PLOS GENETICS

RESEARCH ARTICLE

Pan-genomic open reading frames: A potential supplement of single nucleotide polymorphisms in estimation of heritability and genomic prediction

Zhengcao Li[1,2]*, **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 Sciences, Sun Yat-Sen University, Guangzhou, China

* lizhc7@mail.sysu.edu.cn

Check for updates

OPEN ACCESS

Citation: Li Z, Simianer H (2020) Pan-genomic open reading frames: A potential supplement of single nucleotide polymorphisms in estimation of heritability and genomic prediction. PLoS Genet 16(8): e1008995. https://doi.org/10.1371/journal.pgen.1008995

Editor: Wen Huang, Michigan State University, UNITED STATES

Received: December 27, 2019

Accepted: July 15, 2020

Published: August 24, 2020

Copyright: © 2020 Li, Simianer. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All files are available from the 1002 yeast genome website http://1002genomes.u-strasbg.fr/files/.

Funding: 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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Pan-genomic open reading frames (ORFs) potentially carry protein-coding gene or coding variant information in a population. In this study, we suggest that pan-genomic ORFs are promising to be utilized in estimation of heritability and genomic prediction. A *Saccharomyces cerevisiae* dataset with whole-genome SNPs, pan-genomic ORFs, and the copy numbers of those ORFs is used to test the effectiveness of ORF data as a predictor in three prediction models for 35 traits. Our results show that the ORF-based heritability can capture more genetic effects than SNP-based heritability for all traits. Compared to SNP-based genomic prediction (GBLUP), pan-genomic ORF-based genomic prediction (OBLUP) is distinctly more accurate for all traits, and the predictive abilities on average are more than doubled across all traits. For four traits, the copy number of ORF-based prediction(CBLUP) is more accurate than OBLUP. When using different numbers of isolates in training sets in ORF-based prediction, the predictive abilities for all traits increased as more isolates are added in the training sets, suggesting that with very large training sets the prediction accuracy will be in the range of the square root of the heritability. We conclude that pan-genomic ORFs have the potential to be a supplement of single nucleotide polymorphisms in estimation of heritability and genomic prediction.

Author summary

The properties of single nucleotide polymorphisms (SNPs) as a main source of genetic variability for estimation of heritability and genomic prediction have been widely studied over the past years. This data type remarkably accelerated the development of medical diagnosis in human genetics and prediction of breeding values in livestock and crop breeding field. However, due to the inherent pitfalls of SNP-based prediction, e.g. imperfect LD between markers and causal variants, seeking new genomic datasets of causal variants has become imperative. Our study point out some of the superiorities of pan-

Competing interests: The authors have declared that no competing interests exist.

genomic open reading frames as independent variables in estimation of heritability and genomic prediction.

Introduction

Genome-wide single nucleotide polymorphisms (SNPs) were first proposed in 2001 to be used for predicting genetic values [1]. Implementation in practice became pervasive due to the large amount of SNPs that swiftly became available [2]. By utilizing genome-wide SNP data, 'genomic selection' based on genomically predicted breeding values has triggered a revolution in animal and plant breeding. It improved the genetic progress by reducing generation intervals or increasing predictive ability of breeding values [3–5]. In human genetics, genomic prediction aimed at accurately quantifying disease risk so that preventative measures can be taken earlier [6]. However, SNP markers are normally not causal variants. In genomic prediction the causal variant effects are estimated indirectly by modeling SNP makers that are in linkage disequilibrium (LD) with them [2]. 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 [7]. For prediction of distantly related individuals, even if high density SNP markers were used, the prediction accuracy still can be very low [8]. Likewise, genome-wide SNP data are also used for estimation or dissection of genetic parameters, such as SNP-based heritability [9]. Several factors inevitably cause the 'still missing heritability' problem when using common SNPs exceeding a certain minor allele frequency (MAF) to estimate narrow sense heritability [10]: for instance, the causal variants may not be 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 [11, 12]

Pan-genomic open reading frames (ORFs) potentially hold whole-genome protein-coding gene or coding variant information in a population. An ORF is commonly defined as a sequence that has a length divisible by three and begins with a translation start codon and ends at a stop codon. However, a review paper suggests it is bounded by stop codons, since such difinition distinguishes precisely between ORF, exon, and coding sequence (CDS) [13]. An ORF is a sequence region that is 'open' for translation, and an indicator for a potential protein-coding gene [13]. The detection of ORFs is of central importance in finding protein-coding genes in genomic sequences. Different individuals may carry partially different sets of genes or ORFs. The 'pan-genome' denotes the set of all genes or ORFs present in the genomes of a group of organisms, usually a species [14, 15]. The concept has been applied to bacterial [16], viral [17], plant [18–20], fungal [21], and human genome studies [22]. Series of pan-genomic studies were performed when studying genomic dynamics [23], pathogenesis and drug resistance [24, 25], and species evolution [26].

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" [27]. It has previously been shown to be a good tool for exploring the genotype–phenotype relationship via linkage mapping [28], and the study of "missing heritability" [29]. Importantly, *S. cerevisiae* is an informative predictor of human gene function: nearly 50% of human genes implicated in heritable diseases have a yeast homologue [30], which makes *S. cerevisiae* a model species in the studies for prediction of human disease [31]. 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 [32]. In this study, we used *S. cerevisiae* pan-genomic open reading

frames in genomic prediction, which represent 7,796 non-redundant ORFs, 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) 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, and (2) genomic prediction capitalizing on ORF data and CNO data performed substantially better than that using genome-wide SNP data.

Results and discussion

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 [33]. Based on the first three principal components, each type of dataset showed a diverse genetic structure of the *S. cerevisiae* isolates (Fig 1). 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%). There are only few isolates for both ORF data which are outliers compared to SNP data, and less than 1% of the phenotypic variance accounted for by either class of ORF genotypes is driven by these outlier strains. When we excluded these outlier strains, the predictive abilities remained the same as with all strains. Analogously, three neighbor-joining trees based on the three types of data were constructed (Fig 1). 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. Five of the 'outlier' isolates in ORF-based and CNO-based neighbor-joining trees overlapped. For the SNP-based neighbor-joining tree, the genetic distances among most isolates were relatively large. The heat maps of genetic covariance matrices: *G*, *O*, *C* constructed using the three types of SNP and ORF data are shown in Fig 1, where the yeast strains were in the same order on the basis of their geographical origins in the three matrices. The purple color blocks, indicating high covariance, in the SNP-based genetic covariance matrix were in different positions compared with the purple color blocks in the other two genetic covariance matrices. The purple color blocks in the ORF-based and CNO-based genetic covariance matrices shared similar positions along the diagonal, but compared to the ORF-based genetic covariance matrix, the CNO-based genetic covariance matrix has more blocks indicating high similarity in the off-diagonal regions.

Capturing "still missing heritability"

'Missing heritability' has been identified as 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 [34]. 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 prefiltering for significance in GWAS could capture a larger part, but still not all of the missing heritability [11]. However, the estimation of SNP-based heritability depended on the extent 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

