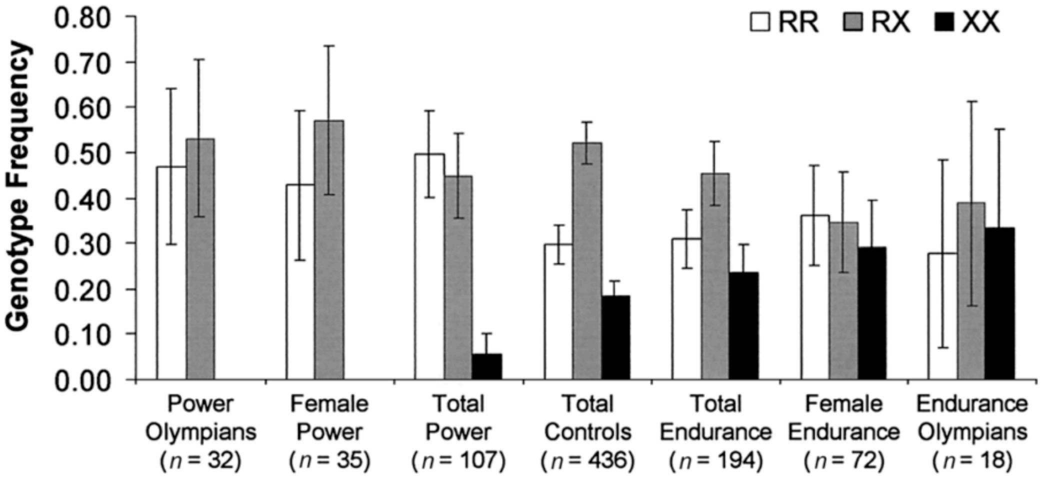
ゲノムと表現型情報の関連を利用する
～GWASとゲノミックプレディクション～

岩田洋佳

生物測定学研究室

足の速さのアソシエーション解析

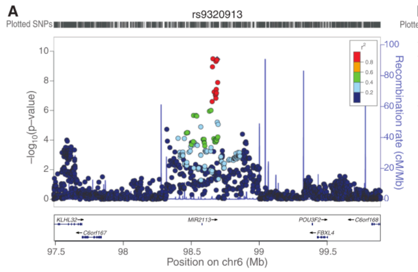


Yang et al. (2003) Am. J. Hum. Genet 73: 627

Figure 1 *ACTN3* genotype frequency in controls, elite sprint/power athletes, and endurance athletes. Compared with healthy white controls, there is a marked reduction in the frequency of the *ACTN3* 577XX genotype (associated with α-actinin-3 deficiency) in elite white sprint athletes; remarkably, none of the female sprint athletes or sprint athletes who had competed at the Olympic level (25 males and 7 females) were α-actinin-3 deficient. Conversely, there is a trend toward an increase in the 577XX genotype in endurance athletes, although this association reaches statistical significance only in females. Error bars indicate 95% CIs.

- ACTN3は、骨格筋のアクチン結合タンパク質α-ctinin-3をコードしている
- アリルXは、終止コドンによってα-ctinin-3が作ることができない
- a-actinin-3の働きはa-actinin-2によって補完されるが、ACTN3の保存性の高さから機能が異なると言われていた

126,559人のデータに基づく最終学歴のGWAS



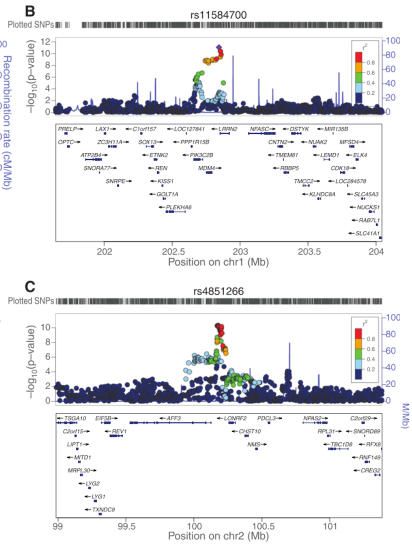


Fig. 1. Regional association plots of replicated loci associated with educational attainment. (**A**) rs9320913, (**B**) rs11584700, (**C**) rs4851266. The plots are centered on the SNPs with the lowest *P* values in the discovery stage (purple diamonds). The R^2 values are from the CEU HapMap 2 samples. The CEU HapMap 2 recombination rates are indicated in blue on the right *y* axes. The figures were created with LocusZoom (http://csg.sph.umich.edu/locuszoom/). Mb, megabases.

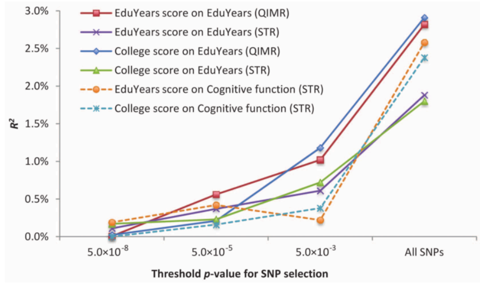


Fig. 2. Explanatory power of the linear polygenic scores estimated for EduYears or College. Solid lines show results from regressions of EduYears on linear polygenic scores in a set of unrelated individuals from the QIMR (*n* = 3526) and STR (*n* = 6770) cohorts. Dashed lines show results from regressions of cognitive function on linear polygenic scores in a sample from STR (*n* = 1419). The scores are constructed from the meta-analysis for either EduYears or College, excluding the cohort (either QIMR or STR) subsequently used as the prediction sample.

Rietveld et al. (2013) Science 340: 1467

- Fig. 1の3つのSNPsは、その後、認識能力などとのアソシエーションでも検出された（Rietveld et al. 2014, PNAS 13790, Ward et al. 2014 PLoS ONE e100248)
- しかし、説明力は非常に低い（R^2 約0.02%）
- 多数の遺伝子に支配され、環境の影響も大きい　（Fig. 2）

最近の遺伝資源は。。。

3K RGP *GigaScience* 2014, 3:7
http://www.gigasciencejournal.com/content/3/1/7

(GIGA)n SCIENCE

DATA NOTE **Open Access**

The 3,000 rice genomes project

The 3,000 rice genomes project[1,2,3*†]

Abstract

Background: Rice, *Oryza sativa* L., is the staple food for half the world's population. By 2030, the production of rice must increase by at least 25% in order to keep up with global population growth and demand. Accelerated genetic gains in rice improvement are needed to mitigate the effects of climate change and loss of arable land, as well as to ensure a stable global food supply.

