

# SNPタイピングデータを用いた ゲノム育種

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

- 人口の増大に伴い、食糧の増産や品質向上のために、作物と家畜のゲノム情報を用いて、効率的な育種改良を行うことが急務となっている。
- そのためには、これら作物や家畜の遺伝的な多様性を理解し、収量や味などの食料としての特徴と関連する遺伝子を調べる必要がある。
- Golden Helix社SNP & Variation Suite (SVS)ソフトウェアでは、SNPジェノタイピングデータによるゲノム情報を用いた、育種に役立つ解析ツールが搭載されている。



## 1. ゲノムワイド関連解析 (GWAS)

### 表現型と関連するSNPの同定

- ゲノム全域を網羅したSNPマイクロアレイや、次世代シーケンサー解析で取得したSNPデータを用いて、多数のサンプルデータを比較し、表現型と関連するSNPの同定を行う。
- ケース/コントロールのようなバイナリデータの他に、量的形質 (Quantitative trait) も表現型データとして扱うことができる。
- 近交系サンプルの血縁関係による偏りを除外するために、線形混合モデルが使用される。

## 2. Genomic Prediction

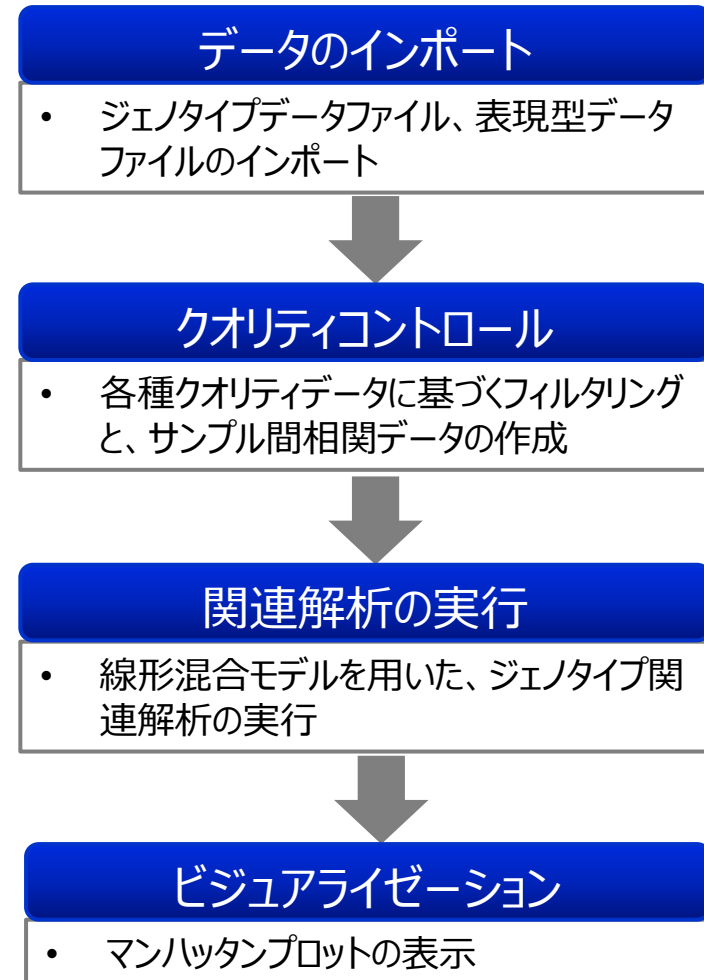
### サンプルごとの形質やゲノム育種価の予測

- SNPジェノタイプデータを用いた、個体の育種価の予測。
- 表現型を予測するモデルを作成し、表現型が未知のサンプルデータをモデルに当てはめ、そのサンプルの表現型を推測することができる。
- 予測した育種価データに基づく個体の選抜に応用が可能。
- GBLUP法、Bayes CやBayes C-pi法などを使用する。

# 1. ゲノムワイド関連解析 (GWAS)

## 使用するジェノタイプデータ:

- 生物種 : ウシ (Bos taurus)
- サンプル数 : 472例
- 解析プラットフォーム :  
Illumina Bovine BeadChip
- SNPマーカー数 : 52,890個



Import Download Resources Window

Text

Third Party

PED/TPED/BED

Golden Helix DSF

Golden Helix Legacy GHD

Public Data

Affymetrix

Illumina

Agilent Files

NimbleGen Data Summary Files

Family Pedigree

Import VCFs and Variant Files

Import Complete Genomics Var Files

Import Impute2 GWAS Files

HapMap

MACH Output

RNA-Seq Tabularized Quantification

## 表現型データ

Phenotypes [4]

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

All: 472 x 9  
Active: 472 x 9

Unsort	Map	Sample Label	B 5 Phenotype1	B 6 Phen2	R 7 Phen3	R 8 Phen4	R 9 Phen5
1		WG0099889-DNAD04_ANG000027	0	1	-0.0358	16.6761427265423	24.8389532585512
2		WG0099889-DNAA02_ANG000008	0	1	0.0362	23.3265398984327	21.7506929572886
3		WG0099889-DNAA03_ANG000016	0	1	0.0061	19.0803796478002	17.3037717691138
4		WG0099889-DNAB03_ANG000017	1	1	-0.0382	19.9554299755303	26.6011234866098
5		WG0099889-DNAB04_ANG000025	1	1	-0.0334	27.6744146632411	23.4778765840422
6		WG0099889-DNAC02_ANG000010	1	1	-0.0278	22.5351447396607	27.0589945698839
7		WG0099889-DNAC03_ANG000018	0	0	0.0884	15.5430766061975	15.7245152305321
8		WG0099889-DNAC04_ANG000026	0	0	0.0432	21.7196917403443	6.34720842916488
9		WG0099889-DNAD02_ANG000011	0	1	0.0217	18.8330050819465	23.8891657707582
10		WG0099889-DNAD03_ANG000019	0	1	-0.0785	20.7770585758938	16.2528640600238
11		WG0099889-DNAE01_ANG000004	0	1	-0.0504	18.6132833636024	25.5057310813833
12		WG0099889-DNAE02_ANG000012	0	1	0.0438	20.6256893411594	29.8261600049517

Phenotypes

## SNPジェノタイプデータ

Bovine HapMap Genotypes [7]

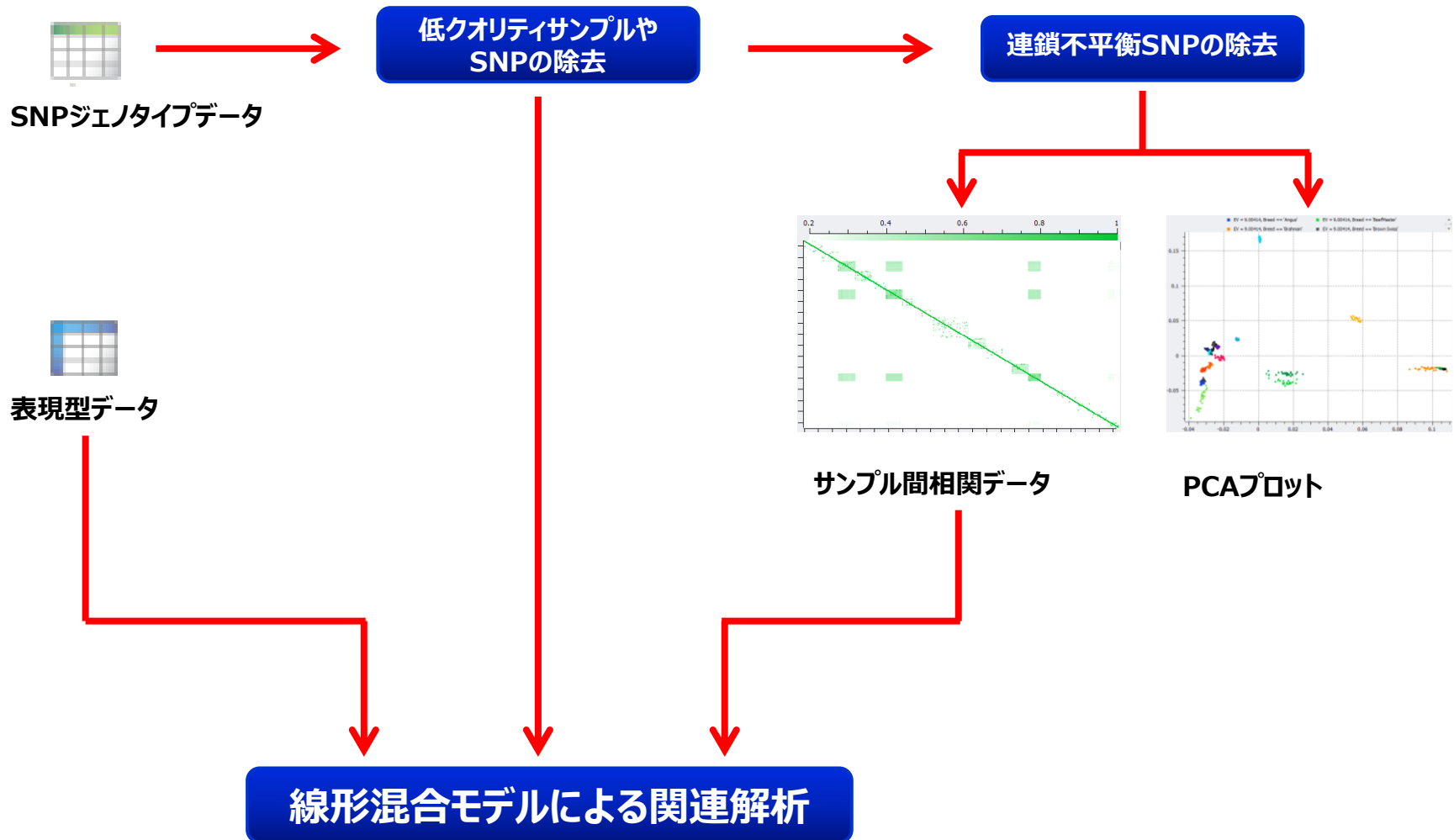
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

