Genetics and Genomics
Computational and Systems Biology

Haplotype Function Score improves biological interpretation and cross-ancestry polygenic prediction of human complex traits

Weichen Song, **Yongyong Shi**, **Guan Ning Lin**

Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, School of Bioengineering, Shanghai Jiaotong University, Shanghai, China • Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China • Biomedical Sciences Institute of Qingdao University (Qingdao Branch of SJTU Bio-X Institutes), Qingdao University, Qingdao, China

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
March 21, 2024 (this version)

Reviewed preprint version 1
December 5, 2023

Sent for peer review
September 28, 2023

Posted to preprint server
August 9, 2023

Abstract

We propose a new framework for human genetic association studies: at each locus, a deep learning model (in this study, Sei) is used to calculate the functional genomic activity score for two haplotypes per individual. This score, defined as the Haplotype Function Score (HFS), replaces the original genotype in association studies. Applying the HFS framework to 14 complex traits in the UK Biobank, we identified 3,619 independent HFS-trait associations with a significance of $p<5\times10^{-8}$. Fine-mapping revealed 2,699 causal associations, corresponding to a median increase of 63 causal findings per trait compared with SNP-based analysis. HFS-based enrichment analysis uncovered 727 pathway-trait associations and 153 tissue-trait associations with strong biological interpretability, including “circadian pathway-chronotype” and “arachidonic acid-intelligence”. Lastly, we applied LASSO regression to integrate HFS prediction score with SNP-based polygenic risk scores, which showed an improvement of 16.1% to 39.8% in cross-ancestry polygenic prediction. We concluded that HFS is a promising strategy for understanding the genetic basis of human complex traits.

eLife assessment

This **valuable** paper presents a new approach for association testing, using the output of neural networks that have been trained to predict functional changes from DNA sequences. As such, the approach is an interesting addition to statistical genetics, and the evidence for the presented method being able to identify trait-associations in regions where GWASs are typically underpowered is **solid**. A limitation is, however, that it is unclear how the quality of these associations compares to those detected using conventional methods. Additional work assessing this method's power and characterizing false positives / false negative regions would be critical to ensure that the method is broadly adopted by the field.

Introduction

Genome-wide association studies (GWAS) have witnessed remarkable advancements over recent years, both in terms of sample size and genetic discovery. However, the elucidation of downstream mechanisms and subsequent applications still face certain limitations(Visscher et al., 2017). One caveat is that the statistical power of GWAS on a variant relies on its population frequency(Li et al., 2020; Null et al., 2022; Zhou et al., 2022), whereas most variants with large effect size are rare(Zeng et al., 2021), leading to insufficient discoveries. Moreover, Linkage Disequilibrium (LD) among neighboring variants can significantly inflate false positive results(Nowbandegani et al., 2022). The variability of LD structure among different populations further compounds the challenges associated with training predictive models and discovering causal genes. Lastly, most trait-relevant variants reside in non-coding regions(Watanabe et al., 2019), which lack direct functional annotations as coding variants. The prevalent approach to addressing this issue is to annotate each variant based on its location within functionally significant regions(Finucane et al., 2015; Grotzinger et al., 2022; Iotchkova et al., 2019; Weissbrod et al., 2020; Zheng et al., 2022), such as transcription factor binding sites or enhancers. While this strategy has considerably advanced the analysis, it's not optimal, as a variant's placement within a functionally important region does not inherently signify that the variant has substantial functional impacts.

The central dogma, proposing that DNA alterations' effects on phenotype are mediated via RNA and protein changes, offers a novel strategy to address these challenges. More precisely, by replacing the original genotypes in association studies with the aggregated impact of variants on transcription or functional genomics, the central dogma ensures the preservation of the majority of genetic information. This 'aggregated impact' offers several benefits for GWAS analysis: it provides direct biological interpretations, bypasses the effects of LD and population genetic history, and amalgamates information from both common and rare variants. One successful implementation of this strategy is Polygenic Transcriptome Risk Scores (PTRS)(Hu et al., 2022; Liang et al., 2022), which employ genetically determined transcription levels rather than genotypes to predict complex trait, and achieved remarkable portability. Nonetheless, the accuracy of imputing transcription levels from genotypes, given the sample size of currently available cohorts (such as the Genotype-Tissue Expression project, GTEx(Aguet et al., 2020)), remains limited (R2 around 0.1 for most genes)(Barbeira et al., 2018). Thus, the performance of PTRS is yet to reach its optimal potential.

Following the success of PTRS, we made one step forward to utilize functional genomics in this strategy. Compared with transcription levels, predicting genetically-determined functional genomic levels has achieved much higher accuracy by multiple recent deep learning (DL) studies(Avsec et al., 2021; Chen et al., 2022; Kelley, 2020; Yan et al., 2021; Zhou et al., 2018). These DL models utilize segments of the human reference genome as training samples, substantially increasing the sample size. Furthermore, functional genomics serve as a mediator between DNA and transcription, thus lessening the influence of non-genic factors such as the environment. Given these advancements, we propose that using the outputs of one of the state-of-the-art DL models, Sei(Chen et al., 2022), as the 'aggregated impact' in this novel strategy could effectively address the challenges aforementioned. Sei accepts a DNA sequence and computes multiple sequence class scores that represent different facets of the functional genomic activities of that sequence. This score integrates impacts from all variants, even those as rare as singletons, into one continuous variable, and is, in theory, unaffected by LD. In line with this notion, a recent similar strategy called cistrome-wide association study (CWAS) integrated variant-chromatin activity and variant-phenotype association to boost power of genetic study of cancer(Baca et al., 2022).

In this study, we present an analytical framework founded on this strategy (**Figure 1**) and implement it on complex traits in the UK Biobank to pinpoint causal loci and genes, decipher biological mechanisms, and devise cross-ancestry prediction models. We segmented the human reference genome into multiple 4,096 bp loci, generated DNA sequences for each locus for two haplotypes per individual, and employed Sei to compute the functional genomic activities of these sequences. We designated this activity score as the Haplotype Function Score (HFS) and analyzed the association between the HFS and each trait. Our findings confirm that the HFS framework offers a unique improvement in the biological interpretation and polygenic prediction of complex traits compared to classic SNP-based methods, thereby demonstrating its value in genetic association studies.

Result

Overview of genome-wide HFS

We used the HFS framework to analyze imputed genotype data from the UK Biobank (**Figure 1**). We segmented the human genome (hg38) into 617,378 discrete, non-overlapping loci, each 4,096 base pairs long. Of these, 590,959 loci carried at least one non-reference haplotype in the UKB cohort (see Method and Table S1). After quality control, these loci contained approximately 1.2 billion haplotypes, with a median count of 819 per loci (Figure S1). We then employed the deep learning framework, sei(Chen et al., 2022), to compute sequence class scores for each haplotype. In its sequence-mode, sei accepts DNA sequences in fasta format and produces multiple distinct sequence class scores, 39 of which were included in our study (Method). Our analysis identified significant variation in sequence class scores across different loci. In fact, 49.7% of loci housed haplotypes whose sequence class (as defined by the maximum of the 39 sequence class scores) differed from the reference haplotype sequence class. Using the reference sequence class as a benchmark, we noted that 16.8% of loci showed a difference between the maximum and minimum haplotype scores that surpassed the score of the reference haplotype. Moreover, the correlation between sequence class scores of adjacent loci was low, with a median R^2 value of 0.013 (Figure S2), effectively reducing the impact of LD in association studies. Further evaluation indicated that this low LD was led by two factors: integration of rare variant impacts and segmentation. Firstly, excluding rare variants from HFS caused the LD raised to median=0.14 (Method; Figure S2C). Secondly, median LD of SNPs from adjacent loci was 0.06, which was significantly higher than HFS LD (paired Wilcoxon $p=1.76\times10^{-5}$) but significantly lower than HFS LD without rare variants (paired Wilcoxon $p<2.2\times10^{-16}$).

Expanding on the sequence class scores, we defined HFS for each locus. Specifically, we computed the mean sequence class score of two haplotypes per individual, reflecting an additive model. We selected the score corresponded to the sequence class of reference sequence as the HFS of the corresponding locus, and its association with each trait was computed using a generalized linear model. Simulation analysis revealed that when a non-reference sequence class score was associated the trait, reference class score could still capture median 70% of HFS-trait association R^2. We applied this framework to 14 polygenic traits in the UKB British ancestry training set (n=350,587; Table S2 and Method), identifying 16,597 significant HFS-trait associations at a threshold of $p<5\times10^{-8}$ (n=15 for insomnia, n=7,573 for height; Table S2), equating to roughly 3,619 independent associations. The most significant associations were between the “promotor” score of chr7:121327898-121331994 (WNT16) and bone mineral density (BMD; regression beta=-0.02, $p<10^{-300}$), and the “promotor” score of chr9:4760952-4765048 (AK3) and platelet count (beta=3.20, $p=2.79\times10^{-262}$; Table S3).

When comparing HFS association with the standard SNP-based GWAS on the same data, we found that 98% of significant HFS loci also harbored a significant SNP. There were a few cases (n=0∼5) where significant HFS loci did not harbored even marginal SNP association (GWAS p>0.01), which

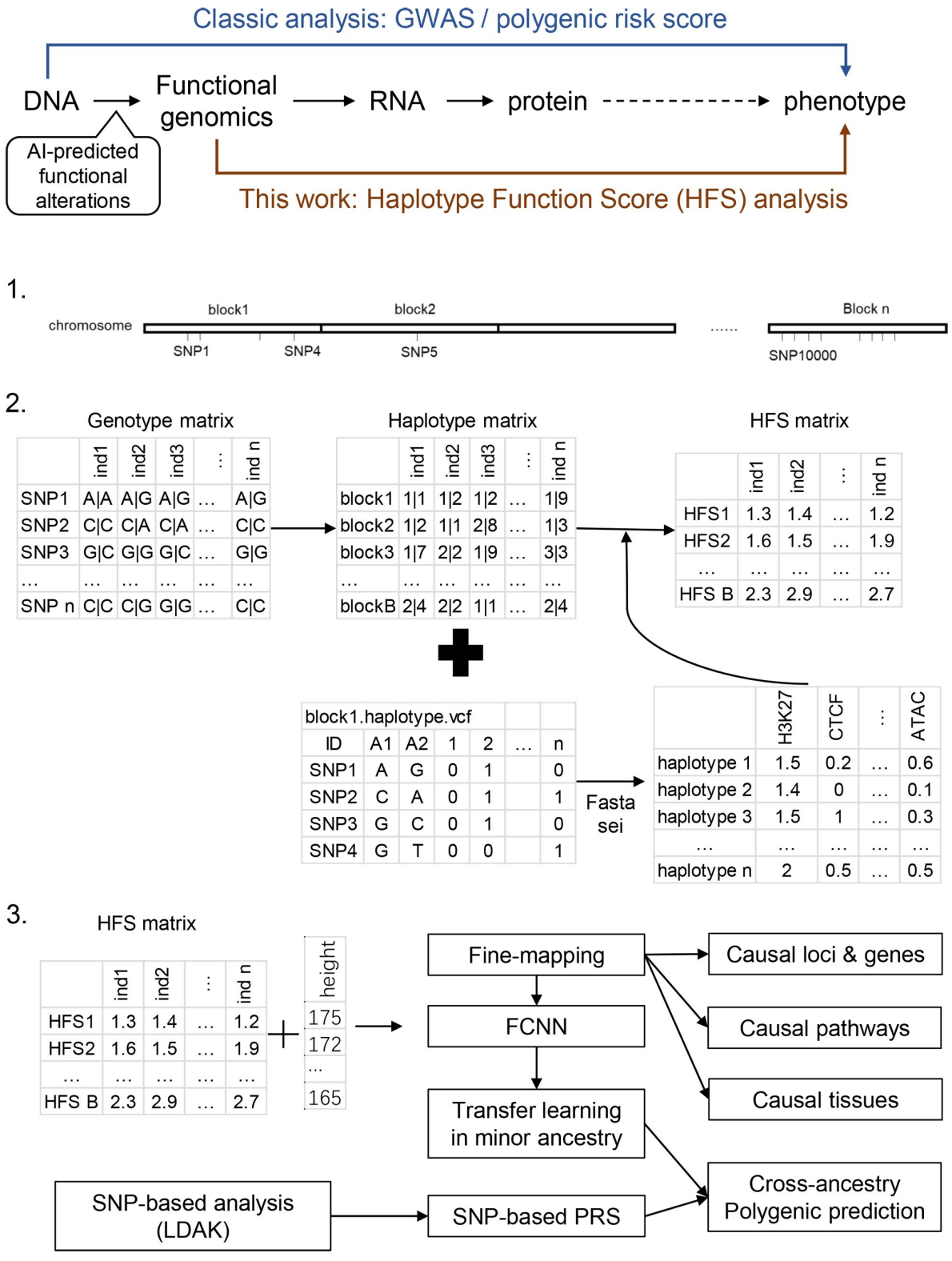


Figure 1

Flowchart of the study.

Ind: individual.

were due to the lack of common SNP in these loci. HFS association p value was higher than GWAS p value in 95 % of significant loci, suggested that HFS did not improve power to detect marginal effect. The genomic control inflation factor (λ_{GC}) for the HFS association test varied between 0.99 for asthma and 1.50 for height, closely resembling the SNP GWAS (Pearson Correlation Coefficient [PCC]=0.91, paired t-test p=0.16; Method and Figure S3). We concluded that HFS-based association tests had adequate power and do not introduce additional p-value inflation.

Fine-mapping based on HFS

Based on these data, we applied SUSIE to fine-map the causal loci that were associated with each of the 14 traits. We divided hg38 genome into 1,361 independent blocks as defined by MacDonald et al.(MacDonald et al., 2022), and applied SUSIE to loci HFS in each of these blocks (number of loci per block=4∼2392). As shown in **Figure 2** and Table S4, we identified a total of 2,699 causal loci-trait associations at the threshold of posterior inclusion probability (PIP)>0.95, hereafter referred to as "causal loci". Compared with SNP-based functionally-aware fine-mapping methods (PolyFun(Weissbrod et al., 2020) and SbayesRC(Zheng et al., 2022)), HFS-based SUSIE detected −11 to 334 more causal signals (median=63, Table S5) for each trait. We cautioned that these methods use summary statistics as input and are by nature less sensitive than individual data-based methods. Yet, we suggested that such impact would be mild, since we used in-sample LD reference (from UKB European sample).