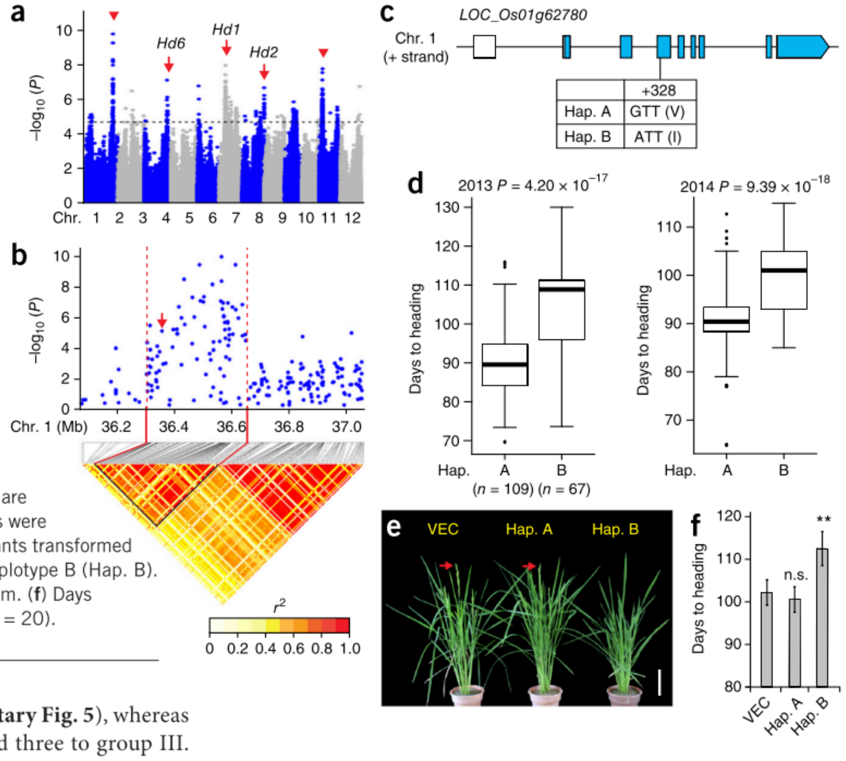
Findings: We resequenced a core collection of 3,000 rice accessions from 89 countries. All 3,000 genomes had an average sequencing depth of 14×, with average genome coverages and mapping rates of 94.0% and 92.5%, respectively. From our sequencing efforts, approximately 18.9 million single nucleotide polymorphisms (SNPs) in rice were discovered when aligned to the reference genome of the temperate *japonica* variety, Nipponbare. Phylogenetic analyses based on SNP data confirmed differentiation of the *O. sativa* gene pool into 5 varietal groups – *indica*, aus/boro, basmati/sadri, tropical *japonica* and temperate *japonica*.

Conclusions: Here, we report an international resequencing effort of 3,000 rice genomes. This data serves as a foundation for large-scale discovery of novel alleles for important rice phenotypes using various bioinformatics and/or genetic approaches. It also serves to understand the genomic diversity within *O. sativa* at a higher level of detail. With the release of the sequencing data, the project calls for the global rice community to take advantage of this data as a foundation for establishing a global, public rice genetic/genomic database and information platform for advancing rice breeding technology for future rice improvement.

Keywords: *Oryza sativa*, Genetic resources, Genome diversity, Sequence variants, Next generation sequencing

新規遺伝子の検出

Figure 2 GWAS for days to heading and identification of the causal gene for the peak on chromosome 1. (**a**) Manhattan plot for days to heading. Dashed line represents the significance threshold ($-\log_{10} P = 4.77$). Arrowheads indicate the position of strong peaks that did not localize with the known *Hd* genes investigated in this study. (**b**) Local Manhattan plot (top) and LD heatmap (bottom) surrounding the peak on chromosome 1. Arrow indicates the position of nucleotide variation in *LOC_Os01g62780*. Dashed lines indicate the candidate region for the peak. (**c**) Exon-intron structure of *LOC_Os01g62780* and DNA polymorphism in that gene. (**d**) Boxplots for days to heading based on the haplotypes (Hap.) for *LOC_Os01g62780* in 2013 (left) and 2014 (right). Box edges represent the 0.25 quantile and the 0.75 quantile with the median values shown by bold lines. Whiskers extend to data no more than 1.5 times the interquartile range, and remaining data are indicated by dots. Differences between the haplotypes were analyzed by Welch's *t*-test. (**e**) Image of transgenic plants transformed with empty vector (VEC), haplotype A (Hap. A) and haplotype B (Hap. B). Red arrows indicate panicle exsertion. Scale bar, 15 cm. (**f**) Days to heading of the transgenic plants. Error bars, s.d. ($n = 20$). $^{**}P < 0.01$; n.s., not significant (Welch's *t*-test).