All: 472 x 52,890  
Active: 472 x 52,890

Unsort	Map	SampleID	G 1 Hapmap43437-BTA-101873	G 2 ARS-BFGL-NGS-16466	G 3 ARS-BFGL-NGS-19289	G 4 ARS-BFGL-NGS-105096
	Chromosome		1	1	1	1
	Position		135098	267940	305793	353745
	Top Strand		[A/G]	[A/G]	[A/G]	[A/G]
	Forward Strand		[A/G]	[T/C]	[A/G]	[T/C]
	Design Strand		[T/C]	[A/G]	[T/C]	[T/C]
1		WG0099889-DNAD04_ANG000027	G_G	C_C	A_A	T_T
2		WG0099889-DNAA02_ANG000008	G_G	C_T	A_A	C_T
3		WG0099889-DNAA03_ANG000016	G_G	C_T	A_A	C_T
4		WG0099889-DNAB03_ANG000017	G_G	C_C	A_A	T_T
5		WG0099889-DNAB04_ANG000025	G_G	C_C	A_A	T_T
6		WG0099889-DNAC02_ANG000010	G_G	C_C	A_A	T_T
7		WG0099889-DNAC03_ANG000018	G_G	C_C	A_A	T_T

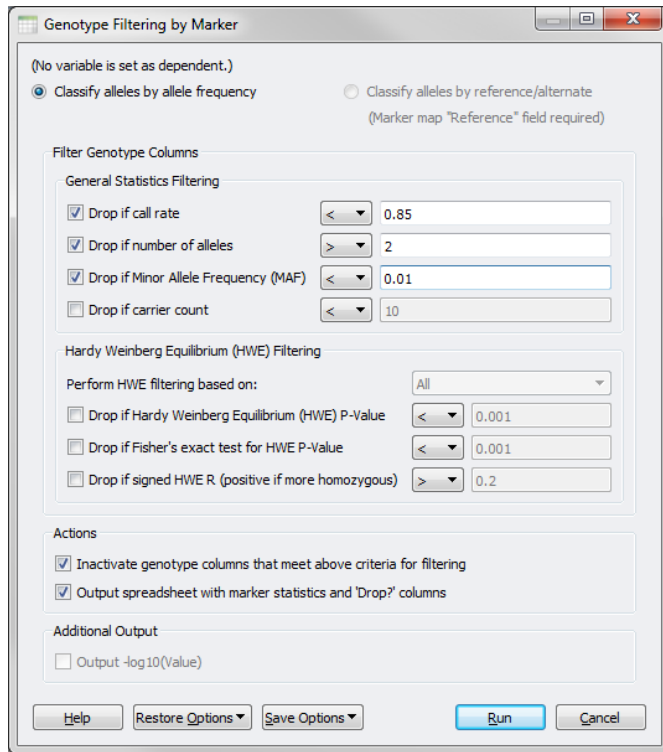
Bovine HapMap Genotypes

Bovine HapMap Genotypes - Sheet 2



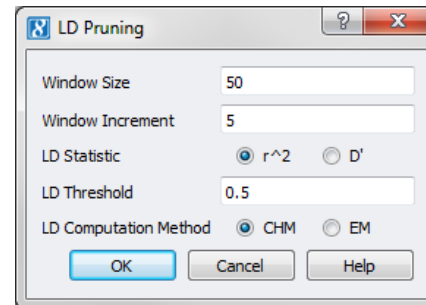
## 低クオリティ、低頻度SNPなどの除去

- Call Rate - 検出されたSNPの割合
- Number of allele - 検出されたアレル数
- Alternate allele frequency - 変異アレルの頻度



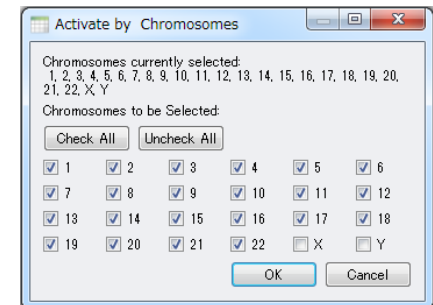
## 連鎖不平衡SNPの除去

- LD Pruning



## 常染色体以外のSNPの除去

- Activate by Chromosomes





Pruned SNP Subset [21]

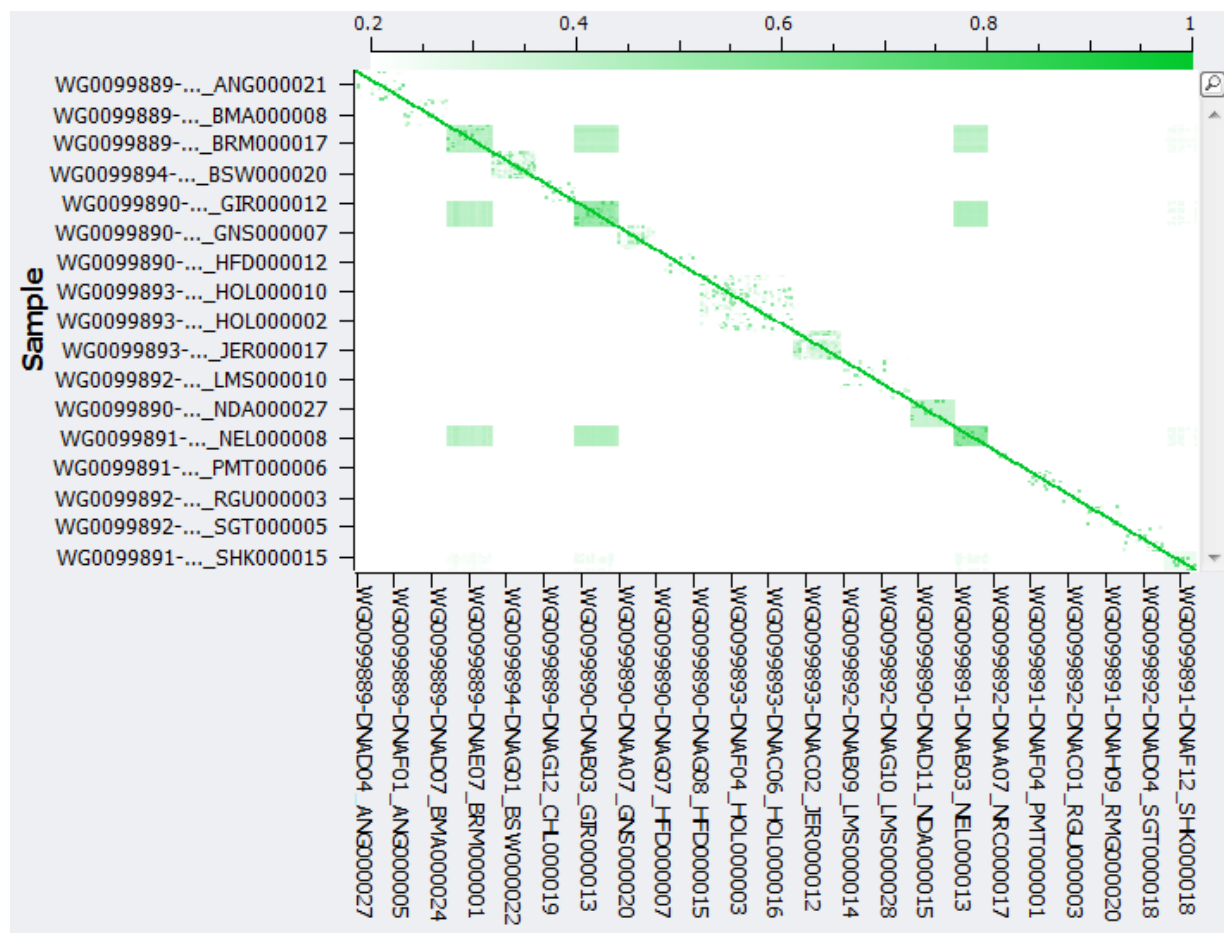
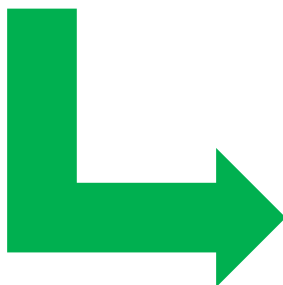
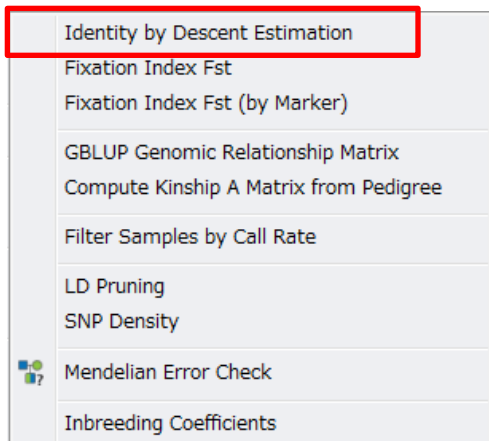
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

