PNAS

Supplemental Information for

Quality assessment and refinement of chromatin accessibility data using a sequence-based predictive model

Seong Kyu Han, Yoshiharu Muto, Parker C. Wilson, Benjamin D. Humphreys, Matthew G. Sampson, Aravinda Chakravarti,* Dongwon Lee*

*Corresponding authors.
Email: dongwon.lee@childrens.harvard.edu, aravinda.chakravarti@nyulangone.org

This PDF file includes:

Supplementary Figs. S1 to S12
Legends for Datasets S1 to S6

Other Supplementary Materials for this manuscript include the following:

Datasets S1 to S6

Supplemental Figures and legends

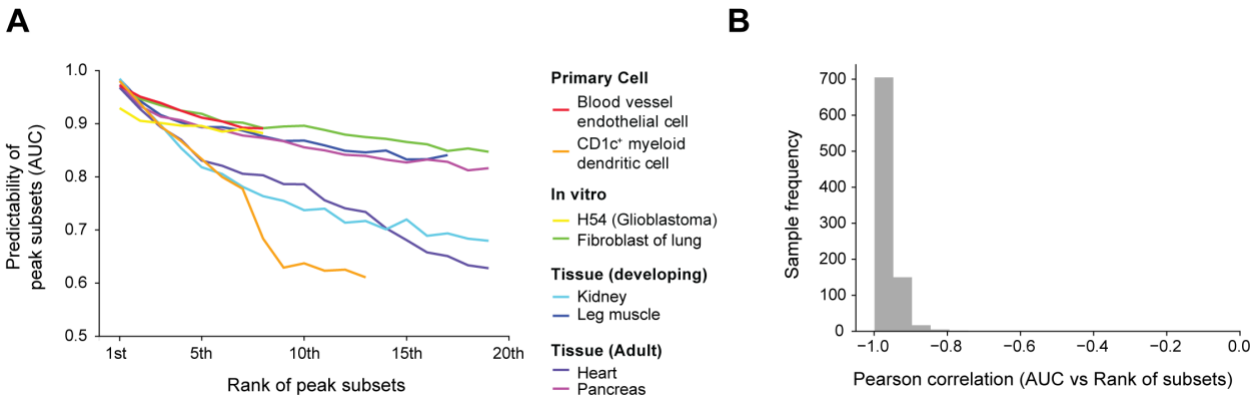


Fig. S1. Correlation between predictability and signal strength of peak subsets. (A) A plot of predictability (AUC, Y-axis) and the rank of signal strengths of peak subsets (X-axis). Each line represents a distinct sample. A total of 8 representative examples, including primary/in vitro differentiated cells and developing/adult tissues, are presented. **(B)** The distribution of Pearson correlations between peak predictability and signal strength across ENCODE datasets; correlations between AUC and the rank of signal strengths were computed per sample.

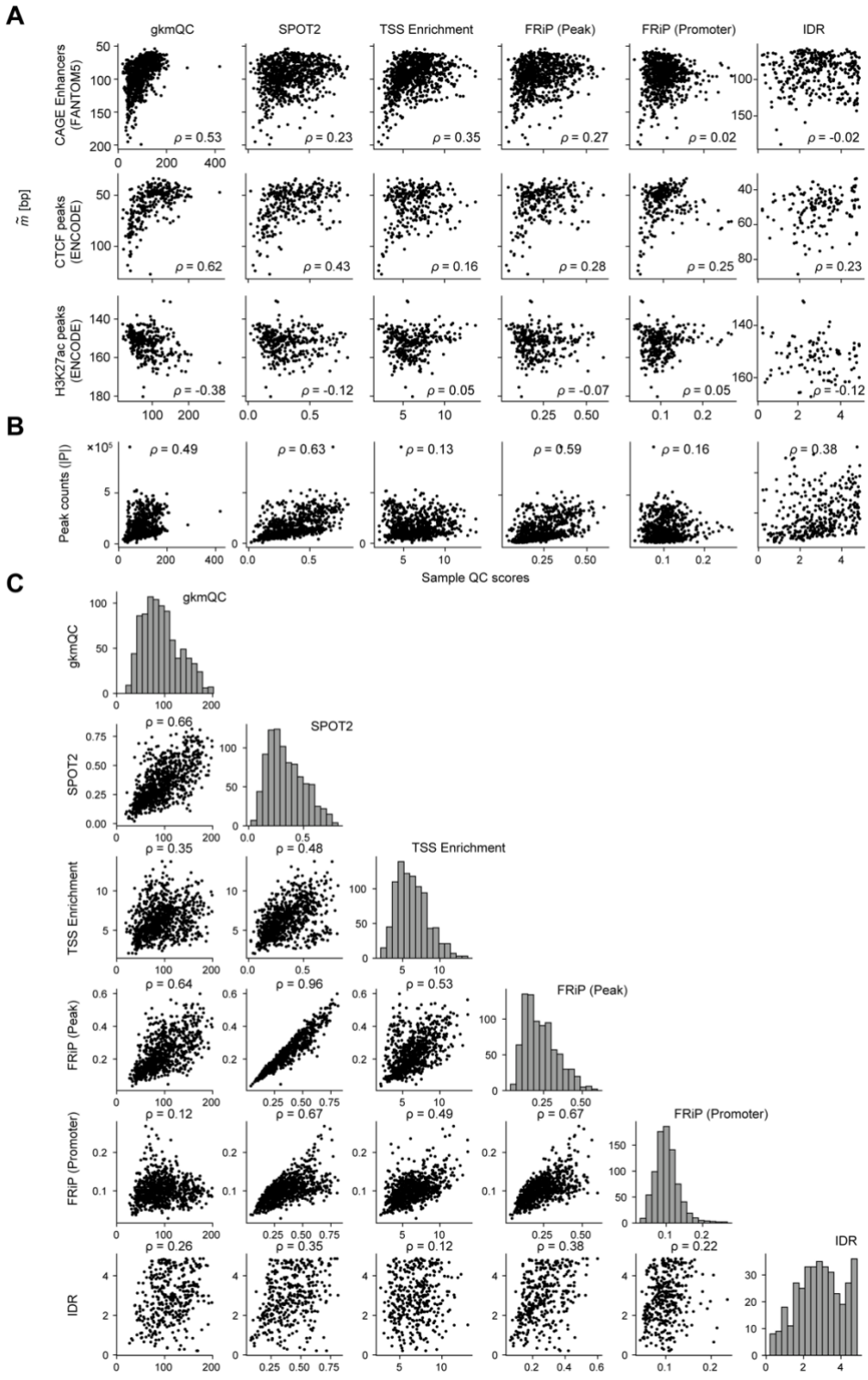


Fig. S2. gkmQC outperforms other QC methods in identifying high-quality samples. (A) For each of the six different QC metrics, correlation plots compare the quality scores to the precision of peak locations using CAGE Enhancers from FANTOM5 (top), CTCF peaks (middle), and H3K27ac peaks (bottom), across the ENCODE samples. (B) Similar to (A), the quality scores are compared with the peak counts. (C) The six different quality metrics are compared against each other in the pairwise correlation plots. The histograms on the diagonal show the distribution of the corresponding quality scores.

