博士論文

効率的育種のための全ゲノム情報を活用する統計遺伝学的手法に関する研究

Research on statistical genetic methods utilizing whole-genome information for efficient breeding

平成 27 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 24 年博士課程入学

小野木　章雄

目次

1　序論

育種とは遺伝的な改良を通じて望ましい性質を持つ品種を作出することを指す。育種の歴史が始まったのは人類が狩猟・採集を中心とした生活から農耕・牧畜を中心とした生活へと移行した時と考えられ、その時期は概ね 1 万年前とされる。初期の頃は自然に生じた望ましい性質を持つ変異体を無意識的に或いは意識的に広げることで育種は進められたと考えられ、その速度は緩やかであった。それでも長い年月をかけた順化・家畜化（domestication）により栽培植物或いは家畜の形態や生理は着実に改良されてきた。例えばイネやオオムギなどの穀物における非脱粒性やマメ類の非休眠性（ラディジンスキー 2000）、家禽における就巣性の喪失などが知られる。

近代的な育種が始まったのは 1900 年にメンデルの法則が再発見されて以降とされる。その進展は統計学と密接に関わってきた。当初メンデルの法則は観察値が不連続となる形質、つまり質的な形質の遺伝様式を説明する理論としては受け入れられたが、観察値が連続的な形質、つまり量的形質の遺伝を説明できるかについては激しい議論があった（Provine 2001）。しかし Fisher（1918）はメンデル遺伝する要素を無数に考慮することで、血縁者間で観察された量的形質の相関を再現できることを示した。メンデルの法則の再発見以降、統計学とともに近代遺伝学、つまり集団遺伝学、量的遺伝学、及び進化遺伝学が発展することになる。遺伝学の進展は植物育種においては交雑により新たに変異を生じさせる交雑育種に理論的背景をもたらし、その後の利用と発展を加速させた。なお先の Fisher の論文は分散（variance）を初めて定義したことでも知られる。

Fisher の想定した無数のメンデル遺伝要素による形質の支配は微小要素モデル（infinitesimal model）と呼ばれ、今日まで特に動物育種における基礎的なモデルとなっている。Wright（1922）は Fisher の研究とは独立してこのモデルを想定し、量的形質の個体間における相関係数として血縁者間の遺伝的関係、つまり血縁係数（coefficient of relationship）を定義した。血縁係数を用いれば、血縁でつながった個体間の量的形質における共分散構造を定義できることになる。血縁係数は後に線形混合モデル（母数効果と変量効果を含む線形回帰モデル）と結びつくことで、育種に多大な貢献をすることとなった（Hill 1996）。この回帰モデルでは量的形質の表現型値つまり観察値を被説明変数とし、それに対する個体の相加的な遺伝的能力（育種価）を変量効果として含み、その共分散構造を血縁

係数で定義する。一方で季節や性別など表現型値に影響を与える環境要因を母数効果として含む。この混合モデルを解くために導き出されたのが混合モデル方程式（mixed model equation）である（Henderson 1949; Henderson 1984）。この方程式は変量効果の分散成分が既知のもとで、個体の育種価に対する最も良い解、最良線形不偏予測（best linear unbiased predictor、BLUP）を与えることがその後示された（Henderson 1975）。混合モデル方程式と分散成分推定のための制限付き最尤推定法（restricted maximum likelihood、REML）（Patterson and Thompson 1971）は、動物及び樹木など一部植物の育種に大きく寄与している。また混合モデルをベイズ統計の枠組みで捉えマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo、MCMC）で解を推定する手法も提案され（Wang et al. 1993; Sorensen and Gianola 2002）、特に REML の収束が困難な複雑なモデルに対して利用されている。混合モデルに基づく育種価の推定及び予測は BLUP 法と呼ばれるが、BLUP 法の普及には計算機の能力の向上が不可欠であった。また多くのソフトウェアが開発され提供されたこともその普及に寄与した（Boldman 1993; Gilmour et al. 1998; Misztal et al. 2002; Madsen et al. 2006）。

BLUP 法に基づいた育種は微小要素モデルを想定し、個々の遺伝子には関心を払わないため遺伝子の機能に対する理解は深まらない。量的形質を支配する遺伝子座（quantitative trait locus、QTL）がいくつ染色体上のどこにあり、その表現型への寄与はどの程度か、ということが詳しく解析できるようになったのは、分子生物学の進展により多数の DNA マーカーが利用できるようになってからである。DNA マーカーは比較的容易に検出可能な塩基配列上の遺伝的変異を指し、制限酵素断片長多型（restriction fragment length polymorphism、RFLP）や反復配列などをもとにしたマーカーがまず実用化された。2 つの遺伝子座が連鎖しているか否か、つまり染色体上での位置関係は直接には観察できないことであり、統計的推定が行われてきた（Fisher 1935; Morton 1955）。しかし多数のマーカー遺伝子型が入手可能となり、それらの相対的位置、つまり連鎖地図（linkage map）を正確に推定する必要性が生じた。Elston and Stewart（1971）は複数の連鎖した遺伝子座における家系内での遺伝子型の出現について尤度関数を定義したが、非常に計算負荷が高かったため、Lander and Green（1987）は隠れマルコフモデルに基づく尤度計算と Expectation-maximization（EM）アルゴリズム（Dempster et al. 1977）を用いた連鎖地図の推定手法を開発した。連鎖地図はヒトを端緒として（Botstein et al. 1980）、シロイヌナズナ（Chang et al. 1988）、トウモロコシ及びトマト（Helentjaris et al. 1986）、イネ（McCouch et al. 1988）、及びウシ（Bishop et al. 1994）など様々な動植物において作成された。さらに得ら

れた連鎖地図への QTL マッピング手法として、区間マッピング法（interval mapping）（Lander and Botstein 1986）や複合区間マッピング法（composite interval mapping）（Zeng 1993; Zeng 1994）など統計的手法が提案された。これらの手法はマーカー間に仮想的に QTL を配置し、家系内で生じる遺伝子座間の相関関係、つまり連鎖不平衡（linkage disequilibrium、LD）を利用して QTL の遺伝子型を推定し、尤度に基づきその存在を検定する。家系内の組換えにより生じる LD を利用したマッピング手法は連鎖解析と呼ばれ、トマトの農業形質に対して初めて適用された（Paterson et al. 1988）。連鎖地図の作成及び連鎖解析は、多くの手法及びソフトウェアの開発（Ott 1976; Lander et al. 1987; Basten et al. 1994; Broman et al. 2003; Iwata and Ninomiya 2006; Yandell et al. 2007）に支えられ、育種においては耐病性に関わる QTL などが多く同定されてきた（Young 1996）。これらの知見が特に作物育種に貢献する一方で、多数の QTL が支配する量的形質においては、検出された QTL だけに基づく選抜やその集積（pyramiding）だけでは改良に限界があること、連鎖解析に用いた家系以外では QTL の効果が保障されないことなどが指摘されている（Bernardo 2008; Heffner et al. 2009）。

分子生物学はさらに進展の速度を上げ、現在は DNA アレイに代表されるような検出技術により数千から数十万座位の 1 塩基多型（single nucleotide polymorphism、SNP）について一度に遺伝子型の決定が可能となった。これによりマーカー密度が大幅に向上した。動植物においても多くの個体・系統で高密度のマーカー遺伝子型が決定されるようになったことから、それら全ゲノム情報に基づく育種手法、いわゆるゲノミックセレクションが実用可能な技術として注目を集めている（Meuwissen et al. 2001; Hayes et al. 2009; Heffner et al. 2009; Jannink et al. 2010; Heslot et al. 2014）。この手法は個体や系統の育種価や表現型をゲノムワイドに密に配置されたマーカーから予測し（ゲノムワイド予測、genomic prediction または whole-genome prediction）、選抜する。ゲノムワイド予測では学習用のデータセットにおいて推定育種価や評価された表現型を被説明変数、ゲノムワイドマーカーの遺伝子型を説明変数とした回帰モデル（ゲノムワイド回帰モデル）を構築し、予測したい個体・系統のゲノムワイドマーカー遺伝子型からその育種価・表現型を予測する（Meuwissen et al. 2001）。高密度のゲノムワイドマーカーを同時に用いることの狙いは、QTL マッピングでは検出できない効果の小さな QTL を、それらと集団内で強い LD にあるマーカーに肩代わりさせることにある。ゲノミックセレクションの大きな利点は、植物であれば表現型を評価する前、動物であれば後代検定（progeny test）により育種価を推定する前にゲノムワイドマーカーの遺伝子型のみから選抜が可能であるため、選抜サイクルを短縮することができ

改良速度の向上が期待できる点である（Hayes et al. 2009; Heffner et al. 2009）。また BLUP 法による育種価推定は血縁情報に基づくため、全きょうだいの優劣は後代検定なしに評価できなかったが、ゲノムワイド予測であればメンデリアンサンプリングを考慮できるために評価可能となる。しかし一方で考慮すべき問題もある。1 つは説明変数（ゲノムワイドマーカー）の数（*p*）が被説明変数（個体・系統）の数（*n*）よりはるかに多い状況で、その効果をどのように学習するか、いわゆる“large *p*, small *n*”と呼ばれる問題である。またゲノミックセレクションあるいはゲノムワイド予測を従来の育種手法とどのように組み合わせるか（Hayes et al. 2009）、さらに長期に継続した場合の選抜反応や近交係数の推移（Daetwyler et al. 2007; Goddard 2009; Jannink et al. 2010）などの問題もある。

高密度マーカーの遺伝子型が利用可能になったことにより、QTL マッピングの手法にも新たな可能性が提示されている。マーカーが密になったため、マーカーと QTL との集団内での LD を利用し QTL を検出することが可能となった。この手法はゲノムワイド関連解析（genome-wide association mapping）と呼ばれ、主にヒトの疾患関連遺伝子の探索に広く用いられているが（Wellcome Trust Case Control Consortium 2007; Barrett et al. 2008; McCarthy et al. 2008）、動植物の経済・農業関連形質にも適用されてきている（Barendse et al. 2007; Huang et al. 2010; Zhao et al. 2011; Sukumaran et al. 2012; Morris et al. 2013; McDaneld et al. 2014）。ゲノムワイド関連解析では集団において QTL を検出するために、連鎖解析と異なり特定の家系に依存しない利点がある。また実験的に家系を作出する必要もない。統計学手法としては表現型値をマーカーに 1 つずつ単回帰し多重検定の補正を行う手法が中心であるが（例えば Purcell et al. 2007）、ゲノムワイド予測で用いられるような多数のマーカーに同時に回帰する手法も提案されている（Xu 2003; Karkkainen and Sillanpaa 2012b）。またマーカーを 1 つずつ検定する場合に、検定対象のマーカー以外をバックグラウンド効果として回帰モデル取り込む方法も提案されている（Yu et al. 2006; Yang et al. 2014）。この手法ではバックグラウンド効果の個体間の共分散構造をマーカーで定義するため、混合モデルの一種となる。なお共分散構造を定義するために全マーカーを用いた場合は、バックグラウンド効果はゲノム情報を用いて推定した育種価と捉えることができる。この手法はゲノミック BLUP（genomic BLUP、GBLUP）と呼ばれ、ゲノムワイド予測における中心的手法となっている（de los Campos et al. 2013）。GBLUP については第 3 章及び第 4 章で再び触れる。

本研究では全ゲノム情報を利用する新たな育種技術、つまりゲノミックセレクショ

ン及びゲノムワイド関連解析を、多くの作物や家畜においてより実用的な手法とすることを目的とし、以下の 4 つの統計学的研究を行った。まず第 2 章では MCMC ではなくより高速な変分ベイズ法に基づくゲノムワイド回帰ソフトウェアを開発した。ゲノムワイド回帰はこれまでその柔軟さからベイズ統計の枠組みで多くの手法が提案されてきているが（Karkkainen and Sillanpaa 2012a）、それらは多くの場合 MCMC でパラメータの推定を行うため、多数のマーカーやサンプル数を扱うことがしばしば困難であった。そのため手法間の比較などが比較的小さなデータセットにおいてしか行うことができず、ゲノムワイド回帰を含めた手法選択を行う上での障害となっていた。第 2 章において提供するソフトウェアはこの従来のソフトウェアの欠点を解消するものと期待される。第 3 章ではアジア栽培イネ（*Oryza sativa* L.）においてゲノムワイド予測の有用性評価と、シミュレーションデータも交えて様々な予測手法の比較と適用範囲を探索した。これまでゲノムワイド予測による表現型予測は、トウモロコシやコムギ（例えば Crossa et al. 2010; Albrecht et al. 2011）など様々な作物でその有用性が示唆されてきているが、アジア栽培イネにおいてはまだその有用性が確認されていなかった。またゲノムワイド予測では“large *p*, small *n*”問題に対処する適切な予測手法の選択が必要となるものの、イネにおいては適切な手法選択の検証が行われていなかったため、第 3 章では実データをもとにこれらの課題について検討した。第 4 章では黒毛和種におけるゲノムワイド予測の有用性を検証した。ゲノム情報を用いた育種価の予測はホルスタイン種などで既に実用化されているものの、未だ黒毛和種では試みられていなかった。この章では従来の育種手法、つまり血縁情報に基づく BLUP 法にゲノム情報を直接結びつけた手法、single-step GBLUP（Legarra et al. 2009、Aguilar et al. 2010）を用いてその有用性を検証するとともに、その手法を運用する際に生じる諸課題についても検証を行った。第 4 章の研究は黒毛和種の育種において有益な情報を与えるとともに、これからゲノム情報に基づく育種を計画する品種・家畜に対しても有益となると考えられる。第 5 章では環境情報をもとに表現型を予測する作物モデル（crop model）とゲノムワイド予測を組み合わせた新たな統計モデルを提案し、アジア栽培イネの出穂期についてその予測能力を評価した。ゲノムワイド予測の欠点の 1 つは、環境の情報を用いないために未試験の環境下での作物の表現型を予測できないことにある。この欠点は環境情報から表現型を予測できる作物モデルと組み合わすことで解消できる可能性があるが、そのような統計モデルはこれまでどの作物及び形質においても試みられていなかった。第 5 章の試みはゲノムワイド予測の可能性を広げる端緒となると考えられる。最後に第 6 章において本研

究を統括するとともに、今後の育種における統計的手法の可能性について論じた。

2　ゲノムワイド回帰のための変分ベイズ法に基づくソフトウェア「VIGoR」の開発

2.1　諸論

ゲノムワイド回帰では、特にベイズ統計の枠組みでモデルを構築する手法がこれまで多く注目を集め、様々な手法が提案されてきている（Karkkainen and Sillanpaa 2012a）。ベイズ回帰手法は通常、パラメータ推定を MCMC で行う場合が多い（例えば Meuwissen et al. 2001; de los Campos et al. 2009a; Habier et al. 2011）。そのため現在ベイズ回帰手法を行うことのできる公開ソフトウェアは、例えば GenSel（Fernando and Garrick 2008）、BLR（de et al. 2009）、BGLR（Perez and de 2014）AlphaBayes（Hickey and Tier 2009）、GS3（Legarra et al. 2010）、また BayeZ（Janss 2010）などは、主に MCMC に依存している。しかしながらその計算量の多さから、多数のマーカーから成り立つようなデータセットを現実的な時間内で解析することは困難である。さらにゲノムワイド予測でよく行われる交差検証やそれを用いたハイパーパラメータの最適化など、集中的に行う繰り返し計算はより規模の小さいデータセットについてもしばし困難となる。この MCMC への依存はベイズ回帰手法の育種への適用を阻む要因の 1 つとなっている。そこで本研究ではベイズ回帰手法を実装し、かつ MCMC に依存しない新たなソフトウェア VIGoR（variational Bayesian inference for genome-wide regression）を開発し、全ゲノム情報の育種への活用をより容易にすることを目的とした。VIGoR は MCMC より高速な変分ベイズ法を用いてパラメータ推定を行う。VIGoR には代表的な 6 つのベイズ回帰手法が実装されている。すなわち Bayesian Lasso（Blasso）（Park and Casella 2008）、extended Bayesian Lasso（EBlasso）（Mutshinda and Sillanpaa 2010）、weighted Bayesian shrinkage regression（wBSR）（Hayashi and Iwata 2010）、BayesC（Habier et al. 2011）、stochastic search variable selection（SSVS）（George and McCulloch 1993）、及び Bayesian mixture regression（MIX）（Luan et al. 2009）である。

本章は以下の段落より成り立つ。「2.2　推定手法」では変分ベイズ法についての一般的な説明を行う。「2.3　ベイズ回帰手法」では VIGoR に実装されているベイズ回帰手法とその変分ベイズアルゴリズムを解説する。「2.4　VIGoR の主機能」では主な 3 つの機能について述べる。「2.5　シミュレーション解析」では VIGoR を用いた解析例をシミュレー

ションを用いて示す。この章ではまた、変分ベイズによって推定されるマーカー効果の事後不確実性（posterior uncertainty）が関連解析の有意性の判断に有用であるかどうかを並べ替え検定（permutation test）と比較することで検証した。「2.6　摘要」において本章の概略を述べる。

VIGoR は Linux または Mac のターミナルから実行可能なプログラムとして、また統計処理用言語 R（R Development Core Team 2011）のパッケージとしても利用可能である。

2.2　変分ベイズ法

変分ベイズ法はデータ（y）の周辺尤度をその下限を最大化することにより近似する。周辺尤度の下限はヤンセンの不等式により

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log \int q(\boldsymbol{\theta}) \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &\geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}\end{aligned}$$

と表すことができる。ここで q はパラメータ θ についての任意の確率分布とする。変分ベイズ法では q として θ の事後分布を互いに独立になるように分解した分布を用いる。つまり

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{P} q_i(\theta_i \mid \boldsymbol{y})$$

を用いる。ここで P はパラメータの数を表す。これ以降この分解した事後分布、つまり q_i を近似事後分布と呼ぶ。周辺尤度の下限は q_i に関して

$$q_i(\theta_i \mid \boldsymbol{y}) \propto \exp\left(E_{q_j, j\neq i}\left[\log p(\boldsymbol{y},\boldsymbol{\theta})\right]\right)$$

とすることで最大化できる。なぜならば周辺尤度の下限は

$$\begin{aligned}\int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} &= \int q_i(\theta_i \mid \boldsymbol{y}) \prod_{j\neq i} q_j(\theta_j \mid \boldsymbol{y}) \left[\log p(\boldsymbol{y},\boldsymbol{\theta}) - \log q_i(\theta_i \mid \boldsymbol{y})\right] d\boldsymbol{\theta} \\ &\quad - \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \sum_{j\neq i} \log q_j(\theta_j \mid \boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int q_i(\theta_i \mid \boldsymbol{y}) \left\{ \int \prod_{j\neq i} q_j(\theta_j \mid \boldsymbol{y}) \log p(\boldsymbol{y},\boldsymbol{\theta}) d\theta_{j\neq i} - \log q_i(\theta_i \mid \boldsymbol{y}) \right\} d\theta_i + const. \\ &= -KL\left\{ q_i(\theta_i \mid \boldsymbol{y}) \,\|\, \exp\left(E_{q_j, j\neq i}\left[\log p(\boldsymbol{y},\boldsymbol{\theta})\right]\right) \right\} + const.\end{aligned}$$

と展開できるからである。ここで KL はカルバック・ライブラー距離（KL 距離）を表す。

ここでは

$$\int q_j\left(\theta_j \mid \boldsymbol{y}\right) d\theta_j = 1$$

となることを用いた。周辺尤度の下限を最大化することは事後分布と近似事後分布との KL 距離を最小化することに等しい。これは

$$\begin{aligned}\log p(\boldsymbol{y}) &= \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} - \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} + KL\left[q(\boldsymbol{\theta} \mid \boldsymbol{y}) \| p(\boldsymbol{\theta} \mid \boldsymbol{y})\right]\end{aligned}$$

と展開できることからわかる。変分ベイズ法は例えば Bishop（2006）や Murphy（2012）の解説が詳しい。

2.3　ベイズ回帰手法

VIGoR が想定する線形回帰モデルは個体 i について以下のように表される。

$$y_i = \sum_{j=1}^{F} z_{ij}\alpha_j + \sum_{p=1}^{P} \gamma_p x_{ip}\beta_p + \varepsilon_i$$

ここで F はモデルに含まれるマーカー以外の変数の数、z_{ij} は変数 j の値、α_j は変数 j の効果、P はマーカー数、γ_p はマーカーp がモデルに含まれるか（1）否か（0）を表す指示変数、x_{ip} はマーカーp の遺伝子型、β_p はマーカー効果、ε_i は残差を表す。γ_p は wBSR 以外では 1 に固定し、wBSR のみにおいて推定対象とした。残差は平均 0、分散$1/\tau_0^2$の正規分布に従うと想定し、τ_0^2 の事前分布は$1/\tau_0^2$と想定した。α_j の事前分布は定数に比例するとした。β_p の事前分布は手法により異なる（表 2.1）。以下では各回帰手法における変分ベイズアルゴリズムについて述べる。

表 2.1　VIGoR に実装されたベイズ回帰手法のモデル構造 [a]

	Hierarchical level		
	1st	2nd	3rd
	Marker effect and indicator	Effect variance and indicator	Shrinkage magnitude
Blasso	$\beta_p \sim N\left(0, \frac{1}{\tau_0^2 \tau_p^2}\right)$	$\tau_p^2 \sim Inv-G\left(1, \frac{\lambda^2}{2}\right)$	$\lambda^2 \sim G(\varphi, \varpi)$
EBlasso	$\beta_p \sim N\left(0, \frac{1}{\tau_0^2 \tau_p^2}\right)$	$\tau_p^2 \sim Inv-G\left(1, \frac{\delta^2 \eta_p^2}{2}\right)$	$\delta^2 \sim G(\varphi, \varpi)$ $\eta_p^2 \sim G(\psi, \theta)$
wBSR	$\beta_p \sim N(0, \sigma_p^2)$ $\gamma_p \sim Bernoulli(\kappa)$	$\sigma_p^2 \sim \chi^{-2}(\nu, S^2)$	
BayesC	$\beta_p \sim N(0, \sigma^2)$ if $\rho_p = 1$ $\beta_p = 0$ if $\rho_p = 0$	$\sigma^2 \sim \chi^{-2}(\nu, S^2)$ $\rho_p \sim Bernoulli(\kappa)$	
SSVS	$\beta_p \sim N(0, \sigma^2)$ if $\rho_p = 1$ $\beta_p \sim N(0, c\sigma^2)$ if $\rho_p = 0$	$\sigma^2 \sim \chi^{-2}(\nu, S^2)$ $\rho_p \sim Bernoulli(\kappa)$	
MIX	$\beta_p \sim N(0, \sigma_A^2)$ if $\rho_p = 1$ $\beta_p \sim N(0, \sigma_B^2)$ if $\rho_p = 0$	$\sigma_A^2 \sim \chi^{-2}(\nu, S^2)$ $\sigma_B^2 \sim \chi^{-2}(\nu, cS^2)$ $\rho_p \sim Bernoulli(\kappa)$	

[a] 確率分布；　*N*、正規分布；　*Inv-G*、逆ガンマ分布；　*G*、　ガンマ分布、　χ^{-2}、尺度付き逆カイ二乗分布；　*Bernoulli*、ベルヌーイ分布

Blasso, Bayesian lasso; EBlasso, extended Bayesian lasso; wBSR, weighted Bayesian shrinkage regression; SSVS, stochastic search variable selection; MIX, Bayesian mixture regression.

2.3.1 Bayesian lasso（Blasso）

Blasso は 1 次正則化項（L_1 項）を持つ回帰手法の Lasso（Tibshirani 1996）をベイズの枠組みで捉えた手法である（Park and Casella 2008）。なお Lasso については次章の「3.2.4.3 Lasso 及び ENet」で説明する。Blasso の変分ベイズ法は Li and Sillanpaa（2012a）で提案されている。この著者らのパラメータ表現ではマーカー効果は残差分散と独立であるが、本研究では Blasso を提案した Park and Casella（2008）に従い、Blasso のマーカー効果の事前分布を残差分散に条件づけられるように変更した。後述するように EBlasso についても同様の変更を行った。このモデル構造の方が収束が速い傾向にあった（結果非掲載）。

Blasso の対数同時事後分布は

$$\frac{N}{2}\log\tau_0^2+\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2+\frac{P}{2}\log\tau_0^2+\frac{1}{2}\sum_{p=1}^{P}\log\tau_p^2-\frac{\tau_0^2}{2}\sum_{p=1}^{P}\tau_p^2\beta_p^2-2\sum_{p=1}^{P}\log\tau_p^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\lambda^2}{\tau_p^2}+(\phi-1)\log\lambda^2-\varpi\lambda^2+Const.$$

と表される。ここで Const はパラメータと独立な定数項を表す。前述のように各パラメータの近似事後分布はこの同時事後確率をそれ以外のパラメータの近似事後分布について期待値をとることにより得られる。

α_j の近似事後分布は正規分布となり、

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

また

$$V\left[\alpha_j\right]=\Lambda_j$$

となる。ここで $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$ である。

β_p の近似事後分布も正規分布となり、

$$E\left[\beta_p\right]=\mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

また

$$V\left[\beta_p\right]=\mathrm{H}_p$$

となる。ここで $\mathrm{H}_p^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+E\left[\tau_p^2\right]E\left[\tau_0^2\right]$ である。

τ_p^2 の近似事後分布は逆ガウシアン分布となり、

$$E\left[\tau_p^2\right]=\mu_p$$

及び

$$E\left[\frac{1}{\tau_p^2}\right]=\frac{1}{\mu_p}+\frac{1}{\xi_p}$$

となる。ここで $\mu_p=\sqrt{\frac{E\left[\lambda^2\right]}{E\left[\beta_p^2\right]E\left[\tau_0^2\right]}}$ 及び $\xi_p=E\left[\lambda^2\right]$ となる。この導出には

$E\left[X^r\right]=\mu^r\sum_{s=0}^{r-1}\frac{(r-1+s)!}{s!(r-1-s)!}\left(\frac{2\xi}{\mu}\right)^{-s}$ 及び $E\left[X^{-r}\right]=\frac{E\left[X^{r+1}\right]}{\mu^{2r+1}}$ を用いた（Chhikara and Folks 1989）

（ただし X はパラメータ μ と ξ で定義される逆ガウシアン分布に従うとする）。

τ_0^2 の近似事後分布はガンマ分布であり、

$$E\left[\tau_0^2\right]=\frac{a_1}{b_1}$$

となる。ここで

$$a_1=\frac{1}{2}(N+P)$$

及び

$$b_1=\frac{1}{2}\left\{\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)^2+\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{p=1}^{P}V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2+\sum_{p=1}^{P}E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right\}$$

である。λ^2 の近似事後分布もガンマ分布であり、

$$E\left[\lambda^2\right]=\frac{a_2}{b_2}$$

となる。ここで $a_2=P+\phi$ 及び $b_2=\frac{1}{2}\sum_{p=1}^{P}E\left[\frac{1}{\tau_p^2}\right]+\varpi$ となる。データの周辺尤度の下限は以下

のように表される。

$$\sum_{p=1}^{P}\left[-\frac{1}{2}E\left[\delta^2\right]E\left[\frac{1}{\tau_p^2}\right]-\frac{1}{2}\log E\left[\lambda^2\right]+\frac{1}{2}\log V\left[\beta_p\right]-\frac{1}{2}E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right]$$

$$-\varpi E\left[\lambda^2\right]-a_1\log b_1+\log\Gamma\left(a_1\right)-a_2\left(\log b_2-1\right)+\log\Gamma\left(a_2\right)+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]$$

$$-\frac{N-P-F}{2}\log 2\pi+\phi\log\varpi-\log\Gamma\left(\phi\right)+\frac{2P+F}{2}-P\log 2$$

ここで Γ はガンマ関数を示す。

2.3.2 Extended Bayesian lasso（EBlasso）

EBlasso は Blasso の拡張として提案された（Mutshinda and Sillanpaa 2010）。Blasso が回帰係数（マーカー効果）の縮約（shrinkage）程度を単一のパラメータ、λ^2、で調節しているのに対し、EBlasso では全体を調節する δ^2 と、個々のマーカーを調節する η_p^2 の 2 つを持つ（表 2.1）。そのため Blasso より柔軟にマーカー毎に縮約程度を調節できる。EBlasso の変分ベイズ法は Li and Sillanpaa（2012a）で提案されている。Blasso 同様に本研究ではマーカー効果が残差分散に条件づけられるように変更したアルゴリズムを導出した。EBlasso の対数同時事後分布は

$$\frac{N}{2}\log\tau_0^2+\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2+\frac{P}{2}\log\tau_0^2+\frac{1}{2}\sum_{p=1}^{P}\log\tau_p^2-\frac{\tau_0^2}{2}\sum_{p=1}^{P}\tau_p^2\beta_p^2-2\sum_{p=1}^{P}\log\tau_p^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\delta^2\eta_p^2}{\tau_p^2}$$
$$+(\phi-1)\log\delta^2-\varpi\delta^2+(\psi-1)\sum_{p=1}^{P}\log\eta_p^2-\theta\sum_{p=1}^{P}\eta_p^2+Const.$$

と表される。

α_j の近似事後分布は正規分布であり、

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

また

$$V\left[\alpha_j\right]=\Lambda_j$$

となる。ここで $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$ である。

β_p の近似事後分布も正規分布となり、

$$E\left[\beta_p\right]=\mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

また

$$V\left[\beta_p\right]=\mathrm{H}_p$$

となる。ここで $\mathrm{H}_p^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+E\left[\tau_p^2\right]E\left[\tau_0^2\right]$ である。

τ_p^2 の近似事後分布は逆ガウシアン分布となり、

$$E\left[\tau_p^2\right]=\sqrt{\frac{E\left[\delta^2\right]E\left[\eta_p^2\right]}{E\left[\beta_p^2\right]E\left[\tau_0^2\right]}}$$

また

$$E\left[\frac{1}{\tau_p^2}\right]=\sqrt{\frac{E\left[\beta_p^2\right]E\left[\tau_0^2\right]}{E\left[\delta^2\right]E\left[\eta_p^2\right]}}+\frac{1}{E\left[\delta^2\right]E\left[\eta_p^2\right]}$$

となる。

τ_0^2 の近似事後分布はガンマ分布であり、

$$E\left[\tau_0^2\right]=\frac{a_1}{b_1}$$

となる。ここで $a_1=\frac{1}{2}(N+P)$ 及び

$$b_1=\frac{1}{2}\left\{\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)^2+\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{p=1}^{P}V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2+\sum_{p=1}^{P}E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right\}$$

である。

δ^2 及び η_p^2 の近似事後分布はいずれもガンマ分布であり、

$$E\left[\delta^2\right]=\frac{a_2}{b_2}$$

及び

$$E\left[\eta_p^2\right]=\frac{a_3}{b_{3,p}}$$

となる。ここで

$$a_2=P+\phi$$

及び

$$b_2=\frac{1}{2}\sum_{p=1}^{P}E\left[\eta_p^2\right]E\left[\frac{1}{\tau_p^2}\right]+\varpi$$

であり、また

$$a_3=1+\psi$$

及び

$$b_{3,p} = \frac{1}{2}E\left[\delta^2\right]E\left[\frac{1}{\tau_p^2}\right] + \theta$$

となる。データの周辺尤度の下限は以下のように表される。

$$\sum_{p=1}^{P}\left[-\frac{1}{2}E\left[\delta^2\right]E\left[\eta_p^2\right]E\left[\frac{1}{\tau_p^2}\right] - \frac{1}{2}\log\left(E\left[\delta^2\right]E\left[\eta_p^2\right]\right) + \frac{1}{2}\log V\left[\beta_p\right] - \frac{1}{2}E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right]$$
$$+\sum_{p=1}^{P}\left[-\theta E\left[\eta_p^2\right] - a_3\left(\log b_{3,p} - 1\right) + \log\Gamma\left(a_3\right)\right]$$
$$-\varpi E\left[\delta^2\right] - a_1\log b_1 + \log\Gamma\left(a_1\right) - a_2\left(\log b_2 - 1\right) + \log\Gamma\left(a_2\right) + \frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]$$
$$-\frac{N-P-F}{2}\log 2\pi + \phi\log\varpi - \log\Gamma\left(\phi\right) + P\left[\psi\log\theta - \log\Gamma\left(\psi\right)\right] + \frac{2P+F}{2} - P\log 2$$

2.3.3 Weighted Bayesian shrinkage regression（wBSR）

wBSR は尤度関数にマーカーをモデルに加えるか否かを決める指示変数 γ_p を持つ（表 2.1）。指示変数が全てのマーカーにおいて 1 で固定された場合、ゲノミックセレクションの概念を初めて提示した Meuwissen et al.（2001）で提案された BayesA と同じモデルとなる。一方ベルヌーイ分布を事前分布として想定しこれを推定した場合は、Meuwissen et al.（2001）の BayesB と等価のモデルとなる。wBSR の変分ベイズ法は Hayashi and Iwata（2013）で提案されている。対数同時事後分布は

$$\frac{N}{2}\log\tau_0^2 - \frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F}z_{ij}\alpha_j - \sum_{p=1}^{P}\gamma_p x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2 - \frac{1}{2}\sum_{p=1}^{P}\log\sigma_p^2 - \frac{1}{2}\sum_{p=1}^{P}\frac{\beta_p^2}{\sigma_p^2} + \left(-\frac{\nu}{2} - 1\right)\sum_{p=1}^{P}\log\sigma_p^2 - \sum_{p=1}^{P}\frac{\nu S^2}{2\sigma_p^2}$$
$$+\left[\sum_{p=1}^{P}\gamma_p\right]\log\kappa + \left[P - \sum_{p=1}^{P}\gamma_p\right]\log\left(1-\kappa\right) + Const.$$

となる。

α_j の近似事後分布は正規分布となり、

$$E\left[\alpha_j\right] = \Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i - \sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik} - \sum_{p=1}^{P}E\left[\gamma_p\right]E\left[\beta_p\right]x_{ip}\right)$$

また

$$V\left[\alpha_j\right] = \Lambda_j$$

となる。ここで$\Lambda_j^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} z_{ij}^2$である。

β_pの近似事後分布も正規分布となり、

$$E\left[\beta_p\right] = \mathrm{H}_p E\left[\gamma_p\right] E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{k\neq p}^{P} E\left[\gamma_k\right] E\left[\beta_k\right] x_{ik}\right)$$

また

$$V\left[\beta_p\right] = \mathrm{H}_p$$

となる。ここで$\mathrm{H}_p^{-1} = E\left[\gamma_p^2\right] E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}^2 + E\left[\frac{1}{\sigma_p^2}\right]$である。

σ_p^2の近似事後分布は尺度付き逆カイ二乗分布となり

$$E\left[\sigma_p^2\right] = \frac{\nu_p S_p^2}{\nu_p - 2}$$

及び

$$E\left[\frac{1}{\sigma_p^2}\right] = \frac{1}{S_p^2}$$

となる。ここで$\nu_p = \nu + 1$及び$S_p^2 = \frac{E\left[\beta_p^2\right] + \nu S^2}{\nu + 1}$である。

τ_0^2の近似事後分布はガンマ分布となり

$$E\left[\tau_0^2\right] = \frac{a_1}{b_1}$$

である。ここで$a_1 = \frac{N}{2}$及び

$$b_1 = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{p=1}^{P} E\left[\gamma_p\right] E\left[\beta_p\right] x_{ip}\right)^2$$
$$+ \frac{1}{2}\sum_{j=1}^{F} V\left[\alpha_j\right]\sum_{i=1}^{N} z_{ij}^2 + \frac{1}{2}\sum_{p=1}^{P}\left(E\left[\gamma_p^2\right] E\left[\beta_p^2\right] - E\left[\gamma_p\right]^2 E\left[\beta_p\right]^2\right)\sum_{i=1}^{N} x_{ip}^2$$

となる。

γ_pの近似事後分布はベルヌーイ分布となり、

$$E\left[\gamma_p\right] = \frac{\kappa \exp\left(R\right)}{\kappa \exp\left(R\right) + \left(1-\kappa\right)\exp\left(R^*\right)}$$

となる。ここで

$$R=-\frac{E\left[\tau_0^2\right]}{2}\left\{\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\gamma_k\right]E\left[\beta_k\right]x_{ik}-E\left[\beta_p\right]x_{ip}\right)^2+W_p+V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2\right\}$$

及び

$$R^*=-\frac{E\left[\tau_0^2\right]}{2}\left\{\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\gamma_k\right]E\left[\beta_k\right]x_{ik}\right)^2+W_p\right\}$$

であり、

$$W_p=\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{k\neq p}^{P}\left(E\left[\gamma_k^2\right]E\left[\beta_k^2\right]-E\left[\gamma_k\right]^2E\left[\beta_k\right]^2\right)\sum_{i=1}^{N}x_{ik}^2$$

となる。

データ周辺尤度の下限は以下のように表される。

$$\sum_{p=1}^{P}\left[-\frac{\nu_p}{2}\log\frac{\nu_pS_p^2}{2}+\log\Gamma\left(\frac{\nu_p}{2}\right)+\frac{1}{2}\log V\left[\beta_p\right]+E\left[\gamma_p\right]\log\frac{\kappa}{E\left[\gamma_p\right]}+\left(1-\kappa\right)\log\frac{\left(1-\kappa\right)}{\left(1-E\left[\gamma_p\right]\right)}\right]$$
$$-a_1\log b_1+\log\Gamma\left(a_1\right)+P\left[\frac{\nu}{2}\log\frac{\nu S^2}{2}-\log\Gamma\left(\frac{\nu}{2}\right)\right]+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]$$
$$-\frac{N-F}{2}\log 2\pi+\frac{P+F}{2}$$

2.3.4 BayesC

BayesC はマーカー効果の事前分布として正規分布と 0 の混合分布、いわゆる“spike and slab prior”を用いている。“spike and slab prior”を用いた線形回帰モデルは Ishwaran and Rao（2005）や後述の SSVS にも見ることができる。BayesC と同様の混合分布を用いたモデルに対する変分ベイズアルゴリズムは Carbonetto and Stephens（2012）で提案されている。Carbonetto and Stephens（2012）では重点サンプリングによりσ^2、τ_0^2、及びκを推定しているが、ここでは他のパラメータ同様にσ^2及びτ_0^2を変分ベイズ法で推定し、κをあらかじめ与えた値に固定するアルゴリズムを提案する。BayesC の対数同時事後分布は

$$\frac{N}{2}\log\tau_0^2-\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2+\sum_{p=1}^{P}\rho_p\left[-\frac{1}{2}\log\sigma^2-\frac{\beta_p^2}{2\sigma^2}\right]+\left(-\frac{\nu}{2}-1\right)\log\sigma^2-\frac{\nu S^2}{2\sigma^2}$$
$$+\left[\sum_{p=1}^{P}\rho_p\right]\log\kappa+\left[P-\sum_{p=1}^{P}\rho_p\right]\log(1-\kappa)+Const.$$

となる。

α_j の近似事後分布は正規分布となり、

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

及び

$$V\left[\alpha_j\right]=\Lambda_j$$

である。ここで $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$ となる。

β_p 及び ρ_p の同時事後分布は

$$q(\beta_p,\rho_p)\propto-\frac{E\left[\tau_0^2\right]}{2}\left[\beta_p^2\sum_{i=1}^{N}x_{ip}^2-2\beta_p\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}z_{ij}E\left[\alpha_j\right]-\sum_{k\neq p}^{P}x_{ik}E\left[\beta_k\right]\right)\right]$$
$$+\rho_p\left[\frac{1}{2}\Phi\left(\frac{\tilde{\nu}}{2}\right)-\frac{1}{2}\log\frac{\tilde{\nu}\tilde{S}^2}{2}-\frac{\beta_p^2}{2\tilde{S}^2}\right]+\rho_p\log\kappa+(1-\rho_p)\log(1-\kappa)$$

となる。ここで $\tilde{\nu}=\nu+\sum_{j=1}^{P}E\left[\rho_j\right]$、$\tilde{S}^2=\dfrac{\nu S^2+\sum_{j=1}^{P}E\left[\beta_j^2\right]}{\tilde{\nu}}$、となり Φ はディガンマ関数、つまり $\Phi(x)=\dfrac{d\Gamma(x)}{dx}\Gamma(x)^{-1}=\Gamma'(x)\Gamma(x)^{-1}$ を表す。このディガンマ関数は尺度付き逆カイ二乗分布、$\chi^{-2}\left(x;\nu,S^2\right)$、に関する積分、つまり

$$\int\chi^{-2}\left(x;\nu,S^2\right)\log x\,dx=-\Phi\left(\frac{\nu}{2}\right)+\log\frac{\nu S^2}{2}\quad(2.1)$$

から生じる。この導出は以下に示す通りである。尺度付き逆カイ二乗分布の確率密度関数 $q(x)$ は

$$q(x)=\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}x^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)$$

となる。ここで以下のように$q(x)$の一部を抜き出しνについて微分を行う。

$$\frac{d}{d\nu}\left[x^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)\right]=x^{-\frac{\nu}{2}-1}\left(-\frac{S^2}{2x}\right)\exp\left(-\frac{\nu S^2}{2x}\right)-\frac{1}{2}x^{-\frac{\nu}{2}-1}\log x\exp\left(-\frac{\nu S^2}{2x}\right)$$

この両辺に$D=\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}$をかけ$x$について積分を行うと、

$$\begin{aligned}D\frac{d}{d\nu}\left[x^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)\right]&=-\frac{S^2}{2}\int\frac{1}{x}Dx^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)dx-\frac{1}{2}\int\log xDx^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)dx\\&=-\frac{S^2}{2}\int q(x)\frac{1}{x}dx-\frac{1}{2}\int q(x)\log xdx\end{aligned}$$

となる。ここで$\int D^{-1}q(x)dx=D^{-1}$を用いて左辺を変形することで

$$D\frac{dD^{-1}}{d\nu}=-\frac{S^2}{2}\int q(x)\frac{1}{x}dx-\frac{1}{2}\int q(x)\log xdx$$

を得る。これから

$$\begin{aligned}\int q(x)\log xdx&=-2D\frac{dD^{-1}}{d\nu}-S^2\int q(x)\frac{1}{x}dx\\&=-2D\frac{dD^{-1}}{d\nu}-1\end{aligned}\qquad(2.2)$$

となる。D^{-1}のνに関する微分は以下のように解くことができる。

$$\begin{aligned}\frac{dD^{-1}}{d\nu}&=\frac{d}{d\nu}\left[\Gamma\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}\right]\\&=\frac{1}{2}\Gamma'\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}+\Gamma\left(\frac{\nu}{2}\right)\frac{d}{d\nu}\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}\\&=\frac{1}{2}\Gamma'\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}-\frac{1}{2}\Gamma\left(\frac{\nu}{2}\right)\left(\log\frac{\nu S^2}{2}+1\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}\end{aligned}$$

よって

$$
\begin{aligned}
\int q(x)\log x dx &= -2D\frac{dD^{-1}}{d\nu} - 1 \\
&= -2\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}\left[\frac{1}{2}\Gamma'\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}} - \frac{1}{2}\Gamma\left(\frac{\nu}{2}\right)\left(\log\frac{\nu S^2}{2}+1\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}\right] - 1 \\
&= -\Gamma\left(\frac{\nu}{2}\right)^{-1}\Gamma'\left(\frac{\nu}{2}\right) + \log\frac{\nu S^2}{2} + 1 - 1 \\
&= -\Gamma\left(\frac{\nu}{2}\right)^{-1}\Gamma'\left(\frac{\nu}{2}\right) + \log\frac{\nu S^2}{2} \\
&= -\Phi\left(\frac{\nu}{2}\right) + \log\frac{\nu S^2}{2}
\end{aligned}
$$

を得る。

ρ_p の近似事後分布は β_p との同時事後分布を β_p について積分することで得られる。つまり

$$
\begin{aligned}
q(\rho_p = 1) &= \int q(\beta_p, \rho_p = 1) d\beta_p \\
&\propto \frac{H_p}{2}\left[E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}\left(y_i - \sum_{j=1}^{F} z_{ij}E\left[\alpha_j\right] - \sum_{k\neq p}^{P} x_{ik}E\left[\beta_k\right]\right)\right]^2 \\
&\quad + \frac{1}{2}\log H_p + \frac{1}{2}\Phi\left(\frac{\tilde{\nu}}{2}\right) - \frac{1}{2}\log\frac{\tilde{\nu}\tilde{S}^2}{2} + \log\kappa \\
&\propto F_p + \log\kappa
\end{aligned}
$$

となる。ここで $H_p^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}^2 + \frac{1}{\tilde{S}}$ である。同様に $q(\rho_p = 0) \propto \log(1-\kappa)$ が得られる。結果として

$$
E\left[\rho_p\right] = \frac{\kappa\exp\left(F_p\right)}{\kappa\exp\left(F_p\right) + (1-\kappa)}
$$

となる。

β_p の近似事後分布は ρ_p との同時事後分布を ρ_p について積分することで得られる。つまり

$$
\begin{aligned}
q(\beta_p) &= q(\beta_p, \gamma_p = 1) + q(\beta_p, \gamma_p = 0) \\
&= q(\beta_p \mid \gamma_p = 1)q(\gamma_p = 1) + q(\beta_p \mid \gamma_p = 0)q(\gamma_p = 0)
\end{aligned}
$$

となる。したがって、

$$E\left[\beta_p\right]=E\left[\beta_p \mid \rho_p=1\right]E\left[\rho_p\right]$$

及び

$$E\left[\beta_p^2\right]=E\left[\beta_p^2 \mid \rho_p=1\right]E\left[\rho_p\right]$$

となる。$q\left(\beta_p \mid \rho_p=1\right)$は正規分布となり

$$E\left[\beta_p \mid \rho_p=1\right]=\mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

及び

$$V\left[\beta_p \mid \rho_p=1\right]=\mathrm{H}_p$$

となる。また

$$E\left[\beta_p^2 \mid \rho_p=1\right]=V\left[\beta_p \mid \rho_p=1\right]+E\left[\beta_p \mid \rho_p=1\right]^2$$

となる。

σ^2の近似事後分布は尺度付き逆カイ二乗分布となり、

$$E\left[\frac{1}{\sigma^2}\right]=\frac{1}{\tilde{S}^2}$$

となる。またτ_0^2の近似事後分布はガンマ分布となり、

$$E\left[\tau_0^2\right]=\frac{a_1}{b_1}$$

となる。ここで$a_1=\frac{N}{2}$及び

$$b_1=\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)^2+\frac{1}{2}\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2$$
$$+\frac{1}{2}\sum_{p=1}^{P}E\left[\rho_p\right]\left(E\left[\beta_p^2 \mid \rho_p=1\right]-E\left[\rho_p\right]E\left[\beta_p \mid \rho_p=1\right]^2\right)\sum_{i=1}^{N}x_{ip}^2$$

である。データ周辺尤度の下限は

$$\sum_{p=1}^{P}\left[\frac{E\left[\rho_p\right]}{2}\log V\left[\beta_p \mid \rho_p=1\right]+E\left[\rho_p\right]\log\frac{\kappa}{E\left[\rho_p\right]}+(1-\kappa)\log\frac{(1-\kappa)}{\left(1-E\left[\rho_p\right]\right)}\right]$$
$$-a_1\log b_1+\log\Gamma(a_1)+\frac{\nu}{2}\log\frac{\nu S^2}{2}-\log\Gamma\left(\frac{\nu}{2}\right)-\frac{\tilde{\nu}}{2}\log\frac{\tilde{\nu}\tilde{S}^2}{2}+\log\Gamma\left(\frac{\tilde{\nu}}{2}\right)+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]$$
$$-\frac{N-F}{2}\log 2\pi+\frac{\sum_{p=1}^{P}E\left[\rho_p\right]+F}{2}$$

