Integrating Omics Data into Genomic Prediction

Dissertation

for the Doctoral Degree

at the Faculty of Agricultural Sciences,

Department of Animal Sciences,

Center for Integrated Breeding Research,

Georg-August-University Göttingen

presented by

Zhengcao Li

born in Bai Cheng, Ji Lin, China

Göttingen, May 2019

1st Referee: Prof. Dr. Henner Simianer

Animal Breeding and Genetics Group

Department of Animal Sciences

Georg-August-University Göttingen

2nd Referee: Prof. Dr. Armin Schmitt

Breeding Informatics group

Department of Animal Sciences

Georg-August-University Göttingen

3rd Referee: Prof. Dr. Thomas Kneib

Faculty of Business and Economic Sciences

Georg-August-University Göttingen

Date of disputation: 01, 07, 2019

Table of contents

SUMMARY

Prediction of genetic values plays a central role in quantitative genetics and breeding. Genomic prediction making use of genome-wide single nucleotide polymorphisms (SNPs) was widely adopted to predict breeding values in animal and plant breeding, and to accurately quantify individual disease risk early in human genetics. In the multi-omics era, as omics data (genome, transcriptome, proteome, metabolome, epigenome etc.) increasingly became available during recent years, exploring multi-layer omics data to be predictors in prediction models has been an accessible way to improve predictive abilities in phenotype prediction.

Gene expression profiles potentially hold valuable information for the prediction of breeding values and phenotypes. The *Drosophila melanogaster* Genetic Reference Panel (DGRP) is a community resource for analysis of population genomics and quantitative traits. It consists of more than 200 fully sequenced inbred lines (include 185 lines with whole genome gene expression data) derived from the Raleigh population, USA. In ***Chapter 2***, the utility of transcriptome data for phenotype prediction was tested with 185 inbred lines of *Drosophila melanogaster* for 9 traits in two sexes. In total, 2,863,909 SNPs and 18,140 genome-wide annotated genes and novel transcribed regions (NTRs) were used for all the analyses. We incorporated the transcriptome data into genomic prediction via two kernel methods: GTBLUP and GRBLUP, both combining single nucleotide polymorphisms and transcriptome data. The genotypic data was used to construct the common additive genomic relationship, which was used in genomic best linear unbiased prediction (GBLUP) or jointly in a linear mixed model with a transcriptome-based linear kernel (GTBLUP), or with a transcriptome-based Gaussian kernel (GRBLUP). We studied the predictive ability of the models and discuss a concept of “omics-augmented broad sense heritability” for the multi-omics era. There was one trait (olfactory perceptions to Ethyl Butyrate in females) in which the predictive ability of GRBLUP was significantly higher (0.23) than the predictive ability of GBLUP (0.21). Nonetheless, for most traits, GRBLUP and GBLUP provided similar predictive abilities, while GRBLUP explained more of the phenotypic variance. The better goodness of fit of GRBLUP in general did not translate into a better predictive ability. A possible explanation was suggested that sample size was small and gene expression was not measured at one time point and in one specific

tissue which is functionally linked to the trait of interest.

It is well known that gene expression and regulation may extensively vary among different tissues. However, the transcripts abundance of *Drosophila melanogaster* used was quantified from the entire flies. To test whether tissue-specific transcriptome data can substantially improve predictive abilities, in ***Chapter 3***, we used tissue-specific transcriptome data from the three mice brain tissues: hippocampus (HIP), prefrontal cortex (PFC), and striatum (STR) for phenotype prediction on four novel behavioral traits and four muscle weight traits with low to medium heritability. There were 1063 mice individuals with pedigree information from a multigenerational outbred population which had been sequenced with the reduced-representation genotyping method genotyping-by-sequencing (GBS). After quality control, 523,028 SNPs were used in the analyses. All analyses were conducted in three groups of mice with pedigree, genotype, gene expression and phenotype data, which contained 208 (HIP), 185 (PFC) and 169 (STR) individuals, respectively. The abundances of RNA products from three tissues encompassed 16,533 genes in HIP, 16,249 genes in PFC and 16,860 genes in STR. For the muscle weight traits, the tissue-specific transcriptome data-based prediction (TBLUP) showed high predictive abilities, and the predictive abilities overall were remarkably higher than the pedigree-based prediction (BLUP) and the SNP-based prediction (GBLUP). For the four behavioral traits, the increase of predictive abilities of the transcriptome data-based prediction (TBLUP) were lower than that for the muscle weight traits. When combining transcriptome data with SNPs or pedigree information as predictors, predictive abilities overall were not improved. To study whether the numbers of genes has impact on transcriptome-based prediction, we randomly chose different number of genes for the prediction with TBLUP. The differences among predictive abilities were negligible. Our results suggested that making use of transcriptome data has the potential to improve phenotype predictions if transcriptome data can be sampled in a specific tissue.

In contrast to phenotype prediction, multi-omics data are not ideal candidates for prediction of genetic value and estimation of heritability, since they are not causal variants but intermediate products between causal variants and phenotypes. During the transfer process of genetic information from DNA to phenotype, multi-omics data are inevitably affected by genetic and environmental effects, and the interaction between both. The ‘pan-genome’ denotes the set of

all genes or open reading frames (ORFs) present in the genomes of a group of organisms. Pan-genomic open reading frames potentially carry genome-wide protein-coding genes or causal variant information in a population. The 1002 Yeast Genome project comprised 1,011 *S. cerevisiae* isolates that maximized the breadth of their ecological and geographical origins. In ***Chapter 4***, we used 787 diploid *S. cerevisiae* isolates with 1,625,809 high-quality reference-based SNPs, 7,796 ORFs, copy number of ORFs (CNO) and 35 traits with linear models in the genomic prediction and estimation of heritability. Our results showed that compared to SNP-based genomic prediction (GBLUP), pan-genomic ORF-based genomic prediction (OBLUP) was distinctly more accurate for all the traits, and the predictive abilities were improved by 132% on average across all traits. In addition, the ORF-based heritability can capture more additive effects than SNP-based heritability for all traits. When we combined two subsets of total SNP data (MAF ≥ 0.01 and MAF ≥ 0.05) which contained 311,447 SNPs and 102,253 SNPs, respectively, to pan-genomic ORFs with GOBLUP, the predictive abilities remained the same with OBLUP only using pan-genomic ORFs data. For the second combined method GCBLUP, the predictive abilities remained the same as with CBLUP for all traits, suggesting that ORF data or CNO data covered all causal variant information which SNP data carried. When using three different numbers of isolates in training sets in ORF-based prediction, the predictive abilities of all traits increased as the number of isolates in the training set increased, showing that increasing the training set size could more accurately estimate ORF effects. We demonstrated that pan-genomic ORFs have the potential to be a substitution of single nucleotide polymorphisms in estimation of heritability and genomic prediction under certain conditions. However, in our study there was still a big gap between traits' heritability estimates and prediction accuracy for all the traits. We provide evidence that if larger sample sizes can be used in training set, the prediction accuracy will be further improved.

1st CHAPTER

General introduction

Genomic prediction

Prediction of breeding values has been of central importance in animal breeding. Since best linear unbiased prediction (BLUP) was introduced in animal breeding (Henderson, 1975), it has been a milestone in the development of breeding models. A big advantage of this method is that it can comprehensively utilize pedigree information across many generations to calculate the similarity matrix $\boldsymbol{A}$ between individuals in a population, which has led to genetic gains in most farmed species (Van Vleck et al., 1986; Havenstein et al., 1994). Meuwissen et al. (2001) proposed to use whole genome single nucleotide polymorphisms (SNPs) to replace the traditional prediction of breeding values using pedigree. The concept of “genomic selection” (GS) has revolutionized animal and plant breeding. Implementation of GS became feasible thanks to the large number of SNPs discovered by genome sequencing and new methods to efficiently genotype large number of SNPs. In order to accurately estimate SNP effects, a number of statistical approaches have been proposed such as genomic best linear unbiased prediction (GBLUP), the “Bayesian Alphabet” and Reproducing kernel Hilbert space regression (RKHS).

The BLUP framework model is:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{a} + \boldsymbol{e}$$

where $\boldsymbol{a}$~$N(0, \boldsymbol{A}\sigma_a^2)$ and $\boldsymbol{e}$~ $N(0, \boldsymbol{I}\sigma_e^2)$ are vectors containing random breeding values and residual effects, respectively and where μ is the overall mean.

Based on the BLUP framework, it was suggested that replacing $\boldsymbol{A}$ in BLUP by $\boldsymbol{G}$, a SNP-based relationship matrix, then under centain circumstances the predictive ability could be improved (VanRaden, 2008), and this method was called genomic BLUP (GBLUP).

The $\boldsymbol{G}$ matrix is:

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ\prime}}{2\Sigma p_i(1-p_i)},$$

where p_i denotes the minor allele frequency (MAF) of marker i. Moreover, $\boldsymbol{Z}$ denotes the minor allele frequency adjusted marker matrix with entries $(0-2p_i)$, $(1-2p_i)$ and $(2-2p_i)$ for genotypes AA, Aa and aa, respectively. The better predictive ability of GBLUP was perhaps because markers provided a better representation of genetic relatedness between individuals than a pedigree (Habier et al., 2007). For example, on the basis of pedigree information only, all full-sibs have the same expected relatedness. By contrast, the realized relatedness based on SNP information varies among full-sibs.

Compared to traditional BLUP, GBLUP assumes that all markers have equal effects, in accordance with the infinitesimal model. However, genome wide association studies (GWAS) indicated that this assumption (that the effect of each SNP comes from a normal distribution, with the same variance across all SNP) may not always be reasonable, e.g. some markers may not have effects, and some markers may have relatively big effects. To better accommodate different effects of SNPs in GS, a variety of prior distributions have to be considered. Thus, some methods under the Bayesian framework were applied in practice, called “Bayesian alphabet”(Gianola, 2013). For instance, Bayes A assumes a Student's t distribution of SNP effects, which may have large effects; Bayes B assumes a mixture distribution with a number of SNPs with no effects and a Student's t distribution for the effects of remaining SNPs (Meuwissen et al., 2001).

The general model for Bayes A and Bayes B is:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{a}_m + \boldsymbol{e},$$

where $\boldsymbol{a}_m$ is a m x 1 vector of normally distributed and independent ORF or CNO effects. The variance of the ith ORF effect, σ_{mi}^2 , is assigned a scaled inverted chi-square distribution $\mathcal{X}^{-2}(v,\ S)$.

The mixture distributions in Bayes B is given by

$$\sigma^2_{\mathrm{mi}} \begin{cases} = 0 & \text{with probability } \pi \\ = \mathcal{X}^{-2}(\mathrm{v}, \mathrm{S}) & \text{with probability } (1-\pi) \end{cases}$$

where σ^2_{mi} , the variance of the ith marker effect, is assigned a scaled inverted chi-square distribution $\mathcal{X}^{-2}(v,\ S)$, where S is a scale parameter and is the number of degrees of freedom.

All these methods only capture additive gene effects. However, some evidences proved that there are substantial or extensive epistatic interactions between genes (Huang et al., 2012; Mackay, 2014; Taylor and Ehrenreich, 2014). Reproducing kernel Hilbert space regression (RKHS), a semi-parametric prediction method, was introduced to the field of animal breeding (Gianola et al., 2006). It was promoted as an alternative option to capture the complicated interactions between genes. RKHS regression proceeds by searching a function and uses the residual sum of squares as a loss function, and assigns the squared norm of **g** under a Hilbert space as a penalty. The objective function to be minimized with respect to **g** is:

$$l(\mathbf{g}|\lambda) = ||y - g||^2 + \lambda||\mathbf{g}||^2_H$$

where λ is a regularization parameter and H represents a Hilbert space, very rich class of functions. In ***Chapter 2***, we chose the Gaussian kernel to calculate the genetic covariance between *Drosophila* inbreed lines by

$$K_{ij} = k(r_i, r_j) = exp\left(-\frac{\|r_i - r_j\|^2}{h}\right)$$

Here, h is a bandwidth parameter which controls how fast the covariance function drops as points get further apart as measured by $k(r_i, r_j)$. The vector r_i gives the vector of standardized expression levels of line i across all genes, and r_j is the vector of standardized expression levels of line j across all genes.

Estimation of heritability

The GBLUP method was also proposed for the estimation of the proportion of the phenotypic variance explained by SNP markers (Yang et al., 2010), called SNP-based heritability. Estimation of the variance explained by all common SNPs used in a genome-wide association study (GWAS) was initially motivated by the 'missing heritability' problem: SNPs significantly associated with human height that were discovered by GWASs only explained 5% of phenotypic variance, which is much smaller than the narrow sense heritability (80%) from within family studies (Maher, 2008). Several factors of the missing heritability were provided, including the causal variants each explaining such a small amount of variation that their effects do not reach stringent significance thresholds: rare variants of large effect were not tagged by common SNPs on genotyping arrays, and the pedigree-based narrow sense heritability may include environmental effects (Yang et al., 2017). How much of the proportion of variance explained by SNPs can be attributable to phenotype? Yang et al. (2010) used all common SNPs (defined here as those with minor allele frequency, MAF ≥ 0.01) to quantify SNP-based heritability for human height with unrelated individuals, and demonstrated that common SNPs on a genotyping array explain a large proportion (45%) of variance in height. Given the small $\hat{h}^2_{GWAS}$ (5%) and relatively large $\hat{h}^2_{SNP}$ (45%), it was concluded that, for complex traits like height, there are likely a large number of common variants with too small effect sizes to pass the stringent GWAS threshold ($P < 5 \times 10^{-8}$) in GWAS (Yang et al., 2010).

The SNP-based heritability is defined as

$$\hat{h}_G^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}$$

where $\hat{\sigma}_g^2$ denotes the proportion of the additive genetic variance explained by the common SNPs; $\hat{\sigma}_e^2$ denotes the residuals.

Narrow sense heritability estimates play a key role in predicting or assessing the effectiveness of artificial selection in that they provide a way to measure the extent to which additive genetic variance is related to phenotypic variance in a specific population (Visscher et al., 2008). However, for the prediction of phenotypes, especially when multi-omics data (the transcriptome, proteome, metabolome, epigenome, metagenome etc.) was used, in ***Chapter 1***, we defined a new concept “omics-augmented broad sense heritability” for the prediction of phenotype which not only includes the effects at the genome level (both additive and non-additive), but also includes the effects of downstream biological regulation captured by one or several omics layers (Li et al., 2019).

The omics-augmented broad sense heritability was defined as the proportion of phenotypic variance explained by whole genome SNP marker and other omics data,

$$\hat{H}_o^2 = \frac{\hat{\sigma}_g^2 + \hat{\sigma}_{omics}^2}{\hat{\sigma}_g^2 + \hat{\sigma}_{omics}^2 + \hat{\sigma}_e^2}$$

where $\hat{\sigma}_g^2$ denotes the proportion of additive genetic variance explained by the whole genome SNP markers and $\hat{\sigma}_{omics}^2$ denotes the variances explained by one or several omics data layers which can be the transcriptome, proteome, metabolome, epigenome, metagenome etc.

Gene expression data

The advent of microarrays in the mid-1990s heralded a new era wherein it became possible to measure the abundances of large numbers of transcripts simultaneously (Skelly et al., 2009). It has been demonstrated that there are widespread variations in gene expression levels between individuals within natural populations (Oleksiak et al., 2002; Cheung et al., 2003). The heritability of gene expression variation on a genome-wide scale was first estimated in a cross between a laboratory and a wild strain of *Saccharomyces cerevisiae* (Brem et al., 2002), indicating a substantial genetic component to transcriptional variation in yeast (Skelly et al., 2009). Most quantitative phenotypes have proved to be genetically complex. As intermediate products between DNA and phenotypes, transcript abundances exhibit substantial genetic complexity, despite their close connection to DNA sequence. Several studies have investigated the prevalence of non-additivity, where gene expression in F1 heterozygotes differs from the mid-value of the homozygous parents (Gibson et al., 2004; Vuylsteke et al., 2005; Swanson-Wagner et al., 2006). It has been proven that non-additivity is common in *D. melanogaster* (Huang et al., 2012), *A. thaliana* and maize (Vuylsteke et al., 2005), and that its extreme forms, overdominance and under dominance, are not rare (Gibson et al., 2004). Genetic interactions have been observed in several studies, and a systematic scan for interacting QTLs found non-additive interactions among loci for roughly half of all transcripts (Brem et al., 2005). Gene expression studies have been used for the identification of expression QTL (eQTL) which regulate the transcription levels of individual genes, and gene expression information was used as the phenotype for eQTL mapping based on genetic markers (Brem et al., 2002; West et al., 2007; Nica and Dermitzakis, 2013). In contrast to the major utility of gene expression as phenotypes, several studies in recent years have directly used them as explanatory variables for predicting complex trait phenotypes. The *Drosophila melanogaster* Genetic Reference Panel (DGRP) is a community resource for analysis of population genomics and quantitative traits. It consists of more than 200 fully sequenced inbred lines (including 185 lines with whole genome gene expression data) derived from the Raleigh population, USA. In ***Chapter 2***, we tested the utility of transcriptome data for phenotype prediction with the 185 inbred lines of *Drosophila melanogaster* for 9 traits in two sexes. In total, 2,863,909 SNPs, and 18,140 genome-wide annotated genes and novel transcribed regions (NTRs) were used for all analyses. We constructed a semiparametric prediction model (GRBLUP) with two kernels combining SNP and transcriptome data. The parametric G kernel was used to capture the additive genetic part, and the Gaussian kernel is a non-parametric kernel which was used to pick up non-additive genetic effects and biological regulation effects regardless of the underlying genetic architecture. In our results, GRBLUP and GBLUP provided similar predictive ability, but GRBLUP could capture more phenotypic variance components explained by transcriptome data. The better goodness of fit of GRBLUP in general did not translate into a better predictive ability. It should be noted, though, that sample size was small, and gene expression was not measured at one time point and in one specific tissue that functionally linked to the trait of interest.

The effects of genetic variation on gene expression are condition-dependent, and gene by environment interactions have been shown in comparisons of inbred strains across conditions (Jin et al., 2001; Chen et al., 2005; Whitehead and Crawford, 2005). In multicellular organisms, the local conditions differ in each tissue, and genetic variation with a cell-type dependent influence on gene expression represents a special case of gene-by-environment interaction. Studies of gene expression in mouse brain (Chesler et al., 2005), hematopoietic, stem cells (Bystrykh et al., 2005), fat and liver (Schadt et al., 2003; Yang et al., 2006), and in rat kidney and fat (Hubner et al., 2005), have found that the genetic basis of variation in a gene's expression is sometimes shared between different tissues but is often unique to each tissue (Cotsapas et al., 2006). Studies in flies and mice have also shown extensive sex dependence of gene expression (Wang et al., 2006). To test whether tissue-specific transcriptome data can substantially improve predictive abilities, in ***Chapter 3***, we used tissue-specific transcriptome data from three mice brain tissues: hippocampus, prefrontal cortex, and striatum for phenotype prediction on four novel behavioral traits and four muscle weight traits with low to medium heritability. For the muscle weight traits, the tissue-specific transcriptome data-based prediction (TBLUP) showed high predictive abilities, and the predictive abilities overall were remarkably higher than the pedigree-based prediction (BLUP), and single nucleotide polymorphisms-based prediction (GBLUP). For the four behavioral traits, the increase of predictive abilities of the transcriptome data-based prediction (TBLUP) were lower than that for the muscle weight traits. When combining transcriptome data with SNPs or pedigree information as predictors, predictive abilities overall were not improved.

Pan-genomic open reading frames

Although genome-wide SNPs were the most mainstream data type for genomic selection and estimation of narrow sense heritability, a question arises: are SNPs the ultimate source of genomic data for prediction of genetic value or estimation of heritability?

Pan-genomic open reading frames potentially hold whole-genome protein-coding genes or causal variant information. The ‘pan-genome’ denotes the set of all genes or open reading frames (ORFs) present in the genomes of a group of organisms, usually a species (Lapierre and Gogarten, 2009; Vernikos et al., 2015). There are three subsets within the concept: the core genome that contains genes shared by all individuals within the populations; the dispensable genome made of genes shared by a subset of the individuals and contributes to the species diversity (Tettelin et al., 2005); and individual-specific genes (Vernikos et al., 2015). The concept has been applied to bacterial (Tettelin et al., 2005), viral (Aherfi et al., 2013), plant (Cao et al., 2011; Li et al., 2014; Zhao et al., 2018) , fungal (Dunn et al., 2012), and human genome studies (Sherman et al., 2019). A series of pan-genomic studies were performed when studying genomic dynamics (Donati et al., 2010), pathogenesis and drug resistance (D'Auria et al., 2010; Hu et al., 2011), bacterial toxins (Fang et al., 2011), and species evolution (Konstantinidis et al., 2006). An open reading frame (ORF) is defined as a sequence that has a length divisible by three and is bounded by stop codons. For a particular reading frame, an ORF is a region that is not interrupted by a stop codon. It is a sequence region that is ‘open’ for translation, and an indicator for a potential protein-coding gene (Sieber et al., 2018). One common use of ORFs is as one piece of evidence to assist in gene prediction. Long ORFs are often used along with other evidences to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence (Deonier et al., 2005). The detection of ORFs is of central importance in finding protein-coding genes in genomic sequences. On the other hand, pan-genomic ORFs provide an opportunity to accommodate the phenotypic variation caused by the potential protein-coding sequences in a population. We hypothesize that pan-genomic ORFs can be viewed as a representation of a whole genomic gene set. Directly using this gene set in genomic prediction can capture more genetic variance than SNP-based prediction. There is an increasing understanding that variation in gene presence/absence and copy number of genes play an essential role in the heritability of complex traits. however, there have been no studies utilizing this information in genomic prediction and estimation of heritability (Marroni et al., 2014). In ***Chapter 4***, we used *S. cerevisiae* pan-genomic ORFs which represent 7,796 non-redundant ORFs in genomic prediction, accounting either for the presence/absence of a specific ORF or copy number of ORF (CNO). We exploited a new source of genome-wide potential gene set for genomic prediction and estimation of heritability, and demonstrated (1) genomic prediction using ORF data and CNO data performed better than that using genome-wide SNP data, and (2) the estimation of narrow sense heritability based on pan-genomic ORF data and CNO data can capture parts of the “missing heritability” that appears when using SNP data.

References

Aherfi, S., Pagnier, I., Fournous, G., Raoult, D., La Scola, B., and Colson, P. (2013). Complete genome sequence of Cannes 8 virus, a new member of the proposed family "Marseilleviridae". *Virus Genes* 47(3)**,** 550-555.

Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436(7051)**,** 701-703.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296(5568)**,** 752-755.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., et al. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using'genetical genomics'. *Nature Genetics* 37(3)**,** 225-232.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics* 43(10)**,** 956-963.

Chen, W.J., Chang, S.H., Hudson, M.E., Kwan, W.-K., Li, J., Estes, B., et al. (2005). Contribution of transcriptional regulation to natural variations in Arabidopsis. *Genome Biology* 6(4)**,** R32.

Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics* 37(3)**,** 233-242.

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.-Y., Morley, M., et al. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics* 33(3)**,** 422-425.

Cotsapas, C.J., Williams, R.B., Pulvers, J.N., Nott, D.J., Chan, E.K., Cowley, M.J., et al. (2006). Genetic dissection of gene regulation in multiple mouse tissues. *Mammalian Genome* 17(6)**,** 490-495.

D'Auria, G., Jiménez-Hernández, N., Peris-Bondia, F., Moya, A., and Latorre, A. (2010). Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics* 11(1)**,** 181.

Deonier, R.C., Tavaré, S., and Waterman, M.S. (2005). *Computational genome analysis: an introduction.* Springer-Verlag Berlin, Heidelberg.

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., et al. (2010). Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology* 11(10)**,** R107.

Dunn, B., Richter, C., Kvitek, D.J., Pugh, T., and Sherlock, G. (2012). Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Research* 22, 908–924.

Fang, Y., Li, Z., Liu, J., Shu, C., Wang, X., Zhang, X., et al. (2011). A pangenomic study of Bacillus thuringiensis. *Journal of Genetics and Genomics* 38(12)**,** 567-576.

Georges, M., Charlier, C., and Hayes, B. (2018). Harnessing genomic information for livestock improvement. *Nature Reviews Genetics* 20,135–56.

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 90, 525–540.

Gianola, D., Fernando, R.L., and Stella, A. (2006). Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics* 173,1761–1776.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S., et al. (2004). Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics* 167(4)**,** 1791-1799.

Habier, D., Fernando, R.L., and Dekkers, J.C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177(4)**,** 2389-2397.

Havenstein, G., Ferket, P., Scheideler, S., and Larson, B. (1994). Growth, livability, and feed conversion of 1957 vs 1991 broilers when fed "typical" 1957 and 1991 broiler diets. *Poultry Science* 73(12)**,** 1785-1794.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics***,** 423-447.

Hu, P., Yang, M., Zhang, A., Wu, J., Chen, B., Hua, Y., et al. (2011). Comparative genomics study of multi-drug-resistance mechanisms in the antibiotic-resistant Streptococcus suis R61 strain. *PLoS One* 6(9)**,** e24988.

Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R., Ayroles, J.F., et al. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences* 109(39)**,** 15553-15559.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* 37(3)**,** 243-253.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* 29(4)**,** 389-395.

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 361(1475)**,** 1929-1940.

Lapierre, P., and Gogarten, J.P. (2009). Estimating the size of the bacterial pan-genome. *Trends in genetics* 25(3)**,** 107-110.

Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* 32(10)**,** 1045-1052.

Li, Z., Simianer, H., and Martini, J.W. (2019). Integrating gene expression data into genomic prediction. *Frontiers in Genetics* 10**,** 126-137.

Mackay, T.F. (2014). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics* 15(1)**,** 22-33.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News* 456(7218)**,** 18-21.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature* 542(7640)**,** 186-190.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* 18**,** 31-36.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4)**,** 1819-1829.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368(1620), 20120362.

Oleksiak, M.F., Churchill, G.A., and Crawford, D.L. (2002). Variation in gene expression within and among natural populations. *Nature Genetics* 32(2)**,** 261-266.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422(6929)**,** 297-302.

Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics* 51(1)**,** 30-35.

Sieber, P., Platzer, M., and Schuster, S. (2018). The Definition of Open Reading Frame Revisited. *Trends in Genetics* 34(3)**,** 167-170.

Skelly, D.A., Ronald, J., and Akey, J.M. (2009). Inherited variation in gene expression. *Annual Review of Genomics and Human Genetics* 10**,** 313-332.

Swanson-Wagner, R.A., Jia, Y., DeCook, R., Borsuk, L.A., Nettleton, D., and Schnable, P.S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences* 103(18)**,** 6805-6810.

Taylor, M.B., and Ehrenreich, I.M. (2014). Genetic interactions involving five or more genes contribute to a complex trait in yeast. *PLoS Genetics* 10(5)**,** e1004324.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102(39)**,** 13950-13955.

Van Vleck, L.D., Westell, R., and Schneider, J. (1986). Genetic change in milk yield estimated from simultaneous genetic evaluation of bulls and cows. *Journal of Dairy Science* 69(11)**,** 2963-2965.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91(11)**,** 4414-4423.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics* 9(4)**,** 255-266.

Vuylsteke, M., Van Eeuwijk, F., Van Hummelen, P., Kuiper, M., and Zabeau, M. (2005). Genetic analysis of variation in gene expression in Arabidopsis thaliana. *Genetics* 171(3)**,** 1267-1275.

Wang, S., Yehya, N., Schadt, E.E., Wang, H., Drake, T.A., and Lusis, A.J. (2006). Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics* 2(2)**,** e15.

West, M.A., Kim, K., Kliebenstein, D.J., Van Leeuwen, H., Michelmore, R.W., Doerge, R., et al. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* 175(3)**,** 1441-1450.

Whitehead, A., and Crawford, D.L. (2005). Variation in tissue-specific gene expression among natural populations. *Genome Biology* 6(2)**,** R13.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42(7)**,** 565-569.

Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics* 49(9)**,** 1304-1310.

Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., et al. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Research* 16(8)**,** 995-1004.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics* 50(2)**,** 278-284.