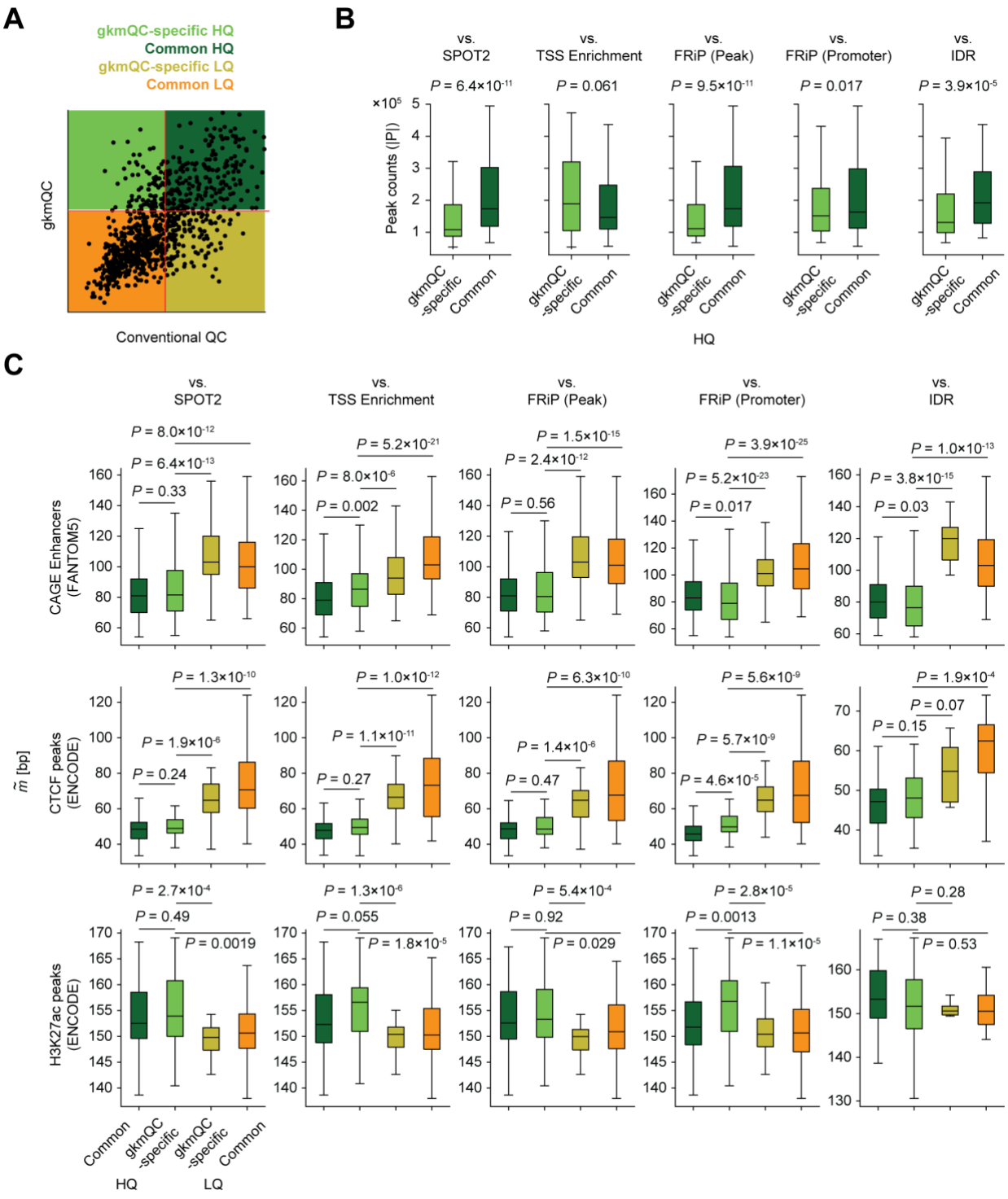


Fig. S3. Analyses of high-quality samples exclusively determined by gkmQC. (A) samples are classified into four different groups based on gkmQC scores and another quality metric (SPOT2 scores are shown as an example). HQ and LQ stand for high and low quality, respectively. We use the 50th percentile as a cut-off for the classification. (B) Peak counts are compared between gkmQC-specific and common HQ groups for each combination of the QC metrics. (C) The precision of peak locations ($\tilde{m}$) is compared across the four sample groups. *P*-values are calculated between the gkmQC HQ group and each of the others with the Mann-Whitney *U* test.