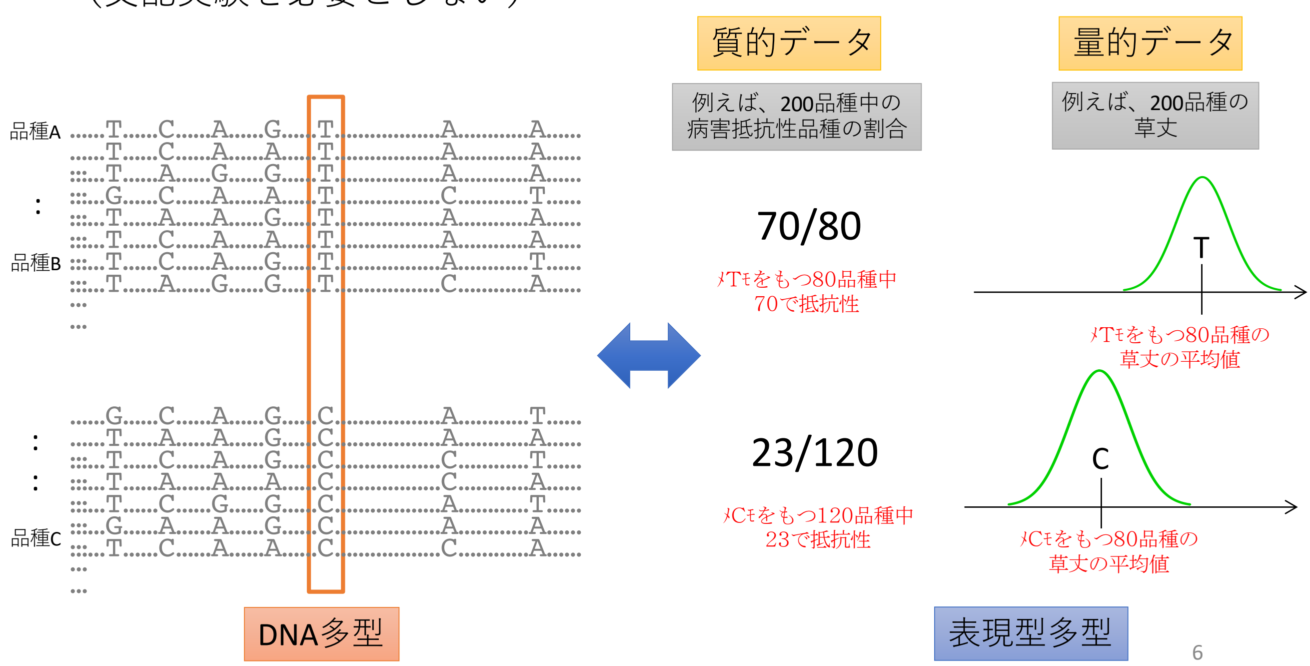
only one polymorphism to group I (**Supplementary Fig. 5**), whereas we assigned six polymorphisms to group II and three to group III.

Yano et al. (2016) Nature Genetics 48: 927

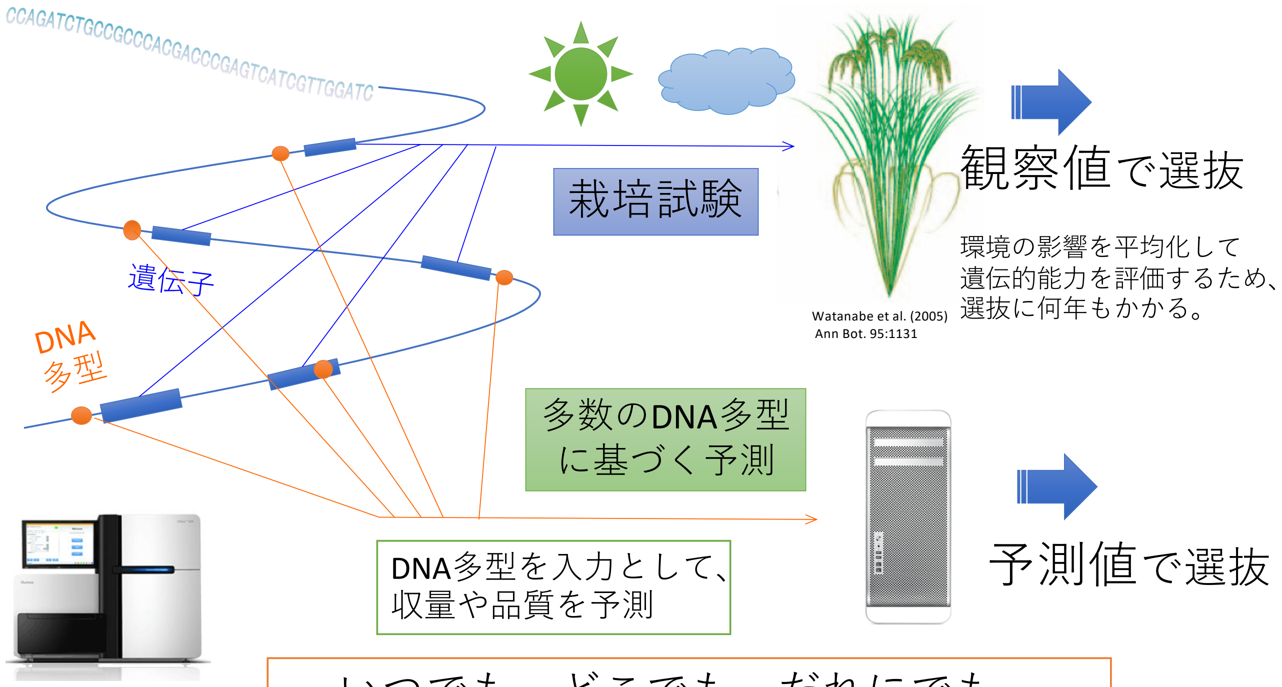
アソシエーション解析 (association analysis)
アソシエーション研究 (association study)

ﾒゲノムワイド～ﾓの場合は、ゲノム全体に分布する多数のDNA多型についてアソシエーションを調べる

- 遺伝資源や育種素材に含まれる品種や系統を用いて、それらで観察されるDNA多型と表現型多型間の関係から、原因遺伝子の検出を試みる方法（交配実験を必要としない）



ゲノミックセレクション



いつでも、どこでも、だれにでも、
栽培試験を行わずに選抜ができる

乳牛育種における成功

"The most dramatic response to genomic selection was observed for the lowly heritable traits DPR, PL, and SCS. Genetic trends changed from close to zero to large and favorable, resulting in rapid genetic improvement in fertility, lifespan, and health in a breed where these traits eroded over time."

