博士論文

ゲノミック予測を用いた育種の効率化・最適化に関する理論的研究

平成 31 年 3 月

東京大学大学院農学生命科学研究科
生産・環境生物学専攻生物測定学研究室
平成 28 年博士課程進学

田中　凌慧

目次

1. 序論

育種とは、動植物が望ましい表現型（phenotype）を示すように、その遺伝子型（genotype）を改良することである。育種学とは、遺伝子型と表現型の関係性について理解し、それを活用すること、あるいはそのための理論・方法を開発する学問分野である。もっとも、遺伝子の存在が認識されるよりも遥かに昔から、育種という営みは表現型選抜（phenotypic selection）として存在した。すなわち、栽培した植物や家畜のうち、優れた性質をもつ個体を選抜し、そうでない個体を淘汰することにより、人類は野生生物を栽培に適した作物・家畜へと改良してきた。

遺伝子型と表現型の関係性を明確化し、集団における遺伝子型の変化と表現型の変化の関係性を記述する理論体系は、量的遺伝学（quantitative genetics）によって築かれ、これが近代的な育種学の礎をなした（Falconer and Mackay, 1996）。量的遺伝学では、メンデルの遺伝法則を集団（population）というレベルに拡張するとともに、表現型が多数の、効果の小さな遺伝子によって支配されることを仮定する（infinitesimal model）。これにより、血縁関係による表現型の相関関係や選抜の効果に関する定量的考察が可能となる。量的遺伝学は、Henderson により開発された混合モデルに基づく統計的推論（Henderson, 1984）と結びつき、とりわけ動物育種を支える理論的基盤となった。

いっぽう、植物育種はやや異なる方向性で発展を遂げた。まず、作物種ごとの遺伝的特性や育種目標を反映した、種々の育種法が開発された。他殖性作物で主に用いられる集団選抜法や、自殖性作物で主に用いられる系統選抜法など、合理的な方法論が確立され実用化されていった（Brown and Caligari, 2013）。これらは、1 回の交配で多数の種子を得られること、自殖や栄養生殖によって遺伝的に同一な個体を多数得られること、種子による個体の長期保存が可能なことなど、動物育種では不可能な多様な実験的操作を前提に開発されたものだといえる。

さらに植物育種に大きな影響を与えたのは、DNA マーカーと QTL（量的遺伝子座：quantitative trait loci）解析に基づくマーカー選抜（marker assisted selection）である（Mauricio, 2001）。QTL 解析は、DNA マーカーの遺伝子型（マーカー遺伝子型；marker genotype）と表現型の関係性を統計モデルに基づき検定することで、表現型に関与する遺伝領域を特定する解析手法である。QTL 解析によって、ある DNA マーカーが目的形質を支配する QTL と強く連鎖していることがわかれば、別の個体の当該マーカー遺伝子型を調べることで、表現型を観測せずとも遺伝子型を改良することができる。

いっぽう、ゲノミック予測（genomic prediction; Meuwissen et al., 2001）は、ゲノム上に配置された DNA マーカー遺伝子型を全て用いて、遺伝子型値を説明する予測モデルを構築する。ゆえに、QTL が多数ある場合には、それら QTL の遺伝子型と相関の強い複数のマーカー遺伝子型を用いて遺伝子型値を算出する統計モデルが得られる。ゲノミック予測を用いて選抜・淘汰を行う育種法はゲノミック選抜（genomic selection）と総称される。マーカー選抜とゲノミック選抜はともに DNA マーカーを用いた選抜法であるが、マーカー選抜が QTL と連鎖する少数の DNA マーカーを用いた選抜法であるのに対し、ゲノミック選抜は過去のデータから構築された、全 DNA マーカーに基づく予測モデルを用いる選抜法である。

ゲノミック予測が提案されてから、研究における主な関心事は、ゲノミック予測が可能なのか、ということであった。DNA マーカーから遺伝子型値を予測する複数の予測モデルが提案され、どのような予測モデルが高い予測精度を実現できるのかが検討される（e.g. Heslot et al., 2012）とともに、モデル構築に用いる系統数、DNA マーカーの種類や密度による予測精度への影響が検証された（Crossa et al., 2013; Hickey et al., 2014）。また、実データに基づく検証は、イネ（Onogi et al., 2014; Spindel et al., 2015）、コムギ（Crossa et al., 2010; Heffner et al., 2011）、トウモロコシ（Zhao et al., 2012）などの主要な穀物に限らず、テンサイ（Würschum et al., 2013）や普通ソバ（Yabe et al., 2018）といった作物、また、果樹（Kumar et al., 2012; Minamikawa et al., 2017）、林木（Grattapaglia et al., 2018）、トマト（Yamamoto et al., 2016）などでも行われている。さらに、米国の乳牛育種では 2008 年からゲノミック選抜が実用化され、2015 年までの統計データによる検証がなされた（García-Ruiz et al., 2016）。検証の結果は、遺伝率が低く改良が困難であった形質を中心に、ゲノミック予測によって年次あたりの改良効率が劇的に向上したことを示すものであった。現在でもゲノミック予測に関する種々の検証・実証は道半ばだが、これまでの多くの研究報告は、ゲノミック予測の幅広い生物種・形質に対する有効性を強く示唆するものである。

ゲノミック予測の有用性が確認されるにつれて、理論・手法研究は新たな段階、ゲノミック予測の実装や活用における最適化へと移りはじめた。ゲノミック予測はあくまで遺伝子型値を予測する道具にすぎない。したがって、ゲノミック選抜を育種法として確立するためには、どのような集団で予測モデルを構築するか、予測モデルに基づきどのように選抜・交配を行うのか、といった一連の意思決定にも何らかの指針を与える必要がある。近年の訓練集団最適化（Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Rincent et al., 2017a）や選抜戦略の最適化に関する研究（Daetwyler et al., 2015; Akdemir and Sánchez, 2016; Lehermeier et al., 2017; Han et al., 2017; Müller et al., 2018）は、シミュレーションや数理最適化の技法を駆使してこれらの課題に取り組むものである。本論文の第２章では、ゲノミック予測で用いられる代表的な統計モデルについて解説するとともに、予測精度に影響を及ぼす因子に関する知見をまとめた。さらに、ゲノミック選抜の最適化に関する先行研究について、一部は理論的背景も含めて解説した。

本論文は、ゲノミック予測に基づく育種を効率化・最適化することを目的とした４つの研究から構成される。第３章と第５章では、それぞれ、訓練集団最適化に関する新規手法の開発および既存手法の拡張を行った。また、第４章ではゲノミック予測における新たな選抜法を提案した。第６章では、育種戦略の最適化を見据えた予測モデルの評価を行った。

第３章では、能動学習（active learning; Settles, 2009）と呼ばれる機械学習分野で開発された手法を、訓練集団最適化に応用した。従来の訓練集団最適化は、混合モデルに基づく回帰により遺伝子型値を予測すること前提としたものであったが、本研究では回帰と並んで重要な分類問題を扱った。４つの実データと１つの仮想データによりシミュレーションを行い、能動学習に基づく訓練集団最適化の有効性を示した。

第４章では、ゲノミック予測を用いて多数の遺伝資源系統から有用系統をスクリーニングする

状況を想定した。想定される状況を black-box 最適化問題と結びつけ、ベイズ最適化（Bayesian optimization; Mockus, 1994）と呼ばれる最適化アルゴリズムを応用することで、予測の不確実性を考慮した、より効率的な選抜法を提案した。提案された指標の有効性を、４つの実データに基づくシミュレーションによって示した。

第５章では再び訓練集団の最適化を扱ったが、ここでは多環境試験における最適化を目標とした。つまり、どの系統を栽培して訓練データとするか、だけでなく、どの環境でどの系統を栽培して訓練データとするか、という多環境試験デザインの最適化に取り組んだ。既存の訓練集団最適化手法を多環境データの解析で用いられるゲノミック予測モデルへ適切に拡張するとともに、４つの実データを用いて評価を行なった。その結果、遺伝子型値の環境間相関（直感的には、当該表現型に注目した場合の、環境間の類似性）に応じて最適な多環境試験のデザインが異なることが示された。

第６章では、後代分離の予測精度という、これまでにほとんど評価されてこなかった尺度で予測モデルの比較を行った。ある交配組み合わせから生じる後代遺伝子型値の分離を予測することは、近年開発が進むゲノミック予測を用いた育種の最適化指標の多くで必要である。したがって、提案された最適化指標を用いてゲノミック選抜を実装する場合には、分離を精度よく予測できる統計モデルを用いなければならない。仮想データを用いた解析により、ゲノミック予測で標準的に用いられるベイズリッジ回帰が、分離予測においては必ずしも優れたモデルではないことを明らかにした。

最後に第７章で本研究の総括を行うとともに、ゲノミック予測に基づく育種学研究が進むべき方向性について議論した。

2. ゲノミック予測の概略

ゲノミック予測の概念と基礎的なモデル化手法は、2001 年に Meuwissen らによって提案された（Meuwissen et al., 2001）。それから今日に至るまで、多くのモデル化手法が提案されるとともに、理論・応用の両面において、様々な研究課題が発見され議論されてきた。本章では、はじめにゲノミック予測の代表的なモデル化手法について説明する。次いで、予測精度に影響する因子（対象形質の遺伝率や DNA マーカーの密度など）について検討した先行研究を概観する。その後は、本論文との関連の深い分野である訓練集団最適化と育種戦略の最適化についてやや詳しく説明する。なかでも訓練集団最適化は本研究との関連が深いため、数式を交えて丁寧に解説する。最後に、複数の目的変数がある場合（遺伝的に相関することが期待される複数の形質を扱う場合や、遺伝子型・環境間交互作用を考慮する場合）のモデル化について簡単に議論する。

以下では各所で必要に応じて原著論文を引用するが、多くの箇所で参考になる review 論文と教科書を、ここであらかじめ列挙しておく。まず、量的遺伝学の基礎的概念については（Falconer and Mackay, 1996）が優れた教科書である。2-1 節で議論されるモデル化手法の多くは（de los Campos et al., 2013b）や（Gianola, 2013）で解説されている。数式の展開を含め、必要なベイズ推論に関する基礎知識は（Bishop, 2006）にほぼ全て網羅されている。また、混合モデル（後述）に関しては（Gianola and Sorensen, 2002）で極めて詳細に議論されている。また、2-2 節では（Desta and Ortiz, 2014）の review 論文を執筆の参考にした。また、（Crossa et al., 2017）では、特に遺伝子型と環境の交互作用に関するモデル化に詳しく触れつつ、ゲノミック予測に関する手法研究や応用例を紹介している。

2-1. 代表的なモデル化手法

2-1-1. マーカー遺伝子型の表記と量的遺伝学における解釈

ゲノミック予測における説明変数 $\mathbf{x}$ と応答変数 y は、それぞれ遺伝子型と表現型値である。応答変数である表現型値は連続変数のこともあれば、順序つき変数であったり、カテゴリカル変数であったりする。ただし、以下では特に断らない限り連続変数であるとする。本論文では、3 章でのみ応答変数が二値変数である場合を扱い、他では全て連続変数の場合を扱う。

古典的な遺伝学において中心的に扱われるのは 2 対立遺伝子（bi-allelic gene）の遺伝子型である。つまり、ある遺伝子は 2 通りの状態 A, B だけを持つと仮定する。ゲノミック予測において典型的に用いられる bi-allelic な SNP（Single Nucleotide Polymorphism; 一塩基多型）マーカー遺伝子型も、同様に 2 つの状態のみをもつ。以下では、bi-allelic な遺伝子型の符号化法と、その遺伝学的解釈について簡潔に述べる。

生物のゲノムは一般に 4 通りの塩基から構成される 2 本鎖の塩基配列で構成される。いま、ある塩基対が 4 通りの塩基のうち 2 通りの塩基だけで構成される bi-allelic な場合を考える。このとき、2 通りの塩基を A, B と表現するならば、その組み合わせである SNP マーカー遺伝子型は

{AA, AB, BB} という３通りで記述できる。

ある SNP の３通りの状態 {AA, AB, BB} には複数の表現が考えられるが、主に遺伝学的な解釈を容易にするため、SNP が表現型に及ぼす効果を相加効果と優性効果に分割する。以下のように、ある SNP の３つの状態に対して、相加効果を捉えるためのマーカー遺伝子型 x_a を

$$x_a \in \{-1, 0, +1\} \equiv \{\mathrm{AA}, \mathrm{AB}, \mathrm{BB}\} \tag{2.1}$$

と定め、優性効果を捉えるための遺伝子型 x_d を

$$x_d \in \{0, 1, 0\} \equiv \{\mathrm{AA}, \mathrm{AB}, \mathrm{BB}\} \tag{2.2}$$

と定めることでこれは実現できる。なお、当然ながら 1 つの個体で遺伝子型が $x_a = 0$ かつ $x_d = 0$ などとなることはない。このとき、注目している集団において、以下のように表現型値を遺伝子型に回帰することを考える。

$$y = a \cdot x_a + d \cdot x_d + e \tag{2.3}$$

ここで、表現型値を y, 遺伝子型を x_a および x_d, 環境効果による残差を e と表記した。このとき得られる回帰係数 a および d は、その遺伝子型を持つ個体の遺伝子型値を定める。すなわち、遺伝子型 AA をもつ個体の遺伝子型値は $-a$ であり、遺伝子型 AB をもつ個体の遺伝子型値は d であり、遺伝子型 BB をもつ個体の遺伝子型値は a となる。少々厳密さを欠く表現ではあるが、本論文の範囲では、表現型値のうち遺伝的に定まる値を遺伝子型値と考えて差し支えない。

なお、古典的な量的遺伝学で定義される「遺伝子置換の平均効果」は、以下のように、相加的な遺伝子型の表現のみによって回帰したときの回帰係数 α であり、上式の遺伝子型値とは異なる。

$$y = \mu + \alpha \cdot x_a + e \tag{2.4}$$

また、この回帰式における切片 μ は遺伝子型値の集団平均と呼ばれる。例えば、遺伝子型 AA, AB, BB の個体の集団内での頻度が p^2, $2pq$, q^2 であるとき、遺伝子型値の集団平均は $\mu = a(p-q) + 2dpq$ であり、遺伝子置換の平均効果は $\alpha = a + d(p-q)$ である。このように、遺伝子置換の平均効果は、アリル頻度という集団の性質に依存する量である。

古典的な集団遺伝学や量的遺伝学においては、DNA 配列から遺伝子型 x_a, x_d を実際に確認することは想定されておらず、ある集団に対して適切な実験を行うことによって表現型を測定し、その表現型と遺伝学的な理論式から遺伝子置換の平均効果を測定することなどが可能な操作だと想定されている。したがって、式(2.4)で定義される集団平均や遺伝子置換の平均効果を用いて議論を進めることが普通である。現在のシーケンス技術を用いれば、我々は SNP マーカーの遺伝子型を実際に確認することができるため、必ずしもこの表式にこだわる必要はないと考えられる。しかしながら、ゲノミック予測の多くの研究においては依然として式(2.4)を出発点とすることが多い。量的遺伝学で古くから議論されている種々の状況、例えば表現型に基づく上位個体の選抜と

無作為交配に基づく遺伝改良、では遺伝子置換の平均効果を用いた理論に一日の長があることも理由の 1 つであろう。

本論文でも、以下では式(2.4)によって定義される遺伝子置換の平均効果を暗に用いて議論を進める。表記を簡単にするため、これ以降では SNP マーカーの相加的な遺伝子型 x_a を単に x と書くことにする。また、回帰式によって計算される遺伝子型の回帰係数 α を、単にマーカー効果と呼ぶことにする。さらに、遺伝子型値という用語を、古典的な量的遺伝学の定義にとらわれず、表現型値を遺伝子型値によって統計的にモデル化したとき、遺伝的に決まる値を指して用いる。これらはいずれも、ゲノミック予測における諸研究の慣例に従ったものである。

2-1-2. マーカー回帰

ゲノムワイドな SNP マーカーの状態と表現型値との関係を適切にモデル化することにより、SNP マーカー遺伝子型から個体の遺伝子型値を推定・予測することが可能だと考えられる。我々の対象とする生物ゲノム中には多数の一塩基多型が存在することから、染色体上で密に配置された SNP マーカーの一部は表現型を支配する遺伝子座の近傍に位置することが期待される。染色体上で近い位置にある塩基は連鎖して子に伝わる確率が高く、したがって、真の遺伝子座とその近傍にある SNP マーカーの遺伝子型は、集団内で強い相関を持つことが期待される。このような 2 つの SNP（あるいは遺伝子）間の関係を連鎖不平衡（linkage disequilibrium）と呼ぶ。なお、2 つの SNP が異なる染色体上に存在し、遺伝子型に相関がない場合は連鎖平衡の関係にあると呼ぶ。すなわち、ゲノミック予測は、形質を支配する真の遺伝子座と連鎖不平衡の関係にある SNP マーカーを用いて表現型を予測する方法だということができる（ただし 2-2-1 節で述べるように、この説明は厳密には正しくない）。

いま、ゲノミック予測の対象となる M 系統について、全部で N 個体の表現型値を計測したとする。系統数 M と個体数 N の大小関係は自由であることをあらかじめ指摘しておく。予測したい M 系統の一部について 1 個体ずつ栽培して表現型を得た場合には $M > N$ である（あるいは、複数個体の表現型値の平均値だけをモデル構築に用いる場合にも同様である）し、複数個体の表現型を測定した場合には $M < N$ になりうる。なお、同じ系統は同じマーカー遺伝子型を持つため、遺伝子型値は系統ごとに定めれば十分である。このとき、N 次元の表現型ベクトルを $\mathbf{y}$、M 次元の遺伝子型値ベクトルを $\mathbf{u}$、N 次元の残差ベクトルを $\mathbf{e}$、集団平均（スカラー）を μ とすると

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{2.5}$$

というモデルを考えることができる。なお、特に断らない限り、ベクトルは縦ベクトルとする。ここで、$\mathbf{1}$ はすべての要素が 1 である N 次元ベクトルであり、行列 $\mathbf{Z}$ は $\mathbf{y}$ と $\mathbf{u}$ の要素の対応を表す計画行列である。

さらに、ゲノミック予測では異なる環境や栽培ブロックで計測された表現値を扱うことが普通である。そこで、こうした環境効果を考慮した

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{2.6}$$

という式が標準的に用いられる。ここで、$\boldsymbol{\beta}$ は環境効果のベクトルである。行列 $\mathbf{X}$ は $\mathbf{y}$ と $\boldsymbol{\beta}$ の要素の対応を表す計画行列である。環境効果のベクトルは L 次元であるとする。なお、行列 $\mathbf{X}$ の要素はマーカー遺伝子型 x ではないことに注意せよ。

ゲノミック予測では、環境効果の総数 L は個体数 N や系統数 M よりも小さく、環境効果 $\boldsymbol{\beta}$ は推定可能な母数効果と考えることが一般的である。いっぽう、先述の通り系統数 M は個体数 N と同じかそれよりも多いことが想定されるため、遺伝子型値 $\mathbf{u}$ は何らかの分布に従う変量効果としてモデル化される。このとき、式(2.6)は母数効果と変量効果の両方を含むため、混合モデル（mixed model）と呼ばれる。

残差がすべての個体に対して独立に同一の正規分布に従うことを仮定すると、混合モデルの尤度は、平均 $\mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$、分散共分散行列 $\mathbf{I}\sigma_e^2$ を持つ多変量正規分布

$$p(\mathbf{y}|\mu, \mathbf{u}, \boldsymbol{\beta}, \sigma_e^2) = N(\mathbf{y}|\mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I}\sigma_e^2) \tag{2.7}$$

となる。ここで、$\mathbf{I}$ は単位行列であり、σ_e^2 は残差分散である。以下、単位行列はその次元によらず $\mathbf{I}$ と表記する。次元を明記する必要性が高い場合には右下に添字で表記する。残差分散（および、後述するマーカー効果の分散や遺伝分散）は最尤法や制限付き最尤法（REML; REstricted Maximum Likelihood）による推定が行われることもあるが、事前分布を採用して他のパラメータと同様にベイズ推定することもしばしば行われる。例えば、共役事前分布として逆カイ二乗分布

$$p(\sigma_e^2) = \chi^{-2}(\sigma_e^2|\nu_e, \tau_e) \tag{2.8}$$

を採用することが一般的である。ここで、ν_e および τ_e は自由度およびスケールに対応する超パラメータである。

マーカー回帰と総称されるモデル群では、通常、遺伝子型値がマーカー効果の線形和によって定まることを仮定する。すなわち、ある系統 i の遺伝子型値 u_i $(i = 1,2,..., n)$について

$$u_i = \sum_{j=1}^{P} \alpha_j x_{ij} \tag{2.9}$$

というモデルを考える。ここで、P は SNP マーカーの総数であり、α_j は j 番目のマーカーの効果、x_{ij} は系統 i の j 番目のマーカーの遺伝子型である。

なお、マーカー遺伝子型には各 SNP について平均 0、分散 1 に標準化された値を用いることもできるが、その場合には先に述べたような遺伝学的解釈に影響を及ぼすと考えられる。また、事前分布やその超パラメータの設計にも標準化の有無は影響すると思われる。このようにマーカー

回帰におけるマーカー遺伝子型の符号化法とモデル化手法の関係はやや難しい問題を孕んでおり、その考察は本研究の範囲を超える。なお、本論文では 6 章でのみマーカー回帰を用いるが、そこでは標準化を行わずにマーカー遺伝子型を用いた。

マーカー回帰モデルは、マーカー効果 α_j の従う事前分布によって特徴付けられる。なお、同じ名称のモデルであっても、超パラメータの設計や推定方法、あるいはモデルそのものが論文や年代によってわずかに異なることがある。以下では、ベイズ回帰を行う R パッケージ {BGLR} における名称と実装に基づき解説を行う（Perez et al., 2014）。これにより、後述する BayesB では、同名のモデルが初めて提案された Muewissen らの論文における同じ呼称のモデルとは異なるので、特に注意されたい。はじめに提案された BayesB が異なるモデルに置き換わった背景については、（Gianola at al., 2009）による議論を参照されたい。

最もよく使われるマーカー回帰モデルとして、ベイズリッジ回帰（BRR; Bayesian Ridge Regression）が挙げられる。BRR では、マーカー効果が正規分布に従うことを仮定する。すなわち、

$$p(\alpha_j|\sigma_\alpha^2) = N(\alpha_j|0, \sigma_\alpha^2) \tag{2.10}$$

$$p(\sigma_\alpha^2) = \chi^{-2}(\sigma_\alpha^2|\nu_\alpha, \tau_\alpha) \tag{2.11}$$

であるとする。このとき、回帰によって得られるマーカー効果はほぼすべて非ゼロの値をとる。言い換えれば、このモデルの下では多数の SNP マーカーが微小な効果を持ち、それが相加的に遺伝子型値に寄与する。これは、古くから量的遺伝学で想定されてきた infinitesimal model の想定に当てはまる。なお、名称の示す通り、本モデルはリッジ回帰のベイズ的表現になっており、リッジ回帰における正則化パラメータを交差検証ではなく階層ベイズ法により推定している。

あるいは、Bayesian LASSO と呼ばれるモデルでは、マーカー効果に二重指数分布

$$p(\alpha_j|\tau_j^2) = N(\alpha_j|0, \sigma_e^2 \cdot \tau_j^2) \tag{2.12}$$

$$p(\tau_j^2) = \mathrm{Exp}\left(\tau_j^2|\frac{\lambda^2}{2}\right) \tag{2.13}$$

を仮定する。ここで、Exp(·)は指数分布である。なお、この λ を超パラメータとすることもあれば、さらに超事前分布を採用することもある（Perez et al., 2012; Gianola, 2013）。Bayesian LASSO はその名前の通り、LASSO（Least Absolute Shrinkage and Selection Operator）回帰をベイズ的に記述したモデルとなっており、BRR と比べて、比較的少数のマーカーに大きな効果を、多くのマーカーに 0 に近い効果を与える傾向にある。このため、Bayesian LASSO は、効果の大きな主働遺伝子の存在を暗に仮定したモデルと考えることができる。

同様に BayesA, BayesB と呼ばれるモデルも、主働遺伝子の存在を想定したモデル化手法である。BayesA は、マーカー効果の周辺分布が t 分布となるような以下の階層ベイズモデルを用い

る。

$$p\left(\alpha_j|\sigma^2_{\alpha_j}\right) = N\left(\alpha_j|0, \sigma^2_{\alpha_j}\right) \tag{2.14}$$

$$p\left(\sigma^2_{\alpha_j}\right) = \chi^{-2}\left(\sigma^2_{\alpha_j}|\nu_\alpha, \tau_\alpha\right) \tag{2.15}$$

BRR と異なり、BayesA ではマーカー効果の分散にも添字 j が用いられており、異なるマーカーが異なる分散を持つ正規分布から生成される。これにより、一部のマーカーが大きな効果を持ちやすくなっている。

BayesA は、マーカー効果の周辺分布が t 分布であることからもわかるように、主働遺伝子の存在を仮定しながらも、ほぼすべてのマーカーに非ゼロな効果が与えられる。いっぽう、BayesB では、マーカー効果の事前分布を以下のような混合分布で表現する。

$$p\left(\alpha_j|\sigma^2_{\alpha_j}, \pi\right) = \pi \cdot N\left(\alpha_j|0, \sigma^2_{\alpha_j}\right) + (1-\pi) \cdot \delta\left(\alpha_j\right) \tag{2.16}$$

$$p\left(\sigma^2_{\alpha_j}\right) = \chi^{-2}\left(\sigma^2_{\alpha_j}|\nu_\alpha, \tau_\alpha\right) \tag{2.17}$$

$$p(\pi) = \mathrm{Beta}(\pi|p_0, \pi_0) \tag{2.18}$$

ここで、$\delta(\cdot)$ はディラックのデルタ（超）関数であり、$\mathrm{Beta}(\cdot|p_0, \pi_0)$ は期待値を π_0、分散を π_0*(1 $-\pi_0$) / (p_0+1)とするベータ分布である。式(2.16)より、BayesB では混合割合 π でマーカーの効果が非ゼロになることがわかる。

2-1-3. カーネル回帰

上述したマーカー回帰では、マーカー効果をパラメータとし、その線形和として遺伝子型値をモデル化した。いっぽうで、GBLUP（Genomic Best Linear Unbiased Predictor）や RKHS 回帰（再生核ヒルベルト空間回帰；Reproducing Kernel Hilbert Space Regression）と呼ばれる手法では、マーカー効果を経由することなく、系統ごとの遺伝子型値を推定する。ゲノミック予測の主目的は遺伝子型値の推定・予測であり、マーカー効果の推定はゲノミック予測における興味の 1 つであるものの、必須ではないことも多い。これらの手法はともにカーネル関数（後述）を用いて予測モデルが記述されるため、まとめてカーネル回帰と呼称する。カーネル回帰は Gianola らによってゲノミック予測に導入された（Gianola et al., 2006）。他の手法との関連など、本節よりも詳細な解説は（Morota and Gianola, 2014）を参照せよ。

GBLUP は、マーカー回帰における BRR のカーネル回帰版とも言えるモデルである。式(2.9), (2.10)が成り立つとき、遺伝子型値のベクトル $\mathbf{u}$ が従う分布は、以下のような多変量正規分布になることが簡単な計算によってわかる。

$$p(\mathbf{u}|\sigma_\alpha^2) = N(\mathbf{u}|0, \mathbf{M}\mathbf{M}^{\mathrm{T}}\sigma_\alpha^2) \tag{2.19}$$

ここで、$\mathbf{M}$ はマーカー遺伝子型行列であり、その i, j 要素は x_{ij} である。すなわち、BRR モデルにおいて、遺伝子型値の分散共分散行列はマーカー遺伝子型行列の積の形で定まる。GBLUP でも同様に、遺伝子型値が正規分布に従うことを仮定する。

$$p(\mathbf{u}|\sigma_u^2) = N(\mathbf{u}|0, \mathbf{G}\sigma_u^2) \tag{2.20}$$

このとき、分散共分散を定める行列 $\mathbf{G}$ をマーカー遺伝子型に基づき以下のように計算する。まず、j 番目のマーカーのアリル頻度を p_j とする。すなわち、j 番目のマーカー遺伝子型が−1 である系統の数を m_1、0 である系統の数を m_2 、+1 である系統の数を m_3 として

$$p_j = \frac{m_2 + 2m_3}{2(m_1 + m_2 + m_3)} \tag{2.21}$$

によってアリル頻度を計算する。このアリル頻度とマーカー遺伝子型を用いて、行列 $\mathbf{G}$ の i, j 成分 G_{ij} を

$$G_{ij} = \frac{x_{ij} + 1 - 2p_j}{2\sum_j p_j(1 - p_j)} \tag{2.22}$$

このように定める。このとき、行列 $\mathbf{G}$ をゲノム関係行列と呼び、分散パラメータ σ_u^2 を（genomic かつ相加的な）遺伝分散と呼ぶ。GBLUP と BRR の違いは、マーカー遺伝子型に対して適切な標準化を行うことで、行列 $\mathbf{G}$ および分散パラメータ σ_u^2 が古典的な量的遺伝学における血縁（家系）行列や遺伝分散に対応づけられることにある（Endelman and Jannink, 2013）。

実際には、式(2.6)の遺伝子型値ベクトルが従う分散共分散として血縁行列を用いて、遺伝子型値の BLUP を点推定することが長らく研究されてきたという歴史がある（これを pedigree-BLUP と呼称する）。なお、BLUP は本来であれば推定量そのものを指す用語であるが、しばしば「BLUP 法により推定する」「BLUP モデルを用いる」などと、混合モデルの変量効果の BLUP を推定することを表現する。本稿でも、以降はこのような表現を用いることがある。

なお、マーカー回帰の場合と同じく、遺伝分散の推定には REML 法や共役事前分布

$$p(\sigma_u^2) = \chi^{-2}(\sigma_u^2|\nu_u, \tau_u) \tag{2.23}$$

を用いた階層ベイズモデルが用いられる。GBLUP は残差分散と遺伝分散を既知とした場合、パラメータのベクトルが正規分布に従う。そのため数学的な取り扱いが容易であり、多くの統計量が

解析的に求まる(例えば2-2-2節など)。量的遺伝学との対応関係が明瞭であることもあり、GBLUPはゲノミック予測の理論研究においてしばしば利用される。本論文においても４章、５章においてGBLUPに準ずるモデル化を用いる。

もう１つのカーネル回帰として、ゲノミック予測においてRKHS回帰と呼称されるモデルについて最後に解説する。RKHS回帰でも式(2.20)と同様の形式

$$p(\mathbf{u}|\sigma_u^2) = N(\mathbf{u}|0, \mathbf{K}\sigma_u^2) \tag{2.24}$$

によって遺伝子型値の事前分布を定める。ただし、RKHS回帰で分散共分散を指定する行列 $\mathbf{K}$ の成分は、カーネル関数と呼ばれる適切な関数 $k(\mathbf{x}_i, \mathbf{x}_j)$ によって

$$\mathrm{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{2.25}$$

と定まる。ここで $\mathbf{x}_i$ は系統 i のマーカー遺伝子型ベクトルであり、$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP})^{\mathrm{T}}$ と定義される。カーネル関数によって構成される行列 $\mathbf{K}$ はグラム行列と呼ばれる。カーネル関数は、グラム行列が半正定値となる（つまり逆行列が存在する）ように設計される。最もよく用いられるカーネル関数はガウスカーネルであり

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{h}\right\} \tag{2.26}$$

と定義される。ここで、$\|\mathbf{x}\|^2$ はベクトルの要素の二乗和である。また、h は超パラメータである。式(2.22)で与えられる行列 $\mathbf{G}$ とは異なり、ガウスカーネルはマーカー遺伝子型に対して線形ではない。したがって、RKHS回帰は、エピスタシス（マーカー間の交互作用）を考慮することができるモデルだと解釈することができる。ただし、単に式(2.26)でカーネルを構成したとしても、エピスタシスをうまく取り込めるとは限らず、エピスタシスを考慮するにはより洗練された方法を用いることが望ましい（Jiang and Reif, 2015）。

なお、機械学習においてはガウスカーネルの他にも様々なカーネルが提案され用いられており、その多くはゲノミック予測においても適用可能であると推察される。したがって、式(2.26)を用いたモデルは「ガウスカーネルを用いたRKHS回帰」と常に明記するべきであるが、今日のゲノミック予測では、RKHS回帰としてもっぱらガウスカーネルが用いられる。また、GBULPもカーネル関数の形で書くことができるため、広い意味ではRKHS回帰と呼んで差し支えないが、他の多くの文献と同様に区別して用いる。GBLUPはマーカーに対して線形なモデルであり、RKHS回帰は非線形なモデルであり、両者を区別するのは妥当な表現だと考える。本論文では、RKHS回帰という用語を、非線形なカーネルを用いたRKHS回帰の意味で用いる。

ここで取り上げた予測モデルの他にも、BayesCやBayesRなどのBayesian alphabetと総称されるマーカー回帰手法が提案されている（Gianola, 2013）ほか、Random Forest（Breiman, 2001）やSVM（Support Vector Machine; Cortes and Vapnik, 1995）などの機械学習手法などがしばしばゲノミック予測に用いられている。

2-2. 代表的な研究課題

2-2-1. 予測精度に影響する諸因子

モデルの予測精度が高いほどゲノミック予測に基づく選抜は確からしいものとなる。したがって、多くの研究で予測モデルの精度に関わる因子やその影響について検討されてきた。なお、実データを用いた解析における予測精度の評価は、交差検証法を用いて予測値と実測値の間のピアソン相関係数を計算することによって行われる。以下でも、特に断らない限り、実データ解析における予測精度は交差検証法により得られたこの相関係数を指すものとする。ただし、仮想データ（シミュレーションデータ）の場合には、解析者が真の遺伝子型値を知ることができるため、予測値と真値の相関係数が用いられる。

モデル構築に用いる DNA マーカーの数については、経験的に、多ければ多いほど予測精度が向上するとされている。ただし、複数の研究結果を比較すると、その度合いは対象とする作物や形質によって異なることがわかる。コムギを用いた Heffner らの研究では、予測に用いる DNA マーカーの数を 1158 から 192 に減らしたとしても、予測精度は 10%程度の低下に留まった（Heffner et al., 2011）。トウモロコシを用いた Zhao らの研究でも同様で、マーカー数による予測精度の変化は僅かであった（Zhao et al., 2011）。いっぽう、エンバクのデータを用いた研究例では、マーカー数による予測精度の応答は形質によって大きく異なっていた（Asoro et al., 2011）。例えば出穂期や収量の予測ではマーカー数の増加が予測精度の向上に大きく寄与したが、β グルカン含有量や草丈の予測精度では向上の度合いが小さかった。

モデル構築に用いる訓練集団のサイズ（集団に含まれる系統の数）についても、大きければ大きいほど予測精度が向上する。上に挙げた 3 つの研究でもこの傾向は確かめられており、すべての作物で訓練集団のサイズが大きくなるにつれて予測精度が向上した（Heffner et al., 2011; Zhao et al., 2011; Asoro et al., 2011）。その影響の度合いは殆どの形質でマーカー数の影響よりも大きかった。しかしながら、マーカー数の場合と同様に、形質によって影響の大きさは異なった。例えば上述のエンバクに関する研究例では、マーカー数を増加させても予測精度が向上しなかった β グルカン含有量などの形質であっても、訓練集団サイズの増加に伴って予測精度が大きく改善した。いっぽう、草丈や収量でも予測精度の向上が見られたが、その度合いは他の形質に比べて小さかった（Asoro et al., 2011）。

また、訓練集団と予測の対象となる集団との遺伝的背景の類似性も予測精度に影響を及ぼすことが示唆されており、類似性が大きいほど予測精度が高くなると考えられている。例えば、上述したエンバクに関する研究では、いくつかの分集団を定義して分集団間での予測精度を検証した。その結果、遺伝的な類似性の高い分集団間では予測精度が高く、類似性の低い分集団間では予測精度が低かった（Asoro et al., 2011）。また、テンサイにおいても複数の分集団を用いた類似の解析が行われ、遺伝的多様性の高い集団でモデルと構築することで、集団構造の異なる複数の集団に対して比較的安定した予測精度を得られることが示された（Würschum et al., 2013）。

予測に用いるモデルもまた、言うまでもなく予測精度に影響する。ただし、モデル間の優劣は、モデルを適用するデータの性質に依存する。各モデルにおいて想定されている遺伝様式（主働遺伝子やエピスタシスの有無など）が実際のデータに当てはまれば、予測精度は高くなりやすいと考えられる。したがって、予測モデルの決定にあたっては、実際のデータに対して複数のモデルを適用して交差検証によって予測精度を評価することが望ましいだろう。モデル間の予測精度を比較検討した研究例としては、例えば（Heslot et al., 2012）や（Onogi et al., 2014）などが挙げられる。

以上にあげた例をはじめ、多くの研究で実データを用いた予測精度の評価が行われている。いっぽうで、理論的な側面を強く持つ研究も少数ながら存在する。それらの理論研究では、主にGBLUP モデルにおける解析計算や、シミュレーションデータを用いた検証が行われている。以下では特に興味深い研究例を取り上げる。

マーカー数の増加による予測精度の向上は、形質に関与する真の遺伝子（または QTL）と強い連鎖不平衡にあるマーカーが存在する可能性が高くなっていることが理由の 1 つと考えられる。しかしながら、理論的に連鎖不平衡が予測精度に及ぼす影響を検討した例はほとんど見当たらない。唯一の例として、de los Campos らは、SNP マーカーが原因 QTL と完全連鎖している場合と連鎖不平衡がある場合について、GBLUP による予測精度（この研究では、相関係数ではなく決定係数）の差に関する理論式を導出している（de los Campos et al., 2013a）。ただし、訓練集団では QTL ベースの遺伝共分散を正確に推定できていること、QTL ベースの遺伝共分散とマーカーによる遺伝共分散が線形な関係を持つこと、などのやや極端な単純化・仮定を採用しており、実際のデータで生じる予測精度の向上と比較することは難しいと考えられる。

ところで、DNA マーカーと原因 QTL の連鎖不平衡が全くない（連鎖平衡の関係にある）場合であっても、GBLUP による遺伝子型値の予測精度は 0 にはならないことがある。これは、SNP マーカーによって個体間の血縁関係が推定できることに起因する。つまり、予測対象の系統と血縁関係にある系統が訓練集団に含まれていれば、その系統の表現型を参照することで一定の予測ができる。より具体的には、じゅうぶん多くのマーカーがあれば、マーカーと QTL が連鎖していなくても、ゲノム関係行列の期待値が血縁行列の線形な式で記述できる（Habier et al., 2007）。このように、ゲノミック予測が連鎖不平衡を用いた予測であると断言するのは適切ではなく、血縁関係と連鎖不平衡の両方を用いた予測だと捉えるべきである。実際に、これより後に Habier らにより行われた解析では、QTL の存在する染色体にマーカーが存在しない場合をシミュレーションしているが、予測精度は 0 ではない（Habier et al., 2013）。同論文ではモデル間の比較や訓練データ数との関係も検討されており、ゲノミック予測において連鎖不平衡とゲノム関係行列の果たす役割について、シミュレーションに基づき詳細な考察がなされている。

最後に、このように相関係数ベースで定義された予測精度について多くの研究がなされているが、相関係数が絶対的な評価指標とは限らないことを補足しておく。例えば Gonzalez と Forni（2011）は、複数の予測手法について、遺伝子型値と予測値のピアソン相関係数、および、 遺伝子型値上位と下位の二値分類問題を考えた場合の AUC（Area Under the receiving operating characteristic Curve）による評価を行い、相関係数と AUC で異なる予測モデルが支持されうる

ことを指摘している。実際の育種で重要なのは選抜の正確さであることも多く、選抜と淘汰の二値分類精度こそが重要とも考えられる。本論文の３章においては、この点を踏まえて研究を行った。

2-2-2. 訓練集団の最適化

ゲノミック予測の精度が訓練集団のサイズに依存することは先に述べたが、同じ系統数でも、どの系統を訓練集団として用いるかによって、予測精度は異なる。いま、マーカー遺伝子型が得られている系統が 3,000 系統あるが、予算や労力の制約により、育種プログラムで圃場試験できるのは 100 系統だけだしよう。なお、これは現実にも十分に起こりうる。例えばイネ遺伝資源について約 3,000 系統のマーカー遺伝子型が公開されており、これは誰でも利用可能である (Wang et al., 2018)。このとき、栽培試験される系統の組み合わせは ${}_{3000}C_{100} \fallingdotseq 10^{189}$ 通りであり、中には予測精度の高いモデルを構築できる組み合わせもあれば、予測精度の低いモデルが得られてしまう組み合わせもある。

このような状況を踏まえて、近年、マーカー遺伝子型 (あるいは、ゲノム関係行列や血縁行列) をもとに予測精度を高めるような訓練集団を選ぶ研究が進展している（Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Rincent et al., 2017a)。この研究は訓練集団最適化（training set optimization）と総称される。本論文では、５章において訓練集団最適化の研究を多環境試験の実験計画に応用する。また、３章で議論される能動学習とも深い関わりをもち、４章で議論されるベイズ最適化ともわずかながら関連性を持つ。その詳細は各章で議論することにして、ここでは、訓練集団最適化に関する基礎的理論について解説する。

訓練集団最適化の理論では「遺伝子型値 $\mathbf{u}$ の推定値 $\mathbf{u}^*$を得たとき、真の遺伝子型値 $\mathbf{u}$ が、推定値 $\mathbf{u}^*$によってどのくらい説明されるか」という問いを考える。ここまではベイズ的な立場で予測モデルについて紹介したが、本節では頻度論的な視点で議論を進める必要があるので注意されたい。ここで、予測モデルとして GBLUP または pedigree-BLUP を仮定する。また、遺伝分散 σ_u^2 や環境分散 σ_e^2 は既知である、あるいは、REML 法などで点推定された値が十分に確からしいため定数とみなせると仮定する。このとき、GBLUP モデルを、改めて以下のように書き下すことができる。

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \tag{2.27}$$

$$p(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) = N(\mathbf{y}|\mathbf{X\beta} + \mathbf{Zu}, \mathbf{R_0}) \tag{2.28}$$

$$p(\mathbf{u}) = N(\mathbf{u}|0, \mathbf{G_0}) \tag{2.29}$$

ここで、式(2.6)では明示していた集団平均を $\boldsymbol{\beta}$ に含めた。また、分散成分は既知とするため、関係行列と合わせて１つの分散共分散行列として表現した。なお、残差の分散共分散行列は必ずしも対角行列である必要はないため、行列 $\mathbf{R_0}$ は分散共分散行列と解釈できる任意の行列である。

このとき、母数効果（環境効果のベクトル）$\boldsymbol{\beta}$ の最良線形不偏推定量（BLUE; Best Linear Unbiased Estimator）を $\boldsymbol{\beta}^*$とし、および、変量効果（遺伝子型値のベクトル）$\mathbf{u}$ の最良線形不偏予測子（BLUP; Best Linear Unbiased Predictor）を $\mathbf{u}^*$とすると、以下の混合モデル方程式（MME; Mixed Model Equation）が成り立つことが知られている（Henderson, 1984）。

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{R}_{\mathbf{0}}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{R}_{\mathbf{0}}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}_{\mathbf{0}}^{-1}\mathbf{X} & \mathbf{G}_{\mathbf{0}}^{-1} + \mathbf{Z}^{\mathrm{T}}\mathbf{R}_{\mathbf{0}}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{u}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{R}_{\mathbf{0}}^{-1}\mathbf{y} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}_{\mathbf{0}}^{-1}\mathbf{y} \end{bmatrix} \tag{2.30}$$

この方程式を $\mathbf{u}^*$について解くと

$$\mathbf{u}^* = \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{y} \tag{2.31}$$

となる。ただし、

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1} \tag{2.32}$$

$$\mathbf{V} = \mathbf{R}_{\mathbf{0}} + \mathbf{Z}\mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}} \tag{2.33}$$

とおいた。なお、行列 $\mathbf{P}$ は $\mathbf{PVP} = \mathbf{P}$ を満たす。いま、このモデルにおける表現型ベクトル $\mathbf{y}$ の、$\mathbf{u}$ に関する周辺分布は

$$p(\mathbf{y}|\boldsymbol{\beta}) = N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \tag{2.34}$$

と表すことができる。よって、推定量 $\mathbf{u}^*$の分散は

$$\mathrm{V}[\mathbf{u}^*] = \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P} \cdot \mathrm{V}[\mathbf{y}] \cdot \mathbf{P}\mathbf{Z}\mathbf{G}_{\mathbf{0}} = \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G}_{\mathbf{0}} \tag{2.35}$$

と計算でき、推定量 $\mathbf{u}^*$と真の遺伝子型値 $\mathbf{u}$ との共分散は

$$\mathrm{Cov}[\mathbf{u}, \mathbf{u}^*] = \mathrm{Cov}[\mathbf{u}, \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{y}] = \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P} \cdot \mathrm{Cov}[\mathbf{u}, \mathbf{y}] = \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G}_{\mathbf{0}} \tag{2.36}$$

と計算できる。

遺伝子型値の予測誤差分散（PEV; Prediction Error Variance）とは、推定された遺伝子型値 $\mathbf{u}^*$で条件づけられた遺伝子型値 $\mathbf{u}$ の分散 $\mathrm{V}[\mathbf{u}|\mathbf{u}^*]$ と定義される（Henderson, 1984; Laloë,1993）。以上の結果を用いることで、GBLUP における PEV は

$$\mathrm{PEV} = \mathrm{V}[\mathbf{u}|\mathbf{u}^*] = \mathrm{V}[\mathbf{u}] - \mathrm{Cov}[\mathbf{u}, \mathbf{u}^*] \cdot \mathrm{V}[\mathbf{u}^*]^{-1} \cdot \mathrm{Cov}[\mathbf{u}^*, \mathbf{u}] = \mathbf{G}_{\mathbf{0}} - \mathbf{G}_{\mathbf{0}}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G}_{\mathbf{0}} \tag{2.37}$$

このように解析的に計算することができる。PEV は表現型ベクトル $\mathbf{y}$ に依存しないため、PEV はどの系統で遺伝子型値を推定するかを決めれば、表現型値を得ることなしに計算することができる。

PEV は、表現型から推定した遺伝子型値を所与の量（入力変数）として、未知の量である真の遺伝子型値（応答変数）を

$$\mathbf{u} = \Gamma \cdot \mathbf{u}^* + \boldsymbol{\varepsilon} \tag{2.38}$$

このように回帰したときに残差ベクトル $\boldsymbol{\varepsilon}$ が従う分散共分散行列である。なお、簡単な計算により、この回帰の切片は 0 であること、Γ が単位行列であることが確認できる（ゆえに、切片はあらかじめ省略している）。残差の分散が小さければ小さいほど、推定された遺伝子型値が真の遺伝子型値によく当てはまる。よって、式(2.37)で与えられる行列の対角成分の平均が小さいほど良い。したがって

$$\mathrm{PEVmean} = \frac{1}{M} \cdot \mathrm{Trace}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0}) \tag{2.39}$$

このように定義される PEVmean を考えることで、推定された遺伝子型値の良し悪しを表現するスカラー指標とすることができる（Akdemir et al., 2015）。

なお、育種ではある系統と別の系統の遺伝子型値の差に注目することがある。そのような場合には、contrast vector と呼ばれるベクトル $\mathbf{c}$ を導入すると便利である。ここで、contrast vector は要素の和が 0 である M 次元ベクトルと定義される。例えば、系統 i と系統 j の差 u_i - u_j に興味がある場合には、contrast vector の i 番目の要素を 1 に、j 番目の要素を-1 に、それ以外の要素を 0 に定めればよい。このとき、注目している遺伝子型値の差に関する PEV は

$$\mathrm{PEV} = \mathrm{V}\left[u_i - u_j | \mathbf{u}^*\right] = \mathrm{V}[\mathbf{c}^{\mathrm{T}}\mathbf{u} | \mathbf{u}^*] = \mathbf{c}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0})\mathbf{c} \tag{2.40}$$

と表記できる。複数の比較したい要素がある場合については、複数の contrast vector についての平均をとり

$$\mathrm{PEVmean} = \sum_s \mathbf{c}_s{}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0})\mathbf{c}_s \tag{2.41}$$

とすればよく、こちらを PEVmean の定義とすることも多い（Rincent et al., 2012）。２つの遺伝子型の差に興味がある場合には共分散成分も考慮しなければならないため、contrast vector を用いた定義を採用すべきだと考えられる。なお、contrast vector の定数倍による影響を無視するため、$\mathbf{c}^{\mathrm{T}}\mathbf{c}$ で割った表現を採用することもある。いま、どの系統を（どの環境で）栽培試験するかという情報は計画行列 $\mathbf{X}, \mathbf{Z}$ によって表現されているため、PEVmean による訓練集団最適化は

$$\underset{\mathbf{X},\mathbf{Z}}{\mathrm{argmin}}\, \mathrm{PEVmean} = \underset{\mathbf{X},\mathbf{Z}}{\mathrm{argmin}} \sum_s \mathbf{c}_s{}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0})\mathbf{c}_s \tag{2.42}$$

のように書くことができる。実際には、行列 $\mathbf{X}$ と $\mathbf{Z}$ には栽培できる系統数などに応じた適切な制約条件が必要であるので、それを踏まえた制約付きの最適化問題となる。

予測誤差分散に並んでしばしば利用される指標は、式(2.38)の回帰における決定係数（CD; Coefficient of Determination）である（Laloë,1993）。こちらは contrast vector による表記で定義される必要があるが、PEV の場合と同様に、以下のように解析的に記述できる。

$$\mathrm{CD} = 1 - \frac{\mathrm{V}[\mathbf{c}^{\mathrm{T}}\mathbf{u}|\mathbf{u}^{*}]}{\mathrm{V}[\mathbf{c}^{\mathrm{T}}\mathbf{u}]} = 1 - \frac{\mathbf{c}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0})\mathbf{c}}{\mathbf{c}^{\mathrm{T}}\mathbf{G_0}\mathbf{c}} \tag{2.43}$$

また、PEVmean の場合と同様に、CDmean は

$$\mathrm{CDmean} = \sum_{s}\left\{1 - \frac{\mathbf{c}_s{}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0})\mathbf{c}_s}{\mathbf{c}_s{}^{\mathrm{T}}\mathbf{G_0}\mathbf{c}_s}\right\} \tag{2.44}$$

と定義される（Rincent et al., 2012）。CD は回帰の当てはまりがよいほど大きいため、CDmean を最大化することによって最適な訓練集団を決定できると考える。したがって、CDmean に基づく訓練集団最適化は

$$\underset{\mathbf{X},\mathbf{Z}}{\operatorname{argmin}}\,\mathrm{CDmean} = \underset{\mathbf{X},\mathbf{Z}}{\operatorname{argmin}}\sum_{s}\left\{1 - \frac{\mathbf{c}_s{}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}\mathbf{G_0})\mathbf{c}_s}{\mathbf{c}_s{}^{\mathrm{T}}\mathbf{G_0}\mathbf{c}_s}\right\} \tag{2.45}$$

と表すことができる。

式(2.40)および(2.43)より、ある contrast vector について

$$\mathrm{CD} = 1 - \frac{\mathrm{PEV}}{\mathbf{c}^{\mathrm{T}}\mathbf{G_0}\mathbf{c}} \tag{2.46}$$

である。したがって、PEV の最小化と CD の最大化は、ある 1 つの contrast vector については同じ結果をもたらす。いっぽう、複数の contrast vector を考えると

$$\mathrm{CDmean} = \sum_{s}\left(1 - \frac{\mathrm{PEV}}{\mathbf{c}_s{}^{\mathrm{T}}\mathbf{G_0}\mathbf{c}_s}\right) \tag{2.47}$$

のように、CDmean は PEV の重み付き和のような形になる。よって、CDmean の最大化と PEVmean の最小化は、一般には一致しない。

Rincent らの解析では、CDmean の最大化は PEVmean の最小化よりも優れた訓練集団を与えた（Rincent et al., 2012）。しかしながら、Yu らの解析（この研究では、式(2.39)のような対角和が CDmean と PEVmean の両方に用いられた）では、CDmean よりも PEVmean のほうが予測精度との関連性が強いことが示唆されている（Yu et al., 2018）。このように、いずれの指標を用いるべきかについては、十分な知見が得られていないのが現状である。

なお、PEV と CD の定義や導出では、モデルが正しいことが暗に仮定されていることに注意す

る必要がある。途中の計算において、真の遺伝子型値 **u** が平均 0、分散共分散行列 $\mathbf{G_0}$ の多変量正規分布に従うことが用いられているためである。実際には、主働遺伝子やエピスタシス、優性効果などの存在により、遺伝子型値が $\mathbf{G_0}$ に従わないことが容易に想定される。

現在のゲノミック予測の立場では、式(2.29)のように与えられる遺伝子型値 **u** の分布は事前分布と捉えるほうが自然である。しかし、PEV や CD は、ゲノミック予測が提案されるよりも古くから、頻度論的な視点で議論されてきたものであるため、こうした視点で説明・理解することは難しい。私見ではあるが、ゲノミック予測やベイズ主義的な視点から PEV や CD を再解釈・あるいは必要に応じて再定義することも重要だと考えられる。例えば、式(2.27)から式(2.29)で定義されるモデルについて、遺伝子型値の事後分散 $\mathrm{V}[\mathbf{u}|\mathbf{y}]$ は PEV と一致することが、簡単な計算によって確認できる。残念ながら現段階ではこれ以上の詳細な議論を発見するには至っていないが、ベイズ的な立場からの解釈をさらに進めることにより、ゲノミック予測に対して有用な知見が得られるものと考える。

2-2-3. 育種戦略の最適化

ゲノミック予測において、予測精度の向上が重要な研究課題とされるのは、それが典型的な集団選抜において遺伝的改良に直結するためである。目的形質が正規分布するという仮定のもと、適当な閾値を超える個体を選抜し、それらの無作為交配によって後代を展開するという集団選抜を考えると、1 回の選抜による遺伝的獲得量 Δ について以下の方程式が成立する（Falconer and Mackay, 1996; Desta and Ortiz, 2014）。

$$\Delta = ir\sigma_u \tag{2.48}$$

ここで、i は選抜強度であり、上位何%の個体を選抜するかによって定まる。また、r はゲノミック予測の予測精度（予測された遺伝子型値と真の遺伝子型値との相関係数）である。これは、育種家の方程式（breeder's equation）をゲノミック予測に拡張した式になっている。式(2.48)より、予測精度に比例して遺伝的獲得量が向上することがわかる。また、選抜強度が強い（少ない数の上位個体を選抜する）ほど、遺伝分散が大きいほど遺伝的獲得量は大きい。

通常、育種では複数回の選抜と交配を行う。この場合には、選抜強度と遺伝分散の間にトレードオフが存在するため、強すぎる選抜は最終的な遺伝的獲得量（複数回の選抜による遺伝的獲得量の総和）を減少させることがある。これは、選抜によって遺伝分散が減少することが原因である（Falconer and Mackay, 1996）。したがって、長期的な遺伝的獲得量を最大化するには、選抜強度を適切に調節する必要があり、遺伝分散の減少を防ぐような選抜や交配の戦略が、ゲノミック予測よりも古くから、BLUP 法の枠組みで検討されてきた（Meuwissen, 1998; Avendaño et al., 2004）。

なお、ゲノミック予測を用いる場合には、選抜によって集団内の遺伝的構成が変化することによる予測精度の変化（多くの場合、予測精度の低下）にも注意する必要がある。例えば、Habier

らによれば、GBLUP は BayesB に比べて選抜に伴う世代の変化による予測精度の低下が顕著であった（Habier et al., 2007）。また、例えば Yabe らのシミュレーションでも、選抜によって予測精度が大きく低下することが指摘されており、予測モデルを適切に更新することが必要だとしている（Yabe et al., 2013）。

以上のように、集団選抜という最も単純な育種法に限定しても、ゲノミック予測における最適な育種戦略を導出することは極めて困難である。この原因として、選抜強度と遺伝分散のようなトレードオフの関係が随所に存在すること、選抜系統数や展開する後代の個体数などの選択肢の組み合わせが膨大であること、親から子への遺伝が確率的なゆらぎを持つこと、真の遺伝子型値が観測できないことなどが挙げられる。

さらに、こと植物育種においては、このように単純な集団選抜ばかりが想定されるわけではない。植物育種では、最終的に純系品種を作出することが求められることが多く、例えば自殖を繰り返して遺伝子型を固定しながら純系品種を作出する。また、一部の作物では倍加半数体を用いることで純系品種を得ることもできる。他にも代表的な育種の目標として、優れた F_1 系統を与える両親を育種することも挙げられる。あるいは、多数の遺伝資源をスクリーニングして、有用な育種母本を探索するという pre-breeding も、植物育種においては重要である。いずれにせよ、多くの場面において、『1 つの（ないしは、少数の）優れた系統を得る』ことが植物育種の目的となる。これは、自殖や種子による系統の保存が困難な動物育種とは対照的である。本論文でも、4 章ではこのような視点に立って選抜手法を評価する。

このような植物育種に特有の問題設定についても、育種戦略の最適化に向けたいくつかの研究例が存在する。ここで重要なアイデアは、「選抜強度」と「遺伝分散」という集団レベルでの議論から、「どの個体とどの個体を交配するか」と「個体の保持する QTL」という個体レベルでの議論へとシフトすることである。例えば、強すぎる選抜は、選抜された個体群における QTL の欠失を生ずることで、遺伝分散を低下させていると解釈することができる。つまり、遺伝分散を低下させない選抜戦略とは、複数存在する有用 QTL をできるだけ失わないように選抜する個体を選ぶ、あるいは交配をデザインすることに相当する。このような議論が実際の育種において可能になったのは、DNA マーカーが活用できるようになったことが大きく、その意味で、より現代的な発想と言えるだろう。以下では、植物育種に特有の育種戦略に関する代表的な研究例を挙げる。

van Berloo と Stam は、自殖性作物の RIL（ricombinant inbred line; 組換え近交系）集団を材料とするマーカー選抜において、ある交配組み合わせが分離世代（F_2 世代）において生じうる最良の後代個体の遺伝子型値を考え、交配組み合わせを選ぶ方法を提案した（van Berloo and Stam, 1998）。彼らの選抜戦略は「1 つの優れた系統を得ること」を目標としており、複数のアリルをポジティブに固定しようとする発想を直接的に採用した最初の研究例だと思われる。

Daetwyler らは、倍加半数体による純系の作出を想定して、ゲノミック予測に基づく個体選抜の指標を考案した。彼らはマーカー回帰を用いてマーカー効果を推定した結果を用いて、最良の半数体を得られうる個体を親として選ぶ方法を提案し、その有効性をシミュレーションにより検証した（Daetwyler et al., 2015）。

van Beloo らや Daetwyler らの選抜基準は、最終的に作出されうる純系の遺伝的能力の最大値が大きいもの（possible best）を選んでいると解釈できる。いっぽうで、そうしたベストな系統が得られる確率（probability of possible best）は考慮していない。そこで、Han らは、ベストな系統が得られる確率まで考慮した選抜基準を提案した(Han et al., 2017)。ただし、Daetwyler らがゲノムワイドマーカーを用いた遺伝的能力の予測を考慮しているのに対して、Han らは比較的少数のアリルを固定すること（マーカー選抜による QTL の集積）を想定している点には注意が必要である。ゲノムワイドマーカーを用いたアプローチとしては、Müller らが、交配によって得られる後代集団における遺伝子型値の最大値の期待値を計算し、その値が大きい交配組み合わせを選ぶことを提案している（Müller et al., 2018)。そのほか、genomic mating（Akdemir and Sánchez, 2016）や usefulness index（Lehermeier et al., 2017）などの、集団選抜に近い立場から優良系統を効率的に作出しようとする研究も活発に行われている。

