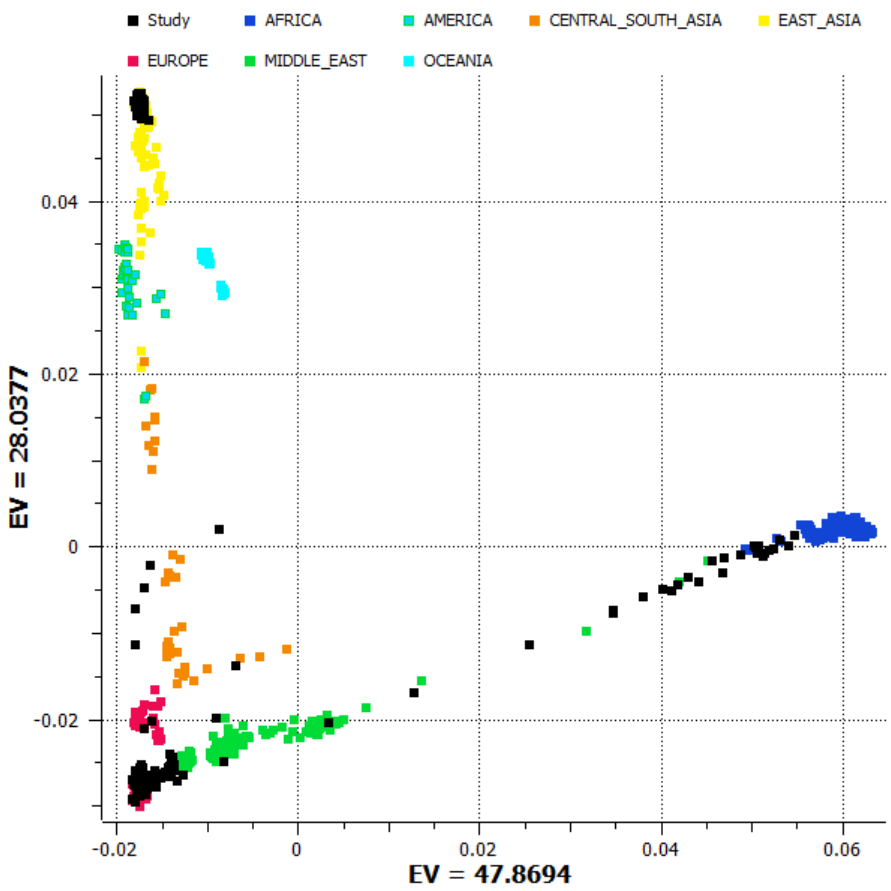
GWASにおけるバイアス調整

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

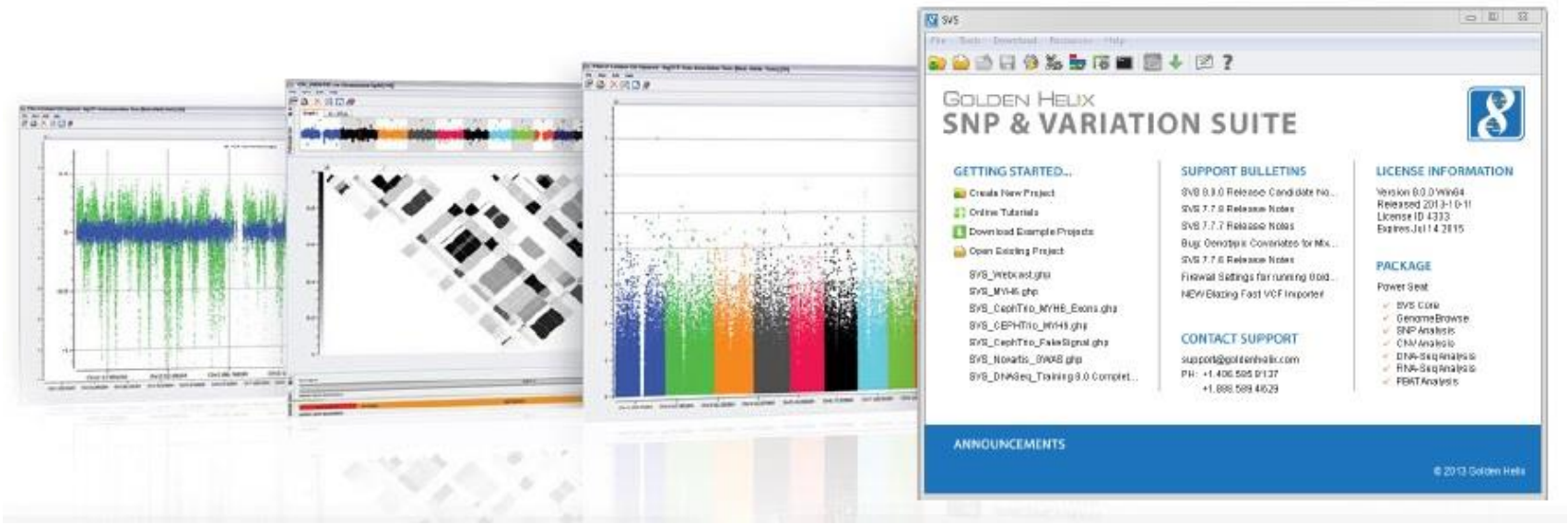


- ゲノムワイド関連解析（GWAS）では、大規模サンプルデータを扱うことになるため、サンプルの遺伝的背景や隠れた近縁関係、バッチエフェクトなどがバイアスとなることがあり、この補正を行うことが必要となる

- Golden Helix社SNP & Variation Suite (SVS)では、基本的なGWAS計算用のツールに加え、これらバイアスに関する調整用のツールが多数搭載されており、サンプルデータや解析結果にあわせて使い分けることができる

SNP & Variation Suite (SVS)