Among these causal loci, only 22% were also lead loci in association analysis (loci with the lowest p-value in 200kb region), and 58% had association p-value > 5×10^{-8}. In line with previous SNP-based analysis(Weissbrod et al., 2020), this result highlighted the importance of using causal signals instead of lead signals in post-GWAS analysis. We found 67 causal loci showing pleiotropic effects on at least two independent traits, including "CTCF-Cohesin" score of chr9:89596537-89600633 that was associated with age at menarche, body mass index (BMI) and height (PIP>0.97; Table S4). We also found that rare variants played an important role in the good find-mapping performance of HFS: when variants with MAF<0.01 were removed, 55.3% of the causal signals would be missed in HFS+SUSIE analysis.

When looking at the reference sequence class of loci, those with functional importance were more likely to be causal loci, including "Promoter" (Odds Ratio [OR]=2.33, $p=1.41\times10^{-14}$) and "Bivalent stem cell enhancer" (OR=2.22, $p=1.11\times10^{-8}$) and "Transcribed region 1" (OR=1.71, $p=1.581\times 10^{-10}$, **Figure 2**). Such functional enrichment was even higher for pleiotropic loci ("Promoter": OR=7.20, $p=3.35\times10^{-5}$). We also observed trait-specific patterns of such sequence class enrichment, such as "CEBPB binding site" (Insomnia: OR=5.25, p=0.01) and "FOXA1/AR/ESR1 binding site" (intelligence: OR=4.69, p=0.01, **Figure 2** and Table S6). These results demonstrated the expected functional patterns of causal loci, and indicated that HFS-based fine-mapping was biologically interpretable and reliable.

Despite the functional enrichment, we applied several secondary analyses to verify the reliability of HFS-based SUSIE result. Firstly, we took causal SNP fine-mapped by PolyFun(Weissbrod et al., 2020) as positive control, and find that compared with genomic region-matched control loci, causal loci were significantly enriched for causal SNP (OR=1.33 to 5.08, Fisher test p=0.12 to 4.72×10^{-52}, Table S5). Secondly, we calculated the heritability tagged by causal loci and PolyFun causal SNP in independent test set (defined as the R^2 of linear regression; Method), and found that causal loci tagged 38% to 251% more heritability than causal SNP (median=151%; Table S5). This was not an artifact of larger number of causal loci, since the Akaike information criterion (AIC) was similar between causal loci and causal SNP (paired t-test p=0.36; Table S5). Thirdly, for traits with sufficient causal loci coverage, we also applied Linkage Disequilibrium Score regression (LDSC) on independent GWAS summary statistic to evaluate heritability enrichment in causal loci. On average, causal loci showed 124-fold enrichment of heritability, significantly larger than genomic region-matched control loci (124-fold vs 101-fold; p=0.0002, Method and Figure S4). Lastly,

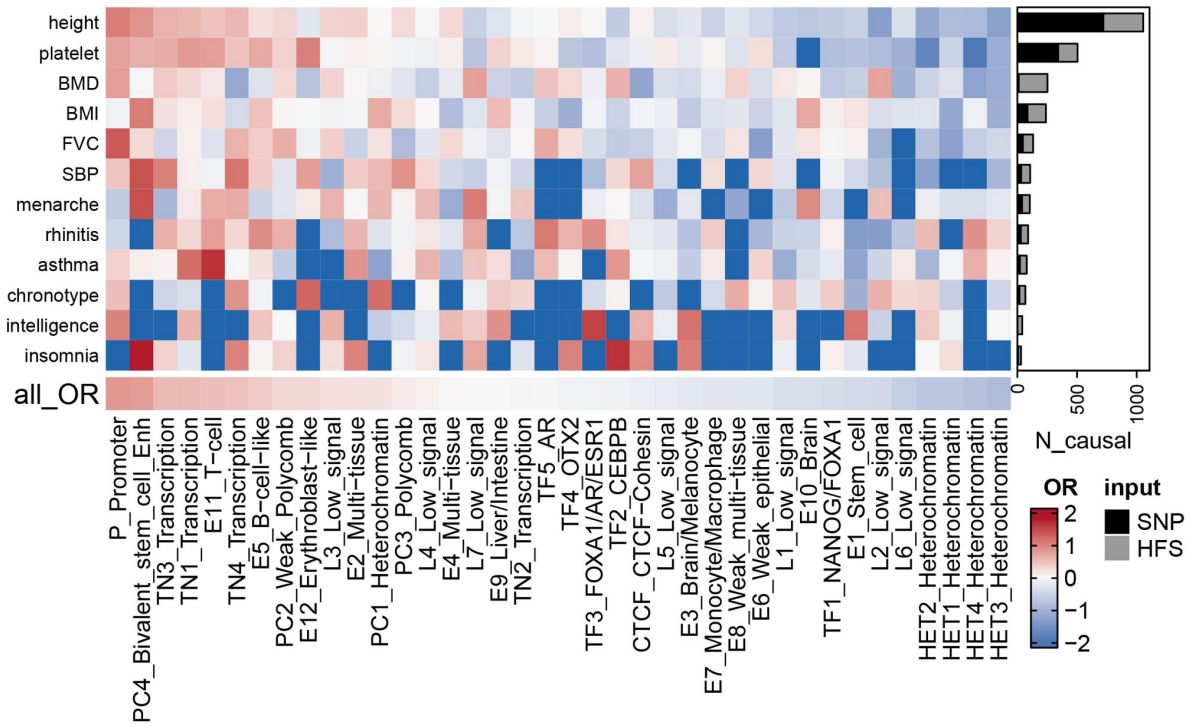


Figure 2

Fine-mapping result summary.

Grey bar plots indicated the number of loci with posterior inclusion probability (PIP)>0.95 in HFS+SUSIE (causal loci). Black bar plots indicated number of SNP with PIP>0.95 in PolyFun or SbayesRC analysis (the larger number was shown). Each grid of heatmap showed the Odds ratio of each sequence class loci being causal loci for each trait. “All_OR” indicated Odds ratio for pooling all traits together. Enh: enhancer. TF: transcription factor binding site.

we applied simulation analysis and found that HFS+SUSIE showed similar advantages over SNP-based methods as in real data, with high accuracy and low false positive rate (Supplementary materials).

We further applied a sliding-window analysis (step=2048 bp, Method) to test whether HFS-based result is robust against the choice of sequence interval. 29.4% of causal loci (PIP>0.95) in the original analysis were still causal in sliding window analysis. 31.1% and 29.3% of causal loci whose 5' and 3' overlapping locus had PIP>0.95 in sliding window analysis, respectively, while themselves were no longer causal. Besides, HFS+SUSIE was also robust when the predefined number of causal loci (L=2 to 10) was changed, and the number of detected loci were not changed. Lastly, removing insertion and deletion would reveal 9% more significant association ($p<5\times10^{-8}$) but 4.7% less causal association (PIP>0.95), and slightly increased inflation factor (Wilcoxon p=0.0001, Figure S4). Taken together, HFS-based SUSIE is a powerful and robust strategy for individual data-based genetic fine-mapping.

Biological interpretation based on HFS

Pinpointing causal loci of complex traits provides the opportunity of analyzing the biological mechanism of them. Thus, based on the HFS-based fine-mapping result, we applied a linear regression model to analyze the underlying pathways, cell types and tissues of each complex trait. For each locus, we annotated its relevance to a pathway by combined SNP to Gene (CS2G) strategy(Gazal et al., 2022), and regressed the PIP against this annotation, with a set of baseline annotations included as covariates, similar to the LDSC framework(Finucane et al., 2018) (Method). After p-value correction and recurrent pathway removal (Method), we detected a total of 727 pathway-trait associations (**Figure 3A** and Table S7). The most significant associations were "megakaryocyte differentiation" with platelet count ($p=2.26\times10^{-34}$), "Insulin-like growth factor receptor signaling pathway", "Endochondral ossification" with height ($p=4.95\times10^{-33}$ and 1.17×10^{-27}), "PD-1 signaling" with allergic disease ($p=5.55\times10^{-25}$), and "major histocompatibility complex pathway" with asthma ($p=1.22\times10^{-23}$). In fact, asthma and allergic disease were predominantly associated with more than 80 immune-related pathways. These associations were all in line with existing knowledge of trait mechanism, and extended the understanding of their genetic basis. For example, PD-1 has recently been suggested as potential targets of allergic diseases like atopic dermatitis(Morales et al., 2021), but such association has not been highlighted by previous genetic association studies.

For other traits, the most significant associations also replicated known mechanisms, such as "osteoblast differentiation", "Wnt ligand biogenesis and trafficking" with BMD ($p=4.59\times10^{-13}$ and 2.78×10^{-12}); "circadian pathway" with chronotype ($p=4.25\times10^{-12}$); "calcium regulated exocytosis of neurotransmitter", "Arachidonic acid metabolism" with intelligence ($p=5.52\times10^{-7}$ and 2.78×10^{-6}); "GPCR pathway" and "adipogenesis" with BMI ($p=4.97\times10^{-10}$ and 2.02×10^{-7}) and "physiological cardiac muscle hypertrophy" with systolic blood pressure ($p=6.32\times10^{-11}$). We also highlighted less significant association which provided novel insights, such as "synaptic vesicle docking" and "neuron migration" with chronotype ($p=4.00\times10^{-7}$ and 4.55×10^{-7}), "Prostaglandins synthesis" with insomnia ($p=5.30\times10^{-9}$), "behavioral response to cocaine" with alcohol intake ($p=3.39\times10^{-8}$) and "roof of mouth development" and "glycoside metabolism" with FVC ($p=2.19\times10^{-12}$ and 5.73×10^{-11}).

For cell type and tissue analysis (**Figure 3B** and Table S8), we applied the same linear model to evaluate whether causal loci enriched in active chromatin regions of each cell type (Method). We found 153 biologically interpretable associations with complex traits. For example, fetal megakaryocyte ($p=5.67\times10^{-22}$) and child spleen ($p=2.15\times10^{-13}$) were found to be key cell type and tissue of platelet count. Systolic blood pressure was significantly associated with multiple heart and artery tissues and fetal cardiomyocyte ($p<1.63\times10^{-5}$), whereas allergic disease was associated with multiple immune cells including natural killer, Treg and B cells ($p<4.79\times10^{-16}$). For brain-related traits, we found 21 significant associations, 14 of which were from central nervous system.

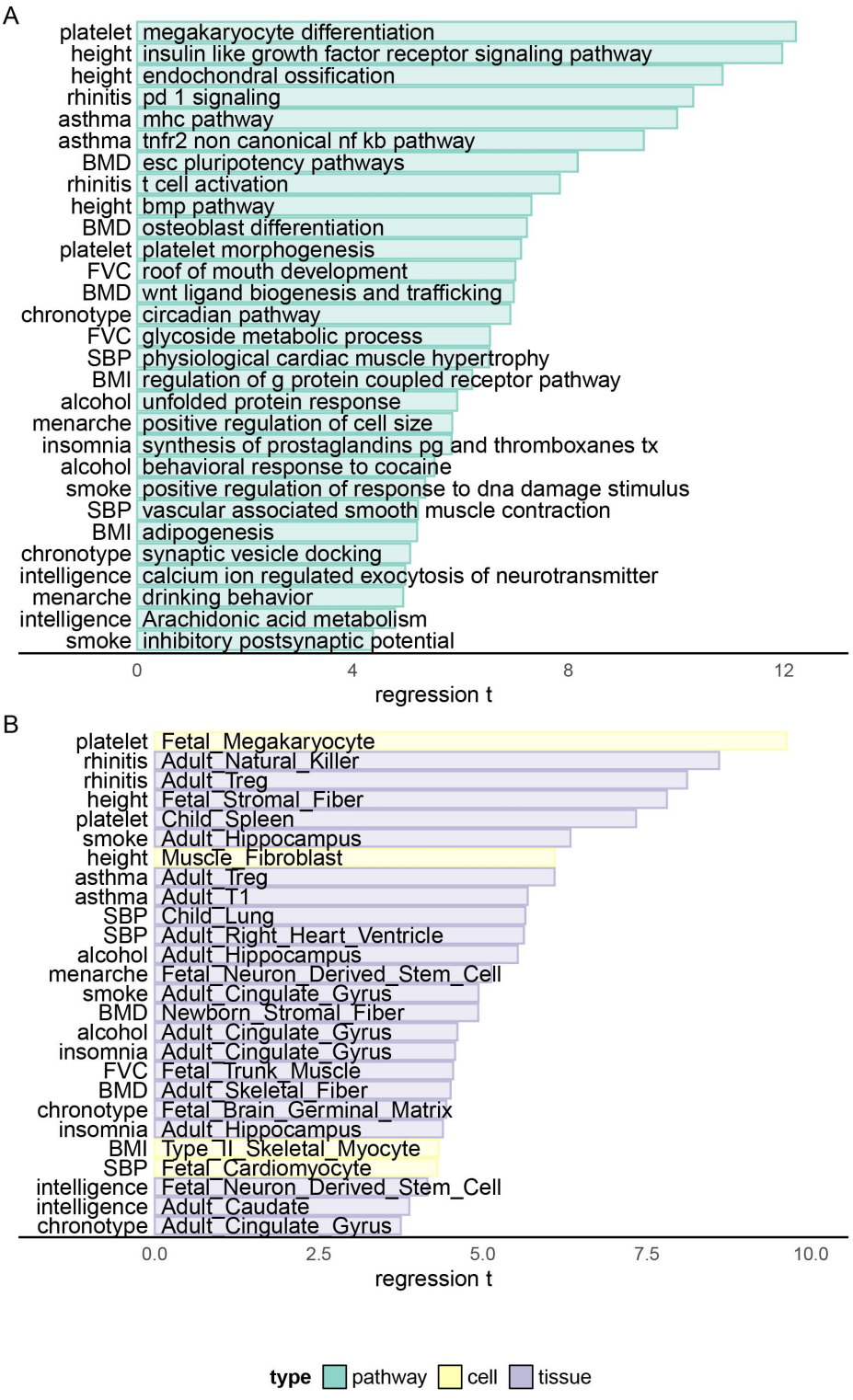


Figure 3

Biological enrichment analysis based on HFS fine-mapping.

X-axis indicated t statistics of the analyzed term in a multivariate linear regression (Method). Cell: single-cell ATAC peak for 222 cell types from Zhang et al(K. Zhang et al., 2021). Tissue: active chromatin regions of 222 tissues from epimap(Boix et al., 2021). For each trait, we showed the most significant term plus one or two terms with high biological interpretation that also passed significance threshold. Full enrichment result was shown in Table S7 and S8.