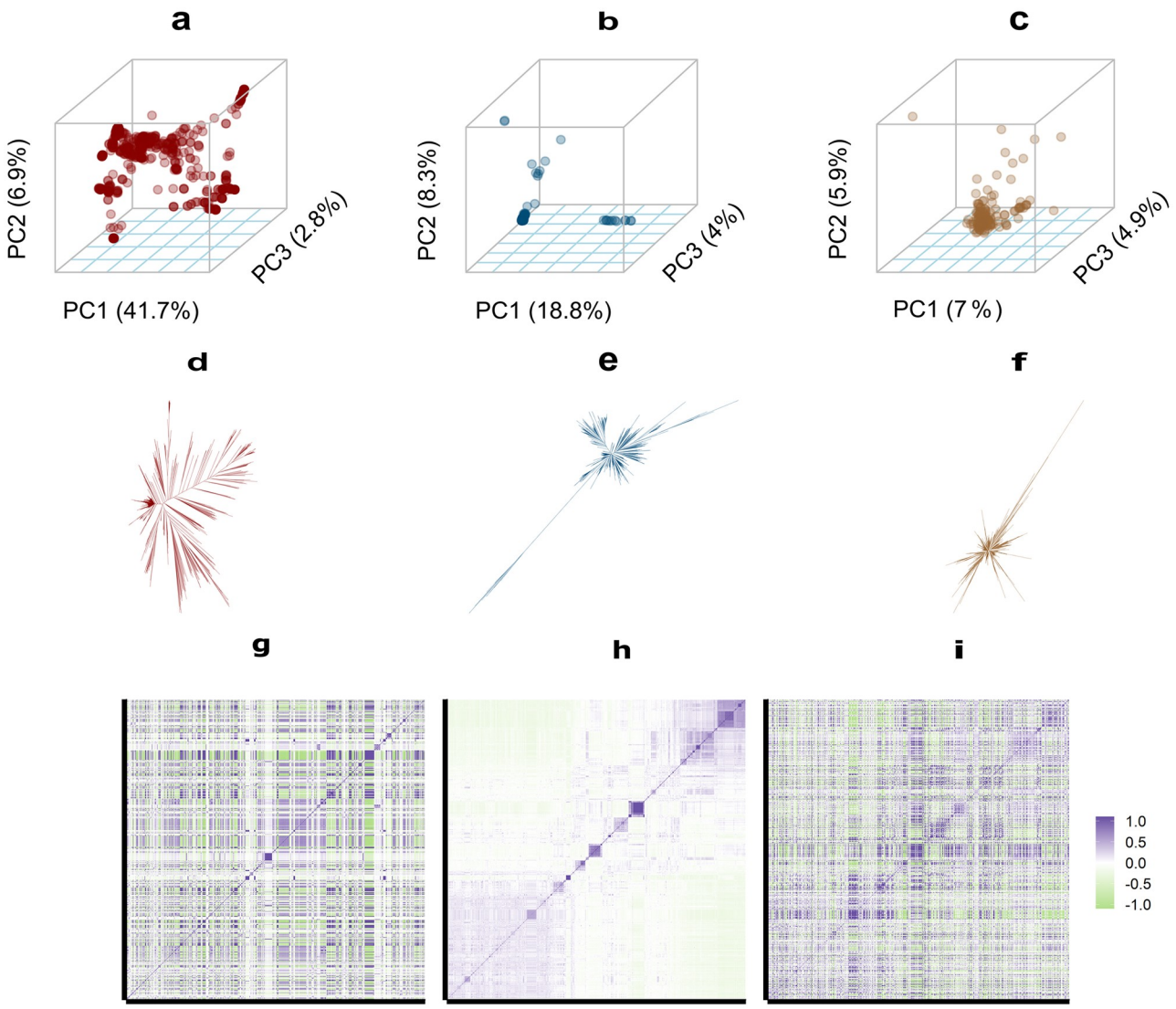

Fig 1. Principal components analysis, neighbor-joining trees and heatmaps of genetic covariance matrices. Panels **a**, **b** and **c** represent principal component (PC) analysis for all common SNPs, pan-genomic ORFs, and the copy numbers of pan-genomic ORFs on 787 diploid *S. cerevisiae* isolates, respectively. PC1, PC2, and PC3 denote the first three principal components. Panels **d**, **e** and **f** represent the neighbor-joining trees of 787 diploid *S. cerevisiae* constructed using the three types of dataset. Panels **g**, **h** and **i** display heatmaps of genetic covariance matrices 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.

https://doi.org/10.1371/journal.pgen.1008995.g001

heritability" [10]. we used SNPs with MAF $\geq$ 0.01, ORFs with frequency $\geq$ 0.05 and CNOs with frequency $\geq$ 0.05 to estimate heritabilities, respectively. Our results show that the SNP-based heritability ($\hat{h}^2_G$) was 0.281 on average across all traits, ranging from 0.004 to 0.67 (Fig 2 and S1 Table), and the ORF-based heritability ($\hat{h}^2_O$) on average across all traits was 0.761, ranging from 0.623 to 0.9, which indicates that pan-genomic ORFs hold more causal variant information than common SNPs in the population. Besides, pan-genomic ORFs were able to capture a major part of the "still missing heritability" for all studied traits, and encompass most of the repertoire of genes or coding variants accessible in the yeast population. This provides evidence that most of the genetic variation of complex traits is additive by nature and

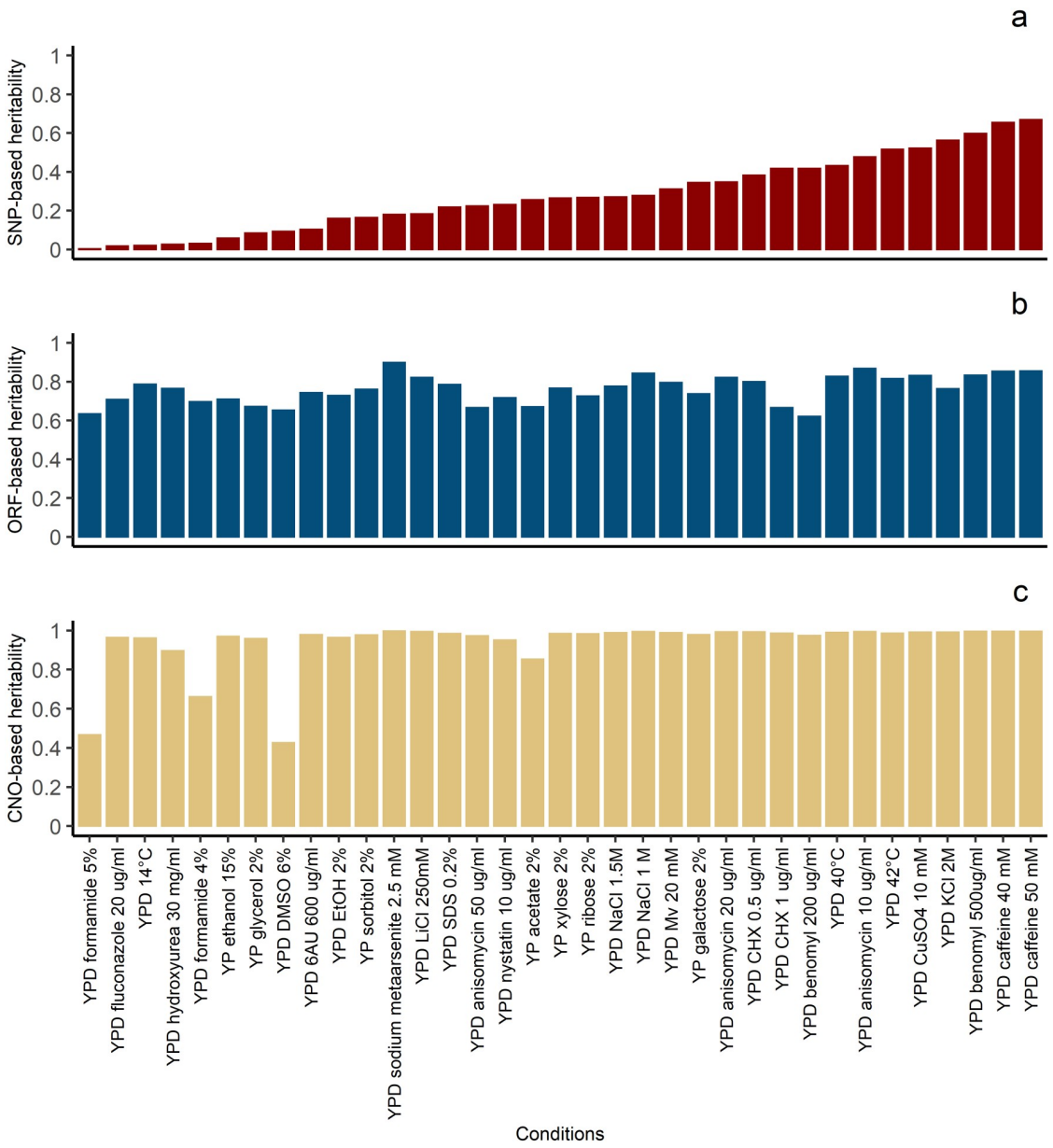

Fig 2. Heritability estimates for all 35 traits estimated based on three data types. Panel **a**, **b**, and **c** depict heritability estimates for all 35 traits estimated based on SNPs with MAF $\geq$ 0.01, pan-genomic open reading frames with frequency $\geq$ 0.05, and the copy numbers of pan-genomic open reading frames with frequency $\geq$ 0.05, respectively. All standard errors were below 0.02.

https://doi.org/10.1371/journal.pgen.1008995.g002

can be captured by a linear model [35]. The CNO-based heritability ($\hat{h}^2_C$) was 0.935 on average across all traits, ranging from 0.445 to 0.996 (Fig 2), and $\hat{h}^2_C$ captured more "missing heritability" compared with $\hat{h}^2_O$ in 32 of 35 traits. The reason could be that ORF copy numbers reflect a variable number of repeats of some genes with a direct effect on the expression intensity of the related gene product. An example of a complete gene repeat is the copy number polymorphism of human alpha-amylase 1 gene (AMY1), which is directly associated with the amount of salivary amylase and significantly varied between populations with different diets [36]. Another example is the correlation between the copy number of the chemokine gene CCL3L1 and

susceptibility to HIV/AIDS, based on significant inter-individual and inter-population differences in the copy number of a segmental duplication encompassing the gene encoding CCL3L1 (MIP-1αP) [37]. In addition, $\hat{h}^2_C$ exceeded 0.98 for 20 of the 35 traits, showing that 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 $\hat{h}^2_C$ was substantially lower than $\hat{h}^2_O$. One possible explanation is that for these three traits ORF repeats were not functional, and thus using the copy number of ORF data presumably added noise in the estimation of genetic variance (S1 Table).

