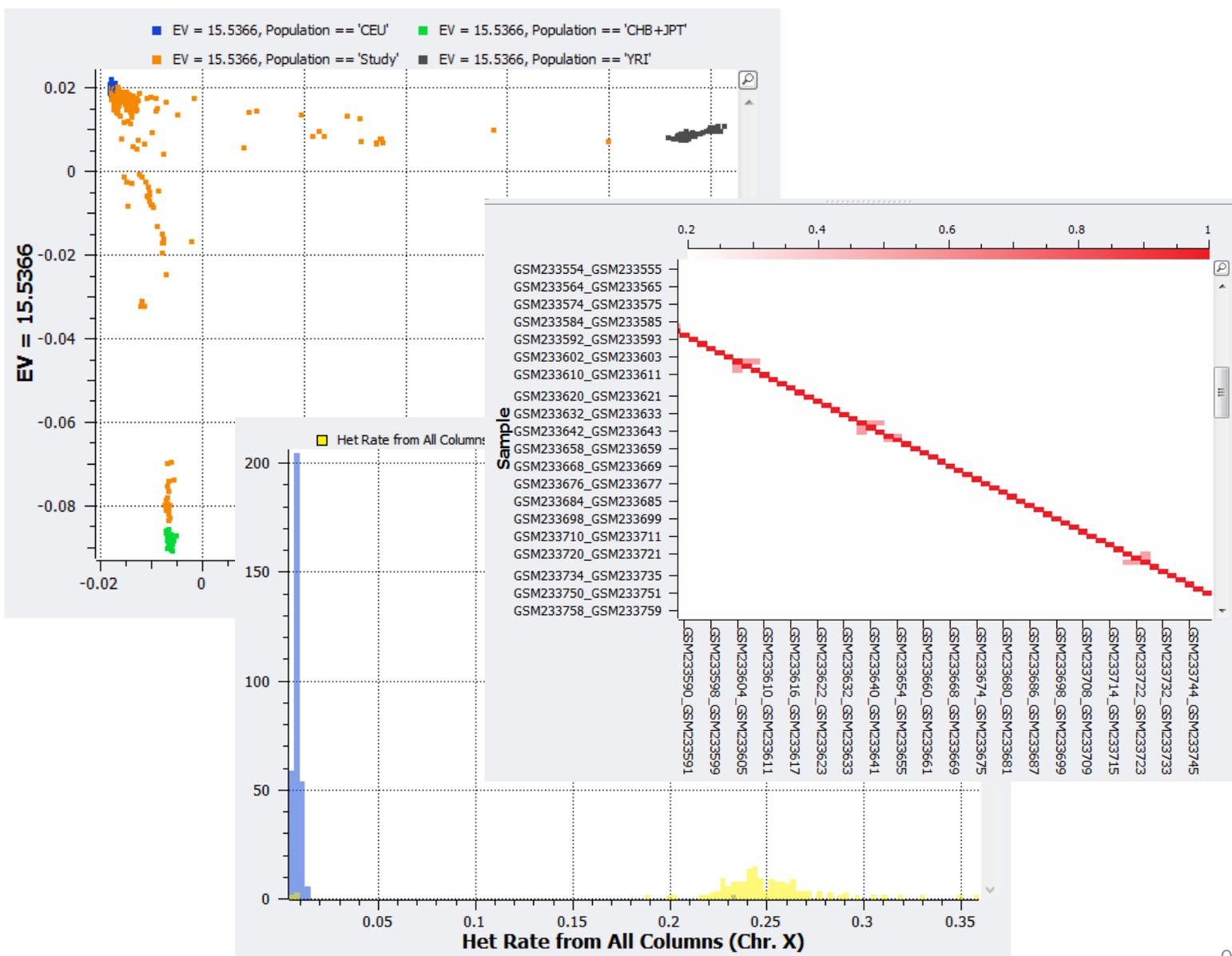
	A	B	C	D	E	F	G
1	Sample	Phenotype 1	Phenotype 2	Phenotype 3	SBP	BMI	Gender
2	GSM233256	1	8.96	1	128	29.9	M
3	GSM233258	0	12.18	0	137	32.8	M
4	GSM233260	1	11.45	0	138	31.6	F
5	GSM233262	0	11.01	1	117	27.4	M
6	GSM233264	0	9.25	0	140	29.7	M
7	GSM233266	1	6.85	0	113	27.7	M
8	GSM233268	1	8.77	1	121	27.7	F
9	GSM233270	1	11.65	0	125	29.2	M
10	GSM233272	0	13.67	1	120	29	M
11	GSM233274	0	10.61	0	149	34.1	M
12	GSM233276	0	10.82	1	88	20	M
13	GSM233278	1	9.62	0	131	31.2	M
14	GSM233280	1	9.02	1	124	26.2	M
15	GSM233282	0	7.94	0	142	31.3	M
16	GSM233284	0	11.73	1	111	22.8	M
17	GSM233286	0	12.81	1	114	26.1	F
18	GSM233288	0	9.46	1	107	23.5	M
19	GSM233290	1	8.55	0	146	37.4	M
20	GSM233292	1	13.28	0	122	27.6	M