Filgen®
biosciences & nanosciences



Core Features

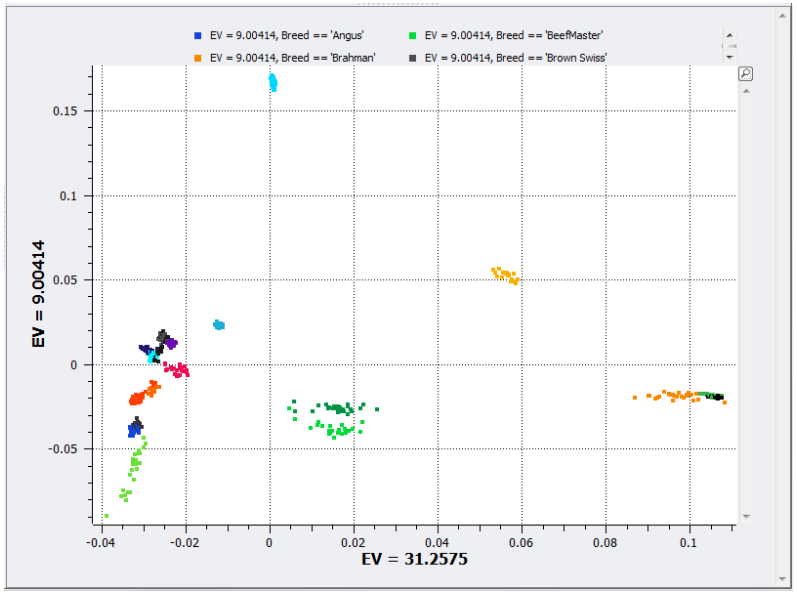
- Powerful Data Management
- Rich Visualizations (GenomeBrowse)
- Robust Static
- Flexible

Applications

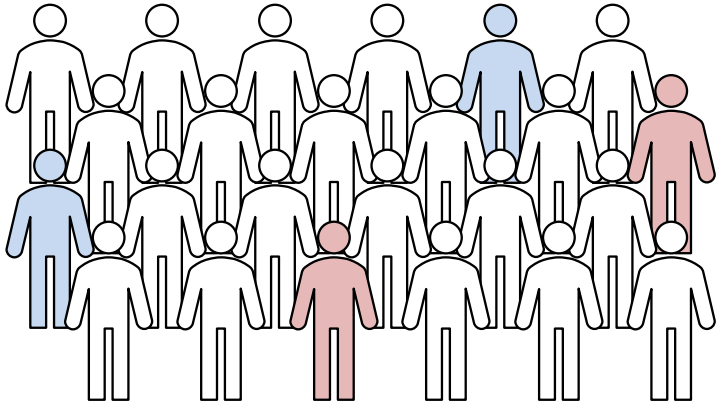
- Genotype Analysis
- Agrigenomics Analysis
- DNA/RNA Sequence Analysis
- CNV Analysis

GWASにおけるバイアス

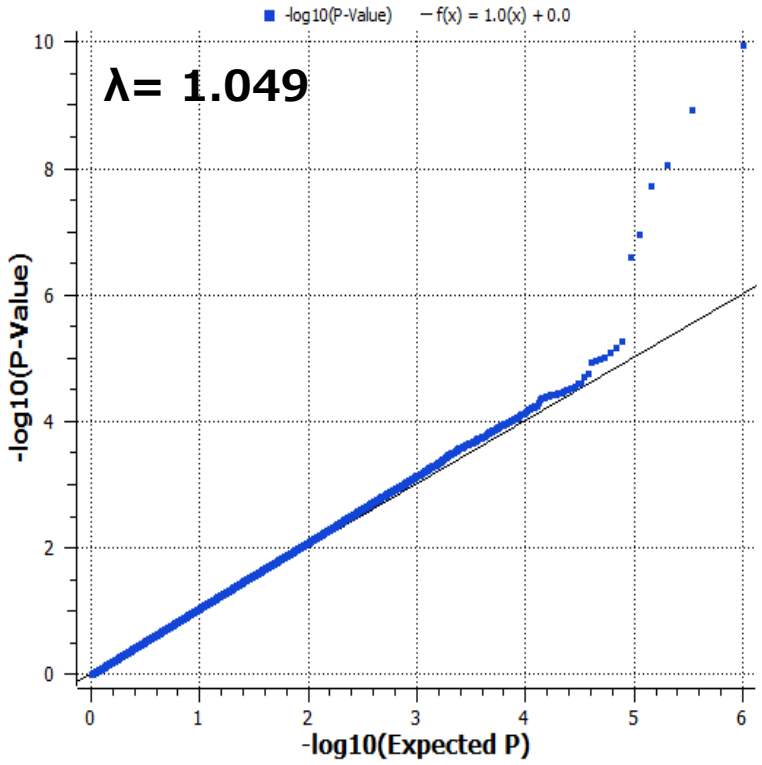
- GWASにおいて大規模なサンプル集団のデータを用いる場合、様々なバイアス要因によって結果がゆがめられ、正しい結果が得られないことがある
  - 実験手技などによるバイアス
    - ✓ バッチエフェクト
  - SNPデータのバイアス
    - ✓ 集団の構造化（人種などの遺伝的背景の違い）
    - ✓ 隠れた血縁関係