2nd CHAPTER

Integrating gene expression data into genomic prediction

Zhengcao Li[1], Ning Gao[2], Johannes W. R. Martini[3], Henner Simianer[1*]

[1]Animal Breeding and Genetics Group, Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

[2]State Key Laboratory of Biocontrol, School of Life Science, Sun Yat-sen University, North Third Road, Guangzhou Higher Education Mega Center, Guangzhou, China

[3]KWS SAAT SE, Einbeck, Germany.

Keywords: GRBLUP, transcriptome, phenotype prediction, *Drosophila melanogaster,* epistasis

Abstract

Gene expression profiles potentially hold valuable information for the prediction of breeding values and phenotypes. In this study, the utility of transcriptome data for phenotype prediction was tested with 185 inbred lines of *Drosophila melanogaster* for 9 traits in two sexes. We incorporated the transcriptome data into genomic prediction via two methods: GTBLUP and GRBLUP, both combining single nucleotide polymorphisms and transcriptome data. The genotypic data was used to construct the common additive genomic relationship, which was used in genomic best linear unbiased prediction (GBLUP) or jointly in a linear mixed model with a transcriptome-based linear kernel (GTBLUP), or with a transcriptome-based Gaussian kernel (GRBLUP). We studied the predictive ability of the models and discuss a concept of "omics-augmented broad sense heritability" for the multi-omics era. For most traits, GRBLUP and GBLUP provided similar predictive abilities, but GRBLUP explained more of the phenotypic variance. There was only one trait (olfactory perceptions to Ethyl Butyrate in females) in which the predictive ability of GRBLUP (0.23) was significantly higher than the predictive ability of GBLUP (0.21). Our results suggest that accounting for transcriptome data has the potential to improve genomic predictions if transcriptome data can be included on a larger scale.

Introduction

Prediction of genetic values has been a key problem in quantitative genetics. Since Meuwissen et al. (2001) published the landmark article, which uses whole genome single nucleotide polymorphisms (SNPs) to modify the traditional prediction of breeding values using family relationship, the concept of "genomic selection" has revolutionized animal and plant breeding. A number of statistical approaches have been applied in practice, such as genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008), ridge regression (Whittaker et al., 1999), or the "Bayesian Alphabet" (Gianola, 2013, Gianola et al., 2009).

These approaches utilizing genome-wide SNP data have been used to increase the genetic progress of breeding programs by increasing predictive accuracy of breeding values, reducing generation intervals or shortening the breeding cycles. In plant line breeding, genomic prediction focuses on breeding values in early generations of a breeding program, while the genomic prediction of phenotypes may be attractive when estimating the commercial value of cultivars (Crossa et al., 2017). Broad sense heritability is the relevant genetic parameter for phenotypic prediction, which is defined as the ratio of genetic variance over the phenotypic variance. It reflects all genetic contributions to a population's phenotypic variance including additive and non-additive effects such as dominance, and epistasis. It was demonstrated that epistasis explains noticeable fractions of variation in human gene expression (Brown et al., 2014). One of the critically important issues for phenotypic prediction and the estimation of broad sense heritability is how to model non-additive effects. There is plenty of literature illustrating an improved prediction of phenotypes when using non-additive relationships (Crossa et al., 2010, Forsberg et al., 2017, Gao et al., 2017, Martini et al., 2016). However, epistatic effects can arise from various interactions between alleles or genotypes at different loci. For more than two genes, higher order interactions may be included, which makes the estimation of epistatic effects very difficult by using typically parametric regression methods. Another problem for the prediction of phenotypes is that from DNA sequences to phenotypes there are complex biological processes that may affect the phenotypes. Even with complete whole sequence information, genomic prediction may not capture multiple interactions between genes and downstream in the biological regulation. The inclusion of additional layers of omics data in the prediction machinery may provide a partial solution for this problem, since for instance transcriptome data may be “closer” to the phenotype, and since an epistatic interaction on the genotype level may be captured by an additive effect on -for instance- the transcriptome level. In the context of defining the respective broad sense heritability for the combination of genotypic data and omics data, the classical concept only covers the proportion of genetic factors including additive or dominance effects and interactions (Lush, 1940). We discuss the concept of "omics-augmented broad sense heritability" to be used in the context of the prediction of phenotypes not only based on effects at the genome level, but also accounting for effects of downstream biological regulation captured by omics data.

Recently, several studies have proposed to exploit transcriptome data as explanatory variables for prediction of traits. Other than nuclear DNA-based SNP data, gene expression levels are affected by several factors, like choice of tissue, time of sampling and experimental conditions, and using only gene expression data in prediction of phenotypes may not be as robust as using SNP markers. Utilizing both genomic marker information and gene expression data could be a promising option. Modeling gene expression data as a predictor into genomic prediction is expected to explain more epistatic variance or complex biological regulation processes and potentially increases predictive accuracy. González-Reymúndez et al. (2017) integrated whole-omics data (including whole-genome gene expression profiles) into breast cancer prediction, and demonstrated that omics and omic-by-treatment interactions explain a sizable fraction of the variance of survival time, and further suggested that whole-omic profiles could be used to improve prognosis prediction accuracy among breast cancer patients. Guo et al. (2016) showed that gene expression levels provided reduced predictive abilities compared to those based on genetic markers. When combing gene expression data with SNPs, the predictive abilities are either greater than or comparable to those with GBLUP alone. Loh et al. (2011) found when comparing genotype markers to gene expression data to predict soybean plant resistance to the pathogen Phytophthora sojae, using gene expression data performed better than genotype markers. (Zarringhalam et al., 2018) obtained robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. Moreover, different types of omics data have been used for hybrid prediction in Maize

(Schrag et al., 2018, Westhues et al., 2017).

Reproducing kernel Hilbert space regression (RKHS), a semi-parametric prediction method, was introduced by Gianola et al. (2006) to the field of animal breeding. It was promoted as an alternative option to capture the complicated interactions between genes. Jiang and Reif (2015) illustrated that the Gaussian kernel models interaction effects implicitly. More importantly, RKHS provides a simple framework to incorporate information on pedigrees, markers, or any other form of data characterizing the genetic background of individuals (de los Campos et al., 2009). Hu et al. (2015) used RKHS for evaluating the utility of methylation information in prediction of plant height, and demonstrated that epigenetic variation accounted for 65% of the phenotypic variance. In the present study, we used five kernel-based methods: GBLUP, TBLUP, RKHS, GTBLUP and GRBLUP. Genomic best linear unbiased prediction (GBLUP) using SNP data is set to be a benchmark model. TBLUP and RKHS are used for transcriptomic prediction, where the first uses a linear kernel and the latter uses a Gaussian kernel. Moreover, we define GTBLUP (combining GBLUP and TBLUP) and GRBLUP (combining GBLUP and RKHS) utilizing both transcriptome data and whole-genome sequence data.

Drosophila melanogaster is a widely used model organism for biological research in genetics, physiology, microbial pathogenesis, and life history evolution, and it has been demonstrated that the architecture of *Drosophila* quantitative traits is dominated by extensive epistasis (Huang et al., 2012). Making use of *Drosophila* omics data stands a chance to capture the prevalent epistasis for phenotype prediction. The *Drosophila melanogaster* Genetic Reference Panel (DGRP) is a community resource for analysis of population genomics and quantitative traits. It consists of more than 200 fully sequenced inbred lines derived from the Raleigh population, USA (Mackay et al., 2012). We used whole-genome SNP data and gene expression data of 185 Drosophila inbred lines from

DGRP in this study. The objective was (1) to combine transcriptome data with whole-genome sequence data for genomic-transcriptomic prediction using GTBLUP and GRBLUP, (2) to assess whether GTBLUP and GRBLUP can capture substantial proportions of phenotypic variances explained by transcriptome data, and (3) to test whether accounting for transcriptome data can improve phenotype prediction.

Materials and methods

Data

Whole-Genome Sequence data

The *Drosophila melanogaster* Genetic Reference Panel (DGRP) is a community resource for analysis of population genomics and quantitative traits. It consists of 205 fully sequenced inbred lines derived from 20 generations of full sibling inbreeding of a single outbred population in Raleigh, North Carolina, USA (Mackay et al., 2012). Whole genome sequence data of all lines were downloaded from the DGRP2 website. SNPs called with a call rate of less than 95% or minor allele frequency (MAF) smaller than 0.01 and individuals with a call rate less than 95% were excluded. In total, 2,863,909 SNPs of the 185 *Drosophila* lines for which transcriptome data were also available were used for this study. Beagle 4.0 (https://faculty.washington.edu/browning/beagle/b4_0.html) was used for the imputation of missing SNP genotypes (Browning and Browning, 2013).

Transcriptome data

The abundances of RNA products of 18,140 genome-wide annotated genes and novel transcribed regions (NTRs) in 185 DGRP lines was quantified using Affymetrix *Drosophila* 2.0 genome-tiling arrays, with two biological replicates for each sex. Since the correlation coefficient between the two replicates on average across all lines reached 0.95, we randomly chose one replicate for this study. The mated 3- to 5-d-old flies were collected between 1:00 and 3:00 PM, and RNA was extracted from the flies homogenized with 1 mL of QIAzol lysis reagent (Qiagen) and two 0.25-in ceramic beads (MP Biomedical). For details on fly husbandry, RNA extraction, RNA sequence annotation and quality control see (Huang et al., 2015).

Phenotype data

In total, 9 Phenotypes, which were measured on females and males separately were used: startle response (STR), starvation resistance (STV), alcohol sensitivity and tolerance (AST), food intake (FI), and olfactory perceptions to 5 chemical odorants: olfactory perceptions to 2-Heptanone (OP2H), Methyl Salicylate (OPMS), l-Carvone (OPIC), 1-Hexanol (OP1H), Ethyl Butyrate (OPEB). These phenotypes are line means or medians of repeated measurements in different ways, and are treated as response variables in our statistical model. For startle response (starvation resistance), there were on average 40±4 (52±11) measurements for females, and 40±4 (52±11) measurements for males, the line medians were taken in several replicates for each trait (Mackay et al., 2012). The line mean of alcohol sensitivity and tolerance was calculated from two replicated measurements for each sex per line (Morozova et al., 2015). The line mean of food intake was measured from 6 replicate assays per sex per DGRP line (Garlapow et al., 2015). For olfactory perceptions to 5 chemical odorants, the average of 10 measurements was calculated as the response score of each individual trial and the averages of 10 trials on the same genotype and sex were recorded as the line means (Arya et al., 2015).The line means and variances are shown in Table 1.

Availability of Supporting Data

The whole genome sequence data, gene expression data of 185 DGRP lines, and phenotype data of 9 traits are available on *Drosophila melanogaster* Genetic Reference

Panel (DGRP, http://dgrp2.gnets.ncsu.edu).

Statistical models

To remove the gender effect in prediction, we performed the subsequent analyses with female and male data separately. Predictions of phenotypes were done with 3 basic approaches and 2 combined methods. The basic approaches were genomic BLUP (GBLUP) to predict phenotypes using genotype data, transcriptomic BLUP (TBLUP) predicting phenotypes using transcriptome data with a linear kernel, and RKHS predicting phenotypes using transcriptome data with a Gaussian kernel (Gianola and van Kaam, 2008). The combined methods, integrating genomic and transcriptome data, were GTBLUP (combining GBLUP and TBLUP) and GRGLUP (combining GBLUP and RKHS).

GBLUP

As a baseline, we used SNP data of 185 *Drosophila* lines to conduct the benchmark GBLUP (VanRaden, 2008). The statistical model for GBLUP is

$$\boldsymbol{y} = 1\mu + \mathbf{g} + \boldsymbol{e} \qquad (1),$$

where $\mathbf{g}{\sim}N(0, \boldsymbol{G}\sigma_g^2)$ and $\boldsymbol{e}{\sim}\, N(0, \boldsymbol{I}\sigma_e^2)$ are vectors containing random breeding values and residual effects, respectively and where μ is the overall mean. The genomic relationship matrix $\boldsymbol{G}$ was calculated as $\boldsymbol{G} = \frac{\boldsymbol{ZZ}\prime}{2\Sigma p_i(1-p_i)}$ (VanRaden, 2008), where p_i denotes the minor allele frequency (MAF) of marker i. Moreover, $\boldsymbol{Z}$ denotes the MAF adjusted marker matrix with entries $(0-2p_i)$ and $(2-2p_i)$ for genotypes AA and aa, respectively.

TBLUP

In this approach, transcriptome data of the 185 *Drosophila* lines were used as predictor variables. The statistic model is:

$$\boldsymbol{y} = 1\mu + \boldsymbol{t} + \boldsymbol{e} \qquad (2)$$

where $\boldsymbol{t} \sim N(0, \boldsymbol{E}\sigma_t^2)$ is a transcriptomic line effect. The corresponding variance-covariance matrix is $\boldsymbol{E} = \boldsymbol{RR}'$ which is a linear kernel calculated from an n x m matrix $\boldsymbol{R}$ of standardized gene expression levels from n lines and m genes. The standardization of gene expression levels was conducted by calculating $r_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$, where x_{ij} is the expression level of gene j in line i, $\bar{x}_j$ is the mean expression level of gene j across all lines, and s_j is the standard deviation of gene expression level of gene j.

Reproducing Kernel Hilbert Space Regression (RKHS)

Analogously, to the previously described approaches, the statistical model was:

$$\boldsymbol{y} = 1\mu + \boldsymbol{v} + \boldsymbol{e} \qquad (3)$$

where $\boldsymbol{v} \sim N(0, \boldsymbol{K}\sigma_v^2)$ is a random effect measured by transcriptome data with $\boldsymbol{K}$ being the genetic covariance matrix (Gianola et al., 2006). We chose the Gaussian kernel to calculate the genetic covariance between lines by

$$K_{ij} = k(r_i, r_j) = exp\left(-\frac{\|r_i - r_j\|^2}{h}\right) \qquad (4),$$

Here, h is a bandwidth parameter, which controls how fast the covariance function drops as points get further apart. The vector r_i gives the vector of standardized expression levels of line i across all genes, and r_j is the vector of standardized expression levels of line j across all genes. The bandwidth parameter h was chosen using a grid search approach under cross-validation, aiming at finding a suitable value that maximized the predictive correlation within a model setting (Gianola and Schön, 2016, Jones et al., 1996).

GTBLUP

In GTBLUP, transcriptome data was integrated into genomic prediction. SNP data and transcriptome data of 185 *Drosophila* lines were treated as predictor variables. The prediction model was:

$$\boldsymbol{y} = 1\mu + \mathbf{g} + \boldsymbol{t} + \boldsymbol{e} \qquad (5),$$

where all variables are defined as described above.

GRBLUP

The statistical model for GRBLUP can be expressed as

$$\boldsymbol{y} = 1\mu + \mathbf{g} + \boldsymbol{v} + \boldsymbol{e} \qquad (6).$$

The only difference between GTBLUP and GRBLUP is that in GRBLUP we replace $\boldsymbol{t}$~ $N(0, \boldsymbol{E}\sigma_t^2)$ of GTBLUP with $\boldsymbol{v}$~$N(0, \boldsymbol{K}\sigma_v^2)$ of RKHS. Again $\boldsymbol{K}$ is the genetic covariance matrix constructed by the Gaussian kernel (4) and the optimum bandwidth parameter h is found by grid-search and cross-validation.

Estimation of the omics-augmented broad sense heritability based on the between line

effects

The omics-augmented broad sense heritability was defined as the proportion of phenotypic variance explained by whole genome SNP marker and other omics data,

$$\hat{H}_o^2 = \frac{\hat{\sigma}_g^2+\hat{\sigma}_{omics}^2}{\hat{\sigma}_g^2+\hat{\sigma}_{omics}^2+\hat{\sigma}_e^2} \qquad (7)$$

where $\hat{\sigma}_g^2$ denotes the proportion of additive genetic variance explained by the whole genome SNP markers and $\hat{\sigma}_{omics}^2$ denotes the variances explained by one or several omics data layers which can be the transcriptome, proteome, metabolome, epigenome, metagenome etc.

(1) SNP-based genomic narrow sense heritability for GBLUP ($\hat{h}_G^2$)

The SNP-based genomic narrow sense heritability is defined as the proportion of phenotypic variance explained by SNP marker effects. This SNP-based heritability is calculated as

$$\hat{h}_G^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2+\hat{\sigma}_e^2} \qquad (8) .$$

(2) SNP and gene expression data-augmented broad sense heritability for GTBLUP ($\hat{H}_{GT}^2$) and GRBLUP ($\hat{H}_{GR}^2$)

The proportion of phenotypic variance explained by SNP data and gene expression data in GTBLUP ($\hat{H}_{GT}^2$) is calculated as

$$\hat{H}_{GT}^2 = \frac{\hat{\sigma}_g^2+\hat{\sigma}_t^2}{\hat{\sigma}_g^2+\hat{\sigma}_t^2+\hat{\sigma}_e^2} \qquad (9).$$

and in GRBLUP ($\hat{H}_{GR}^2$) are calculated as

$$\widehat{H}^2_{GR} = \frac{\widehat{\sigma}^2_g + \widehat{\sigma}^2_v}{\widehat{\sigma}^2_g + \widehat{\sigma}^2_v + \widehat{\sigma}^2_e} \qquad (10).$$

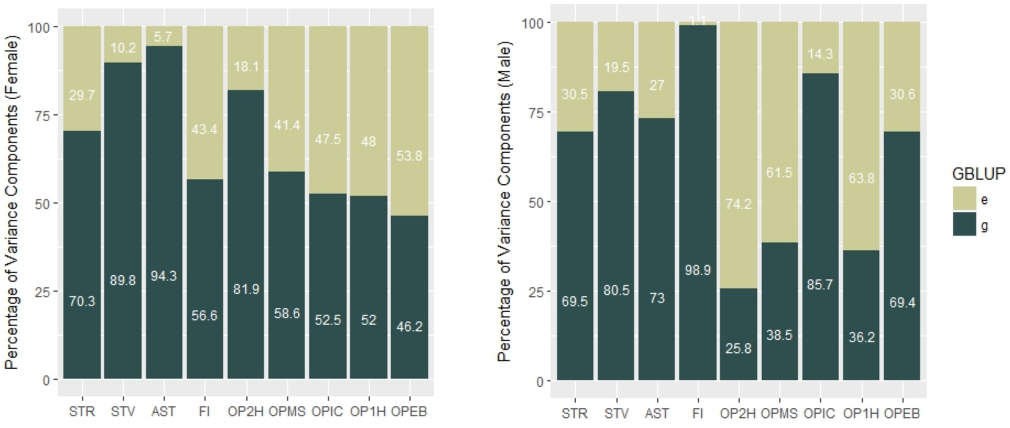
The variance components $\widehat{\sigma}^2_g$, $\widehat{\sigma}^2_t$, $\widehat{\sigma}^2_v$, $\widehat{\sigma}^2_e$ from models (1), (5), and (6) were estimated from the entire data sets, using the R package "regress" (Clifford and McCullagh, 2014), which also provided predictions of random effects.

Comparison of predictive abilities

The different approaches were assessed using 20 replicates of a 5-fold cross-validation (Erbe et al., 2013). Predictive abilities were defined as the Pearson's correlation coefficients between predicted genetic values and observed phenotypes in the test sets. The final predictive ability of each model was the mean of the predictive abilities across 100 estimates. Overall predictive abilities among the five models implemented in the study were compared using a Tukey's honest significant difference test (Tukey, 1949).

Results

Estimation of "omics-augmented broad sense heritability" based on the between line effects and variance components



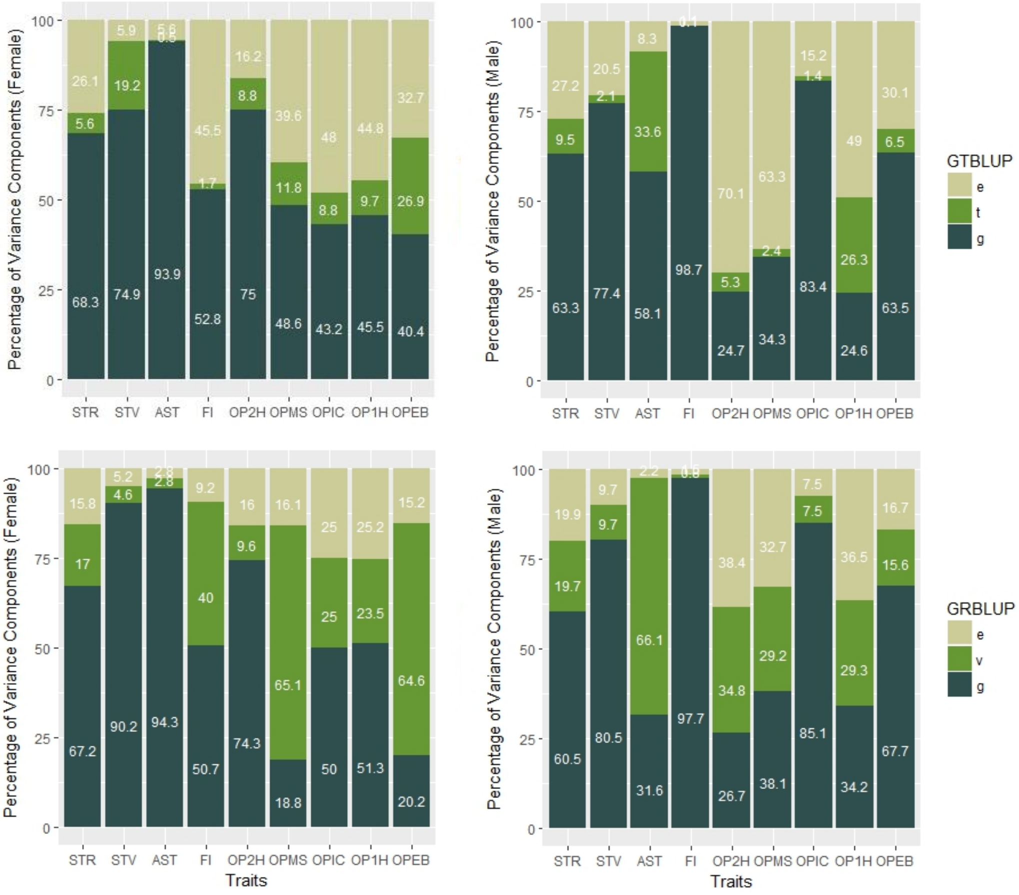


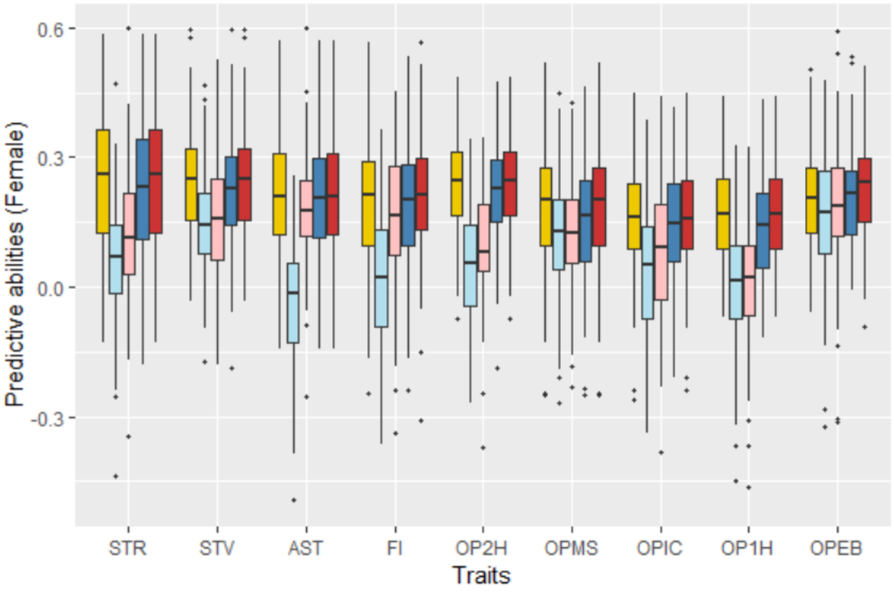
Figure 1: Percentages of variance components of GBLUP, GTBLUP and GRBLUP for 9 traits for females (left) and males (right). e is the residual; t is the transcriptomic line effect in GTBLUP; v is the transcriptomic line effect in GRBLUP, and g is the additive genetic effect captured by SNP data.

Genomic heritabilities obtained with model (1) ranged from 0.25 to 0.99 and are generally high. On average across all traits, they are slightly higher for females ($\hat{h}^2_{Gf} = 0.66 \pm 0.059$) than for males ($\hat{h}^2_{Gm} = 0.63 \pm 0.081$) (see Fig. 1 and Table 1). It should be noted, though, that these values pertain to the average performance of many replications of inbred individuals, and thus should not be compared to narrow sense heritability estimates on an individual base.

In GTBLUP and GRBLUP, we integrated transcriptome data into genomic prediction. The only difference between these two methods is that two different kernels were used to construct the relationship matrix based on transcriptome data. For the SNP and gene expression data-augmented heritability, $\widehat{H}^2_{GR}$ was higher than $\widehat{H}^2_{GT}$ for almost all traits and in both sexes (Table 1). Only the trait FI did not show this pattern for males. Across all traits, $\widehat{H}^2_{GR}$ had a mean of 0.85±0.050 for females and 0.81± 0.080 for males compared to $\widehat{H}^2_{GT}$ 0.71±0.025 for females, and 0.69±0.049 for males. Compared to GTBLUP, GRBLUP captured more genetic variance explained by gene expression data for some traits, especially for some traits with relatively low SNP-based genomic heritability h^2_G, such as FI, OPMS, OPIC, OP1H, and OPEB in females and AST, OP2H, OPMS, and OP2H in males.

Overall predictive ability

The predictive abilities of the 9 traits obtained with the 5 statistical models for females and males are shown in Figure 2 and Supplementary Table 1.



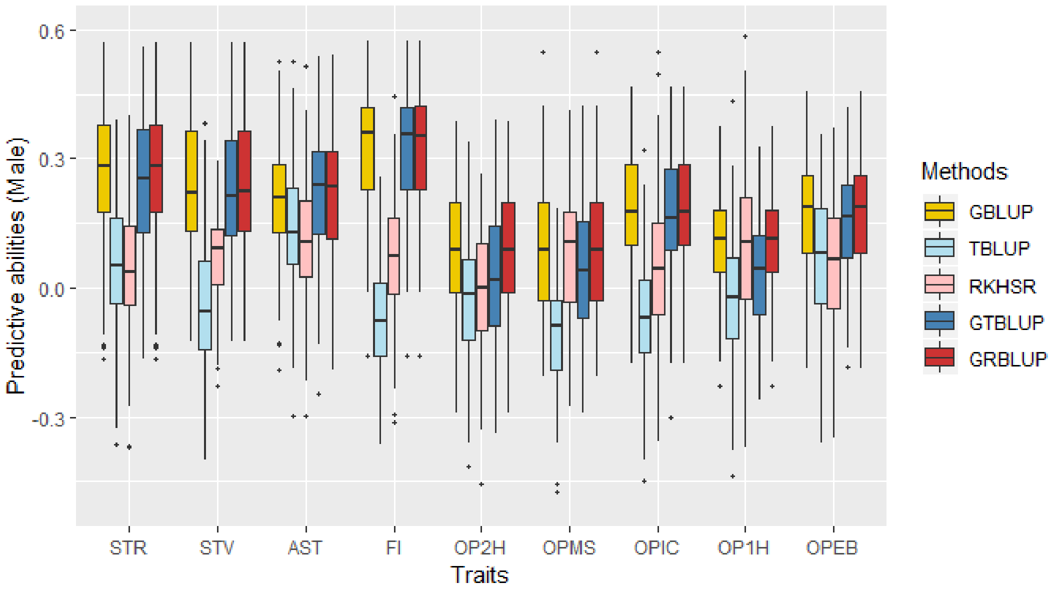


Figure 2: Predictive abilities for 9 traits with 5 statistical models in females and males.