When 1'625'809 SNPs were used in estimation of $\hat{h}^2_G$, $\hat{h}^2_G$ increased for 25 traits (S1 Fig), predictive abilities increased in only 8 of 35 traits (S2 Fig). For the remaining 27 traits, the predictive abilities decreased, which suggests that rare SNPs might have caused overestimation of heritability for the majority of traits in this population. This phenomenon is also observed in estimation of $\hat{h}^2_O$. When all ORFs were included, $\hat{h}^2_O$ for all traits increased (S1 Fig), but the predictive abilities for all traits remained the same as with ORFs with frequency ≥ 0.05 (S3 Fig), which suggests that rare ORFs inflate heritability estimates. The square root of $\hat{h}^2_G$ and SNP-based predictive abilities across traits are highly positively correlated. Similarly, for ORF data, the square root of heritability and predictive ability are also remarkably correlated (Fig 3). Nevertheless, the square root of ORF-based heritabilities and ORF-based predictive abilities have a more linear relationship than the result with SNP data (Fig 3), which might be because relationships between training and test set are better explained with ORFs. In addition, the predictive abilities of OBLUP and CBLUP for many traits are higher than the square root of SNP-based heritability (Fig 3). This can be seen as another piece of evidence that ORFs and CNOs capture a part of "missing heritability". We notice that the values of $\hat{h}^2_C$ for most traits are similar and very high, leaving little room for residuals. To verify whether this phenomenon was caused by rare CNOs, we excluded 1471 CNOs with frequency <0.05. However, the $\hat{h}^2_C$ and CNO-based predictive abilities for all traits remained the same as with all CNOs (S4 Fig), suggesting that rare CNOs are not causal for the apparent bias in heritability estimation. Interestingly, we see "missing" heritability when using SNPs in the model, using ORF or CNO data seems to generate "phantom" heritability, which, however, doesn't appear to have an adverse effect on predictive ability of the models.

Improvement of predictive abilities

Precision of SNP-based genomic prediction depends on two factors: SNP-based heritability and the accuracy with which the SNP marker effects are estimated [38]. The square root of the SNP-based heritability provides the upper bound of predictive ability for SNP-based genomic prediction, and this upper bound can be approached when big sample sizes are used for model training [39]. However, the inherent limitation of SNP-based genomic prediction is the extent of LD between SNP markers and causal variants. If causal variants are in low LD with the used set of SNPs, additive genetic effects would be underestimated [11] [40], and the SNP-based heritability would be lower than narrow-sense heritability whose square root is the ultimate upper bound of predictive ability when genetic variance explained by all additive effects is captured. Since there is no perfect LD between causal variants and SNPs, e.g. when rare variants are not captured by common SNPs [10], 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

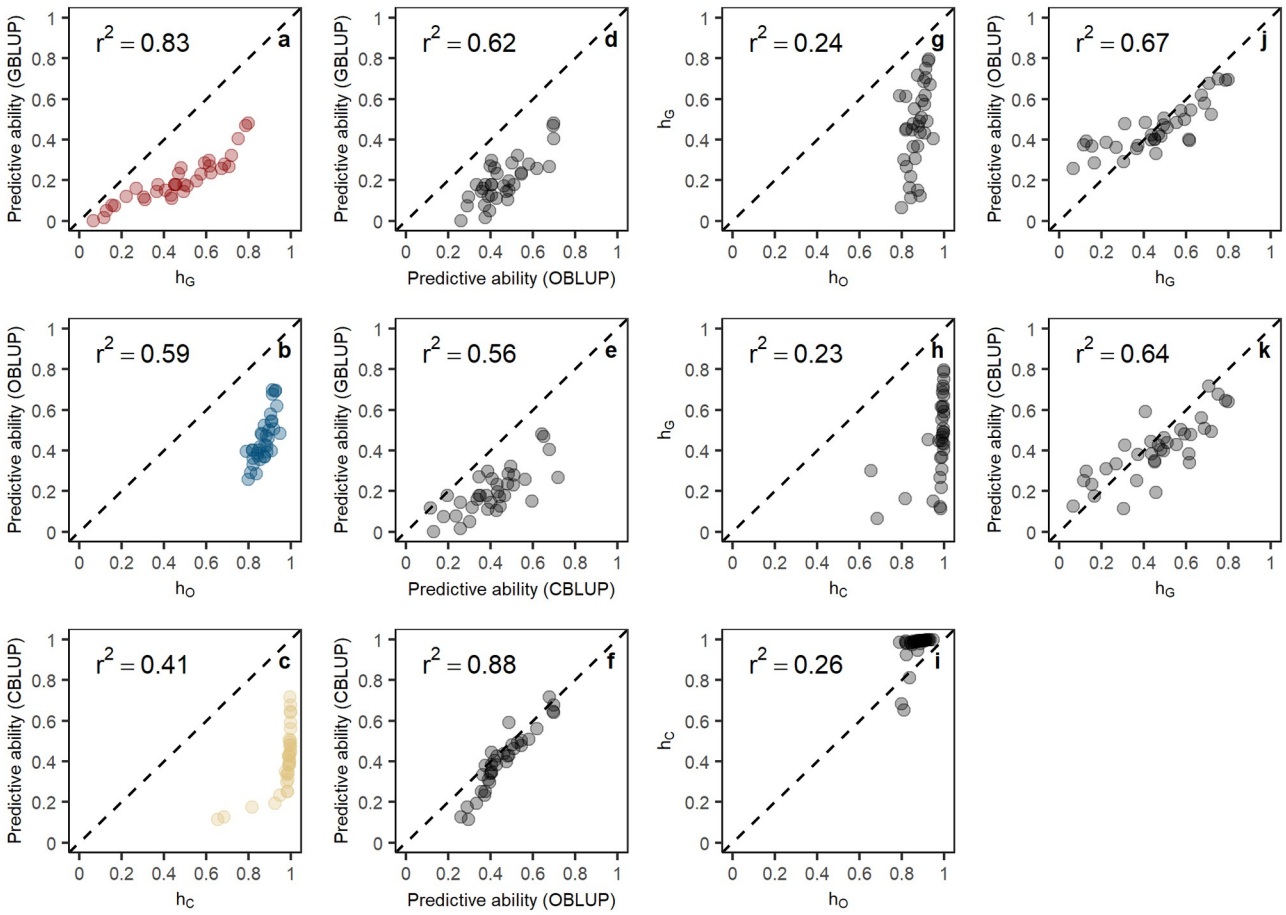

Fig 3. Correlations between predictive abilities and the square root of heritabilities. **(a)** The correlation between predictive abilities of GBLUP and the square root of SNP-based heritabilities across all traits; **(b)** The correlation between predictive abilities of OBLUP and the square root of ORF-based heritabilities across all traits; **(c)** The correlation between predictive abilities of CBLUP and the square root of CNO-based heritabilities across all traits; **(d)** The correlation between predictive abilities of GBLUP and predictive abilities of OBLUP across all traits; **(e)** The correlation between predictive abilities of GBLUP and predictive abilities of CBLUP across all traits; **(f)** The correlation between predictive abilities of CBLUP and predictive abilities of OBLUP across all traits; **(g)** The correlation between the square root of SNP-based heritabilities and the square root of ORF-based heritabilities across all traits; **(h)** The correlation between the square root of SNP-based heritabilities and the square root of CNO-based heritabilities across all traits; **(i)** The correlation between the square root of CNO-based heritabilities and the square root of ORF-based heritabilities across all traits; **(j)** The correlation between predictive abilities of OBLUP and the square root of SNP-based heritabilities across all traits;**(k)** The correlation between predictive abilities of CBLUP and the square root of SNP-based heritabilities across all traits; h_G, h_O, h_C represent the square root of SNP-based heritabilities, ORF-based heritabilities, CNO-based heritabilities, respectively. r^2 depicts the coefficient of determination. The dots in the 9 panels depict the 35 traits.

https://doi.org/10.1371/journal.pgen.1008995.g003

prediction accuracy by increasing the training set size [41]. Thus, it is necessary to explore new sources of predictors to overcome the imperfection.