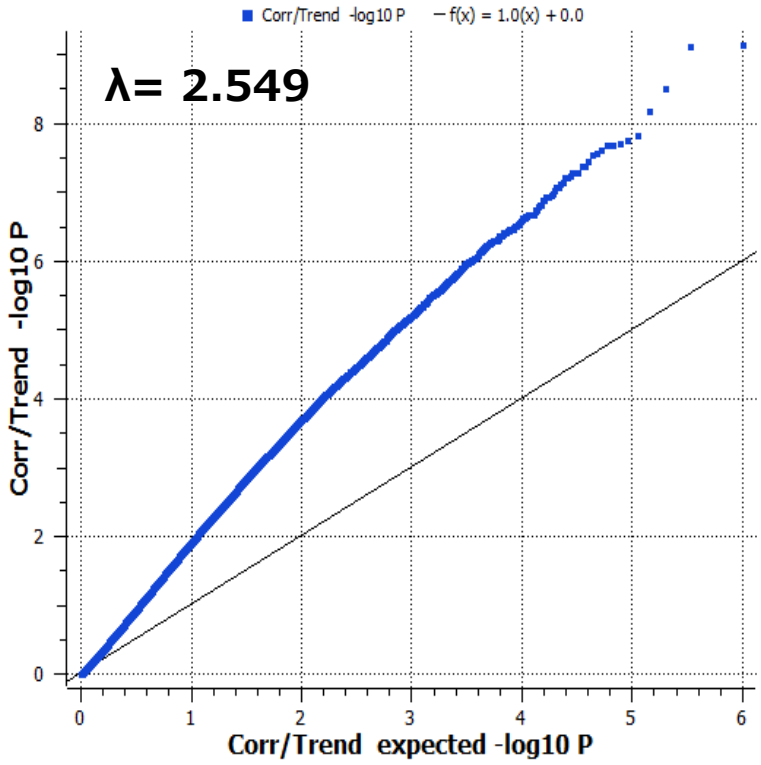


- 一部のバイアスはQCで除外することができるが、SNPデータのバイアスの中には、除外できないものもある
  - ✓ Call Rate
  - ✓ マイナーアレル頻度
  - ✓ Hardy–Weinberg平衡



GWASにおけるバイアス





- サンプル集団内にバイアスが存在すると、GWAS結果の統計量のインフレーションが起こり、偽陽性が出やすくなる
- 統計量のインフレーションは、Q-Qプロットやλ（ラムダ）値より評価が可能
- ただしλ値は、ポリジェニック効果によっても増大するので注意が必要

バイアス補正の手法

Genomic Control法

- ✓ λ値を補正係数として用いてGWAS計算を実行
- ✓ 保守的な補正方法のため、偽陰性が多くなる場合がある

PCA法

- ✓ SNPデータをもとに計算された主成分の値を用いて補正を行う
- ✓ GWAS計算には回帰分析を使用し、共変量として主成分の値を用いる
- ✓ いくつまでの主成分の値を補正に用いるかは、ユーザー自身で検討が必要

混合モデル法

- ✓ サンプル間のゲノム関係行列データを用いて補正を行う
- ✓ 計算に用いる変数を、固定効果と変量効果に分けて行う混合モデル法を使い、ゲノム関係行列を変量効果として用いる
- ✓ ゲノム関係行列やGWAS計算に用いる手法を、様々なアルゴリズムから選択が可能