ゲノミック予測の理論研究における関心事は、予測精度の向上を目的とした予測モデルの新規開発から、既存の予測モデルの活用へと、徐々に移ろいでる印象がある。ゲノミック予測によって表現型を予測できることが周知の事実となるにつれ、これまでの植物育種学が築いてきた多様な育種法に、どのようにゲノミック予測を活用するべきかが議論されるようになったとも言えるだろう。本節で取り上げた研究を皮切りに、表現型をゲノミック予測の予測値で置き換えるだけでなく、ゲノミック予測の特性を活用した新たな育種法・選抜法を開発しようという流れが、今後ますます加速するものと考えられる。本論文では育種戦略そのものを新規に開発することはないが、４章で取り上げるベイズ最適化の応用では、統計モデルを用いるゲノミック予測だからこそ可能になった選抜基準を提案する。また、６章で扱う分離の予測は、育種戦略の最適化に関わる重要な要素である。

2-2-4. 複数の形質や GxE のモデル化

2-1 節では、1 つの表現型をモデル化する方法について議論した。言うまでもなく、実際には複数の形質が育種目標とされるほうが普通である。また、植物育種において無視することのできない要素に遺伝子型・環境交互作用 (GxE; Genotype-by-Environment interaction) が挙げられる。すなわち、遺伝子型値は植物の置かれる環境によって変化するものと考える必要がある。式(2.6)のように複数の環境について同一の遺伝子型値を推定する混合モデルでは、これに対応することはできない。

GxE や複数の形質をモデル化する方法の 1 つは、環境間、あるいは形質間の遺伝共分散を予測モデルに組み込むことである（Burgueño et al., 2012)。このモデルについては、５章で数式を交えて解説する。直感的には、異なる環境について同一の遺伝子型値を推定するモデルと、環境ごとに独立に遺伝子型値を推定するモデルの、ちょうど中間に位置するようなモデルになっている。

本論文では扱わないが、もう 1 つの興味深いアプローチは、ゲノミッック予測と作物モデルを組み合わせる方法である。作物モデルとは、ある品種の環境に対する応答を、生理学的な知見や実験をもとにモデル化したものである（Soltani and Sinclair, 2012)。例えば、イネの出穂に対す

る DVR モデルでは、1 日ごとの日長と気温の影響が累積することで出穂に至るというモデル化がなされており、異なる環境（異なる日長と気温）での出穂を予測することができる（Nakagawa et al., 2005; Yin et al., 1997）。作物モデルには品種固有のパラメータが設計されていることが普通であり、例えば DVR モデルにも、日長や気温への応答性に関わるパラメータを品種ごとに定めることができ、これにより、系統間で異なる環境応答が表現される。作物モデルのパラメータは、通常、その品種に対して適切な実験を行うことにより定められる。つまり、作物モデルを当てはめるには、当てはめたい品種についてコストをかけて実験を行う必要があり、育種で扱われるような多数の系統に対して作物モデルを適用することは現実的ではなかった。しかし、すでに品種固有のパラメータが推定されている系統とそのマーカー遺伝子型を訓練データとして予測モデルを構築することで、マーカー遺伝子型を取得するだけで、作物モデルを適用した環境応答の予測が可能になる（Technow et al., 2015; Onogi et al., 2016）。すなわち、環境応答に関しては作物モデルが、系統間の違いに関してはゲノミック予測が、それぞれモデル化を担当することで、GxE のモデル化を実現する。

複数の形質に関する議論では、適切なモデル化手法の検討だけでなく、モデルに基づく育種戦略における扱い方を議論することも重要だと考えられる。例えば、２つの形質を共に大きくしたい場合に、両方の和を予測してその値が大きい系統を選抜するのか、それとも、形質ごとに優れた系統を選抜して交配するのか、といった問題に、合理的な解答を与える必要がある。おそらく、最適な選抜戦略とは、形質間の予測精度の違いを反映したものになるだろう。さらに興味深い問題設定は、形質間で表現型の取得コストが異なる場合について、多形質を予測するモデルを用いる場合である。もし、表現型の取得コスト（費用や労力）が低い形質の情報によって、取得コストの高い形質の予測精度を改善することができれば、それぞれの表現型を、どのくらいの個体数だけ、どの系統について測定するべきか、という問題を最適化することで、表現型の取得コストを削減することができるだろう。

以上のように、本章の冒頭では単形質についての代表的なモデル化手法を扱ったが、実際の植物育種で入手される多形質・多環境のデータをどのように扱うかについては議論が続いている状況にある。また、本論文の主題とは大きく離れるため紙面を割かなかったが、既存の生物学的・農学的な知見をいかにモデル化に反映させるべきかについても、しばしばモデル化における課題として挙げられている。本論文では、モデルの高精度化を実現する訓練データの選択や、モデルをより適切に活用した選抜法などを開発するが、そこでは比較的単純な、既存のモデル化手法を前提とする。育種の効率化・最適化は、モデル化手法と、その活用法の両方を開発することによってなされるはずである。本論文は後者に重きをおくが、前者もまた同じように重要であることを、ここで強調しておく。

3. 能動学習に基づくゲノミック選抜モデルの効率的構築

3-1. 序論

ゲノミック選抜では、収量や開花までの日数などの連続変数（量的形質）の「値」そのものを予測することが求められる場合が多く、したがって回帰モデルを用いることが主流である。モデルの評価にも、真の遺伝子型値と予測された遺伝子型値とのピアソン相関係数がしばしば用いられる。いくつかの合理的な仮定の下で単純な集団育種法を想定したとき、この相関係数が遺伝的獲得量に比例する（Falconer and Mackay, 1996; Desta and Ortiz, 2014）ことは、この評価基準の合理性の根拠の 1 つである。

いっぽうで、ゲノミック予測の役割を個体選抜に限定すれば、選抜・淘汰という二値分類問題となる。この立場からは、モデルの評価には分類精度や AUC、あるいは κ 係数などの指標が用いられる（González-Recio and Forni, 2011; Ornella et al., 2014）。例えば、分類器としてサポートベクトルマシン（SVM）を用いることで、回帰モデルを上回る選抜・淘汰の分類精度を達成できた例が報告されている（Ornella et al., 2014）。また、特定の病害について明瞭な感受性・非感受性が見られる場合のように、初めから目的形質が二値変数、あるいは少数クラスの順序つき変数で表現される場合（質的形質）には、閾値を用いて表現される分類モデルを利用するのが自然である（Montesinos-Lopez et al., 2015）。

構築するのが回帰モデルであれ分類モデルであれ、ゲノミック予測における重要な研究課題は、できるだけ少ない訓練データ数で精度の高い予測モデルを構築することである。ゲノミック予測における 1 サンプルの訓練データは、系統のマーカー遺伝子型（説明変数：入力 $\mathbf{x}$）と表現型（応答変数：出力 y）の組であるが、これを得るには DNA のシーケンスと圃場試験を実施しなければならない。このうち入力 $\mathbf{x}$ については、シーケンス技術の進歩によって 1 サンプルあたりのマーカー遺伝子型の取得コストが大きく下がり、大量のサンプルについて同時にマーカー遺伝子型を取得することが可能になった。他方、表現型については、植物の生育にかかる時間が長いこと、栽培や評価を行う人的資源が有限であることなどから、多数の系統についてハイスループットに取得することは非常に困難である。すなわち、モデルの入力 $\mathbf{x}$ と出力 y のうち、とりわけ出力の取得コストが高い。

同様の状況は一般の機械学習課題においても見られる。例えば自然言語処理では、人間の音声の意味を判別することが重要な研究課題の 1 つである。このとき、訓練データは音声の波形と、その言語的内容の組となる。入力 $\mathbf{x}$ を得るには適当な音声を録音するだけでよいが、出力 y を得るには、実際に人間が録音された音声を聞いて、その内容について記述する必要がある。このように、機械学習のほとんどの応用例では、出力の情報を取得するコストは入力のそれよりも高い。尤も、応答変数は簡単に得られないからこそ予測したい対象となるのだから、これは至極当然のことである。

このような状況をふまえ、能動学習（active learning；Settles, 2009）と呼ばれる、入力 $\mathbf{x}$ が既知である場合に、その出力 y についてコストをかけて調査し、訓練データに追加するか否かを適切に判断するための方法について検討が行われてきた。これは、言い換えれば、訓練データを

母集団から無作為に（受動的に）抽出するのではなく、何らかの基準に基づき「能動的に」選択することである。能動学習の応用例は枚挙にいとまがないが、例えば 2010 年までの比較的初期の研究例として、画像解析（Hoi et al., 2006; Tuia et al., 2009）や自然言語処理（Lewis and Gale, 1994; Zhu et al., 2010）、あるいは創薬（Warmuth et al., 2003）への応用などが挙げられる。また、近年では深層学習を用いた画像分類に特化した能動学習フレームワークも議論されている（Wang et al., 2017）。

ゲノミック予測における類似の問題設定として、訓練集団最適化と総称される研究を挙げることができる。そこでは、GBLUP および pedigree-BLUP を予測モデルとして用いる場合について、訓練データの候補となる系統の入力 $\mathbf{x}$（あるいは、サンプル間の関係行列 $\mathbf{G}$）が既知である場合に、どの系統について圃場調査を行って出力 y を得るべきかについて研究されてきた（Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Rincent et al., 2017a）。上記の論文で基本的なアイデアとなるのは、同モデルにおいて古くから量的遺伝学的に定義されてきた予測誤差分散や決定係数と呼ばれる指標を最小化または最大化することである。これらの統計量はモデル構築に用いる表現型値には依存せず、系統間の関係行列 $\mathbf{G}$ に基づき計算される。よって、例えば予測誤差分散を最小化するような訓練集団はマーカー遺伝子型だけで決定することができ、そのような訓練集団を用いることで予測精度が向上することが期待される。

これらの研究成果は、ゲノミック予測において最もよく用いられる GBLUP モデルを用いる場合に適用可能であり、現時点では最も有用な訓練データの選択法の 1 つだと言えるだろう。また、上記の論文中では言及されていないが、GBLUP と同じ形式で記述される単一のカーネルを用いた RKHS 回帰にも適用することが可能だと推察できる。いっぽうで、冒頭で述べたような分類問題に対しては、異なる手法を用いることが望ましいと考えられる。モデルの精度向上に貢献する訓練データは、予測モデルそのものに依存するからである。

また、訓練集団最適化に関するいずれの先行研究でも、逐次的に訓練集団を更新する状況については十分に検討されていない。能動学習では、逐次的な訓練集団の更新が多くの場合に想定される。この想定はゲノミック予測においても考慮するに値する。例えば、多数のシーケンスされた遺伝資源について、1 作期ごとにその一部を栽培評価する試験を何度か繰り返し、病害抵抗性を持つか否かをゲノムから分類する予測モデルを構築したい、といった状況は十分に想定される。このとき、2 期目に栽培評価するべき系統は 1 期目の試験を踏まえて決定するべきである。つまり、それまでの結果を踏まえて逐次的に訓練データを選ぶことで、一度にまとめて訓練データを選ぶよりも、より効率的なデータの選択が可能になると考えられる。

以上のように、限られた数の訓練データで精度の高いモデルを実現することはゲノミック選抜における重要課題であり、能動学習はその解決策となることが期待される。しかしながら現状では、代表的な回帰モデルである GBLUP に関して能動学習に近い発想で研究が行われているのみであり、能動学習が分類モデルを用いたゲノミック予測において有効であるかどうかは検証されていない。能動学習は幅広い対象について有用であることが経験的に示されているものの、ゲノミック予測で用いられる高次元小標本のマーカー遺伝子型データや、ノイズの大きい表現型データに対しても有効に機能するかは未知数である。そこで、本研究では、分類器を用いてゲノミッ

ク予測を行う場合の訓練集団の選択に能動学習を応用した。分類器として先行研究でゲノミック予測における有効性が経験的に確かめられた SVM を用い、複数の実データ、および 1 つの仮想データを用いてシミュレーションを行うことで、能動学習によって訓練集団のサイズを節約しながら分類精度を向上できるかについて検証を行った。

3-2. 材料・方法

3-2-1. 分類問題としてのゲノミック予測と SVM

遺伝的改良を行うためには、優れた系統を選抜し、そうでない系統を淘汰する必要がある。ゲノミック予測においても、連続量の表現型を回帰モデルによって予測したのちに、望ましい予測値をもつ系統を選抜することがしばしば行われる。この場合、ある系統を選抜するか淘汰するかの二値分類だけが最終的な問題となっている。したがって、単純な個体選抜を行うには、淘汰されるべき系統のラベルを $y = -1$ に、選抜されるべき系統のラベルを $y = +1$ にして、二値分類器を構成すれば十分である。

選抜と淘汰の基準は育種目標によって大きく異なる。例えば出穂期については、極端な早生・晩成は望ましくないとされ淘汰されることも多い。また、複数の形質を同時に改良する場合（例えば果実の糖度と酸度）には、複数の表現型値のバランスが問われることもある。さらには、何らかの閾値が予め決まっており、それを超えない個体を全て淘汰するような選抜もしばしば行われる。こうした多様な選抜様式について検討することも極めて重要であるが、本研究では単純のため 1 つの形質のみを考え、用いているデータセットにおける表現型上位 20%が選抜の、残りの 80%が淘汰の対象であると仮定した。

本研究では、既にゲノミック予測における有用性が示されている SVM を分類器として採用した。以下、(Bishop, 2006）に基づき、簡単に SVM について解説する。なお、本研究では、各系統が 1 つの表現型値を持つ状況のみを扱う。これには、ある系統の表現型値として複数個体の表現型値の代表値（e.g. 算術平均や中央値）を用いる場合が該当する。よって、以下では 2 章のように系統数 M と個体数 N を区別せず、すべて N で表現する。

ある系統 i ($i = 1, 2, ..., N$)の二値ラベルを y_i とし、入力（マーカー遺伝子型）を P 次元ベクトル $\mathbf{x}_i$ として、以下のような（Q 次元の特徴量ベクトル $\boldsymbol{\varphi}_i = [\varphi_1(\mathbf{x}_i), \varphi_2(\mathbf{x}_i), ..., \varphi_Q(\mathbf{x}_i)]$についての）線形な分離境界 f を考える。

$$y_i = \mathrm{sign}(f(\mathbf{x}_i)) \tag{3.1}$$

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi} + b \tag{3.2}$$

ここで、sign(z)は z の符号が正であれば 1 を、負であれば−1 を与える関数である。なお $f(\boldsymbol{x}_i)$は常に 0 以外の値を取るものと仮定する。また、$\mathbf{w}$ は $\boldsymbol{\varphi}(\mathbf{x})$に対応する長さの重みベクトルであり、$b$

は分離境界の位置を決めるスカラーである。ゆえに、分類境界を決定することは、訓練データにしたがって最適な $\mathbf{w}$ および b を決定することと等価である。

SVM では、以下の制約付き最適化問題によって $\mathbf{w}$ および b を定める。

$$\underset{\mathbf{w},b}{\operatorname{argmin}}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i\right\} \tag{3.3}$$

$$\text{subject to} \quad (\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_i + b)y_i \geq 1 - \xi_i,\ \xi_i > 0$$

$$\xi_i = \begin{cases} 0 & (\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_i + b)y_i \geq 1 \text{ のとき} \\ |y_i - f(\mathbf{x}_i)| & (\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_i + b)y_i < 1 \text{ のとき} \end{cases} \tag{3.4}$$

ここで、C はコストパラメータと呼ばれる正の値をとる超パラメータであり、誤分類をどの程度まで許容するかを調節する。また、ξ はスラック変数と呼ばれる。

式(3.3)の制約付き最適化問題は、ラグランジュの未定乗数法により、α_i と β_i をラグランジュ乗数として以下のように定義される L を最小化する問題に置き換えられる。

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_i + b)y_i - (1 - \xi_i)\} - \sum_{i=1}^{N}\beta_i\xi_i \tag{3.5}$$

ただし、不等式制約に関する Karush-Kuhn-Tucker 条件として、任意の i について

$$\alpha_i\{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_i + b)y_i - (1 - \xi_i)\} = 0 \tag{3.6}$$

$$\beta_i\xi_i = 0 \tag{3.7}$$

$$(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_i + b)y_i - (1 - \xi_i) \geq 0 \tag{3.8}$$

を満たす必要がある。

L を各変数で微分した

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i\, y_i\boldsymbol{\varphi}_i \tag{3.9}$$

$$0 = \sum_{i=1}^{N}\alpha_i\, y_i \tag{3.10}$$

$$\alpha_i = C - \beta_i, \qquad \forall i \tag{3.11}$$

これらの結果を式(3.5)に用いることで、 $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1,\ldots,N}$ に関する最適化問題

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{\varphi}_i{}^{\mathrm{T}} \boldsymbol{\varphi}_j \tag{3.12}$$

を得ることができる。これを解くことで $\boldsymbol{\alpha}$ が求まり、得られた $\boldsymbol{\alpha}$ を 式(3.6)、式(3.8) に代入することによって他の未知変数を求めることができる。

式(3.12)において、$\boldsymbol{\varphi}$ は内積 $\boldsymbol{\varphi}_i{}^{\mathrm{T}}\boldsymbol{\varphi}_j$ の形でのみ $\boldsymbol{\alpha}$ の推定に寄与する。さらに、分類境界 $f(\boldsymbol{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi} + b$ について、式(3.9)より

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \, y_i \boldsymbol{\varphi}_i{}^{\mathrm{T}} \boldsymbol{\varphi} + b \tag{3.13}$$

が成り立つ。すなわち、分類境界もまた特徴量ベクトルの内積にのみ依存する。このような特徴ベクトルの内積だけで記述できるモデルは、カーネル法の枠組みで一般化できることが知られている。任意の 2 つの入力ベクトルに対して定まる関数 $k(x_i,x_j)$ の値を i,j 要素に持つグラム行列 $\mathbf{K}$ が半正定値となるとき、その関数をカーネル関数と呼ぶ。SVM は、特徴量を明示的に定めなくとも、カーネル関数だけを定めれば定義することができる。よく利用されるカーネル関数はガウスカーネルであり

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{h}\right) \tag{3.14}$$

と定義される。ここで、h は正の値をとる超パラメータである。本研究でも、このガウスカーネルを用いた SVM を採用した。なお、ガウスカーネルの計算に先立って、マーカー遺伝子型はすべてのマーカーにおいて平均を 0 に、分散を 1 にする標準化を行なった。

ゲノミック予測を分類問題と捉える場合には、正例と負例が 50%ずつにならず、一方に偏ることが問題となる（不均衡データ; imbalanced data)。例えば本研究では、選抜される系統は全体の 20%しかないため、負例が正例の 4 倍多く存在する。このような状況で分類器を訓練してしまうと、分類境界が適切に定まらない可能性が指摘されており、様々な対処法が検討されている (He and Garcia, 2009; Blagus and Lusa, 2010)。本研究では、SVM において最も簡便な不均衡データへの対処法だと思われる、Error Cost の調節を行った。これは、もとの最適化問題

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \right\}$$

の代わりに、正例と負例に異なるコストパラメータを導入した

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C_+ \sum_{\{i|y_i=+1\}}^{N_+} \xi_i + C_- \sum_{\{i|y_i=-1\}}^{N_-} \xi_i \right\} \tag{3.15}$$

を考えるように SVM を修正する方法である（Akbani et al., 2004; Huang and Du, 2005）。ただし、訓練データにおける正例の数を N_+ とし、負例の数を N_- とした。コストパラメータが大きいほど、そのクラスの誤分類には厳しいペナルティがかかる。よって、ラベルが少ない方のクラスについてのみコストパラメータを大きく設定することで、そのクラスの正答率を意図的に高めることができる。なお、具体的なコストパラメータの設定については、経験則として正例と負例の逆比を用いることが、先に挙げた論文で提案されている。通常の SVM では、コストパラメータの default 値を 1 に定める（Chang and Lin, 2011）ため、本研究では $C_+ = N_-/N_+$, $C_- = 1$ とパラメータを設定し、多数派クラスに対するコストパラメータを既定値に保ちつつ、少数派クラスの誤分類を厳しく制限した。なお、ゲノミック予測における Ornella らの先行研究では、真の正例と負例の割合は 15:85 であるところを、分類器の学習では 40:60 に訓練データを分割して学習させるという ad hoc な工夫を行っている（Ornella et al., 2014）。

3-2-2. 能動学習の概要

ここでは、能動学習について、Settles の総説（Settles, 2009）を主に参照して概説する。はじめに、能動学習の意義を簡単な例で確認する。次に、能動学習の全体像を、想定される状況に応じた能動学習の類型とともに説明し、今回のゲノミック予測における問題設定がそのどれに当てはまるかを述べる。さらに、能動学習において中心的な役割を果たす質問戦略（query strategy）のうち、本研究で用いた uncertainty sampling と呼ばれる方法について説明する。

多くの統計的推論では、訓練データは、母集団からの無作為標本であることが想定される。無作為抽出された訓練データは、いくつかの学習タスクには都合が良い。例えば母集団の平均と分散を推定する場合には、訓練データは無作為抽出されているほうが良いであろう。

しかしながら、学習タスクによっては、より能動的に訓練データの抽出をコントロールすることにより、ある予測精度を実現するために必要なデータの数を減らす、あるいは、同じ数のデータで予測性能を向上させることができる。Settles は、図 3-1 のような状況を例示している。最上段の図は、全データの分布とラベルを示している。この例では、各クラスの入力 $\boldsymbol{x}$ は、２つの平均が異なる２変量正規分布から生成されている。中段の図は、全データから無作為に訓練データを抽出して学習した分類境界であり、最下段の図は、真の決定境界に近いデータ点を作為的に抽出して学習した分類境界である。中段と下段を比べると、分類器の訓練に用いられるデータによって、分類性能は大きく異なることが理解できる。このように、通常の無作為抽出による訓練データと能動的に集められた訓練データでは、得られる学習器の性能が異なる。

能動学習では、少ない訓練データで学習器の性能を向上させることを目的に、質問戦略と呼ばれる訓練データの選択基準に基づき訓練データを追加し、学習器を更新することを繰り返す。Settles は、想定される状況や求められるタスクによって、能動学習を以下の３つに類型している。

- Membership query synthesis

このタイプでは、新たに追加すべきデータを、何らかの方法で生成する。１～９の手書き数字の識別を例にすれば、追加する訓練データは学習器が生成した“曖昧な数字”になる。生成は、質問戦略によって自動的に行われる。人間は、そのようにして生成された“曖昧な数字”が１～９のどれに当てはまるかを判断していく。

・ Stream-based selective sampling

このタイプでは、自動的、かつ、連続的にデータの入力が収集できるような場合にその出力を調べて訓練データとするか否かを判断する。上述の手書き数字の例では、絶えず新たな「１～９のどれかが判定されていない画像」が与えられる。それを人間によって判定し新たな訓練データにするか否かは、質問戦略によって自動的に判定される。人間は、訓練データにすべきだと判断された画像のみについて、１～９のどれに当てはまるかを判断していく。

・ Pool-based sampling

このタイプでは、事前に訓練入力が全て与えられていることを想定する。そのうちの一部について出力を調べることを繰り返し、学習器の性能を向上させていく。手書き数字の例では、始めから大量の手書き数字の画像が与えられており、質問戦略によって、どの画像について人間が１～９のどれに当てはまるかを判断するべきなのかを決定する。

典型的なゲノミック予測における訓練集団の決定では、多数の系統（遺伝資源や、圃場試験の候補である種子または発芽実生）のマーカー遺伝子型が与えられ、そのうちどの系統を圃場試験するかを決定することが求められる。これは pool-based sampling にあてはまる。実際には、段階的にシーケンスする、育種の途中で新たな材料が候補となる、などの理由から、逐次的に訓練データの候補が与えられる stream-based selective sampling と pool-based sampling の組み合わせのような状況も十分に想定されるが、本研究では簡単のため、事前に全ての系統のマーカー遺伝子型（入力）が既知であるとした。

最後に、能動学習における質問戦略のうち、uncertainty sampling について解説する。質問戦略は、能動学習の成否を決める最も重要なものである。用いるべき質問戦略は、扱うデータの性質や用いる分類器などによって異なるが、最も広く用いられているものの 1 つが uncertainty sampling である。二値分類問題に対する uncertainty sampling では、以下の式で表される $U(\mathbf{x})$ が大きいものを次の訓練データとして選ぶ。

$$U(\mathbf{x}) = \min\{\Pr(y = +1|\mathbf{x}), \Pr(y = -1|\mathbf{x})\} \tag{3.16}$$

定義より、この $U(\boldsymbol{x})$は、ある入力 $\boldsymbol{x}$ について片方のクラスに所属する確率が 1、もう片方のクラスに所属する確率が 0 と予測されたとき最小値 $U(\boldsymbol{x}) = 0$ をとり、両方のクラスについて所属確率が 0.5 と予測されたとき最大値 $U(\boldsymbol{x}) = 0.5$ をとる。すなわち、分類が確からしい入力については小さな値に、分類が不確かな入力については大きな値になる。この性質が、$U(\boldsymbol{x})$に基づく訓練データの選択が uncertainty sampling とよばれる理由である。

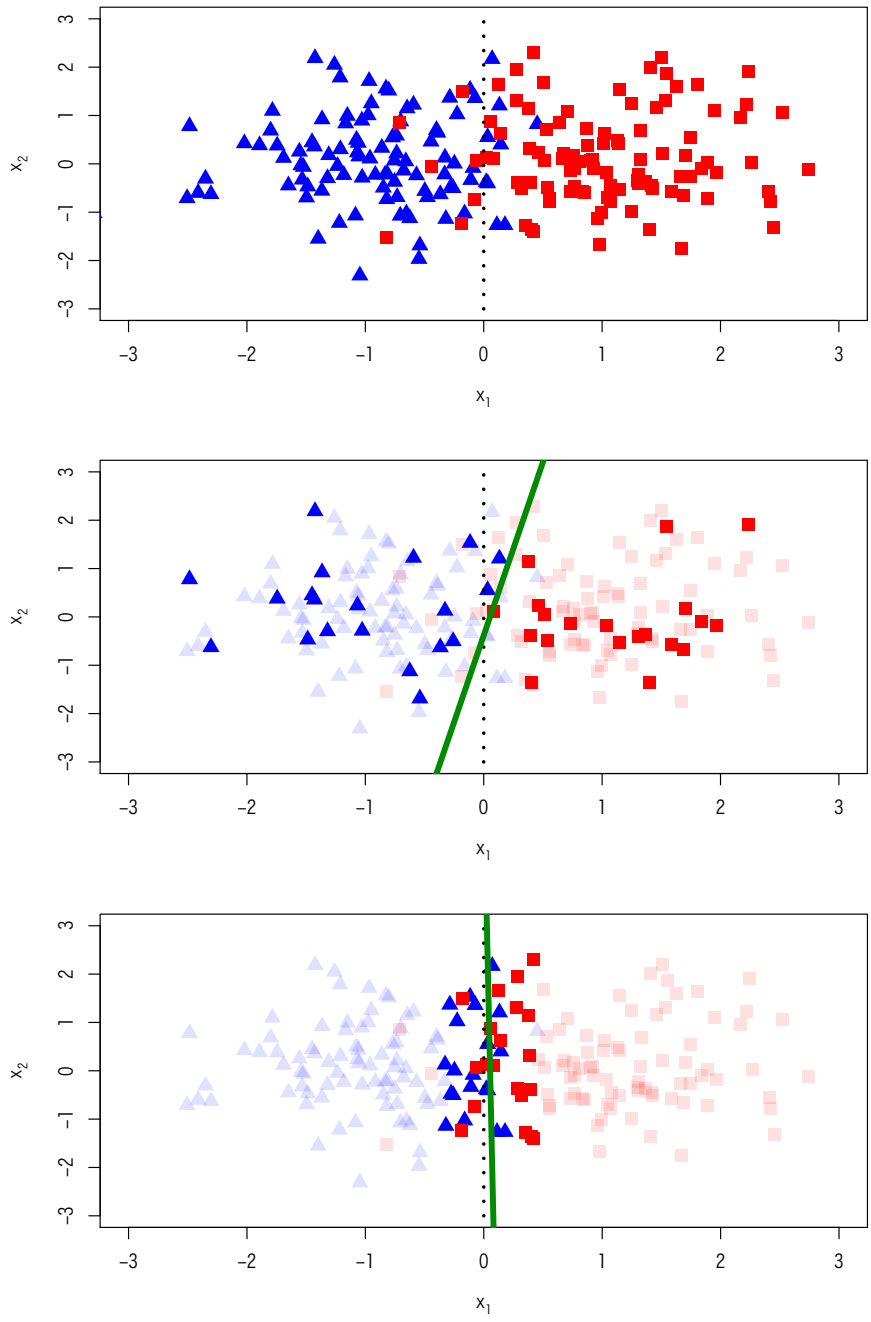


図 3-1. 訓練データによる分類境界の違い

赤と青の 2 クラスのラベルを持つ仮想データを 2 変量正規分布から 100 点ずつ生成した（上段）。このデータにおける最適な分類境界は $x_1 = 0$ に設計されている（灰色点線）。各クラスから無作為に 20 点ずつを訓練データとして分類器を作成したのが中段であり、各クラスで真の境界に近い 20 点ずつを訓練データとして分類器を作成したのが下段である。この例では、下段のほうが真の境界に近い結果を示している。なお、この図では分類器として Fisher の線形判別分析を用いた。

3-2-3. ２クラス SVM を分類器とする uncertainty sampling

本研究では、ゲノミック選抜の分類器として２クラス SVM を用い、質問戦略として uncertainty sampling を用いた。本節では、SVM による能動学習に uncertainty sampling を適用する方法について説明する。

SVM は非確率的分類器であり、uncertainty sampling において $U(\boldsymbol{x})$を求めるために必要なクラス所属確率を与えない。しかし、２クラス分類かつ全てのデータが線形分離可能な場合には、推定した分類関数により与えられる未知データの分類予測

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i \boldsymbol{\varphi}_i{}^{\mathrm{T}} \boldsymbol{\varphi} + \hat{b} \tag{3.17}$$

の絶対値が 0 に近いものを次の訓練データとして選ぶという uncertainty sampling が提案されている（Tong and Koller, 2001）。これは、SVM の分離平面と新たな入力 $\mathbf{x}$ の距離が最も小さいものを次の訓練データとして選ぶことを意味する。

あるいは別の方法として、SVM の予測において事後確率に相当するものを推定し、それを式 (3.16) に用いることも考えられる。SVM における事後確率は、ロジスティックシグモイド関数を用いて推定を行うことができるが、推定値が妥当な値であるとは限らないことが指摘されている（Tipping, 2001）。

R の {kernlab} パッケージには、ロジスティックシグモイド関数を用いた事後確率の推定法（Lin et al., 2007）が実装されている。しかし上述したように必ずしもこの方法によって得られる事後確率は妥当ではないことから、本研究では Tong と Koller により提案された方法を用いて uncertainty sampling を行った。本研究における分類問題は線形分離できない可能性もあるが、それを無視した。

3-2-4. κ 係数による分類精度の評価

通常、分類器の性能は以下の式で定義される分類精度によって評価される。

$$Accuracy = \frac{n_{TP} + n_{TN}}{N} \tag{3.18}$$

ただし

$$N = n_{TP} + n_{FP} + n_{TN} + n_{FN} \tag{3.19}$$

である。ここで $n_{TP}, n_{FP}, n_{TN}, n_{FN}$ はそれぞれ真陽性、偽陽性、真陰性、偽陰性のサンプル数を表す。

しかし、このように定義された分類精度は、各クラスに所属するサンプル数が大きく異なる不均衡データに対しては、良い指標とならないことがある（He and Garcia, 2009）。その例を図 3-2 に示した。分類精度は左右の表で 0.9 となっている。しかし、左の表は、全てのデータを多数派のクラスに分類したにすぎず、明らかに予測器として妥当ではない。

κ 係数（Cohen's kappa）と呼ばれる分類精度の指標を使うことで、この問題をある程度回避することができる（Ornella et al., 2014; Cohen, 1960）。κ 係数は

$$\kappa = \frac{A_{obs} - A_{by\ chance}}{1 - A_{by\ chance}} \tag{3.20}$$

と定義される。ここで、A_{obs} は実際の一致率（通常の分類精度）、$A_{\mathrm{by\ chance}}$ は偶然の一致率であり、次のように定義される。

$$A_{obs} = \frac{n_{TP} + n_{TN}}{N} \tag{3.21}$$

$$A_{by\ chance} = \frac{n_{TP} + n_{FP}}{N} \times \frac{n_{TP} + n_{FN}}{N} + \frac{n_{TN} + n_{FP}}{N} \times \frac{n_{TN} + n_{FN}}{N} \tag{3.22}$$

κ 係数は、通常の分類精度を、偶然の一致率を考慮して補正したものと解釈できる。図 3-2 において、左の表における κ 係数の値は 0、右の表における値は 0.286 となっており、κ 係数を使うことで妥当な相対評価ができることがわかる。本研究では、この κ 係数を用いて SVM の予測性能を評価した。

		真のクラス	
		クラス1	クラス2
予測されたクラス	クラス1	90	10
	クラス2	0	0

分類精度＝0.9
κ 係数＝0

		真のクラス	
		クラス1	クラス2
予測されたクラス	クラス1	85	5
	クラス2	5	5

分類精度＝0.9
κ 係数＝0.286

図 3-2. 分類精度と κ 係数

２つの異なる分類予測の結果に対応する分類精度と κ 係数を示した。左表の結果と右表の結果で、式(3.18)によって定義される分類精度はともに 0.9 であるが、κ 係数は左表の場合 0 であり、右表の場合 0.286 である。

3-2-5. シミュレーションの設定

能動学習によって訓練データを選ぶことで効率的に分類精度が向上するかを検証するため、シミュレーションによる評価を行った。シミュレーションでは、以下のように、データの追加と予測モデルの更新を繰り返した。

- 表現型値の大きいほうから 20%の系統が選抜の対象であり、それ以外の 80%の系統が淘汰の対象であるとする 2 クラス分類問題を定義した。
- 分類精度を検証するためのテストデータとして、実データセットのサンプル数によらず、100 系統を無作為に選んだ。また、初期訓練データとして 50 系統を無作為に選んだ。
- はじめは、初期訓練データとして無作為に選ばれた 50 系統についてのみ、表現型がわかっているとした。
- 表現型未知の系統を、逐次圃場試験を行って評価し、訓練データに加えることができるとした。ここで、一度に圃場試験できる系統数は 50 系統だけであるとし、全ての系統を圃場試験するまで圃場試験と予測モデルの更新を繰り返した。
- 合計 1,000 反復のシミュレーションを行った。実データの解析においてはテストデータと初期訓練集団を無作為に 1,000 回生成する反復を設けた。シミュレーションデータの解析においては、遺伝子型値および表現型値の乱数生成に 10 回の反復をとり、生成された各データについて初期訓練集団の取り方に 100 回の反復を設けた。

例えば、330 系統の実データであれば、テストデータ 100 系統と初期 50 系統をまず無作為に選び、その後は未試験系統のうち 50 系統を選んで試験することを 3 回繰り返し、4 回目に残りの 30 系統を試験して、1 回のシミュレーションが終了する。

圃場試験を用いて評価を行う系統を選ぶ基準として、表現型未知の系統から無作為に評価する系統を選ぶ場合（受動学習）と、前節で述べた uncertainty sampling に基づき評価する系統を選ぶ場合（能動学習）の 2 通りを用いて、シミュレーションによる比較・検討を行った。学習を行う系統の選択法の違いによる分類精度は、テストデータの予測を行なったときの κ 係数を用いて評価した。

シミュレーションの予測モデルには、ガウスカーネルを用いた SVM を用いた。カーネル関数の超パラメタ h は、モデル構築に用いる DNA マーカー数に固定した。この設定は、SVM の標準的なライブラリである libsvm においてデフォルトに設定されている値である（Chang and Lin, 2011）。不均衡データに対応するために、先に述べた方法で Error Cost の調節を行なった。

3-2-6. 使用したデータセット

シミュレーションによる解析には、以下に示す３つの実データと、１つの仮想データを使用した。

・　イネ遺伝資源データ（RiceDiversity dataset）

RiceDiversity（http://ricediversity.org/index.cfm）が公開しているイネ遺伝資源の遺伝子型・表現型データセットのうち、系統間の遺伝的多様性の評価に用いられた 395 系統、1,311 SNP マーカーの遺伝子型データ（Zhao et al., 2010）および、対応する表現型データ（Zhao et al., 2011）を用いた。表現型データには 34 形質が記録されていたが、本研究では一穂穎果数（Florets per bpanicle）、Aberdeen における開花期（Flowering time at Aberdeen）、Arkansas における開花期（Flowering time at Arkansas）、Faridpur での開花期（Flowering time at Faridpur）、稔実率（Panicle fertility）、穂数（Panicle number per plant）、草丈（Plant height）、種子長（Seed length）、一穂籾数（Seed number per panicle）、種子面積（Seed surface area）、種子幅（Seed Width）の 11 形質を解析に用いた。なお、395 系統のうち、対応する表現型が利用可能であった系統は 374 系統であった。ただし、形質によっては一部の系統の表現型値が欠測していたため、シミュレーションで用いられる系統数は、形質ごとにわずかに異なる。

・　CIMMYT コムギデータ（CIMMYT_wheat dataset）

CIMMYT のコムギ育種において記録された 599 系統、1,279 DArT（Diversity Array Technology）マーカーのデータを用いた（Crossa, 2010）。表現型値として、異なる環境（E1~E4、詳細不明）における収量を平均 0、分散 1 に基準化されたものが記録されている（yield_E1, E2, E3, および E4）。なお、DArT マーカーは優性マーカーであるため、SNP マーカーとは異なり $x \in \{0, 1\}$ という二値表現を用いて表記されるが、本研究ではマーカー遺伝子型を標準化してガウスカーネルを計算するため、二値入力変数についての表記法は結果に影響しないと考えられる。

・　Perez コムギデータ（Perez's_wheat dataset）

２つめのコムギのデータとして、Perez らが使用した 306 系統、1,717 DArT マーカーのデータを使用した（Perez et al., 2012）。このデータセットでは、収量（yield）が５つの異なる条件（乾燥条件かつ畝あり：drought-bed、乾燥条件かつ畝なし：drought-flat、灌漑条件かつ畝あり：irrigation-bed、灌漑条件かつ畝なし：irrigation-flat、加温条件かつ畝あり：heat-bed）で 2010 年に測定されている。このうち、乾燥条件かつ畝あり、灌漑条件かつ畝なしの２条件では、2009 年にも収量が測定されていたため、これらは２年間の平均収量を１つの表現型値として用いた。

・　トウモロコシデータ（Maize dataset）

CIMMYT コムギデータと同じ論文で使用されたトウモロコシのデータを使用した。こちらも CIMMYT が実施しているトウモロコシ育種においてデータが収集されたものであり、264 系統、1,135 SNP からなるデータセットである。表現型として２つの条件（乾燥条件：SS, Severe Stress、灌漑条件：WW, Well-Watered）での収量を解析に用いた。

・　仮想データ（RiceSim dataset）

上述したイネ遺伝資源データの SNP マーカー遺伝子型を用いて、QTL の数や遺伝率の異なる仮想形質をシミュレーションにより作成した。全 1,311 SNP マーカーから無作為に選んだマーカーに正規分布に従う効果を与え、遺伝率に応じてマーカー効果と環境効果の大きさを調節した。ここでは全マーカーに占める QTL の割合（p.QTL; percentage of QTL）が 1%、5%、10%のそれぞれの場合について、遺伝率（h2）を 0.2、0.5、0.8 として、合計 9 通りの条件で仮想データを作成した。ただし、本シミュレーションでは、QTL として効果を与えたマーカーもモデル構築に使用できるとした。実際には QTL と完全に連鎖しているマーカーが存在するとは考えづらく、したがって、通常のゲノムワイドマーカーに基づく予測は、本シミュレーションに比べて精度が低下すると予想される。

仮想データを用いた解析においては、真の遺伝子型値がわかっているため、モデルは表現型値を用いて生成し、予測精度は遺伝子型値をもとに評価した。QTL の効果をシミュレーションした場合と、実データに基づく解析との重要な違いは、前者では遺伝子型値そのものを用いて精度評価を行っているのに対し、後者では、表現型値を遺伝子型値と仮定して精度評価を行っている点である。後者の場合、環境分散が大きく遺伝率が低ければ、表現型値と遺伝子型値の隔たりが大きくなるため、精度評価の結果にも影響すると考えられる。本解析で用いた実データでは個体ごとの表現型値が公開されていないため、遺伝率の計算はできず、この点に関して実データをもとに考察を加えるのは困難であった。そのため、実データで得られた結果を補足することを主目的に、仮想データを用いた解析を行った。

3-3. 結果

・　実データにおける能動学習の有効性

実データにおいて、受動学習と能動学習により訓練集団を追加したときの分類精度（κ 係数）の変化を図 3-3 から図 3-6 に示した。いずれのデータセットにおいても、能動学習によって得られる κ 係数の値は、受動学習によって得られるものよりも、おしなべて高い値を示した。いっぽう、イネ遺伝資源データの Seed length や CIMMYT コムギデータの yield_E1 など、能動学習によってデータを選択することにより、かえって κ 係数が減少している形質もあった。

受動学習に対する能動学習による κ 係数の上昇幅は形質によって異なった。能動学習による κ 係数の増加（あるいは減少）について統計的有意性を検定するため、1 回目のデータ選択（初期集団 50 系統に基づき新たな 50 系統を追加し、100 系統の訓練データを用いた場合の κ 係数）に注目した場合の κ 係数の差を Wilcoxson の順位和検定によって両側検定した（多重検定に関する補正は用いていない）結果を表 1 に示した。例えばイネ遺伝資源データにおいて、能動学習と受動学習の κ 係数の差は、最小値−0.019（Seed length）、中央値+0.044（Seed width）、最大値+0.0575（Florets per panicle）であった。能動学習を用いた場合の κ 係数の上昇は、全 22 形質のうち 17 形質において 5%有意水準で確認された。同じ有意水準において、能動学習によって κ 係数が下降した形質は 3 形質のみであった。以上のことから、能動学習によって訓練データを選択することで、同じデータ数でも分類精度の高いモデルが構築できることが示唆された。

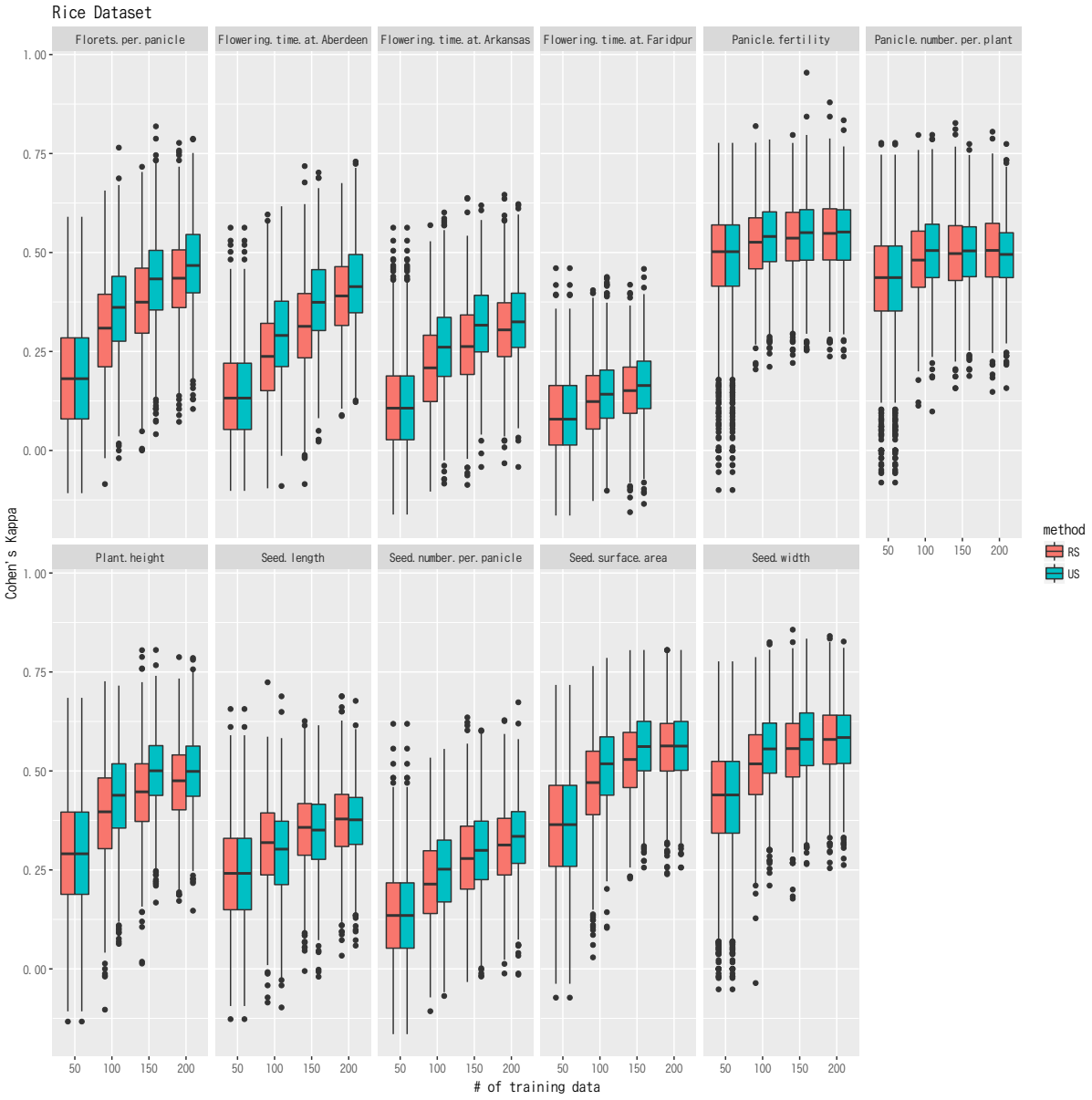


図 3-3. イネ遺伝資源データにおける κ 係数の変化

横軸にシミュレーションにおける訓練データ数を、縦軸に κ 係数をとった箱ひげ図により、能動学習（US; uncertainty sampling）と受動学習（RS; random sampling）による分類精度の推移を比較した。イネ遺伝資源データでは、多くの形質で能動学習により得られる κ 係数が受動学習のそれを上回った。

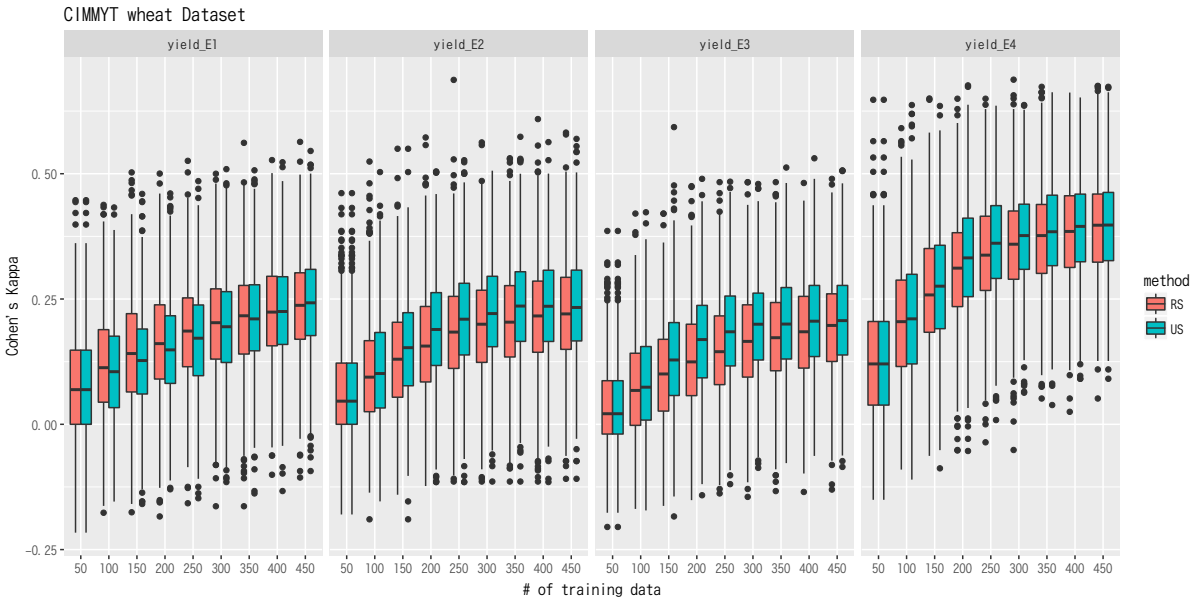


図 3-4. CIMMYT コムギデータにおける κ 係数の変化

横軸にシミュレーションにおける訓練データ数を、縦軸に κ 係数をとった箱ひげ図により、能動学習（US; uncertainty sampling）と受動学習（RS; random sampling）による分類精度の推移を比較した。CIMMYT コムギデータでは、yield_E2 や yield_E3 では能動学習により κ 係数が高くなったいっぽうで、能動学習と受動学習にほとんど差が見られない形質（yield_E4）や、能動学習を行うことにより分類精度が減少してしまう形質（yield_E1）も見られた。

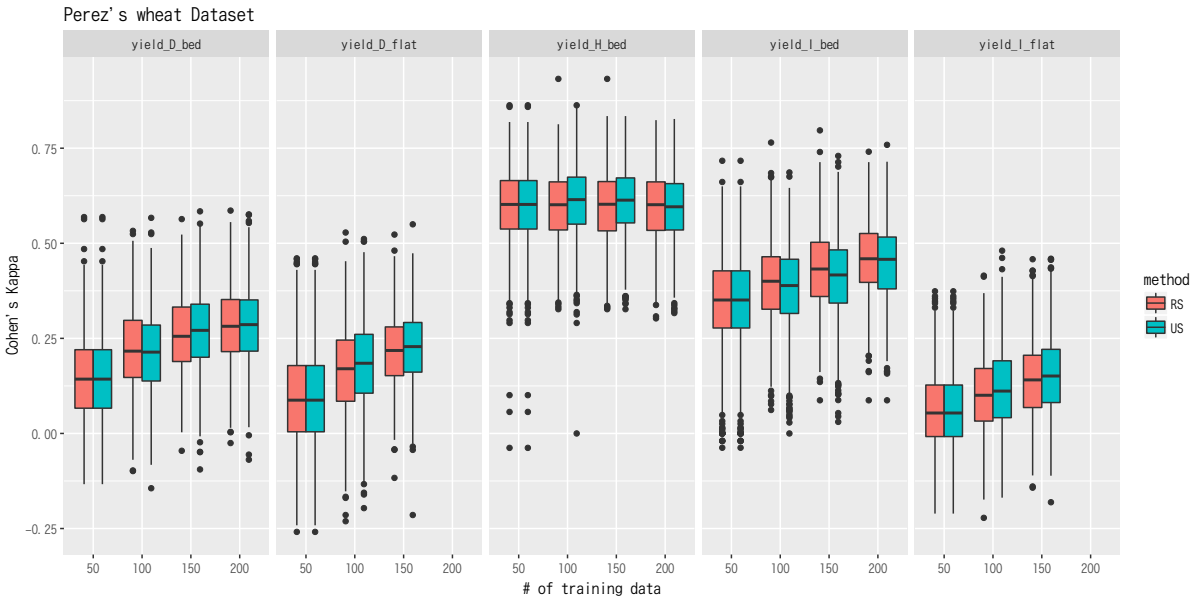


図 3-5. Perez コムギデータにおける κ 係数の変化

横軸にシミュレーションにおける訓練データ数を、縦軸に κ 係数をとった箱ひげ図により、能動学習（US; uncertainty sampling）と受動学習（RS; random sampling）による分類精度の推移を比較した。Perez コムギデータの形質は全て収量であるが、能動学習によって κ 係数が増加している環境と、逆に減少している環境が見られた。

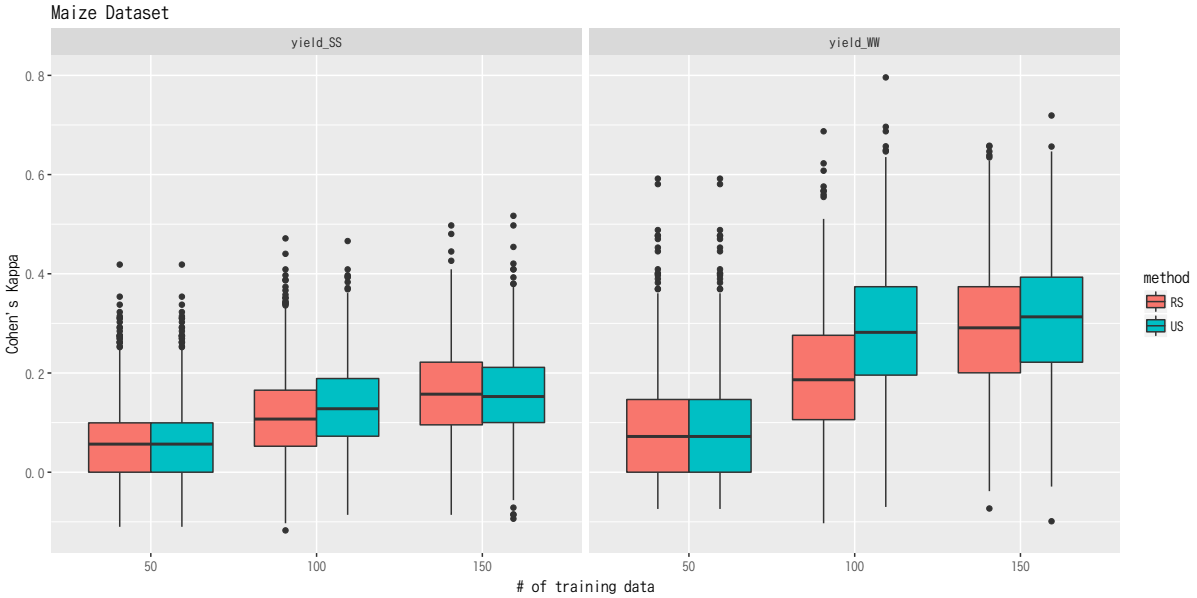


図 3-6. トウモロコシデータにおける κ 係数の変化

横軸にシミュレーションにおける訓練データ数を、縦軸に κ 係数をとった箱ひげ図により、能動学習（US; uncertainty sampling）と受動学習（RS; random sampling）による分類精度の推移を比較した。トウモロコシデータでは、yield_WW では能動学習による精度の向上が明瞭に確認された。また、yield_SS でも 1 度目の選択（初期集団 50 系統から、新たに 50 系統を加えた場合）では能動学習によって κ 係数が向上した。

表 3-1. 実データにおける能動学習と受動学習による κ 係数の差

初期データ 50 系統に基づき新たな訓練データ 50 系統を選んだ後の κ 係数について、能動学習と受動学習の差の、1000 反復のシミュレーションにおける平均値を計算した（3 列目；US-RS)。また、κ 係数の差が有意に異なるかを Wilcoxson の符号和検定によって両側検定し、5%有意な形質を * で、1%有意な形質を ** で、0.1%有意な形質を *** で示した。

Dataset	Trait	US-RS	*p*-value
RiceDiversity	Florets.per.panicle	0.057	< 2.2E-16 ***
	Flowering.time.at.Aberdeen	0.055	< 2.2E-16 ***
	Flowering.time.at.Arkansas	0.051	< 2.2E-16 ***
	Flowering.time.at.Faridpur	0.019	1.51E-05 ***
	Panicle.fertility	0.015	5.94E-04 ***
	Panicle.number.per.plant	0.020	3.33E-06 ***
	Plant.height	0.045	2.32E-14 ***
	Seed.length	-0.019	2.11E-04 ***
	Seed.number.per.panicle	0.029	2.46E-08 ***
	Seed.surface.area	0.045	< 2.2E-16 ***
	Seed.width	0.044	< 2.2E-16 ***
CIMMYT_wheat	yield_E1	-0.010	0.0390 *
	yield_E2	0.010	0.0444 *
	yield_E3	0.014	0.0065 **
	yield_E4	0.006	0.2914
Perez's_wheat	yield_D_bed	-0.005	0.3577
	yield_D_flat	0.015	0.0048 **
	yield_H_bed	0.011	0.0032 **
	yield_I_bed	-0.015	0.0073 **
	yield_I_flat	0.013	0.0083 **
Maize	yield_SS	0.020	1.16E-07 ***
	yield_WW	0.093	< 2.2E-16 ***

・　仮想データにおける能動学習の有効性

異なる QTL 数と遺伝率を仮定して表現型を生成した仮想データに対しても、同様に能動学習と受動学習による κ 係数の変化を図示し（図 3-7)、1 回目のデータ選択を行なった後の κ 係数に関する手法間差の検定結果を表にまとめた（表 3-2)。解析の結果、仮想データを生成する設定を変えた場合でも、能動学習により得られる κ 係数が受動学習のそれを下回ることはなく、遺伝率や QTL の数に対して、能動学習は頑健に機能すると考えられた。

なお、遺伝率や QTL の割合が能動学習に及ぼす影響については、図と表から明瞭な傾向・解釈可能な結果は読み取れなかった。例えば QTL の割合が 1%の場合に注目すると、遺伝率の増加とともに能動学習と受動学習の差に関する p 値は増加しているが、QTL の割合が 10%になるとこの傾向は逆になっている。

