Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method

Byung-Ju Kim[a,b,c] **and Sung-Hou Kim**[a,b,c,d,1]

[a]Department of Chemistry, University of California, Berkeley, CA 94720; [b]Department of Integrative Omics for Biomedical Sciences, Yonsei University Graduate School, Seoul, Korea; [c]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; and [d]Center for Computational Biology, University of California, Berkeley, CA 94720

Contributed by Sung-Hou Kim, December 21, 2017 (sent for review October 13, 2017; reviewed by Young Seok Ju and Min S. Lee)

Prevention and early intervention are the most effective ways of avoiding or minimizing psychological, physical, and financial suffering from cancer. However, such proactive action requires the ability to predict the individual's susceptibility to cancer with a measure of probability. Of the triad of cancer-causing factors (inherited genomic susceptibility, environmental factors, and lifestyle factors), the inherited genomic component may be derivable from the recent public availability of a large body of whole-genome variation data. However, genome-wide association studies have so far showed limited success in predicting the inherited susceptibility to common cancers. We present here a multiple classification approach for predicting individuals' inherited genomic susceptibility to acquire the most likely phenotype among a panel of 20 major common cancer types plus 1 "healthy" type by application of a supervised machine-learning method under competing conditions among the cohorts of the 21 types. This approach suggests that, depending on the phenotypes of 5,919 individuals of "white" ethnic population in this study, (*i*) the portion of the cohort of a cancer type who acquired the observed type due to mostly inherited genomic susceptibility factors ranges from about 33 to 88% (or its corollary: the portion due to mostly environmental and lifestyle factors ranges from 12 to 67%), and (*ii*) on an individual level, the method also predicts individuals' inherited genomic susceptibility to acquire the other types ranked with associated probabilities. These probabilities may provide practical information for individuals, heath professionals, and health policymakers related to prevention and/or early intervention of cancer.

genomic/environmental factors | *k* nearest neighbor method | SNP syntax | multiple assortment model | cancer risk

Prevention and early diagnosis are the two most effective ways of avoiding or minimizing psychological, physical, and financial suffering from cancer. However, we have only limited knowledge of predicting the susceptibility for various common cancers. There are two broadly defined groups of cancers. Only 5–10% of all known human cancer types are caused primarily by one or a few genes or genomic elements inherited from parents [such as breast invasive carcinoma 1 (*BRCA1*) and *BRCA2* in "hereditary" breast cancer]). However, most common cancer types (over 90%) are caused by the accumulation of uninherited (i.e., acquired) somatic mutations triggered by a complex interaction of three types of causal elements (1, 2), the cancer triad (Fig. 1): environmental factors (the factors imposed on individuals from outside), lifestyle factors (the factors chosen or accepted by individuals), and an individual's inherited genomic susceptibility factors (1).

Types of Inherited Genomic Variations

Of the triad, there are no extensive experimental data or methods, at present, to gather and quantitatively describe the data on an individual's environmental and lifestyle history accumulated during the individual's life. However, the inherited genomic information can be experimentally obtained from the variation of the whole-genome sequences of an individual's noncancerous "germ-line" cells, which change little during one's lifetime. A typical human diploid chromosome contains about 6 billion bp, of which a fraction of 0.1% [about 4–5 million as of 2015 (3)] of diploid locations show genotypes occasionally different from those of the "reference" human genome at an experimentally detectable frequency, and this number is expected to increase. Among many types of variations inherited from ancestors, about 90% are of one type of variation, SNPs (single-nucleotide polymorphisms), caused by mutational events at single-nucleotide positions that resulted in a single base pair change in the ancestor's genomes. The remaining types of variations, such as insertions and deletions of various lengths, copy number variations, inversions, and others, are the results of a much smaller number of mutational events, although they affect more base pairs per event, and it is more challenging, at present, to accurately determine experimentally at the whole-genome level. SNPs can be experimentally identified for an individual as genotypes at the SNP loci spread throughout the entire length of the individual's genome. It is assumed that some of these variants are more likely to contain information about the individual's inherited traits or susceptibility for many kinds of phenotypes, including complex diseases, such as common cancer. Beyond the inherited genomic variations, somatic cancer cells have additional acquired genomic changes (4) resulting from environmental and/or lifestyle factors that are causal to cancer initiation.

Significance

Prevention and early diagnosis of cancer are the most effective ways of avoiding psychological, physical, and financial suffering from cancer. We present a machine-learning method for statistically predicting individuals' inherited susceptibility (and environmental/lifestyle factors by inference) for acquiring the most likely type among a panel of 20 major common cancer types plus 1 "healthy" type. The results show that, depending on the type, about 33–88% of a cancer cohort have acquired its cancer type primarily due to inherited genomic susceptibility factors and that the rest are primarily due to environmental/lifestyle factors. These personal genomic susceptibilities with associated probabilities may provide practical information for individuals, heath professionals, and health policymakers related to prevention and/or early intervention of cancer.