GBLUP as the reference method provided predictive abilities ranging from 0.162 ± 0.012 to 0.240 ± 0.013 in females and from 0.095 ± 0.015 to 0.325 ± 0.013 in males across all traits. For GBLUP, the proportion of phenotypic variance explained by SNP data and genomic predictive abilities were highly positively correlated. The correlation coefficients were 0.731 and 0.885 for females and males, respectively. Transcriptome-based prediction alone was not accurate for most traits: observed predictive abilities were 0.001 ± 0.013 to 0.182 ± 0.011 for females, and 0.036 ± 0.014 to 0.107 ± 0.014 for males with RKHS and -0.035 ± 0.011 to 0.165±0.014 for females and -0.113±0.013 to 0.13±0.015 for males with TBLUP. The correlation between female and male predictive abilities with RKHS and TBLUP were low with correlation coefficients of 0.077 and -0.189 respectively.

Except for one trait (OPEB) in females, there was no significant difference of predictive abilities between GRBLUP and GBLUP. For the trait OPEB in female, GRBLUP (0.23 ± 0.012) gave a higher predictive ability than GBLUP (0.208 ± 0.012). Both GRBLUP (female 0.21, male 0.187) and GBLUP (female 0.205, male 0.184) provided better predictive abilities on average in all traits than GTBLUP (female 0.187, male 0.156) for female and male. It is worth noting that predictive abilities between males and females for all models were found to be remarkably different for 6 out of 9 traits (AST, FI, OP2H, OPMS, OPIC, OP1H). In females, the predictive abilities of three models (GBLUP, GTBLUP and GRBLUP) varied slightly among all 9 traits with a range between 0.139 ± 0.012 (OP1H in GTBLUP) and 0.24 ± 0.013 (STV in GRBLUP), while in males the predictive abilities of these three models have a more significant variation ranging from 0.045 ± 0.014 (OPMS in GTBLUP) to 0.326 ± 0.014 (FI in GRBLUP). The correlation coefficient between predictive abilities in females and males across all traits and models is 0.623 (Fig. 3).

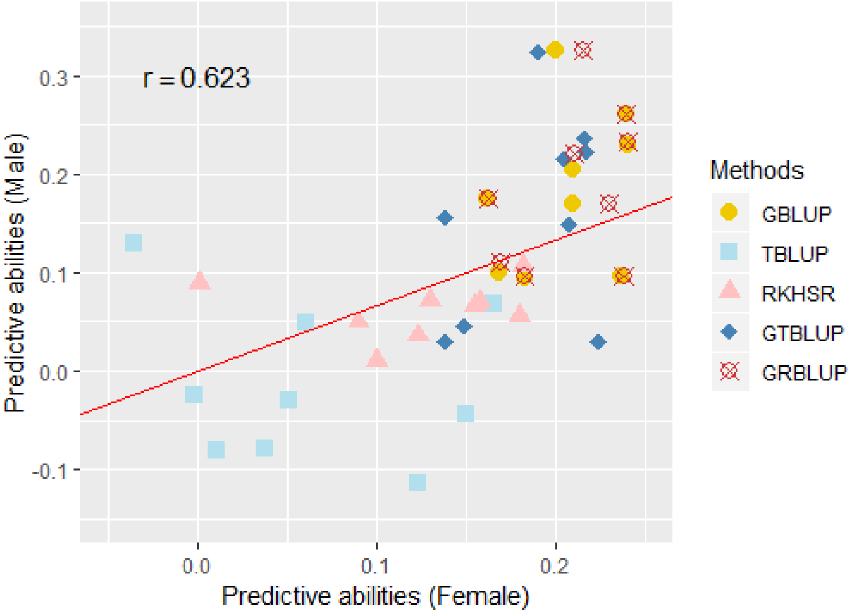


Figure 3: The correlation between predictive abilities in females and males across 9 traits and 5 statistical models. r denotes the Pearson correlation coefficient between female and male predictive abilities across all traits and all statistical models. The red line denotes a standardized major axis regression line.

The correlation coefficients between heritabilities $\hat{h}^2_G$, $\hat{H}^2_{GT}$, $\hat{H}^2_{GR}$ and predictive abilities for GBLUP, GTBLUP, GRBLUP across all traits and both sexes are 0.823, 0.821 and 0.832 respectively (Fig. 4). The bandwidth parameter h in the Gaussian kernel varied dramatically from 0.7 to 270'000, and choosing the right value had great impact on predictive abilities of RKHS and GRBLUP.

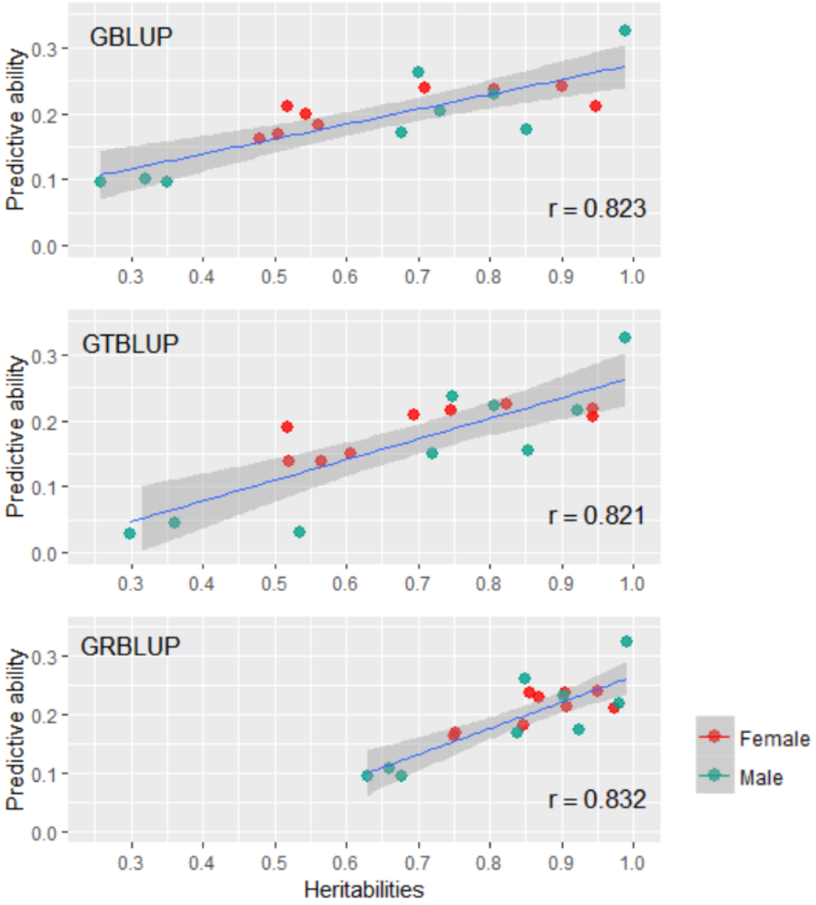


Figure 4: The correlation between heritabilities $\hat{h}^2_G$, $\hat{H}^2_{GT}$, $\hat{H}^2_{GR}$ and predictive abilities for GBLUP, GTBLUP, and GRBLUP across all traits and both sexes. r denotes the Pearson correlation coefficient. The blue lines denote the overall linear regression lines. The grey shadow denotes the 0.95 confidence interval.

Discussion

Previous *Drosophila* genomic prediction studies have shown that there is a high degree of genotype by sex interaction in some traits. Ober et al. (2012) showed that given the significant sex by line interaction variance in starvation resistance, the prediction is more accurate in females than in males (0.254 vs. 0.203), and in chill coma recovery time the predictive ability is very low for female and zero for male. It has also been found that 42% of the *Drosophila* transcriptome is genetically variable between males and females, including the novel transcribed regions (NTRs) (Huang et al., 2015). We also found expression patterns to be clearly separated between males and females (see Supplementary Figure 1) and thus we performed all analyses on females and males separately in order to remove the gender effect in prediction.

Omics-augmented broad sense heritability

Yang et al. (2010) showed that 45% of the variance for human height can be explained by considering all SNPs simultaneously when using GBLUP to estimate the narrow sense heritability, the proportion of phenotypic variance due to additive genetic variance. Two explanations for the “missing heritability” were provided: (1) the causal variants each explain such a small amount of variation that their effects do not reach stringent significance thresholds, or (2) the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Speed et al. (2012) argued that GBLUP may not be capable to provide unbiased estimates of the genomic heritability, and a main reason is that in the computation of the G matrix the LD between SNPs and QTL is ignored. Kim et al. (2017) proposed that the main problem of estimating genomic heritability does not reside in the manner the G matrix is computed, but rather in the use of massive numbers of markers that are in LD with QTL. Since there is probably no complete linkage disequilibrium between SNPs and all causal variants, which e.g. also can be structural variants, using SNP data may not provide accurate estimates of narrow sense heritability. Narrow sense heritability estimates play a key role in predicting or assessing the effectiveness of artificial selection in that they provide a way to measure the extent to which additive genetic variance is related to phenotypic variance in a specific population (Visscher et al., 2008). However, for the prediction of phenotypes, the concept of broad sense heritability is more useful than the concept of narrow sense heritability, because it reflects all the genetic contributions to a population's phenotypic variance including additive and non-additive effects, which provides upper limits to estimates of transmissible genetic variance (Lush, 1940, Stoltenberg, 1997). Nevertheless, as mentioned, even if all SNPs were used, only part of the genetic effects can be captured. The inclusion of additional layers of genomic information in the prediction machinery may provide a partial solution for this problem. When DNA information is transcribed into RNA and then expressed as protein products, abundance of gene expression products is one of the intermediate layers in this process. We assume that the missing additive variance in estimation of narrow sense heritability by using SNP data, and some non-additive effects may be captured by the gene expression data. In this case, utilizing both SNP data and gene expression data to estimate broad sense heritability can be a promising approach. The classical definition of broad sense heritability is the ratio of genetic variance to the phenotypic variance, which implicitly assumes that all genetic variation must be encoded at the genome level. However, gene expression data may be inevitably affected by some external regulation which belongs to environment effects in terms of the classical genetic model, where the phenotype is considered to be affected by genetic and environmental effects, and the interaction between both. In the multi-omics era, the input information for the phenotypic prediction machinery is not restricted to gene or genome layer. Multi-omics data reflecting the transcriptome, proteome, metabolome, epigenome, metagenome etc. are increasingly exploited as input data for the phenotypic prediction (Acharjee et al., 2016, Xu et al., 2016). Thus, we discuss the concept “omics-augmented broad sense heritability” for the prediction of phenotype which not only includes the effects at the genome level (both additive and non-additive), but also includes the effects of downstream biological regulation captured by one or several omics layers. In phenotype prediction this concept can help to measure the extent to which the information in the different layers of multi-omics data is related to phenotypic variance in a specific population. For some traits substantially affected by non-additivity and downstream biological regulation effects, or with poor LD between SNPs and QTL, the estimated genomic heritabilities may be low so that they may be inadequate as a measure of predictive ability. In this case the omics-based broad sense heritability may be more informative than narrow or broad sense heritability because of the inclusion of non-additive effects and biological regulation effects in the numerator of $\widehat{H}_o^2$, and it can be seen as the potential upper limit of the predictive ability of phenotypic prediction when utilizing multi-omics data. This method was used to measure the increased heritabilities of 11 traits when incorporating gene expression and metabolic data into phenotypic prediction in maize, however, without discussing the reasonability (Guo et al., 2016). It must be highlighted that the “omics - augmented broad sense heritability” is just available in the context of phenotype prediction, while in the genomic prediction for breeding values this concept is of limited usefulness because the biological regulation variance in the numerator of $\widehat{H}_o^2$ is not fully heritable. The approach should be seen as a complement or partial substitution to the classical narrow sense heritability when using multi-omics data to predict phenotypes.

Assessment of predictive abilities

Due to the transmission of genetic information from DNA sequence to transcripts, information at the gene expression layer (transcriptome) is “closer” to phenotypes than genomic information, and thus should help providing better predictions of phenotypes than genomic information. However, unlike the DNA sequence, the transcriptome information is not stably inherited and measurements of transcriptome abundance are affected by choice of tissue, time of sampling and experimental conditions. In this study, predictive abilities of RKHS obtained on 9 traits were relatively low (0.001 to 0.182 in female, 0.036 to 0.107 in male), and were much lower than predictive abilities obtained with GBLUP using SNP data. A similar result was also shown in maize, where predictive abilities of transcriptomic prediction were always lower than the genomic prediction when comparing both using eight statistical models (Xu et al., 2017). RKHS and GRBLUP performed significantly better than TBLUP and GTBLUP, indicating that RKHS with a Gaussian should be preferred when conducting transcriptome-based prediction.

For GBLUP, we found predictive ability and the phenotypic variance component explained by SNP data to be highly positively correlated with correlation coefficients of 0.73 and 0.89 for females and males respectively. However, the phenotypic variance explained by SNP data was exceedingly high (> 0.8) for some traits, such as STV, AST, OP2H in females and STV, AST, FI, OPIC in males, while the predictive abilities for these traits were relatively low. The reason could be the small sample size of lines and this result was consistent with the previous study for starvation resistance and startle response which the predictive abilities were 0.239 ± 0.012 and 0.23 ± 0.012 respectively. Ober et al. (2012) showed that the predictive ability could reach 0.58 if the number of sequenced lines for training was increased to 1000 (Ober et al., 2012).

We incorporated transcriptome data with genomic prediction using GRBLUP which combine the standard GBLUP and the RKHS method. From an RKHS point of view, the genomic relationship matrix G in GBLUP can be viewed as a parametric kernel that only captures genetic values based on an additive genetic relationship among individuals. The Gaussian kernel is a non-parametric kernel which may pick up genetic signals regardless of the underlying genetic architecture. Choosing the most suitable bandwidth parameter h can provide an optimal $\frac{\sigma_k^2}{\sigma_k^2+\sigma_e^2}$ ratio, which gives an appropriate weight to the phenotypic variance explained by transcriptome data, leading to an optimized predictive performance. GRBLUP can be considered as a case of RKHS with two kernels. For the comparison between GTBLUP and GRBLUP, the only difference between these two methods is that two different kernels were used to construct a relationship matrix based on transcriptome data. In GTBLUP, we replaced the Gaussian kernel used in GRBLUP with a linear kernel. Compared with GBLUP, the SNP and gene expression data-based broad sense heritability H_{GT}^2 of GTBLUP was higher than the SNP-based genomic heritability h_G^2 of GBLUP at all 9 traits in both male and female, but GTBLUP slightly decreased the combined predictive ability for most traits. This result suggests that there may be an overfitting problem when using GTBLUP to model the combined data. Xu et al. (2017) observed an analogical result which decreased the predictive ability when combining transcriptome data and metabolic data into genomic prediction for six yield-related traits in maize (Xu et al., 2017). Compared to GTBLUP, GRBLUP captured more genetic variance explained by gene expression data for some traits, especially for traits with relatively lower genomic heritability h_G^2 in GBLUP, such as FI, OPMS, OPIC, OP1H, OPEB in female; and AST, OP2H, OPMS, OP2H in male. For the omics-based broad sense heritability based on the between line effects, $\widehat{H}_{GR}^2$ was higher than $\widehat{H}_{GT}^2$ for all 9 traits in both males and females, and GRBLUP provided a superior predictive ability than GTBLUP across all traits. This demonstrated that the Gaussian kernel is superior to the linear kernel $E = RR^T$ for modeling transcriptome data in genomic prediction.

In our result, there was only one trait (OPEB in females) for which the predictive ability of GRBLUP (0.23) was higher than the predictive ability of GBLUP (0.21). This indicated that predictive ability can be improved when combining transcripts with SNPs using GRBLUP, but it depends on the traits. For the rest of the traits for both males and females, the SNP and gene expression data-based heritability H_{GR}^2 was remarkably increased compared to the SNP-based heritability h_G^2 of GBLUP. However, there is no significant difference in predictive ability between GRBLUP and GBLUP, which might be caused by the small sample size and may be changing with increased sample sizes.

Conclusion

We constructed a semiparametric prediction model (GRBLUP) with two kernels combining SNP and transcriptome data. The parametric G kernel was used to capture the additive genetic part, and the Gaussian kernel is a non-parametric kernel which was used to pick up non-additive genetic effects and biological regulation effects regardless of the underlying genetic architecture. In our study, GRBLUP and GBLUP provided similar predictive ability, but GRBLUP could capture more phenotypic variance components explained by transcriptome data. The better goodness of fit of GRBLUP in general did not translate into a better predictive ability. It should be noted, though, that sample size was small and gene expression was not measured at one time point and in one specific tissue functionally linked to the trait of interest. However, including transcriptomic data can increase predictive ability, as was shown for the trait OLED in females. We conclude that adding more specifically collected transcriptome data has the potential to improve genomic predictions in larger scale applications.

Author contributions

All authors were involved in the design of the study. ZCL performed the model validations and wrote the manuscript. NG and JWRM participated in discussing the statistical models. All authors commented on the manuscript and read and approved the final version.

Conflict of Interest Statement

All authors declare to have no competing interests.

Acknowledgments

ZCL thanks China Scholarship Council for financial support.

References

Acharjee, A., Kloosterman, B., Visser, R.G., and Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics* 17(5)**,** 180-191.

Arya, G.H., Magwire, M.M., Huang, W., Serrano-Negron, Y.L., Mackay, T.F., and Anholt, R.R. (2015). The genetic basis for variation in olfactory behavior in Drosophila melanogaster. *Chemical Senses* 40(4)**,** 233-243.

Brown, A.A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J.B., et al. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* 3**,** e01381.

Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity by descent detection in population data. *Genetics* 194(2), 459-471.

Clifford, D., and McCullagh, P. (2014). The regress package. *R News* 6**,** 6.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science* 22(11)**,** 961-975.

de los Campos, G., Gianola, D., and Rosa, G.J. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* 87(6)**,** 1883-1887.

Erbe, M., Gredler, B., Seefried, F.R., Bapst, B., and Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS One* 8(12)**,** e81046.

Forsberg, S.K., Bloom, J.S., Sadhu, M.J., Kruglyak, L., and Carlborg, Ö. (2017). Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature Genetics* 49(4)**,** 497-503.

Gao, N., Martini, J.W., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., et al. (2017). Incorporating Gene Annotation into Genomic Predictionof Complex Phenotypes. *Genetics* 207, 489–501.

Garlapow, M.E., Huang, W., Yarboro, M.T., Peterson, K.R., and Mackay, T.F. (2015). Quantitative genetics of food intake in Drosophila melanogaster. *PLoS One* 10(9)**,** e0138129..

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194(3)**,** 573-596.

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183(1)**,** 347-363.

Gianola, D., Fernando, R.L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173(3)**,** 1761-1776.

Gianola, D., and Schön, C.-C. (2016). Cross-validation without doing cross-validation in genome-enabled prediction. *G3: Genes, Genomes, Genetics* 6(10)**,** 3107-3128.

Gianola, D., and van Kaam, J.B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178(4)**,** 2289-2303. doi: 10.1534/genetics.107.084285.

González-Reymúndez, A., de los Campos, G., Gutiérrez, L., Lunt, S.Y., and Vazquez, A.I. (2017). Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *European Journal of Human Genetics* 25(5)**,** 538-544. doi: 10.1038/ejhg.2017.12.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics* 129(12)**,** 2413-2427. doi: 10.1007/s00122-016-2780-5.

Hu, Y., Morota, G., Rosa, G.J., and Gianola, D. (2015). Prediction of plant height in Arabidopsis thaliana using DNA methylation data. *Genetics* 201(2)**,** 779-793.

Huang, W., Carbone, M.A., Magwire, M.M., Peiffer, J.A., Lyman, R.F., Stone, E.A., et al. (2015). Genetic basis of transcriptome diversity in Drosophila melanogaster. *Proceedings of the National Academy of Sciences* 112(44)**,** E6010-E6019. doi: 10.1073/pnas.1519159112.

Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R., Ayroles, J.F., et al. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences* 109(39)**,** 15553-15559. doi: 10.1073/pnas.1213423109.

Jones, M.C., Marron, J.S., and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 91(433)**,** 401-407. doi: Doi 10.2307/2291420.

Kang, T., Ding, W., Zhang, L., Ziemek, D., and Zarringhalam, K. (2017). A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. *BMC Bioinformatics* 18(1)**,** 565-576.

Kim, H., Grueneberg, A., Vazquez, A.I., Hsu, S., and de los Campos, G. (2017). Will big data close the missing heritability gap? *Genetics* 207(3)**,** 1135-1145.

Loh, P.-R., Tucker, G., and Berger, B. (2011). Phenotype prediction using regularized regression on genetic data in the DREAM5 Systems Genetics B Challenge. *PloS One* 6(12)**,** e29095.

Lush, J.L. (1940). Intra-sire correlations or regressions of offspring on dam as a method of estimating heritability of characteristics. *Proceedings of the American Society of Animal Nutrition* 1940(1)**,** 293-301.

Mackay, T.F., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., et al. (2012). The Drosophila melanogaster genetic reference panel. *Nature* 482(7384)**,** 173-178. doi: 10.1038/nature10811.

Martini, J.W., Wimmer, V., Erbe, M., and Simianer, H. (2016). Epistasis and covariance: how gene interaction translates into genomic relationship. *Theoretical and Applied Genetics* 129(5)**,** 963-976.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4)**,** 1819-1829.

Morozova, T.V., Huang, W., Pray, V.A., Whitham, T., Anholt, R.R., and Mackay, T.F. (2015). Polymorphisms in early neurodevelopmental genes affect natural variation in alcohol sensitivity in adult drosophila. *BMC Genomics* 16(1)**,** 865-881. doi: 10.1186/s12864-015-2064-5.

Ober, U., Ayroles, J.F., Stone, E.A., Richards, S., Zhu, D., Gibbs, R.A., et al. (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. *PLoS Genetics* 8(5)**,** e1002685. doi: ARTN e1002685 10.1371/journal.pgen.1002685.

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., et al. (2018). Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics* 208, 1373–1385..

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics* 91(6)**,** 1011-1021. doi: 10.1016/j.ajhg.2012.10.010.

Stoltenberg, S.F. (1997). Coming to terms with heritability. *Genetica* 99(2-3)**,** 89-96.

Tukey, J.W. (1949). Comparing individual means in the analysis of variance. *Biometrics***,** 99-114.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91(11)**,** 4414-4423. doi: 10.3168/jds.2007-0980.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics* 9(4)**,** 255-266.

Westhues, M., Schrag, T.A., Heuer, C., Thaller, G., Utz, H.F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theoretical and Applied Genetics* 130(9)**,** 1927-1939.

Whittaker, J., Thompson, R., and Denham, M. (1999). Marker-assisted selection using ridge regression. *Annals of Human Genetics* 63(4)**,** 366-366.

Xu, S., Xu, Y., Gong, L., and Zhang, Q. (2016). Metabolomic prediction of yield in hybrid rice. *The Plant Journal* 88(2)**,** 219-227. doi: 10.1111/tpj.13242.

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity* 119(3)**,** 174-184. doi: 10.1038/hdy.2017.27.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42(7)**,** 565-569. doi: 10.1038/ng.608.

Zarringhalam, K., Degras, D., Brockel, C., and Ziemek, D. (2018). Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. *Scientific Reports* 8(1)**,** 1237-1247.

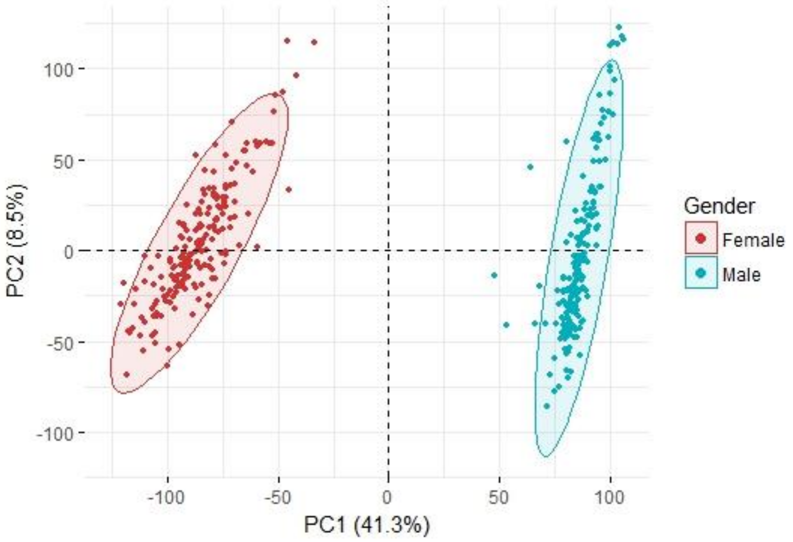
Table 1: Line means (M) and variances (V) of phenotypes and heritability estimates for the 9 traits in males and females. $\hat{h}_G^2$ denotes the SNP-based genomic heritability calculated with GBLUP; $\hat{H}_{GT}^2$ denotes the SNP and gene expression data-based broad sense heritability calculated with GTBLUP; $\hat{H}_{GR}^2$ denotes the SNP and gene expression data-based broad sense heritability calculated with GRBLUP. r denotes the phenotypic correlation between female and male phenotypes across lines.

	Female					**Male**					
Traits	M	V	$\hat{h}_G^2$	$\hat{H}_{GT}^2$	$\hat{H}_{GR}^2$	M	V	$\hat{h}_G^2$	$\hat{H}_{GT}^2$	$\hat{H}_{GR}^2$	r
STR	28.75±0.44	40.29	0.703	0.739	0.842	28.29±0.50	41.22	0.701	0.749	0.801	0.958
STV	60.61±0.89	159.06	0.898	0.943	0.948	45.65±0.67	90.39	0.805	0.807	0.903	0.684
AST	17.36±0.28	14.03	0.943	0.944	0.972	16.49±0.24	10.45	0.730	0.923	0.978	0.685
FI	0.99±0.04	0.36	0.566	0.545	0.908	1.02±0.05	0.50	0.989	0.988	0.980	0.674
OP2H	3.10±0.04	0.28	0.819	0.823	0.840	3.04±0.04	0.28	0.258	0.299	0.616	0.760
OPMS	3.40±0.03	0.15	0.586	0.605	0.839	3.32±0.03	0.17	0.385	0.361	0.673	0.582
OPIC	3.50±0.03	0.20	0.525	0.520	0.750	3.39±0.03	0.21	0.851	0.853	0.925	0.697
OP1H	2.30±0.04	0.28	0.520	0.565	0.748	2.34±0.04	0.28	0.362	0.536	0.635	0.794
OPEB	3.51±0.03	0.18	0.462	0.673	0.848	3.57±0.03	0.16	0.694	0.719	0.833	0.594

supplementary material

Supplementary table 1: Empirical prediction accuracy ± standard deviation of 9 traits in 5 statistical models for both females and males.

	Female					**Male**				
Traits	GBLUP	TBLUP	RKHS	GTBLUP	GRBLUP	GBLUP	TBLUP	RKHS	GTBLUP	GRBLUP
STR	0.239 ± 0.015	0.061±0.013	0.123 ± 0.014	0.216 ± 0.014	0.239 ±0 .015	0.261 ± 0.015	0.049±0.014	0.036 ± 0.014	0.237 ± 0.016	0.261 ± 0.015
STV	0.240 ± 0.013	0.150±0.014	0.155 ± 0.013	0.217 ± 0.013	0.240 ± 0.013	0.230 ± 0.014	-0.044±0.013	0.067 ± 0.011	0.222 ± 0.014	0.233 ± 0.014
AST	0.210 ± 0.013	-0.035±0.011	0.182 ± 0.011	0.205 ± 0.013	0.211 ± 0.013	0.204 ± 0.014	0.130±0.015	0.107 ± 0.014	0.215 ± 0.015	0.220 ± 0.015
FI	0.200 ± 0.015	0.011±0.015	0.158 ± 0.016	0.190 ± 0.015	0.215 ± 0.014	0.325 ± 0.013	-0.081±0.015	0.070 ± 0.013	0.324 ± 0.013	0.326 ± 0.014
OP2H	0.237 ± 0.012	0.051±0.012	0.100 ± 0.011	0.224 ± 0.012	0.238 ± 0.012	0.096 ± 0.014	-0.030±0.013	0.010 ± 0.013	0.029 ± 0.016	0.096 ± 0.014
OPMS	0.183 ± 0.015	0.123±0.015	0.130 ± 0.013	0.149 ± 0.013	0.183 ± 0.015	0.095 ± 0.015	-0.113±0.013	0.072 ± 0.014	0.045 ± 0.014	0.096 ± 0.015
OPIC	0.162 ± 0.012	0.038±0.015	0.090 ± 0.015	0.139 ± 0.013	0.163 ± 0.012	0.175 ± 0.014	-0.078±0.011	0.050 ± 0.015	0.155 ± 0.015	0.175 ± 0.014
OP1H	0.168 ± 0.011	-0.002±0.013	0.001 ± 0.013	0.139 ± 0.012	0.169 ± 0.011	0.100 ± 0.012	-0.025±0.011	0.090 ± 0.015	0.030 ± 0.012	0.110 ± 0.012
OPEB	0.210 ± 0.012	0.165±0.014	0.180 ± 0.014	0.208 ± 0.011	0.230 ± 0.012	0.170 ± 0.013	0.068±0.015	0.056 ± 0.014	0.149 ± 0.012	0.170 ± 0.013



Supplementary Figure 1: PC analysis of female (red) and male lines (green) for gene expression data. The variances explained by PC 1 (x-axis) and PC 2 (y-axis) are shown in the respective captions.