- Sample QC
 - Call Rate
 - 性別不一致のチェック
- Marker QC
 - Call Rate
 - ハーディー・ワインベルグ平衡
 - マイナーアレル頻度
- Principle Component Analysis (PCA)
- Identity by Descent (IBD)

✓ QC項目例

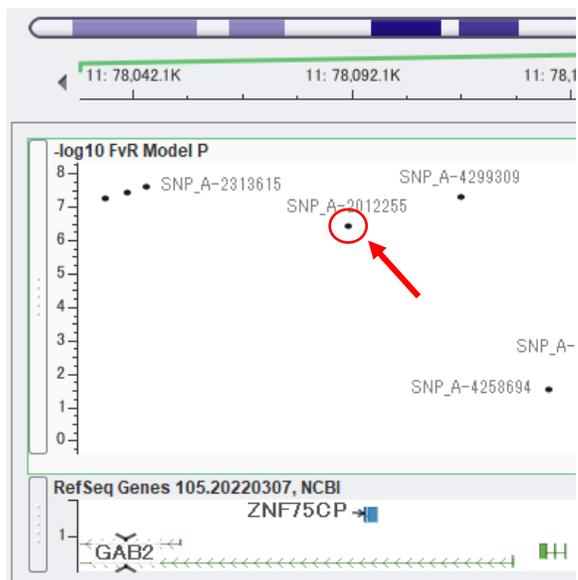
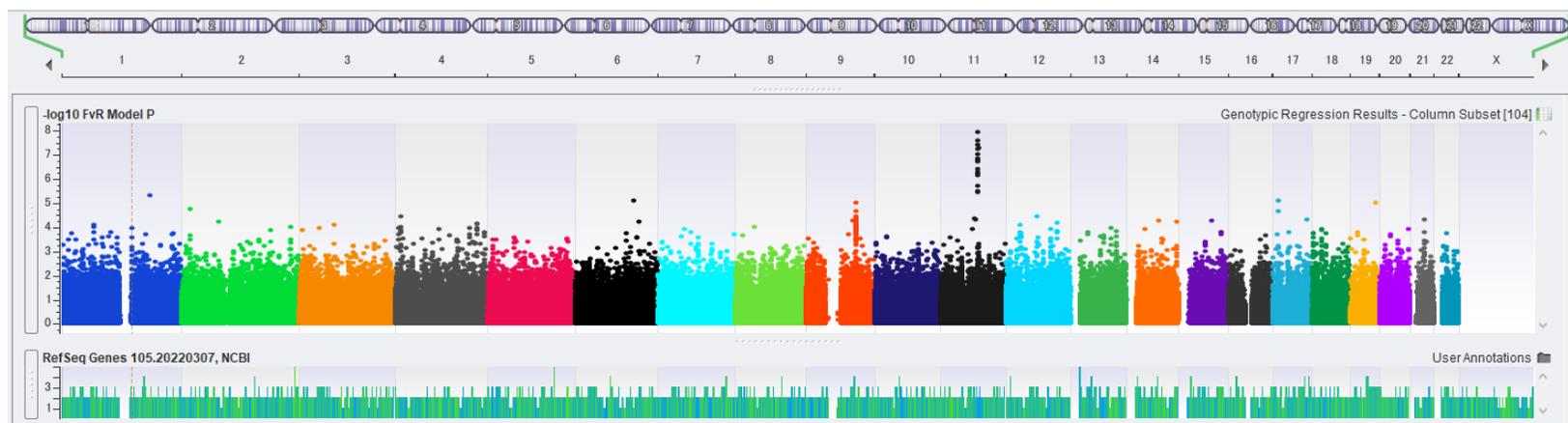
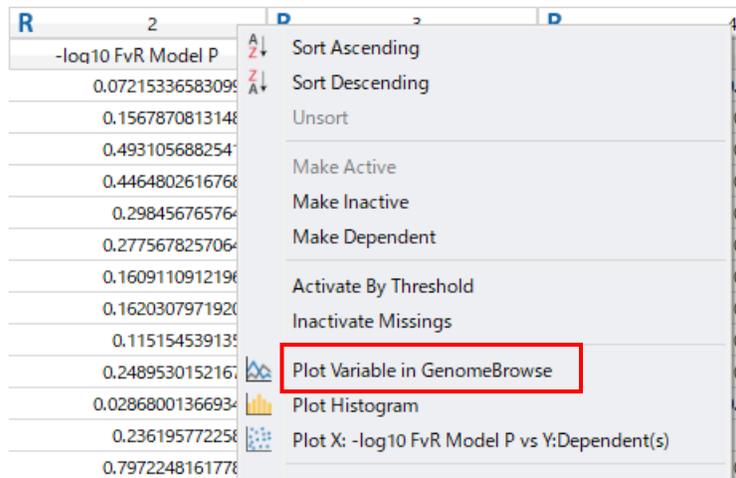


- ✓ 統計解析の実施前に、データのQCのステップを実行
- ✓ 各種のQC項目でデータのフィルタリングを実施したり、集計データのグラフ表示が可能



The image shows two overlapping software windows. The background window is titled "Genotype Association Tests" and shows options for "Case/Control dependent variable: Phenotype 1 (237 cases and 227 controls)" and "Additive model: (dd) -> (Dd) -> (DD)". It has tabs for "Association Test Parameters", "PCA Parameters", and "Overall Marker Statistics". The foreground window is titled "Genotypic Regression Analysis" and shows "Quantitative dependent variable: Phenotype 2 (464 active samples)" and "Linear Regression". It has tabs for "Regression Parameters", "Output Parameters", and "Genotypic Parameters". The "Regression Parameters" tab is active, showing "Selection Parameters" with options like "Regress on each of the 384255 genotypic columns" and "Regress on a moving window with parameters:". It also shows "Full Model Covariates" and "Reduced Model Covariates" lists with buttons to "Add Covariate", "Add Interaction", "Remove Selected", and "Clear List".