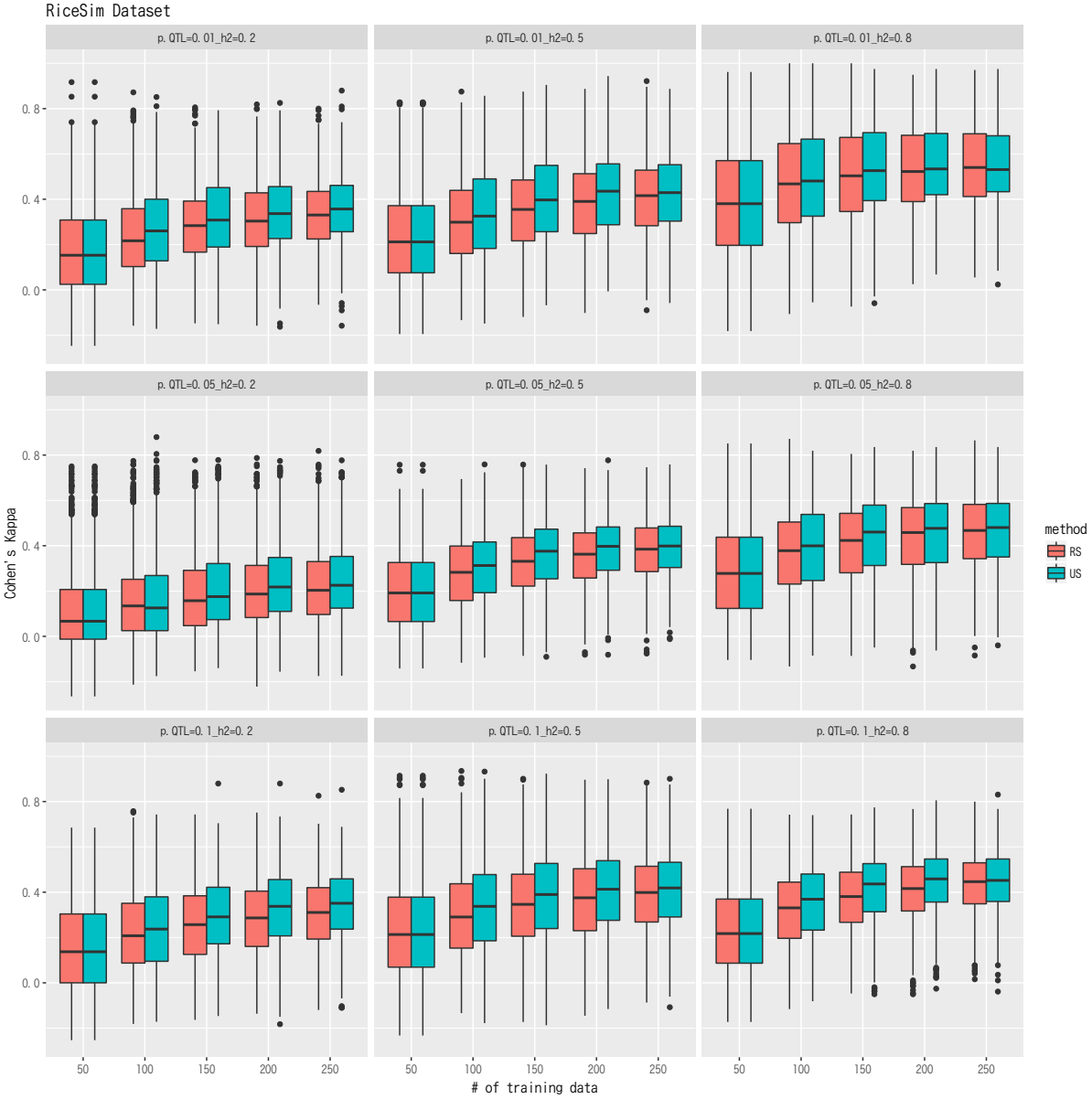


図 3-7. 仮想データにおける κ 係数の変化

横軸にシミュレーションにおける訓練データ数を、縦軸に κ 係数をとった箱ひげ図により、能動学習（US; uncertainty sampling）と受動学習（RS; random sampling）による分類精度の推移を比較した。最上段に QTL の割合が 1%の場合、中段に 5%の場合、最下段に 10%の場合を図示し、左列に遺伝率が 0.2 の場合を、２列目に 0.5 の場合を、３列目に 0.8 の場合を図示した。

表 3-2. 仮想データにおける能動学習と受動学習による κ 係数の差

初期データ 50 系統に基づき新たな訓練データ 50 系統を選んだ後の κ 係数について、能動学習と受動学習の差の、1000 反復のシミュレーションにおける平均値を計算した（3 列目；US-RS)。また、κ 係数の差が有意に異なるかを Wilcoxson の符号和検定によって両側検定し、5%有意な形質を * で、1%有意な形質を ** で、0.1%有意な形質を *** で示した。

Ratio of QTL	Heritability	US-RS	*p*-value
1%	0.2	0.030	0.00045 ***
	0.5	0.032	0.00110 **
	0.8	0.023	0.04025 *
5%	0.2	0.008	0.63304
	0.5	0.023	0.00229 **
	0.8	0.021	0.00815 **
10%	0.2	0.017	0.04143 *
	0.5	0.034	0.00023 ***
	0.8	0.032	0.00001 ***

・　能動学習により選ばれやすい系統

能動学習によって高頻度で選ばれる系統があるのかを検証するために、実データにおける 1 回目のデータ選択に注目して、1000 回のシミュレーションのうち何回その系統が選ばれたかをヒストグラムを用いて示した（図 3-8 から図 3-11）。また、異なる形質間でも共通の系統が選ばれているかどうかを検証するために、選ばれる頻度が高かった 50 系統を形質ごとに抽出し、ベン図によってその共通性を示した（図 3-12 から図 3-15）。ただし、イネ遺伝資源データは 11 形質を含むため、ベン図の視認性を高めるため、恣意的に形質を 3 つのグループに分割して作図した。

ヒストグラムをみると、ほぼすべてのデータセット・形質でヒストグラムの左側に大きな度数が分布しており、選ばれる回数が極端に少ない系統があったことが読み取れる。いっぽうで、ヒストグラムの右裾にもいくつかの系統が分布しており、系統ごとの選ばれやすさに大きな違いがあったことがわかる。例えばイネ遺伝資源データの Florets per panicle では、1,000 回中 100 回未満しか選ばれなかった系統が多数ある一方で、一部の系統は 300 回以上選ばれている（図 3-9）。すべての系統から無作為に選択したとすると、系統ごとの被選択回数は二項分布 $Bin(n=N-100, p=50/(N-100))$に従うはずである（ここで、$N$ はデータセットの全系統数）が、図示されたヒストグラムのほとんどは、明らかにそのような二項分布には従っていない。これは、uncertainty sampling が機能し、能動的に未試験の系統を選択できていることを示唆している。

なお、必ずしも同じ系統を選ぶことで分類精度の改善に繋がるとは限らない。追加されるべき系統は、その時点で既知であるデータに依存するべきだからである。実際に、イネ遺伝資源データの Seed length では選ばれやすいデータとそうでないデータが明確に分かれていたが、能動学習によりむしろ κ 係数が低下していた（図 3-8, 表 3-1）、逆に、トウモロコシデータでは選ばれた回数が極端に多い、あるいは少ない系統は見られなかったが、能動学習によって有意に κ 係数の値が増加していた（図 3-11, 表 3-1）。

ベン図を用いて形質ごとに選択されやすかった系統の重なりを調べたところ、類似の形質であっても同じ系統が選ばれやすいわけではないことがわかった。例えばイネ遺伝資源データでは、3 つの異なる地点の開花日数についての能動学習で優先的に選ばれた系統のうち、3 地点全てにおいて共通して選ばれやすかった系統は 6 系統のみであった（図 3-12, a）。ただし、CIMMYT コムギデータでは、4 つの環境で測定された収量について、比較的同じ系統を選ぶ傾向も見られた（図 3-13）。

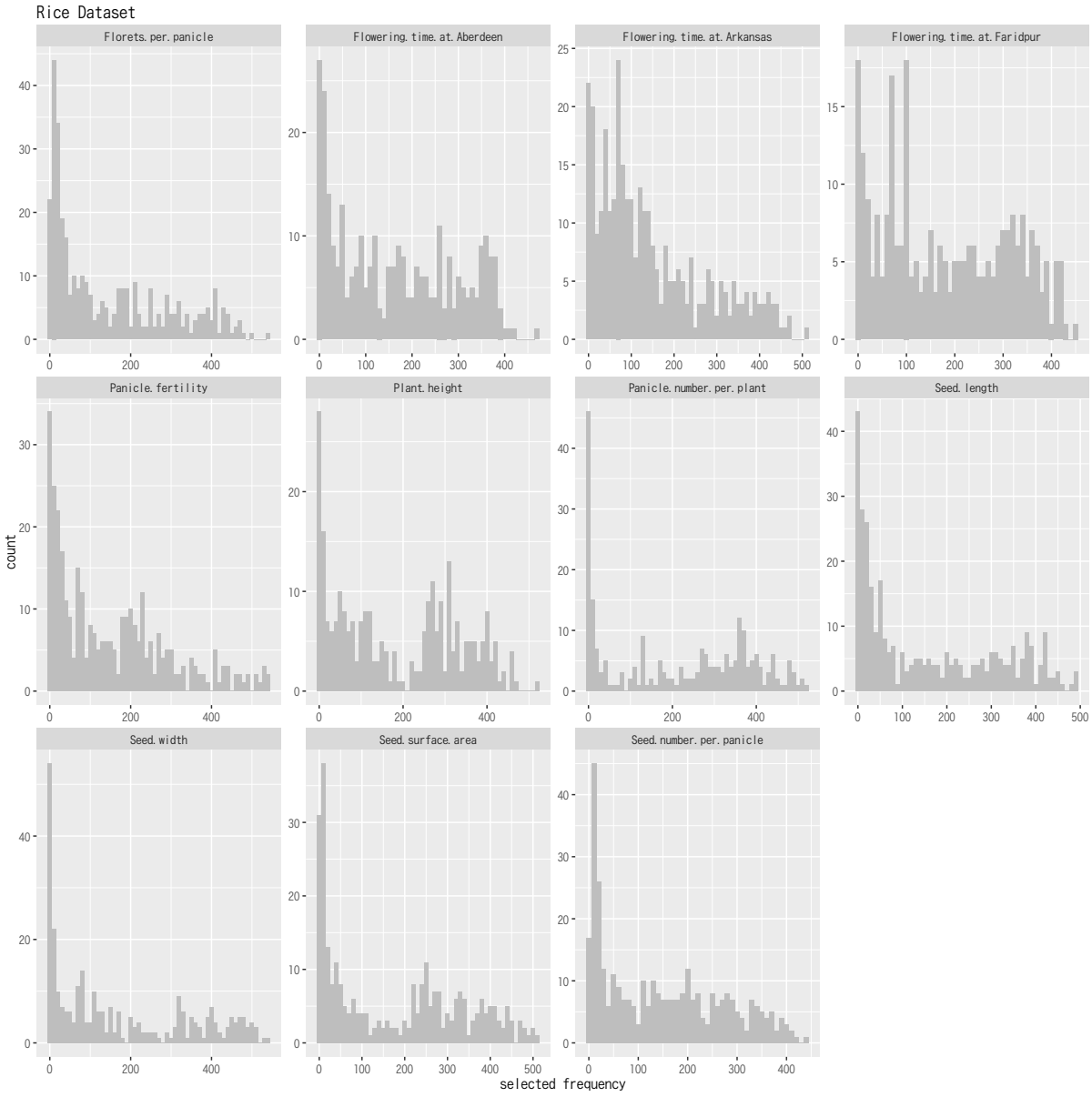


図 3-8. イネ遺伝資源データにおける被選択回数のヒストグラム

イネ遺伝資源データについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において選ばれた回数を系統ごとに数え上げ、ヒストグラムで示した。すべての形質で、選ばれやすい系統や選ばれにくい系統が存在したことがわかる。

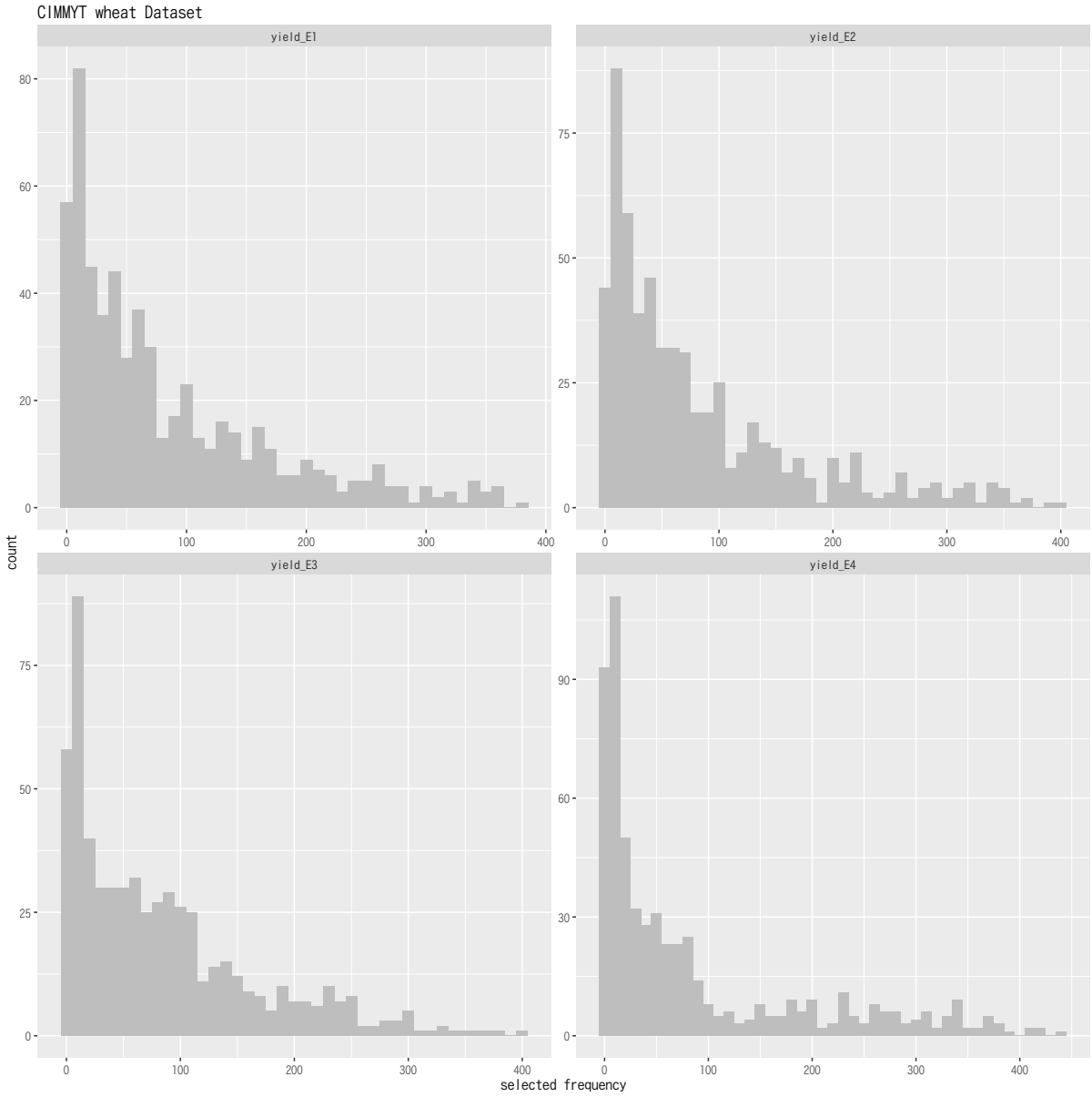


図 3-9. CIMMYT コムギデータにおける被選択回数のヒストグラム

CIMMYT コムギデータについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において選ばれた回数を系統ごとに数え上げ、ヒストグラムで示した。すべての形質で、選ばれやすい系統や選ばれにくい系統が存在したことがわかる。

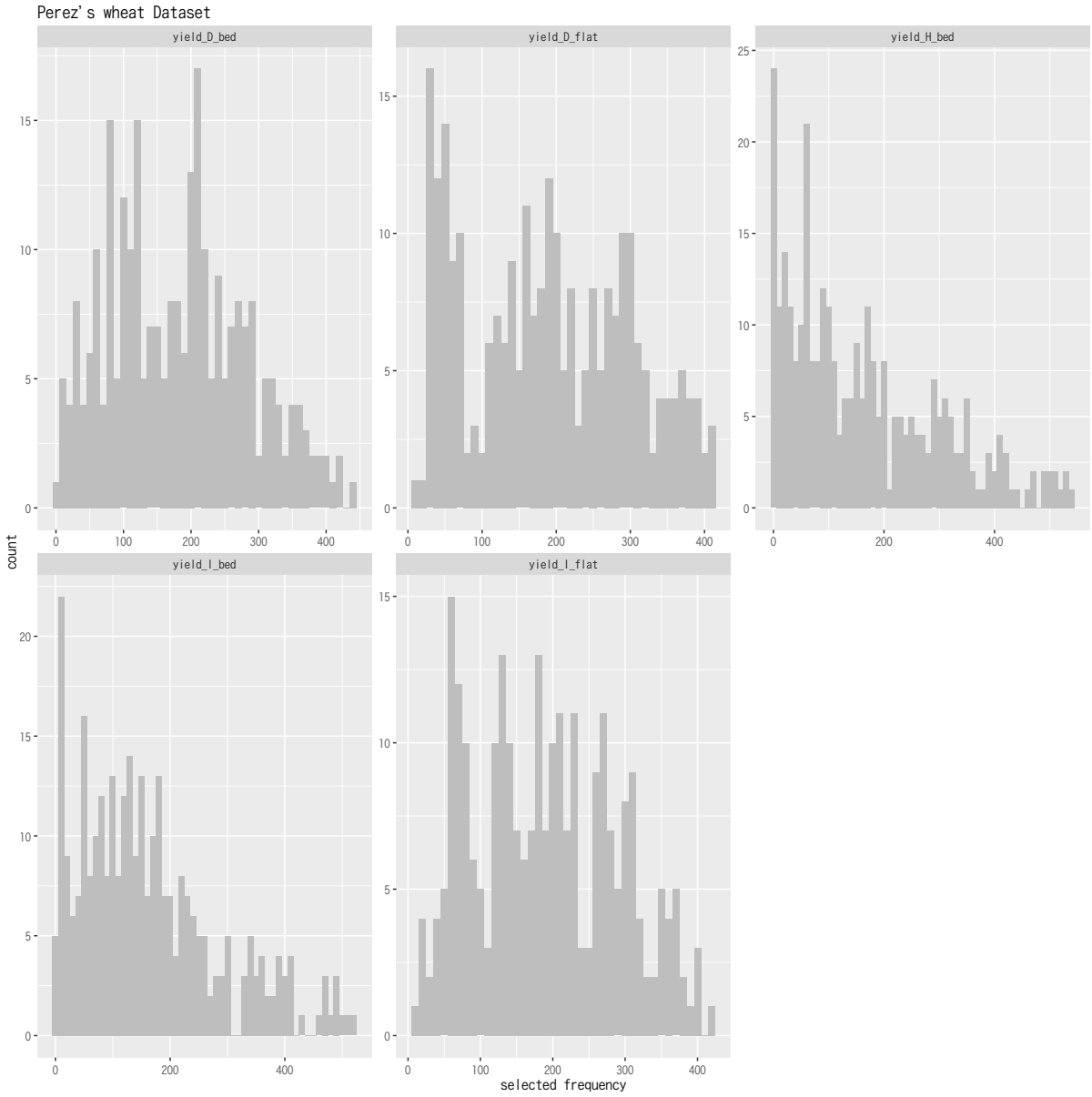


図 3-10. Perez コムギデータにおける被選択回数のヒストグラム

Perez コムギデータについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において選ばれた回数を系統ごとに数え上げ、ヒストグラムで示した。このデータセットでは、形質によってヒストグラムの概形が大きく異なっていた。

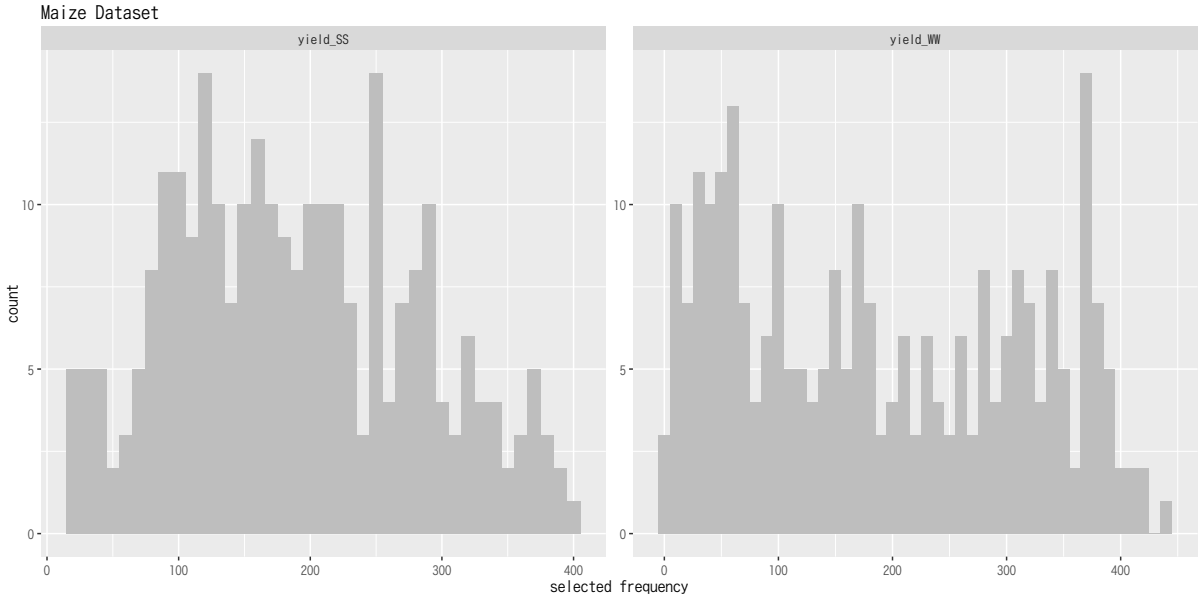
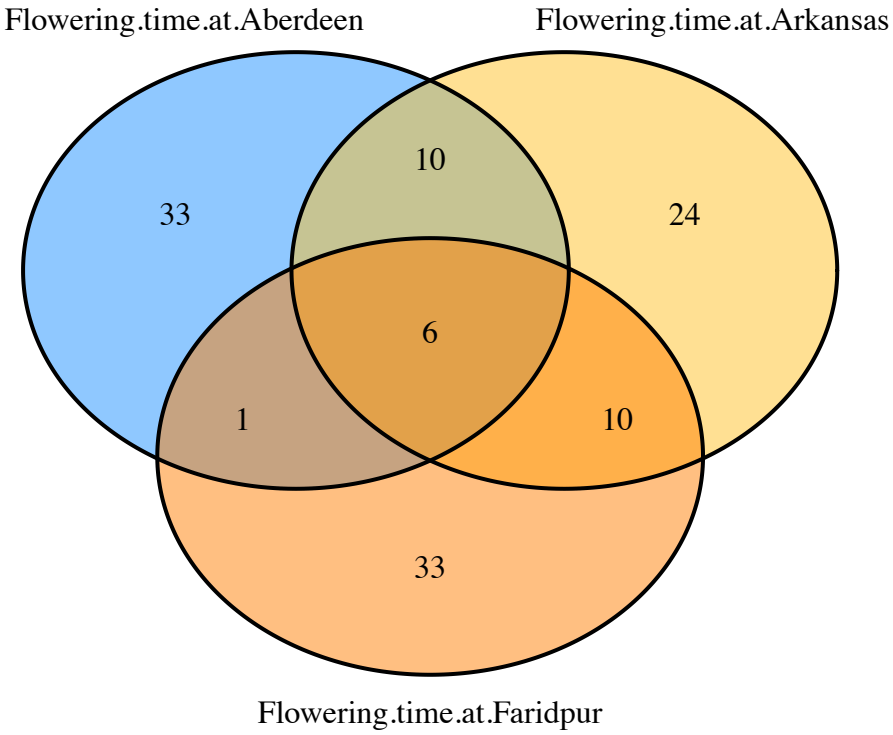


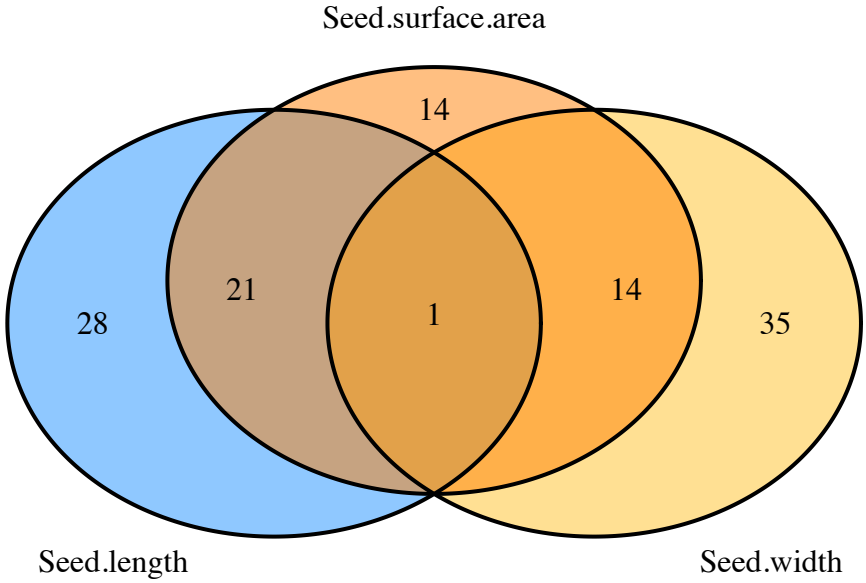
図 3-11. トウモロコシデータにおける被選択回数のヒストグラム

トウモロコシデータについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において選ばれた回数を系統ごとに数え上げ、ヒストグラムで示した。このデータセットでは、他のデータセットに比べると、選ばれやすい系統とそうでない系統が明瞭に分かれていなかった。

(a)


Flowering.time.at.Aberdeen
Flowering.time.at.Arkansas
33
10
24
6
1
10
33
Flowering.time.at.Faridpur

(b)


Seed.surface.area
14
28
21
1
14
35
Seed.length
Seed.width

(c)

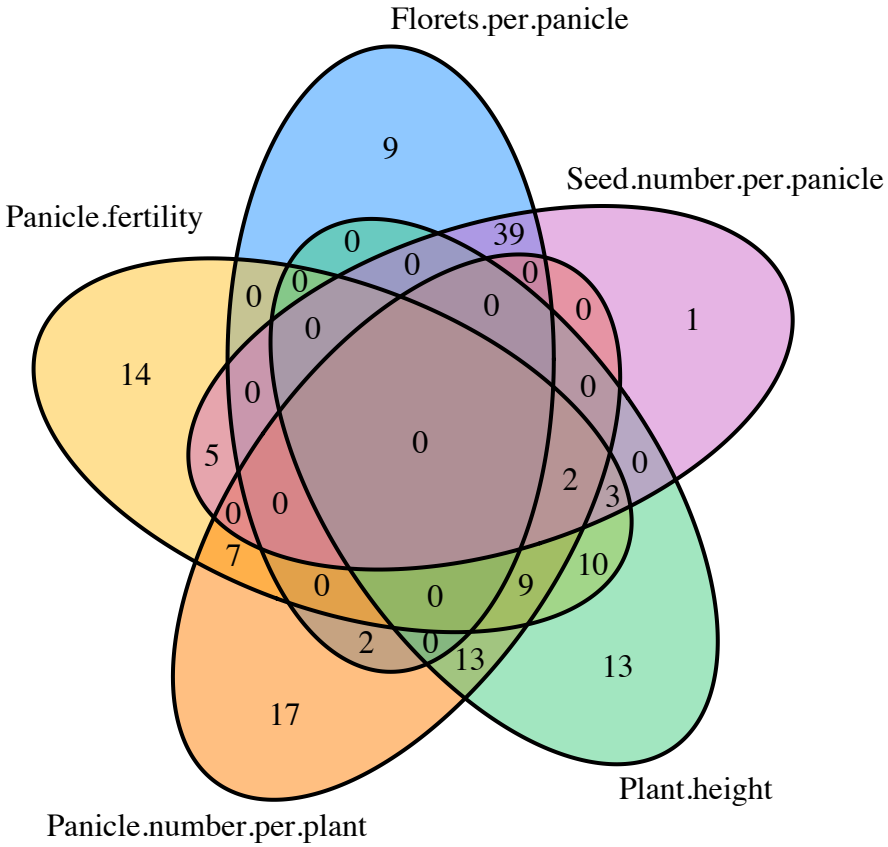


図 3-12. イネ遺伝資源データにおいて選ばれた系統の重複

イネ遺伝資源データについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において、選ばれた回数の多かった上位 50 系統を形質ごとに抽出し、その重なりをベン図によって示した。ただし、イネ遺伝資源データには 11 の形質が含まれるため、恣意的に形質を 3 つに分けて図示した。(a)では開花までの日数に関する 3 形質を、(b)では種子の形状・サイズに関する 3 形質を、(c)ではそれらに含まれない 5 形質を図示した。

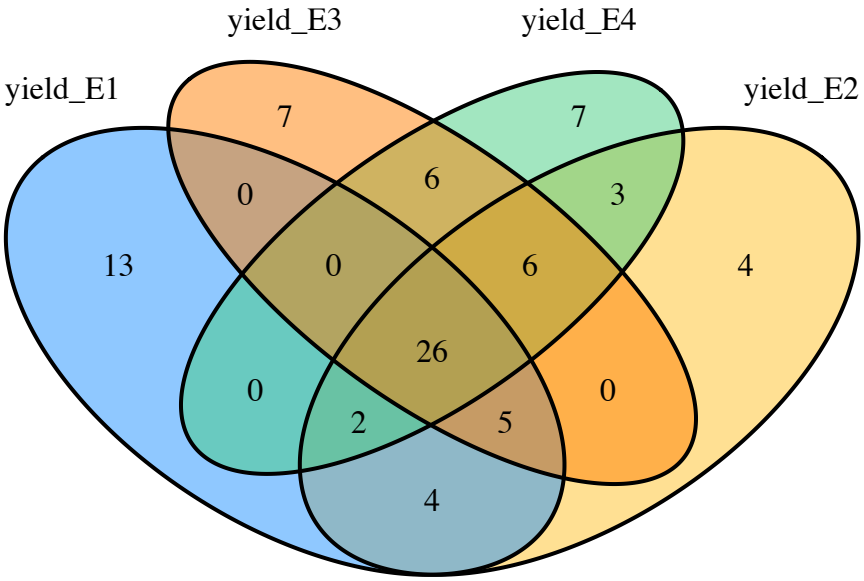


図 3-13. CIMMYT コムギデータにおいて選ばれた系統の重複

CIMMYT コムギデータについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において、選ばれた回数の多かった上位 50 系統を形質ごとに抽出し、その重なりをベン図によって示した。

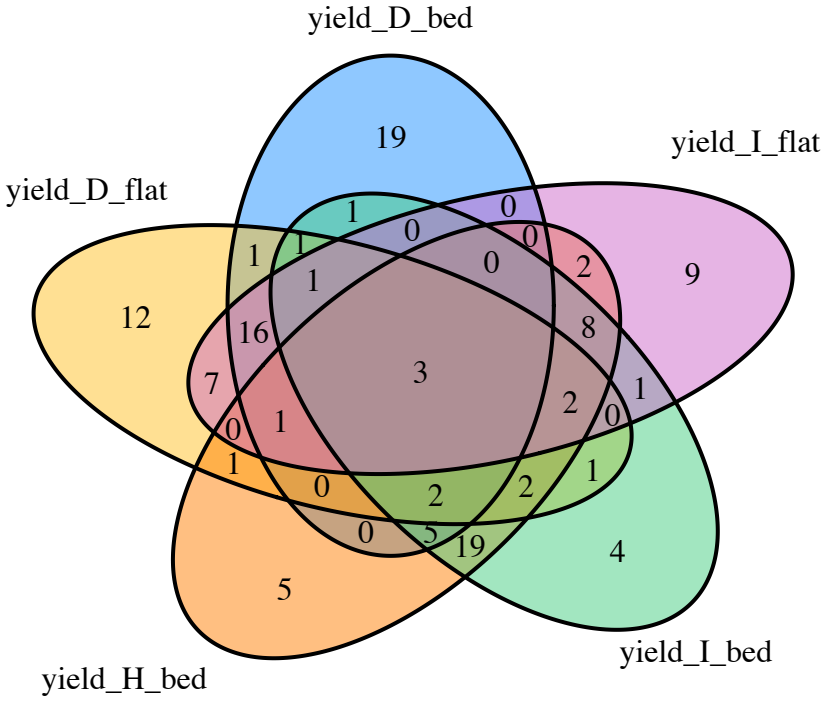


図 3-14. Perez コムギデータにおいて選ばれた系統の重複

Perez コムギデータについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において、選ばれた回数の多かった上位 50 系統を形質ごとに抽出し、その重なりをベン図によって示した。

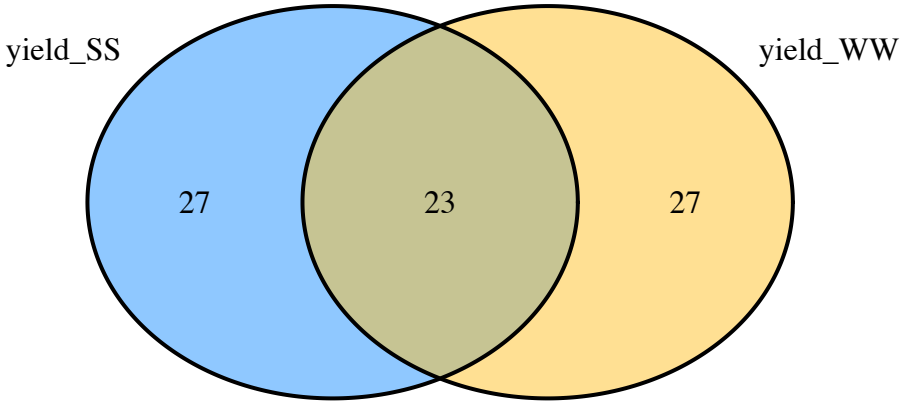


図 3-15. トウモロコシデータにおいて選ばれた系統の重複

トウモロコシデータについて、初期データ 50 系統に基づき新たな訓練データ 50 系統を選ぶ初回のデータ追加において、選ばれた回数の多かった上位 50 系統を形質ごとに抽出し、その重なりをベン図によって示した。

3-4. 考察

本研究では、分類器を用いたゲノミック予測において、訓練データを uncertainty sampling によって能動的に追加することで分類器の精度が向上するかどうかを、複数の実データおよび 1 つの仮想データを用いたシミュレーションにより検証した。実施したシミュレーションの範囲では、uncertainty sampling による訓練データの追加は、ほぼ全てのデータセット・形質について分類精度（κ 係数）を効率的に改善した。ゲノミック予測で扱われるデータや状況は、これまでに能動学習の応用研究が行われてきた画像解析や自然言語処理とは違った性質を持つと考えられ、必ずしも先行研究と同様に能動学習が機能することは保証されていなかった。本研究で得られた結果は、ゲノミック予測においても、能動学習を活用して効率よく精度の高いモデルが構築できることを強く示唆するものである。

シミュレーションでは、多数の系統についてマーカー遺伝子型が既知であるいっぽうで、表現型は未知であると仮定した。この仮定は、今後ますます現実のものとなると予想される。マーカー遺伝子型の取得コストは現在でも低下しており、約 3,000 系統のイネ遺伝資源についてゲノム情報が公開された（Wang et al., 2018）ことに代表されるように、遺伝資源や主要な育成系統のゲノム情報が利用可能になりつつある。いっぽうで、表現型の取得コストも高速フェノタイピング技術によって下がりつつある（e.g. Araus and Carins, 2014）が、栽培管理そのもののコストを下げることは容易ではない。また、目的に応じて同じ表現型であっても異なる地域や条件で測定する必要もある。あるいは気候変動を例に考えれば、数年前の表現型値と現在の表現型値も同じとは言えず、絶えず表現型の取得が継続される必要があるだろう。こうした状況を鑑みれば、多数の系統のゲノム情報と、その一部に関する既知の表現型をもとに、どの系統を栽培試験するかを適切に決定するという問題は、多くの場面で生じるものと考えられる。

能動学習において、選ばれやすい系統は、同じデータセットでも形質によって大きく異なっていた。訓練集団最適化に関する先行研究（Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Rincent et al., 2017a）では、ゲノムデータだけを基に訓練データを決定するため、どのような形質に対しても同じ系統が選ばれる。これは問題設定の違いによるところが大きい。訓練集団最適化では、全く表現型が取得されていない場合に、どの系統を栽培試験するかを議論しているため、事前に得られた一部の系統に関する表現型値を用いることが想定されておらず、そのような方法は未だ提案されていない。ただし、5-4 節で触れるように、既存の訓練集団最適化の研究を拡張することで、今回の能動学習で想定したような、取得済みの表現型情報を活用して次に試験するべき系統を選ぶ方法を構築することは可能だと考えられる。本研究のシミュレーションで想定された状況を踏まえて訓練集団最適化に関する手法を発展させることは、興味深い研究テーマの 1 つとなる。

実際の育種では、ある形質は質的であり別の形質は量的であるといったことも多く、したがって、形質ごとに異なる予測モデルを用いることが不可欠な場合も多い。能動学習であれ訓練集団最適化であれ、今までに研究された理論は「あるモデルにおける」最適な訓練データの選択を実現しようとするものであり、「複数のモデル」あるいは「複数の育種目標」に対して最適な訓練データを選ぶための理論にも取り組んでいく必要がある。ただし、その道程は容易ではないことが

予想される。予測モデルが違えば、最適な訓練データも、それを得るための最適化指標（能動学習における、質問戦略）も異なるためである。つまり、複数の目的関数（個々の予測モデル・対象形質に対して定められた、質問戦略や最適化指標）を同時に最適化する問題を扱わなければならない。このような問題は多目的最適化（multi-objective optimization）として知られており、選抜や交配計画の最適化へ応用する動きも見られる（Hunter and McClosky, 2016; Akdemir et al., 2018）。多目的最適化が有効な解決策となりうるかは現時点で不明だが、いずれにせよ、現実に生じる複雑な状況で最適な訓練集団を決定するには、より高度な数理科学的手法・複雑なアルゴリズムの導入が必要だと考えられる。

能動学習の有効性は強く示唆されたものの、κ 係数の増加量はそれほど劇的ではないように思われた（表 3-1, 3-2）。1,000 回のシミュレーション結果について詳細に調べると、κ 係数の増加量が最も大きかったイネ遺伝資源データの Florets per panicle の場合でも、κ 係数が減少したシミュレーションが 1,000 回中 320 回あることがわかった。すなわち、能動学習によりデータを選択することが必ず良い結果を招くことは保証されず、あくまで期待値の意味での有効性しか本研究では示されていない。ただし、能動学習そのものが常に精度の向上を保証するものではないため、この結果は妥当なものと言えるだろう。

本研究が採用した 2 クラス SVM と uncertainty sampling は、能動学習の最も単純な方法にすぎず、適用する能動学習の手法を変更することによって、さらに効率よく分類精度を高めることも可能だと考えられる。分類モデルは機械学習において良く研究されており、本研究で用いた SVM のほか、Random Forest（Breiman, 2001）や XgBoost（Chen and Guestrin, 2016）をはじめとして、近年その利用が加速する深層学習などの手法も利用を検討するべきだろう。ただし、学習器ごとに最適な能動学習アルゴリズムが異なることには注意が必要である。なお、Random Forest については query by committee と呼ばれる質問戦略が古くから提案されている（Seung et al., 1992）。

また、本研究のように 50 系統をまとめて選択する（batch mode sampling）場合には、1 サンプルずつ追加する場合とは異なる戦略を採用することが望ましいであろう。例えば、Hoi らは複数の訓練データに対して、それらが追加された場合の情報量を計算するという方針を提案している（Hoi et al., 2006）。あるいは、入力変数を用いてクラスター解析を行い、その結果を活用するという方法も検討されている（Patra and Bruzzone, 2012）。植物育種で扱われるデータでは、複数のクラスター（分集団）の存在が想定されることも多いため、この方法が機能する可能性は十分にあると予想される。

遺伝率や形質を支配する QTL の数は、ゲノミック予測の精度に影響する因子だと考えられている。本研究では、これらの因子が異なる複数の仮想データを用いてシミュレーションを行い、能動学習の成否に関する影響を検証した。今回のシミュレーションで得られた結果によれば、これらの因子は能動学習の効果に大きな影響を及ぼすとは言えず、能動学習が形質の遺伝様式に対して頑健な手法であることが示唆された。ただし、実データにおいては　一部の形質で能動学習が κ 係数を低下させた例も見られた。能動学習がどのような条件で特に有効なのか、あるいは、有効ではないのかについては、より詳細な検討によって結論づける必要がある。本研究で用いた仮

想データは比較的マーカー密度が低く、また、系統数も多くはない。より大規模なデータセットを用いることは重要であろう。また、本研究では検討しなかった因子、例えば主働遺伝子の有無や集団構造の強弱、非相加的効果（優性効果やエピスタシス）の存在についても検討する価値がある。

本研究では、二値分類問題として表現されるゲノミック予測に能動学習を用いた世界で初めての例である。得られた結果は能動学習の有効性を強く示唆するものであり、能動学習の応用研究を推進することで、ゲノミック予測の運用コストを大きく左右する表現型の取得コストを低減できることが期待される。ただし、本研究で検証されていない要素も数多くあり、また、能動学習アルゴリズムにも改善の余地がある。より具体的な状況でのシミュレーションや実証研究、より洗練されたアルゴリズムの適用と開発が、これからの重要な研究課題と考えられる。

4. ベイズ最適化に基づく優良系統の効率的発見

4-1. 序論

近代品種の多くは栽培化に伴い有用な遺伝変異、とりわけ病害抵抗性や生理的ストレスへの耐性に関連する遺伝子を失っていると考えられており、それゆえ、遺伝資源のもつ多様な遺伝的変異を近代品種に導入することは育種における重要課題の 1 つである（Tanksly and McCouch, 1997; Jordan et al., 2011; McCouch et al., 2013）。したがって、遺伝資源の収集・保存活動は世界規模で継続されており、例えば USDA-ARS（United States Department Agriculture – Agriculture Research Service; 米国農業研究事業団）の所管する NPGS（National Plant Germplasm System; 国立植物遺伝資源システム）だけでも、様々な植物種について延べ 57 万点を超える遺伝資源を保存している（Byrne et al., 2018）。

このように膨大な数の遺伝資源が育種に利用できる状態にあるが、この「候補の多さ」こそが、育種家が遺伝資源を利用することを難しくしてきた。例えば、あるストレス環境に対する適応性の高い系統を遺伝資源からスクリーニングしたいとしよう。もし、あらゆる遺伝資源系統を当該ストレス環境下で（十分な個体反復や年次反復をとって）評価できれば、最も生育の良かった系統を有望な育種母本として選ぶことができる。しかし、現実には限られた人的・金銭的資源で育種を行う必要があり、数千〜数万点にも及ぶ遺伝資源系統を評価することは不可能に近い。したがって、育種家は、遺伝資源のどの系統を評価するべきかを、何らかの合理性をもって決断する必要がある。

そこで、コア・コレクション（遺伝資源の 10%未満の系統数）やミニ・コア・コレクション（遺伝資源の 1%未満の系統数）と呼ばれる、できるだけ幅広く遺伝資源の多様性をカバーするように設計された集団が構築されてきた。コア・コレクションの設計には様々な方法が提案されている（Odong et al., 2013）が、一般には、地理的起源や DNA マーカーの情報、および、既知の表現型情報を考慮しながら作成される（e.g. Kaga et al., 2012）。あるいは、FIGS（Focused Identification of Germplasm Strategy）と呼ばれる、遺伝資源の表現型情報と地理的情報（起源地域とその環境に関する情報）を結びつけることによって、注目している表現型について有望な形質を有することが期待される系統群を抽出するといった試みも研究されている（Bari et al., 2012; Khazaei et al., 2013）。現在では、このように何らかの事前情報をもとにして小規模な集団を抽出し、それを評価・スクリーニングするのが一般的な遺伝資源の活用方法だといえる。

これらは現実的な制約の中で遺伝資源を活用する優れた枠組みを提供しているが、当然ながら問題もある。例えばコア・コレクションの設計は表現型の多様性を幅広くカバーするように行われるが、表現型として利用されるのは早晩性や種子形態などの基本的な形質が主である。これら形質の重要性は言うまでもないが、個々の育種目標に対応する形質に関する表現型の多様性は保証されないため、コア・コレクションに有用遺伝子を含む系統が含まれているとは限らない。いっぽう、FIGS ではゲノム情報を活用していない点が懸念される。

以上の議論を踏まえれば、現実的なスケール（評価される系統数）で、しかし、特定の系統群に限定されることのない、遺伝資源の活用法が必要であることがわかる。そこで注目されつつあ

るのが、ゲノミック予測を用いて有用遺伝資源を探索する方法である（Gorjanc et al. 2016; Yu et al., 2016）。表現型に比べて、マーカー遺伝子型は遥かに高速かつ安価に取得することができる。よって、多数の遺伝資源をシーケンスしてマーカー遺伝子型を取得しておけば、その一部について表現型を評価して予測モデルを構築し、表現型を評価しなかった系統の遺伝子型値を予測することで、シーケンスした系統の全てを対象にスクリーニングを行うことができる。あるいは、これを繰り返し行うことで、予測モデルを更新しながら有用遺伝資源を逐次的に探索することもできる。なお、ゲノミック予測は、通常の育成系統や育種集団だけでなく遺伝資源に対しても、予測に基づき優れた系統を発見できることが経験的に確かめられている（e.g. Pace et al. 2015; Chang et al. 2016）。

ゲノミック予測（特に GBLUP モデル）が多数の遺伝子に支配される量的形質に有効な手法であることは、遺伝資源探索においても強調するに値する。遺伝資源の活用は病害抵抗性やストレス耐性を対象とすることが多いため、主働遺伝子の同定を行い近代品種へ導入するというマーカー選抜のほうがゲノミック選抜よりも優れている場合もあるだろう。しかしながら、ストレス耐性であっても微働遺伝子の存在が無視できるわけではなく、とりわけ「ある地域・環境への適応性」のような育種目標が設定される場合には、複合的なストレス環境への適応性が求められると考えられる。単一遺伝子座だけでそのような適応性を改善できるとは限らず、むしろ複数のストレス耐性遺伝子座を考えるべきであろう。また、病害抵抗性についても、1 つの遺伝子座による抵抗性はウイルスや細菌の進化によって打破されやすいことが問題視されている（Palloix et al., 2009）。また、Chang らのタバコ輪点ウイルスに対する耐性に関する研究では、1 つの主働遺伝子が GWAS で検出されたものの、ゲノミック予測はこの主働遺伝子だけでは説明できないウイルス耐性を説明できたとされている（Chang et al. 2016）。このような観点からは、ゲノミック予測に基づき有望な育種母本をスクリーニングし、近代品種との交配と選抜を通してストレス環境への適応性や病害抵抗性を高めることも、極めて有効なアプローチだと思われる。

ゲノミック予測を用いて効率よく優れた系統を探索するという目標は、ある種の最適化問題と解釈できる。すなわち、ゲノムから遺伝子型値が定まる生物学的なプロセスを、遺伝子型 $\mathbf{x}$ を入力として遺伝子型値 u を出力とする関数 $u(\mathbf{x})$と捉えれば、優れた系統を探索することは、この関数 $u(\mathbf{x})$の最大値を与える $\mathbf{x}$ を探索することに相当する。したがって、数理科学で提案された最適化アルゴリズムを応用することで、優れた遺伝資源の探索戦略を構築できると考えられる。

ただし、この最適化においては、関数 $u(\mathbf{x})$が未知であるということに注意する必要がある。我々に可能な操作は、真の関数から得られる遺伝子型値 u と環境誤差 e の和である表現型 y から、適当なモデル $f(\mathbf{x})$を用いて関数 $u(\mathbf{x})$を推定することだけである。もし我々のモデル $f(\mathbf{x})$が真の関数 $u(\mathbf{x})$を含むと仮定しても、有限の訓練データから推定されたモデルは真の関数と一致するとは限らない。このように、ゲノミック予測を用いた遺伝資源探索は、black-box 最適化と呼ばれる、最適化したい関数の式が与えられていない状況で最大値を探索する最適化問題となる。

また、遺伝資源探索への応用を考えた場合には、最適化に用いるデータ点（マーカー遺伝子型と表現型の組）ができるだけ少ないアルゴリズムでなければならない。例えば遺伝的アルゴリズムや粒子群最適化は良く知られた black-box 最適化アルゴリズムであるが、これらは比較的多く

の入出力を並列評価することで高速に最適解を探索することにその長所がある。通常の最適化問題では、入力 **x** から出力 y を得るコストは無視できるほど小さいと考えるのが普通であるため、このような最適化アルゴリズムを用いることもできる。しかし、いま考えている遺伝資源探索では、出力 y は圃場試験を行わなければ得ることができない。よって、評価される系統数をできる限り抑え、効率よく最大値に迫ることが極めて重要である。

ベイズ最適化（Bayesian optimization）は、まさにこのような状況に適合するアルゴリズムだと考えられる。ベイズ最適化の理論や概念は古くから提案されており、いくつかの入出力をもとに、最適化したい未知の関数をベイズ推定しつつ、推定された関数の事後分布を用いて次に出力を得るべき入力を決定することを繰り返すことで、少数のデータから効率よく black-box 最適化を実現するアルゴリズムである（Mockus, 1994; Jones et al., 1998; Shahriari et al., 2016）。ベイズ最適化は、入力に対応する出力を得るコストが大きい場合には極めて有効であり、例えば Seko らは、融点の高い物質を効率よくスクリーニングするという問題にベイズ最適化を応用している（Seko et al., 2014）。また、深層学習をはじめとする機械学習の超パラメータの調節にもベイズ最適化は利用されている（Snoek and Larochelle, 2012）。前者の例では物質の融点をコンピュータで理論計算するため、後者の例では超パラーメタの良し悪しを交差検証によって評価するために長い計算時間がかかることから、ベイズ最適化が用いられる。このような特性は、遺伝資源探索への応用において大きな長所になると考えられる。

以上のように、ゲノミック予測による遺伝資源の探索は、マーカー遺伝子型が利用可能な全ての遺伝資源系統を対象とした効率的なスクリーニング法として注目を集めている。遺伝資源の一部を圃場評価し、予測モデルの更新と選抜を行うことを繰り返すことにより、多数の遺伝資源全てを圃場試験することなく優れた系統を発見できる。本研究では、ゲノミック予測に基づく遺伝資源探索を最適化問題と捉え、ベイズ最適化を適切に応用することにより、探索をさらに加速・効率化するための選抜戦略を提案する。実データを用いて、異なる戦略に基づく遺伝資源探索をシミュレーションし、予測値の大きな系統を選抜するという通常の選抜法に比べて、ベイズ最適化に基づく選抜法が、より効率よく遺伝資源から優れた系統を発見できることを示す。

4-2. 材料・方法

4-2-1. 問題の整理

ここでは、ゲノミック予測を用いた遺伝資源探索を、最適化問題として定式化する。ある系統 i のマーカー遺伝子型ベクトルを $\mathbf{x}_i$、遺伝子型値を u_i とすると

$$u_i = f(\mathbf{x}_i) + \varepsilon_i \tag{4.1}$$

が成り立つ。ここで、$f(\mathbf{x})$は何らかのゲノミック予測モデルである。ε はゲノミック予測モデルで説明できない遺伝子型値であり、例えば、DNA マーカーが真の原因遺伝子と完全連鎖していないこと、エピスタシスや優性効果が存在するにも関わらず GBLUP などの加法的なモデルを用いて

いることなどが原因となって生じる。いま、1つの系統あたり複数個体の表現型値を得られるとすれば、系統 i の j 番目の個体の表現型値 y_{ij} は遺伝子型値 u_i と環境誤差 e_{ij} の和として表されるので、式(4.1)を踏まえれば、

$$y_{ij} = u_i + e_{ij} = f(\mathbf{x}_i) + \varepsilon_i + e_{ij} \tag{4.2}$$

と表せる。いま、遺伝子型値が大きいほど優れた系統であると仮定すれば、遺伝資源探索の（厳密な意味での）目標は

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}} \left[f(\mathbf{x}) + \varepsilon\right] \tag{4.3}$$

で表される、最良の系統 **x***を発見すること、あるいは、$f(\mathbf{x}^*) + \varepsilon^*$ にできるだけ近い値を与える **x** を、決められた制約の中で発見することである（ε*は最良の系統における ε とする）。ここで、$\mathcal{X}$ はマーカー遺伝子型を持つ遺伝資源系統全てからなる集合とする。

ただし、本研究では単純のため、1つの系統に対して1つの表現型値だけを得ることを考える。また、ゲノミック予測モデルが説明できない遺伝子型値についても無視する。これにより、本研究における最適化問題を

$$y_i = f(\mathbf{x}_i) + e_i \tag{4.4}$$

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\ f(\mathbf{x}) \tag{4.5}$$

と書くことができる。なお、前者の仮定は、複数環境で測定された表現型値の割合（例えば、2つの異なる施肥条件での収量の差や比）などを考える場合には必然として生じうる。また、後述する実データでは、表現型値が（おそらくは複数個体や複数ブロックの平均値として）1つの系統・形質につき1つしか与えられていないため、この仮定が必要であった。後者の仮定は、前者の仮定のもとでは ε と e を統計的に分割することができないために必須となる。したがって、式(4.4)の e_i は、モデルで説明できない遺伝子型値と環境効果の両方を含むが、それは最適化の対象とはみなさないことにする。

実際には、ゲノミック予測モデル $f(\mathbf{x})$は訓練データによって変化する。したがって、原理的には、最良の系統 **x***を定義することは困難である。言い換えれば、もし「正しい」ゲノミック予測モデル f^*があるとすれば、最良の系統とは

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\ f^*(\mathbf{x}) \tag{4.6}$$

と表現されるべきものであるが、このようなゲノミック予測モデル f^*は、遺伝資源を全て調べ尽くしてモデル $f(\mathbf{x})$を構築したとしても得られるものではない。しかし、本研究では、現実的な仮定として、全ての遺伝資源（シミュレーションでは、データセットに含まれる全系統）を用いて構築されたモデルが最も確からしいと考え、それを f^*と仮定した。つまり、シミュレーションにおいては、全てのデータを用いて予測モデルを構築した場合の遺伝子型値の推定値を計算し、その最大値を与える系統を最良の系統と定義した。これにより、ある系統が最良の系統であることを

判定することは実際の遺伝資源探索では不可能だが、シミュレーションで手法を評価するにあたっては、最良の遺伝子型値を持つ系統が得られたかどうかを判断できるとした。

いま、全ての遺伝資源の数は有限なので、集合 $\mathcal{X}$ に含まれる全ての系統について総当たり的に表現型を測定すれば、最良の系統 $\mathbf{x}^*$を求めることができる。本研究の目標は、そうした総当たり探索よりも効率よく最良の系統を発見する戦略を構築することである。つまり、全ての系統について表現型を調べることなく最良の系統を発見する、あるいは、ある決まった系統数だけを圃場評価した場合に、評価した系統のうち最も遺伝子型値が大きい系統が、最良の系統の遺伝子型値 $f^*(\mathbf{x}^*)$ に近いことが目標となる。

4-2-2. ベイズ最適化

ベイズ最適化は、black-box 最適化に用いられるアルゴリズムの中でも、少ない数のデータ点で未知の関数の最大値を探索できるという特長をもつ。ここでは、ベイズ最適化のアルゴリズムについて解説し、遺伝資源探索への応用についても最後に議論する。なお、ベイズ最適化の説明においても、随所でゲノミック予測や遺伝資源探索への応用を意図した表現を用いる。ベイズ最適化のアルゴリズムを追うことで、遺伝資源探索との対応関係は非常に明快であることがわかる。

まず、ベイズ最適化の初期状態として、1つ以上のデータ点（入力 $\mathbf{x}$ と出力 y の組）が既知であるものとする。また、出力 y がわからない入力 $\mathbf{x}$（これを未知データと呼ぶことにする）が多数あると仮定する。本研究では $\mathbf{x}$ のとりうる値（探索候補の系統数）が有限個しかない状況を扱うが、通常の応用では $\mathbf{x}$ が連続変数であることが想定されるため、未知データは無限個あると言える。初期データ点をいくつにするか、どのように決めるかは任意である。

ベイズ最適化は、2つのステップを繰り返すことで関数の最大値を効率よく探索する。1つめのステップは未知である目的関数 f を既知のデータ（初期データ、または、次のステップで追加されたデータを加えたデータ集合）からベイズ的に推定することであり、2つめのステップは獲得関数 (acquisition function) と呼ばれる関数をもとに追加すべき未知データを選ぶことである。選ばれた未知データについて、入力に対応する出力を調べることで、既知データに加える。以下、これらのステップについて、順を追って解説する。ベイズ最適化を1次元の入力 x について実行する場合の模式図を図 4-1 に示す。

目的関数をベイズ的に推定するには、ガウス過程回帰（Gaussian process regression）と呼ばれる方法が用いられる。これは RKHS 回帰をベイズ的に表現・計算する場合の呼称であり、したがってゲノミック予測で用いられる GBLUP も、ガウス過程回帰の一種と捉えることができる（Morota and Gianola, 2014）。ガウス過程回帰のモデル（尤度関数）は、式 (4.4)について

$$p(\mathbf{y}|\mu, \sigma_e^2) = N(\mathbf{y}|\mathbf{1}\mu, \mathbf{K} + \mathbf{I}\sigma_e^2) \tag{4.7}$$

と書くことができる。ここで、$\mathbf{y}$ は出力（表現型）のベクトル、μ は全平均（集団平均）、σ_e^2 は残

差 e の分散、**K** は入力変数 **x**（マーカー遺伝子型）から計算される分散共分散行列（関係行列）である。このモデルで、μ、σ_e^2、および行列 **K** に含まれるパラメータをデータから推定すれば、未知の入力 **x** に対する出力 y の予測分布を得ることができる。本研究ではガウス過程の計算に R の {BGLR} パッケージを用いるが、そこでは逆カイ２乗分布を残差 σ_e^2 の事前分布として用い、全平均 μ には非常に大きな分散をもつ正規分布を flat prior として事前分布に用いる（Perez and de los Campos, 2014）。

通常のベイズ最適化では、入力 **x** と出力 y の関係について線形な関係を仮定することはできないため、行列 **K** はガウスカーネルや指数カーネルなどの非線形カーネル、および、それらを混合したような複雑なカーネル関数により計算される。しかし、ゲノミック予測では、マーカー遺伝子型の相加効果が支配的であるという仮定のもと、GBLUP のように線形カーネルに基づく分散共分散行列を用いることがしばしばある。本研究でも、ゲノミック予測の慣例に習い、分散共分散行列を

$$p(\mathbf{u}|\sigma_u^2) = N\left(\mathbf{u}|0, \frac{\mathbf{M}\mathbf{M}^{\mathrm{T}}}{P}\sigma_u^2\right) \tag{4.8}$$

という線形カーネルによって指定した。ここで、**M** は各列を平均 0、分散 1 に基準化したマーカー遺伝子型行列であり、P はマーカー数（すなわち **M** の列数）である。また、σ_u^2 は遺伝分散に対応するパラメータであり、σ_e^2 と同様、逆カイ２乗分布を事前分布として採用する。式(4.8)による分散共分散の指定は、行列の対角成分を全て 1 にするという意味を持つ。SNP マーカーを用いる場合や、古典的な血縁による BLUP との対応をするためには GBLUP による定式化が優れている可能性があるが、本実験では DArT マーカーにより遺伝子型が記述されているデータを使用すること、また、モデル比較を行うわけではないことから、全てのデータについて式(4.8)を用いた。この方法は、本実験で解析に用いる R パッケージ {BGLR} のサンプルコードでも採用されている（Perez and de los Campos, 2014）。なお、線形カーネルの基準化に関する議論は（VanRaden, 2008）を参照されたい。

ここで、分散パラメータ σ_e^2 および σ_u^2 のベイズ推論が十分確からしいことを仮定する、あるいは、分散パラメータを点推定し、それを定数として扱うことを仮定すると、未知の入力 **x** についての事後予測分布は正規分布となり、以下のように記述できる。

$$p(y|\mathbf{x}) = N\big(y|m(\mathbf{x}), s^2(\mathbf{x})\big) \tag{4.9}$$

ベイズ最適化の次なるステップは、こうした事後予測分布から獲得関数を計算することである。獲得関数は、未知データ点の「良さ」を表現した関数だといえる。古くから提案されている有名な獲得関数は期待改善量（expected improvement）と呼ばれるものであり、既知データの出力推定値の最大値を $\hat{y}_M$ とすると

$$\mathrm{EI}(\mathbf{x}) = \int_{\hat{y}_M}^{\infty} (y - \hat{y}_M) \cdot p(y|\mathbf{x})\, dy \tag{4.10}$$