3rd CHAPTER

Utilizing tissue-specific gene expression data for phenotype prediction in mice

Zhengcao Li[1], Henner Simianer[1*]

[1]Animal Breeding and Genetics Group, Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

Keywords:

Phenotype prediction, tissue-specific transcriptome data, mice, behavioral traits, muscle weight traits

Abstract

Transcriptome potentially holds valuable information for the prediction of phenotypes. However, gene expression and regulation may extensively vary among different tissues. In this study, the effectiveness of tissue-specific transcriptome data from three mice brain tissues: hippocampus (HIP), prefrontal cortex (PFC), and striatum (STR) was tested for phenotype prediction on four novel behavioral traits and four muscle weight traits with low to medium heritability. The abundances of RNA products from three tissues encompassed 16,533 genes in HIP, 16,249 genes in PFC, and 16,860 genes in STR. For the muscle weight traits, the tissue-specific transcriptome data-based prediction (TBLUP) showed a high level of statistical robustness, and the predictive abilities overall were remarkably higher than the pedigree-based prediction (BLUP), and single nucleotide polymorphisms based genomic prediction (GBLUP). For the four behavioral traits, the improvement of predictive abilities with TBLUP was lower than that for the muscle weight traits. When different numbers of genes were randomly chosen for prediction with TBLUP, the differences among predictive abilities were negligible. Combining transcriptome data with SNPs or pedigree information as predictors did not improve predictive abilities. Our results suggest that inclusion of transcriptome data has the potential to improve phenotype predictions if transcriptome data can be sampled in a specifically relevant tissue.

Introduction

Genomic selection (GS) making use of genome-wide single nucleotide polymorphisms (SNPs) has been widely adopted to replace the pedigree-based prediction of breeding values using in animal and plant breeding (Meuwissen et al., 2001). GS is a form of marker-assisted selection which can improve the breeding progress by increasing predictive accuracy of breeding values, or reducing generation intervals (Schaeffer, 2006). In plant line breeding, genomic prediction mainly focuses on breeding values in early generations of a breeding program, while the genomic prediction of phenotypes may be attractive when estimating the commercial value of cultivars (Crossa et al., 2017). Likewise, in human genetics, phenotype prediction aims at accurately quantifying disease risk so that preventative measures may be taken earlier, (Abraham and Inouye, 2015). However, under certain circumstances, the predictive ability of SNP-based phenotype prediction often remains low. e.g., for some traits with low heritability in humans, such as psychiatric illnesses (Bouchard Jr, 2004), reproductive fitness traits (Kosova et al., 2010), tinnitus (Kvestad et al., 2010), or behavioral problems (Pappa et al., 2015). For some traits in livestock, such as litter weight gain in pigs (Thekkoot et al., 2016) or lamb survival in sheep (Hatcher et al., 2010), SNP-based phenotype prediction also has limited predictive abilities. Although some studies illustrated an improved prediction of phenotypes when using linear or non-linear kernels to model epistatic effects (Su et al., 2012; Vitezica et al., 2013; Akdemir and Jannink, 2015; Jiang and Reif, 2015), the improvement was still inappreciable especially for accurate prediction of human disease risk which requires more precision than prediction in livestock and crops (Wray et al., 2013b).

On the other hand, sample size has a significant impact on prediction accuracy (Kim et al., 2017). It has been shown in a dairy cattle application, that the predictive ability increases as the number individuals in the training set increases (Erbe et al., 2013). Furthermore, for prediction of distantly related individuals, the prediction accuracy is lower than prediction of closely related individuals, because the extent of LD between SNP and causal variants depends on the relatedness of the sample of individuals used (Wray et al., 2013a). If closely related individuals are included in the sample, long-range LD is generated even between SNPs and QTLs on different chromosomes (Wray et al., 2013b).

Another problem for phenotype prediction is that there are complex biological processes from DNA sequences to observable phenotypes. Only using information from the genome level may not capture such complex downstream effects, which often encompass linear or non-linear interactions between different genetic and regulatory complexes (Mackay et al., 2009). The inclusion of transcriptome data in the prediction model may provide a partial solution for this problem, since transcriptome data may be “closer” to the phenotype, and causal variants influence phenotypes by causing variation in protein sequence and/or the abundance of transcripts. Variation in the transcripts abundance has significant impact on quantitative traits (Mackay et al., 2009).

Recently, transcriptome data have been employed for prediction of complex traits in several studies. In human disease prediction, it was demonstrated that whole-genome gene expression profiles increased the prediction accuracy when they were used in breast cancer prediction (Vazquez et al., 2016). In maize complex traits prediction, transcriptome data were found to have similar predictive ability as SNP markers, and it was stressed that the use of transcript information may be important for unveiling the contribution of regulatory variation to the genetic architecture of traits (Azodi et al., 2019). In other studies gene expression data was considered as complementary information, and was combined with sequence data in the prediction of traits. E.g integrating transcriptome data into prediction of rice yield lead to a higher prediction accuracy of the combined method compared to only using a single type of predictors (Hu et al., 2019).

However, a challenge of gene expression data-based phenotype prediction stems from the fact that, other than nuclear DNA-based SNP data, the mRNA transcript abundance is affected by several factors, such as time of sampling and experimental conditions. In addition, gene expression levels may be variable among different tissues. Assessing gene expression in the specific tissue at the specific time is critical for the success of gene expression data-based phenotype prediction.

An advanced intercross line (AIL) of mice is the simplest possible outbred population (Darvasi and Soller, 1995). It is produced by intercrossing two inbred strains beyond the F2 generation, and has been demonstrated a powerful tool for genetic analysis (Gonzales et al., 2018). The LG/J x SM/J advanced intercross line (AIL) of mice is a multigenerational outbred population, which was derived from the LG and SM inbred strains (Ehrich et al., 2005). In this paper, we used AIL of mice (generation 50–56) with pedigree, SNP, phenotype data, and gene expression quantified from three brain tissues. For prediction we used five kernel-based linear models: best linear unbiased prediction (BLUP) with pedigree data, genomic BLUP (GBLUP) with SNP data, transcriptomic BLUP (TBLUP) with tissue-specific transcriptome data, GTBLUP combining SNP data and tissue-specific transcriptome data, and PTGLUP combining pedigree data and tissue-specific transcriptome data for phenotype prediction. The objective was to test whether using tissue-specific transcriptome data with a linear model can improve phenotype prediction compared to BLUP and GBLUP for traits with medium to low heritability in the studied mouse population.

Materials and methods

Genotype and pedigree data

The 1063 mouse individuals (530 female, 533 male) with pedigree information were from a multigenerational outbred population which had been sequenced with the reduced-representation genotyping method genotyping-by-sequencing (GBS) (Elshire et al., 2011). The GBS data yielded 38,238 high-quality autosomal SNPs (Gonzales et al., 2018). X chromosomal SNPs were excluded to avoid potential problems with genotyping accuracy, statistical power, and other complications that had been discussed elsewhere (Wise et al., 2013). Beagle 4.1 had been used in conjunction with haplotypes of LG and SM lines obtained from whole genome sequencing data to impute 4.3 million additional SNPs into the 1063 mice (Browning and Browning, 2007; Nikolskiy et al., 2015; Browning and Browning, 2016). SNPs with MAF < 0.1, and Hardy–Weinberg Equilibrium violations were removed. Finally, 523,028 SNPs were used in the analysis.

Gene expression data

The generation of expression data is described in Gonzales et al. (2018), and provided tissue-specific gene expression data of three brain tissues: hippocampus (HIP), prefrontal cortex (PFC), and striatum (STR) were quantified with mRNA transcript abundances. These data were used to map expression quantitative trait loci QTL (eQTL) contributing to mammalian behavior and physiological traits (Gonzales et al., 2018). Three groups of 208 (HIP), 185 (PFC) and 169 (STR) individuals, respectively, sampled among the 1063 phenotyped and genotyped mice, were used to generate gene expression data. The tissues from each brain had been dissected within five minutes, aiming at limiting stress-induced changes in gene expression. All brain tissues were dissected by the same experimenter and subsequently stored at−80°C until extraction. The abundances of RNA products from three tissues encompassed 16,533 genes in HIP, 16,249 genes in PFC and 16,860 genes in STR, respectively. For more details, see Gonzales et al. (2018). The overlap of individuals and the overlap of genes among the three subsets is shown in Figure 1.

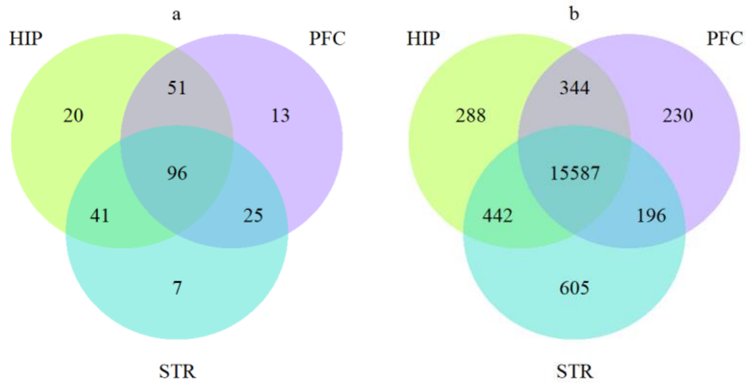


Figure 1 The Venn diagrams show the overlap of individuals (a) and the overlap of genes (b) among the three groups of mice

Phenotype data

We used 8 traits in this study which had been phenotyped previously, including four novel behavioral traits, three of which were conditioned place preference (CPP) for methamphetamine, and the other one was the number of side changes which was a trait measuring locomotor activity. CPP is an associative learning paradigm that had been used to measure the motivational properties of drugs in humans and rodents (Tzschentke, 1998; Mayo et al., 2013), and it was defined as the number of seconds spent in a drug-associated environment relative to a neutral environment over the course of 30 min. The full procedure takes eight days, which were referred to as D1–D8. The baseline preference was measured after administration of vehicle (0.9% saline, i.p.) on D1. On D2 and D4, mice were administered methamphetamine (1 mg kg−1, i.p.) and restricted to one visually and tactically distinct environment; on D3 and D5 mice were administered vehicle and restricted to the other, contrasting environment. The locomotor activity trait measured during the CPP test on D1 and D8. CPP and locomotor traits were measured across six five-minute intervals and summed them to generate a total phenotype for each day.

Further four hindlimb muscle weight traits relevant to exercise physiology were chosen. The four phenotyped muscles include two dorsiflexors: tibialis anterior (TA), and extensor digitorum longus (EDL), and two plantar flexors: gastrocnemius and plantaris. Individual muscles were isolated under a dissection microscope and weighed to 0.1 mg precision on a Pioneer balance.

For more details about phenotyping, statistical description and heritability of phenotypes see table 1 and (Gonzales et al., 2018).

Table 1. Descriptions, means (M), standard deviations (SD), minimum (MIN), maximum (MAX) and SNP-based heritabilities ($\hat{h}^2_{SNP}$) of phenotypes.

Category	Traits	Trait descriptions	$\hat{h}^2_{SNP}$	M	SD	MIN	MAX
Muscle weight traits	TGW	Tibialis anterior weight (mg)	0.379	48.602	7.558	28.9	70.8
	EDLW	Extensor digitorum longus weight (mg)	0.429	8.845	1.558	4.9	13.3
	GW	Gastrocnemius weight (mg)	0.309	110.573	20.893	65.9	168.3
	SW	Soleus weight (mg)	0.202	7.754	1.839	3.2	13.5
Behavioral traits	SCD8	Side changes on Day 8 (20-25 min)	0.105	27.606	10.542	0	78
	D1C	Day 1 activity (saline, 25-30 min)	0.093	1381.883	451.429	0	3093
	D2C	Day 2 activity (1 mg/kg meth, 10-15 min)	0.252	389.366	155.803	0	1019
	D3C	Day 3 activity (saline, 0-30 min)	0.221	192.84	84.028	0	623

Data availability

The genotypes, pedigree, phenotypes, and gene expression data of mice population are freely and publicly available on http://palmerlab.org/protocols-data/, and on http://genenetwork.org/.

Statistical models and estimation of predictive ability

Predictions of phenotypes were performed with 5 linear models: best linear unbiased prediction (BLUP) with pedigree data, genomic BLUP (GBLUP) with SNP data, transcriptomic BLUP (TBLUP) with tissue-specific transcriptome data, GTBLUP combined SNP data and tissue-specific transcriptome data, and PTGLUP combined pedigree data and tissue-specific transcriptome data.

BLUP

The statistical model is (Henderson, 1975):

$$\boldsymbol{y} = \boldsymbol{1}\boldsymbol{\mu} + \boldsymbol{a} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ is the vector of phenotypic observations, $\boldsymbol{a}$~$N(0, \boldsymbol{A}\sigma_a^2)$ and $\boldsymbol{e}$~ $N(0, \boldsymbol{I}\sigma_e^2)$ are vectors containing random breeding values and residual effects, respectively and where μ is the overall mean and $\mathbf{1}$ a vector of ones. The numerator relationship matrix $\boldsymbol{A}$ was calculated using AIL pedigree information from generation 1 to 56 with the R package "AGHmatrix" (Amadeu et al., 2016).

GBLUP

We used all available SNPs data to conduct the benchmark GBLUP (VanRaden, 2008). The statistical model for GBLUP is:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{e}$$

where $\mathbf{g}$~$N(0, \boldsymbol{G}\sigma_g^2)$. The genomic relationship matrix $\boldsymbol{G}$ was calculated as $\boldsymbol{G} = \frac{\boldsymbol{ZZ\prime}}{2\Sigma p_i(1-p_i)}$ (VanRaden, 2008), where p_i denotes the minor allele frequency (MAF) of marker i. Moreover, $\boldsymbol{Z}$ denotes the MAF adjusted marker matrix with entries $(0-2p_i)$, $(1-2p_i)$ and $(2-2p_i)$ for genotypes AA, Aa and aa, respectively. All other variables are as defined above.

TBLUP

In this approach, tissue-specific gene expression data were used as predictor variables (Li et al., 2019). The statistic model is:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{t} + \boldsymbol{e}$$

where $\boldsymbol{t}$~$N(0, \boldsymbol{E}\sigma_t^2)$ is a transcriptomic line effect. The corresponding variance-covariance matrix is $\boldsymbol{E} = \boldsymbol{RR}'$ which is a linear kernel calculated from an n x m matrix $\boldsymbol{R}$ of standardized gene expression levels from n lines and m genes. The transcriptomic relationship matrix $\boldsymbol{E}$ used here reflected transcriptomic similarity between individuals based on tissue-specific gene expression data. The standardization of gene expression levels was conducted by calculating $r_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j}$, where x_{ij} is the expression level of gene j in line i, $\bar{x}_j$ is the mean expression level of gene j across all lines, and s_j is the standard deviation of gene expression level of gene

j. All other variables are as defined above.

GTBLUP

In GTBLUP, transcriptome data was integrated into genomic prediction (Li et al., 2019). SNP data and transcriptome data were treated as predictor variables. The prediction model was:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{t} + \boldsymbol{e}$$

where all variables are defined as described above.

PTBLUP

In PTBLUP, pedigree data and transcriptome data were treated as predictor variables. The prediction model was:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{a} + \boldsymbol{t} + \boldsymbol{e}$$

where all variables are defined as described above.

The different approaches were assessed using 20 replicates of a 5-fold cross-validation (Erbe et al., 2013). Predictive abilities were defined as the Pearson's correlation coefficients between predicted genetic values and observed phenotypes in the test sets. The final predictive ability of each model was the mean of the predictive abilities across 100 estimates. Random effects from the five models were estimated using the R package "regress" (Clifford and McCullagh, 2014).

Results

All analyses were conducted in three groups of mice, and each group had a unique type of gene expression data quantified from one of the three brain tissues: hippocampus, prefrontal cortex, and striatum. The predictive abilities of eight traits obtained with the five statistical models for the three groups are shown in Figure 2.

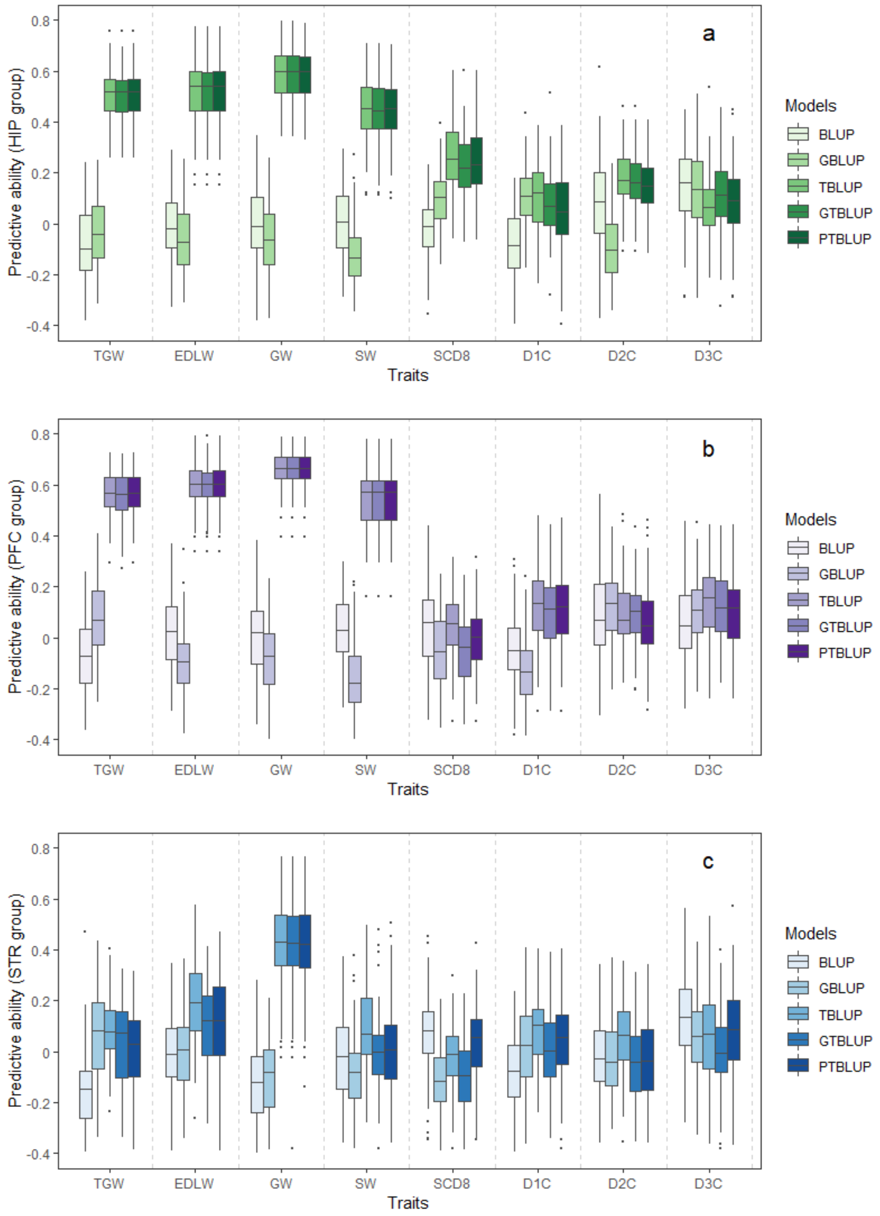


Figure 2. Predictive abilities for 8 traits with 5 statistical models in 3 groups: Panels a, b, and c, refer to the HIP, PFC, and STR group, respectively. Trait names are as in Table 1.

BLUP and GBLUP provided very low predictive abilities for the traits with low to medium heritabilities, the observed predictive abilities being -0.003 and -0.02 on average across eight traits and three groups, respectively. The transcriptome-based prediction (TBLUP) was the most accurate method for which the observed predictive ability was 0.26 on average, and it performed equal to or slightly better than the two combining methods (GTBLUP and PTBLUP) whose observed predictive abilities were 0.235 and 0.246, respectively.

In both the HIP and PFC groups, the predictive abilities of TBLUP, GTBLUP and PTBLUP on the four muscle weight traits (TGW, EDLW, GW, SW) were remarkably higher than the predictive abilities of BLUP and GBLUP. For the two behavioral trait (SCD8, D2C) in HIP group and one behavioral trait (D1C) in PFC group, the predictive abilities of TBLUP, GTBLUP and PTBLUP were also higher than BLUP and GBLUP, while for the remaining behavioral traits (D1C, D3C) in HIP group, and SCD8, D2C, D3C in PFC group, the predictive abilities of all models remained low. In the STR group, the predictive abilities of TBLUP, GTBLUP and PTBLUP were distinctly higher than the predictive abilities of BLUP and GBLUP just for two muscle weight traits (EDLW, GW).

For the four behavioral traits (SCD8, D1C, D2C, D3C) with low heritabilities from 0.1 to 0.25, the five models in the three groups overall provided very low predictive abilities (< 0.1). For the four muscle weight traits (TGW, EDLW, GW, SW) with low to medium heritabilities from 0.2 to 0.43, the predictive abilities of the models (TBLUP, GTBLUP, PTBLUP) with transcriptome data on average were distinctly higher (0.42) than the predictive abilities of the models (BLUP, GBLUP) without transcriptome data (-0.05). The SNP-based heritabilities, the proportion of phenotypic variance explained by the additive effects of 523,028 SNPs for 8 traits had been estimated previously using the restricted maximum likelihood algorithm in GEMMA as described in Gonzales et al. (2018), and are shown in table 1. The heritabilities of 8 traits were highly correlated with the predictive abilities of TBLUP on average across 8 traits and 3 groups with a correlation coefficient of 0.71, while the heritabilities were not resp. unfavorably correlated with the predictive abilities of BLUP or GBLUP on average across 8 traits and 3 groups with correlation coefficient of -0.25 or -0.08, respectively (Fig. 3). To study whether the number of genes has impact on transcriptome-based prediction, we randomly chose 1000, 50000 and 10000 genes for prediction with TBLUP. The differences among predictive abilities using different numbers of genes were negligible (results not shown).

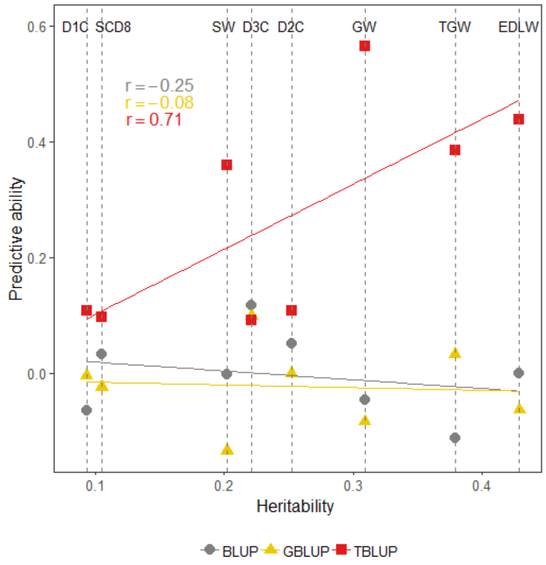


Figure 3. The correlation between heritabilities of 8 traits and the average predictive abilities across 8 traits and 3 groups in BLUP, GBLUP and TBLUP. r denotes the Pearson correlation coefficient. The red, grey and yellow lines denote three standardized major axis regression lines. The dotted lines represent the heritabilities of 8 traits.

Discussion

In this study, we tested gene expression data quantified from three brain tissues in a mouse outbred population for phenotype prediction. The utility of gene expression data for phenotype prediction has also been evaluated in several other species (Acharjee et al., 2016; Guo et al., 2016; Tissier et al., 2018). For prediction of complex traits in *Drosophila melanogaster*, the predictive abilities of transcriptome-based method TBLUP and reproducing kernel Hilbert space regression were significantly lower than GBLUP both in females and males for all traits with high heritability (Li et al., 2019). For this result, the explanation was suggested that gene expression was not measured at one time point and in one specific tissue functionally linked to the trait of interest. It has been demonstrated that gene expression extensively varies among tissues in teleost fish, soybean, mice and humans (Maguire et al., 2002; Oleksiak et al., 2002; Hsieh et al., 2003; Yang et al., 2006; GTEx Consortium, 2015). The low predictive abilities was also found in maize where transcriptome-based prediction did not outperform models that used genotype data for flowering time, height and grain yield (Azodi et al., 2019), and grain dry matter content (Schrag et al., 2018).

In contrast to the studies in *Drosophila melanogaster* that used transcript abundance quantified from entire flies (Li et al., 2019), or in maize that utilized transcriptome data from whole-seedling tissues (Schrag et al., 2018; Azodi et al., 2019), our primary interest was assessing the effectiveness of gene expression data quantified from specific tissues for phenotype prediction of complex traits with medium to low heritability. For the four muscle weight traits, prediction based on tissue-specific transcriptome data performed remarkably better than pedigree-based prediction and SNP-based prediction. A similar study revealed that when using transcriptome data specifically sampled from flag leaves for rice yield prediction, the predictive ability was distinctly higher than SNP-based prediction for yield per plant and grain number per panicle (Hu et al., 2019). In addition, improved predictive abilities were also observed when using gene expression data specifically quantified from immature seeds compared to SNP-based prediction in a maize study for predictions of days to silking, kernel width and ear diameter (Guo et al., 2016). These studies demonstrated that using tissue-specific gene expression data is effective to improve predictive abilities of transcriptome-based phenotype prediction.

However, for the four behavioral traits, transcriptome data did not generally improve prediction relative to pedigree and SNPs based approaches. Analogously, for tiller number per plant and 1000-gain weight in rice (Hu et al., 2019) , and cob weight and plant height in maize (Guo et al., 2016), the transcriptome data-based prediction performed also worse than SNP-based prediction, even though tissue-specific gene expression data were used. This may indicate that the predictive abilities based on tissue-specific transcriptome data depend on the genomic architecture of traits. In our study, the predictive abilities of BLUP and GBLUP were extremely low for all traits in the mouse populations. Such low predictive abilities were not surprising, since all traits we analyzed had relatively low heritabilities and extremely small training set sizes. Similar results were also observed in studies for weight and growth slope with low heritabilities with about 1900 mice across families, while for within-family prediction, the predictive abilities was considerably improved compared to across-family prediction (Legarra et al., 2008; Neves et al., 2012), confirming that predictive abilities are highly dependent on relatedness among individuals.

Some studies indicated that combining transcriptome data with SNP data or pedigree information could improve predictive abilities for several yield and quality-related traits in silage maize (Westhues et al., 2017; Schrag et al., 2018). However, our results have shown that combining gene expression data with pedigree data or SNPs did not improve predictions, and in some cases even decreased predictive abilities. It was also observed in studies of maize and *Drosophila melanogaster* that a combined prediction of transcriptome data and SNP data had similar or slightly lower predictive ability than the superior single type of data (Li et al., 2019). This indicates that combining different sources of data will not always bring improvement of predictive abilities.