となる。

2.3.5 Stochastic search variable selection（SSVS）

SSVS は BayesC 同様に“spike and slab prior”をマーカー効果に用いるが、その混合分布は 2 つの正規分布からなり、一方の正規分布の分散が他方の c 倍で固定されている（表 2.1）。この c を 1 以下の 0 に近い値に設定することにより、被説明変数への貢献が小さい回帰係数を 0 に向けて縮約させる。ただし BayesC はマーカー効果が 0 となることを許容しているのに対し、SSVS は完全に 0 にはしない点に特徴がある。SSVS に対する変分ベイズ法は提案されておらず、本研究で導出を行った。対数同時事後分布は

$$\frac{N}{2}\log\tau_0^2-\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2-\frac{1}{2}\left[\sum_{p=1}^{P}\rho_p\right]\log\sigma^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\rho_p\beta_p^2}{\sigma^2}-\left[P-\sum_{p=1}^{P}\rho_p\right]\frac{1}{2}\log c\sigma^2-\frac{1}{2}\sum_{p=1}^{P}\frac{(1-\rho_p)\beta_p^2}{c\sigma^2}$$
$$+\left(-\frac{\nu}{2}-1\right)\log\sigma^2-\sum_{p=1}^{P}\frac{\nu S^2}{2\sigma^2}+\left(-\frac{\nu}{2}-1\right)\log c\sigma^2-\sum_{p=1}^{P}\frac{\nu S^2}{2c\sigma^2}$$
$$+\left[\sum_{p=1}^{P}\rho_p\right]\log\kappa+\left[P-\sum_{p=1}^{P}\rho_p\right]\log(1-\kappa)+Const.$$

となる。

α_j の近似事後分布は正規分布となり、

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

及び

$$V\left[\alpha_j\right]=\Lambda_j$$

となる。ここで $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$ である。

β_p の近似事後分布も正規分布となり、

$$E\left[\beta_p\right]=\mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

及び

$$V\left[\beta_p\right]=\mathrm{H}_p$$

となる。ここで $\mathrm{H}_p^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+E\left[\frac{1}{\sigma^2}\right]\left[E\left[\rho_p\right]\left(1-\frac{1}{c}\right)+\frac{1}{c}\right]$ である。

σ^2 の近似事後分布は尺度付き逆カイ二乗分布となり、その自由度は $\tilde{\nu}=\nu+P$、尺度パラメータは

$$\tilde{S}^2=\frac{\sum_{p=1}^{P}E\left[\rho_p\right]E\left[\beta_p^2\right]+\frac{1}{c}\sum_{p=1}^{P}\left(1-E\left[\rho_p\right]\right)E\left[\beta_p^2\right]+\nu S^2}{\nu+P}$$

となる。また

$$E\left[\frac{1}{\sigma^2}\right]=\frac{1}{\tilde{S}^2}$$

となる。

ρ_p の近似事後分布はベルヌーイ分布となり、

$$E\left[\rho_p\right]=\frac{\kappa\exp\left(-\frac{1}{2}E\left[\frac{1}{\sigma^2}\right]E\left[\beta_p^2\right]\right)}{\kappa\exp\left(-\frac{1}{2}E\left[\frac{1}{\sigma^2}\right]E\left[\beta_p^2\right]\right)+\frac{(1-\kappa)}{\sqrt{c}}\exp\left(-\frac{1}{2c}E\left[\frac{1}{\sigma^2}\right]E\left[\beta_p^2\right]\right)}$$

となる。

τ_0^2 の近似事後分布はガンマ分布であり

$$E\left[\tau_0^2\right]=\frac{N}{b}$$

となる。ここで

$$b=\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)^2+\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{p=1}^{P}V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2$$

である。データ周辺尤度の下限は

$$\sum_{p=1}^{P}\left[\frac{1}{2}\log V\left[\beta_p\right]+E\left[\rho_p\right]\log\frac{\kappa}{E\left[\rho_p\right]}+\left(1-\kappa\right)\log\frac{\left(1-\kappa\right)}{\left(1-E\left[\rho_p\right]\right)}\right]$$
$$-a_1\log b_1+\log\Gamma\left(a_1\right)+\frac{\nu}{2}\log\frac{\nu S^2}{2}-\log\Gamma\left(\frac{\nu}{2}\right)-\frac{\tilde{\nu}}{2}\log\frac{\tilde{\nu}\tilde{S}^2}{2}+\log\Gamma\left(\frac{\tilde{\nu}}{2}\right)+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]$$
$$-\frac{N-F}{2}\log 2\pi+\frac{P+F}{2}-\frac{1}{2}\left(P-\sum_{p=1}^{P}E\left[\rho_p\right]\right)\log c$$

となる。

2.3.6 Bayesian mixture regression（MIX）

MIX を提案した Luan et al.（2009）ではマーカー効果の事前分布は 2 つの正規分布の混合分布とし、それぞれが効果の大きな及び小さなマーカーの事前分布として働くことをデータから学習させている。本研究ではマーカー効果の大きさによる事前分布への振り分けをより効率的にするために、正規分布の分散の事前分布（尺度付き逆カイ二乗分布）に、分散の期待値が両者で異なるように重み c を加えるように変更した（表 2.1）。c を 0 に近づけることで一方の正規分布の分散の事前期待値が小さくなり、マーカー効果を 0 に向けて縮約することが期待できる。ただし SSVS 同様にマーカー効果を完全に 0 にしないことが特徴となる。MIX に対する変分ベイズ法は提案されておらず、本研究で導出した。対数同時事後分布は

$$\frac{N}{2}\log\tau_0^2-\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2-\frac{1}{2}\left[\sum_{p=1}^{P}\rho_p\right]\log\sigma_A^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\rho_p\beta_p^2}{\sigma_A^2}-\left[P-\sum_{p=1}^{P}\rho_p\right]\frac{1}{2}\log\sigma_B^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\left(1-\rho_p\right)\beta_p^2}{\sigma_B^2}$$
$$+\left(-\frac{\nu}{2}-1\right)\log\sigma_A^2-\sum_{p=1}^{P}\frac{\nu S^2}{2\sigma_A^2}+\left(-\frac{\nu}{2}-1\right)\log\sigma_B^2-\sum_{p=1}^{P}\frac{\nu cS^2}{2\sigma_B^2}$$
$$+\left[\sum_{p=1}^{P}\rho_p\right]\log\kappa+\left[P-\sum_{p=1}^{P}\rho_p\right]\log\left(1-\kappa\right)+Const.$$

となる。

α_j の近似事後分布は正規分布となり、

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

及び

$$V\left[\alpha_j\right]=\Lambda_j$$

である。ここで$\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$となる。

β_pの近似事後分布も正規分布となり、

$$E\left[\beta_p\right]=\mathrm{H}_pE\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

及び

$$V\left[\beta_p\right]=\mathrm{H}_p$$

となる。ここで$\mathrm{H}_p^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+E\left[\frac{1}{\sigma_A^2}\right]E\left[\rho_p\right]+E\left[\frac{1}{\sigma_B^2}\right]\left(1-E\left[\rho_p\right]\right)$である。

σ_A^2及びσ_B^2の近似事後分布はいずれも尺度付き逆カイ二乗分布となる。σ_A^2については自由度が

$$\tilde{\nu}_A=\nu+\sum_{p=1}^{P}E\left[\rho_p\right]$$

尺度パラメータが

$$\tilde{S}_A^2=\frac{\sum_{p=1}^{P}E\left[\rho_p\right]E\left[\beta_p^2\right]+\nu S^2}{\nu+\sum_{p=1}^{P}E\left[\rho_p\right]}$$

となり、

$$E\left[\frac{1}{\sigma_A^2}\right]=\frac{1}{\tilde{S}_A^2}$$

となる。σ_B^2については自由度が

$$\tilde{\nu}_B=\nu+P-\sum_{p=1}^{P}E\left[\rho_p\right]$$

尺度パラメータが

$$\tilde{S}_B^2=\frac{\sum_{p=1}^{P}\left(1-E\left[\rho_p\right]\right)E\left[\beta_p^2\right]+\nu cS^2}{\nu+P-\sum_{p=1}^{P}E\left[\rho_p\right]}$$

となり、

$$E\left[\frac{1}{\sigma_B^2}\right]=\frac{1}{\tilde{S}_B^2}$$

となる。

ρ_p の近似事後分布はベルヌーイ分布となり、

$$E\left[\rho_p\right]=\frac{\kappa D}{\kappa D+(1-\kappa)D^*}$$

となる。ここで

$$D=\sqrt{\frac{2}{\sum_{p=1}^{P}E\left[\rho_p\right]E\left[\beta_p^2\right]+\nu S^2}}\exp\left\{\Phi\left[\frac{1}{2}\left(\nu+\sum_{p=1}^{P}E\left[\rho_p\right]\right)\right]-\frac{1}{2}E\left[\beta_p^2\right]E\left[\frac{1}{\sigma_A^2}\right]\right\}$$

及び、

$$D^*=\sqrt{\frac{2}{\sum_{p=1}^{P}\left(1-E\left[\rho_p\right]\right)E\left[\beta_p^2\right]+\nu c S^2}}\exp\left\{\Phi\left[\frac{1}{2}\left(\nu+P-\sum_{p=1}^{P}E\left[\rho_p\right]\right)\right]-\frac{1}{2}E\left[\beta_p^2\right]E\left[\frac{1}{\sigma_B^2}\right]\right\}$$

である。

τ_0^2 の近似事後分布はガンマ分布であり

$$E\left[\tau_0^2\right]=\frac{N}{b}$$

となる。ここで

$$b=\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)^2+\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{p=1}^{P}V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2$$

である

データ周辺尤度の下限は

$$\begin{aligned}
&\sum_{p=1}^{P}\left[\frac{1}{2}\log V\left[\beta_p\right]+E\left[\rho_p\right]\log\frac{\kappa}{E\left[\rho_p\right]}+(1-\kappa)\log\frac{(1-\kappa)}{\left(1-E\left[\rho_p\right]\right)}\right]\\
&-a_1\log b_1+\log\Gamma\left(a_1\right)-\frac{\tilde{\nu}_A}{2}\log\frac{\tilde{\nu}_A\tilde{S}_A^2}{2}+\log\Gamma\left(\frac{\tilde{\nu}_A}{2}\right)-\frac{\tilde{\nu}_B}{2}\log\frac{\tilde{\nu}_B\tilde{S}_B^2}{2}+\log\Gamma\left(\frac{\tilde{\nu}_B}{2}\right)\\
&+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]+\nu\log\frac{\nu S^2}{2}-2\log\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\log c\\
&-\frac{N-F}{2}\log 2\pi+\frac{P+F}{2}
\end{aligned}$$

となる。

2.3.7　収束の判定

変分ベイズ法、パラメータの期待値、分散、または最尤推定値を相互に初期値から収束するまで繰り返し計算を行う。回帰係数（α_j 及び β_p）の初期値は 0 とし、τ_0^2 の初期値は$100/V[\boldsymbol{y}]$とした。ここで$V[\boldsymbol{y}]$は表現型値の分散を表す。試行は

$$\frac{\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|^2}{\left\|\boldsymbol{\theta}^*\right\|^2} < 10^{-m}$$

となるまで繰り返した。ここで‖ ‖はユークリディアンノルム、$\boldsymbol{\theta}$ は全てのパラメータ値を含むベクター、$\boldsymbol{\theta}^*$はその試行において新たに計算された値を含むベクター、m は収束判定の基準を表し、VIGoR の規定値は 9 とした。

2.3.8　ハイパーパラメータ

VIGoR で実装したベイズ回帰手法にはいずれもハイパーパラメータを与える必要がある。BayesC については全マーカーのうち 0 でない効果を持つマーカーの割合（κ）及び表現型分散のうちマーカーが説明する分散（σ_m^2）についての想定値のもとで、ハイパーパラメータを決定する方法が提案されている（Habier et al. 2011）。本研究ではその方法を他の回帰手法にも適用した。

VIGoR では被説明変数（つまり表現型値）を標準化するためにσ_m^2は 1 以下の値をとる。マーカーj の効果の分散をσ_j^2 とするとマーカー間の連鎖平衡を想定した場合、

$$\sigma_m^2 = \kappa \sum_{j=1}^{P} \sigma_j^2 2(1+f)\, p_j \left(1-p_j\right)$$

と近似できる（Habier et al. 2011）。ここで f は近交係数を、p_j はマーカーj のアリル頻度を表す。近交係数とアリル頻度の積はマーカー遺伝子型のサンプリングにより生じる分散を表している。Blasso ではσ_j^2 を$\frac{1}{\tau_j^2}\left(1-\sigma_m^2\right)$と表すことができるため（表 2.1）、

$$\sigma_m^2 = \kappa \sum_{j=1}^{P} \frac{1}{\tau_j^2}\left(1-\sigma_m^2\right) 2(1+f)\, p_j \left(1-p_j\right)$$

を得る。マーカー効果分散及び λ^2 の期待値はそれぞれ$E\left[\frac{1}{\tau_j^2}\right] = \frac{2}{\lambda^2}$ 及び$\frac{\varphi}{\omega}$ と表されるため、

$$\varpi = \frac{\varphi}{4\kappa(1+f)\sum_{j}^{P} p_j(1-p_j)\left(\frac{1}{\sigma_m^2}-1\right)}$$

となる。これから、κ、σ_m^2、及びφを与えることでωを決定することができる。EBlassoの場合も同様に考えて

$$\theta = \frac{\psi\varphi}{4\kappa\varpi(1+f)\sum_{j}^{P} p_j(1-p_j)\left(\frac{1}{\sigma_m^2}-1\right)}$$

を得る。

wBSRとBayesCではマーカー効果分散の期待値が$\frac{\nu S^2}{(\nu-2)}$であるため、S^2は

$$S^2 = \frac{(\nu-2)\sigma_m^2}{\nu\kappa(1+f)\sum_{j=1}^{P} 2p_j(1-p_j)} \quad (2.1)$$

と表すことができる。

SSVSではσ_m^2は

$$\sigma_m^2 = \sigma^2\sum_{j\in G_1} 2(1+f)p_j(1-p_j) + c\sigma^2\sum_{j\in G_2} 2(1+f)p_j(1-p_j)$$

と表すことができる。ここでG_1及びG_2はそれぞれ大きな及び小さな分散の事前正規分布に振り分けられたマーカーのグループを表す。G_1及びG_2の大きさ、つまり含まれるマーカー数の期待値はκP及び(1−κ)Pであるため、

$$\sigma_m^2 a = \sigma^2\kappa\sum_{j=1}^{P} 2(1+f)p_j(1-p_j)$$

$$\sigma_m^2(1-a) = c\sigma^2(1-\kappa)\sum_{j=1}^{P} 2(1+f)p_j(1-p_j)$$

と表すことができる。ここでaは大きな分散の事前正規分布に振り分けられたマーカーが説明できる分散が、マーカー全体が説明する分散に占める割合である。この等式をcについて解くと

$$c = \frac{1-a}{a}\frac{\kappa}{1-\kappa}$$

となる。このcを用いるとS^2は

$$S^2 = \frac{(\nu-2)\sigma_m^2}{\nu\left[\kappa + c(1-\kappa)\right](1+f)\sum_{j=1}^{P} 2p_j(1-p_j)}$$

と表すことができる。c 及び S^2 に関する同じ式が MIX においても導くことができる。

以上の式で Blasso の ω、EBlasso の θ、wBSR と BayesC の S^2、SSVS と MIX の c 及び S^2 が決定できる。このためには f、σ_m^2、κ、及び a（SSVS と MIX のみ）に加えて、その他のハイパーパラメータ、つまり Blasso では φ、EBlasso では φ、ω、及び ψ、wBSR、BayesC、SSVS、及び MIX では ν を与える必要がある。これらのハイパーパラメータは結果に与える影響が比較的小さいためデータセットに依然せず決定できることが経験上明らかになっている（2.5　シミュレーション解析参照）。f については栽培イネのような自殖性作物や近交系マウスなどでは 1 とし、他殖性の場合は 0 とする。なお以上のハイパーパラメータの計算は VIGoR の R パッケージで提供されている関数 *hyperpara* を用いて実行可能である。

2.4　VIGoR の主機能

コマンドラインプログラム及び R パッケージいずれにおいても、VIGoR は 3 つの主機能、*Estimation*、*Tuning&Estimation*、及び *Evaluation* を提供する（図 2.1）。1 つ目の機能はゲノムワイド関連解析に、2 つ目及び 3 つ目の機能はゲノムワイド予測のために設計されている。コマンドラインプログラムではデフォルトの機能は *Estimation* に設定されており、オプション「-t」及び「-e」によってそれぞれ *Tuning&Estimation* 及び *Evaluation* が実行できる（図 2.1 及び図 2.2）。R パッケージではそれぞれの機能について異なる関数を提供している（図 2.2）。*Estimation* 機能では VIGoR は与えられたハイパーパラメータのもとで選択された回帰手法をデータにあてはめ、マーカー効果を推定する。複数のハイパーパラメータ値が与えられた場合は、それぞれの値のもとでのあてはめを自動的に繰り返す。関連解析において必要なマーカー効果の有意性の判定は変分ベイズ法であれば推定された事後不確実性によってなされる。加えてマーカー効果の帰無分布を得るための並べ替え検定も実行できる。*Tuning&Estimation* 機能はハイパーパラメータ値が複数与えられた時のみ有効となる。この機能では VIGoR は各ハイパーパラメータ値のもとで交差検証を行い、最も小さい平均最小二乗誤差が得られたハイパーパラメータを用いてマーカー効果の推定を行う。この機構は利用できるデータから最も予測誤差の小さいモデルを構築することを意図している。*Evaluation* 機能では、VIGoR は交差検証を行い予測値を返す。複数のハイパーパラメータ

値が与えられた場合は、各交差検証の分割内でさらに交差検証を行い、最も平均最小二乗誤差が小さいハイパーパラメータをその分割において用いる。交差検証におけるサンプルの分割パターンはユーザーが指定することも可能である。これにより異なる手法を同一の分割パターンのもとで比較することができる。指定されない場合は与えられた分割数のもとでランダムに分割を行うが、そのときの分割パターンをファイル（コマンドラインプログラム）あるいはオブジェクト（R パッケージ）として出力するため、その後の解析には同じ分割パターンで交差検証を行うことが可能である。

コマンドラインプログラムは表現型値やマーカー遺伝子型値について独自の入力ファイル形式を持つが、一方で PLINK（Purcell et al. 2007）の入力ファイルである PED ファイル形式も用いることができる。またマーカー遺伝子型補完ソフトウェア Beagle（Browning and Browning 2007）の出力ファイルである dose ファイル形式もマーカー遺伝子型ファイルの代わりに使用可能である。

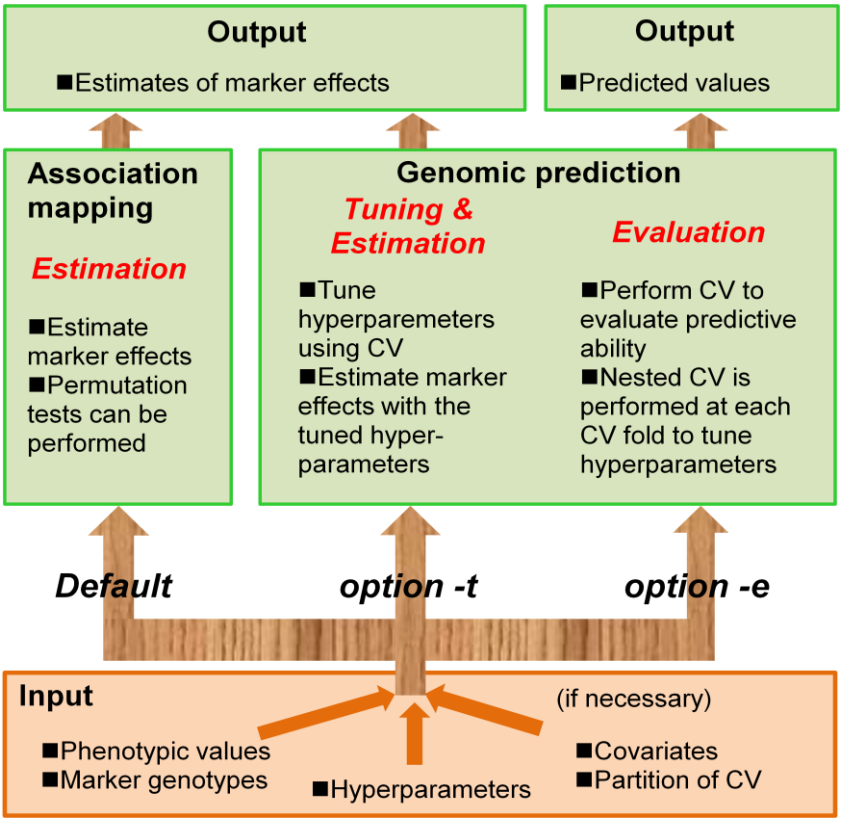

図 2.1　VIGoR のアナリシスフロー。VIGoR は入力情報として表現型値、マーカー遺伝子型、ハイパーパラメータ値を要求する。必要ならばマーカー遺伝子型以外の共変量や交差検証における分割方法も指定できる。VIGoR は 3 つの主機能、*Estimation*、*Tuning&Estimation*、*Evaluation* を提供する。*Estimation* 機能はゲノムワイド関連解析に、残りの 2 つはゲノムワイド予測のために設計されている。*Estimation* 機能では並び替え検定（permutation test）を実行可能である。コマンドラインプログラムでは *Estimation* 機能が既定であり、オプション *t* または *e* により機能選択できる。R パッケージではそれぞれの機能に異なる R 関数を提供している。

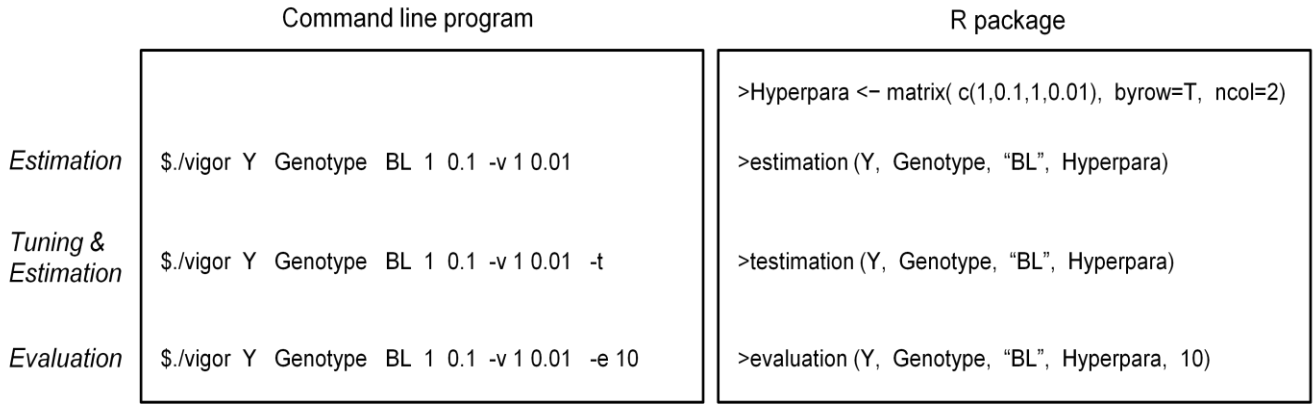

図 2.2　VIGoR の使用例。Y 及び Genotype はそれぞれ表現型値（被説明変数）及びマーカー遺伝子型を格納したファイル名（コマンドラインプログラム）または R オブジェクト（R パッケージ）を表す。この例では Bayesian lasso（BL）を用い 2 組のハイパーパラメータ値、φ = 1.0 と ω = 0.1、及び φ = 1.0 と ω = 0.01 を与えている。コマンドラインプログラム及び R 関数ともに表現型値（Y）、マーカー遺伝子型（Genotype）、回帰手法、及びハイパーパラメータを同じ順番で引数として取る。複数組のハイパーパラメータ値を与える場合はコマンドラインプログラムでは v オプションを用いて与える。R 関数では図中の Hyperpara オブジェクトのように行列として与える。コマンドラインプログラムでは *Tuning&Estimation* 及び *Evaluation* 機能は t 及び e オプションで選択する。「-e 10」は 10 倍の交差検証を表す。R パッケージはそれぞれの機能に対して *estimation*、*testimation*、*evaluation* の 3 つの関数を備える。

2.5　シミュレーション解析

2.5.1　シミュレーションデータの作成

このシミュレーションでは QTLMAS 第 15 回ワークショップで提供されたマーカー遺伝子型データを用いた（Elsen et al. 2012）。このデータは 3,220 個体から成り、そのうち 2,000 個体は学習用セットであり表現型値を持つ。これら 2,000 個体の後代である 1,000 個体は評価用セットで表現型値を持たない。残りの 220 個体はこれらの個体の祖先にあたる。合計 9,990 SNPs がシミュレートされているが、うち学習セットにおいて多型のあった 7121 マーカーを用いた。このマーカー遺伝子型を用いて表 2.2 にあるように 4 つのシナリオに従い表現型値を作成した。シナリオ I では学習用セットの個体数（Ntrain）を 2,000、つまりオリジナルのデータセットと同じとし、QTL の数（Nqtl）を 7、つまりマーカー数のおよそ 1,000 分の 1 とした。シナリオ II では Ntrain は 2,000、Nqtl は 70 とした。シナリオ III 及び IV では Ntrain は 200 とし、Nqtl はそれぞれ 7 及び 70 とした。シナリオ I 及び II において QTL はマーカーの中からランダムに選択し、その相加的効果は標準正規分布から生成した。シナリオ III 及び IV における QTL とその効果はそれぞれシナリオ I 及び II と共通とした。QTL として選択されたマーカーは回帰には用いなかった。従ってマーカー数はシナリオ I 及び III で 7,114、シナリオ II 及び IV で 7,051 であった。狭義の遺伝率、つまり表現型値分散に占める相加的遺伝分散の割合、は 0.5 とし、表現型値は遺伝子型値（つまり QTL 効果の和）にランダムにノイズを加えることで生成した。各シナリオを 20 反復繰り返した。QTL とその効果は各反復において新たにサンプリングし、表現型値もそれに応じて作成し直した。シナリオ III 及び IV において学習セットで用いる個体は各反復においてランダムに選択した。この結果、選択した 200 個体中で多型のないマーカーが生じることになるが、それはそのままデータ中に維持し使用した。ただし QTL には常に多型があることを確認した。

表 2.2　シミュレーションシナリオ [a]

Scenario	*Ntrain*[b]	*Nqtl*[c]
I	2,000	7
II	2,000	70
III	200	7
IV	200	70

[a] データセットは第 15 回 QTLMAS ワークショップにおける公開データセットをもとに、遺伝率 0.5 のもとで作成した

[b] 学習セットの大きさ

[c] QTL 数。QTL 効果は標準正規分布から生成した。

2.5.2　ハイパーパラメータの選択

「2.3.8　ハイパーパラメータ」で示した方法を用いて Blasso の ω、EBlasso の θ、wBSR と BayesC の S^2、SSVS と MIX の c 及び S^2 を決定した。σ_m^2 は 0.5、つまりシミュレーションで想定した遺伝率とした。計算に必要な κ、a（SSVS と MIX のみ）、及びその他のハイパーパラメータの値は表 2.3 に示したものを用いた。また f は 0 とした。

表 2.3　シミュレーション解析において与えたハイパーパラメータ値

	φ	ω	ψ	ν	a	κ
Blasso	1					10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.0
EBlasso	0.1	0.1	1			10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.0
wBSR				4		10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.0
BayesC				4		10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.0
SSVS				4	0.1, 0.3, 0.5, 0.7, 0.9	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
MIX				4	0.1, 0.3, 0.5, 0.7, 0.9	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}

2.5.3　ゲノムワイド予測

評価用セットの遺伝子型値を予測するために、VIGoR の *Tuning&Estimation* 機能を用いた。Blasso、EBlasso、wBSR、及び BayesC については表 2.3 に示した 5 通りの κ を試した。SSVS と MIX については表 2.3 に示した 4 通りの κ と 5 通りの *a* の計 20 組み合わせを試した。*Tuning&Estimation* 機能によりこれら複数の値から、5 分割の交差検証により最も小さい平均二乗誤差を示した値が自動的に選択され、最終的なマーカー効果の推定に用いられた。交差検証での分割はランダムに行った。予測の結果を表 2.4 に示した。EBlasso が概ね最も高いかそれに近い正確さを示した一方、SSVS と MIX は他より劣る傾向があった。SSVS と MIX の低調な結果はおそらく、これらの手法が効果 0 のマーカーを想定していないにも関わらず（表 2.1）、シミュレーションでは多くのマーカー効果を 0 としている点に原因があると考えられた。そのため、これらの手法はより複雑な形質、つまり主働遺伝子に加えてある程度の微働遺伝子効果（多数の小さな QTL 効果の和）が寄与する形質においてより有用であると考えられる。この点はまた後述するゲノムワイド関連解析におけるこれらの手法の低調な結果の理由でもあると考えられた。

表 2.4　各シナリオにおける予測の正確さの平均（標準偏差）[a]

	I	II	III	IV
Blasso	0.92 (0.02)	0.90 (0.01)	0.62 (0.06)	0.64 (0.06)
EBlasso	0.95 (0.02)	0.91 (0.01)	0.74 (0.10)	0.63 (0.06)
wBSR	0.93 (0.03)	0.87 (0.02)	0.64 (0.13)	0.61 (0.07)
BayesC	0.92 (0.04)	0.85 (0.03)	0.68 (0.13)	0.60 (0.07)
SSVS	0.90 (0.04)	0.84 (0.03)	0.58 (0.09)	0.54 (0.08)
MIX	0.84 (0.04)	0.82 (0.03)	0.58 (0.05)	0.61 (0.06)
Average[b]	0.94 (0.02)	0.90 (0.01)	0.74 (0.09)	0.64 (0.06)

[a] 予測の正確さは予測値と評価用セットにおける真の遺伝子型値とのピアソン相関係数として計測した。反復は 20 回行った。

[b]Average は 6 手法の予測値の相加平均による予測を表す

2.5.4　ゲノムワイド関連解析

ここでは *Estimation* 機能を用いた。いずれの手法でも 1 通りのハイパーパラメータの値を用いた。κ はいずれの手法、シナリオにおいても 10^{-3} とした。これは QTL 数を約 7 と想定していることになる。SSVS と MIX についてはいずれのシナリオでも *a* を 0.9 とした。この設定は 7 つの QTL が全マーカーが説明する表現型分散の 90 %を説明すると想定していることになる。

パラメータ推定手法は変分ベイズ法を用いた。変分ベイズ法で推定されるマーカー効果の事後不確実性がマーカー効果の有意性判定においてどの程度信頼できるのか調べるために、並べ替え検定と比較を行った。事後不確実性による判定及び並べ替え検定いずれでも *P* 値は 0.05 とした。事後不確実性による判定は範囲、

$$\left[E\left[\beta_p\right]-1.96\times\sqrt{V\left[\beta_p\right]},\quad E\left[\beta_p\right]+1.96\times\sqrt{V\left[\beta_p\right]}\right]$$

が 0 を含まなければ有意とみなした。wBSR の場合は $E\left[\beta_p\right]$ を $E\left[\beta_p\right]E\left[\gamma_p\right]$、$V\left[\beta_p\right]$ を $V\left[\gamma_p\right]E\left[\beta_p^2\right]+E\left[\gamma_p\right]^2V\left[\beta_p\right]$ と置き換えた。後者は

$$\begin{aligned}V\left[\gamma_p\beta_p\right]&=E\left[\left(\gamma_p\beta_p\right)^2\right]-E\left[\gamma_p\beta_p\right]^2\\&=E\left[\gamma_p^2\right]E\left[\beta_p^2\right]-E\left[\gamma_p\right]^2E\left[\beta_p\right]^2\\&=\left(V\left[\gamma_p\right]+E\left[\gamma_p\right]^2\right)\left(V\left[\beta_p\right]+E\left[\beta_p\right]^2\right)-E\left[\gamma_p\right]^2E\left[\beta_p\right]^2\\&=V\left[\gamma_p\right]E\left[\beta_p^2\right]+E\left[\gamma_p\right]^2V\left[\beta_p\right]\end{aligned}$$

から導かれる。並び替え検定は以下のように行った。表現型値をマーカー遺伝子型に対してシナリオ I 及び II では 200 回、III 及び IV では 1000 回ランダムに並べ替え、それぞれについて回帰手法をあてはめ、最も大きな効果（絶対値）を用いて帰無分布を作成した。検定したいマーカー効果の絶対値が帰無分布の 95 %分位点より上回った場合有意とみなした。wBSR についてはマーカー効果の代わりに $E\left[\beta_p\right]E\left[\gamma_p\right]$ を用いた。

EBlasso と wBSR、BayesC が概ね他より優れたマッピング成績を残した（表 2.5 及び表 2.6）。これらの手法については、検出された QTL 数及び偽陽性数に表されるように、不確実性による有意性判断と並べ替え検定はほぼ同等の結果を残した。Blasso については並び替え検定の方が良好な結果が得られた。MIX についてはいずれの方法でも満足のいく

結果は得られなかった。事後分布の不確実性が過小推定された場合、偽陽性数が並べ替え検定より増加する結果が予想されるが、そのような傾向は検出 QTL 数及び偽陽性数両方を増加させた SSVS 以外には観察されなかった。よって特に EBlasso、wBSR、BayesC において事後不確実性による判断は有用であることが示唆された。しかしながら、変分ベイズ法はマーカー効果の事後分布をマーカー間の依存関係、つまり連鎖不平衡を無視して因数分解し推定するために、もし今回用いたマーカー遺伝子型より連鎖不平衡が強い場合は、この結論は成り立たないとも考えることができる。一方で関連解析の場合はマーカー効果の解は疎になるように事前分布を用いて調節するために、マーカー間の依存関係は大きな影響を持たないということも考えられ、この点は Carbonetto and Stephens（2012）でも指摘されている。並べ替え検定は時間を消費するために、最初は事後不確実性により判定し、多くの陽性が得られた場合に並べ替え検定を実施するのが実用的であろう。

事後不確実性を用いて EBlasso、wBSR、BayesC いずれかで有意と判定されたマーカーを有意と判定した場合、検出された QTL 数及び偽陽性数は、いずれの手法及びシナリオにおいても増加した（表 2.5 及び表 2.6）。反対に事後不確実性を用いて EBlasso、wBSR、BayesC 全ての手法で有意と判定されたマーカーを有意と判定した場合、検出された QTL 数及び偽陽性数はいずれの手法及びシナリオにおいても減少した（表 2.5 及び表 2.6）。以上の結果から前者は QTL の検出がより困難な状況、例えば Ntrain が小さい場合などに有効であろう。一方で後者は確実な陽性シグナルを選定したい場合、または検出された QTL に優先順位を付けたい場合に有効と考えられる。

表 2.5　シナリオ I 及び II において事後不確実性（Un）及び並び替え検定（Pe）により正しく検出された QTL 数（N_Q）及び偽陽性数（N_F）[a]

		Scenario I		Scenario II	
	Test	N_Q	N_F	N_Q	N_F
Blasso	Un	3.6 (0.9)	0.4 (1.2)	6.3 (3.6)	0.0 (0.0)
	Pe	4.0 (0.9)	1.8 (2.4)	14.2 (3.0)	0.4 (0.7)
EBlasso	Un	4.8 (1.2)	0.8 (1.2)	20.1 (4.0)	0.7 (0.9)
	Pe	4.8 (1.2)	0.6 (1.0)	20.4 (3.8)	0.8 (0.9)
wBSR	Un	4.7 (1.1)	2.2 (1.9)	20.9 (3.5)	1.8 (1.4)
	Pe	4.7 (1.1)	2.3 (1.9)	20.9 (3.6)	1.8 (1.4)
BayesC	Un	5.1 (1.2)	3.5 (2.5)	26.9 (4.0)	2.9 (1.7)
	Pe	5.1 (1.2)	3.5 (2.5)	26.9 (4.0)	3.0 (1.7)
SSVS	Un	3.1 (1.2)	10.8 (4.0)	14.3 (3.5)	5.2 (2.8)
	Pe	2.5 (0.8)	1.1 (1.3)	5.8 (3.0)	0.8 (1.0)
MIX	Un	1.2 (0.7)	0.0 (0.2)	0.2 (0.5)	0.0 (0.0)
	Pe	7.0 (0.2)	1667 (691)	70.0 (0.0)	1311 (141)
Or (E, wB, BC)[b]	Un	5.3 (1.2)	5.5 (4.4)	30.8 (5.2)	4.1 (2.1)
And (E, wB, BC)[c]	Un	2.9 (1.0)	0.0 (0.0)	7.5 (4.1)	0.4 (0.7)

[a]Pe での試行回数は 200。Un 及び Pe ともに P 値は 0.05 を用いた。括弧内の数字は標準偏差を表す。

[b]EBlasso、wBSR、BayesC いずれかで有意と判定されたマーカーを有意とする

[c]EBlasso、wBSR、BayesC 全てで有意と判定されたマーカーを有意とする

表 2.6　シナリオ III 及び IV において事後不確実性（Un）及び並び替え検定（Pe）により正しく検出された QTL 数（N_Q）及び偽陽性数（N_F）[a]

		Scenario III		Scenario IV	
	Test	N_Q[b]	N_F[c]	N_Q	N_F
Blasso	Un	0.0 (0.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Pe	1.4 (0.7)	0.2 (0.5)	0.8 (1.5)	0.0 (0.2)
EBlasso	Un	1.7 (1.0)	0.3 (0.6)	2.1 (2.2)	0.0 (0.2)
	Pe	1.6 (0.9)	0.1 (0.3)	1.4 (1.8)	0.0 (0.0)
wBSR	Un	1.6 (1.1)	0.2 (0.4)	1.7 (1.9)	0.0 (0.2)
	Pe	1.6 (1.1)	0.3 (0.5)	1.8 (1.8)	0.2 (0.4)
BayesC	Un	1.8 (1.1)	0.2 (0.6)	2.9 (2.6)	0.2 (0.4)
	Pe	1.8 (1.1)	0.4 (0.7)	3.0 (2.6)	0.2 (0.4)
SSVS	Un	1.4 (1.1)	4.5 (2.3)	4.8 (3.3)	3.4 (2.2)
	Pe	0.3 (0.6)	0.0 (0.2)	0.4 (1.2)	0.1 (0.3)
MIX	Un	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Pe	3.5 (1.6)	82.0 (121.5)	28.9 (16.0)	49.4 (47.8)
Or (E, wB, BC)[b]	Un	2.2 (1.2)	0.8 (1.2)	3.9 (2.7)	0.2 (0.6)
And (E, wB, BC)[c]	Un	0.6 (0.9)	0.0 (0.0)	0.6 (1.4)	0.0 (0.0)

[a]Pe での試行回数は 200。Un 及び Pe ともに P 値は 0.05 を用いた。括弧内の数字は標準偏差を表す。

[b]EBlasso、wBSR、BayesC いずれかで有意と判定されたマーカーを有意とする

[c]EBlasso、wBSR、BayesC 全てで有意と判定されたマーカーを有意とする

この関連解析におけるマーカー効果推定のための平均の計算時間を表 2.7 に示した(並べ替え検定の時間は含まない)。計算時間は主に収束までの試行数を反映しており、wBSR 及び BayesC は他の手法より少ない試行数で収束する傾向があった（結果非掲載)。

表 2.7　各シナリオでの関連解析における計算時間（秒）[a]

	Simulation scenario			
	I	II	III	IV
Blasso	311.9 (92.8)	315.9 (92.3)	9.5 (1.0)	7.9 (0.9)
EBlasso	206.3 (59.0)	202.2 (19.4)	17.1 (1.8)	17.0 (1.8)
wBSR	96.6 (60.5)	47.9 (20.5)	2.3 (0.7)	1.6 (0.4)
BayesC	71.3 (77.7)	37.8 (15.1)	1.8 (0.7)	1.4 (0.4)
SSVS	140.8 (56.7)	119.5 (23.5)	31.8 (6.5)	33.4 (6.8)
MIX	102.4 (16.8)	100.1 (9.0)	22.6 (4.8)	22.7 (6.7)

[a] 計算は 2.66 GHz Intel Xeron processor を搭載した Mac OS X ver. 10.6.8 により行った。反復数は 20 で括弧内の数字は標準偏差を表す。

2.6　摘要

本研究ではゲノムワイド関連解析及びゲノムワイド予測用のソフトウェア VIGoR（variational Bayesian inference for genome-wide regression）を新たに開発した。このソフトウェアの特徴は（1）6 つのベイズ回帰手法を実装していること、（2）マルコフ連鎖モンテカルロ法より高速な変分ベイズ法によりパラメータ推定を行うこと、（3）交差検証やハイパーパラメータの最適化などが自動化されていること、である。またシミュレーションによりゲノムワイド関連解析においてマーカー効果の事後不確実性が経験的に並べ替え検定と同程度に信頼できることを示した。

3　アジア栽培イネ（*Oryza sativa* L.）におけるゲノムワイド予測有用性の検証及び予測手法の適用範囲探索

3.1　諸論

ゲノミックセレクションにおける予測の正確さは直接的に遺伝的改良量に影響を与える。そのため、高い予測能力を持つ手法の開発は、ゲノミックセレクションの可能性を評価したこれまでの研究における主な目標の一つであった。これまで評価されてきた予測手法としては GBLUP やその拡張（VanRaden 2008; Aguilar et al. 2010; Christensen and Lund 2010）、リッジ回帰、Lasso、elastic net（ENet）などの正則化回帰手法（Usai et al. 2009; Li and Sillanpaa 2012b; Ogutu et al. 2012）、BayesA、BayesB などのベイズ回帰モデル（Meuwissen et al. 2001: de los Campos et al. 2009a; Hayashi and Iwata 2010、Habier et al. 2011）、非相加的効果を考慮するための reproducing kernel Hilbert space 回帰（RKHS）などのノンパラメトリック回帰手法（Gianola et al. 2006; Gianola and van Kaam 2008; Long et al. 2010; Ober et al. 2011）、機械学習の分野でよく用いられるサポートベクターマシンやランダムフォレスト（RForest）などの回帰及び分類手法（Long et al. 2011a; Ogutu et al. 2011）、次元削減を目的とした回帰手法（Solberg et al. 2009; Long et al. 2011b）など多岐にわたる。リッジ回帰あるいはそれと同等な GBLUP や、BayesA、BayesB、及び Blasso（Bayesian lasso、第 2 章参照）はよく注目を集める手法であり、多くの研究中で評価されている（de los Campos et al. 2013）。対照的に、Lasso（Tibshirani 1996）、ENet（Zou and Hastie 2005）、RForest（Breiman 2001）は、パターン認識や機械学習では一般的な手法であるが、ゲノミックセレクションの文脈で取り扱われる頻度は低い（de los Campos et al. 2013）。

主にシミュレーションを用いた手法の比較により、手法の相対的な予測能力に影響を与える要因が明らかにされつつある。少ない QTL に支配される形質では、BayesB や Lasso のような変数選択の特徴を持つ手法が、全マーカー（変数）が均等に遺伝分散に貢献すると想定するリッジ回帰や GBLUP のような手法より優れる傾向にある（Colautti et al. 2010; Daetwyler et al. 2010）。変数選択は連鎖不平衡（LD）により悪影響を受けるため、BayesB とリッジ回帰の相対的な予測の正確さはたとえ遺伝的構造（遺伝率や QTL 数）が同じであっても LD の程度により変化する（Wimmer et al. 2013）。リッジ回帰や GBLUP による予測

はマーカーと QTL の LD よりも系統間の遺伝的関係により強く依存するため、学習用セットからの遺伝的乖離の程度の高い系統の予測では BayesB に劣る傾向がある（Habier et al. 2007; Zhong et al. 2009）。RKHS などのノンパラメトリック回帰手法は非相加的な形質では線形モデルより優れる傾向がある（Long et al. 2010; Ober et al. 2011; Gonzalez-Camacho et al. 2012）。これらの要素（遺伝的構造、LD 構造、学習用セットと評価用セットの遺伝的関係）は集団や形質により変化することが考えられ、またこれらの要素間の複雑な相互作用も予測手法の相対的なパフォーマンスに影響を与えると考えられる。そのため、最適な予測手法の選択には、実際のデータを用いた検証が必要となるであろう。

実際、予測手法の比較は作物のデータを用いても行われてきた。例えばコムギ（Crossa et al. 2010; Heffner et al. 2011; Perez-Rodriguez et al. 2012）、トウモロコシ（Crossa et al. 2010; Albrecht et al. 2011; Riedelsheimer et al. 2012a; Zhao et al. 2012; Crossa et al. 2013）、オオムギ（Lorenz et al. 2012; Endelman et al. 2014）などである。Lorenzana and Bernardo（2009）は複数の手法をトウモロコシ、オオムギ、シロイヌナズナで比較した包括的な研究を行う一方、Heslot et al.（2012）らは同様にトウモロコシ、オオムギ、コムギ、シロイヌナズナを用いた比較研究を報告している（このうちいくつかの集団は Lorenzana and Bernardo（2009）と共通であった）。これらの研究は用いられた集団やそれに非常に近い集団のゲノムワイド予測については有用な知見を与える。しかしながら、実際の集団・形質では遺伝的構造が未知であり、手法間の差を生んだ要因が明らかでないことから、得られた結論を一般化しにくいという欠点を持つ。予測手法を正確に特徴づけるためには、実際のデータのみを用いた比較では不十分である。

アジア栽培イネは世界の主用な穀物の 1 つであるものの、コムギやトウモロコシなど前述の穀物と異なり、ゲノムワイド予測の有用性や予測手法選択の検証が未だ行われていない。そこで本研究はアジア栽培イネ品種の集団を用いて、複数のゲノムワイド予測手法の比較を行った。比較した手法は GBLUP、RKHS、Lasso、ENet、RForest、Blasso、EBlasso（第 2 章参照）、BayesA 及び BayesB に等価な wBSR（第 2 章参照）、そして全手法の平均値（Ave）である。さらに実データをもとに作成したシミュレーションデータにおいてもゲノムワイド予測手法の比較を行った。このシミュレーションにおいては、QTL 数（Nqtl）、学習用セットの大きさ（Ntrain）、遺伝率、エピスタシスの有無、さらに LD の範囲をシミュレーション条件として考慮した。このシミュレーションの目的は、１）シミュレーションにおいて各手法の適用可能な範囲を探索し、２）シミュレーション結果を用いて実デー

タで観察された予測手法間の正確さにおける差を生んだ要因を探ることである。シミュレーションは実際のマーカー遺伝子型から作成しているため、その結果は本研究で評価しなかった形質や将来品種数が増加した場合にも有用であろう。さらにシミュレーション結果は本研究で用いたイネ集団同様に、学習用セットが小さく、LD 構造が比較的強い集団におけるゲノムワイド予測にも有用な知見を与え得る。

3.2 材料と手法

3.2.1 イネ品種と表現型評価

主に日本で造成された 110 水稲品種を用いた（表 3.1）。手法比較は到穂日数（DH）、稈長（CL）、穂長（PL）、穂数（PN）、粒長（GL）、粒幅（GW）、玄米長（BL）、玄米幅（BW）の 8 形質について行った。栽培は独立行政法人農業・食品産業技術総合研究所（NARO）近畿中国四国農業研究センター（WARC、広島県福山市）で 2006 年から 2011 まで 6 年間行われた。DH は発芽から全体の半分の穂が出穂した日までの日数として計測された。CL は土壌表面から穂首までの長さ、PL は穂首から芒を含まない穂先までの長さとして計測された。PN は正常な穂の数として計測された。GL、GW、BL、BW はデジタルスライドゲージを用いて計測された。栽培試験及び表現型評価は NARO/WARC の出田収博士、農業生物資源研究所の江花薫博士、神戸大学大学院農学研究科附属食資源教育研究センターの吉岡拓磨氏及び山崎将紀准教授によって行われ本研究に提供された。表現型値は 6 年分を平均して用いた。平均値は予測手法の学習時には平均 0、標準偏差が 1 になるように標準化して用いた。