Garcia-Ruiz et al.
Proc Natl Acad Sci 113(28): E3995-4004

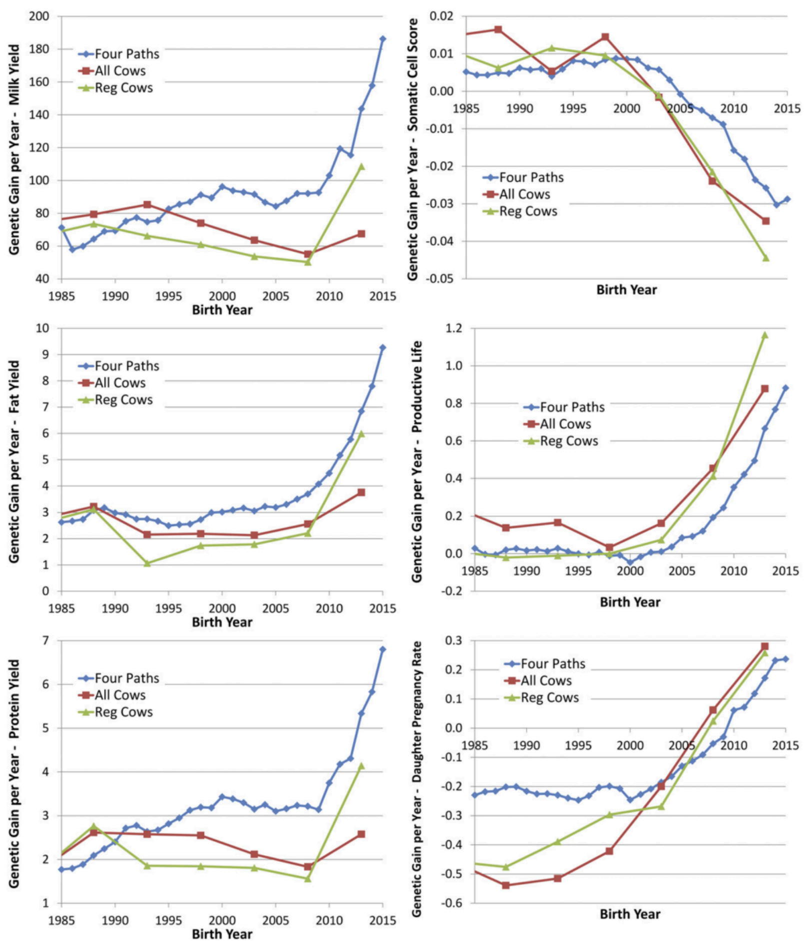


Fig. 3. Genetic gain per year estimates from four paths of selection (Four Paths) and segmented regressions of trait PBV on birth year for all cows (All Cows) or the subset of cows registered in the national herdbook (Reg Cows) for six traits (milk, fat, and protein yields; SCS; PL; and DPR).

ソバのGS実験

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 March 2018
doi: 10.3389/fpls.2018.00276

Check for updates

Potential of Genomic Selection in Mass Selection Breeding of an Allogamous Crop: An Empirical Study to Increase Yield of Common Buckwheat

Shiori Yabe[1†], *Takashi Hara*[2†], *Mariko Ueno*[3], *Hiroyuki Enoki*[4], *Tatsuro Kimura*[4], *Satoru Nishimura*[5], *Yasuo Yasui*[3], *Ryo Ohsawa*[2] and *Hiroyoshi Iwata*[1*]

[1] Graduate School of Agricultural and Life Sciences, University of Tokyo, Tokyo, Japan, [2] Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan, [3] Graduate School of Agriculture, Kyoto University, Kyoto, Japan, [4] Biotechnology and Afforestation Laboratory, Agriculture & Biotechnology Business Division, Toyota Motor Corporation, Miyoshi, Japan, [5] Information System Development Department, X-Frontier Division, Frontier Research Center, Toyota Motor Corporation, Nagoya, Japan

OPEN ACCESS

Edited by:
Bunyamin Tar'an,
University of Saskatchewan, Canada

Reviewed by:
Xuehui Li,
North Dakota State University,
United States
Jessica Elaine Rutkoski,
International Rice Research Institute,
Philippines

***Correspondence:**
Hiroyoshi Iwata
aiwata@mail.ecc.u-tokyo.ac.jp

To evaluate the potential of genomic selection (GS), a selection experiment with GS and phenotypic selection (PS) was performed in an allogamous crop, common buckwheat (*Fagopyrum esculentum* Moench). To indirectly select for seed yield per unit area, which cannot be measured on a single-plant basis, a selection index was constructed from seven agro-morphological traits measurable on a single plant basis. Over 3 years, we performed two GS and one PS cycles per year for improvement in the selection

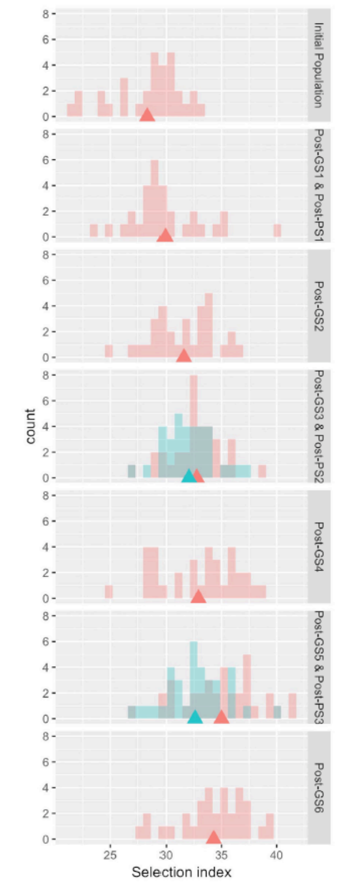


FIGURE 3 | Distribution of the observed values of the selection index in the evaluation of breeding schemes in 2014. Rows 1, 3–7: pink, Initial and Post-GS populations; blue, Post-PS populations. Row 2: both populations are shown in pink. Triangles: population means.

Genomic predictionの遺伝資源スクリーニングへの応用

nature plants

ARTICLES

PUBLISHED: 3 OCTOBER 2016 | ARTICLE NUMBER: 16150 | DOI: 10.1038/NPLANTS.2016.150

Genomic prediction contributing to a promising global strategy to turbocharge gene banks

Xiaoqing Yu[1], Xianran Li[1], Tingting Guo[1], Chengsong Zhu[1], Yuye Wu[2], Sharon E. Mitchell[3], Kraig L. Roozeboom[2], Donghai Wang[2], Ming Li Wang[4], Gary A. Pederson[4], Tesfaye T. Tesso[2], Patrick S. Schnable[1], Rex Bernardo[5] and Jianming Yu[1*]

The 7.4 million plant accessions in gene banks are largely underutilized due to various resource constraints, but current genomic and analytic technologies are enabling us to mine this natural heritage. Here we report a proof-of-concept study to integrate genomic prediction into a broad germplasm evaluation process. First, a set of 962 biomass sorghum accessions were chosen as a reference set by germplasm curators. With high throughput genotyping-by-sequencing (GBS), we genetically characterized this reference set with 340,496 single nucleotide polymorphisms (SNPs). A set of 299 accessions was selected as the training set to represent the overall diversity of the reference set, and we phenotypically characterized the training set for biomass yield and other related traits. Cross-validation with multiple analytical methods using the data of this training set indicated high prediction accuracy for biomass yield. Empirical experiments with a 200-accession validation set chosen from the reference set confirmed high prediction accuracy. The potential to apply the prediction model to broader genetic contexts was also examined with an independent population. Detailed analyses on prediction reliability provided new insights into strategy optimization. The success of this project illustrates that a global, cost-effective strategy may be designed to assess the vast amount of valuable germplasm archived in 1,750 gene banks.