- ✓ QC後は、統計モデルを選択して統計解析を実行
- ✓ 遺伝モデルや統計手法、多重検定などのオプションを選択して解析を実行
- ✓ サンプルの遺伝的背景が異なる場合は、PCAや共変量補正によるバイアス補正を行う



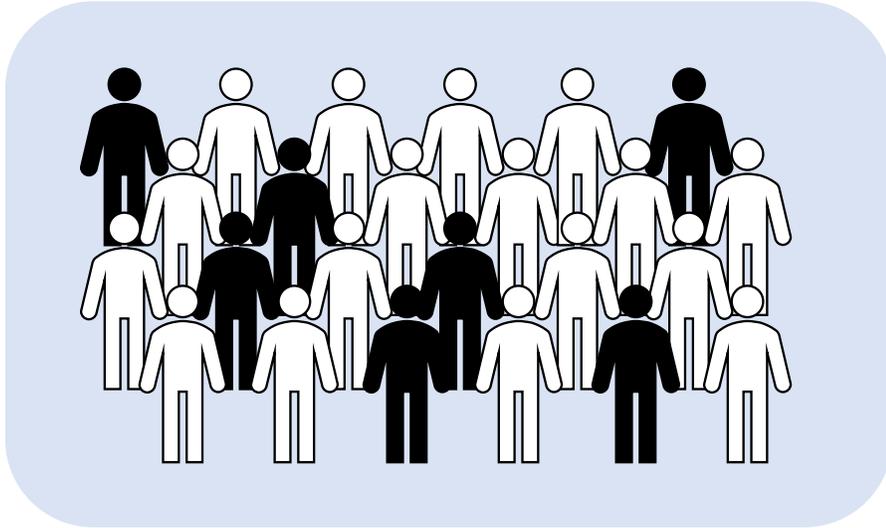
Chr11: 78,091,150 - 78,091,150

-log10 FvR Model P 6.39313312151234

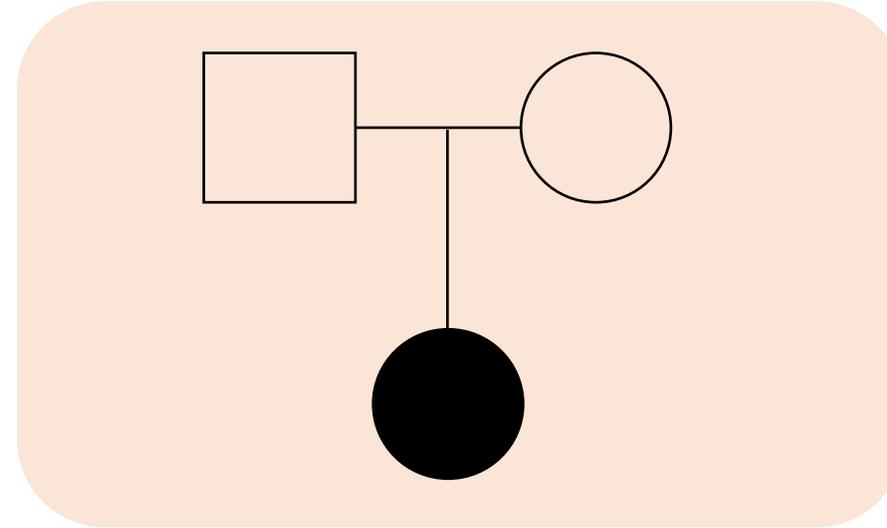
Extended Fields

Name	SNP_A-2012255
dbSNP RS ID	rs2373115
Associated Gene	?
Cytoband	q14.1
Reference Alleles A/B	[G/T]
Top Alleles	[C/A]
Bottom Alleles	[G/T]
Strand	-
Strand Versus dbSNP	same
Probe Count	12
FvR Model P-Value	4.04451898639024e-07
-log10 FvR Model P	6.39313312151234
expected FvR Model P	2.73256040910333e-05
expected -log10 FvR Model P	4.56343022826198

- ✓ 統計解析終了後は、SNPごとのP値のリストより、P値をゲノムブラウザーにプロットしたマンハッタンプロットを作図してデータの可視化を行う
- ✓ マンハッタンプロット上で、有意なSNPの詳細データを確認