All: 468 x 43,589  
Active: 468 x 43,589

Unsort		G 1	G 2	G 3	G 4
Map	SampleID	Hapmap43437-BTA-101873	ARS-BFGL-NGS-16466	ARS-BFGL-NGS-105096	Hapmap34944-BES1_Contig627_1906
	Chromosome	1	1	1	1
	Position	135098	267940	353745	393248
	Top Strand	[A/G]	[A/G]	[A/G]	[A/C]
	Forward Strand	[A/G]	[T/C]	[T/C]	[A/C]
	Design Strand	[T/C]	[A/G]	[T/C]	[A/C]
1	WG0099889-DNAD04_ANG000027	G_G	C_C	T_T	C_C
2	WG0099889-DNAA02_ANG000008	G_G	C_T	C_T	C_C
3	WG0099889-DNAA03_ANG000016	G_G	C_T	C_T	A_C
4	WG0099889-DNAB03_ANG000017	G_G	C_C	T_T	C_C
5	WG0099889-DNAB04_ANG000025	G_G	C_C	T_T	C_C
6	WG0099889-DNAC02_ANG000010	G_G	C_C	T_T	C_C
7	WG0099889-DNAC03_ANG000018	G_G	C_C	T_T	C_C
8	WG0099889-DNAC04_ANG000026	G_G	C_C	T_T	C_C
9	WG0099889-DNAD02_ANG000011	G_G	C_C	T_T	C_C
10	WG0099889-DNAD03_ANG000019	G_G	C_T	C_T	A_C
11	WG0099889-DNAE01_ANG000004	G_G	C_T	C_T	A_C
12	WG0099889-DNAE02_ANG000012	A_G	C_C	C_T	C_C

Pruned SNP Subset

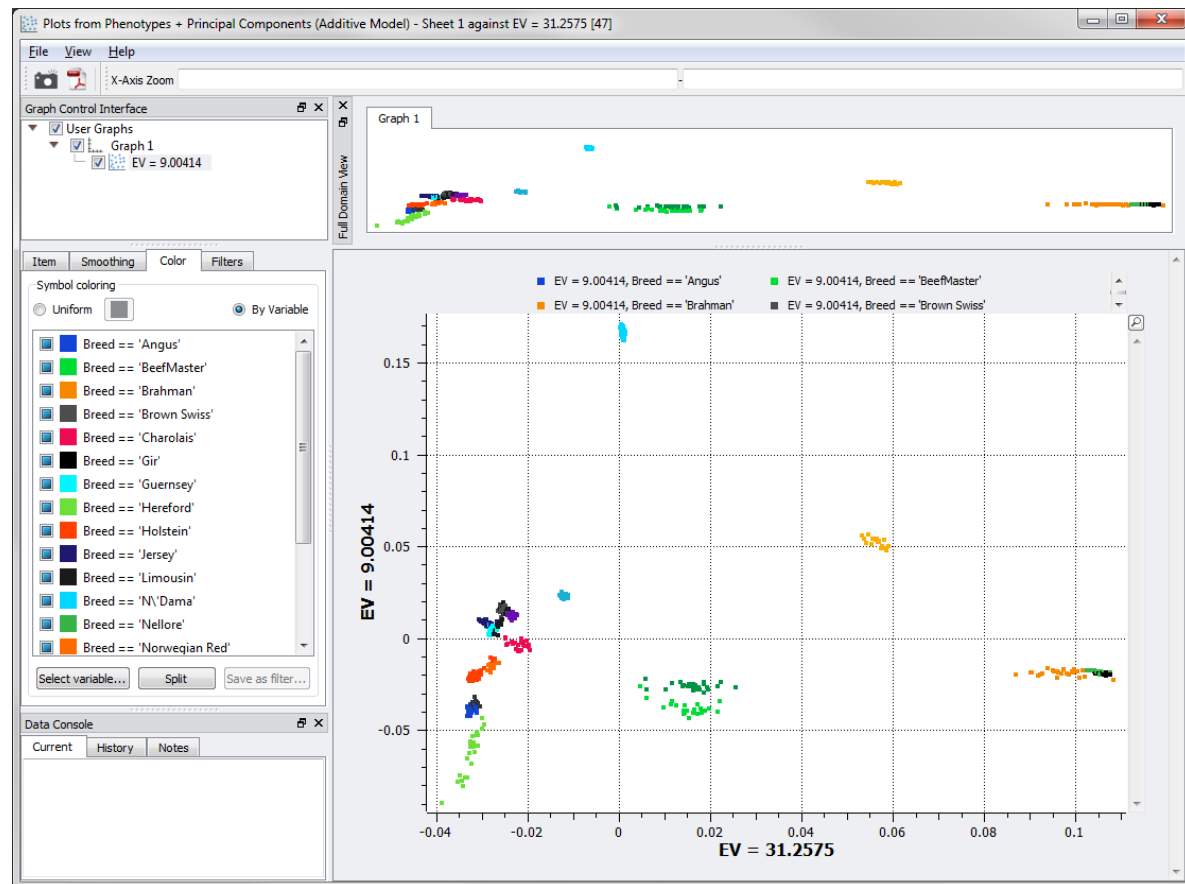
- 各種フィルタリングの実施によって、サンプル数やSNP数が変化する。



- サンプル間相関データの計算には、Identical by State (IBS), Identical by Descent (IBD), GBLUP Genomic Relationship Matrixなど、様々な手法がある。
- ここで計算したサンプル間相関データ、および表現型データとSNPジェノタイプデータを使用して、関連解析を実行する。

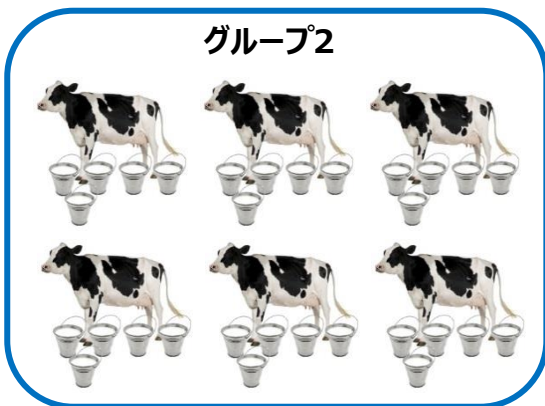
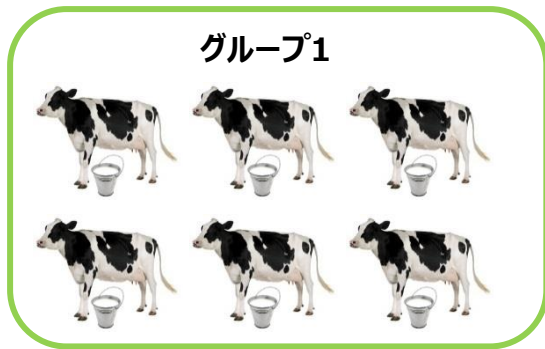
# 主成分分析 (PCA)

- Genotype Statistics by Marker
- Genotype Filtering by Marker
- Genotype Statistics by Sample
- Quality Assurance and Utilities
- LD Reports
- Genotype Principal Component Analysis**
- Create Imputation Reference Panel
- Genotype Imputation with BEAGLE
- PBAT Family-Based QA
- PBAT Genotype Analysis
- Genotype Association Tests
- Haplotype Association Tests
- Haplotype Block Detection
- Haplotype Trend Regression
- Runs of Homozygosity for GWAS
- Bayesian Genomic Prediction
- Compute Genomic BLUP (GBLUP)
- Genetic Correlation of Two Traits using GBLUP
- K-Fold Cross Validation (for Genomic Prediction)
- Predict Phenotypes From Existing Results
- Mixed Linear Model Analysis
- LD Score Regression
- Mixed Linear Model Analysis with Interactions

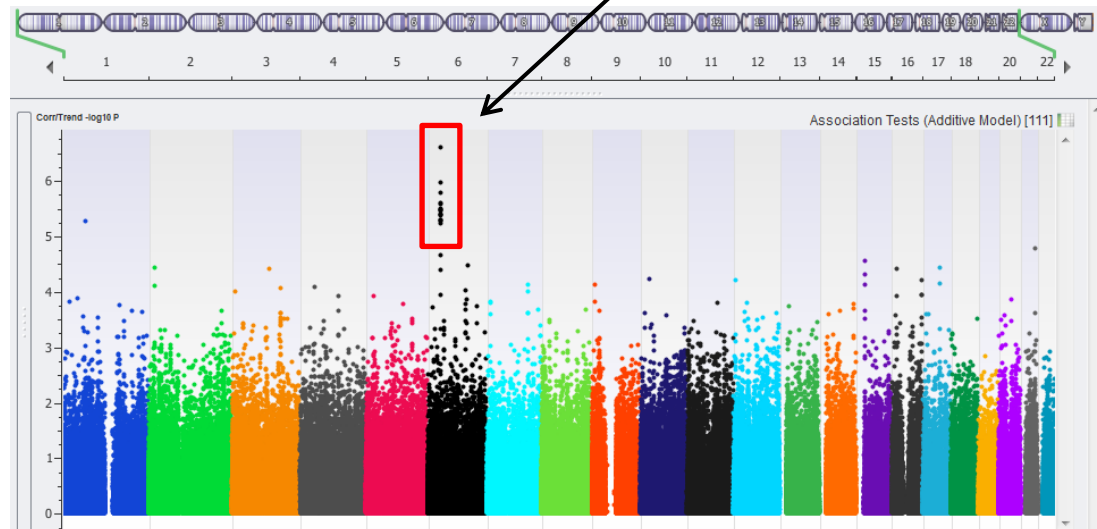


- 主成分分析 (PCA) により、集団の構造化を調べることも可能。

- 関連解析では、サンプルの表現型と関連しているSNPマーカを見つけることができる。
- 関連マーカが分かれば、表現型を規定する遺伝子などが分かり、品種改良などに活用できる。