BO for genomic screening of plant germplasm

Theor Appl Genet (2018) 131:93–105
https://doi.org/10.1007/s00122-017-2988-z

CrossMark

ORIGINAL ARTICLE

Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates

Ryokei Tanaka[1] · Hiroyoshi Iwata[1]

Received: 3 May 2017 / Accepted: 19 September 2017 / Published online: 6 October 2017

Abstract

Key message **A new pre-breeding strategy based on an optimization algorithm is proposed and evaluated via simulations. This strategy can find superior genotypes with less phenotyping effort.**

Abstract Genomic prediction is a promising approach to search for superior genotypes among a large number of accessions in germplasm collections preserved in gene

genotype among candidate genotypes and showed that the EI-based strategy required fewer genotypes to identify the best genotype than the usual and random selection strategy. Therefore, Bayesian optimization can be useful for applying genomic prediction to pre-breeding and would reduce the number of phenotyped accessions needed to find the best accession among a large number of candidates.

データ解析のながれ

	SNP 1	SNP 2	SNP 3
Line 1	A	G	?
Line 2	?	T	A
Line 3	T	T	?
Line 4	?	G	C

欠測データの補完

	SNP 1	SNP 2	SNP 3
Line 1	A	G	C
Line 2	T	T	A
Line 3	T	T	A
Line 4	A	G	C

スコア化

	SNP 1	SNP 2	SNP 3
Line 1	−1	−1	1
Line 2	1	1	−1
Line 3	1	1	−1
Line 4	−1	−1	1

GWAS & GSモデリング

集団構造の解析

GWAS

GSモデリング

予測精度評価

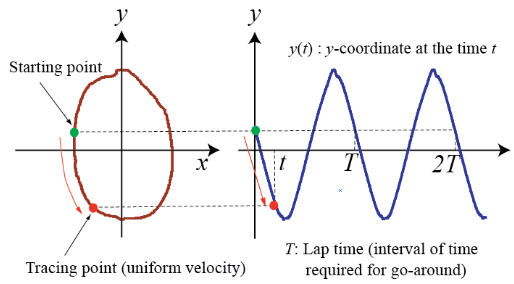
	Trait 1	Trait 2
Line 1	304	13.2
Line 2	449	15.4
Line 3	296	8.3
Line 4	354	6.8