Genomic Control法

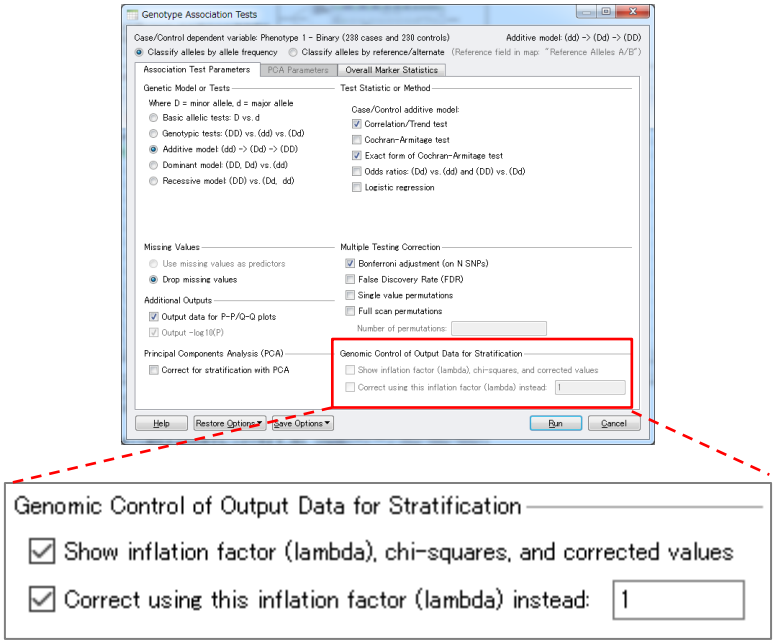
■ Genomic Controlによる補正

- 1ステップ目のGWAS計算結果において、λ値を確認
- 2ステップ目のGWAS計算時に、補正用パラメータとして1ステップ目で計算したλ値を入力

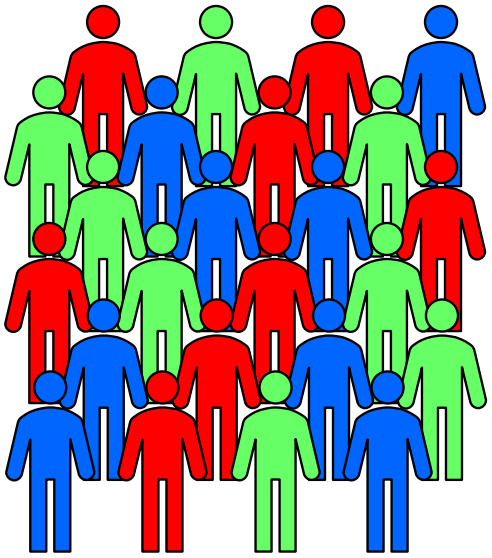
ステップ1
■ GWAS実行結果より、λ値を確認

ステップ2
■ 2回目の計算時に、補正用パラメータとして入力

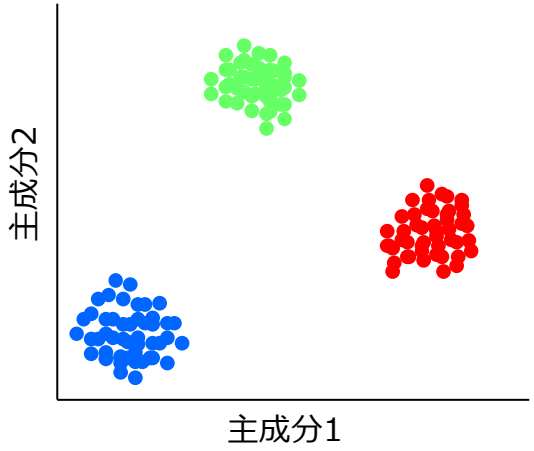
```
Markerwise Genotype Statistics:
  Call Rate: No
  Number of Alleles: No
  Allele Frequencies: Yes
  Carrier Counts: No
  HWE P-Value: No
  Fisher's Exact Test for HWE P-Value: No
  Signed HWE R: No

  Also output -log10(Value): No
  Also output data for P-P/Q-Q plots: No

  Genotype Counts: Yes
  Allele Counts: Yes


Inflation Factor (Lambda) Found for Armitage : 2.66031
```



PCA法



サンプル名	主成分1	主成分2
サンプル1	-0.0091	0.015
サンプル2	-0.013	0.016
サンプル3	-0.017	0.014
サンプル4	-0.0073	-0.079
サンプル5	0.0078	0.0059



補正用パラメータとして利用

■ **PCA (Principal Component Analysis) による補正**

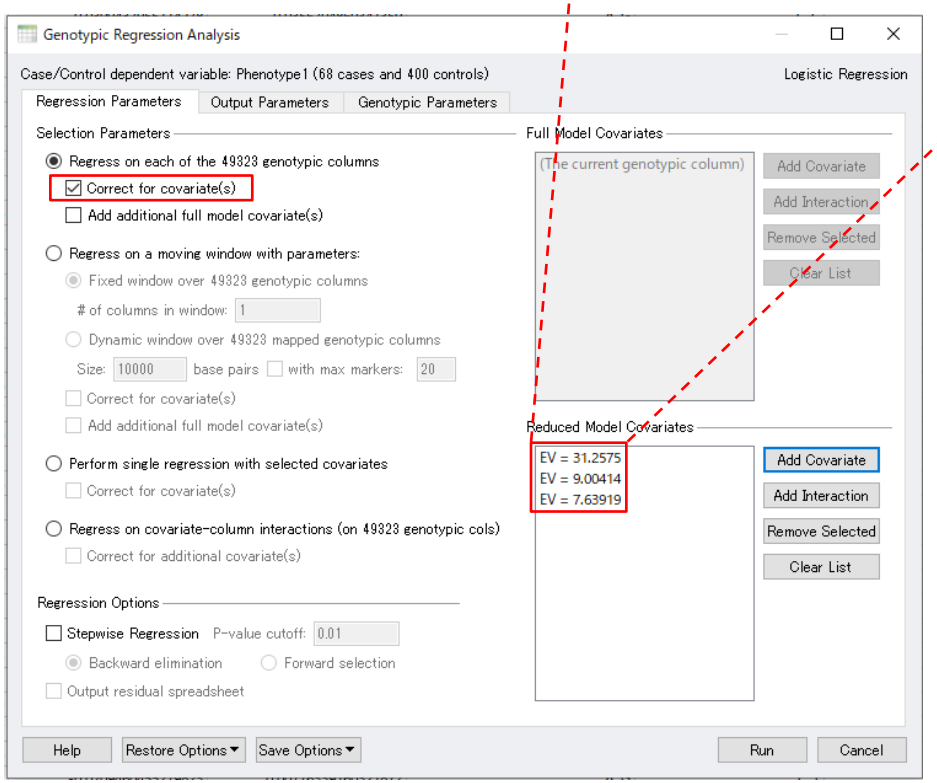
- サンプルごとのSNPデータより、サンプル間の関係の主成分（Principal Component）を計算し、GWAS実行時の補正用パラメータとして利用する

- SVSでは、主成分の計算アルゴリズムにEIGENSTRATを改良したものを使用

PCA法

主成分データ　　SNPデータ

SampleID	Breed	Phenotype1	EV = 31.2575	EV = 9.00414	EV = 7.63919	Hapmap43437-BTA-101873	ARS-BFGL-NGS-16466	ARS-BFGL-NGS-105096
WG0099889-DNAG01 ANG000006	Angus	0	-0.0323107711182121	-0.0390477216669378	-0.0203866456446584	G_G	C_T	C_T
WG0099889-DNAG02 ANG000014	Angus	1	-0.0334511204995828	-0.0368577126539786	-0.0219630320469021	G_G	T_T	C_C
WG0099889-DNAH02 ANG000015	Angus	1	-0.0325235980832336	-0.037030429653628	-0.0218091697179903	A_G	C_T	C_T
WG0099889-DNAH03 ANG000023	Angus	1	-0.0334797162287917	-0.041983240543258	-0.0219867218216778	A_G	C_T	C_T
WG0099889-DNAA05 BMA000005	BeefMaster	0	0.0161309046995375	-0.0399437955114278	-0.0356704860341359	A_G	C_C	C_C
WG0099889-DNAB05 BMA000006	BeefMaster	0	0.01475190110482	-0.0391245884742798	-0.0294125523992397	A_A	C_C	C_T
WG0099889-DNAB07 BMA000022	BeefMaster	0	0.0170611720484291	-0.040533528599169	-0.0335210411521843	A_A	C_C	T_T
WG0099889-DNAD06 BMA000016	BeefMaster	0	0.0186522832107677	-0.0385825915576416	-0.0331631085746 5	A_G	C_T	C_C