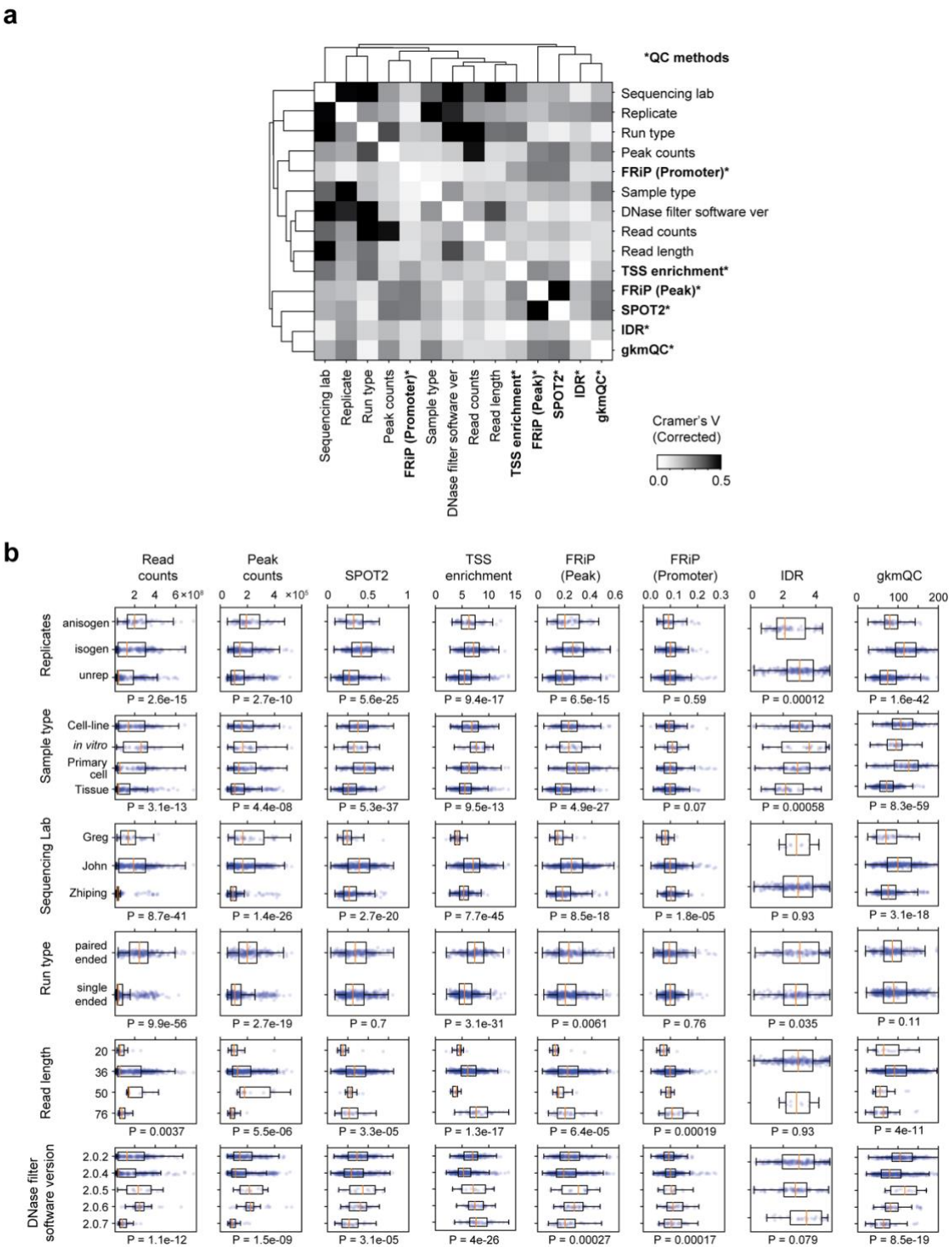
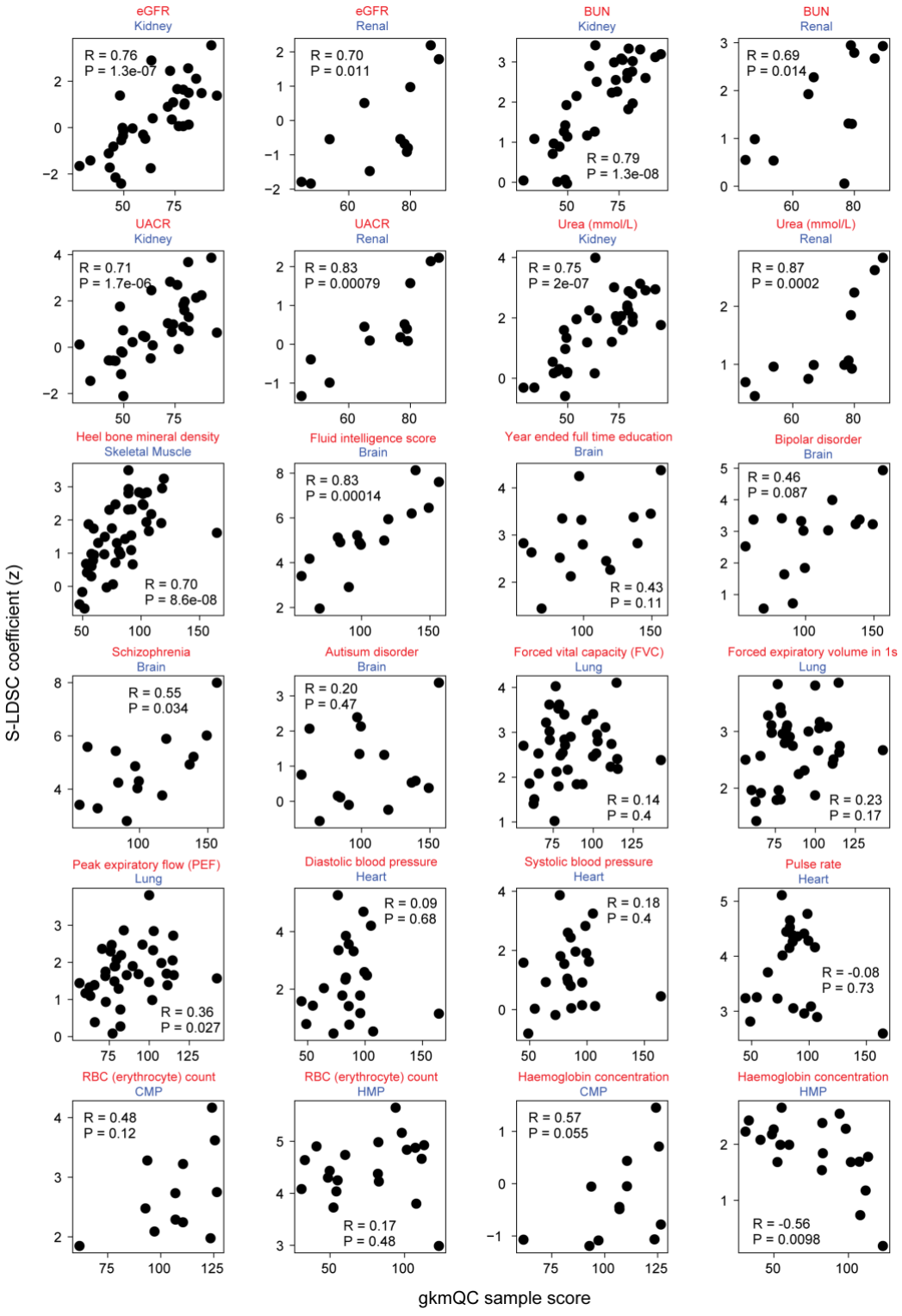


Fig. S4. **Analysis of biological and technical factors affecting quality metrics (A)** Heatmap shows covariation between several technical factors and quality metrics. Cramer's V was used to quantify correlations of continuous and discrete variables. Technical factors were not clustered with quality metrics (bold black). **(B)** Boxplots show differences in quality metrics (x-axis) for several different technical factors. *P*-values were calculated with one-way ANOVA of the quality metric scores with respect to the technical factors.


eGFR
Kidney
R = 0.76
P = 1.3e-07
eGFR
Renal
R = 0.70
P = 0.011
BUN
Kidney
R = 0.79
P = 1.3e-08
BUN
Renal
R = 0.69
P = 0.014
UACR
Kidney
R = 0.71
P = 1.7e-06
UACR
Renal
R = 0.83
P = 0.00079
Urea (mmol/L)
Kidney
R = 0.75
P = 2e-07
Urea (mmol/L)
Renal
R = 0.87
P = 0.0002
Heel bone mineral density
Skeletal Muscle
R = 0.70
P = 8.6e-08
Fluid intelligence score
Brain
R = 0.83
P = 0.00014
Year ended full time education
Brain
R = 0.43
P = 0.11
Bipolar disorder
Brain
R = 0.46
P = 0.087
Schizophrenia
Brain
R = 0.55
P = 0.034
Autisum disorder
Brain
R = 0.20
P = 0.47
Forced vital capacity (FVC)
Lung
R = 0.14
P = 0.4
Forced expiratory volume in 1s
Lung
R = 0.23
P = 0.17
Peak expiratory flow (PEF)
Lung
R = 0.36
P = 0.027
Diastolic blood pressure
Heart
R = 0.09
P = 0.68
Systolic blood pressure
Heart
R = 0.18
P = 0.4
Pulse rate
Heart
R = -0.08
P = 0.73
RBC (erythrocyte) count
CMP
R = 0.48
P = 0.12
RBC (erythrocyte) count
HMP
R = 0.17
P = 0.48
Haemoglobin concentration
CMP
R = 0.57
P = 0.055
Haemoglobin concentration
HMP
R = -0.56
P = 0.0098
S-LDSC coefficient (z)
gkmQC sample score