GSモデリングは来週

単回帰による解析

$$y_i = u + \beta_j x_{ij} + e_i$$

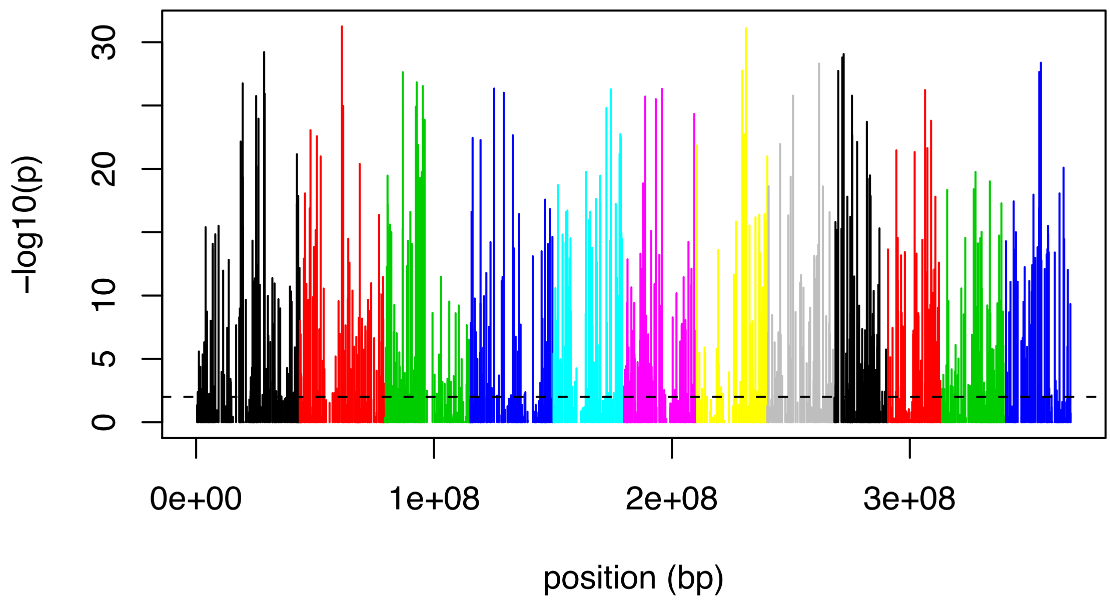
玄米形を
定量化した情報



1311の一塩基多型
（SNPs）

All materials can be
downloaded from
http://ricediversity.org/

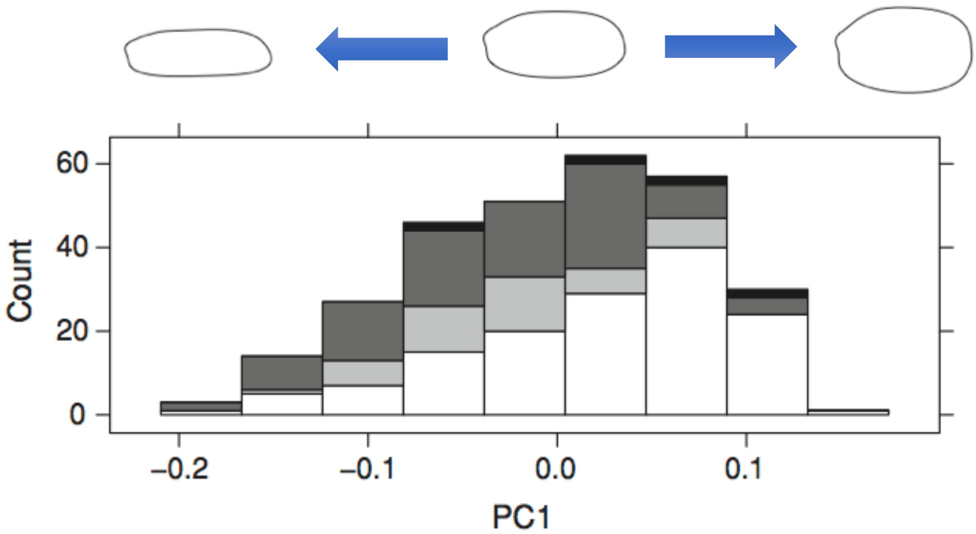
単回帰 結果



ありとあらゆるSNPsが有意に…（まさにマンハッタン？）

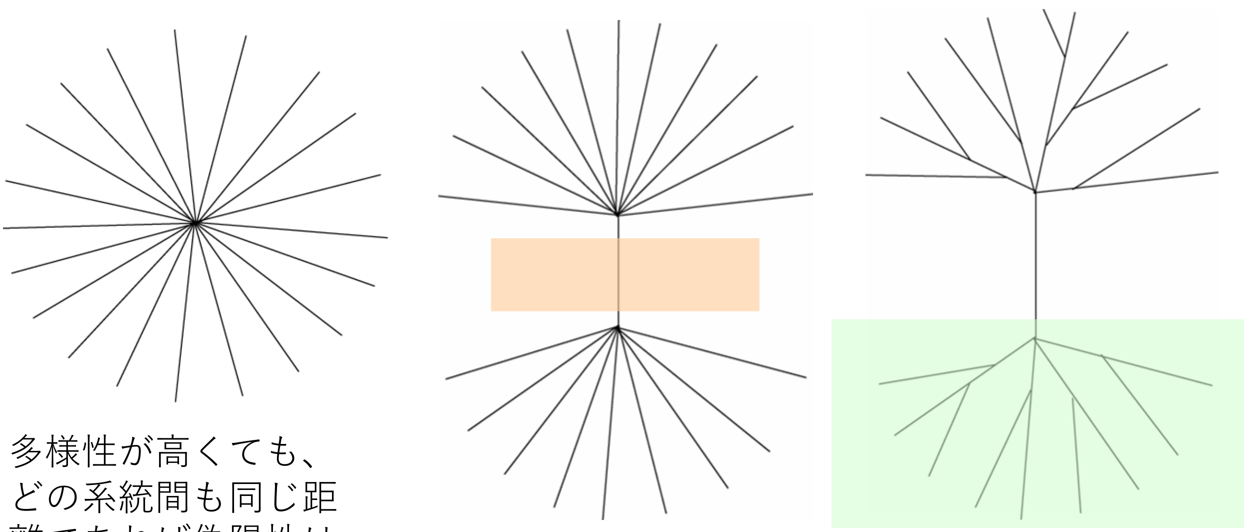
こんなにたくさん遺伝子があるの？

分集団構造と玄米形の関係



- Indica品種（灰色）には細長いタイプが多く、Japonica品種（白）には幅広のタイプが多い

作物にみられる遺伝構造



多様性が高くても、どの系統間も同じ距離であれば偽陽性は生じない。

しかし、実際は…

分集団があったり、

家系構造があったりする

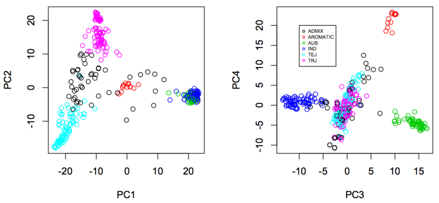
→ 偽陽性を生じる原因となる

分集団と家系構造を考慮した回帰

$$y_i = u + \beta_j x_{ij} + \sum_{k=1}^{K} v_k q_{ik} + \alpha_i + e_i$$

分集団による表現型の違いを吸収させる

血縁関係 **A** をモデルに組み込む



$$\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_\alpha^2)$$

- Yu et al. (2006) Nat. Genet. 38: 203

アソシエーション解析における偽陽性率、偽陰性率、偽発見率

	原因遺伝子として検出されない	**原因遺伝子として検出される**
本当は陰性（原因遺伝子でない）	*A*	*B* (第 1 種の過誤)
本当は陽性（原因遺伝子である）	*C* (第 2 種の過誤)	*D*

偽陽性率（false positive rate: FPR）= B / (A + B)

本当は陰性なのに、陽性（原因遺伝子）として検出される割合

偽陰性率（false negative rate: FNR）= C / (C + D)

本当は陽性（原因遺伝子）なのに、陰性として検出される割合

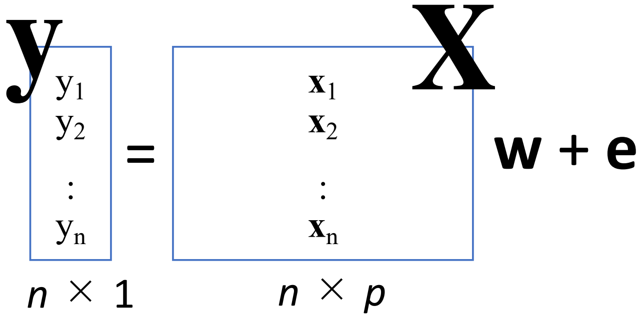
偽発見率（false discovery rate: FDR）= B / (B + D)

陽性（原因遺伝子）として発見されたものの中の、ニセモノ（陰性）の割合

偽陽性率と偽陰性率はトレードオフの関係にある
したがって
偽発見率を一定にコントロールする方法も採られる

large p small n 問題

数千～数万



$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \ (n \times 1), \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \ (n \times p)$$