Author contributions: S.-H.K. designed research; B.-J.K. and S.-H.K. performed research; B.-J.K. coded all new computer programs; B.-J.K. and S.-H.K. analyzed data; and B.-J.K. and S.-H.K. wrote the paper.

Reviewers: Y.S.J., Korea Advanced Institute of Science and Technology; and M.S.L., Diagnomics, Inc.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: sunghou@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717960115/-/DCSupplemental.

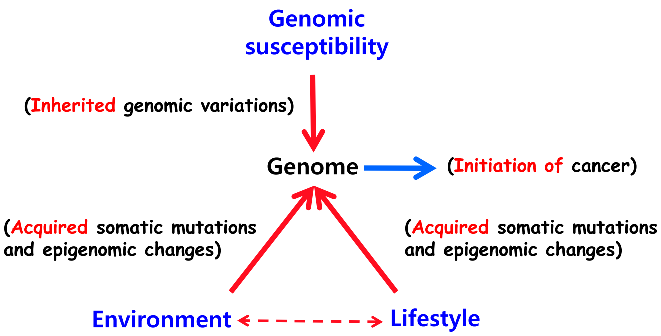
Genomic susceptibility
(Inherited genomic variations)
Genome
(Initiation of cancer)
(Acquired somatic mutations and epigenomic changes)
(Acquired somatic mutations and epigenomic changes)
Environment
Lifestyle

Fig. 1. The cancer triad. A simplified view of interactions among three factors (inherited genomic susceptibility, environmental, and lifestyle factors) that cause somatic mutations and epigenetic modifications of an individual's genome to initiate cancer.

For the purpose of predicting the inherited genomic susceptibility for common cancers, we used only the SNP genotype data obtained from the germ-line cells of cancer patients (most commonly from blood leukocytes or nontransformed somatic cells), because they represent the overwhelming portion of all ancestral genomic variations. The experimental procedure of SNP data acquisition is better established than others at present, and the SNP data descriptors are easier to compare computationally.

Machine Learning

Due to dramatic advances in whole-genome sequencing technology and array technology, a large amount of inherited germ-line SNP information from patients with many different cancer types (4) as well as "healthy" populations (3) became publicly available in recent years. Such availability provides an opportunity to explore developing computational methods to predict an individual's inherited genomic susceptibility for a given cancer type. In addition, since all cancers share common characteristics, such as unregulated cancer cell growth, avoidance of programed cell death, invasion of other normal cells, and others (5), such availability may allow the susceptibility prediction for one cancer type under a competing condition with other cancer types using machine-learning (ML) algorithms (6).

ML algorithms have been developed to build a prediction model from a training set of a large body of observations (i.e., data-driven model) to make predictions on a testing set. Many of them have been shown to be very powerful tools for multiple classification of a large body of data from complex systems, such as handwritten numerical digits, written or spoken words, images including human faces, and others (6). ML methods, thus, are fundamentally different from most genome-wide association studies (GWAS), which can be considered model-driven data selection: the model is the hypothetical polygenic model (7), and the data selected are the low *P*-valued genotypes from the GWAS (8). This approach has so far limited successes in predicting disease susceptibility (9, 10).

In applying ML methods to genomic variations, there is a critically important difference to be recognized in the content of the data used: for the classification of the complex systems mentioned above, the content of the data is "complete" for the description of the systems as, for example, digitized pixels of an image. However, for the prediction of cancer phenotypes, the content of the data available is significantly incomplete to describe the phenotypes, because the data contain only the genomic variation information and no information about environmental and lifestyle factors. This incompleteness of the data content is addressed in *Inherited Genomic Factor vs. Environmental/Lifestyle Factor*.

Previous Studies

There have been many studies to predict susceptibilities for complex diseases, such as common cancers, by GWAS of SNPs with limited success, and the possible reasons for the limited success have been extensively discussed (reviewed in refs. 9 and 10). However, in an earlier study, the feasibility of obtaining much higher susceptibility prediction was shown by using ML algorithms on eight types of cancers plus one "control" type (11). In this study, four different susceptibility prediction methods were explored in two steps. First, the genotypes at whole-genome SNP loci of an individual are described in two different ways: (*i*) by an ordered profile of the SNP genotypes and (*ii*) by an ordered profile of the "SNP genotype syntaxes" (SNP-Ss), where each SNP-S is defined as a linked ordered-SNP genotype of a given length. The former description assumes that each SNP genotype is independent of its neighbor SNP genotypes, and the latter assumes that they may be "linked" beyond what linkage disequilibrium coefficients (12) suggest. Second, the two descriptors were applied to two different supervised ML algorithms: (*i*) *k* nearest neighbor (kNN) method (13) and (*ii*) support vector machine (SVM) method (14). Both algorithms require a minimal number of variable parameters to be optimized using a training dataset to find good prediction models for testing datasets, and thus, it is easier to understand and interpret the results than some other more powerful ML algorithms, such as artificial neural network algorithms of various depth and complexities (15). Of the four sets of results (two descriptors times two algorithms), the best prediction (66% accuracy on average for the training set) for the inherited genomic susceptibility was achieved by using the kNN method on the ordered profiles of SNP-Ss and the next best prediction (62% on average) by using the SVM method on the ordered profiles of SNP genotypes. In both cases, although the validation of the prediction model could be tested on the training dataset, it was not possible to assemble most of the testing sets for all nine phenotypes, because the cohort size available in the public databases at the time of the study was not large enough.