The 'pan-genome' denotes the set of all genes or ORFs present in the genomes of a group of organisms [16, 42], which 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 using this gene level structure variation set as a supplement of SNPs in genomic prediction will 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 [43]. Coding variants falling within a coding region, especially non-synonymous variants, may change amino acid sequences, and then lead to phenotype variations [44]. In our results, GBLUP as a reference method provided predictive abilities ranging from 0 to 0.48

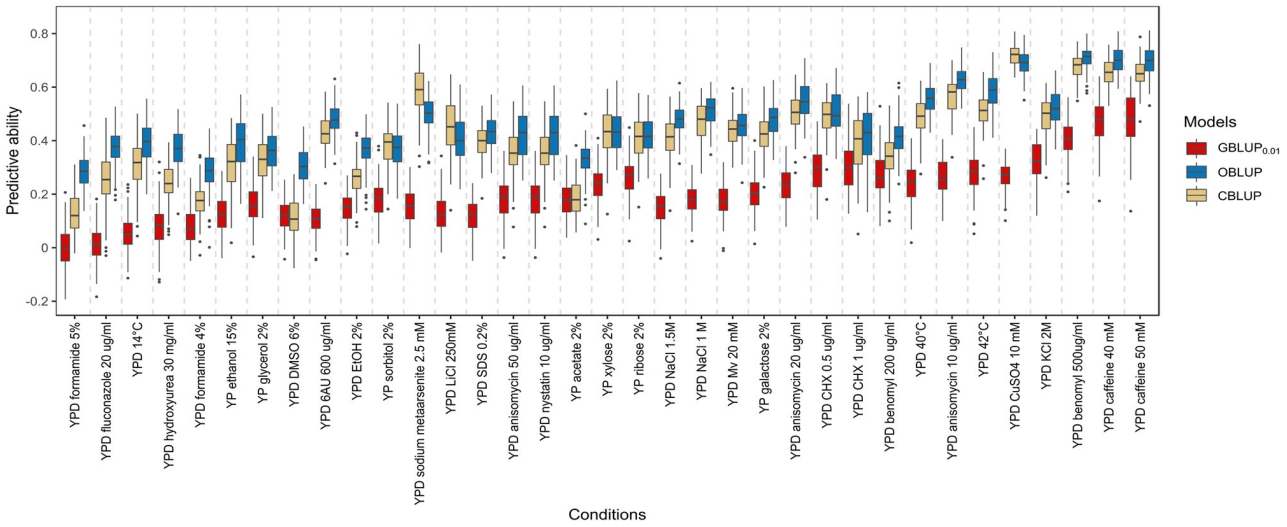

Fig 4. Predictive abilities of three models across 35 traits. $GBLUP_{0.01}$ using all common SNPs, OBLUP using pan-genomic open reading frames, and CBLUP using copy numbers of pan-genomic open reading frames.

https://doi.org/10.1371/journal.pgen.1008995.g004

across the studied traits (Fig 4, S2 Table and S5 Fig). Compared to GBLUP, pan-genomic ORF-based prediction (OBLUP) was more accurate for all traits: observed predictive abilities ranged from 0.28 to 0.71, more than doubled on average across all traits, which manifested the distinct advantage of making use of pan-genomic ORF data in genomic prediction (S6 Fig). When using different numbers of isolates in training sets between $n = 200$ and $n = 600$ in steps of 50 for ORF-based prediction, the predictive abilities of all traits increased as the number of isolates in the training set increased (Fig 5), confirming that increasing the training set size yields more accurately estimated ORF effects. We fitted for each phenotypic trait the function (2): $r = w\sqrt{\frac{n\hat{h}_O^2}{n\hat{h}_O^2+Me}}$ to the ORF data, in which r is the observed predictive ability of OBLUP for this trait, w is the maximum predictive accuracy with infinite training set size, n is the number of isolates in the training set, $\hat{h}_O^2$ is the ORF-based heritability estimate, and Me is the number of independent chromosome segments [41]. The parameter w is the asymptote of the function given above for $n \rightarrow \infty$ and thus an extrapolation revealing an estimate of the maximum achievable predictive ability. This model provided an perfect fit of realized prediction accuracies when applied to dairy cattle data [41]. For 28 of the studied traits, the optimal estimate of w was in the range of the square root of the heritability of the trait. For seven traits the maximum likelihood optimization yielded estimates of w and Me which were far beyond the expected range (Fig 5). The distribution of the estimates for w and Me as well as the correspondence of the estimates of w for the 28 traits where the estimates of r were <1 are in Fig 5. The medians of the estimates of w and Me were 0.71 and 293, respectively. When varying the value of the heritability used in ORF-based prediction we observed that this had almost no effect on the prediction accuracy.

It is noteworthy 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 [45]. Such non-coding variants may have effects in the gene expression process, such as transcription factor binding, DNA methylation, and mRNA

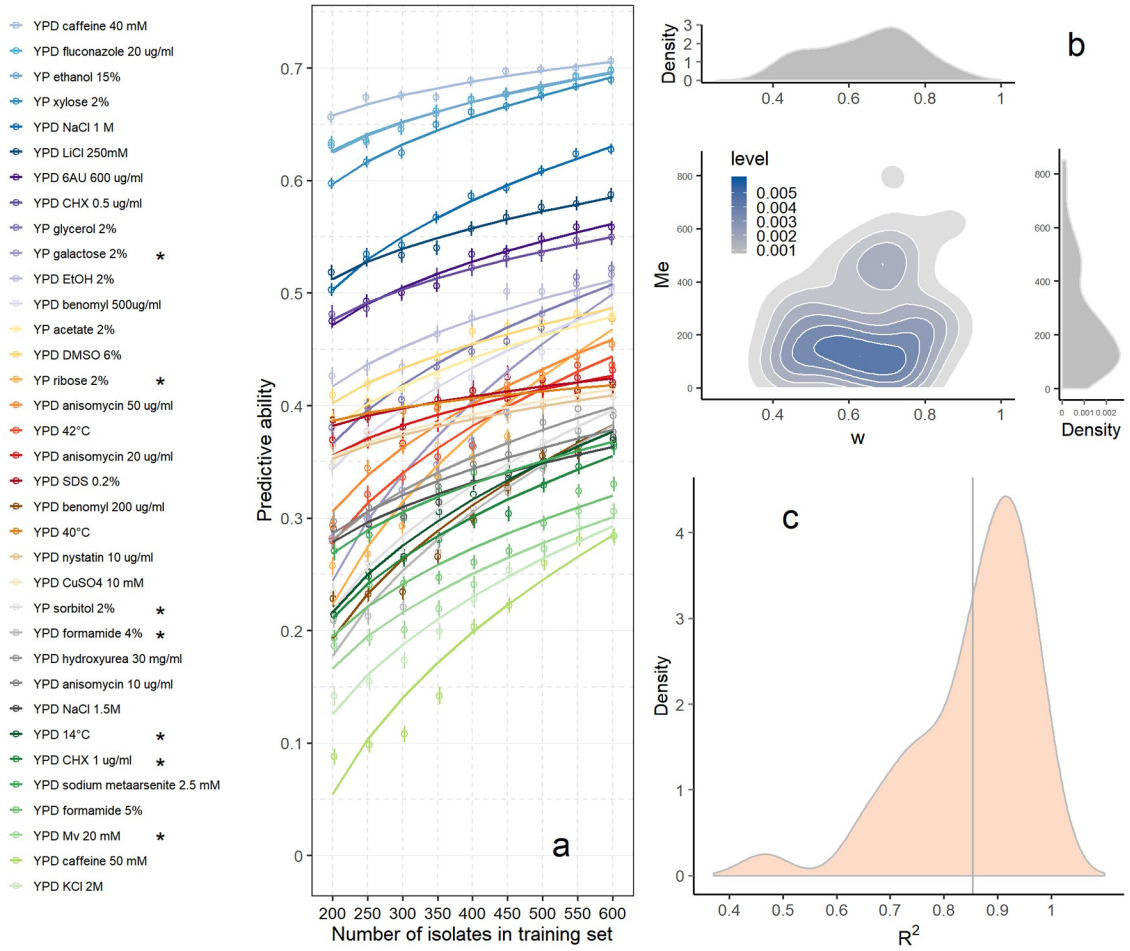

Fig 5. The predictive abilities of ORF-based genomic prediction for 35 traits using different numbers of isolates in training sets. (a) When using different numbers of isolates in training sets between $n = 200$ and $n = 600$ in steps of 50 for ORF-based prediction, the predictive abilities of all traits increased as the number of isolates in the training set increased. The solid curves are fitted lines that correspond to the function (2), where r represents the predictive ability in this study (details see text). Traits for which the fitting of the function produced outliers ($\hat{r} > 1$) are marked by an asterisk (*). **(b)** The joint distributon of *w* and *Me* for 28 traits with $\hat{r} < 1$. **(c)** The R^2 of the fitting of the function (2) for the 35 traits. The vertical line depicts the mean of the R^2 for the 35 traits.