表現型との関連が高いマーカ



## 線形混合モデル解析

おもにサンプルの血縁関係による偏りを除外し、関連解析を行う場合に用いられる手法。近交系サンプルなどの解析に用いられる。

- **Mixed Model GWAS using a single locus (EMMAX)**
  - ジェノタイプデータによるサンプル間の相関データを用いて、血縁関係の偏りを補正する。
  - 1か所のSNPごとに表現型との関連を計算する。
- **Multi-locus mixed model GWAS (MLMM)**
  - ジェノタイプデータによるサンプル間の相関データを用いて、血縁関係の偏りを補正する。
  - 複数か所のSNPをまとめて、表現型との関連を計算する。

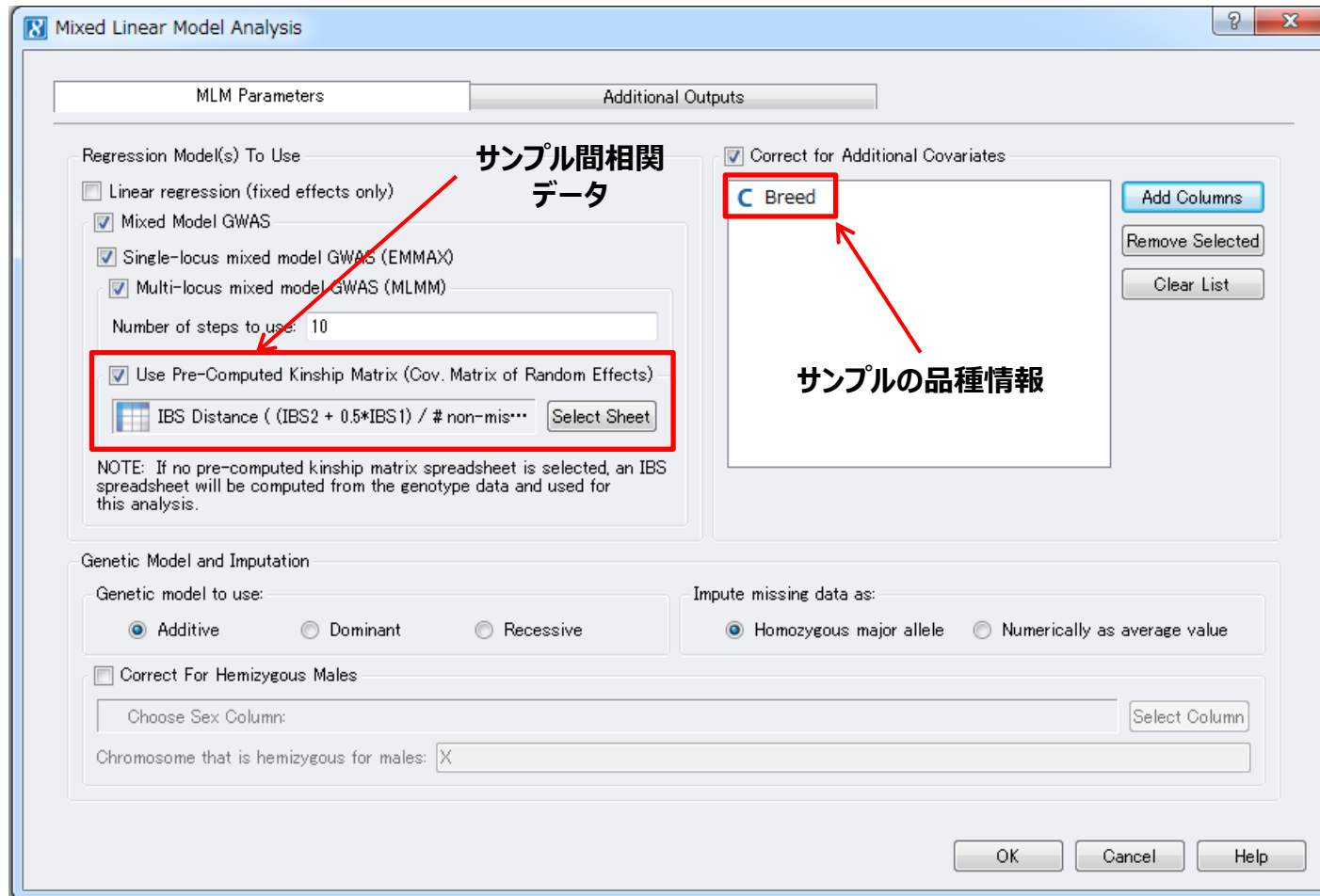
表現型データ

Phenotype + High Quality Data - Sheet 1 [45]

Unsort		B	5	R	6	R	7	R	8	R	9	G	10	G	11
Map	SampleID	Phenotype1	Phen2	Phen3	Phen4	Phen5	Hapmap43437-BTA-101873	ARS-BFGL-NGS-16466							
1	WG0099889-DNAD04_ANG000027	0	1	-0.0358	16.6761427265423	24.8389532585512	G_G	C_C							
2	WG0099889-DNAA02_ANG000008	0	1	0.0362	23.3265398984327	21.7506929572886	G_G	C_T							
3	WG0099889-DNAA03_ANG000016	0	1	0.0061	19.0803796478002	17.3037717691138	G_G	C_T							
4	WG0099889-DNAB03_ANG000017	1	1	-0.0382	19.9554299755303	26.6011234866098	G_G	C_C							
5	WG0099889-DNAB04_ANG000025	1	1	-0.0334	27.6744146632411	23.4778765840422	G_G	C_C							
6	WG0099889-DNAC02_ANG000010	1	1	-0.0278	22.5351447396607	27.0589945698839	G_G	C_C							
7	WG0099889-DNAC03_ANG000018	0	0	0.0884	15.5430766061975	15.7245152305321	G_G	C_C							
8	WG0099889-DNAC04_ANG000026	0	0	0.0432	21.7196917403443	6.34720842916488	G_G	C_C							
9	WG0099889-DNAD02_ANG000011	0	1	0.0217	18.8330050819465	23.8891657707582	G_G	C_C							
10	WG0099889-DNAD03_ANG000019	0	1	-0.0785	20.7770585758938	16.2528640600238	G_G	C_T							
11	WG0099889-DNAE01_ANG000004	0	1	-0.0504	18.6132833636024	25.5057310813833	G_G	C_T							
12	WG0099889-DNAE02_ANG000012	0	1	0.0438	20.6256893411594	29.8261600049517	A_G	C_C							
13	WG0099889-DNAE03_ANG000020	0	1	0.03	18.7853510513897	25.3954670386242	G_G	C_T							
14	WG0099889-DNAF03_ANG000021	0	1	-0.1487	16.0964707795	26.5363421525724	A_G	C_T							
15	WG0099889-DNAG03_ANG000022	0	1	-0.0145	21.0490543187925	20.4968872140319	A_G	C_T							
16	WG0099889-DNAH01_ANG000007	0	0	0.0064	20.8434901894035	27.2988004304254	G_G	C_C							
17	WG0099889-DNAA04_ANG000024	1	1	0.0281	17.5777976474953	14.4559497122448	G_G	C_T							
18	WG0099889-DNAB01_ANG000001	1	1	0.0113	23.6330738091984	11.7640928208919	G_G	C_C							
19	WG0099889-DNAB02_ANG000009	1	1	-0.0649	22.0043692661169	23.478588601197	A_G	C_C							

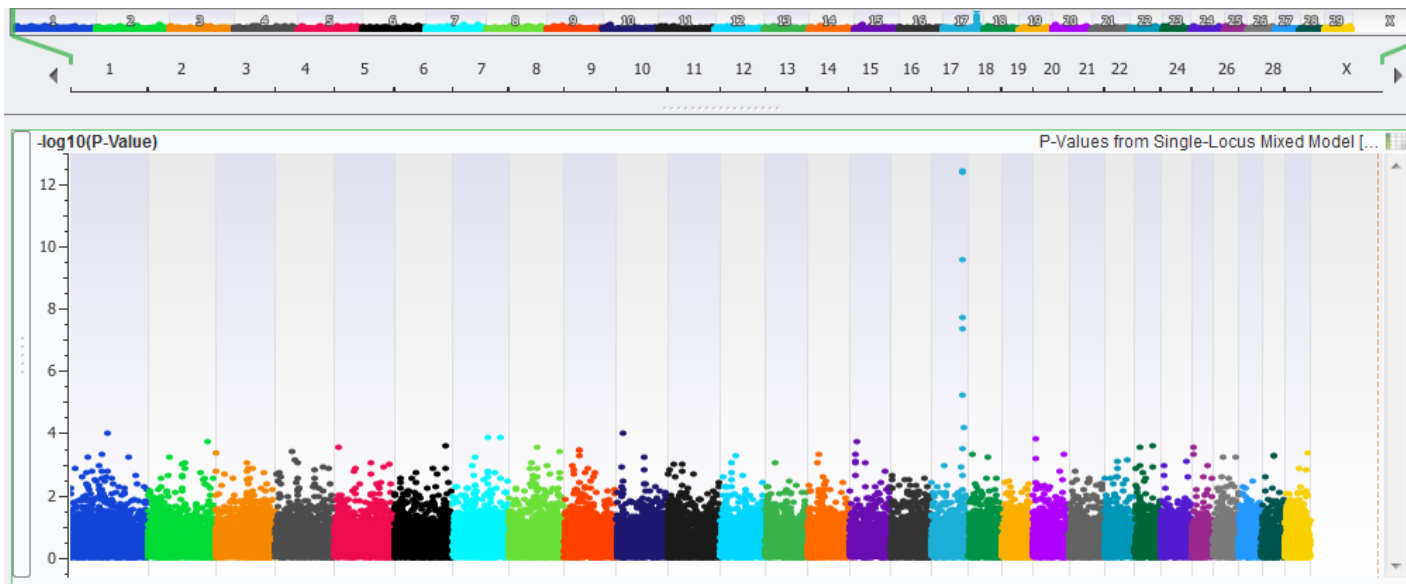
SNPジェノタイプデータ

- 解析のキーとなる、表現型データを指定する。
- 表現型データは、2値のケース/コントロールデータの他に、量的形質（Quantitative trait）も使用できる。



