Supporting Information

Kim and Kim 10.1073/pnas.1318383110

SI Text

For classifying individuals into one of nine phenotypes denoted by $\boldsymbol{T} = \{B, C, H, K, O, L, R, U, N\}$, where each phenotype is labeled by the first letter of the full initials of each trait (class) (Table 2). We use Bayesian inference of prediction results of four methods that are composed of two classification algorithms of the support vector machine (SVM) and k-nearest neighbor (kNN) analysis applied to two different descriptors of SNP and SNP syntax (SNP-S). The method names are abbreviated to *kNN/SNP-S*, *kNN/SNP*, *SVM/SNP-S*, and *SVM/SNP*, and they are mathematically denoted by $\boldsymbol{m^1}$, $\boldsymbol{m^2}$, $\boldsymbol{m^3}$, and $\boldsymbol{m^4}$, respectively. Each method requires training of its own parameters, which attempts to identify the best-performing parameter compositions. Once all methods are optimally fitted on the dataset, for each test individual $\boldsymbol{i}$, we select a trait of having highest posterior probability conditioned on prediction results from trained methods, which can be formulated as $\boldsymbol{P(s_i|M_i^1, M_i^2, M_i^3, M_i^4)}$, where $\boldsymbol{s_i}$ denotes the predicted trait of individual $\boldsymbol{i}$, and $\boldsymbol{M_i^j}$ denotes the trait of individual $\boldsymbol{i}$ predicted by method $\boldsymbol{m^j}$. By Bayes theorem, thus, we write

$$\begin{aligned} t_{max} &= argmax_{t\in T} P(s_i = t|M_i^1, M_i^2, M_i^3, M_i^4) \\ &= argmax_{t\in T} \frac{P(M_i^1, M_i^2, M_i^3, M_i^4|s_i = t) \times P(s_i = t)}{P(M_i^1, M_i^2, M_i^3, M_i^4)}, \end{aligned}$$

where the denominator $\boldsymbol{P(M_i^1, M_i^2, M_i^3, M_i^4)}$ is a normalizing constant. Because the predictive decisions of each method are inherently independent from each other and applying the chain rule (31)

$$= argmax_{t\in T} \prod_{j=1}^{4} P(M_i^j|s_i = t) \times P(s_i = t),$$

where $\boldsymbol{P(M_i^j|s_i = t)}$ and $\boldsymbol{P(s_i = t)}$ can be empirically inferred from the observations during the training phase of each of four methods by maximum likelihood estimation. For example, $\boldsymbol{P(M_i^1 = C|s_i = B)}$ can be estimated by identifying a fraction of true breast invasive carcinoma (BRCA) individuals who were predicted to belong to colon adenocarcinoma (COAD) class by kNN/SNP-S method among entire BRCA samples in the training set. For $\boldsymbol{P(s_i = t)}$, it corresponds to a fraction of samples of trait $\boldsymbol{t}$ of all training individuals, which is identical for each of nine traits, because the same sample size was used for each trait.