https://doi.org/10.1371/journal.pgen.1008995.g005

degradation [46], and further influence phenotypes [47, 48]. Nevertheless, when we combined two subsets of total SNP data (MAF $\geq$ 0.01 and MAF $\geq$ 0.05) which contained 308'137 SNPs and 102'253 SNPs, respectively, with pan-genomic ORFs in GOBLUP, no more phenotypic variance explained by SNPs was captured, and the predictive abilities remained the same as with OBLUP only using pan-genomic ORF data (S7 Fig), 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. Significantly, the yeast genome is highly compact compared with other eukaryotic species, with ~70% of the genome sequence comprising ORFs, and a protein-encoding gene can be found for every 2 kb of the genome [49]. By contrast, the human genome contains a potential protein-encoding gene for every 30 kb, and the functional sequence encoding genes covers only ~30% of the total sequence [50]. The predictive ability of copy numbers of pan-genomic ORF-based prediction (CBLUP) was 0.13 to 0.72, which was significantly higher than the predictive ability of GBLUP (Fig 4 and S8 Fig). 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 (Fig 4). The reason could be that some of CNOs were not simple repeats of causal variants, and these CNOs added noise in the prediction. For the second combined method GCBLUP using SNPs and CNOs, the predictive abilities remained the same as with CBLUP for all traits (S9 Fig and S2 Table), suggesting that CNO data covered all causal variant information which SNP data carried. We used two Bayesian methods: Bayes A and Bayes B for predictions with the three types of dataset [1]. Both methods are resampling methods, where in Bayes A marker effects are sampled from a *t*-distribution, while in Bayes B marker effects are assigned a zero effect with probability 1—π and are sampled from a *t*-distribution with probability π (we used $\pi = 0.05$). By this, Bayes B accounts for the generally assumed genetic trait architecture, that a limited set of genes, and thus regions in the genome, have an effect on the given complex traits. Difference in predictive abilities between GBLUP and SNP-based Bayes A and Bayes B is negligible in our results. Likewise, when using exclusively pan-genomic ORF data, OBLUP gave similar predictive abilities with ORF-based Bayes A and Bayes B for all traits (S10 Fig), yet Bayes B performed slightly better than CBLUP and Bayes A for 22 traits, when only using copy numbers of pan-genomic ORFs, which indicated that some of the copy numbers of ORF information had no genetic effect (S11 Fig).

Recent pan-genome studies for higher mammals, such as human [22], and pig [51], revealed that non-redundant DNA sequence are absent from current reference genomes, but these studies do not provide ORF information for the populations. 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%) [2], rice (54%) [52], *Brassica napus* (62%) [53], bread wheat (64.3%) [54], tomato (74.2%) [18], and the proportions decrease when more individuals are added in the pan-genome populations [16]. A *Brassica oleracea* study shows nearly 20% of genes are influenced by presence/absence variation, and some of these genes are annotated with functions related to important agronomic traits, such as flowering time and disease resistance [55]. It has also been demonstrated in the tomato study that such variation may contribute to phenotypic diversity and crop improvement [18]. In yeast *S. cerevisiae* with genome size ∼120Mb about 6'000 genes are reported [49]. Rice (*Oryza sativa L.ssp.indica*) also has a comparatively compact genome (∼460Mb), but still harbours ∼40'000 to ∼50'000 genes, meaning that ∼70% of the rice genome sequence are transcribed in genes [56]. From this angle, one may speculate that pan-genomic ORFs might play a significant role in prediction of rice agronomic traits. Whether pan-genomic ORF data can be used for human disease 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 affected by 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 [57]. In a human genetics' context, using pan-genomic ORFs as a complement in genomic 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 "missing heritability", but we also see that the captured genetic variance is "phantom" to some degree. The ORF-based and CNO-based genomic prediction are more accurate than SNP-based genomic prediction for all traits in the yeast isolates. We demonstrate that pan-genomic ORFs have a potential to supplement SNPs in estimation of heritability and genomic prediction. However, in our study there still is a major gap between heritability and prediction accuracy for all traits,

but we provide evidence that prediction accuracy will be further improved if larger sample sizes can be used in training sets.

Materials and methods

Whole-Genome SNPs

We used publicly available data from 1,011 *S. cerevisiae* isolates that represent the breadth of their ecological and geographical origins comprised in the 1002 Yeast Genome project. Among these distantly related isolates, 918 had been deep sequenced [33], and the other 93 isolates that had previously been sequenced [58–60]. 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. After filtering out isolates with aneuploidies, we chose 787 diploid *S. cerevisiae* isolates for which SNP, ORF, copy number of ORF and phenotypes were available for all analyses. We removed SNPs with missing rate >0.05, MAF <0.01, and 311'447 SNPs were remained. 3'310 SNPs which violated Hardy–Weinberg Equilibrium (based on a χ^2 test, p $<10^{-6}$) were also removed. The remaining missing genotypes were imputed using Beagle 4.1 [61]. In total, 308'137 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 S12 Fig.

Pan-genomic open reading frames

The *S. cerevisiae* pan-genome had been determined for the 1,011 genomes using *de novo* genome assemblies and detection of non-reference genome material [33], revealing 7,796 non-redundant ORFs. Among them, 4,940 were core ORFs, containing ORFs present in all isolates and 2,856 ORFs showed a presence/absence variability within the population, containing ORFs that were dispensable or isolate-specific genes. The copy number of each ORF (including copy numbers of core ORFs) was assessed by mapping the reads from each strain to the pan-genomic ORFs with BWA [62]. For details of the *de novo* genome assemblies, detection of non-reference genome material, annotation of ORFs, and the assessment of the ORF copy numbers see [33]. The frequency distribution of pan-genomic open reading frames in 787 diploid *S. cerevisiae* isolates is shown in S12 Fig.

Phenotypes

Quantitative high-throughput phenotyping had been performed using end-point colony growth on solid medium [33]. 971 strains were phenotyped in parallel under 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 S3 Table.

Statistical models

GBLUP, OBLUP, and CBLUP: As a baseline, we conduct the benchmark genomic best linear unbiased prediction (GBLUP) [63], using SNP data. Pan-genomic ORF presence/absence, and copy number of ORF (CNO) information are tested with newly defined approaches termed

OBLUP and CBLUP, respectively. The general statistical model is

$$y = 1\mu + g + e \quad (1)$$

where y is the vector of phenotypic observations, μ is the overall mean and 1 is a vector of ones, and $g \sim N(0, \Gamma)$ and $e \sim N(0, I\sigma_e^2)$ are vectors containing random additive genetic effects and residual effects, and GBLUP, OBLUP and CBLUP only differ in the covariance matrix Γ used for the genetic effects.

In GBLUP the covariance structure of additive effects was $G\sigma_g^2$ with $G = \frac{ZZ'}{2\sum p_i(1-p_i)}$, where p_i denotes the minor allele frequency (MAF) of marker i. Moreover, 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.

The ORF-based covariance matrix in OBLUP was calculated as $O\sigma_o^2$ with $O = \frac{WW'}{\sum q_i(1-q_i)}$, where q_i denotes the frequency of ORF i, and W denotes the ORF matrix with entries $(0 - q_i)$ and $(1 - q_i)$ that represented absence and presence of ORF, respectively.