For example, adult Hippocampus and cingulate gyrus were both linked to alcohol intake, smoking and insomnia ($p<1.11 \times 10^{-5}$), whereas chronotype was associated with embryonic brain germinal matrix ($p<8.68\times10^{-6}$) and intelligence with embryonic neuron derived stem cell ($p<6.89\times10^{-7}$).

We also applied other modified strategies for this task but did not get satisfying result. For example, using cS2G to link locus to gene lists specifically expressed in each cell type suffered from scRNA dataset batch effect, whereas linear mix model was less sensitive than standard linear model (Supplementary Materials).

Taken together, our result suggests that fine-mapping results based on HFS could pinpoint the causal pathways, cell types and tissues underlying complex traits, and is valuable for the biological interpretation of genetic association study.

Highlighted genes for complex traits

Enhanced power of fine-mapping and biological enrichment could reveal novel key genes for trait mechanism study. Below we integrated fine-mapping result and their functional annotation in several case studies to find causal signals and trait-relevant genes in regions not resolved by previous genetic association studies.

In our study, platelet count had large number of causal loci (**Figure 2**) which showed significant functional enrichment (**Figure 3**). To find key loci and genes underlying platelet count, we focused on causal loci that overlapped with active regions in "fetal megakaryocyte" and "child spleen tissue", and applied cS2G(Gazal et al., 2022) to link them to two key pathways ("megakaryocyte differentiation" and "platelet morphogenesis", Method and **Figure 4A**). We chose these annotations based on p-value in biological enrichment analysis in **Figure 3**. A total of 25 loci were highlighted (**Figure 4A**), which were recurrently linked to well-known platelet-regulating genes like MEF2C, SH2B3, FLI1, RUNX1, THPO and NFE2. Among them we noticed a less-studied gene RBBP5, a target of key transcriptome factor MEF2C during megakaryopoiesis(Kong et al., 2019). Specifically, in 1q32.1 region, HFS+SUSIE identified two loci with PIP>0.9 (**Figure 4B**). SNP-based association also found significant association in this region, but SNP fine-mapping(Weissbrod et al., 2020) could not resolve this signal and only found seven signals between PIP=0.1 to 0.5. This was unlikely a statistical inflation, since HFS-based association test p-value was actually higher than SNP-based one (Figure S5). One of the causal loci, chr1:47401806-47405902 (PIP=1), overlapped with spleen active chromatin and harbored a cCRE in megakaryocyte, and was linked to RBBP5 and three other genes (Table S9). RBBP5 is known to be involved in megakaryocyte differentiation during megakaryopoiesis and was regulated by MEF2C(Kong et al., 2019), but previous genetic association studies provided little evidence for its association with platelet count.

The major histocompatibility complex (MHC) region has long been a challenge of genetic association study due to its long-range LD, and is often excluded in fine-mapping tools. However, many disorders like Schizophrenia(Sekar et al., 2016) and immune diseases(Nawijn et al., 2011) are robustly associated with MHC region. In our HFS-based fine-mapping of asthma, we found 15 loci within MHC region had PIP>0.95, 11 of which overlapped with active chromatin regions in Treg or natural killer cells (**Figure 4C** and Table S10). This result showed good discrimination between causal and non-causal loci: despite these 15 likely causal loci, only six loci had PIP between 0.25 to 0.95. Since MHC region harbored a large number of genes, these causal loci were linked to as much as 105 potential target genes, which hindered the discovery of true targets. We further filtered them based on the involvement in pathway "TNFR2-NFKB pathway" and "innate lymphocyte [ILC] development", since these pathways were most significantly associated with asthma (**Figure 3**), even after excluding MHC region ($p=2.57\times10^{-13}$ and 1.39×10^{-17}). We found five genes (LTA, LTB, TNF, PSMB8 and PSMB9) that were predicted to be regulated by five causal loci overlapped with active chromatin regions (**Figure 4C**), which could be considered as potential key genes for further validation.

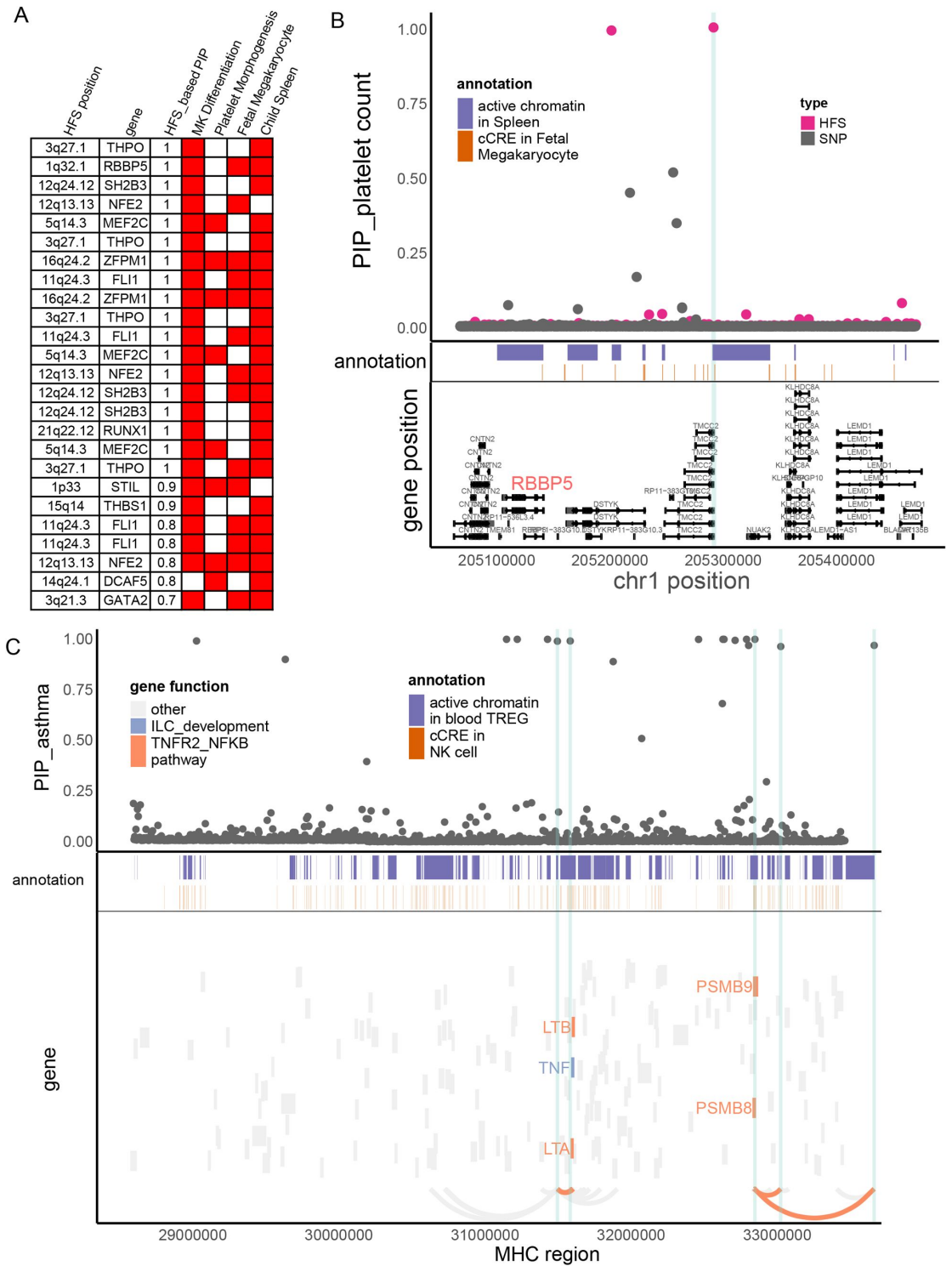


Figure 4

HFS linked trait to causal genes.

A: Target genes of causal loci identified by HFS+SUSIE for platelet count. Only genes that showed functional convergence were shown. B: Regional plot for RBBP5. HFS: loci PIP calculated by HFS+SUSIE. SNP: SNP PIP calculated by PolyFun. cCRE: credible cis-regulation elements. C: Regional plot of MHC region for asthma. Thickened curve linked highlighted causal loci to its target genes predicted by cS2G(Gazal et al., 2022).

Similarly, we fine-mapped MHC region for other allergic diseases (Figure S6 and Table S11) and found potential key genes including HLA family and AGER. We also highlighted other gene-trait association not previously emphasized by GWAS, including GATA4 and NPPA (cardiac muscle hypertrophy) with SBP, ALOX5 (Arachidonic acid metabolism) with intelligence and CRY1 (circadian pathway) with chronotype, as further discussed in Table S12 to S14 and supplementary information.

On the other hand, HFS perform worse than SNP-based fine-mapping on exonic regions. Taking height as an example, PolyFun detected 125 causal SNPs (PIP>0.95) in the exonic regions, but only 16% (20) of loci that harbored them also reached PIP>0. 5 (11 reached PIP>0.95) in HFS+SUSIE analysis. Among the 105 loci that missed such signals (HFS PIP<0.5), 12 had a nearby loci (within 10kb) showing HFS PIP>0.95, which likely reflected false positive led by LD. Thus, SNP-based analysis should be prioritized over HFS in coding regions.

HFS-based polygenic prediction

Lastly, we analyzed the potentiality of HFS in polygenic prediction accuracy. Compared with state-of-the-art SNP-based polygenic risk score (PRS) algorithm LDAK-BOLT(Q. Zhang et al., 2021), HFS-based PRS (weighted by SUSIE posterior effect size) reached 47% to 90% of R^2 in independent European test set (meta-analyzed proportion=75.6%, 95% confidence interval=75.3%~75.8%, Figure S7). The gap between performance of HFS-based and SNP-based PRS reflected the fact that HFS only captured (the majority of) functional genomic alterations and missed the information of amino acid sequence and post-translational modification. We thus proposed that integrating information from HFS and SNP could provide better performance. Specifically, in the large European training set we trained SNP PRS model by LDAK. Then, in a small tuning sample of target ancestry, we calculated per-locus HFS prediction score of height (sum of HFS within this block, weighted by SUSIE posterior effect size), then used machine learning to integrate them with LDAK PRS into a final polygenic prediction score, hereafter referred to as “HFS+LDAK”. To choose the proper machine learning tools to achieve this goal, in British European test set we applied LASSO, ridge regression and elastic net and compared the result (**Figure 5B**). They gave comparable result with only difference of R^2 around 0.01, and all of them were profoundly better than simple linear regression. We chose LASSO as the algorithm in the formal analysis.

Using height as a representative trait, we first estimated the proportion of variance captured by top loci, and found that HFS of loci with PIP>0.4 (n=5,101) captured roughly 80% of variance explained by all genome-wide loci (n=1,200,024 corresponded to sling-window strategy; **Figure 5A**). We then calculated HFS+LDAK in non-British European (NBE), South Asian (SAS), East Asian (EAS) and African (AFR) population in UK Biobank, and observed 17.5%, 16.1%, 17.2% and 39.8% improvement over LDAK alone ($p=3.21\times10^{-16}$, 0.0001, 0.002 and 0.001, respectively. **Figure 5C**). As a comparison, we integrated LDAK with PolyFun-pred(Weissbrod et al., 2022) and SbayesRC(Zheng et al., 2022) using Polypred framework(Weissbrod et al., 2022), but did not observe significant improvement over LDAK alone (difference in R^2<0.01, p=0.001~0.07, **Figure 5C**). Since PolyFun-pred+BOLT-LMM has been shown to significantly outperformed BOLT-LMM alone(Weissbrod et al., 2022), we reasoned that the improvement of LDAK over BOLT-LMM might have attenuated the improvement brought about by PolyFun-pred, making it difficult to reach significance threshold. Taken together, we concluded that HFS could bring about mild but significant improvement to classic SNP-based PRS in the task of cross-ancestry polygenic prediction.

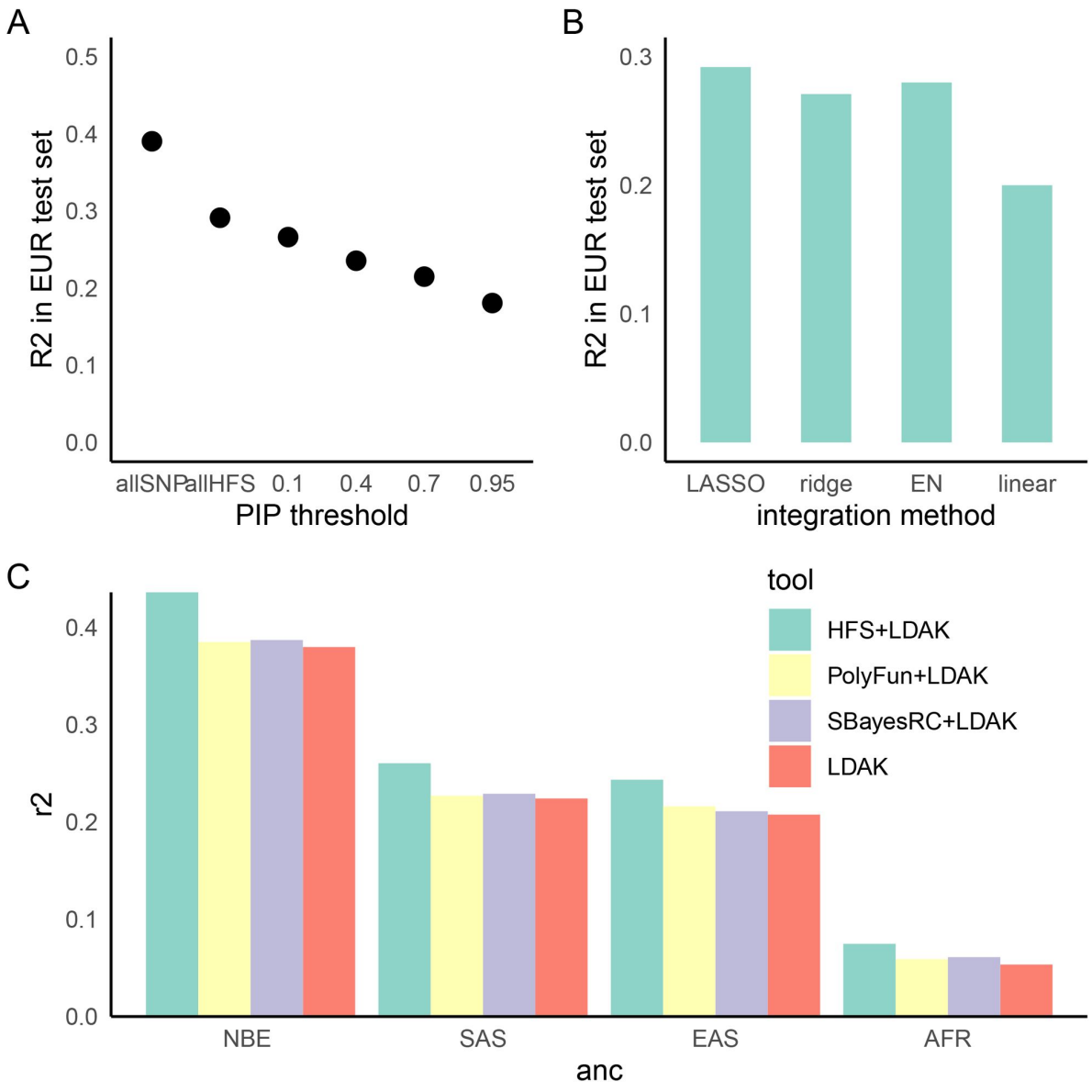


Figure 5

HFS-based polygenic prediction.