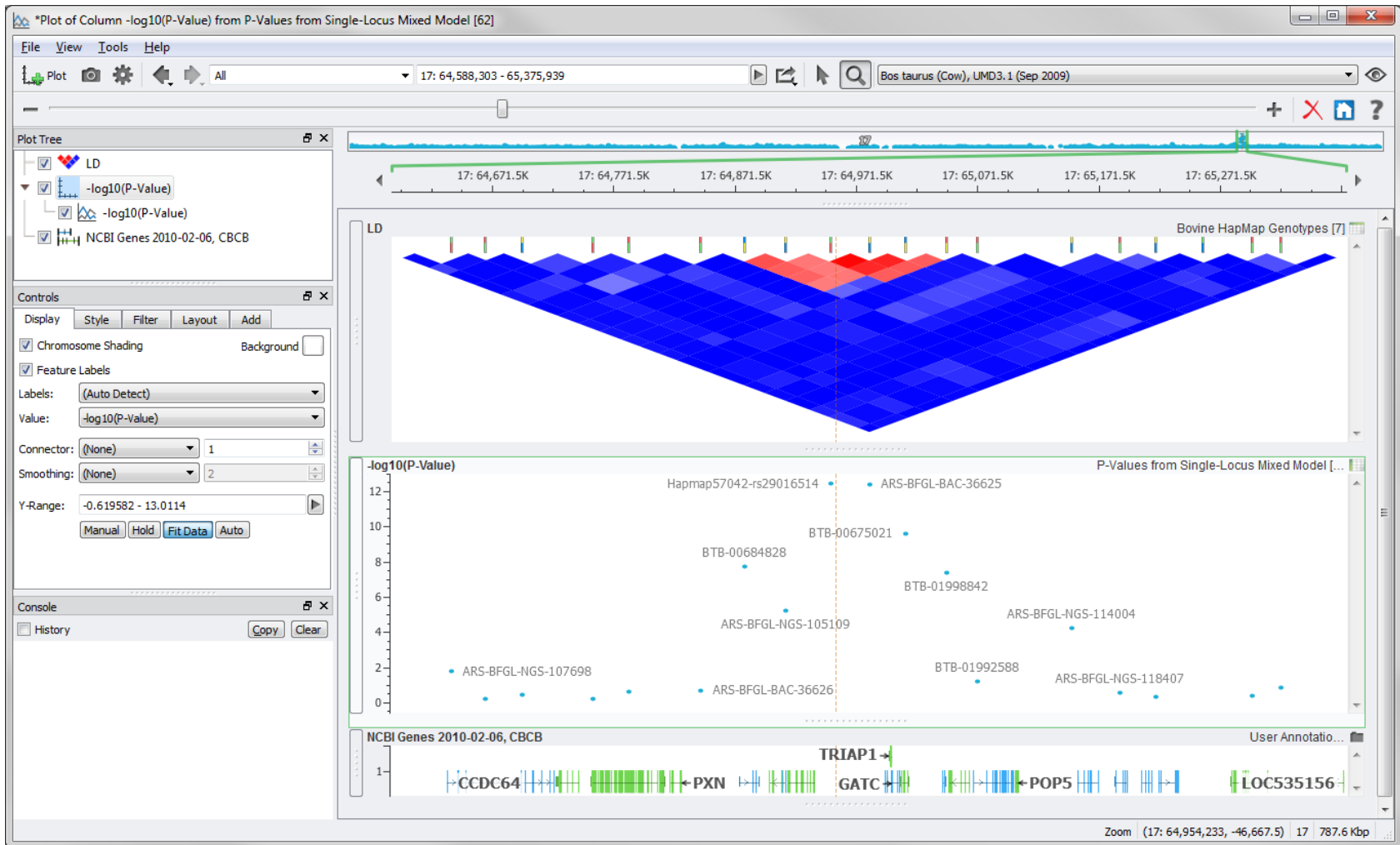
- あらかじめ計算しておいた、サンプル間相関データを指定する。
- 表現型データに、各サンプルの品種情報などが含まれている場合は、その情報を用いて、品種によるバイアスを補正することができる。

Unsort	Marker	R	1	R	2	R	3	R	4	R	5
Map			P-Value		$-\log_{10}(\text{P-Value})$		Regression Beta		Beta Standard Error		Expected P
1	Hapmap43437-BTA-101873		0.634513300818039		0.197559269690104		0.0123242353782897		0.0259070674082485		0.635454047807311
2	ARS-BFGL-NGS-16466		0.925025664620226		0.0338462176952145		0.00281525587295601		0.029899242560134		0.924690306753442
3	ARS-BFGL-NGS-105096		0.987391940894995		0.0055104218188536		-0.000524998104079046		0.033203898161163		0.986953348336476
4	Hapmap34944-BES1_Contig627_1906		0.99008078323066		0.00432936875600361		-0.000384722945111051		0.0309283813981902		0.989528211990349
5	BTA-07251-no-rs		0.199303726822715		0.700484580258834		-0.0297222216077198		0.023122271350621		0.198943697666403
6	ARS-BFGL-NGS-98142		0.895135643960814		0.0481111490862626		-0.00423302864493059		0.0320967254164401		0.894846217788861
7	Hapmap53946-rs29015852		0.177090485154123		0.751804772271634		0.0348059450436607		0.0257460097201309		0.176074042535937



- 計算結果のSNPごとのP-valueをゲノムブラウザーにプロットし、マンハッタンプロットを表示する。





- 興味のあるゲノム上の領域を指定し、連鎖不平衡 (LD) プロットを表示することが可能。

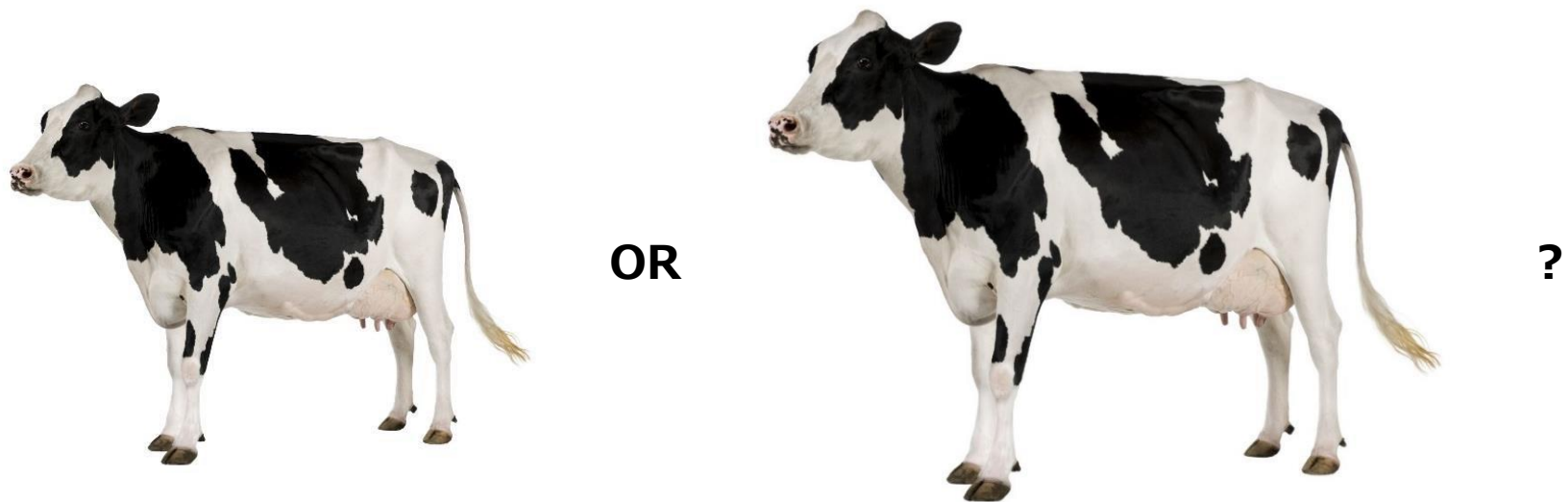
## 2. Genomic Prediction

- **ゲノム育種価（gEBV）の計算**

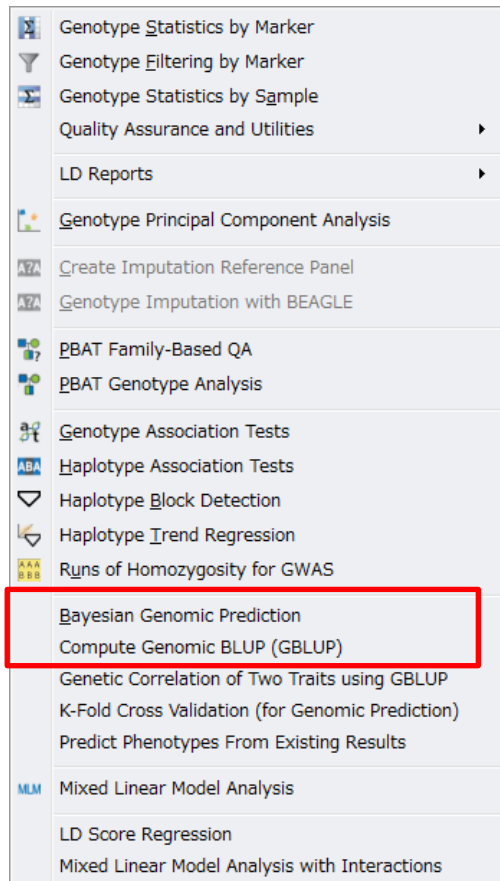
- ジェノタイプデータを使用し、個体の遺伝能力（育種価）の推定を行う。
- 後代検定による血統情報を必要としないため、従来法に比べて、コスト削減とスピードアップが図れる。