数百

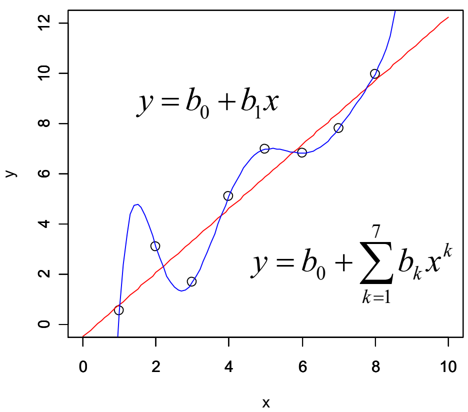
$$p >> n$$

パラメータ数（マーカー）　サンプル数

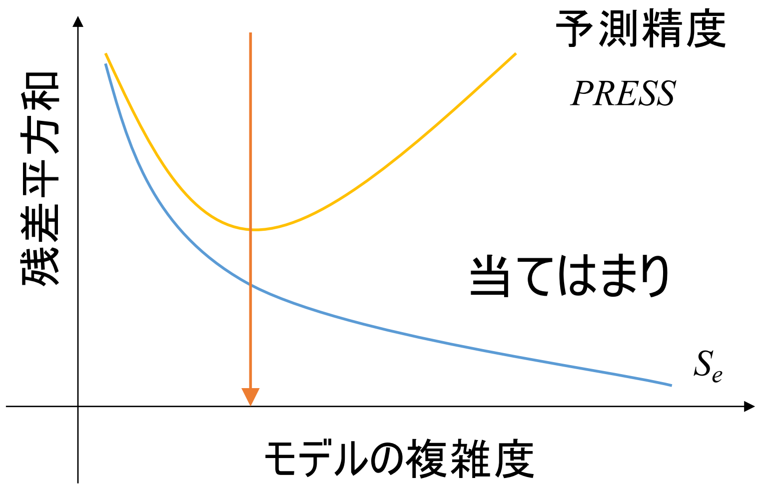
$$\mathbf{w} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

X'Xは特異。重回帰はできない。

当てはまりと予測精度の関係



青線の曲線は明らかに当てはめすぎ！

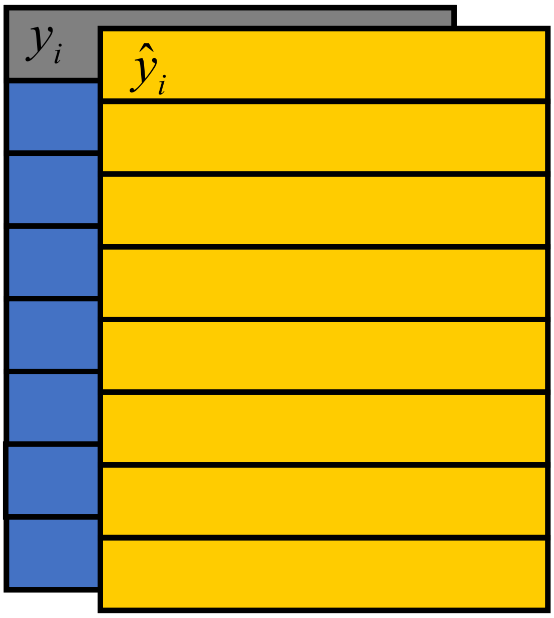


解析中のデータへの当てはまりは、モデルを複雑にすればするほど良くなるが、
未知データに対する予測精度はある複雑度を境に悪化する。

測選抜を行うGSでは、あてはまりの良さではなく、
予測精度の良さを評価することが重要

未知データにおける予測精度の評価 ～交差検証法～

(*n*-fold cross-validation)



1. データをnセットに分割
2. i番目のセットを除いてモデルパラメータを推定
3. i番目のセットについて，2で求めたモデルで予測値 y_i を計算
4. 2, 3をn回繰り返す．
5. 全てのデータについて，予測値 y_i と実測値 y_i を比較して精度を評価する
   精度には予測値と実測値間の相関や、両者の差の2乗和（*PRESS*）などを用いる

nがデータ数のとき，leave-one-out（１個抜き）クロスバリデーションという

リッジ回帰

$$y_i = \sum_j^M x_{ij} w_j + e_i = \mathbf{x}_i^{\mathrm{T}} \mathbf{w} + e_i$$

変数選択によるMLRと異なり、全てのSNPsがモデルに含まれる

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{x}_i^{\mathrm{T}} \mathbf{w})^2 + \lambda \|\mathbf{w}\|^2$$

λが、残差とペナルティのバランスをとる

回帰残差の二乗

ペナルティ

回帰係数が大きくなることにペナルティを課した最小二乗

多数のパラメータが自由に動くのを抑える

$$\|\mathbf{w}\|^2 = \sum_j^M w_j^2$$

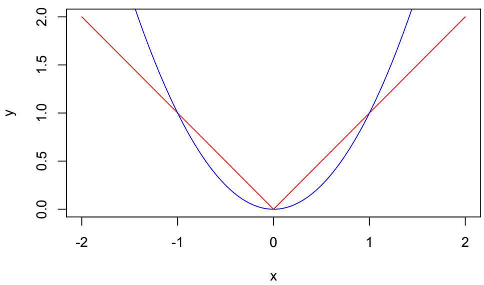
LASSO

$$y_i = \sum_{j}^{M} x_{ij} w_j + e_i = \mathbf{x}_i^{\mathrm{T}} \mathbf{w} + e_i$$

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{x}_i^{\mathrm{T}} \mathbf{w})^2 + \lambda \|\mathbf{w}\|_1$$