A: Prediction R^2 of HFS-based polygenic risk score (PRS) using different threshold of PIP. allSNP: SNP-based PRS calculated by LDAK-BOLT(Q. Zhang et al., 2021). n: number of features included in the corresponding PRS. B: Prediction R^2 of per-block HFS score in British European test set by different methods. EN: elastic net. C: Prediction R^2 of different tools in non-British European (NBE), South Asian (SAS), East Asian (EAS) and African (AFR) groups in UK Biobank.

Discussion

In this study, we designed the new HFS framework for genetic association analysis and demonstrated that it could improve classic SNP-based analysis in terms of causal loci and gene identification, biological interpretation and polygenic prediction. We suggest that HFS is a promising strategy for future genetic studies, but more progresses in algorithm and computation and data resources is still desired.

Compared with SNP, HFS has several compelling features. For instance, LD between adjacent HFS is much lower than SNP, which enhances the precision of statistical fine-mapping. For those false-positive variants caused by LD, they are expected to make little impacts on functional genomics, thus their HFS would be close to reference and would not influence downstream analysis significantly. In line with these advantages, we showed that HFS-based fine-mapping had high statistical power, and downstream enrichment analysis was capable of revealing biologically interpretable mechanisms. As a typical example, our findings of enrichment of intelligence-associated loci in arachidonic acid metabolism pathway is in line with the well-known role of polyunsaturated fatty acid in neurodevelopment(Helland et al., 2003). Nonetheless, previous GWAS provided little evidence on this association. Secondly, HFS could integrate effects of all variants within a locus, regardless of their population frequency. Thus, HFS could capture information from rare variants overlooked by classic association study and improve polygenic prediction, as shown by our result. In fact, HFS framework could directly extend to whole-genomes sequencing data and capture all mutations as rare as singleton, making one step forward to fill in the “missing heritability”.

Despite its potential, the current HFS framework carries several drawbacks and necessitates significant enhancements. A key limitation is the substantial computational cost. In this study, the transformation phase of the genotype-haplotype-sequence for UK Biobank SNP data required hundreds of thousands of CPU core hours. This computation cost would increase exponentially when analyzing whole-genome sequencing (WGS) data or employing a sliding-window strategy. A potential solution could involve developing a new algorithm that bypasses the variant calling stage and directly generates DNA sequences per locus from raw sequencing or SNP array data. For the sequence-to-HFS step, Sei(Chen et al., 2022) required about 1.8 GPU hours per one million sequences. Intriguingly, the majority of Sei’s output is unused in the HFS framework, since Sei predicts over twenty thousand functional genomic features, while the HFS only represents one of their integrated scores. Future development of novel deep learning models that predict functional genomics in a manner more fitting to the HFS framework could considerably reduce computation costs. Lastly, it’s currently unfeasible to incorporate all genome-wide HFS into a single LASSO model. This limitation forced us to first integrate HFS into pre-locus score, which inevitably sacrificed the accuracy.

Another hurdle arises in integrating HFS with other genomic features. Intrinsically, HFS captures only the variant effect mediated by functional genomics, while a genetic variant might also influence amino acids, post-transcriptional modifications (PTMs)(Park et al., 2021), and 3D chromosomal structures(Zhou, 2022). Therefore, HFS alone cannot wholly replace SNP without any loss, as our results demonstrate that the HFS-based prediction model captured approximately 70% of the variance explainable by the SNP-based prediction model. One potential solution is to extend the concept of HFS, applying deep learning to quantify the genetically determined values of PTMs, protein biochemical properties(Pejaver et al., 2020), and protein and chromosomal structures, potentially employing AlphaFold(Jumper et al., 2021)-derived features(Liu et al., 2022). Analyzing HFS in conjunction with these multi-modal function scores could provide a comprehensive depiction of the genetic architecture of complex traits. However, the colossal computational cost is currently prohibitive. As a compromise, we simply performed joint analysis

of HFS with SNP PRS in our prediction model analysis. This approach is far from optimal, as it led to only moderate improvement and did not enhance fine-mapping and biological enrichment analysis.

The challenge of using sequence-based deep learning (DL) models in HFS applications is further compounded by their difficulty in predicting variations between individuals. Recent studies(Huang et al., 2023; Sasse et al., 2023) indicate that DL models, trained on the reference human genome, demonstrate limited accuracy in predicting gene expression levels across different individuals. This limitation is likely due to the models' inability to account for long-range regulatory patterns, which are crucial for understanding the impact of variants on gene expression and vary across genes. In contrast, our study leveraged sequence-determined functional genomic profiles in association studies, which mitigates this issue to an extent. For instance, although sei cannot identify the specific gene regulated by a given input sequence, it can predict changes in the sequence's functional activity. Future improvements in DL models' ability to predict interindividual differences could be achieved by incorporating cross-individual data in the training process. An example of such data is the EN-TEX(Rozowsky et al., 2023) dataset, which aligns functional genomic peaks with the specific individuals and haplotypes they correspond to.

In summary, our results demonstrate that incorporating HFS to represent genetically-determined functional genomic activities in genetic association studies offers robust improvements in both the biological interpretation and polygenic prediction of complex traits. Thus, the application of the HFS framework in future genetic association studies holds considerable promise.

Method

Sample description

This study analyzed UK Biobank data, with application ID 84436. We only included participants with array imputed genotype data in bgen format that passed UKB quality control, and removed related individuals. We randomly selected 350,587 self-identified British ancestry Caucasians as training sample. The remaining participants were grouped according to their ancestry, where Non-British European, South Asian, East Asian and African groups serve as test samples.

All phenotypes analyzed (Table S1) were collected from UKB table browser, which came from self-report or physical measurement. Phenotypes were first adjusted by age, sex, top ten principal components, Townsend index and genotype array quality metrics by linear regression. We then applied inverse-normal transformation on the residuals. Binary phenotypes were adjusted in the same way except by generalized linear regression.

Genotype data processing

We first segmented hg38 genome into 4,096 bp loci. To do so, we downloaded chromatin state annotation of 222 human tissues at different developmental stage (embryo, newborn, adult) from epimap(Boix et al., 2021) database. For each tissue, all chromosomal regions annotated as "transcription start site (TSS), transcription region (TX), enhancer, promoter" in at least half of the samples were marked as active regions. The union of active regions across all tissues was taken, and regions annotated as genomic gaps (centromere, ambiguous base pairs, etc.) in the Hg38 genome were removed. Then, for this series of active regions, if the length is less than 4096bp, the locus is defined as a 4096bp area centered around the active region. If the length is greater than 4096bp, 4096bp length loci are gradually delineated from the midpoint outward. Finally, non-overlapping 4096bp blocks were used to cover the remaining genomic regions. This resulted in about 617,378 genomic regions in total. In the sliding-window analysis, all these blocks were

shifted 2048 bp towards 5' end, generating another 617,378 blocks. We repeated the fine-mapping analysis and applied polygenic analysis on these combined blocks, using height as a representative trait.

For each of the loci, we obtained ID of variants within this locus by bedtools(Quinlan and Hall, 2010), then extracted genotypes from UKB .bgen file by bgenix, finally used Plink(Purcell et al., 2007) to remove all variants with INFO<0.8, Hardy-Weinberg $p<10^{-6}$, allele count <10 or missing rate >10%, and removed individual that missed more than 10% of retained variants in this locus. The output vcf file was liftover to hg38 by Crossmap(Zhao et al., 2014) and phased by SHAPEIT4(Delaneau et al., 2019). Phased vcf was transformed to .haps format by Plink, which in turn gave rise to two files: a vcf file containing information of each haplotype, and an n x 2 matrix in plain text that recorded the id of two haplotypes per individual.

HFS calculation

There has been several DL models that predict functional genomic profiles based on DNA sequence(Avsec et al., 2021; Chen et al., 2022; Kelley, 2020; Yan et al., 2021; Zhou et al., 2018). Among them, we chose sei(Chen et al., 2022) to calculate HFS for the following reasons: 1) the required input length (4,096bp) is moderate. 2) it represents 21,906 functional genomic tracks, more comprehensive than other models. 3) it integrated information of the entire sequence, not only the few bp at the center. For each haplotype at each locus, we generated its corresponding DNA sequence by bcftools(Danecek et al., 2021) consensus option. At each locus, the start point of each sequence was matched to the start point of reference sequence. When insertion variants made the sequence longer than 4,096bp, we discarded base pairs at the 3' end. Likewise, with deletion variants, we added N to the 3' end. We applied sei to predict 21,906 functional genomic tracks for each sequence, without normalizing for histone mark (divided each track score by the sum of histone mark score) as suggested by the sei author. We then used the projection matrix provided by sei to calculate forty sequence class scores, which could be regards as the weighted sum of these tracks and represented different aspect of functional genomic activities. We discarded the last score (heterochromatin 6 [centromere]), since its proportion is too low and is functionally trivial, leading to 39 scores per haplotype.

On each individual, we derived from each sequence class score the mean of two haplotypes, corresponding to additive model. For HFS Linkage Disequilibrium calculation, we extracted the mean value of sequence class score corresponding to reference sequence class of adjacent loci, and calculate R^2 value between them. The sequence class score of the reference sequence class was defined as the HFS for this locus, and was used for downstream trait association analysis.

HFS-trait association

For each locus, we calculated the association between trait-specific HFS and adjusted, normalized trait value by linear regression, without any covariates (this is because all selected covariates have been adjusted at the normalization step). For uniformity, we set the significance threshold at $p<5\times10^{-8}$, even if it was over-stringent for n=590,959 loci. Among significant associations, we defined an independent association as the locus with the lowest p value in the 200kb regions. As a positive control, we applied quantitative and binary GWAS with REGENIE(Mbatchou et al., 2021), using default settings and the same British training sample. The main difference is that we used raw trait values in REGENIE, and provided the same covariates. We calculated the genomic control inflation factor, $\lambda_{GC,}$ as the median of X^2 statistics, separately for HFS association test and GWAS (only those SNPs in hapmap3(Altshuler et al., 2010) project were calculated). We compared the λ_{GC} between HFS and SNP by Pearson correlation analysis and paired t-test.

Fine-mapping analysis

We divided hg38 genome into 1,361 independent blocks as defined by MacDonald et al.(MacDonald et al., 2022), and applied SUSIE to HFS of all loci within each block, separately for each trait (parameters: maximum number of causal signal=10, coverage=0.95). We subtracted reference HFS value for each locus prior to analysis, such that homozygous reference haplotype corresponded to HFS=0. To avoid influence of sei prediction noise, we rounded the HFS value at two decimals. This is due to the fact that even if a variant actually makes no impact on functional genomics, sei would still output a value that are close to but not equal to reference sequence class score. Rounding procedure would set such HFS to zero and remove the random value from sei. Loci whose HFS had PIP>0.95 were defined as causal loci, and loci that had causal association with multiple traits were defined as pleiotropic loci. As a positive control, we applied PolyFun(Weissbrod et al., 2020) and SbayesRC(Zheng et al., 2022) on the GWAS summary statistics by REGENIE on the same training set, and extracted the reported PIP to define causal SNP.

To analyze the functional characteristics of causal loci, we first defined the sequence class of each locus by the maximum sequence class score of reference haplotype. We then tested whether each sequence class contained excess causal loci of each trait by Fisher test. For each causal locus, we also defined a "control" locus as the nearest locus that matched the p-value of this causal locus, and tested whether causal loci carried more PolyFun causal SNP than control loci by Fisher test. Furthermore, For traits whose causal loci covered >0.1% of genome-wide SNP, we applied LDSC(Finucane et al., 2015) to quantify the heritability enrichment in causal and control loci, and compare their difference by jackknife method. To avoid winner's curse, we used external GWAS summary statistics for this analysis(Mikaelsdottir et al., 2021; Yengo et al., 2022). As an alternative method to quantify the heritability captured by causal loci, we ran multivariate linear regression in independent British test set where HFS of causal loci were independent variables and trait value were dependent variable, and calculated the R^2 and AIC. We applied the same analysis on causal SNP, and compared AIC between HFS and SNP multivariate regression.

Functional enrichment analysis

Similar to the idea of LDSC(Finucane et al., 2015), we first generated a series of baseline annotation of each locus, then tested whether locus PIP was associated with functional annotations after controlling the impact of these baseline annotations. Specifically, we defined the following baseline annotations:

1. number of haplotypes, range of HFS distribution of all haplotypes (scaled by reference HFS), and 39 sequence class score of reference haplotype.
2. genomic regions of conserved base, high Phastcons score(Siepel et al., 2005) in mammals, primates and vertebrate, exon, intron, untranslated regions at 3' and 5' and 200bp flanking regions of TSS. We used bedtools *intersect -f 0.1* option to annotate each locus by these annotations.
3. maximum B statistics(McVicker et al., 2009), minimum allele age and $ASMC_{avg}$(Palamara et al., 2018) of all variants within this locus.