表 3.1　本研究に供したイネ品種

ID	Cultivar	Improved[a]	Year[b]
1	Aichinokaori	Improved	1988
2	Aikoku	Landrace	
3	Akage	Landrace	
4	Akanezora	Improved	1993
5	Akebono	Improved	1953
6	Akihikari	Improved	1976
7	Akitakomachi	Improved	1984
8	Asahi1	Landrace	
9	Asahi2	Landrace	
10	Asahinoyume	Improved	2000
11	Asanohikari	Improved	1988
12	Benisengoku	Improved	1953
13	Bozu	Landrace	
14	Chiyohonami	Improved	1987
15	Chiyonishiki	Improved	1986
16	Daichinokaze	Improved	2002
17	Domannaka	Improved	1993
18	Dontokoi	Improved	1997
19	Eiko	Improved	1939
20	Fujihikari	Improved	1977
21	Fujisaka5	Improved	1947
22	Fukei175	Improved	1993
23	Fukuhibiki	Improved	1995
24	Fusaotome	Improved	1999
25	Futaba	Improved	1940
26	Ginbozu	Landrace	
27	Gohyakumangoku	Improved	1957
28	Goropikari	Improved	1994
29	Haenuki	Improved	1993
30	Hanaechizen	Improved	1993
31	Haruru	Improved	2001
32	Hatsuboshi	Improved	1977
33	Hatsunishiki	Improved	1954
34	Hatsushimo	Improved	1950
35	Hatsushizuku	Improved	1998
36	Hayamasari	Improved	1990
37	Himenomochi	Improved	1972
38	Hinohikari	Improved	1989
39	Hitomebore	Improved	1992
40	Hiyokumochi	Improved	1971
41	Hohohonoho	Improved	1993
42	Hoshinoyume	Improved	2000
43	Hounen-wase	Improved	1955
44	Houyoku	Improved	1961
45	Itadaki	Improved	2003
46	Jikkoku	Landrace	
47	Joshu	Landrace	
48	Kameji	Landrace	
49	Kamenoo	Landrace	
50	Kamenoo4	Landrace	1915
51	Kihou	Improved	1968
52	Kinmaze	Improved	1948
53	Kinuhikari	Improved	1989
54	Kirara397	Improved	1990
55	Koganebare	Improved	1981
56	Koganemasari	Improved	1976
57	Koganemochi	Improved	1956
58	Koshihikari	Improved	1956
59	Koshiji-wase	Improved	1953

表 3.1　（続き）

60	Manamusume	Improved	2001
61	Matsuribare	Improved	1995
62	Menkoina	Improved	2001
63	Millenishiki	Improved	2003
64	Mineasahi	Improved	1981
65	Morita-wase	Landrace	1913
66	Mutsuhomare	Improved	1986
67	Nakate-shinsenbon	Improved	1950
68	Natsuhikari	Improved	1984
69	Nihonmasari	Improved	1973
70	Nipponbare	Improved	1963
71	Nishihomare	Improved	1979
72	Norin1	Improved	1931
73	Norin18	Improved	1941
74	Norin22	Improved	1943
75	Norin29	Improved	1945
76	Norin6	Improved	1936
77	Norin8	Improved	1937
78	Notohikari	Improved	1986
79	Ohba	Landrace	
80	Okiniiri	Improved	1996
81	Omachi	Landrace	
82	Otomemochi	Improved	1966
83	Ouu197	Improved	1937
84	Reiho	Improved	1969
85	Reimei	Improved	1966
86	Rikuu132	Improved	1921
87	Rikuu20	Landrace	1915
88	Sasanishiki	Improved	1963
89	Sasashigure	Improved	1952
90	Satojiman	Improved	2005
91	Sekitori	Landrace	
92	Senichi	Landrace	
93	Shinriki	Landrace	
94	Shirosenbon	Landrace	
95	Taichung65	Improved	1927
96	Takenari	Landrace	
97	Todoroki-wase	Improved	1968
98	Toyonishiki	Improved	1969
99	Tsugaruroman	Improved	2000
100	Tsukinohikari	Improved	1986
101	Yamabiko	Improved	1958
102	Yamadanishiki	Improved	1936
103	Yamasenishiki	Improved	1962
104	Yukara	Improved	1962
105	Yukihikari	Improved	1986
106	Yukinosei	Improved	1990
107	Yumeakari	Improved	2002
108	Yumehikari	Improved	1992
109	Yumehitachi	Improved	2000
110	Yumetsukushi	Improved	1995

[a]Improved 及び Landrace はそれぞれ改良された品種及び在来品種を表す

[b]Year は品種として確立された年を表す

3.2.2　マーカー遺伝子型データ

DNA は 1 つの品種につき典型的な 1 個体を選択し、CTAB 法（Murray and Thompson 1980）を用いて抽出した。ゲノムワイドな 3102 座位の 2 対立遺伝子マーカーについて遺伝子型が決定された。これらのうち、3071 座位は日本水稲品種のゲノムシークエンスから得られた SNP マーカーであり（Yamamoto et al. 2010、Nagasaki et al. 2010）、残り 31 マーカーは 2 対立遺伝子から成る単純反復配列（short sequence repeat）マーカーであった（Yamasaki and Ideta 2013）。すべてジェノタイピングは Illumina BeadStation 500G genotyper（Illumina Inc.、San Diego、CA、USA）を用い、マニュアルに従って行われた。DNA 抽出、マーカーの選択、及びジェノタイピングは江花薫博士、吉岡拓磨氏、及び山崎将紀准教授により行われ本研究に提供された。隣接するマーカー間の連鎖及び物理距離はそれぞれ 1.0^{-6} から 11.6 cM、0.079 から 2504 kb の範囲であり、およそ 10 cM が 2500 kb に相当していた。平均値（標準偏差）はそれぞれ 0.49（±0.76）cM 及び 122.3（±167.7）kb であった。平均のマイナーアリル頻度は 0.309（±0.124）であった。同じ染色体上にある全てのマーカーペア間の LD は 0 または 2 で表されたマーカー遺伝子型のピアソン相関係数（r^2）として計算した。

3.2.3　遺伝的集団構造

研究に用いた 110 品種の遺伝的集団構造は階層的クラスタリングを用いて行った。クラスタリングはマーカー遺伝子型から計算した品種間のユークリッド距離をもとに、R 関数 hclust を用いて行った。

3.2.4　予測手法

以下では本研究で比較した予測手法について説明する。全ての回帰モデルは切片を含むが簡略化のため以下の説明からは省いた。

3.2.4.1　GBLUP

GBLUP は血縁情報をもとにした BLUP における分散共分散行列をゲノムワイドマーカーから求めた関係行列 **G** に置き換えることで実行できる（VanRaden 2008）。以下の線形回帰モデルを用いた。

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \mathbf{e}$$

ここで **y** は表現型値を含むベクター、**Z** は計画行列、**u** は相加的効果であり、多変量正規分

布（multivariate normal distribution）、MVN $(\mathbf{u}|\mathbf{0}, \mathbf{G}\sigma_u^2)$ に従うと想定する。また **e** は残差であり MVN $(\mathbf{e}|\mathbf{0}, \mathbf{I_n}\sigma_e^2)$（$\mathbf{I_n}$ は次元数 n の単位行列、n は系統数）に従うと想定する。**u** の解は混合モデル方程式を用いて得ることができる（Henderson 1984）。

$$\mathbf{u} = \left(\mathbf{Z'Z} + \mathbf{G}^{-1}\lambda_G\right)^{-1}\mathbf{Z'y} \quad (3.1)$$

ここで λ_G はσ_e^2/σ_u^2を示す。分散成分（σ_u^2 及び σ_e^2）は REML により R パッケージ rrBLUP（ver. 4.2、Endelman 2011）を用いて推定した。非対角成分を縮約させた **G** の推定も同パッケージにより行った（Endelman and Jannink 2012）。

GBLUP はリッジ回帰に関連している。リッジ回帰では表現型値はマーカー遺伝子型に回帰される。

$$\mathbf{y} = \mathbf{ZX\beta} + \mathbf{e}$$

ここで **X** はマーカー遺伝子型であり、**β** は回帰係数、**Z** は計画行列である。リッジ回帰の目的関数は

$$\min_{\beta} \left\|\mathbf{y} - \mathbf{ZX\beta}\right\|^2 + \lambda_R\mathbf{\beta'\beta}$$

となる。ここでは$\|\ \|$はユークリディアンノルムを、λ_R は正則化パラメータを表す。$\lambda_R\mathbf{\beta'\beta}$ は L_2 正則化項を示す。**β** の解は

$$\mathbf{\beta} = \left(\mathbf{X'Z'ZX} + \lambda_R\mathbf{I_P}\right)^{-1}\mathbf{X'Z'y}$$

で表される。ここで *P* はマーカー数を表す。もし遺伝子型値の合計、つまり **Xβ** に興味があるならば、**β** の解を求める必要はなく、双対解を用いることで **β** の解を回避できる。**β** の解の式は以下のように書きかえることができる。

$$\begin{aligned}\mathbf{\beta} &= \frac{1}{\lambda_R}\mathbf{X'Z'}\left(\mathbf{y} - \mathbf{ZX\beta}\right)\\ &= \mathbf{X'Z'\alpha}\end{aligned}$$

ここで $\mathbf{\alpha}$ は$\frac{1}{\lambda_R}\left(\mathbf{y} - \mathbf{ZX\beta}\right)$を示す。$\mathbf{y} - \mathbf{ZX\beta}$は$\mathbf{y} - \mathbf{ZXX'Z'\alpha}$と表すことができるので、

$$\lambda_R\mathbf{\alpha} = \mathbf{y} - \mathbf{ZXX'Z'\alpha}$$

を得る。これから以下の **α** に関する等式が導かれる。

$$\mathbf{\alpha} = \left(\mathbf{ZXX'Z'} + \lambda_R\mathbf{I}_n\right)^{-1}\mathbf{y}$$

結果的に

$$\begin{aligned}\mathbf{X\beta} &= \mathbf{XX'Z'\alpha} \\ &= \mathbf{XX'Z'}\left(\mathbf{ZXX'Z'} + \lambda_R \mathbf{I_N}\right)^{-1}\mathbf{y}\end{aligned} \quad (3.2)$$

を得る。ここで式 3.1 は逆行列に関する公式を用いて以下のように展開できる。

$$\begin{aligned}\mathbf{u} &= \left(\mathbf{Z'Z} + \mathbf{G}^{-1}\lambda_G\right)^{-1}\mathbf{Z'y} \\ &= \left[\frac{\mathbf{G}}{\lambda_G} - \frac{\mathbf{G}}{\lambda_G}\mathbf{Z'}\left(\mathbf{I_N} + \frac{1}{\lambda_G}\mathbf{ZGZ'}\right)^{-1}\mathbf{Z}\frac{\mathbf{G}}{\lambda_G}\right]\mathbf{Z'y} \\ &= \frac{1}{\lambda_G}\left[\mathbf{GZ'} - \mathbf{GZ'}\left(\lambda_G\mathbf{I_N} + \mathbf{ZGZ'}\right)^{-1}\mathbf{ZGZ'}\right]\mathbf{y} \\ &= \frac{1}{\lambda_G}\left[\mathbf{GZ'}\left(\lambda_G\mathbf{I_N} + \mathbf{ZGZ'}\right)^{-1}\left(\lambda_G\mathbf{I_N} + \mathbf{ZGZ'} - \mathbf{ZGZ'}\right)\right]\mathbf{y} \\ &= \mathbf{GZ'}\left(\lambda_G\mathbf{I_N} + \mathbf{ZGZ'}\right)^{-1}\mathbf{y}\end{aligned}$$

u を **Xβ**、**G** を **XX′**、λ_G を λ_R に置き換えることで式 3.2 を得る。

3.2.4.2 RHKS

GBLUP 及びリッジ回帰では表現型値は線形にマーカー遺伝子型と結びついている。しかしながら、優性効果やエピスタシス効果などの非相加的な QTL 効果はそれらの関係を非線形にする。この非線形性はマーカー遺伝子型を特徴空間に写像することで考慮することができる。$\mathbf{x_i}$ と $\mathbf{x_j}$ を系統 *i* と *j* のマーカー遺伝子型、$\varphi(\mathbf{x_i})$と$\varphi(\mathbf{x_j})$を特徴空間におけるマーカー遺伝子型の写像とする。カーネルは特徴空間での内積として定義される（Shawe-Taylor and Cristianini 2004）。

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \langle\varphi(\mathbf{x_i}), \varphi(\mathbf{x_j})\rangle$$

ここで⟨x, y⟩は x と y の内積を表す。一般的なカーネルの 1 つはガウシアンカーネルであり、

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2 / 4P\theta\right) \quad (3.3)$$

と表される（Endelman 2011）。ここで θ はバンド幅を表す。式 3.2 で示唆されるように、系統 *i* の遺伝子型値は特徴空間が未知であっても内積が評価できれば計算可能である。したがって、式 3.2 の X に関する内積、つまり **XX'** を式 3 に置き換えることで非相加的な効果を考慮にいれることができる。このことは式 3.1 の **G** を式 3.3 に基づくガウシアンカーネル行列 **K** に置き換えることと等価である。この方法では系統間の関係は GBLUP よりも局所的になる。つまり系統間の関係は$\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2$が小さくなるにつれて急速に失われる。RKHS は Gianola et al.（2006）、Gianola and van Kaam（2008）、または de los Campos et al.（2009b）

などに詳しい。RKHS の実行には R パッケージ rrBLUP を用いた。GBLUP の λ_G に相当する正則化パラメータは REML で推定した。バンド幅は 0.1 から 1 までの 0.1 区切りのグリッドから最も高い尤度を与えるものを選択した。これは rrBLUP の初期設定である。

3.2.4.3 Lasso 及び ENet

Lasso は多数の説明変数に対して疎な解を得るために L_1 正則化項を用いる（Tibshirani 1996）。目的関数は

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \mathbf{x_i}' \boldsymbol{\beta}\right)^2 + \lambda_L |\boldsymbol{\beta}|$$

となる。ここで λ_L は正則化パラメータを、$|\ |$ は L_1 ノルムを表す。L_1 項は L_2 項より回帰係数（マーカー効果）をより強く 0 に圧縮する。そのためごく少数の説明変数が 0 でない効果を持つようになる。ENet は L_1 項と L_2 項両方を用いる（Zou and Hastie 2005）。目的関数は

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \mathbf{x_i}' \boldsymbol{\beta}\right)^2 + \lambda_E \left((1-\alpha)/2 \|\boldsymbol{\beta}\|^2 + \alpha |\boldsymbol{\beta}|\right)$$

となる。ここでは λ_E 正則化パラメータを表す。ENet は Lasso とリッジ回帰両方の性質を併せ持つ。つまり Lasso のように変数選択を行うが相関する変数をグループとして選択し、リッジ回帰のようにそれら選択した相関する変数の効果を揃って 0 に縮約する。マーカーは通常 LD 構造、つまり相関構造を持つ。そのため、ENet は Lasso よりゲノムワイド予測に適していることが期待された。Lasso 及び ENet は R パッケージ glmnet（ver. 1.9-5、Friedman et al. 2010）を用いて行った。

3.2.4.4 RForest

RForest は集合学習の一種で複数の決定木（classification and regression trees; CART）を用いる（Breiman et al. 1984）。各決定木はブートストラップサンプルから作成される。あらかじめ決められた数の変数（ゲノムワイド予測の場合はマーカー）を各決定木の各ノードにおいてランダムに選択する。各ノードにおいてサンプル（系統）は変数の 1 つを用いて 2 つに分割されていく。このとき分割に用いる変数とそれによる分割の仕方は残差の二乗和（つまり平均からの差の二乗和）の減少幅に従って決定される。予測されるサンプル（系統）は作成された決定木の上からその変数に従って分類されていき、辿り着いた最終ノードの

平均値がその決定木における予測値となる。これを多数の決定木で重みをつけずに平均することにより最終的な予測値を得る。ブートストラップサンプルから得られる予測値を重みなしで平均し予測値を得る手法は bagging（bootstrap aggregating、Breiman 1996）と呼ばれる。一般に予測の汎化誤差は学習セットのランダム性に基づく分散と、その手法の推定バイアスに分けられることが知られている（Hastie et al. 2009）。bagging の意図はブートストラップサンプルから学習することにより分散を減少させることにある（Breiman 1996）。また各ノードで変数をランダムに選択し予測値間の相関を減らすことで、さらに分散の減少が図られている。決定木では複数の変数が連続してサンプルの分割に用いられるために、変数間の相互作用を検出することが可能となる。RForest は R パッケージの randomForest（ver. 4.6-7、Breiman 2001）を用いて行った。決定木の数は 1000 としたが、その他のパラメータはパッケージの既定値を用いた。

3.2.4.5　Blasso、EBlasso、及び wBSR

これらの手法は第 2 章で述べた。Blasso のハイパーパラメータ φ は 1 とし、ω は 1、10^{-2}、10^{-5} の 3 値を試した。EBlasso のハイパーパラメータ φ、ω、及び ψ は 1 とし、θ は 1、10^{-2}、10^{-5} の 3 値を試した。wBSR のハイパーパラメータ ν は 4 とし、κ は 0.05、0.1、0.2、0.5、0.7、1 の 5 値を試した。S^2 は近交係数（f）を 1 とし式 2.1（2.3.8　ハイパーパラメータ）に従って求めた。なお σ_m^2 は全てのシナリオで 0.5 とした。これらの手法は VIGoR の *Evaluation* 機能を用いて実行した。ハイパーパラメータの最適化は 5 分割のランダムな交差検証を用いて行った。

3.2.4.6　Ave

上述 8 手法で得られた予測値の相加平均を用いた。

3.2.5　シミュレーション

手法間のパフォーマンス差に関連する要因をより詳細に調べるために、実際のマーカー遺伝子型をもとにシミュレーションデータを作成した。考慮したシミュレーション条件は Nqtl、遺伝率、Ntrain、エピスタシス、及び LD の範囲である。シミュレーションを通して QTL はマーカーから選択し、予測には QTL として選択されなかったマーカーのみを用いて行った。Nqtl は 6、12、36、120 の 4 値を想定した。Nqtl が 6 の場合は QTL は異なる染色

体から選択するようにし、それ以外の場合はランダムに選択した。QTL が説明する遺伝分散は均等であると仮定し、QTL の相加的効果はアリル頻度をもとに決定した。相加的効果の符号（正か負か）はランダムに決定した。各品種の相加的効果の和、つまり相加的な遺伝子型値は各品種が持つ QTL アリルの相加的効果の和として計算した。表現型値は、そこにランダムに正規分布から生成したノイズを加えることで作成した。ノイズの分散は相加的遺伝子型値の分散と狭義の遺伝率から決定した。狭義の遺伝率は 0.1、0.3、0.5、0.7、0.9 の 5 値を想定した。後述するように本研究では予測能力の評価に 11 分割の交差検証を用いたために、110 品種での学習用セットの大きさ、つまり Ntrain は 100 となる。

また品種数を 330 及び 550 に増加させたデータセットも作成した。これらはそれぞれ Ntrain が 300 及び 500 に相当する。元となったマーカー遺伝子型での LD 構造を維持するために、Wimmer et al.（2013）に従いマーカー間の相関行列のコレスキー分解を行い、それを利用して品種数を増加させたデータを作成した。まずマーカー間相関行列 **C** を R パッケージ MatrixR で提供されている nearPD 関数で正定値行列に近似した行列 **C***を作成し、**C***をコレスキー分解することで上三角行列 **U** を得た（つまり $\mathbf{C}^* = \mathbf{U}'\mathbf{U}$）。次に要素が−1 または 1 からなる品種数（330 または 550）×マーカー数（3102）の行列 **W** を作成した。**W** における各マーカーの要素（−1 または 1）は実際のマーカー遺伝子型におけるアリル頻度をパラメータとして用いたベルヌーイ試行により作成した。これは全ての品種は近交系であると仮定しているため、アリル頻度が遺伝子型頻度と等しいことによる。この **W** に右側から **U** をかけることで新たなマーカー遺伝子型 **G** を得た。

QTL 間の 2 次の相互作用（エピスタシス）をシミュレートするため、QTL として選択されたマーカーから、Nqtl の 3 分の 1 のペアを重複がないように選択した（つまり相互作用するように選択された QTL 数は Nqtl の 3 分の 2 となる）。各ペアのエピスタシス効果はペアとなった QTL の相加的効果からランダムに選択した。エピスタシスによる遺伝子型値は、ペアとなった QTL 間の遺伝子型の積にエピスタシス効果をかけることにより作成した。これに相加的効果による遺伝子型値を加えることで合計の遺伝子型値を作成した。表現型値は、そこにランダムに正規分布から生成したノイズを加えることで作成した。ノイズの分散は合計の遺伝子型値の分散と広義の遺伝率から決定した。エピスタシス効果による変動を含めた広義の遺伝率は 0.1、0.3、0.5、0.7、0.9 の 5 値を想定した。この手順では遺伝分散のうちおよそ 40 %がエピスタシスによる分散となった。これ以降、シミュレーションの条件として遺伝率に触れる際は、エピスタシスがない場合の狭義の遺伝率、ある場

合の広義の遺伝率に関わらず「遺伝率」と呼ぶことに注意する。

広範囲に渡る LD（long-range LD）が予測能力に及ぼす影響を調べるために、Nqtl が 6 のシミュレーションにおいて、もとの LD 構造が QTL の周囲にしか保存されていないデータセットを、QTL から 10 cM 以上離れたマーカーの遺伝子型をランダムに並べ替えることで作成した。平均して QTL の 10 cM 以内には 325.3（±87.9）マーカーが存在した。

合計して 150 シナリオに渡るシミュレーションを行った。これらのうち 120 シナリオは 4 つの Nqtl 値（6、12、36、120）、5 つの遺伝率（0.1、0.3、0.5、0.7、0.9）、3 つの Ntrain 値（100、300、500）、エピスタシスの有無の組み合わせから成る。残りの 30 シナリオは LD 構造を壊した Nqtl が 6 のもので、残りの条件（遺伝率、Ntrain、エピスタシスの有無）の組み合わせから成る。各 Nqtl 値において 100 QTL セットを作成し、QTL の位置と効果は同じ Nqtl 値で作成されたシナリオ間で共通とした。各シナリオにおいて表現型値は各 QTL セットから 1 度作成した。結果として 1 つのシナリオにつき 100 反復の試行を行った。

3.2.6 予測の正確さの評価

本研究では実データ、シミュレーションの解析に関わらず 11 分割の交差検証を行った。品種数は 110、330、550 の 3 通りであるため、交差検証における学習用セットの大きさ、つまり Ntrain はそれぞれ 100、300、500 となる。交差検証では予測手法間で品種の分割方法は共通とした。実データにおける比較では交差検証を 100 回行い、シミュレーションでは 1 つの反復につき 1 回行った。Ntrain が 300 及び 500 のシミュレーションでは交差検証での品種の分割はランダムに行った。実データと Ntrain が 100 の時のシミュレーションでは学習用セットにおける品種の遺伝的グループの組成が、実データのそれと極力同じになるように分割した。他の研究との予測の正確さの比較を意味あるものにするために、Daetwyler et al.（2013）の示唆に基づき学習用セットと評価用セットとの遺伝的関係を平均二乗遺伝的距離（rel^2）を用いて計算した。rel^2 は rrBLUP パッケージの A.mat 関数を用いて計算した実現ゲノム関係行列から計算した。Lasso、ENet、Blasso、EBlasso、wBSR については 11 分割交差検証の各分割において、さらに 10 分割（Lasso、ENet）または 5 分割（Blasso、EBlasso、wBSR）の交差検証を用いてハイパーパラメータを調整した。

実データでは予測の正確さは予測値と表現型値とのピアソン相関係数で測定した。シミュレーションでは予測値と生成した遺伝子型値とのピアソン相関係数で測定した。エピスタシスを想定した場合は、遺伝子型値は相加的効果とエピスタシス効果の和を用いた。

相関係数の計算は、交差検証の全分割の予測値を用いて行った（つまり分割毎に計算したのではない）。手法間の正確さの差は R の aov 及び TukeyHSD 関数を用いてチューキー法で検定した。

3.2.7　正確さの変動係数

手法間の正確さの差を推定するために、正確さの変動係数を計算した。係数が大きいほど手法間の差が大きいことを示す。その場合手法の選択がより大きな影響を持つことになる。

3.2.8　実データに最も近いシミュレーションシナリオの探索

正確さの手法間順位が実データと最も近いシミュレーションシナリオを探索した。順位の類似性評価にはスピアマン相関係数を用いた。実データに最も近いシナリオにおける狭義または広義の遺伝率と比較するために、実データにおける狭義の遺伝率を rrBLUP による分散成分（相加的遺伝分散と環境分散）の推定値を用いて求めた。

3.3　結果

3.3.1　遺伝的及び LD 構造

本研究で用いた品種群はそれぞれ 61 及び 49 品種からなる 2 つの遺伝的グループに分かれた（図 3.1）。実データの交差検証における学習用セットと評価用セットとの平均 rel^2 は 0.154（±0.03）であった。

日本水稲品種が自殖であり、また新品種の育成に交配親として用いられる品種が比較的限られていることから予想された通り、LD は長い距離に渡って検出された。同じ染色体に位置するマーカーペアの r^2 は、ペアの距離がおよそ 40 cM を超えるまで、異なる染色体間のマーカーペア間のそれ（つまりバックグラウンド）を上回った（図 3.2）。1 cM 毎のウィンドウで計算した平均 r^2 は 12 cM まで 0.1 以上であった。隣接するマーカーペアの平均 r^2 は 0.52（±0.37）であった。

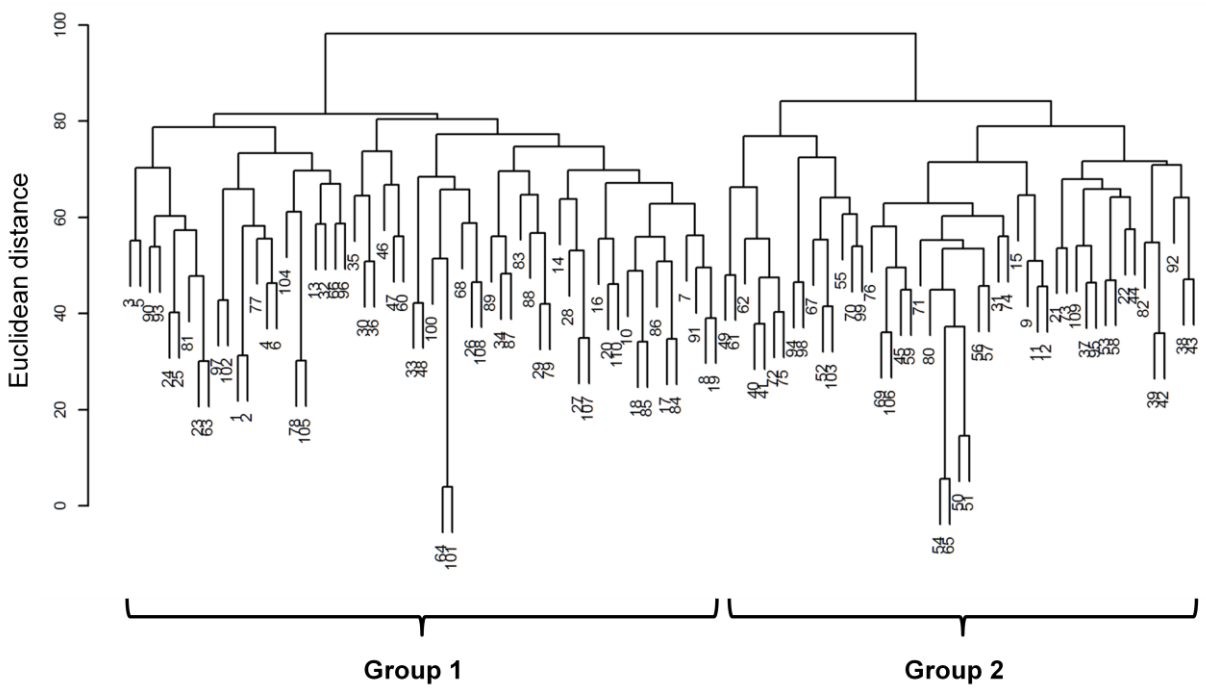

図 3.1　階層的クラスタリングにより推定されたイネ 110 品種の遺伝的グループ。図中の数字は表 3.1 における ID に対応する。

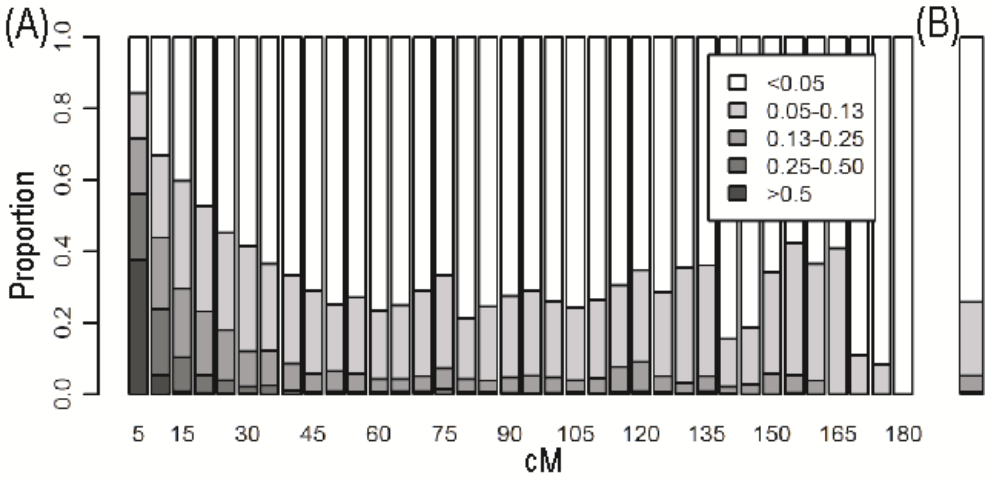

図 3.2　r^2 を用いて計算したイネ 110 品種における連鎖不平衡（LD）。（A）同じ染色体上にあるマーカー間の r^2 をマーカー間の距離 5 cM 毎に 5 分画の割合として図示した。（B）異なる染色体上にあるマーカー間の r^2（バックグラウンドの LD）。

3.3.2　実データにおける予測手法の比較

本研究で用いた 8 形質の表現型値の分布は図 3.3 に示した。これらの形質における予測の正確さは図 3.4 に示した。正確さは概ね DH 及び CL で高く、他の形質では中程度であった。最も正確な予測は DH、PL、PN では RForest で得られた。CL、BL では RKHS で、GW では GBLUP で得られた。GL 及び BW では Ave で得られた。DH、CL、PL、PN、GW、BW では wBSR が最も低く、BL は Lasso が、GL では EBlasso が最も低かった。手法間での正確さの変動係数は 0.05（DH）、0.10（CL）、0.08（PL）、0.13（PN）、0.09（GL）、0.10（GW）、0.11（BL）、及び 0.06（BW）であった。実データの予測における Lasso、ENet、及び wBSR の振舞いは、それぞれゼロでない効果を持つマーカー数、交差検証で選択されたハイパーパラメータ α の値、交差検証で選択されたハイパーパラメータ κ の値で知ることができる。これらの値は表 3.2 に示した。

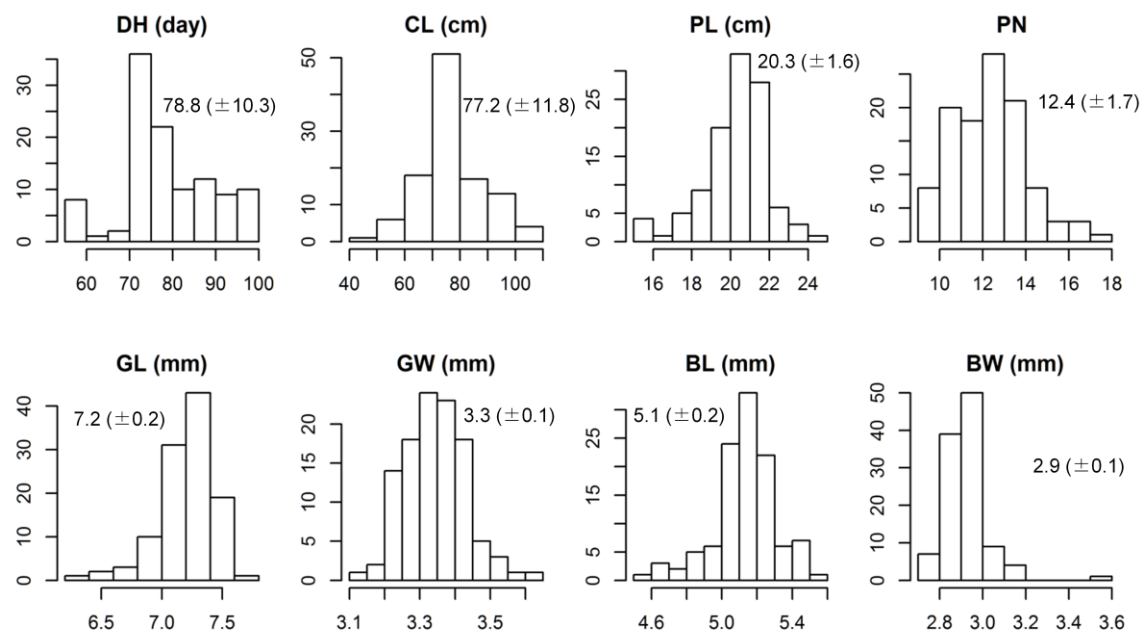

図 3.3　イネ 110 品種の 6 年間の平均表現型値の分布。品種平均値（標準偏差）は図中に示した。DH（到穂日数）、CL（稈長）、PL（穂長）、PN（穂数）、GL（粒長）、GW（粒幅）、BL（玄米長）及び BW（玄米幅）。

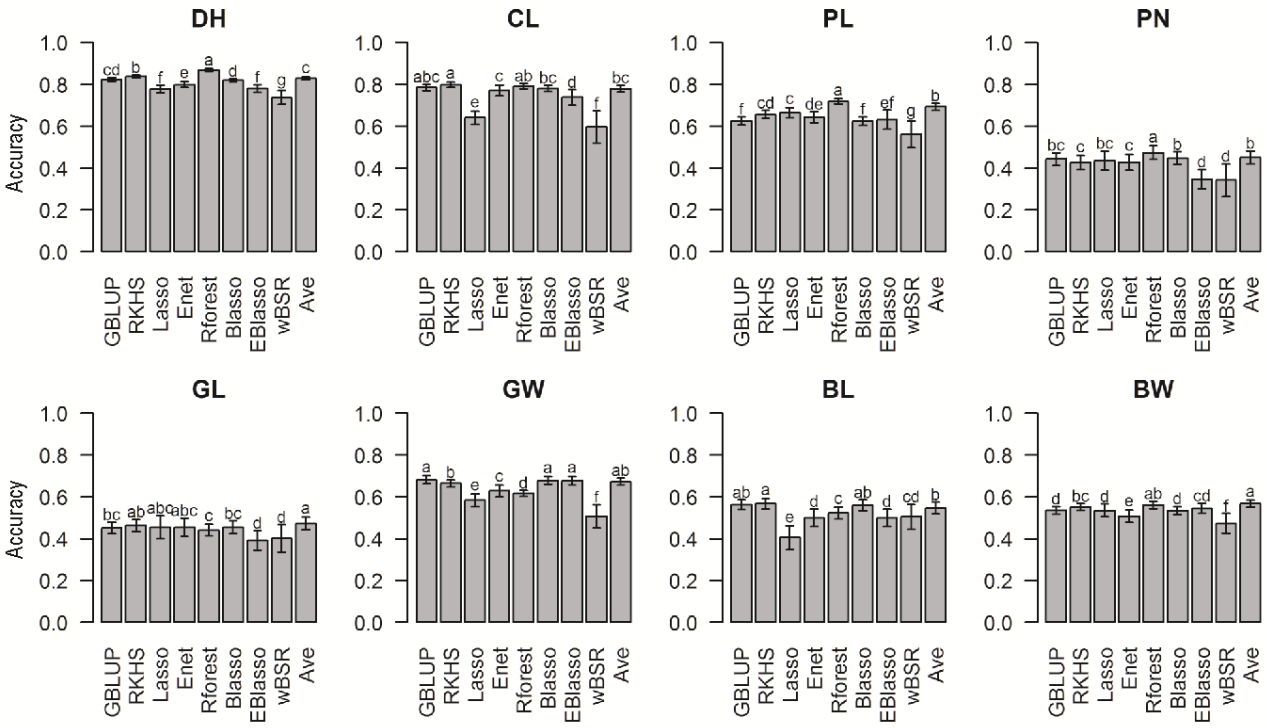

図 3.4　イネ 8 形質における 9 つの予測手法の正確さ。予測の正確さは予測値と表現型値のピアソン相関係数で評価した。異なる小文字は有意差を示す（$P<0.05$）。DH（到穂日数）、CL（稈長）、PL（穂長）、PN（穂数）、GL（粒長）、GW（粒幅）、BL（玄米長）及び BW（玄米幅）。

表 3.2　Lasso、ENet、及び wBSR の振る舞いを表す指標

Trait[a]	Lasso[b]	ENet[c]	wBSR[d]
DH	54.0 (20.0)	0.14 (0.25)	0.79 (0.37)
CL	70.7 (21.8)	0.03 (0.13)	0.84 (0.34)
PL	62.2 (19.4)	0.65 (0.29)	0.11 (0.15)
PN	52.8 (16.2)	0.47 (0.32)	0.95 (0.20)
GL	47.0 (15.3)	0.52 (0.34)	0.93 (0.25)
GW	73.9 (31.2)	0.26 (0.32)	0.74 (0.41)
BL	63.0 (36.8)	0.19 (0.30)	0.95 (0.20)
BW	38.3 (26.3)	0.63 (0.34)	0.70 (0.43)

[a] DH（到穂日数）、CL（稈長）、PL（穂長）、PN（穂数）、GL（粒長）、GW（粒幅）、BL（玄米長）及び BW（玄米幅）

[b] 0 でない効果を持つマーカー数の平均値（標準偏差）

[c] 交差検証で選択された α の平均値（標準偏差）。α が 1 のとき ENet は Lasso と同等に、0 のときはリッジ回帰と同等になる。

[d] 交差検証で選択された κ の平均値（標準偏差）。κ は回帰モデルに含まれるマーカーの割合についての想定値を表す。

3.3.3　シミュレーションにおける予測手法の比較

実際のマーカー遺伝子型をもとに作成したシミュレーションデータを用いて予測手法を比較した。シミュレーションでは合計 120 シナリオを作成した。シミュレーション条件として Nqtl、Ntrain、遺伝率、エピスタシスの有無を考慮した。各シナリオにおける予測手法の正確さの順位は図 3.5 に、正確さは図 3.6、3.7、及び 3.8 に示した。Lasso、ENet、wBSR の振舞いを表す指標はそれぞれ図 3.9、3.10、及び 3.11 に示した。

GBLUP は Nqtl 及び遺伝率が高いシナリオで高く順位づけされる傾向があった。この傾向はエピスタシスがない場合により顕著であった。RKHS はエピスタシスが有り、Nqtl と遺伝率が高い場合に他の手法より優位になる傾向があった。Lasso と ENet は Nqtl が小さく遺伝率が高い場合に高く順位づけされる傾向があった。RForest は Ntrain が最も小さいときに（つまり 100 のとき）他より正確な予測を与える傾向があった。特に、Ntrain が 100 でエピスタシスがあり、Nqtl が 6 の場合、遺伝率に関わらず最も高い正確さを与えた。Blasso と EBlasso は遺伝率が低いときにより高く順位づけされる傾向があった。wBSR は多くのシナリオにおいて他より劣る傾向があった。Blasso と EBlasso の顕著な特徴はシナリオを問わないその安定性にあった。Blasso と EBlasso の平均の順位はそれぞれ 3.4±1.9 と 3.7±1.5 であり、これらは他の手法、GBLUP（5.2±2.2）、RKHS（5.1±2.6）、Lasso（6.5±2.5）、ENet（4.9±1.9）、RForest（6.1±3.1）、wBSR（7.3±1.7）より高かった。Ave（2.8±1.4）もシナリオ間で安定であった。

正確さの手法間での変動係数は、遺伝率または Ntrain が増加するほど減少する傾向があった（表 3.3）。このことは手法の選択は遺伝率または Ntrain が減少するほどより重要になることを示唆した。変動係数は Nqtl 及びエピスタシスの有無には影響を受けなかった（結果非掲載）。

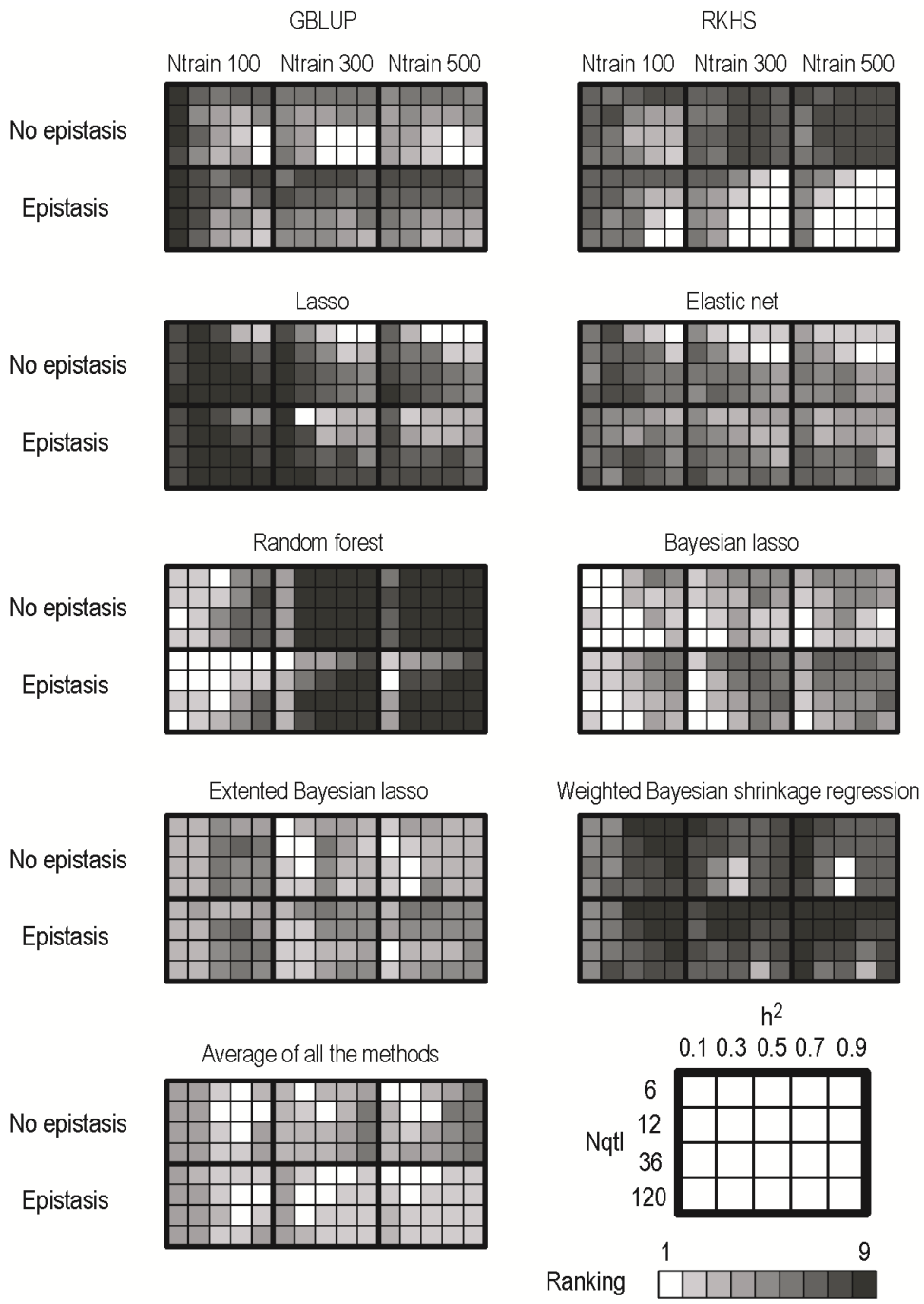

図 3.5　シミュレーションにおける予測手法の順位。合計 120 通りのシミュレーションシナリオは QTL 数（Nqtl、行）、遺伝率（h^2、列）、学習用セットのサイズ（Ntrain）、及びエピスタシスの有無が異なる。順位は高い側から低い側へ、白から黒の濃淡で表している。

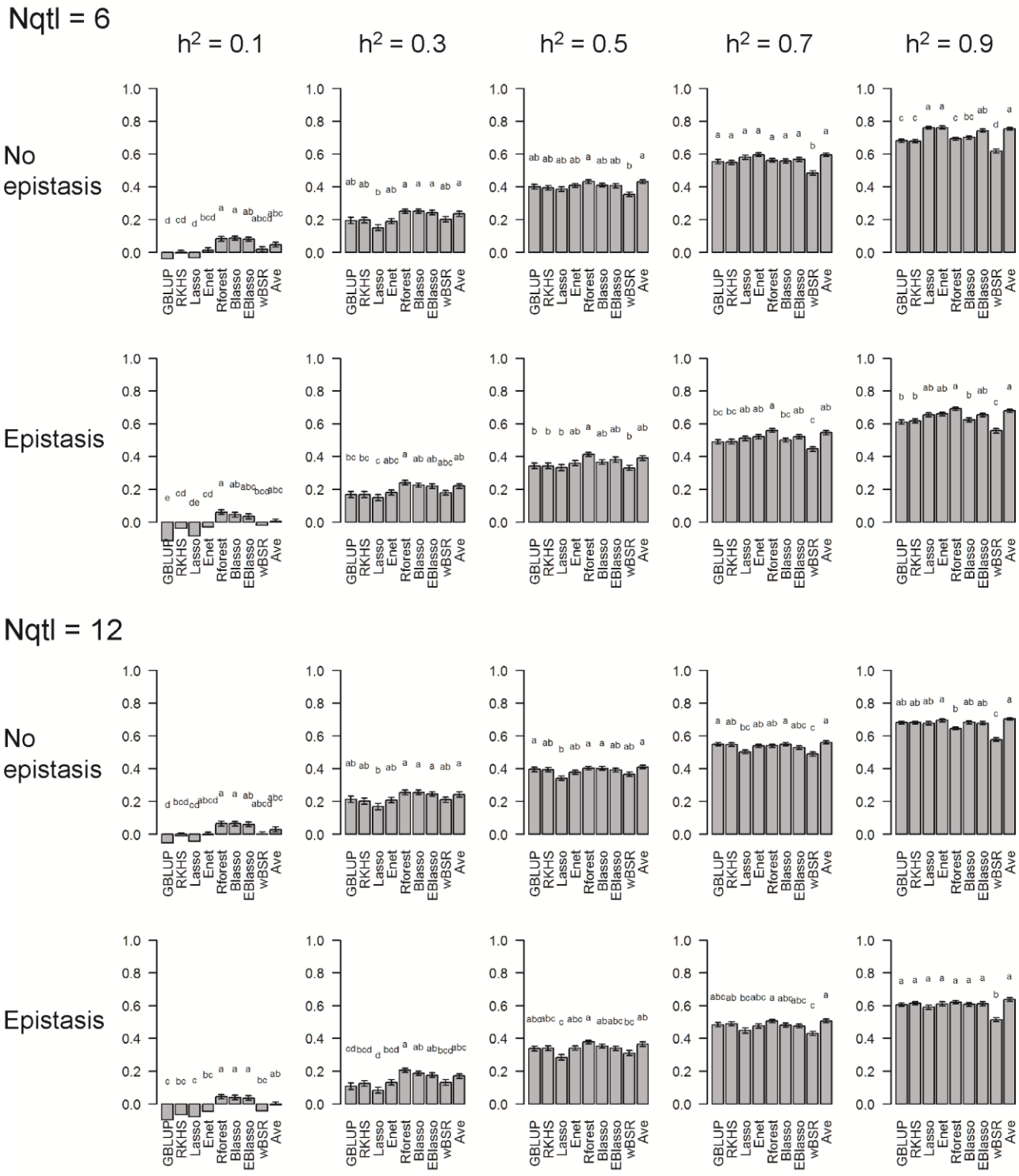

図 3.6 学習用セットのサイズ（Ntrain）が 100 の場合の予測の正確さ。正確さは予測値とシミュレートされた真の遺伝子型値とのピアソン相関係数で評価した。異なる小文字は有意差を表す（P<0.05）。h^2 はエピスタシスがない場合は狭義の、ある場合は広義の遺伝率を表す。手法は左から GBLUP、RKHS、Lasso、ENet、Rforest、Blasso、EBlasso、wBSR、及び Ave。