- **表現型の予測**

- 既存モデルにジェノタイプデータを当てはめることで、表現型未知個体の表現型データの予測を行う。
- ジェノタイプデータさえあれば予測が可能のため、表現型データを得るのに時間がかかる場合などにおいて、早期の個体の選抜などに有効。



- Genomic Predictionでは、サンプルごとの変量効果（ゲノム育種価）とSNPごとのアレル代替効果を計算する。
- ジェノタイプデータによるサンプル間の相関データを用いて、血縁関係の偏りを補正する。



## GBLUP

- 全SNPが表現型に影響を及ぼしていると仮定する。
- 遺伝子と環境因子の相互作用の補正が可能。

## Bayes C

- 個々のSNPの表現型に及ぼす効果が、確率 $\pi$ で0になる。（ $\pi$ の値は固定）

## Bayes C-pi

- 個々のSNPの表現型に及ぼす効果が、確率 $\pi$ で0になる。（ $\pi$ の値は変動する）

Compute Genomic BLUP (GBLUP)
?
×

**Computations**

Estimate the variance explained by all markers.

Compute GBLUP (Genomic Best Linear Unbiased Predictors) of additive genetic merits by sample and of allele substitution effects (ASE) by marker. If you select Gene by Environment Interactions, a GBLUP will be computed for every GRM.

**REML Computation Algorithm**

Use EMMA (faster, works for one GRM)   
  Use AI REML (works for multiple GRM's)

**Impute missing genotypic data as:**

Homozygous major allele   
  Numerically as average value

**Correct For Gender**

Choose Sex Column:  Select Column

Chromosome that is hemizygous for males:

Dosage compensation

Full dosage compensation  
 Equal X-linked genetic variance for males and females  
 No dosage compensation

**Use Pre-Computed Genomic Relationship Matrix for Gender Chromosome (AI REML only)**

Select Sheet

NOTE: If used, the other GRM must be pre-computed and must be for non-gender chromosomes only.

**Use Pre-Computed Genomic Relationship Matrix**

Select Sheet

NOTE: If no pre-computed genomic relationship matrix spreadsheet is selected, a genomic relationship matrix will be computed from the genotype data and used for this analysis.

**Normalization Algorithm (Used or Assumed) for the GRM**

Overall normalization   
  Normalize by individual marker (GCTA method)

**Correct for Additional Covariates**

Add Columns  
Remove Selected  
Clear List

**Correct for Gene by Environment Interactions**

Add Columns  
Remove Selected  
Clear List


**Missing Phenotypes**

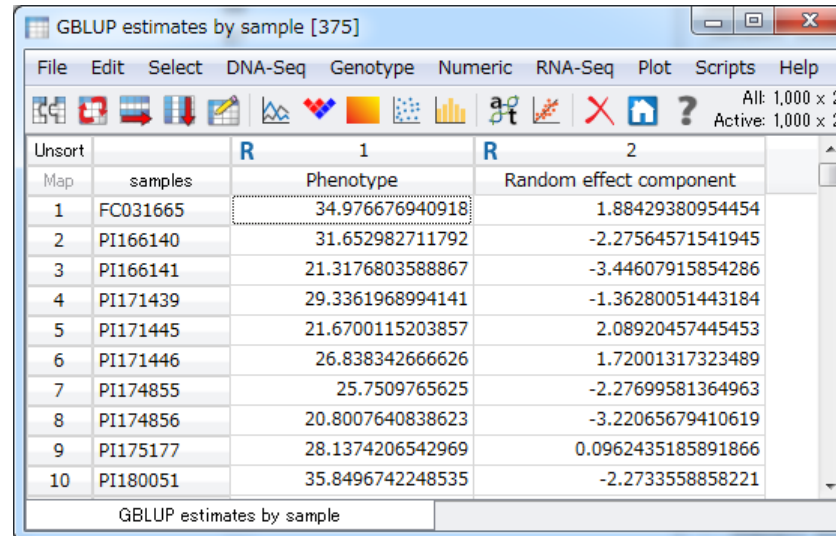
Predict random effects for samples with missing phenotypes  
 Drop samples with missing phenotypes

OK Cancel Help

- 必要に応じて、あらかじめ計算しておいたサンプル間相関データを指定。
- GBLUPでは、環境因子データによる補正も可能。

# 出力データ

 GBLUP estimates by sample  
GBLUP estimates by marker



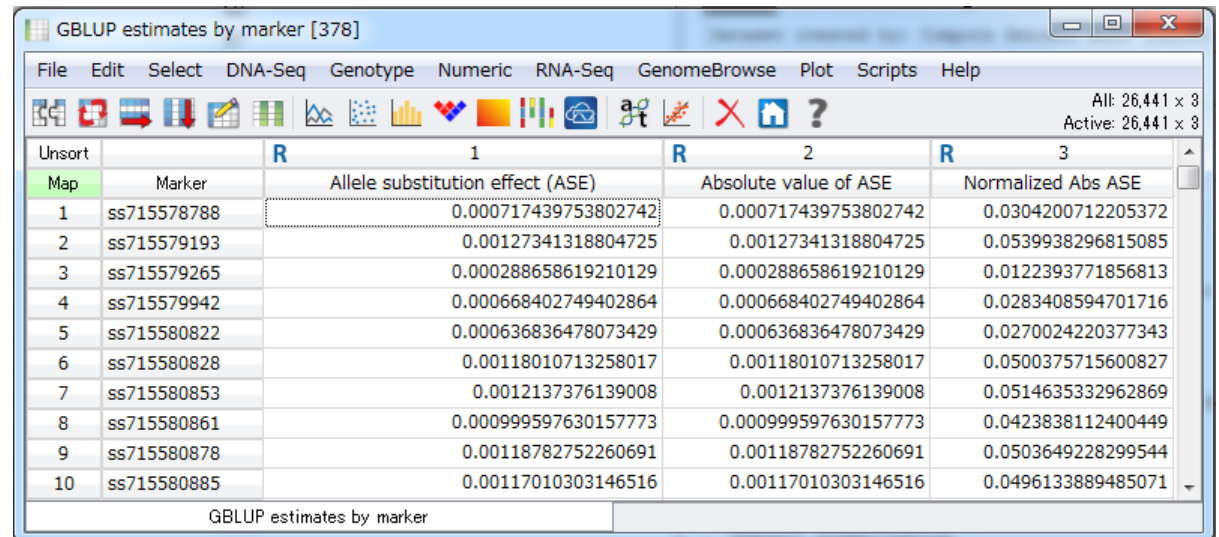
GBLUP estimates by sample [375]

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

All: 1,000 x 2  
Active: 1,000 x 2

Unsort		R	1	R	2
Map	samples		Phenotype		Random effect component
1	FC031665		34.976676940918		1.88429380954454
2	PI166140		31.652982711792		-2.27564571541945
3	PI166141		21.3176803588867		-3.44607915854286
4	PI171439		29.3361968994141		-1.36280051443184
5	PI171445		21.6700115203857		2.08920457445453
6	PI171446		26.838342666626		1.72001317323489
7	PI174855		25.7509765625		-2.27699581364963
8	PI174856		20.8007640838623		-3.22065679410619
9	PI175177		28.1374206542969		0.0962435185891866
10	PI180051		35.8496742248535		-2.2733558858221

GBLUP estimates by sample



GBLUP estimates by marker [378]

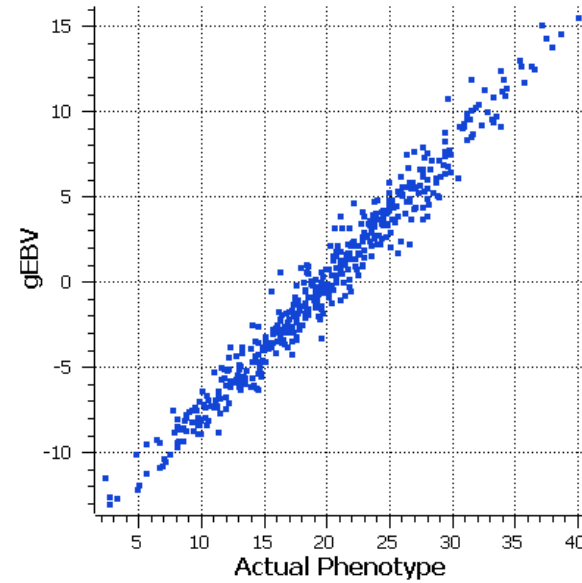
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help