Type two and three annotations were directly obtained from LDSC(Finucane et al., 2015) baseline annotations. We did not include annotations related to functional genomics, since 39 sequence class score were used to capture functional genomic characteristics. Conditioned on these baseline annotations, we analyzed the enrichment of PIP in the following functional annotations:

1. Biological pathways: We downloaded all pathways from MsigDB(Subramanian et al., 2005) C2: canonical pathways category (including Reactome(Fabregat et al., 2018), PID(Schaefer et al., 2009), Biocarta and Wikipathway) and C6: Gene ontology(Ashburner et al., 2000) (biological process) category. We retained only pathways with >5 genes and <500 genes. We generated a gene × pathway binary matrix and applied hierarchical clustering so that similar pathways were placed close to each other. We sequentially compared adjacent pathways, and removed the smaller one if the fraction of overlap > 30%. A total of 3,219 pathways were retained. We then linked each locus to these pathways by cS2G(Gazal et al., 2022) strategy. Specifically, a locus L would be annotated as 1 for pathway P only if L contained a SNP that was link to P with cS2G score>0.5.
2. Tissue-specific chromatin activity: We downloaded chromHMM(Ernst and Kellis, 2012) chromatin state annotation for 833 samples from epimap(Boix et al., 2021), and grouped them according to developmental stages and second-level tissue types. For each group, all chromosomal regions annotated as "transcription start site (TSS), transcription region (TX), enhancer, promoter" in at least half of the samples were marked as active regions. We used bedtools *intersect -f 0.1* option to annotate whether each locus was active in each tissue.
3. cell type-specific open chromatin regions: we downloaded scATAC-seq peak data from Zhang et al.(K. Zhang et al., 2021), and annotated each locus by bedtools *intersect -f 0.1* option.

We applied multivariate linear regression of PIP against baseline annotations + one of the functional annotations. Regression coefficient > 0 and Bonferroni-adjusted regression p value<0.05 were used as significance threshold. From the final results, we manually removed those pathways and cell types that reached significance threshold in more than half of the traits, since these pathways likely reflected unrecognized confounders.

Polygenic prediction

We used the posterior effect size estimated by SUSIE on sliding window strategy (doubling the number of loci) as weights, and calculated the weighted sum of HFS as the polygenic risk score of each trait, and calculated R^2 in independent British test sample with simple linear regression. As a positive control, we applied LDAK-BOLT(Q. Zhang et al., 2021) algorithm on the SNP array data (about seven hundred thousand variants) with tenfold cross-validation and max iteration=200 in the same training sample, and calculated SNP-based PRS with the output SNP weights. Normalized trait values were analyzed, without any covariates provided. Array data were filtered by Plink with option --geno 0.1 --hwe 1e-6 --mac 100 --maf 0.01 --mind 0.1.

To train the refined model that predict height, we first calculated per-block HFS-based prediction score of height as the weighted sum of HFS within this block. Then, within each target ancestry group (non-British European (NBE), South Asian (SAS), East Asian (EA) and African (AFR) participants in UK Biobank), we randomly selected half as tuning sample and half as test sample. In the tuning sample, we applied LASSO regression that included both LDAK PRS and genome-wide per-block HFS score (1,361 in total). The choice of LASSO regression was based on a comparison on British European test set (**Figure 5B**), where LASSO, ridge and elastic net gave similar results and LASSO was relatively better. In the tuning sample of target ancestry, LASSO estimated the weights to combine per-block HFS score and LDAK PRS. We calculated the final prediction score in the test sample using these weights, and evaluated its prediction by linear regression R^2. Since the outcome (height) has already been adjusted and standardized, no covariates were included in this step. Additionally, we applied PolyFun-pred(Weissbrod et al., 2022) and SbayesRC(Zheng et al., 2022) to the summary statistics of height (calculated by REGENIE in the same training sample), and integrated their effect size with LDAK weight in the

tuning sample using Polypred(Weissbrod et al., 2022) method. PRS for LDAK, LDAK+PolyFun and LDAK+SbayesRC were calculated by plink *score* option, excluding variants with INFO<0.8, Hardy-Weinberg $p<10^{-6}$, allele count <2 or missing rate >10% in the target test set.

Statistical analysis

All p-values were two-sided and adjusted by Bonferroni unless otherwise specified. For group comparison, we used Fisher test for count data and paired t-test for continuous data. For R^2 of PRS comparison, we applied r2redux(Momin et al., 2023) R package to estimate 95% confidence interval and its p-value for the difference of R^2.

Availability of data and materials

This work was based on UK Biobank data from project 84436. Code for this analysis is available at *https://github.com/WeiCSong/HFS*

Acknowledgements

This work was supported by grants from the 2030 Science and Technology Innovation Key Program of Ministry of Science and Technology of China (No. 2022ZD020910001), the National Natural Science Foundation of China (No. 81971292, 82150610506) and the Natural Science Foundation of Shanghai (No. 21ZR1428600), the Medical-Engineering Cross Foundation of Shanghai Jiao Tong University (No. YG2022ZD026).

Authors' contributions

W.S designed the study, collected and analyzed the data and wrote the manuscript. G.N.L and Y.S supervised the study. All authors read, revise and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Aguet F *et al.* (2020) **The GTEx Consortium atlas of genetic regulatory effects across human tissues** *Science (80-)* **369**:1318–1330 https://doi.org/10.1126/SCIENCE.AAZ1776

2. Altshuler DM *et al.* (2010) **Integrating common and rare genetic variation in diverse human populations** *Nature* **467**:52–58 https://doi.org/10.1038/nature09298

3. Ashburner M *et al.* (2000) **Gene Ontology: tool for the unification of biology** *Nat Genet* **25**:25–29 https://doi.org/10.1038/75556

4. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR (2021) **Effective gene expression prediction from sequence by integrating long-range interactions** *Nat Methods* **18** https://doi.org/10.1038/S41592-021-01252-X

5. Baca SC *et al.* (2022) **Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation** *Nat Genet* **54** https://doi.org/10.1038/S41588-022-01168-Y

6. Barbeira AN *et al.* (2018) **Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics** *Nat Commun* **9**:1–20 https://doi.org/10.1038/s41467-018-03621-1

7. Boix CA, James BT, Park YP, Meuleman W, Kellis M (2021) **Regulatory genomic circuitry of human disease loci by integrative epigenomics** *Nature* **590**:300–307 https://doi.org/10.1038/s41586-020-03145-z

8. Chen KM, Wong AK, Troyanskaya OG, Zhou J (2022) **A sequence-based global map of regulatory activity for deciphering human genetics** *Nat Genet* **54**:940–949 https://doi.org/10.1038/s41588-022-01102-2

9. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM (2021) **Twelve years of SAMtools and BCFtools** *Gigascience* **10**:1–4 https://doi.org/10.1093/GIGASCIENCE/GIAB008

10. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET (2019) **Accurate, scalable and integrative haplotype estimation** *Nat Commun 2019 101* **10**:1–10 https://doi.org/10.1038/s41467-019-13225-y

11. Ernst J, Kellis M (2012) **ChromHMM: Automating chromatin-state discovery and characterization** *Nat Methods* https://doi.org/10.1038/nmeth.1906

12. Fabregat A *et al.* (2018) **The Reactome Pathway Knowledgebase** *Nucleic Acids Res* **46**:D649–D655 https://doi.org/10.1093/nar/gkx1132

13. Finucane HK *et al.* (2015) **Partitioning heritability by functional annotation using genome-wide association summary statistics** *Nat Genet* **47**:1228–1235 https://doi.org/10.1038/ng.3404

14. Finucane HK *et al.* (2018) **Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types** *Nat Genet* **50**:621–629 https://doi.org/10.1038/s41588-018-0081-4

15. Gazal S *et al.* (2022) **Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity** *Nat Genet* **54**:827–836 https://doi.org/10.1038/s41588-022-01087-y

16. Grotzinger AD, de la Fuente J, Davies G, Nivard MG, Tucker-Drob EM (2022) **Transcriptome-wide and stratified genomic structural equation modeling identify neurobiological pathways shared across diverse cognitive traits** *Nat Commun 2022 131* **13**:1–15 https://doi.org/10.1038/s41467-022-33724-9

17. Helland IB, Smith L, Saarem K, Saugstad OD, Drevon CA (2003) **Maternal Supplementation With Very-Long-Chain n-3 Fatty Acids During Pregnancy and Lactation Augments Children's IQ at 4 Years of Age** *Pediatrics* **111**:e39–e44 https://doi.org/10.1542/PEDS.111.1.E39

18. Hu X *et al.* (2022) **Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program** *Am J Hum Genet* **0** https://doi.org/10.1016/J.AJHG.2022.03.007

19. Huang C, Shuai R, Baokar P, Chung R, Rastogi R, Kathail P, Ioannidis N (2023) **Personal transcriptome variation is poorly explained by current genomic deep learning models** *bioRxiv* https://doi.org/10.1101/2023.06.30.547100

20. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, Timpson NJ, Dunham I, Birney E, Soranzo N (2019) **GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals** *Nat Genet* **51**:343–353 https://doi.org/10.1038/s41588-018-0322-6

21. Jumper J *et al.* (2021) **Highly accurate protein structure prediction with AlphaFold** *Nat 2021 5967873* **596**:583–589 https://doi.org/10.1038/s41586-021-03819-2

22. Kelley DR (2020) **Cross-species regulatory sequence activity prediction** *PLoS Comput Biol* **16** https://doi.org/10.1371/JOURNAL.PCBI.1008050

23. Kong X, Ma L, Chen E, Shaw CA, Edelstein LC (2019) **Identification of the Regulatory Elements and Target Genes of Megakaryopoietic Transcription Factor MEF2C** *Thromb Haemost* **119** https://doi.org/10.1055/S-0039-1678694

24. Xihao Li *et al.* (2020) **Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale** *Nat Genet* **52**:969–983 https://doi.org/10.1038/s41588-020-0676-4

25. Liang Y, Pividori M, Manichaikul A, Palmer AA, Cox NJ, Wheeler HE, Im HK (2022) **Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries** *Genome Biol* **23**:1–18 https://doi.org/10.1186/S13059-021-02591-W/FIGURES/5

26. Liu Z, Pan W, Li W, Zhen X, Liang J, Cai W, Xu F, Yuan K, Lin GN (2022) **Evaluation of the Effectiveness of Derived Features of AlphaFold2 on Single-Sequence Protein Binding Site Prediction** *Biology (Basel)* **11** https://doi.org/10.3390/BIOLOGY11101454/S1

27. MacDonald JW, Harrison T, Bammler TK, Mancuso N, Lindström S (2022) **An updated map of GRCh38 linkage disequilibrium blocks based on European ancestry data** *bioRxiv* https://doi.org/10.1101/2022.03.04.483057

28. Mbatchou J *et al.* (2021) **Computationally efficient whole-genome regression for quantitative and binary traits** *Nat Genet* **53**:1097–1103 https://doi.org/10.1038/s41588-021-00870-7

29. McVicker G, Gordon D, Davis C, Green P (2009) **Widespread genomic signatures of natural selection in hominid evolution** *PLoS Genet* **5** https://doi.org/10.1371/journal.pgen.1000471

30. Mikaelsdottir E *et al.* (2021) **Genetic variants associated with platelet count are predictive of human disease and physiological markers** *Commun Biol* **4** https://doi.org/10.1038/S42003-021-02642-9

31. Momin MM, Lee S, Wray NR, Lee SH (2023) **Significance tests for R2 of out-of-sample prediction using polygenic scores** *Am J Hum Genet* **0** https://doi.org/10.1016/J.AJHG.2023.01.004

32. Morales MAG, Montero-vargas JM, Vizuet-de-rueda JC, Teran LM (2021) **New Insights into the Role of PD-1 and Its Ligands in Allergic Disease** *Int J Mol Sci* **22** https://doi.org/10.3390/IJMS222111898

33. Nawijn MC *et al.* (2011) **Identification of the Mhc Region as an Asthma Susceptibility Locus in Recombinant Congenic Mice** *Am J Respir Cell Mol Biol* **45** https://doi.org/10.1165/RCMB.2009-0369OC

34. Nowbandegani PS, Wohns AW, Ballard JL, Lander ES, Bloemendal A, Neale BM, O'Connor LJ (2022) **Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies** *bioRxiv* https://doi.org/10.1101/2022.09.06.506858

35. Null M, Dupuis J, Sheinidashtegol P, Layer RM, Gignoux CR, Hendricks AE (2022) **RAREsim: A simulation method for very rare genetic variants** *Am J Hum Genet* **0** https://doi.org/10.1016/J.AJHG.2022.02.009

36. Palamara PF, Terhorst J, Song YS, Price AL (2018) **High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability** *Nat Genet* **50**:1311–1317 https://doi.org/10.1038/s41588-018-0177-x

37. Park CY, Zhou J, Wong AK, Chen KM, Theesfeld CL, Darnell RB, Troyanskaya OG (2021) **Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk** *Nat Genet* **53** https://doi.org/10.1038/S41588-020-00761-3

38. Pejaver V *et al.* (2020) **Inferring the molecular and phenotypic impact of amino acid variants with MutPred2** *Nat Commun* **11**:1–13 https://doi.org/10.1038/s41467-020-19669-x

39. Purcell S *et al.* (2007) **PLINK: A tool set for whole-genome association and population-based linkage analyses** *Am J Hum Genet* **81**:559–575 https://doi.org/10.1086/519795

40. Quinlan AR, Hall IM (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–842 https://doi.org/10.1093/BIOINFORMATICS/BTQ033

41. Rozowsky J *et al.* (2023) **The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models** *Cell* **186**:1493–1511 https://doi.org/10.1016/J.CELL.2023.02.018

42. Sasse A, Ng B, Spiro AE, Tasaki S, Bennett DA, Gaiteri C, De Jager PL, Chikina M, Mostafavi S, Allen PG (2023) **Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings** *bioRxiv* https://doi.org/10.1101/2023.03.16.532969

43. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) **PID: the Pathway Interaction Database** *Nucleic Acids Res* **37** https://doi.org/10.1093/NAR/GKN653

44. Sekar A *et al.* (2016) **Schizophrenia risk from complex variation of complement component 4** *Nature* **530** https://doi.org/10.1038/NATURE16549

45. Siepel A *et al.* (2005) **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes** *Genome Res* **15**:1034–1050 https://doi.org/10.1101/GR.3715005

46. Subramanian A *et al.* (2005) **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles** *Proc Natl Acad Sci U S A* **102**:15545–15550 https://doi.org/10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG

47. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) **10 Years of GWAS Discovery: Biology, Function, and Translation** *Am J Hum Genet* **101**:5–22 https://doi.org/10.1016/J.AJHG.2017.06.005