と定義される。この式は事後分布が正規分布でなくても成り立つ、期待改善量の定義式である。期待改善量は、既知のデータから計算された予測分布 $p(y|\mathbf{x})$ のもとで、ある未知データ $\mathbf{x}$ に対応する出力 y が、既知データの最大推定出力値 $\hat{y}_M$ に比べてどの程度大きくなるかを計算した値である。つまり、既知データから期待される、ある未知データを調べた場合の、最大値の改善量を意味している。ベイズ最適化では、期待改善量の大きい点を次のデータとして選ぶ。

なお、式(4.10)右辺の積分は、式(4.9) のように事後予測分布が正規分布である場合には、標準正規分布の確率密度関数 $\phi(z)$ と累積密度関数 $\Phi(z)$ を用いて

$$\mathrm{EI}(\mathbf{x}) = (m - \hat{y}_M) \cdot \Phi(z) + s \cdot \phi(z) \tag{4.11}$$

このように簡単な形式で表現できる。ただし、

$$z = \frac{m - \hat{y}_M}{s} \tag{4.12}$$

である。

期待改善量に基づく未知データの選択について、その意味を理解することは重要である。式(4.11), (4.12)で表される、事後予測分布が正規分布である場合の期待改善量について、値の変化を等高線図によって図 4-2 に示した。図より、事後平均 m が既知データの最大推定出力値 $\hat{y}_M$ よりも大きい場合には、期待改善量は事後分散 s^2 にあまり影響されず、事後平均が大きいほど大きくなることがわかる。逆に、事後平均 m が既知データの最大推定出力値 $\hat{y}_M$ よりも小さい場合には、期待改善量は事後分散に強い影響を受け、分散が大きいほど大きくなる。この性質は直感的にも妥当である。今までに調べたデータの出力を上回る出力を得られることが予想されれば、単にそのデータを選べば良い。いっぽう、今までに調べたデータよりも劣ることが予想される場合には、予測の不確かさ（予測分布の分散）が大きいデータを選ぶことで、既存のデータを上回る確率を高めることができる。

Black-box 最適化の文脈では、探索と活用のジレンマと呼ばれる現象がよく知られている。これは、既存のデータから示唆される「良い」入力を調べる（知識の活用）だけでは、ある局所最適解に陥ってしまう可能性が高いため、今までに調べられていない入力領域を調べる（探索）ことも重要だということを意味する。探索が強すぎると最大値に近づくのが遅くなり、アルゴリズムの収束効率が悪化する。逆に、探索が弱すぎると局所最適解に陥る可能性が高くなるというジレンマが生じる。期待改善量は、単に予測値の大きい未知データを選ぶのではなく、予測の不確かさを考慮して、不確かさが大きい未知データを選びやすくすることにより、活用一辺倒ではなく、未知の入力領域を探索することもできる指標となっている。これは、通常のゲノミック選抜で、予測値の大きな系統から順に選抜を行うのとは対照的である。

以上のように、ベイズ最適化では、ガウス過程回帰による予測と、獲得関数に基づく未知データの選択を繰り返すことで大域的最適化を実現する。このステップは、実世界におけるゲノミック予測を用いた遺伝資源探索と容易に対応づけられる。まず、ガウス過程回帰による予測は、ゲ

ノミック予測そのものである。このとき、ベイズ最適化では非線形カーネルが多く用いられるが、ゲノミック予測で用いられる線形カーネルであってもアルゴリズムの動作には影響がない（性能には影響すると思われるが、それは本研究では議論しないことにする）。次のステップである獲得関数に基づくデータの追加は、ゲノミック予測に基づく未試験系統の選択と考えられる。通常のゲノミック予測では、予測値が大きい系統から順に未試験の系統を選ぶが、ベイズ最適化では、獲得関数を計算し、その値の大きい系統から順に選ぶ。未知データの出力を調べて既知データに追加することは、遺伝資源探索では圃場評価に対応する。したがって、ベイズ最適化を遺伝資源探索に応用する方法は非常に単純であり、選抜法として、単に予測値の大きな系統を選ぶのではなく、未試験の系統について獲得関数（ここでは、期待改善量）の値を計算し、その値の大きな系統を選ぶだけでよい。

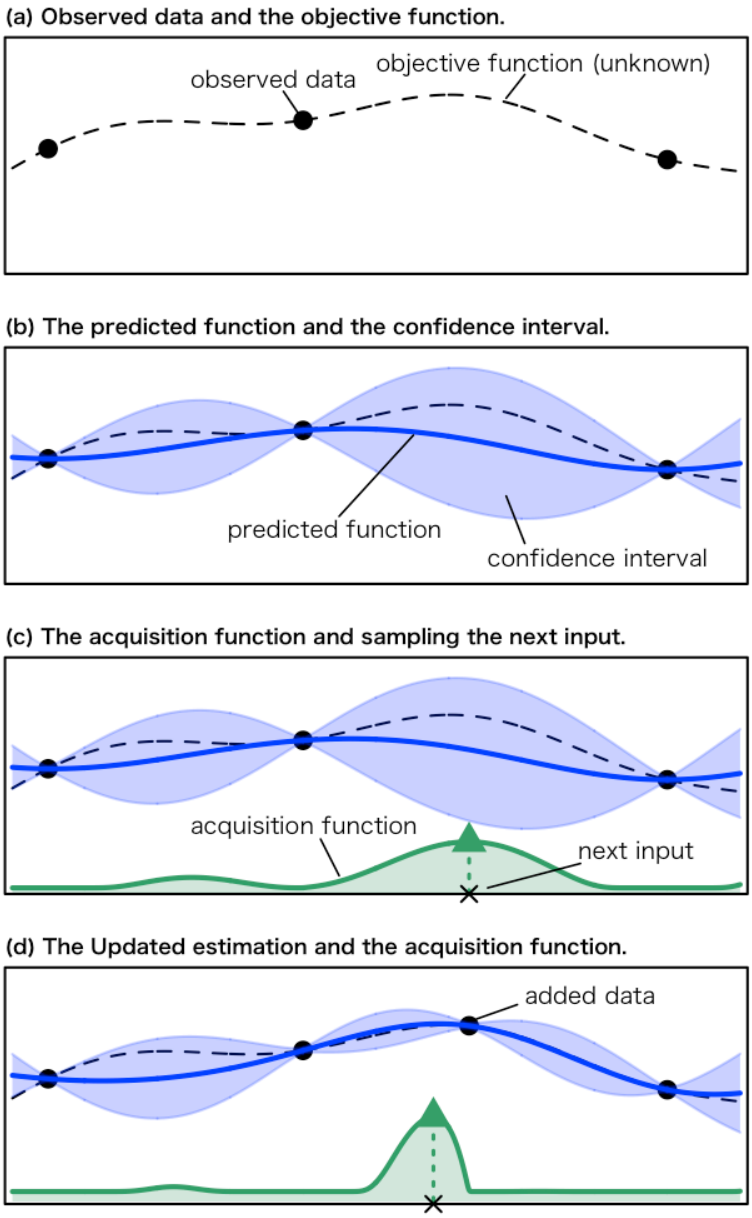


図 4-1. ベイズ最適化の模式図

1 次元の入力変数について、非線形カーネルによるガウス過程回帰を用いた場合のベイズ最適化の模式図。なお、この図ではノイズがないことを仮定した。(a) 初期状態では、いくつかの点について入力と出力が分かっている。真の関数（破線）は未知である。(b) ガウス過程による予測を行うことで、未知の入力に対する平均と標準偏差が計算できる。 (c) その予測に基づき、獲得関数（緑線）が計算される。獲得関数の最大値を与える点（緑色三角印）が次の入力となる。(d) 選ばれた入力に対応する出力を評価して既知データに追加し、(a)-(c)を実行した後の図。このように、予測の更新、獲得関数の再計算、未知データの追加を繰り返すことで最適化が進行する。

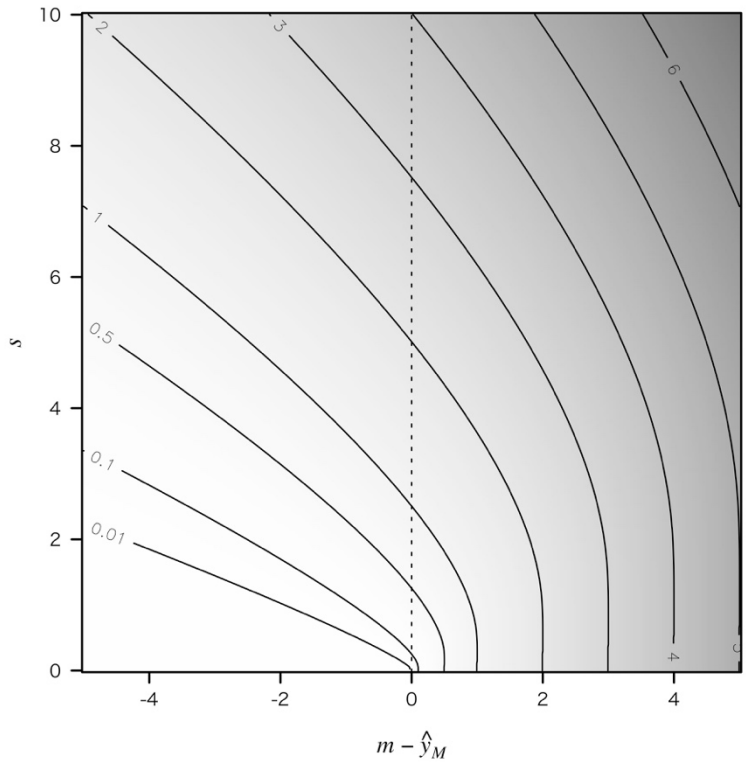


図 4-2. 期待改善量の等高線図

事後予測分布の平均 m と標準偏差 s の変化による期待改善量の変化を投稿線図で示した。標準偏差による期待改善量の増加は、図の左側と右側で異なる傾向を示す。なお、等高線の目盛りが等間隔ではないことに注意せよ。

4-2-3. シミュレーションの設定

ベイズ最適化に基づく遺伝資源探索法の有効性を、実データに基づくシミュレーションによって検証した。シミュレーションの仮定を以下に示す。

- 全ての形質で、表現型値の大きい系統が望ましいと仮定した。
- 全ての系統についてマーカー遺伝子型が予めわかっていると仮定した。
- 全データを用いて推定された遺伝子型値を“真の”遺伝子型値と仮定し、シミュレーションの評価はこの“真の”遺伝子型値によって行なった。
- はじめは、初期訓練データとして無作為に選ばれた系統についてのみ、表現型値がわかっているとした。初期訓練データの系統数（初期系統数）は 10 系統、50 系統、90 系統の 3 通りを試した。
- 表現型未知の系統を、逐次圃場試験を行って評価し、訓練データに加えることができるとした。ここで、一度に圃場試験できる系統数には限りがあると仮定した。1 回あたりに選ばれて試験される系統数（選抜系統数）は 10 系統、20 系統、40 系統の 3 通りを試した。
- 新たに未試験の系統を選抜する方法として、無作為選抜（RS; random selection)、予測値最大の系統の選抜（MP; maximum predicted values)、期待改善量による選抜（EI; expected improvement）の 3 通りを試した。以下、それぞれ RS 戦略、MP 戦略、EI 戦略と略記する。
- MP 戦略と EI 戦略で用いるゲノミック予測モデルは、式(4.7), (4.8)で定まる、線形カーネルに基づくガウス過程回帰を用いた。計算には R の {BGLR} パッケージを用いた（Perez and de los Campos, 2014)。
- 期待改善量の計算は、式(4.11)および(4.12)によって行った。実際には分散パラメータも未知変数でありデータから推定されているため、厳密には事後予測分布を正規分布とみなすことはできないが、ここではその影響を無視し、MCMC の結果をもとに計算される表現型の事後予測分布の平均および標準偏差の期待値を用いて計算を実行した。
- 以上のように、初期訓練データから初めて、予測モデルの更新と新たなデータの追加を繰り返すというステップを、全ての系統を圃場試験するまで繰り返した。
- シミュレーションには、1 つの形質・設定について、初期集団が異なる 100 回の反復を設けた。
- 表現型の取得において、年次の主効果や、年次と遺伝子型の交互作用などはないものとした。つまり、ある系統は、何回目に追加される場合でも、解析に使用した実データの表現型値そのものが得られることとした。

このシミュレーションでは、全系統の一部が逐次的に圃場評価されていく。以下では、ある時点までに選ばれた系統を選抜済み系統、選ばれていない系統を未試験系統と呼ぶことにする。本研究では、各選抜法について、選抜済み系統の“真の”遺伝子型値の最大値および平均値を計算して評価に用いた。遺伝資源探索法としては、選抜済み系統の遺伝子型値の最大値が高いことが最も望ましく、平均的に良い系統を選ぶことは必ずしも重要ではないことに注意されたい。

また、選抜済み系統を用いて構築したゲノミック予測モデルを用いて、未試験系統の予測を行なった場合の予測精度（全データから推定した“真の”遺伝子型値と、選抜済み系統から予測された遺伝子型値との間の相関係数）を計算した。この精度評価はモデル比較などを行う場合の精度評価とは意味合いが異なり、異なる選抜法によって遺伝資源を探索した場合に、探索を進めるうち、残っている遺伝資源系統の予測精度がどのように変動するかを検証する目的で行っていることに注意されたい。すなわち、あらかじめ適当な数の系統をテスト集合として抽出しておき、常にその集団を用いて評価することはせず、あえて、選抜法によって異なるテスト集合を用いている。

4-2-4. 使用したデータセット

４つの実データセットを解析に使用した。２つのコムギデータはともに elite line のデータセットであるが、ある集団の一部を逐次評価して優れた系統を探索すると言う意味では、シミュレーションに用いることができると考えられる。なお、本章で用いるデータセットのうち３つ（CIMMYT コムギデータ; CIMMYT_wheat dataset, Perez コムギデータ; Perez's_wheat dataset, トウモロコシデータ; Maize dataset）は３章と同一であるため、ここでは説明を省く。異なる１つはイネ遺伝資源データであり、表現型は３章で用いたデータと同じソースを用いているが、マーカー遺伝子型を（当時の）最新版データに置き換えている。

・　イネ遺伝資源データ（Rice dataset）

RiceDiversity（http://ricediversity.org/index.cfm）が公開しているイネ遺伝資源の遺伝子型・表現型データセットのうち、全 412 系統、約 44,000 SNP マーカーの遺伝子型データ（Zhao et al., 2011）を用いた。マーカー遺伝子型に含まれる欠測値は、当該 SNP のアリル頻度をもとに遺伝子型の平均値を計算して、その値で補完した。表現型データでは 34 形質が記録されていたが、本研究では Aberdeen における開花期（Flowering time at Aberdeen）、Arkansas における開花期（Flowering time at Arkansas）、Faridpur での開花期（Flowering time at Faridpur）、穂数（Panicle number per plant）、草丈（Plant height）、一穂穎果数（Florets per panicle）、種子長（Seed length）、種子面積（Seed surface area）、種子幅（Seed width）、の 9 形質を解析に用いた。なお、マーカー遺伝子型と表現型の両方が揃っている系統数は 400 系統弱（形質ごとに異なる）であり、これが解析に用いられた。

4-3. 結果

・　選抜済み系統の遺伝子型値の最大値

全ての系統のうち最も遺伝子型値の大きい系統をできるだけ少ない繰り返しで発見する、あるいは、決まった繰り返し回数で、できるだけ遺伝子型値の大きい系統を発見することができれば、それは優れた遺伝資源の探索法だと考えられる。初期系統数が 10 系統、選抜系統数が 10 系統の場合（この設定は現実的ではないが、最も細かい繰り返しを行うため、手法間の違いが最も明瞭に現れることが期待される）について、選抜済み系統の遺伝子型値の最大値の推移を図示した（図 4-3)。なお、図示するにあたって、形質ごとに遺伝子型値を 0 から 1 にスケーリングし、形質ごとの結果を細線で、全形質についての結果を太線で図示した。

図から、特にイネ遺伝資源データで、EI 戦略（赤線、三角印）は、RS 戦略（黒線、四角印）や MP 戦略（青線、丸印）よりも効率的に、選抜済み系統の遺伝子型値の最大値を向上させていることが読み取れる。すなわち、期待改善量によって未試験系統を選ぶことで、同じ選抜回数でも、より優れた系統を発見できることが示された。

なお、異なるシミュレーション設定（初期系統数と選抜系統数）で得られた結果は、章末にて同様に図示した（図 4-S1 から図 4-S8)。これらの図を見ると、初期系統数や選抜系統数が変化しても、EI 戦略が RS 戦略や MP 戦略を上回るパフォーマンスを示すことがわかる。

イネ遺伝資源データでは EI 戦略と MP 戦略に明瞭な違いが見て取れるいっぽうで、トウモロコシデータや CIMMYT コムギデータでは、両者にあまり違いがないように思われる。そこで、初期系統数 10 系統、選抜系統数 10 系統の場合のシミュレーションについて、最良の系統を選ぶまでに選抜された系統数を表 4-1 にまとめ、EI 戦略と MP 戦略で要した系統数に有意な違いがあるかを *t* 検定した。

表 4-1 より、トウモロコシデータの２形質とコムギデータの一部を除く全てのデータセット・形質で、EI 戦略を行うことにより、最良の系統を発見するまでに必要な系統数が、MP 戦略に比べて有意に減少することがわかった。また、全ての形質についての減少率の平均は 30%ほどであった。なお、一部の形質では EI 戦略によってむしろ必要な系統数が増加していたが、イネ遺伝資源データの Seed width を除いて、その差は 100 回のシミュレーションでは有意ではなかった。

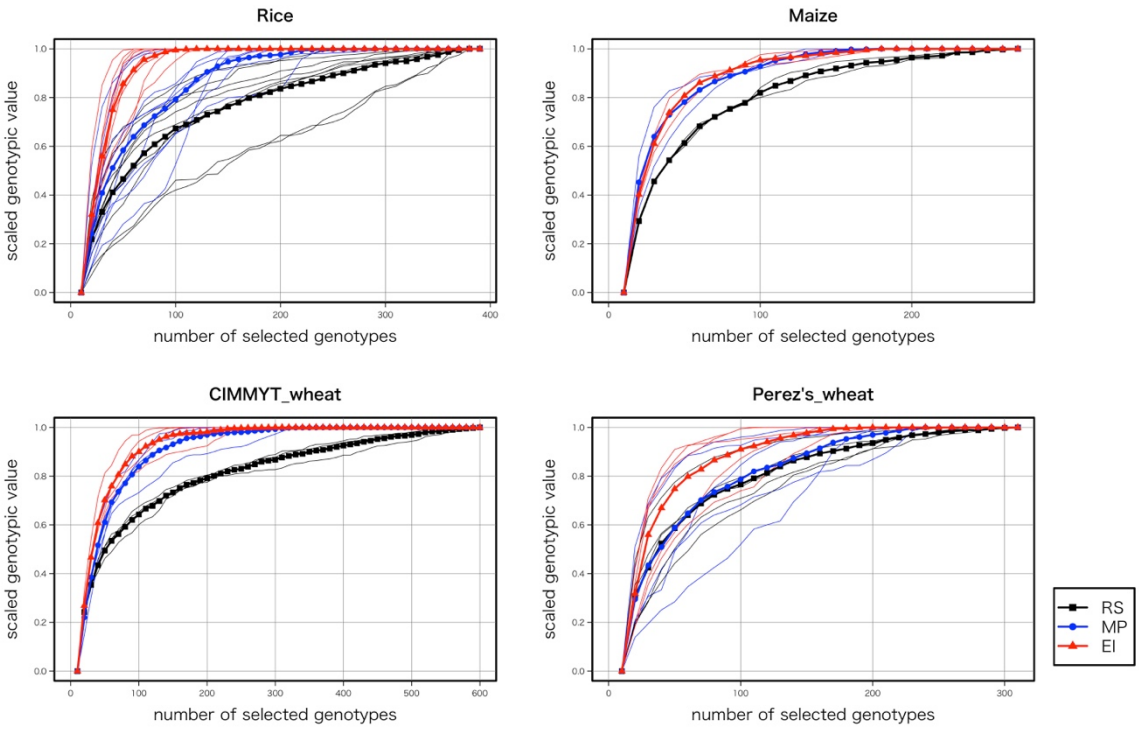


図 4-3. 選抜済み系統の遺伝子型値の最大値

初期系統数 10 系統、選抜系統数 10 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。ここで、1 つの形質に関する結果は細線で、全ての形質に関する平均を太線で示した。また、遺伝子型値は各形質について最小値 0、最大値 1 にスケーリングすることで、同一の目盛りで表現した。EI 戦略（赤線、三角印）が、他の選抜法に比べて、選抜済み系統の遺伝子型値の最大値を高くできることがわかる。

表 4-1. 最良の系統を発見するまでに選抜した系統数

初期系統数 10 系統、選抜系統数 10 系統のシミュレーションについて、最良の系統を発見するまでに選ばれた系統数の平均（SVG）と標準偏差（SD）を示した。また、MP 戦略と EI 戦略について、最良の系統を発見するまでに要した系統数に差があるかを両側 *t* 検定によって検定し、5%有意な形質を * で、1%有意な形質を ** で、0.1%有意な形質を *** で示した（多重比較に関する補正なし）。

Dataset	Trait	RS		MP		EI		
		AVG	SD	AVG	SD	AVG	SD	significance
RiceDiversity	Florets.per.panice	177.5	112.4	86.4	57.4	34.5	12.1	***
	Flowering.time.at.Aberdeen	187.3	100.6	133.5	59.0	70.4	15.7	***
	Flowering.time.at.Arkansas	216.4	111.5	65.3	38.7	37.8	12.0	***
	Flowering.time.at.Faridpur	135.8	83.7	141.3	57.4	64.6	18.9	***
	Panicle.number.per.plant	180.6	107.3	40.5	16.6	30.9	10.6	***
	Plant.height	196.2	104.6	107.8	29.5	86.3	25.5	***
	Seed.length	198.7	117.3	100.5	38.5	36.3	14.0	***
	Seed.width	203.8	108.3	63.4	16.7	72.9	21.3	***
	Seed.surface.area	176.1	107.7	126.8	32.9	35.0	11.1	***
CIMMYT_wheat	yield_E1	295.7	168.8	154.0	99.4	133.7	79.6	*
	yield_E2	299.3	173.8	122.6	51.0	104.4	35.5	***
	yield_E3	300.9	171.3	88.3	42.7	64.6	38.8	***
	yield_E4	331.2	173.6	84.2	48.3	57.8	31.6	***
Perez's_wheat	yield_D_bed	166.0	87.5	59.3	34.7	66.1	24.8	
	yield_D_flat	130.8	75.8	158.2	76.8	87.7	42.7	***
	yield_H_bed	162.0	88.8	121.0	53.4	82.0	51.5	***
	yield_I_bed	170.0	95.1	73.1	46.3	38.3	23.4	***
	yield_I_flat	134.9	75.1	49.2	34.3	41.0	19.6	**
Maize	yield_SS	129.5	83.5	57.5	33.9	57.3	31.9	
	yield_WW	137.3	72.9	112.1	46.2	115.0	51.5	

・　選抜済み系統の遺伝子型値の平均値

最大値の場合と同様に、選抜済み系統の遺伝子型値の平均値を、初期系統数 10 系統、選抜系統数 10 系統の場合について図示した（図 4-4)。遺伝子型値の平均値は、最大値とは異なり、EI 戦略よりも MP 戦略で高くなることがわかった。よって、期待改善量による選抜は「1 つの優れた系統」を探索する場合には優れた戦略であるいっぽうで、「多数の良い系統」を集める場合には、通常のゲノミック選抜で行われるように、予測値の大きな系統を選抜したほうが良いと考えられた。

平均値の推移についても、異なるシミュレーションの設定での結果を図 4-S9 から図 4-S16 に示した。やはり最大値に関する結果と同様に、平均値についても、初期系統数や選抜系統数による大きな傾向の違いは見られなかった。

・　選抜法ごとの予測精度

やはり同様に、初期系統数 10 系統、選抜系統数 10 系統の場合について、シミュレーション中の予測精度（試験済み系統から未試験系統を予測したときの予測精度）の変化を図示した（図 4-5)。予測精度は、おしなべて、RS 戦略の場合に最も高かった。また、EI 戦略の予測精度は、RS 戦略より低いながらも徐々に増加する傾向が見られた。いっぽう、MP 戦略では、予測精度はモデル構築に利用できる系統数が増加しているにも関わらずほぼ横ばいであった。

予測精度の推移についても、異なるシミュレーションの設定について同様の図を作成した（図 4-S17 から図 4-S24）が、やはり図 4-5 と同様の傾向が確認された。

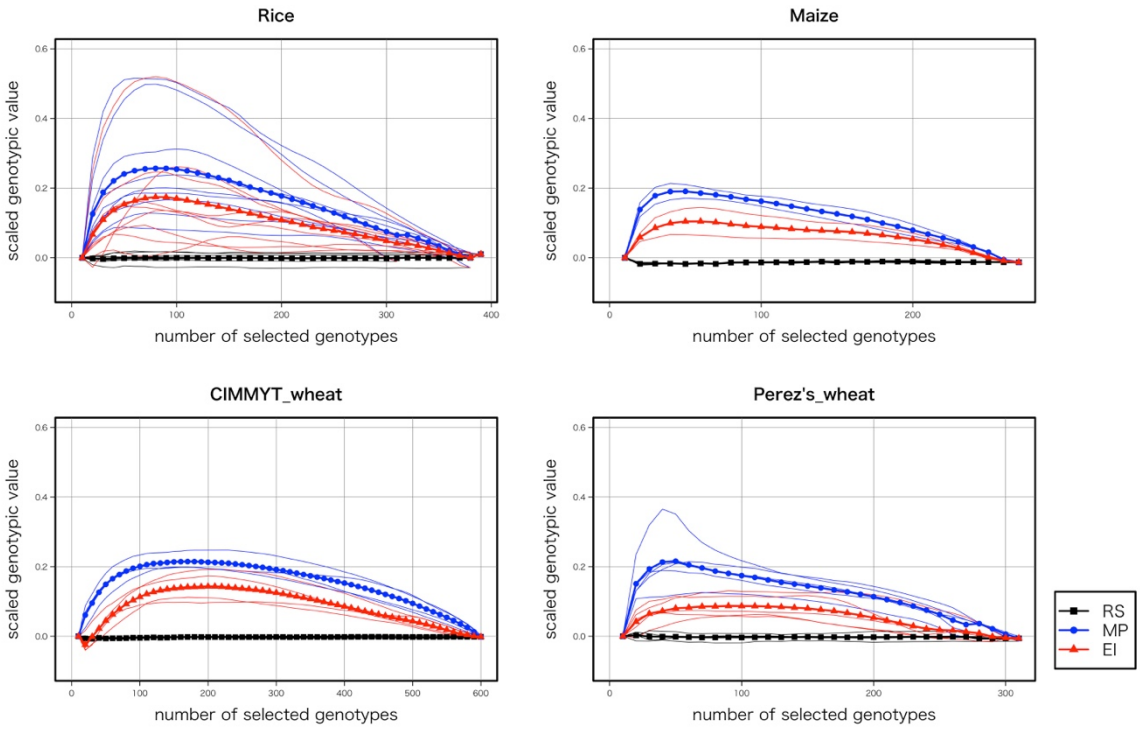


図 4-4. 選抜済み系統の遺伝子型値の平均値

初期系統数 10 系統、選抜系統数 10 系統のシミュレーションについて、異なる選抜法で選ばれた系統の遺伝子型値の平均値を示した。ここで、1 つの形質に関する結果は細線で、全ての形質に関する平均を太線で示した。また、遺伝子型値は各形質について最小値 0、最大値 1 にスケーリングすることで、同一の目盛りで表現した。MP 戦略（青線、丸印）によって、選抜済み系統の遺伝子型値の平均値を高くできることがわかる。

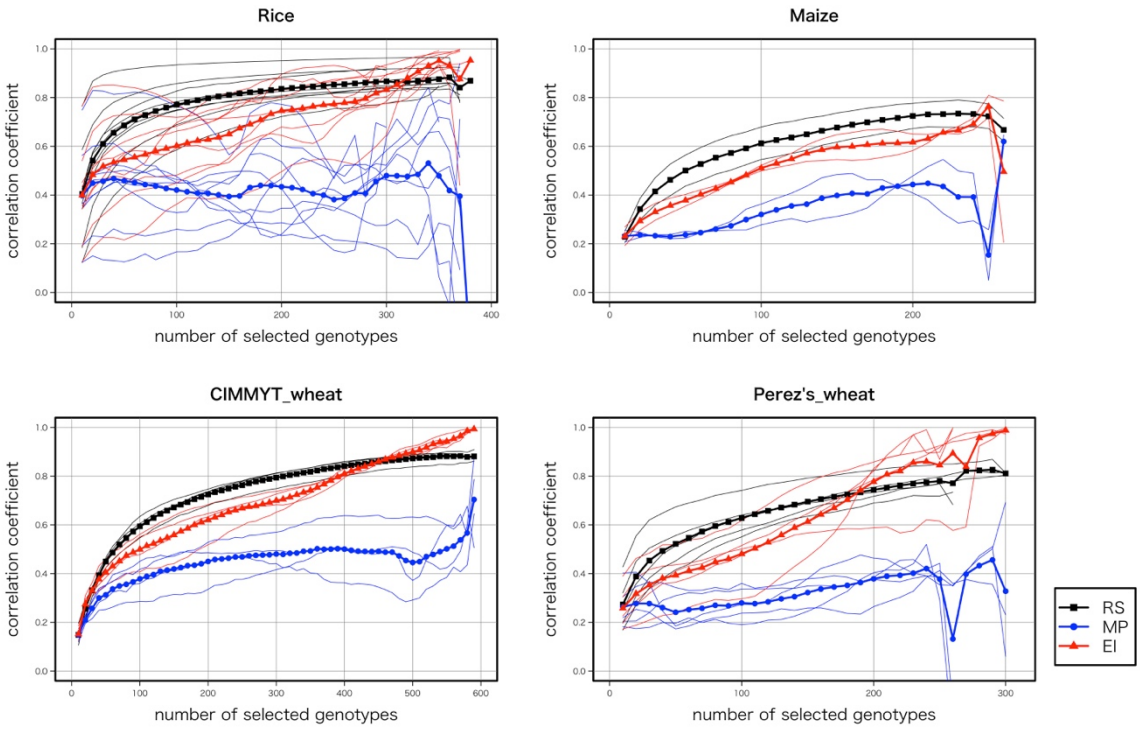


図 4-5. シミュレーションにおける予測精度の推移

初期系統数 10 系統、選抜系統数 10 系統のシミュレーションについて、選抜済みの系統を用いて構築された予測モデルで未選抜系統の遺伝子型を予測した場合の予測精度の変化を図示した。ここで、 1 つの形質に関する結果は細線で、全ての形質に関する平均を太線で示した。予測精度は RS 戦略で最も高く、MP 戦略で最も低いことがわかる。

4-4. 考察

本研究では、ゲノミック予測を用いた遺伝資源探索法を効率化するために、ベイズ最適化を応用する方法を検討し、期待改善量に基づき未試験系統を選抜すること（EI 戦略）を、単純に予測値の大きな系統を選ぶゲノミック選抜（MP 戦略）に代わる、新たな選抜戦略として提案した。シミュレーションの結果、選抜済み系統の遺伝子型値の最大値は EI 戦略において MP 戦略よりも大きくなることが示され、EI 戦略を用いることで、遺伝資源から優れた系統を効率よく探索できることが強く示唆された。

提案された EI 戦略と MP 戦略との最大の違いは、予測の不確かさを活用することにある。予測が不確かであることは、それだけ既存の最大値を上回る可能性が高いことを意味している。当然ながら、不確かさの大きい系統を実際に圃場試験してみると、遺伝子型値は予測値よりもずいぶん悪い可能性もある。それでも、1 つの優れた系統を得るためには、そのような「可能性に賭ける」戦略が優れていると、EI 戦略の意味を直感的に理解することができる。なお、Yu らは、ゲノミック予測における遺伝資源探索の可能性を指摘するとともに、予測の信頼性の上界（upper bound of prediction reliability）が大きい遺伝資源系統が、そうでない系統よりもより高い表現型を示したことを報告している（Yu et al., 2016）。この信頼性の上界は、おそらく予測の不確かさが大きいほど大きいと考えられ、したがって、EI 戦略とも関係が深いと思われる。ただし、Yu らは上界を用いて選抜する指標を具体的に示したわけではない。期待改善量は、式(4.10)から式(4.12)に見られるように、予測された遺伝子型値の平均と分散（標準偏差）のバランスをとるような指標となっている。したがって、本研究は、Yu らの指摘に対して、ベイズ最適化という数理科学的手法を用いて、具体的な選抜の方法論を与えたものと評価することもできる。

選抜済み系統の遺伝子型値の平均に関する結果を踏まえると、EI 戦略は 1 つの優れた遺伝資源系統を探索する場合には優れた戦略となるが、多数の良い系統を遺伝資源集団から集める場合には良い戦略とはいえず、MP 戦略を用いるほうが良いことがわかる。遺伝資源を活用する場合、集団平均を高めたい場面はほとんどないであろう。しかし、例えば多数の系統を含む育種集団があり、その一部を未試験の環境で試験することを繰り返し、対象の育種集団の中から良い系統をできるだけ多く集め、その後に集団選抜法などを用いて集団の育種価を向上させたい、という状況は（これも、あまりありそうもないが）起こりうる。そのような場合には、EI 戦略ではなく MP 戦略を用いるほうが良いであろう。また、1 つではなく「いくつかの」優れた系統を発見したいという場合には、シミュレーションの結果から推察するに、EI 戦略と MP 戦略の中間的な戦略をとることも考慮するに値する。

EI 戦略と MP 戦略の違いについて興味深いのは、予測精度の変化が大きく異なることである。これは、EI 戦略が探索と活用のジレンマをある程度調節していることから生じたと考えられる。EI 戦略は、未試験系統のうち、予測分布の分散が大きい系統を選びやすい戦略である。ここで、ガウス過程回帰において、予測分布の分散が大きい系統とは、自身と類似の遺伝子型を持つ系統が訓練データに含まれない系統である。ゆえに、EI 戦略は、訓練データに含まれる遺伝子型の多様性を高める傾向をもつと予想され、それが予測精度の向上につながった可能性がある（Würschum et al., 2013）。いっぽう、MP 戦略にはこのような特性は全くない。また、MP 戦

略によって選抜済み系統の遺伝子型値の平均が高くなっていることも、予測精度を低下させている理由である可能性がある。なぜならば、選抜済み系統の遺伝子型値の平均が高いということは、それだけ未試験系統の遺伝子型値は低いものが多いからである。つまり、MP 戦略をとると、遺伝子型値の高い系統を中心とする訓練集団を用いて、遺伝子型値の低い系統を中心とする集団を予測することになる。理論的な考察は難しいが、直感的には、MP 戦略によって生じるこの表現型のバイアスが、予測モデルの構築に不利をもたらしたと考えられる。予測精度の低下は選抜精度の低下につながるため、MP 戦略によって優れた系統を探索することを難しくしたと予想される。

EI 戦略と MP 戦略に関する上述の結果は、初期系統数や選抜系統数を変えても同じであった。初期系統数について、もし初めから多くの系統を用いて予測モデルを構築できれば、モデルの予測精度は向上すると期待される（Heffner et al., 2011; Asoro et al., 2011）。よって、優れた系統を正しく予測できる可能性が高まり、予測の不確実性を考慮しない MP 戦略に有利な設定となると推測していたが、実際にはそうではなく、EI 戦略によって不確かさの大きい系統を選ぶことには、常にメリットがあることが示唆された。

EI 戦略の MP 戦略に対する優位性は、データセット間や形質間で大きく異なっていた。しかし、本シミュレーションで用いたデータセット・形質の範囲では、EI 戦略が MP 戦略に比べて（1 つの優れた系統を探索するという意味で）有意に劣った結果を招くことはほとんどなかった。このように、EI 戦略はシミュレーションの設定やデータセットに対して非常に頑健に機能した。これは、EI 戦略が、その時々の予測モデルに応じて、探索と活用のジレンマに柔軟に対処できることが理由だと推測される。期待改善量は、予測モデルの不確実性がおしなべて低ければ、単に予測値の高い系統を選ぶ。逆に、予測の不確実性が高い系統がまだ試験されていなければ、それを選んで試験する。ベイズ最適化を用いて遺伝資源を探索する場合には、このような、ある種の自動的な判断が働いていると考えられる。

ところで、期待改善量は予測分布の分散の大きい系統を選びやすい戦略であることを指摘したが、このような「分散が大きいものを好む」戦略は、最適な交配組み合わせを選ぶ戦略と類似した発想である。例えば usefulness（Lehermeier et al., 2017）や risk（Akdemir and Sánchez, 2016）などと表現される指標で交配組み合わせを選ぶことが提案されているが、これらは遺伝子型の分離に由来する表現型の分離を考慮し、分離の大きい交配組み合わせを積極的に選ぼうとする戦略になっている。本研究とこれらの研究に直接の関係性はないが、いずれも、ゲノミック予測を用いた育種において、単に系統レベルの予測値だけを用いるのではなく、遺伝的な分離や予測の不確実性といった「ばらつき」を考慮することの重要性を示唆している。

本研究では、ベイズ最適化のうち期待改善量に基づく戦略を採用し、また、複数の系統を同時に選ぶ際には、単に期待改善量の大きい系統を順に選抜した。この方法には、大いに改善の余地がある。例えば、獲得関数としては、予測モデルの上側信頼区間（Auer et al, 2002; Srinivas et al., 2010）や Thompson Sampling（Chapelle and Li, 2011）などが有名である。また、複数のデータを同時に追加する場合には、単に獲得関数の値の大きい順に選ぶよりも良い方法があることが知られている（e.g. Hernández-Lobato et al., 2017）。

また、本研究の結果だけでは、ベイズ最適化による遺伝資源探索が実際に優れた戦略となるこ

とは保証されない。最も懸念されるのは、シミュレーションのスケールが、実際の遺伝資源探索とは大きく異なることである。本研究では、数百系統の集団を遺伝資源と見立て、数十系統（データセットによって異なるが、1%から 20%ほど）を圃場試験するというシミュレーションを行った。実際の遺伝資源探索では、数千〜数万系統の遺伝資源から、数十〜数百系統を圃場試験するという状況を考える必要がある。現状では、このような大規模な表現型データは入手することができないため、実データでシミュレーションを行う場合には、本研究のように設定する必要があった。しかし、例えばシミュレーションによって仮想的にデータを生成することで、スケールの意味ではより現実的な状況に近い解析が可能である。もっとも、遺伝資源の活用は、例えば果樹などでも必要であり、その場合には、一度に評価できる樹木の本数はあまり多くない。このように、対象とする植物種や育種計画によっては、本シミュレーションと類似のスケールで遺伝資源の探索を行うこともあるだろう。

また、難しい問題の 1 つは、EI 戦略による遺伝資源の探索をどのくらい続けるか、ということである。すなわち、探索の stopping rule をどのようにするべきか考える必要がある。本シミュレーションでは、おおよそ全系統の 3 分の 1 ほどを試験した段階で、最良の系統が発見されているようである。しかし、今回のシミュレーション結果だけから経験的に stopping rule を定めるのは難しく、今後、シミュレーションによる経験評価、あるいは、可能であれば何らかの理論的考察により、妥当な stopping rule を決める必要があるだろう。

そのほか、複数の目的形質がある場合にどのような方法を用いるべきか、年次効果や遺伝子型と年次の交互作用がある場合にどのような影響があるか、など、本研究で無視した因子については、これから丁寧に検討される必要がある。そのためには、ベイズ最適化の専門家と育種の専門家が協力して議論を進めていくことも重要であろう。