■ Genotype Regression Analysis

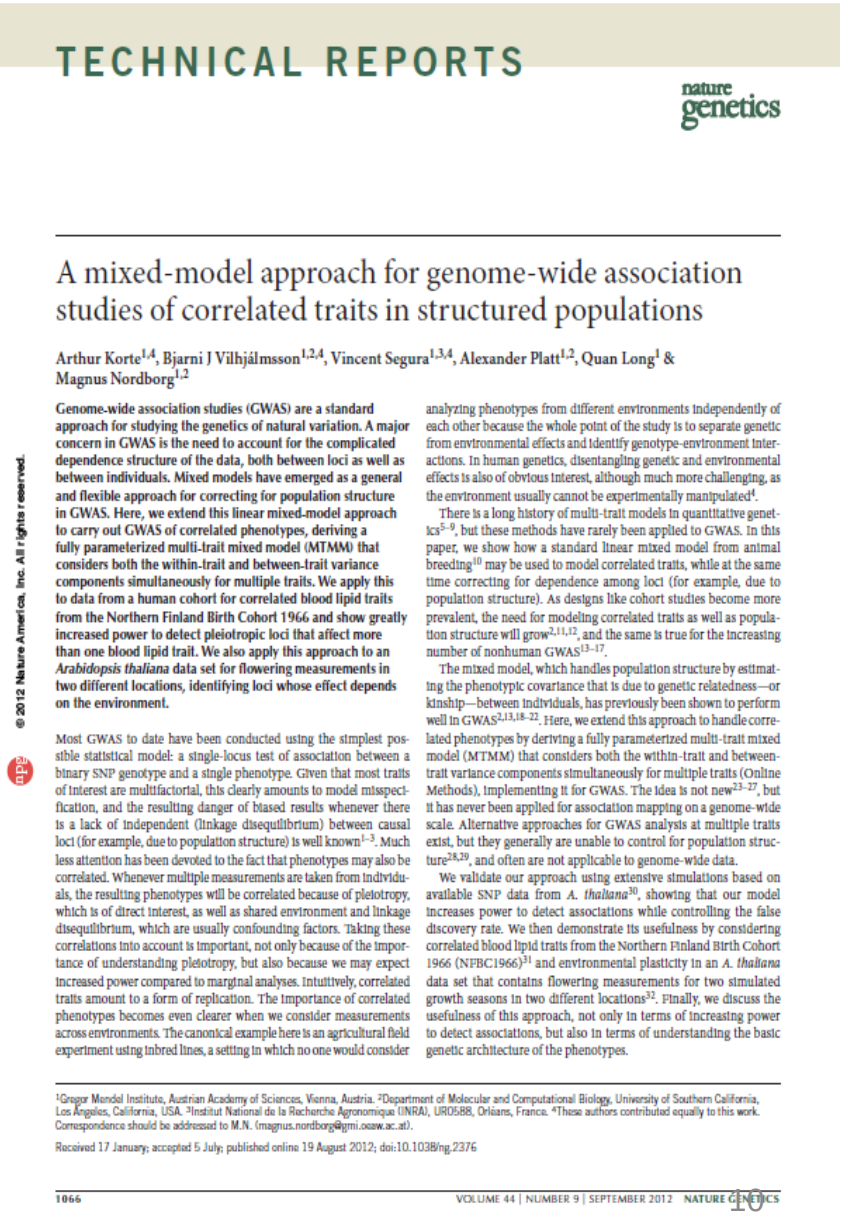
- SNPデータを用いて回帰分析を実行するためのツール
- SNPデータシートに、サンプルごとの共変量（身長、体重、性別など）を加えておくことで、共変量の補正が可能
- 主成分データを共変量として設定することで、集団の構造化の補正を行う

混合モデル法

Filgen® biosciences & nanosciences

■ 混合モデルによる補正

- 共変量などの変数を、**固定効果**と**変量効果**に分けて計算を行う
  - ✓ **固定効果**： 性別、年齢などの固定された因子
  - ✓ **変量効果**： サンプル同士の血縁関係や遺伝的関係、バッチエフェクトなど、ばらつきが存在する因子
- 変量効果として、サンプル間のゲノム関係行列を用いることで、集団の構造化や隠れた血縁関係の補正を行う
- 3種類のアルゴリズムが利用可能
  - ✓ EMMAX
  - ✓ MLMM
  - ✓ GBLUP



TECHNICAL REPORTS

nature genetics

A mixed-model approach for genome-wide association studies of correlated traits in structured populations

Arthur Korte[1,4], Bjarni J Vilhjálmsson[1,2,4], Vincent Segura[1,3,4], Alexander Platt[1,2], Quan Long[1] & Magnus Nordborg[1,2]

Genome-wide association studies (GWAS) are a standard approach for studying the genetics of natural variation. A major concern in GWAS is the need to account for the complicated dependence structure of the data, both between loci as well as between individuals. Mixed models have emerged as a general and flexible approach for correcting for population structure in GWAS. Here, we extend this linear mixed-model approach to carry out GWAS of correlated phenotypes, deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits. We apply this to data from a human cohort for correlated blood lipid traits from the Northern Finland Birth Cohort 1966 and show greatly increased power to detect pleiotropic loci that affect more than one blood lipid trait. We also apply this approach to an *Arabidopsis thaliana* data set for flowering measurements in two different locations, identifying loci whose effect depends on the environment.