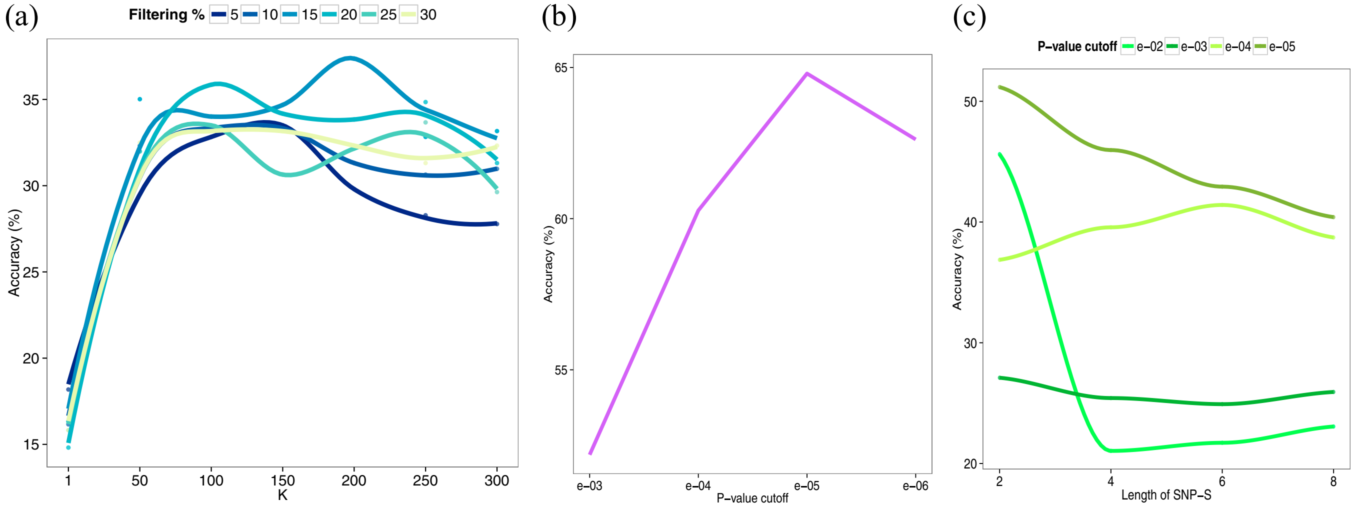

Fig. S1. (*A*) Optimization of parameters for the process of applying the kNN algorithm to the profiles of SNPs. The kNN/SNP method has two parameters: (*i*) filtering percentage for selecting rare features below specified frequency threshold, (e.g., for 1% filtering, the features below 1% frequency are selected for analysis) and (*ii*) k for selecting number of nearest neighbors of a test individual. (*B*) Optimization of parameters for the process of applying SVM algorithm to the profiles of SNPs. The SVM/SNP method has one parameter of the P value threshold for selecting the features with their P values below a specified value. (*C*) Optimization of parameters for the process of applying SVM algorithm to the profiles of SNP-Ss. The SVM/SNP-S method has two parameters: (*i*) P value threshold for filtering out features whose P values are greater than a specified value and (*ii*) the length of SNP-S.

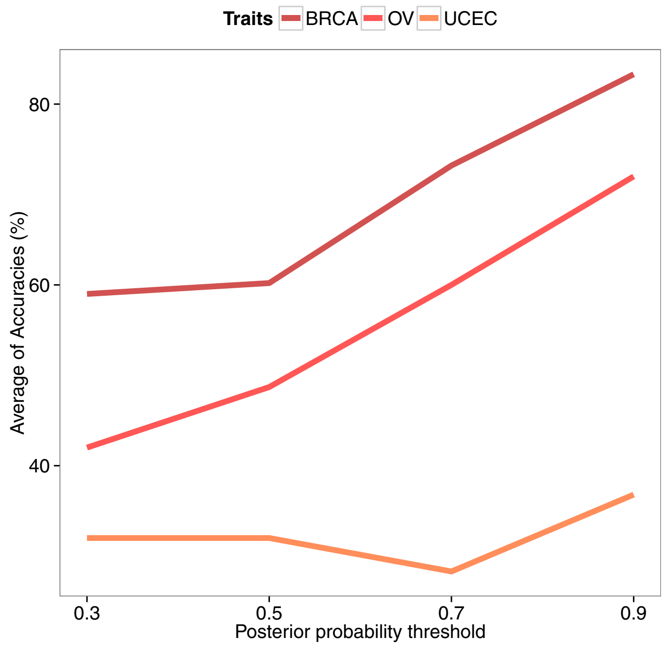

Fig. S2. Confidence of individual prediction. In our framework, confidence for the prediction result for an individual is equivalent to the highest posterior probability of Bayes inference for the individual. To investigate how the degree of the individual confidence relates to the predictive capacity, we plotted the overall accuracy of all test individuals as a function of the posterior probability threshold. For example, if the threshold is 0.7, then we calculate the average accuracy for those who have the posterior probability over 0.7 only.