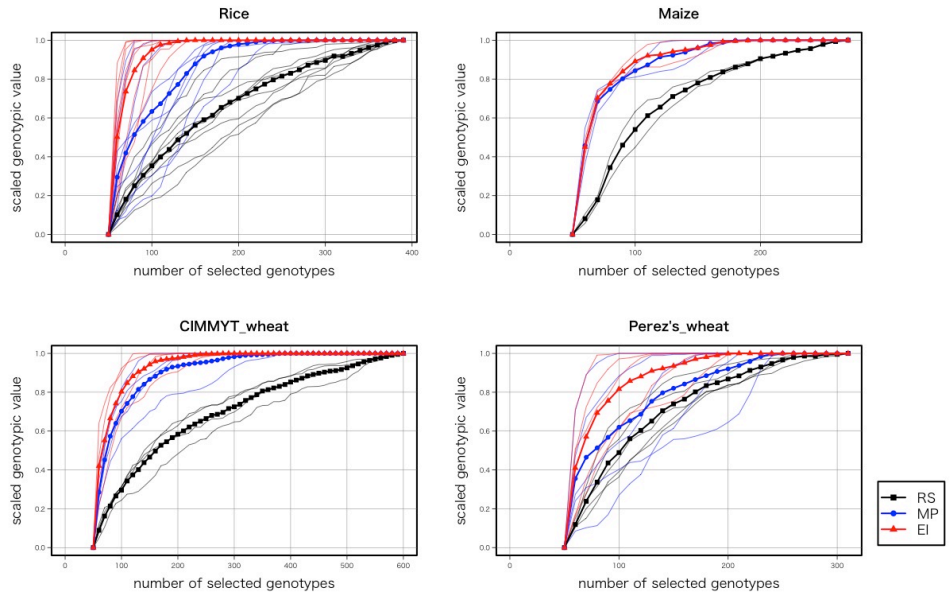


図 4-S1. 選抜済み系統の遺伝子型値の最大値（初期系統数 50 系統、選抜系統数 10 系統）

初期系統数 50 系統、選抜系統数 10 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

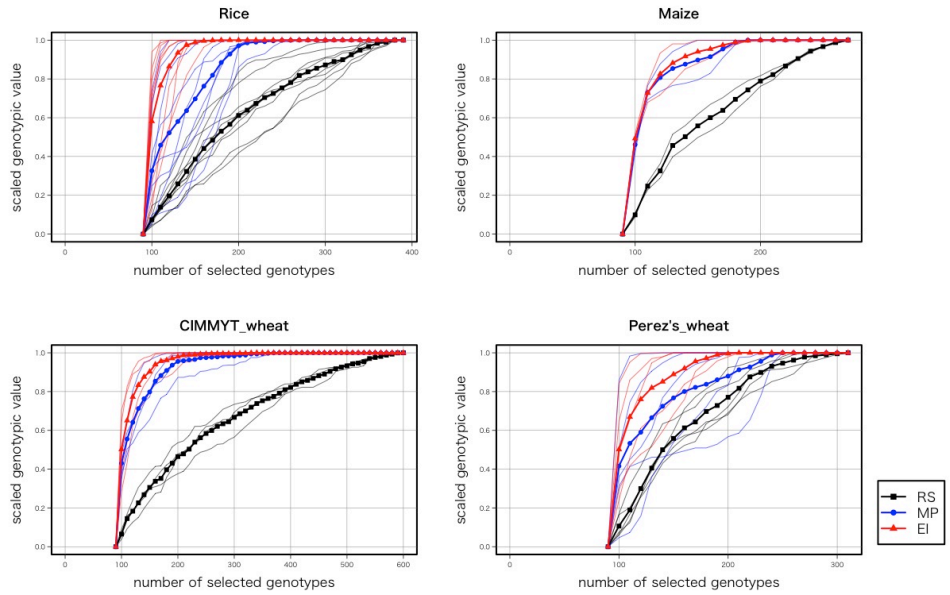


図 4-S2. 選抜済み系統の遺伝子型値の最大値（初期系統数 90 系統、選抜系統数 10 系統）

初期系統数 90 系統、選抜系統数 10 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

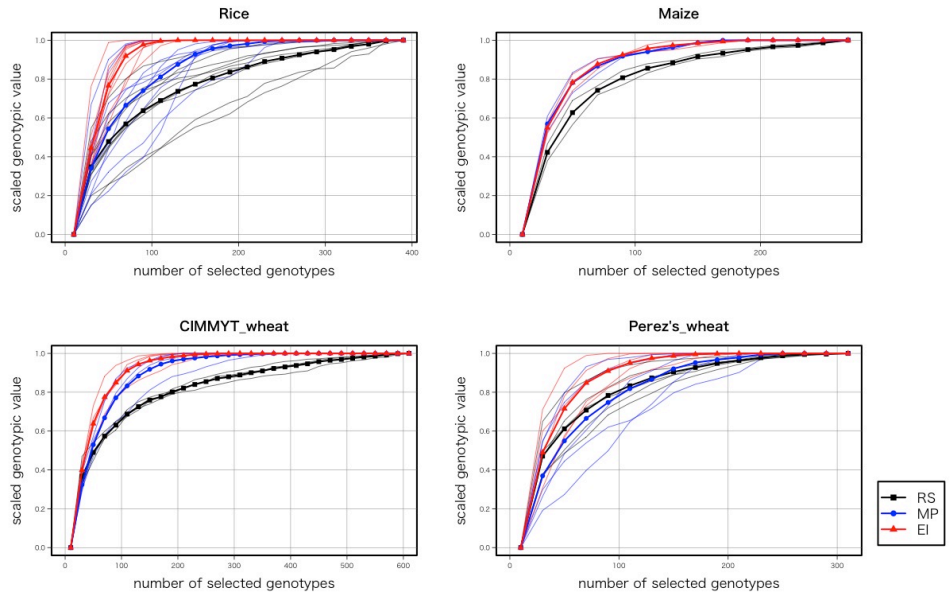


図 4-S3. 選抜済み系統の遺伝子型値の最大値（初期系統数 10 系統、選抜系統数 20 系統）

初期系統数 10 系統、選抜系統数 20 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

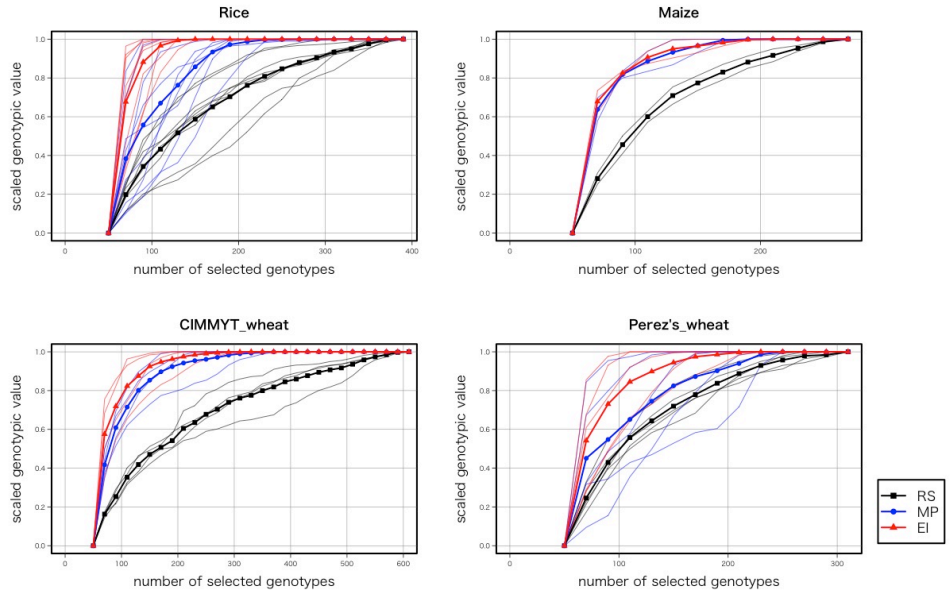


図 4-S4. 選抜済み系統の遺伝子型値の最大値（初期系統数 50 系統、選抜系統数 20 系統）

初期系統数 50 系統、選抜系統数 20 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

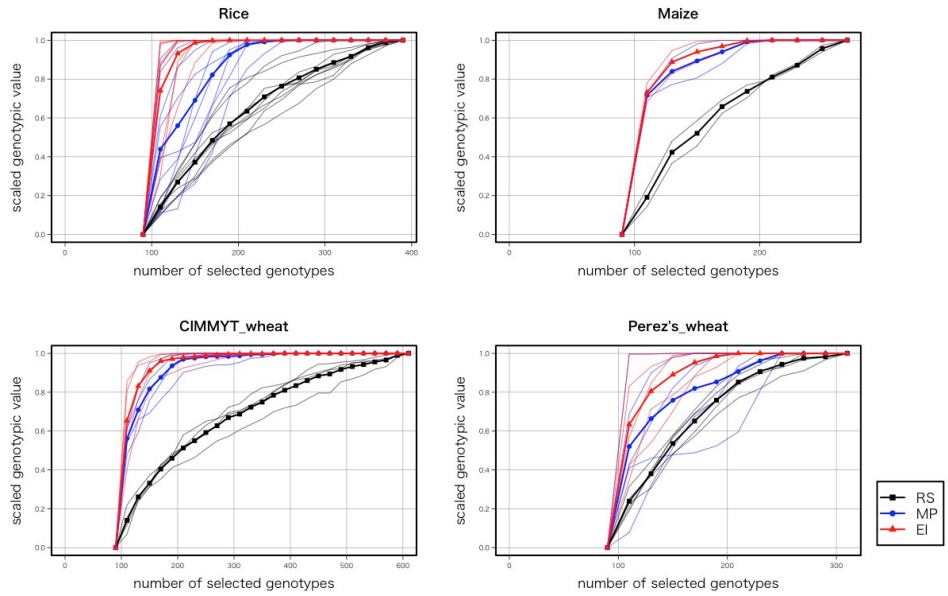


図 4-S5. 選抜済み系統の遺伝子型値の最大値（初期系統数 90 系統、選抜系統数 20 系統）

初期系統数 90 系統、選抜系統 20 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

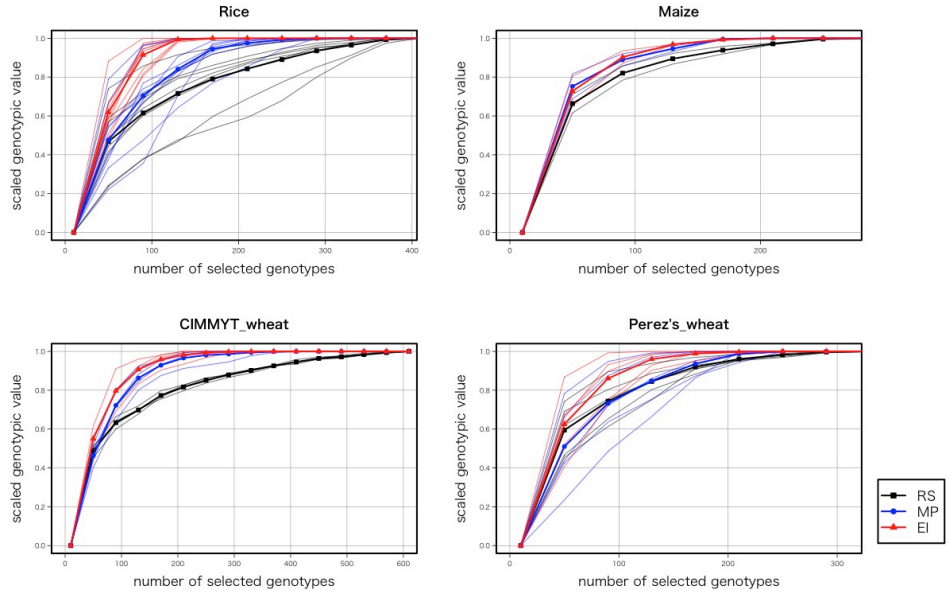


図 4-S6. 選抜済み系統の遺伝子型値の最大値（初期系統数 10 系統、選抜系統数 40 系統）

初期系統数 10 系統、選抜系統数 40 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

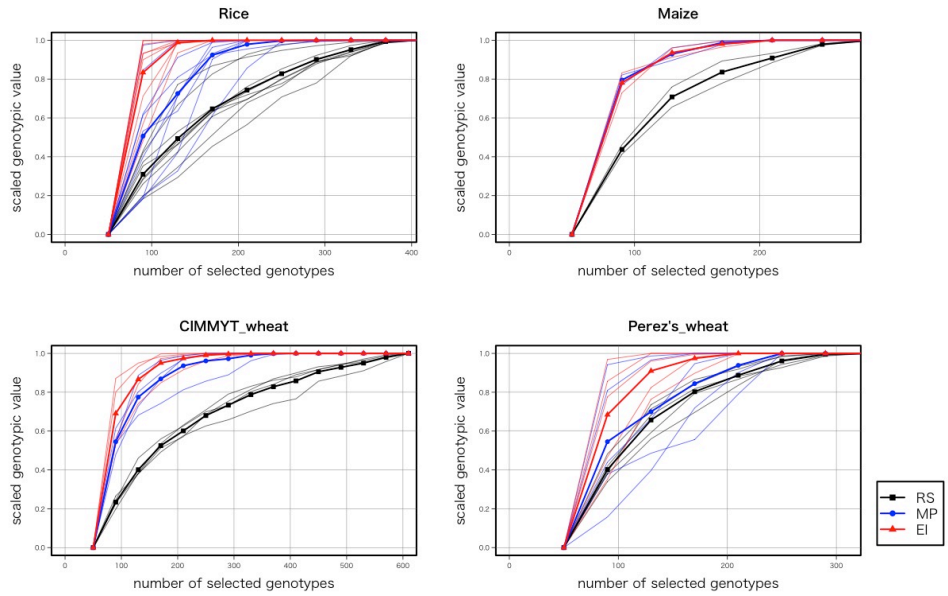


図 4-S7. 選抜済み系統の遺伝子型値の最大値（初期系統数 50 系統、選抜系統数 40 系統）

初期系統数 50 系統、選抜系統 40 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

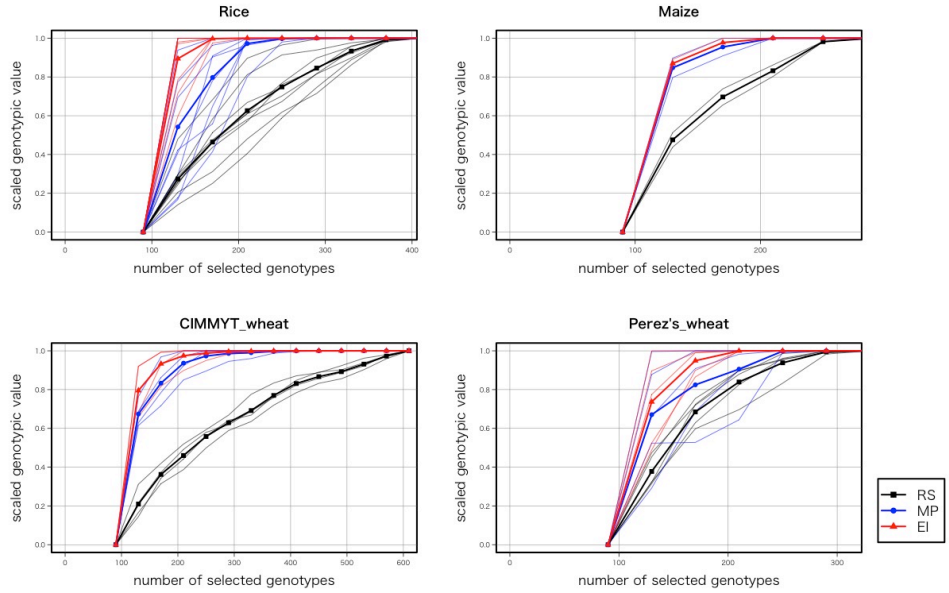


図 4-S8. 選抜済み系統の遺伝子型値の最大値（初期系統数 90 系統、選抜系統数 40 系統）

初期系統数 90 系統、選抜系統数 40 系統のシミュレーションについて、選抜済み系統の遺伝子型値の最大値を示した。作図法は図 4-3 と同じである。

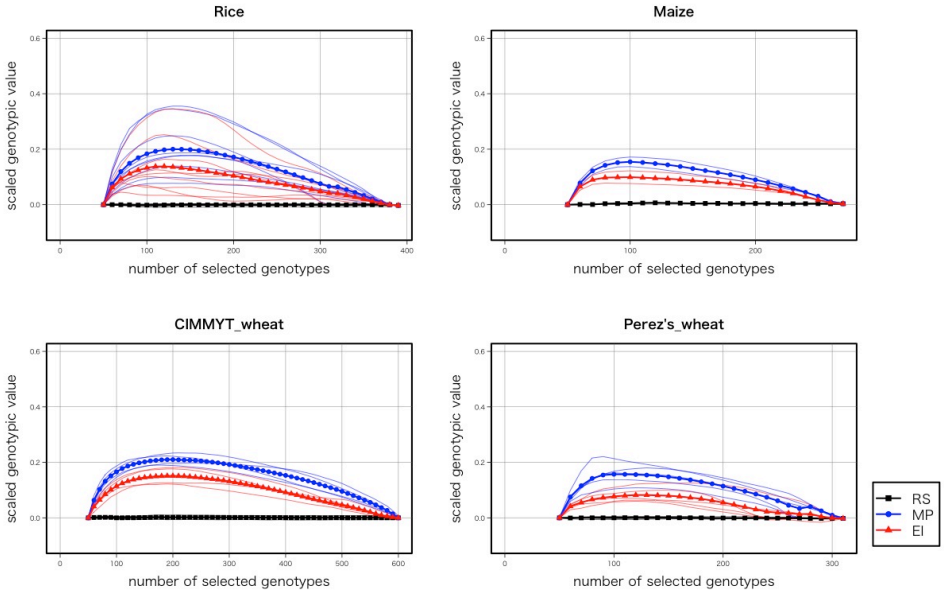


図 4-S9. 選抜済み系統の遺伝子型値の平均値（初期系統数 50 系統、選抜系統数 10 系統）

初期系統数 50 系統、選抜系統数 10 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

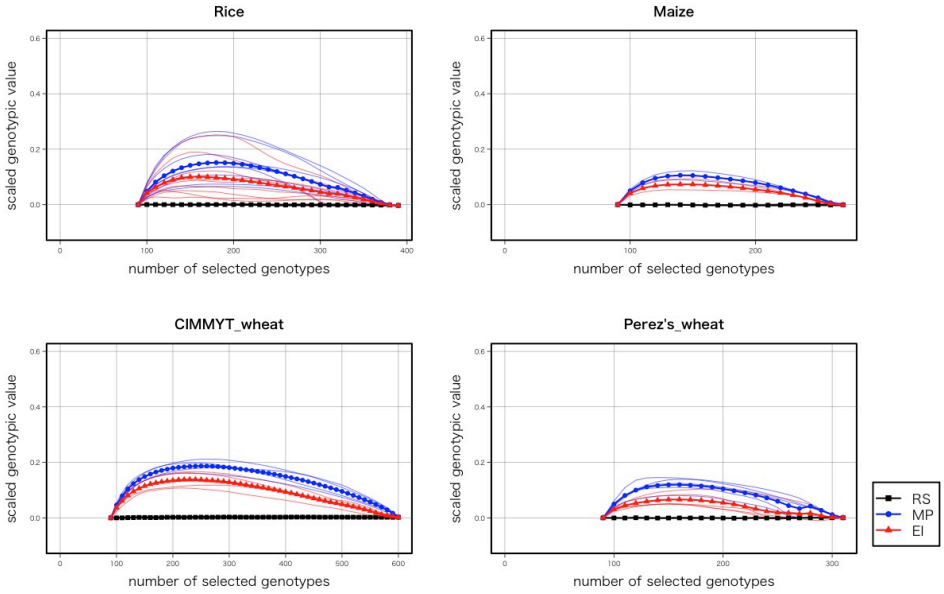


図 4-S10. 選抜済み系統の遺伝子型値の平均値（初期系統数 90 系統、選抜系統数 10 系統）

初期系統数 90 系統、選抜系統数 10 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

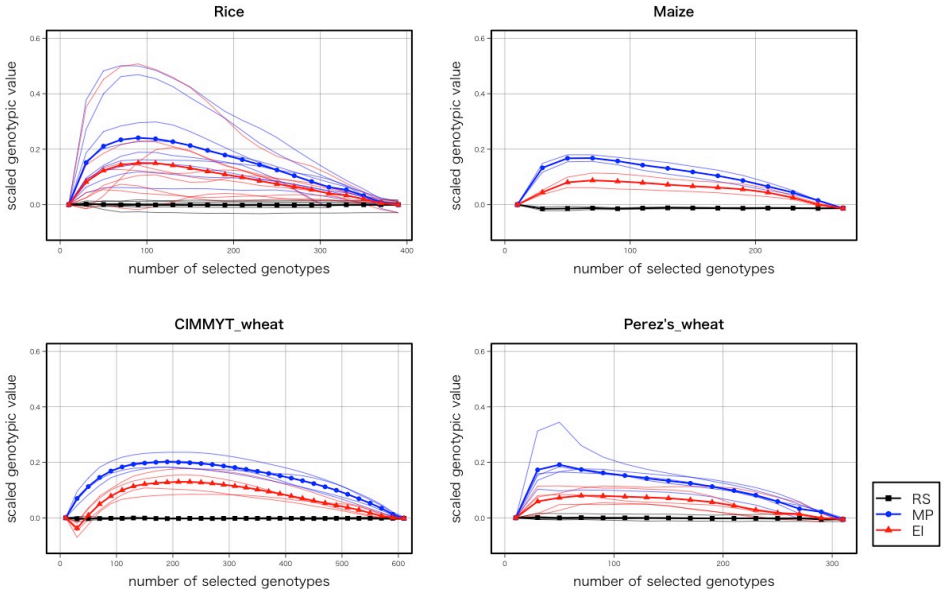


図 4-S11. 選抜済み系統の遺伝子型値の平均値（初期系統数 10 系統、選抜系統数 20 系統）

初期系統数 10 系統、選抜系統数 20 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

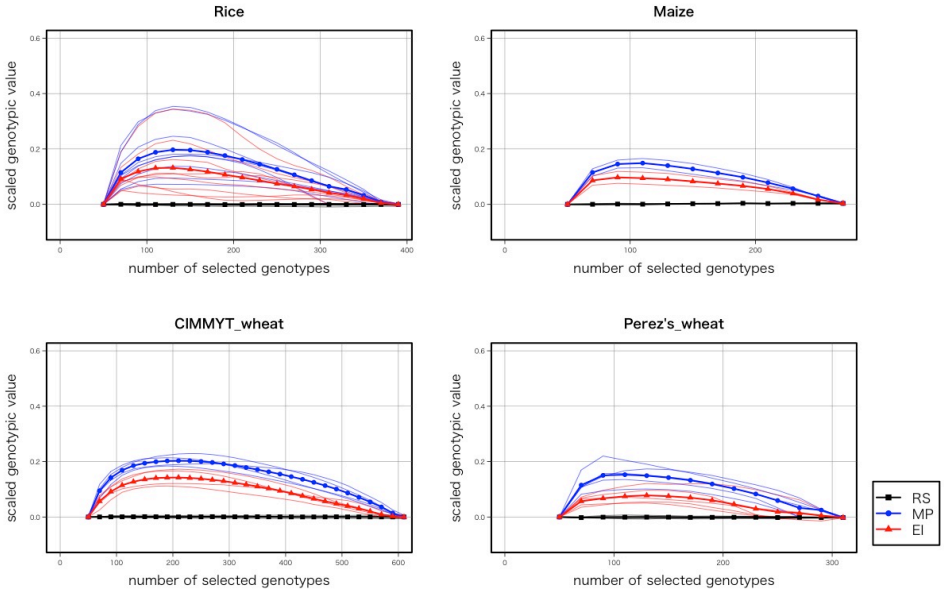


図 4-S12. 選抜済み系統の遺伝子型値の平均値（初期系統数 50 系統、選抜系統数 20 系統）

初期系統数 50 系統、選抜系統数 20 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

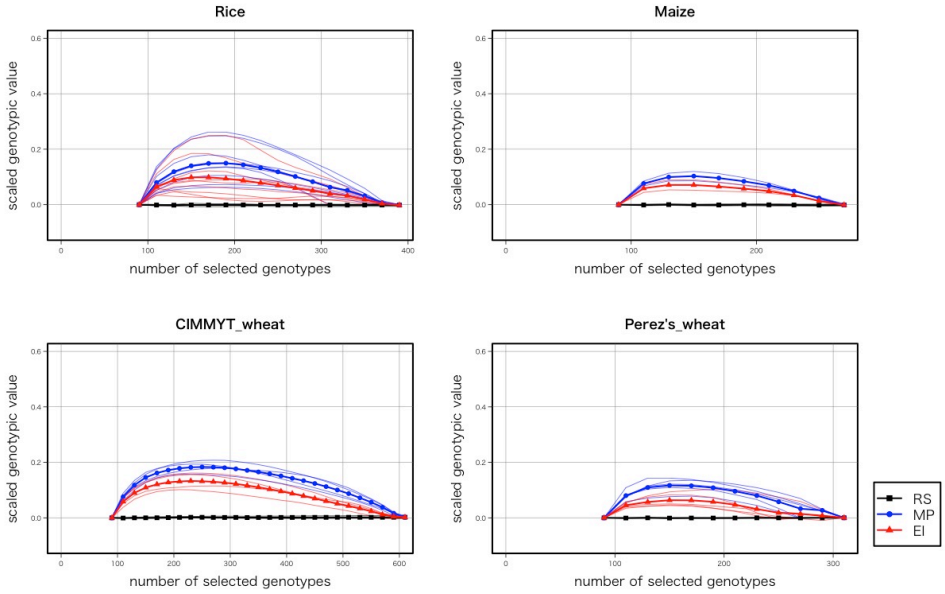


図 4-S13. 選抜済み系統の遺伝子型値の平均値（初期系統数 90 系統、選抜系統数 20 系統）

初期系統数 90 系統、選抜系統数 20 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

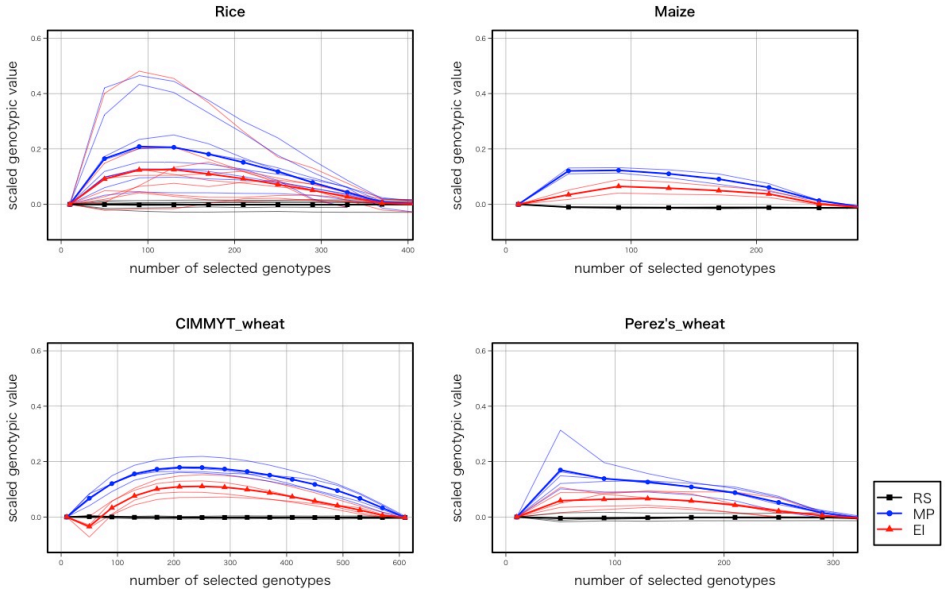


図 4-S14. 選抜済み系統の遺伝子型値の平均値（初期系統数 10 系統、選抜系統数 40 系統）

初期系統数 10 系統、選抜系統数 40 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

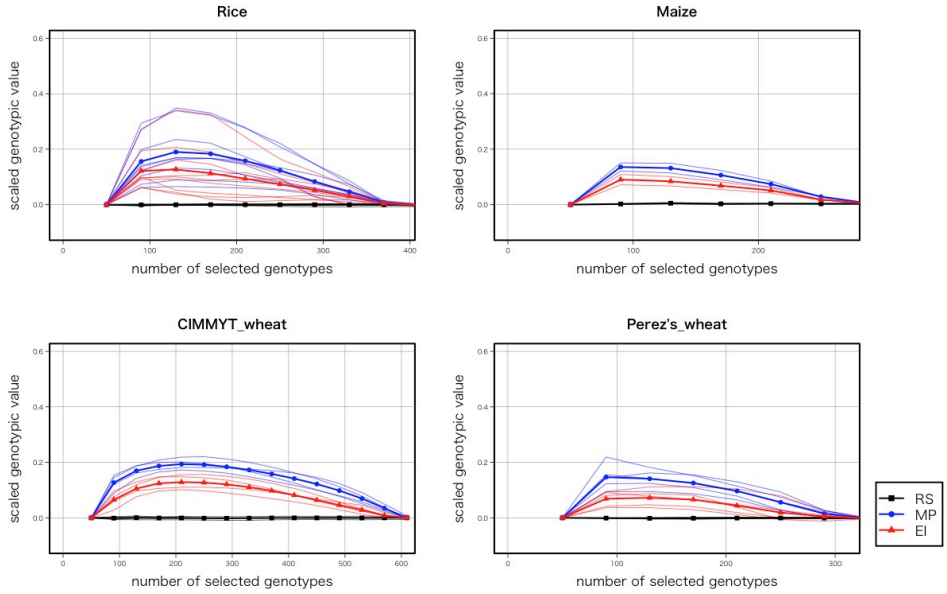


図 4-S15. 選抜済み系統の遺伝子型値の平均値（初期系統数 50 系統、選抜系統数 40 系統）

初期系統数 50 系統、選抜系統数 40 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

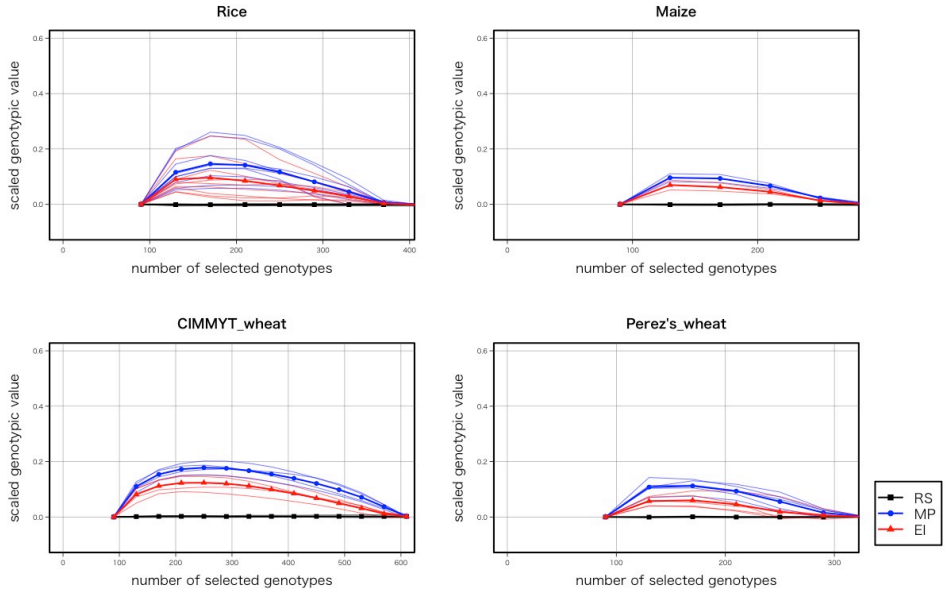


図 4-S16. 選抜済み系統の遺伝子型値の平均値（初期系統数 90 系統、選抜系統数 40 系統）

初期系統数 90 系統、選抜系統数 40 系統のシミュレーションについて、選抜済み系統の遺伝子型値の平均値を示した。作図法は図 4-4 と同じである。

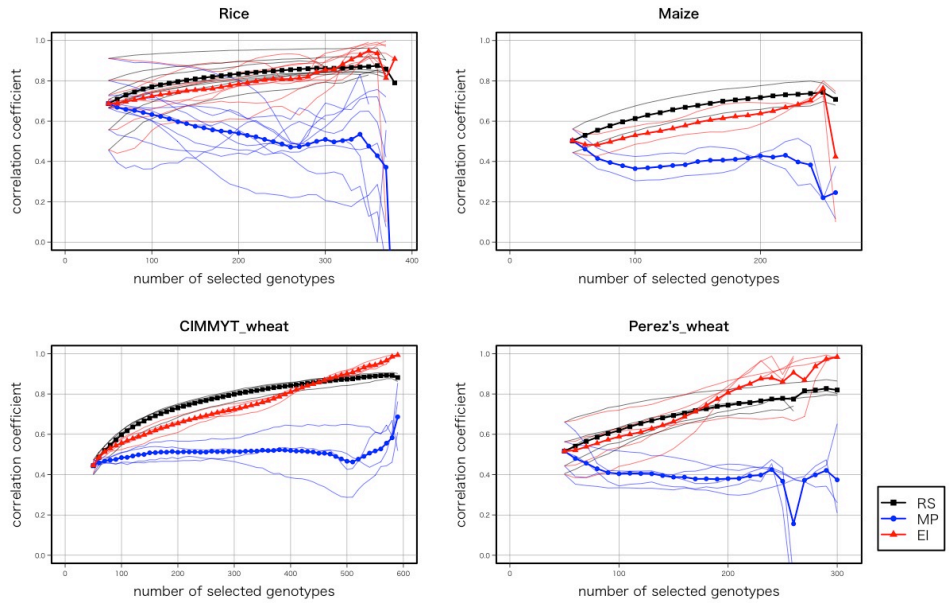


図 4-S17. シミュレーションにおける予測精度の推移(初期系統数 50 系統、選抜系統数 10 系統)

初期系統数 50 系統、選抜系統数 10 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

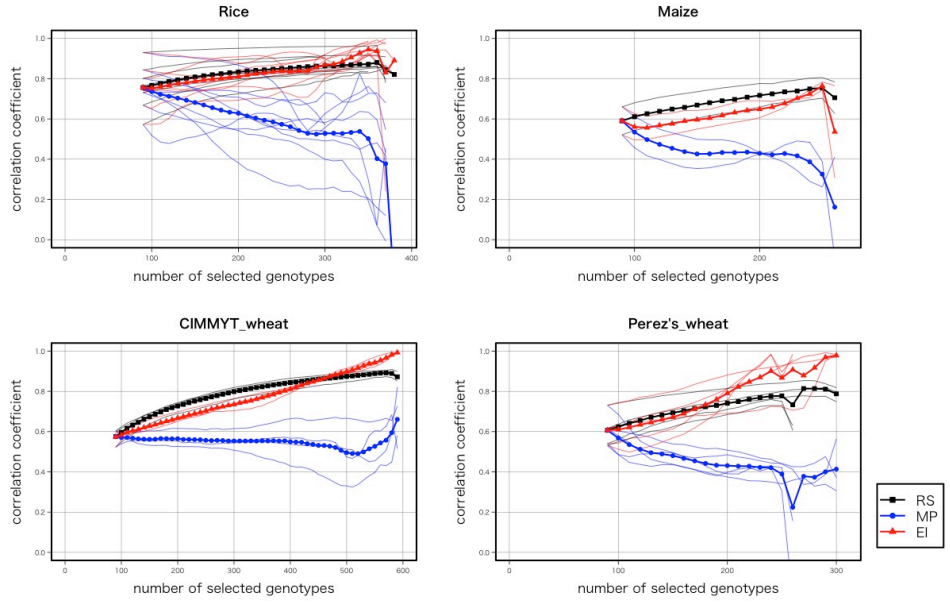


図 4-S18. シミュレーションにおける予測精度の推移(初期系統数 90 系統、選抜系統数 10 系統)

初期系統数 90 系統、選抜系統数 10 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

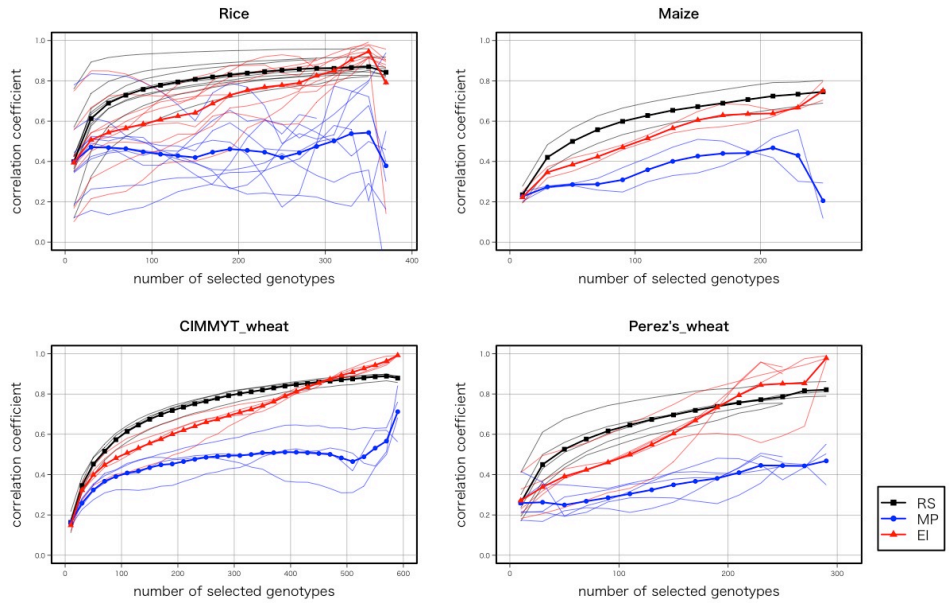


図 4-S19. シミュレーションにおける予測精度の推移(初期系統数 10 系統、選抜系統数 20 系統)

初期系統数 10 系統、選抜系統数 20 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

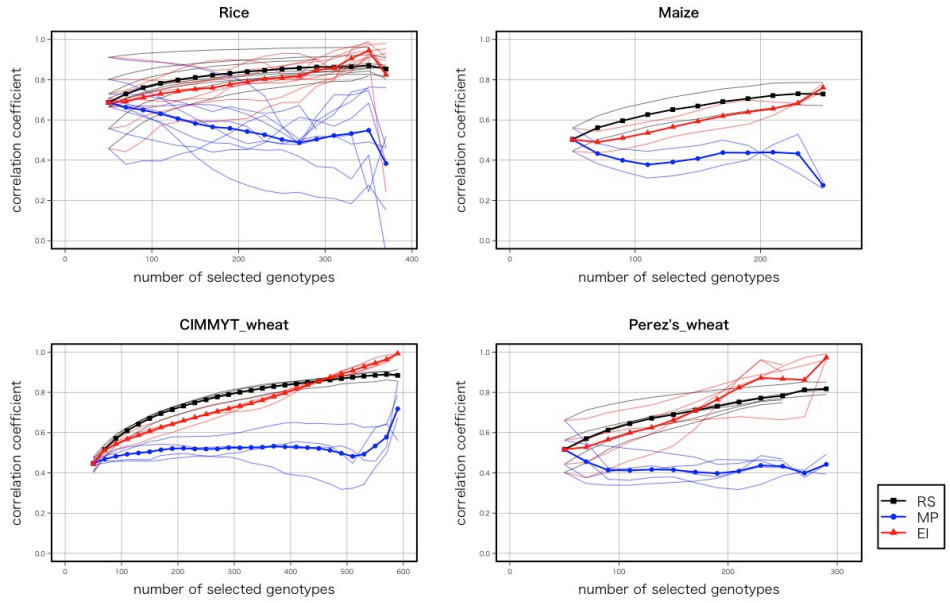


図 4-S20. シミュレーションにおける予測精度の推移(初期系統数 50 系統、選抜系統数 20 系統)

初期系統数 50 系統、選抜系統数 20 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

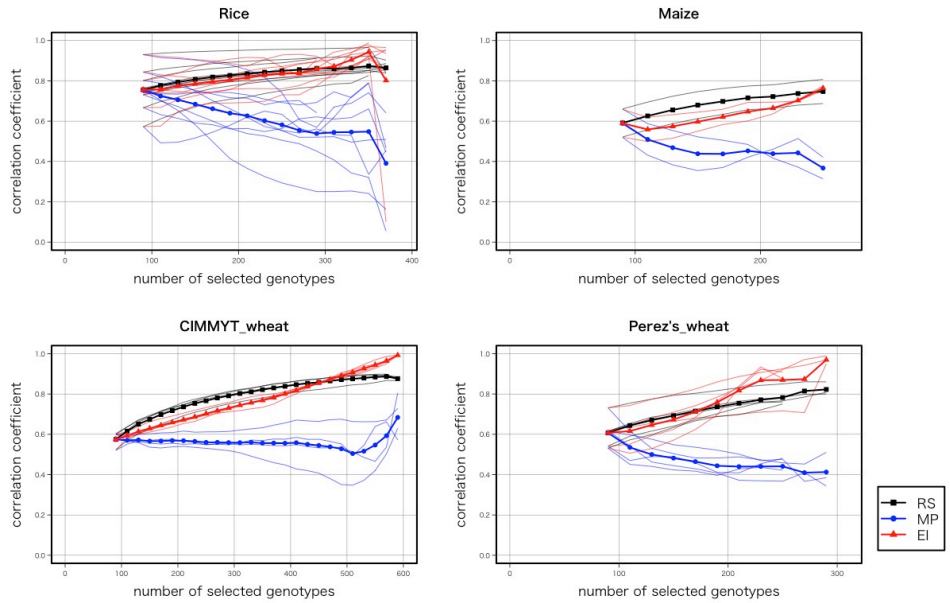


図 4-S21. シミュレーションにおける予測精度の推移(初期系統数 90 系統、選抜系統数 20 系統)

初期系統数 90 系統、選抜系統数 20 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

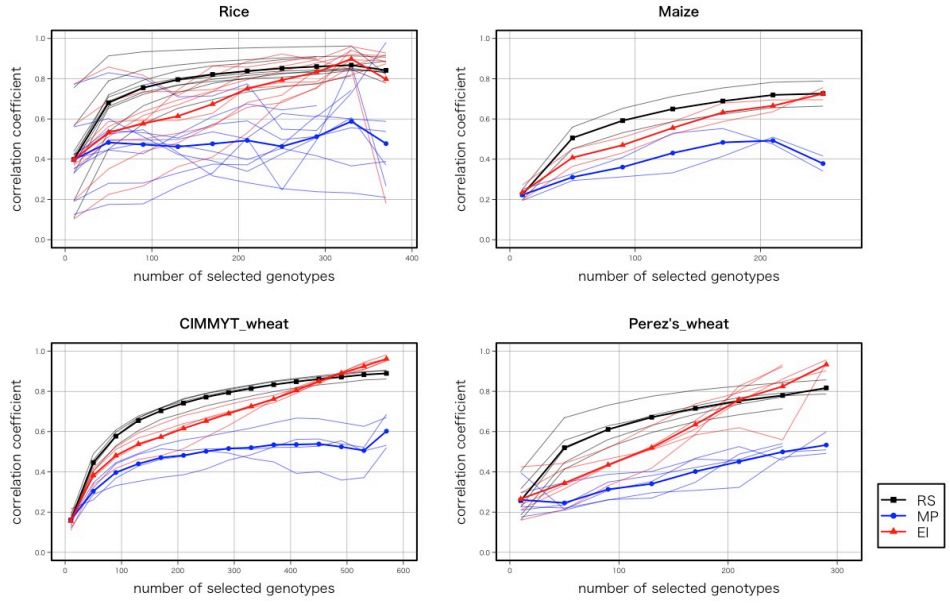


図 4-S22. シミュレーションにおける予測精度の推移(初期系統数 10 系統、選抜系統数 40 系統)

初期系統数 10 系統、選抜系統数 40 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

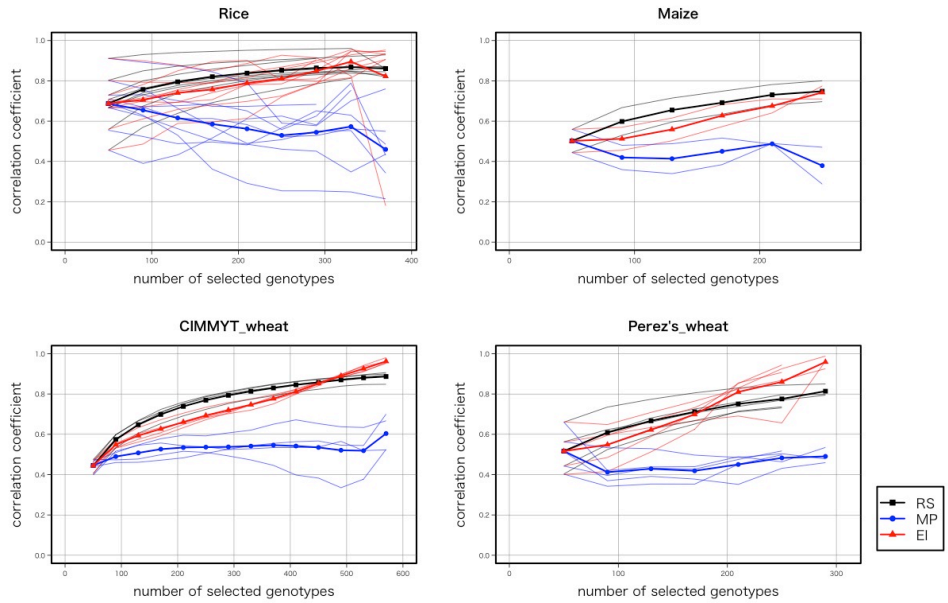


図 4-S23. シミュレーションにおける予測精度の推移（初期系統数 50 系統、選抜系統数 40 系統）

初期系統数 50 系統、選抜系統数 40 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

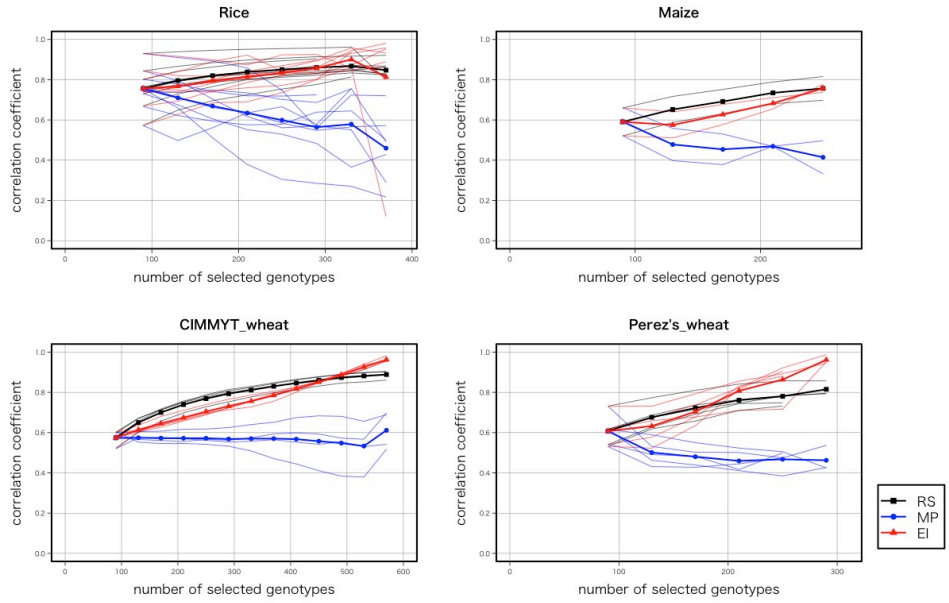


図 4-S24. シミュレーションにおける予測精度の推移（初期系統数 90 系統、選抜系統数 40 系統）

初期系統数 90 系統、選抜系統数 40 系統のシミュレーションについて、予測精度の変化を示した。作図法は図 4-5 と同じである。

5. ゲノミック予測における多環境試験デザインの最適化

5-1. 序論

植物の表現型は、遺伝子型値、環境効果、そして、遺伝子型と環境の交互作用（GxE; genotype-by-environment interaction）、および環境誤差によって定まる。複数の系統を複数の環境や地域で栽培評価する多環境試験は GxE についての知見を得る上で不可欠であり、ほとんどの育種プログラムで実施されている。幅広い環境で栽培できる品種を作出したければ遺伝子型値が大きく GxE が小さい系統を選抜することが望ましいが、環境ごとに異なる品種をリリースすることを許すのであれば、特定の環境で GxE が大きい系統を積極的に選抜することも考えられる。このように GxE への対応は育種目標によって異なるが、いずれの場合にも、多環境試験によって複数環境・複数品種の表現型データを取得し、統計解析を行うことが必須である。

多環境試験から得られるデータは GxE についての貴重な情報を育種家に提供する。しかし、多環境試験を大規模に（多数の系統について、多数の試験地・栽培環境で）実施するのは極めて困難である。4 章でも考えたような遺伝資源のスクリーニングを例にとれば、1 つの環境でさえ全ての系統を栽培することは難しく、複数環境での試験ともなれば、なおさら困難を極めることは言うまでもないであろう。ゆえに、多環境試験を行うに先立って多数（数百〜数千）の系統を小規模に（ある典型的な条件で、1 系統あたり数個体ほど）pre-screening することで有望な系統や特徴的な数十〜数百系統を選抜し、それらについて多環境試験を行う、などの方法が実践的には行われている。つまり、何らかの方法で選ばれた少数の系統についてのみ多環境試験が実施され、試験された系統についてのみ GxE についての知見が得られる。

ゲノミック予測は、多環境試験データに基づく予測（あるいは、多環境試験データの補完）にも有用な手法である。マーカー遺伝子型を持つ系統について、異なる環境ごとにその一部が測定されているとする。このとき、多環境データに対するゲノミック予測（以下、多環境ゲノミック予測とよぶ）を用いることで、測定されなかった系統・環境のペアについても、遺伝子型値（環境によらない遺伝子型値と、その環境における GxE 効果の和）を推定することが可能である（Burgueño et al., 2012）。すなわち、実際に試験して得られる「部分的なデータ」から、全系統を全環境で栽培試験しなければ得られない「完全データ」を推定し、復元することができる。マーカー遺伝子型の取得コストは、多環境試験を大規模に実施するコストに比べれば小さい。ここでの完全データはあくまで予測に基づき得られたデータであるため、その信頼性などを統計的に判断して適切に利用しなければならないが、試験しなかった系統・環境についても何らかの情報を得られることの価値は非常に高いと考えられる。したがって、ゲノミック予測に基づく多環境試験データの予測・復元は、多環境試験データに対する重要な解析手法の 1 つであり、現実的な制約のもとでは、優れた次善の策だといえる。

ここで興味深い研究課題は、多環境試験をどのように計画（デザイン）するべきか、ということである。ここでの多環境試験デザインとは、いわゆる圃場設計（ブロック数や個体数、畝間を何 cm にするかなど）ではなく、『どの系統を、どの環境で試験するか』を意味する。例えば、先述した典型的な多環境試験計画では、「全系統の一部を、全ての環境で試験する」という試験デザインを採用している。別の見方をすれば、選ばれなかった系統は一切試験されないデザインとな

っている。古くから開発されてきた GxE に関する統計解析手法では、ゲノム情報の利用は想定されておらず、系統数×環境数の表現型値の行列から GxE についての知見を得るものであり、特定の系統群を全ての環境で試験することが前提となってきた（e.g. Malosetti et al., 2013）ことが、こうした試験デザインが採用される理由であろう。しかし、ゲノミック予測では、マーカー遺伝子型によって系統間の関連性が定義されることで、ある系統の情報から、別の系統について推論することができる。したがって、環境ごとに全く異なる系統を栽培試験することも、ゲノミック予測の文脈では不適切なデザインとは言えなくなる。

ゲノミック予測の精度がモデル構築に用いる訓練集団に依存する（Asoro et al., 2011; Würschum et al., 2013）ことを考えれば、多環境ゲノミック予測において、異なる実験デザインが異なる予測精度を与えることは想像に難くない。各環境でどの系統を栽培しても、環境ごとの栽培系統数が同じであれば、試験にかかるコストはほぼ同じだと言える。したがって、環境ごとに何系統を栽培するかという条件が与えられたとき、どの環境でどの系統を栽培試験するべきかを適切に決定することで、栽培試験のコストを変えることなく、多環境ゲノミック予測の精度を改善できると考えられる。

この問題設定は、ゲノミック予測における訓練集団最適化を、多環境試験へと拡張したものと捉えることができる。訓練集団最適化とは、（GxE を考慮しない）ゲノミック予測において、どの系統を栽培試験して予測モデル構築に用いるべきかを、マーカー遺伝子型（あるいは、そこから計算されるゲノム関係行列）から最適化しようとする研究の総称である（Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Rincent et al., 2017a）。これまでに、予測誤差分散（PEV; Prediction Error Variance）や決定係数（CD; Coefficient of Determination）といった指標を最小化（PEV の場合）または最大化（CD の場合）することにより、GBLUP モデルの予測精度を高めるように試験するべき系統を選ぶ方法が提案されてきた。これら PEV や CD は、栽培試験を実施する候補となる系統について、マーカー遺伝子型のみから（表現型を用いることなく）計算可能なため、栽培試験を実施する前に PEV や CD を計算して、試験するべき系統の選択に用いることができる。

よって、この PEV や CD を多環境ゲノミック予測に拡張することで、多環境試験デザインを最適化できると考えられる。次節以降で詳述するように、多環境ゲノミック予測に用いられる典型的な予測モデルについては、PEV や CD を自然に拡張することができる。ただし、通常の GBLUP モデルとはことなり、これらの指標を計算するために事前に定めるべき超パラメータが複雑化してしまうことが問題となる。GBLUP モデルにおける PEV や CD は、対象形質の遺伝率だけを超パラメータに持つ。この超パラメータの設定は選ばれる系統の組み合わせにほとんど影響しないことが確かめられており、例えば $h^2 = 0.5$ と設定して計算すればよいと考えられている（Akdemir et al., 2015）。しかし、後述するように、多環境ゲノミック予測で PEV や CD を計算するには、環境ごとの遺伝率と、環境間の遺伝相関行列を指定しなければならなくなる。特に、従来の訓練集団最適化に関する研究では存在しなかった遺伝相関行列は、多環境ゲノミック予測における PEV や CD に基づく最適化に影響を及ぼす可能性が否定できず、詳細に検討されるべきだと考えられる。

なお、PEV の最小化や CD の最大化といった最適化問題は、いわゆる組合せ最適化と呼ばれる、厳密解を求めることが極めて困難な最適化問題である。例えば、500 系統から 100 系統を 1 つの試験地あたり栽培試験できるとすると、1 つの試験地あたりの組み合わせ数は ${}_{500}C_{100} \fallingdotseq 10^{107}$ 通りにもなる。試験地が増えると、それに比例して総組み合わせ数はさらに増加する。したがって、すべての可能な多環境試験デザインについて PEV や CD を計算し尽くして最適なデザインを探すことは不可能であり、何らかの heuristic なアルゴリズムを利用して、(局所) 最小値や最大値を与えるデザインを求める必要がある。先行研究では遺伝的アルゴリズムがこの最適化に用いられており (Akdemir et al., 2015)、わずかな改変を行うだけで多環境ゲノミック予測における PEV や CD の最適化にも適用できると考えられた。

以上のように、多環境試験は GxE について知るためには不可欠であるが、多数の系統を用いた大規模試験を実施するのは現実的ではなく、多環境ゲノミック予測を用いて未試験のデータを補完することが次善の策となる。このとき、ゲノミック予測による補完精度はできる限り高いほうが望ましい。そこで、本研究では、既存の訓練集団最適化で提案された PEV および CD を多環境ゲノミック予測モデルに拡張することで、多環境試験デザインの最適化を試みた。PEV や CD が優れた多環境試験デザインを導けるかどうかを、設定する必要のある超パラメータの影響とあわせて、複数の実データに基づくシミュレーションによって検証した。

5-2. 材料・方法

5-2-1. 混合モデルにおける PEV と CD

はじめに、先行研究で用いられてきた、GxE を考慮しない場合の PEV と CD について説明する。PEV や CD は、GBLUP を含む混合モデルにおいて歴史的に定義されてきた (Henderson, 1984; Laloë, 1993)。いま、ある系統のある環境での表現型値として、複数反復・複数個体の表現型の平均値を用いることを仮定する。このとき、N 次元の表現型ベクトルを $\mathbf{y}$、L 次元の環境効果ベクトルを $\boldsymbol{\beta}$、M 次元の遺伝子型値ベクトルを $\mathbf{u}$、N 次元の残差ベクトルを $\mathbf{e}$ とすると、混合モデルは

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{5.1}$$

とかける。ここで、行列 $\mathbf{X}$ は $\mathbf{y}$ と $\boldsymbol{\beta}$ の要素の対応を表す $N \times L$ 次元の計画行列であり、行列 $\mathbf{Z}$ は $\mathbf{y}$ と $\mathbf{u}$ の要素の対応を表す計画行列である。このとき、遺伝子型値ベクトル $\mathbf{u}$ は

$$p(\mathbf{u}) = N(\mathbf{u}|0, \mathbf{G_0}) \tag{5.2}$$

で表される、行列 $\mathbf{G_0}$ を分散共分散行列とする多変量正規分布に従うことが仮定される。また、尤度関数を

$$p(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) = N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R_0}) \tag{5.3}$$

とする。行列 $\mathbf{R_0}$ は残差の分散共分散行列である。

どの系統がどの環境で測定されるかという実験デザインは、計画行列に反映されることを強調しておく。例えば $\mathbf{y}$ の n 番目の要素が m 番目の系統の l 番目の環境における表現型であったとすると、行列 $\mathbf{X}$ の n 行目は、l 番目の要素は 1 であり、それ以外の要素は 0 である L 次元ベクトルになる。また、行列 $\mathbf{Z}$ の n 行目は、m 番目の要素は 1 であり、それ以外の要素は 0 である m 次元ベクトルになる。環境効果 $\boldsymbol{\beta}$ や遺伝子型ベクトル $\mathbf{u}$ をより柔軟に設計する場合にはこの限りではないが、1 つの環境に 1 つの環境効果を、1 つの系統に 1 つの遺伝子型値を設定することを仮定すれば、この説明で十分である。

PEV と CD は、ともにゲノミック予測の精度に関連する統計量であり、実験デザイン（計画行列）を指定すれば、表現型ベクトル $\mathbf{y}$ によらず定まる。もし混合モデルが正しければ、PEV が小さいほど、CD が大きいほど、予測精度が高くなることが期待される。なお、PEV の定義式は研究ごとにわずかに異なる（Hickey et al., 2009）ようだが、ここでは訓練集団最適化において最も一般的に用いられる、contrast vector と呼ばれるベクトル $\mathbf{c}$ を用いた定義を採用する。訓練集団最適化で最もよく用いられる PEV と CD は以下のように表される(e.g. Rincent et al., 2012)。

$$\mathrm{PEVmean}(\mathbf{X},\mathbf{Z}) = \sum_{m=1}^{M} \mathbf{c}_m{}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P_0}\mathbf{Z}\mathbf{G_0})\mathbf{c}_m \tag{5.4}$$

$$\mathrm{CDmean}(\mathbf{X},\mathbf{Z}) = \sum_{m=1}^{M} \left\{1 - \frac{\mathbf{c}_m{}^{\mathrm{T}}(\mathbf{G_0} - \mathbf{G_0}\mathbf{Z}^{\mathrm{T}}\mathbf{P_0}\mathbf{Z}\mathbf{G_0})\mathbf{c}_m}{\mathbf{c}_m{}^{\mathrm{T}}\mathbf{G_0}\mathbf{c}_m}\right\} \tag{5.5}$$

ただし

$$\mathbf{P_0} \equiv \mathbf{V_0}^{-1} - \mathbf{V_0}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{V_0}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V_0}^{-1} \tag{5.6}$$

$$\mathbf{V_0} \equiv \mathbf{R_0} + \mathbf{Z}\mathbf{G_0}\mathbf{Z}^{\mathrm{T}} \tag{5.7}$$

とした。また、$\mathbf{c}_m$ は第 m 要素を $1-1/M$ とし、それ以外の全ての要素を $-1/M$ とする M 次元ベクトルである。この contrast vector では $\mathbf{c}_m{}^{\mathrm{T}}\mathbf{u} = u_m - \bar{u}$ となり（ただし $\bar{u}$ は遺伝子型の平均値）、ある系統の遺伝子型値の、集団平均からの偏差を考えていることがわかる。なお、PEVmean や CDmean が実験デザインに依存することを明示するために、左辺の括弧に行列 $\mathbf{X}$ および $\mathbf{Z}$ を記した。

5-2-2. 多環境ゲノミック予測における PEV と CD

多環境ゲノミック予測で最も典型的に用いられるモデルは、環境間の遺伝分散共分散行列 $\mathbf{K}$ を用いるものである（Burgueño et al., 2012）。すなわち、環境ごとに異なる遺伝子型値を考え、異なる環境における遺伝子型値が正または負に相関することを認めることで、この相関を環境間の

予測に利用するモデル化を行う。いま、ある環境 l において N_l 系統を測定したとして、その表現型ベクトルを $\mathbf{y}_l$ とし、表現型ベクトルを $\boldsymbol{Y} = \left(\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_L^{\mathrm{T}}\right)^{\mathrm{T}}$ とする。同様に残差ベクトルも $\boldsymbol{E} = \left(\mathbf{e}_1^{\mathrm{T}}, \ldots, \mathbf{e}_L^{\mathrm{T}}\right)^{\mathrm{T}}$ と定める。これらのベクトルは長さ N とする（したがって $N_1 + N_2 + \ldots + N_L = N$ である）。また、遺伝子型値のベクトルを $\boldsymbol{U} = \left(\mathbf{u}_1^{\mathrm{T}}, \ldots, \mathbf{u}_L^{\mathrm{T}}\right)^{\mathrm{T}}$ とする。遺伝子型値ベクトルの長さは $L \times M$ である。このとき、多環境ゲノミック予測モデルは、混合モデルと同じ形式で

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{U} + \boldsymbol{E} \tag{5.8}$$

とかける。計画行列 $\mathbf{X}$ は混合モデルの場合と同じサイズであるが、計画行列 $\mathbf{Z}$ は N 行 $L \times M$ 列であることに注意せよ。また、このモデルでは行列 $\mathbf{Z}$ だけで実験デザインが表現され、行列 $\mathbf{X}$ は $\mathbf{Z}$ を決めれば一意に定まることも付け加えておく。多環境ゲノミック予測では、遺伝子型値の従う分布を

$$p(\boldsymbol{U}) = N(\boldsymbol{U}|0, \mathbf{K} \otimes \mathbf{G}) \tag{5.9}$$

と定める。ここで、記号⊗はクロネッカー積を意味する。行列 $\mathbf{K}$ は環境間の遺伝分散共分散行列（L 次元正方行列）であり、通常はデータから推定される。行列 $\mathbf{G}$ は系統間の（ゲノム）関係行列であり、マーカー遺伝子型から計算される。多環境ゲノミック予測モデルの尤度関数は

$$p(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{\beta}) = N(\boldsymbol{Y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{U}, \mathbf{R}) \tag{5.10}$$

と、やはり混合モデルと同様に書くことができる。残差の分散共分散行列 $\mathbf{R}$ は任意に定めても構わないが、本研究では妥当な仮定として、ある環境の残差は独立同分布であり、かつ、環境間でも残差は独立であるとする。このとき、行列 $\mathbf{R}$ は対角行列を要素とするブロック対角行列となり

$$\mathbf{R} = blockdiag\{\mathbf{R}_l\} \tag{5.11}$$

$$\mathbf{R}_l = \mathbf{I}\sigma_{e_l}^2 \tag{5.12}$$

と表せる。ただし、記号 $blockdiag\{\mathbf{A}_l\}$ で、左上から順に正方行列 $\mathbf{A}_1, \mathbf{A}_2, \ldots$ を持つブロック対角行列を表すものとする。また、分散 $\sigma_{e_l}^2$ は、l 番目の環境における残差分散であり、これもデータから推定されるべきパラメータである。したがって、$\boldsymbol{\sigma}_e^2 = \left(\sigma_{e_1}^2, \ldots, \sigma_{e_L}^2\right)$ のようにまとめて表記するならば、尤度関数(5.10)は

$$p(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{\beta}, \mathbf{K}, \boldsymbol{\sigma}_e^2) \tag{5.13}$$

と表現するべきものである。行列 $\mathbf{K}$ および残差分散 $\boldsymbol{\sigma}_e^2$ の推定には様々な方法が考えられるが、本研究で推定を行う場合には、R の {MTM} パッケージを用いてベイズ推定を行った（de los Campos and Grüneberg, 2014）。同パッケージでは、共役事前分布として、分散共分散行列には逆ウィシャート分布が、残差分散には逆カイ２乗分布がそれぞれ用いられる。

多環境ゲノミック予測の式(5.8), (5.9), (5.10)は、混合モデルと全く同じ形式である。したがっ

て、単純に $\mathbf{K}\otimes\mathbf{G}\rightarrow\mathbf{G_0}$ および $\mathbf{R}\rightarrow\mathbf{R_0}$ と代入することで、PEV や CD を多環境ゲノミック予測へと拡張できる。よって、ある contrast vector について

$$\mathrm{PEV}(\mathbf{Z})=\mathbf{c}^{\mathrm{T}}\{(\mathbf{K}\otimes\mathbf{G})-(\mathbf{K}\otimes\mathbf{G})\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}(\mathbf{K}\otimes\mathbf{G})\}\mathbf{c} \tag{5.14}$$

$$\mathrm{CD}(\mathbf{Z})=1-\frac{\mathbf{c}^{\mathrm{T}}\{(\mathbf{K}\otimes\mathbf{G})-(\mathbf{K}\otimes\mathbf{G})\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}(\mathbf{K}\otimes\mathbf{G})\}\mathbf{c}}{\mathbf{c}^{\mathrm{T}}(\mathbf{K}\otimes\mathbf{G})\mathbf{c}} \tag{5.15}$$

ただし

$$\mathbf{P}\equiv\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1} \tag{5.16}$$

$$\mathbf{V}\equiv\mathbf{R}+\mathbf{Z}(\mathbf{K}\otimes\mathbf{G})\mathbf{Z}^{\mathrm{T}} \tag{5.17}$$

とすればよい。

ここで、contrast vector を適切に設定し平均をとれば、最小化または最大化するべき指標を得ることができる。本研究では、環境ごとに系統の遺伝子型値の（その環境における）集団平均からの偏差を正確に推定することが重要であると考え、これを反映した contrast vector として $\mathbf{c}_{lm}$ を以下のように設計した。

$$\mathbf{c}_{lm}=\mathbf{c}_{\mathrm{ind}}+\mathbf{c}_{\mathrm{mean}} \tag{5.17}$$

ここで $\mathbf{c}_{\mathrm{ind}}$ は第 $M\times(l-1)+m$ 要素だけを 1 とし、その他の要素が全て 0 であるベクトルである。また、$\mathbf{c}_{\mathrm{mean}}$ は第 $M\times(l-1)+1$ 要素から第 $M\times l$ 要素までの全ての要素が $-1/M$ であり、その他の要素が全て 0 であるベクトルである。このとき、m 番目の系統の l 番目の環境における遺伝子型値を u_{ml} とおき、l 番目の環境における遺伝子型値の集団平均を $\bar{u}_l$ とおけば

$$\mathbf{c}_{lm}{}^{\mathrm{T}}\boldsymbol{U}=u_{ml}-\bar{u}_l \tag{5.18}$$

が成り立つ。この contrast vector を用いて、多環境試験デザインの最適化は

$$\mathrm{PEVmean}(\mathbf{Z})=\sum_{l=1}^{L}\sum_{m=1}^{M}\mathbf{c}_{lm}{}^{\mathrm{T}}\{(\mathbf{K}\otimes\mathbf{G})-(\mathbf{K}\otimes\mathbf{G})\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}(\mathbf{K}\otimes\mathbf{G})\}\mathbf{c}_{lm} \tag{5.19}$$

$$\mathrm{CDmean}(\mathbf{Z})=\sum_{l=1}^{L}\sum_{m=1}^{M}\left\{1-\frac{\mathbf{c}_{lm}{}^{\mathrm{T}}\{(\mathbf{K}\otimes\mathbf{G})-(\mathbf{K}\otimes\mathbf{G})\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{Z}(\mathbf{K}\otimes\mathbf{G})\}\mathbf{c}_{lm}}{\mathbf{c}_{lm}{}^{\mathrm{T}}(\mathbf{K}\otimes\mathbf{G})\mathbf{c}_{lm}}\right\} \tag{5.20}$$

このように表現される PEVmean の最小化、または、CDmean の最大化によって実現することができる。なお、行列 $\mathbf{X}$ は $\mathbf{Z}$ を決めれば一意に定まるため、左辺の括弧から省いた。

5-2-3. 多環境ゲノミック予測における PEV と CD の超パラメータ

多環境ゲノミック予測を実際に行う場合には、データに基づき環境間の遺伝分散共分散行列 **K** および残差分散ベクトル $\boldsymbol{\sigma}_e^2$ を推定する。しかし、多環境試験デザインの最適化は表現型データを取得する前に行わなければならないため、これらは何らかの方法で事前に指定するべき超パラメータとみなす必要がある。

本研究では、PEVmean および CDmean の最適化において、遺伝分散共分散行列 **K** および残差分散ベクトル $\boldsymbol{\sigma}_e^2$ を直接指定するのではなく、環境間の遺伝相関行列 **S** と、全ての環境で共通する遺伝率 h^2 を超パラメータとして設定する。さらに、遺伝分散や環境分散が環境間で共通であると仮定する。この仮定は実際には正しくないが、事前に遺伝分散などについて情報を得ることは一般に不可能なため、妥当な単純化だと考えられる。以下では、PEVmean と CDmean の最適化において、このように超パラメータを設定すれば十分であることを示す。

環境間の遺伝分散共分散は、環境間の遺伝相関および遺伝率を用いて表現できる。すなわち

$$\mathbf{K} = \mathbf{S} \cdot \frac{h^2}{1-h^2}\sigma_e^2 = \tilde{\mathbf{S}}\sigma_e^2 \tag{5.21}$$

が成り立つ。ただし

$$\tilde{\mathbf{S}} = \frac{h^2}{1-h^2}\mathbf{S} \tag{5.22}$$

とおいた。ここで、残差共分散について、環境間で遺伝分散が共通であるという仮定より

$$\mathbf{R} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}\sigma_e^2 \tag{5.23}$$

が成り立つことを利用すると

$$\mathbf{V} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}\sigma_e^2 + \mathbf{Z}\big(\tilde{\mathbf{S}}\sigma_e^2 \otimes \mathbf{G}\big)\mathbf{Z}^{\mathrm{T}} = \mathbf{Z}\big(\mathbf{I} + \tilde{\mathbf{S}} \otimes \mathbf{G}\big)\mathbf{Z}^{\mathrm{T}}\sigma_e^2 \tag{5.24}$$

が成り立つ。ここで、**S** と h^2 にのみ依存する項を

$$\tilde{\mathbf{V}} = \mathbf{Z}\big(\mathbf{I} + \tilde{\mathbf{S}} \otimes \mathbf{G}\big)\mathbf{Z}^{\mathrm{T}} \tag{5.25}$$

とおくと

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1} = \left\{\tilde{\mathbf{V}}^{-1} - \tilde{\mathbf{V}}^{-1}\mathbf{X}\big(\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1}\right\}\frac{1}{\sigma_e^2} \tag{5.26}$$

も成り立つ。再び **S** と h^2 にのみ依存する項を

$$\tilde{\mathbf{P}} = \tilde{\mathbf{V}}^{-1} - \tilde{\mathbf{V}}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1} \tag{5.27}$$

とおく。すると、PEVmean は

$$\mathrm{PEVmean}(\mathbf{Z}) = \sigma_e^2 \sum_{l=1}^{L}\sum_{m=1}^{M} \mathbf{c}_{lm}{}^{\mathrm{T}}\left\{\left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right) - \left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right)\mathbf{Z}^{\mathrm{T}}\tilde{\mathbf{P}}\mathbf{Z}\left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right)\right\}\mathbf{c}_{lm} \tag{5.28}$$

このように書くことができる。つまり、相関行列 **S** と遺伝率 h^2 を定めると、PEVmean は任意の計画行列について定数倍 σ_e^2 を除いて計算することができる。定数倍は PEVmean の順序に影響しないため、どのように残差分散 σ_e^2 を定めても、PEVmean の最大値を与える計画行列 **Z** は変わらない。すなわち、行列 **S** と遺伝率 h^2 を定めれば PEVmean の最適化が可能である。同様に CDmean についても

$$\mathrm{CDmean}(\mathbf{Z}) = \sum_{l=1}^{L}\sum_{m=1}^{M} \left\{1 - \frac{\mathbf{c}_{lm}{}^{\mathrm{T}}\left\{\left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right) - \left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right)\mathbf{Z}^{\mathrm{T}}\tilde{\mathbf{P}}\mathbf{Z}\left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right)\right\}\mathbf{c}_{lm}}{\mathbf{c}_{lm}{}^{\mathrm{T}}\left(\tilde{\mathbf{S}} \otimes \mathbf{G}\right)\mathbf{c}_{lm}}\right\} \tag{5.29}$$

が成り立つため、CDmean は環境間の遺伝相関行列 **S** と遺伝率 h^2 を定めれば計算できることがわかる。

なお、上記はあくまで PEVmean および CDmean を最適化して多環境試験デザインを決定する場合に用いる仮定である。後述するシミュレーションにおいて得られた多環境試験デザインを評価する場合には、多環境試験デザインに対応して取得される表現型データを用いて、環境間の遺伝分散共分散 **K** や残差分散ベクトル $\boldsymbol{\sigma}_e^2$ が推定され、遺伝子型値の推定が行われることに注意せよ。

5-2-4. 最適化に用いる遺伝的アルゴリズム

最適な多環境試験デザインの探索、すなわち式(5.28)の最小化や式(5.29)の最大化を行うために、本研究では遺伝的アルゴリズムを用いる。遺伝的アルゴリズムによって得られる最適デザイン **Z** は大域的最適解である保証はないが、無作為に解を探索するよりははるかに効率よく優れた（PEVmaen が小さい、または、CDmean が大きい）デザインを得ることができる。本節では、本研究で用いた遺伝的アルゴリズムについて説明する。このアルゴリズムは（Akdemir et al., 2015）で用いられた遺伝的アルゴリズムに着想を得て設計された。なお、説明では PEVmean の場合だけを表現するが、CDmean の場合にも最小を最大に読み替えれば同じである。

説明に先立って、多環境試験デザインを表現する $M \times L$ 次元の行列 **D** を新たに定義する。ここで、m 番目の系統が l 番目の環境で測定されている場合に **D** の m, l 要素を 1 とし、測定されていない場合には 0 とする。いま、元の仮定（ある系統のある環境での表現型値として、複数反復・複数個体の表現型の平均値を用いること）の下では、行列 **D** を定めれば、計画行列 **Z** および **X** は一意に定まる。ただし、アルゴリズムについて考える場合には、この行列 **D** を用いたほうがわか

りやすい。

アルゴリズムの初期状態として複数の行列 **D** を生成する。これは遺伝的アルゴリズムにおいて個体生成と呼ばれる。ここで、環境ごとに測定できる系統数があらかじめ決まっているため、行列 **D** の列和が栽培系統数に一致するように生成される。初期生成する行列数は、本研究では 800 通りとした。

まず、これらの行列について、PEVmean を計算し（適応度の評価）、最小の PEVmean を持つ行列から順に一定数（本研究では 10 通り）を保持し、残りは全て破棄する。すなわち、いま考えている 800 通りの実験デザインのうち、PEVmean が小さいデザインだけを残す。これは、遺伝的アルゴリズムにおける適応度に基づく選抜のステップに対応する。

次に、保持された行列をもとに、再び 800 通りの行列を生成する。ここでは第一に、交叉または交配と呼ばれるステップを以下のように行う。はじめに、無作為に 2 つの行列を選ぶ。そして、選ばれた 2 つの行列に $\mathbf{D}_1$、$\mathbf{D}_2$ の i,j 要素について、共通の値を持つ（ともに 0 または 1 である）場合には、その値を保持する。そうでない要素については、列ごとにそれら要素を無作為に並べ替える。すなわち、2 つの実験デザインを選んだとき、それぞれの環境について、両方で栽培されている系統はそのままに、どちらか一方で栽培されている系統は無作為に選ぶ。この操作を 800 回繰り返す。次に、交叉によって得られた行列に対し、一定の確率（本研究では 50%）で突然変異を起こす。ここでは、まず無作為に 1 つの列を選択する。そして、その列で 1 の値を持つ要素と 0 の値を持つ要素を 1 つずつ選び、値を入れ替える。これは、ある環境で栽培される系統を、無作為に 1 つだけ交換することを意味する。

上述した選抜、交叉、突然変異を繰り返すことで、PEVmean の小さい実験デザインを効率よく探索することができると考えられる。繰り返し数は、アルゴリズムの収束を、繰り返しごとに得られる PEVmean の値の減少率をもとに定性的に判断して決定することが望ましい。しかし、本研究では、現実的な計算時間でデザインを決定するために、繰り返し数を 100 回に固定した。

5-2-5. シミュレーションの設定

多環境試験デザインの最適化にも PEVmean および CDmean が利用できるのか、また、超パラメータの設定が得られる最適デザインにどのように影響するのかを検討するために、以下のように、実データに基づくシミュレーションを行った。

- M 系統が L 環境で測定された実データを用意した。元の実データに欠測値がある場合には、欠測値を含む系統を全て削除し、欠測値を一切含まない表現型データ行列を作成した。
- 全ての系統についてマーカー遺伝子型がわかっていると仮定した。
- 1 つの環境につき、M 系統のうち 10%, 30%（既定値），または 50%が測定できると

した（以下、栽培系統数 10%の場合、などと記述する）。

- PEVmean および CDmean の計算に必要な環境間の遺伝相関行列 **S** について、実データから計算した表現型相関を用いる場合（既定値）、環境間相関を全て 0.25 にする場合、全て 0.5 にする場合、全て 0.75 にする場合の 4 通りを試した。
- PEVmean および CDmean の計算に必要な遺伝率 h^2 について、$h^2 = 0.25, 0.5, 0.75$ の 3 通りを試した。既定値は $h^2 = 0.5$ とした。
- PEVmean および CDmean の評価および最適デザインの探索は、5-2-2 節から 5-2-4 節で説明した方法で行った。
- 遺伝的アルゴリズムによって得られる解には、初期値依存性が存在する。したがって、本研究では 5 通りの乱数シードを用いて 5 通りの解（最適デザイン）を得た。
- PEVmean 最小化、CDmean 最大化と比較するために、環境ごとに無作為に栽培系統を選ぶ多環境試験デザイン（RANDOM_ver1）および、無作為に選んだ系統を全ての環境で栽培する多環境デザイン（RANDOM_ver2）をそれぞれ 20 通り作成した。
- 得られた多環境デザインに対応する表現型をデータから抽出し、式(5.8)から式(5.12)で定義される多環境ゲノミック予測モデルを用いて遺伝子型値および表現型値を推定・予測した。