When we randomly chose different numbers of genes for prediction with TBLUP, the difference among predictive abilities was negligible. This is in accordance with a maize hybrid prediction study using 1000 and 10000 randomly chosen mRNAs (Zenke-Philippi et al., 2016), where only minor differences were observed in predictive abilities. This indicates that high numbers of genes are not necessarily required for transcriptome-based prediction, and that transcription profiling with limited resources might result in prediction accuracies that can be successfully used for indirect selection (Zenke-Philippi et al., 2016).

In this study, modeling transcriptome data with linear models was shown to have the potential to improve trait prediction. The reasons for this improvement could be that transcriptome data might be “closer” to the phenotype, and harbor more information than genotypes, e.g. multiple interactions between different genes and between genes and environmental factors. The heritable part of genome-wide gene expression variation was first assessed in a cross population of *Saccharomyces cerevisiae* (Brem et al., 2002), indicating a substantial genetic component in transcriptional variation in yeast (Skelly et al., 2009). Furthermore, it has been proven that non-additivity is common in *D. melanogaster* (Huang et al., 2012), *A. thaliana* and maize (Vuylsteke et al., 2005), and that its extreme forms, overdominance and underdominance, are common (Gibson et al., 2004). The expression level of genes may be a complicated non-linear function of genetic effects and environmental effects, but this complicated function could be linearly captured by the transcript abundances.

Gene expression can be greatly affected by the tissue sampled and time of measurement. Thousands of genes are differentially expressed between tissues or show tissue preferential expression (Melé et al., 2015). In addition, some gene expression products have “housekeeping” functions, and are therefore expressed in all cells, while other genes are expressed in a tissue-specific manner (Fagerberg et al., 2014). It has been found that variation in gene expression is even far greater among tissues (47% of total variance in gene expression) than among individuals (4% of total variance) (Melé et al., 2015). Hence, it is important to use gene expression data from specific tissues for phenotype prediction. In this study, we demonstrated that the predictive abilities of tissue-specific transcriptome data-based prediction were remarkably higher than the pedigree-based and SNP-based prediction for certain traits. However, since the gene expression data from three brain tissues were quantified from three different subsets of the mouse population, we could not compare the predictive abilities among predictions with transcriptome data from different tissues. More studies with larger sample size and with different types of tissue-specific gene expression data need to be performed to further explore the potential benefits.

Acknowledgments

We are grateful to Abraham A. Palmer and Natalia M. Gonzales at University of Chicago for providing the mice dataset. ZCL thanks China Scholarship Council for financial support. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

References

Abraham, G., and Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. *Current Opinion in Genetics and Development* 33**,** 10-16.

Acharjee, A., Kloosterman, B., Visser, R.G., and Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics* 17(5)**,** 180-191.

Akdemir, D., and Jannink, J.-L. (2015). Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics* 199(3)**,** 857-871.

Amadeu, R.R., Cellon, C., Olmstead, J.W., Garcia, A.A., Resende, M.F., and Muñoz, P.R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. *The Plant Genome* 9(3), 1-10.

Azodi, C.B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S.-H. (2019). Transcriptome-based prediction of complex traits in maize. *BioRxiv***,** 587121.

Bouchard Jr, T.J. (2004). Genetic influence on human psychological traits: A survey. *Current Directions in Psychological Science* 13(4)**,** 148-151.

Browning, B.L., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics* 98(1)**,** 116-126.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* 81(5)**,** 1084-1097.

Clifford, D., and McCullagh, P. (2014). The regress package. *R News* 6**,** 6.

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348(6235)**,** 648-660.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science* 22(11)**,** 961-975.

Darvasi, A., and Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141(3)**,** 1199-1207.

Ehrich, T.H., Hrbek, T., Kenney-Hunt, J.P., Pletscher, L.S., Wang, B., Semenkovich, C.F., et al. (2005). Fine-mapping gene-by-diet interactions on chromosome 13 in a LG/J× SM/J murine model of obesity. *Diabetes* 54(6)**,** 1863-1872.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6(5)**,** e19379.

Erbe, M., Gredler, B., Seefried, F.R., Bapst, B., and Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS One* 8(12)**,** e81046. doi: 10.1371/journal.pone.0081046.

Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics* 13(2)**,** 397-406.

Gonzales, N.M., Seo, J., Cordero, A.I.H., Pierre, C.L.S., Gregory, J.S., Distler, M.G., et al. (2018). Genome wide association analysis in a mouse advanced intercross line. *Nature Communications* 9(1)**,** 5162-5174.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics* 129(12)**,** 2413-2427. doi: 10.1007/s00122-016-2780-5.

Hatcher, S., Atkins, K., and Safari, E. (2010). Lamb survival in Australian Merino sheep: a genetic analysis. *Journal of Animal Science* 88(10)**,** 3198-3205.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics***,** 423-447.

Hsieh, W.-P., Chu, T.-M., Wolfinger, R.D., and Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165(2)**,** 747-757.

Hu, X., Xie, W., Wu, C., and Xu, S. (2019). A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnology Journal* pp. 1–10.

Jiang, Y., and Reif, J.C. (2015). Modeling epistasis in genomic selection. *Genetics* 201(2)**,** 759-768.

Kim, H., Grueneberg, A., Vazquez, A.I., Hsu, S., and de los Campos, G. (2017). Will big data close the missing heritability gap? *Genetics* 207(3)**,** 1135-1145.

Kosova, G., Abney, M., and Ober, C. (2010). Heritability of reproductive fitness traits in a human population. *Proceedings of the National Academy of Sciences* 107(suppl 1)**,** 1772-1778.

Kvestad, E., Czajkowski, N., Engdahl, B., Hoffman, H.J., and Tambs, K. (2010). Low heritability of tinnitus: results from the second Nord-Trøndelag health study. *Archives of Otolaryngology–Head & Neck Surgery* 136(2)**,** 178-182.

Legarra, A., Robert-Granié, C., Manfredi, E., and Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics* 180(1)**,** 611-618.

Li, Z., Simianer, H., and Martini, J.W. (2019). Integrating gene expression data into genomic prediction. *Frontiers in Genetics* 10**,** 126-137.

Mackay, T.F., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10(8)**,** 565-577.

Maguire, T.L., Grimmond, S., Forrest, A., Iturbe-Ormaetxe, I., Meksem, K., and Gresshoff, P. (2002). Tissue-specific gene expression in soybean (Glycine max) detected by cDNA microarray analysis. *Journal of Plant Physiology* 159(12)**,** 1361-1374.

Mayo, L.M., Fraser, D., Childs, E., Momenan, R., Hommer, D.W., De Wit, H., et al. (2013). Conditioned preference to a methamphetamine-associated contextual cue in humans. *Neuropsychopharmacology* 38(6)**,** 921-929.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., et al. (2015). The human transcriptome across tissues and individuals. *Science* 348(6235)**,** 660-665.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4)**,** 1819-1829.

Neves, H.H., Carvalheiro, R., and Queiroz, S.A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC Genetics* 13(1)**,** 100-117.

Nikolskiy, I., Conrad, D.F., Chun, S., Fay, J.C., Cheverud, J.M., and Lawson, H.A. (2015). Using whole-genome sequences of the LG/J and SM/J inbred mouse strains to prioritize quantitative trait genes and nucleotides. *BMC Genomics* 16(1)**,** 415-427.

Oleksiak, M.F., Churchill, G.A., and Crawford, D.L. (2002). Variation in gene expression within and among natural populations. *Nature Genetics* 32(2)**,** 261-266.

Pappa, I., Fedko, I.O., Mileva-Seitz, V.R., Hottenga, J.-J., Bakermans-Kranenburg, M.J., Bartels, M., et al. (2015). Single nucleotide polymorphism heritability of behavior problems in childhood: genome-wide complex trait analysis. *Journal of the American Academy of Child & Adolescent Psychiatry* 54(9)**,** 737-744.

Schaeffer, L. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of animal breeding and Genetics* 123(4)**,** 218-223.

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., et al. (2018). Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics* 208, 1373–1385.

Su, G., Christensen, O.F., Ostersen, T., Henryon, M., and Lund, M.S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PloS One* 7(9)**,** e45293.

Thekkoot, D., Kemp, R., Rothschild, M., Plastow, G., and Dekkers, J. (2016). Estimation of genetic parameters for traits associated with reproduction, lactation, and efficiency in sows. *Journal of Animal Science* 94(11)**,** 4516-4529.

Tissier, R., Houwing-Duistermaat, J., and Rodríguez-Girondo, M. (2018). Improving stability of prediction models based on correlated omics data by using network approaches. *PloS One* 13(2)**,** e0192853.

Tzschentke, T.M. (1998). Measuring reward with the conditioned place preference paradigm: a comprehensive review of drug effects, recent progress and new issues. *Progress in Neurobiology* 56(6)**,** 613-672.

Verhoeven, K. J. F., Vanhala, T. K., Biere, A., Nevo, E. and Van Damme, J. M. M (2004). The genetic basis of adaptive population differentiation: a quantitative trait locus analysis of fitness traits in two wild barley populations from contrasting habitats. *Evolution* **58**, 270–283.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91(11)**,** 4414-4423. doi: 10.3168/jds.2007-0980.

Vazquez, A.I., Veturi, Y., Behring, M., Shrestha, S., Kirst, M., Resende, M.F., et al. (2016). Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. *Genetics* 203(3)**,** 1425-1438.

Vitezica, Z.G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195(4)**,** 1223-1230.

Westhues, M., Schrag, T.A., Heuer, C., Thaller, G., Utz, H.F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theoretical and Applied Genetics* 130(9)**,** 1927-1939.

Wise, A.L., Gyi, L., and Manolio, T.A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. *The American Journal of Human Genetics* 92(5)**,** 643-647.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013a). Author reply to A commentary on Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14(12)**,** 894.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013b). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14(7)**,** 507-515.

Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., et al. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Research* 16(8)**,** 995-1004.

Zenke-Philippi, C., Thiemann, A., Seifert, F., Schrag, T., Melchinger, A.E., Scholten, S., et al. (2016). Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. *BMC Genomics* 17(1)**,** 262-270.

4th CHAPTER

Pan-genomic Open Reading Frames: A Potential Substitution of Single Nucleotide Polymorphisms in Estimation of Heritability and Genomic Prediction

Zhengcao Li[1], Henner Simianer[1*]

[1]Animal Breeding and Genetics Group, Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, Goettingen, Germany

Keywords

Pan-genome, Open reading frames, Single nucleotide polymorphisms, Heritability, Genomic Prediction

Abstract

Pan-genomic open reading frames potentially carry genome-wide protein-coding genes or coding variant information in a population. In this study, we used 1,011 S. cerevisiae isolates with 1,625,809 SNPs, 7,796 pan-genomic ORFs, and the copy numbers of ORFs in genomic prediction and estimation of heritability for 35 traits. Our results show that compared to SNP-based genomic prediction, pan-genomic ORF-based genomic prediction (OBLUP) was distinctly more accurate for all traits, and the prediction was improved by 132% on average across all traits. When using different numbers of isolates in training sets in ORF-based prediction, the predictive abilities for all traits increased as more isolates were added in the training sets. In addition, the ORF-based heritability can capture more genetic effects than SNP-based heritability for all traits. Using copy numbers of pan-genomic ORF information to estimate heritability accounts for more "missing heritability" compared to ORF-based heritability in all 32 traits. For four traits (YP sorbitol 2%, YPD sodium metaarsenite 2.5 mM, YPD LiCl 250mM, YPD CuSO4 10 mM), using the copy numbers of pan-genomic ORFs-based prediction was more accurate than pan-genomic ORF-based prediction. When combining pan-genomic ORFs or the copy numbers of pan-genomic ORFs with common SNPs in prediction models, no increase in phenotypic variance explained was observed. When using exclusively pan-genomic ORF data, OBLUP had similar predictive abilities as ORF-based Bayes A and Bayes B prediction for all traits. However, when only using copy numbers of pan-genomic ORFs, Bayes B performed slightly better than a linear model accounting for copy numbers and Bayes A for 22 of the traits. We demonstrate that pan-genomic ORFs have the potential to be a substitution of single nucleotide polymorphisms in estimation of heritability and genomic prediction under certain conditions.

Introduction

Genome-wide single nucleotide polymorphisms (SNPs) were first proposed in 2001 to be used for predicting genetic values (Meuwissen et al., 2001). Implementation in practice became pervasive due to the large amount of single nucleotide polymorphisms (SNP) that became available in recent years (Goddard and Hayes, 2007). By utilizing genome-wide SNP data, 'genomic selection' based on genomically predicted breeding values has triggered a revolution of estimating genetic value in animal and plant breeding. It improved the breeding progress by reducing generation intervals or increasing predictive ability of breeding values (Schaeffer, 2006; Goddard et al., 2010; Crossa et al., 2017). In human genetics, genomic prediction aimed at accurately quantifying disease risk so that preventative measures can be taken earlier (Abraham and Inouye, 2015). However, SNP markers are normally not causal variants, but in genomic prediction the causal variant effects are estimated indirectly by modeling SNP makers that are in linkage disequilibrium (LD) with them (Goddard and Hayes, 2007). The prediction accuracy highly depends on the level of LD between SNP markers and causal variants, and the level of LD depends on the relatedness of the individuals used (Wray et al., 2013a). For prediction of distantly related individuals, even if high density SNP or whole-genomic SNP markers were used, the prediction accuracy still can be very low (de los Campos et al., 2013). Likewise, genome-wide SNP data are also used for estimation or dissection of genetic parameters, such as SNP-based heritability (Evans et al., 2018). Several factors inevitably caused the ’still missing heritability’ problem when using common SNPs with minor allele frequency (MAF) ≥ 0.01 to estimate narrow sense heritability (Wray et al., 2013b). e.g. the causal variants are not in complete LD with the SNPs that have been genotyped, or rare variants of large effect are not tagged by common SNPs on genotyping arrays (Yang et al., 2010; Yang et al., 2017).

Pan-genomic open reading frames potentially hold whole-genome protein-coding genes or coding variant information. The ‘pan-genome’ denotes the set of all genes or open reading frames (ORFs) present in the genomes of a group of organisms, usually a species (Lapierre and Gogarten, 2009; Vernikos et al., 2015). The concept has been applied to bacterial (Tettelin et al., 2005), viral (Aherfi et al., 2013), plant (Cao et al., 2011; Li et al., 2014; Zhao et al., 2018) , fungal (Dunn et al., 2012), and human genome studies (Sherman et al., 2019). Series of pan-genomic studies were performed when studying genomic dynamics (Donati et al., 2010), pathogenesis and drug resistance (D'Auria et al., 2010; Hu et al., 2011), bacterial toxins (Fang et al., 2011), and species evolution (Konstantinidis et al., 2006). An open reading frame (ORF) is defined as a sequence that has a length divisible by three and is bounded by stop codons (Sieber et al., 2018). It is a sequence region that is ‘open’ for translation, and an indicator for a potential protein-coding gene (Sieber et al., 2018). The detection of ORFs is of central importance in finding protein-coding genes in genomic sequences.

The budding yeast *Saccharomyces cerevisiae* is a model organism which is not only a premier model for eukaryotic cell biology, but also the pioneer organism for the establishment of the new fields “functional genomics” and “systems biology” (Botstein and Fink, 2011). It has previously been shown to be a good tool for exploring the genotype–phenotype relationship via linkage mapping (Fay, 2013), and the study of “missing heritability” (Bloom et al., 2013). Importantly, *S. cerevisiae* is an informative predictor of human gene function: nearly 50% of human genes implicated in heritable diseases have a yeast homologue (Kumar and Snyder, 2001), which makes *S. cerevisiae* a suitable model species for studies of accurate prediction of human disease (Märtens et al., 2016). It further has a compact genome: ~70% of its total (non-ribosomal) DNA sequence is protein-coding, and the yeast genome is reported to encode ~6,200 genes (Goffeau et al., 1996).

Structural variants (SVs) such as presence/absence variants (PAVs) and copy number variants (CNVs) have been proven to substantially influence genetic variation and phenotypic diversity (Marroni et al., 2014). In this study, we used *S. cerevisiae* pan-genomic open reading frames which represent 7,796 non-redundant ORFs in genomic prediction, accounting either for the presence/absence of a specific ORF or its copy number (CNO). With this we exploited a new source of genome-wide variability for genomic prediction and estimation of heritability, and demonstrated (1) genomic prediction using ORF data and CNO data performed substantially better than that using genome-wide SNP data, and (2) the estimation of heritability based on pan-genomic ORF data and CNO data can capture parts of the “missing heritability” that appears when using SNP data.

Data and Methods

Whole-Genome SNP data

We used 1,011 *S. cerevisiae* isolates that maximized the breadth of their ecological and geographical origins comprised in the 1002 Yeast Genome project. In these distantly related isolates, 918 of the isolates had been deep sequenced (Peter et al., 2018), and the other 93 isolates that had previously been sequenced (Skelly et al., 2013; Bergström et al., 2014; Strope et al., 2015). A total of 1,625,809 high-quality SNPs was reported across the 1,011 genomes. Most of these SNPs were present at very low frequency, with 31.3% of the polymorphic positions being singletons and 93% with a minor allele frequency (MAF) < 0.1. We chose a subset of 787 diploid *S. cerevisiae* isolates for which SNP, ORF, copy number of ORF and phenotypes were available for all analyses. The SNPs with missing rate > 0.05, MAF < 0.01, and Hardy–Weinberg Equilibrium violations (based on a Chi-squared test, $p < 10^{-6}$) were removed. The remaining missing genotypes were imputed using Beagle 4.1 (Browning and Browning, 2013). In total, 311’447 SNPs were used in the analysis. The distribution of minor allele frequency of all common SNPs in 787 diploid *S. cerevisiae* isolates is shown in Supplementary Figure 1.

Pan-genomic open reading frame data the *S. cerevisiae* pangenome had been determined by the 1,011 genomes using de novo genome assemblies and detection of non-reference genome material, and represented by 7,796 non-redundant ORFs. Among them, 4,940 were core ORFs, containing ORFs present in all isolates and 2,856 ORFs had a presence/absence variability within the population, containing ORFs that were dispensable or isolate-specific genes. For annotating ORFs in non-reference materials, an integrative yeast gene annotation pipeline had been set up previously by combining different existing annotation approaches, which gave rise to an evidence-leveraged protein-coding gene annotation (Yue et al., 2017). Three individual components: RATT package (Otto et al., 2011), yeast genome annotation pipeline(YGAP) (Proux-Wéra et al., 2012), and Maker pipeline(v2.31.8) (Holt and Yandell, 2011) were independently run for gene annotation, and their results were subsequently integrated using EVidenceModeler(EVM) (Haas et al., 2008). Proteomes of the *Saccharomyces* species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. arboricolus, S. uvarum* and *S. eubayanus*) were retrieved and used in the annotation pipeline to provide protein alignment support for annotated gene models. For details of the de novo genome assemblies, detection of non-reference genome material, and annotation of ORFs see (Peter et al., 2018). The frequency distribution of pan-genomic open reading frames in 787 diploid *S. cerevisiae* isolates is shown in Supplementary Figure 1.

The copy number of each ORF of the pangenome (include copy numbers of core ORFs) was assessed by mapping the reads from each strain to the pan-genomic ORFs with BWA (Li and Durbin, 2009), using default parameters. The median coverage for each ORF was taken as coverage for the ORF in the specific isolate. The ratio between the values of individual ORFs and the values of genome coverage on the reference of the isolate was considered as the copy number for the haploid genome. After removing ORFs with missing value, 7708 ORFs and the copy numbers for 7708 ORFs were left and used in the analysis. For more information about copy number variation distribution across isolates and ORFs see (Peter et al., 2018).

Phenotype data

Quantitative high-throughput phenotyping had been performed using end-point colony growth on solid medium (Peter et al., 2018). In parallel, 971 strains were phenotyped in different conditions that affect various physiological and cellular responses. Strains were pregrown in flat-bottom 96-well microplates containing liquid yeast extract peptone dextrose (YPD) medium. Each phenotype value was normalized using the growth ratio between 35 stress conditions and standard YPD medium at 30°C. Pairwise Pearson's correlations of fitness trait values between replicates were calculated for each condition. In total, 35 fitness traits were used in the present study. The overall statistical description of the 35 traits is shown in Supplementary Table 1, and the correlation matrix of the 35 traits is shown in Supplementary Figure 2.

Statistical models

GBLUP: As a baseline, we conduct the benchmark GBLUP (VanRaden, 2008), using all 311'447 common SNPs (MAF ≥ 0.01) of 787 diploid *S. cerevisiae* isolates. The statistical model for GBLUP is

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{g} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of phenotypic observations, μ is the overall mean and $\mathbf{1}$ is a vector of ones, and $\mathbf{g} \sim N(0, \boldsymbol{G}\sigma_{\mathrm{g}}^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ are vectors containing random additive genetic effects and residual effects. The genomic relationship matrix $\boldsymbol{G}$ was calculated as $\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2\Sigma p_i(1-p_i)}$, where p_i denotes the minor allele frequency (MAF) of marker i. Moreover, $\boldsymbol{Z}$ denotes the MAF adjusted marker matrix with entries $(0-2p_i)$, $(1-2p_i)$ and $(2-2p_i)$ for genotypes 0, 1 and 2, respectively, where the coding refers to the number of reference alleles observed in the genotype.

OBLUP: The model for OBLUP is

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{o} + \boldsymbol{e},$$

where $\boldsymbol{o} \sim N(0, \boldsymbol{O}\sigma_o^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ are vectors containing random additive genetic effects modeled by pan-genomic ORFs and residual effects, respectively. The ORF-based covariance matrix $\boldsymbol{O}$ was calculated as $\boldsymbol{O} = \frac{\boldsymbol{WW\prime}}{\Sigma q_i(1-q_i)}$, where q_i denotes the frequency of ORF i, and $\boldsymbol{W}$ denotes the ORF matrix with entries $(0-q_i)$ and $(1-q_i)$ that represented absence and presence of ORFs, respectively. All other variables are defined as in the GBLUP model.

CBLUP: The model for CBLUP is

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{c} + \boldsymbol{e},$$

where $\boldsymbol{c} \sim N(0, \boldsymbol{C}\sigma_c^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ are vectors containing random additive genetic effects modeled by copy numbers of pan-genomic ORFs and residual effects, respectively. The covariance matrix based on the copy numbers of pan-genomic ORFs $\boldsymbol{C}$ was calculated as $\boldsymbol{C} = \frac{\boldsymbol{SS'}}{f}$, where $\boldsymbol{S}$ denotes the copy numbers of ORFs matrix with entries $(b_{ij} - u_i)$ where $0 \leq b_{ij} \leq 296$ represents the copy number of the ith ORF in jth isolate, and u_i denotes the mean of copy numbers of ORF i in all isolates. f is a scalar which denotes the median of the diagonal of $\boldsymbol{SS'}$. All other variables are defined as in the GBLUP model.

GOBLUP and GCBLUP: The linear model for GOBLUP is

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{o} + \boldsymbol{e},$$

and the linear model for GCBLUP is

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{c} + \boldsymbol{e},$$

where all variables are defined as described above.

ORF or CNO-based Bayes A and Bayes B: The model of ORF or CNO -based Bayes A is

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{a_m} + \boldsymbol{e},$$

where $\boldsymbol{a_m}$ is a m x 1 vector of normally distributed and independent ORF or CNO effects. The variance of the ith ORF effect, σ_{mi}^2 , is modeled as a scaled inverted chi-square distribution $\mathcal{X}^{-2}(v,\ S)$, where S = 0.002, and v = 5. $\boldsymbol{y},\ \boldsymbol{\mu},\ \boldsymbol{e}$ are defined as described above. Gibbs-sampling chains for 50,000 iterations were run, and the first 45,000 burn-in iterations were discarded. The model of ORF or CNO-based Bayes B is the same as with ORF-based Bayes A, but the prior distribution of the variance of ORF effect is a mixture of distributions which is given by

$$\sigma_{mi}^2 \begin{cases} = 0 & with\,probability\,\pi \\ = \mathcal{X}^{-2}(v,S) & with\,probability\,(1-\pi) \end{cases}$$

ORF or ORF-based Bayes A and Bayes B were implemented in an R package 'BGLR' (Pérez and de Los Campos, 2014).

Estimation of heritability

The SNP-based heritability was defined as the proportion of phenotypic variance explained by SNP marker effects and calculated as $\hat{h}_G^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2+\hat{\sigma}_e^2}$. All common SNPs (defined here as those with MAF ≥ 0.01) were used for the estimation (Yang et al., 2017).

The ORF-based heritability was defined as the proportion of phenotypic variance explained by ORF effects. It was calculated as $\hat{h}_O^2 = \frac{\hat{\sigma}_O^2}{\hat{\sigma}_O^2+\hat{\sigma}_e^2}$. All variable ORFs without missing values were used for the estimation. The copy number of ORF (CNO)-based heritability was defined as the proportion of phenotypic variance explained by the copy number of ORF effects. It was calculated as $\hat{h}_C^2 = \frac{\hat{\sigma}_C^2}{\hat{\sigma}_C^2+\hat{\sigma}_e^2}$. The copy numbers of 7,708 pan-genomic ORFs without missing values were used for the estimation. The ORF-SNP-based heritability was defined as the proportion of phenotypic variance explained by ORF and SNP effects. It was calculated as $\hat{h}_{GO}^2 = \frac{\hat{\sigma}_O^2+\hat{\sigma}_g^2}{\hat{\sigma}_O^2+\hat{\sigma}_g^2+\hat{\sigma}_e^2}$. All common

SNPs and variable ORFs without missing values were used for the estimation. The CNO-SNP-based heritability was defined as the proportion of phenotypic variance explained by CNO and SNP effects. It was calculated as $\hat{h}^2_{GC} = \frac{\hat{\sigma}^2_c+\hat{\sigma}^2_g}{\hat{\sigma}^2_c+\hat{\sigma}^2_g+\hat{\sigma}^2_e}$. All common SNPs and copy numbers of 7708 pan-genomic ORFs without missing values were used for the estimation. The variance components $\hat{\sigma}^2_g$, $\hat{\sigma}^2_o$, $\hat{\sigma}^2_c$, $\hat{\sigma}^2_e$ from models above were estimated from the entire data sets, using the R package “regress” (Clifford and McCullagh, 2014), which also provided predictions of random effects.

Comparison of predictive abilities

The predictive abilities of these models were measured with 20 replicates of a 5-fold cross-validation (Erbe et al., 2013). We defined predictive abilities as the Pearson’s correlation coefficients between predicted genetic values and observed phenotypes in the test sets. The mean of the predictive abilities across 100 estimates was the final predictive ability of each model.

Principal component analysis

Principal components analysis (PCA) of all common SNPs, pan-genomic open reading frames, and copy number of pan-genomic open reading frames on 787 diploid *S. cerevisiae* isolates was performed using R package ‘factoextra’.

Genomic and genetic distances

Three neighbor-joining trees were constructed with the R package ‘ape’ using all common SNPs, pan-genomic open reading frames, and copy number of pan-genomic open reading frames, respectively (Paradis and Schliep, 2018). Isolate dissimilarities were estimated via “Euclidean distance” for each pair of isolates with the dist.gene function.

Linkage disequilibrium

The extent of linkage disequilibrium was measured for two subsets of total SNPs: (1) MAF ≥ 0.01, (2) MAF ≥ 0.05, which contained 311’447 SNPs and 102’253 SNPs, respectively. The software PLINK 1.9 was used to calculate r^2 as a standard measure of association for linkage disequilibrium between syntenic pairwise SNPs (Purcell et al., 2007). The average r^2 between all pairwise SNPs on each chromosome represented the extent of linkage disequilibrium on the three subsets of total SNPs.

Data availability

ALL data used in this study are available in the 1002 Yeast Genome website http://1002genomes.u-strasbg.fr/files/.