48. Wang G, Sarkar A, Carbonetto P, Stephens M (2020) **A simple new approach to variable selection in regression, with application to genetic fine mapping** *J R Stat Soc Ser B (Statistical Methodol* **82**:1273–1300 https://doi.org/10.1111/RSSB.12388

49. Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D (2019) **A global overview of pleiotropy and genetic architecture in complex traits** *Nat Genet* **51**:1339–1348 https://doi.org/10.1038/s41588-019-0481-0

50. Weissbrod O *et al.* (2020) **Functionally informed fine-mapping and polygenic localization of complex trait heritability** *Nat Genet* **52**:1355–1363 https://doi.org/10.1038/s41588-020-00735-5

51. Weissbrod O *et al.* (2022) **Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores** *Nat Genet* **54**:450–458 https://doi.org/10.1038/s41588-022-01036-9

52. Yan J *et al.* (2021) **Systematic analysis of binding of transcription factors to noncoding variants** *Nature* **591**:147–151 https://doi.org/10.1038/s41586-021-03211-0

53. Yang J *et al.* (2015) **Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index** *Nat Genet* **47**:1114–1120 https://doi.org/10.1038/ng.3390

54. Yang J *et al.* (2010) **Common SNPs explain a large proportion of the heritability for human height** *Nat Genet* **42**:565–569 https://doi.org/10.1038/ng.608

55. Yengo L *et al.* (2022) **A saturated map of common genetic variants associated with human height** *Nature* **610**:704–712 https://doi.org/10.1038/s41586-022-05275-y

56. Zeng J *et al.* (2021) **Widespread signatures of natural selection across human complex traits and functional genomic categories** *Nat Commun* **12** https://doi.org/10.1038/s41467-021-21446-3

57. Zhang K *et al.* (2021) **A single-cell atlas of chromatin accessibility in the human genome** *Cell* **184**:1–17 https://doi.org/10.1016/j.cell.2021.10.024

58. Zhang Q, Privé F, Vilhjálmsson B, Speed D (2021) **Improved genetic prediction of complex traits from individual-level data or summary statistics** *Nat Commun* **12**:1–9 https://doi.org/10.1038/s41467-021-24485-y

59. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L (2014) **CrossMap: a versatile tool for coordinate conversion between genome assemblies** *Bioinformatics* **30**:1006–1007 https://doi.org/10.1093/BIOINFORMATICS/BTT730

60. Zheng Z *et al.* (2022) **Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries** *bioRxiv* https://doi.org/10.1101/2022.10.12.510418

61. Zhou J (2022) **Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale** *Nat Genet* **54**:725–734 https://doi.org/10.1038/s41588-022-01065-4

62. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG (2018) **Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk** *Nat Genet* **50**:1171–1179 https://doi.org/10.1038/s41588-018-0160-6

63. Zhou W, Bi W, Zhao Z, Dey KK, Jagadeesh KA, Karczewski KJ, Daly MJ, Neale BM, Lee S (2022) **SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests** *Nat Genet* **54**:1466–1469 https://doi.org/10.1038/s41588-022-01178-w

Article and author information

Weichen Song

Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, School of Bioengineering, Shanghai Jiaotong University, Shanghai, China, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China

For correspondence: goubegou@sjtu.edu.cn

Yongyong Shi

Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China, Biomedical Sciences Institute of Qingdao University (Qingdao Branch of SJTU Bio-X Institutes), Qingdao University, Qingdao, China

For correspondence: shiyongyong@gmail.com

Guan Ning Lin

Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, School of Bioengineering, Shanghai Jiaotong University, Shanghai, China

For correspondence: nickgnlin@sjtu.edu.cn

Editors

Reviewing Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Tübingen, Germany

Senior Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Tübingen, Germany

Reviewer #1 (Public Review):

Summary:

In this paper, Song, Shi, and Lin use an existing deep learning-based sequence model to derive a score for each haplotype within a genomic region, and then perform association tests between these scores and phenotypes of interest. The authors then perform some downstream analyses (fine-mapping, various enrichment analyses, building polygenic scores) to ensure that these associations are meaningful. The authors find that their approach allows them to find additional associations, the associations have biologically interpretable enrichments in terms of tissues and pathways, and can slightly improve polygenic scores when combined with standard SNP-based PRS.

Strengths:

- I found the central idea of the paper to be conceptually straightforward and an appealing way to use the power of sequence models in an association testing framework.

- The findings are largely biologically interpretable, and it seems like this could be a promising approach to boost power for some downstream applications.

Weaknesses:

- While not a weakness of the manuscript, the proposed method is computationally intensive.

https://doi.org/10.7554/eLife.92574.2.sa1

Reviewer #2 (Public Review):

Summary:

In this work, Song et al. propose a locus-based framework for performing GWAS and related downstream analyses including finemapping and polygenic risk score (PRS) estimation. GWAS are not sufficiently powered to detect phenotype associations with low-frequency variants. To overcome this limitation, the manuscript proposes a method to aggregate variant impacts on chromatin and transcription across a 4096 base pair (bp) loci in the form of a haplotype function score (HFS). At each locus, an association is computed between the HFS and trait. Computing associations at the level of imputed functional genomic scores enables

integration of information across variants spanning the allele frequency spectrum and bolster the power of GWAS.

The HFS for each locus is derived from a sequence-based predictive model - Sei. Sei predicts 21,907 chromatin and TF binding tracks, which can be projected onto 40 pre-defined sequence classes (representing promoters, enhancers etc.). For each 4096 bp haplotype in their UKB cohort, the proposed method uses the Sei sequence class scores to derive the haplotype function score (HFS). The authors apply their method to 14 polygenic traits, identifying ~16,500 HFS-trait associations. They finemap these trait-associated loci with SuSie, as well perform target gene/pathway discovery and PRS estimation.

Strengths:

Sequence-based deep learning predictors of chromatin status and TF binding have become increasingly accurate over the past few years. Imputing aggregated variant impact using Sei, and then performing an HFS-trait association is therefore an interesting approach to bolster power in GWAS discovery. The manuscript demonstrates that region-level associations can be identified at the level of an aggregated functional score using sequence-based deep learning models. The finemapping and pathway identification analyses suggest that HFS-based associations identify relevant causal pathways and genes from an association study. Identifying associations at the level of functional genomics increases portability of PRSs across populations. Imputing functional genomic predictions using a sequence-based deep learning model does not suffer from the limitation of TWAS where gene expression is imputed from a limited size reference panel such as GTEx and is an interesting direction to bolster discovery power.

However, a few limitations to this method in its current form are:

(1) HFS-based association is going to miss coding variation as well as noncoding regulatory variants such as splicing variants/polyadenylation variants which are not modeled by Sei. This will lead to false negatives in the HFS-based association and additionally false negatives + associated false positives in the finemapping. Going forward, it'll therefore be important to characterize how this influences the genome-wide finemapping.

(2) Sei predicts chromatin status / ChIP-seq peaks in the center of a 4kb region. It is thus not clear therefore whether the functional effects of variants not in the center of the 4kb region would be captured in a single Sei score. It also remains unclear how much the choice of window affects the association tests / finemapping.

(3) There are going to be cases where there's an association driven by a variant that is correlated with a Sei prediction in a neighboring window. These would represent false positives for the method, it would be useful to identify or characterize these cases.

Minor Concerns:
(1) Sequence based deep learning model predictions can be miscalibrated for insertions and deletions (INDELs) as compared to SNPs. It'll be important to note that model INDEL scores may not be calibrated, which might also lead to false positives / false negatives in the finemapping.

https://doi.org/10.7554/eLife.92574.2.sa0

Author Response

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public Review):

Summary:

In this paper, Song, Shi, and Lin use an existing deep learning-based sequence model to derive a score for each haplotype within a genomic region, and then perform association tests between these scores and phenotypes of interest. The authors then perform some downstream analyses (fine-mapping, various enrichment analyses, and building polygenic scores) to ensure that these associations are meaningful. The authors find that their approach allows them to find additional associations, the associations have biologically interpretable enrichments in terms of tissues and pathways, and can slightly improve polygenic scores when combined with standard SNP-based PRS.

Strengths:

- *I found the central idea of the paper to be conceptually straightforward and an appealing way to use the power of sequence models in an association testing framework.*
- *The findings are largely biologically interpretable, and it seems like this could be a promising approach to boost power for some downstream applications.*

Weaknesses:

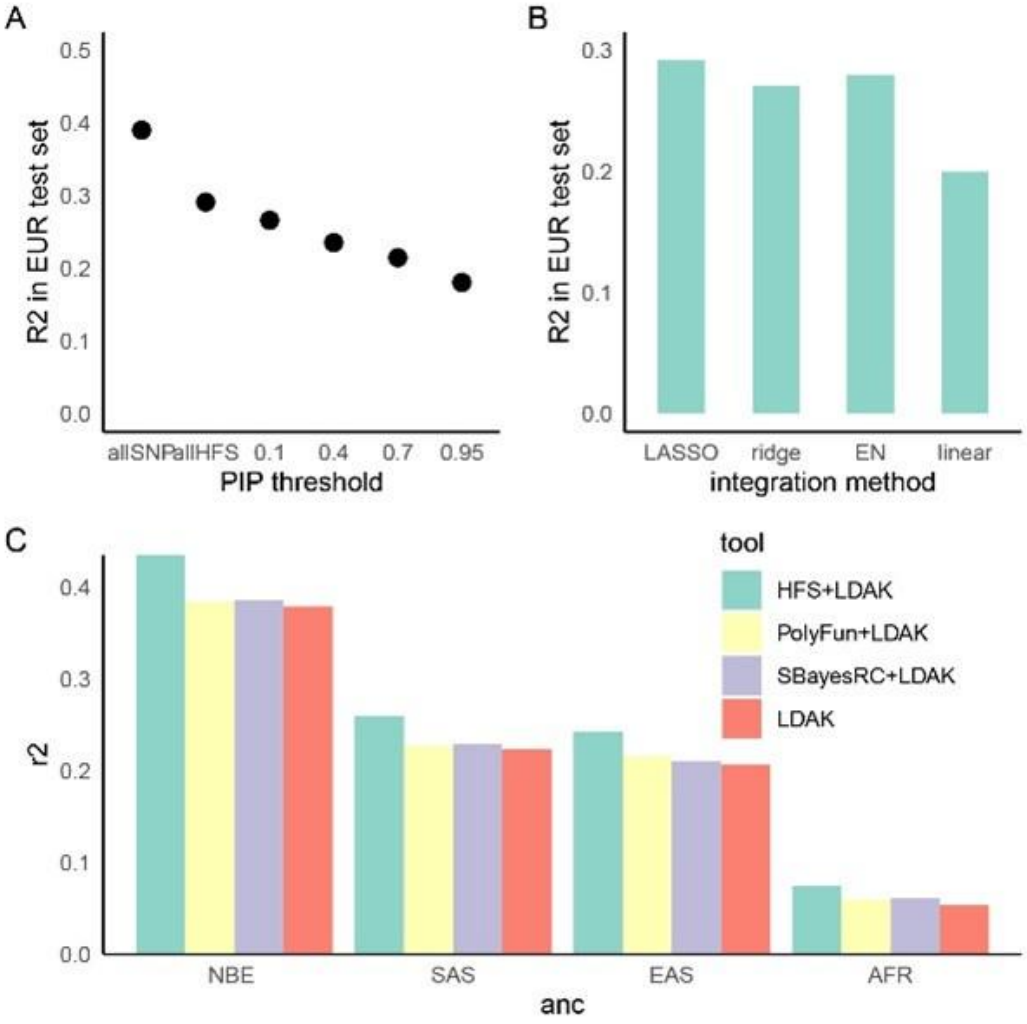
- *The methods used to generate polygenic scores were difficult to follow. In particular, a fully connected neural network with linear activations predicting a single output should be equivalent to linear regression (all intermediate layers of the network can be collapsed using matrix-multiplication, so the output is just the inner product of the input with some vector). Using the last hidden layer of such a network for downstream tasks should also be equivalent to projecting the input down to a lower dimensional space with some essentially randomly chosen projection. As such, I am surprised that the neural network approach performs so well, and it would be nice if the authors could compare it to other linear approaches (e.g., LASSO or ridge regression for prediction; PCA or an auto-encoder for converting the input to a lower dimensional representation).*

Response: We thank the reviewer for the recognition and valuable suggestion on our work. Just as the reviewer suggested, our polygenic prediction procedure is equivalent to linear transformation and in this revision, we indeed found that it was unnecessary to use neural network framework to replace linear model. Indeed, both our result and previous work indicated that linear model fitted polygenic traits better than non-linear one, which was also the reason we chose linear activation for neural network in the original manuscript.

In this revision, we followed the reviewer's suggestion to apply a more straightforward linear framework for polygenic prediction. We first calculated weighted sum of HFS for each block (1,361 independent blocks in total), then, in each target ancestry, we used LASSO regression to integrate them with SNP PRS into one final score. We also conducted comparative analysis in British European test set and found that LASSO, ridge and elastic net gave similar result, and LASSO performed slightly better. By applying this straightforward framework and sliding window strategy, we moderately improved the prediction performance.

Line 349: “Using height as a representative trait, we first estimated the proportion of variance captured by top loci, and found that HFS of loci with PIP>0.4 (n=5,101) captured roughly 80% of variance explained by all genome-wide loci (n=1,200,024 corresponded to sling-window strategy; Figure 5A). We then calculated HFS+LDAK in non-British European (NBE), South Asian (SAS), East Asian (EAS) and African (AFR) population in UK Biobank, and observed 17.5%, 16.1%, 17.2% and 39.8% improvement over LDAK alone (p=3.21×10-16, 0.0001, 0.002 and 0.001, respectively. Figure 5C).”

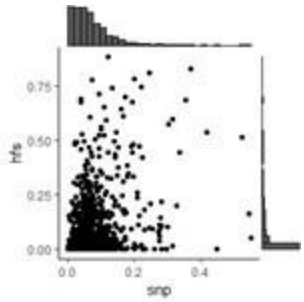
Author response image 1.