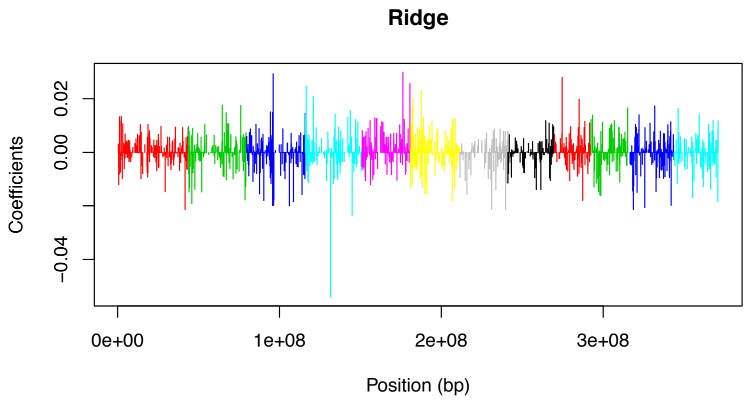
リッジ回帰とよく似るが、絶対値の和に比例したペナルティをかける

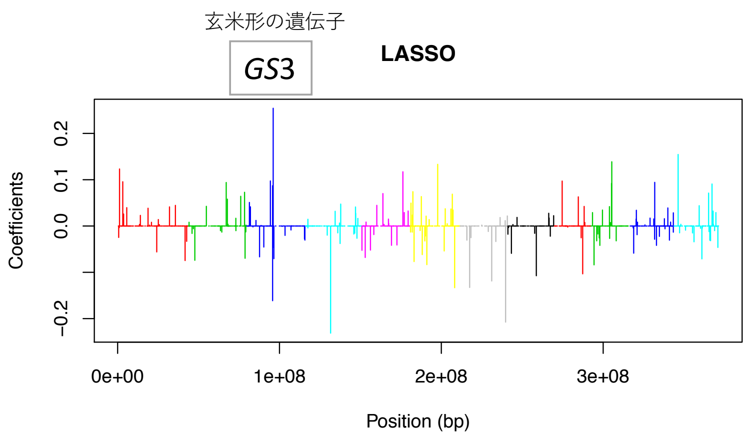
ペナルティ



$$\|\mathbf{w}\|_1 = \sum_{j}^{M} |w_j|$$

ridge, LASSO 回帰係数の比較



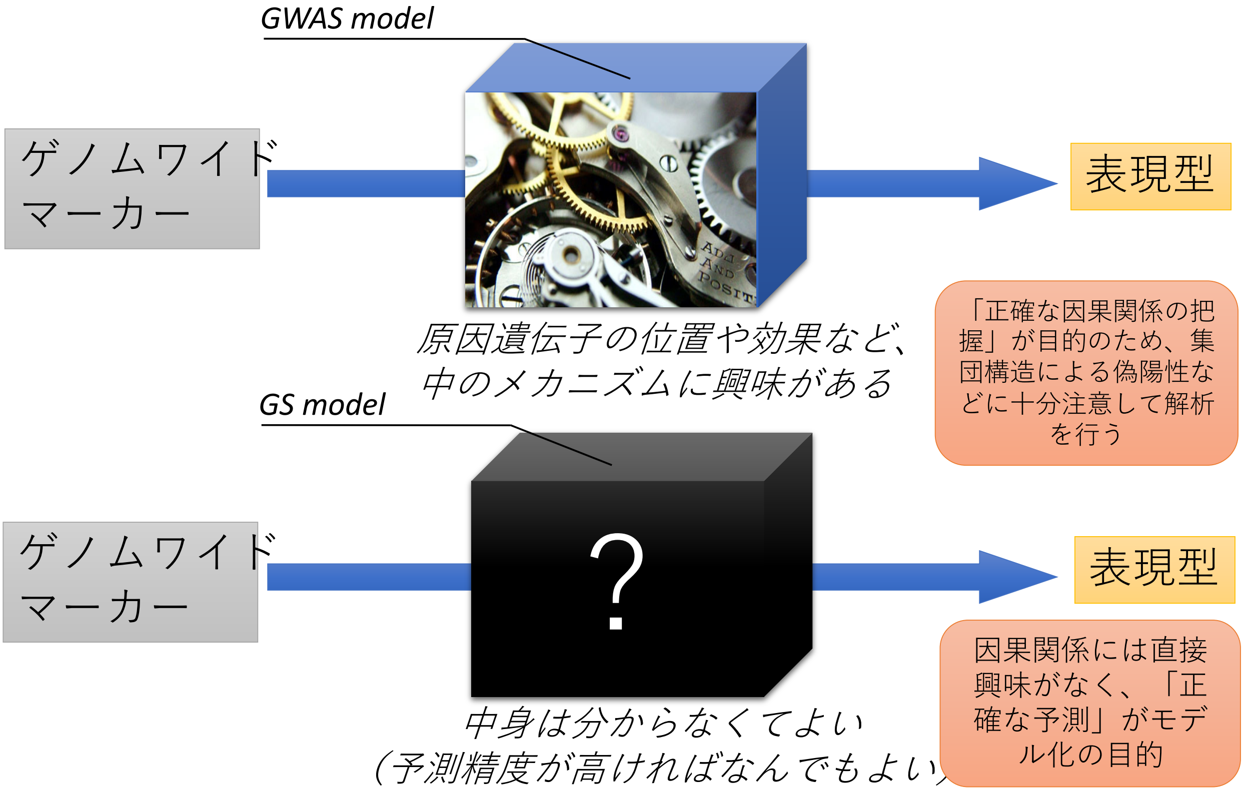


ridge 回帰に比べLASSOはより強く0に収縮（shrink）していることが分かる

また、LASSOの結果をみるとGS3の効果は大きいがその他の領域も予測に利用されている
↓
ゲノムワイドマーカーを利用した予測の必要性を示唆

ridge 回帰は効果の小さな遺伝子がたくさんある場合に、LASSOは比較的効果の大きな遺伝子がある場合に適している

解釈のためのモデル化と
予測のためのモデル化の違い



GS用統計手法とRパッケージ

正則化線形回帰
- ridge regression, LASSO, elastic net
  - glmnet etc.

混合モデル
- BLUP
  - rrBLUP

機械学習
- SVM, RVM （カーネル法）
  - kernlab etc.
- random forest
  - randomForest etc

作物におけるGS手法精度比較に関しては
Zhong et al. (2009) Genetics 182: 355
Crossa et al. (2010) Genetics 186: 713
Iwata and Jannink (2011) Crop Sci 51: 1915
Heffner et al. (2011) Crop Sci 51: 2597
など

ベイズ法
- Bayesian linear regression (Bayesian ridge, Bayesian LASSO)
  - BLR etc
- RKHS regression （カーネル法）
  - RKHSw
    （Crossa et al. (2010) Genetics 186: 713のオンライン資料にRプログラムあり）