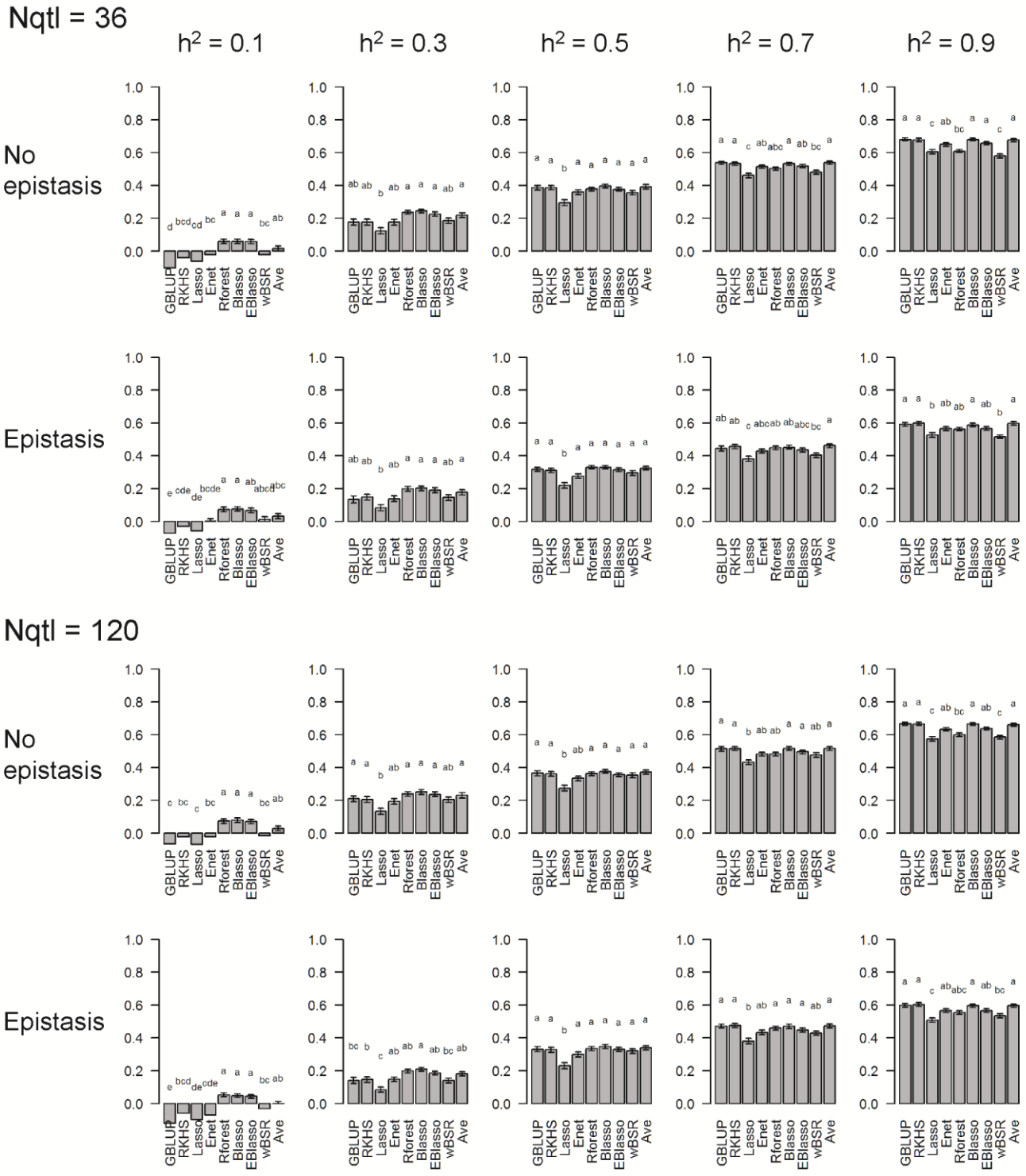

図 3.6　学習用セットのサイズ（Ntrain）が 100 の場合の予測の正確さ（続き）

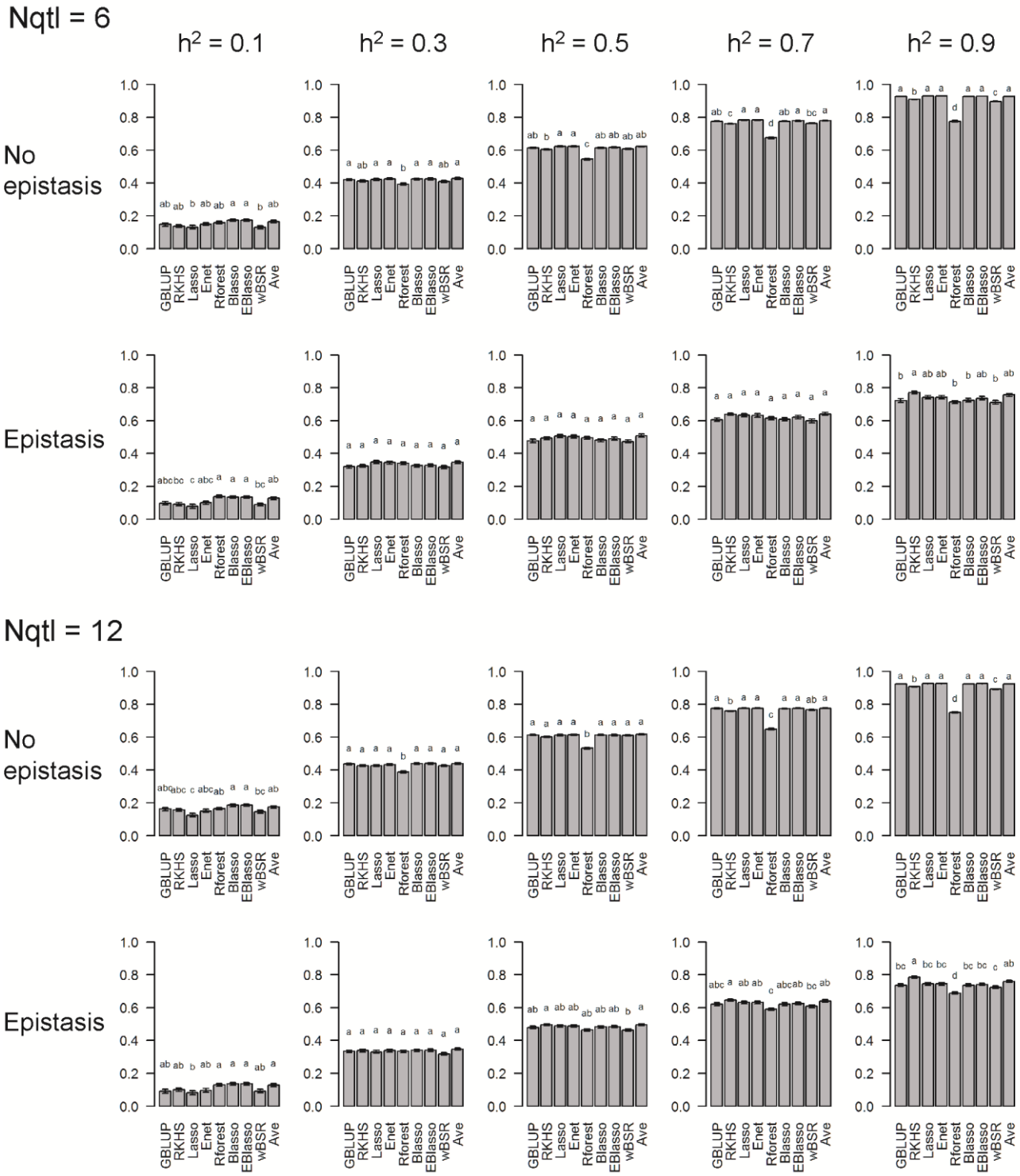

図 3.7 学習用セットのサイズ（Ntrain）が 300 の場合の予測の正確さ。正確さは予測値とシミュレートされた真の遺伝子型値とのピアソン相関係数で評価した。異なる小文字は有意差を表す（P<0.05）。h^2 はエピスタシスがない場合は狭義の、ある場合は広義の遺伝率を表す。手法は左から GBLUP、RKHS、Lasso、ENet、Rforest、Blasso、EBlasso、wBSR、及び Ave。

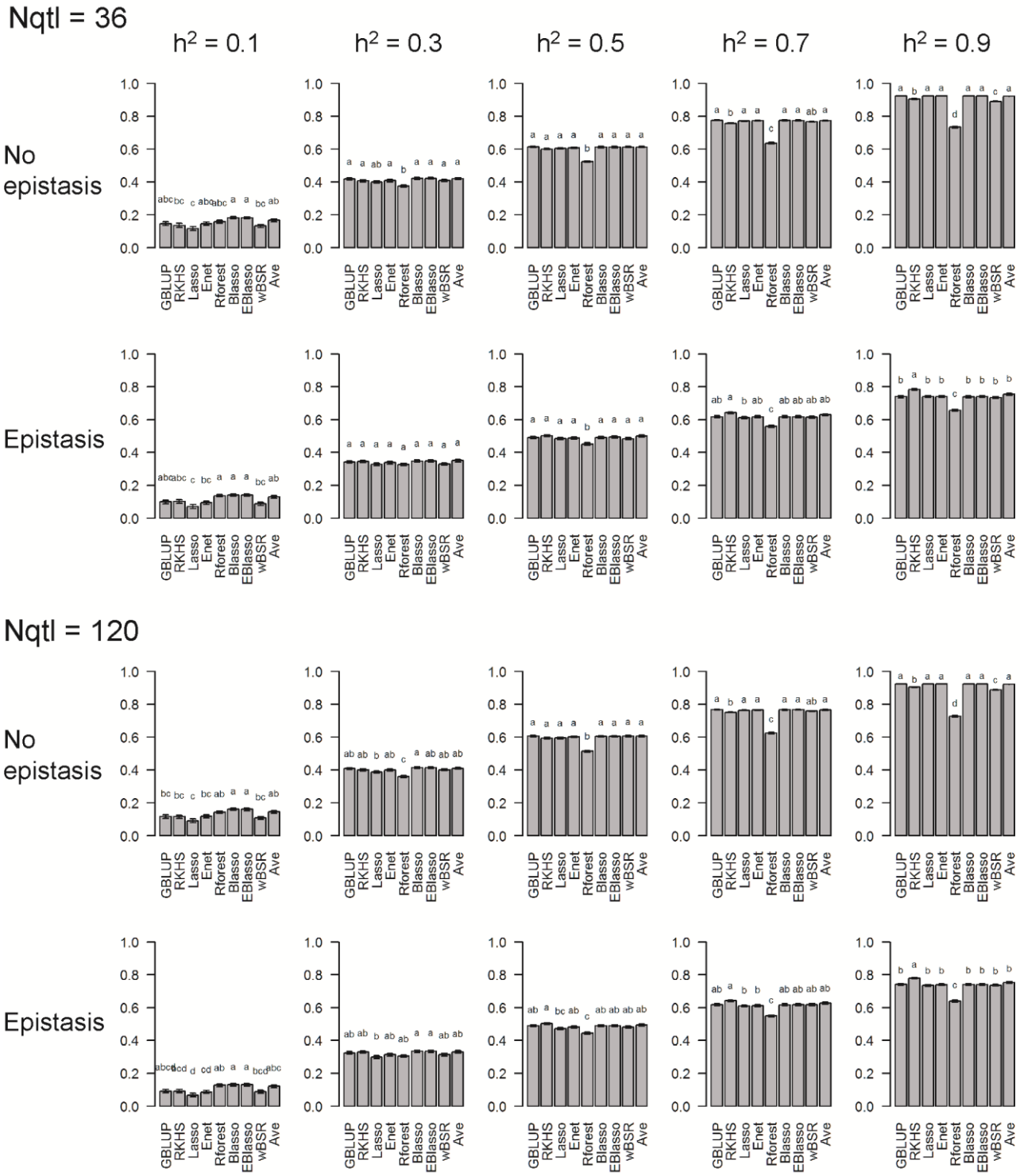

図 3.7　学習用セットのサイズ（Ntrain）が 300 の場合の予測の正確さ（続き）

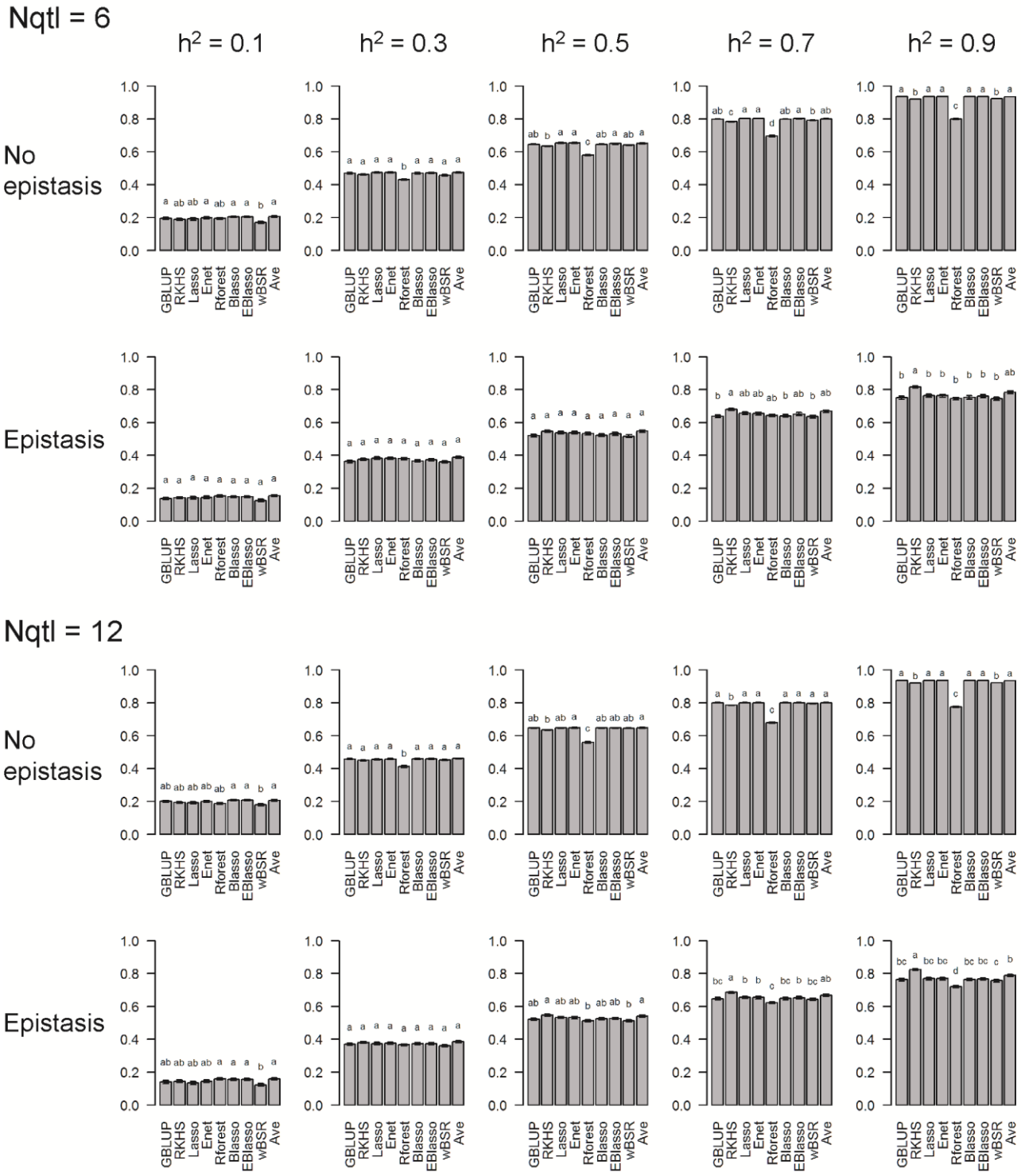

図 3.8 学習用セットのサイズ（Ntrain）が 500 の場合の予測の正確さ。正確さは予測値とシミュレートされた真の遺伝子型値とのピアソン相関係数で評価した。異なる小文字は有意差を表す（P<0.05）。h^2 はエピスタシスがない場合は狭義の、ある場合は広義の遺伝率を表す。手法は左から GBLUP、RKHS、Lasso、ENet、Rforest、Blasso、EBlasso、wBSR、及び Ave。

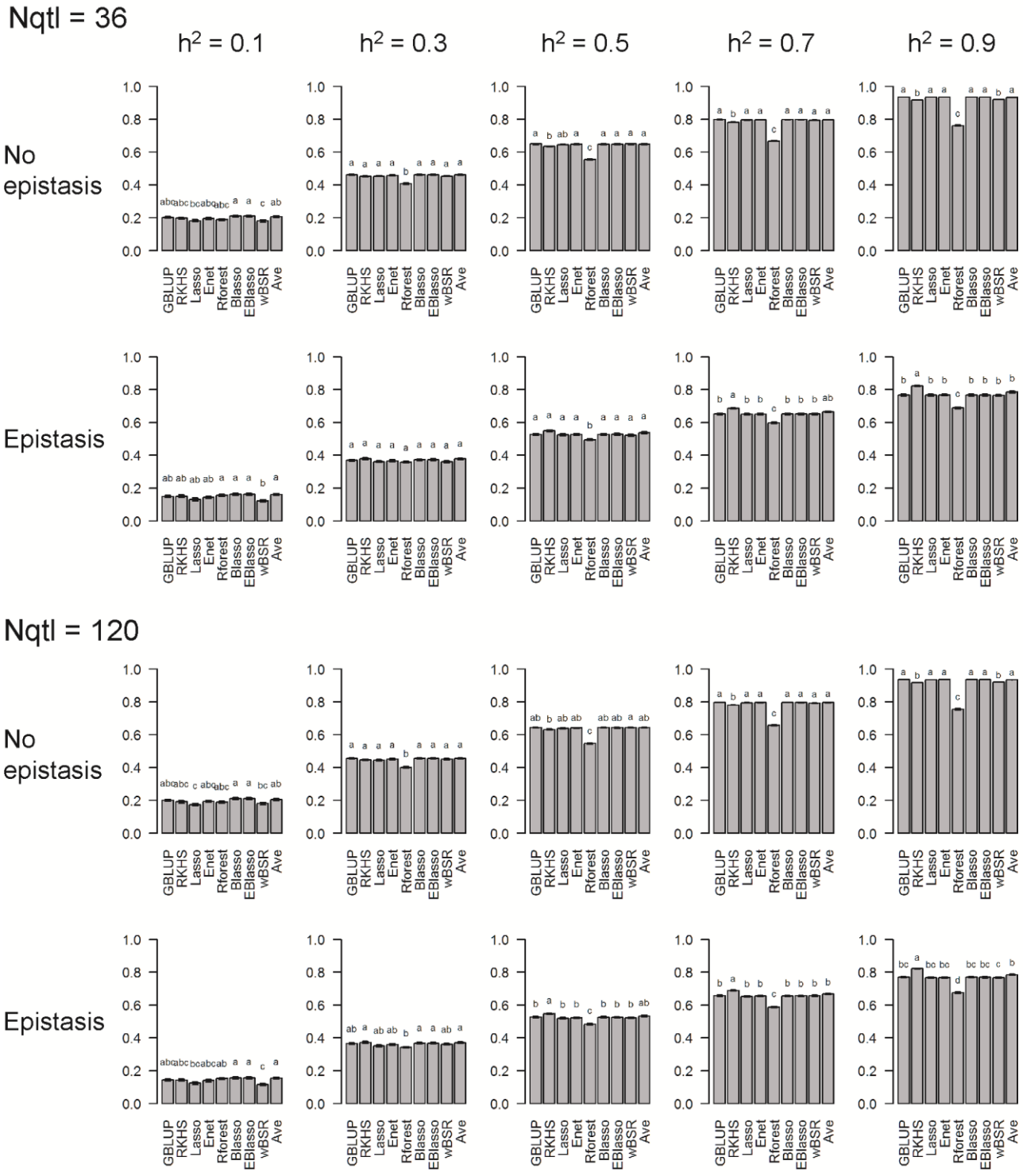

図 3.8　学習用セットのサイズ（Ntrain）が 500 の場合の予測の正確さ（続き）

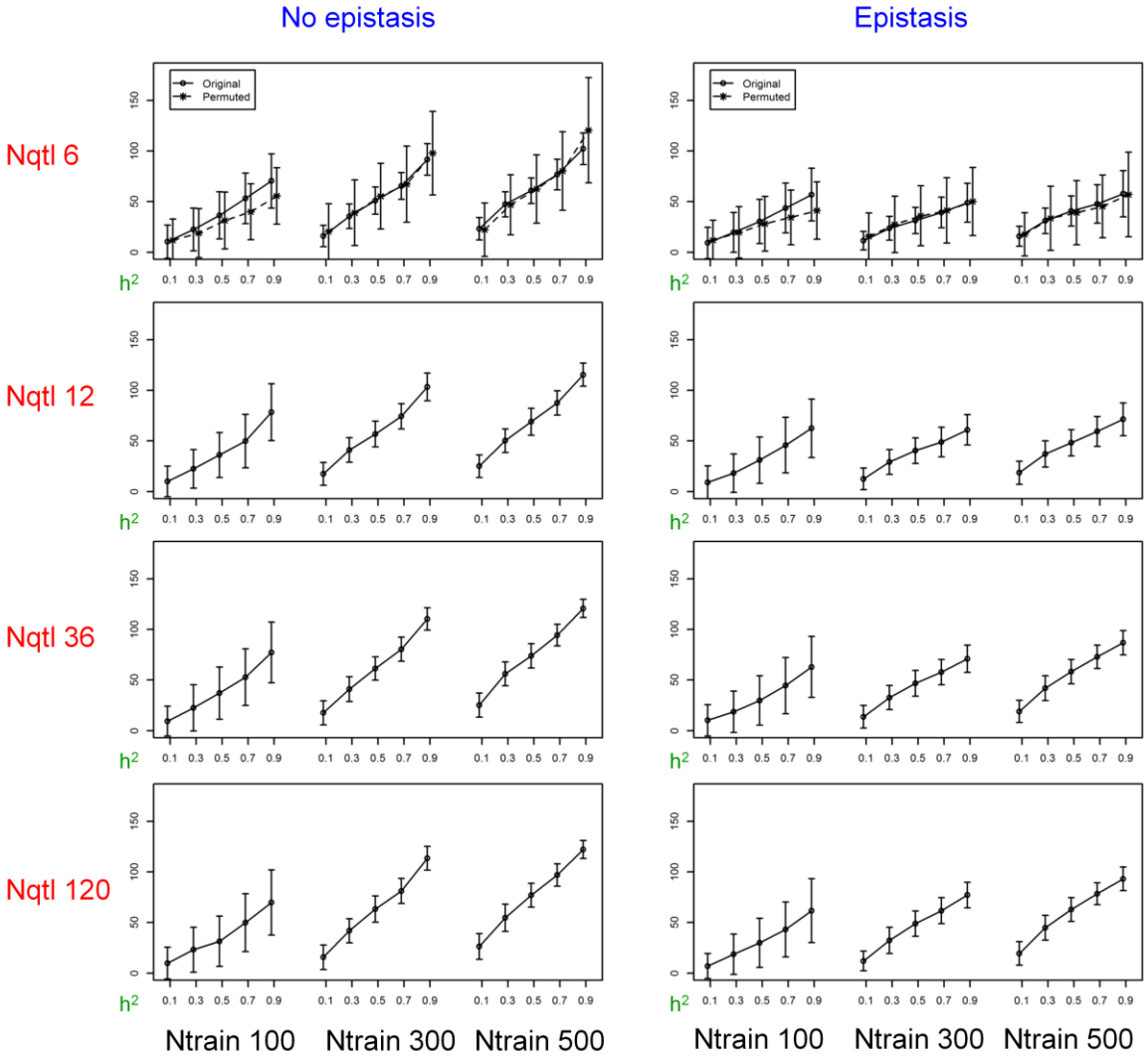

図 3.9　Lasso が選択するマーカー数に遺伝率（h^2）、QTL 数（Nqtl）、学習用セットのサイズ（Ntrain）、及びエピスタシスが与える影響。Nqtl が 6 の場合は長距離の連鎖不平衡（LD）をマーカーのランダムな並び替えにより壊した際の結果（Permutated）も併記。

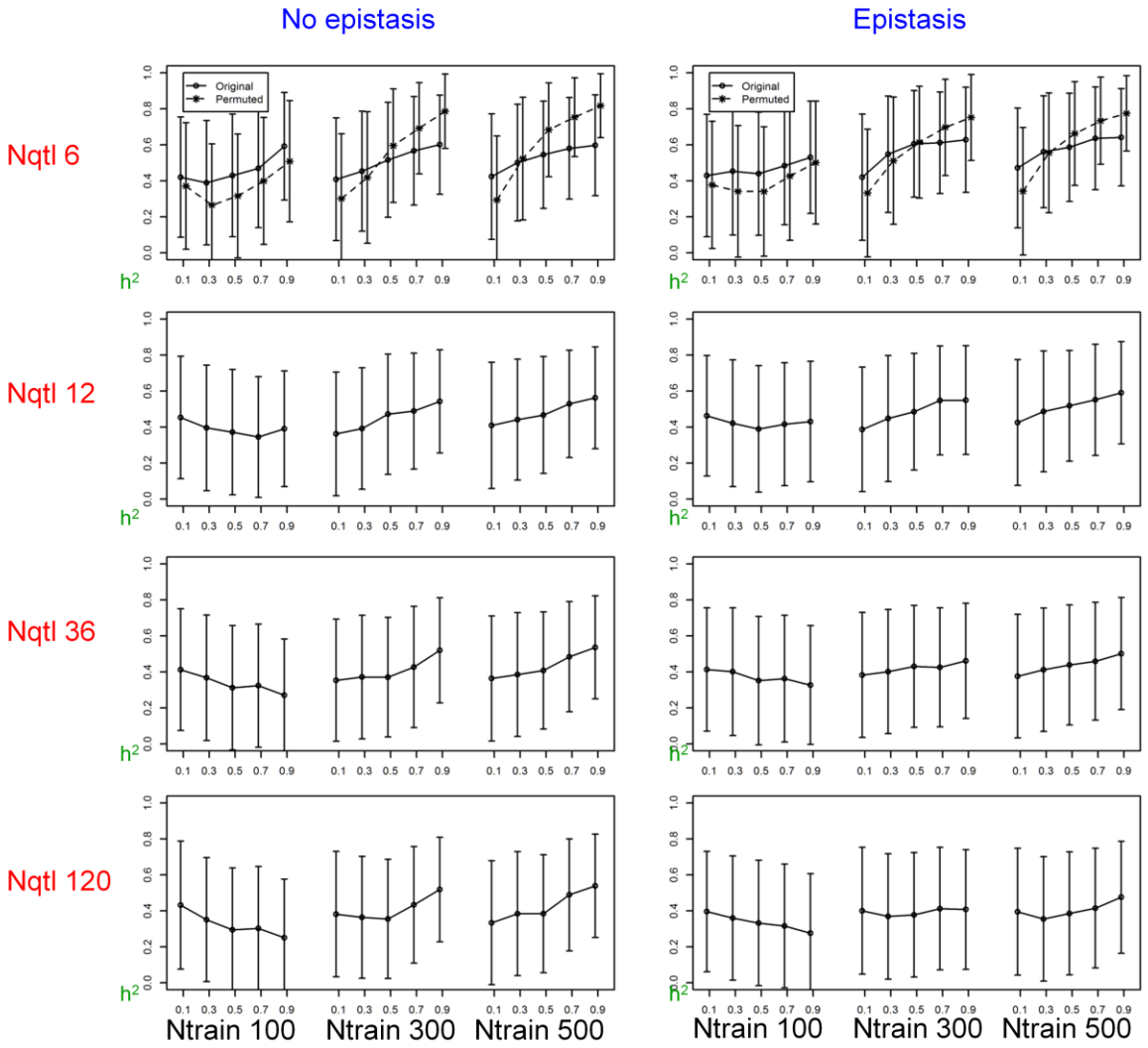

図 3.10　ENet において交差検証で選択する α に遺伝率（h^2）、QTL 数（Nqtl）、学習用セットのサイズ（Ntrain）、及びエピスタシスが与える影響。α が 1 のときは ENet は Lasso に、0 のときはリッジ回帰と等価となる。Nqtl が 6 の場合は長距離の連鎖不平衡（LD）をマーカーのランダムな並び替えにより壊した際の結果（Permutated）も併記。

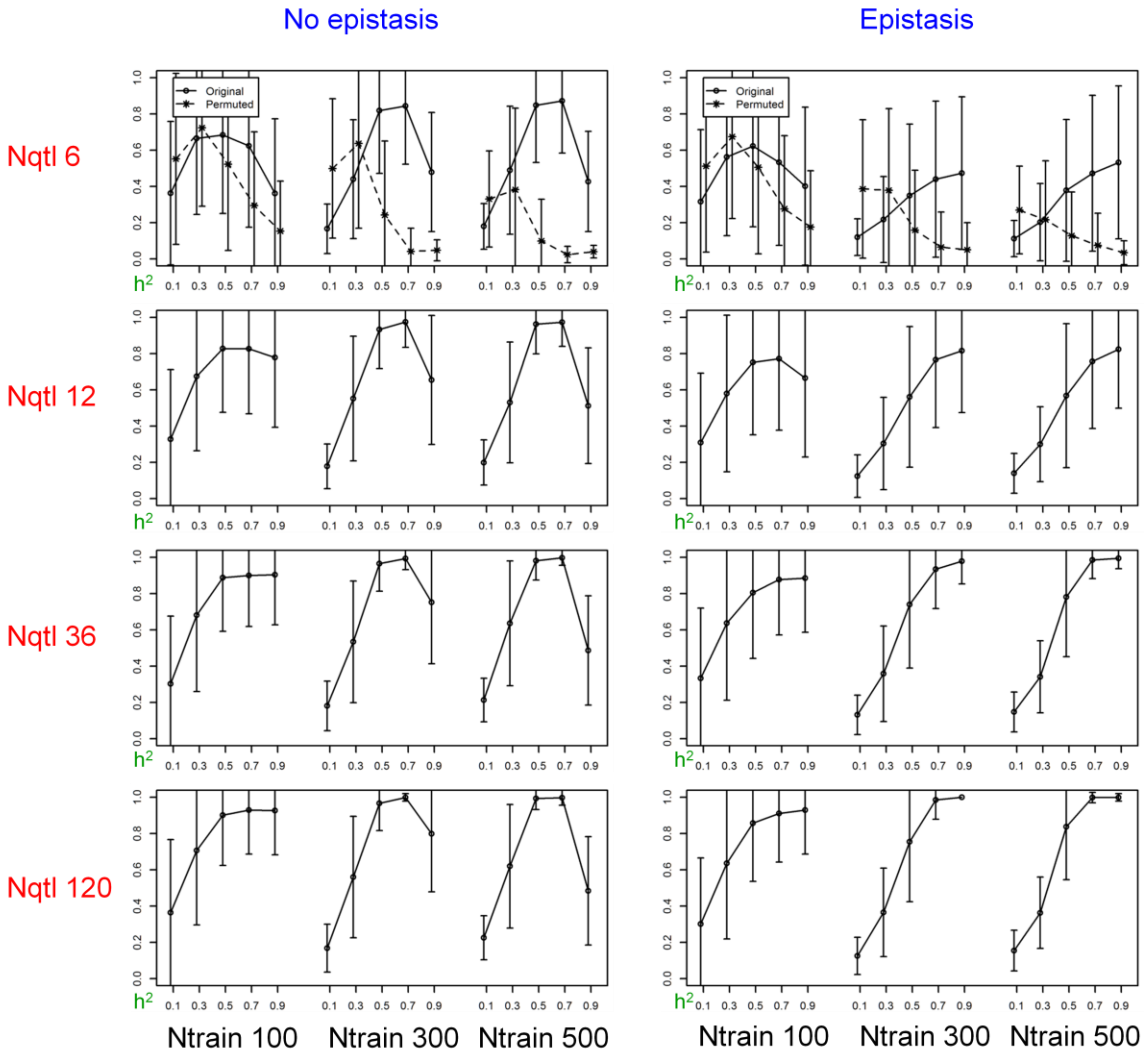

図 3.11　wBSR において交差検証で選択する κ に遺伝率（h^2）、QTL 数（Nqtl）、学習用セットのサイズ（Ntrain）、及びエピスタシスが与える影響。κ は回帰モデルに含まれるマーカーの割合に対する想定値を表す。Nqtl が 6 の場合は長距離の連鎖不平衡（LD）をマーカーのランダムな並び替えにより壊した際の結果（Permutated）も併記。

表 3.3 QTL 数（Nqtl）が 6 でエピスタシスがない場合の 9 つの予測手法間の予測の正確さの変動係数 [a].

	Heritability				
Ntrain	0.1	0.3	0.5	0.7	0.9
100	3.02 (6.39)	0.85 (2.23)	0.20 (0.43)	0.11 (0.07)	0.09 (0.05)
300	0.54 (1.16)	0.05 (0.03)	0.04 (0.02)	0.05 (0.02)	0.06 (0.02)
500	0.20 (0.58)	0.04 (0.02)	0.04 (0.01)	0.04 (0.02)	0.05 (0.01)

[a] 変動係数は反復毎に計算し平均値（標準偏差）を表示

3.3.4　LD 構造を壊したシミュレーションにおける予測手法の比較

長く続く LD 構造が予測手法の選択に与える影響を調べるために、QTL から 10 cM 以上離れたマーカー遺伝子型をランダムに並び替えたデータセットを作成した。Nqtl は 6 とした。手法の正確さの順位は図 3.12 に、正確さは図 3.13 に示した。大きく 2 つの傾向が観察された。1 つ目は、エピスタシスがあり Ntrain が高く（300 または 500）、遺伝率が高いシナリオにおける RKHS の優位性が失われ、2 つ目は Ntrain が高く、遺伝率が低いシナリオにおいて EBlasso の優位性が失われる傾向にあった。Lasso、ENet、wBSR の振舞いを表す指標はそれぞれ図 3.9、3.10、及び 3.11 に示した。

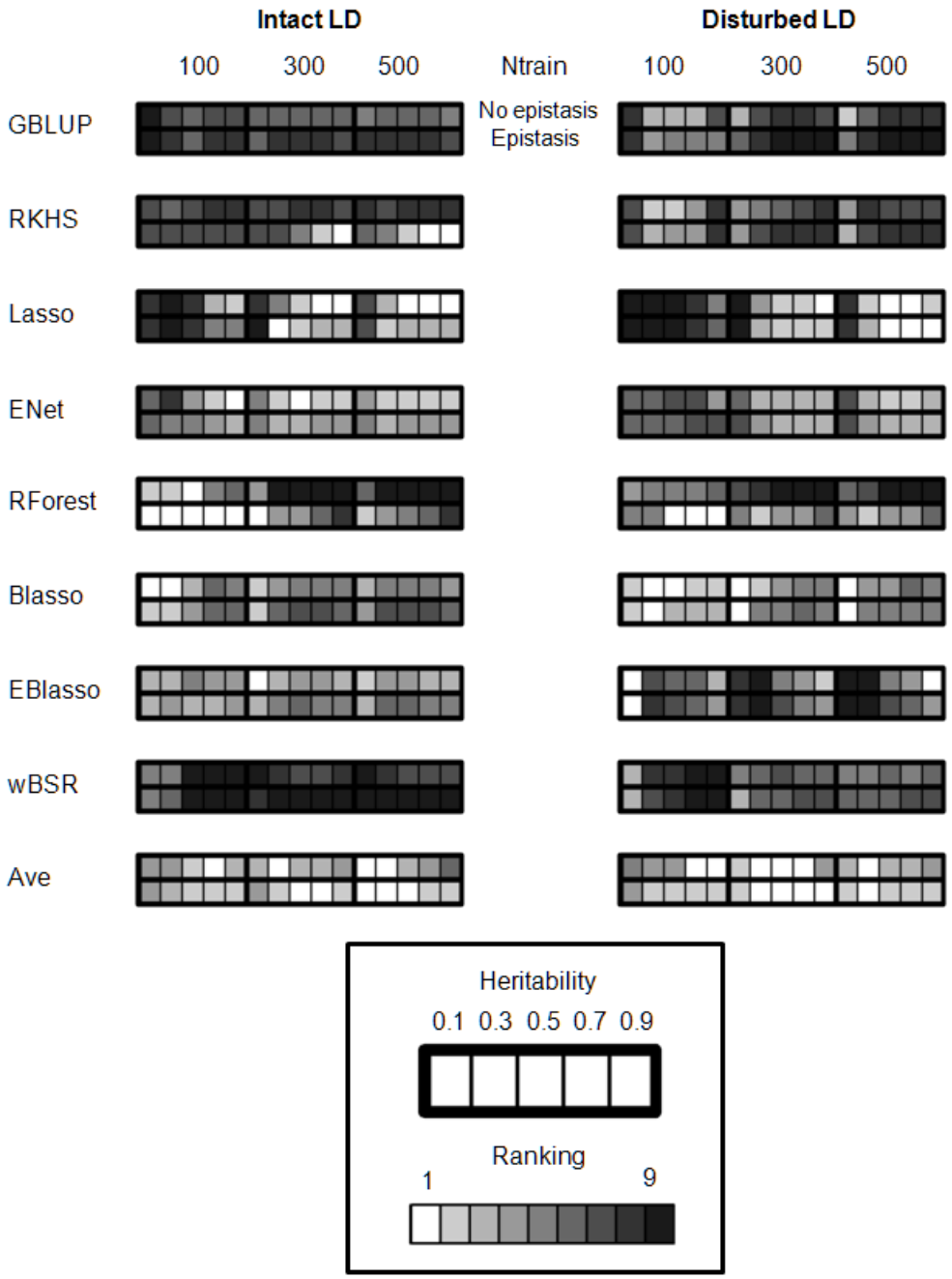

図 3.12　QTL から 10 cM 以上離れたマーカーをランダムに並べ替えることで長距離の連鎖不平衡（LD）を壊したシミュレーションデータにおける予測手法の順位（Disturbed LD）。Intact LD は LD 構造がもとの遺伝子型データと同じ場合の結果であり、図 3.5 の抜粋となる。QTL 数は 6。

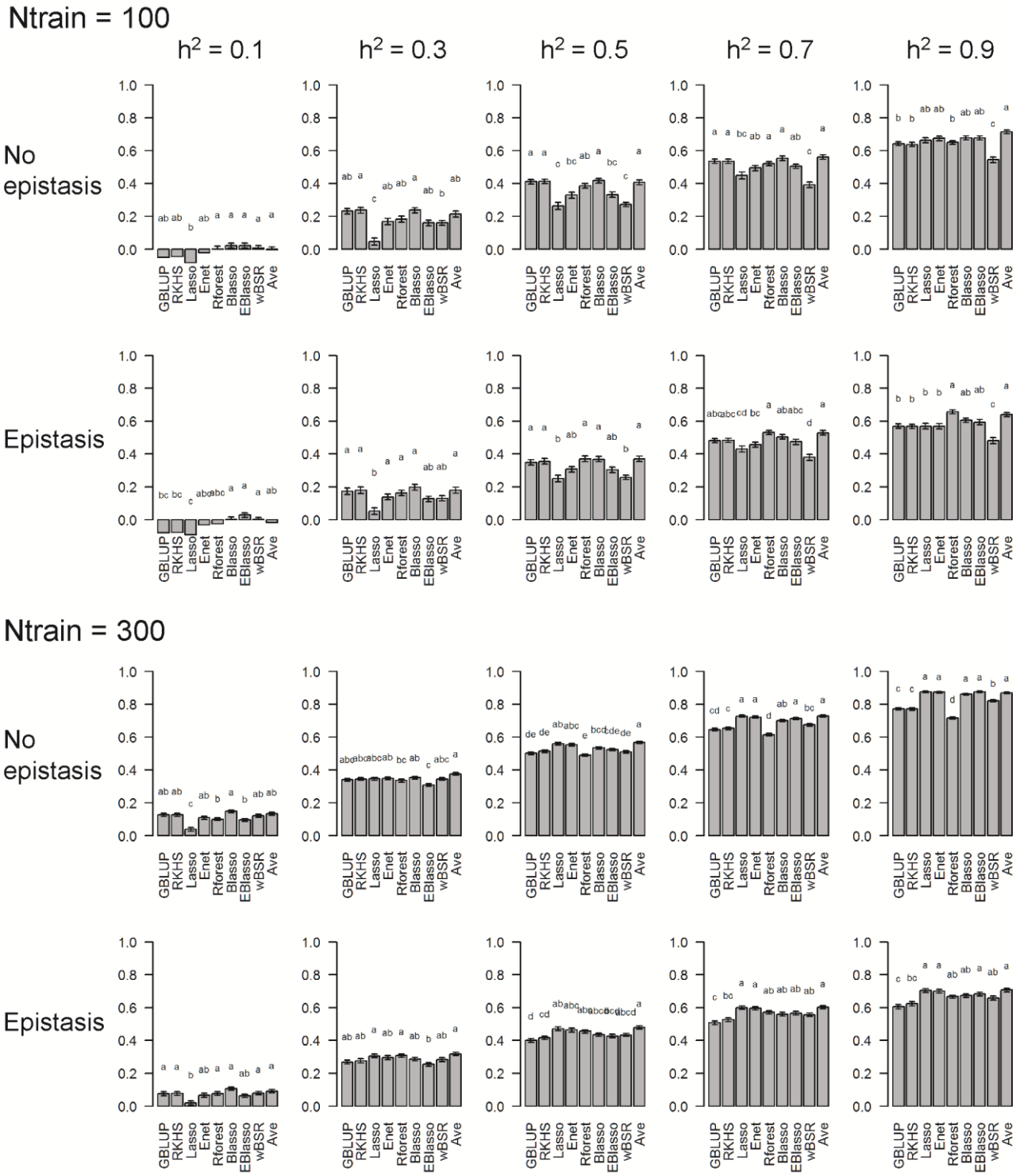

図 3.13 QTL から 10 cM 以上離れたマーカーをランダムに並べ替えることで長距離の連鎖不平衡（LD）を壊したシミュレーションデータにおける予測手法の正確さ。正確さは予測値とシミュレートされた真の遺伝子型値とのピアソン相関係数で評価した。異なる小文字は有意差を表す（P<0.05）。h^2 はエピスタシスがない場合は狭義の、ある場合は広義の遺伝率を表す。手法は左から GBLUP、RKHS、Lasso、ENet、Rforest、Blasso、EBlasso、wBSR、及び Ave。

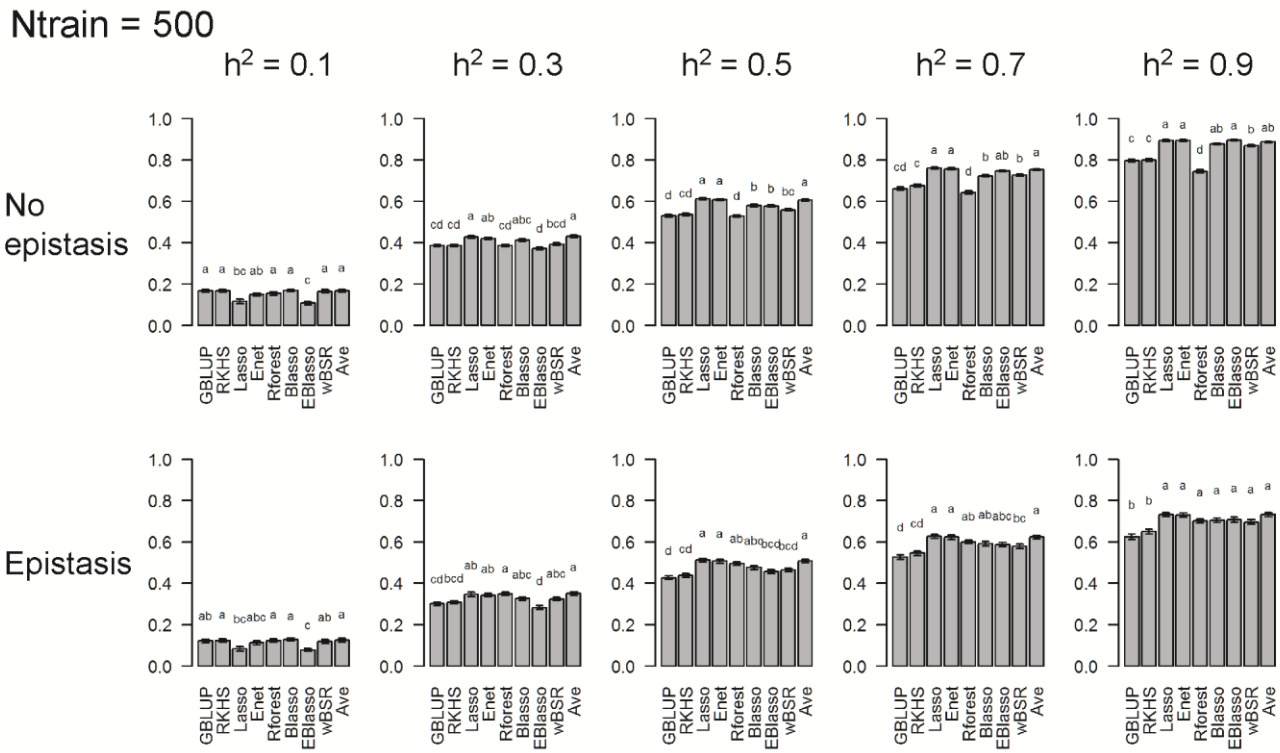

図 3.13 QTL から 10 cM 以上離れたマーカーをランダムに並べ替えることで長距離の連鎖不平衡（LD）を壊したシミュレーションデータにおける予測手法の正確さ（続き）

3.3.5　実形質に最も近いシナリオの探索

実データの解析で観察された手法間の正確さの差を生む要因を調べるために、予測の順位において実形質に最も近いシミュレーションシナリオを探索した（表 3.4 及び図 3.14）。いずれの形質においても、幅広いシナリオにおいて中程度のスピアマン相関係数が観察されたものの、最も高い係数は 0.95 を超えており考慮したシミュレーション条件で予測手法の順位がよく再現できている可能性が示唆された。いずれの形質でも、予測手法の順位において最も近いと示唆されたシナリオにおける遺伝率は中程度かそれ以上に高かった（0.5–0.9）。Nqtl は GW と BL を除いて比較的小さかった（6 または 12）。また 5 形質においてエピスタシスが示唆された。GBLUP を用いて推定された狭義の遺伝率を表 3.4 に示した。

表 3.4　予測手法の順位に関して各形質に最も近いシミュレーションシナリオ

Trait[a]	Scenario			ρ[c] (*p*-value)	Vu/(Vu+Ve)[d]
	Nqtl	h^{2} [b]	Epistasis		
DH	12	0.7	+	0.95 (3.5×10^{-4})	1.00
CL	12	0.7	+	0.82 (1.2×10^{-2})	1.00
PL	6	0.9	+	0.78 (1.7×10^{-2})	0.71
PN	6	0.5	−	0.78 (1.7×10^{-2})	0.51
GL	12	0.9	−	0.67 (5.9×10^{-2})	0.40
GW	36	0.9	−	0.93 (7.5×10^{-4})	0.82
BL	120	0.7	+	0.88 (3.1×10^{-3})	0.55
BW	12	0.9	+	0.85 (6.1×10^{-3})	0.52

[a] DH（到穂日数）、CL（稈長）、PL（穂長）、PN（穂数）、GL（粒長）、GW（粒幅）、BL（玄米長）及び BW（玄米幅）

[b] エピスタシスをシミュレートした（していない）場合は広義（狭義）の遺伝率

[c] スピアマン相関係数

[d] 狭義の遺伝率の推定値. 相加的遺伝分散 （Vu）と残差分散（Ve）は GBLUP で推定した。関係行列は rrBLUP パッケージで提供されている A.mat 関数を用いて作成した（Endelman 2011）。