We fitted for each phenotypic trait the function

$$r = w\sqrt{\frac{n\hat{h}_O^2}{n\hat{h}_O^2 + Me}} \quad (2)$$

in which r is the observed predictive ability of OBLUP for this trait, w is the maximum predictive accuracy with infinite training set size, n is the number of isolates in the training set, $\hat{h}_O^2$ is the ORF-based heritability estimate, and Me is the number of independent chromosome segments [41]. The two model parameters w and Me were empirically determined for each of the 35 traits with a maximum likelihood approach using the function "optim" in R [64].

The CNO-based covariance matrix in CBLUP was calculated as $C\sigma_c^2$ with $C = \frac{SS'}{f}$, where S denotes the copy numbers of ORFs matrix with entries $(b_{ij} - u_i)$ where $0 \leq b_{ij} < 297$ represents the copy number of the i th ORF in j th 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 SS'.

Further, we used two models combining SNP and ORF information (GOBLUP), and combining SNP and CNO information (GCBLUP)

$$y = 1\mu + g + h + e \quad (3)$$

where $g \sim N(0, G\sigma_g^2)$ is a vector containing random additive genetic effects modeled by SNPs, and $h \sim N(0, H)$ is a vector containing random additive genetic effects where the covariance matrix is derived from pan-genomic ORFs ($H = O\sigma_o^2$) in GOBLUP or from CNOs ($H = C\sigma_c^2$) in GCPLUP, respectively. All other variables are defined as described above.

SNP, ORF or CNO-based Bayes A and Bayes B: The model of SNP, ORF or CNO-based Bayes A [1] is

$$y = 1\mu + a_m + e \quad (4)$$

where a_m is a m x 1 vector of normally distributed and independent SNP, ORF or CNO effects. The variance of the i th variant effect, σ_{mi}^2 is modeled as a scaled inverted chi-square distribution χ^2 (v, S), where $S = 0.002$, and $v = 5$. y, μ, 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 SNP, ORF or CNO-based Bayes B [1] is the same as with Bayes A, but the prior

distribution of the variance of variant effect is a mixture of distributions which is given by

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

SNP, ORF or CNO-based Bayes A and Bayes B were implemented in an R package 'BGLR' [65].

Estimation of heritability

The SNP-based heritability was defined as the proportion of phenotypic variance explained by SNP marker effects and calculated as $\hat{h}^2_G = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \hat{\sigma}^2_e}$. All SNPs with MAF $\geq$ 0.01 were used for the estimation [12]. The ORF-based heritability was defined as the proportion of phenotypic variance explained by ORF effects. It was calculated as $\hat{h}^2_O = \frac{\hat{\sigma}^2_o}{\hat{\sigma}^2_o + \hat{\sigma}^2_e}$, using variable ORFs with frequency $\geq$ 0.05. 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 $\hat{h}^2_C = \frac{\hat{\sigma}^2_c}{\hat{\sigma}^2_c + \hat{\sigma}^2_e}$ using copy numbers of pan-genomic ORFs with frequency $\geq$ 0.05. Analogously, an ORF-SNP-based heritability $\hat{h}^2_{GO} = \frac{\hat{\sigma}^2_g + \hat{\sigma}^2_o}{\hat{\sigma}^2_g + \hat{\sigma}^2_o + \hat{\sigma}^2_e}$ and a CNO-SNP-based heritability $\hat{h}^2_{GC} = \frac{\hat{\sigma}^2_g + \hat{\sigma}^2_c}{\hat{\sigma}^2_g + \hat{\sigma}^2_c + \hat{\sigma}^2_e}$ was calculated. The variance components $\hat{\sigma}^2_g$, $\hat{\sigma}^2_o$, $\hat{\sigma}^2_c$, $\hat{\sigma}^2_e$ from the models above were estimated from the entire data sets, using the R package "regress" [66].

Comparison of predictive abilities

The predictive abilities of the defined models were measured with 20 replicates of a 5-fold random cross-validation [41]. 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 [67]. Isolate dissimilarities were estimated via "Euclidean distance" for each pair of isolates with the "dist.gene" function.

Supporting information

S1 Fig. Heritability estimates for all 35 traits estimated based on all SNPs, all ORFs, and all CNOs, respectively.
(TIF)

S2 Fig. Predictive abilities of $GBLUP_{all}$, $GBLUP_{0.01}$. $GBLUP_{all}$ using all SNPs, and $GBLUP_{0.01}$ using SNPs with MAF $\geq$ 0.01.
(TIF)

S3 Fig. Predictive abilities of $OBLUP_{all}$, $OBLUP_{0.01}$,and $OBLUP_{0.05}$. $OBLUP_{all}$ using all ORFs, $OBLUP_{0.01}$ using ORFs with frequency ≥ 0.01, and $OBLUP_{0.05}$ using ORFs with frequency ≥ 0.05, respectively.
(TIF)

S4 Fig. Predictive abilities of $CBLUP_{all}$, $CBLUP_{0.01}$ and $CBLUP_{0.05}$. $CBLUP_{all}$ using all CNOs, $CBLUP_{0.01}$ using CNOs with frequency ≥ 0.01, and $CBLUP_{0.05}$ usig CNOs with frequency ≥ 0.05, respectively.
(TIF)

S5 Fig. Predicted (y-axis) *vs.* observed (x-axis)phenotypes in SNP-based prediction for 35 traits. *r* represents the correlation coefficient between predicted and observed phenotypes across 35 traits.
(TIF)

S6 Fig. Predicted (y-axis) *vs.* observed (x-axis)phenotypes in ORF-based prediction for 35 traits. *r* represents the correlation coefficient between predicted and observed phenotypes across 35 traits.
(TIF)

S7 Fig. Box plots for predictive abilities of $GBLUP_{0.01}$, $GBLUP_{0.05}$, OBLUP, $GOBLUP_{0.01}$, $GOBLUP_{0.05}$. $GBLUP_{0.01}$ 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.
(TIF)

S8 Fig. Predicted (y-axis) *vs.* observed (x-axis)phenotypes in CNO-based prediction for 35 traits. *r* represents the correlation coefficient between predicted and observed phenotypes across 35 traits.
(TIF)

S9 Fig. Box plots for predictive abilities of $GBLUP_{0.01}$, CBLUP, $GCBLUP_{0.01}$. $GBLUP_{0.01}$ 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.
(TIF)

S10 Fig. Box plots for predictive abilities of OBLUP, $BayesA_{ORF}$ and $BayesB_{ORF}$ using pan-genomic open reading frames across 35 traits.
(TIF)

S11 Fig. Box plots for predictive abilities of CBLUP, $BayesA_{CNO}$ and $BayesB_{CNO}$ using copy numbers of pan-genomic open reading frames across 35 traits.
(TIF)

S12 Fig. 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 (blue).
(TIF)

S1 Table. 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.
(PDF)

S2 Table. Predictive abilities estimated from five models across 35 traits. GBLUP, OBLUP, CBLUP, GOBLUP and GCBLUP.
(PDF)

S3 Table. Statistical description of phenotype data.
(PDF)

Author Contributions

Conceptualization: Zhengcao Li.

Methodology: Zhengcao Li, Henner Simianer.

Supervision: Henner Simianer.

Validation: Zhengcao Li, Henner Simianer.

Writing – original draft: Zhengcao Li.

Writing – review & editing: Henner Simianer.