Results

Population structure based on different genetic variants

Three types of datasets: all common SNPs, pan-genomic open reading frames, and copy numbers of pan-genomic open reading frames were used for principal components analysis (PCA) on the 787 diploid *S. cerevisiae* isolates. Based on the first four principal components, each type of dataset showed a diverse genetic structure of the *S. cerevisiae* isolates (Supplementary Figure 3). Compared to the PCA with SNPs where most isolates scattered into a shape of triangle, most isolates in PCA with ORFs and CNOs gathered, but isolates in PCA with CNOs were more scattered than isolates in PCA with ORFs. The first principal component (PC1) in the PCA with SNPs caught 41.7% of the total variance which was much more than PC1 in PCA with ORFs (18.8%) and PCA with CNOs (7.04%). Likewise, three neighbor-joining trees based on the three types of data were shown in Supplementary Figure 4. The ORF-based and CNO-based neighbor-joining trees had similar shapes in which the genetic distances among most isolates were close, and only a few isolates were far away from the other isolates in terms of genetic distance. The ‘outlier’ isolates in ORF-based and CNO-based neighbor-joining trees partly overlapped. For the SNP-based neighbor-joining tree, the genetic distances among most isolates were relatively large, and the isolates clustered into groups that were clearly separated from each other. The heat maps of genetic covariance matrixes: $\boldsymbol{G}$, $\boldsymbol{O}$, $\boldsymbol{C}$ constructed using three types of datasets are shown in Supplementary Figure 5, where the yeast strains were in the same order on the basis of their geographical origins in the three matrices. The red color blocks, indicating high covariance, in the SNP-based genetic covariance matrix were in different positions compared with the red color blocks in the other two genetic covariance matrixes. The red color blocks in the ORF-based and CNO-based genetic covariance matrixes shared similar positions along the diagonal region, but compared to the ORF-based genetic covariance matrix, the CNO-based genetic covariance matrix has more red color blocks indicating high similarity in the off-diagonal regions.

Estimation of heritability

Narrow sense heritability was estimated using three datasets with three models: all common SNPs (GBLUP), pan-genomic open reading frames (OBLUP), or copy numbers of pan-genomic open reading frames (CBLUP). The SNP-based heritability ($\hat{h}_G^2$) was the lowest on average across all traits (0.281 ± 0.005), ranging from 0.004 ± 0.002 to 0.67 ± 0.003 (Supplementary table 1). The ORF-based heritability ($\hat{h}_O^2$) on average across all traits was 0.847 ± 0.002, ranging from 0.766 ± 0.004 to 0.919 ± 0.001, and notably captured more phenotypic variance attributable to the additive genetic variation than the SNP-based heritability in all traits. The CNO-based heritability ($\hat{h}_C^2$) was the highest on average across all traits (0.935 ± 0.002), ranging from 0.445 ± 0.021 to 0.996 ± 0 (Figure 1).

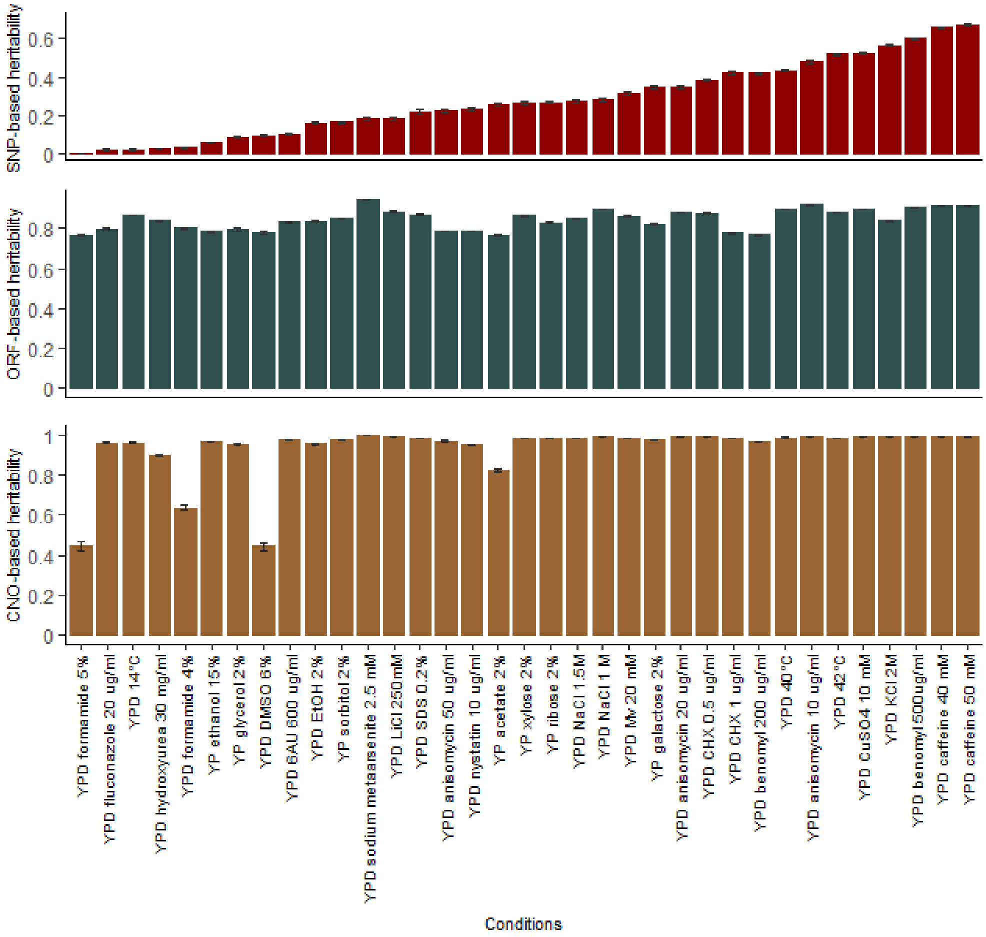


Figure 1. Heritability estimates for all 35 traits estimated based on all common SNPs, pan-genomic open reading frames, and the copy numbers of pan-genomic open reading frames, respectively. Each error bar indicates the standard error of the estimate.

When using copy numbers of pan-genomic ORF information to estimate narrow sense heritability, the $\hat{h}_C^2$ captured more "missing heritability" compared with $\hat{h}_O^2$ in 32 traits, and only for three traits (YPD formamide 5%, YPD formamide 4%, YPD DMSO 6%) $\hat{h}_C^2$ was lower than $\hat{h}_O^2$. Among the 32 traits, there were 20 traits for which $\hat{h}_C^2$ exceeded 0.98. We combined all common SNPs with pan-genomic ORFs to estimate the SNP-ORF-based heritability ($\hat{h}_{GO}^2$) using GOBLUP, and combined all common SNPs with pan-genomic CNOs to estimate the SNP-CNO-based heritability ($\hat{h}_{GC}^2$) using GCBLUP. The $\hat{h}_{GO}^2$ and $\hat{h}_{GC}^2$ were consistent with $\hat{h}_O^2$ and $\hat{h}_C^2$ for all traits, and no more additive genetic variance explained by SNPs was captured (Supplementary table 2).

Assessment of predictive abilities

The predictive abilities of the 35 traits obtained with the 3 models: GBLUP, OBLUP, CBLUP are shown in Figure 2 and Supplementary Table 3.

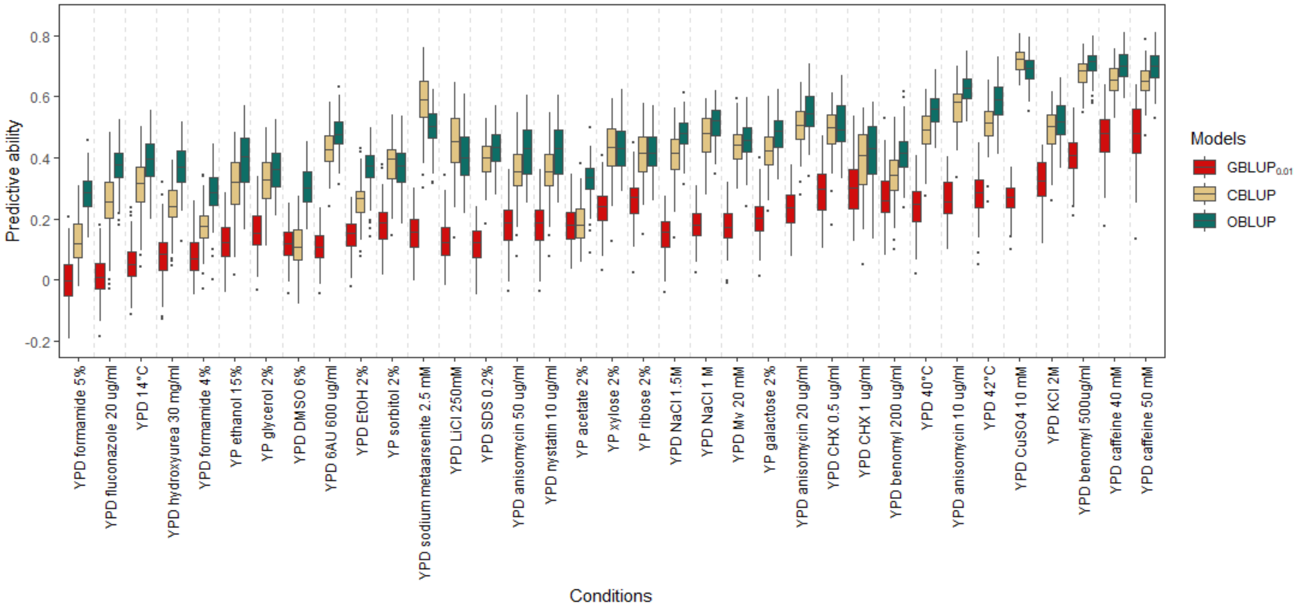


Figure 2. Predictive abilities of three models across 35 traits: GBLUP$_{0.01}$ using all common SNPs, OBLUP using pan-genomic open reading frames, and CBLUP using copy numbers of pan-genomic open reading frames.

GBLUP as the reference method provided predictive abilities ranging from 0.002 ± 0.007 to 0.482 ± 0.009 across the studied traits. For GBLUP, the SNP-based heritability and genomic predictive ability were highly positively correlated with r= 0.935 (Figure 3).

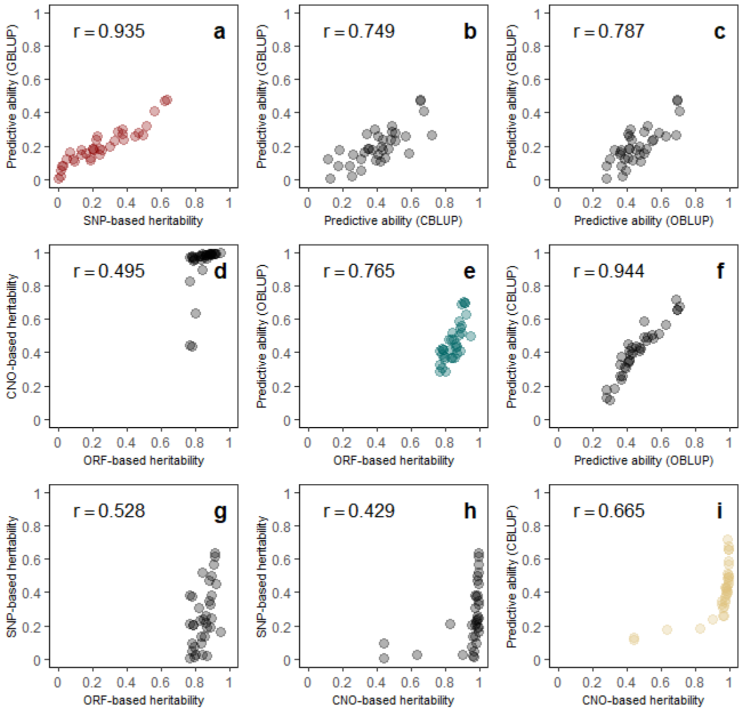


Figure 3. Panel a depicts the correlation between predictive abilities of GBLUP and SNP-based heritabilities across all traits; b depicts the correlation between predictive abilities of GBLUP and predictive abilities of CBLUP across all traits; c depicts the correlation between predictive abilities of GBLUP and predictive abilities of OBLUP across all traits; d depicts the correlation between CNO-based heritabilities and ORF-based heritabilities across all traits; e depicts the correlation between predictive abilities of OBLUP and ORF-based heritabilities across all traits; f depicts the correlation between predictive abilities of CBLUP and predictive abilities of OBLUP across all traits; g depicts the correlation between SNP-based heritabilities and ORF-based heritabilities across all traits; h depicts the correlation between SNP-based heritabilities and CNO-based heritabilities across all traits; i depicts the correlation between predictive abilities of CBLUP and CNO-based heritabilities across all traits . r depicts the Pearson correlation coefficients. The dots in the 9 panels depict the 35 traits.

Compared to GBLUP, pan-genomic ORF-based prediction (OBLUP) was more accurate for all traits: observed predictive abilities ranged from 0.284 ± 0.006 to 0.706 ± 0.004. The correlation coefficient between SNP-based predictive abilities and ORF-based predictive abilities was 0.787, and the correlation coefficient between the ORF-based heritability and ORF-based predictive ability was 0.765. When using different numbers of isolates in training sets in ORF-based prediction, the predictive abilities of all traits increased as the number of isolates in the training set increased(Figure 4), showing that increasing the training set size could more accurately estimate ORF effects. The curves in Figure 4 corresponding to a function, $r = w\sqrt{\frac{nh^2}{nh^2+M_e}}$ (Erbe et al., 2013), of the heritability ($\hat{h}_O^2$), the number of isolates (n) and the number of independent chromosome segments (M_e) was used to fit the predicted points by least squares, where r represents the predictive ability in this study. The two parameters w and M_e across 35 traits were determined with a maximum likelihood approach which was done using the function “optim” in R (Team, 2013).

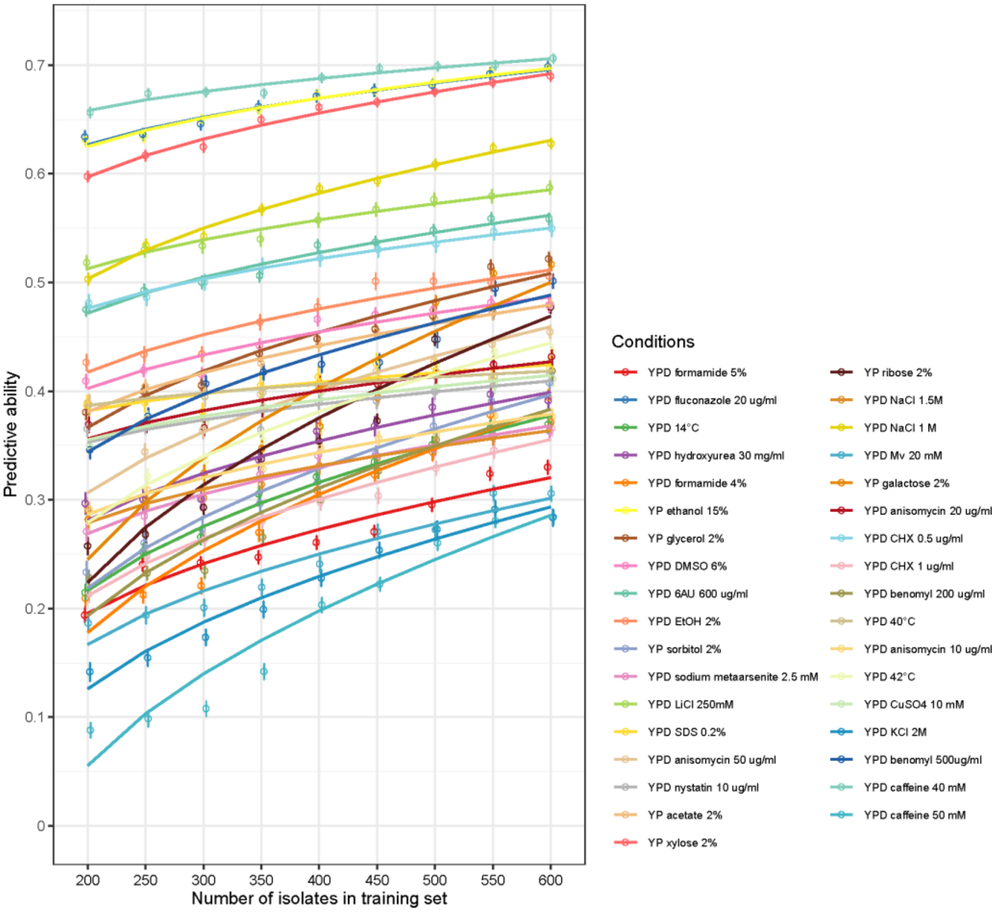


Figure 4. The predictive abilities of ORF-based genomic prediction for 35 traits using different number of isolates in training sets. The solid curves are fitted lines that correspond to the function, $r = w\sqrt{\frac{nh^2}{nh^2+M_e}}$ *(Erbe et al., 2013), where r represents the predictive ability in this study.*

The predictive ability of copy numbers of pan-genomic ORF-based prediction (CBLUP) was 0.13 ± 0.008 to 0.72 ± 0.004, which was significantly higher than the predictive ability of GBLUP. For four traits (YP sorbitol 2%, YPD sodium metaarsenite 2.5 mM, YPD LiCl 250mM, YPD CuSO4 10 mM), CBLUP was more accurate than OBLUP, while for the remaining 31 traits, CBLUP was slightly less accurate than OBLUP. The reason could be that some of CNOs were not simple repeats of causal variants, and these CNOs added noise in the prediction. The correlation coefficient between the CNO-based heritability and CNO-based predictive ability was 0.633. When we combined two subsets of total SNP data (MAF ≥ 0.01 and MAF ≥ 0.05) which contained 311’447 SNPs and 102’253 SNPs, respectively, to pan-genomic ORFs with GOBLUP, the predictive abilities remained the same with OBLUP only using pan-genomic ORFs data. The average r^2 between all pairwise SNPs on each chromosome for the two subsets of all SNPs (MAF ≥ 0.01, MAF ≥ 0.05) were 0.034 and 0.119, respectively. For the second combined method GCBLUP, the predictive abilities remained the same as with CBLUP for all traits (Supplementary Figure 6, 7 and Supplementary Table 3), suggesting that ORF data or CNO data covered all causal variant information which SNP data carried. When using exclusively pan-genomic ORF data, OBLUP had similar predictive abilities with ORF-based Bayes A and Bayes B for all traits (Supplementary Figure 8). However, when only using copy numbers of pan-genomic ORFs, Bayes B performed slightly better than CBLUP and Bayes A for 22 traits, which indicated that some of the copy numbers of ORF information had no genetic effect (Supplementary Figure 9).

Discussion

Capture of “still missing heritability”

‘Missing heritability’ has been a critical problem in quantitative genetics: causal variants discovered using genome-wide association studies (GWAS) only explain a small proportion of the phenotypic variation of human height (Maher, 2008). When using all common SNPs simultaneously in a linear model, 45% of phenotypic variance of human height can be explained, which demonstrated that SNP data without any pre-filtering for significance in GWAS could capture a larger part, but still not all of the missing heritability (Yang et al., 2010). However, the estimation of SNP-based heritability depended on the extend of LD between SNP markers and causal variants. If SNPs were in low LD with causal variants, which might occur if common SNPs are used but causal variants have low MAF, genomic variants cannot be well tagged by SNPs. Thus, a part of the heritability could still be missing, which was termed "still missing heritability" (Wray et al., 2013b).

Our results show that the ORF-based heritability ($\hat{h}_O^2$) was able to capture a major part of the "still missing heritability" for all traits. On average across all traits 84.7% of phenotypic variance was explained by ORF-based additive genetic effects, while only 25.4% of phenotypic variance can be explained by SNP-based additive genetic effects. This indicates that pan-genomic open reading frames hold more causal variant information than common SNPs in the population, and pan-genomic ORFs encompass most of the repertoire of genes or coding variants accessible to the yeast population. On the other hand, it also provides evidence that most of the genetic variation of complex traits is additive in nature. In other words, additive genetic variance accounts for most of total genetic variance, and this genetic variation can be captured by a linear model (Hill et al., 2008). Furthermore, the CNO-based heritability ($\hat{h}_C^2$) was higher than $\hat{h}_O^2$ for 32 of the 35 traits, which indicates that copy number variation of pan-genomic ORFs can further explain more of the missing variance of additive genetic effects. The reason could be that part of copy numbers of ORFs reflect a variable number of repeats of some complete genes. An example of a complete gene repeat was that the copy number of human alpha-amylase 1 gene (AMY1), which is directly associated with the amount of salivary amylase (Walker, 2007), significantly varied between different populations with different diets. Another example is the correlation between the copy number of the chemokine gene CCL3L1 and susceptibility to HIV/AIDS. There are significant interindividual and interpopulation differences in the copy number of a segmental duplication encompassing the gene encoding CCL3L1 (MIP-1αP) (Gonzalez et al., 2005). In addition, the $\hat{h}_C^2$

exceeded 0.99 for 19 of the 35 traits, which showed copy numbers of pan-genomic ORFs harbored almost all causal variant information in the yeast population for these traits. However, there were three traits (YPD formamide 5%, YPD formamide 4%, YPD DMSO 6%) for which $\hat{h}_C^2$ was substantially lower than $\hat{h}_O^2$. It showed the causal variants of these three traits were not repeated by copy numbers, and using copy number of ORF data presumably added noise in the estimation of genetic variance.

Improvement of predictive ability

Precision of SNP-based genomic prediction depends on two factors: SNP-based heritability and the accuracy with which the SNP marker effects are estimated (Goddard et al., 2009). The SNP-based heritability provides the upper bound of predictive ability for SNP-based genomic prediction, this upper bound can be reached when big sample sizes are used for model training (Kim et al., 2017). However, the biggest inherent limitation of SNP-based genomic prediction is the extend of LD between SNP markers and causal variants. When causal variants are in low LD with SNPs, additive genetic effects will be underestimated (Yang et al., 2010; Speed et al., 2012), and the SNP-based heritability can be much lower than narrow-sense heritability which is the ultimate upper bound of predictive ability when genetic variance explained by all additive effects are captured. Since there is no perfect LD between causal variants and SNPs, e.g. when rare variants are not captured by common SNPs (Wray et al., 2013b), the ultimate upper bound (narrow-sense heritability) can never be reached when only using SNPs in genomic prediction. Due to this limitation, genomic prediction suffers from diminishing improvements when trying to increase prediction accuracy by increasing the training set. It is necessary to explore new sources of predictors to overcome the imperfection. Recently, multi-omics data (transcriptome, metabolome, proteome etc.) appeared to be possible complements to SNP markers in genomic prediction (Guo et al., 2016; González-Reymúndez et al., 2017; Li et al., 2019). However, these types of data also have inherent limitations for prediction of genetic value, since they are not causal variants but intermediate products between causal variants and phenotypes (Rockman and Kruglyak, 2006). During the transfer process of genetic information from DNA to phenotype, multi-omics data will be inevitably affected by environmental effects , or the interaction effects between genes and environments (Gibson et al., 2004).

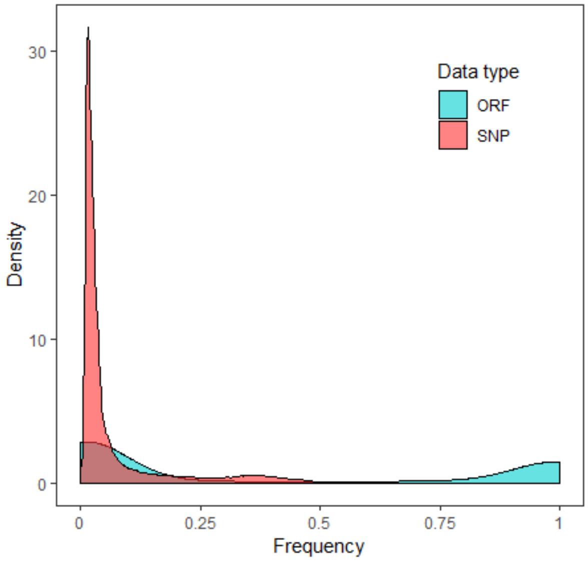
The ‘pan-genome’ denotes the set of all genes or ORFs present in the genomes of a group of organisms (Tettelin et al., 2005; Bentley, 2009). It provides an opportunity to accommodate the phenotypic variation caused by the potential protein-coding sequences in a population. We hypothesize that pan-genomic ORFs can be viewed as a representation of a pan-genomic gene set, and directly using this pan-genomic structure variation (presence/absence) set at gene level in genomic prediction can capture more genetic variance than SNP-based prediction. Furthermore, pan-genomic ORFs can also be viewed as a representation of a coding variant set. Causal variants are either coding or regulatory (Georges et al., 2018). Coding variants falling within a coding region, especially non-synonymous variants, may change amino acid sequences, and then lead to phenotype variations (Marouli et al., 2017). In our results, compared to SNP-based genomic prediction, pan-genomic ORF-based genomic prediction was substantially more accurate for all traits, and the predictive abilities were improved by 132% on average across all traits, which manifested the distinct advantage of making use of pan-genomic ORF data in genomic prediction. However, it should be noted that the pan-genomic ORFs excluded most of non-coding causal variants which are regulatory variants located in non-coding regions. It has been proven that the majority of disease and trait associated variants emerging from genome-wide association analysis studies (GWAS) in humans lie within noncoding sequence that are not in linkage disequilibrium with coding exons (Maurano et al., 2012). Such noncoding variants have substantial effects in gene expression (Albert and Kruglyak, 2015), and may further influence phenotypes (Yan et al., 2002; Kleinjan and van Heyningen, 2005) Nevertheless, when we combined pan-genomic ORFs with common SNPs in the model, no more phenotypic variance explained by SNPs was captured, which suggests the noncoding variants have limited impact on the variation of phenotypes in the yeast population, or are not in sufficient LD with the used SNP set.

A recent study (Sherman et al., 2019), showed that the African pan-genome encompasses ~10% more DNA than the current human reference genome, but this study did not provide ORF information for the population. To our knowledge, there are no other higher animals’ pan-genomes reported so far. In plants, a range of pan-genome studies have shown gene presence/absence variation in many species. Different species present various proportions of core genes: *Brachypodium distachyon* (35%) (Gordon et al., 2017), rice (54%) (Wang et al., 2018), *Brassica napus* (62%) (Hurgobin et al., 2018), bread wheat (64.3%) (Montenegro et al., 2017), and tomato (74.2%) (Gao et al., 2019). Whether pan-genomic ORF data can be used for human risk prediction or for animal or plant breeding remains unverified, but one advantage of ORF-based genomic prediction is obvious: ORF-based genomic prediction is not involved in the 'insufficient LD' problem which appears in SNP-based estimation of heritability and genomic prediction. Relative to livestock and crops, predicting genotypes or phenotypes using SNPs in humans may be more challenging because the extent of LD in human populations is lower than in domesticated species, which have a long and intensive history of selection and smaller effective population size. In a human genetics context, ORF based prediction may have the potential to more accurately identify individuals that are at risk for diseases, and to improve the preventive medicine strategies and clinical decision making.