関連解析

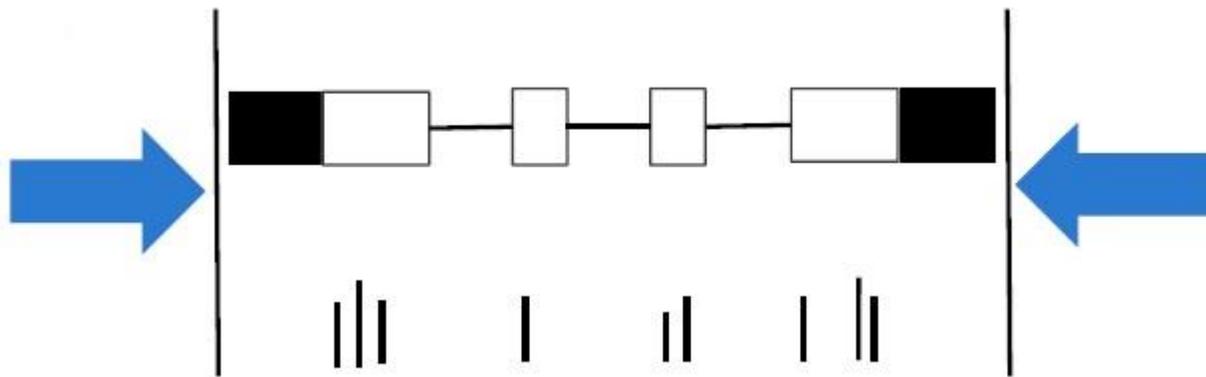


Trio解析

- ✓ レアバリエント解析では、コモンバリエントと同様に多数のサンプルを扱う関連解析と、家系情報と少数のサンプルを利用する Trio解析の2種類の手法が用いられる
- ✓ 基本的に、関連解析は多因子疾患に、 Trio解析は遺伝性疾患の研究に利用される



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



- ✓ 通常のGWASでは検出力が足りず、レアバリエントの解析には使用できないため、バリエントを遺伝子ごとに集約し、遺伝子単位で関連解析を行う

- ✓ バリエントの効果の方向に応じて、2種類の手法が存在

Burden Tests

- 遺伝子ごとにバリエントの効果を集計して検定を行う
- それぞれのバリエントの効果は、すべて同方向に作用することが前提となる

- バリエントの重み付けなし : **CAST, CMC**
- バリエントの重み付けあり : **KBAC**

Kernel Tests

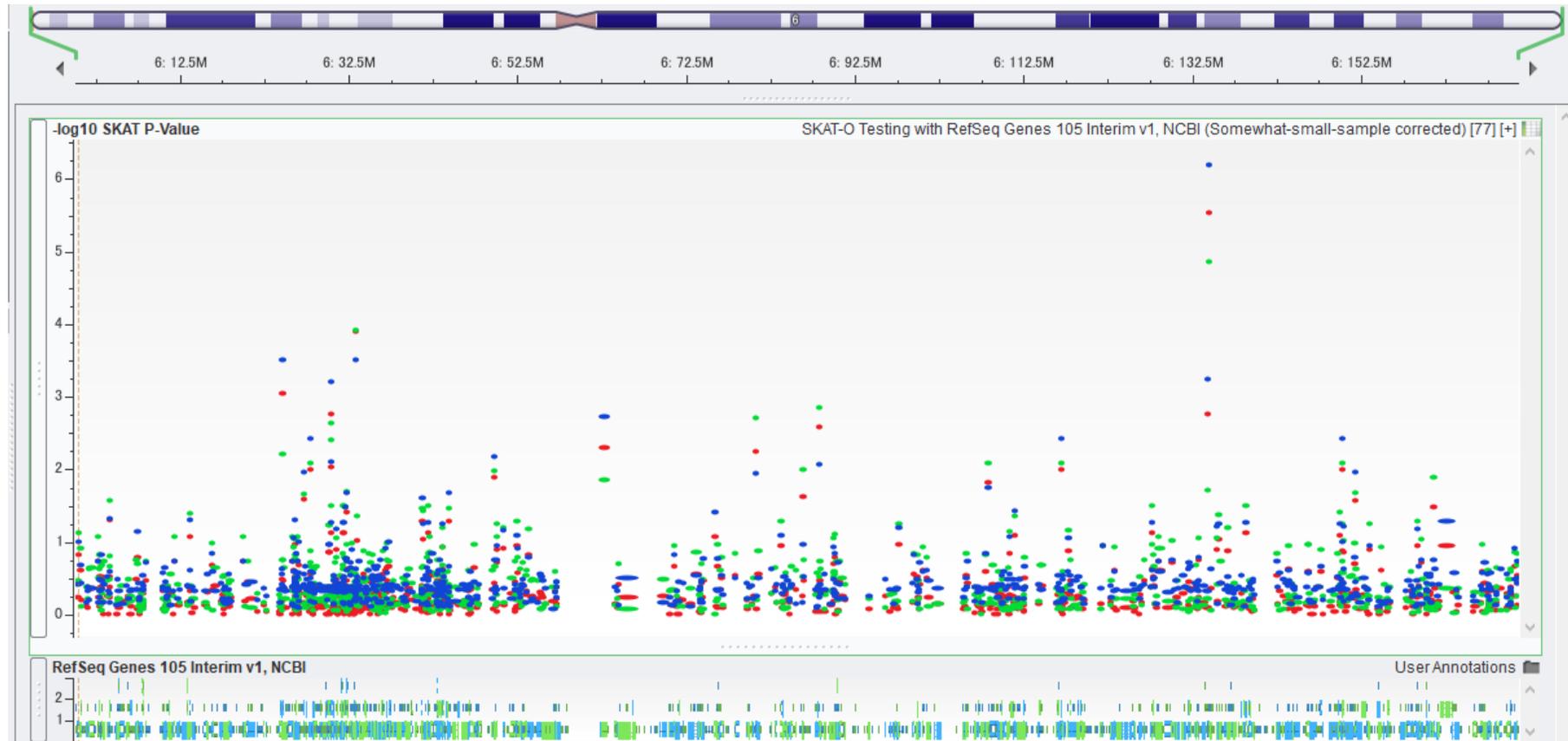
- 遺伝子ごとに、各バリエントの統計スコアの2乗を集計して検定を行う
- 各バリエントの効果は、様々な方向に作用する場合に用いられる