References

1. Meuwissen Theo HE and Hayes Ben J and Goddard Michael E. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819–1829.
2. Goddard ME and Hayes BJ. Genomic selection. Journal of Animal breeding and Genetics. 2007; 124 (6):323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x PMID: 18076469
3. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. Journal of animal Breeding and genetics. 2006; 123(4):218–223. https://doi.org/10.1111/j.1439-0388.2006.00595.x
4. Goddard Michael E and Hayes Ben J and Meuwissen Theo HE. Genomic selection in livestock populations. Genetics research. 2010; 92(5-6):413–421. https://doi.org/10.1017/S0016672310000613 PMID: 21429272
5. Crossa José and Pérez-Rodríguez Paulino and Cuevas Jaime and Montesinos-López Osval and Jarquín Diego and de los Campos Gustavo and Burgueño Juan and González-Camacho Juan M and Pérez-Elizalde Sergio and Beyene Yoseph and others. Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science. 2017; 22(11):961–975. https://doi.org/10.1016/j.tplants.2017.08.011 PMID: 28965742
6. Abraham Gad and Inouye Michael. Genomic risk prediction of complex human disease and its clinical application. Current opinion in genetics & development. 2015; 33:10–16. https://doi.org/10.1016/j.gde.2015.06.005
7. Wray Naomi R and Yang Jian and Hayes Ben J and Price Alkes L and Goddard Michael E and Visscher Peter M. Author reply to A commentary on Pitfalls of predicting complex traits from SNPs. PLoS genetics. 2013; 14(12):894.
8. de los Campos Gustavo and Vazquez Ana I and Fernando Rohan and Klimentidis Yann C and Sorensen Daniel. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS genetics. 2013; 9(7):e1003608. https://doi.org/10.1371/journal.pgen.1003608 PMID: 23874214
9. Evans Luke M and Tahmasbi Rasool and Vrieze Scott I and Abecasis Gonçalo R and Das Sayantan and Gazal Steven and Bjelland Douglas W and De Candia, Teresa R and Goddard Michael E and Neale Benjamin M and others. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nature genetics. 2018; 50(5):737–745. https://doi.org/10.1038/s41588-018-0108-x
10. Wray Naomi R and Yang Jian and Hayes Ben J and Price Alkes L and Goddard Michael E and Visscher Peter M. Pitfalls of predicting complex traits from SNPs. Nature Reviews Genetics. 2013; 14(7):507–515. https://doi.org/10.1038/nrg3457 PMID: 23774735
11. Yang Jian and Benyamin Beben and McEvoy Brian P and Gordon Scott and Henders Anjali K and Nyholt Dale R and Madden Pamela A and Heath Andrew C and Martin Nicholas G and Montgomery

Grant W and others. Common SNPs explain a large proportion of the heritability for human height. Nature genetics. 2010; 42(7):565–569. https://doi.org/10.1038/ng.608 PMID: 20562875

12. Yang Jian and Zeng Jian and Goddard Michael E and Wray Naomi R and Visscher Peter M. Concepts, estimation and interpretation of SNP-based heritability. Nature genetics. 2017; 49(9):1304. https://doi.org/10.1038/ng.3941 PMID: 28854176
13. Sieber P, Platzer M, Schuster S. The definition of open reading frame revisited. Trends in Genetics. 2018; 34(3):167–170.
14. Lapierre Pascal and Gogarten J Peter. Estimating the size of the bacterial pan-genome. Trends in genetics. 2009; 25(3):107–110. https://doi.org/10.1016/j.tig.2008.12.004 PMID: 19168257
15. Vernikos George and Medini Duccio and Riley David R and Tettelin Herve. Ten years of pan-genome analyses. Current opinion in microbiology. 2015; 23:148–154. https://doi.org/10.1016/j.mib.2014.11.016 PMID: 25483351
16. Tettelin Hervé and Masignani Vega and Cieslewicz Michael J and Donati Claudio and Medini Duccio and Ward Naomi L and Angiuoli Samuel V and Crabtree Jonathan and Jones Amanda L and Durkin A Scott and others. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences. 2005; 102(39):13950–13955. https://doi.org/10.1073/pnas.0506758102
17. Aherfi Sarah and Pagnier Isabelle and Fournous Ghislain and Raoult Didier and La Scola Bernard and Colson Philippe. Complete genome sequence of Cannes 8 virus, a new member of the proposed family "Marseilleviridae". Virus Genes. 2013; 47(3):550–555. https://doi.org/10.1007/s11262-013-0965-4 PMID: 23912978
18. Gao Lei and Gonda Itay and Sun Honghe and Ma Qiyue and Bao Kan and Tieman Denise M and Burzynski-Chang Elizabeth A and Fish Tara L and Stromberg Kaitlin A and Sacks Gavin L and others. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nature genetics. 2019; 51(6):1044–1051. https://doi.org/10.1038/s41588-019-0410-2 PMID: 31086351
19. Li Ying-hui and Zhou Guangyu and Ma Jianxin and Jiang Wenkai and Jin Long-guo and Zhang Zhouhao and Guo Yong and Zhang Jinbo and Sui Yi and Zheng Liangtao and others. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology. 2014; 32(10):1045. https://doi.org/10.1038/nbt.2979 PMID: 25218520
20. Zhao Qiang and Feng Qi and Lu Hengyun and Li Yan and Wang Ahong and Tian Qilin and Zhan Qilin and Lu Yiqi and Zhang Lei and Huang Tao and others. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nature genetics. 2018; 50(2):278–284. https://doi.org/10.1038/s41588-018-0041-z PMID: 29335547
21. Dunn Barbara and Richter Chandra and Kvitek Daniel J and Pugh Tom and Sherlock Gavin. 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. 2012; 22(5):908–924. https://doi.org/10.1101/gr.130310.111 PMID: 22369888
22. Sherman Rachel M and Forman Juliet and Antonescu Valentin and Puiu Daniela and Daya Michelle and Rafaels Nicholas and Boorgula Meher Preethi and Chavan Sameer and Vergara Candelaria and Ortega Victor E and others. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nature genetics. 2019; 51(1):30–39. https://doi.org/10.1038/s41588-018-0273-y PMID: 30455414
23. Donati Claudio and Hiller N Luisa and Tettelin Hervé and Muzzi Alessandro and Croucher Nicholas J and Angiuoli Samuel V and Oggioni Marco and Hotopp Julie C Dunning and Hu Fen Z and Riley David R and others. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome biology. 2010; 11(10):R107. https://doi.org/10.1186/gb-2010-11-10-r107 PMID: 21034474
24. D'Auria Giuseppe and Jiménez-Hernández Nuria and Peris-Bondia Francesc and Moya Andrés and Latorre Amparo. Legionella pneumophila pangenome reveals strain-specific virulence factors. BMC genomics. 2010; 11(1):181–194. https://doi.org/10.1186/1471-2164-11-181 PMID: 20236513
25. Hu Pan and Yang Ming and Zhang Anding and Wu Jiayan and Chen Bo and Hua Yafeng and Yu Jun and Chen Huanchun and Xiao Jingfa and Jin Meilin. Comparative genomics study of multi-drug-resistance mechanisms in the antibiotic-resistant Streptococcus suis R61 strain. PLoS One. 2011; 6(9):e24988. https://doi.org/10.1371/journal.pone.0024988 PMID: 21966396
26. Konstantinidis Konstantinos T and Ramette Alban and Tiedje James M. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B: Biological Sciences. 2006; 361(1475):1929–1940. https://doi.org/10.1098/rstb.2006.1920
27. Botstein David and Fink Gerald R. Yeast: an experimental organism for 21st Century biology. Genetics. 2011; 189(3):695–704. https://doi.org/10.1534/genetics.111.130765 PMID: 22084421

28. Fay Justin C. The molecular basis of phenotypic variation in yeast. Current opinion in genetics & development. 2013; 23(6):672–677. https://doi.org/10.1016/j.gde.2013.10.005

29. Bloom Joshua S and Ehrenreich Ian M and Loo Wesley T and Lite Thúy-Lan Võ and Kruglyak Leonid. Finding the sources of missing heritability in a yeast cross. Nature. 2013; 494(7436):234–237. https://doi.org/10.1038/nature11867 PMID: 23376951

30. Kumar Anuj and Snyder Michael. Emerging technologies in yeast genomics. Nature Reviews Genetics. 2001; 2(4):302–312. https://doi.org/10.1038/35066084 PMID: 11283702

31. Märtens Kaspar and Hallin Johan and Warringer Jonas and Liti Gianni and Parts Leopold. Predicting quantitative traits from genome and phenome with near perfect accuracy. Nature communications. 2016; 7:11512–11520. https://doi.org/10.1038/ncomms11512 PMID: 27160605

32. Marroni Fabio and Pinosio Sara and Morgante Michele. Structural variation and genome complexity: is dispensable really dispensable?. Current Opinion in Plant Biology. 2014; 18:31–36.