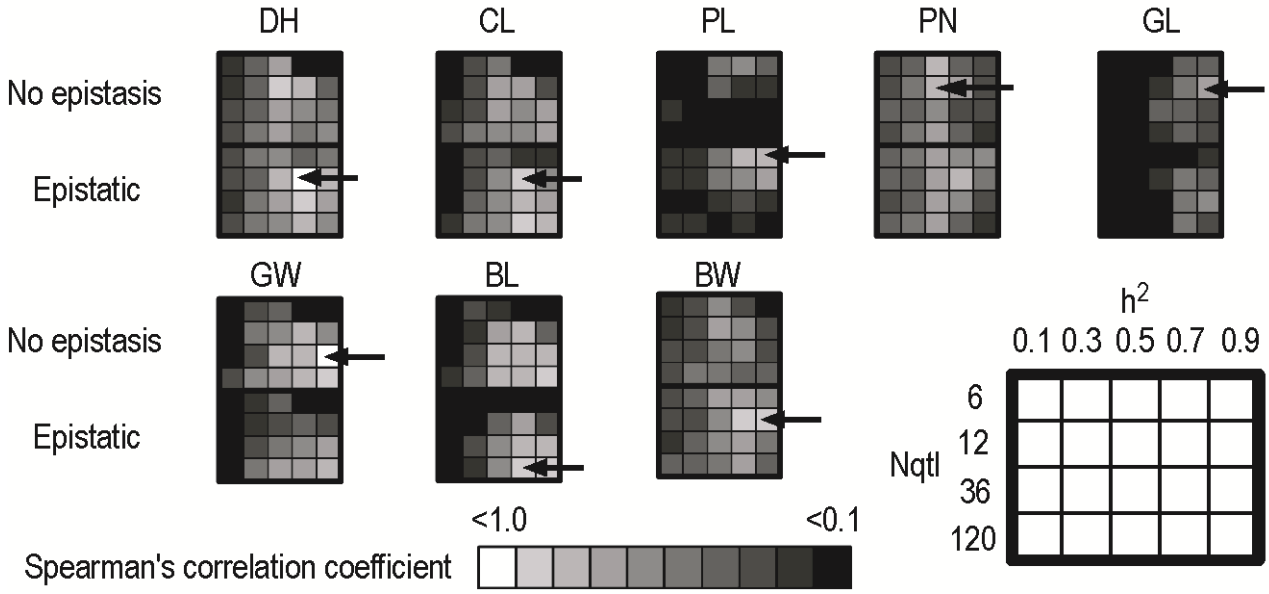

図 3.14　学習用セットのサイズ（Ntrain）が 100 のシミュレーションシナリオと実際の形質間の予測手法順位のスピアマン相関係数。矢印は最も相関係数が高かったシナリオを示す。DH（到穂日数）、CL（稈長）、PL（穂長）、PN（穂数）、GL（粒長）、GW（粒幅）、BL（玄米長）及び BW（玄米幅）。

3.4 考察

本研究では 9 つのゲノムワイド予測手法をアジア栽培イネの 8 つの形質、また実際のマーカー遺伝子型データから作成したシミュレーションデータにおいて比較した。GBLUP、RKHS、及び RForest は実データの 8 形質のうちそれぞれ 1、2、3 つの形質において最も正確な予測を示した。手法の平均をとることにより（つまり Ave により）2 つの形質で最も正確な予測を得た。各形質において最も正確であった手法を用いた場合の表現型と予測値の相関係数は 0.47～0.87 と比較的高く、アジア栽培イネにおいてもゲノムワイド予測が有用である可能性が示唆された。シミュレーション結果から GBLUP、RKHS、RForest、Lasso、及び ENet がある特定のシナリオで特に正確な予測を与える「specialist」としての性格を、Blasso、EBlasso、及び Ave はシナリオ間で安定した正確さを示すことから「generalist」としての性格を持つことが示唆された。以下では本研究から示唆された各予測手法の特徴とその適用可能範囲について述べる。それから実データの解析で観察された予測間の正確さの差を生む要因について議論する。最後にアジア栽培イネにおけるゲノムワイド予測について手法選択に関する議論を行う。

3.4.1 GBLUP

GBLUP は遺伝分散に対してすべてのマーカーが等しく貢献すると想定しているため、予想された通り、GBLUP は Nqtl が増えるにつれて高く順位づけられる傾向にあった。GBLUP またはリッジ回帰の Nqtl が大きいシミュレーションにおける優位性は先行研究でも報告されている（Daetwyler et al. 2010; Jia and Jannink 2012）。ただ本研究の結果は Nqtl が大きくても Ntrain または遺伝率が低下するにつれて GBLUP の優位性が失われ、Blasso や EBlasso、RForest に劣ることが示唆された（図 3.5）。おそらく REML に基づく GBLUP のパラメータ推定は Ntrain や遺伝率が小さい（低い）ほど困難となる一方、Blasso や EBlasso では情報の減少を事前分布が補填していることが原因と考えられた。この解釈が正しければそういった条件下では GBLUP のパラメータ推定をベイズの枠組みの中で行う手法（Legarra et al. 2008; Makowsky et al. 2011）や、ベイジアンリッジ回帰（Crossa et al. 2010; Perez-Rodriguez et al. 2012）が REML に基づく GBLUP より頑強であるかもしれない。

3.4.2 RKHS

シミュレーション結果は RKHS が遺伝率及び Nqtl が高く（大きく）、エピスタシスが作用

する場合に有用であることを示した。エピスタシスが存在する場合に、RKHS が非線形性を考慮しない回帰手法より優位であることは RKHS の理論と一致しており、また先行研究によっても確認されている（Gonzalez-Camacho et al. 2012）。遺伝率が RKHS の順位に影響を与える理由は GBLUP の場合と同様と考えられた。つまり REML に基づく手法はベイズ階層化モデルより情報の減少に弱いと考えられる。RKHS をベイズの枠組みでパラメータ推定すれば（Gonzalez-Recio et al. 2008; de los Campos et al. 2010）、MCMC による計算量の増加が伴うものの、おそらく RKHS はより頑強になるであろう。RKHS は全ゲノムマーカーを用いて作られるカーネル行列に基づくため、RKHS が Nqtl が大きい場合に優位性を示すことは直感的に理解しやすい。RKHS が長い LD 構造を壊した場合にその優位性を失うことも、おそらくこのことと関係があるのと考えられる。つまり長い LD 構造が存在する場合、QTL 効果は多くのマーカーに分散され、それが RKHS の想定によく合うことが考えられる。一方で長い LD 構造が失われ、QTL 効果が近傍の限られた数のマーカーにしか分散されない場合、RKHS は優位性を失うと考えられた。

3.4.3 Lasso

Lasso は Nqtl が小さく、遺伝率が高く、Ntrain が大きい場合に高く順位づけられる傾向があった。Lasso は L_1 正則化項を持つために変数選択の特徴を持ち、一方でリッジ回帰は L_2 正則化項を持つためにすべての回帰係数を 0 に縮約させる性質を持つ。そのため予想された通り Lasso は GBLUP や RKHS といったリッジ回帰と同様の性質を持つ手法より Nqtl が小さいシナリオで優位となった。同様の傾向は先行研究でも報告されている（Usai et al. 2009; Ogutu et al. 2012）。この優位性は Nqtl が増加する、または遺伝率や Ntrain が減少すると急速に失われた。Donoho and Stodden（2006）によって Lasso による変数選択は、真に 0 でない変数の数（k）、サンプル数（n）、変数の数（p）間の比率に非常に敏感であることが示されている。つまり、ある与えられた n 対 p 比（n/p）のもとでは、変数選択は k 対 n 比（k/n）がある点（崩壊点）を超えると働かないことが示されている。この崩壊点は n 対 p 比が小さくなるほど、あるいはノイズが大きくなるほど早く訪れる。ゲノムワイド予測では、n 対 p 比は通常 1 より非常に小さく、また QTL 効果は LD により多くのマーカーに分散されているため k が Nqtl より大きいことが予想され、Lasso による変数選択がうまく働かない場合が多いことが考えられる。そのため、Lasso のゲノムワイド予測における適用可能範囲は狭いことが考えられた。

変数が互いに相関している場合、Lasso による変数選択がうまく働かないことが知られている（Buhlmann and van de Geer S 2011）。Wimmer et al.（2013）らは LD が強くなるにつれて Lasso による変数選択がより困難になることを示した。そのため本研究では QTL から 10 cM 以上離れたマーカー遺伝子型をランダムに並び替えることにより長い LD 構造を壊したシミュレーションを作成し、それにより Lasso による予測の正確さが上昇することを期待した。しかしながら正確さは僅かに減少し、少なくとも本研究での *n* 対 *p* 比（約 100/3000）と *k* 対 *n* 比（6/100 以上、*k* は Nqtl より大きいことが予想されるため）では、用いたイネ集団における LD の強さは Lasso の予測能力に大きな影響を与えていないことが示唆された。

3.4.4 ENet

ENet の順位は Lasso と同様の傾向を示した。これはおそらく両手法とも変数選択の特徴を有しているからであろう。エピスタシスがない場合、ENet は Ntrain が 300 または 500、遺伝率が 0.7 または 0.9、Nqtl が 12 のときに最も正確な予測を与えた。これらの条件は、Nqtl が 6 ではなく 12 であることを除けば、Lasso にとって最適な条件とよく似ていた。この結果は ENet の適用可能範囲が Lasso のそれより少し Nqtl が大きな場合にあることを示唆している。ENet は L_1 及び L_2 正則化項を両方持つため、Lasso のような変数選択の特徴を持つと同時に、リッジ回帰のように（選択された）変数の回帰係数を 0 に縮約させる特徴も持つ（Zou and Hastie 2005; Hastie et al. 2009）。ENet のゲノムワイド予測における Lasso には無い利点は、ENet は相関した 0 でない効果を持つマーカーをグループとして選択できること、また Ntrain より多くのマーカー数を選択できることであろう（Zou and Hastie 2005）。後者に関しては Lasso は最大で Ntrain と同数のマーカーしか選択できないことから生じる。したがって ENet では Lasso より多くのマーカーが選択される傾向があり、それは本研究でも観察され（結果非掲載）、先行研究でも報告されている（Li and Sillanpaa 2012b）。この特徴は Lasso より大きな Nqtl の場合にある ENet の適用可能範囲を定義するのに関連していると考えられる。Nqtl が 36 または 120 のときは ENet は GBLUP、つまりリッジ回帰に劣る傾向があった。同様の結果は Wimmer et al.（2013）によっても報告されている。この先行研究によれば ENet の正確さは Lasso とリッジ回帰の中間に位置するか、いずれにも劣る傾向があり、両方より優れることはほとんどなかった。おそらくこの結果はゲノムワイド予測における ENet の狭い適用可能範囲を反映していると考えられた。

3.4.5 RForest

RForest の適用可能範囲は遺伝率が低く Nqtl と Ntrain が小さい場合に観察された。遺伝率が低い場合の RForest の優位性は、ノイズが大きなデータに有用であることが知られている bagging（Breiman 1996; Dietterich 2000）によることが考えられた。遺伝率と Ntrain に関わらず、RForest の正確さは Nqtl が増加するとともに僅かに減少する傾向にあった。この原因は、大きな Nqtl と強い LD 構造により多くのマーカーが互いに似たような大きさの効果を持つようになり、決定木のノードにおいてランダムに変数（マーカー）選択をしても、決定木の残差間の相関が大きく減少しなかったことにあるかもしれない。または Nqtl の増加による正確さの減少は、ノードにおいて選択するマーカー数（*mtry*）が不足した可能性もある。なお *mtry* はいずれのシナリオでもマーカー数の 3 分の 1（約 1000）で固定している（R パッケージ randomForest の規定値）。Ntrain が増加するにつれて RForest の正確さは上昇するものの、順位は低下する傾向があった。これは他の GBLUP や ENet などの回帰手法が RForest より Ntrain 増加の恩恵を大きく受けていることを示している。これは RForest がサンプルをブートストラップして学習するため、ブートストラップの性質からサンプルの約 63 %しか一度の学習には用いていないことに原因があるのかもしれない（Kohavi 1995）。

RForest はエピスタシスが作用するシナリオにおいて高く順位づけられる傾向があった。RForest に用いられる決定木ではマーカーを逐次的に使用してサンプルを分割することからこの特徴は予想された。相互作用する SNPs（QTLs）を検出するために RForest を用いた先行研究も報告されている（Bureau et al. 2005; Jiang et al. 2009; Yao et al. 2013）。しかし RForest のエピスタシスを検出する能力は 2 つの要因により制限されるであろう。1 つ目は、少なくともエピスタシスに関わるマーカーの 1 つが決定木の分割で用いられるように検出可能な相加的効果を持つことである。2 つ目はエピスタシスに関わるマーカーのすべてが 1 本の決定木の異なるノードでノード分割に用いられる変数の候補として選択されることである。前者は決定木の適切なサイズ（深さ）に関する議論の中で Breiman et al.（1984）によって指摘されている。後者は RForest が決定木の残差間の相関を減らすために、各ノードにおいてランダムに *mtry* 個マーカーをランダムに選択することから生じる（Breiman 2001）。高次の相互作用を検出するためには *mtry* は十分に大きくする必要があるものの、*mtry* の増加は残差間の相関の増加をもたらし bagging の効率を下げることになる。加えて、決定木の深さも十分に大きくする必要があるだろう。本研究では 2 つの QTL の相加的効果間の相互

作用を考慮したが、もしエピスタシスがより複雑な相互作用の上に成り立っていたならば、RForest は優位性を示さなかったかもしれない。

Ntrain が大きいとき（300 または 500）エピスタシスが存在するシナリオでも、RForest はしばしば線形回帰モデル（Blasso や ENet）に劣る傾向があった。同様の傾向は遺伝率が低い場合の RKHS にも見られた。線形回帰モデルの RForest（RKHS）に対する優位性は、相加的遺伝子型値をより正確に予測することで達成されていた（結果非掲載）。したがって、もしエピスタシス分散の割合が本研究（約 40 %）より大きくなると、RForest（RKHS）と線形回帰モデルとの順位は変わる可能性もある。

3.4.6 Blasso

Blasso はシナリオに関わらず安定した順位を示した。Park and Casella（2008）が示したように Blasso（及び EBlasso）は Lasso とリッジ回帰の折衷物と言うことができる。正則化パラメータ λ^2 が増加するとともに、Blasso はリッジ回帰よりも速く回帰係数（マーカー効果）を 0 に引き寄せるが、Lasso が行うように回帰係数を完全に 0 に圧縮することはしない。この特徴と、おそらく事前分布によりもたらされる情報、さらに λ^2 のハイパーパラメータのグリッド探索により、Blasso は遺伝的構造に関わらず安定した予測結果を残したと考えられる。ENet も Lasso とリッジ回帰の折衷物と言えるが、ENet は Lasso 同様に変数選択の特徴を持つため（Hastie et al. 2009）、Blasso と ENet の順位の傾向は著しく異なった。第 12 回 QTLMAS ワークショップで提供された主働遺伝子が大きな影響を与える（Nqtl が小さい）形質を模したデータセットでは（Lund et al. 2009）、Blasso の正確さは Lasso や ENet を下回ったが（Li and Sillanpaa 2012b）、リッジ回帰より上であった（Usai et al. 2009）。本研究のシミュレーション結果でも同様の傾向が観察された。すなわち、Nqtl が 6、Ntrain が 300 または 500、遺伝率が 0.5 以上の場合、Blasso は Lasso 及び ENet とリッジ回帰の中間の正確さを示した。Li and Sillanpaa（2012b）はこの原因を Lasso とリッジ回帰との折衷であるが故に、Lasso より回帰係数を過小推定しやすいためと考察している。

3.4.7 EBlasso

EBlasso ではマーカー毎に縮約の度合いを調節することができ、そのため Blasso より QTL マッピングにより適していることが示唆されている（Mutshinda and Sillanpaa 2010）。シミュレーションでは EBlasso は Blasso 同様に遺伝的構造や Ntrain に関わらず安定した予測結果

を残した。予想した通り EBlasso は Nqtl が 6、エピスタシスがないシナリオにおいて Blasso より高く順位づけられる傾向があった。このことは EBlasso が Blasso より QTL とマーカーの関連シグナルを検出するのに向いていることを反映していると考えられた。しかしながら Nqtl が 6 で遺伝率が高い場合は、EBlasso は Lasso に劣る傾向があった。このことは Lasso と比較して EBlasso がいまだマーカー効果を過小推定する傾向にあることを示唆している。長い LD 構造を壊すことにより、EBlasso による予測は遺伝率や Ntrain に大きく依存するようになった。この傾向は Lasso とよく似ていたものの、その原因は推測できなかった。

3.4.8 wBSR

wBSR の順位はシナリオに関わらず一様に低く、明確な適用可能範囲を示さなかった。この結果はおそらく 3 つの要因によると考えられる。1 つ目は学習セットの大きさ（Ntrain）が小さいことである。wBSR と等価な手法である BayesA または BayesB はシミュレートされた主働遺伝子によって制御される形質において、GBLUP またはリッジ回帰より優位であることが報告されている（Zhang et al. 2010; Clark et al. 2011; Sun et al. 2012; Daetwyler et al. 2013）。BayesB は比較的微働遺伝子による影響の大きな形質においても GBLUP より優位であることが報告されている（Zhang et al. 2010; Sun et al. 2012）。これらの先行研究では Ntrain は 1000 以上であり、本研究の Ntrain の最大値（500）より大きい。2 つ目は強い LD 構造である。Lasso による変数選択同様に、BayesB による変数選択も強い LD 構造により阻害され、それにより同手法の予測能力も影響を受けることが報告されている（Wimmer et al. 2013）。図 3.11 に示したように、交差検証により選択された、0 でない効果を持つマーカーの割合を示す κ 値は、Ntrain が 300 または 500、遺伝率が 0.3 より大きい場合に、長い LD 構造を壊すことにより著しく減少した。このことは LD 構造を壊すことによって、より少ないマーカーがより大きな重みを持ったことを示しており、変数選択が効率よく働くようになったことを示している。しかしながら wBSR の正確さはそれでも Lasso や ENet を下回っており、長い LD 構造だけが要因では無いことが伺われた。3 つ目の要因はマーカー効果の分散の事前分布を定義するハイパーパラメータ ν 及び S^2 が最適化されていないことである。本研究では κ をグリッド探索により決定したが、ν は 4 に固定し、S^2 は κ と ν を用いてマーカーが説明する表現型分散の割合する想定に基づき決定した。マーカーが説明可能な表現型分散の割合は未知であるため、全てのシナリオで 0.5 を用いた。この選択が wBSR の予測能力に影響を与えたかもしれない。また S^2 を求めるために用いた式はマーカー間の

連鎖平衡を想定しており（Habier et al. 2011）、これが本研究で用いた LD 構造が強い集団においては問題となったかもしれない。Gianola et al.（2009）が指摘するように、マーカー効果の事前分布の影響は wBSR のモデル構造においては避けられない。これは個々のマーカー効果の分散が、たった 1 つの観察値（そのマーカーの効果）からしか学習されないからである。ハイパーパラメータで規定される事前分布の影響は、ハイパーパラメータを未知変数として推定することによりある程度は緩和できる可能性が指摘されている（Nadaf et al. 2012）。

3.4.9 Ave

Ave は集合学習の最も単純な適用方法であると言える。集合学習が成功するための 1 つの鍵は、集合したメンバーの多様性である。Krogh and Vedelsby（1995）は「集合のあいまいさ（ensemble ambiguity）」、つまり集合メンバーの予測値間の分散が大きくなるほど汎化誤差が小さくなることを示した。Ave は Daetwyler et al.（2013）においても Hickey and Gorjanc (2012)のベンチマークデータに対して試みられているが、特に主働遺伝子が大きな影響を与える形質においては正確さを改善することにはつながらなかった。この結果の原因はおそらく集合のあいまいさが十分でなかったことかもしれない。これは集合メンバーが変数選択を行う線形回帰モデル（BayesB、BayesC、Lasso、Bayesian stochastic search variable selection）、及びリッジ回帰とそれと等価である GBLUP によって成り立っていたことによる。特にオリゴジェニックな形質では、変数選択を行う線形回帰モデルをメンバーに加えてもそれらの手法はいくつかの主働 QTL をもとに同じような予測値を返すことが予想されるため、集合のあいまいさを増加させることにはならないだろう。この考えに基づくと、集合学習または複数手法の平均をとる方法は、より複雑で非相加的効果が優勢な形質に対して有効なのかもしれない。本研究のシミュレーションでは Ave はエピスタシスが作用する場合に僅かだがより正確な結果が得られているので、この仮説は正しいようにみえる。Heslot et al.（2012）では正確さの向上は 18 形質中 2 形質にしか認められなかった。しかしながら RKHS がほとんどの形質において最も高い正確さを与えていることから、非相加的効果の影響が強いことを示唆しており、この結果は先の仮説に反するようにみえる。Ave をはじめとする集合学習のゲノムワイド予測における特徴を知るためには、適切なメンバーの選択などについてさらなる研究が必要であろう。

3.4.10　実データで観察された予測手法間の正確さの差を生んだ要因

実データを用いた比較研究では、手法間の正確さの差を生んだ要因は通常明らかにできない。本研究では実データでの手法の順位に最も近いシミュレーションシナリオを探すことによりその要因を推測した。この推測は遺伝的構造、つまり形質を支配する QTL の数やその効果の分布、及び遺伝率を推定することと深く関連するが、遺伝的構造の推定は一般に困難であり、Agarwala et al.（2013）の研究が示唆するように非常に多数のサンプル数が必要となる。本研究ではサンプル数が限られていたために、広い範囲のシナリオに渡って中程度のスピアマン相関係数が観察された（図 3.14）。加えて 2 種類の遺伝率、すなわち最も近いシナリオでの狭義または広義の遺伝率と、実データの分散成分推定から得られた狭義の遺伝率、の間での大きなかい離が特に GL と BW で観察された。このかい離は当然ながら狭義・広義の定義の違いにも起因するであろうが、他に単純化されたシミュレーション条件や分散成分推定の不確かさにも起因したと考えられる。それにも関わらず、多くの形質で実データとそれに手法間順位において最も近いシナリオとでは高いスピアマン相関係数が観察され、シミュレーション条件に実データにおける手法間の正確さの差を生んだ要因が含まれていたことが示唆された。今回扱った形質のうち、DH についてはこれまで集中的に QTL 解析が行われ、日本水稲イネ集団において出穂を制御している遺伝子とその多型が数多く検出されている。代表的な遺伝子として *Hd1*（Yano et al. 1997; Yano et al. 2000）、*Hd6*（Takahashi et al. 2001）、*Hd16*（Matsubara et al. 2008）、*Hd17*（Matsubara et al. 2008）、*Ghd7*（Xue et al. 2008）などが挙げられる。これらの遺伝子を含む比較的少数のネットワークにより出穂が制御されていることが明らかにされてきており（Izawa 2007; Tsuji et al. 2011）、これは Nqtl が 12 と少なく、かつエピスタシスの関与を示唆した本研究の結果（表 3.4）と概ね一致する。しかし一方で、6 年間計測された DH の年次間相関係数は平均 0.98 以上であることを考慮すると、表 3.4 で示すように今回示唆された（広義の）遺伝率 0.7 は過小推定されていると考えられる。

この実験に関して、実データとそれに順位において最も近いシナリオ間で、2 つ大きな違いが観察された。1 つ目は DH と CL で観察された正確さである。DH 及び CL の正確さは 0.60（wBSR による CL の予測）から 0.87（RForest による DH の予測）に渡っていたが、一方で最も近かったシナリオでは 0.43（wBSR）から 0.51（RForest）と、正確さが予測値と真の遺伝子型値とのピアソン相関で測られていたにも関わらず、実データよりかなり低かった。この結果は手法間の順位に影響を与えずに正確さ（ピアソン相関係数）に

影響を与える要因が存在し、それがシミュレーションで考慮されていないことを示唆しているのかもしれない。2 つ目は GL と BL のそれぞれ最も近いシナリオにおける Nqtl が 12 と 120 と非常に大きいことである。この両形質が全く異なる遺伝的支配を受けているとは考えにくく、実際表現型値間のピアソン相関係数は 0.87 と高かった。しかしながら、測定誤差など非遺伝性の要素が GL の表現型値に影響を与えている可能性も考えられた。その根拠は GL に対する予測の正確さが BL より低い傾向があったこと、実データに最も近いシナリオは GL においてスピアマン相関係数が最も低かったこと、そして分散成分推定により得られた狭義の遺伝率が GL で最も低かったことである。

3.4.11 アジア栽培イネにおける予測手法の選択

エピスタシスが作用することが予想される場合は、RKHS または RForest が推奨される。これら両手法は相補的な適用可能範囲を示し互いを補いあえる特徴を持つ。エピスタシスが無い場合は GBLUP または Blasso が推奨される。Lasso や ENet は適用可能範囲が狭く、今後 Ntrain が非常に大きくならない限り実用性は低いであろう。Blasso はその頑強性から魅力的であるが、頑強性という観点からは複数手法を平均する方法が優れるであろう。特に、手法の選択は Ntrain や遺伝率が小さい（低い）ときには重要性を増すことに留意する必要がある。RKHS のようなカーネル回帰手法や RForest や Ave のような集合学習は今後さらに改良する余地があると考えられる。

3.5 摘要

ゲノムワイド予測は表現型の評価なしに選抜を可能とする新たな植物育種手法であるが、これまでアジア栽培イネにおいてその有用性は検証されていない。そこで本研究ではアジア栽培イネの 8 つの農業形質について 9 つのゲノムワイド予測手法を比較した。また実際の遺伝子型データから作成したシミュレーションデータにおいても比較を行った。本研究の目的はアジア栽培イネの品種集団において最適な予測手法を探ることと、シミュレーションデータで比較することで各手法の適用可能な範囲を探索し、実データで観察された予測手法間の差を生んだ要因を探ることである。実データにおいてはランダムフォレスト（RForest）が 3 つの形質で最も正確な予測を示し、reproducing kernel Hilbert space 回帰（RKHS）と全手法の平均（Ave）が 2 形質で、Genomic BLUP（GBLUP）が 1 形質において最も正確な予測を行った。各形質において最も正確であった手法を用いた場合の表現型

と予測値の相関係数は 0.47～0.87 と中程度以上であり、アジア栽培イネにおいてもゲノムワイド予測が有用である可能性が示唆された。シミュレーションにおいては Bayesian lasso、Extended Bayesian lasso 及び Ave が様々なシミュレーションシナリオに渡り安定した正確さを示す一方、GBLUP、RKHS、Lasso、Elastic net、RForest はそれぞれ固有の適用範囲を示した。weighted Bayesian Shrinkage Regression は明確な適用範囲を示さなかった。手法の正確さの順位に着目すると実データの各形質において、観察された手法間の順位に非常に近いシミュレーションシナリオが存在していた。このことは今回シミュレーションにおいて考慮した条件、QTL の数、遺伝率、エピスタシスの有無、学習セットのサイズが手法間の相対的正確さに強い影響を与えていた可能性を示している。

4　黒毛和種におけるゲノムワイド予測有用性の検証及びゲノムと血縁情報両方に基づく育種価予測手法に関する諸検証

4.1　諸論

黒毛和種は日本固有の肉専用品種であり、細い筋繊維や密な脂肪沈着に表される高い肉質で知られる（内藤 1978）。近年は国内だけでなく国外でも需要が高まっており、黒毛和種の生産が日本の主要な輸出産業として発展していくことが望まれる。そのためには肉質や肉量、あるいは飼料効率などの様々な形質における継続的な遺伝的改良が重要となる。これまでゲノミックセレクションの有用性は、ホルスタインやジャージーなどの乳用種だけでなく肉用種においても活発に研究されてきており（Garrick 2011; Saatchi et al. 2011; Mujibi et al. 2011）、黒毛和種においてもゲノム情報を利用した育種手法の研究が、高品質な畜産物の生産力及び国際的な競争力向上のためにも必要であるものの、実際にゲノムワイド予測の有用性を検証した報告は未だない。

第 1 章で述べたように、ゲノミックセレクションには既存の育種プログラムとどのように組み合わせるかという課題がある。現時点ですでにゲノミックセレクションが利用されているホルスタイン種では、従来の BLUP 法とゲノミックセレクションを平行して用いている（Hayes et al. 2009）。すなわち BLUP 法により種雄牛の育種価を推定し、その推定育種価からゲノムワイドマーカーの効果を推定した後、種雄牛候補の若い雄牛を予測し選抜する。つまり BLUP 法は選抜するためではなく学習用の育種価を推定するために用いられている。この方式は現行の BLUP 法を利用した育種価推定方法を変更する必要がないという利点の一方で、必要な操作が増えること、選抜された（つまり能力の高い）種雄牛の育種価からマーカー効果を学習することによるバイアスが生じること、BLUP 法とゲノムワイド予測を別に行うために情報損失の可能性があることなどの欠点もある。また第 3 章で明らかなように予測の正確さは学習セットのサイズに大きく依存するため、この手法では BLUP 法から推定した信頼度の高い育種価を持つ種雄牛が多数存在する必要がある。しかしながら、黒毛和種はホルスタインより集団が非常に小さいため高い信頼度の育種価を持つ種雄牛も少ない。このことは世界的な規模で育種が進むホルスタインのような品種とは異なる育種手法が黒毛和種に必要であることを示唆している。

一方で BLUP 法とゲノムワイド予測を組み合わせた手法も提案されている（Legarra et al. 2009; Aguilar et al. 2010; Christensen and Lund 2010; Garrick et al. 2014; Liu et al. 2014; Fernando et al. 2014）。これらの手法は推定の基礎を血縁情報に置きながら、それにマーカー遺伝子型の情報も加えることでメンデリアンサンプリングも考慮できる手法となっている。Single-step genomic BLUP（ssGBLUP）は血縁情報に基づいた育種価の分散共分散行列（分子血縁係数行列、一般に A 行列と呼ばれる）に、ゲノムワイドマーカーから推定した分散共分散行列（ゲノム関係行列または G 行列）を混ぜ合わせた H 行列を作成し、それを A 行列の代わりに用いることで育種価を推定する（Legarra et al. 2009、Aguilar et al. 2010）。この手法は現状の育種方式を大きく変更することなしにゲノム情報を活用できるという利点の他、（１）育種価の高い個体のみから学習することにより生じるバイアスを回避できること、（２）ゲノム情報は血縁情報に追加される形で用いられるので、マーカー遺伝子型を持つ個体が少数の場合でも血縁情報だけに基づく予測に比べ正確さが減少しないこと、及び（３）血縁情報だけから育種価を推定し、そこから訓練する場合より情報の損失が少ないこと、が期待される。血縁情報から育種価を推定しそこからさらに学習する、という複数のステップを踏まないためにシングルステップ法またはワンステップ法と呼ばれる。

ssGBLUP については実用上まだ幾つかの点で明確でないことがある。1 つは A と G 行列を混合する際の割合（混合パラメータ）であり、それが予測の正確さと偏りに影響を及ぼすことが報告されている（Aguilar et al. 2010）。2 つ目は効率的に予測の正確さを向上させるために、どのような個体を選択的にジェノタイピングすることが必要か、つまり選択的ジェノタイピングの方法である。3 つ目はマーカー遺伝子型を持たない個体への予測能力も向上するか、という点である。これらの点を明確にすることは、今後黒毛和種のゲノム育種を ssGBLUP に基づき行っていく上で重要となるだろう。また一つ目が ssGBLUP の手法に特徴的な課題である反面、2 及び 3 つ目の課題はシングルステップ法共通の課題と考えられ、今後 ssGBLUP に限らずシングルステップ法によりゲノムワイド予測を行う上でも重要な知見となり得る。

本研究では黒毛和種におけるゲノムワイド予測の有用性を検証し、またシングルステップ法によるゲノムワイド予測の諸課題を解決するために、黒毛和種集団において ssGBLUP と BLUP の予測能力の比較を行った。形質としては代表的な枝肉形質である Beef marbling score（BMS）、枝肉重量（CW）、ロース芯面積（REA）を対象とした。

4.2 材料と手法

4.2.1 表現型記録

一般社団法人家畜改良事業団（LIAJ）が行っている後代検定事業において、これまで 616 頭の種雄牛が選抜された。種雄牛の生年は 1980 年から 2008 年であった。後代検定に関わる 17,347 頭の肥育牛に関する BMS、CW、REA の表現型値は公益社団法人日本食肉格付協会（JMGA）により計測され、LIAJ より家系情報とともに本研究に提供された。これら肥育牛のと畜年月は 2006 年 1 月から 2013 年 5 月であった。これらの肥育牛は合計 466 頭の種雄牛の産子であり、そのうち 240 頭が LIAJ の、残り 226 頭が LIAJ 以外により造成された種雄牛であった。LIAJ の残りの種雄牛 376 頭（616－240）についてもその産子の表現型記録が収集されていたが、それらの記録は古く現在とは異なる方法で評価されていたため本研究からは除外した。これらに加えて、山形県農業総合研究センター（YPARC）から 3,089 頭の肥育牛の表現型記録及び家系情報も提供を受けた。YPARC の肥育牛は全て LIAJ 種雄牛（84 頭）の産子であった。前述の 240 頭の LIAJ 種雄牛との重複を除くと、合計 274 頭の LIAJ 種雄牛が表現型記録を持つ産子を有した。これら 2 種類のデータセットから肥育牛 20,436（17,347＋3,089）頭の表現型記録及び家系情報を用いた。

ssGBLUP と BLUP の予測能力比較のために、表現型記録を LIAJ 種雄牛の後代検定時期により 2 つに分割した（図 4.1A）。1 つ目のセット（Data I）は 2011 年前期までに検定された種雄牛（種雄牛群 B、図 4.1A）196 頭の後代肥育牛 12,064 頭から成る。この肥育牛には YPARC のものも含まれる。2 つ目のセット（Data II）は 2011 年後期以降に検定された種雄牛（種雄牛群 C）78 頭の後代肥育牛 8,372 頭から成る。種雄牛群 A（342 頭）は後代を持たない。種雄牛群 B は Data I 及び II に後代を有するが、種雄牛群 C は Data I には後代を持たず Data II のみに持つ。前述のように種雄牛 616 頭は全てマーカー遺伝子型を持つ。肥育牛のうち、Data I に含まれる YPARC の肥育牛 952 頭（肥育牛群 D）と、Data II に含まれる種雄牛群 C の後代肥育牛 370 頭（肥育牛群 E）の合計 1,322 頭については後述のようにマーカー遺伝子型が得られた。選択的ジェノタイピング検証のために、ssGBLUP に用いるマーカー遺伝子型のセットを図 4.1B のように 8 通り試み、それぞれにおいて肥育牛群 E の表現型値を予測した（表現型値予測、後述）。なおセット 7 及び 8 は予測される肥育牛群 E のマーカー遺伝子型を用いない場合であるため、この検証はマーカー遺伝子型を持たない個体に対する ssGBLUP の予測能力評価も兼ねている。

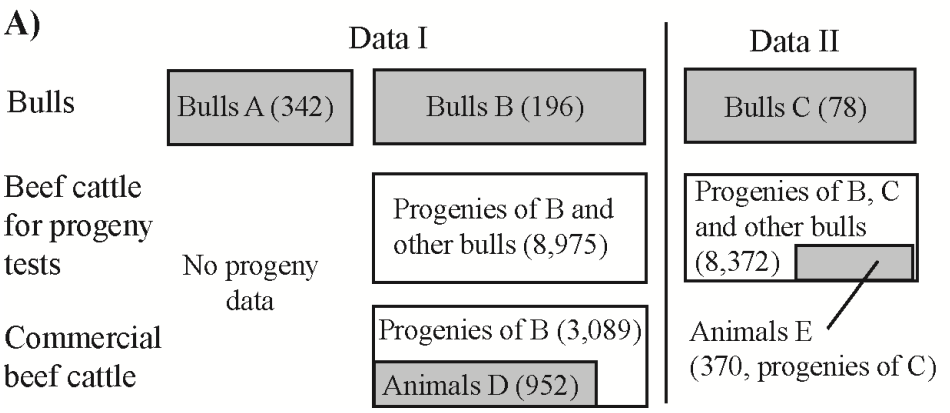

図 4.1　(A) 本研究で用いたデータの概略。データは種雄牛の後代検定時期により 2 つに分割した (Data I と II)。白と灰のボックスはそれぞれマーカー遺伝子型を持たない個体、及び持つ個体を表す。ボックス内の数字は個体数を表す。種雄牛は 3 グループに分割した。種雄牛群 A（Bulls A）は今回用いたデータにおいて表現型記録を持つ後代を有しない。種雄牛群 B（Bulls B）は Data I 及び II いずれにも表現型記録を持つ後代を有する。種雄牛群 C（Bulls C）は表現型記録を持つ後代を Data II には有するが Data I には有しない。手法の予測能力は種雄牛群 C の育種価及び肥育牛群 E（Animals E）の表現型値を Data I に含まれる表現型記録を用いて予測することで評価した。(B) H 行列を作成するために用いた遺伝子型セット。A から E は種雄牛群 A から肥育牛群 E までのグループに対応し、どのグループのマーカー遺伝子型を用いて H 行列を作成したかを＋記号で示している。

4.2.2　マーカー遺伝子型

LIAJ 種雄牛 616 頭のうち、32 及び 584 頭はそれぞれ Illumina BovineSNP50 beadchip version 1 及び 2（Illumina Inc.）を用いて LIAJ 家畜改良技術研究所によりジェノタイピングされた。種雄牛のコールレイトは 0.98 以上であった。常染色体上のマーカーを、コールレイト（0.9 以上）、マイナーアリル頻度（0.01 以上）、及びハーディ・ワインベルグ平衡のカイ二乗検定 P 値（0.01 以上）により選定し、38,755 マーカーにおける遺伝子型を得た。肥育牛の一部は Illumina BovineLD beadchip で LIAJ 家畜改良技術研究所によりジェノタイピングされた。このチップは 50K チップ上の 6,412 マーカーを含む。コールレイト 0.9 以上の肥育牛について、種雄牛のマーカー遺伝子型をリファレンスとして Beagle（Browning and Browning 2009）で補完し、38,755 マーカーの遺伝子型を得た。このうち父親との間でメンデル遺伝様式の確認を行い、それに従わないマーカーを 100 以上持つ個体を除いた。最終的に 1,576 頭の肥育牛についてのマーカー遺伝子型が得られた。このうち 952 頭は肥育牛群 D、370 頭は肥育牛群 E であり、残りの 254 頭は YPARC から提供された肥育牛であったものの、表現型記録が後述する基準を満たさなかったためにマーカー遺伝子型のみ使用された。

4.2.3　予測手法

ssGBLUP は血縁情報により定義される個体間の関係行列（分子血縁係数行列、A 行列）とゲノムワイドなマーカー遺伝子型により定義されるゲノム関係行列（G 行列）を次式のように混合した H 行列を、育種価の分散共分散行列として用いる。

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau\left(\alpha\mathbf{G} + \beta\mathbf{A}_{22}\right)^{-1} - \varpi\mathbf{A}_{22}^{-1} \end{bmatrix} \quad (4.1)$$

ここで τ、ω、α、β は A 行列と G 行列の混合割合を決定する混合パラメータであり、A_{22} は A 行列のうちマーカー遺伝子型が得られた個体間の A 行列の要素を表す。この式は以下のようにして導かれる（Aguilar et al. 2010）。まずマーカー遺伝子型が得られていない個体の育種価を $\mathbf{u}_1$、得られている個体の育種価を $\mathbf{u}_2$ とする。さらにマーカー遺伝子型が得られていない個体間の分子血縁係数行列を $\mathbf{A}_{11}$、マーカー遺伝子型が得られていない個体と得られている個体間の行列を $\mathbf{A}_{12}$ とする。$\mathbf{u}_1$ と $\mathbf{u}_2$ の同時分布は

$$\begin{aligned} p\left(\mathbf{u}_1,\ \mathbf{u}_2\right) &= p\left(\mathbf{u}_1 \mid \mathbf{u}_2\right) p\left(\mathbf{u}_2\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2\right)^{\mathrm{T}} \mathbf{A}_{11}'\left(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2\right)\right] \exp\left[-\frac{1}{2}\mathbf{u}_2^{\mathrm{T}}\mathbf{A}_{22}^{-1}\mathbf{u}_2\right] \end{aligned}$$

となる。ここで$\mathbf{A}'_{11}$はA行列の逆行列（$\mathbf{A}^{-1}$）の$\mathbf{A}_{11}$に対応する成分を表す。ここで$\mathbf{u_2}$の分散共分散行列をG行列で置き換えると、

$$\exp\left[-\frac{1}{2}\left(\mathbf{u}_1-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2\right)^{\mathrm{T}}\mathbf{A}'_{11}\left(\mathbf{u}_1-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2\right)\right]\exp\left[-\frac{1}{2}\mathbf{u}_2^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{u}_2\right]$$

$$=\exp\left\{\left[-\frac{1}{2}\begin{bmatrix}\mathbf{u}_1, & \mathbf{u}_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{A}'_{11} & -\mathbf{A}'_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}'_{11} & \mathbf{G}^{-1}+\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}'_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1, & \mathbf{u}_2\end{bmatrix}\right]\right\}$$

$$=\exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{u}_1, & \mathbf{u}_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{A}'_{11} & \mathbf{A}'_{12}\\ \mathbf{A}'_{21} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}+\mathbf{A}'_{22}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1, & \mathbf{u}_2\end{bmatrix}\right\}$$

$$=\exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{u}_1, & \mathbf{u}_2\end{bmatrix}^{\mathrm{T}}\left(\mathbf{A}^{-1}+\begin{bmatrix}0 & 0\\ 0 & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}\end{bmatrix}\right)\begin{bmatrix}\mathbf{u}_1, & \mathbf{u}_2\end{bmatrix}\right\}$$

を得る。つまり$\mathbf{u}_1$と$\mathbf{u}_2$の分散共分散行列の逆行列が

$$\mathbf{A}^{-1}+\begin{bmatrix}0 & 0\\ 0 & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}\end{bmatrix}$$

で表されることになる。式4.1ではこれに4つの混合パラメータを加えている。τとωは混合時のA_{22}行列とG行列の重みを表す。αとβはG行列が非特異行列になるのを防ぐ目的でA_{22}行列と加算されるときの相対的な重みを表す。BLUPの場合はA行列を育種価の分散共分散行列として用いる（Henderson 1984）。BLUP及びssGBLUPともにA行列は表現型値またはマーカー遺伝子型を持つ個体から5世代遡り作成した。H行列の計算はpreGSf90プログラム（Aguilar et al. 2010; Aguilar et al. 2011）を用いた。

G行列はVanRaden (2008)で提案されている1番目の手法を用いて作成した。この手法では基礎集団、つまり選抜が始まる前の任意交配が仮定できる集団におけるマーカーアリル頻度が必要となるが、これは未知の値であるため、現在の集団におけるアリル頻度を用いた。A行列の基礎集団は血縁をそれ以上遡ることができない個体の集団と仮定するが、この手法で求めたG行列の基礎集団は、必ずしもA行列のものとは一致しない。H行列を用いた場合、これに起因して遺伝分散や推定育種価の偏りが起きることが知られており、その解決法としてG行列とA_{22}行列の対角及び非対角要素の平均値を合わせる方法が提案されているため（Chen et al. 2011; Forni et al. 2011; Vitezica et al. 2011）、本研究でもこれに従った。αとβはそれぞれ0.95と0.05に設定した。τとωは同じ値を用いて、1、0.75、0.5、0.25、0.1の5つの値を試した。

ssGBLUPの比較手法としてBLUPも用いた。これ以降BLUPによる推定育種価をEBV（estimated breeding value）、ssGBLUPによる推定値をGEBV（genomic estimated breeding

value）として区別する。

4.2.4 線形混合モデル

以下の線形混合モデルを用いた。

$$Y_{ijkl} = S_i + YM_j + F_k + u_{ijkl} + a_1 t_{ijkl} + a_2 t_{ijkl}^2 + e_{ijkl}$$

ここで Y_{ijkl} は肥育牛 *ijkl* の表現型記録、S_i は *i* 番目の性の効果（*i* は 1 または 2）、YM_j は *j* 番目のと畜年月日の効果（*j* は 1 から 64）、F_k は *k* 番目の農家の効果（*k* は 1 から 118）、u_{ijkl} は育種価、t_{ijkl} は肥育牛 *ijkl* の月齢偏差、a_1 及び a_2 は回帰係数、e_{ijkl} は残差である。これらのうち u_{ijkl} と e_{ijkl} 以外は母数効果として扱い、u_{ijkl} は変量効果とした。u_{ijkl} は ssGBLUP の場合は $\mathrm{MVN}\left(\mathbf{0},\ \mathbf{H}\sigma_a^2\right)$、BLUP の場合は $\mathrm{MVN}\left(\mathbf{0},\ \mathbf{A}\sigma_a^2\right)$ に従うと想定した。ここで σ_a^2 は相加的遺伝分散を示す。e_{ijkl} は $N\left(0,\ \sigma_e^2\right)$ に従うとした。σ_e^2 は残差分散であり、狭義の遺伝率は $\hat{\sigma}_a^2 \big/ \left(\hat{\sigma}_a^2 + \hat{\sigma}_e^2\right)$ として推定される。BMS については予備調査において月齢偏差の二次項がないモデルで低い AIC が得られたため（結果非掲載）、二次項を除いたモデルを用いた。ssGBLUP と BLUP ともに、母数効果、変量効果、及び分散成分（σ_a^2 と σ_e^2）の推定は airemlf90 プログラム（Misztal et al. 2002）を用い、REML 法で行った。

4.2.5 予測能力の比較・評価

本研究では肥育牛表現型値の予測及び種雄牛育種価の予測の両方を試みた。肥育牛表現型値の予測では、まず全表現型記録を用いて母数効果を推定し、肥育牛群 E の表現型記録から母数効果を除いた後、この補正済み表現型記録を Data I の記録のみから予測した。全表現型記録からの母数効果の推定は、BLUP 及び ssGBLUP いずれでも可能だが、どの τ（ω）値であっても ssGBLUP から推定した母数効果は、BLUP によるそれとほぼ等しかったため（結果省略）、BLUP での推定値により補正を行った。混合パラメータ（τ）の影響、選択ジェノタイピングの検証、及びマーカー遺伝子型を持たない個体に対する予測能力の検証は、この補正済み表現型値の予測を通じて行った。

種雄牛育種価の予測は、まず全表現型記録（Data I＋II）を用いて種雄牛群 C の育種価を推定し、その推定育種価を Data I の表現型記録のみを用いて予測した。この場合、種雄牛群 C の育種価の推定にそれらの後代記録（肥育牛群 E）を用いないために、後代の

ない若い候補種雄牛の育種価予測を模すことができる。ただし、予測対象となる全記録からの推定育種価もあくまで推定値であり、真の育種価ではない点に注意が必要である。種雄牛群 C は平均 21.6（±4.8）頭の後代肥育牛を有した。全表現型記録からの育種価の推定は BLUP 及び ssGBLUP いずれでも行い、予測値は両手法の推定値と比較した。

予測能力の測定は予測育種価と、補正済み表現型値（表現型値予測）または推定育種価（種雄牛育種価予測）とのピアソン相関係数を用いた

4.3 結果

4.3.1 分散成分の推定

全表現型記録（Data I 及び II）から、BLUP 及び ssGBLUP を用いて推定した遺伝分散及び残差分散を表 5.1 に示した。ssGBLUP についてはマーカー遺伝子型セット 1 を用い、各 τ（ω）値（1、0.75、0.5、0.25、0.1）のもとで推定した分散成分推定値を示した。

表 4.1 全データから推定した相加的遺伝分散（Vu）及び残差分散（Ve）[a]

Method	τ[b]	BMS[c]			CW[3]			REA[3]		
		Vu	Ve	$h^{2\,d}$	Vu	Ve	h^2	Vu	Ve	h^2
BLUP	–	2.91	1.47	0.66	1273.4	1007.3	0.56	30.87	40.59	0.43
ssGBLUP	1.0	2.45	1.78	0.58	1279.6	1000.4	0.56	29.46	41.44	0.42
	0.75	2.84	1.54	0.65	1458.7	889.6	0.62	34.49	39.32	0.47
	0.5	3.07	1.39	0.69	1519.5	852.3	0.64	36.05	37.36	0.49
	0.25	3.18	1.31	0.71	1481.6	875.7	0.63	35.37	37.77	0.48
	0.1	3.10	1.35	0.70	1387.7	934.6	0.60	33.41	38.99	0.46