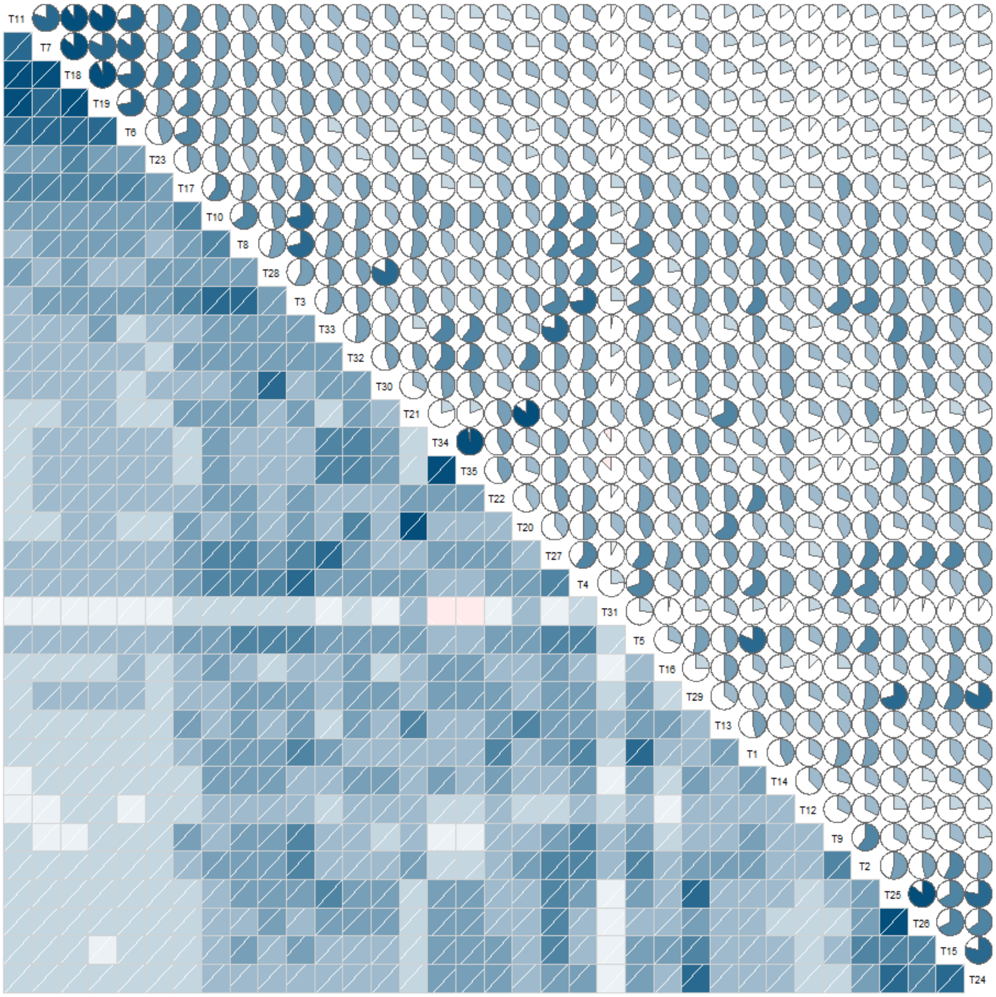
In conclusion, the ORF-based and CNO-based heritability can capture a major part of the "still missing heritability", and ORF-based and CNO-based genomic prediction were more accurate than SNP-based genomic prediction for all traits in the distantly related yeast isolates. We demonstrated that pan-genomic ORFs explained more causal variance than common SNPs in the population, and so ORFs have potential to substitute or complement SNPs in estimation of heritability and genomic prediction under certain conditions. However, in our study there still was a major gap between heritability and prediction accuracy for all traits. We provide evidence that prediction accuracy will be further improved if larger sample sizes can be used in training sets.

Acknowledgments

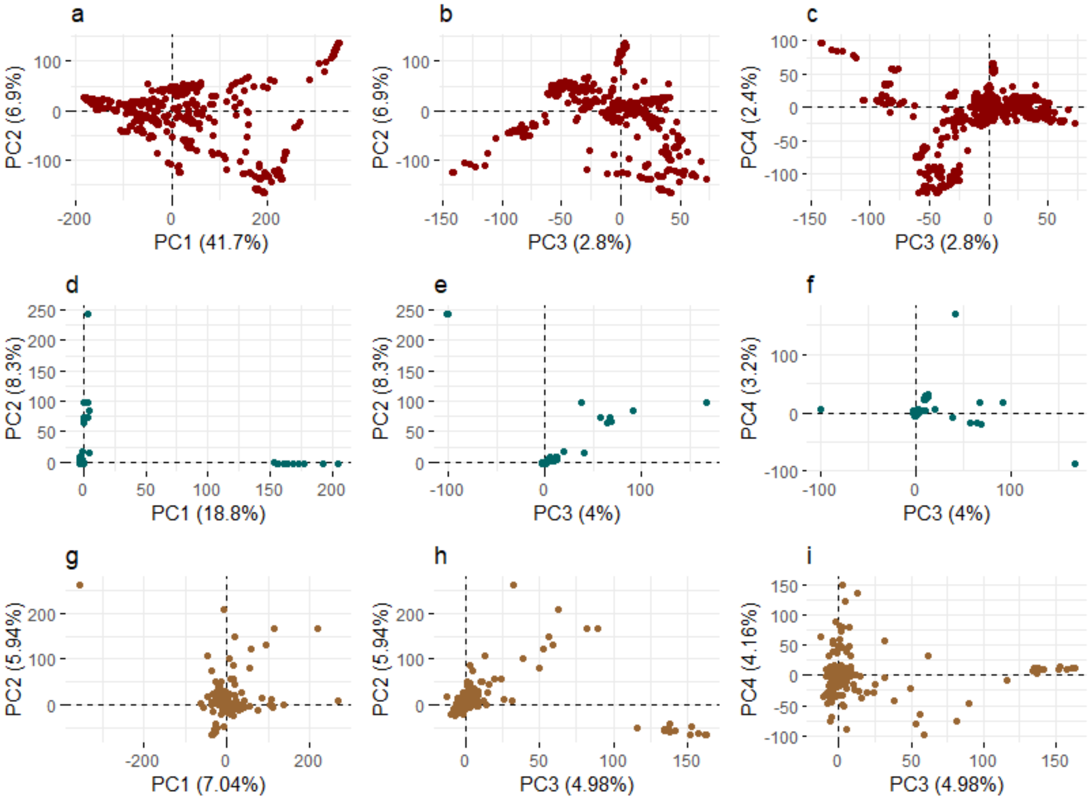
ZCL thanks China Scholarship Council for financial support.



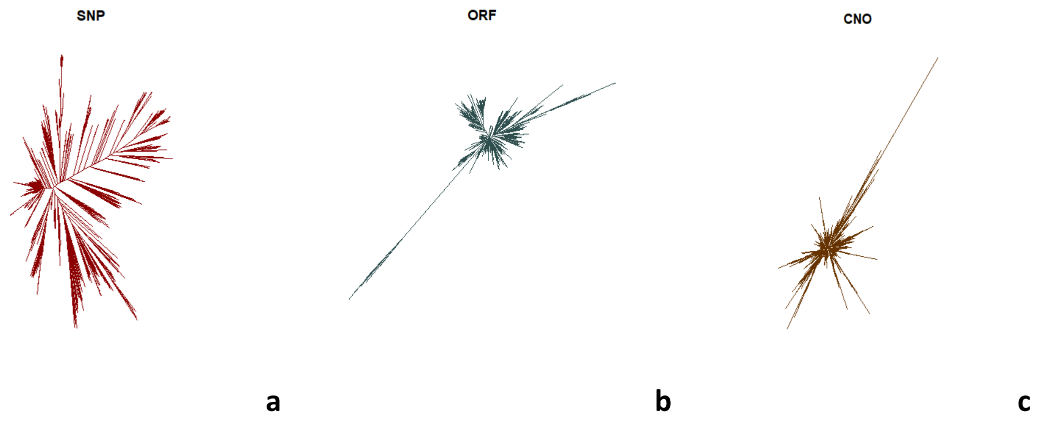
Supplementary Figure 1. Distribution of minor allele frequency of all common SNPs (red), and distribution of frequency of occurrence of variable ORFs among 787 diploid S. cerevisiae isolates (green).



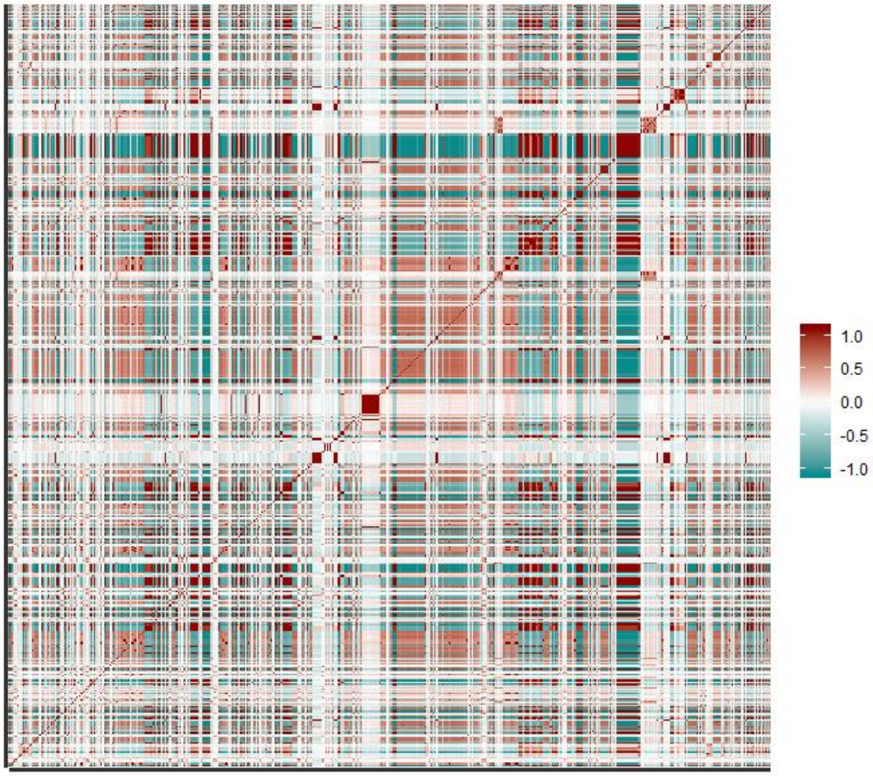
Supplementary Figure 2. Phenotype correlation matrix of 35 traits. Traits were sorted according to the principal component ordering. The blue and pink color denote positive and negative correlation, respectively. The scale ranges from r = 0.977 for combination T34 and T35 to -0.126 for combination T35 and T31. T1 to T35 represent the 35 traits which were shown in Supplementary Table 1.



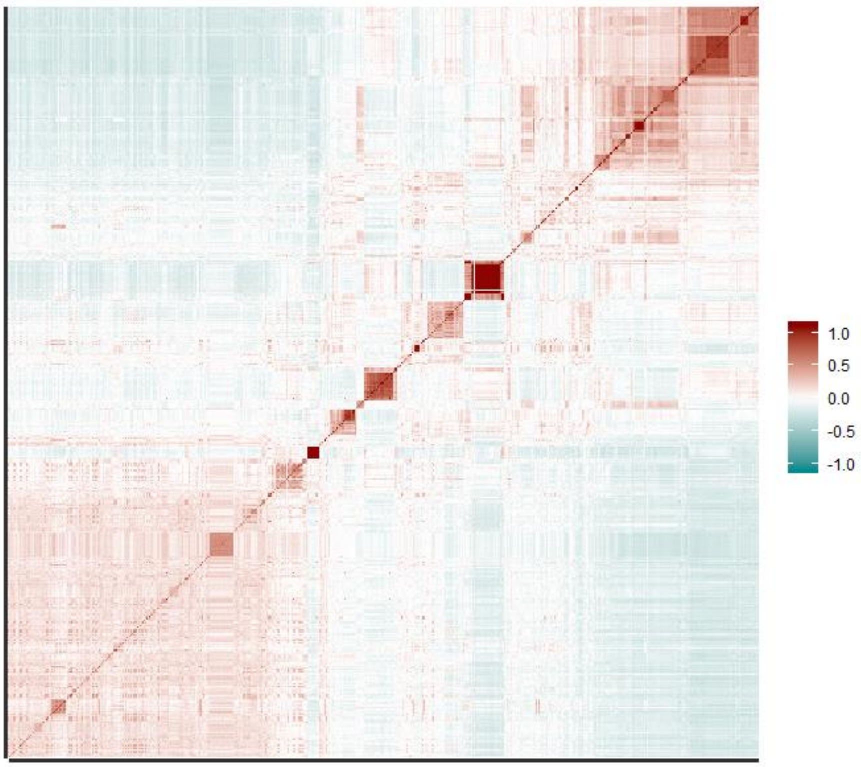
Supplementary Figure 3. Panels a, b and c represent principal component (PC) analysis for all common SNPs on 787 diploid S. cerevisiae isolates. Panels d, e and f represent PC analysis for pan-genomic open reading frames on 787 diploid S. cerevisiae isolates. Panels g, h and i represent PC analysis for the copy numbers of pan-genomic open reading frames on 787 diploid S. cerevisiae isolates. PC1, PC2, PC3 and PC4 denote the first four principal components.



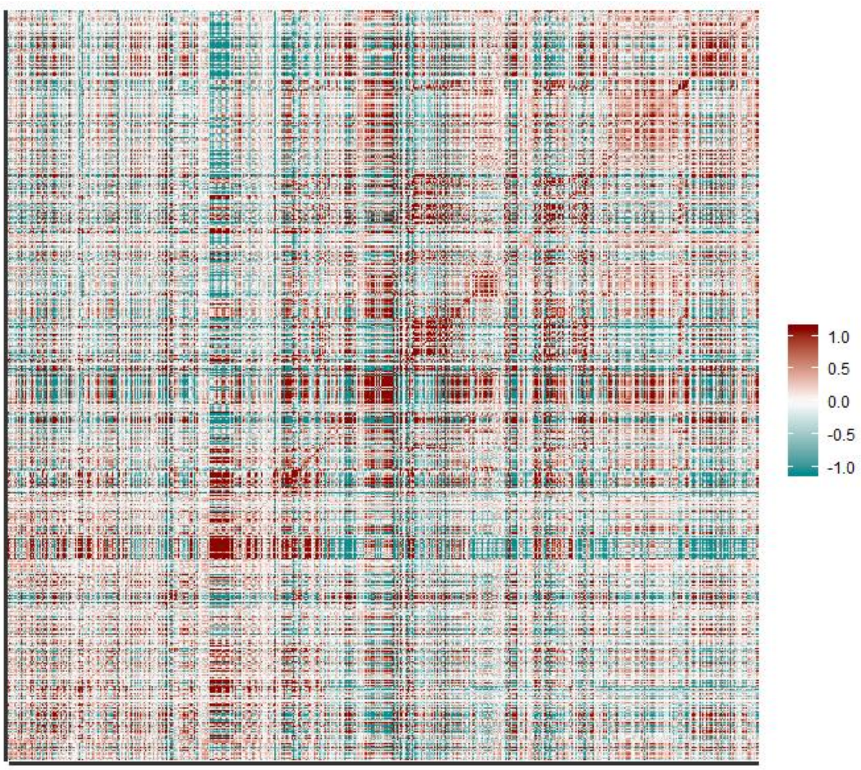
Supplementary Figure 4. Panels a, b and c represent the Neighbor-joining trees of 787 diploid S. cerevisiae constructed using all common SNPs, pan-genomic open reading frames, and copy numbers of pan-genomic open reading frames, respectively.



(a)

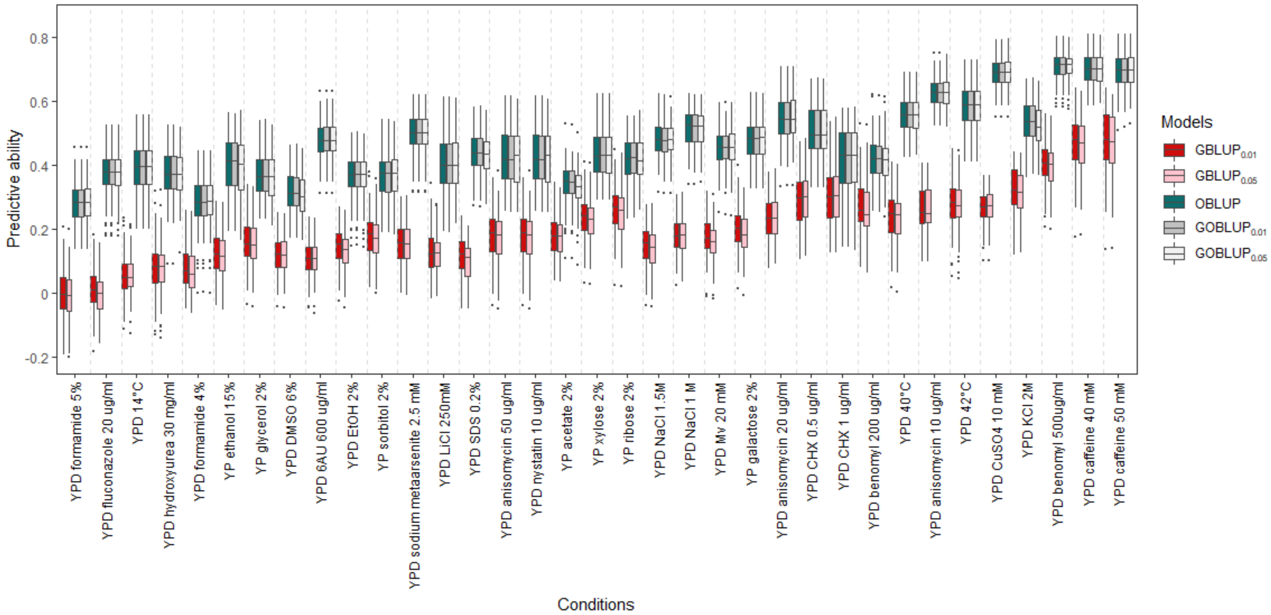


(b)

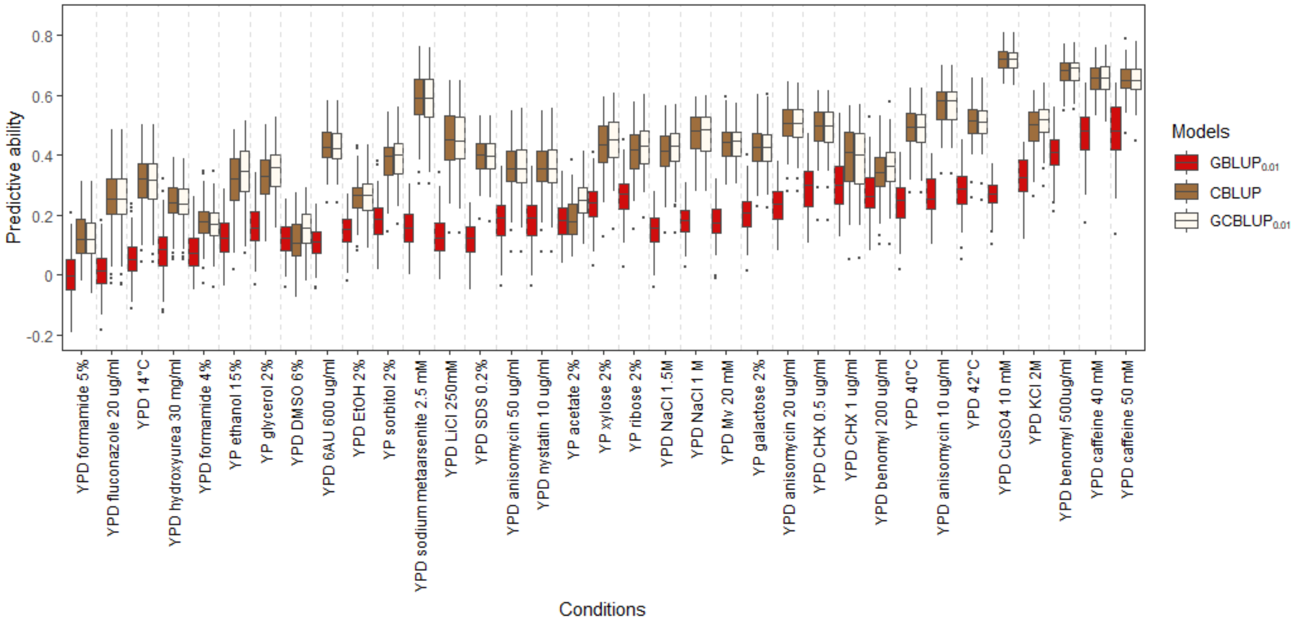


(c)

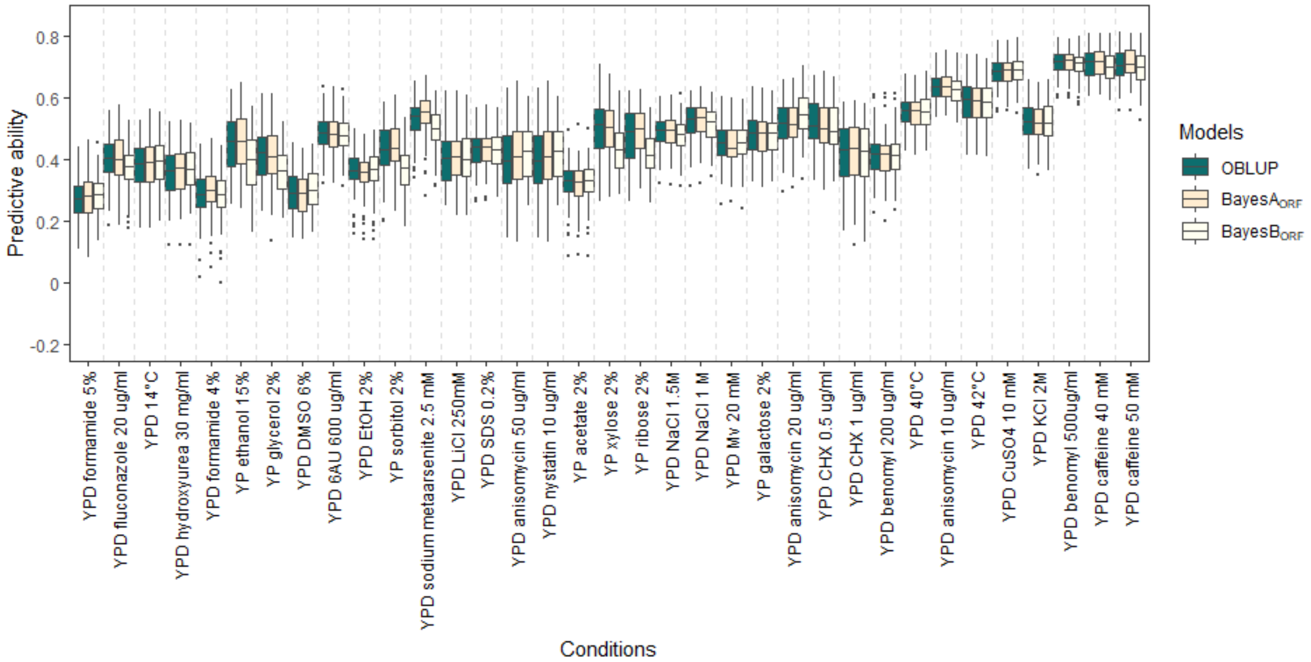
Supplementary Figure 5. Panels a, b and c display heatmaps of genetic covariance matrixes of 787 diploid S. cerevisiae isolates based on all common SNPs, pan-genomic open reading frames, and copy numbers of pan-genomic open reading frames, respectively. Isolates are in the same order in all three panels.



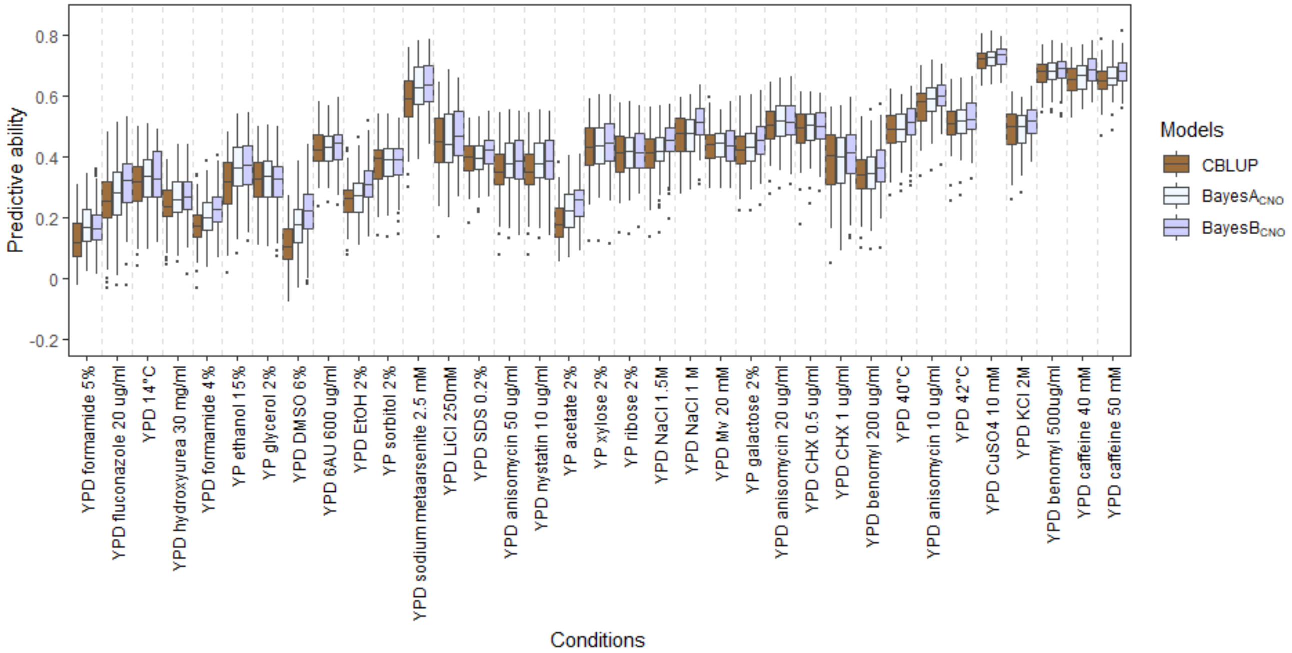
Supplementary Figure 6. Box plots for predictive abilities of GBLUP$_{0.01}$ using SNPs with MAF ≥ 0.01, GBLUP$_{0.05}$ using SNPs with MAF ≥ 0.05, OBLUP using pan-genomic open reading frames, GOBLUP$_{0.01}$ using both SNPs with MAF ≥ 0.01 and pan-genomic open reading frames, GOBLUP$_{0.05}$ using both SNPs with MAF ≥ 0.05 and pan-genomic open reading frames.



Supplementary Figure 7. Box plots for predictive abilities of $GBLUP_{0.01}$ using SNPs with MAF ≥ 0.01, CBLUP using copy numbers of pan-genomic open reading frames, $GCBLUP_{0.01}$ using both SNPs with MAF ≥ 0.01 and copy numbers of pan-genomic open reading frames.



Supplementary Figure 8. Box plots for predictive abilities of OBLUP, BayesA$_{ORF}$ and BayesB$_{ORF}$ using pan-genomic open reading frames across 35 traits.



Supplementary Figure 9. Box plots for predictive abilities of OBLUP, BayesA$_{ORF}$ and BayesB$_{ORF}$ using copy numbers of pan-genomic open reading frames across 35 traits.

Supplementary Table 1. Statistical description of phenotype data.