Table S1. Codes for SNP genotype

Genotype	AA	CC	GG	TT	AC/CA	AG/GA	AT/TA	CG/GC	CT/TC	GT/TG
Genotype code	A	B	C	D	E	F	G	H	I	J

Table S2. Sample quality control of Affymetrix 6.0 SNP genotype data

		Before quality control		After quality control	
Study	Trait/cohort	Male/female	Total	Male/female	Total
TCGA	BRCA	6/694	700	0/511	511
	COAD	179/159	338	101/86	187
	HNSC	106/38	144	95/34	129
	KIRC	47/30	77	43/25	68
	LGG	38/36	74	34/32	66
	OV	0/427	427	0/379	379
	READ	72/59	131	54/41	95
	UCEC	0/301	301	0/237	237
HapMap	CEU	80/85	165	31/38	69
Total		528/1,829	2,357	358/1,383	1,741

BRCA, breast invasive carcinoma; CEU, Caucasians from Utah; COAD, colon adenocarcinoma; HapMap, Haplotype Map Project; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LGG, brain lower grade glioma; OV, ovarian serous cystadenocarcinoma; READ, rectum adenocarcinoma; TCGA, The Cancer Genome Atlas project initiated by the National Institute of Health; UCEC, uterine corpus endometrioid carcinoma.

Table S3. Training performance of *k*NN algorithm applied to profiles of SNPs

Actual trait	Predicted trait: BRCA	COAD	HNSC	KIRC	LGG	OV	READ	UCEC	CEU	Sample size	Accuracy (%)
BRCA	2	13	10	2	4	12	6	8	9	66	3.0%
COAD	0	27	12	0	6	11	4	3	3	66	40.9%
HNSC	0	15	36	1	2	1	3	2	6	66	54.5%
KIRC	1	20	9	7	4	6	8	4	7	66	10.6%
LGG	0	19	17	4	9	4	3	2	8	66	13.6%
OV	2	4	5	1	3	45	2	2	2	66	68.2%
READ	2	20	12	6	4	4	13	0	5	66	19.7%
UCEC	4	8	9	1	2	21	2	15	4	66	22.7%
CEU	0	0	1	1	0	0	0	0	66	66	100%
										Sum 594	Overall 37.0%

For the abbreviations, refer to Table S2 legend.

Table S4. Training performance of the SVM algorithm applied to profiles of SNPs

Actual trait	Predicted trait: BRCA	COAD	HNSC	KIRC	LGG	OV	READ	UCEC	CEU	Sample size	Accuracy (%)
BRCA	35	0	3	3	0	0	0	24	1	66	53.0%
COAD	8	33	1	12	2	0	4	5	1	66	50.0%
HNSC	1	0	52	9	1	0	0	1	2	66	78.8%
KIRC	5	1	4	45	1	0	0	9	1	66	68.2%
LGG	1	0	4	1	57	0	0	3	0	66	86.4%
OV	14	0	0	2	0	24	0	26	0	66	36.4%
READ	9	0	1	17	4	0	28	6	1	66	42.4%
UCEC	17	0	1	3	2	0	0	41	2	66	62.1%
CEU	0	0	0	0	0	0	0	0	66	66	100%
										Sum 594	Overall 64.1.0%

For the abbreviations, refer to Table S2 legend.

Table S5. Training performance of SVM algorithm applied to profiles of SNP-Ss

Actual trait	Predicted trait: BRCA	COAD	HNSC	KIRC	LGG	OV	READ	UCEC	CEU	Sample size	Accuracy (%)
BRCA	31	1	2	13	0	3	2	3	11	66	47.0%
COAD	5	11	1	28	1	2	3	1	14	66	16.7%
HNSC	1	2	39	7	2	0	7	0	8	66	59.1%
KIRC	5	10	1	31	2	3	3	4	7	66	47.0%
LGG	1	3	2	7	37	0	4	1	11	66	56.1%
OV	13	2	0	8	1	31	0	4	7	66	47.0%
READ	7	4	1	17	1	2	23	1	10	66	34.8%
UCEC	6	5	1	10	3	1	0	34	6	66	51.5%
CEU	0	0	0	0	0	0	0	0	66	66	100%
										Sum 594	Overall 51.1.0%

For the abbreviations, refer to Table S2 legend.