- *A very interesting point of the paper was the low R^2 between the HFS scores in adjacent windows, but the explanation of this was unclear to me. Since the HFS scores are just deterministic functions of the SNPs, it feels like if the SNPs are in LD then the HFS scores should be and vice versa. It would be nice to compare the LD between adjacent windows to the average LD of pairs of SNPs from the two windows to see if this is driven by the fact that SNPs are being separated into windows, or if sei is somehow upweighting the importance of SNPs that are less linked to other SNPs (e.g., rare variants).*

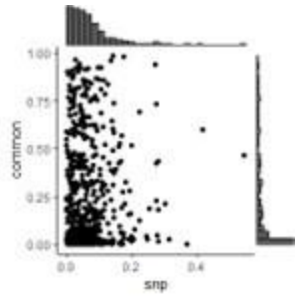
Response: We thank the reviewer for the suggestion on understanding LD mechanism. In this revision, we used chromosome 1 as an example and calculate the pairwise LD among all SNPs within two adjacent loci. As shown in Figure S1 (below), although HFS-based LD is still significantly lower than median SNP-based LD (paired Wilcoxon test p=1.76e-5), we found that median SNP LD between loci was still lower than what typically observed between adjacent SNPs in GWAS (histogram of x axis; median =0.06). We reasoned that dividing SNPs into block is one of the reasons that HFS suffer less LD than standard GWAS, but not the whole story.

Author response image 2.



We agree with the reviewer that the effect of rare variants could also play an important role. In fact, sei author has also found that rare variants tended to have larger sei-predicted effects. We conducted an approximate analysis that remove all rare variants and repeated HFS calculation. Indeed, here HFS LD has profoundly raised to median=0.14, indicating that involving rare variants was vital for low LD.

Author response image 3.



Line 123: “Further evaluation indicated that this low LD was led by two factors: integration of rare variant impacts and segmentation. Firstly, excluding rare variants from HFS caused the LD raised to median=0.14 (Method; Figure S2C). Secondly, median LD of SNPs from adjacent loci was 0.06, which was significantly higher than HFS LD (paired Wilcoxon p=1.76×10-5) but significantly lower than HFS LD without rare variants (paired Wilcoxon p<2.2×10-16).”

> - *There were also a number of robustness checks that would have been good to include in the paper. For instance, do the findings change if the windows are shifted? Do the findings change if the sequence is reverse-complemented?*

Response: Following the reviewer’s suggestion, we conducted a sliding window analysis where all loci were shifted 2048 bp, thereby doubling the total number of loci. In fine-

mapping analysis, more than 90% of the causal loci were reproduced in sliding window analysis, either by themselves or by a overlapping locus:

Line 207: “29.4% of causal loci (PIP>0.95) in the original analysis were still causal in sliding window analysis. 31.1% and 29.3% of causal loci whose 5’ and 3’ overlapping locus had PIP>0.95 in sliding window analysis, respectively, while themselves were no longer causal.”

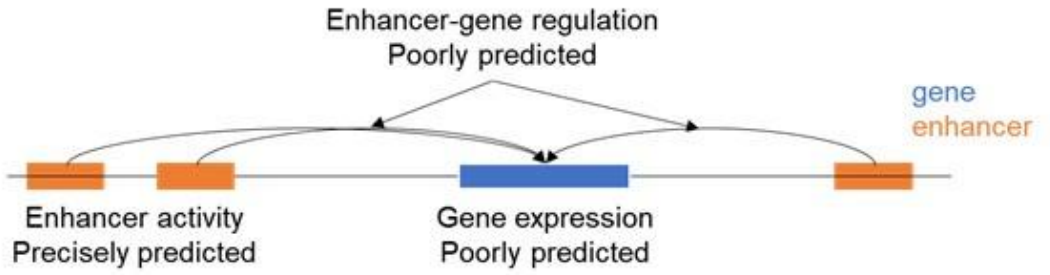
In polygenic prediction analysis, sliding window strategy significantly improved prediction accuracy, as we discussed in question 1.

As for the issue of reverse complement, the nature of sei input layer is to encode both strand in a symmetric manner, such that the output for both strands would be the same. We have also run sei on the reverse complement (generated by seqkit seq -r -p) to verify that original sequence and reverse complement give the same output.

> - *It was also difficult to contextualize the present work in terms of recent results showing that sequence models tend to not do very well at predicting cross-individual expression changes (and such results presumably hold for predicting cross-individual chromatin changes). In particular, it would be good for the authors to contrast their findings with the work of Alex Sasse and colleagues (https://www.biorxiv.org/content/10.1101/2023.03.16.532969.abstract) and Connie Huang and colleagues (https://www.biorxiv.org/content/10.1101/2023.06.30.547100.abstract).*

Response: Following the reviewer’s suggestion, we added a new discussion paragraph on the issue of sequence model performance on interindividual variations. In brief, we suggest that although the drawback of lack of cross-individual training sets exists and future improvement is necessary, chromatin changes could be better predicted than gene expression. This is because the latter task requires information on long range interaction, which varies among genes and are difficult to be captured by using reference genome as training set. We made a schematic to clarify this:

Author response image 4.



We also noticed a few recent studies that directly validated sei predictions by experiments and showed significant accuracy, such as https://doi.org/10.1016/j.neuron.2022.12.026. Taken together, while we agreed that it is necessary to improve sequence model by adding more cross-individual training samples, the current SOTA model sei could still provide unique value to our study.

Line 423: “The challenge of using sequence-based deep learning (DL) models in HFS applications is further compounded by their difficulty in predicting variations between individuals. Recent studies(Huang et al., 2023; Sasse et al., 2023) indicate that DL models, trained on the reference human genome, demonstrate limited accuracy in predicting gene

expression levels across different individuals. This limitation is likely due to the models' inability to account for long-range regulatory patterns, which are crucial for understanding the impact of variants on gene expression and vary across genes. In contrast, our study leveraged sequence-determined functional genomic profiles in association studies, which mitigates this issue to an extent. For instance, although sei cannot identify the specific gene regulated by a given input sequence, it can predict changes in the sequence's functional activity. Future improvements in DL models' ability to predict interindividual differences could be achieved by incorporating cross-individual data in the training process. An example of such data is the EN-TEX(Rozowsky et al., 2023) dataset, which aligns functional genomic peaks with the specific individuals and haplotypes they correspond to."

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *In this work, Song et al. propose a locus-based framework for performing GWAS and related downstream analyses including finemapping and polygenic risk score (PRS) estimation. GWAS are not sufficiently powered to detect phenotype associations with low-frequency variants. To overcome this limitation, the manuscript proposes a method to aggregate variant impacts on chromatin and transcription across a 4096 base pair (bp) loci in the form of a haplotype function score (HFS). At each locus, an association is computed between the HFS and trait. Computing associations at the level of imputed functional genomic scores should enable the integration of information across variants spanning the allele frequency spectrum and bolster the power of GWAS.*
>
> *The HFS for each locus is derived from a sequence-based predictive model. Sei. Sei predicts 21,907 chromatin and TF binding tracks, which can be projected onto 40 pre-defined sequence classes (representing promoters, enhancers, etc.). For each 4096 bp haplotype in their UKB cohort, the proposed method uses the Sei sequence class scores to derive the haplotype function score (HFS). The authors apply their method to 14 polygenic traits, identifying ~16,500 HFS-trait associations. They finemap these trait-associated loci with SuSie, as well as perform target gene/pathway discovery and PRS estimation.*
>
> *Strengths:*
>
> *Sequence-based deep learning predictors of chromatin status and TF binding have become increasingly accurate over the past few years. Imputing aggregated variant impact using Sei, and then performing an HFS-trait association is, therefore, an interesting approach to bolster power in GWAS discovery. The manuscript demonstrates that associations can be identified at the level of an aggregated functional score. The finemapping and pathway identification analyses suggest that HFS-based associations identify relevant causal pathways and genes from an association study. Identifying associations at the level of functional genomics increases the portability of PRSs across populations. Imputing functional genomic predictions using a sequence-based deep learning model does not suffer from the limitation of TWAS where gene expression is imputed from a limited-size reference panel such as GTEx.*
>
> *However, there are several major limitations that need to be addressed.*
>
> *Major concerns/weaknesses:*
>
> *(1) There is limited characterization of the locus-level associations to SNP-level associations. How does the set of HFS-based associations differ from SNP-level associations?*

Response: We thank the reviewer for the recognition and the valuable suggestion on our manuscript. Following the reviewer's suggestion, in this revision we added a paragraph to compare the basic characteristics between HFS-based and SNP-based association study. These comparisons suggested that HFS had no advantage in testing marginal association, but performed better in detecting causal associations.

Line 144: "When comparing HFS association with the standard SNP-based GWAS on the same data, we found that 98% of significant HFS loci also harbored a significant SNP. There were a few cases (n=0~5) where significant HFS loci did not harbored even marginal SNP association (GWAS p>0.01), which were due to the lack of common SNP in these loci. HFS association p value was higher than GWAS p value in 95 % of significant loci, suggested that HFS did not improve power to detect marginal effect. The genomic control inflation factor (λGC) for the HFS association test varied between 0.99 for asthma and 1.50 for height, closely resembling the SNP GWAS (Pearson Correlation Coefficient [PCC]=0.91, paired t-test p=0.16; Method and Figure S3). We concluded that HFS-based association tests had adequate power and do not introduce additional p-value inflation."

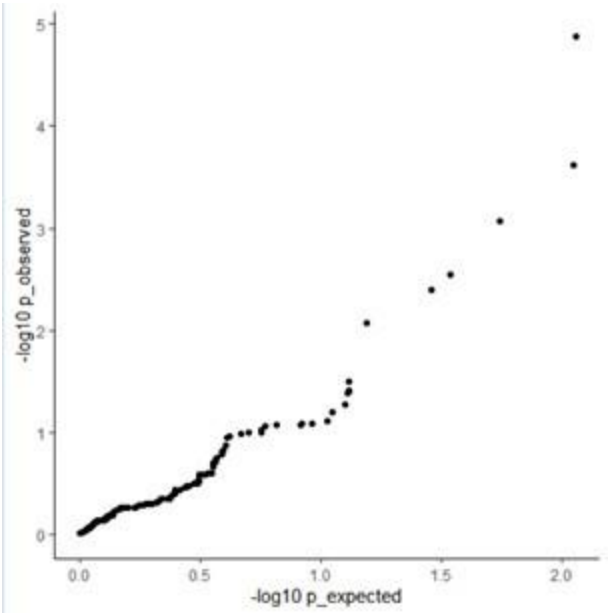
> *(2) A clear advantage of performing HFS-trait associations is that the HFS score is imputed by considering variants across the allele frequency spectrum. However, no evidence is provided demonstrating that rare variants contribute to associations derived by the model. Similarly, do the authors find evidence that allelic heterogeneity is leveraged by the HFS-based association model? It would be useful to do simulations here to characterize the model behavior in the presence of trait-associated rare variants.*

Response: Following the reviewer's suggestion, we conducted a sensitivity analysis that removed all rare (MAF<0.01) variants and repeated the HFS analysis (HFScommon) on chromosome 1. In linear association analysis, we found that 10.6% of HFS signals (p<5×10-8) were missed by HFScommon. In fine-mapping, 55.3% of HFS causal signals (PIP>0.95) were missed by HFScommon. We concluded that rare variants played an important role in the performance of HFS, especially its advantages in fine-mapping.

Line 175: "We also found that rare variants played an important role in the good find-mapping performance of HFS: when variants with MAF<0.01 were removed, 55.3% of the causal signals would be missed in HFS+SUSIE analysis."

We then attempted to conduct a simulation analysis where rare variants were causal to the phenotype, and the association statistics were the same as real GWAS of height. However, such simulation seemed not to properly reflect real scenario: no matter how we changed the association between rare variants and the phenotype, HFS association p-value could hardly reached the significance level of SNP association. We proposed that this is because simulation could not properly reflect how variants impact functional genomics: in fact, when randomly selected a rare variant as causal variant, there is high possibility that it had no impact on functional genomics, therefore its HFS would be close to zero. When such a variant was set as causal (which is unlikely in real scenario), HFS would not properly capture the association. We reasoned that it might be difficult to evaluate HFS by simulation, since the nonlinear relation between SNP and HFS as well as among SNPs were difficult to be properly simulated.

Author response image 5.



> *(3) Sei predicts chromatin status / ChIP-seq peaks in the center of a 4kb region. It would therefore be more relevant to predict HFS using overlapping sequence windows that tile the genome as opposed to using non-overlapping windows for computing HFS scores. Specifically, in line 482, the authors state that "the HFS score represents overall activity of the entire sequence, not only the few bp at the center", but this would not hold given that Sei is predicting activity at the center for any sequence.*

Response: We thank the reviewer for the suggestion on sliding window design. In this revision, we shifted all loci 2,048 bp to double the number of loci and repeated the fine-mapping and polygenic prediction analysis. For fine-mapping, we found that the result was generally robust with regard to sliding window procedure, and the majority of the causal associations were retained:

Line 207: “29.4% of causal loci (PIP>0.95) in the original analysis were still causal in sliding window analysis. 31.1% and 29.3% of causal loci whose 5’ and 3’ overlapping locus had PIP>0.95 in sliding window analysis, respectively, while themselves were no longer causal.”

In polygenic prediction, sliding window analysis provided a significantly improved performance compared with previous analysis on non-overlapping loci:

However, since in this revision we have several updates on the polygenic prediction procedure, it was difficult to quantify how much improvement was led by sliding window design. Thus, we directly showed the new result in figure 5 but did not compare it with the original result.

We also modified the previously imprecise statement to:

Line 490: “…it integrated information of the entire sequence, not only the few bp at the center.”

> *(4) Is the HFS-based association going to miss coding variation and several regulatory variants such as splicing variants? There are also going to be cases where there's an*

> *association driven by a variant that is correlated with a Sei prediction in a neighboring window. These would represent false positives for the method, it would be useful to identify or characterize these cases.*

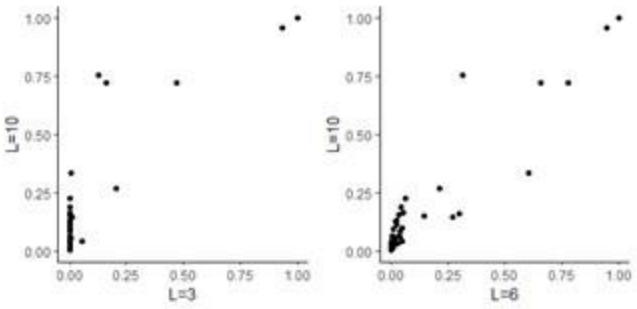
Response: As the reviewer suggested, sei captured only functional genomic features and is by nature prone not to perform well when the causal variants impact protein sequences. In this revision, we characterized this by focusing on causal exonic variants (SNP PIP>0.95):

Line 322: “On the other hand, HFS perform worse than SNP-based fine-mapping on exonic regions. Taking height as an example, PolyFun detected 125 causal SNPs (PIP>0.95) in the exonic regions, but only 16% (20) of loci that harbored them also reached PIP>0. 5 (11 reached PIP>0.95) in HFS+SUSIE analysis. Among the 105 loci that missed such signals (HFS PIP<0.5), 12 had a nearby locus (within 10kb) showing HFS PIP>0.95, which likely reflected false positive led by LD. Thus, SNP-based analysis should be prioritized over HFS in coding regions.”