なお、栽培系統数、環境間の遺伝相関、および遺伝率については、すべて既定値の場合を標準のシナリオ（scenario.1）とし、そこから 1 つの条件だけを変化させて、合計で 8 通りのシナリオを作成した（表 5.1）。senario.1 から 4 に注目することで、遺伝子型値の環境間相関行列 **S** による影響を、senario.1, 5, 6 に注目することで遺伝率 h^2 による影響を評価することができる。また、senario.1, 7, 8 に注目することで、環境ごとの測定系統数が変わった場合に関する頑健性を考えることができる。

遺伝相関行列の既定値は、現実には不可能な方法であることを強調しておく。表現型データは実験を行わなければ得られないためである。この方法を規定値としたのは、遺伝相関行列については最も理想的だと思われる条件のもとで、他の条件について比較・検討するためである。

また、遺伝相関行列のデフォルト以外の設定では、全ての環境間相関が同じ値に固定されることにも注意されたい。つまり、この設定において、「どの環境で」栽培するべきかについては全くランダムに決定される。しかし、どの系統を多数試験するか、あるいは試験しないかといった系統ごとの試験回数や、どの系統とどの系統を同時に試験するか、といったことは自由に調節することができる。

ゲノミック予測の精度は、以下の 2 通りの異なる評価方法に基づき行った。予測精度の指標として、以下で比較対象とされる 2 つの値について相関係数または RMSE（root mean square error）を環境ごとに計算し、その算術平均をとった。

- 方法１：実測データ（PEVmean などにより得られる、一部の系統を試験した結果）に基づき多環境ゲノミック予測モデルから推定・予測された遺伝子型値と、完全データ（全系統の全環境における表現型値）に基づきモデルから推定された遺伝子型値とを比較する。つまり、それぞれの環境における遺伝子型値 $\boldsymbol{U}$ の推定に注目する。
- 方法２：実測データに基づきモデルから推定・予測された表現型値と、完全データとを比較する。つまり、表現型値の予測に注目する。

いずれの評価方法でも、測定された系統と測定されなかった系統の両方を精度評価に用いたことに注意されたい。これは、選ばれる系統が手法間（PEVmean, CDmean, random_ver1, random_ver2）で異なるため、選ばれなかった（実際に“予測”された）系統だけを精度評価に用いるのが好ましくないと考えられるためである。方法１と方法２では、ゲノミック予測に求められることがわずかに異なる。方法１は、どのようなデータについてもゲノミック予測モデルによって（相加的）遺伝子型値を推定することを想定しており、推定される遺伝子型値が完全データを用いた場合とできるだけ近いことが望ましいという想定に立つ。いっぽう、方法２では、ゲノミック予測モデルにより表現型値を補完・復元するという目的を想定している。

5-2-5. 使用したデータセット

解析には、以下の４つのデータセットを用いた。日本水稲データを除くデータは、3 章または４章で用いたデータと同じである。ただし、欠測値の除去などにより、データの系統数がこれまでとは異なることがある。

なお、ゲノム関係行列は、SNP マーカーの場合には式(2.20)から式(2.22)によって作成し、DArT マーカーの場合には式(4.8)によって作成した。また、トウモロコシデータでは、データセット内にあらかじめ含まれていた関係行列を用いた。ただし、トウモロコシデータを除く３つのデータセットでは、ゲノム関係行列に基づく種々の逆行列演算が数値的に不安定になることがあった。そこで、ゲノム関係行列の対角成分に 10^{-10} を加えることで計算を安定させるという ad hoc な対応を行った。

・　日本水稲データ（JapanRice dataset）

日本水稲 112 品種について、国内４試験地（北海道札幌市、秋田県大曲市、兵庫県加西市、岡山県笠岡市、千葉県つくば市）で 2015 年に記録された到穂日数データを用いた。マーカー遺伝子型として、minor allele frequency が 2.5%より大きくなるように filtering した 257,879 SNPs を用いてゲノム関係行列を作成した。欠測値のある系統を除き、最終的に 95 系統、５試験地を解析に用いた。このデータセットにおける環境間の表現型相関は、最小値 0.88、中央値 0.95、最大値 0.99 であった。

・　イネ遺伝資源データ（RiceDiversity dataset）

4章で用いた同名のデータセットと同じであるが、ここでは多環境試験を想定するため、解析対象形質として、Aberdeen における開花期（Flowering time at Aberdeen）、Arkansas における開花期（Flowering time at Arkansas）、Faridpur における開花期（Flowering time at Faridpur）の3つに注目した。表現型値の欠測を除外し、最終的に 301 系統、3環境のデータを解析に用いた。なお、本研究ではマーカー遺伝子型の filtering と imputation を実行して得られた 29,564 SNP を用いてゲノム関係行列を作成した。このデータにおける表現型値の環境間相関は 0.49（Faridpur と Arkansas）、0.66（Arkansas と Aberdeen）、および 0.48（Aberdeen と Faridpur）であった。

・　コムギデータ（Wheat dataset）

3章および4章で用いた Perez コムギデータと同一のデータセットであるが、ここでは 2010 年に取得された収量のみを用いた。環境（栽培条件）数は全部で5つであり、乾燥条件かつ畝あり：drought-bed、乾燥条件かつ畝なし：drought-flat、灌漑条件かつ畝あり：irrigation-bed、灌漑条件かつ畝なし：irrigation-flat、加温条件かつ畝あり：heat-bed となっている。欠測値を除外し、最終的に 266 系統、5 環境のデータを解析に用いた。ゲノム関係行列は、minor allele frequency が 5%より大きい 1,666 DArT マーカーに基づき計算した。このデータにおける表現型値の環境間相関は、最小値−0.10、中央値 0.17、最大値 0.46 であった。

・　トウモロコシデータ（Maize dataset）

3章および4章で用いたトウモロコシデータ（Crossa et al., 2010）とほぼ同一のデータセットだと思われるが、本研究では、より新しい Crossa et al.（2013）において用いられたマーカー遺伝子型データ、および対応する表現型データを用いた。この論文では genotyping-by-sequencing 法によってマーカー遺伝子型を得ており、複数の imputation 手法について比較が行われた。ここでは、比較された手法のうち、本研究で用いる集団において安定して高い予測精度を実現した、隣接 100 マーカーによる imputaion で得られたマーカー遺伝子型に基づくゲノム関係行列を用いた。なお、GBS により取得された SNP マーカーには欠測値が多く含まれるため、ゲノム関係行列の作成には僅かな工夫をすることが望ましい（Crossa et al., 2013; VanRaden, 2008）。そのため、本研究では Crossa らが作成したゲノム関係行列をそのまま用いた。

本データに含まれる系統数は 253 系統であり、異なる条件（乾燥条件：SS, Severe Stress、灌漑条件：WW, Well-Watered）での収量を表現型データとして持つ。表現型値の環境間相関は 0.27 であった。

表 5-1. シミュレーションのシナリオ設定

異なる 8 通りのパラメータ設定でシミュレーションを行った。遺伝子型値の環境間相関行列 **S** は表現型値の環境間相関（Phenotypic Correlation）を既定値とし、環境間で共通の遺伝率（Heritability）は 0.5 を既定値とし、栽培系統数（Ratio.Select）は 30%を既定値とした。

	S matrix	Heritability	Ratio.Select
scenario.1	Phenotypic Correlation	0.50	30%
scenario.2	0.25 for all	0.50	30%
scenario.3	0.50 for all	0.50	30%
scenario.4	0.75 for all	0.50	30%
scenario.5	Phenotypic Correlation	0.25	30%
scenario.6	Phenotypic Correlation	0.75	30%
scenario.7	Phenotypic Correlation	0.50	10%
scenario.8	Phenotypic Correlation	0.50	50%

5-3. 結果

・ PEVmean および CDmean による予測精度の向上

PEVmean, CDmean および２通りの無作為デザインに基づく予測精度(相関係数または RMSE)を表 5-2 から表 5-5 にまとめた。表 5-2 と表 5-3 には遺伝子型値について計算した相関係数および RMSE（方法１）を、表 5-4 と表 5-5 は表現型値について計算した相関係数および RMSE（方法２）を示した。PEVmean と CDmean については、5 通りの最適デザインから得られる実測データに基づく予測精度の平均（AVG）および標準偏差（SD）を示し、さらに、RANDOM_ver1 との予測精度の差を対応のない両側 t 検定によって検定したときの $-\log_{10} P$ 値を表に示した。無作為デザイン（RANDOM_ver1, RANDOM_ver2）については、20 通りの無作為デザインから得られる予測精度の平均と標準偏差を示した。

検定の有意水準について、ここで補足しておく。まず、多重検定についての補正を考えない場合には $-\log_{10}P = 1.31$ が 5%有意水準の、$-\log_{10}P = 2.00$ が 1%有意水準の閾値となる。いま、１つのデータセットについて 8 シナリオの独立な検定を行なっているとして Bonferroni 補正を適用すると、5%有意水準は $-\log_{10}P = 2.21$ になり、1％有意水準は $-\log_{10}P = 2.91$ になる。あるいは４通りの評価方法についても補正を行う（つまり合計で 32 通りの検定があると考えて補正を実施する）場合には、5%有意水準は $-\log_{10}P = 2.81$ になり、1％有意水準は $-\log_{10}P = 3.51$ になる。これらを考慮し、表の視認性を高めるため、対応する $-\log_{10}P$ 値が 3 を超える場合の予測精度については下線（精度が向上した場合には一重下線、低下した場合は二重下線）を引いて強調した。また、以下では特に断らない限り、$-\log_{10}P$ 値が 3 を超えることを「有意」と表現する。

まず、全てのデータセットについて、どのような評価方法を用いた場合でも、PEVmean および CDmean を用いた場合に、予測精度が有意に低下した例は２例（表 5-2 と表 5-4；ともにイネ遺伝資源データ, scenario.2, PEVmean の場合）のみであった。いっぽう、データセットや評価方法によって程度は異なるものの、PEVmean および CDmean を用いて多環境デザインを決めた場合に予測精度が向上した例が多数確認された。

評価方法の違いについて、表 5-2 から表 5-5 を見比べると、表現型について予測精度を計算した表 5-4 と表 5-5 では、PEVmean と CDmean によって有意に精度が向上した例が、相対的に少なくなっている。有意ではない条件についても $-\log_{10}P$ 値は全体的に小さくなっており、表現型の復元という意味では、PEVmean や CDmean によって多環境試験デザインを決定しても、復元の精度は、環境ごとに無作為に試験する系統を選ぶ場合と同程度であることが示唆された。ただし、日本水稲データについてはこの限りではなく、例えば RMSE ベースで 0.5~1 日ほど高い精度で完全データを復元できた。なお、表 5-2 と表 5-3、あるいは表 5-4 と表 5-5 を比較すると、相関係数と RMSE による違いは多少あるものの、共通して有意な条件も多く認められた。

以下では主に表 5-2 に注目し、データセット間の違い、PEVmean と CDmean の違い、およびパラメータの影響について結果を確認する。まず、データセット間で、PEVmean および CDmean による予測精度の向上が顕著なものとそうでないものがあることがわかる。例えば、日本水稲デ

ータでは、PEVmean や CDmean により最適デザインを決めることで、ほとんどのパラメータ設定で予測精度が改善した。同データにおける予測精度は RANDOM_ver1 の場合でも非常に高かった（全体の 30%を試験する場合に 0.959）が、それがさらに高まった（表 5-2)。これを RMSE ベースで確認すると、3.770 日の RMSE が、最も良い scenario.1 の場合で 2.701 日まで減少したことがわかる (表 5-3)。いっぽう、同じイネのデータでも、イネ遺伝資源データでは、PEVmean や CDmean による予測精度の変化はほとんど有意ではなかった。

PEVmean と CDmean で、予測精度に顕著な違いは見られなかった。表 5-2 や表 5-3 で予測精度が有意に向上したケースを数えれば PEVmean の方が多かったが、予測精度そのものについて確認すると、日本水稲データでは CDmean のほうが PEVmean より高く、逆にコムギデータやトウモロコシデータでは PEVmean のほうが CDmean より高い傾向にあった。ただし、その差はほとんどの場合で非常にわずかであった。

全て既定値の scenario.1 では、イネ遺伝資源データを除く 3 つのデータで、PEVmean、CDmean の少なくとも一方を用いた場合に相関係数が有意に向上した (表 5-2)。そこから遺伝子型の環境間相関に関するパラメータを変化させた scenario.2 から scenario.4 の結果を見ると、予測精度の平均値は scenario.1 に比べてわずかに悪化する例がほとんどであった。例えば、表 5-2 のコムギデータにおける CDmean の結果に注目すると、scenario.1 での相関係数は 0.761 であるが、scenario.2 では 0.752 に、scenario.3 では 0.759 に、scenario.4 では 0.719 にそれぞれ低下していた。scenario.1 では、遺伝子型の環境間相関行列 **S** に、本来は実験結果からしか得られないはずの表現型相関を用いている。つまり、この超パラメータ **S** の値をかなり正しく与えた場合の予測精度が scenario.1 であるため、この結果は妥当なものと推察される。

さらに詳細に結果を確認すると、遺伝子型の環境間相関 **S** や遺伝率 h^2 の設定が、正しい値に近いほど、得られる多環境試験デザインは高精度になることが示唆された。例えば、トウモロコシデータは表現型相関が 0.27 と低いデータセットであったが、ここで表 5-2 を確認すると、scenario.2（環境間相関 0.25 の場合）や scenario.5（遺伝率が 0.25 の場合）において、最適デザインに基づく予測精度が高かったことがわかる。いっぽうで、日本水稲データは環境間相関が極めて高い (中央値 0.95) データであったが、再び表 5-2 を確認すると、このデータでは scenario.4（環境間相関 0.75 の場合）や scenario.6（遺伝率 0.75 の場合）に、他の設定よりも相関係数が高くなっていた。さらに、逆に scenario.2 の場合には相関係数が有意に低下していた。このように、遺伝子型の環境間相関 **S** に対して実際の相関とあまりに異なるパラメータを入力してしまうと、得られる最適デザインが不適切なものになり、かえって精度の悪化を招きうることがわかった。なお、環境間相関 **S** の違いによってどのように最適デザインが異なるかについては、すぐ後で改めて検証する。

栽培系統数（全系統に対する、試験される系統割合）が変わっても、PEVmean や CDmean により得られる最適デザインは、やはり無作為なデザインに比べて優れた予測精度を与えることが示唆された（表 5-2, scenario.7 および scenario.8)。栽培系統数の違いは予測精度には大きく影響するが、例え全系統の 10%しか栽培できず予測精度の高いモデルを得ることが困難であっても、逆に全体の 50%を栽培可能であり予測精度の高いモデルを得られる場合でも、PEVmean や

CDmean によって試験デザインを決めることで、得られるモデルをより良いものにできることが示唆された。

最後に、２つの異なる無作為デザイン RANDOM_ver1 と RANDOM_ver2 の比較を行うと、ほぼ全ての場合でRANDOM_ver1 のほうが予測精度の高い試験デザインを導くことがわかった。特に日本水稲データでは違いが顕著であり、栽培系統数 30%の場合に RANDOM_ver1 での相関係数が 0.959 であるのに対し、RANDOM_ver2 での相関係数は 0.832 にとどまった。このことから、ゲノミック予測により実測データから完全データを復元することだけを考えれば、ある系統に絞って多環境試験を実施するよりは、それぞれの環境で無作為に系統を選んで試験を行なうほうが望ましいと考えられる。

表 5-2. 実測データに基づく遺伝子型値と完全データに基づく遺伝子型値の間の相関係数（方法 1 による相関係数）。ここで、AVG は平均値、SD は標準偏差を表す。また、CDmean および PEVmean により得られる予測精度と RANDOM_ver1 により得られる予測精度との差を両側 t 検定して得られる P 値の負の常用対数を–log(p)として示した。また、–log(p)の値が 3 を超えるものについて、PEVmean や CDmean によって精度が向上したものを下線で、低下したものを二重下線で示した。詳細は本文を参照せよ。

		CDmean			PEVmean			RANDOM_ver1		RANDOM_ver2	
		AVG	SD	-log(p)	AVG	SD	-log(p)	AVG	SD	AVG	SD
Wheat Dataset	scenario.1	0.761	0.015	4.47	0.747	0.014	3.93	0.699	0.021	0.685	0.023
	scenario.2	0.752	0.012	5.03	0.743	0.016	3.14				
	scenario.3	0.759	0.019	3.20	0.747	0.010	5.27				
	scenario.4	0.719	0.011	1.80	0.746	0.013	4.00				
	scenario.5	0.721	0.012	2.05	0.730	0.013	2.79				
	scenario.6	0.759	0.024	2.64	0.730	0.021	1.61				
	scenario.7	0.519	0.020	4.00	0.530	0.019	4.96	0.449	0.040	0.441	0.094
	scenario.8	0.872	0.026	1.57	0.875	0.006	7.45	0.834	0.019	0.814	0.018
Maize Dataset	scenario.1	0.723	0.016	2.24	0.731	0.008	5.22	0.692	0.023	0.689	0.025
	scenario.2	0.721	0.010	3.37	0.731	0.014	3.22				
	scenario.3	0.722	0.016	2.06	0.727	0.019	2.06				
	scenario.4	0.696	0.018	0.18	0.699	0.023	0.29				
	scenario.5	0.716	0.017	1.59	0.734	0.003	7.24				
	scenario.6	0.735	0.019	2.48	0.732	0.017	2.68				
	scenario.7	0.530	0.030	2.97	0.528	0.022	3.51	0.446	0.073	0.458	0.086
	scenario.8	0.844	0.012	1.72	0.838	0.006	2.00	0.825	0.015	0.815	0.021
RiceDiversity Dataset	scenario.1	0.861	0.018	0.75	0.881	0.025	1.45	0.847	0.016	0.815	0.022
	scenario.2	0.851	0.019	0.15	0.858	0.026	0.37				
	scenario.3	0.880	0.015	2.27	0.874	0.024	1.18				
	scenario.4	0.877	0.012	2.71	0.876	0.008	3.90				
	scenario.5	0.874	0.018	1.55	0.854	0.023	0.22				
	scenario.6	0.862	0.012	1.31	0.869	0.010	2.36				
	scenario.7	0.602	0.210	0.37	0.657	0.181	0.14	0.687	0.098	0.593	0.095
	scenario.8	0.941	0.012	1.81	0.944	0.005	5.49	0.921	0.011	0.895	0.015
JapanRice Dataset	scenario.1	0.980	0.001	7.45	0.976	0.004	4.84	0.959	0.011	0.832	0.065
	scenario.2	0.959	0.003	0.02	0.944	0.005	3.15				
	scenario.3	0.976	0.002	6.08	0.966	0.001	2.07				
	scenario.4	0.976	0.003	5.53	0.976	0.002	5.93				
	scenario.5	0.977	0.005	4.51	0.974	0.002	5.10				
	scenario.6	0.979	0.001	7.01	0.979	0.002	7.18				
	scenario.7	0.904	0.009	2.01	0.899	0.004	1.84	0.825	0.122	0.550	0.163
	scenario.8	0.992	0.002	3.31	0.994	0.001	6.78	0.987	0.004	0.906	0.030

表 5-3. 実測データに基づく遺伝子型値と完全データに基づく遺伝子型値の間の RMSE（方法 1 による RMSE）。ここで、AVG は平均値、SD は標準偏差を表す。また、CDmean および PEVmean により得られる予測精度と RANDOM_ver1 により得られる予測精度との差を両側 t 検定して得られる P 値の負の常用対数を–log(p)として示した。また、–log(p)の値が 3 を超えるものについて、PEVmean や CDmean によって精度が向上したものを下線で、低下したものを二重下線で示した（この表には、精度が低下した例はない）。詳細は本文を参照せよ。

		CDmean			PEVmean			RANDOM_ver1		RANDOM_ver2	
		AVG	SD	-log(p)	AVG	SD	-log(p)	AVG	SD	AVG	SD
Wheat Dataset	scenario.1	0.455	0.014	3.37	0.467	0.010	3.58	0.498	0.016	0.510	0.017
	scenario.2	0.465	0.010	3.93	0.466	0.014	2.51				
	scenario.3	0.464	0.018	1.97	0.466	0.003	7.40				
	scenario.4	0.481	0.008	2.20	0.468	0.012	2.66				
	scenario.5	0.483	0.012	1.33	0.473	0.009	2.99				
	scenario.6	0.479	0.026	0.76	0.491	0.014	0.43				
	scenario.7	0.604	0.010	3.19	0.603	0.011	3.05	0.631	0.018	0.628	0.033
	scenario.8	0.359	0.034	0.80	0.352	0.011	3.45	0.385	0.019	0.406	0.018
Maize Dataset	scenario.1	0.326	0.008	2.32	0.329	0.008	1.99	0.343	0.014	0.340	0.013
	scenario.2	0.331	0.011	1.08	0.322	0.012	1.95				
	scenario.3	0.325	0.013	1.53	0.321	0.007	3.51				
	scenario.4	0.331	0.013	0.97	0.339	0.011	0.32				
	scenario.5	0.327	0.011	1.63	0.319	0.006	4.47				
	scenario.6	0.333	0.008	1.21	0.324	0.011	2.02				
	scenario.7	0.417	0.004	2.13	0.409	0.007	3.20	0.430	0.018	0.428	0.019
	scenario.8	0.257	0.013	0.86	0.257	0.009	1.18	0.268	0.013	0.269	0.017
RiceDiversity Dataset	scenario.1	6.107	0.369	0.23	5.714	0.769	0.65	6.212	0.339	6.614	0.381
	scenario.2	6.301	0.478	0.15	5.820	0.562	0.70				
	scenario.3	5.574	0.426	1.61	5.813	0.613	0.65				
	scenario.4	5.919	0.607	0.46	5.898	0.358	0.90				
	scenario.5	5.868	0.461	0.75	6.149	0.481	0.10				
	scenario.6	6.094	0.318	0.31	6.049	0.643	0.22				
	scenario.7	8.752	1.443	0.08	8.835	1.072	0.17	8.603	0.790	9.335	0.792
	scenario.8	4.126	0.333	1.20	3.992	0.095	5.27	4.502	0.333	5.074	0.361
JapanRice Dataset	scenario.1	2.712	0.088	8.68	2.974	0.339	2.64	3.770	0.463	7.000	1.160
	scenario.2	3.700	0.132	0.25	4.288	0.206	2.71				
	scenario.3	2.857	0.092	7.52	3.413	0.061	2.51				
	scenario.4	2.963	0.216	4.31	2.908	0.080	7.15				
	scenario.5	2.942	0.347	2.66	3.039	0.113	5.74				
	scenario.6	2.767	0.073	8.29	2.770	0.106	8.07				
	scenario.7	5.739	0.400	2.76	5.873	0.163	2.68	7.141	1.576	10.386	1.376
	scenario.8	1.637	0.220	2.34	1.416	0.067	7.74	2.095	0.334	5.386	0.682

表 5-4. 実測データに基づく表現型の予測値と完全データ（実測値）の間の相関係数（方法２による相関係数）。ここで、AVG は平均値、SD は標準偏差を表す。また、CDmean および PEVmean により得られる予測精度と RANDOM_ver1 により得られる予測精度との差を両側 t 検定して得られる P 値の負の常用対数を−log(p)として示した。また、−log(p)の値が 3 を超えるものについて、PEVmean や CDmean によって精度が向上したものを下線で、低下したものを二重下線で示した。詳細は本文を参照せよ。

		CDmean			PEVmean			RANDOM_ver1		RANDOM_ver2	
		AVG	SD	-log(p)	AVG	SD	-log(p)	AVG	SD	AVG	SD
Wheat Dataset	scenario.1	0.617	0.012	3.83	0.609	0.007	5.36	0.577	0.018	0.563	0.017
	scenario.2	0.612	0.009	4.54	0.608	0.012	2.97				
	scenario.3	0.620	0.016	2.85	0.610	0.011	3.37				
	scenario.4	0.592	0.012	1.30	0.613	0.012	3.22				
	scenario.5	0.594	0.011	1.64	0.600	0.010	2.54				
	scenario.6	0.616	0.018	2.45	0.602	0.011	2.75				
	scenario.7	0.412	0.014	4.23	0.417	0.008	7.14	0.360	0.028	0.354	0.066
	scenario.8	0.727	0.025	0.89	0.729	0.006	4.21	0.706	0.017	0.687	0.015
Maize Dataset	scenario.1	0.662	0.013	2.02	0.661	0.012	2.11	0.638	0.024	0.635	0.023
	scenario.2	0.661	0.010	2.29	0.659	0.013	1.75				
	scenario.3	0.656	0.013	1.41	0.659	0.021	1.07				
	scenario.4	0.642	0.023	0.13	0.642	0.023	0.14				
	scenario.5	0.651	0.015	0.79	0.670	0.006	4.71				
	scenario.6	0.673	0.019	2.05	0.664	0.015	1.92				
	scenario.7	0.466	0.023	2.86	0.462	0.021	2.83	0.404	0.055	0.412	0.063
	scenario.8	0.783	0.017	0.74	0.778	0.012	0.61	0.770	0.017	0.764	0.019
RiceDiversity Dataset	scenario.1	0.650	0.018	0.43	0.674	0.019	0.85	0.658	0.014	0.630	0.019
	scenario.2	0.640	0.011	1.87	0.666	0.020	0.36				
	scenario.3	0.673	0.013	1.19	0.665	0.016	0.42				
	scenario.4	0.669	0.022	0.44	0.667	0.010	0.93				
	scenario.5	0.664	0.015	0.36	0.652	0.012	0.46				
	scenario.6	0.654	0.012	0.24	0.664	0.015	0.33				
	scenario.7	0.462	0.151	0.40	0.497	0.129	0.19	0.527	0.068	0.449	0.064
	scenario.8	0.726	0.009	0.26	0.727	0.004	0.68	0.723	0.009	0.703	0.014
JapanRice Dataset	scenario.1	0.967	0.002	7.40	0.962	0.003	4.92	0.945	0.011	0.820	0.064
	scenario.2	0.944	0.004	0.18	0.928	0.006	3.49				
	scenario.3	0.962	0.002	5.87	0.951	0.001	1.68				
	scenario.4	0.962	0.003	5.38	0.962	0.002	5.60				
	scenario.5	0.963	0.005	4.28	0.960	0.002	4.69				
	scenario.6	0.964	0.001	6.70	0.965	0.002	6.98				
	scenario.7	0.889	0.008	1.96	0.883	0.004	1.78	0.811	0.122	0.541	0.159
	scenario.8	0.981	0.002	3.28	0.982	0.001	5.51	0.975	0.004	0.895	0.029

表 5-5. 実測データに基づく表現型の予測値と完全データ（実測値）の間の RMSE（方法 2 による RMSE)。ここで、AVG は平均値、SD は標準偏差を表す。また、CDmean および PEVmean により得られる予測精度と RANDOM_ver1 により得られる予測精度との差を両側 t 検定して得られる P 値の負の常用対数を−log(p)として示した。また、−log(p)の値が 3 を超えるものについて、PEVmean や CDmean によって精度が向上したものを下線で、低下したものを二重下線で示した（この表には、精度が低下した例はない)。詳細は本文を参照せよ。

		CDmean			PEVmean			RANDOM_ver1		RANDOM_ver2	
		AVG	SD	-log(p)	AVG	SD	-log(p)	AVG	SD	AVG	SD
Wheat Dataset	scenario.1	0.794	0.011	2.44	0.800	0.005	3.31	0.817	0.015	0.827	0.011
	scenario.2	0.798	0.006	3.56	0.800	0.012	1.58				
	scenario.3	0.802	0.017	0.96	0.802	0.003	3.80				
	scenario.4	0.811	0.011	0.54	0.804	0.012	1.15				
	scenario.5	0.804	0.007	2.06	0.801	0.007	2.55				
	scenario.6	0.815	0.020	0.08	0.819	0.011	0.10				
	scenario.7	0.929	0.014	2.00	0.927	0.011	2.77	0.954	0.016	0.941	0.029
	scenario.8	0.717	0.029	0.11	0.711	0.010	0.11	0.712	0.018	0.725	0.014
Maize Dataset	scenario.1	0.474	0.012	0.82	0.475	0.009	0.94	0.485	0.015	0.482	0.014
	scenario.2	0.479	0.013	0.36	0.467	0.013	1.49				
	scenario.3	0.473	0.014	0.86	0.467	0.007	2.93				
	scenario.4	0.474	0.014	0.75	0.483	0.008	0.14				
	scenario.5	0.477	0.013	0.50	0.470	0.008	1.96				
	scenario.6	0.479	0.011	0.43	0.470	0.011	1.37				
	scenario.7	0.614	0.026	1.03	0.592	0.013	0.21	0.588	0.026	0.578	0.025
	scenario.8	0.399	0.014	0.38	0.398	0.012	0.54	0.405	0.014	0.401	0.017
RiceDiversity Dataset	scenario.1	11.637	0.322	0.35	11.356	0.593	0.22	11.510	0.252	11.727	0.303
	scenario.2	11.762	0.349	0.73	11.239	0.249	1.14				
	scenario.3	11.230	0.296	0.98	11.389	0.456	0.23				
	scenario.4	11.417	0.433	0.18	11.517	0.252	0.02				
	scenario.5	11.517	0.328	0.02	11.679	0.298	0.54				
	scenario.6	11.556	0.355	0.10	11.539	0.463	0.05				
	scenario.7	13.588	1.177	0.10	13.550	0.664	0.12	13.440	0.671	14.087	0.671
	scenario.8	10.147	0.266	0.10	10.113	0.153	0.32	10.184	0.311	10.433	0.323
JapanRice Dataset	scenario.1	3.397	0.069	7.99	3.657	0.302	2.51	4.329	0.448	7.652	1.341
	scenario.2	4.279	0.099	0.19	4.818	0.193	2.71				
	scenario.3	3.522	0.100	6.76	4.005	0.041	2.33				
	scenario.4	3.592	0.185	4.53	3.588	0.084	6.30				
	scenario.5	3.597	0.310	2.66	3.754	0.114	4.45				
	scenario.6	3.521	0.089	6.87	3.480	0.129	6.65				
	scenario.7	6.348	0.414	2.70	6.705	0.390	1.83	7.767	1.621	11.372	1.630
	scenario.8	2.370	0.193	2.23	2.240	0.063	6.67	2.755	0.290	5.934	0.761

・　遺伝子型値の環境間相関が最適デザインに及ぼす影響

パラメータとして事前に定める遺伝子型値の環境間相関行列 **S** や遺伝率 h^2 が異なれば、得られる最適デザインは異なる。行列 **S** の影響が顕著に確認できる図として、それぞれのデータにおける scenario.2 から scenario.4 の場合について、系統の選ばれた頻度を棒グラフによって示した（図 5-1 から 5-4)。また、遺伝率 h^2 の影響が確認できる scenario.1, scenario.5, scenario.6 についても同様に示した（図 5-5 から 5-8)。

まず、図 5-1 から図 5-4 より、遺伝子型値の環境間相関の大小によって、得られる最適デザインの性質が大きく異なることがわかる。コムギデータの CDmean の場合を例にとると、環境間相関が 0.25 と低い scenario.2 では、全く試験されない系統が約 60 系統あるいっぽうで、3 回以上試験される系統が約 50 系統あった。逆に、環境間相関が 0.75 と高い scenario.4 では、全く試験されない系統は約 20 系統しかなく、また、3 回以上試験される系統もほとんどかった。つまり、大部分の系統は 1 回または 2 回試験されていた。この大まかな傾向は PEVmean でも CDmean でも同じであった。すなわち、遺伝子型値の環境間相関が低い場合には、集中的に試験される系統と全く試験されない系統に別れるような実験デザインが選ばれ、環境間相関が高い場合には、あらゆる系統を満遍なく試験するような実験デザインが選ばれていた。

図 5-1 から図 5-4 の上下に注目すると、PEVmean と CDmean の違いも認められた。環境間相関の低い scenario.2 に注目すると、PEVmean は CDmean よりも、全く試験しない系統の数が明らかに多い。他の scenario.3 および scenario.4 でも、これは同様である。すなわち、PEVmean のほうが、CDmean よりも、特定の系統を選んで試験する傾向が強いことが示唆された。予測精度の意味では両者に大きな差異は見られなかったが、どのような試験デザインを志向するかについては、両者に比較的明瞭な違いがあると考えられる。

図 5-5 から図 5-8 を見ると、遺伝率もまた、最適デザインの性質に影響することがわかる。遺伝率が低い（scenrio.5）場合には、環境間相関が低い場合と同様に、評価しない系統が多数あった。いっぽう、遺伝率が高い場合（scenrio.6）には、試験されない系統の数が減少し、比較的多くの系統を満遍なく試験する傾向が見られた。ただし、遺伝率に対する応答にはデータセット間でも違いが見られた。コムギデータとトウモロコシデータでは、遺伝率が低い場合とそうではない場合に大きな差が見られたが、イネ遺伝資源データではそのような劇的な違いはなく、遺伝率の増加とともに実験デザインが緩やかに変化した。また、日本水稲データでは、遺伝率による試験デザインの違いはほとんど見られなかった。

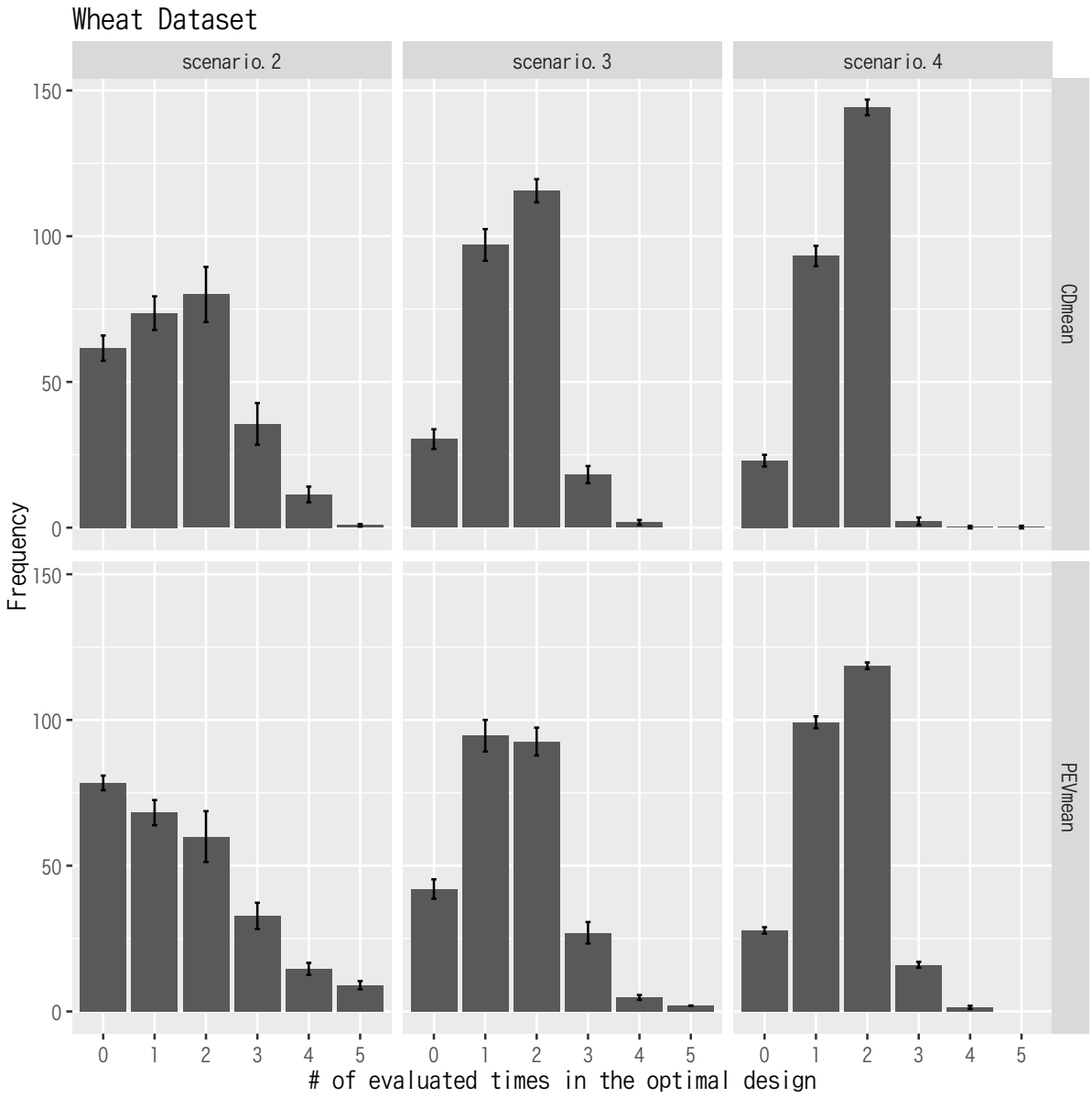


図 5-1. 環境間相関と最適化指標が最適デザインに及ぼす影響（コムギデータ）

設定した遺伝子型値の環境間相関が異なる scenario.2 から scenario.4 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。環境間相関が低い場合には試験されない系統が多数ある一方で、環境間相関が高い場合にはほとんどの系統が 1 回以上試験されたことがわかる。

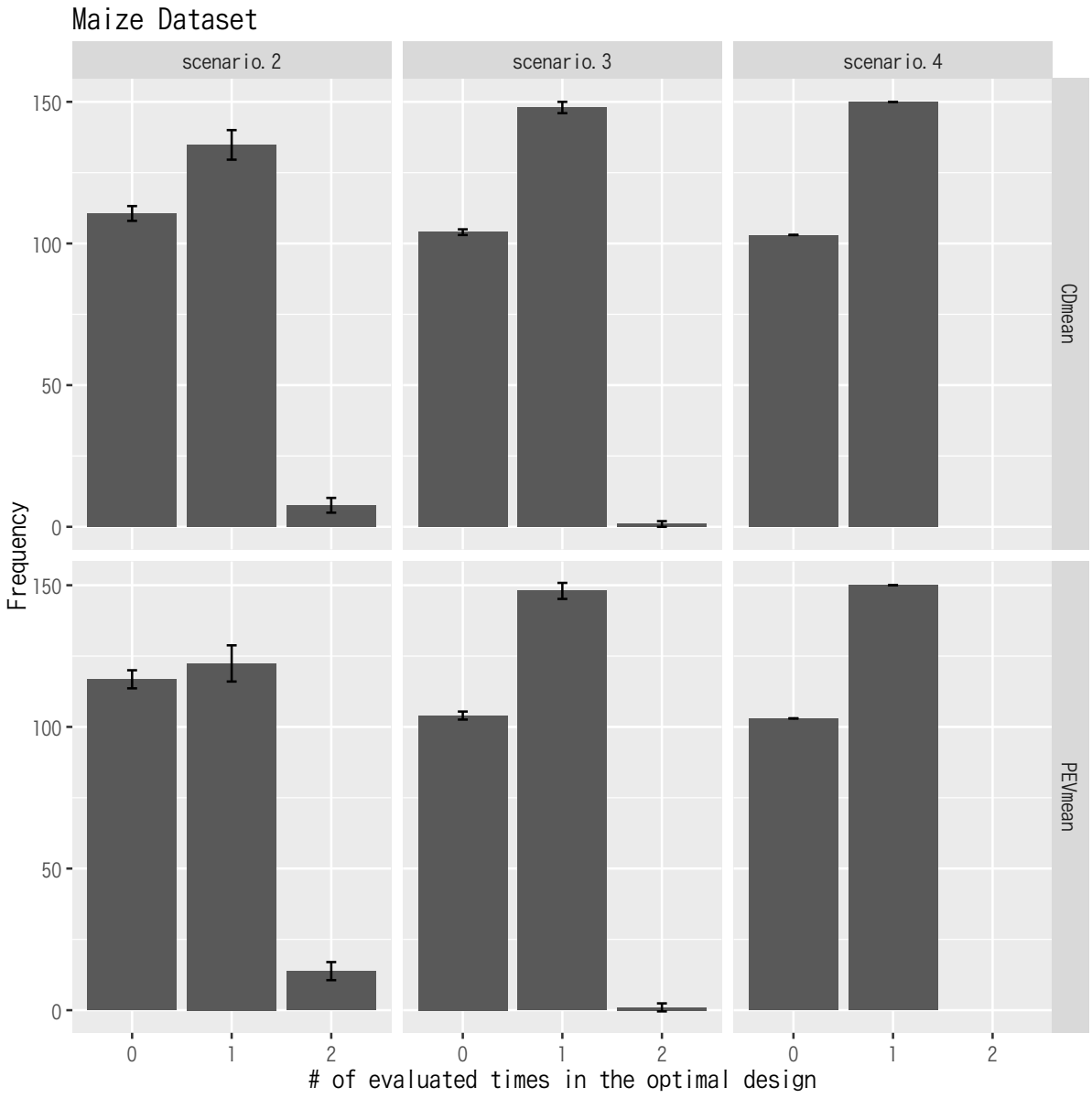


図 5-2. 環境間相関と最適化指標が最適デザインに及ぼす影響（トウモロコシデータ）

設定した遺伝子型値の環境間相関が異なる scenario.2 から scenario.4 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。トウモロコシデータは環境数が２であるため、他のデータほど傾向は明瞭でないが、環境間相関に対する応答は同じである。

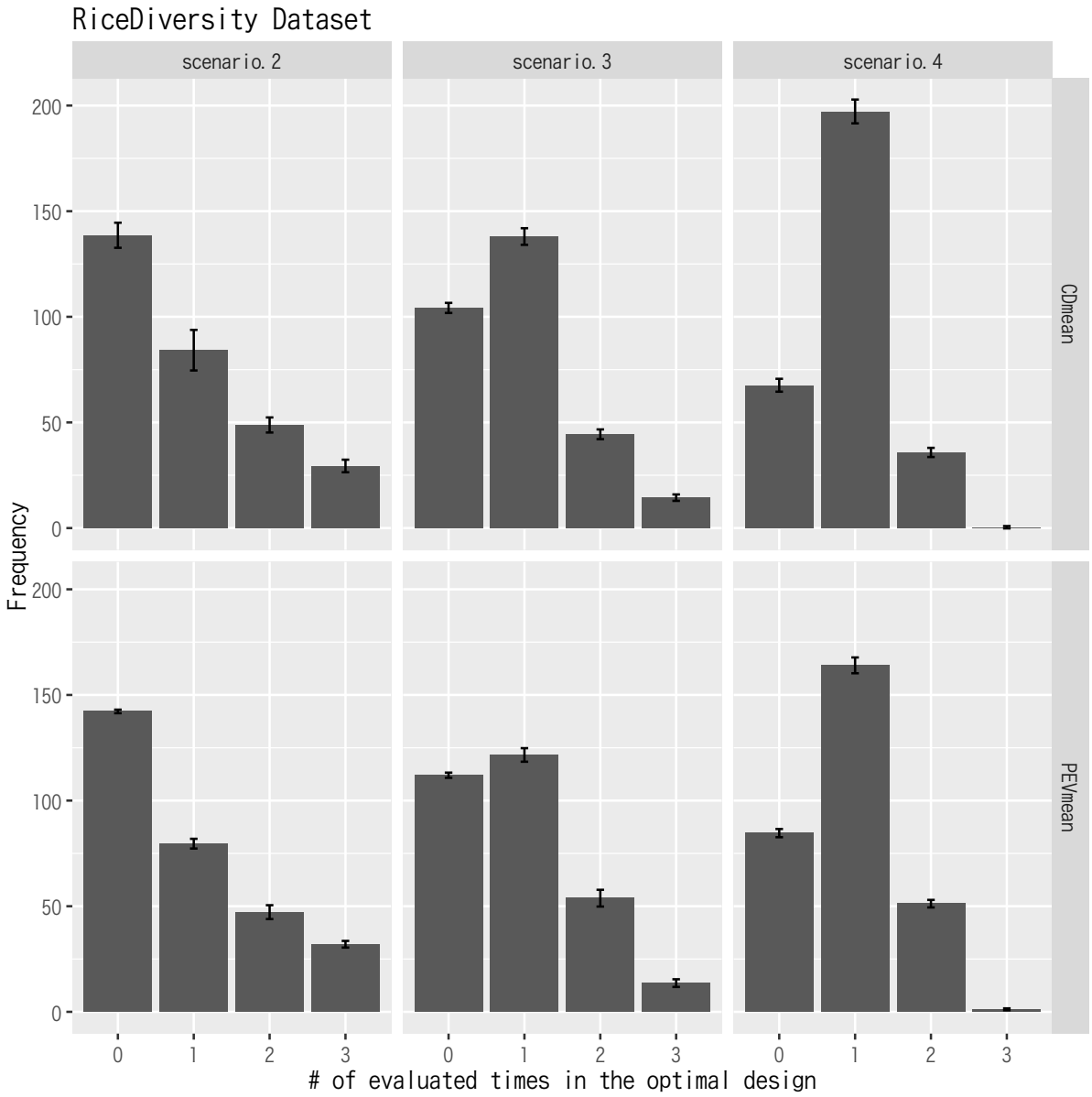


図 5-3. 環境間相関と最適化指標が最適デザインに及ぼす影響（イネ遺伝資源データ）

設定した遺伝子型値の環境間相関が異なる scenario.2 から scenario.4 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。環境間相関が低い場合には試験されない系統が多数ある一方で、環境間相関が高い場合には、より多くの系統を栽培試験するデザインが選ばれたことがわかる。

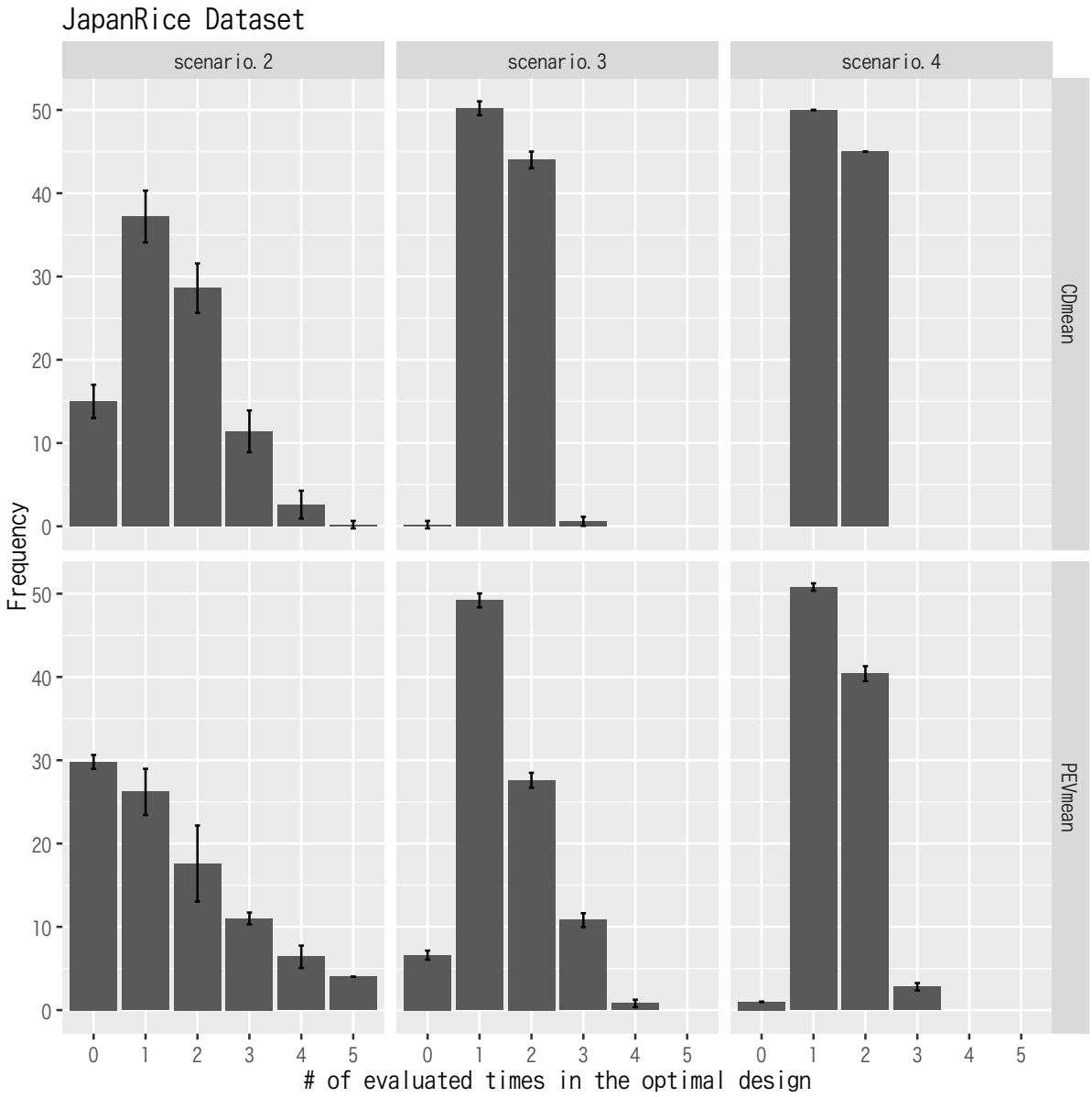


図 5-4. 環境間相関と最適化指標が最適デザインに及ぼす影響（日本水稲データ）

設定した遺伝子型値の環境間相関が異なる scenario.2 から scenario.4 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。環境間相関が低い場合には試験されない系統が多数ある一方で、環境間相関が高い場合には、ほぼ全ての系統が 1 回あるいは 2 回試験されるデザインが最適デザインとして選ばれたことがわかる。

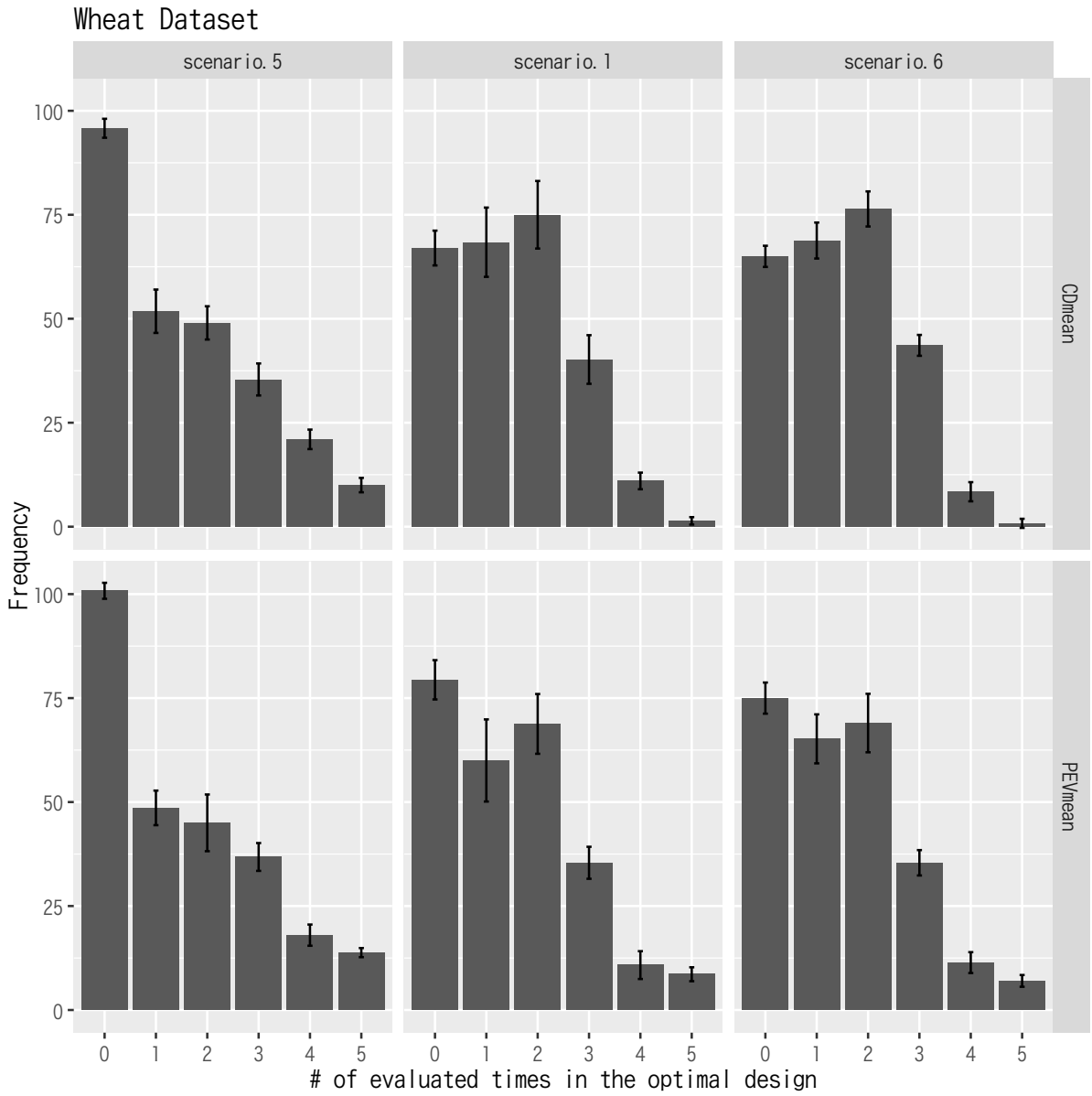


図 5-5. 遺伝率と最適化指標が最適デザインに及ぼす影響（コムギデータ）

設定した遺伝率の異なる scenario.1, scenario.5, および scenario.6 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。なお、図の左から右に向かって遺伝率が高くなるように図示している。図より、遺伝率に応じて最適デザインの傾向が異なり、特に遺伝率が低い scenario.5 においては他と大きく異なるデザインが導かれたことがわかる。

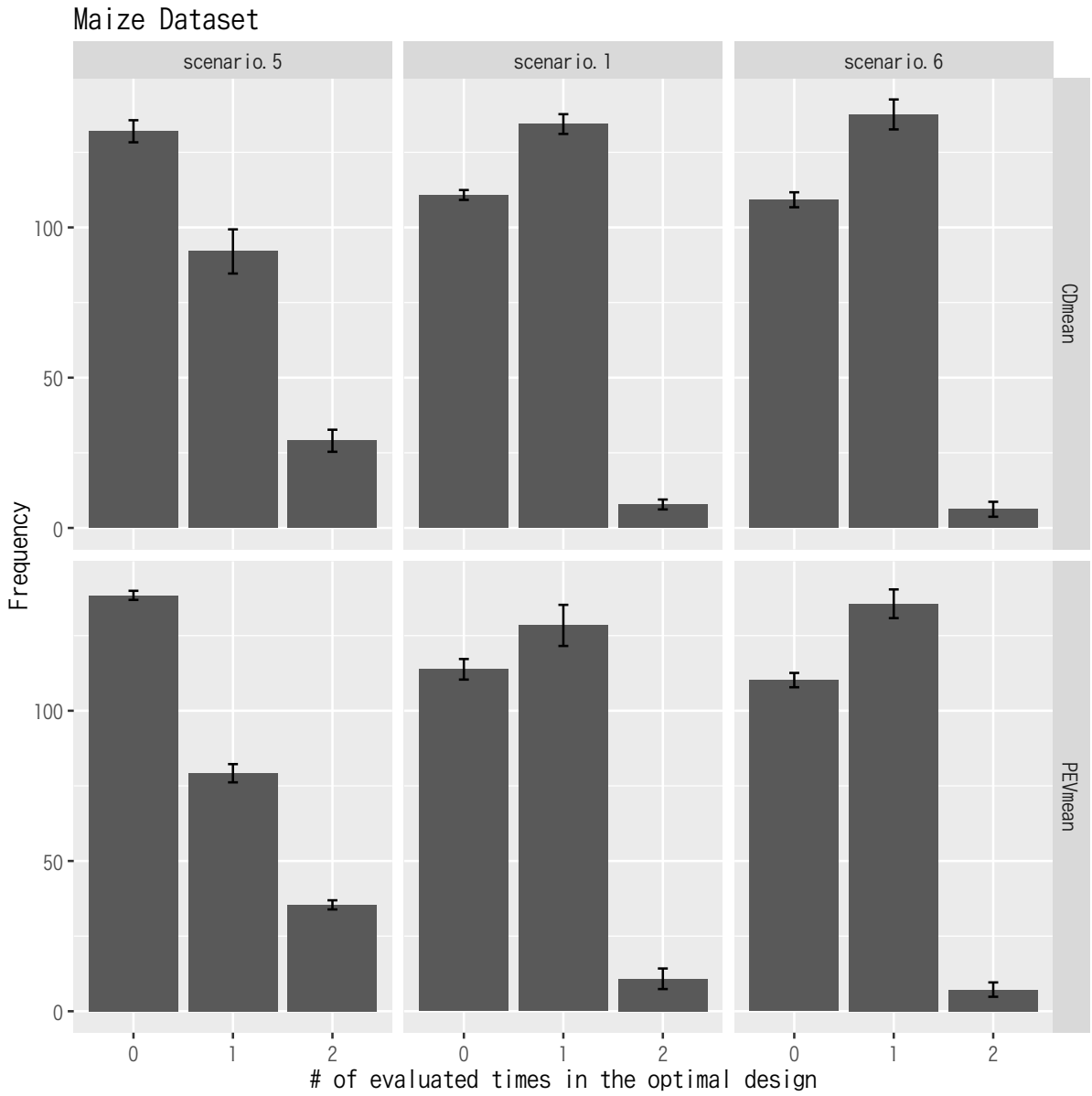


図 5-6. 遺伝率と最適化指標が最適デザインに及ぼす影響（トウモロコシデータ）

設定した遺伝率の異なる scenario.1, scenario.5, および scenario.6 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。なお、図の左から右に向かって遺伝率が高くなるように図示している。図より、遺伝率に応じて最適デザインの傾向が異なり、特に遺伝率が低い scenario.5 においては他と大きく異なるデザインが導かれたことがわかる。