33. Peter Jackson and De Chiara Matteo and Friedrich Anne and Yue Jia-Xing and Pflieger David and Bergström Anders and Sigwalt Anastasie and Barre Benjamin and Freel Kelle and Llored Agnès and others. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature. 2018; 556(7701):339–344. https://doi.org/10.1038/s41586-018-0030-5 PMID: 29643504

34. Maher Brendan. Personal genomes: The case of the missing heritability. Nature News. 2008; 456 (7218):18–21. https://doi.org/10.1038/456018a

35. Hill William G and Goddard Michael E and Visscher Peter M. Data and theory point to mainly additive genetic variance for complex traits. PLoS genetics. 2008; 4(2):e1000008. https://doi.org/10.1371/journal.pgen.1000008 PMID: 18454194

36. Walker Francis O. Huntington's disease. The Lancet. 2007; 369(9557):218–228. https://doi.org/10.1016/S0140-6736(07)60111-1

37. Gonzalez Enrique and Kulkarni Hemant and Bolivar Hector and Mangano Andrea and Sanchez Racquel and Catano Gabriel and Nibbs Robert J and Freedman Barry I and Quinones Marlon P and Bamshad Michael J and others. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science. 2005; 307(5714):1434–1440. https://doi.org/10.1126/science.1101160 PMID: 15637236

38. Goddard Michael E and Wray Naomi R and Verbyla Klara and Visscher Peter M and others. Estimating effects and making predictions from genome-wide marker data. Statistical Science. 2009; 24(4):517–529. https://doi.org/10.1214/09-STS306

39. Kim Hwasoon and Grueneberg Alexander and Vazquez Ana I and Hsu Stephen and de los Campos Gustavo. Will big data close the missing heritability gap?. Genetics. 2017; 207(3):1135–1145. https://doi.org/10.1534/genetics.117.300271 PMID: 28893854

40. Speed Doug and Hemani Gibran and Johnson Michael R and Balding David J. Improved heritability estimation from genome-wide SNPs. The American Journal of Human Genetics. 2012; 91(6):1011–1021. https://doi.org/10.1016/j.ajhg.2012.10.010 PMID: 23217325

41. Erbe Malena and Gredler Birgit and Seefried Franz Reinhold and Bapst Beat and Simianer Henner. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS One. 2013; 8(12):e81046. https://doi.org/10.1371/journal.pone.0081046 PMID: 24339895

42. Bentley Stephen. Sequencing the species pan-genome. Nature Reviews Microbiology. 2009; 7:258–259.

43. Georges Michel and Charlier Carole and Hayes Ben. Harnessing genomic information for livestock improvement. Nature Reviews Genetics. 2019; 20(3):135–156. https://doi.org/10.1038/s41576-018-0082-2 PMID: 30514919

44. Marouli Eirini and Graff Mariaelisa and Medina-Gomez Carolina and Lo Ken Sin and Wood Andrew R and Kjaer Troels R and Fine Rebecca S and Lu Yingchang and Schurmann Claudia and Highland Heather M and others. Rare and low-frequency coding variants alter human adult height. Nature. 2017; 542(7640):186–190. https://doi.org/10.1038/nature21039 PMID: 28146470

45. Maurano Matthew T and Humbert Richard and Rynes Eric and Thurman Robert E and Haugen Eric and Wang Hao and Reynolds Alex P and Sandstrom Richard and Qu Hongzhu and Brody Jennifer and others. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337(6099):1190–1195. https://doi.org/10.1126/science.1222794 PMID: 22955828

46. Albert Frank W and Kruglyak Leonid. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nature Reviews Genetics. 2015; 16(4):197–212. https://doi.org/10.1038/nrg3891 PMID: 25707927

47. Yan Hai and Dobbie Zuzana and Gruber Stephen B and Markowitz Sanford and Romans Kathy and Giardiello Francis M and Kinzler Kenneth W and Vogelstein Bert. Small changes in expression affect

predisposition to tumorigenesis. Nature genetics. 2002; 30(1):25–26. https://doi.org/10.1038/ng799 PMID: 11743581

48. Kleinjan Dirk A and van Heyningen Veronica. Long-range control of gene expression: emerging mechanisms and disruption in disease. The American Journal of Human Genetics. 2005; 76(1):8–32. https://doi.org/10.1086/426833 PMID: 15549674

49. Goffeau André and Barrell Bart G and Bussey Howard and Davis RW and Dujon Bernard and Feldmann Heinz and Galibert Francis and Hoheisel JD and Jacq Cr and Johnston Michael and others. Life with 6000 genes. Science. 1996; 274(5287):546–567. https://doi.org/10.1126/science.274.5287.546 PMID: 8849441

50. Es Lander and Lm Linton and others. Initial sequencing and analysis of the human genome. Nature. 2001; 409(6822):860. https://doi.org/10.1038/35057062

51. Li Mingzhou and Chen Lei and Tian Shilin and Lin Yu and Tang Qianzi and Zhou Xuming and Li Diyan and Yeung Carol KL and Che Tiandong and Jin Long and others. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. Genome research. 2017; 27(5):865–874. https://doi.org/10.1101/gr.207456.116 PMID: 27646534

52. Wang Wensheng and Mauleon Ramil and Hu Zhiqiang and Chebotarov Dmytro and Tai Shuaishuai and Wu Zhichao and Li Min and Zheng Tianqing and Fuentes Roven Rommel and Zhang Fan and others. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018; 557 (7703):43–49. https://doi.org/10.1038/s41586-018-0063-9 PMID: 29695866

53. Hurgobin Bhavna and Golicz Agnieszka A and Bayer Philipp E and Chan Chon-Kit Kenneth and Tirnaz Soodeh and Dolatabadian Aria and Schiessl Sarah V and Samans Birgit and Montenegro Juan D and Parkin Isobel AP and others. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant biotechnology journal. 2018; 16(7):1265–1274. https://doi.org/10.1111/pbi.12867 PMID: 29205771

54. Montenegro Juan D and Golicz Agnieszka A and Bayer Philipp E and Hurgobin Bhavna and Lee Huey-Tyng and Chan Chon-Kit Kenneth and Visendi Paul and Lai Kaitao and Doležel Jaroslav and Batley Jacqueline and others. The pangenome of hexaploid bread wheat. The Plant Journal. 2017; 90 (5):1007–1013. https://doi.org/10.1111/tpj.13515 PMID: 28231383

55. Golicz Agnieszka A and Bayer Philipp E and Barker Guy C and Edger Patrick P and Kim HyeRan and Martinez Paula A and Chan Chon Kit Kenneth and Severn-Ellis Anita and McCombie W Richard and Parkin Isobel AP and others. The pangenome of an agronomically important crop plant Brassica oleracea. Nature communications. 2016; 7:13390.

56. Jun Yu and Songnian Hu and Jun Wang. A Draft Sequence of the Rice Genome (Oryza sativa L. Ssp. Indica). Science. 2002; 296(5565):79–91. https://doi.org/10.1126/science.1068037 PMID: 11935017

57. Wray Naomi R and Kemper Kathryn E and Hayes Benjamin J and Goddard Michael E and Visscher Peter M. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. Genetics. 2019; 211(4):1131–1141. https://doi.org/10.1534/genetics.119.301859 PMID: 30967442

58. Skelly Daniel A and Merrihew Gennifer E and Riffle Michael and Connelly Caitlin F and Kerr Emily O and Johansson Marnie and Jaschob Daniel and Graczyk Beth and Shulman Nicholas J and Wakefield Jon and others. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome research. 2013; 23(9):1496–1504. https://doi.org/10.1101/gr.155762.113 PMID: 23720455

59. Bergström Anders and Simpson Jared T and Salinas Francisco and Barré Benjamin and Parts Leopold and Zia Amin and Nguyen Ba Alex N and Moses Alan M and Louis Edward J and Mustonen Ville and others. A high-definition view of functional genetic variation from natural yeast genomes. Molecular biology and evolution. 2014; 31(4):872–888. https://doi.org/10.1093/molbev/msu037 PMID: 24425782

60. Strope Pooja K and Skelly Daniel A and Kozmin Stanislav G and Mahadevan Gayathri and Stone Eric A and Magwene Paul M and Dietrich Fred S and McCusker John H. The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome research. 2015; 25(5):762–774. https://doi.org/10.1101/gr.185538.114 PMID: 25840857

61. Browning Brian L and Browning Sharon R. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013; 194(2):459–471. https://doi.org/10.1534/genetics.113.150029 PMID: 23535385

62. Li Heng and Durbin Richard. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics. 2009; 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

63. VanRaden Paul M. Efficient methods to compute genomic predictions. Journal of dairy science. 2008; 91(11):4414–4423.

64. Team, R Core and others. R: A language and environment for statistical computing. Computing. 2013.

65. Pérez Paulino and de Los Campos Gustavo. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014; 198(2):483–495. https://doi.org/10.1534/genetics.114.164442 PMID: 25009151

66. Clifford David and McCullagh Peter. Package 'regress'. 2013.

67. Paradis Emmanuel and Schliep Klaus. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018; 35(3):526–528. https://doi.org/10.1093/bioinformatics/bty633