- **SKAT**
- **SKAT-O** (Burden Testsとの組み合わせ)

赤 : SKAT-O

青 : Burden Test

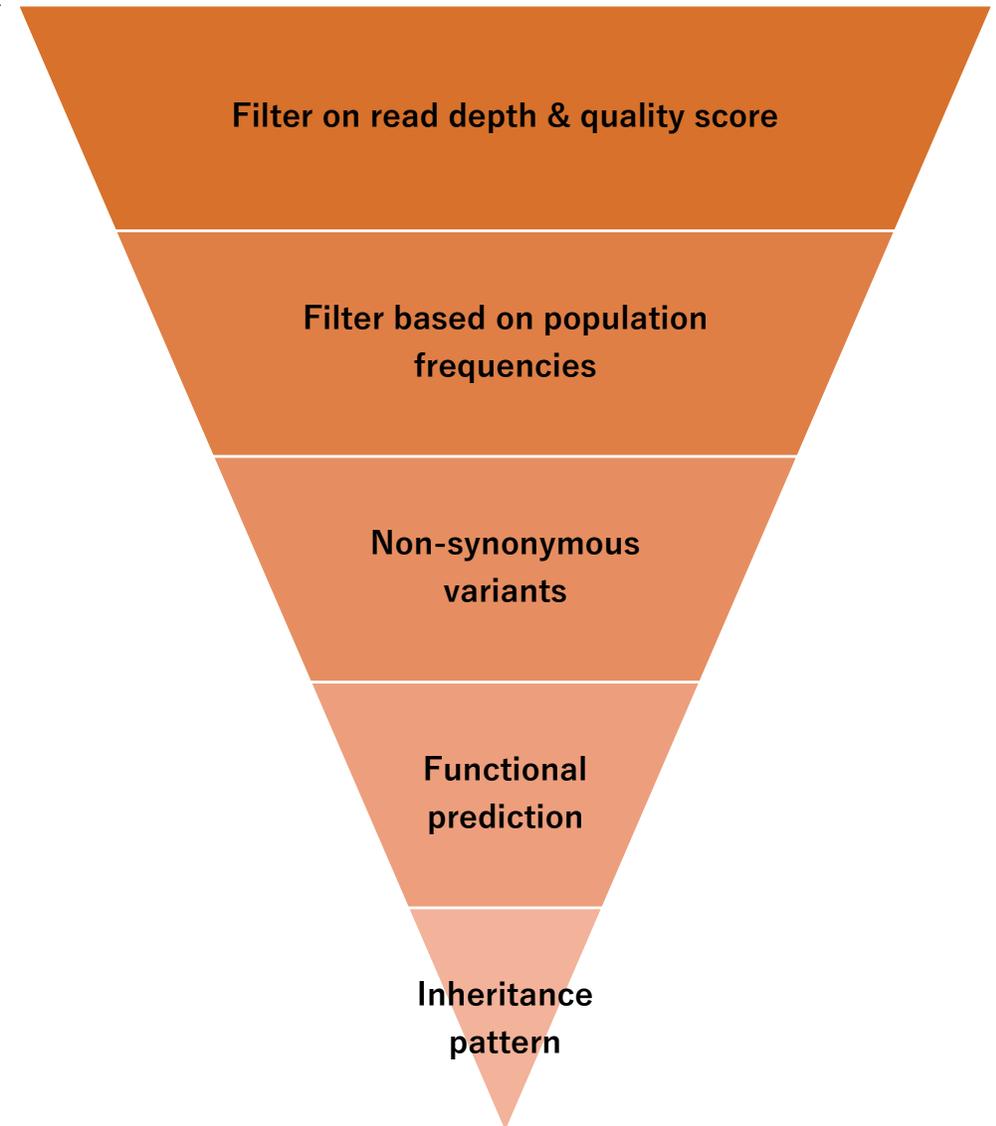
緑 : SKAT



- ✓ 一般的にバリアントの効果は様々な方向に作用するため、遺伝子ごとにBurden Testの方が検出力が高い場合と、SKATの方が高い場合がある
- ✓ SKAT-OはBurden TestとSKATを組み合わせたもので、両方のバランスを最適な値に調整して計算が可能のため、バリアントの効果が多様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