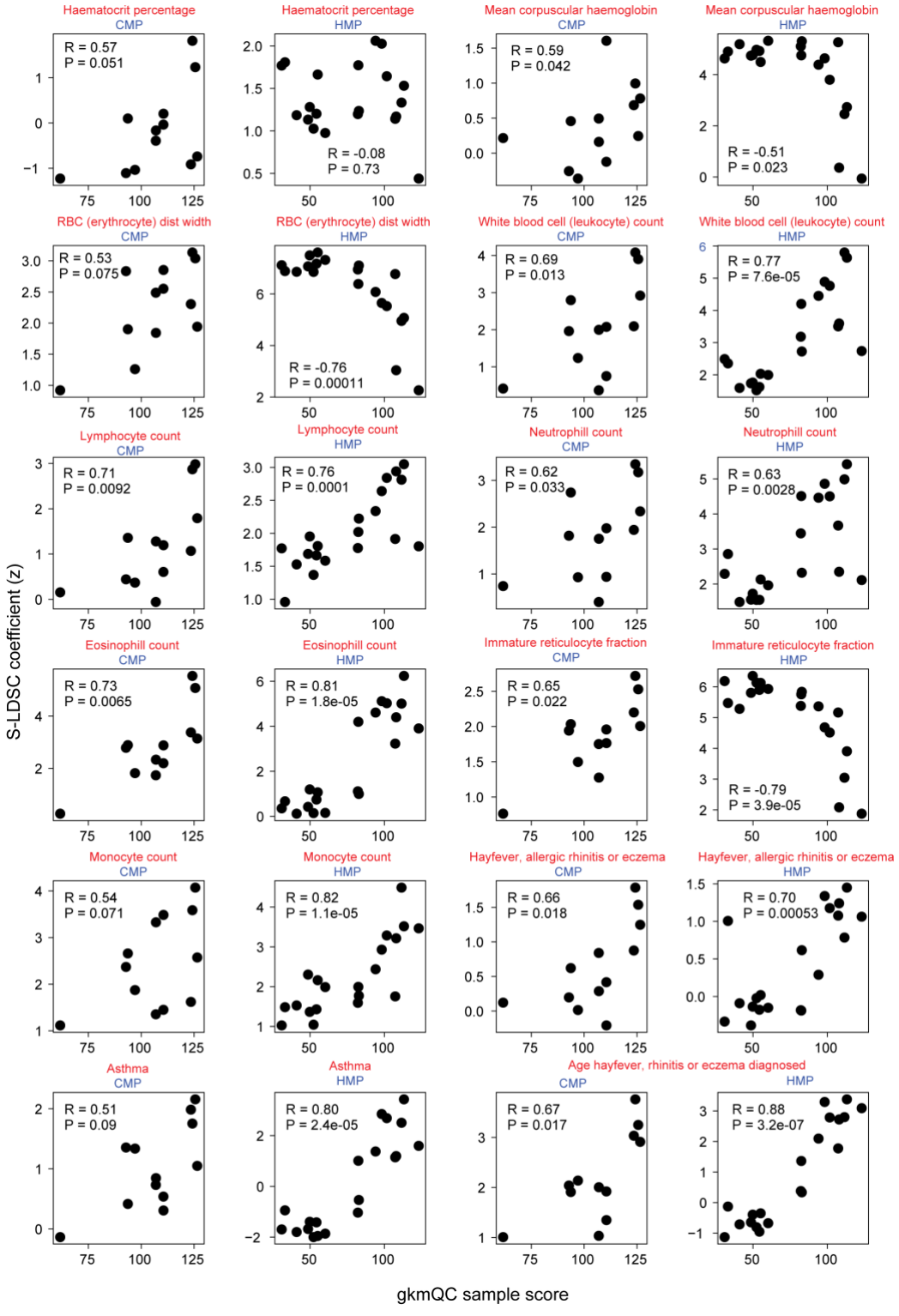


Fig. S5. Correlation analyses of gkmQC sample scores and S-LDSC coefficients for 48 pairs of relevant tissues and phenotypes (2 pages). Dots in the scatterplots represent samples of chromatin accessibility data. The X-axis is the gkmQC sample scores, and the Y-axis is the S-LDSC coefficient. Correlations are Pearson's correlation coefficients. The title of each scatterplot shows a tissue- or cell-type for chromatin-accessibility data (Blue) and a relevant GWAS trait (Red).

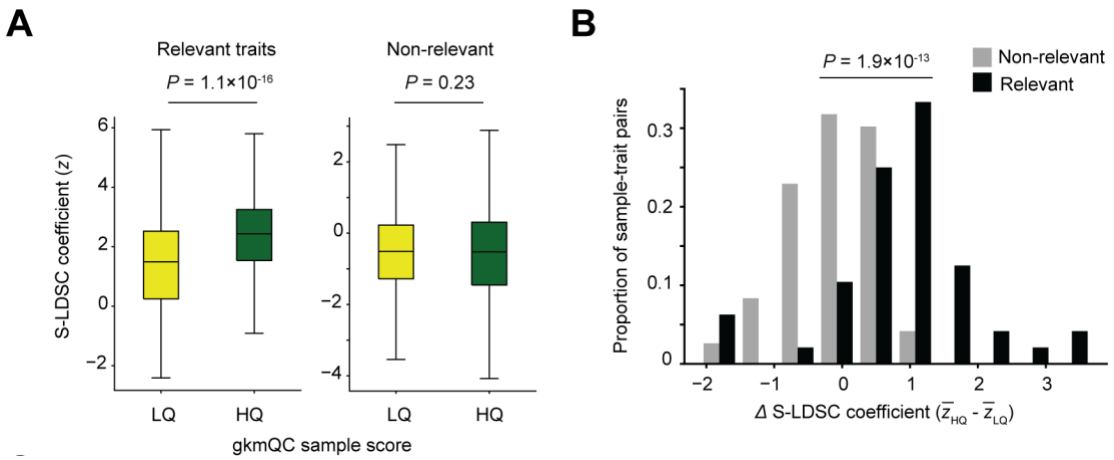


Fig. S6. **Contribution of tissue-specific peaks in high-quality samples to relevant traits. (A)** Boxplots show distributions of S-LDSC coefficients of high- and low-quality samples paired with relevant (left) and non-relevant traits (right). Mann-Whitney *U* test is used to test the significance of the differences. **(B)** Histograms show differences in mean S-LDSC coefficients between high- and low-quality samples for relevant and non-relevant traits. Paired *t*-test is used to test the significance of the differences.