> *Additional minor concerns:*
>
> *(1) It's not clear whether SuSie-based finemapping is appropriate at the locus level, when there is limited LD between neighboring HFS bins. How does the choice of the number of causal loci and the size of the segment being finemapped affect the results and is SuSie a good fit in this scenario?*

Response: Following the reviewer’s suggestion, we reran SUSIE under different predefined causal loci number (from 2 to 10), and found that the identified causal loci were consistent.

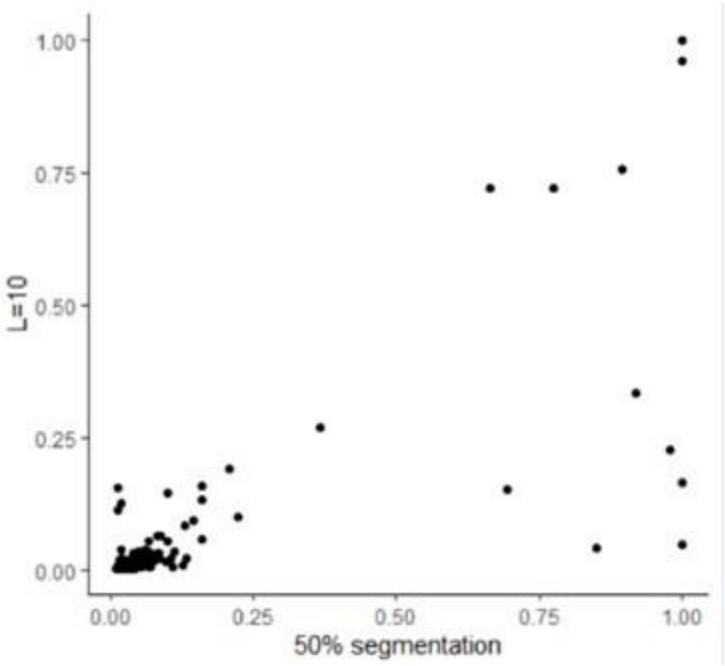
Author response image 6.



Line 211: “Besides, HFS+SUSIE was also robust when the predefined number of causal loci (L=2 to 10) was changed, and the number of detected loci were not changed.”

As for the size of segmentation, we divided the predefined segmentations (independent blocks detected by LDetect) into two half and reran SUSIE, and found that three additional causal loci emerged in one half. This suggested that using too small segmentation might increase the false positive rate. However, since there is no LD between independent blocks (which was guaranteed by LDetect), it is not necessary to use even longer blocks.

Author response image 7.

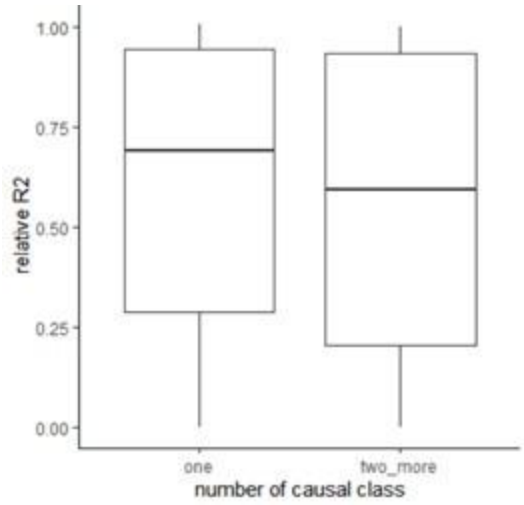


Line 133: “Simulation analysis revealed that when a non-reference sequence class score was associated the trait, reference class score could still capture median 70% of HFS-trait association R2.”

> *(2) It is not clear how a single score is chosen from the 117 values predicted by Sei for each locus. SuSie is run assuming a single causal signal per locus, an assumption which may not hold at ~4kb resolution (several classes could be associated with the trait of interest). It's not clear whether SuSie, run in this parameter setting, is a good choice for variable selection here.*

Response: As we discussed below (question 3), in this revision we no longer applied SUSIE to find one sequence class score for each locus due to the impact of overfitting, and use the reference sequence class uniformly for all loci. As reviewer suggested, we applied simulation to evaluate how this procedure influence HFS performance, especially when multiple sequence class of the same locus is causal to the phenotype. We found that reference sequence class score could capture median 69.1% of phenotypic R2 when the causal sequence class is not the reference, and captured median 59.2% of R2 when there was 2~5 non-reference causal class. We concluded that the loss led by skipping sequence class selection is mild, and it is necessary to do so in consideration of the risk of overfitting.

Author response image 8.



> *(3) A single HFS score is being chosen from amongst multiple tracks at each locus independently. Does this require additional multiple-hypothesis correction?*

Response: We agree with the reviewer that choosing the sequence class for each locus represented multiple testing, and with additional experiments we indeed observed some evidences of overfitting of this procedure. Thus, in this revision, we no longer applied the per-locus feature selection procedure, but instead used the sequence class corresponded to the reference (hg38) sequence. Consequently, additional multiple-testing correction is avoided with this procedure. We admitted that such simplification missed certain information, but as mentioned above, such lost is moderate, and is necessary to ensure statistical robustness and reduce false positive. In fact, with such simplification we better controlled the inflation factor of HFS GWAS and got better portability in polygenic prediction.

> *(4) The results show that a larger number of loci are identified with HFS-based finemapping & that causal loci are enriched for causal SNPs. However, it is not clear how the number of causal loci should relate to the number of SNPs. It would be really nice to see examples of cases where a previously unresolved association is resolved when using HFS-based GWAS + finemapping.*

Response: In this revision, we did not observe a clear relation between causal loci number and causal gene number. The only trend is that SNP-based fine-mapping seemed to perform better at coding regions, in accordance with the fact that HFS capture functional genomic signals. We also added new interpretations to highlight some examples where HFS resolve previously unresolved association signals. For example,

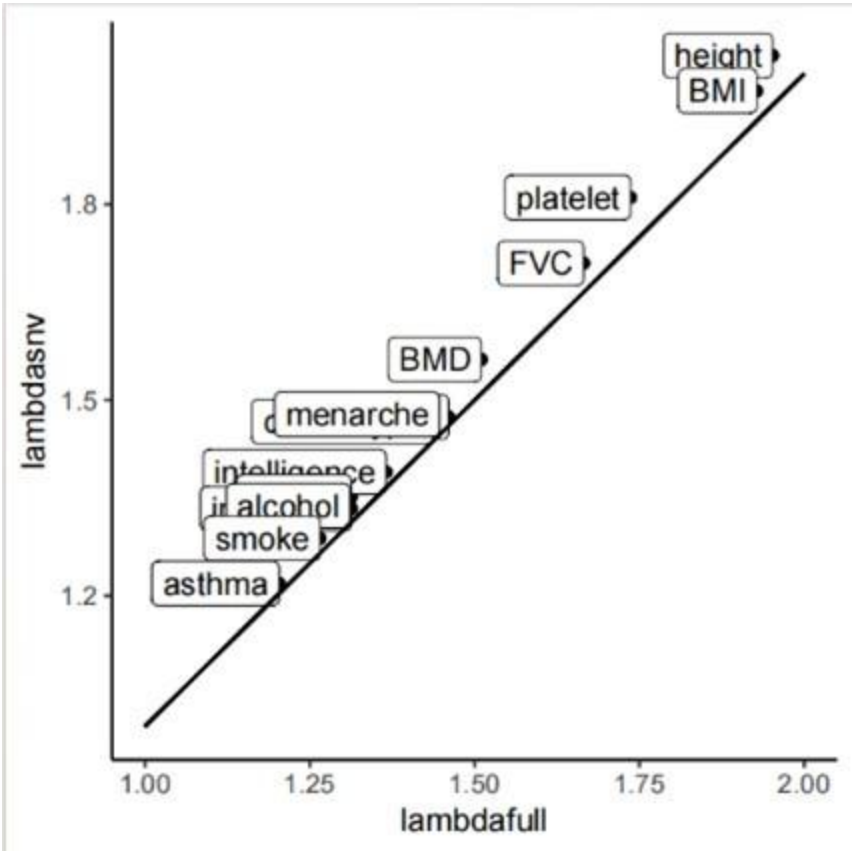
Line 287: “Specifically, in 1q32.1 region, HFS+SUSIE identified two loci with PIP>0.9 (Figure 4B). SNP-based association also found significant association in this region, but SNP fine-mapping(Weissbrod et al., 2020) could not resolve this signal and only found seven signals between PIP=0.1 to 0.5.”

> *(5) Sequence-based deep learning model predictions can be miscalibrated for insertions and deletions (INDELs) as compared to SNPs. Scaling INDEL predictions would likely improve the downstream modeling.*

Response: Following the reviewer's suggestion, we conducted a sensitivity analysis that removed all indel on chromosome 1 and repeated HFS analysis. Removing indel has indeed increased the number of significant (p<5e-8) association by 9%, but also slightly increased inflation factor (paired wilcox test p=0.0001). In fine mapping analysis, removing indel caused a 4.7% decrement in the number of detected causal association (PIP>0.95). We reasoned that the potential miscalibration on indel has indeed impacted the statistical power of HFS, but the proper approach to control this impact might not be direct and is still await optimizing. In this revision, we still kept all indels in the analysis, since we proposed that the power of fine-mapping is more important than the power of marginal association.

Line 213: "Lastly, removing insertion and deletion would reveal 9% more significant association (p<5×10-8) but 4.7% less causal association (PIP>0.95), and slightly increased inflation factor (Wilcoxon p=0.0001, Figure S4)."

Author response image 9.



> ***Reviewer #1 (Recommendations For The Authors):***
>
> *It was unclear to me why the sei output was rounded to two decimal places to "avoid influence of sei prediction noise". Wouldn't rounding introduce additional noise?*

Response: We thank the reviewer for pointing out our inadequate description. The rounding procedure is used to mask the low value that likely did not reflect any real change. The idea is that, even if a variant actually does not bring about any functional changes, sei would still output a very low HFS value that is not equal to, but close to, zero. By rounding procedure, such low values would be set to zero, which could avoid noise. We have added this rationale to the method section:

Line 529: “This is due to the fact that even if a variant actually makes no impact on functional genomics, sei would still output a value that are close to but not equal to reference sequence class score. Rounding procedure would set such HFS to zero and remove the random value from sei.”

> *Minor comments / typos:*
>
> - *There are many typos in the abstract.*

Response: We have revised the typo and grammar issues in the abstract in this revision.

> - *I believe "Arachnoid acid-intelligence" should be "Arachidonic acid-intelligence".*
> - *Consistently there is no space between text and parenthetical citations. For example, "sei(Chen et al., 2022)" should be "sei (Chen et al., 2022)".*
> - *Line 110: "at least one non-reference haplotypes" --> "at least one non-reference haplotype".*
> - *Line 155: "data-based method" --> "data-based methods".*
> - *Lines 165-166: "functionally importance" --> "functional importance".*

Response: We have made these revisions accordingly.

> - *Line 210: the sentence containing "this annotation on conditioned of a set of baseline annotations" is unclear.*

Response: We have revised this sentence as “…regressed the PIP against this annotation, with a set of baseline annotations included as covariates, similar to the LDSC framework.”

> - *Line 213: "association" --> "associations".*
> - *Line 219: "association" --> "associations".*
> - *Line 251: "result" --> "results".*
> - *Line 269: "result" --> "results".*
> - *Line 289: "known to involved" --> "known to be involved".*
> - *Line 356: "LDAK along" --> "LDAK alone".*
> - *Line 362: "BOLT-LMM along" --> "BOLT-LMM alone".*
> - *Supplement: "Hihglighted" --> "Highlighted".*

Response: We have made these revisions accordingly.

- *Line 444: Were "British ancestry Caucasians" defined as individuals that self-identified as "white British"? If so, then they should be described as "self-identified "white British"".*

Response: As the reviewer pointed out, we have changed the description as self-identified British ancestry Caucasians.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *(1) A 2022 cistrome-wide association study (CWAS) computed associations between genetically-predicted chromatin activity and phenotypes. Adding a reference to this paper would be helpful. https://pubmed.ncbi.nlm.nih.gov/36071171/*

Response: Following the reviewer's suggestion, we discussed the similarity between CWAS and our study:

Line 89: "In line with this notion, a recent similar strategy called cistrome-wide association study (CWAS) integrated variant-chromatin activity and variant-phenotype association to boost power of genetic study of cancer. (Baca et al., 2022)."

> *(2) Line 487 states: "We applied sei to predict 21,906 functional genomic tracks for each sequence, without normalizing for histone mark." It's not clear what normalization is being referred to here.*

Response: We have revised the sentence to:

Line 495: "We applied sei to predict 21,906 functional genomic tracks for each sequence, without normalizing for histone mark (divided each track score by the sum of histone mark score) as suggested by the sei author."

> *(3) The figures are extremely low resolution, they need to be updated.*

Response: In this revision, we uploaded separate pdf file for each figure to provide high resolution graphs.

> *(4). The results section was difficult to follow and would benefit from being written more clearly.*

Response: In this revision, we re-arranged some of the result section to better clarify the main idea. We moved all statistical results to the bracket and focused our main text on the interpretation. For example,

Line 123: "Further evaluation indicated that this low LD was led by two factors: integration of rare variant impacts and segmentation. Firstly, excluding rare variants from HFS caused the LD raised to median=0.14 (Method; Figure S2C). Secondly, median LD of SNPs from adjacent loci was 0.06, which was significantly higher than HFS LD (paired Wilcoxon p=1.76×10-5) but significantly lower than HFS LD without rare variants (paired Wilcoxon p<2.2×10-16)."

> *(5) "Along" is used several times in the final results section (PRS estimation), this should be "alone".*

Response: We have modified all misused "along" by "alone" in this revision.

> *(6) Instead of using notation identifying genomic location, it might be clearer to provide gene names when illustrating examples of trait-associated promoters.*

Response: In this revision, we added gene name of the corresponding promoters to the main text to better clarify the findings.