VCFファイル →

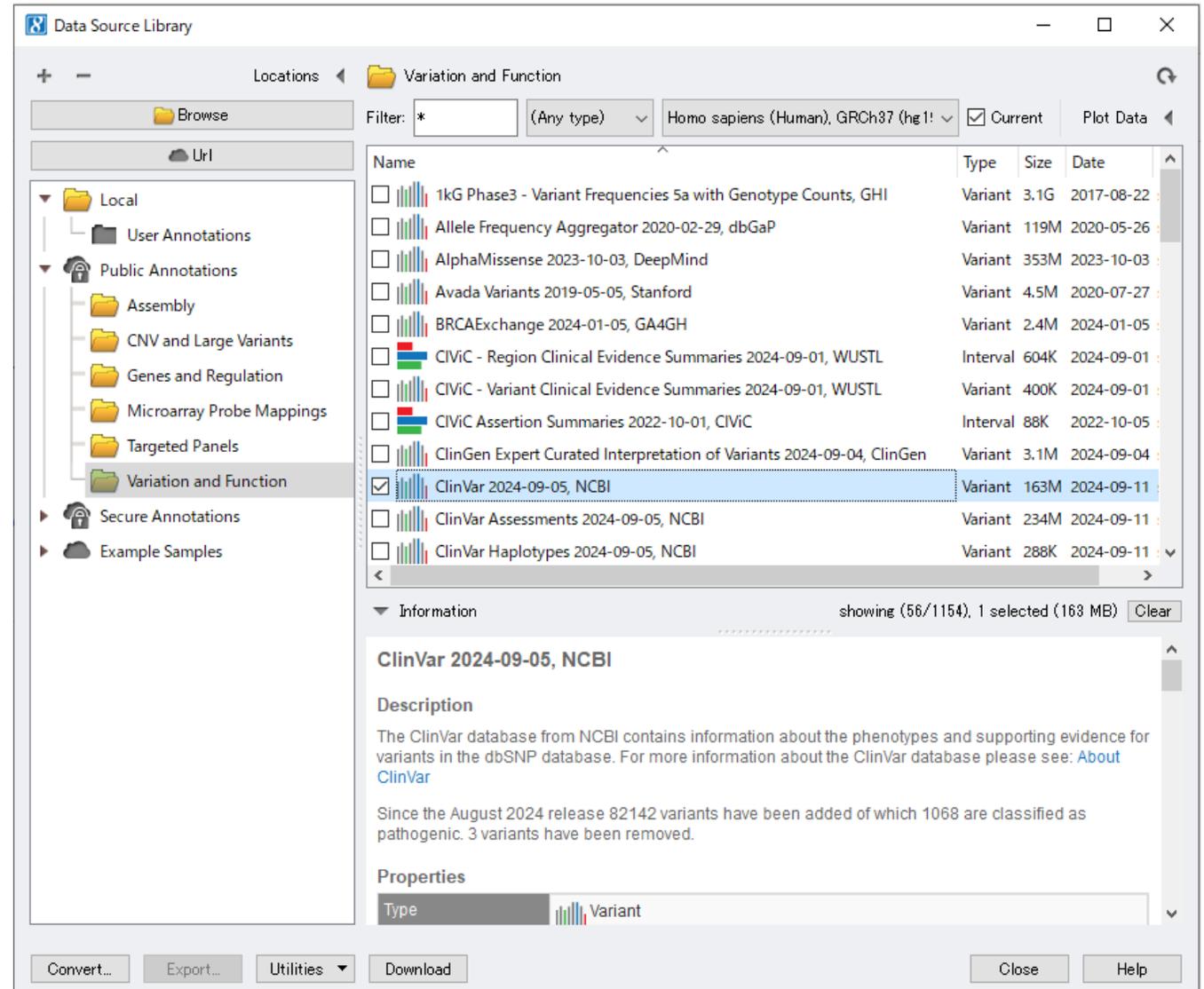
- ✓ トリオ解析では、統計的手法を用いず、配列データを直接比較して病的バリエントを検出する
- ✓ 解析には次世代シーケンサーの2次解析パイプラインより出力されたバリエントをまとめたファイル（VCFファイル）を用いる
- ✓ VCFファイルには多数のバリエントが含まれているため、様々なアノテーションリソースを活用し、バリエントの絞り込みを行う必要がある



✓ アノテーションリソースは、様々なデータベースより最新のデータをアノテーションリソースライブラリーよりダウンロード

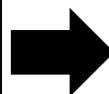
- RefSeq Genes
- dbSNP
- ClinVar
- CIViC
- ICGC / TCGA
- MSK Impact
- Orphanet
- BRCA Exchange
- PMKB
- dbNSFP
- REVEL
- AlphaMissense
- 1000 Genomes
- NHLBI 6500 Exomes
- ExAC Variant
- gnomAD Exomes/Genomes
- GenomeAsia 100K
- 各種遺伝子パネルのターゲットデータ
...など

✓ 外部VCFファイルやBEDファイルなども、アノテーションデータに変換して使用することが可能



バリエーションフィルタリング

- gnomAD Exomes Variant Frequencies 4.0, BROAD
- dbNSFP Functional Predictions 3.0, GHI
- RefSeq Genes 105.20220307, NCBI