Most GWAS to date have been conducted using the simplest possible statistical model: a single-locus test of association between a binary SNP genotype and a single phenotype. Given that most traits of interest are multifactorial, this clearly amounts to model misspecification, and the resulting danger of biased results whenever there is a lack of independent (linkage disequilibrium) between causal loci (for example, due to population structure) is well known[1–3]. Much less attention has been devoted to the fact that phenotypes may also be correlated. Whenever multiple measurements are taken from individuals, the resulting phenotypes will be correlated because of pleiotropy, which is of direct interest, as well as shared environment and linkage disequilibrium, which are usually confounding factors. Taking these correlations into account is important, not only because of the importance of understanding pleiotropy, but also because we may expect increased power compared to marginal analyses. Intuitively, correlated traits amount to a form of replication. The importance of correlated phenotypes becomes even clearer when we consider measurements across environments. The canonical example here is an agricultural field experiment using inbred lines, a setting in which no one would consider analyzing phenotypes from different environments independently of each other because the whole point of the study is to separate genetic from environmental effects and identify genotype-environment interactions. In human genetics, disentangling genetic and environmental effects is also of obvious interest, although much more challenging, as the environment usually cannot be experimentally manipulated[4].

There is a long history of multi-trait models in quantitative genetics[5–9], but these methods have rarely been applied to GWAS. In this paper, we show how a standard linear mixed model from animal breeding[10] may be used to model correlated traits, while at the same time correcting for dependence among loci (for example, due to population structure). As designs like cohort studies become more prevalent, the need for modeling correlated traits as well as population structure will grow[2,11,12], and the same is true for the increasing number of nonhuman GWAS[13–17].

The mixed model, which handles population structure by estimating the phenotypic covariance that is due to genetic relatedness—or kinship—between individuals, has previously been shown to perform well in GWAS[2,13,18–22]. Here, we extend this approach to handle correlated phenotypes by deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits (Online Methods), implementing it for GWAS. The idea is not new[23–27], but it has never been applied for association mapping on a genome-wide scale. Alternative approaches for GWAS analysis at multiple traits exist, but they generally are unable to control for population structure[28,29], and often are not applicable to genome-wide data.

We validate our approach using extensive simulations based on available SNP data from *A. thaliana*[30], showing that our model increases power to detect associations while controlling the false discovery rate. We then demonstrate its usefulness by considering correlated blood lipid traits from the Northern Finland Birth Cohort 1966 (NFBC1966)[31] and environmental plasticity in an *A. thaliana* data set that contains flowering measurements for two simulated growth seasons in two different locations[32]. Finally, we discuss the usefulness of this approach, not only in terms of increasing power to detect associations, but also in terms of understanding the basic genetic architecture of the phenotypes.

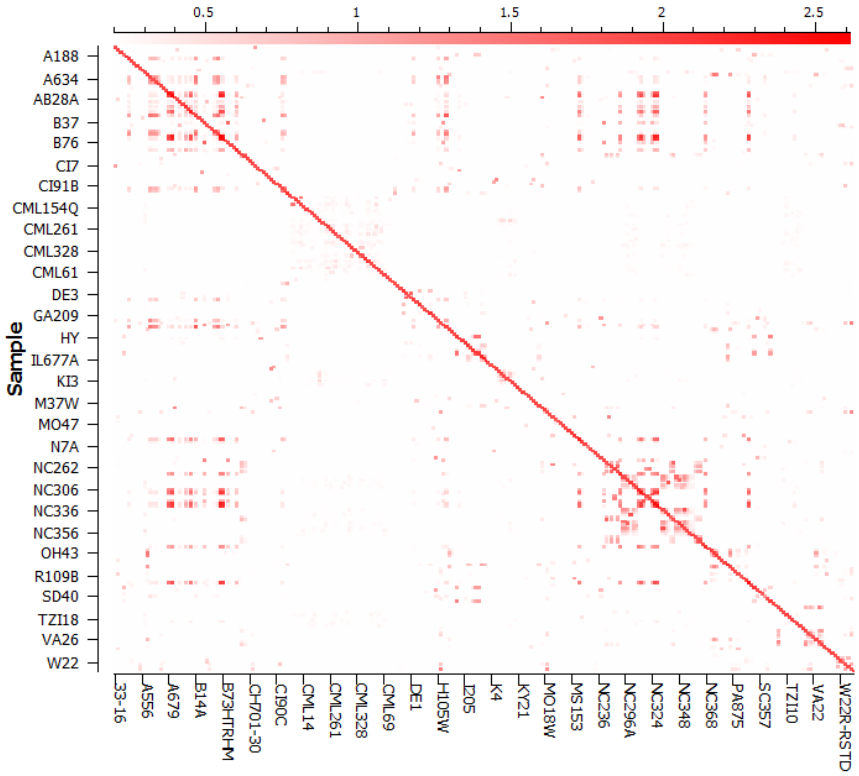
[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria. [2]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. [3]Institut National de la Recherche Agronomique (UNRA), UR0588, Orléans, France. [4]These authors contributed equally to this work. Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oeaw.ac.at).

Received 17 January; accepted 5 July; published online 19 August 2012; doi:10.1038/ng.2376