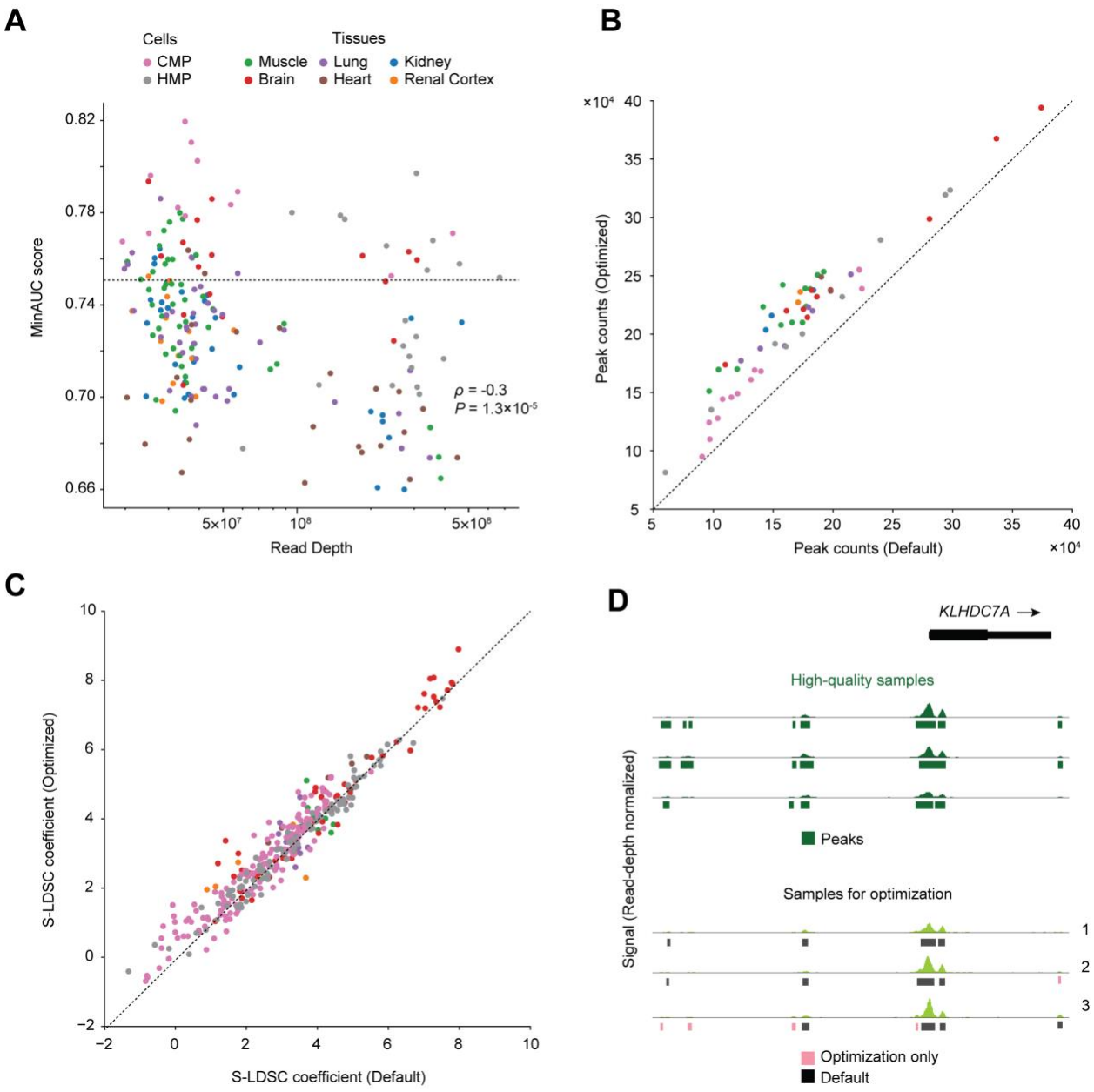


Fig. S7. Optimizing peak-calling from bulk chromatin accessibility data. (A) Scatterplot comparing read depth (total read counts) with MinAUC using Spearman rank correlation coefficients. Each dot is a distinct sample with colors representing tissues and cell types. **(B)** Peak counts before and after gkmQC optimization. **(C)** S-LDSC coefficients from the heritability analysis of relevant GWAS traits before and after optimization. **(D)** Newly found peaks after optimization recapitulate peaks observed in high-quality samples at the *KLHDC7A* locus (**Fig. 2B**).