① Select Tracks
② Annotation and Filter Options
③ Review

Select options for annotating and filtering for RefSeq Genes 105.20220307, NCBI

① Select Tracks
② Annotation and Filter Options
③ Review

Select options for annotating and filtering for RefSeq Genes 105.20220307, NCBI

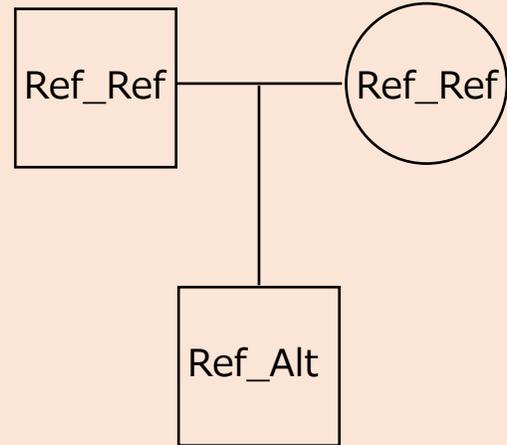
Annotate Transcript Options

- Only annotate verified mRNA transcripts
- Amino Acid Notation: 3 Letter 1 Letter
- Splice Site Boundaries
 - Splice Donor Distance: Splice Region Exonic Distance:
 - Splice Acceptor Distance: Splice Region Intronic Distance:
- Splice Site Predictions from MaxEntScan, GeneSplicer, NNSplice, and PWM (Requires Clinical Variant Scoring Feature)
 - Include Splice Site Predictions Include Novel Splice Site Predictions
- Output Options
 - Variant Report
 - Variant Interactions Report Include Auxillary Transform Fields
 - Include Intergenic Variants in Output
 - Report Upstream and Downstream Genes (within bp window):
- Optional Filters:
 - Effect (Combined) is one of LoF, Missense
 - Filter on: Effect (Combined) *The highest priority of the effect annotations found among...*
 - Is one of
 - LoF
 - Missense**
 - Other
 - Unknown
 - Invalid
 - Missing