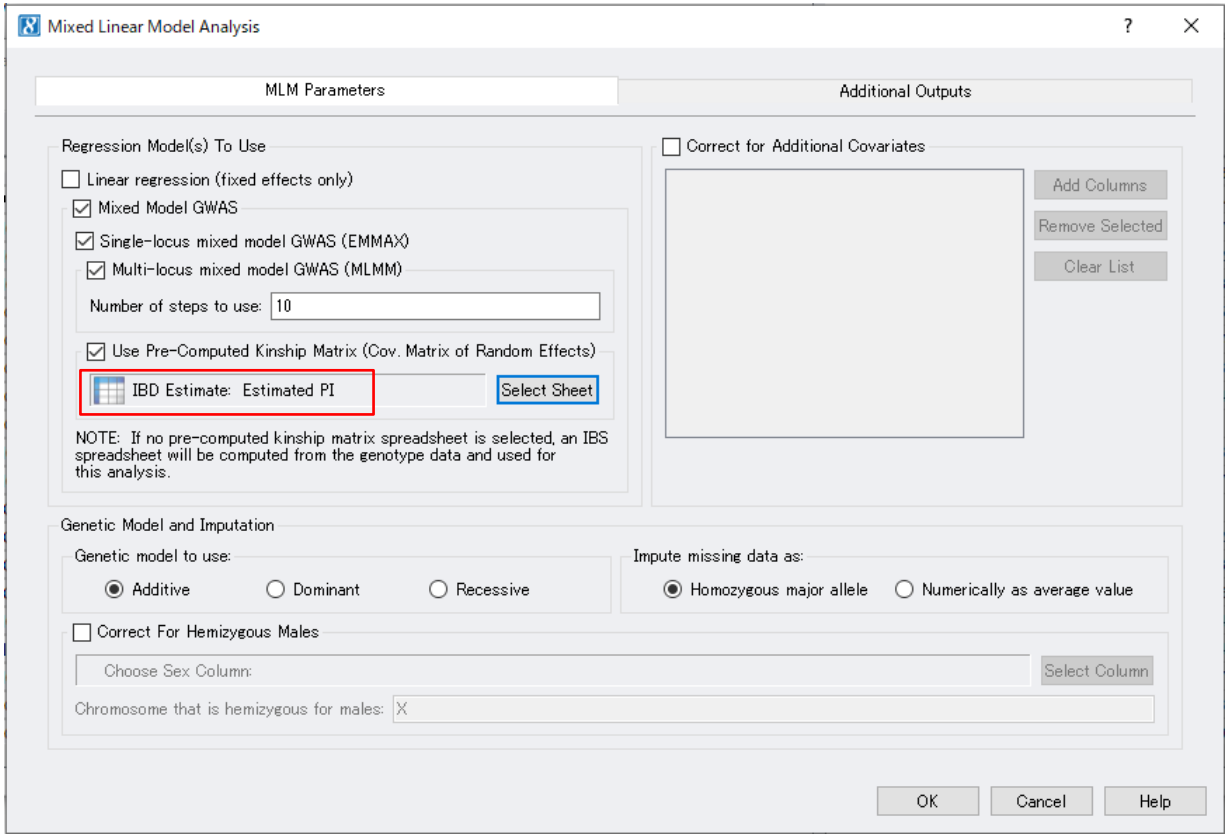
混合モデル法

■ ゲノム関係行列

- SNPデータをもとに、サンプル同士の関係を数値化したもの
- すべてのサンプル同士で計算を行うため、“サンプル数” x “サンプル数” の行列データとなる
- 計算アルゴリズムには、Identical by State (IBS), Identical by Descent (IBD), GBLUPを選択可能
- 計算した行列データは、ワンクリックでヒートマップ表示に切り替え、関係の強度をグラフィカルに表示することが可能



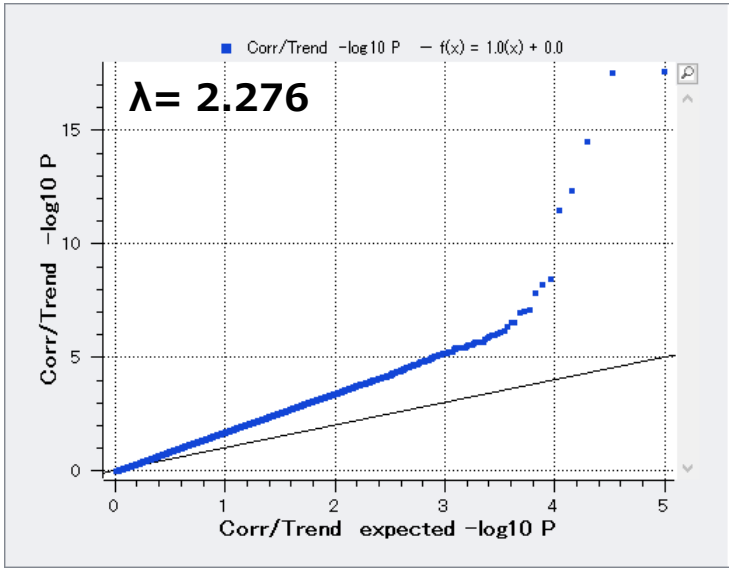
混合モデル法



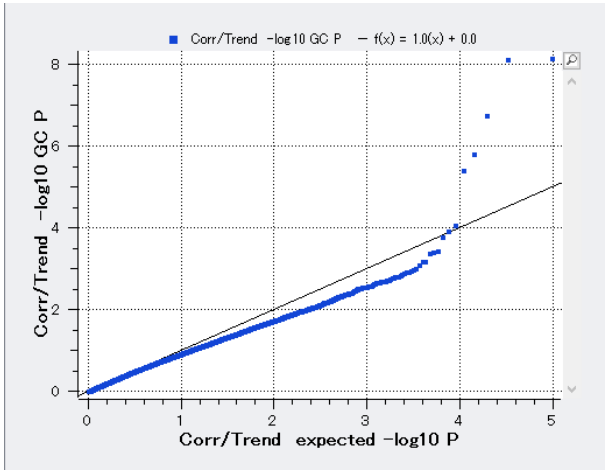
■ Mixed Linear Model Analysis

- 混合モデルでGWAS計算を行うためのツール
- 変量効果として、ゲノム関係行列データを選択して計算を実行
- 共変量データは、固定効果としてパラメータ選択することが可能