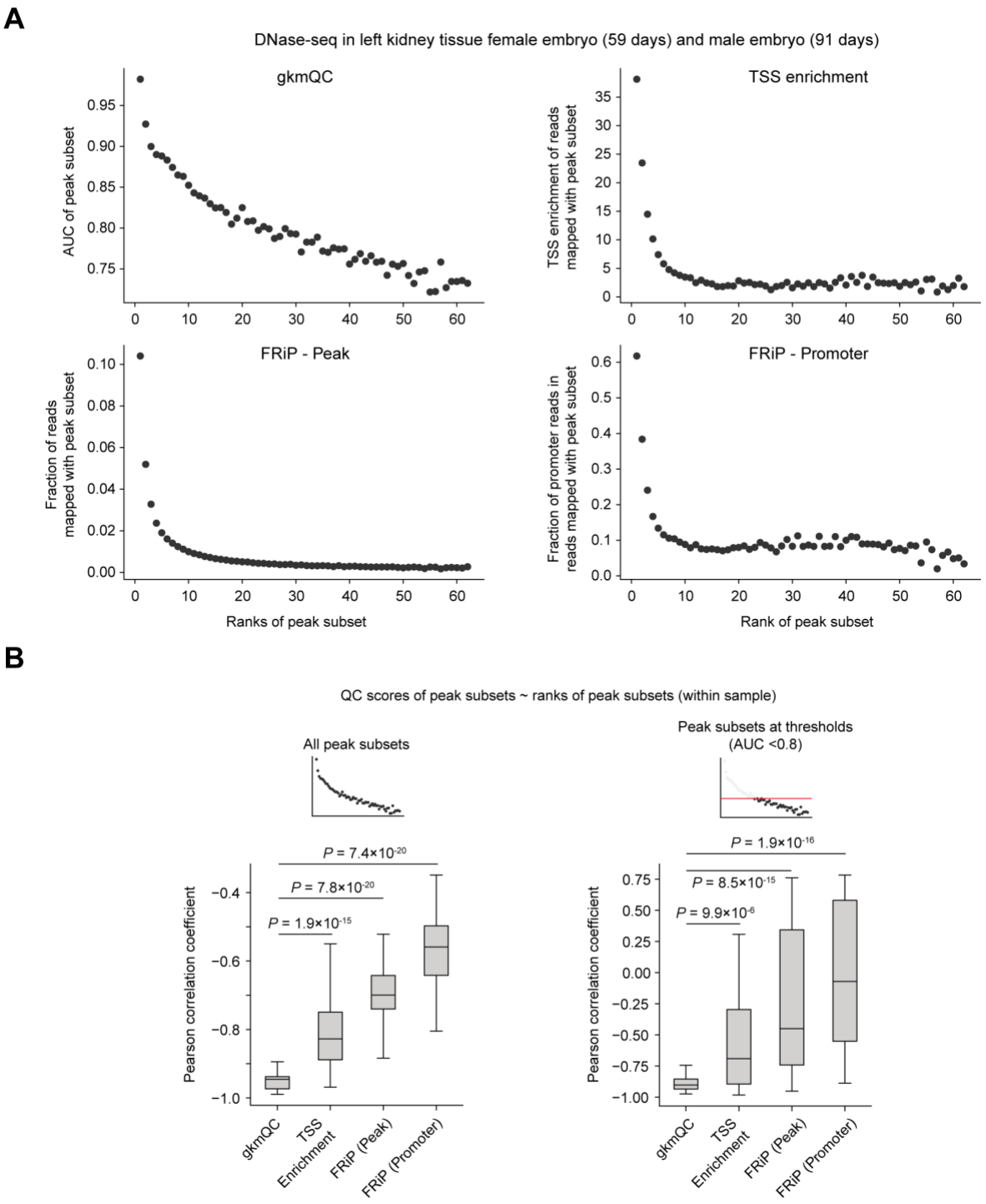
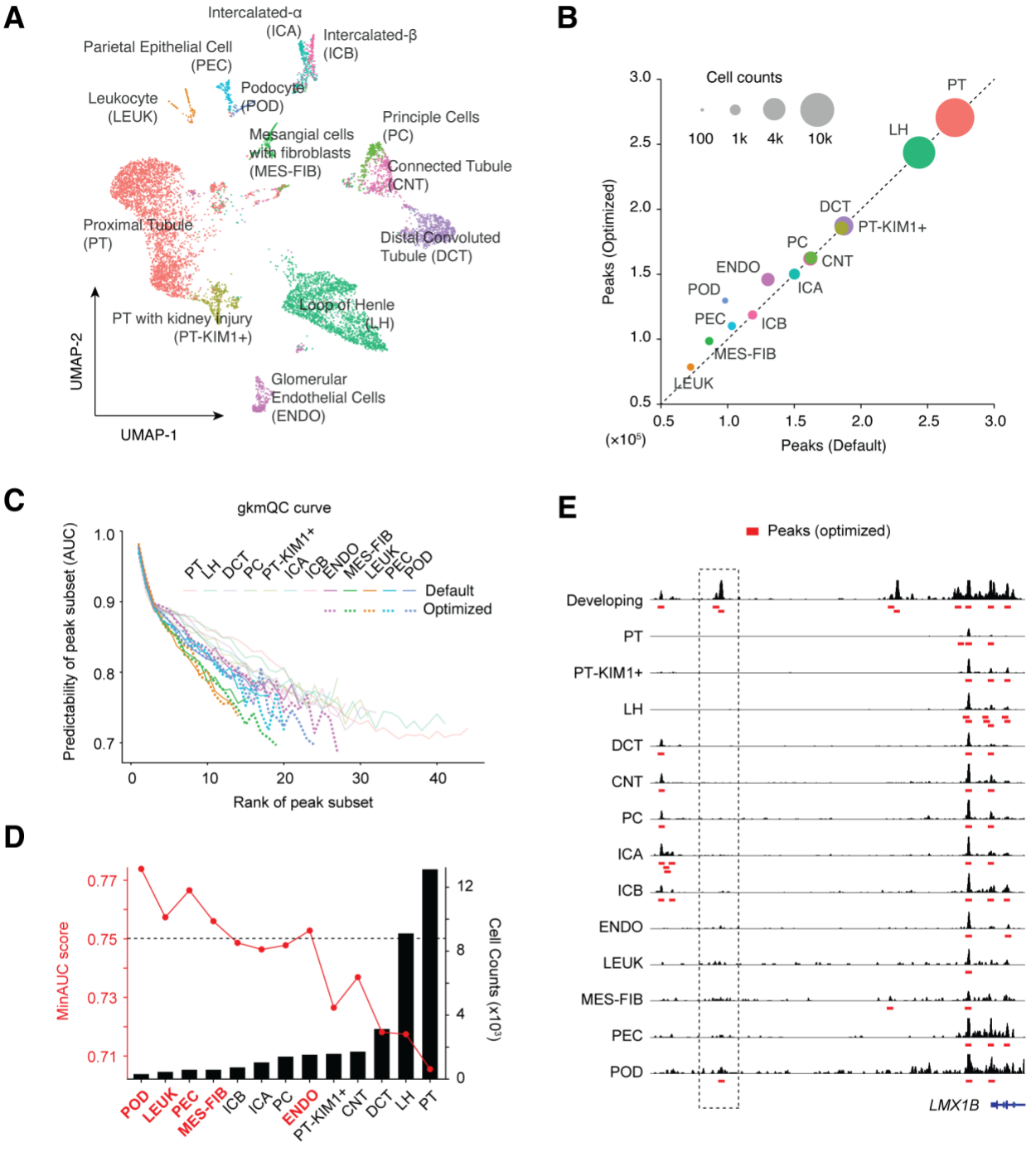


Fig. S8. gkmQC scores are most strongly correlated with ranks of peak subsets compared to other quality metrics. (A) For each of the four QC methods (gkmQC, TSS enrichment, and FRiP – Peak/Promoter), the quality scores for peak subsets are compared to their ranks, using DNase-seq from embryonic kidney tissue as an example. While AUCs from gkmQC significantly correlate with peak subset ranks, other methods do not, especially when ranks are greater than 10. **(B)** The distributions of the correlations between quality scores and ranks of peak subsets for all ENCODE samples are compared across the four QC methods. P-values are calculated between the gkmQC and each of the other methods with the Mann-Whitney U test.



		UACR					eGFR					SCZ				
		POD	PEC	ENDO	MES-FIB	LEUK	POD	PEC	ENDO	MES-FIB	LEUK	POD	PEC	ENDO	MES-FIB	LEUK
Optimized	S-LDSC z	3.357	0.781	3.200	3.019	1.062	0.797	1.012	2.026	2.735	1.377	0.010	0.169	-2.287	-0.822	-1.745
	Pr[h_g^2]	9.3%	2.0%	7.4%	6.6%	2.7%	4.1%	2.2%	5.3%	8.5%	5.4%	2.3%	0.9%	-1.4%	0.4%	-0.8%
	S-LDSC z	2.306	2.726	2.922	1.074	-1.305	4.411	3.534	3.823	2.673	2.126	-1.263	0.092	-0.053	-2.391	-0.935
	Pr[h_g^2]	25.4%	27.4%	29.1%	19.6%	10.5%	47.1%	45.7%	51.1%	28.3%	26.0%	4.5%	7.5%	9.3%	1.9%	2.8%
Default	S-LDSC z	0.873	-1.359	0.042	0.447	0.635	1.932	1.164	-1.004	0.570	0.864	0.712	-1.192	0.632	-0.769	0.883
	Pr[h_g^2]	0.6%	-0.5%	0.1%	0.6%	0.8%	1.1%	1.9%	-0.3%	1.0%	1.3%	0.2%	-0.3%	0.3%	-0.1%	0.5%

Fig. S9. Peak-calling optimization of kidney snATAC-seq data identifies more functional peaks for rare cell types. (A) UMAP plot of kidney snATAC-seq data. Color is based on the annotation of known kidney cell types. **(B)** Comparison of peak counts before and after optimization where each dot represents kidney cell type in (**A**). Dot sizes represent cell counts of the corresponding cell types. **(C)** gkmQC curves for peaks from pseudo-bulk reads of kidney cells. Dashed lines are gkmQC curves for optimized peak-calling. The five cell types with MinAUC >0.75 were optimized. **(D)** MinAUC scores of kidney cell types (red line with dots; left Y-axis) are anti-correlated with cell counts (black bars; right Y-axis), demonstrating more significant optimization in rarer cell types; cell types with MinAUC >0.75 are highlighted (red). **(E)** A representative locus upstream of *LMX1B* containing a podocyte-specific peak. All kidney cell types in the snATAC-seq and developing (DNase-seq) are shown. **(F)** Heritability is compared between optimized and default peaks for five rare cell types with MinAUC >0.75. Similar to Fig. 5D, the table presents heritability for three disjoint peak subsets; optimization-only (top), commonly found before/after optimization (middle), and only with default values (bottom).