All: 26,441 x 3  
Active: 26,441 x 3

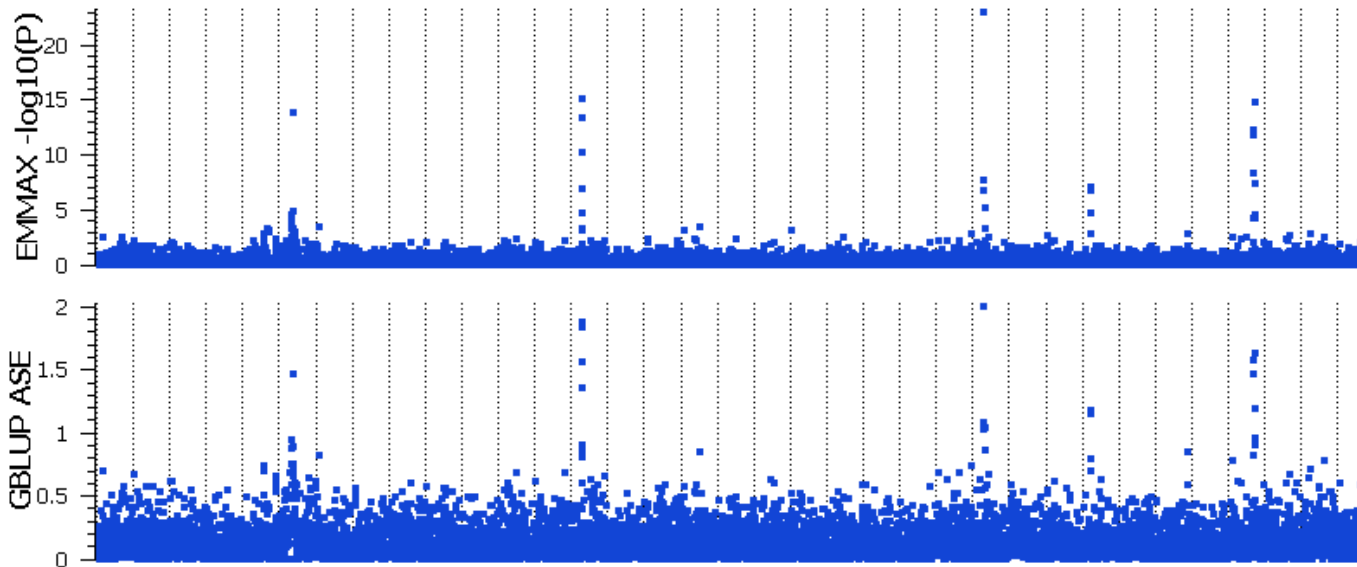
Unsort		R	1	R	2	R	3
Map	Marker		Allele substitution effect (ASE)		Absolute value of ASE		Normalized Abs ASE
1	ss715578788		0.000717439753802742		0.000717439753802742		0.0304200712205372
2	ss715579193		0.00127341318804725		0.00127341318804725		0.0539938296815085
3	ss715579265		0.000288658619210129		0.000288658619210129		0.0122393771856813
4	ss715579942		0.000668402749402864		0.000668402749402864		0.0283408594701716
5	ss715580822		0.000636836478073429		0.000636836478073429		0.0270024220377343
6	ss715580828		0.00118010713258017		0.00118010713258017		0.0500375715600827
7	ss715580853		0.0012137376139008		0.0012137376139008		0.0514635332962869
8	ss715580861		0.000999597630157773		0.000999597630157773		0.0423838112400449
9	ss715580878		0.00118782752260691		0.00118782752260691		0.0503649228299544
10	ss715580885		0.00117010303146516		0.00117010303146516		0.0496133889485071

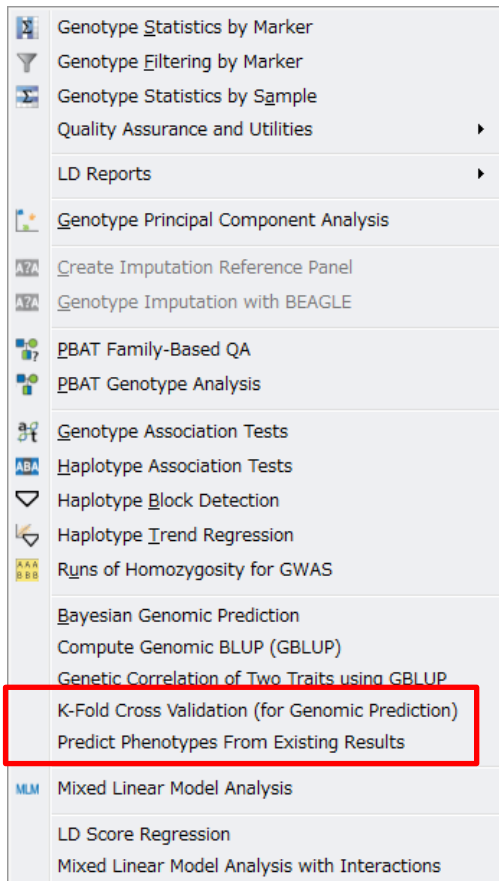
GBLUP estimates by marker

Actual Phenotype vs  
Random effect component (gEBV)



Allele substitution effect vs GWAS result





## K-Fold Cross Validation

- 交差検証法 (Cross Validation) を用いて、各手法 (GBLUP, Bayes C, Bayes C-pi) のパフォーマンスの評価を行う。
- 表現型予測に使用するモデルの計算を行う。

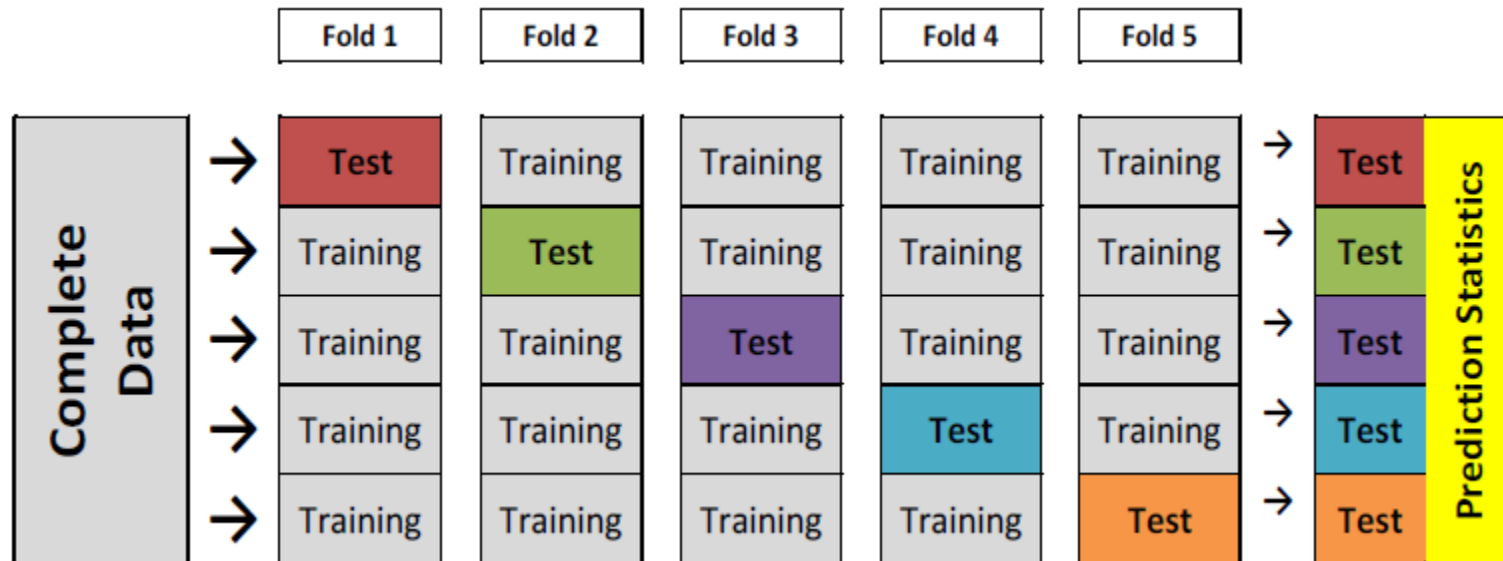
## Predict Phenotypes From Existing Results

- K-Fold Cross Validationで計算したモデルを使用し、表現型未知サンプルのジェノタイプデータから表現型の予測を行う。



# K-Fold Cross Validation

- データセットをK個に分割し、そのうち1個をテストセット、残りのK-1個をトレーニングセットとし、K回の検証を行う。
- 計算に使用するサンプルデータは、表現型データとジェノタイプデータの両方が必要。
- 検証結果として、実際の表現型データと、ジェノタイプデータから予測された表現型データの相関係数などをまとめたレポートが出力される。
- 表現型予測用のモデルデータも同時に出力される。



**K-Fold Cross Validation (for Genomic Prediction)**

Computations  
Perform k-fold cross validation on GBLUP and Bayes C\C-pi

**Method(s)**

- Genomic Best Linear Unbiased Predictors (GBLUP)
- Bayes C-pi
- Bayes C

Bayesian Options

Number of Iterations: 50000

Burn-in: 0

Thinning: 0

Initial Pi (for Bayes C this will be the fixed value): 0.5

Computation Method:  As Is  Centered

Correct For Gender

Choose Sex Column: [ ] Select Column

Chromosome that is hemizygous for males: X

Use Pre-Computed Genomic Relationship Matrix

Pre-computed genomic relationship matrix spreadsh... [ ] Select Sheet

NOTE: If no pre-computed genomic relationship matrix spreadsheet is selected, a genomic relationship matrix will be computed from the genotype data and used for this analysis.