[a] ssGBLUP では遺伝子型セット 1 を用いた。最も高い推定値は下線で示した。

[b] G と A 行列の混合割合

[c] BMS（beef marbling score）、 CW（枝肉重量）、REA（ロース芯面積）

[d] 狭義の遺伝率（相加的遺伝分散を相加的遺伝分散と残差分散の和で割った値）

4.3.2　表現型値予測

肥育牛群 E の表現型値はまず全データから BLUP を用いて推定した母数効果で補正した。これら補正した肥育牛群 E の表現型値を Data I の表現型記録から BLUP 及び ssGBLUP を用いて予測した（表 4.2）。マーカー遺伝子型セットは 1 を用い、τ（ω）は 5 つの値（1、0.75、0.5、0.25、0.1）を試した。最も正確な予測は BMS では τ が 0.5 のとき、CW と REA では τ が 1 のとき得られた（それぞれ 0.39、0.44、0.42）。いずれの形質でも、全ての τ 値において ssGBLUP で BLUP より高い正確さが得られた。

次に τ 値を BMS においては 0.5、CW と REA においては 1.0 とし、1 以外のマーカー遺伝子型セットを用いて表現型値予測を行った（表 4.3）。セット 2（種雄牛群 A の遺伝子型が用いられていない）及びセット 4（種雄牛群 C の遺伝子型がない）ではセット 1 とほぼ同等の結果が得られた。このことから種雄牛群 A（つまり表現型記録を持つ後代がいない種雄牛）、及び種雄牛群 C（予測された肥育牛の父）のマーカー遺伝子型情報の予測に対する貢献度が小さいことが示唆された。肥育牛群 D（表現型記録とマーカー遺伝子型両方を持つ肥育牛）のマーカー遺伝子型を除いた場合（セット 6）、予測の正確さは全種雄牛の遺伝子型を除いた場合（セット 5）より減少した。予測された肥育牛（肥育牛群 E）の遺伝子型を除いた場合（セット 7）、正確さは減少したがそれでも BLUP よりは高く、ssGBLUP による予測がマーカー遺伝子型を持たない個体に対しても有効であることが示唆された。しかしながら、さらに種雄牛群 C（予測された肥育牛の父）の遺伝子型を除くと、BLUP とほぼ同程度の正確さとなった。

表 4.2　異なる G と A 行列の混合割合を用いた場合の表現型値予測 [a]

Method	τ^b	BMS[c]	CW[c]	REA[c]
BLUP	–	0.29	0.27	0.30
ssGBLUP	1.00	0.38	<u>0.44</u>	<u>0.42</u>
	0.75	0.38	0.42	0.41
	0.50	<u>0.39</u>	0.40	0.40
	0.25	0.37	0.35	0.37
	0.10	0.34	0.30	0.34

[a] 全データから BLUP で推定した母数効果で補正した肥育牛群 E の表現型を Data I に含まれる表現型記録から予測した。ssGBLUP では遺伝子型セット 1 を用いた。最も高い相関係数は下線で示した。

[b] G と A 行列の混合割合

[c] BMS（beef marbling score）、 CW（枝肉重量）、REA（ロース芯面積）

表 4.3　異なる遺伝子型セットを用いた場合の表現型値予測 [a]

Method	Set[b]	BMS[d]	CW[d]	REA[d]
BLUP[c]	–	0.29	0.27	0.30
ssGBLUP	1[c]	0.39	<u>0.44</u>	<u>0.42</u>
	2	0.39	0.44	0.41
	3	0.37	0.41	0.39
	4	<u>0.39</u>	0.44	0.42
	5	0.37	0.40	0.38
	6	0.35	0.36	0.38
	7	0.30	0.33	0.41
	8	0.29	0.26	0.30

[a] 全データから BLUP で推定した母数効果で補正した肥育牛群 E の表現型を Data I に含まれる表現型記録から予測した。ssGBLUP における G と A 行列の混合割合（τ）は BMS では 0.5、CW と REA では 1.0 とした。最も高い相関係数は下線で示した。

[b] H 行列作成に用いた遺伝子型セット

[c] 表 4.2 の値を再記載

[d] BMS（beef marbling score）、 CW（枝肉重量）、REA（ロース芯面積）

4.3.3　育種価予測

まず全表現型記録を用いて BLUP 及び ssGBLUP により種雄牛群 C の育種価を推定した。ssGBLUP の τ 値は BMS については 0.5、CW と REA については 1 とした。マーカー遺伝子型はセット 1 を用いた。この 2 種類の推定育種価（EBV と GEBV）を Data I の表現型記録のみを用いて BLUP 及び ssGBLUP で予測した。ssGBLUP で予測した際の τ 値及びマーカー遺伝子型セットは、育種価推定時と同じものを用いた。これら 2 種類の推定育種価と 2 種類の予測手法により、4 通りの比較を行った（表 4.4）。つまり EBV を BLUP または ssGBLUP で予測する場合と、GEBV を BLUP または ssGBLUP で予測する場合である。CW と REA については ssGBLUP は、どちらの推定育種価においても BLUP より高い正確さを示した。一方で、BMS については ssGBLUP は GEBV についてのみ BLUP より高い正確さを示した。

表 4.4 育種価予測の正確さ [a]

Prediction	BMS[b]		CW[b]		REA[b]	
Method	EBV	GEBV[c]	EBV	GEBV[c]	EBV	GEBV[c]
BLUP	<u>0.58</u>	0.60	0.75	0.75	0.62	0.61
ssGBLUP[d]	0.57	<u>0.63</u>	<u>0.81</u>	<u>0.84</u>	<u>0.71</u>	<u>0.79</u>
Cor.[e]	0.99		0.99		0.98	

[a] 全データを用いて推定した種雄牛群 C の育種価を Data I に含まれる表現型記録から予測した。最も高い相関係数は下線で示した。

[b] BMS（beef marbling score）、 CW（枝肉重量）、REA（ロース芯面積）

[c, d] ssGBLUP における G と A 行列の混合割合（τ）は BMS では 0.5、CW と REA では 1.0 とした。遺伝子型セットは 1 を用いた。

[e] 予測された種雄牛（種雄牛群 C）における EBV と GEBV のピアソン相関係数

4.3.4　遺伝的趨勢

図 4.2 に 1990 年以降に生まれた 595 頭の LIAJ 種雄牛育種価（EBV）の遺伝的趨勢を図示した。それ以前に生まれた残りの 21 頭については、趨勢を見るには生年が疎らであったため図から除外した。育種価の増加は BMS と REA で顕著であったが、CW では観察されなかった。このことは、CW がこの後代検定事業において重要視されていなかったことを示唆している。もうひとつ顕著な傾向は BMS における育種価の分散が、全年代を通じて他の形質より小さいことが挙げられる。これらの傾向は GEBV でも観察された（結果非掲載）。

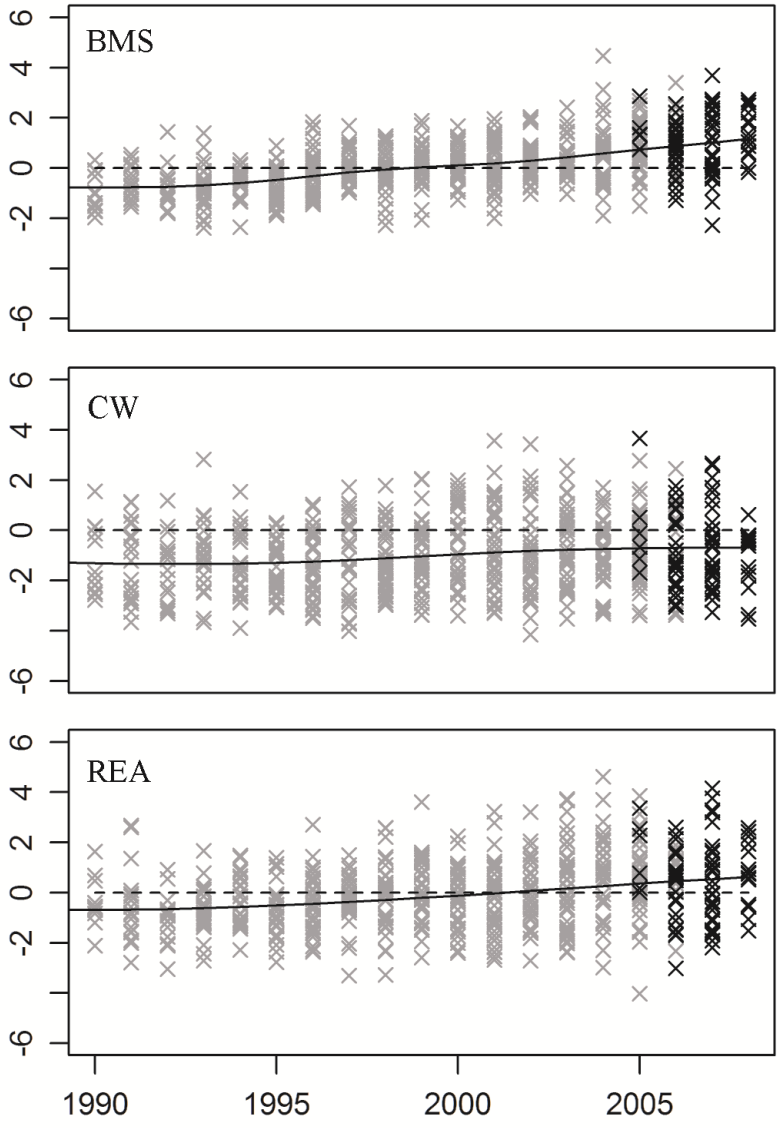

図 4.2　種雄牛育種価の遺伝的趨勢。全データから ssGBLUP で推定した GEBV を生年に沿って図示した。上段は BMS（Beaf marbling score）、中段は CW（枝肉重量）、下段は REA（ロース芯面積）。G と A 行列の混合割合（τ）は BMS で 0.5、それ以外で 1.0 とした。遺伝子型セットは 1 を用いた。GEBV は標準偏差でスケーリングして集団平均（破線）を 0 として示した。黒点は予測された種雄牛群 C を表す。実線は R 関数「smooth.spline」で描いた。

4.4 考察

ゲノム情報を利用した ssGBLUP は従来の血縁のみに基づく BLUP より、BMS の育種価予測の 1 例を除き正確な予測を与え、黒毛和種育種におけるその有用性が示唆された。これはゲノムワイド予測の有用性を黒毛和種において確認した最初の報告である。本研究では ssGBLUP の BLUP に対する優位性を検証することを目的としたが、ホルスタイン種で実用されている予測手法（BLUP により学習セットの育種価推定を行いそれに基づき GBLUP による予測を実行する手法）といずれの形式がより黒毛和種集団に有用であるかについては、まだ検証の余地がある。この BLUP と GBLUP を併用した手法は北米のホルスタイン集団では ssGBLUP とほぼ同等の予測能力を示しているものの（Aguilar et al. 2010）、これは 6,000 頭以上の種雄牛の育種価から学習した結果である。今回マーカー遺伝子型を持つ黒毛和種種雄牛 616 頭のうち、表現型記録を有する後代を持つ、つまり信頼度がある程度高い育種価が得られる可能性があるのは種雄牛群 B 及び C の 274 頭であり、このデータ規模で ssGBLUP と同程度の正確さで予測可能なのかは、今後検証していく必要がある。ただし仮に予測能力がほぼ同等であった場合、ssGBLUP には予測のバイアスや従来手法（BLUP）からの移行の容易さといった利点があると考えられる。

BLUP または ssGBLUP を用いた遺伝率の推定値は 3 形質いずれについても、これまで報告されている BLUP を用いた推定値と同程度であった。例えば BMS では 0.64（Inoue et al. 2011）または 0.51（Nogi et al. 2011）、CW では 0.43（Arakawa et al. 2009）または 0.61（Nogi et al. 2011）、REA では 0.52（Arakawa et al. 2009）と報告されている。本研究においては BMS の遺伝率が最も高く（0.58 から 0.71）、REA で最も低かった（0.42 から 0.49）。遺伝率が高いほど正確に予測できることが期待されたが、いずれの検証方法（表現型値予測及び育種価予測）及び手法（BLUP と ssGBLUP）を用いても BMS における正確さが最も低く、CW が最も高い傾向にあった。BMS における低い正確さの一因は、近年の世代において育種価の分散が小さいことにあると考えられた。例えば、全表現型記録を用いて推定された種雄牛の GEBV の変動係数は、1.22（BMS）、2.24（CW）、3.73（REA）であった（マーカー遺伝子型セット 1 を用いて計算、τ 値は BMS では 0.5、それ以外の形質では 1.0）。BMS は最も経済的価値の高い形質であるため、BMS における小さな分散はおそらく最も強く選抜がかかった結果であると考えられた。一方で CW と REA における BMS より高い正確さは、おそらくより高い育種価の分散に起因すると考えられた。

ssGBLUP による遺伝率と遺伝分散の推定値は τ 値に依存して変化した。τ 値は 1 の

とき、BMS の遺伝分散は BLUP による推定値から減少する傾向にあったが、CW と REA ではおおむね一定であった。考えられる原因の一つは、G 行列によって捉えられる遺伝分散の大きさが、A 行列によるものより少ないことが挙げられる。これはマーカーと QTL との連鎖不平衡がマーカー密度の不足などで十分でないことによると考えられる。この考察は、τ 値を 1 から下げていくと（つまり G に対する A の割合を大きくすると）遺伝分散の推定値が上昇することからも支持される。また BMS における遺伝分散減少の別の要因として、この形質に対する強い選抜が関与しているかもしれない。G 行列と A 行列を結合する際の一つの問題は、両者の基礎集団が同等でないことにある。A 行列の場合血縁がそれ以上遡れない個体集団を基礎集団とし、その集団の育種価を 0 と想定する。一方でマーカー遺伝子型を持つ個体（つまり G 行列に関与する個体）は基本的に新しい世代であるため、G 行列の基礎集団の育種価は、A 行列の基礎集団の育種価、つまり 0 より大きくなることが予想される。これを経験的に解決するために「4.2.3　予測手法」で述べたよう G 行列の対角及び非対角の要素平均を A_{22} 行列のそれらと同じとする調整を行っているが、形質によりこれまでの選抜強度に差があれば、G 行列の基礎集団の育種価の A 行列のそれ（0）からの逸脱程度は形質により異なっていても不思議ではない。そのためこのスケーリングでは BMS の場合 G 行列と A 行列の基礎集団が揃っていない可能性があり、それが遺伝分散の推定値に影響を与えた可能性がある。

遺伝率の推定値と予測の正確さの関係については、最も高い遺伝率または尤度が得られた τ 値で、表現型値予測において必ずしも最も高い正確さが得られたわけではなかった（表 4.2）。それより、A 行列を用いて得られた遺伝率に最も近い推定値が得られた τ 値で、最も高い正確さが得られる傾向にあった（表 4.1 及び 4.2）。このことは遺伝率が正確な予測のために τ 値を調節するよい指標となることを示唆している。しかしながら、τ 値により予測の正確さが変化するとはいえ、その差はごく僅かであるということも観察された。例えば BMS の表現型値予測では、τ 値が 0.5 のときもっとも高い正確さが得られたが、τ 値が 1 のときと比較してわずか 0.007 しか差がなかった。同様に先行研究においても、A 行列と G 行列の混合比の調節による正確さの向上は観察されたとしても小さい。例えばアメリカのホルスタインにおける育種価予測では、最も高い決定係数は τ 値が 0.7 及び 0.8 のときに得られているが、τ 値が 1 のときと比較して 0.02 の差しかなかった（Aguilar et al. 2010）。また Christensen et al.（2012）はブタの日増体重及び飼料要求率において α 及び β 値を調節し、α（β）値がそれぞれ 0.7（0.3）及び 0.85（0.15）のときに正確さが最も高かったものの、

0.95（0.05）のときと比較し 0.01 以下の差しかなかったと報告している。Harris et al.（2012）はウシ複数品種からなる集団の泌乳量において 3 つの τ 値（0.3、0.5、0.7）を試したものの、正確さに与える影響は非常に小さかったと報告している。結果として、τ 値を推定遺伝率をもとに調節することは予測能力を最大化するために有用であろうが、それによる正確さの増加は小さいことが考えられる。

表現型値予測において、種雄牛群 A（表現型記録を持つ後代がいない種雄牛グループ）または C（表現型値を予測された肥育牛群 E の父親であるグループ）のマーカー遺伝子型を除くことは、予測の正確さにあまり影響を与えなかった（表 4.3）。加えて、種雄牛群 B（表現型記録を持つ後代を持つ種雄牛グループ）のマーカー遺伝子型を除くと、すべての種雄牛の遺伝子型を除いた場合とほぼ同じ正確さとなった。これらの結果は、種雄牛群 B のマーカー遺伝子型が予測において非常に重要であり、一方で A と C の影響が非常に小さいことを示している。種雄牛群 B には主に黒毛和種集団において非常に影響力の大きい、つまり後代が多数いる種雄牛から成り立つ。今回の結果はそういった種雄牛のマーカー遺伝子型が予測の上で非常に重要であることを示唆する。種雄牛群 C の影響が小さいことは予想外であったが、このことは種雄牛群 B が頻繁に C の父、または予測された肥育牛群 E の母の父であったことに起因するのかもしれない。そういった場合はマーカー遺伝子型の情報がすでに飽和しており、種雄牛群 C の遺伝子型が追加の情報をもたらさなかった可能性が考えられる。肥育牛群 D（表現型記録とマーカー遺伝子型情報いずれも持つグループ）のマーカー遺伝子型を除いた場合、正確さはすべての形質で減少し、このことは表現型記録を持つ個体のマーカー遺伝子型を決定する重要性を示唆している。種雄牛群 A 及び C の影響が小さかったことを考慮すると、表現型記録を持つ個体のジェノタイピングは、後代の少ない種雄牛のそれより優先度が高いことが示唆されたといえる。ただしこの結論には、マーカー遺伝子型の補完については考慮していない。つまり、マーカー遺伝子型を補完した個体について予測する場合、補完の正確さが予測の正確さに影響を与えることが指摘されているが（Pszczola et al. 2011）、本研究では予測された肥育牛群 E の遺伝子型を補完するために、その父である種雄牛群 C の遺伝子型を用いた。そのため、もし種雄牛群 C の遺伝子型がなければ、肥育牛群 E の遺伝子型補完がどの程度の正確さで可能であるかは考慮しておらず、補完の正確さが減少するならば、予測の正確さもやはり減少する可能性があることには留意する必要がある。

ssGBLUP のマーカー遺伝子型を持たない個体に対する予測能力は、肥育牛群 E の

マーカー遺伝子型を用いないことで評価した（表 4.3）。その結果すべての形質で正確さは減少したが、それでも BLUP による予測よりは上回っており、ssGBLUP がマーカー遺伝子型を持たない個体の予測についても有効であると示唆された。しかしながら、さらに種雄牛群 C のマーカー遺伝子型も除去すると、正確さは BLUP とほぼ同程度となった。このことは ssGBLUP によるマーカー遺伝子型を持たない個体の予測は、その個体の父母または近い血縁にある個体がジェノタイプされている場合に限り、BLUP より有用であることが示唆された。

育種価予測については、予測対象となる種雄牛群 C の育種価を推定するために、まず複数の選択肢がある。つまり BLUP と異なる τ 値から得られた H 行列を用いる ssGBLUP である。もし種雄牛群 C が多数の表現型記録を持つ後代を有していれば手法間での推定値の違いは小さくなるはずだが、今回は種雄牛 1 頭あたりの後代数が少頭数（平均 21.6 頭）であったため、推定育種価が手法により異なることが予想された。本研究では現在育種価推定の標準的手法である BLUP と、BMS については τ 値を 0.5 とした ssGBLUP、CW と REA については τ 値が 1 の ssGBLUP を用いて全表現型記録から育種価を推定した。ssGBLUP におけるこれら τ 値の選択は、これらの値のもと表現型値予測で最も高い正確さが得られたためである。また同じ τ 値は Data I の表現型記録から予測する際にも用いられた。育種価予測の結果は、ssGBLUP が 1 つの例外、すなわち BMS における EBV の予測、を除いて BLUP より高い正確さを示した（表 4.4）。BMS における EBV 予測では、BLUP がわすかに ssGBLUP より高い正確さを示した。しかしながら EBV より GEBV の方が表現型値予測においてより高い正確さを示したことから、GEBV が真の育種価に近いと考えられた。つまり GEBV を予測した結果の方がより適切に手法に予測能力を反映しており、この結果は育種価予測においても ssGBLUP が BLUP より優れていることを示していると考えられた。

結論として ssGBLUP を用いることにより、従来の育種システムを大きく変更することなく、全ゲノム情報を利用可能なことが示された。ssGBLUP によるゲノムワイド予測は後代検定前の予備選抜に効果を発揮すると考えられる。これにより選抜強度の増加或いは後代検定に必要な後代数の減少とそれによる費用及び労力の減少が期待できる。加えてゲノムワイド予測は農家において母親候補を選抜する際にも有効となるだろう。ssGBLUP を用いることにより、その個体がジェノタイピングされていなくてもその父母がジェノタイピングされていれば、予測の正確さの向上が期待できるため、農家はジェノタイピングのコスト負担なしにその恩恵を受けられる可能性がある。また本研究はジェノタイピング

の優先順位についても明確な基準を与えた。すなわち、後代数の多い影響力の大きな種雄牛のジェノタイピングを最優先とし、さらに表現型記録を持つ肥育牛のジェノタイピングを影響力の小さい種雄牛より優先すべきである。

4.5 摘要

黒毛和種は日本固有の肉用牛であるが、肉質が高いことから今後国内だけではなく国外でもその畜産製品の消費増加が見込まれる。現在は血縁に基づいた BLUP 法が育種の中心であるが、全ゲノム情報を利用したより効率的な手法を構築していくことは、国内畜産業の国際的競争力を高める観点からも重要となってくる。しかしながら、全ゲノム情報を利用した育種価予測の有用性は未だに検証されていない。そこで本研究では血縁情報に全ゲノム情報を加えた single-step genomic BLUP（ssGBLUP）法について、その有用性を BLUP 法と比較した。表現型の予測において ssGBLUP 法は全ての形質で BLUP 法より正確な予測を与え、育種価の予測においてはほぼ同等な結果となった 1 形質をのぞき BLUP より正確な予測を与えた。これらの結果は全ゲノム情報が黒毛和種育種においても有用であることを示唆している。また ssGBLUP における 3 つ課題、すなわち血縁情報とゲノム情報の混合割合の最適化、選択的ジェノタイピングの方法、及びゲノム情報を持たない個体に対する予測能力についても検証を行った。混合割合は推定される遺伝率には影響を与えたものの、予測能力には大きな影響を与えなかった。選択的ジェノタイピングについては、後代数の多い種雄牛、また表現型記録を持つ肥育牛を優先的にジェノタイピングすることが予測の正確さ向上に重要であることがわかった。また ssGBLUP はゲノム情報を持たない個体に対しても BLUP より正確な予測を行うことが示されたが、それはその父あるいは近親がジェノタイピングされている場合に限られた。

5　ゲノムと環境情報両方に基づく表現型値予測モデルの構築と検証：イネ出穂期予測への適用

5.1　諸論

農業においては収量や開花期、ストレス耐性や形態など様々な形質を予測することが望まれるが、これらの形質はしばしば非常に複雑で、遺伝子や環境、管理方法間の相互作用を含む多数の因子の影響を受ける。このような複雑な形質の予測において鍵となる技術の 1 つに、作物モデルあるいは生態生理学的モデルと呼ばれる数理モデルがある（Yin et al. 2004; Hammer et al. 2006）。作物モデルは一般に生理学的なプロセスを表す複数の関数からなり、様々な環境・管理の条件、例えば気温、日長、降水量、或いは播種日を入力値として植物の生長や環境への応答を記述する（Tardieu 2003）。作物モデルのパラメータは遺伝的な調節を受けることが広く認識されているため、遺伝的係数（genetic coefficients）やモデル入力形質（model input traits）とも呼ばれる（Yin et al. 2003）。そのため作物モデルを用いることで遺伝子と環境・管理方法との相互作用を考慮出来ることが知られている（例えば Chapman et al. 2002; Chapman et al. 2003）。

作物モデルのパラメータに対する遺伝学的解剖、つまり QTL マッピングも行われている。例えばオオムギの生長や開花期（Yin et al. 2000; Hunt and Pararajasingham 1995）、トウモロコシの葉の生長（Reymond et al. 2003）、モモ果実の生理学的品質（Quilot et al. 2005）、イネ及びコムギの出穂期（Nakagawa et al. 2005; Bogard et al. 2014）、乾燥ストレス下でのイネ収量（Gu et al. 2014）、アブラナ属の葉の発達や開花期（Uptmoor et al. 2008; Uptmoor et al. 2009; Uptmoor et al. 2012）などの作物モデルのパラメータに対して QTL マッピングが行われている。推定された QTL 効果は次にパラメータ値を推定するために用いることができる。実際先に挙げた研究の一部では、全 QTL 効果の和をとることにより得られたパラメータ値を用いて、新たな遺伝子型（つまり系統・品種）の表現型値または既知の遺伝子型の未試験環境下での表現型値を予測している。例えば新しい遺伝子型を未試験の環境下で予測した場合、観察値と予測値間の決定係数はトウモロコシの葉の伸長率では 0.74（Reymond et al. 2003）、イネの収量では 0.20 から 0.21 であり（Gu et al. 2014）、またコムギの出穂期予測における平均二乗誤差（root mean squared error、RMSE）は 6.3 日であった（Bogard et al. 2014）。

これらの研究は有意と判定された QTL（マーカー）のみを用いて作物モデルのパラメータを予測している。しかしながら、検出される QTL は通常遺伝分散の一部しか説明しないため、作物モデルを用いた予測の正確さはゲノムワイドマーカーを全て予測変数として取り込むことにより、つまりゲノムワイド予測手法を用いることにより向上する可能性がある。第 3 章で述べたように、これまで多数の統計的手法がゲノムワイド予測に適用されておりその性質が明らかとなってきている。よってこれらの手法は上述の作物モデルを基礎とした表現型の予測にも有用であると考えられる。ゲノムワイド予測の簡単な適用方法は、先に挙げた QTL を利用した予測研究のように、作物モデルのパラメータを形質として扱いこれを予測することであろう（以降 2 段階法と呼ぶ)。しかしながらこの手法では、作物モデルのパラメータ推定時の不確実性がゲノムワイド予測モデルの構築時に考慮されない、また作物モデルのパラメータ推定は通常遺伝子型（系統）毎に行われるため（例えば Nakagawa et al. 2005)、他の遺伝子型の情報を利用できない、といった欠点が考えられる。一方で作物モデルをゲノムワイド予測モデルと合わせ、マーカー効果と作物モデルパラメータを同時に推定するような統合的なモデリングも考えることができる。統合モデルでは作物モデルパラメータ推定時の不確実性を予測モデル構築の際に考慮することができ、またマーカーを通じて遺伝子型間で情報がやり取りされるため、推定がより頑強になることが期待される。しかしながら、統合モデルはいまだにどの作物及び形質においても開発されておらず、その可能性は評価されていない。

本研究では統合モデルの予測能力を作物モデル及び 2 段階法とイネ（*Oryza sativa*, L）の出穂期において比較した。イネ出穂期を選択したのはそれが農業上重要な形質であることと、予測のための有用な作物モデル、つまり発達割合（developmental rate、DVR）タイプの開花期モデル（DVR モデル)、が既に開発されているためである（Yin et al. 1997 及び Nakagawa et al. 2005)。2 段階法で用いるゲノムワイド予測手法としては EBlasso と RKHS を比較した。本研究では 2 種類の統合モデルを提案する。1 つ目はゲノムワイドマーカーを DVR モデルのパラメータの共分散構造を定義するために用いるモデルである。ここでは非相加的遺伝子効果を考慮するためにガウシアンカーネル行列を用いた。2 つ目は DVR モデルのパラメータをゲノムワイドマーカーに回帰し、パラメータとそれに対するマーカー効果を同時に推定するモデルである。回帰手法としては EBlasso を用いた。この統合モデルのために変分ベイズと MCMC を組み合わせた推定手法も新たに開発した。このモデルは Malosetti et al.（2006）で提案された非線形の混合モデルと類似するが、本研究の手法はモ

デル選択の必要がないため、多数のゲノムワイドマーカーに対して適用可能という点でより洗練されていると言える。これらの手法の予測能力は緯度の大きく異なる 6 つの試験地で栽培された戻し交雑系統（backcross inbred line、BIL）を用いて比較した。

5.2 材料と手法

5.2.1 DVR モデル

Nakagawa et al.（2005）の DVR モデルを出穂期予測に用いた。このモデルは Yin et al.（1997）によって提案された「three-stage Beta model」に基づく。このモデルは出穂前の 3 つの発達段階を想定する。最初の段階は「幼若段階（juvenile phase）」であり植物は光に感応性を持たない。次の段階は「光感応段階（photo sensitive phase）」であり、光刺激が発達割合の蓄積に影響を与える。最後が「後光感応段階（post-photo sensitive phase）」であり、この段階は光感応段階から出穂までの光に感応性のない段階を指す。DVR は出芽日から蓄積していき、発達ステージ（developmental stage、DVS）を押し上げる。

$$DVS = \sum_{d=1}^{D} DVR_d$$

ここで D は出芽日からの日数を指す。出穂は DVS が 1 に達すると起こる。出芽後 d 日目における DVR は以下のように定義される。

$$DVR_d = \begin{cases} f(T_d)/G & \text{if } DVS_d < DVS_1 \text{ or } DVS_d > DVS_2 \\ f(T_d)g(P_d)/G & \text{if } DVS_1 \le DVS_d \le DVS_2 \end{cases}$$

ここで T_d は日平均気温（℃）、P_d は理論日長（h）、f 及び g はそれぞれ T_d 及び P_d を DVR_d と結びつけるための関数、G（≥ 0）は気温と日長が最適であったときの到穂日数（days to heading、DH）、DVS_1 及び DVS_2 はそれぞれ幼若段階と光感応段階が終了する時期を表す。f 及び g は以下のように定義される。

$$f(T_d) = \begin{cases} \left[\left(\frac{T_d - T_b}{T_o - T_b} \right) \left(\frac{T_c - T_d}{T_c - T_o} \right)^{(T_c - T_o)/(T_o - T_b)} \right]^{\alpha} & \text{if } T_b \le T_d \le T_c \\ 0 & \text{if otherwise} \end{cases}$$

$$g(P_d) = \begin{cases} \left[\left(\frac{P_d - P_b}{P_o - P_b} \right) \left(\frac{P_c - P_d}{P_c - P_o} \right)^{(P_c - P_o)/(P_o - P_b)} \right]^{\beta} & \text{if } P_o \le P_d \\ 1 & \text{if } P_o > P_d \end{cases}$$

ここで T_b、T_o、T_c はそれぞれ生長における最低、最適、最高気温を指し、P_b、P_o、P_c も同様に最低、最適、最高日長を表す。Nakagawa et al. (2005)に従い T_b、T_o、T_c、P_b、P_o、及び P_c はそれぞれ 8℃、30℃、42℃、0 h、10 h、及び 24 h とした。パラメータ α（≥0）及び β（≥0）はそれぞれ温度感応性及び日長感応性を表し、値が大きいほど感応性が上昇し、最適値に近い温度或いは日長でないと *DVR* の蓄積が鈍化する。パラメータ値によりどのように感応性が変化するかは Nakagawa et al. (2005)の図 1 に例示されている。DVS_1 及び DVS_2 は Nakagawa et al. (2005)に従い

$$DVS_1 = 0.145 + 0.005G$$
$$DVS_2 = 0.345 + 0.005G$$

とした。あらかじめ値を定めたパラメータのうち幾つか、例えば T_o などは遺伝子型により変動する可能性があるものの、本研究では *G*、α、及び β の遺伝的差異を考慮した。従ってこれ以降遺伝子型 *i* の DVR モデルパラメータを G_i、α_i、及び β_i として表し、全遺伝子型の値を含むベクターを **G**、**α**、及び **β** として表す。これらのパラメータはいずれも 0 以上という制限があることに留意する。

5.2.2　予測手法

表 5.1 に本研究で用いた予測手法の一覧を表示した。DVR モデルのパラメータ推定手法として Nelder-Mead 法（NM）と、モデルパラメータ（*G*、α、β）の共分散として単位行列を用いたベイズ法（Bayes_I）を用いた。これらの手法はゲノム情報を用いていないために、これ自体では新たな遺伝子型に対する予測ができない。2 段階法として、NM と回帰手法、RKHS 及び EBlasso と組み合わせた手法を用いた（それぞれ NM+R 及び NM+E と呼ぶ）。これらの回帰手法は NM によって推定された *G*、α、及び β を被説明変数とする。3 つの統合モデル、Bayes_G、Bayes_Glog、及び Bayes_Elog は Bayes_I をもとにベイズ統計の枠組みで開発した。Bayes_G では新たな遺伝子型について予測するために、*G*、α、及び β の共分散をゲノムワイドマーカーから計算したガウシアンカーネル行列で定義した。Bayes_Glog もまたガウシアンカーネル行列を用いるが、推定を自然対数変換した *G*、α、及び β のもとで行う。この変換により *G*、α、及び β の事前分布（後述）の平均及び分散を推定することが容易となる。Bayes_Elog では自然対数変換した *G*、α、及び β を EBlasso を用いてゲノムワイドマーカーに回帰し、DVR モデルパラメータとそれに対するマーカー効果を同時に推定する。これらの手法は以下で説明する。

表 5.1　用いた予測手法の一覧

Genomic information	Method type[a]	Method abbreviation[b]	Log[c]	Usage of genomic information	Cross-validation schemes[d]
Not use		NM	−		LOEO
		Bayes_I	−		LOEO
Use	2-step	NM+R	−	Regression (RKHS)	LOEO, LOGO, LOEGO
		NM+E	−	Regression (EBlasso)	LOEO, LOGO, LOEGO
	Unified	Bayes_G	−	Covariance	LOEO, LOGO, LOEGO
		Bayes_Glog	+	Covariance	LOEO, LOGO, LOEGO
		Bayes_Elog	+	Regression (EBlasso)	LOEO, LOGO, LOEGO

[a]「2-step」はDVRモデルパラメータをNelder-Mead（NM）により推定し、その推定値を被説明変数としてReproducive kernel Hilbert space regression（RKHS）またはextended Bayesian lasso（EBlasso）によってゲノムワイドマーカーに回帰すること示す。「Unified」はマーカーを共分散（covariance）または回帰（regression）の形で事前分布に組み込みDVRモデルパラメータを推定する方法を指す。

[b]手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

[c]Bayes_GlogとBayes_ElogではDVRモデルパラメータを自然対数変換した

[d] 手法は 1 個抜き交差検証（leave-one-out cross-validation、LOO）を用いて比較した。LOEO、leave-one environment-out LOO ; LOGO、leave-one genotype-out LOO ; LOEGO、leave-one combination of environments and genotypes-out LOO

5.2.2.1 Nelder-Mead 法（NM）

目的関数は

$$\mathrm{argmin}_{G_i,\alpha_i,\beta_i}\sum_{j=1}\left|H_{ij}-h\left(G_i,\alpha_i,\beta_i,\mathbf{T}_j,\mathbf{P}_j\right)\right|$$

であり、ここで H_{ij} は遺伝子型 *i* の環境 *j* における DH、*h* は与えられた G_i、α_i、β_i、環境 *j* における日平均気温 $\mathbf{T_j}$、及び理論日長 $\mathbf{P_j}$ のもとでの DH を返す関数である。事前調査において二乗誤差に基づく最適化も試みたが、予測の正確さの点で絶対誤差による方法が上回ったためにこちらを採用した（結果非掲載）。最適化は個々の遺伝子型について行われた。最適化の際の初期値は *G* については 40、55、80 の 3 値、α と β については 0.01、5、10 の 3 値を使用した。これらの値は Nakagawa et al.（2005）が報告した日本晴とカサラスの BIL 集団において推定された各パラメータ値の分布を覆うように選択した。従って最適化は異なる初期値の組み合わせから 27 回（3×3×3）行い、27 通りの推定値の組み合わせを得た。この中で、最も目的関数を最小化させた組み合わせを推定値として選択した。複数の組み合わせが目的関数を最小化した場合はそれらの推定値の平均を用いた。NM は R 関数 optim を用いて行った。NM を行う R コードは東京大学大学院の渡部真哉氏により作成され本研究に提供された。

5.2.2.2 RKHS 及び EBlasso

これら回帰手法については第 2 章（EBlasso）及び 3 章（RKHS）で述べた。RKHS は R パッケージの rrBLUP（Endelman 2011）、EBlasso は VIGoR（第 2 章参照）を用いて行った。RKHS ではガウシアンカーネルを用いバンド幅は 1 とした。EBlasso のハイパーパラメータ、φ、ω、ψ、及び θ はそれぞれ 0.1、0.1、1、0.1 とし、*G*、α、及び β で共通とした。これらの回帰手法は NM で推定された *G*、α、及び β を被説明変数として学習した。

5.2.2.3 ベイズ手法

全てのベイズ手法（Bayes_I、Bayes_G、Bayes_Glog、及び Bayes_Elog）において、DVR モデルは被説明変数が 1 で固定された回帰モデルとみなした。つまり

$$1 = \sum_{d=1}^{H_{ij}} DVR_{ij,d} + e_{ij}$$

と想定した。$DVR_{ij,d}$ は遺伝子型 *i* の環境 *j* における出芽後 *d* 日目の DVR を表し、e_{ij} は残差を表す。残差は正規分布 $N\left(0,\ \sigma_e^2\right)$ に従うと想定し、その分散の事前分布は $\sigma_e^2 \sim 1/\sigma_e^2$ とした。尤度関数は

$$L\left(H_{ij} \mid G_i, \alpha_i, \beta_i, \sigma_e^2\right) = \left(2\pi\sigma_e^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_e^2}\left(1 - \sum_{d=1}^{H_{ij}} DVR_{ij,d}\right)^2\right]$$

と定義される。本研究で提案するベイズ手法は後述のように *G*、α、及び β の事前分布において異なる。以下では主にパラメータ *G* について述べるが、いずれの手法においても α 及び β について同様に事前分布及び推定手順を定義できる。

5.2.2.4 Bayes_I 及び Bayes_G

Bayes_I では、**G** は切断多変量正規分布（truncated multivariate normal distribution、TrMVN）に従うと想定する。

$$\mathbf{G} \sim \mathrm{TrMVN}\left(\mathbf{1}\mu_G,\ \ \mathbf{I}\sigma_G^2\right)$$

ここで **1** は要素が全て 1 のベクターを表す。μ_G はモード、σ_G^2 は分散を、**I** は単位行列を表す。Bayes_G では

$$\mathbf{G} \sim \mathrm{TrMVN}\left(\mathbf{1}\mu_G,\ \ \mathbf{K}\sigma_G^2\right)$$

と想定した。ここで **K** は遺伝子型間の共分散行列を定義するゲノム関係行列を表す。ここでは RKHS で用いたカーネル行列と同じものを用いた。TrMVN のモード及び分散の推定は容易ではないため、$\mu_G = 55$ 及び $\sigma_G^2 = 100$ とした。α 及び β の場合は $\mu_\alpha = 5$、$\mu_\beta = 5$、$\sigma_\alpha^2 = 9$、及び $\sigma_\beta^2 = 9$ とした。これらの値は日本晴とカサラスの BIL 集団における推定値の分布をもとに決定した（Nakagawa et al. 2005）。**G** の事後分布の推定はメトロポリス法により行った（Metropolis et al. 1953）。G_i の推定において提案値 G_i*は正規分布

$$G_i^* \sim N\left(G_i, \quad \delta_G^2\right)$$

から生成した。ここで δ_G^2 は提案分布の分散である。提案値は 0 以下であれば棄却し、そうでなければ採択確率に従って採択した。採択確率は Bayes_G の場合は

$$\min\left[1, \quad \frac{P\left(G_i^* \mid \mathbf{G}_{-i}, \mu_G, \sigma_G^2, \mathbf{K}\right)\prod_{j=1} L\left(H_{ij} \mid G_i^*, \alpha_i, \beta_i, \sigma_e^2\right)}{P\left(G_i \mid \mathbf{G}_{-i}, \mu_G, \sigma_G^2, \mathbf{K}\right)\prod_{j=1} L\left(H_{ij} \mid G_i, \alpha_i, \beta_i, \sigma_e^2\right)}\right]$$

となり、ここで

$$P\left(G_i \mid \mathbf{G}_{-i}, \mu_G, \sigma_G^2, \mathbf{K}\right) \propto \exp\left\{-\frac{K_{ii}'}{2\sigma_G^2}\left[G_i - \left(\mu_G - \frac{1}{K_{ii}'}\sum_{l \neq i}\left(G_l - \mu_G\right)K_{il}'\right)\right]^2\right\}$$

となる。K_{ij}'は **K** の逆行列における i 行 j 列目の要素を指す。$\mathbf{G}_{-\mathbf{i}}$ は i 以外の **G** の要素を表す。Bayes_I の場合は **K** を **I** に置き換えればよい。δ_G は 2.0 とした。また α 及び β の場合は $\delta_\alpha = \delta_\beta = 0.3$ とした。

残差分散 σ_e^2 の事後分布もメトロポリス法を用いて推定した。提案値 σ_e^2 *は正規分布、

$$\sigma_e^{2*} \sim N\left(\sigma_e^2, \quad \delta_e^2\right)$$

から生成した。提案値は 0 以下であれば棄却し、そうでなければ以下の採択確率に従って採択した。

$$\min\left[1, \quad \frac{P\left(\sigma_e^{2*}\right)\prod_{i=1}\prod_{j=1} L\left(H_{ij} \mid G_i, \alpha_i, \beta_i, \sigma_e^{2*}\right)}{P\left(\sigma_e^2\right)\prod_{i=1}\prod_{j=1} L\left(H_{ij} \mid G_i, \alpha_i, \beta_i, \sigma_e^2\right)}\right]$$

δ_e は 0.005 とした。

MCMC は 60,000 回繰り返しとし、サンプリング間隔は 50 とした。得られた 1,200 サンプルのうち最初の 200 回はバーンイン（burn-in）として削除し、事後分布は残りの 1,000 サンプルから推定した。DH の事後分布は得られた各サンプルにおいて DH を予測すること

により得た。

5.2.2.5 Bayes_Glog

この手法では **G** を自然対数変換した $\tilde{\mathbf{G}}$ に対して事前分布を設けた。

$$\tilde{\mathbf{G}} \sim \mathrm{MVN}\left(\mathbf{1}\mu_G, \quad \mathbf{K}\sigma_G^2\right)$$

ここで MVN は多変量正規分布を表す。このモデルでは事前分布の平均（μ_G）及び分散（σ_G^2）にそれぞれ以下のような無情報事前分布を想定し推定した。

$$\mu_G \sim \text{const.}$$
$$\sigma_G^2 \sim \text{Inv-}\chi^2\left(-2, \quad 0\right)$$

ここで const.は定数を、Inv-χ^{-2} は尺度付き逆カイ二乗分布を表す。$\tilde{\mathbf{G}}$ の事後分布はメトロポリス法により推定した。$\tilde{G}_i$ の推定において、提案値 $\tilde{G}_i$*は正規分布

$$\tilde{G}_i^* \sim N\left(\tilde{G}_i, \quad \delta_G^2\right)$$

から生成した。提案値は以下の採択確率に従って採択した。

$$\min\left[1, \quad \frac{P\left(\tilde{G}_i^* \mid \tilde{\mathbf{G}}_{-i}, \mu_G, \sigma_G^2, \mathbf{K}\right)\prod_{j=1} L\left(H_{ij} \mid G_i = \exp\left(\tilde{G}_i^*\right), \alpha_i = exp\left(\tilde{\alpha}_i\right), \beta_i = \exp\left(\tilde{\beta}_i\right), \sigma_e^2\right)}{P\left(\tilde{G}_i \mid \tilde{\mathbf{G}}_{-i}, \mu_G, \sigma_G^2, \mathbf{K}\right)\prod_{j=1} L\left(H_{ij} \mid G_i = \exp\left(\tilde{G}_i\right), \alpha_i = \exp\left(\tilde{\alpha}_i\right), \beta_i = \exp\left(\tilde{\beta}_i\right), \sigma_e^2\right)}\right]$$

ここで

$$P\left(\tilde{G}_i \mid \tilde{\mathbf{G}}_{-i}, \mu_G, \sigma_G^2, \mathbf{K}\right) \propto \exp\left\{-\frac{K_{ii}'}{2\sigma_G^2}\left[\tilde{G}_i - \left(\mu_G - \frac{1}{K_{ii}'}\sum_{l\neq i}\left(\tilde{G}_l - \mu_G\right)K_{il}'\right)\right]^2\right\}$$

となる。δ_G は 0.07 とした。また $\tilde{\alpha}$ 及び $\tilde{\beta}$ の場合、δ_α は 0.03 及び δ_β は 0.2 とした。

μ_G 及び σ_G^2 の推定はギブズサンプリングを用いて行った。それぞれの事後分布は

$$N\left(\frac{\sum_i \tilde{G}_i \sum_j K_{ij}'}{\sum_i \sum_j K_{ij}'}, \quad \frac{\sigma_G^2}{\sum_i \sum_j K_{ij}'}\right)$$

及び

$$\text{Inv-}\chi^2\left(N-2, \quad \frac{\left(\tilde{\mathbf{G}} - \mathbf{1}\mu_G\right)^{\mathrm{T}}\mathbf{K}^{-1}\left(\tilde{\mathbf{G}} - \mathbf{1}\mu_G\right)}{N-2}\right)$$

となる。ここで N は遺伝子型数（系統数）を表す。残差分散 σ_e^2 の事後分布は Bayes_I 及び Bayes_G と同様に行った。MCMC の条件（試行数やサンプリング間隔）もそれらと同様に行った。

5.2.2.6 Bayes_Elog

この手法では自然対数変換した $\tilde{G}_i$ に対して事前分布

$$\tilde{G}_i \sim \mathrm{N}\left(\mu_G + \sum_{p=1}^{P} g_p x_{ip},\quad {}^{1}\!\!\diagup\!\!_{\tau_{0,G}^2}\right)$$

を想定した。ここで μ_G は切片、g_p はマーカーp の $\tilde{G}$ に対する効果、x_{ip} は−1（AA）、0（AK）、及び 1（KK）とコードされたマーカーp の遺伝子型を表す。なお A はカサラス、K はコシヒカリ由来のアリルを表している。マーカー効果は EBlasso 同様にモデリングした。よって第 2 章で述べたように g_p の事前分布は

$$\begin{aligned}
g_p &\sim \mathrm{N}\left(0, \frac{1}{\tau_{0,G}^2 \tau_{p,G}^2}\right) \\
\tau_{p,G}^2 &\sim \text{Inv-G}\left(1, \frac{\delta_G^2 \eta_{p,G}^2}{2}\right) \\
\delta_G^2 &\sim \mathrm{G}(\varphi, \varpi) \\
\eta_{p,G}^2 &\sim \mathrm{G}(\psi, \theta)
\end{aligned}$$

となり、φ、 ω、 ψ、及び θ がハイパーパラメータとなる。

このモデルについては変分ベイズ法を用いた。第 2 章で述べたように変分ベイズ法では事後分布をパラメータ毎に因数分解して推定を行う。ここで $\boldsymbol{\Theta}$ をマーカー効果など EBlasso に関するパラメータを全て含んだベクターとする。DVR モデルパラメータと $\boldsymbol{\Theta}$ の同時事後分布、$q\left(\tilde{\mathbf{G}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \sigma_e^2, \boldsymbol{\Theta} \mid \boldsymbol{y}\right)$ を以下のように分解する。

$$q\left(\tilde{\mathbf{G}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \sigma_e^2, \boldsymbol{\Theta} \mid \boldsymbol{y}\right) \propto q\left(\tilde{\mathbf{G}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \sigma_e^2 \mid \boldsymbol{y}\right) \prod_{m=1}^{M} q(\theta_m \mid \boldsymbol{y})$$

ここで M は EBlasso に関わる全パラメータ数を表す。変分ベイズ法ではデータの周辺尤度をその下限を最大化することにより近似する。下限は任意のパラメータ $\bullet$ について以下のように置くことで最大化できる。

$$q(\bullet \mid \boldsymbol{y}) \propto \exp\left(E_{q(\circ \mid \boldsymbol{y})}\left[\log p\left(\boldsymbol{y}, \tilde{\mathbf{G}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \sigma_e^2, \boldsymbol{\Theta}\right)\right]\right)$$