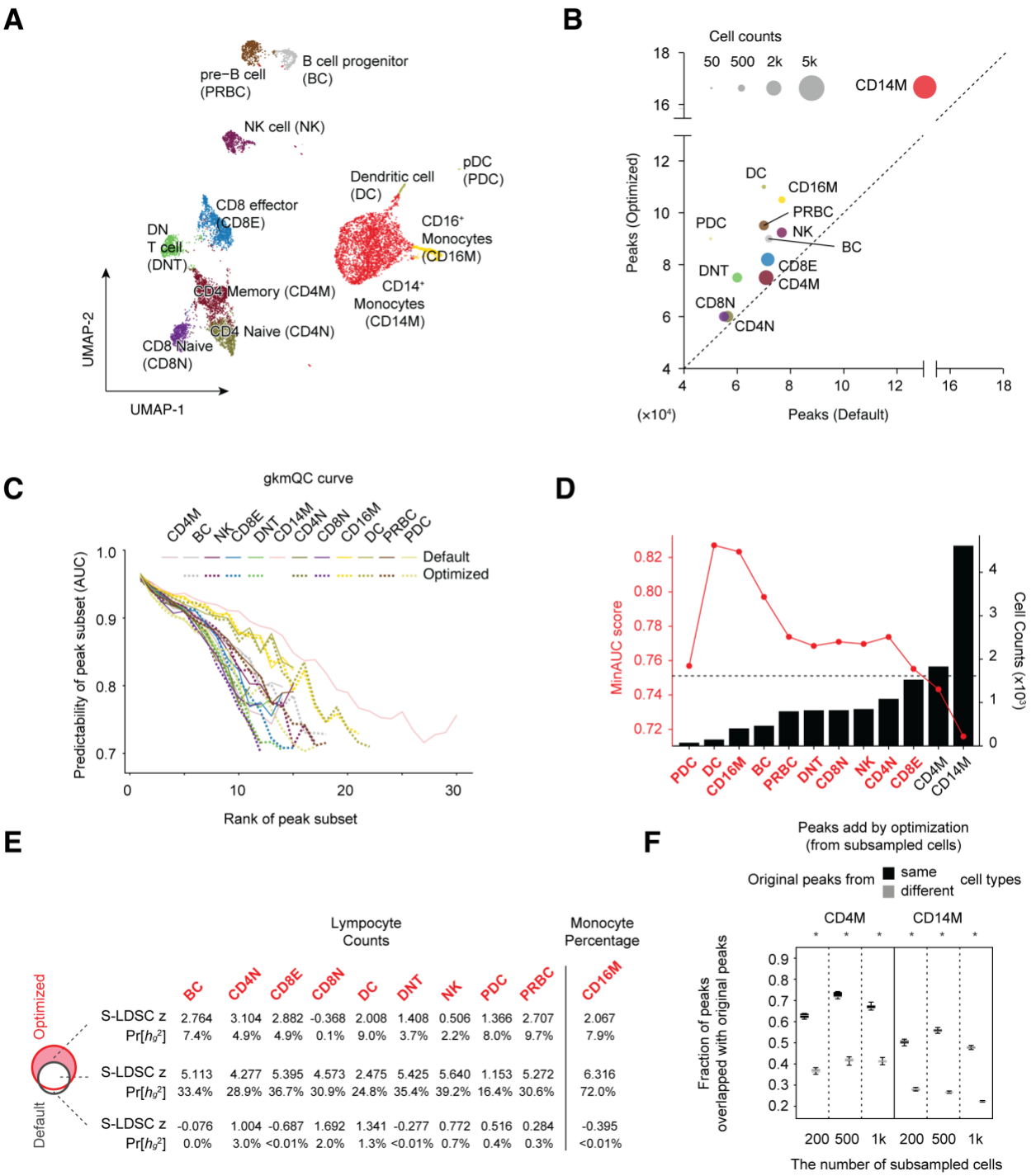


Fig. S10. Peak-calling optimization of PBMC snATAC-seq data. Figs. A-E are analogous to Figs. S7A-D and F, except that ten cell types were optimized for PBMC snATAC-seq data. Fig. F is analogous to Fig. 6D. Here, two major cell types (CD14+ Monocyte and CD4+ Memory T cells) were analyzed.