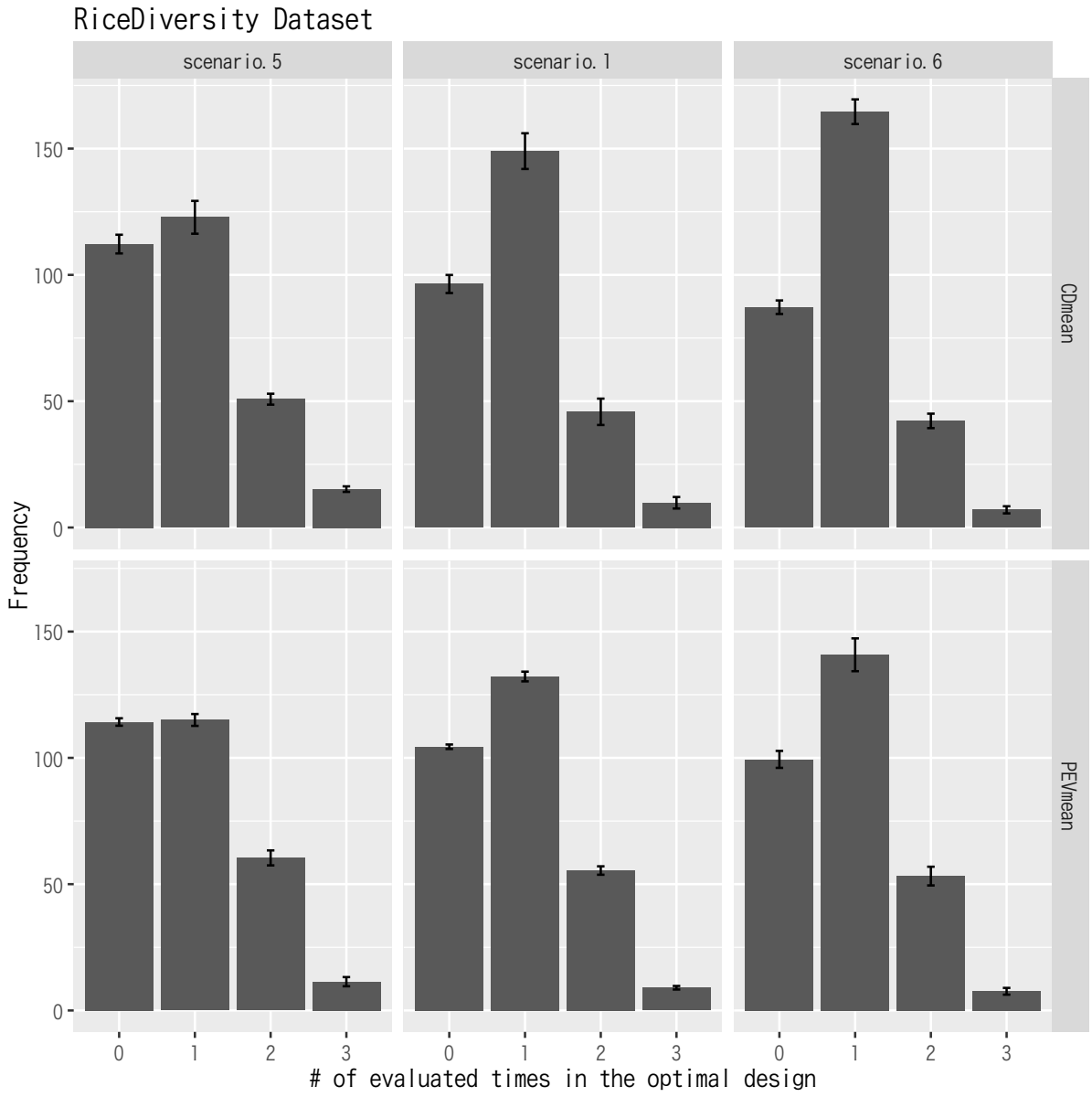


図 5-7. 遺伝率と最適化指標が最適デザインに及ぼす影響（イネ遺伝資源データ）

設定した遺伝率の異なる scenario.1, scenario.5, および scenario.6 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。なお、図の左から右に向かって遺伝率が高くなるように図示している。図より、遺伝率に応じて最適デザインの傾向が異なることがわかる。コムギデータやトウモロコシデータとは異なり、イネ遺伝資源データでは遺伝率が低い場合とそうでない場合に明瞭な違いがあるわけではなかった。

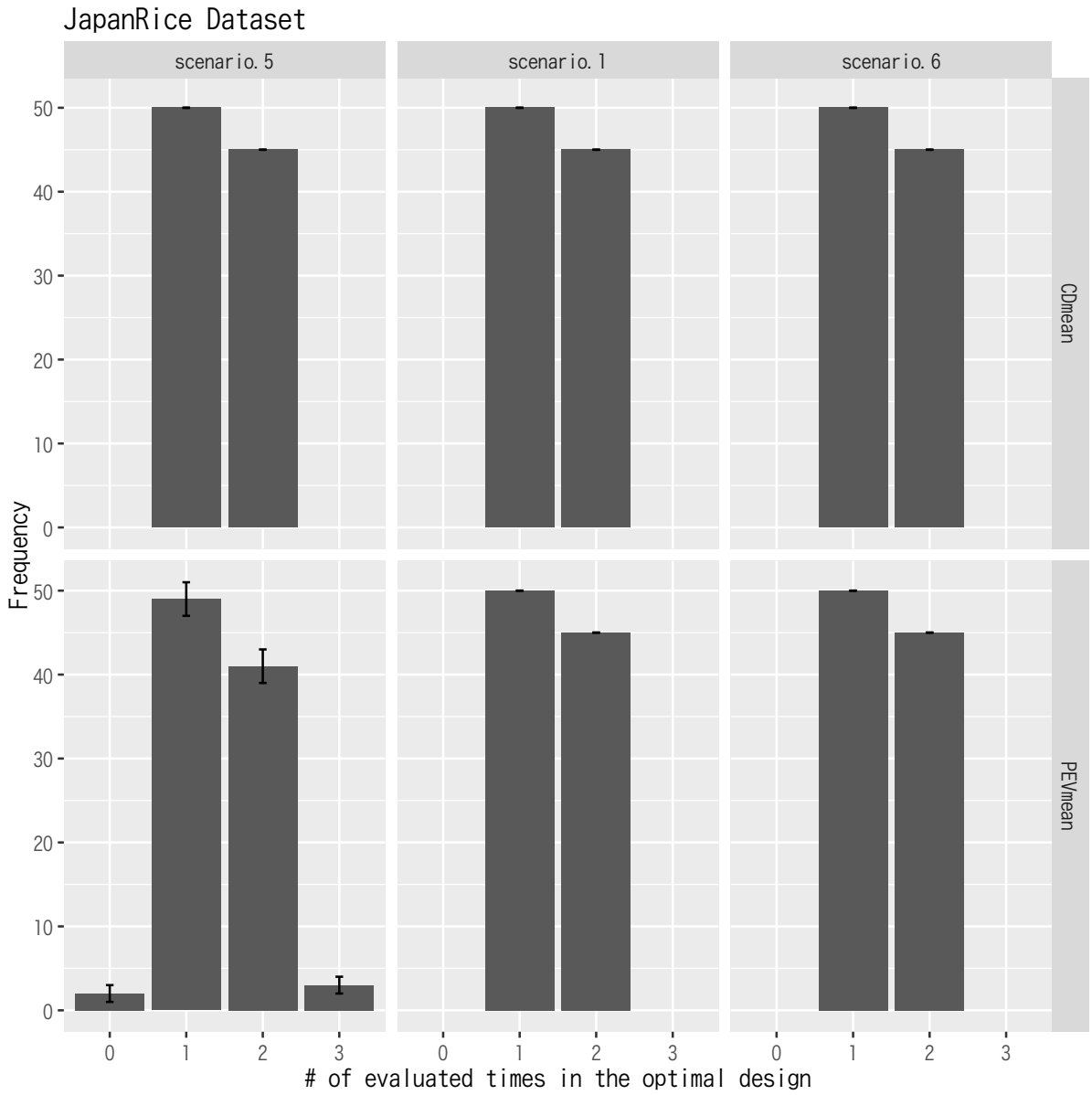


図 5-8. 遺伝率と最適化指標が最適デザインに及ぼす影響（日本水稲データ）

設定した遺伝率の異なる scenario.1, scenario.5, および scenario.6 について、それぞれの最適化指標が導く最適デザインの傾向を、最適デザインがそれぞれの系統を試験する回数を横軸にとり、縦軸にその頻度（系統数）をとった棒グラフで示した。なお、図の左から右に向かって遺伝率が高くなるように図示している。日本水稲データでは、遺伝率によるデザインの違いはほとんど見られなかった。

・　PEVmean および CDmean で高頻度に選ばれる系統

PEVmean や CDmean がどのような系統をより多く試験しているかを確かめるために、特定の系統に集中する傾向が最も強く現れる scenario.2 に注目して、3 回以上試験された系統を抽出し、マーカー遺伝子型に基づく第 1、第 2 主成分平面上に赤色のシンボルで図示した（図 5-9）。ここで、PEVmean および CDmean には異なる初期値から得られる 5 通りの最適デザインが存在するため、5 通りのうち何通りのデザインでその系統が 3 回以上選ばれたかによって、赤色シンボルの大きさを変えて図示した。シンボルが大きいほど、5 通りのうち多くで 3 回以上選ばれていることを意味する。なお、トウモロコシデータは 2 環境のデータセットであるため、ある系統が 3 回以上選ばれることはなく、また、2 回選ばれた系統は多数ある。したがってトウモロコシデータでは明瞭な傾向を読み取ることが困難であったことから、ここでは図から除外した。

図より、PEVmean と CDmean ともに、選ばれやすい系統は主成分平面の特定の領域に偏っているわけではなく、平面を幅広く網羅するように栽培系統を選んでいると考えられた。また、どのデータでも、主成分平面の「端」に位置する系統が選ばれていることも示唆された。各主成分の寄与率はデータセットごとに大きく異なるにも関わらず、得られた図の大局的な様子は類似しており、PEVmean と CDmean により得られる最適デザインの傾向は、遺伝的な集団構造の強弱には、あまり影響されていない可能性が示唆された。

・　遺伝的アルゴリズムの収束

遺伝的アルゴリズムによる PEVmean の最小化および CDmean の最大化について、その収束を、アルゴリズムの iteration ごとの PEVmean の最小値、または CDmean の最大値を図示することで確認した。ここでは、コムギデータにおける scenario.1 の場合の結果を、代表例として図 5-10 および図 5-11 に示した。

これらの図より、遺伝的アルゴリズムによって PEVmean と CDmean の最適化が適切に進行し、100 回の繰り返しによってほぼ plateau に達していることがわかる。しかし、90 回目以降の PEVmean および CDmean の変化を詳細に確認すると、依然としてアルゴリズムの繰り返しが進むにつれて PEVmean が減少（CDmean が増加）していた。すなわち、遺伝的アルゴリズムは、本研究の解析では収束に達していないことがわかった。表 5-2 から 5-5 において、PEVmean や CDmean を用いた場合にも予測精度の標準偏差（SD）は 0 ではなかったが、これは異なる局所解に陥っている可能性のみならず、遺伝的アルゴリズムによる解探索が十分に収束するまで繰り返されていないことが理由に挙げられる。

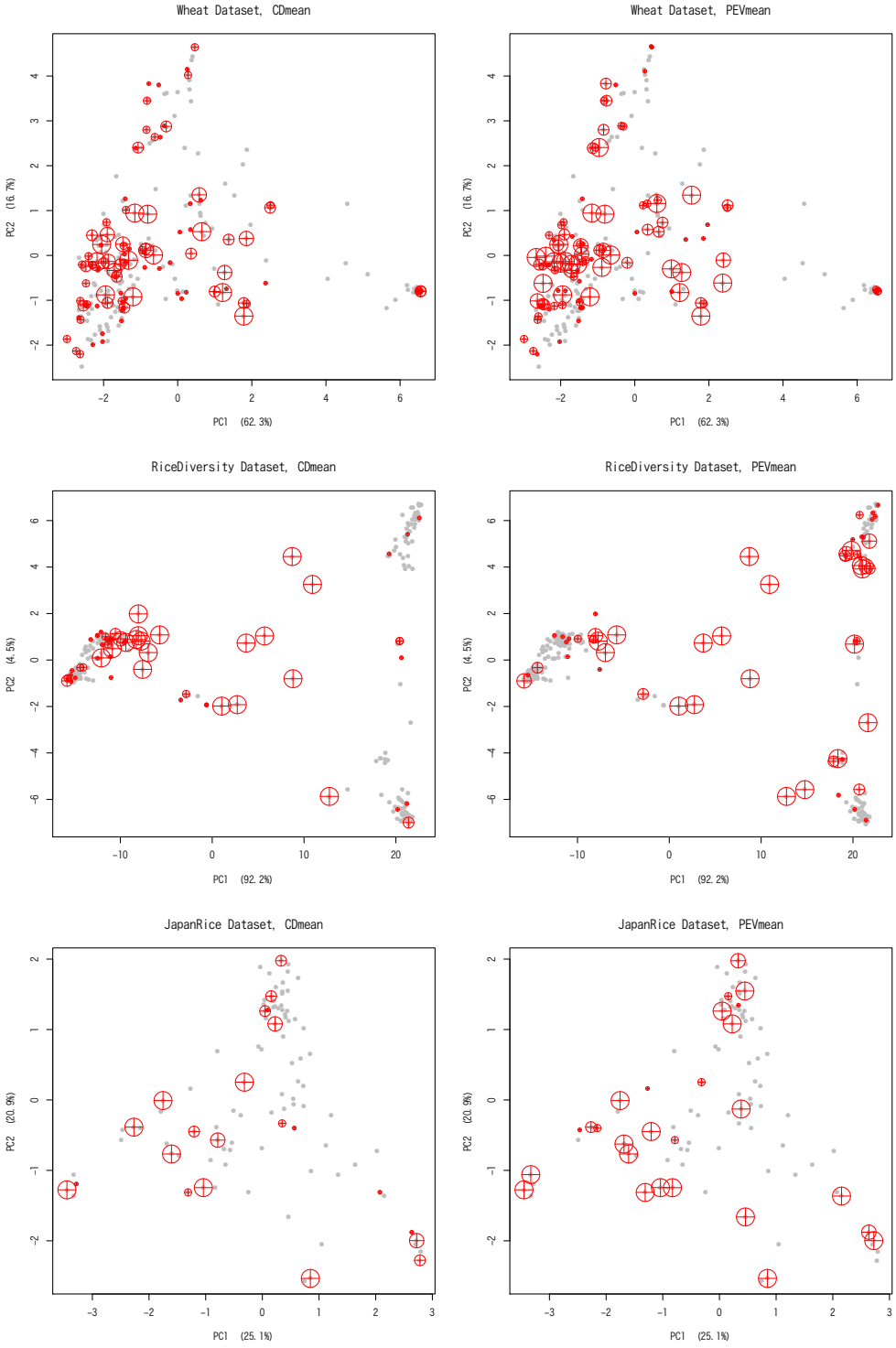


図 5-9. 最適デザインが高頻度に試験する系統

マーカー遺伝子型に基づく主成分平面に、最適デザインが高頻度で試験する系統を赤色のシンボルで図示した。詳細は本文を参照せよ。

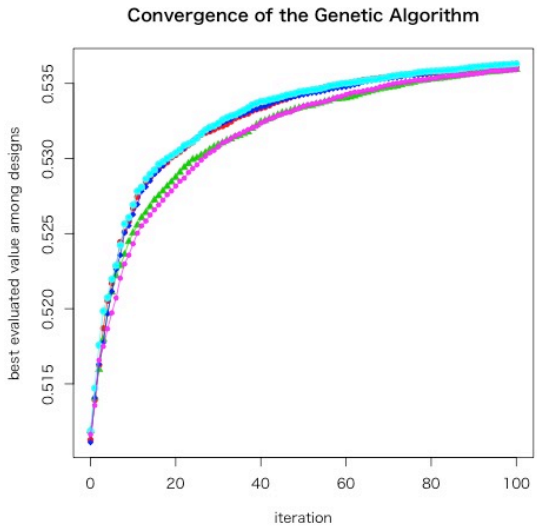


図 5-10. 遺伝的アルゴリズムによる CDmean の最大化

コムギデータの scenario.1 の場合について、遺伝的アルゴリズムによる CDmean の値の変化を図示した。異なる線は異なる乱数シードによる結果を表す。図より、遺伝的アルゴリズムによって CDmean は増加しているものの、100 回の繰り返しでは最適解に収束していないことがわかる。

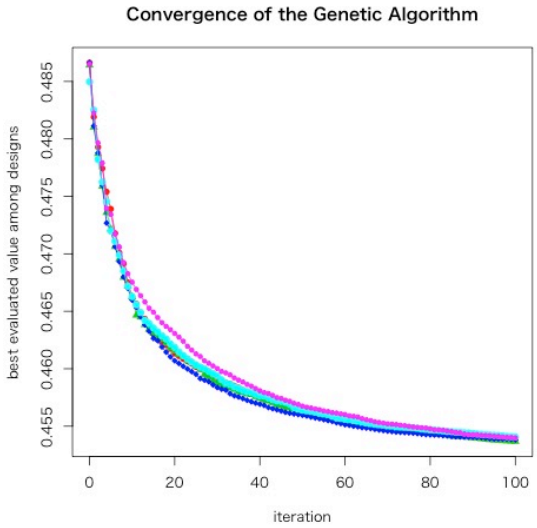


図 5-11. 遺伝的アルゴリズムによる PEVmean の最小化

コムギデータの scenario.1 の場合について、遺伝的アルゴリズムによる PEVmean の値の変化を図示した。異なる線は異なる乱数シードによる結果を表す。図より、遺伝的アルゴリズムによって PEVmean は減少しているものの、100 回の繰り返しでは最適解に収束していないことがわかる。

5-4. 考察

本研究では、訓練集団最適化を多環境ゲノミック予測モデルに拡張し、多環境試験デザインの最適化に PEVmean および CDmean を利用することを提案した。複数の実データによるシミュレーションの結果、PEVmean および CDmean に基づき試験する系統・環境を選ぶことで、無作為に多環境試験デザインを決めるよりも予測精度の高いモデルが得られる可能性が高いことが示唆された。また、PEVmean と CDmean を多環境ゲノミック予測モデルに拡張した場合には、計算に必要な超パラメータが複雑化してしまうことを示し、遺伝率と遺伝子型値の環境間相関の 2 つを超パラメータとして設定することを提案した。このとき、GxE を考慮しない通常の訓練集団最適化とは異なり、そうした超パラメータ、とりわけ遺伝子型値の環境間相関行列 **S** の設定によって、得られる最適デザインの性質が大きく異なることも明らかになった。いっぽう、PEVmean と CDmean による予測精度の差は顕著ではなく、本研究の範囲では、いずれの指標が優れているかについて明確な結論を得ることは困難であった。

遺伝子型値の環境間相関行列 **S** を表現型相関によってほぼ正しい値に設定した場合には、PEVmean や CDmean による最適デザインを用いることで、イネ遺伝資源データ以外のデータセットで、遺伝子型値の予測精度（完全データに基づく推定値と、実測データに基づく推定値の相関係数）が有意に向上した。イネ遺伝資源データでは有意な差は見られなかったが、予測精度の平均値は減少したわけではなく、遺伝的アルゴリズムを改善して最適デザインをより適切に探索できれば、他のデータセットと同様の結果が得られる可能性が高いと考えられる。ただし、PEVmean が小さい（または CDmean が大きい）ことは、必ずしも予測精度が高いことを保証するものではない（Yu et al., 2018）ため、大域的最適解に収束しても、その実験デザインが予測精度を最大化するとは限らないことに注意する必要がある。

PEVmean や CDmean は、与えられた遺伝子型値の環境間相関の強弱に応じて、最適デザインを柔軟に変化させた（図 5-1 から図 5-4）。環境間相関が弱い場合（scenario.2）には、栽培されない系統を増やしてでも、一部の系統を何回も栽培試験するデザインが選ばれた。いっぽう、環境間相関が強い場合 (scenario.4) には、ほぼ全ての系統を満遍なく試験する最適デザインが選ばれた。この結果は極めて妥当なものである。いま、ある環境 i におけるある系統 j の遺伝子型値 u_{ij} を予測したいとする。このとき、当然ながら最も有益な情報は、その系統のその環境における表現型値 y_{ij} に他ならない。では、それに次いで有益な情報とは何であろうか。もし、遺伝子型値の環境間相関が強いのであれば、別の環境 k における同一系統の表現型値 y_{ik} を用いても u_{ij} の予測ができる。しかし、遺伝子型値の環境間相関が弱く、環境が変われば遺伝子型値も大きく異なる（つまり GxE が支配的である）場合には、y_{ik} は u_{ij} の推定にほとんど貢献しない。このような場合には、むしろ、遺伝的関係の強い系統 l (SNP マーカー遺伝子型による関係行列 **G** の i, l 成分の絶対値が大きい系統）の同一環境における表現型 y_{lj} を用いてゲノミック予測を行うことで u_{ij} を予測することが有利だと考えられる。このような定性的考察から、環境間相関が強い場合には、どの系統も少なくとも 1 回以上栽培試験を行ったほうがよく、環境間相関が低い場合には、集団を代表する系統をどの環境でも栽培しておき、それぞれの環境でゲノミック予測を活用するほうがよいことが示唆される。PEVmean と CDmean は、与えられた遺伝子型値の環境間相関 (行列 **S**) と、集団の遺伝的関係 (ゲノム関係行列 **G**) とを総合的に判断しながら、系統数に関する制約の下で、

できるだけ有益な情報を得られるように、多環境試験のデザインを最適化していると考えられる。

遺伝率もまた多環境試験デザインに影響を与えた（図 5-5 から図 5-8）。1 つの環境だけを考える訓練集団最適化では、遺伝率をどのように設定しても得られる最適デザインはほとんど変わらなかった（e.g. Akdemir et al., 2015）が、多環境試験デザインへの応用では、遺伝率もまた慎重に指定しなければならないパラメータであることが示唆された。通常の訓練集団最適化については、1 つの環境だけを考えれば、遺伝率をどのように設定しても系統間の相対的な重要性は変わらないため、選ばれる系統は同じになると理解することができる。多環境試験について考えると、遺伝率が低いときには、遺伝子型の環境間相関（行列 **S**）に比べて、相対的に系統間相関（ゲノム関係行列 **G**）の重要性が高まると推察される。なぜならば、遺伝率の低下は環境間の遺伝共分散を低下させ、相対的に、予測における系統間の相関関係の重要性を高めるためである。したがって、遺伝率が低い場合には、集団中の代表的な系統をどの環境でも測定し、ゲノム関係行列を用いてそれぞれの環境でゲノミック予測を行うことが予測精度の向上に繋がると思われる。

このように、遺伝子型値の環境間相関や遺伝率といった超パラメータを適切に与えることの重要性が明らかになったが、当然ながら、妥当な値を設定するには何らかの事前情報が不可欠である。最も簡単な方法は、育種家の事前知識によって決めることであろう。多くの場合、育種家は対象の形質や環境に関する何らかの知識や経験、感覚を持っているため、例えば本研究で設定した複数のパラメータのどれが尤もらしいかを選ぶことができる。本研究のシミュレーションによれば、超パラメータの設定は最適デザインに大きく影響するものの、良いパラメータと劇的に異なるパラメータを設定する（e.g. 日本水稲データ、scenario.2）ような誤りを避けられれば、PEVmean や CDmean によって得られるデザインは、無作為に栽培系統を選ぶ方法に劣ることはないようである。よって、育種家の経験によってパラメータを設定しても、大きな問題は生じないと考えられる。

あるいは、過去に収集された表現型データがあれば、それを用いてパラメータを決めることも考えられる。いま注目している集団と概ね遺伝的構成が似通った系統群について、概ね同じ環境での栽培記録があれば、その表現型相関や遺伝率を参考にすることで、妥当な遺伝子型値の環境間相関を決めることが可能であろう。この推定が必ずしも正しい保証はないが、それでも大きな問題にはならないことは上述の通りであろう。

本研究では、PEVmean と CDmean を多環境ゲノミック予測モデルに拡張するにあたり、元の定義をできるだけ保存するようにした。しかし、これ以外の拡張を議論することも興味深い話題である。PEVmean や CDmean は、遺伝子型値 **u** の推定に注目した指標になっており、環境効果 **β** の推定については（少なくとも明示的には）考慮していない。もし、評価方法 2 で想定したように、多環境ゲノミック予測の目的が遺伝子型値の推定だけでなく表現型値の予測にもあるとすれば、環境効果も正しく推定し、最終的な表現型の予測ができるだけ精度よくできることが望ましい。おそらく、PEVmean や CDmean を適切に修正することで、このような目的を想定した最適化指標を得ることが可能であろう。表 5-2 と表 5-4、あるいは表 5-3 と表 5-5 を比較すると、特に PEVmean の場合に、方法 2（表現型値ベースでの評価）より方法 1（遺伝子型値ベースでの評価）による $-\log_{10}$P 値が大きい傾向にあった。これは、そもそも PEVmean や CDmean が、遺

伝子型値の推定を想定して構築された指標であることと関係しているかもしれない。

PEVmean や CDmean による訓練集団最適化は、より一般の統計的推論・機械学習で研究されているおける最適実験計画（OED; optimal experimental design）と、ほとんど同じ問題設定であると考えられ、実際に Akdemir らはそれを論文中で指摘している（Akdemir et al., 2015）。特に興味深いのはベイズ推論へと OED を拡張した Bayesian OED と呼ばれる研究分野であり、そこでは興味のあるパラメータについて、事前分布と事後分布の Kulbuck-Leiblier 情報量を最大化する方法などが最適化指標として用いられているようである（Chaloner and Verdinelli, 1995）。Bayesian OED の特長は、ベイズ推論の強みである事後分布の逐次更新により、過去のデータによる推論結果を自然な形で反映できることにある。これは sequential design と呼ばれる問題設定である（e.g. Pauwels et al., 2014）。つまり、一部の測定結果（e.g. 昨年までの試験結果）が既知の場合に、どの系統をどの環境で追加試験するべきか、という問題を自然に扱うことができ、これは応用上非常に重要である。PEV や CD は、ゲノミック予測が提案されるより前から、頻度論的な発想によって定義され、その後に PEVmean や CDmean としてゲノミック予測に応用された。本研究では先行研究に倣ってこれらの指標を用いたが、ゲノミック予測はベイズ推論の枠組みで考えることが主流であり、Bayesian OED の応用は十分に可能であると予想される。

本研究における問題点の 1 つに、遺伝的アルゴリズムが収束していなかったことが挙げられる（図 5-10, 5-11）。Akdemir らの論文における遺伝的アルゴリズムの設定は、個体数 800（本研究と同じ）、選抜個体数 5（本研究では 10）、突然変異率 0.5（本研究と同じ）、繰り返し数 300（本研究では 100）であり、違いは選抜個体数と繰り返し数にある（Akdemir et al., 2015）。一般に、選抜個体数は多いほど収束までにかかる繰り返し数は増加するが、局所解に陥る可能性は低くなる。実際にはこの選抜個体数（および個体数、突然変異率）について複数の場合を試し、図 5-6 のようなグラフを目視で確認して、局所解に陥らない程度に選抜個体数を小さくすることが望ましいだろう。繰り返し数は当然ながら多い方が望ましいが、繰り返し数に単純に比例して計算時間が増大する。

このアルゴリズムの収束に関連して、PEVmean や CDmean の計算コストを、本研究において無視できない課題の 1 つとして挙げておく。本研究で実際に用いた PEVmean および CDmean の実装では、その計算に $L \times M$ 次元正方行列の逆行列演算を含んでおり（この実装は本文における数式とは異なる）、それが計算のボトルネックとなっていた。2018 年 11 月 25 日現在、AWS（Amazon Web Service）の m5.12xlarge インスタンス（vCPU=48 コア、メモリ 192GB）を用いて 40 コアによる並列計算を行なった場合でも、100 回（初期集団を含めて 101 回×800 通りの行列演算）の繰り返しに約 77 分の計算時間を要した。本研究で繰り返し数を少なく設定したのは、単にこの計算時間の長さによる。なお、Akdemir らは、PEVmean の定義をやや修正し、かつ、特異値分解に基づく低次元近似を用いることで計算コストを大幅に削減していた（Akdemir et al., 2015）が、その場合には contrast vector を自由にデザインすることは難しいと考えられる。どのような contrast vector についても高速に PEVmean や CDmean を計算する（近似）計算法の開発は重要な課題である。

多環境試験データに基づき植物の環境応答を解析する方法として、作物モデルの利用も頻繁に

行われる。近年では、ゲノミック予測と作物モデルを組み合わせることで、未知の環境における未試験系統の表現型を予測しようとする試み（Technow et al., 2015; Onogi et al., 2016）が行われるとともに、その場合にどのような系統を栽培試験するべきかを最適化する研究も行われている（Rincent et al., 2017b）。本研究との（方法論の意味での）関連性は明らかではないが、このようなアプローチとの融合も検討されるべき課題の 1 つであることは間違いないだろう。

本研究は PEVmean および CDmean を多環境試験デザインの最適化へ適用した最初の例である。PEVmean と CDmean の拡張は簡単であり、その有効性が示唆されるとともに、超パラメータに応じて合理的な最適デザインが選ばれていることが明らかとなった。いっぽう、超パラメータの設定方法や最適化指標の計算効率などが課題として残った。特に超パラメータの設定については、より具体的な状況を踏まえて議論する必要がある。例えば、適当な数系統について過去の実験データがある場合について、どのように翌年の多環境試験をデザインするか、といった問題は実用上重要であろう。この際、本研究では考慮できなかった Bayesian OED などの手法を取り入れることも検討されるべきだと考える。

6. ゲノミック予測に基づく交配後代の分離予測に関するシミュレーション研究

6-1. 序論

ゲノミック予測の最も単純な活用法は、まず適当な集団や系統群を訓練集団として予測モデルを構築し、ある別の集団や系統の遺伝子型値をマーカー遺伝子型と予測モデルに基づき予測することで、表現型を測定する手間を省くことである。例えば、栽培試験により多数の系統を評価することが難しい果樹育種を考えると、多数の実生をシーケンスすれば、その段階で予測モデルによって選抜を行うことが可能になる。これにより選抜候補となる個体数が劇的に増加し、優れた個体だけを圃場評価できるため、遺伝的獲得量が大きくなると考えられる。このように、ゲノミック予測はある「個体」のもつゲノム情報から、その個体の遺伝子型値を予測する方法である。したがって、基本的には個体や系統レベルでの選抜に利用することが想定される。

しかし、遺伝は確率的なプロセスとして近似することができるため、両親の DNA 配列（正しくハプロタイプが推定されたマーカー遺伝子型）と DNA マーカー間の組替え価が分かっていれば、交配により得られる後代個体のマーカー遺伝子型をシミュレーションすることができる (e.g. Iwata et al., 2013)。両親が純系である場合にはより単純であり、全ての塩基対がホモ接合型であるためハプロタイプの推定は不要となり、マーカー間の組替え価が分かっていれば、分離集団のマーカー遺伝子型をシミュレーションすることが可能である。例えば R の {qtl} パッケージには、RIL 集団の遺伝子型を生成する機能が実装されている（Broman et al., 2003)。

後代マーカー遺伝子型のシミュレーションをゲノミック予測と組み合わせることで、交配後代集団における遺伝子型値の分離（segregation）を予測することができる。つまり、ある両親から得られる後代個体のマーカー遺伝子型を多数シミュレーションし、それにゲノミック予測モデルを当てはめれば、それは後代個体の遺伝子型値の仮想サンプルとみなすことができる。よって、例えばある両親を交配して得られる後代集団の遺伝子型値の分布（平均値や分散）を計算することができる。

分離を考慮するべき重要な例の 1 つは、イネをはじめとする純系系統を系統選抜法によって育種する場合である。このとき、２系交雑集団を作出するための両親系統は、自殖により遺伝子型の固定が進んだ F_5, F_6 世代における遺伝子型値の最大値が高くなるように選ばれるべきである。後代遺伝子型値が正規分布に従うとすると、分布から抽出された標本最大値（つまり、展開した後代個体の遺伝子型値の最大値）の期待値は平均と分散の増加関数であるから、後代遺伝子型値の平均だけでなく、分散が大きいことが望ましい。このような立場から、ある両親の交配と自殖の繰り返しによって得られる RIL 集団の遺伝子型をシミュレーションによって生成し、遺伝子型値の分離をゲノミック予測に基づき予測する研究が、過去に何例か報告されている (Mohammadi et al., 2015; Tiede et al., 2015; Lado et al., 2017)。

分離の予測が重要なのは、決して純系品種の育種に限ったことではない。Iwata らは、果樹育種においてゲノミック予測とシミュレーションによる後代分離の予測を行った（Iwata et al., 2013)。果樹の場合には接ぎ木などの栄養生殖が可能であることなどから、純系の親を出発点に

RIL 集団を作るのではなく、ヘテロな遺伝子型をもつ親品種同士を交配して得られる F_1 集団において、１つの優良個体を得ることが重要である。この場合にも、F_1 集団の遺伝子型値の最大値を高めること、あるいは Iwata らの研究で注目されているように、ある閾値を超える後代を多数生じること（e.g. 既存品種より早生であり、かつ果実重が重いこと）が重要となる。いずれの目的においても、後代の遺伝子型値の平均だけでなく、分散も予測しなければならない。

高精度な分離予測は、長期的な育種戦略の最適化においても不可欠である。複数回の選抜と交配を繰り返す循環選抜を考えた場合、初期の選抜サイクルにおいて強すぎる選抜をかけると、後代における遺伝分散が著しく減少し、その後の改良が進みづらくなるというトレードオフが存在する。したがって、特に初期の選抜では、後代の分散が低下しすぎないような交配計画を立案する必要がある。このような背景から、ゲノミック予測に基づく選抜・交配計画を最適化しようという試みが活発化しているが、そこでも後代の遺伝分散、つまり分離の予測に基づく交配計画の策定が行われる（Lehermeier et al., 2017; Müller et al., 2018）。

以上のように、後代集団における遺伝子型値の分離をゲノミック予測に基づき予測することは、交配組み合わせの選定や長期的な育種戦略の最適化において不可欠である。しかしながら、後代分離の予測精度については、依然として十分な検討が行われていないのが現状である。

まず、多くの先行研究において、後代分離の計算方法が厳密な意味では正しくない。ゲノミック予測において推定されるマーカー効果には、予測の不確実性が存在する。ベイズ推論の表現を借りれば、マーカー効果は事後分布に従う確率変数として扱う必要がある。しかし、ほとんどの先行研究はマーカー効果の点推定量を用いて後代の分離を推定しており（Mohammadi et al., 2015; Tiede et al., 2015; Lado et al., 2017）、マーカー効果の不確実性を考慮して計算を行っている例はごくわずかである（Iwata et al., 2013; Zhong and Jannink, 2007）。

次に、ゲノミック予測において提案された複数のモデルについて、後代分離の予測精度を比較した例が非常に少ない。マーカー効果の不確実性が考慮された２つの研究では、いずれも BayesA に相当する手法のみを用いている（Iwata et al., 2013; Zhong and Jannink, 2007）。いっぽう、RIL 集団に関するほとんどの研究はベイズリッジ回帰を採用しており（Mohammadi et al., 2015; Tiede et al., 2015; Lado et al., 2017）、やはり他の予測手法については検討されていない。

昨夏、ようやく後代の分離予測（RIL 集団における分離予測）に関するモデル比較を行った論文が発表され、目的形質を支配する QTL 数や遺伝率の異なる仮想形質を用いた詳細な検討がなされた（Yao et al., 2018）。しかしながら、残念ながらこの論文でもマーカー効果の点推定量だけを用いて計算を行っていた。また、この研究では、後代分散の推定精度が、真の分散と推定された分散との間の相関係数によって評価された。相関係数による評価では、分散が一律に過小・過大予測されていたとしても精度評価に影響を及ぼさないことが問題だと思われる。ゲノミック予測において相関係数が評価に用いられるのは、個体選抜においては個体間の相対比較だけが重要であるという暗黙の了解に基づくものである。分離予測の場合、交配組み合わせ間で分散を直接比較するよりは、推定された分散と平均を適切に組み合わせた指標を計算し、それを交配組み合わせの評価に用いることが多い（e.g. Lehermeier et al., 2017）。よって、分散の予測精度は真値との絶対誤差を考慮して行うべきだと思われる。

以上の状況を鑑み、本研究では、ゲノミック予測に基づく後代分離の予測について独自に検証を行った。ここでは、ヘテロな親集団において予測モデルを構築し、その集団から交配組み合わせを選び、F_1 分離集団の予測を行う状況を想定した。予測モデルとして、ベイズリッジ回帰と BayesA という性質の異なる２つのマーカー回帰を用い、分離の予測精度を比較した。また、形質に寄与する遺伝率、回帰に用いられる SNP マーカーの数についてもそれぞれ２通りの条件を設定し、これらの条件によって分離の予測精度がどのように変わるかについても検証した。

6-2. 材料・方法

6-2-1. 後代分離の計算方法

本節では、後代の分離をシミュレーションによって計算する方法について説明する。まず、両親から子へとマーカー遺伝子型 **x** が受け継がれる際の確率的過程を定義する。次に、マーカー回帰について簡単に説明する。最後に、それら２つを用いて後代の分離を計算する方法を示す。

いま、DNA が二本鎖であることを考えると、ある系統のマーカー遺伝子型ベクトル **x** は、２つのハプロタイプベクトル **h1**, **h2** の和として

$$\mathbf{x} = \mathbf{h1} + \mathbf{h2} \tag{6.1}$$

このように表現できる。それぞれのベクトルはマーカー数（P とする）の次元を持つ。ハプロタイプベクトル **h1**, **h2** の要素は-1/2 または 1/2 であり、それぞれ遺伝子型が A または B であることに対応する（2-1-1 節も参照せよ）。いま、ハプロタイプは完璧に推定できる、すなわち、式(6.1)における左辺から右辺への分解が、いかなる **x** についても正しく実行可能であると仮定する。

後代個体のマーカー遺伝子型 **x** について、両親のマーカー遺伝子型が $\mathbf{x}_k$ と $\mathbf{x}_l$ であるとする。後代個体のマーカー遺伝子型 **x** は、両親のマーカー遺伝子型、および、マーカー間の組換え価 **θ** に基づき確率的に定まると仮定する。つまり、条件付き分布

$$p(\mathbf{x}|\mathbf{x}_k, \mathbf{x}_l, \boldsymbol{\theta}) \tag{6.2}$$

が存在する。以下ではこの条件付き分布について解説するが、式(6.2)を書き下すことはせず、条件付き分布から **x** を生成する手続きを順に述べる。

マーカー遺伝子型が後代に受け継がれるときには、減数分裂によって、ある親の持つ２つのハプロタイプ **h1**, **h2** から新たなハプロタイプ **h** が確率的に生じる。最も単純化された減数分裂の過程は、組替え価に基づくマルコフ連鎖として表現される。まず、二値変数で表現される状態 s_j を定義する。ここで、添字 $j = 1, 2, ..., P$ はマーカーの位置を表す。この状態 s_j について、$s_j = 1$ のときハプロタイプ **h1** の要素を新たなハプロタイプ **h** の要素とし、$s_j = 2$ のときハプロタイプ **h2** の要素を新たなハプロタイプ **h** の要素とする。すなわち

$$p(h_j = h1_j | s_j = 1) = p(h_j = h2_j | s_j = 2) = 1 \tag{6.3}$$

とする。

状態 s_j は、組換え価によって遷移確率が定義されるマルコフ連鎖に従う。すなわち

$$p(s_{j+1} = 2 | s_j = 1) = p(s_{j+1} = 1 | s_j = 2) = \theta_{j,j+1} \tag{6.4}$$

$$p(s_{j+1} = 1 | s_j = 1) = p(s_{j+1} = 2 | s_j = 2) = 1 - \theta_{j,j+1} \tag{6.5}$$

である。なお、初期状態 s_1 は確率 1/2 で乱択される。

以上の式を組み合わせることで、1 つの親系統から新規ハプロタイプを生成することができる。まず、ある親のマーカー遺伝子型を式(6.2)によって 2 つのハプロタイプに分割し、式(6.3) , (6.4), (6.5)に基づき新規ハプロタイプを生成する。これを両親について実行することで 2 つの新規ハプロタイプが得られる。最後に再び式(6.1)により新規ハプロタイプの和を取ることで、子のマーカー遺伝子型 $\mathbf{x}_{kl}$ を得ることができる。条件付き確率分布(6.2)は、これら一連の手続きによって定義される。標準的な伝承サンプリングの技術を用いることにより、条件付き確率分布(6.2)に従う T 個（$t = 1, 2, ..., T$）のサンプル

$$\mathbf{x}^{[t]} \sim p(\mathbf{x}|\mathbf{x}_k, \mathbf{x}_l, \boldsymbol{\theta}) \tag{6.6}$$

を得ることは容易である。

なお、以上の過程は、現在の生物学的な知見を全て反映したものではなく、あくまで単純化されたモデルであることを強調しておく。例えば、突然変異や多重乗り換えの抑制といった比較的よく知られた現象も、ここでは考慮されていない。

また、ハプロタイプ **h1**, **h2** や組替え価については、実際に真の値を観測することは難しい量である。しかし、観測されたマーカー遺伝子型をはじめとする諸情報に基づき、ハプロタイプの推定を行うことが可能なことも多い（e.g. Browing and Browing, 2007）。また、組換え価については、主要な作物では連鎖地図が作成されており、DNA マーカー間の遺伝距離が分かっていることも多い。その場合には、組換え価を Haldane の式や Kosambi の式によって推定することができる。ハプロタイプや組換え価の推定は単純ではなく、実際にはここにも誤りや不確実性が存在するが、本研究ではこれらの推定は全て確からしいと仮定し、これ以上は議論しないことにする。

次に、マーカー効果の推定について説明する。本研究では、遺伝子型値はマーカー回帰モデルによって表現されると仮定する。すなわち、ある系統 i の遺伝子型値 u_i ($i = 1,2,..., N$)について

$$u_i = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i \tag{6.7}$$

が成り立つとする。ここで、$\boldsymbol{\alpha}$ はマーカー効果の P 次元ベクトル、$\mathbf{x}_i$ は系統 i のマーカー遺伝子型ベクトルである。実際には優性効果やエピスタシスなどの存在により、このモデルが厳密に正しいことはほとんどないが、それを無視する。

本章で用いるベイズリッジ回帰（BRR; Bayesian ridge regression）、BayesA のいずれのモデルも、マーカー効果に（階層的な）事前分布を定める。BRR における事前分布は

$$p(\boldsymbol{\alpha}|\nu_\alpha, \tau_\alpha) = N(\boldsymbol{\alpha}|\mathbf{0}, \mathbf{I}\sigma_\alpha^2) \cdot \chi^{-2}(\sigma_\alpha^2|\nu_\alpha, \tau_\alpha) \tag{6.8}$$

であり、BayesA における事前分布は

$$p(\boldsymbol{\alpha}|\nu_\alpha,\tau_\alpha)=\prod_{j=1}^{P}N\left(\alpha_j|0,\sigma^2_{\alpha_j}\right)\cdot\chi^{-2}\left(\sigma^2_{\alpha_j}|\nu_\alpha,\tau_\alpha\right) \tag{6.9}$$

である（Perez and de los Campos, 2014）。BRR ではマーカー効果の周辺分布が正規分布であるのに対し、BayesA では t 分布である（Gianola, 2013）。したがって、BRR ではマーカー効果が強く制約され、全てのマーカー効果が 0 に近い値をとるが、BayesA では一部のマーカー効果が大きな値をとることが許容される。

単純化のため、表現型が系統ごとに 1 つの値だけをとることを仮定するとともに、1 つの環境だけでモデル化を行うことを仮定する。このとき、マーカー回帰の尤度関数は

$$p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2_e)=N(\mathbf{y}|\mathbf{1}\mu+\mathbf{X}\boldsymbol{\alpha},\mathbf{I}\sigma^2_e) \tag{6.10}$$

と定めることが普通である。ただし、$\mathbf{y}$ は表現型値の N 次元ベクトル、$\mathbf{X}$ は $N\times P$ 次元のマーカー遺伝子型行列、μ は表現型値の平均値である。ここで、残差分散 σ^2_e については、事前分布を

$$p(\sigma^2_e|\nu_e,\tau_e)=\chi^{-2}(\sigma^2_e|\nu_e,\tau_e) \tag{6.11}$$

と定める。

このとき、R のパッケージ {BGLR} を用いることで、BRR、BayesA のいずれのモデルに対しても、マーカー効果の事後分布に従う Z 個（$z=1,2,...,Z$）のサンプル

$$\boldsymbol{\alpha}^{[z]}\sim p(\boldsymbol{\alpha}|\mathbf{X},\mathbf{y})\propto p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2_e)\cdot p(\boldsymbol{\alpha}|\nu_\alpha,\tau_\alpha)\cdot p(\sigma^2_e|\nu_e,\tau_e) \tag{6.12}$$

を、Gibbs sampling に基づき得ることが可能である（Perez and de los Campos, 2014）。なお、Gibbs sampling によって事後分布からのサンプルを得るには適切に burn-in を定める必要があるが、本研究では経験則に基づき、MCMC の chain 数を 12,000、burn-in サンプル数を 2,000 と定めた。

仮想後代マーカー遺伝子型のサンプル $\mathbf{x}^{[t]}$ とマーカー遺伝子型の事後分布からのサンプル $\boldsymbol{\alpha}^{[z]}$ を用いることで、後代遺伝子型値の平均と分散の推定量を得ることができる。いま、後代平均の推定量を m とし、分散の推定量を v とすると、これらの値は

$$m=\frac{1}{Z}\sum_z m_z=\frac{1}{Z}\sum_z\left(\frac{1}{T}\sum_t\boldsymbol{\alpha}^{[z]\mathrm{T}}\mathbf{x}^{[t]}\right)=\frac{1}{Z\cdot T}\sum_z\sum_t\boldsymbol{\alpha}^{[z]\mathrm{T}}\mathbf{x}^{[t]} \tag{6.13}$$

$$v=\frac{1}{Z}\sum_z v_z=\frac{1}{Z}\sum_z\left\{\frac{1}{T}\sum_t\left(\boldsymbol{\alpha}^{[z]\mathrm{T}}\mathbf{x}^{[t]}-\frac{1}{Z}\sum_z\boldsymbol{\alpha}^{[z]\mathrm{T}}\mathbf{x}^{[t]}\right)^2\right\} \tag{6.14}$$

と計算される。つまり、マーカー効果を nuisance parameter とみなし、事後分布を用いて期待値をとる。よって、マーカー遺伝子型の確率的な変動だけが直接的に後代の遺伝子型値の分離に影響する。

なお、多くの先行研究で用いられている、マーカー効果の平均値を点推定量として後代遺伝子型値の分離予測に用いる計算法では、

$$m' = \frac{1}{T}\sum_{t} m_t = \frac{1}{T}\sum_{t}\left(\frac{1}{Z}\sum_{\boldsymbol{z}} \boldsymbol{\alpha}^{[z]\mathrm{T}}\right)\mathbf{x}^{[t]} = \frac{1}{Z\cdot T}\sum_{z}\sum_{t} \boldsymbol{\alpha}^{[z]\mathrm{T}}\mathbf{x}^{[t]} \tag{6.15}$$

$$v' = \frac{1}{T}\sum_{t}\left\{\left(\frac{1}{Z}\sum_{\boldsymbol{z}} \boldsymbol{\alpha}^{[z]\mathrm{T}}\right)\mathbf{x}^{[t]} - m'\right\}^2 \tag{6.16}$$

このように平均と分散を求めている。よって、平均値の推定量はいずれの計算法を用いても全く同じだが、分散については一致しない。

6-2-2. 仮想データの生成と解析手法

本研究では、ゲノミック予測を用いた育種のシミュレーションを支援する目的で開発された R パッケージ {BreedingSchemeLanguage} を用いて仮想データを生成した（Yabe et al., 2017）。はじめに、同パッケージの defineSpecies 関数と initializePopulation 関数を用いて 1,000 個体からなる仮想集団を作成した。ここでは、染色体数を 7 本、染色体の長さ（全ての染色体に共通）を 150 cM、有効集団サイズを 100 個体、SNP マーカーの数を 5,000 SNPs、マイナーアリル頻度を 0.1 以上に設定した。QTL の個数は 20 とし、全ての QTL は相加効果のみを持つとした。生成された集団について、遺伝率が 0.2 または 0.6 になるように、環境誤差を分散の異なる正規分布から無作為抽出して遺伝子型値に加え、表現型値を生成した。また、遺伝率が 1 である（環境誤差を一切加えず、遺伝子型値を表現型値とする）理想的な場合についても検証を行った。なお、染色体数、染色体の長さ、有効集団サイズの設定は、{BreedingSchemeLanguage} パッケージの既定値を採用したものである。また、同パッケージでは、QTL の効果は、当該アリルの頻度によらず、正規分布からの無作為抽出によって生成される。

この仮想集団 1,000 個体すべてを用いて予測モデルを構築した。予測モデルには BRR および BayesA を用いた。ここで、5,000 SNP 全てを用いる場合と、無作為に選ばれた 500 SNP のみを用いる場合の 2 通りでモデル構築を行った。ここでは {BGLR} パッケージを用いて、Gibbs sampling に基づくマーカー効果の MCMC 標本を得た。MCMC の繰り返し数は 12,000 回とし、burn-in として冒頭の 2,000 回分のサンプルを破棄するとともに、10 回に 1 回の頻度でマーカー効果を抽出した。これにより、1 通りの回帰につき 1,000 個の事後確率分布に従うマーカー効果の標本を得て、以後の計算に用いた。

仮想集団 1,000 個体の全通り交配は 499,500 通りであるが、そのうち 4,950 通りを無作為に

選んで後代個体を仮想生成した。1 通りの交配組み合わせについて、それぞれ 1,000 個体の後代マーカー遺伝子型をシミュレーションにより生成した。シミュレーションは式(6.1)から(6.6)に基づき行なった。ここで、組換え価は {BreedingSchemeLanguage} によって集団を生成した際に与えられる地図距離に基づき、Haldane の式によって計算した。

以上の手続きに従って得られた仮想後代個体、および、事後分布に従うマーカー効果のサンプルから、式(6.13), (6.14)によって後代遺伝子型値の平均値と分散を推定した。また、先行研究における推定法を式(6.15), (6.16)によって再現し、マーカー効果の不確実性を考慮しない場合に生じる問題点についても検証した。

まとめると、本研究では遺伝率の異なる仮想形質について、マーカー密度と予測モデルを変えてマーカー効果を推定した。さらに、MCMC サンプルを用いた厳密計算法と、事後平均を用いた計算法の２通りで後代集団の遺伝子型値の平均値と分散を予測した。これにより、遺伝率、マーカー密度、予測モデル、計算法の違いによる分離の推定精度を検証した。なお、遺伝率については 0.6 を、マーカー数については 5,000 を既定値とし、それぞれの値を動かした。

上述の手続きでは親系統を予測モデルの訓練集団に組み込んだが、実際の育種では親系統をモデル構築に利用できる場合とそうでない場合の両方が生じうる。例えば収量や籾数など、交配後にしか得られない表現型が育種目標の場合には、親系統をモデル構築に用いることが難しい。しかし、生育の初期から中期に特異的に問題となる病害抵抗性や生理的特性が育種目標であれば、交配前に表現型を取得し、親系統を訓練集団に用いて交配計画を立てることができる。また、当然ながら、前年までに表現型が取得された系統が親候補であれば、やはりその表現型を利用して予測モデルを構築できる。

6-3. 結果

・　後代平均の予測精度について

後代遺伝子型値の平均値について、遺伝率と予測モデルを変えた場合の予測結果を図 6-1 に、マーカー数と予測モデルを変えた場合の予測結果を図 6-2 に示した。後代平均の予測については、理論計算の通り、MCMC サンプルを用いた厳密計算法（赤色、十字）とマーカー効果の事後平均値に基づく計算法（青色、丸印）は一致する。

２つの予測モデルの間には、それほど明瞭な違いは見られなかった。遺伝率の低下は、一般のゲノミック予測における推論と同様、予測精度の顕著な悪化を引き起こした（図 6-1）。また、マーカー密度についても同様に、マーカー密度の低下による予測精度の悪化が確認された（図 6-2）。遺伝率やマーカー密度の変化が予測精度に与える影響は、２つのモデルでほぼ共通であった。

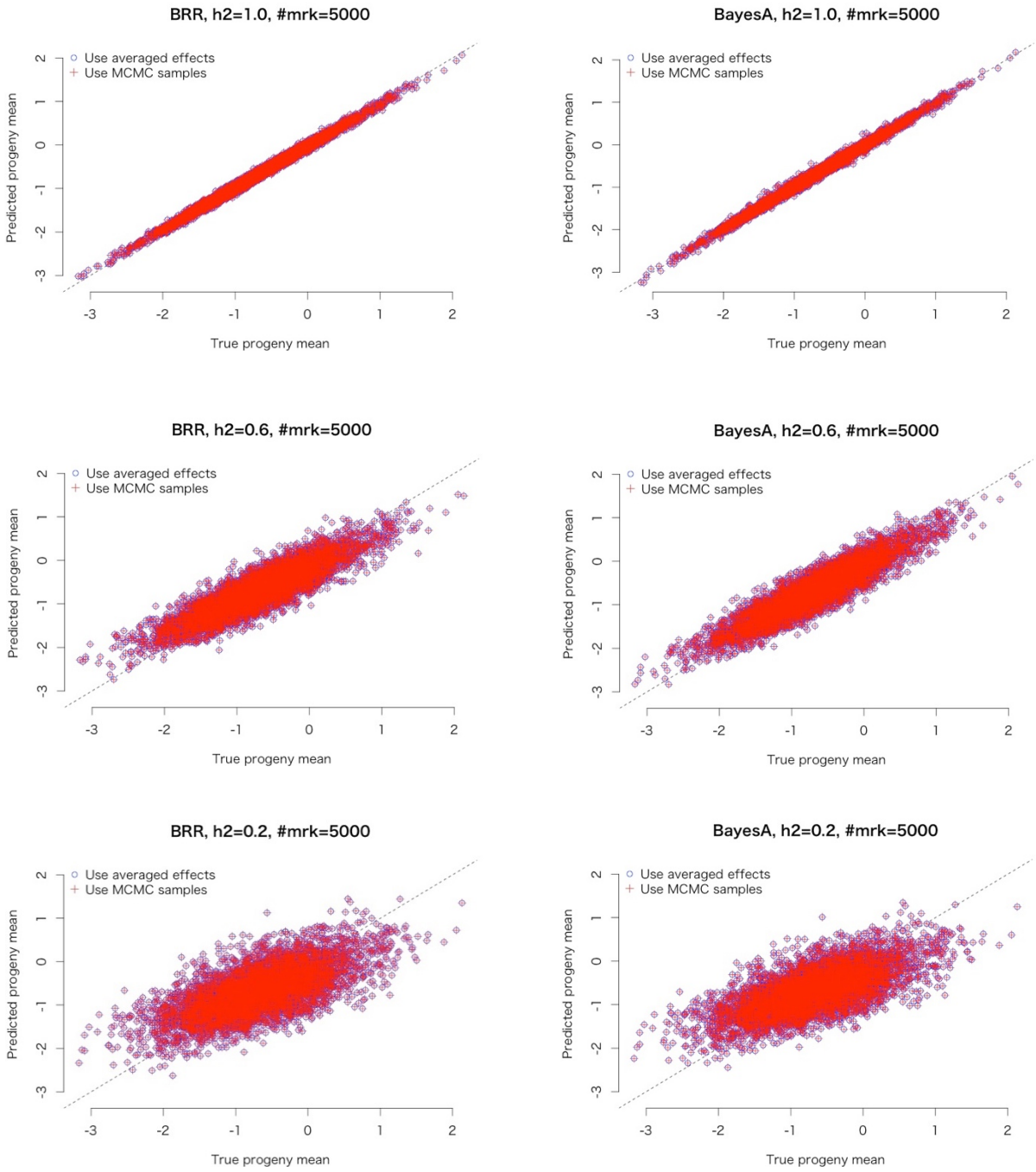
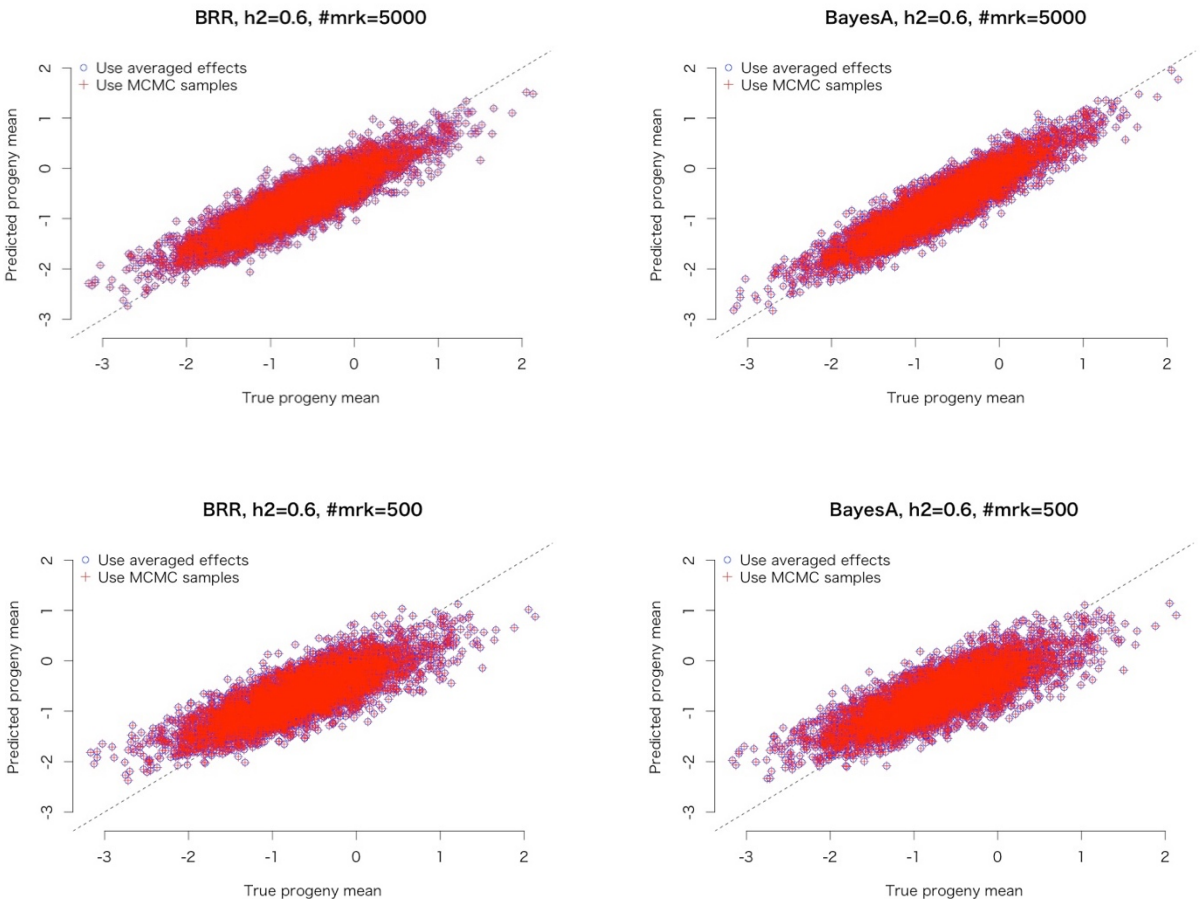


図 6-1. 遺伝率と予測モデルによる後代平均の予測精度の変化

横軸に真の後代平均を、縦軸にゲノミック予測に基づく後代平均の予測値をとり、予測結果を図示した。ここで、マーカー効果の事後平均を点推定量とする方法を青色・丸印で、MCMC サンプルを用いた厳密計算法を赤色・十字で示した。左列が BRR による予測を、右列が BayesA による予測であり、上段から順に、遺伝率が 1 の場合、0.6 の場合、0.2 の場合である。



、

図 6-2. マーカー密度と予測モデルによる後代平均の予測精度の変化

横軸に真の後代平均を、縦軸にゲノミック予測に基づく後代平均の予測値をとり、予測結果を図示した。ここで、マーカー効果の事後平均を点推定量とする方法を青色・丸印で、MCMC サンプルを用いた厳密計算法を赤色・十字で示した。左列が BRR による予測を、右列が BayesA による予測であり、上段はマーカー数が 5,000 の場合であり、下段はマーカー数が 500 の場合である。なお、上段の図は図 6-1 の中段の図と同一であるが、比較を容易にするためにここでも示した。