ここで$q(\circ \mid \boldsymbol{y})$は$\bullet$以外のパラメータに関する近似事後分布を表す。Bayes_Elog では

$$\begin{aligned} q(\theta_m \mid \boldsymbol{y}) &\propto \exp\left(E_{q(\tilde{\mathbf{G}},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\sigma_e^2 \mid \boldsymbol{y}) \text{ and } q(\theta_k \mid \boldsymbol{y}),k\neq m}\left[\log p\left(\boldsymbol{y},\tilde{\mathbf{G}},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\sigma_e^2,\boldsymbol{\Theta}\right)\right]\right) \text{ for } 1 \le m \le M \\ q\left(\tilde{\mathbf{G}},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\sigma_e^2 \mid \boldsymbol{y}\right) &\propto \exp\left(E_{q(\theta_m \mid \boldsymbol{y}),1\le m\le M}\left[\log p\left(\boldsymbol{y},\tilde{\mathbf{G}},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\sigma_e^2,\boldsymbol{\Theta}\right)\right]\right) \end{aligned} \tag{5.1}$$

となる。最初の項は EBlasso のパラメータに関して既知の形をした事後分布を生成する。その事後分布の期待値及び分散は以下のように表すことができる。

$$E\left[\mu_G\right] = \Lambda E\left[\tau_{0,G}^2\right]\sum_{i=1}^{N}\left(E\left[\tilde{G}_i\right] - \sum_{p=1}^{P} E\left[g_p\right]x_{ip}\right)$$

ここで$V\left[\mu_G\right] = \Lambda$及び$\Lambda^{-1} = E\left[\tau_{0,G}^2\right]N$、

$$E\left[g_p\right] = \mathrm{H}_{p,G} E\left[\tau_{0,G}^2\right]\sum_{i=1}^{N} x_{ip}\left(E\left[\tilde{G}_i\right] - \mu_G - \sum_{k\neq p}^{P} E\left[g_k\right]x_{ik}\right)$$

ここで$V\left[g_p\right] = \mathrm{H}_{p,G}$及び$\mathrm{H}_{p,G}^{-1} = E\left[\tau_{0,G}^2\right]\sum_{i=1}^{N} x_{ip}^2 + E\left[\tau_{p,G}^2\right]E\left[\tau_{0,G}^2\right]$、

$$E\left[\tau_{p,G}^2\right] = \sqrt{\frac{E\left[\delta_G^2\right]E\left[\eta_{p,G}^2\right]}{E\left[g_{p,G}^2\right]E\left[\tau_{0,G}^2\right]}}$$

及び

$$E\left[\frac{1}{\tau_{p,G}^2}\right] = \sqrt{\frac{E\left[g_p^2\right]E\left[\tau_{0,G}^2\right]}{E\left[\delta_G^2\right]E\left[\eta_{p,G}^2\right]}} + \frac{1}{E\left[\delta_G^2\right]E\left[\eta_{p,G}^2\right]}$$

また

$$E\left[\tau_{0,G}^2\right] = \frac{a_1}{b_{G1}}$$

ここで$a_1 = \frac{1}{2}\left(N+P\right)$及び

$$b_{G1} = \frac{1}{2}\left\{\begin{array}{l}\sum_{i=1}^{N}\left(E\left[\tilde{G}_i\right] - \mu_G - \sum_{p=1}^{P} E\left[g_p\right]x_{ip}\right)^2 + \\ \qquad NV\left[\mu_G\right] + \sum_{p=1}^{P} V\left[g_p\right]\sum_{i=1}^{N} x_{ip}^2 + \sum_{p=1}^{P} E\left[\tau_{p,G}^2\right]E\left[g_p^2\right] + \sum_{i=1}^{N} V\left[\tilde{G}_i\right]\end{array}\right\}$$

また

$$E\left[\delta_G^2\right] = \frac{a_2}{b_{G2}}$$

ここで$a_2 = P + \phi$、及び$b_{G2} = \frac{1}{2}\sum_{p=1}^{P} E\left[\eta_{p,G}^2\right] E\left[\frac{1}{\tau_{p,G}^2}\right] + \varpi$ である。さらに

$$E\left[\eta_{p,G}^2\right] = \frac{a_3}{b_{3G,p}}$$

ここで$a_3 = 1 + \psi$、及び$b_{3G,p} = \frac{1}{2} E\left[\delta_G^2\right] E\left[\frac{1}{\tau_{p,G}^2}\right] + \theta$となる。

式 5.1 の 2 つ目の項は$\tilde{\mathbf{G}}$，$\tilde{\boldsymbol{\alpha}}$，$\tilde{\boldsymbol{\beta}}$に関する未知の事後分布を生成する。つまり

$$q\left(\tilde{\mathbf{G}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \sigma_e^2 \mid \boldsymbol{y}\right)$$

$$\propto \prod_{i=1} \left[\begin{array}{l} q\left(\tilde{G}_i \mid \mu_G, g_p, \tau_{0,G}^2\right) q\left(\tilde{\alpha}_i \mid \mu_\alpha, a_p, \tau_{0,\alpha}^2\right) q\left(\tilde{\beta}_i \mid \mu_\beta, b_p, \tau_{0,\beta}^2\right) q\left(\sigma_e^2\right) \\ \qquad \times \prod_{j=1} L\left(H_{ij} \mid G_i = \exp\left(\tilde{G}_i\right)_i, \alpha_i = \exp\left(\tilde{\alpha}_i\right)_i, \beta_i = \exp\left(\tilde{\beta}_i\right), \sigma_e^2\right) \end{array} \right]$$

ここで

$$q\left(\tilde{G}_i \mid \mu_G, g_p, \tau_{0,G}^2\right) \propto \exp\left\{ -\frac{E\left[\tau_{0,G}^2\right]}{2} \left[\tilde{G}_i - \left(E\left[\mu_G\right] + \sum_{p=1}^{P} E\left[g_p\right] x_{ip} \right) \right]^2 \right\}$$

であり、また$q\left(\sigma_e^2\right) \propto \frac{1}{\sigma_e^2}$となる。この事後分布の推定にはメトロポリス法を用いた。$\tilde{G}_i$の推定において、提案値$\tilde{G}_i$*は正規分布

$$\tilde{G}_i^* \sim N\left(\tilde{G}_i, \quad \delta_G^2\right)$$

から生成した。提案値は以下の採択確率に従って採択した。

$$\min\left[1, \quad \frac{q\left(\tilde{G}_i^* \mid \mu_G, g_p, \tau_{0,G}^2\right) \prod_{j=1} L\left(H_{ij} \mid G_i = \exp\left(\tilde{G}_i^*\right), \alpha_i = \exp\left(\tilde{\alpha}_i\right), \beta_i = \exp\left(\tilde{\beta}_i\right), \sigma_e^2\right)}{q\left(\tilde{G}_i \mid \mu_G, g_p, \tau_{0,G}^2\right) \prod_{j=1} L\left(H_{ij} \mid G_i = \exp\left(\tilde{G}_i\right), \alpha_i = \exp\left(\tilde{\alpha}_i\right), \beta_i = \exp\left(\tilde{\beta}_i\right), \sigma_e^2\right)}\right]$$

生成した MCMC サンプルから$\tilde{G}_i$の期待値及び分散を計算し、それを EBlasso パラメータの推定に用いた。δ_G（及びδ_αとδ_β）は Bayes_Glog と同じ値を用いた。また残差分散σ_e^2の事後分布は Bayes_I 及び Bayes_G と同様に行った。

EBlasso 及び DVR モデルパラメータの事後分布の期待値及び分散の推定は以下の手順で行った。

①$\tilde{G}_i$（$\tilde{\alpha}_i$、$\tilde{\beta}_i$）の事前分布の平均（$\mu_G + \sum_{p=1}^{P} g_p x_{ip}$）及び分散（$1/\tau_{0,G}^2$）の初期値をそれぞれ

ξ_{G1}及びξ_{G2}^2（$\xi_{\alpha 1}$と$\xi_{\alpha 2}^2$及び$\xi_{\beta 1}$と$\xi_{\beta 2}^2$)とおく。ξ_{G1}及びξ_{G2}^2は 4 及び 0.04 とした。これらの値は Nakagawa et al.（2005）により報告された日本晴とカサラスの BIL 集団での推定値から決定した。同様に$\xi_{\alpha 1}=\xi_{\beta 1}=1.61$及び$\xi_{\alpha 2}^2=\xi_{\beta 2}^2=1.0$とした。$\tilde{G}_i$（$\tilde{\alpha}_i$、$\tilde{\beta}_i$）の初期値はこの事前分布からランダムに生成する。

②$\tilde{\mathbf{G}}$，$\tilde{\boldsymbol{\alpha}}$，$\tilde{\boldsymbol{\beta}}$及び$\sigma_e^2$の MCMC サンプリングを 600 回行う。サンプリング間隔は 10 とし、得られた 60 サンプルから各パラメータの期待値及び分散を計算する。

③$\tilde{\mathbf{G}}$をゲノムワイドマーカーに回帰し、EBlasso に関係するパラメータを繰り返し計算する。この繰り返し計算は以下の基準を満たした時点で終了する。

$$\frac{\left\|\Theta_{G,current}-\Theta_{G,previous}\right\|^2}{\left\|\Theta_{G,previous}\right\|^2}<10^{-9}$$

ここで$\Theta_{G,current}$及び$\Theta_{G,previous}$はそれぞれ現在及び直前の試行におけるパラメータの期待値を含むベクターを表す。また‖ ‖はユークリディアンノルムを指す。計算は概ね 100 から 200 回の試行で終了した。

④$\tilde{\boldsymbol{\alpha}}$をゲノムワイドマーカーに回帰し、EBlasso に関係するパラメータを繰り返し計算する。収束判定は③と同様に行う。

⑤$\tilde{\boldsymbol{\beta}}$をゲノムワイドマーカーに回帰し、EBlasso に関係するパラメータを繰り返し計算する。収束判定は③と同様に行う。

⑥ステップ②～⑤を 100 サイクル繰り返す。最後のサイクルでは MCMC サンプルの数を 300（MCMC 試行数を 3,000）に増やし、その MCMC サンプルから *G*、α、及びβの事後分布を推定する。最後のサイクルにおけるマーカー効果の期待値を推定値として用いる。

ステップ③から⑤において収束を加速させるためには、また EBlasso のハイパーパラメータ（φ、 ω、 ψ、及びθ）をスケールの異なるパラメータ *G*、α、及びβ間で共通とするためには、各ステップの際に$\tilde{G}_i$（$\tilde{\alpha}_i$、$\tilde{\beta}_i$）を標準化することが望ましい。しかし標準化は$\tilde{G}_i$の事前分布を *i* 以外の遺伝子型のパラメータ$\tilde{G}_j$によって条件づけることを意味するため、事後分布を変えることになる。そのため、ここでは$\tilde{G}_i$（$\tilde{\alpha}_i$、$\tilde{\beta}_i$）のスケールを定数、ξ_{G1}及びξ_{G2}（$\xi_{\alpha 1}$と$\xi_{\alpha 2}$及び$\xi_{\beta 1}$と$\xi_{\beta 2}$)によって以下のように変換した。

$$\frac{\tilde{G}_i - \xi_{G1}}{\xi_{G2}}$$

この手法によりパラメータ間でスケールが完全に一致するわけではないが、共通のハイパーパラメータを使用できる程度には近くなることが期待できる。φ、ω、ψ、及び θ は NM+E 同様に 0.1、0.1、1、及び 0.1 とした。

5.2.3　表現型記録とゲノムワイドマーカー遺伝子型

コシヒカリとカサラスに由来する BIL（コシヒカリ/カサラス//コシヒカリ）174 系統と親系統の合計 176 系統を用いた。この BIL 系統は農業生物資源研究所イネゲノムリソースセンターから提供されている（Ma et al. 2002）。この BIL 系統における 162 座位の制限酵素断片長多型の遺伝子型と連鎖地図は http://www.rgrc.dna.affrc.go.jp/jp/ineKKBIL182.html において公開されている。マーカーは全て 2 つの対立遺伝子から成っていた。これらの系統は 2007 年から 2009 年において 6 試験地､合計 9 環境において栽培され DH が記録された（表 5.2）。これら栽培試験は九州大学大学院の望月俊宏教授、NARO の中川博視博士、農業環境技術研究所の長谷川利拡博士によって行われ、表現型及び気象記録が本研究のために提供された。栽培試験は各環境において 2 反復ずつ行われたが、記録の欠損がしばし観察され、2 反復の平均値と 1 反復のみの記録を同様に扱うと含まれるノイズの大きさが系統間で異なることが考えられたため、各環境において欠損値がより少ない反復を表現型として使用した。得られた DH の分布は図 5.1 に示した。

表5.2　栽培試験の概要

Location	Year[a]	Ave. mean daily temperature (°C)/ photoperiod (h)			Emergence date	# evaluated genotypes[b]
		June	July	Aug.		
Tsukuba	2007	21.8 / 14.5	22.9 / 14.3	27.4 / 13.5	17, May	176
	2008E	20.6 / 14.5	25.6 / 14.3	25.4 / 13.5	2–11, Apr.	176
	2008L	20.6 / 14.5	25.6 / 14.3	25.4 / 13.5	25, June	176
	2009	21.4 / 14.5	25.0 / 14.3	24.9 / 13.5	30, Apr.–2, May	176
Ishikawa	2008	20.8 / 14.6	26.7 / 14.3	26.4 / 13.5	24-27, Apr.	176
Fukuoka	2008	22.2 / 14.3	29.1 / 14.1	27.5 / 13.3	5, June	171
Ishigaki	2008	28.6 / 13.6	29.5 / 13.5	29.5 / 13.0	27, June	166
Thai Nguyen	2008	28.1 / 13.4	28.4 / 13.3	28.2 / 12.9	2, July	148
Ha Noi	2008	28.6 / 13.4	29.3 / 13.3	28.1 / 12.8	18, June	174

a栽培年。EとLはそれぞれ早植及び晩殖を表す。

b 表現型記録を持つ遺伝子型数

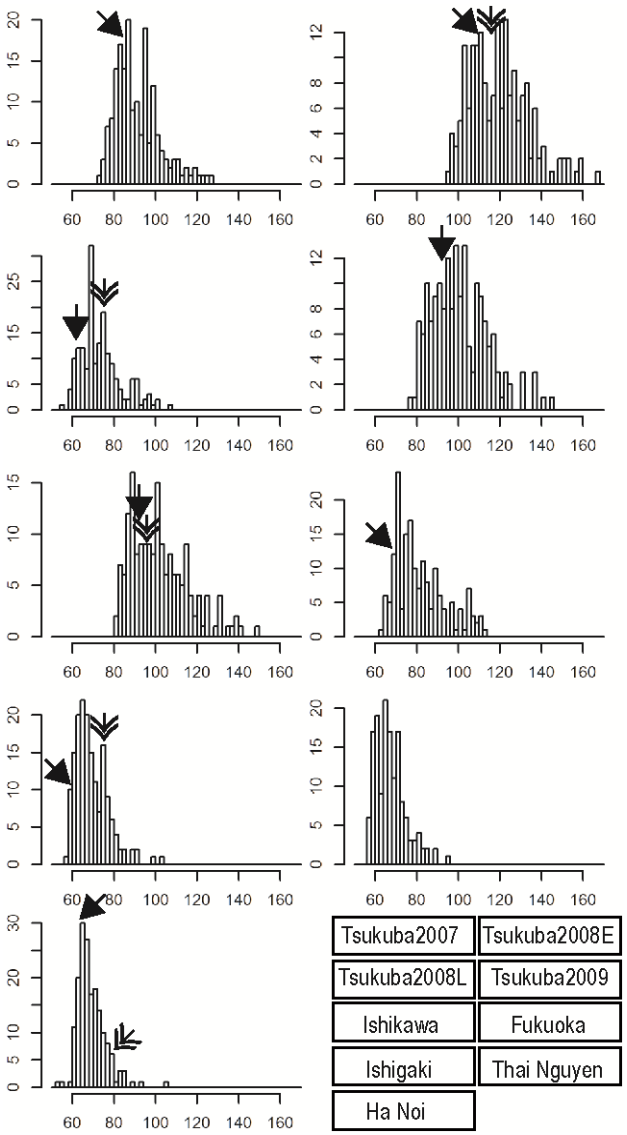

図5.1　各環境における到穂日数（days to heading、DH）の分布。黒及び二重の矢印はそれぞれ親品種であるコシヒカリ及びカサラスのDHを示す。Tsukuba2007及びTsukuba2009でのカサラスのDHはコシヒカリと同じであった。両品種の記録はThai Nguyenで欠損しており、カサラスの記録はFukuokaで欠損している。

5.2.4　全データ解析と交差検証

手法の振る舞いを比較するためにまず全データを用いた解析を行った。次に、手法間で予測能力を比較するために、3 種類の 1 個抜き交差検証（leave-one-out cross validation、LOO）を行った。1 つ目は環境抜きの LOO（leave-one environment-out、LOEO）である。この LOO では 9 つの環境のうち 1 つをデータから除き、その除かれた環境における全遺伝子型の DH を、残った 8 環境における全遺伝子型の DH から予測する。これを全環境について行う。2 つ目は遺伝子型抜きの LOO（leave-one genotype-out、LOGO）である。この LOO では 176 遺伝子型のうち 1 つをデータから除き、その除かれた遺伝子型の全環境における DH を残ったデータから予測する。これを全遺伝子型について行う。最後は遺伝子型と環境の組み合わせ抜きの LOO（leave-one combination of environments and genotypes-out、LOEGO）である。この LOO では、9 環境のうち 1 環境と 176 遺伝子型のうち 1 遺伝子型をデータから除き、残った 8 環境 175 遺伝子型のデータから、除いた環境における除いた遺伝子型の DH を予測する。これを環境と遺伝子型全ての組み合わせについて行う。LOEO、LOGO、及び LOEGO はそれぞれ、既知遺伝子型の未試験環境下での、新しい遺伝子型の試験済み環境下での、及び新しい遺伝子型の未試験環境下での予測能力の検証に対応する。

各手法の予測能力は観察された DH と予測された値間の RMSE により計測した。また予測値を観察値に回帰した際の回帰係数も計算した。1 より小さい係数は予測値の縮約（shrinkage）を示唆するために、これを縮約の指標として報告した。

上述したように NM 及び Bayes_I はそれ自体では新しい遺伝子型への予測ができないために LOEO においてのみ検証した。NM+R 及び NM+E を LOEO で検証した場合は、NM で推定した *G*、α、及び β を回帰し、得られた当てはめ値を用いて予測を行った。

5.3　結果

5.3.1　全データ解析

観察された DH とモデルの当てはめにより得られた DH の推定値間の RMSE は NM で最も小さく、NM+E で最も大きかった（表 5.3）。ベイズ手法で推定された残差分散 σ_e^2 は非常に小さく、当てはまりの良さを示唆していた。

各手法で推定された *G*、α、β、及び Tsukuba2008E での DH（当てはめ値）の分布を図 5.2 で比較した。ベイズ手法では事後分布の期待値を推定値として用いた。*G* の推定値の分布は手法間で似通っており、NM による推定値が最も遺伝子型間の差異が大きかった（SD

= 8.5)。α と β の分布は手法間でやや異なっており、特に Bayes_I 及び Bayes_G と、その他の手法間で顕著な差が観察された。Bayes_I と Bayes_G による α の分布は二峰性で、比較的大きな遺伝子型間差異が見られた（それぞれ SD = 0.53 及び 0.67)。β の推定値が 0.5 より小さかった遺伝子型の割合は NM、NM+R、NM+E、Bayes_Glog、及び Bayes_Elog でそれぞれ 11.4 %、6.3 %、11.9 %、4.5 %、及び 17.6 %であったが、Bayes_I 及び Bayes_G では全く観察されなかった。一方で推定値の手法間のピアソン及びスピアマン相関係数は α と β であっても概ね高かった（表 5.4～5.6)。このことはいずれのパラメータでも遺伝子型間の推定値の大きさの順番は手法間で概ね一定であることを示唆している。Tsukuba2008E での DH の遺伝的差異は NM+R と NM+E で明らかに小さかった（図 5.2)。同様の傾向は他の環境でも観察された（結果非掲載)。

表 5.3　全データ解析における推定値と観察値間の平均二乗誤差（root mean squared error、RMSE）、推定値の観察値に対する回帰係数（slope）、及びベイズ手法における残差分散の事後平均

	NM	NM+R	NM+E	Bayes_I	Bayes_G	Bayes_Glog	Bayes_Elog
RMSE (d)	3.6	4.8	6.4	3.9	4.5	4.0	4.0
Slope	0.96	0.90	0.87	0.96	0.95	0.95	0.95
σ_e^2 $(\times 10^{-3})$				2.40	2.73	2.34	2.82

手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及び log（自然対数変換）をそれぞれ表す。

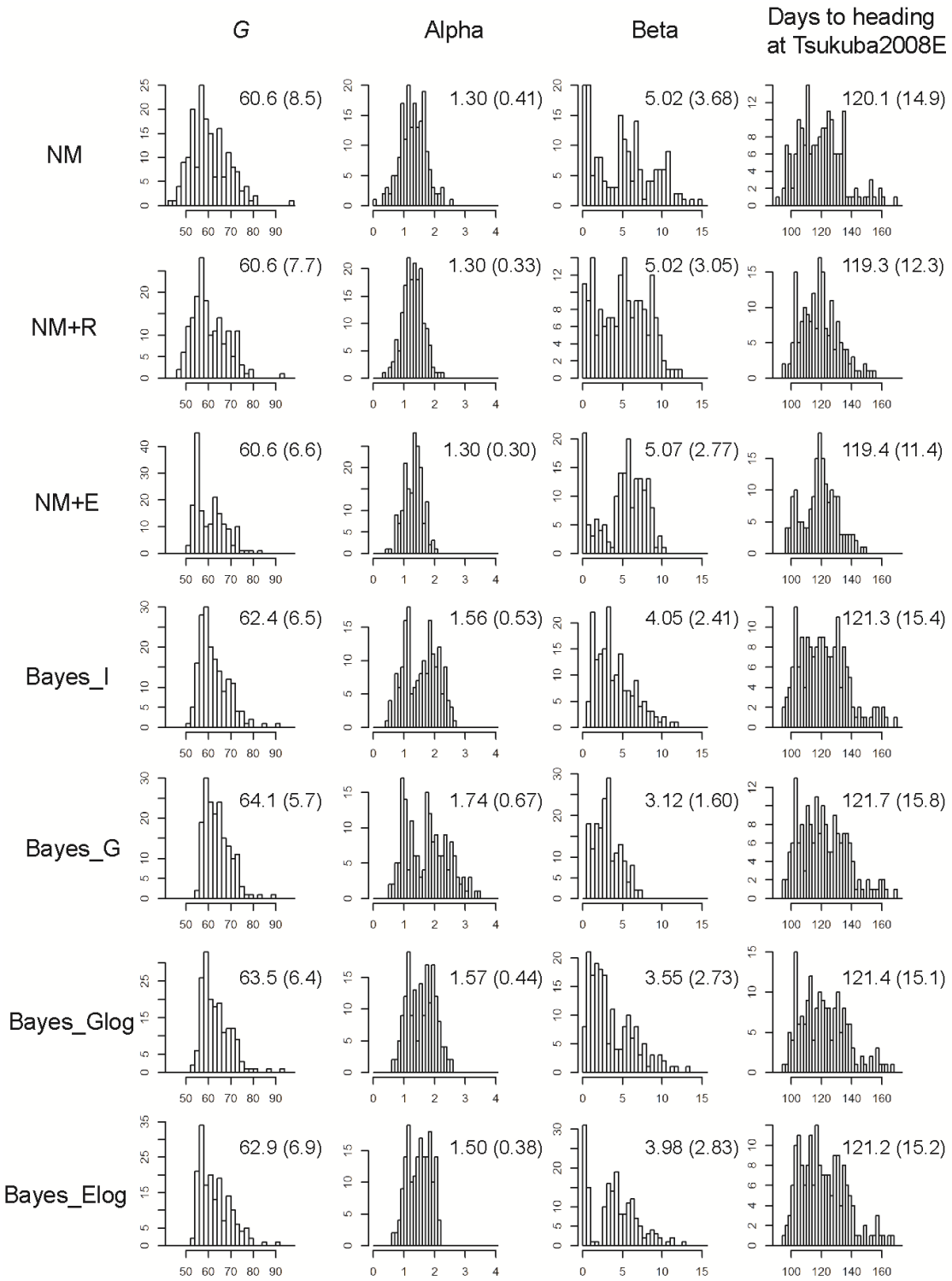

図5.2　全データから推定されたDVRモデルパラメータとTsukuba2008Eにおける到穂日数の分布。*G*、α、及びβはそれぞれ基本生長性、温度感応性、及び日長感応性を表す。平均値（標準偏差）を図中に示した。手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

表5.4　全データから推定された*G*の手法間でのスピアマン（上側）及びピアソン（下側）相関係数[a]

	NM	NM+R	NM+E	Bayes_I	Bayes_G	Bayes_Glog	Bayes_Elog
NM		0.99	0.82	0.91	0.76	0.91	0.88
NM+R	0.99		0.87	0.91	0.79	0.93	0.90
NM+E	0.86	0.9		0.82	0.78	0.87	0.90
Bayes_I	0.93	0.93	0.82		0.91	0.96	0.96
Bayes_G	0.83	0.85	0.78	0.95		0.91	0.92
Bayes_Glog	0.93	0.94	0.85	0.98	0.94		0.96
Bayes_Elog	0.92	0.93	0.90	0.97	0.94	0.97	

[a]*G*は基本生長性を表す。ベイズ手法では事後期待値を用いた。手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

表5.5　全データから推定されたαの手法間でのスピアマン（上側）及びピアソン（下側）相関係数[a]

	NM	NM+R	NM+E	Bayes_I	Bayes_G	Bayes_Glog	Bayes_Elog
NM		0.96	0.83	0.80	0.61	0.81	0.76
NM+R	0.96		0.92	0.82	0.67	0.87	0.82
NM+E	0.83	0.92		0.81	0.73	0.89	0.86
Bayes_I	0.80	0.81	0.8		0.90	0.93	0.95
Bayes_G	0.57	0.62	0.69	0.87		0.84	0.90
Bayes_Glog	0.80	0.86	0.88	0.93	0.80		0.96
Bayes_Elog	0.75	0.80	0.85	0.94	0.87	0.96	

[a]αは温度感応性を表す。ベイズ手法では事後期待値を用いた。手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

表5.6　全データから推定されたβの手法間でのスピアマン（上側）及びピアソン（下側）相関係数[a]

	NM	NM+R	NM+E	Bayes_I	Bayes_G	Bayes_Glog	Bayes_Elog
NM		0.97	0.9	0.95	0.87	0.94	0.92
NM+R	0.97		0.95	0.95	0.91	0.98	0.95
NM+E	0.87	0.94		0.91	0.89	0.95	0.96
Bayes_I	0.94	0.92	0.84		0.92	0.95	0.93
Bayes_G	0.83	0.89	0.87	0.89		0.92	0.88
Bayes_Glog	0.93	0.93	0.85	0.96	0.88		0.96
Bayes_Elog	0.92	0.94	0.92	0.93	0.87	0.95	

[a]βは日長感応性を表す。ベイズ手法では事後期待値を用いた。手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

Bayes_I、Bayes_G、及び Bayes_Glog のマルコフ連鎖は、図 5.3 に示した対数尤度の推移に示されるように速やかに収束していた。*G*、α、及び β の推定で用いたメトロポリス法の採択確率は 0.38 から 0.72（Bayes_I）、0.43 から 0.74（Bayes_G）、及び 0.24 から 0.81（Bayes_Glog）と良好な混合を示唆した。Bayes_Elog で用いた変分ベイズと MCMC を組み合わせた推定手法の再現性を確認するために、$\tilde{G}$、$\tilde{\alpha}$、及び$\tilde{\beta}$ の異なる初期値から 5 回解析を繰り返した。$\tilde{G}$、$\tilde{\alpha}$、及び$\tilde{\beta}$ の推定値は反復間でほぼ一致しており、平均のピアソン相関係数は 0.99 以上であった。またそれらパラメータに対するマーカー効果の推定値もほぼ一致しており、ピアソン相関係数は 0.98 以上であった。

ベイズ手法の中では、Bayes_Elog だけは$\tilde{G}$、$\tilde{\alpha}$、及び$\tilde{\beta}$ の分散を 2 つの成分、つまりマーカーによって説明される分散と、されない分散（残差分散）に分割する。マーカーによって説明された分散は、経験的に全マーカー効果の和、つまり $\tilde{G}$ の場合は$\sum_{p=1}^{P} g_p x_{ip}$、の分散として求めた。この分散成分のパラメータの分散に対する割合は 0.74（$\tilde{G}$）、0.83（$\tilde{\alpha}$）、及び 0.99（$\tilde{\beta}$）であった。一方で NM+E においてもパラメータの分散に対するマーカーが説明した分散の割合を同様に求めると、0.60（*G*）、0.51（α）、及び 0.56（β）であり、NM+E の方が Bayes_Elog より推定値が縮約傾向にあることが示唆された。また NM+R では 0.82（*G*）、0.64（α）、及び 0.69（β）であった。

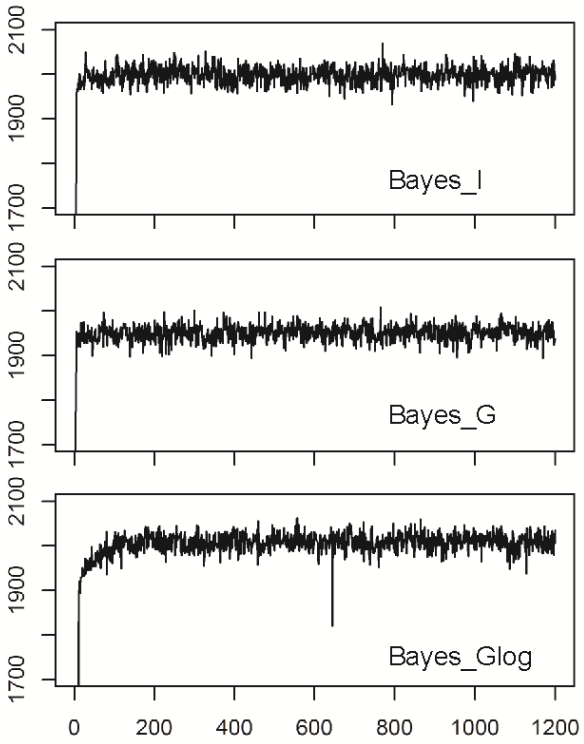

図5.3　Markov chain Monte Carlo（MCMC）サンプルにおける対数尤度の推移。MCMCサンプル数は1,200で最初の200サンプルはバーンインとして推定から除いた。

Bayes_Elog で推定されたマーカー効果を図 5.4 に示した。また NM+E による推定値も合わせて図示した。いずれの手法でも出穂期の主働遺伝子の近傍に強い関連シグナルを検出した。*Hd1*（Yano et al. 1997; Yano et al. 2000）と *Hd2*（Yano et al. 1997）は全てのパラメータに関与することが、*Hd5*（Yano et al. 1997）と *Hd6*（Takahashi et al. 2001）はそれぞれ *G*（基本生長性）とβ（日長感応性）、α（温度感応性）とβに、*Hd3*（Yano et al. 1997; Monna et al. 2002）はαに関与することがいずれの手法においても示唆された。さらに NM+E においては 1 番染色体にβ（関連するマーカーは C14085）、4 番染色体に *G*（C12132S と C933）、12 番染色体にα及びβ（R1684）に関連シグナルが観察された。

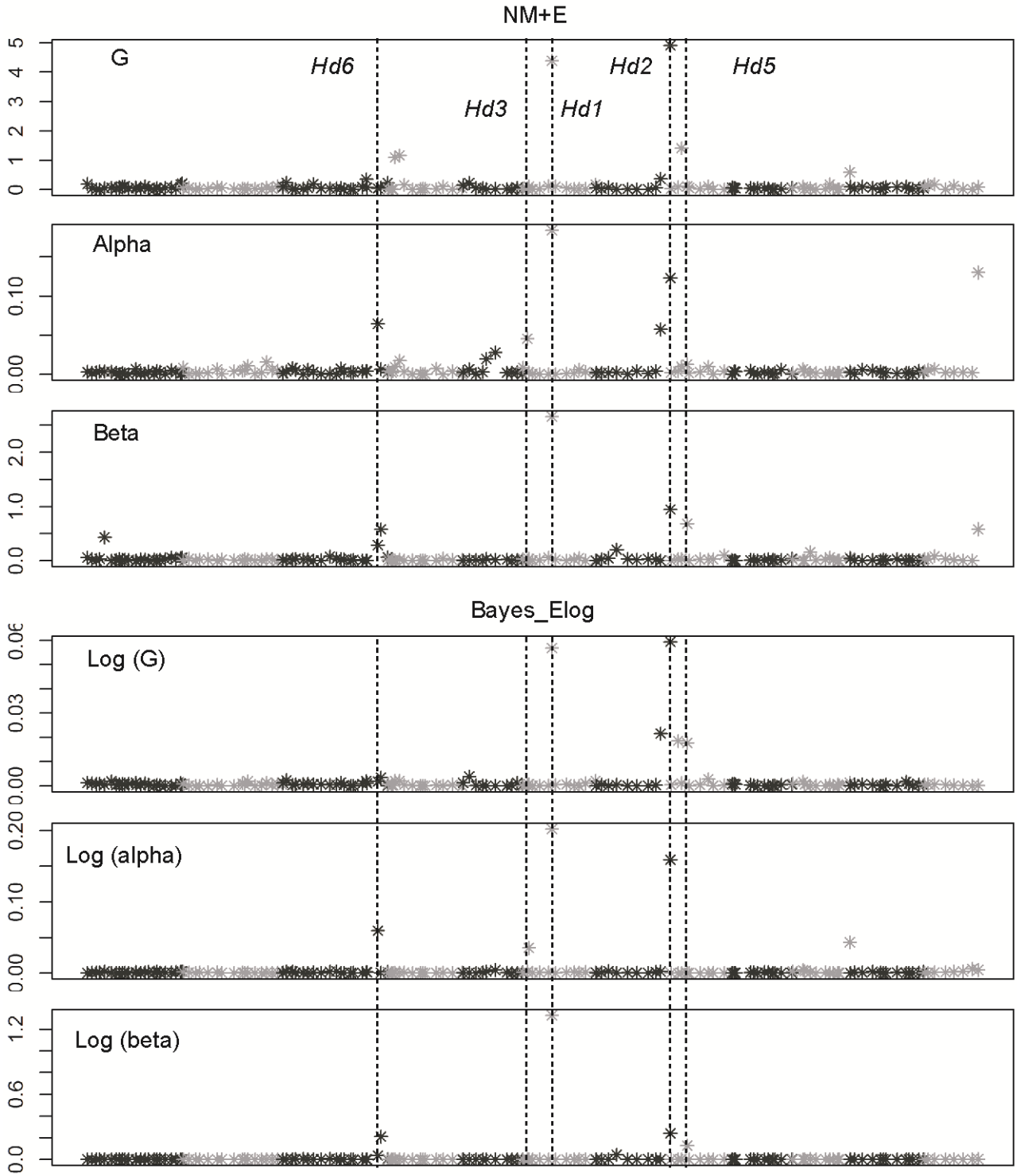

図5.4　NM+E及びBayes_Elogによって推定されたDVRモデルパラメータに対するマーカー効果。染色体により点の濃さを変えている（左から右へ1番から12番染色体を表す）。破線は既知の主働遺伝子を示す。Bayes_ElogはDVRモデルパラメータを対数変換しているため、y軸の尺度がNM+Eと異なっていることに留意する。*G*、α、及びβはそれぞれ基本生長性、温度感応性、及び日長感応性を表す。

5.3.2　交差検証

LOEO の結果を図 5.5 に示した。最小の RMSE は Bayes_Elog で観察された。全てのベイズ手法は NM に基づく手法（NM、NM+R、及び NM+E）より小さな RMSE を示した。回帰係数（slope）が最も 1 に近いことから表されるように、Bayes_I が最も縮約程度の小さな予測を行った。NM は Tsukuba2008E においてときに大きな予測誤差が観察された。図 5.1 に示したように Tsukuba2008E での DH は他環境より長く、そのためこの環境に対する予測は外挿にあたり、他環境より予測が困難であることが推察できる。LOGO 及び LOEGO では、Bayes_Elog が最小の RMSE 及び 1 に最も近い回帰係数を示した（図 5.6 及び 5.7）。一方で Bayes_G 及び Bayes_Glog は NM に基づく手法より大きな RMSE を示した。興味深いことに、NM+E は LOEO では最も大きな RMSE を示したが、LOGO と LOEGO では 2 番目に小さな RMSE を返した。全ての手法において LOGO における RMSE は LOEO より大きく、試験済みの環境下で新しい遺伝子型に対する予測を行うことは、未試験の環境下で既知の遺伝子型に対する予測を行うことより困難であることが示唆された。

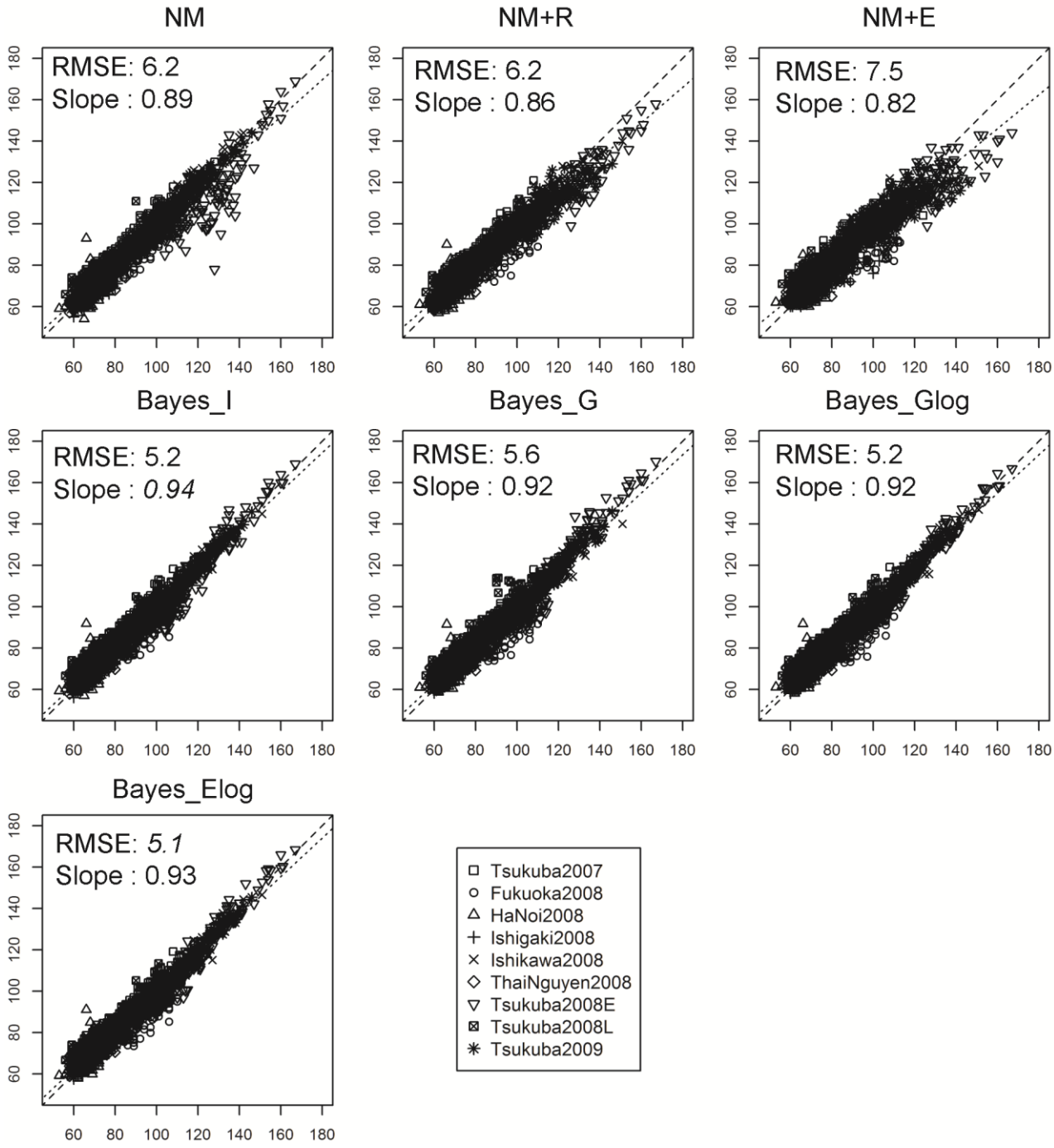

図5.5　1環境抜き交差検証（leave-one environment-out cross-validation、LOEO）の結果。各手法について観察値（x軸）と予測値（y軸）をプロットした。破線は1対1の線を、ドット線は回帰直線を示す。平均二乗誤差（RMSE）と予測値の観察値に対する回帰係数（slope）を図中に示した。手法で用いられている略号は、NM（Nelder-Mead）、I（単位行列）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

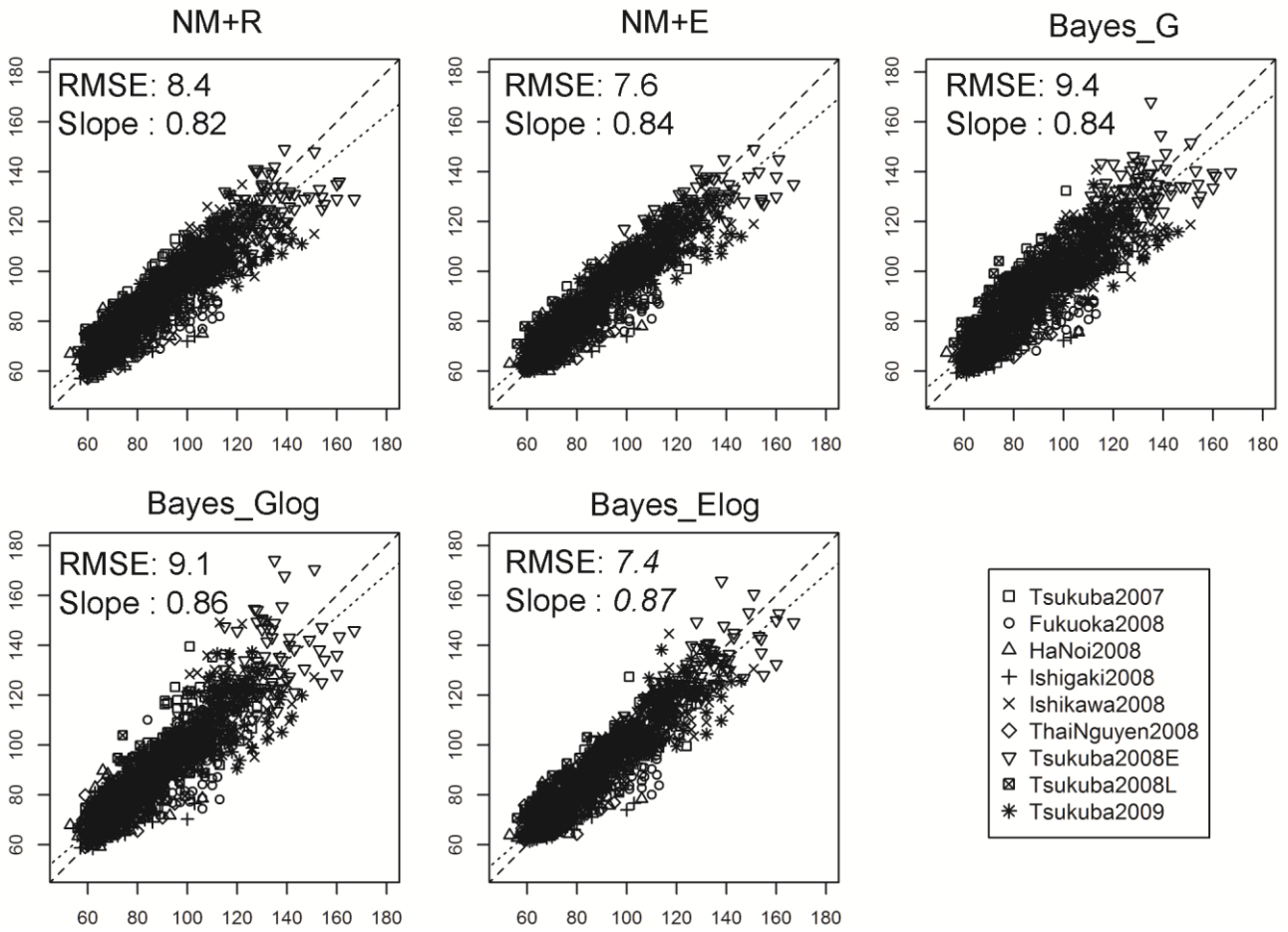

図5.6　1遺伝子型抜き交差検証（leave-one genotype-out cross-validation、LOGO）の結果。各手法について観察値（x軸）と予測値（y軸）をプロットした。破線は1対1の線を、ドット線は回帰直線を示す。平均二乗誤差（RMSE）と予測値の観察値に対する回帰係数（slope）を図中に示した。手法で用いられている略号は、NM（Nelder-Mead）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

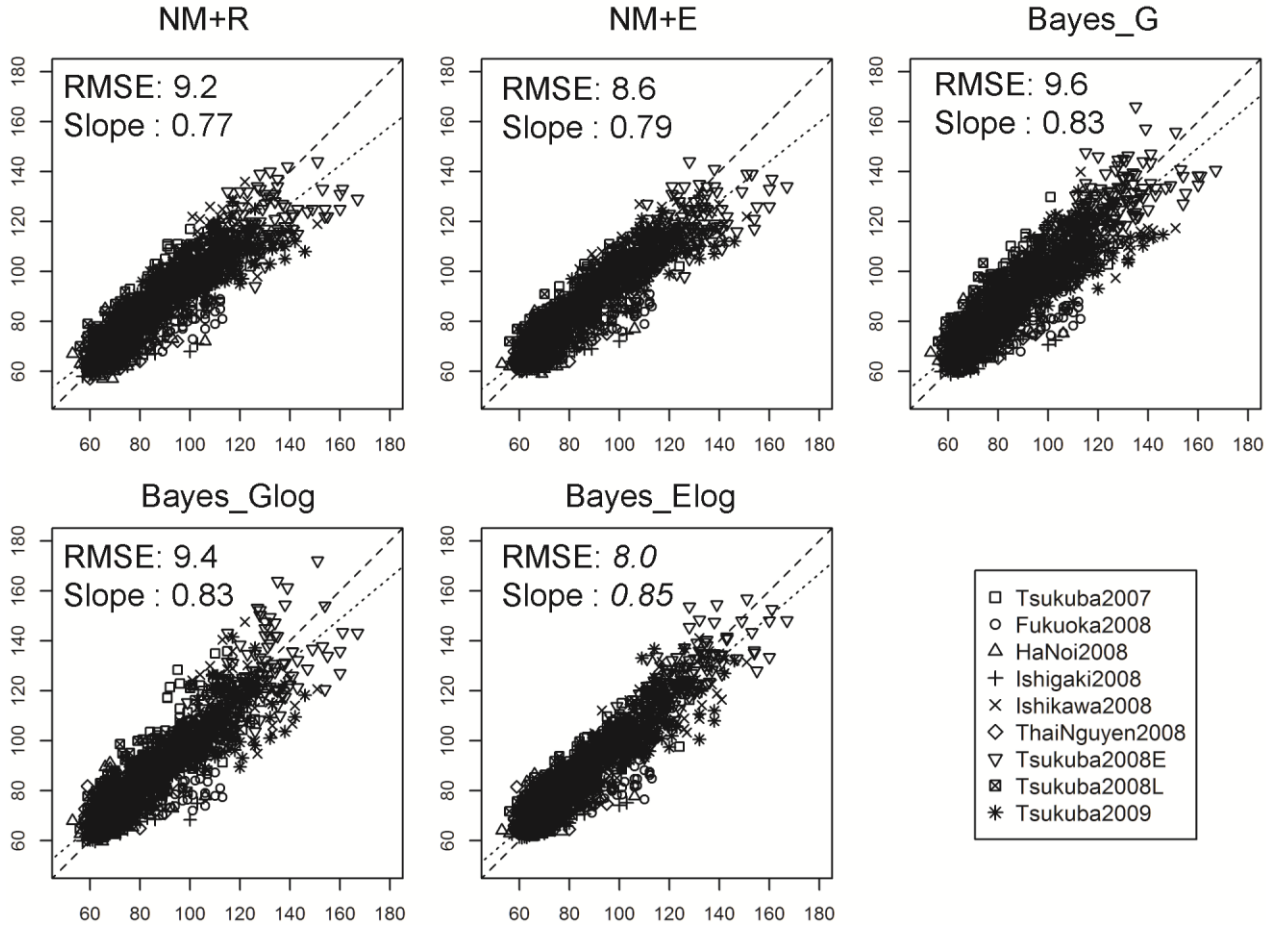

図5.7　1環境・遺伝子型組み合わせ抜き交差検証（leave-one combination of environments and genotypes-out cross-validation、LOEGO）の結果。各手法について観察値（x軸）と予測値（y軸）をプロットした。破線は1対1の線を、ドット線は回帰直線を示す。平均二乗誤差（RMSE）と予測値の観察値に対する回帰係数（slope）を図中に示した。手法で用いられている略号は、NM（Nelder-Mead）、R（RKHS）、E（EBlasso）、G（ガウシアンカーネル行列）、及びlog（自然対数変換）をそれぞれ表す。