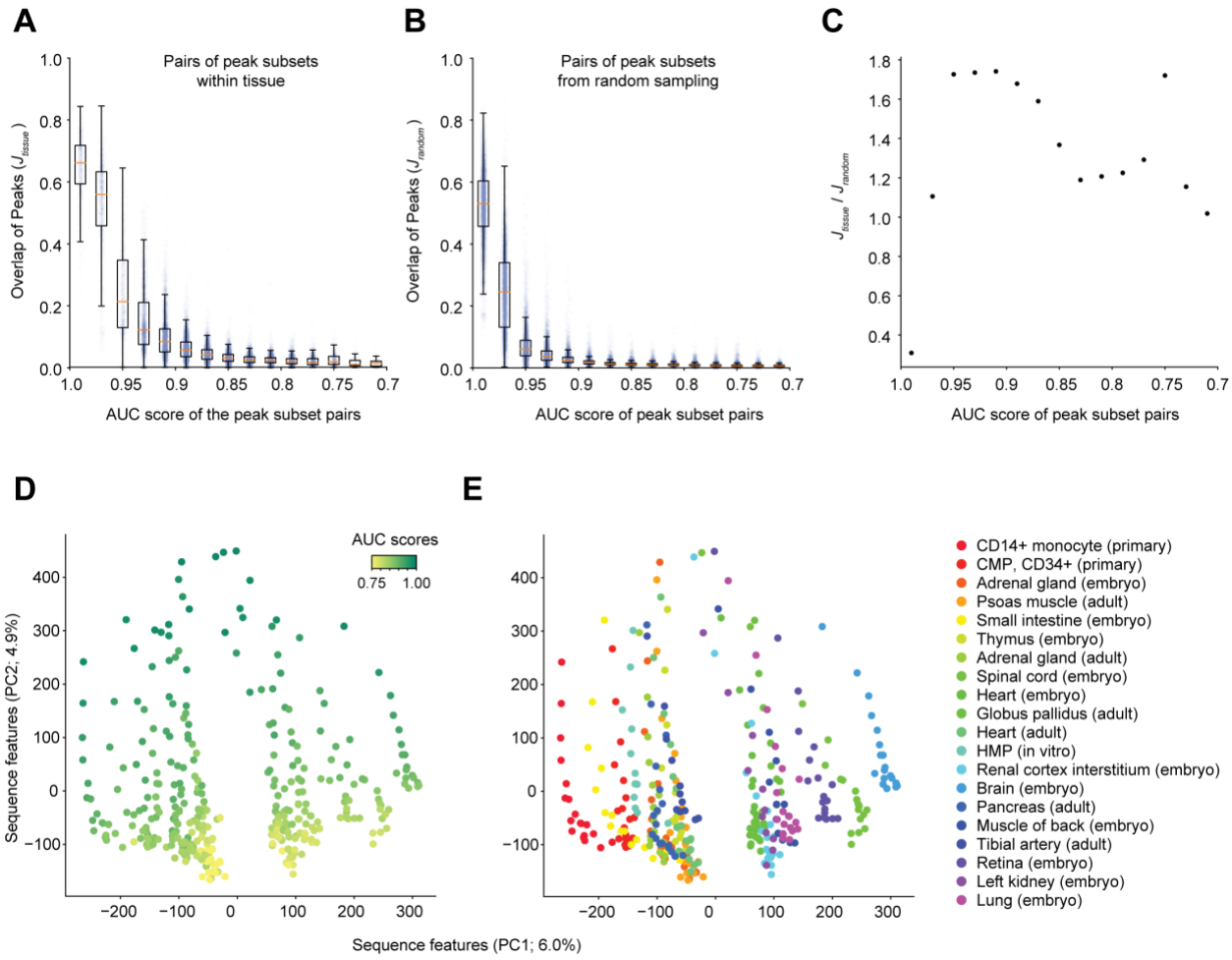


Fig. S11. Differences in peak-calling signals between peak subsets reflect differences in tissue-specificity. Figs. A-C depict relative degrees of tissue-specificity of peak subsets as a function of their peak predictability (AUC). We calculated overlaps of peak-subset pairs in a similar AUC range **(A)** from the same tissue and **(B)** from different tissues via random sampling. Jaccard index coefficients were used to quantify overlap. **(C)** The overlap ratios between **(A)** and **(B)** (J_{tissue} / J_{random}) are calculated for each of the AUC ranges. **(D** and **E)** Principal component 1 (PC1) and PC2 from PCA analysis of the trained sequence features are shown for several different tissues. Peak subsets are represented as dots, color-coded for **(D)** AUCs and **(E)** tissues. Peak subsets in a medium range of predictability scores (0.8< AUC <0.95) have more tissue-specific sequence features.

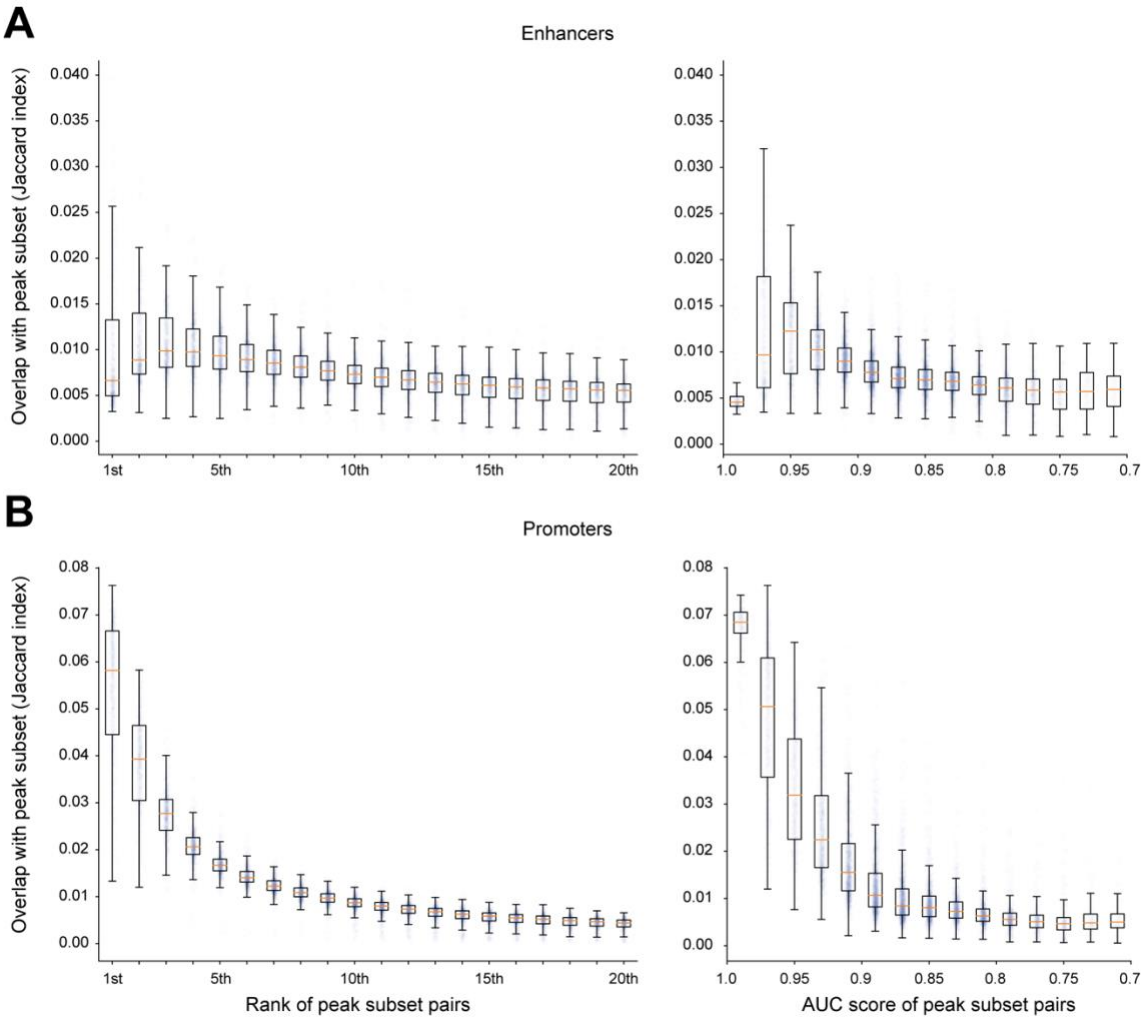


Fig. S12. Enrichment analysis of enhancers and promoters with respect to peak signal and predictability. (A) FANTOM5 (distal) enhancers and **(B)** promoters are compared to peak subsets according to their ranks and AUCs. Overlaps between two peak sets are calculated by the Jaccard index. Boxplots represent overlap distributions across different samples.

Legends of Datasets

Dataset S1 High-quality samples classified by gkmQC and other metrics.

Dataset S2 Metadata and quality metric statistics of 886 ENCODE DNase-seq samples.

Dataset S3 Results from partitioned heritability analysis using 200 ENCODE DNase-seq samples and relevant GWAS traits.

Dataset S4 Archived files including the BED files of optimized peaks for 58 DNase-seq data. (https://osf.io/download/9pez8/)

Dataset S5 Archived files including the BED files of optimized peaks for kidney and PBMC snATAC-seq data. (https://osf.io/download/yx9p8/)

Dataset S6 Computing speed of gkmQC with respect to the number of peaks.