	Conditions	Mean ± Standard error	Variance	Max value	Minimum value
T1	YPD formamide 5%	0.258 ± 0.004	0.01	1	0.026
T2	YPD fluconazole 20 ug/ml	0.405 ± 0.005	0.02	1.142	0.005
T3	YPD 14°C	0.462 ± 0.003	0.008	0.81	0.021
T4	YPD hydroxyurea 30 mg/ml	0.358 ± 0.004	0.01	0.849	0.02
T5	YPD formamide 4%	0.406 ± 0.004	0.013	0.964	0.026
T6	YP ethanol 15%	0.585 ± 0.006	0.033	1.517	0.013
T7	YP glycerol 2%	0.546 ± 0.006	0.031	1.632	0.004
T8	YPD DMSO 6%	0.533 ± 0.004	0.013	1.277	0.049
T9	YPD 6AU 600 ug/ml	0.31 ± 0.004	0.015	1.213	0.002
T10	YPD EtOH 2%	0.394 ± 0.004	0.011	0.756	0.008
T11	YP sorbitol 2%	0.431 ± 0.005	0.021	1.455	0.01
T12	YPD sodium metaarsenite 2.5 mM	0.256 ± 0.008	0.048	1.576	0.002
T13	YPD LiCl 250mM	0.2 ± 0.004	0.014	0.971	0.003
T14	YPD SDS 0.2%	0.244 ± 0.006	0.025	0.84	0.002
T15	YPD anisomycin 50 ug/ml	0.222 ± 0.005	0.018	1.123	0.006
T16	YPD nystatin 10 ug/ml	0.138 ± 0.003	0.009	0.705	0.001
T17	YP acetate 2%	0.432 ± 0.004	0.016	1.125	0.034
T18	YP xylose 2%	0.397 ± 0.004	0.016	1.345	0.01
T19	YP ribose 2%	0.413 ± 0.005	0.016	1.113	0.006
T20	YPD NaCl 1.5M	0.151 ± 0.003	0.005	0.48	0.003
T21	YPD NaCl 1 M	0.241 ± 0.003	0.009	0.645	0.014
T22	YPD Mv 20 mM	0.171 ± 0.003	0.007	0.641	0.002
T23	YP galactose 2%	0.92 ± 0.01	0.082	1.903	0.075
T24	YPD anisomycin 20 ug/ml	0.378 ± 0.009	0.061	1.464	0.003
T25	YPD CHX 0.5 ug/ml	0.301 ± 0.005	0.023	1.135	0.001
T26	YPD CHX 1 ug/ml	0.141 ± 0.004	0.01	1.132	0.002
T27	YPD benomyl 200 ug/ml	0.239 ± 0.003	0.007	0.679	0.007
T28	YPD 40°C	0.655 ± 0.007	0.044	1.318	0.066
T29	YPD anisomycin 10 ug/ml	0.623 ± 0.009	0.067	1.318	0.001
T30	YPD 42°C	0.418 ± 0.007	0.038	1.555	0.005
T31	YPD CuSO4 10 mM	0.617 ± 0.016	0.2	1.756	0.01
T32	YPD KCl 2M	0.194 ± 0.004	0.01	0.585	0.007
T33	YPD benomyl 500ug/ml	0.278 ± 0.004	0.015	0.857	0.009
T34	YPD caffeine 40 mM	0.211 ± 0.005	0.022	0.739	0.002
T35	YPD caffeine 50 mM	0.15 ± 0.004	0.013	0.627	0.002

Supplementary Table 2. Heritabilities estimated from five models across 35 traits: GBLUP, OBLUP, CBLUP, GOBLUP and GCBLUP. $\hat{h}^2_G$ *denoted the SNP-based heritability;* $\hat{h}^2_O$ *the ORF-based heritability;* $\hat{h}^2_C$ *the CNO-based heritability;* $\hat{h}^2_{GO}$ *the SNP-ORF-based heritability;* $\hat{h}^2_{GC}$ *the SNP-CNO-based heritability*

Conditions	$\hat{h}^2_G$	$\hat{h}^2_O$	$\hat{h}^2_C$	$\hat{h}^2_{GO}$	$\hat{h}^2_{GC}$
YPD formamide 5%	0.004 ± 0.002	0.766 ± 0.004	0.445 ± 0.021	0.767 ± 0.004	0.447 ± 0.021
YPD fluconazole 20 ug/ml	0.019 ± 0.003	0.799 ± 0.003	0.966 ± 0.003	0.8 ± 0.003	0.966 ± 0.003
YPD 14°C	0.021 ± 0.003	0.868 ± 0.002	0.965 ± 0.001	0.868 ± 0.002	0.962 ± 0.001
YPD hydroxyurea 30 mg/ml	0.027 ± 0.001	0.841 ± 0.002	0.9 ± 0.003	0.846 ± 0.002	0.91 ± 0.003
YPD formamide 4%	0.032 ± 0.003	0.802 ± 0.004	0.638 ± 0.01	0.803 ± 0.004	0.63 ± 0.01
YP ethanol 15%	0.059 ± 0.002	0.784 ± 0.004	0.971 ± 0.001	0.79 ± 0.004	0.973 ± 0.003
YP glycerol 2%	0.094 ± 0.003	0.796 ± 0.004	0.957 ± 0.001	0.798 ± 0.004	0.955 ± 0.001
YPD DMSO 6%	0.104 ± 0.004	0.78 ± 0.004	0.44 ± 0.021	0.79 ± 0.004	0.441 ± 0.021
YPD 6AU 600 ug/ml	0.094 ± 0.005	0.832 ± 0.002	0.98 ± 0	0.834 ± 0.002	0.981 ± 0.003
YPD EtOH 2%	0.165 ± 0.004	0.839 ± 0.002	0.96 ± 0.001	0.843 ± 0.002	0.963 ± 0.001
YP sorbitol 2%	0.162 ± 0.005	0.852 ± 0.003	0.976 ± 0	0.851 ± 0.003	0.972 ± 0
YPD sodium metaarsenite 2.5 mM	0.181 ± 0.006	0.946 ± 0.001	0.999 ± 0	0.946 ± 0.001	0.993 ± 0.005
YPD LiCl 250mM	0.184 ± 0.006	0.887 ± 0.002	0.994 ± 0	0.889 ± 0.002	0.993 ± 0
YPD SDS 0.2%	0.219 ± 0.011	0.871 ± 0.002	0.988 ± 0	0.877 ± 0.002	0.987 ± 0
YPD anisomycin 50 ug/ml	0.226 ± 0.007	0.788 ± 0.003	0.974 ± 0.001	0.796 ± 0.003	0.976 ± 0.001
YPD nystatin 10 ug/ml	0.312 ± 0.007	0.788 ± 0.003	0.955 ± 0.001	0.796 ± 0.003	0.956 ± 0.003
YP acetate 2%	0.232 ± 0.005	0.767 ± 0.003	0829 ± 0.006	0.771 ± 0.003	0827± 0.006
YP xylose 2%	0.258 ± 0.006	0.865 ± 0.003	0.985 ± 0	0.861 ± 0.003	0.985 ± 0.001
YP ribose 2%	0.266 ± 0.005	0.829 ± 0.003	0.983 ± 0	0.822 ± 0.003	0.981 ± 0
YPD NaCl 1.5M	0.279 ± 0.009	0.851 ± 0.002	0.989 ± 0	0.861 ± 0.002	0.985 ± 0.001
YPD NaCl 1 M	0.268 ± 0.005	0.898 ± 0.001	0.995 ± 0	0.903 ± 0.001	0.991 ± 0
YPD Mv 20 mM	0.272 ± 0.006	0.863 ± 0.002	0.989 ± 0	0.868 ± 0.002	0.985 ± 0.002
YP galactose 2%	0.349 ± 0.007	0.823 ± 0.002	0.978 ± 0.001	0.823 ± 0.002	0.979 ± 0.001
YPD anisomycin 20 ug/ml	0.346 ± 0.006	0.886 ± 0.001	0.993 ± 0	0.887 ± 0.001	0.9943 ± 0
YPD CHX 0.5 ug/ml	0.384 ± 0.004	0.88 ± 0.002	0.992 ± 0	0.881 ± 0.002	0.992 ± 0.003
YPD CHX 1 ug/ml	0.419 ± 0.007	0.779 ± 0.003	0.983 ± 0.001	0.783 ± 0.004	0.986 ± 0.001
YPD benomyl 200 ug/ml	0.419 ± 0.005	0.77 ± 0.003	0.973 ± 0	0.772 ± 0.003	0.977 ± 0
YPD 40°C	0.433 ± 0.005	0.897 ± 0.001	0.99 ± 0	0.898 ± 0.001	0.991 ± 0
YPD anisomycin 10 ug/ml	0.479 ± 0.006	0.92 ± 0.001	0.995 ± 0	0.921 ± 0.001	0.991 ± 0.001
YPD 42°C	0.517 ± 0.005	0.883 ± 0.001	0.986 ± 0	0.883 ± 0.001	0.982 ± 0
YPD CuSO4 10 mM	0.523 ± 0.005	0.897 ± 0.001	0.991 ± 0	0.898 ± 0.001	0.991 ± 0
YPD KCl 2M	0.563 ± 0.005	0.84 ± 0.001	0.992 ± 0	0.856 ± 0.002	0.99 ± 0.001
YPD benomyl 500ug/ml	0.599 ± 0.003	0.909 ± 0.001	0.996 ± 0	0.909 ± 0.001	0.991 ± 0
YPD caffeine 40 mM	0.656 ± 0.003	0.916 ± 0.001	0.996 ± 0	0.917 ± 0.001	0.992 ± 0
YPD caffeine 50 mM	0.67 ± 0.003	0.919 ± 0.001	0.996 ± 0	0.919 ± 0.001	0.997 ± 0

Supplementary Table 3. Predictive abilities estimated from five models across 35 traits: GBLUP, OBLUP, CBLUP, GOBLUP and GCBLUP.

Conditions	**GBLUP**	**OBLUP**	**CBLUP**	**GOBLUP**	**GCBLUP**
YPD formamide 5%	0.002 ± 0.007	0.284 ± 0.006	0.13 ± 0.008	0.281 ± 0.007	0.126 ± 0.008
YPD fluconazole 20 ug/ml	0.017 ± 0.007	0.376 ± 0.007	0.258 ± 0.01	0.375 ± 0.007	0.257 ± 0.01
YPD 14°C	0.053 ± 0.007	0.394 ± 0.008	0.308 ± 0.009	0.394 ± 0.008	0.306 ± 0.009
YPD hydroxyurea 30 mg/ml	0.078 ± 0.008	0.369 ± 0.007	0.241 ± 0.007	0.373 ± 0.007	0.238 ± 0.007
YPD formamide 4%	0.077 ± 0.007	0.284 ± 0.008	0.177 ± 0.007	0.282 ± 0.008	0.172 ± 0.007
YP ethanol 15%	0.121 ± 0.007	0.391 ± 0.009	0.311 ± 0.009	0.402 ± 0.009	0.338 ± 0.009
YP glycerol 2%	0.163 ± 0.007	0.363 ± 0.008	0.33 ± 0.008	0.371 ± 0.007	0.353 ± 0.008
YPD DMSO 6%	0.119 ± 0.006	0.306 ± 0.007	0.114 ± 0.007	0.316 ± 0.007	0.152 ± 0.007
YPD 6AU 600 ug/ml	0.106 ± 0.005	0.48 ± 0.006	0.426 ± 0.006	0.479 ± 0.006	0.424 ± 0.006
YPD EtOH 2%	0.146 ± 0.006	0.366 ± 0.007	0.26 ± 0.007	0.367 ± 0.007	0.26 ± 0.007
YP sorbitol 2%	0.179 ± 0.007	0.371 ± 0.007	0.377 ± 0.008	0.371 ± 0.007	0.389 ± 0.008
YPD sodium metaarsenite 2.5 mM	0.154 ± 0.006	0.501 ± 0.007	0.588 ± 0.009	0.501 ± 0.007	0.586 ± 0.009
YPD LiCl 250mM	0.129 ± 0.007	0.408 ± 0.009	0.45 ± 0.009	0.407 ± 0.009	0.45 ± 0.009
YPD SDS 0.2%	0.112 ± 0.006	0.436 ± 0.006	0.396 ± 0.006	0.44 ± 0.007	0.392 ± 0.007
YPD anisomycin 50 ug/ml	0.18 ± 0.007	0.418 ± 0.009	0.356 ± 0.008	0.421 ± 0.009	0.355 ± 0.008
YPD nystatin 10 ug/ml	0.18 ± 0.007	0.418 ± 0.009	0.356 ± 0.008	0.421 ± 0.009	0.355 ± 0.008
YP acetate 2%	0.18 ± 0.006	0.33 ± 0.006	0.185 ± 0.007	0.346 ± 0.006	0.248 ± 0.006
YP xylose 2%	0.236 ± 0.007	0.431 ± 0.007	0.433 ± 0.008	0.432 ± 0.007	0.448 ± 0.008
YP ribose 2%	0.261 ± 0.007	0.417 ± 0.007	0.409 ± 0.008	0.421 ± 0.007	0.424 ± 0.007
YPD NaCl 1.5M	0.147 ± 0.006	0.477 ± 0.006	0.409 ± 0.008	0.48 ± 0.006	0.419 ± 0.008
YPD NaCl 1 M	0.181 ± 0.006	0.516 ± 0.006	0.471 ± 0.007	0.518 ± 0.006	0.47 ± 0.007
YPD Mv 20 mM	0.174 ± 0.006	0.454 ± 0.006	0.438 ± 0.006	0.456 ± 0.006	0.437 ± 0.006
YP galactose 2%	0.199 ± 0.007	0.479 ± 0.006	0.429 ± 0.008	0.478 ± 0.006	0.427 ± 0.007
YPD anisomycin 20 ug/ml	0.232 ± 0.007	0.55 ± 0.007	0.507 ± 0.007	0.548 ± 0.007	0.504 ± 0.007
YPD CHX 0.5 ug/ml	0.286 ± 0.008	0.506 ± 0.008	0.489 ± 0.007	0.505 ± 0.008	0.488 ± 0.007
YPD CHX 1 ug/ml	0.299 ± 0.008	0.421 ± 0.01	0.387 ± 0.011	0.419 ± 0.01	0.382 ± 0.011
YPD benomyl 200 ug/ml	0.273 ± 0.007	0.411 ± 0.007	0.341 ± 0.008	0.421 ± 0.007	0.356 ± 0.007
YPD 40°C	0.237 ± 0.008	0.558 ± 0.005	0.486 ± 0.007	0.558 ± 0.006	0.484 ± 0.007
YPD anisomycin 10 ug/ml	0.26 ± 0.007	0.628 ± 0.005	0.57 ± 0.006	0.627 ± 0.005	0.57 ± 0.006
YPD 42°C	0.28 ± 0.007	0.587 ± 0.006	0.511 ± 0.007	0.586 ± 0.006	0.508 ± 0.007
YPD CuSO4 10 mM	0.268 ± 0.005	0.69 ± 0.004	0.72 ± 0.004	0.689 ± 0.004	0.719 ± 0.004
YPD KCl 2M	0.323 ± 0.007	0.522 ± 0.006	0.492 ± 0.007	0.538 ± 0.006	0.512 ± 0.007
YPD benomyl 500ug/ml	0.407 ± 0.007	0.706 ± 0.004	0.674 ± 0.005	0.708 ± 0.004	0.677 ± 0.005
YPD caffeine 40 mM	0.471 ± 0.009	0.698 ± 0.005	0.655 ± 0.005	0.698 ± 0.005	0.654 ± 0.005
YPD caffeine 50 mM	0.482 ± 0.009	0.697 ± 0.005	0.655 ± 0.005	0.695 ± 0.005	0.654 ± 0.005

References

Abraham, G., and Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. *Current Opinion in Genetics and Development* 33**,** 10-16.

Aherfi, S., Pagnier, I., Fournous, G., Raoult, D., La Scola, B., and Colson, P. (2013). Complete genome sequence of Cannes 8 virus, a new member of the proposed family "Marseilleviridae". *Virus Genes* 47(3)**,** 550-555.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* 16(4)**,** 197-212.

Bentley, S. (2009). Sequencing the species pan-genome. *Nature Reviews Microbiology* (7), 258–259.

Bergström, A., Simpson, J.T., Salinas, F., Barré, B., Parts, L., Zia, A., et al. (2014). A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution* 31(4)**,** 872-888.

Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.-L.V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature* 494(7436)**,** 234-237.

Botstein, D., and Fink, G.R. (2011). Yeast: an experimental organism for 21st Century biology. *Genetics* 189(3)**,** 695-704.

Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity by descent detection in population data. *Genetics* 84, 210–223.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics* 43(10)**,** 956-963.

Clifford, D., and McCullagh, P. (2014). The regress package. *R News* 6**,** 6.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science* 22(11)**,** 961-975.

D'Auria, G., Jiménez-Hernández, N., Peris-Bondia, F., Moya, A., and Latorre, A. (2010). Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics* 11(1)**,** 181-194.

de los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genetics* 9(7)**,** e1003608.

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., et al. (2010). Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology* 11(10)**,** R107.

Dunn, B., Richter, C., Kvitek, D.J., Pugh, T., and Sherlock, G. (2012). Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Research* 22, 908–924.

Erbe, M., Gredler, B., Seefried, F.R., Bapst, B., and Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS One* 8(12)**,** e81046. doi: 10.1371/journal.pone.0081046.

Evans, L.M., Tahmasbi, R., Vrieze, S.I., Abecasis, G.R., Das, S., Gazal, S., et al. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics* 50(5)**,** 737-745.

Fang, Y., Li, Z., Liu, J., Shu, C., Wang, X., Zhang, X., et al. (2011). A pangenomic study of Bacillus thuringiensis. *Journal of Genetics and Genomics* 38(12)**,** 567-576.

Fay, J.C. (2013). The molecular basis of phenotypic variation in yeast. *Current Opinion in Genetics and Development* 23(6)**,** 672-677.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics* 51, 1044–1051 doi: 10.1038/s41588-019-0410-2.

Georges, M., Charlier, C., and Hayes, B. (2018). Harnessing genomic information for livestock improvement. *Nature Reviews Genetics* 20, 135–156.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S., et al. (2004). Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics* 167(4)**,** 1791-1799.

Goddard, M., and Hayes, B. (2007). Genomic selection. *Journal of Animal breeding and Genetics* 124(6)**,** 323-330.

Goddard, M.E., Hayes, B.J., and Meuwissen, T.H. (2010). Genomic selection in livestock populations. *Genetics Research* 92(5-6)**,** 413-421.

Goddard, M.E., Wray, N.R., Verbyla, K., and Visscher, P.M. (2009). Estimating effects and making predictions from genome-wide marker data. *Statistical Science* 24(4)**,** 517-529.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274(5287)**,** 546-567.

González-Reymúndez, A., de los Campos, G., Gutiérrez, L., Lunt, S.Y., and Vazquez, A.I. (2017). Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *European Journal of Human Genetics* 25(5)**,** 538-544. doi: 10.1038/ejhg.2017.12.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307(5714)**,** 1434-1440.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., et al. (2017). Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature Communications* 8(1)**,** 2184-2195.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics* 129(12)**,** 2413-2427. doi: 10.1007/s00122-016-2780-5.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 9(1)**,** R7.

Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* 4(2)**,** e1000008.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12(1)**,** 491-505.

Hu, P., Yang, M., Zhang, A., Wu, J., Chen, B., Hua, Y., et al. (2011). Comparative genomics study of multi-drug-resistance mechanisms in the antibiotic-resistant Streptococcus suis R61 strain. *PLoS One* 6(9)**,** e24988.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.K.K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnology Journal* 16(7)**,** 1265-1274.

Kim, H., Grueneberg, A., Vazquez, A.I., Hsu, S., and de los Campos, G. (2017). Will big data close the missing heritability gap? *Genetics* 207(3)**,** 1135-1145.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *The American Journal of Human Genetics* 76(1)**,** 8-32.

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 361(1475)**,** 1929-1940.

Kumar, A., and Snyder, M. (2001). Emerging technologies in yeast genomics. *Nature Reviews Genetics* 2(4)**,** 302-312.

Lapierre, P., and Gogarten, J.P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics* 25(3)**,** 107-110.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14)**,** 1754-1760.

Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* 32(10)**,** 1045-1052.

Li, Z., Simianer, H., and Martini, J.W. (2019). Integrating gene expression data into genomic prediction. *Frontiers in Genetics* 10**,** 126-137.

Märtens, K., Hallin, J., Warringer, J., Liti, G., and Parts, L. (2016). Predicting quantitative traits from genome and phenome with near perfect accuracy. *Nature Communications* 7**,** 11512-11520.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News* 456(7218)**,** 18-21.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature* 542(7640)**,** 186-190.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* 18**,** 31-36.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099)**,** 1190-1195.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4)**,** 1819-1829.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.K.K., et al. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal* 90(5)**,** 1007-1013.

Otto, T.D., Dillon, G.P., Degrave, W.S., and Berriman, M. (2011). RATT: rapid annotation transfer tool. *Nucleic Acids Research* 39(9)**,** e57-e57.

Pérez, P., and de Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198(2)**,** 483-495.

Paradis, E., and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3)**,** 526-528.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* 556(7701)**,** 339-344.

Proux-Wéra, E., Armisén, D., Byrne, K.P., and Wolfe, K.H. (2012). A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics* 13(1)**,** 237-249.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81(3)**,** 559-575.

Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics* 7(11)**,** 862-872.

Schaeffer, L. (2006). Strategy for applying genome‐wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* 123(4)**,** 218-223.

Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics* 51(1)**,** 30-39.

Sieber, P., Platzer, M., and Schuster, S. (2018). The Definition of Open Reading Frame Revisited. *Trends in Genetics* 34(3)**,** 167-170.

Skelly, D.A., Merrihew, G.E., Riffle, M., Connelly, C.F., Kerr, E.O., Johansson, M., et al. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Research* 23, 1496–1504.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics* 91(6)**,** 1011-1021.

Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M., et al. (2015). The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research* 25, 762–774

Team, R.C. (2013). R: A language and environment for statistical computing.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences* 102(39)**,** 13950-13955.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91(11)**,** 4414-4423. doi: 10.3168/jds.2007-0980.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23**,** 148-154.

Walker, F.O. (2007). Huntington's disease. *The Lancet* 369(9557)**,** 218-228.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557**,** 43-49.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013a). Author reply to A commentary on Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14(12)**,** 894.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013b). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14(7)**,** 507-15.

Yan, H., Dobbie, Z., Gruber, S.B., Markowitz, S., Romans, K., Giardiello, F.M., et al. (2002). Small changes in expression affect predisposition to tumorigenesis. *Nature Genetics* 30(1)**,** 25-26.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42(7)**,** 565-569.

Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics* 49(9)**,** 1304-1310.

Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., et al. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics* 49(6)**,** 913-924.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics* 50(2)**,** 278-284.

5th CHAPTER

General discussion

General discussion

It was long presumed that single nucleotide polymorphisms (SNPs) represent the majority of genetic variation across individuals. Animal and plant breeding increasingly benefit from the implementation of genomic selection (GS), and the increasing availability of SNPs supports the advanced method. SNPs have also been used to trace genes that are undergoing selective sweeps or to observe population structure variation (Cavanagh et al., 2013). However, the critical role of structural variations (SVs) is becoming increasingly acknowledged (Wendel et al., 2016). SVs are defined as large sequence variation (> 1 kb) such as insertions, duplications, copy number variants, deletions and translocations in the genome (Feuk et al., 2006). In this study, we utilized two types of SVs which are deemed to prevailingly contribute to genomic and phenotype variation: copy number variants (CNVs), sequences that are present in different copy numbers among individuals, and presence/absence variants (PAVs), sequences that are present in some individuals but absent in others (Marroni et al., 2014). Although the contribution of CNVs and PAVs to genome and phenotype diversity is significant, these structural variants in many genomic sequences have no significant phenotypic consequence (Sebat et al., 2004). Nevertheless, gene dosage can cause genetic diseases, either alone or in combination with other genetic or environmental factors (Inoue and Lupski, 2002). We used pan-genomic presence/absence of ORFs and copy number of ORFs that combined both SVs and gene dosage information in genomic prediction, which excluded non-causal SVs in the process of prediction. Our results demonstrate that presence/absence of ORFs and copy number of ORFs have a dominant impact on phenotype variation. Similar conclusions have also been drawn in a *Brachypodium distachyon* pan-genome study where differentially present genes contribute substantially to the understanding of population genetics and phenotypic variation within a eukaryotic species (Gordon et al., 2017). When using pan-genomic ORFs in genomic prediction, we exclusively picked dispensable ORFs as predictors, since core ORFs present in all isolates will not affect prediction accuracy. A recent study compared the predicted biological functions of core and dispensable pan-genes, and revealed that core genes are enriched for essential cellular processes (e.g. glycolysis), whereas the dispensable genes are not indispensable for survival, since they could be absent in at least one individual (Marroni et al., 2014). The dispensable genes are enriched for functions that may be advantageous in some environments (e.g. disease resistance, gene regulation). The observed enrichment of dispensable genes with putative adaptive functions in that study suggests that dispensable genes are preferentially retained when they acquire functions that confer benefits under certain circumstances. Therefore, they may contribute to phenotypic variation that could be of particular interest for animal and plant breeding and evolutionary studies of adaptive traits (Marroni et al., 2014). A *Brassica napus* pan-genome study proved that the main cause of gene presence/absence variation is homoeologous exchange (HE), and demonstrated their considerable association with agronomic traits (Hurgobin et al., 2018). The meiotic chromosome pairing that occurs between homoeologous chromosomes leads to increased homoeologous exchanges and gene conversion events. These HE-related PAV events are useful to understand the association between genomic structural rearrangement and phenotypic variation, particularly the role of genome duplications or deletions spanning genes with trait-related dosage effects (Hurgobin et al., 2018).

Gene expression data has been suggested to be a valuable resource for phenotype prediction (Guo et al., 2016). The heritable part of genome-wide gene expression variation was first assessed in a cross population of *Saccharomyces cerevisiae* (Brem et al., 2002), indicating a substantial genetic component in transcriptional variation in yeast (Skelly et al., 2009). Furthermore, it has been proven that non-additivity is common in *D. melanogaster* (Huang et al., 2012), *A. thaliana* and maize (Vuylsteke et al., 2005), and that its extreme forms, overdominance and underdominance, are common (Gibson et al., 2004). These heritable components have potential to be utilized for complex traits prediction. However, gene expression can be greatly affected by the tissue sampled and time of measurement. Thousands of genes are differentially expressed between tissues or show tissue preferential expression (Melé et al., 2015). In addition, some gene expression products have "housekeeping" functions, and are therefore expressed in all cells, while other genes are expressed in a tissue-specific manner (Fagerberg et al., 2014). It has been found that variation in gene expression is even far greater among tissues (47% of total variance in gene expression) than among individuals (4% of total variance) (Melé et al., 2015). Hence, it is important to use gene expression data from specific tissues for phenotype prediction.

Overall, the thesis focuses on two critical problems in quantitative genetics: prediction of genetic values or phenotypes, and estimation of heritability. In the multi-omics era, we verified that gene expression data, especially tissue-specific gene expression data, can be integrated into genomic prediction, can be regarded as a complementary information for prediction of phenotype. At the gene level, we first explored pan-genomic ORFs to be a potential substitution of SNPs in prediction of genetic value and estimation of heritability. The valuable resources will play an important role in understanding the diversity of the genome and the genetic architecture of complex traits, and then accelerate the breeding process. In a human genetics' context, omics data-based prediction may have the potential to more accurately identify individuals that are at risk for diseases, and to improve the preventive medicine strategies and clinical decision making.

References

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296(5568)**,** 752-755.

Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics* 13(2)**,** 397-406.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. *Nature Reviews Genetics* 7(2)**,** 85-97.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S., et al. (2004). Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics* 167(4)**,** 1791-1799.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., et al. (2017). Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature communications* 8(1)**,** 2184-2197.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics* 129(12)**,** 2413-2427. doi: 10.1007/s00122-016-2780-5.

Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R., Ayroles, J.F., et al. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences* 109(39)**,** 15553-15559. doi: 10.1073/pnas.1213423109.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.K.K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant biotechnology journal* 16(7)**,** 1265-1274.

Inoue, K., and Lupski, J.R. (2002). Molecular mechanisms for genomic disorders. *Annual review of genomics and human genetics* 3(1)**,** 199-242.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* 18**,** 31-36.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., et al. (2015). The human transcriptome across tissues and individuals. *Science* 348(6235)**,** 660-665.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science* 305(5683)**,** 525-528.

Skelly, D.A., Ronald, J., and Akey, J.M. (2009). Inherited variation in gene expression. *Annual review of genomics and human genetics* 10**,** 313-332.

Vuylsteke, M., Van Eeuwijk, F., Van Hummelen, P., Kuiper, M., and Zabeau, M. (2005). Genetic analysis of variation in gene expression in Arabidopsis thaliana. *Genetics* 171(3)**,** 1267-1275.

Wendel, J.F., Jackson, S.A., Meyers, B.C., and Wing, R.A. (2016). Evolution of plant genome architecture. *Genome biology* 17(1)**,** 17-37.

Acknowledgements

Prof. Dr. Henner Simianer. Many thanks for supervising my PhD project. Your patience for my academically immature thoughts gives me valuable experience in many aspects of life.

Prof. Dr. Armin Schmitt and Prof. Dr. Thomas Kneib. Thanks for being my second and third supervisors.

Ms. Döring. Thanks for always helping me whenever I need help.

Frau Amudha. Really happy and lucky for sharing office with you.

My colleagues. Thank you for all the nice discussions and supports during my PhD project.

My friends. Mengyu Tu, Shasha Shen, Ning gao, Yao Wang, Quan Liu, Yixuan Xing, Shuwen Shan, Mengmeng Han, Yi Zhang, Haitao Wang, Fangzheng Xu, and my football team. Thanks for making my life wonderful in Göttingen.

My parents. Your endless love will always be my courage to face challenges.