・　後代分散の推定精度について

後代遺伝子型値の分散に関しても同様に、遺伝率と予測モデルを変えた場合の予測結果を図 6-3 に、マーカー数と予測モデルを変えた場合の予測結果を図 6-4 に示した。

まず、後代分散の予測については、MCMC サンプルを用いた厳密計算法（赤色・十字）とマーカー効果の事後平均値に基づく計算法（青色・丸印）で大きな違いが見られた。マーカー効果の点推定量だけを用いた場合には、分散の予測が過小予測（分散の予測値が、真値に比べてほぼ一様に小さくなる）された。また、図 6-3 より、この過小予測は遺伝率が低くなるにつれて顕著になることもわかった。すなわち、後代分散の予測では、マーカー効果の不確実性を考慮しなければならないことが明らかとなった。そこで、以下では厳密計算法の結果のみに注目し、その他の因子についての結果を述べる。

遺伝率の低下は、後代分散の予測に大きな影響を与えた。図 6-2 を縦方向に比較すると、いずれの予測モデルを用いた場合でも、遺伝率が低下するにつれて、分散の予測が極端に縮小する（ほとんどの分散の予測値が同じ値に近づく）傾向が見られた。とりわけ遺伝率が 0.2 と低い場合には、後代の分散についてはほとんど予測できていなかった。

図 6-2 のうち、特に上段と中段について横方向に比較を行うと、後代分散の予測精度が、用いる予測モデルにも大きく依存しうることがわかる。BRR を用いた場合には、遺伝率が 1 である場合でさえ、分散はやや過小予測されるとともに、わずかに縮小している。遺伝率が 0.6 の場合に両モデルの違いは最も顕著であり、BRR では遺伝率の低下による分散の縮小予測がはっきりと確認できるが、BayesA ではその度合いが小さく、遺伝率の低下に対して頑健に後代分散の予測が可能であった。

予測に用いる SNP マーカー数の減少は、後代分散をやや縮小予測させるとともに過小予測させた。また、BayesA では、一部の組み合わせから生じた後代集団については、特異的に分離の推定が縮小かつ過小予測してしまう傾向が観察された。ただし、4,950 通りの交配組み合わせ全体については、BRR と BayesA による分散の予測精度に顕著な違いは見られなかった。

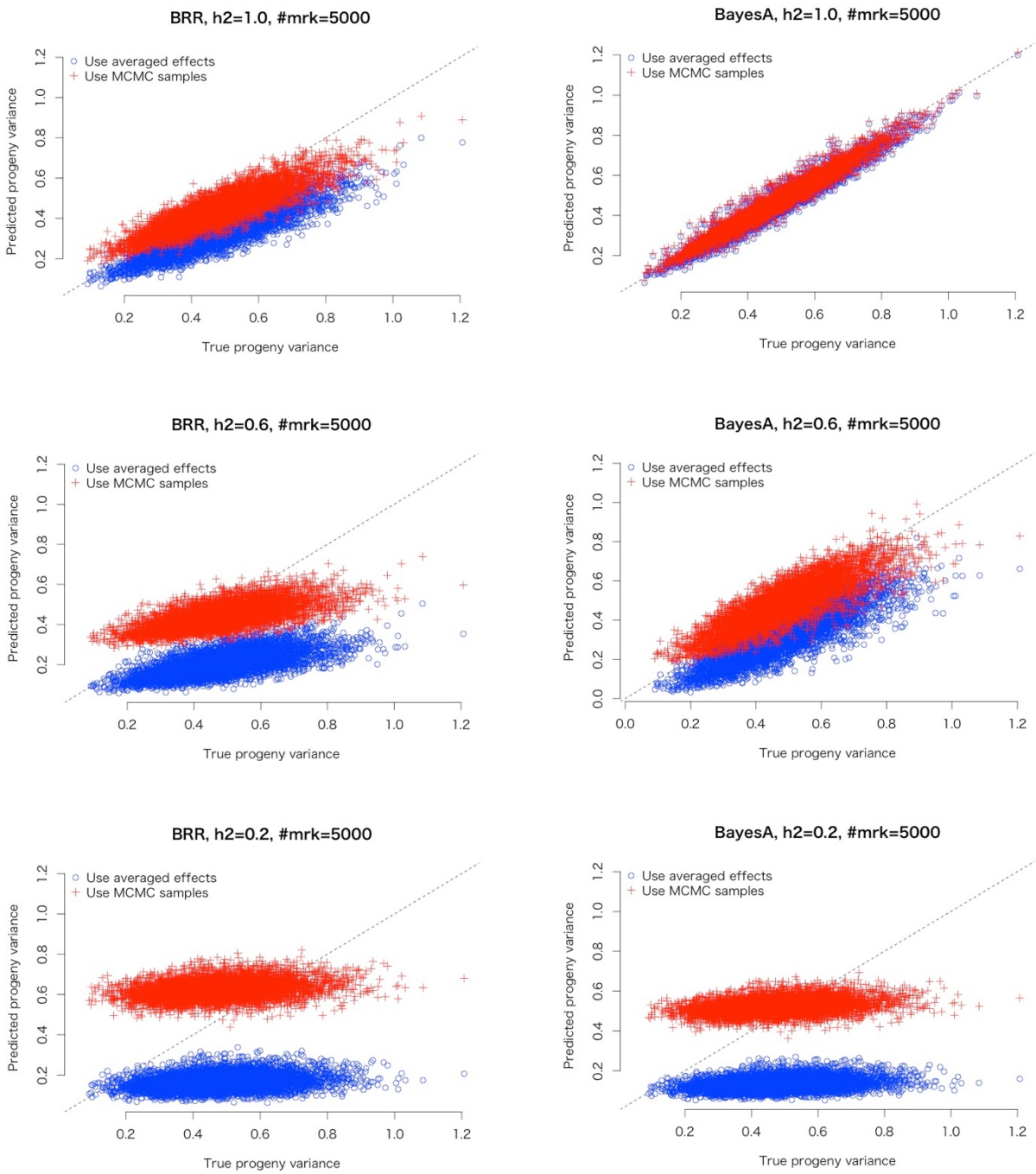


図 6-3. 遺伝率と予測モデルによる後代分散の予測精度の変化

横軸に真の後代分散を、縦軸にゲノミック予測に基づく後代分散の予測値をとり、予測結果を図示した。ここで、マーカー効果の事後平均を点推定量とする方法を青色・丸印で、MCMC サンプルを用いた厳密計算法を赤色・十字で示した。左列が BRR による予測を、右列が BayesA による予測であり、上段から順に、遺伝率が 1 の場合、0.6 の場合、0.2 の場合である。

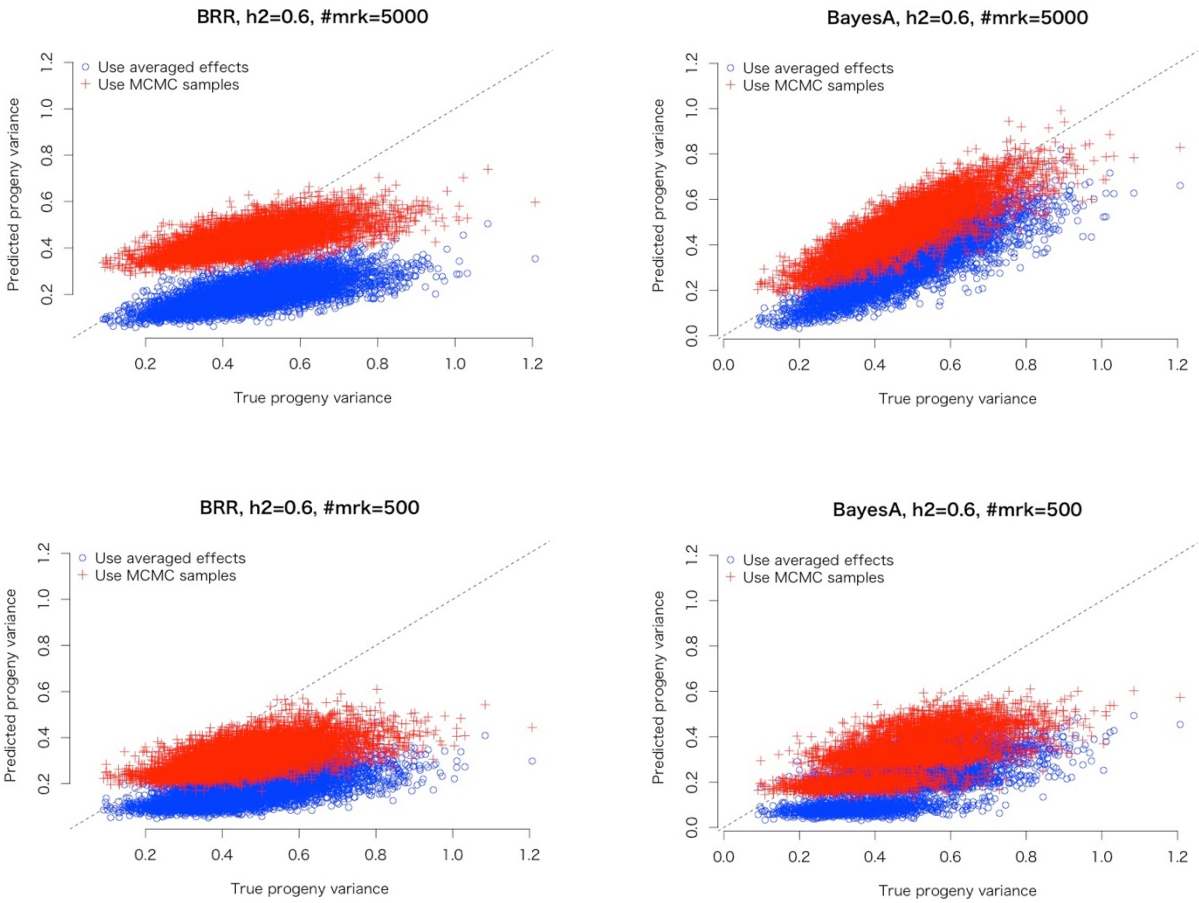


図 6-4. マーカー密度と予測モデルによる後代分散の予測精度の変化

横軸に真の後代分散を、縦軸にゲノミック予測に基づく後代分散の予測値をとり、予測結果を図示した。ここで、マーカー効果の事後平均を点推定量とする方法を青色・丸印で、MCMC サンプルを用いた厳密計算法を赤色・十字で示した。左列が BRR による予測を、右列が BayesA による予測であり、上段はマーカー数が 5,000 の場合であり、下段はマーカー数が 500 の場合である。なお、上段の図は図 6-3 の中段の図と同一であるが、比較を容易にするためにここでも示した。

6-4. 考察

交配後代の遺伝子型値について、後代平均の予測精度は、遺伝率やマーカー数によらず、BRR と BayesA の間に大きな違いがなかった。よって、後代平均を予測するだけを目的とする場合には、いずれの予測モデルを用いても問題ないことが示唆された。いっぽうで、後代分散の予測精度は予測モデルによって大きく異なり、BayesA を用いるほうが BRR に比べて優れた予測を与えることが示唆された。

後代分散の予測精度は、遺伝率の低下やマーカー数の減少によって悪化したが、そこには２つの異なる変化が確認された。１つは縮小予測であり、もう１つは過小予測である。遺伝率の低下は主に縮小予測を、マーカー数の減少は主に過小予測を引き起こした。また、このような視点では､BRR は BayesA に比べて分散の縮小予測を引き起こしやすいモデルであることが示唆された。

過小予測については、単純に予測モデルが説明できる遺伝分散が減少したことが原因だと推察される。マーカー数の減少は、真の QTL との連鎖不平衡を低下させ、QTL によって説明される遺伝分散を捉えることを困難にする。つまり、500 個の SNP マーカーだけで捉えられる遺伝的変動は真の遺伝分散よりも小さく、それにより、後代集団における遺伝分散も一様に過小評価されたと予想される。いっぽうで、分散の縮小予測については、現段階で適切な理論的説明は見出せない。

BayesA や BayesB は、BRR に比べて、QTL の近傍に位置する一部の DNA マーカーを活用する傾向が強いモデル化手法だと考えられており、このため、ゲノム全体にわたる連鎖不平衡のパターンが世代の変化によって変わっても、頑健に予測が可能であると考えられている（Habier et al., 2007; Habier et al., 2013）。本研究の結果は、これら先行研究とは異なる視点から BayesA モデルの優位性を示唆したものである。上記の先行研究では、それぞれの後代集団における個体レベルでの予測精度（真の遺伝子型値と予測値の相関係数）が計算され評価に用いられた。この予測精度が高いほど後代平均も後代分散も正確に予測できることが期待されるものの、必ずしもそうではない場合がある。例えば、この相関係数は全ての予測値に定数和や正の定数倍を行なっても変わらないため、個々の遺伝子型値が過小・縮小予測されていても相関係数は高い可能性がある。しかし、個々の遺伝子型値が過小予測されれば後代平均は過小予測され、縮小予測されれば後代分散は過小予測される。つまり、Habier らの先行研究は家系内の予測精度について、本研究は家系間の予測精度について論じたものであり、異なる視点でモデルの精度評価を行なっている。本研究と Habier らの先行研究を踏まえれば、世代を超えて予測モデルを適用したい場合には、BRR ではなく BayesA を利用することに大きな利点があると結論づけられる。

なお、本研究の結果は、必ずしも直近の研究結果（Yao et al., 2018）と矛盾するものではない。Yao らもまた BayesA、BayesB、BRR を含む複数のモデルで後代分散の予測精度を比較し、モデル間にほとんど差はないと結論づけている。しかし、彼らは評価基準として真の分散と予測された分散の間の相関係数を用いており、分散の過小・縮小推定については一切考慮していない。また、Yao らは独自に持つ育種材料の利用を想定し、予測モデルの構築を 57 系統という非常に少ない系統数で実施している。その他にも、本研究と Yao らのシミュレーションには条件に隔たりがあり、結果を直接は比較することはできない。Yao らの結果と本研究の結果を踏まえると、

BayesA と BRR で分離の予測精度は異なる可能性があるが、条件によってはその差はわずかであると結論づけるべきであろう。例えば、本研究のシミュレーションデータでも、遺伝率が低い場合には、BRR と BayesA の違いは明瞭ではない。

本研究では、BayesA が BRR よりも後代の分離予測において優れたモデルであることが示唆された。本研究で比較を行わなかった他のモデル化手法についても同様の比較を行うことは重要である。しかし、さらに興味深いのは、後代の分離、特に分散の予測にも優れた性質（不偏性や漸近一致性など）を持つ予測モデルの開発である。BRR だけでなく BayesA でも、遺伝率やマーカー数によっては分散が過小予測された。後代平均の予測について不偏性を満たしつつ、後代分散の予測も不偏にするような推定法は存在するのか、存在するならばどのように構成すれば良いか、存在しないならばどのようなトレードオフがあるのか、などの疑問に答えることは重要な研究課題である。そのためには、より数理統計学的なアプローチが必須だと考えられる。

7. 総合考察

・　本研究の総括

本研究では、ゲノミック予測を用いた育種の効率化・最適化を目標とし、３章と５章では訓練データの最適な選択法について、４章では予測モデルに基づく選抜法について検討した。また、６章では、育種戦略の最適化を行うために不可欠な後代分離の予測についてモデル比較を行なった。

３章では、能動学習をゲノミック予測における訓練集団最適化に応用し、uncertainty sampling によって、SVM による選抜・淘汰の二値分類を高精度化できることを示した。高次元データであるマーカー遺伝子型を入力データとすること、また、表現型は環境誤差が大きいため選抜・淘汰の二値ラベルに誤りが多いと推察されることなど、ゲノミック予測に特有の難しさが能動学習の効果に影響を及ぼすことが懸念されたが、本研究の範囲では顕著な影響は確認されなかった。

５章では、既存の訓練集団最適化指標を多環境試験デザインの最適化に応用した。既存手法の拡張は難しくないものの、その場合には設定するべき超パラメータが複雑化し、かつ、設定されたパラメータによって得られる多環境試験デザインが大きく異なることを明らかにした。定性的考察より、遺伝率や遺伝子型値の環境間相関が低い場合には、集団を代表するような遺伝子型を持つ一部の系統を複数の環境で繰り返し試験すべきであり、逆にそれらが高い場合には、全ての系統をほぼ同じ回数だけ試験するべきであることがわかる。最適化指標は、与えられた超パラメータに従って、上述のような多環境試験デザインにおけるトレードオフを調節していた。

３章および５章で扱った能動的な訓練データ選択法の開発は、今後のゲノミック予測における重要な課題の１つである。すでに、表現型の入手にかかる時間的・金銭的コストはマーカー遺伝子型のそれよりも大きく、育種におけるボトルネックになっている。表現型測定の技術も日進月歩で進化を続けているが、この状況を覆すことは容易ではない。なぜなら、遺伝子型が実在し時間的に安定であるいっぽう、表現型は確率的に変動し、かつそれは時間的に非定常な可能性があるからだ。遺伝子型は、現状のシーケンス技術の精度や、わずかに固定されていない遺伝子座、突然変異などによる変動はあるが、１つの系統にほぼ１つ定まり、高精度に測定されれば、その値が時間変化することはほとんどない。しかし、表現型値は環境の影響によって変動する。地球の気候変動を考えれば、環境が同一であることはほとんどなく、表現型の測定に終わりはない。農業が植物工場のような閉鎖系だけで行えるようになるまでは、我々は変動する環境の影響を、表現型の測定を通して育種に反映させていく必要があるだろう。どの系統を測定し、どの系統の測定を放棄するか、それをどのような環境で行うかを合理的に決定することは、こうした中長期的な目線に立てば、育種が経済的な営みとして利益を生むために必須の課題だといえる。

４章では、ベイズ最適化を応用し、ゲノミック予測に基づく優良遺伝資源の探索を効率化する選抜法を提案した。提案された期待改善量に基づく選抜戦略は、予測モデルの不確実性を考慮し、予測が不確実な系統を優先して選抜し圃場試験することで、既存のデータに頼っていては発見することの難しい優良系統を早期に発見できた。遺伝資源の重要性は広く認知されながらも、その効率的な活用法が存在しないために、実際の育種における利用はコア・コレクションなどの一部

の系統に限定されていた。ゲノミック予測はジーンバンクの利用を加速する手立てとして注目を集めつつあり（Yu et al., 2016; Crossa et al., 2017）、期待改善量を用いた選抜戦略は、その効率をさらに高める方法として期待される。

6章では、ゲノミック予測を用いて、ある両親から生じる後代個体の遺伝子型値の分離を予測する問題を扱った。ゲノミック予測におけるモデル評価は、主に適当な集団を用いた交差検証によって実施されており（e.g. Heslot et al., 2012）、分離世代の予測精度について考慮した例はほとんどなく、近年の研究例も厳密には正しくない計算法に基づくものであった（Yao et al., 2018）。本研究の結果は、BRR モデルに対する BayesA モデルの優位性を強く示唆するものであった。BayesA と類似の性質を持つと考えられる BayesB は、ある集団に注目した場合に、世代の更新に伴う予測精度の低下が BRR に比べて小さいことが知られている（Habier et al., 2007）。本研究の結果と合わせて、交配を伴う育種へゲノミック予測を利用する場合には、BRR を用いるのではなく、BayesA モデルを用いることが強く推奨される。

3章から5章までの研究は、いずれも、交配による新たな遺伝子型の作出を想定せず、ある事前に定められた集団の中で、予測モデルの精度を高めること（3章、5章）、あるいは、集団に含まれる優良系統を効率よく探索すること（4章）を目的とした。遺伝資源の探索や、既存系統を別の環境に導入する導入育種法においては、これらの章で提案された方法を直接的に利用することができる。しかし、多くの育種では、交配や自殖によって新たな遺伝子型が生じ、それらを選抜・評価することが目標となる。その場合にも、新たに得られた系統のマーカー遺伝子型を取得すれば、モデルに基づき予測を行い、その uncertainty や expected improvement を計算することができると考えられるため、必ずしも本研究の手法が利用できないとは限らないが、その有効性は担保されていない。6章で議論したような後代の予測を踏まえて、3章から5章で開発・評価した各手法をさらに発展させることは、非常に興味深い話題だと思われる。ここで注意すべきは、能動学習もベイズ的最適化も、ある予測モデルを前提として、訓練データや選抜の最適化を行なっているという点である。したがって、用いる予測モデルが妥当なものでなければ、これら手法の結果もまた妥当でないものになる可能性がある。例えば、4章では GBLUP を予測モデルとして用いたが、これは BRR と類似の性質をもつ予測モデルである。よって、6章で得られた結果を踏まえれば、交配を含む育種にベイズ最適化を適用したい場合には、GBLUP では不適切な可能性が高く、BayesA やそれに準じたモデルを用いるべきかもしれない。

なお、能動学習やベイズ最適化に関連の深い研究分野として、強化学習（reinforcement learning）が挙げられる。本研究では、特定の状況（訓練データの構築や、優良系統の選抜）、特定の予測モデル（SVM や GBLUP）に限定することで、uncertainty sampling や PEVmean / CDmean、また期待改善量といった明確で計算可能な最適化基準を導出した。しかし、実際の育種はより複雑であり、このような最適化基準を導出することは一般に不可能である。強化学習は、能動学習やベイズ最適化の拡張に当たる、ある目的関数からの返り値（例えば、育種によって得られた系統の良さ）だけを前提として、試行の繰り返しによって最適な戦略を発見する機械学習の1分野である。強化学習の成功例である alpha 碁（Silver et al., 2017）に倣って表現すれば、育種というゲームに勝利する（優れた系統を作出する）ための最善手（最善な実験計画、交配組み合わせや選抜されるべき個体）を、強化学習によって導けるようになるのではないだろうか。

ただし、alpha 碁の学習にかかった試合数を考えれば、最適な育種戦略を得るためには、育種というゲームを繰り返しプレイしなければならない。コンピュータ上で高速かつ厳密な試行錯誤が可能な囲碁とは異なり、現実世界で何千（あるいは何億か、それ以上かもしれない）通りもの育種を試行錯誤することは不可能であるので、何かしらの妥当な「育種シミュレータ」が必要であろう。そのためには、育種に対するじゅうぶん妥当な理解がなくてはならず、しかも、それが、少なくともシミュレーションとして実装可能な形で、定式化されている必要がある。したがって、育種のモデル化そのものが、強化学習の活用に向けても重要な意味を持つと思われる。このような難しさが想定されるものの、強化学習が育種の効率化に向けて有効なアプローチとなる可能性は高い。

・　育種における不確実性の定量化と活用

本論文で扱った４つの研究に共通する視点は、「予測の不確実性」をいかに考慮するかということである。３章の能動学習や５章の多環境試験最適化では、予測の不確実性が小さくなるように試験するべき系統・環境を設計することで、予測モデルの精度を高めることができた。また、４章のベイズ最適化では、予測の不確実性が大きな未試験系統を優先して選抜することで、より効率的に優れた系統を発見できることを明らかにした。上記３つの研究とは少し意味合いが異なるものの、６章の分離予測では、マーカー効果の不確実性を正しく考慮する計算法を用いることで、点推定を用いる（予測を考慮しない）場合に生じる過小予測を防げることがわかった。

訓練集団最適化のようなごく一部の理論研究を除けば、ゲノミック予測に関するほとんどの研究で、こうした予測の不確実性はしばしば見過ごされてきた。しかし、本論文の研究結果が示すように、ゲノミック予測の運用（選抜や後代分離の予測）において、予測の不確実性は極めて重要な役割を果たす。「不確実」という言葉は、ともするとネガティブな印象を与える。だが、４章で不確実性の高い系統こそが選抜すべき系統であったように、必ずしも予測が不確実であることが悪いことであるとは限らない。実験・測定によって得られるデータは有限であり、そこから得られるいかなる結論も、必然的に不確実性を持つ。つまり、不確実性とは常に存在するものであり、我々の得たデータを正しく表現し、適切な意思決定を行うためには、目を背けてはならないものである。本論文で扱った４つの研究を通して、ゲノミック予測によって育種の不確実性を定量化することの重要性と、それを活用することの利点が強く示唆されたといえる。

ゲノミック予測は、表現型の実測をマーカー遺伝子型と予測モデルで肩代わりすることで、育種サイクルの促進や選抜候補系統数の増大、低遺伝率な形質の正確な評価を実現し、育種効率を向上する育種技術だと一般には捉えられている。それは誤りではないが、ゲノミック予測の１つの側面を切り取ったものに過ぎない。ゲノミック予測を用いる本質的な利点は、遺伝子型と表現型の関係が定量的・数学的に記述されることある。マーカー遺伝子型が取得できず、ゲノミック予測ができなかった時代において、未試験の系統は（家系に関する情報がない限りは）等しく「わからない」対象であった。マーカー遺伝子型によって系統間の関係性が表現され、ゲノミック予測モデルによって既存の表現型データが統計的に活用されることで、未試験の系統に関する「不確実性」が定量的に記述されたことで、本研究のようにモデル構築や選抜の効率化が可能になっ

たのである。本論文の研究を支えた「予測の不確実性」は、まさにゲノミック予測による定量化の賜物と言える。

このゲノミック予測による定量化は、あくまで予測モデルを通して既存のデータを活用することによって行われる。裏を返せば、予測モデルに組み込まれなかったデータや生物学的な知見は、少なくとも予測には反映されず、予測モデルから導出される選抜法も、あくまで予測モデルに活用されたデータの範囲で合理的なものにすぎない。したがって、多様なデータや既存の生物学的知見を予測モデルに活用する方法の開発は極めて重要である。ゲノム情報については、例えば既知の QTL を明示的にモデルに組み込むこと（Rutkoski et al., 2014）、バイオインフォマティクス的な解析（配列の相同性に基づく解析など）によって得られたアノテーション情報を活用すること（Morota et al., 2014; Gao et al., 2017）などが検討されている。いっぽう、環境の情報については作物モデルとゲノミック予測の融合が有力な方法だと考えられる（Technow et al., 2015; Onogi et al., 2016）。これらの研究をさらに活発化させることはもちろん、より抽象的・概念的なレベルで、異なる種類（ゲノム、環境、表現型など）やスケール（細胞、組織、個体、群落など）の情報・知見を統合的に処理する数理的な枠組みを検討するべきだと考える。多様な情報が統合的にモデル化されれば、その予測モデルへと再び本研究の議論を適用することで（これが可能かどうかは予測モデルに依存するが、能動学習もベイズ的最適化も、幅広い機械学習手法・確率モデルに対して何らかの拡張が可能である）、やはり効率的な育種法を構築できると考えられる。

以上のように、本研究はゲノミック予測に基づく育種の効率化・最適化を目的とし、予測の不確実性を活用することでそれを実現した。ゲノミック予測は育種を客観化する強力な道具であり、本研究で扱った能動学習やベイズ最適化は、その道具を合理的に運用する手法だといえる。育種の効率化・最適化に向けては、道具の発明（より優れた予測モデルの開発・検証）と、道具の使い方の洗練（本研究を始めとする、数理科学的手法による育種の最適化）の両方が必要である。本研究はその一端を明らかにしたものに過ぎないが、得られた結果や導入された視点は、今後の多くの研究に活かされるものと期待される。

8. 摘要

2050 年までに地球の人口は 90 億人を突破すると試算されており、持続可能な食料供給を実現するには育種を加速する技術の開発は不可欠である。ゲノミック予測とは、ゲノムワイドマーカー遺伝子型をもとに表現型値を予測することをいい、この予測値をもとに選抜を行ことをゲノミック選抜という。ゲノミック選抜は、育種において大きなコストを要している表現型評価を、予測モデルを用いて大幅に省略でき、育種の効率化に貢献する重要技術と考えられている。本研究では、ゲノミック予測を用いた育種の効率化・最適化を目的とした新規手法の開発・予測モデルの評価を行なった。

1. 能動学習に基づくゲノミック予測モデルの効率的構築

ゲノミック選抜は、ある系統を選抜するか淘汰するかの二値分類問題として捉えることもできる。このとき、予測モデルの分類精度は選抜効率に直結する重要な因子であり、できる限り高い分類精度を実現する予測モデルを構築することが望ましい。一般に、ゲノミック予測の精度は、モデル構築に用いる系統に依存することが知られており、同じ系統数でも、どの系統を用いるかを適切に選択することで、高精度な予測モデルを構築できると考えられる。無作為抽出された訓練データを受動的に用いて予測モデルを学習するのではなく、既存のデータと予測モデルを踏まえて能動的に訓練データを選択し学習する方法は能動学習とよばれ、機械学習の分野を中心に開発が進み、創薬等の分野では既に実用化されている。能動学習は幅広い研究分野でその有効性が確かめられているが、ゲノミック予測で扱う高次元かつノイズの大きいデータでも機能するかは未知数であった。

そこで、本研究では、能動学習をゲノミック予測に応用することで、予測精度を効率よく向上できるかをシミュレーションを用いて検証した。遺伝子型の選抜・淘汰を分類問題としてあつかい、サポートベクトルマシンを用いて分類モデルを構築した。能動学習には、最も標準的なアルゴリズム uncertainty sampling を採用した。4 つの実データセットを用いて検証した結果、延べ 22 形質のうち 17 形質で能動学習により有意に分類精度が向上し、3 形質で低下した。以上の結果から、能動学習がゲノミック予測の訓練データ選択法として有用であることが示された。

2. ベイズ最適化に基づく優良系統の効率的発見

遺伝資源の持つ多様な有用変異を活用することは育種における重要課題である。しかし、現状では、遺伝資源の利用はコア・コレクションに含まれるごく一部の系統に限られている。ゲノミック予測を用いれば、一部の系統の表現型とマーカー遺伝子型から構築された予測モデルをもとに、マーカー遺伝子型が取得された全ての系統について遺伝子型値を予測できる。一部系統の表現型評価-モデル更新-未試験系統の選抜、というサイクルを繰り返すことで、マーカー遺伝子型をもつ全ての遺伝資源系統を対象に、有用系統の探索を行うことができる。

本研究では、ゲノミック予測を用いた優良遺伝資源系統の探索を black-box 最適化の枠組みで定式化するとともに、ベイズ最適化とよばれる最適化アルゴリズムに基づき、期待改善量という新たな選抜基準をもとに選抜を行う方法を提案した。期待改善量は予測モデルの不確実性を考慮した選抜基準であり、不確実性が高い、すなわち、非常に劣った系統である可能性もあるが、既存の系統よりも優れた系統である可能性も高い系統を優先的に選ぶ戦略を与える。実データを用いたシミュレーションによる検証の結果、期待改善量に基づく選抜戦略は、通常のゲノミック予測で行われる選抜戦略に比べて、平均で 30%ほど少ない試験系統数で、遺伝資源内の優良系統を発見できた。通常の選抜戦略は、単純に予測値の大きな系統から順に選抜するが、その場合、一部の系統だけで構築された予測モデルを盲信してしまう。期待改善量に基づく選抜戦略は、不確実性を考慮することにより、予測モデルの誤りを適宜修正しつつ、既存の系統を上回る系統を選抜できると考えられた。この結果から、期待改善量に基づく選抜戦略により、ゲノミック予測を用いた優良遺伝資源系統の探索をさらに加速できると期待される。

3. ゲノミック予測における多環境試験デザインの最適化

多環境試験は、遺伝子型と環境の交互作用（GxE; genotype-by-environment interaction）に関する情報を得るために必須である。しかし、多数の系統を用いた大規模な多環境試験を行うには大きな金銭的・労力的に大きなコストが必要であり、通常は、主要な系統に絞って多環境試験を実施し、それらの系統についてのみ GxE を評価する。しかし、ゲノミック予測を多環境の表現型データに拡張する（多環境ゲノミック予測）ことで、一部の表現型データをもとに、試験しなかった表現型を補完することも可能である。つまり、ゲノム情報によって系統間の類似性が定義されていれば、必ずしも一部の系統を選んで多環境試験を行う必要はなく、それぞれの環境で異なる系統を試験しても、GxE に関する知見を得ることができる。

この場合にも、能動学習により試験するべき系統を選んだ場合と同様に、各環境で試験すべき系統を適切に選ぶことにより、得られる予測モデルの精度が向上する可能性がある。本研究では、ゲノミック予測のモデル構築のためにどの系統を用いるか、という訓練集団の最適化のために提案された予測誤差分散（PEV; prediction error variance）および決定係数（CD; coefficient of determination）を多環境におけるゲノミック予測にも拡張し、どの系統をどこで試験すべきか、という多環境試験のデザインの最適化に用いた。PEV や CD を多環境試験のデザインの最適化に用いる場合には、遺伝子型値の環境間相関や、対象形質の遺伝率を超パラメータとして事前に設定する必要がある。本研究では、この超パラメータの設定が、PEV や CD によって得られる多環境試験のデザインを大きく左右することを明らかにした。例えば、環境間相関が低い場合には、すべての候補系統を満遍なく試験するようなデザインが選ばれ、逆に、環境間相関が高い場合には、一部の代表的な系統を複数の環境で試験するようなデザインが選ばれた。また、多環境試験のデザインを PEV や CD によって最適化する場合には、これら超パラメータを妥当な値に設定する必要があることを明らかにした。例えば、表現型値に GxE の影響がほとんどなく表現型値の環境間相関が 0.9 を超えるような場合に、環境間相関を 0.25 と設定してしまうと、PEV や CD を用いることにより、予測精度が逆に悪化した。しかし、真の状態と大きく異なる設定をしない限

り、PEV や CD による最適多環境試験デザインを用いて予測精度を改善できることがわかった。

4. ゲノミック予測に基づく交配後代の分離予測に関するシミュレーション研究

ゲノミック予測は、ある個体の持つマーカー遺伝子型をもとに、その個体の遺伝子型値を予測する手法である。そして、個体や系統を、表現型値ではなく予測値をもとに選抜するのが基本的な利用法である。しかし、両親のマーカー遺伝子型とマーカー間の組換え価が与えられれば、その後代個体のマーカー遺伝子型をシミュレーションによって生成できる。こうして生成される仮想のマーカー遺伝子型にゲノミック予測を適用すれば、後代個体のもつ遺伝子型値の平均値や分散の予測値を得ることができる。これら予測値は育種家が交配組合せを選定する際に有益な情報となる。

後代の分離予測はゲノミック予測の重要な活用手段であるにも関わらず、分離予測の精度（つまり、後代遺伝子型値の平均値や分散の予測精度）については十分な検討がなされてこなかった。本研究では、ベイズリッジ回帰（BRR; Bayesian ridge regression）および BayesA とよばれる 2 つの代表的な予測モデルについて、後代分離の予測精度に注目したモデル比較を行なった。

まず、数少ない先行研究のほとんどが、後代分散を厳密に正確な方法で計算していないことを明らかにした。具体的には、訓練データから構築された予測モデルには常に不確実性がともなうため、後代遺伝子型値の分散を予測する場合には、その不確実性を考慮しなければならない。考慮しない場合には、後代分散が大きく過小予測される可能性があることを理論式の導出により明らかにした。さらに、BRR と BayesA が、後代平均の予測についてはほとんど同程度の予測精度を与えるにも関わらず、後代分散の予測については、BayesA のほうがはるかに優れた予測精度を与えることを明らかにした。BRR によって予測された後代分散は、ほとんどの交配組み合わせで似通った値を示す縮小予測になる傾向が、BayesA に比べて強く見られた。すなわち、交配組み合わせ間で後代遺伝子型値の分散の大小を予測・比較したい場合には、BayesA を使用することが強く推奨される。BayesA に類似する BayesB が、別の観点からも、交配による世代の変化に頑健な予測ができることが先行研究で示されており、本研究の結果と合わせて、BRR に対する BayesA の優位性が示された。

ゲノミック予測は、育種家によりともすると主観的に行われる育種を客観化するための強力な道具であり、本研究で扱った能動学習やベイズ最適化は、その道具を合理的に運用するための手法である。本研究で提案した手法により、いくつかの単純化された条件のもとでは、ゲノミック予測の優れた運用法が与えられることがわかった。しかし、実際の植物育種は非常に複雑であり、本研究はそのごく一部を切り取って最適化したものにすぎない。持続的な食料供給の実現に向けて、本研究のようなアプローチをもとにした育種のモデル化と最適化を、さらに推し進めることが望まれる。

9. 謝辞

本研究の遂行および論文の執筆にあたり、指導教員である東京大学大学院農学生命科学研究科生産・環境生物学専攻の岩田洋佳准教授からは多大なるご助言をいただいた。岩田洋佳准教授には、東京大学農学部４年次の研究室配属後より一貫してご指導の労を賜った。その長年のご指導、ご助言の全てに、この場を借りて深く感謝の意を表したい。同じく研究室配属後よりご指導をいただいた同研究室の岸野洋久教授、大森宏助教、また修士課程進学後からご指導いただいた鐘ヶ江弘美特任助教にも深く感謝する。本研究の遂行に必要な統計学、機械学習、量的遺伝学、バイオインフォマティクスの基礎的知識を身につけることができたのは、岸野洋久教授、岩田洋佳准教授、大森宏助教、鐘ヶ江弘美特任助教のご指導の賜物である。

また、東京大学大学院農学生命科学研究科の井澤毅教授と二宮正士特任教授、東京大学大学院新領域創成科学研究科メディカル情報生命専攻の津田宏治教授にも謝意を表する。井澤毅教授、二宮正士特任教授、津田宏治教授には、岸野洋久教授とともに本論文の副査を務めていただき、多数の有益なご指摘を頂戴した。

本研究の５章で解析に用いた日本水稲データの表現型データは、神戸大学大学院農学研究科附属色資源教育研究センターの山崎将紀教授にご提供いただいた。また、同データのマーカー遺伝子型は、農業・食品産業技術総合研究機構（以下、農研機構）遺伝資源センターの江花薫子博士、農研機構農業環境変動研究センターの中川博規博士、農研機構次世代作物開発センターの矢部志央理博士、鐘ヶ江弘美特任助教らによって所得されたものをご提供いただいた。大変貴重なデータの使用をご快諾いただいたことに、この場を借りて謝意を表したい。

上記の日本水稲のデータを除けば、本研究で使用したデータは全て公開データであり、私自身が実際に圃場で取得に携わったデータは解析に用いられていない。しかしながら、岩田洋佳准教授の計らいで、在学中に複数の圃場試験・調査に参加させていただいた。全ての方々の名前をここで取り上げることはできないが、圃場でともに作業していただいた先生やスタッフの皆様、そして学生の皆様にも改めて感謝したい。複雑系である圃場での現象を扱う農学において、データを取得することは何よりも重要であり、それを現場で学ぶ機会を設けていただいたことは、本研究の遂行に大いに役に立った。

また、生物測定学研究室のスタッフ、学生、研究員の方々にも感謝したい。佐々木三枝子事務員には、研究活動を側面から支えていただいた。研究室の OB・OG である小野木章雄博士（現・農研機構次世代作物開発研究センター）と矢部志央理博士は、心強い先輩として多くの質問に答えていただいた。動植物のゲノミック予測に関する理論と実践の双方に造詣の深い両氏が先輩として在籍されていたことは、私にとって大変に幸運なことであった。また、多くの学生のうち、特に堀智明氏とは数多くの機械学習・量的遺伝学の教科書を一緒に勉強し、議論を重ねた。研究室における日々の活発な議論の積み重ねなくして、本研究を遂行することは叶わなかった。この場を借りて謝意を表する。

10. 参考文献

Akbani, R., S. Kwek and N. Japkowicz. 2004. Applying support vector machines to imbalanced datasets. In: European conference on machine learning (pp. 39–50). Springer, Berlin, Heidelberg.

Akdemir, D. and J.I. Sánchez. 2016. Efficient breeding by genomic mating. Front. Genet. 7: 210.

Akdemir, D., J.I. Sanchez and J.L. Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. 47: 38.

Akdemir, D., W. Beavis, R. Fritsche-Neto, A.K. Singh and J. Isidro-Sánchez. 2018. Multi-objective optimized breeding strategies for sustainable food improvement. Heredity. doi: 10.1038/s41437-018-0147-1.

Araus, J.L. and J.E. Cairns. 2014. Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. 19: 52–61.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott and J.L. Jannink. 2011. Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. Plant Genome. 4: 132–144.

Auer, P., N. Cesa-Bianchi and P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. Mach. Learn. 47: 235–256.

Avendaño, S., J.A. Woolliams and B. Villanueva. 2004. Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. Genet. Res. (Camb.) 83: 55–64.

Bari, A., K. Street, M. Mackay, D.T.F. Endresen, E.D. Pauw and A. Amri. 2012. Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. Genet. Resour. Crop Evol. 59: 1465–1481.

Bishop, C.M. 2006. Pattern Recognition and Machine Learning. Springer.

Blagus, R. and L. Lusa. 2010. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics. 11: 523.

Breiman, L. 2001. Random Forests. Mach. Lern. 45: 5–32.

Broman, K.L., H. Wu, S. Sen and G.A. Chrchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinfomatics. 19: 889–890.

Brown, J. and P. Calligari. 2008. An introduction to plant breeding. Oxford: Blackwell Publishing.

Browning, S.R. and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084–1097.

Burgueño, J., G. de los Campos, K. Weigel and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52: 707–719.

Byrne, P.F., G.M. Volk, C. Gardner, M.A. Gore, P.W. Simon and S. Smith. 2018. Sustaining the future of plant breeding: the critical role of the USDA-ARS national plant germplasm system. Crop Sci. 58: 451–468.

Chaloner, K. and I. Verdinelli. 1995. Bayesian experimental design: a review. Stat. Sci. 10: 273–304.

Chang, C.C. and C.J. Lin. 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2: 27.

Chang, H.X., P.J. Brown, A.E. Lipka, L.L. Domier and G.L. Hartman. 2016. Genome-wide association and genomic prediction identifies associated loci and predicts the sensitivity of Tobacco ringspot virus in soybean plant introductions. BMC Genom. 17: 153.

Chapelle, O. and L. Li. 2011. An empirical evaluation of Thompson sampling. Adv. Neural. Inf. Process. Syst. pp.2249-2257.

Chen, T. and C. Guestrin. 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20: 37–46.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. Mach. Learn. 20: 273–297.

Crossa J., Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueño, V.S. Windhausen, E. Buckler, J.L. Jannink, M.A.L. Cruz and R. Babu. 2013. Genomic prediction in maize breeding populations with genotypeing-by-sequencing. G3 (Bethesda). 3: 1903–1926.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, J.M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, S. Dreisigacker, R. Singh, X. Zhang, M. Gowda, M. Roorkiwal, J. Rutkoski and R.K. Varshney. 2017. Genomic selection in plant breeding: methods, models and perspectives. Trends Plant Sci. 22: 961–975.

Daetwyler, H.D., M.J. Hayden, G.C. Spangenberg and B.J. Hayes. 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics. 200: 1341–1348.

de los Campos, G. and A. Grüneberg. 2014. MTM pakage: Fits a (Bayesian) Multivariate Gaussian Mixed Effects Model using a Gibbs Sampler. URL: http://quantgen.github.io/MTM/vignette.html (accessed November 26, 2018).

de los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis and D. Sorensen. 2013a. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet. 9: e1003608

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler and M.P.L. Calus. 2013b. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 193: 327–345.

Desta, Z.A. and R. Ortiz. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19: 592–601.

Endelman, J.B. and J.L. Jannink. 2013. Shrinkage estimation of the realized relationship matrix. G3 (Bethesda). 2: 1405–1413.

Falconer, D.S. and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. 4th edn, Longman Group, London.

Gao, N., J.W.R. Martini, Z. Zhang, X. Yuan, H. Zhang, H. Simianer and J. Li. 2017. Incorporating gene annotation into genomic prediction of complex phenotypes. Genetics. 207: 489–501.

García-Ruiz, A., J.B. Cole, P.M. VanRaden, G.R. Wiggans, F.J. Ruiz-López and C.P.V. Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. U.S.A. 113: E3995–4004.

Gianola, D. 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 194: 573–596.

Gianola, D., G. de los Campos, W.G. Hill, E. Mandredi and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics. 183: 347–363.

Gianola, D., R. Fernando and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics. 173: 1761–1776.

González-Recio, O. and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43: 7.

Gorjanc, G., J. Jenko, S.J. Hearne and J.M. Hickey. 2016. Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genom. 17: 30.

Grattapaglia, D., O.B. Silva-Junior, R.T. Resende, E.P. Cappa, B.S.F. Muller, B. Tan, F. Isik, B. Ratcliffe and Y.A. El-Kassaby. 2018. Quantitative genetics and genomics converge to accelerate forest tree breeding. Front. in Plant Sci. 9: 1963.

Habier, D., R.L. Fernando and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177: 2389–2397.

Habier, D., R.L. Fernando and J.C.M. Dekkers. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 194: 597–607.

Han, Y., J.N. Cameron, L. Wang and W.D. Beavis. 2017. The predicted cross value for genetic introgression of multiple alleles. 205: 1409–1423.

He, H. and E.A. Garcia. 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21: 1263–1284.

Heffner, E.L., J.L. Jannink and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome. 4: 65–75.

Henderson, C.R. 1984. Applications of Linear Models in Animal Breeding. Guelph Ontario, CA: Univ. Guelph.

Hernández-Lobato, J.M., J. Requeima, E.O. Pyzer-Knapp and A. Aspuru-Guzik. 2017. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. Proc. Int. Conf. Mach. Learn. pp.1470-1479.

Heslot, N., H.P. Yang, M.E. Sorrells and J.L. Jannink. 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52: 146–160.

Hickey, J.M., R.F. Veerkamp, M.P.L. Calus, H.A. Mulder and R. Thompson. 2009. Estimation of prediction error variances via Monte Carlo sampling methods using different formulations of the prediction error variance. Genet. Sel. Evol. 41: 23.

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Prasanna, M. Grondona, A. Zambelli, V.S. Windhausen, K. Mathews and G. Gorjanc. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54: 1476–1488.

Hoi, S.C.H., R. Jin, J. Zhu and M.R. Lyu. 2006. Batch mode active learning and its application to medical image classification. Proc. 23rd Int. Conf. Mach. Learn, 417–424.

Huang, Y.M. and S.X. Du. 2005. Weighted support vector machine for classification with uneven training class sizes. In: Conf. Proc. IEEE Int. Conf. Syst. Mach. Cybern. 7: 4365–4369.

Hunter, S.R. and B. McClosky. 2016. Maximizing quantitative traits in the mating design problem via simulation-based Pareto estimation. IIE Trans. 48: 565–578.

Isidro, J., J.L. Jannink, D. Akdemir. J. Poland, N. Heslot and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. 128: 145–158.

Iwata, H., T. Hayashi, S. Terakami, N. Takada, T. Saito and T. Yamamoto. 2013. Genomic prediction of trait segregation in a progeny population: a case study of Japanese pear (*Pyrus pyrifolia*). BMC Genet. 14: 81.

Jiang, Y. and J.C. Reif. 2015. Modeling epistasis in genomic selection. Genetics. 201: 759–768.

Jones, D.R., M. Schonlau and W.J. Welch. 1998. Efficient global optimization of expensive black-box functions. J. Glob. Optim. 13: 455–492.

Jordan, D.R., E.S. Mace, A.W. Cruickshank, C.H. Hunt and R.G. Henzell. 2011. Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. Crop Sci. 51: 1444–1457.

Kaga, A., T. Shimizu, S. Watanabe, Y. Tsubokura, Y. Katayose, K. Harada, D.A. Vaughan and N. Tomooka. 2012. Evaluation of soybean germplasm conserved in NIAS genebank and development of mini core collections. Breed. Sci. 61: 566¬592.

Khazaei, H., K. Street, A. Bari, M. Mackay and F.L. Stoddard. 2013. The FIGS (Focused Identification of Germplasm Strategy) approach identifies traits related to drought adaptation in *Vicia faba* genetic resources. PLoS ONE. 8: e63107.

Kumar, S., D. Chagne, M.C.A.M. Bink, R.K. Volz, C. Whitworth and C. Carlisle. 2012. Genomic selection for fruit quality traits in apple (*Malus×domestica* Borkh.). PLoS ONE. 7: e36674.

Lado, B., S. Battenfield, C. Guzman, M. Quincke, R.P. Singh, S. Dreisigacker, R.J. Pena, A. Fritz, P. Silva, J. Poland and L. Gutierrez. 2017. Strategies for selecting crosses using genomic prediction in two wheat breeding programs. Plant Genome. 10.

Laloë, D. 1993. Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 557–576.

Lehermeier, C., S. Teyssèdre and C.C. Schön. 2017. Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. Genetics. 207: 1651–1661.

Lewis, D.D. and W.A. Gale. 1994. A sequential algorithm for training text classifiers. Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, pp.3–12.

Lin, H.T., C.J. Lin, and R.C. Weng. 2007. A note on Platt's probabilistic outputs for support vector machine. Mach. Learn. 68: 267–276.

Malosetti, M., J.M. Ribaut and F.A. van Eeuwijk. 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Front. Physiol. 4: 44.

Mauricio, R. 2001. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat. Rev. Genet. 2: 370–381.

McCouch, S., G. J. Baute, J. Bradeen, P. Bramel, P. K. Bretting, E. Buckler, J. M. Burke, D. Charest, S. Cloutier, G. Cole, H. Dempewolf, M. Dingkuhn, C. Feuillet, P. Gepts, D. Grattapaglia, L. Guarino, S. Jackson, S. Knapp, P. Langridge, A. Lawton-Rauh, Q. Lijua, C. Lusty, T. Michael, S. Myles, K. Naito, R. L. Nelson, R. Pontarollo, C. M. Richards, L. Rieseberg, J. Ross-Ibarra, S. Rounsley, R.S. Hamilton, U. Schurr, N. Stein, N. Tomooka, E. van der Knaap, D. van Tassel, J. Toll, J. Valls, R.K. Varshney, J. Ward, R. Waugh, P. Wenzl and D. Zamir. 2013. Agriculture: feeding the future. Nature. 499: 23–24.

Meuwissen, T.H.E. 1998. Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 76: 2575–2583.

Meuwissen, T.H.E., B.J. Hayes and M.E. Goddard. 2014. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157: 1819–1829.

Minamikawa, M.F., K. Nonaka, E. Kaminuma, H. Kajiya-Kanegae, A. Onogi, S. Goto, T. Yoshioka, A. Imai, H. Hamada, T. Hayashi, S. Matsumoto, Y. Katayose, A. Toyoda, A. Fujiyama, Y. Nakamura, T. Shimizu and H. Iwata. 2017. Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. Sci. Rep. 7: 4721.

Mockus, J. 1994. Application of Bayesian approach to numerical methods of global and stochastic optimization. J. Glob. Optim. 4: 347–365.

Mohammadi, M., T. Tiede and K.P. Smith. 2015. PopVar: a genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. Crop Sci. 55: 2068–2077.

Montesinos-López, O.A., A.M. Montesinos-López P. Pérez-Rodríguez, G. de los Campos, K. Eskridge and J. Crossa. 2015. Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. G3 (Bethesda). 5: 291–300.

Morota, G. and D. GIanola. 2014. Kernel-based whole-genome prediction of complex traits: a review. Front. Plant Sci. 5: 363.

Morota, G., R. Abdollahi-Arpanahi, A. Kranis and D. Gianola. 2014. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genomics. 15: 109.

Müller, D., P. Schopp and A.E. Melchinger. 2018. Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. G3 (Bethesda). 8: 1173–1181.

Nakagawa, H., J. Yamagishi, N. Miyamoto, M. Motoyama, M. Yano and K. Nemoto. 2005. Flowering response of rice to photoperiod and temperature: a QTL analysis using a phenological model. Theor. Appl. Genet. 110: 778–786.

Odong, T.L., J. Jansen, F.A. van Eeuwijk, T.J.L van Hintum. 2013. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor. Appl. Genet. 126: 189–305.

Onogi A, Watanabe M, Mochizuki T, Hayashi T, Nakagawa H, Hasegawa T, Iwata H. 2016. Towards integration of genomic selection with crop modeling: the development of an integrated approach to predicting rice heading dates. Theor. Appl. Genet. 129: 805–817.

Onogi, A., O. Ideta, Y. Inoshita, K. Ebana, T. Yoshioka, M. Yamasaki and H. Iwata. 2015. Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). 128: 41–53.

Ornella, L., P. Pérez, E. Tapia, J.M. González-Camacho, J. Burgueño, X. Zhang, F.S. Vicente, D. Bonnett, S. Dreisigacker, R. Singh, N. Long and J. Crossa. 2014. Genomic-enabled prediction with classification algorithms. Heredity. 112: 616–626.

Pace, J., X. Yu and T. Lübberstedt. 2015. Genomic prediction of seedling root length in maize (*Zea mays* L.). Plant J. 83: 903–912.

Palloix, A., V. Ayme and B. Moury. 2009. Durability of plant major resistance genes to pathogens depends on the genetic background, experimental evidence and consequences for breeding strategies. New Pytol. 183: 190–199.

Patra, S. and L. Bruzzone. 2012. Cluster-assumption based batch mode active learning technique. Pattern Recognit. Lett. 33: 1042–1048.

Pauwels, E., C. Lajaunie and J.P. Vert. 2014. A Bayesian active learning strategy for sequential experimental design in systems biology. BMC Syst. Biol. 8: 102.

Pérez-Rodríguez, P., D. Gianola, J.M. González-Camacho, J. Crossa, Y. Manès and S. Dreisigacker. 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2: 1595–1605.

Perez, P. and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 198: 483–495.

Rincent, R., A. Charcosset and L. Moreau. 2017a. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theor. Appl. Genet. 130: 2231–2247.

Rincent, R., D. Laolë, S. Nicolas, T. Altman, D. Brunel, P. Revilla, V.M. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C.C. Schoen, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics. 192: 715–728.

Rincent, R., E. Kuhn, H. Monod, F.X. Oury, M. Rousset, V. Allard, and J.L. Gouis. 2017b. Optimization of multi-environment trials for genomic selection based on crop models. Theor. Appl. Genet. 130: 1735–1752.

Rutkoski, J.E., J.A. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, H. Barbier, M.N. Rouse, J.L. Jannink and M.E. Sorrells. 2014. Genomic selection for quantitative adult plant stem rust resistance in wheat. Plant Genome. 7.

Seko, A. T. Maekawa, K. Tsuda and I. Tanaka. 2014. Machine learning with systematic density-functional theory calculations: application to melting temperatures of single- and binary-component solids. Phys. Rev. B. 89: 054303.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report. University of Wisconsin-Madison. 1648.

Seung, H.S., M. Opper and H. Sompolinsky. 1992. Query by committee. Proc. 5th Annu. Work. Comput. Learn. Theory, 287–294

Shahriari, B., K. Swersky, Z. Wang, R.P. Adams, N. de Freitas. 2016. Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE. 104: 148–175.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel and D. Hassabis. 2017. Mastering the game of Go without human knowledge. Nature. 550: 354–359.

Snoek, J. and H. Larochelle. 2012. Practical Bayesian optimization of machine learning algorithms. Adv. Neural. Inf. Process. Syst. pp.2951-2959

Soltani, A. and T.R. Sinclair. 2012. Modelling physiology of crop development, growth and yield. UK: CABI publishing

Sorensen, D and D. Gianola. 2002. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redona, G. Atlin, J.L. Jannink and S.R. McCouch. 2015. Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. 11: e1004982.

Srinivas, N., A. Krause, S. Kakade and M. Seeger. 2010. Gaussian process optimization in the bandit setting: No regret and experimental design. Proc. Int. Conf. Mach. Learn. pp.1015–1022.

Tanksley, S.D. and McCouch S.R. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science. 227: 1063–1066.

Technow, F., C.D. Messina, L.R. Totir and M. Cooper. 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS ONE 10: e0130855.

Tiede, T., L. Kumar, M. Mohammadi and K.P. Smith. 2015. Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. Mol. Breed. 35: 199.

Tipping, M.E. 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1: 211–244.

Tong, S. and Koller D. 2001. Support vector machine active learning with applications to text

classification. J. Mach. Learn. Res. 2: 45–66.

Tuia, D., F. Ratle, F. Pacifici, M.F. Kanevski and W.J. Emery. 2009. Active learning methods for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 47: 2218–2232.

van Berloo, R. and P. Stam. 1998. Marker-assisted selection in autogamous RIL populations: a simulation study. Theor. Appl. Genet. 96: 147–154.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Wang, K., D. Zhang, Y. Li, R. Zhang and L. Lin. 2017. Cost-effective learning for deep image classification. IEEE Trans. Circuits. Syst. Video Technol. 27: 2591–2600.

Wang, W et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 567: 43–49.

Warmuth, M.K., J. Liao, G. Rätsch, M. Mathieson, S. Putta and C. Lemmen. 2003. Active learning with support vector machines in the drug discovery process. J. Chem. Inf. Comput. Sci. 43: 667–673.

Würschum, T., J.C. Reif, T. Kraft, G. Janssen and Y. Zhao. 2013. Genomic selection in sugar beet breeding populations. BMC Genet. 14: 85.

Yabe, S. H. Iwata and J.L. Jannink. 2017. A simple package to script and simulate breeding schemes: the breeding scheme language. Crop Sci. 57: 1347–1354.

Yabe, S. R. Ohsawa and H. Iwata. 2013. Potential of genomic selection for mass selection breeding in annual allogamous crops. Crop Sci. 53: 95–105.

Yabe. S., T. Hara, M. Ueno, H. Enoki, T. Kimura, S. Nishimura, Y. Yasui, R. Ohsawa and H. Iwata. 2018. Potential of genomic selection in mass selection breeding of an allogamous crop: an empirical study to increase yield of common buckwheat. Front. Plant Sci. 9: 276.

Yamamoto, E., H. Matsunaga, A. Onogi, H. Kajiya-Kanegae, M. Minamikawa, A. Suzuki, K. Shirasawa, H. Hirakawa, T. Nunome, H. Yamaguchi, K. Miyatake, A. Ohyama, H. Iwata and H. Fukuoka. 2016. A simulation-based breeding design that uses whole-genome prediction in tomato. Sci. Rep. 6: 19454.

Yao, J., D. Zhao, X. Chen, Y. Zhang and J. Wang. 2018. Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). Crop J. 6: 353–365.

Yin, X., M.J. Kropff, T. Horie, H. Nakagawa, H.G.S. Centeno, D. Zhu and J. Goudriaan. 1997. A model for

photothermal responses of flowering in rice I. model description and parameterization. Field Crops Res. 51: 189–200.

Yu, H., M.L. Spangler, R.M. Lewis and G. Morota. 2018. Do stronger measures of genomic connectedness enhance prediction accuracies across management units? J. Anim. Sci. sky316

Yu., X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T. Tesso, P.S. Schnable, R. Bernardo and J. Yu. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants. 2: 16150.

Zhao, K., C.W. Tung, G.C. Eizenga, M.H. Wright, M.L. Ali, A.H. Price, G.J. Norton, M.R. Islam, A. Reynolds, J. Mezey, A.M. McClung, C.D. Bustamante and S.R. McCouch. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat. Commun. 2: 467.

Zhao, K., M. Wright, J. Kimball, G. Eizenga, A. McClung, M. Kovach, W. Tyagi, M.L. Ali, C.W. Tung, A. Reynolds, C.D. Bustamante and S.R. McCouch. 2010. Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. PLoS ONE 5: e10780.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc and J.C. Reif. 2012. Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. 124: 769–776.

Zhong, S. and J.L. Jannink. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics. 177: 567–576.

Zhu, J., H. Wang, B.K. Tsou and M. Ma. 2010. Active learning with sampling by uncertainty and density for data annotations. IEEE Trans. Audio Speech Lang. Proc. 18: 1323–1331.