ベイズ手法では予測された DH の事後分散を推定することができる。この事後分散の経験的な信頼性を調べるために、各交差検証法において、観察された（真の）DH が信用区間に含まれた頻度を計算した（表 5.7）。いずれの交差検証法においても、全ての手法が事後分散を過小推定していることが示唆された。過小推定は LOEO で顕著であったが、LOGO と LOEGO では過小評価は大幅に緩和された。

表 5.7　1 個抜き交差検証（leave-one-out cross validation、LOO）において観察された（真の）到穂日数（days to heading、DH）が信用区間に含まれた頻度 [a]

LOO	Bayesian	mean	Credible interval				
schemes[b]	methods[c]	Posterior SD[d]	95 %	90 %	80 %	70 %	60 %
LOEO	G	1.9 (±0.9)	0.46	0.40	0.32	0.25	0.18
	Glog	1.7 (±0.7)	0.44	0.37	0.29	0.22	0.18
	Elog	1.6 (±0.5)	0.42	0.35	0.26	0.19	0.14
LOGO	G	8.6 (±4.9)	0.86	0.78	0.69	0.60	0.51
	Glog	9.7 (±5.8)	0.93	0.88	0.78	0.69	0.57
	Elog	6.0 (±2.5)	0.86	0.78	0.69	0.58	0.49
LOEGO	G	8.6 (±4.9)	0.83	0.77	0.67	0.58	0.49
	Glog	9.9 (±6.0)	0.91	0.86	0.76	0.65	0.56
	Elog	6.1 (±2.5)	0.84	0.76	0.65	0.55	0.46

[a] どの LOO 検証においても、各環境における各遺伝子型の DH の事後分布が予測される。観察された（真の）DH が事後分布の信用区間に含まれた頻度を計算した。

[b]LOEO、leave-one environment-out cross-validation；LOGO、leave-one genotype-out cross- validation；LOEGO、leave-one combination of environments and genotypes-out cross-validation

[c]G、Bayes_G；Glog、Bayes_Glog；Elog、Bayes_Elog

[d]DH の事後標準偏差の平均値（標準偏差）

5.4 考察

本研究ではイネのDHをDVRモデルとゲノミックセレクションにおいて発展してきたゲノムワイド予測手法とを組み合わせた統合モデルを開発し予測した。統合モデル（Bayes_G、Bayes_Glog、及びBayes_Elog）を2段階法（NM+R、NM+E）及びモデルパラメータの推定方法が異なる2つのDVRモデル（NM及びBayes_I）と比較した結果、DVRモデルとゲノムワイド予測手法（EBlasso）をベイズ統計の枠組みで統合したモデル（Bayes_Elog）の結果が、予測の正確さ、縮約の程度、及び交差検証法間での安定性から最も好ましかった。Bayes_ElogはDVRモデルに基づき出穂期予測に特化した手法であるが、DVRモデルパラメータの推定を柔軟なメトロポリス法で行っているために、DVRモデルとは全く異なる関数から成り立つ他の作物モデルに対してもBayes_Elogと同様の手法は適用可能であろう。またDVRモデルパラメータに対するゲノムワイドマーカー効果を高速な変分ベイズ法で推定しているために、パラメータ数やマーカー数が増加してもその計算コストは大きくは上昇しない。そのため作物モデルとゲノムワイド予測手法を組み合わせたBayes_Elogのような統合モデリングは今後さらなる研究や適用を行うに値するであろう。一方でBayes_Elogと同じ回帰手法（EBL）を用いた2段階法、つまりNM+Eはより縮約傾向にあるにも関わらず（例えば図 5.7）、関連シグナル（大きな効果を持つマーカー）が多く観察される傾向にあった（図 5.4）。しかし現在入手できる情報からはこれらのシグナルが真か否かを判断することは難しい。本研究では予測に焦点を合わせたが、統合モデルのQTLマッピングにおける適性や能力も興味深い問題である。

LOEOにおけるNMの予測は9環境中Tsukuba2008Eで最も悪かった（この環境でのRMSEは12.6）。この環境下でのNMの予測はRKHS（つまりNM+R）またはEBlasso（NM+E）と組み合わせることで改善された（Tsukuba2008Eでの RMSEはそれぞれ9.4及び11.0）。この改善はおそらくDVRモデルパラメータのNMによる推定値に含まれていたノイズを除去したためであろう。このことは図5.2に示されるように、NM+R及びNM+Eによる推定値において遺伝子型間の分散が減少したことからも示唆される。しかしながら他環境での予測はゲノム情報を用いることで悪化し、NM+R及びNM+Eの全環境でのRMSEは改善されなかった（図5.5）。Yin et al.（2000）はオオムギの収量及びバイオマスを既知の遺伝子型について未試験の環境下で予測した場合、QTLから推定したパラメータを用いることで正確さが向上したと報告している。しかし一方Nakagawa et al.（2005）ではイネの出穂期の予測において同様の予測を行ったところ正確さが減少した。アブラナ属の開花期についても

同様の結果が報告されている（Uptmoor et al. 2012）。Uptmoor et al.（2012）は正確さ減少の理由として、検出されていない小さな効果の QTL が存在することと、QTL 効果の推定が不正確であることを挙げている。ゲノムワイドマーカーを予測変数として用いることで前者はある程度軽減できたであろうが、EBlasso 及び RKHS も予測値を縮約させる正則化回帰手法であるため、後者はやはり NM+R 及び NM+E の予測能力に影響したであろう。

NM に基づく手法とベイズ手法の相対的な予測の正確さは LOO の種類により異なった。LOEO では NM+R と NM+E は全てのベイズ手法より大きな RMSE を示したが、両手法は LOGO 及び LOEGO においては Bayes_G 及び Bayes_Glog より小さいかほぼ同等の RMSE を示した。NM+R と NM+E の予測値の観察値に対する回帰係数（slope）は全 LOO でベイズ手法より小さく、DH の予測値がより縮約していることが示唆された。第 3 章のランダムフォレストの説明において述べたように、汎化誤差は学習セットのランダム性に基づく分散と、その手法の推定バイアスに分けられる。縮約手法は前者の分散を減少させることにより汎化誤差を減らすため（Hastie et al. 2009）、LOGO 及び LOEGO における NM+R と NM+E の相対的に良好な予測結果は縮約によることが考えられた。ただし DVR モデルパラメータの縮約は新しい遺伝子型に対する予測（つまり LOGO 及び LOEGO）には有効であるが、未試験の環境下における既知遺伝子型に対する予測（LOEO）には有効ではない。これは予測されるモデルパラメータが環境にではなく遺伝子型に固有であることから明らかであろう。未試験環境下での既知遺伝子型に対する予測で、ノイズの除去を行う場合は、汎化誤差のバイアスを減らすような手法、例えば CART（第 3 章のランダムフォレストの説明を参照）などがより適切かもしれない。

Bayes_Elog は交差検証法に関わらず Bayes_G 及び Bayes_Glog より正確な予測を示した。これら手法間の RMSE の差は LOGO と LOEGO、つまり新しい遺伝子型に対して予測を行う場合により顕著であった。Bayes_Elog と 2 手法の差の 1 つは、ゲノム情報の用い方にある（表 5.1）。Bayes_Elog とその他 2 手法との差は回帰手法、EBlasso と RKHS、の差と捉えることができる。NM+E がやはり LOGO 及び LOEGO で NM+R より予測が正確であることを考えると、EBlasso の方が RKHS より本研究で用いた集団での DVR モデルパラメータの予測には適しているのであろう。第 3 章で示したように、EBlasso はエピスタシスがない場合は RKHS より概ね高い正確さを示し、エピスタシスがある場合は、QTL が少なく学習セットのサイズが小さい（100 から 300）場合に RKHS より高い正確さを示す傾向がある。アジア栽培イネの出穂期については、幾つかの主働遺伝子がその調節に重要な役割を

果たしていること、及びそれらの間の相互作用がその調節と局所適応に重要であることが示唆されている（Izawa 2007; Tsuji et al. 2011）。出穂期の調節にエピスタシスが存在すれば、EBlasso の優位性は比較的少数の QTL 及び小さなデータサイズによることが考えられた。一方で、出穂期に対する QTL 効果の非線形性が DVR モデルにおいてある程度考慮されているならば、パラメータに対する QTL 効果は DH に対するそれよりも相加的である可能性も考えられた。ソルガムにおけるシミュレーション研究では、相加的効果のみを持つ QTL から作物モデルを介して遺伝子間の相互作用のモデリングが可能であることが示唆されている（Chapman et al. 2003）。Bayes_Elog と Bayes_G 及び Bayes_Glog とのもう 1 つの相違点は、Bayes_Elog のみが DVR モデルパラメータに対して残差分散を想定している点にある。Bayes_Elog では各パラメータにおいてマーカーによりその分散が高い割合で説明されたものの、この違いが予測能力に与える影響は無視できないかもしれない。

DVR モデルパラメータ、*G*、α、及びβは 0 以上に制限されているために、統合モデルにおいては TrMVN を事前分布として用いることと、対数変換の 2 つの方法を試みた。前者の欠点はそのモード及び分散の推定が容易ではないために、本研究で行ったようにあらかじめ値を与えるならば、ある程度の情報を持つ事前分布を用いる必要がある点である。事実、Bayes_I において *G*、α、及びβの事前分布の分散を 100 倍にして無情報な事前分布に近づけたところ、LOEO における RMSE は 5.2 から 12.1 に増加した。Bayes_G は Bayes_I よりは頑強であったが、やはり同じ処理で LOEO における RMSE は 5.6 から 6.0 に増加した。加えて、TrMVN は正規分布より大きく逸脱した分布を持つパラメータの事前分布には不向きであろう。図 5.2 で示したように、Bayes_I 及び Bayes_G で推定したβの分布は、他手法のそれとは明らかに異なっていた。もし NM や Bayes_Elog で推定されたようなβの分布が真の分布に近いのであれば、Bayes_I や Bayes_G で 0 に近いβを正確に推定することが難しいことは納得しやすい。逆に当然ながら対数変換の欠点は、変換後の値が正規分布から大きく逸脱する場合に不向きな点である。本研究では 1 つのモデルの中ではどのパラメータも同じ種類の事前分布を用いたが、本来はパラメータ毎に設定可能である。例えば、対数変換はβには適切かもしれないが、*G* やαには TrMVN がより適切かもしれない。TrMVN や対数変換に代わる方法としては、代理（surrogate）パラメータを用いる方法が考えられる。例えば代理パラメータ$\tilde{\theta}$がもとのパラメータ θ に以下のような関数で変換されるとする。

$$\theta = g\left(\tilde{\theta}\right) = \begin{cases} \tilde{\theta} \text{ if } \tilde{\theta} > 0 \\ 0 \text{ if } \tilde{\theta} \leq 0 \end{cases}$$

このような代理パラメータは 0 から 1 の間に制限されるアリル頻度をモデルするために用いられている（Coop et al. 2010）。この方法の利点はパラメータの値が 0 になり得る点で、これは TrMVN や対数変換では不可能なことである。

ベイズ推定の利点の 1 つは予測値の事後分散が推定できる点にある。表 5.7 で示されるように、LOEO において得られた DH の事後分散は非常に過小評価されており、実用上の信頼性は低いと考えられる。一方で LOGO や LOEGO では特に Bayes_Glog において事後分散はある程度合理的な幅を有していた。Bayes_Elog は他のベイズ手法より事前分布に多くのパラメータを含んでおり、予測の不確実性が上昇することが考えられたが、DH の事後標準偏差は他の 2 手法と同等か小さかった。しかしながらより多くのマーカーを用いた場合は不確実性（事後分散）が上昇する可能性もある。

5.5 摘要

ゲノムワイド予測では環境情報を考慮しないために、未試験の環境下における植物の表現型を予測することがしばしば困難である。一方で作物モデルを用いた場合、環境情報から表現型を予測できるものの、未試験の遺伝子型に対する予測を行うことはできない。本研究では両者の欠点を補うために、イネ出穂期予測のための作物モデルである Developmental rate（DVR）モデルにゲノムワイドマーカー情報を統合した新たな予測モデルを開発した。9 つの環境で栽培された戻し交雑自殖系統集団を用いて、その予測能力を DVR モデル及び DVR モデルパラメータをゲノムワイドマーカーに回帰する 2 段階法と比較した。手法間比較は 3 種類の交差検証、1 環境抜き、1 遺伝子型抜き、及び 1 環境と 1 遺伝子型の組み合わせ抜きの交差検証で行った。その結果いずれの交差検証においても、DVR モデルパラメータをゲノムワイドマーカーに回帰し、そのマーカー効果を Extended Bayesian lasso でモデルパラメータと同時に推定する統合モデルが最も正確な予測を与えた。このことは作物の表現型予測における統合モデリングの可能性を示唆しており、今後のさらなる研究及び適用が期待される。

6　総合考察

本研究では全ゲノム情報を利用した新たな育種手法、ゲノミックセレクション及びゲノムワイド関連解析をより実用的な手法とするために、第 2 章ではゲノムワイド予測及びゲノムワイド関連解析のための高速ソフトウェアの開発、第 3 章ではイネ表現型予測のための予測手法に関する検証、第 4 章では黒毛和種におけるゲノム情報を用いた育種価予測に関する検証、そして第 5 章では環境と全ゲノム情報両方に基づく表現型予測のためのモデル開発を行った。序論で述べたように育種は統計学的な手法に支えられ発展してきた。そして統計的手法の発展は特定の作物や品種に限らず育種分野全体に恩恵をもたらすことができる。そのため本研究で得られた知見は実際にデータを扱ったイネやウシに限らず有用な知見となり得ると考えられる。

第2章では変分ベイズ法に基づいたゲノムワイド回帰を行うソフトウェアを開発し、シミュレーションを通してその性能を評価した。ゲノムワイド回帰において変分ベイズ法によるパラメータ推定がどの程度妥当か、という点については Carbonetto and Stephens（2012）に詳しい。著者らは BayesC と同じ回帰モデルにおいて、変分ベイズ法と MCMC を比較している。その結果、変数間の相関が低い（相関係数 0.2 の）場合は変分ベイズ法で近似した事後分布と MCMC による正確な事後分布は非常に近いものの、相関が高い（相関係数 0.8 の）場合は、変分ベイズによる近似事後分布は変数間の相関の高さに由来する不確実性を反映しないことを指摘している。このことが特にゲノムワイド予測において与える影響は明らかではないものの、本研究の第 2 章で用いた QTLMAS 第 15 回ワークショップのデータセットでは、隣り合うマーカー間の相関係数の絶対値は平均 0.40 であり、その影響はゲノムワイド予測及び関連解析いずれにおいても比較的小さかったと考えられる。その一方、第3章で用いたイネ品種集団では隣接マーカー間の相関係数の絶対値は平均 0.72（r^2 = 0.52）であり、このことが予測の正確さに影響を与えた可能性は考えられる。しかしながら変分ベイズ法に基づいた Blasso や EBlasso の予測能力は同じ正則化回帰手法である Lasso や ENet と遜色はなく、シミュレーション条件によっては上回っていたため、変数間の相関が変分ベイズによる予測に与える影響は本研究では示唆されなかった。また変分ベイズ法を用いる場合の難点の一つは局所解への収束であるが、第 5 章で開発した作物モデルと EBlasso との統合モデル Bayes_Elog において、異なる初期値から繰り返し計算した場合もほぼ同じ

解が得られたこと、また Bayes_Elog による予測が試みた全ての交差検証で最も正確であったことから、局所解への収束はゲノムワイド回帰においては大きな問題にならない可能性が示唆された。しかしながらマーカー数やサンプル数の増加により局所解への収束がより大きな問題となる可能性もある。そのため第 2 章で開発した VIGoR においても初期値をランダムに生成するオプションを追加し、異なる初期値から最適化が可能なようにする必要があると考えられる。なお近年分子生物学の発展は目覚ましく、次世代シークエンサーにより決定できる塩基配列数は飛躍的に増え（Mardis 2008）、これを利用した高密度マーカーの探索と遺伝子型決定は育種でも用いられ始めている（Poland and Rife 2012）。またウシではヒトの 1000 ゲノムプロジェクト（Kaiser 2008）にならい 1,000 頭の種雄牛についてその全塩基配列を決定するプロジェクトも進行している（Daetwyler et al. 2014）。そのため膨大な数のマーカーを統計的に利用した表現型や育種価の予測、また QTL の探索は今後も重要な課題となる。VIGoR の利用及び普及はこうした課題の解決に役立つことが期待される。一方、高速かつ正確な手法とそれを実装したソフトウェアの開発については今後も続けていく必要があると考えられる。

第 3 章ではアジア栽培イネにおいてゲノムワイド予測が育種において有用な技術となり得ることを示した。また 8 つのゲノムワイド予測手法と、その平均値の合計 9 つの手法を比較し、シミュレーションデータを通してそれらの手法の適用範囲を明らかにし、それによりイネだけでなく他の作物においても有益な情報を提供した。なお本研究では上述の 9 手法を取り上げたが、統計学及び機械学習の分野においては様々な予測或いは分類手法が次々と提案されている。本研究で取り上げなかった手法のうち、今後遺伝学においても重要性が増す可能性があるものにニューラルネットワークが挙げられる。比較的近年提案された Deep learning（Hinton et al. 2006）や Extreme learning machine（Huang et al. 2006）などは、特に画像認識において良好な成績を残している。本研究で扱ったイネの実データでは非線形性回帰手法である RKHS や RForest が比較的良好な成績を示しており、同様に非線形性を持つこれら手法のゲノムワイド予測への適用は非常に興味深い。

第 4 章では黒毛和種においてゲノムワイド予測が従来手法である BLUP より有用であることを示した。動物育種においては作物や果樹の育種のように、自殖や接ぎ木などで優秀な個体のクローンを増殖させ利用することができないため、必然的に遺伝的に異なる個体が育種過程において植物よりはるかに多く生じることになる。そのため表現型記録は持つがマーカージェノタイピングがされていない、つまり全ゲノム情報は持たない個体が

多く存在する。一般的なゲノムワイド予測ではゲノム情報と表現型情報両方を持つ個体しか利用することができないため、このような表現型記録しか持たない個体をどのように利用するか、ということは動物育種固有の課題と言える。本研究ではこの課題を解決するために ssGBLUP を用いたが、他にもベイズ統計の枠組みで明示的に回帰を行う手法（Fernando et al. 2014）も提案されており、今後はこれら手法間の比較も必要となるだろう。本研究では 3 つの枝肉形質、BMS、CW、及び REA を取り上げたが、このうち特に BMS は黒毛和種においてこれまで重点的に改良されてきた形質であり、既に多くの種雄牛が高い能力を有していると言える。一方で繁殖性や飼養効率性など、農家の長期的な利益に直結する重要なこれら形質は表現型測定の困難さもあってあまり改良が進んでいない。そのためゲノム情報を用いた育種の高速化はこのような形質に特に有用であると考えられる。またこのような新たな形質について改良する場合、従来改良を進めてきた形質についても配慮する必要があるが、そのためには 1 つの形質のみを考慮するのではなく、多形質を扱うゲノムワイド予測モデルも試みる必要があるだろう。

第 5 章ではゲノムと環境情報両方から表現型を予測する手法を新たに提案し、それがイネ出穂期予測において有用であることを示した。環境への応答とそれによる表現型の変化とくに植物において顕著であり、環境とゲノム情報の両方を考慮した統計学的予測モデルは、気候変動が大きくなると予想される今後、より重要性を増していくと考えられる。また近年、次世代シークエンサーや質量分析法の進歩により、DNA、RNA、タンパク質、代謝物、或いは無機物などについての情報が豊富に得られるようになってきた。ゲノム（genome）と表現型（phenome）だけではないこれら多様な情報はオミックス（omics）データと総称され、その統合的な利用は複雑な生命現象を理解するための鍵として注目を集めている（Shinozaki and Sakakibara 2009; Berry et al. 2011; Berg 2013; Deshmukh et al. 2014）。Nagano et al.（2012）は圃場で栽培されたイネの mRNA 発現量の変動が、気温や湿度などの環境情報の関数を用いて予測できることを示した。このことはオミックスデータにさらに環境情報を組み合わせた包括的かつ動的な予測モデルの可能性を示唆している。また代謝産物量を形質としたゲノムワイド関連解析も例えばトウモロコシ（Riedelsheimer et al. 2012b; Wen et al. 2014）やイネ（Chen et al. 2014）で報告されている。一方で表現型予測については、オミックスデータの中では表現型に近い代謝産物に基づく予測がイネ（Redestig et al. 2011）やトウモロコシ（Riedelsheimer et al. 2012c）、及びブタ（Rohart et al. 2012）の経済形質で報告されている。こういった様々な種類の情報の蓄積によりそれらを統合的に育

種へ利用できる可能性が広がる一方で、これらの研究は 2 種類の情報、例えば環境と発現情報、あるいは代謝産物と表現型情報、に基づいた比較的単純なモデリングを試みている。本研究では 3 種類の情報、つまりゲノム、環境、及び表現型情報を統合したモデルを開発・提案し、その有用性を示した。3 種類の情報を同時にモデリングする試みはこれまでにほとんどなく、画期的である。今後はこうしたアプローチを発展させ、環境情報、ゲノム情報、及び表現型にその他オミックスデータを取り込んだより包括的なモデリングを開発することで、特に植物育種が大きな進展を果たす可能性が考えられる。

本研究では全ゲノム情報を育種で活用するためのソフトウェアを第 2 章で開発し、第 3 章及び 4 章においてそれぞれイネ及びウシの実データを用いて全ゲノム情報を育種に活用する有用性を示し、そして第 5 章では全ゲノム情報に環境情報を加えた新たな統計モデルの可能性を示した。第 3 章及び 4 章で示されたその有用性を考慮しても、今後全ゲノム情報が動植物の育種を支えていくことは確実と言えるだろう。それに加え第 5 章で提案したような環境や他のオミックス情報を加えた統合モデリングが今後は活発に研究されていくと思われる。しかしながらそういった新たなモデルや手法を実際に育種へ利用していくためには、第 2 章で開発したような常に使いやすく高速かつ安定したソフトウェアが必須となる。かつて生物は表現型のみが観察できた。そこから統計学者や遺伝学者は見えない事実、つまり遺伝子や環境の表現型への作用を科学的に推論するために統計的手法を発展させ育種へと応用してきた。次に遺伝子型を観察できるようになり、連鎖解析やゲノムワイド予測及び関連解析といった統計的手法が利用されるようになった。生物において観察可能な事柄は確実に増えている。それら手に入る多種多様な生物情報から育種において有用な知識を得るためには、統計学的モデル及び推定手法、さらにはそれを汎用化するためのソフトウェアの開発が今後ますます重要になるであろう。

7　謝辞

本研究を遂行し論文を執筆するにあたり、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・環境生物学専攻生物測定学研究室の岩田洋佳准教授からは多くのご指導及びご助言を頂いた。この場を借りて深く感謝の意を表したい。また同研究室の岸野洋久教授及び大森宏助教にも研究を進めるにあたり適宜ご助言を頂いたことにここで謝意を表する。本論文の副査を務めて頂いた岸野洋久教授、東京大学大学院新領域創成科学研究科複雑理工学専攻の杉山将教授、東京大学大学院農学生命科学研究科附属生態調和農学機構の二宮正士教授、及び筑波大学大学院生命環境科学研究科先端農業技術科学専攻の林武司教授には有益なご指摘を数多く頂いたことに深く感謝する。第 3 章で用いた日本水稲データを提供して頂いた農業・食品産業技術総合研究機構近畿中国四国農業研究センターの出田収博士、農業生物資源研究所の江花薫博士、神戸大学大学院農学研究科附属食資源教育研究センターの山崎将紀准教授及び吉岡拓磨氏にもこの場を借りて謝意を表したい。山崎将紀准教授にはイネの出穂期やその遺伝的構造についても度々ご教授頂いた。第 4 章で使用した黒毛和種のデータを提供して頂いた一般社団法人家畜改良事業団及び山形県農業総合研究センターにもこの場を借りて謝意を表する。両団体からの全面的な協力なしに研究を完遂することは不可能であった。第 5 章で用いた BIL データを提供して頂いた九州大学大学院農学研究院資源生物科学部門の望月俊宏教授、農業・食品産業技術総合研究機構の中川博視博士、及び農業環境技術研究所の長谷川利拡博士にも謝意を表する。長谷川利拡博士にはデータ提供だけでなく、形質調査を通じてイネの作物学的側面についてもご教授頂いた。また研究を側面から常に支援して頂いた生物測定学研究室の佐々木三枝子事務員にもこの場を借りて深く感謝の意を表したい。最後にその優秀な頭脳で常に良い刺激を与え続けてくれた同研究室の学生及び研究員の方々にもここで謝意を表する。

8 参考文献

Agarwala V, Flannick J, Sunyaev S, Altshuler D, GoT2D C (2013) Evaluating empirical bounds on complex disease genetic architecture. Nat Genet 45:1418-1429.

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93:743-752.

Aguilar I, Misztal I, Legarra A, Tsuruta S (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J Anim Breed Genet 128:422-428.

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schon CC (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339-350.

Arakawa A, Iwaisaki H, Anada K (2009) Estimation of breeding values from large-sized routine carcass data in Japanese Black cattle using Bayesian analysis. Anim Sci J 80:617-623.

Barendse W, Reverter A, Bunch RJ, Harrison BE, Barris W, Thomas MB (2007) A validated whole-genome association study of efficient food conversion in cattle. Genetics 176:1893-1905.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van GA, Zelenika D, Franchimont D, Hugot JP, de VM, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ (2008) Genome-wide association defines more than 30 distinct

susceptibility loci for Crohn's disease. Nat Genet 40:955-962.

Basten CJ, Weir BS, Zeng ZB (1994) Zmap–a QTL cartographer Smith C, Gavora JS, Chesnais BBJ, Fairfull W, Gibson JP, Kennedy BW, Burnside EB (Eds.), Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software, Volume 22, Guelph, Ontario, Canada, pp. 65–66. Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production.

Berg EL (2013) Systems biology in drug discovery and development. Drug Discov Today 19:113-125.

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649-1664.

Berry DP, Meade KG, Mullen MP, Butler S, Diskin MG, Morris D, Creevey CJ (2011) The integration of 'omic'disciplines and systems biology in cattle breeding. Animal 5:493-505.

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York.

Bishop MD, Kappes SM, Keele JW, Stone RT, Sunden SL, Hawkins GA, Toldo SS, Fries R, Grosz MD, Yoo J, et (1994) A genetic linkage map for cattle. Genetics 136:619-639.

Bogard M, Ravel C, Paux E, Bordes J, Balfourier F, Chapman SC, Le GJ, Allard V (2014) Predictions of heading date in bread wheat (Triticum aestivum L.) using QTL-based parameters of an ecophysiological model. J Exp Bot 65:5849-5865.

Boldman KL (1993) A Manual for Use of MTDFREML. A Set of Programs to Obtain Estimates of Variances and Covariances [DRAFT] US Department of Agriculture, Agricultural Research Service.

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man

using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees CRC press, Boca Raton, FL.

Breiman L (1996) Bagging predictors. Machine learning 24:123-140.

Breiman L (2001) Random forests. Machine learning 45:5-32.

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210-223.

Browning SR, Browning BL (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 81:1084-97.

Buhlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications Springer, Berlin.

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van EP (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171-182.

Carbonetto P, Stephens M (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal 7:73-108.

Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM (1988) Restriction fragment length polymorphism linkage map for Arabidopsis thaliana. Proc Natl Acad Sci U S A 85:6856-6860.

Chapman S, Cooper M, Hammer GL (2002) Using crop simulation to generate genotype by environment interaction effects for sorghum in water-limited environments. Crop Pasture Sci 53:379-389.

Chapman S, Cooper M, Podlich D, Hammer G (2003) Evaluating plant breeding strategies by simulating gene action and dryland environment effects. Agron J 95: 99-113.

Chen CY, Misztal I, Aguilar I, Legarra A, Muir WM (2011) Effect of different genomic relationship matrices on accuracy and scale. J Anim Sci 89:2673-2679.

Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H, Zhang W, Zhang L, Yu S, Wang G, Lian X, Luo J (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 46:714-721.

Chhikara RS, Folks JL (1989) The inverse Gaussian distribution MARCEL DEKKER, NewYork.

Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2.

Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G (2012) Single-step methods for genomic evaluation in pigs. Animal 6:1565-1571.

Clark SA, Hickey JM, van JH (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18.

Colautti RI, Eckert CG, Barrett SC (2010) Evolutionary constraints on adaptive evolution during range expansion in an invasive plant. Proc Biol Sci 277:1799-1806.

Coop G, Witonsky D, Di RA, Pritchard JK (2010) Using environmental correlations to identify loci

underlying local adaptation. Genetics 185:1411-1423.

Crossa J, Campos GL, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713-724.

Crossa J, Beyene Y, Kassa S, Perez P, Hickey JM, Chen C, de G, Burgueno J, Windhausen VS, Buckler E, Jannink JL, Lopez CMA, Babu R (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903-1926.

Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection J Anim Breed Genet 124:369-376.

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

Daetwyler HD, Calus MP, Pong-Wong R, de LG, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347-365.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van BR, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, Esquerre D, Bouchez O, Rossignol MN, Klopp C, Rocha D, Fritz S, Eggen A, Bowman PJ, Coote D, Chamberlain AJ, Anderson C, VanTassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B, Lund MS, Veerkamp RF, Boichard DA, Fries R, Hayes BJ (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-865.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009a) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375-385.

de los Campos G, Gianola D, Rosa GJ (2009b) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87:1883-1887.

de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb.) 92:295-308.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B Met 39:1-38.

Deshmukh R, Sonah H, Patil G, Chen W, Prince S, Mutava R, Vuong T, Valliyodan B, Nguyen HT (2014) Integrating omic approaches for abiotic stress tolerance in soybean. Front Plant Sci 5:244.

Dietterich TG (2000) Ensemble methods in machine learning Multiple classifier systems (pp. 1-15). Springer Berlin Heidelberg.

Donoho D, Stodden V (2006) Breakdown point of model selection when the number of variables exceeds the number of observations. Proceedings of the International Joint Conference on Neural Networks: 1916-1921.

Elsen JM, Tesseydre S, Filangi O, Le RP, Demeure O (2012) XVth QTLMAS: simulated dataset. BMC Proc 6 Suppl 2:S1.

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Human hered 21:523-542.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package

rrBLUP. Plant Genome 4: 250-255.

Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 (Bethesda) 2:1405-1413.

Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink JL (2014) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. 54:48-59.

Fernando RL, Dekkers JC, Garrick DJ (2014) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol 46:50.

Fernando R, Garrick DJ (2008) GenSel: User Manual for a Portfolio of Genomic Selection Related Analyses Animal Breeding and Genetics, Iowa State University, Ames, IA. Available at http://bigs.ansci.iastate.edu/bigsgui.

Fisher RA (1918) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the royal society of Edinburgh 52:399-433.

Fisher RA (1935) The detection of linkage with “dominant” abnormalities. Annals of Eugenics 6:187-201.

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33: 1-22.

Garrick DJ (2011) The nature, scope and impact of genomic prediction in beef cattle in the United States. Genet Sel Evol 43:17.

Garrick DJ, Dekkers JCM, Golden BL, Fernando RL (2014) Bayesian prediction combining genotyped and non-genotyped individuals. 10th World Congress on Genetics Applied to Livestock Production.

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88: 881-889.

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761-1776.

Gianola D, van Kaam JB (2008) Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289-2303.

Gianola D, de G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347-363.

Gilmour AR, Cullis BR, Welham SJ, Thompson R (1998) ASReml user's manual New South Wales Agriculture, Orange, Australia, 185.

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Gonzalez-Camacho JM, de LG, Perez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759-771.

Gonzalez-Recio O, Gianola D, Long N, Weigel KA, Rosa GJ, Avendano S (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178:2305-2313.

Gu J, Yin X, Zhang C, Wang H, Struik PC (2014) Linking ecophysiological modelling with quantitative genetics to support marker-assisted crop design for improved yields of rice (*Oryza sativa*) under drought stress. Ann Bot 114:499-511.

Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, van EF, Chapman S, Podlich D (2006) Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci 11:587-593.

Harris BL, Winkelman AM, Johnson DL (2012) Large-scale single-step genomic evaluation for milk production traits. Interbull Bulletin (46).

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York.

Hayashi T, Iwata H (2010) EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet 11:3.

Hayashi T, Iwata H (2013) A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics 14:34.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433-443.

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1-12.

Heffner EL, Jannink JL, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65-75.

Helentjaris T, Slocum M, Wright S, Schaefer A, Nienhuis J (1986) Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor Appl Genet 72:761-769.

Henderson CR (1949) Estimation of changes in herd environment. J Dairy Sci 32:706.

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-447.

Henderson CR (1984) Applications of Linear Models in Animal Breeding. Third Edition edited by Schaeffer LR University of Guelph, Guelph, Ontario.

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146-160.

Heslot N, Jannink JL, Sorrells ME (2014) Perspectives for genomic selection applications and research in plants. Crop Sci doi:10.2135/cropsci2014.03.0249.

Hickey JM, Tier B (2009) AlphaBayes (Beta): Software for polygenic and whole genome analysis. User Manual. University of New England, Armidale, Australia.

Hickey JM, Gorjanc G (2012) Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 (Bethesda) 2:425-427.

Hill WG (1996) Sewall Wright's "Systems of Mating". Genetics 143:1499-1506.

Hinton G, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural comput 18:1527-1554.

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70:489-501.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961-967.

Hunt LA, Pararajasingham S (1995) CROPSIM-WHEAT: a model describing the growth and development of wheat. Can J Plant Sci 75: 619-632.

Inoue K, Kobayashi M, Shoji N, Kato K (2011) Genetic parameters for fatty acid composition and feed efficiency traits in Japanese Black cattle. Animal 5:987-994.

Ishwaran H, Rao JS (2005) Spike and slab variable selection: Frequentist and Bayesian strategies. Ann Stat 33:730-773.

Iwata H, Ninomiya S (2006) AntMap: constructing genetic linkage maps using an ant colony optimization algorithm. Breeding Sci 56:371-377.

Izawa T (2007) Adaptation of flowering-time by natural and artificial selection in Arabidopsis and rice. J Exp Bot 58:3091-3097.

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomic Proteomic 9:166-177.

Janss LLG (2010) Bayz manual version version 2.03 Janss Biostatistics, Leiden, The Netherlands.

Available at http://www.bayz.biz/.

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513-1522.

Jiang R, Tang W, Wu X, Fu W (2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10 Suppl 1:S65.

Kaiser J (2008) A plan to capture human diversity in 1000 genomes. Science 319:395.

Karkkainen HP, Sillanpaa MJ (2012a) Back to basics for Bayesian model building in genomic selection. Genetics 191:969-987.

Karkkainen HP, Sillanpaa MJ (2012b) Robustness of Bayesian multilocus association models to cryptic relatedness. Ann Hum Genet 76:510-523.

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. The International Joint Conference on Artificial Intelligence (IJCAI) 14:1137-1145.

Krogh A, Vedelsby J (1995) Neural network ensembles, cross validation, and active learning Tesauro G, Touretzky DS, and Leen TK (eds), Advances in neural information processing systems 7:231-238, MIT Press, Cambridge MA.

Lander ES, Botstein D (1986) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map Cold Spring Harbor symposia on quantitative biology (Vol. 51, pp. 49-62). Cold Spring Harbor Laboratory Press.

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-181.

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84:2363-2367.

Legarra A, Robert-Granie C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611-618.

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. J Dairy Sci 92:4656-4663.

Legarra A, Ricardi A, Filangi O (2010) GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCπ and Bayesian Lasso) Available at http://snp.toulouse.inra.fr/~alegarra/.

Li Z, Sillanpaa MJ (2012a) Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190:231-249.

Li Z, Sillanpaa MJ (2012b) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125:419-435.

Liu Z, Goddard ME, Reinhardt F, Reents R (2014) A single-step genomic model with direct estimation of marker effects. J Dairy Sci 97:5833-5850.

Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A, Gonzalez-Recio O (2010) Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetic Res. 92:209-225.

Long N, Gianola D, Rosa GJ, Weigel KA (2011a) Application of support vector regression to genome-assisted prediction of quantitative traits. Theor Appl Genet 123:1065-1074.

Long N, Gianola D, Rosa GJ, Weigel KA (2011b) Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. J Anim Breed Genet

128:247-257.

Lorenz AJ, Smith KP, Jannink JL (2012) Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci 52:1609-1621.

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151-161.

Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen TH (2009) The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119-1126.

Lund MS, Sahana G, de KDJ, Su G, Carlborg O (2009) Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proc 3 Suppl 1:S1.

Ma JF, Shen R, Zhao Z, Wissuwa M, Takeuchi Y, Ebitani T, Yano M (2002) Response of rice to Al stress and identification of quantitative trait Loci for Al tolerance. Plant Cell Physiol 43:652-659.

Madsen P, Sørensen P, Su G, Damgaard LH, Thomsen H, Labouriau R (2006) DMU-a package for analyzing multivariate mixed models 8th World Congress on Genetics Applied to Livestock Production (Vol. 247).

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de G (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7:e1002051.

Malosetti M, Visser RG, Celis-Gamboa C, van EFA (2006) QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. Theor Appl Genet 113:288-300.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends in genetics 24:133-141.

Matsubara K, Kono I, Hori K, Nonoue Y, Ono N, Shomura A, Mizubayashi T, Yamamoto S, Yamanouchi U, Shirasawa K, Nishio T, Yano M (2008) Novel QTLs for photoperiodic flowering revealed by using reciprocal backcross inbred lines from crosses between japonica rice cultivars. Theor Appl Genet 117:935-945

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369.

McCouch SR, Kochert G, Yu ZH, Wang ZY, Khush GS, Coffman WR, Tanksley SD (1988) Molecular mapping of rice chromosomes. Theor Appl Genet 76:815-829.

McDaneld TG, Kuehn LA, Thomas MG, Snelling WM, Smith TP, Pollak EJ, Cole JB, Keele JW (2014) Genomewide association study of reproductive efficiency in female cattle. J Anim Sci 92:1945-1957.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087-1092.

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH (2002) BLUPF90 and related programs Proc. 7th World Congr. Genet. Appl. Livest, Prod. Montpellier, France.

Monna L, Lin X, Kojima S, Sasaki T, Yano M (2002) Genetic dissection of a genomic region for a quantitative trait locus, Hd3, into two loci, Hd3a and Hd3b, controlling heading date in rice. Theor Appl Genet 104:772-778.

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A 110:453-458.

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318.

Mujibi FD, Nkrumah JD, Durunna ON, Stothard P, Mah J, Wang Z, Basarab J, Plastow G, Crews DHJ, Moore SS (2011) Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. J Anim Sci 89:3353-3361.

Murphy KP (2012) Machine learning: a probabilistic perspective MIT press, London.

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321-4325.

Mutshinda CM, Sillanpaa MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186:1067-1075.

Nadaf J, Riggio V, Yu TP, Pong-Wong R (2012) Effect of the prior distribution of SNP effects on the estimation of total breeding value. BMC Proc 6 Suppl 2:S6.

Nagano AJ, Sato Y, Mihara M, Antonio BA, Motoyama R, Itoh H, Nagamura Y, Izawa T (2012) Deciphering and prediction of transcriptome dynamics under fluctuating field conditions. Cell 151:1358-1369.

Nagasaki, H., Ebana, K., Shibaya, T., Yonemaru, J. I, Yano, M. (2010) Core single-nucleotide polymorphisms—a tool for genetic analysis of the Japanese rice population. Breeding Sci. 60:648-655.

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

Nogi T, Honda T, Mukai F, Okagaki T, Oyama K (2011) Heritabilities and genetic correlations of fatty acid compositions in longissimus muscle lipid with carcass traits in Japanese Black cattle. J Anim Sci 89:615-621.

Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H (2011) Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188:695-708.

Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc 5 Suppl 3:S11.

Ogutu JO, Schulz-Streeck T, Piepho HP (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc 6 Suppl 2:S10.

Ott J (1976) A computer program for linkage analysis of general human pedigrees. Am J Hum Genet 28:528-529.

Park T, Casella G (2008) The Bayesian lasso. J Ame Stat Assoc 103:681-686.

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721-726.

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545-554.

Perez P, de LG (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package.

Genetics 198:483-495.

Perez-Rodriguez P, Gianola D, Gonzalez-Camacho JM, Crossa J, Manes Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2:1595-1605.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92-102.

Pszczola M, Mulder HA, Calus MP (2011) Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. J Dairy Sci 94:431-441.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de BPIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559-575.

Provine WB (2001) The Origins of Theoretical Population Genetics: With a New Afterword University of Chicago Press (London).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de BPIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559-575.

Quilot B, Kervella J, Genard M, Lescourret F (2005) Analysing the genetic control of peach fruit quality through an ecophysiological model combined with a QTL approach. J Exp Bot 56:3083-3092.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Redestig H, Kusano M, Ebana K, Kobayashi M, Oikawa A, Okazaki Y, Matsuda F, Arita M, Fujita N, Saito K (2011) Exploring molecular backgrounds of quality traits in rice by predictive models based on high-coverage metabolomics. BMC Syst Biol 5:176.

Reymond M, Muller B, Leonardi A, Charcosset A, Tardieu F (2003) Combining quantitative trait Loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol 131:664-675.

Riedelsheimer C, Technow F, Melchinger AE (2012a) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452.

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012b) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci U S A 109:8872-8877.

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012c) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217-220.

Rohart F, Paris A, Laurent B, Canlet C, Molina J, Mercat MJ, Tribout T, Muller N, Iannuccelli N, Villa-Vialaneix N, Liaubet L, Milan D, San CM (2012) Phenotypic prediction based on metabolomic data for growing pigs from three main European breeds. J Anim Sci 90:4729-4740.

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JC, Fernando RL, Schnabel RD, Garrick DJ, Taylor JF (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet Sel Evol 43:40.

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis Cambridge University Press, New York.

Shinozaki K, Sakakibara H (2009) Omics and bioinformatics: an essential toolbox for systems analyses of plant functions beyond 2010. Plant Cell Physiol 50:1177-1180.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH (2009) Reducing dimensionality for prediction of genome-wide breeding values. Genet Sel Evol 41:29.

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York.

Sukumaran S, Xiang W, Bean SR, Pedersen JF, Kresovich S, Tuinstra MR, Tessoa TT, Hamblin MT, Yu J (2012) Association mapping for grain quality in a diverse sorghum collection. Plant Genome 5:126-135.

Sun X, Qu L, Garrick DJ, Dekkers JC, Fernando RL (2012) A fast EM algorithm for BayesA-like prediction of genomic breeding values. PLoS One 7:e49157.

Takahashi Y, Shomura A, Sasaki T, Yano M (2001) Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. Proc Natl Acad Sci U S A 98:7922-7927.

Tardieu F (2003) Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. Trends Plant Sci 8:9-14.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B Met 58:267-288.

Tsuji H, Taoka K, Shimamoto K (2011) Regulation of flowering in rice: two florigen genes, a

complex gene network, and natural variation. Curr Opin Plant Biol 14:45-52.

Uptmoor R, Schrag T, Stützel H, Esch E (2008) Crop model based QTL analysis across environments and QTL based estimation of time to floral induction and flowering in Brassica oleracea. Mol Breed 21:205-216.

Uptmoor R, Osei-Kwarteng M, Gürtler S, Stützel H (2009) Modeling the effects of drought stress on leaf development in a Brassica oleracea doubled haploid population using two-phase linear functions. J Am Soc Hortic Sci 134:543-552.

Uptmoor R, Li J, Schrag T, Stützel H (2012) Prediction of flowering time in Brassica oleracea using a quantitative trait loci-based phenology model. Plant Biology 14:179-189

Usai MG, Goddard ME, Hayes BJ (2009) LASSO with cross-validation for genomic selection. Genetic Res 91:427-436.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414-4423.

Vitezica ZG, Aguilar I, Misztal I, Legarra A (2011) Bias in genomic predictions for populations under selection. Genet Res (Camb) 93:357-366.

Wang CS, Rutledge JJ, Gianola D (1993) Marginal Inferences About Variance Components in a Mixed Linear Model Using Gibbs Sampling Genet Select Evol 25:41–62.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, Yan J (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun 5:3438.

Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schon CC (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195:573-587.

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330-338.

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801.

Xue W, Xing Y, Weng X, Zhao Y, Tang W, Wang L, Zhou H, Yu S, Xu C, Li X, Zhang Q (2008) Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. Nat Genet 40:761-767.

Yamamoto T, Nagasaki H, Yonemaru J, Ebana K, Nakajima M, Shibaya T, Yano M (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11:267.

Yamasaki M, Ideta O (2013) Population structure in Japanese rice population. Breed Sci 63:49-57.

Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, von SR, Yi N (2007) R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. Bioinformatics 23:641-643.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. Nat Genet 46:100-106.

Yano M, Harushima Y, Nagamura Y, Kurata N, Minobe Y, Sakaki T (1997) Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. Theor Appl Genet 95:1025-1032.

Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, Baba T, Yamamoto K, Umehara Y, Nagamura Y, Sasaki T (2000) Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. Plant Cell 12:2473-2484.

Yao C, Spurlock DM, Armentano LE, Page CDJ, Vandehaar MJ, Bickhart DM, Weigel KA (2013) Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. J Dairy Sci 96:6716-6729.

Yin X, Kropff MJ, Horie T, Nakagawa H, Centeno HG, Zhu D, Goudriaan J (1997) A model for photothermal responses of flowering in rice I. Model description and parameterization. Field Crop Res 51:189-200.

Yin X, Chasalow SD, Dourleijn CJ, Stam P, Kropff MJ (2000) Coupling estimated effects of QTLs for physiological traits to a crop growth model: predicting yield variation among recombinant inbred lines in barley. Heredity (Edinb) 85:539-549.

Yin X, Stam P, Kropff MJ, Schapendonk AH (2003) Crop modeling, QTL mapping, and their complementary role in plant breeding. Agron J 95: 90-98.

Yin X, Struik PC, Kropff MJ (2004) Role of crop physiology in predicting gene-to-phenotype relationships. Trends Plant Sci 9:426-432.

Young ND (1996) QTL mapping and quantitative disease resistance in plants. Annu rev phytopathol 34:479-501.

Yu J, Pressoir G, Briggs WH, Vroh BI, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208.

Zeng ZB (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci U S A 90:10972-10976.

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457-1468.

Zhang Z, Liu J, Ding X, Bijma P, de KDJ, Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5: e12648.

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2:467.

Zhao Y, Gowda M, Liu W, Wurschum T, Maurer HP, Longin FH, Ranc N, Reif JC (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769-776.

Zhong S, Dekkers JC, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. Genetics 182:355-364.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Stat Soc B 67:301-320.

内藤元男（1978）原色図鑑世界の牛．養賢堂（東京）．

ラディジンスキー G.（2000） 栽培植物の進化　自然と人間がつくる生物多様性　藤巻宏訳. 農山漁村文化協会（東京）．