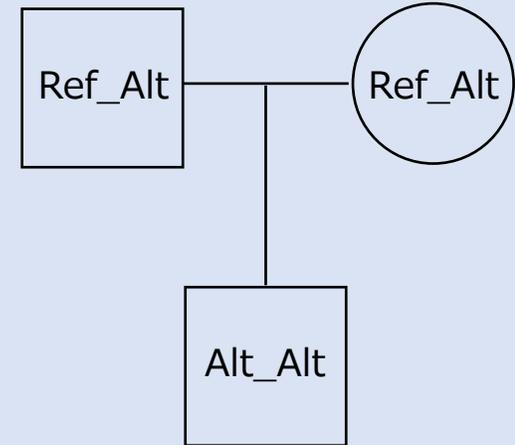
Help < Back Next > Cancel

✓ データベースを選択して、フィルタリングの条件を指定

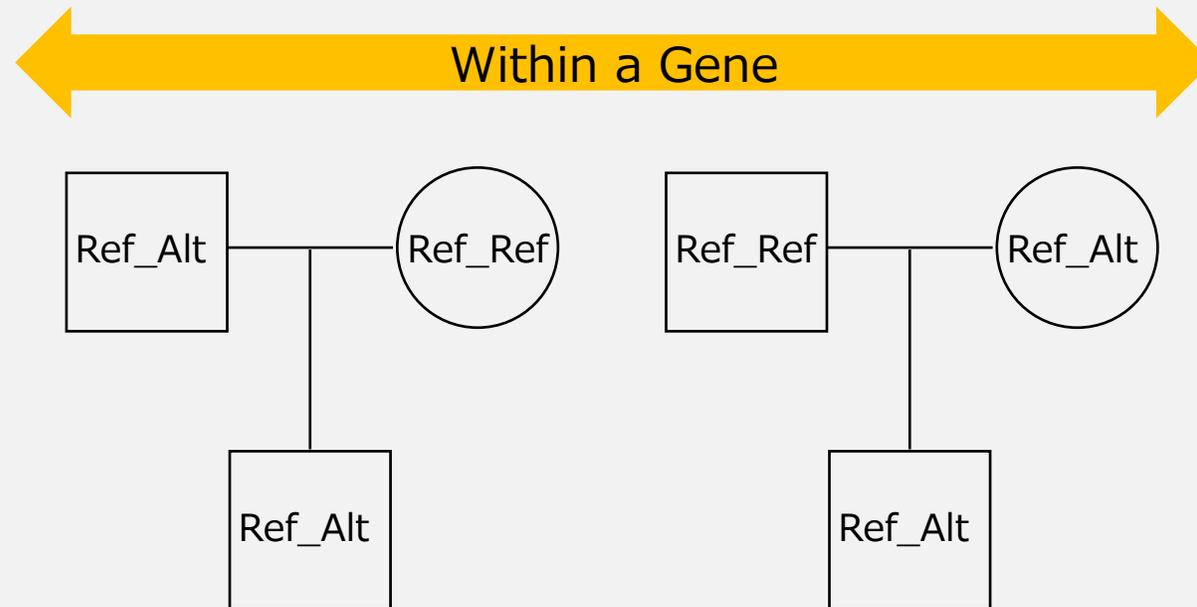
de Novo Mutation



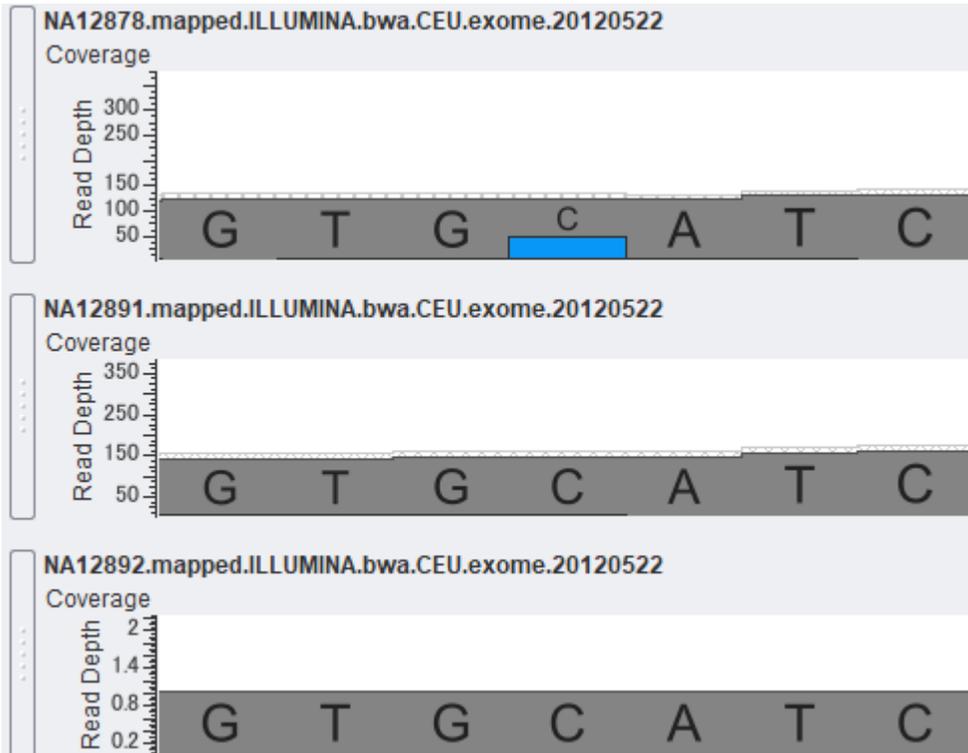
Rare Recessive Mutation



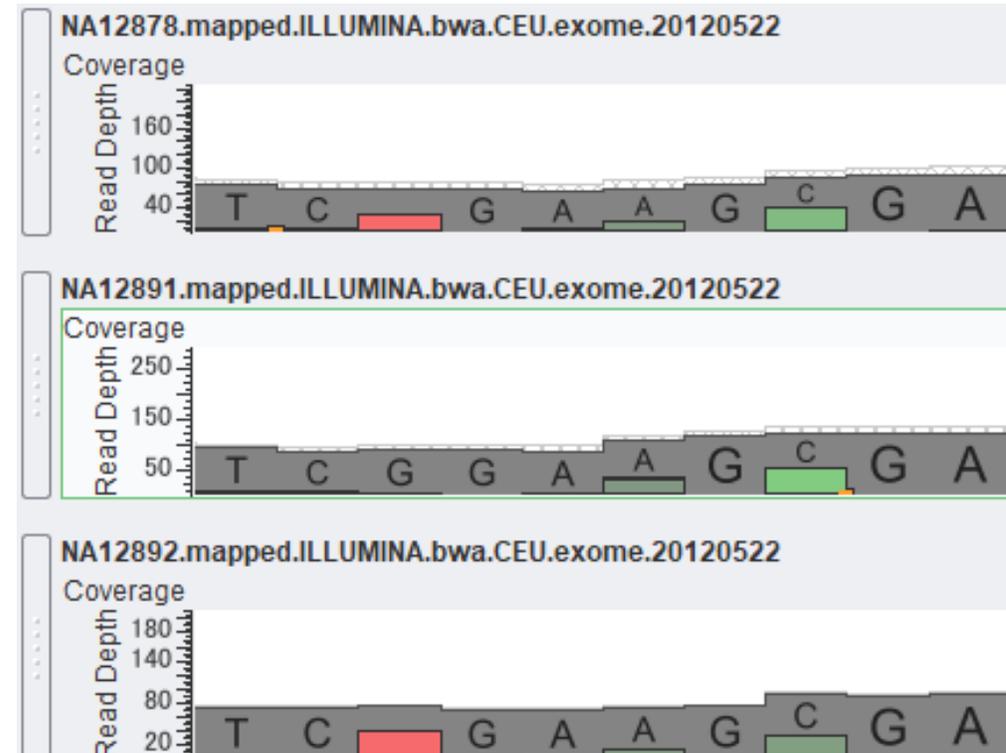
Compound Heterozygous Mutation



de Novo Mutation



Compound Heterozygous Mutation



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: support@filgen.jp