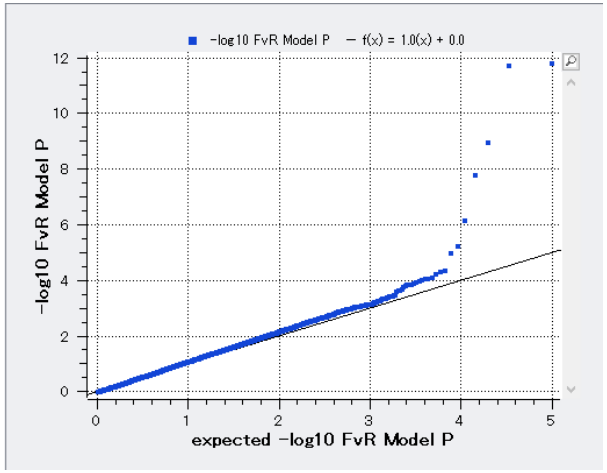
補正結果



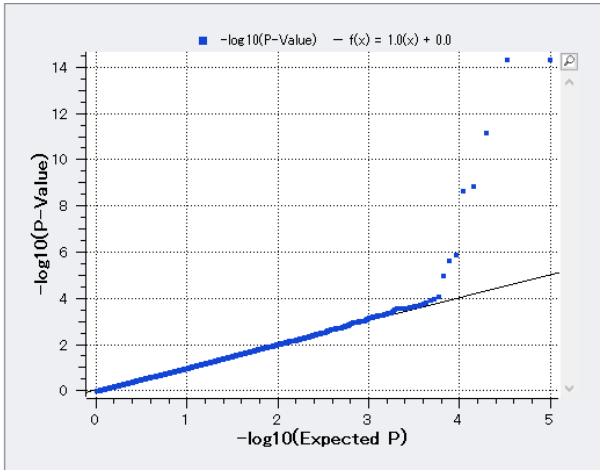
Naïve GWAS



Genomic Control法

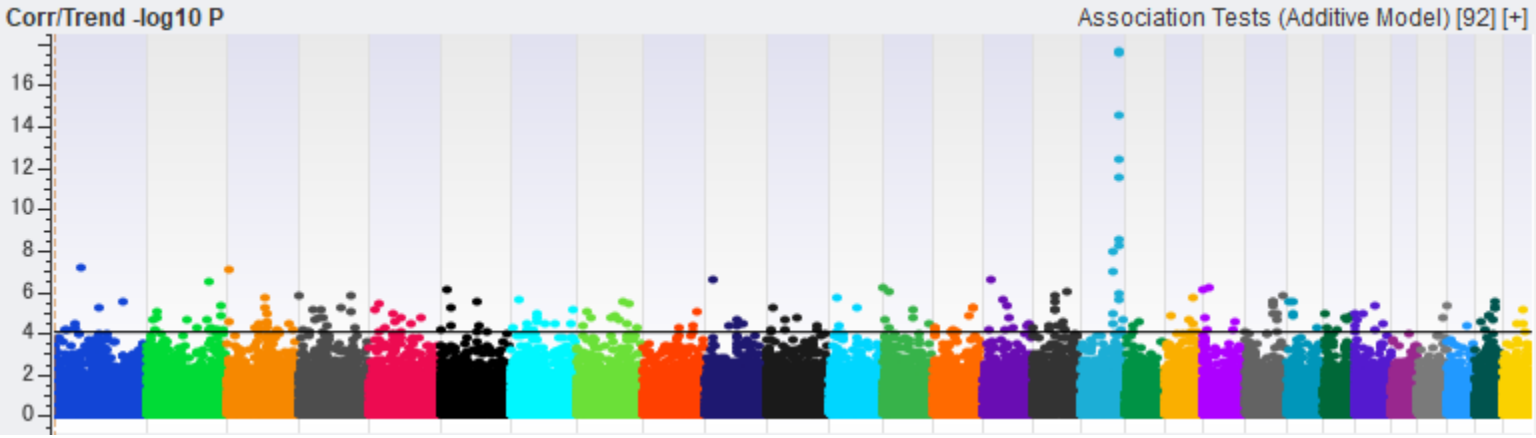


PCA法

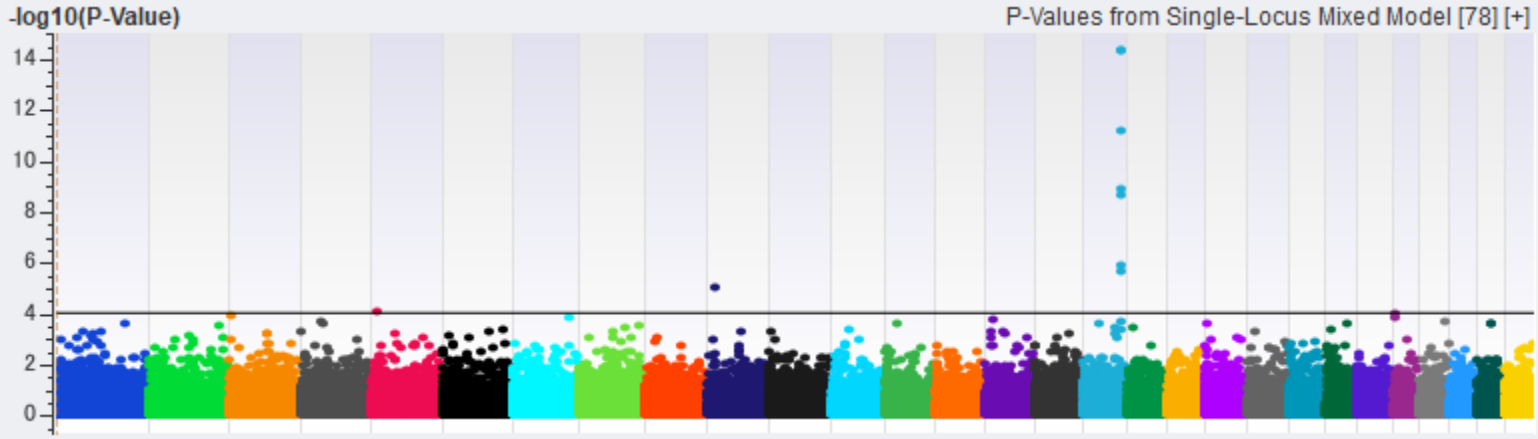


混合モデル法

補正結果



Naïve GWAS



混合モデル法

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