Correct for Additional Covariates

Add Columns  
Remove Selected  
Clear List

Impute Missing Genotypic Data As

Homozygous major allele  Numerically as average value

Stratify Folds by

Grouping... [ ] Select Column

**K-Fold Options**

Number of Folds: 5

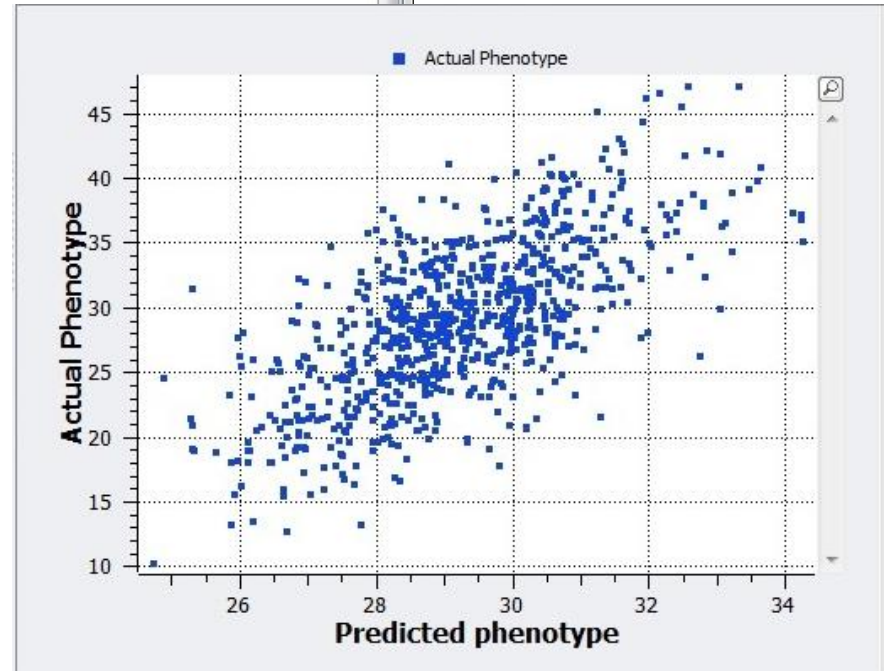
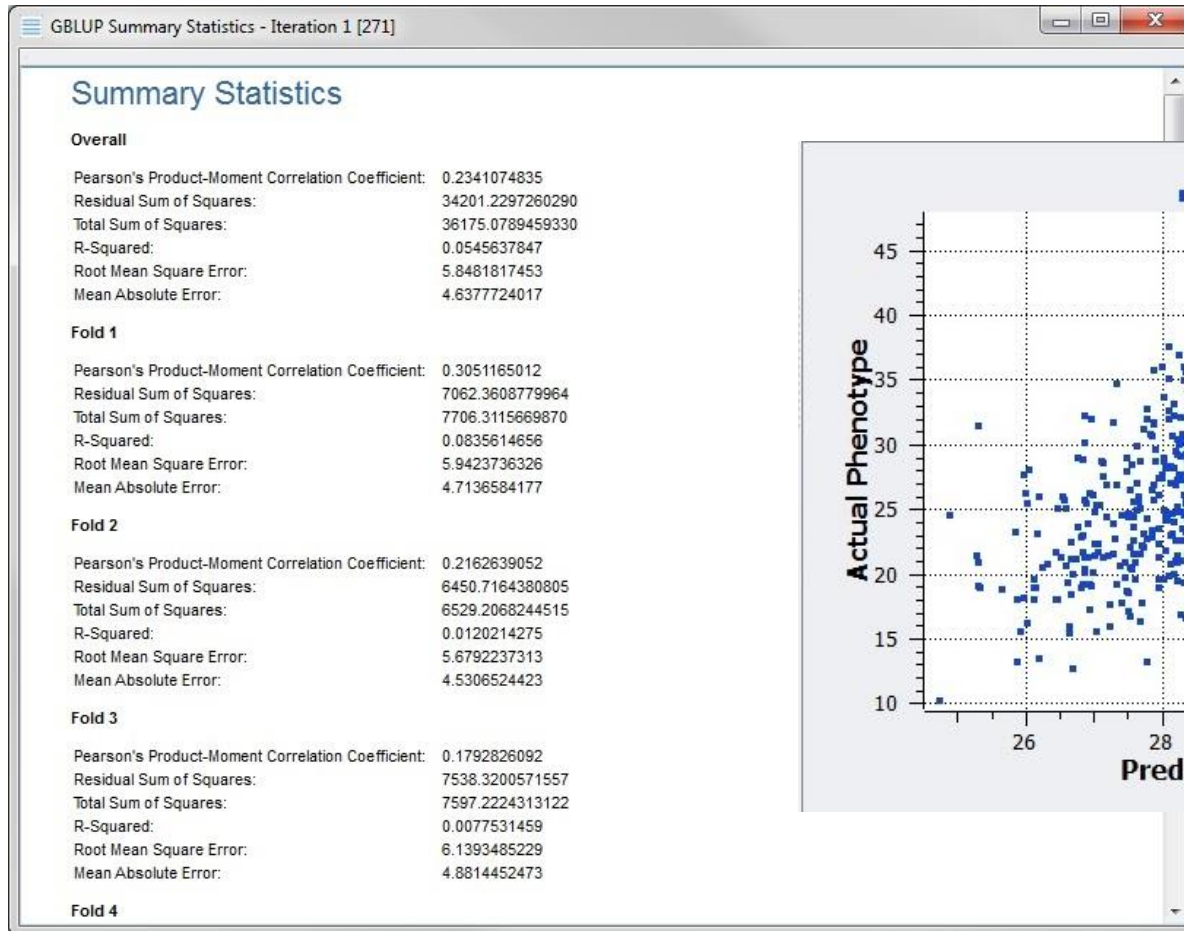
Number of Iterations: 1

Spreadsheet Options

Delete intermediate spreadsheets with results for each fold?

OK Cancel Help

- 評価を行う手法の選択や、データセットの分割数、繰り返し数などを指定する。



- SNPごとのアリル代替効果とサンプルごとの変量効果、固定効果に加え、検証結果をまとめたレポートが出力される。

Predict Phenotypes From Existing Results

Computations  
Predict Phenotypes using existing ASE and Fixed Effect Coefficients

Computation Method(s)

As is (genotype values will be 0, 1, or 2)  
(Recommended for Bayesian Results)

Centered (genotype values will be coded as 0, 1, or 2, then centered by the mean)  
(Recommended for GBLUP Results)

Impute Missing Genotypic Data As:

Homozygous major allele  Numerically as average value

Correct For Gender

Choose Sex Column:  Select Column

Chromosome that is hemizygous for males:

Homozygous Markers:

Include (Recommended for GBLUP Results)  Remove (Recommended for Bayesian Results)

Correct for Additional Covariates

Add Columns  
Remove Selected  
Clear List

Transformed Data

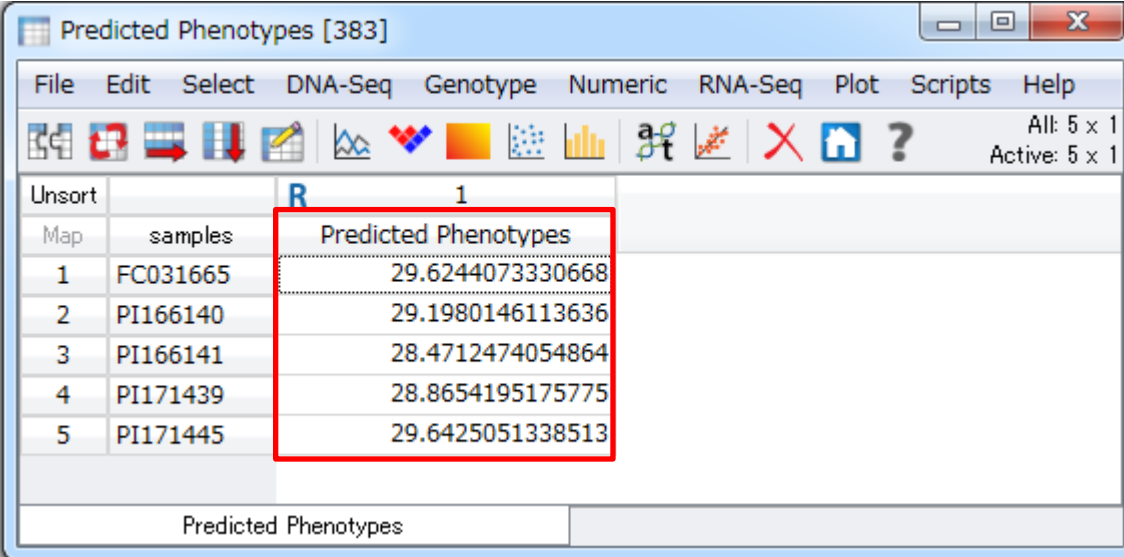
Mean   
Standard Deviation

Model Values

Allele Substitution Effects   
 Fixed Effect Coefficients

OK Cancel Help

- 計算に使用するサンプルデータは、ジェノタイプデータのみを必要とする。
- K-Fold Cross Validationから出力された、アレル代替効果と固定効果のデータを指定する。



Map	samples	Predicted Phenotypes
1	FC031665	29.6244073330668
2	PI166140	29.1980146113636
3	PI166141	28.4712474054864
4	PI171439	28.8654195175775
5	PI171445	29.6425051338513

- 各サンプルごとに予測された表現型データが出力される。

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](mailto:biosupport@filgen.jp)