Objective of This Study

Since our feasibility study (11), the numbers of cancer types and cohort sizes in The Cancer Genome Atlas (TCGA) database (4) have increased substantially, such that we now can improve and optimize the prediction model using a sufficient number of testing sets for 20 major cancer types, which account for most of the common cancer occurrences in the United States (16). Two specific objectives of this study are (*i*) to estimate the relative proportion of a cancer cohort who acquired cancer primarily by the inherited genomic susceptibility for the cancer (or its corollary: the proportion due to uninherited factors; i.e., the combined environmental and lifestyle factors) and (*ii*) to estimate the rank-ordered probabilities of an individual's susceptibilities for one or more phenotypes. On a practical side, it is hoped that the information of an individual's highest susceptibility toward 1 of the 21 phenotypes (20 cancer types plus 1 control type) may be useful for the individual to decide to minimize the cancer risk through environment and lifestyle improvement if the individual's inherited genomic susceptibility for the cancer is predicted low, or to proactively monitor for early detection and intervention of the cancer to which the individual is most susceptible if the genomic susceptibility is predicted high.

Results

Accuracy of an ML Prediction for Inherited Genomic Susceptibility.

The results from the ML method of kNN suggest that, depending on the phenotypes of 5,919 individuals of "white" ethnic population in this study representing 20 cancer types and one healthy phenotype, the prediction accuracy for each cohort ranges from about 33–88% (Fig. S1). Fig. 2 highlights the multiclass prediction

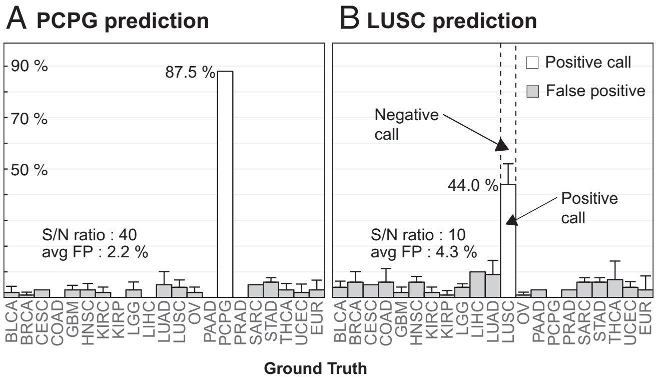

Fig. 2. The prediction of the inherited genomic susceptibility for two cancer types (*A*) PCPG and (*B*) LUSC, corresponding to the highest (87.5%) and median (44.0%) prediction accuracies among the 21 phenotypes. The top of the panel indicates the predicted phenotype, and the *x* axis lists all of the observed ("ground truth") phenotypes of testing sets. (*B*) For example, the testing samples of the LUSC cohort show that 44% of testing samples with the LUSC phenotype (ground truth) are predicted correctly as having acquired LUSC [positive call (PC) or true positive (TP) call as the white bar] and that 56% missed prediction of LUSC [negative call or false negative (FN) call as the dotted bar; negative call is defined here for the LUSC testing samples not predicted as having LUSC phenotype by the kNN model]. All of the gray bars are false positives (defined as non-LUSC testing samples predicted to have LUSC phenotypes), with an average false prediction rate ("average error") of 4.3%, thus giving the S/N ratio of about 10-fold (44 of 4.3) for the PC. The SD for the multiple testing sets is shown as the T on the top of each bar (for the number of testing sets for each phenotype) (Table S1). The interpretation of negative call marked by the dotted bar is in *Inherited Genomic Factor vs. Environmental/Lifestyle Factor*. FP, false positive.

for the cohorts of two cancer types: (*i*) pheochromocytoma and paraganglioma (PCPG) and (*ii*) lung squamous cell carcinoma (LUSC), corresponding to the highest (88%) and median (44%) prediction accuracies of the 21 phenotypes. The signal-to-noise (S/N) ratio for LUSC (Fig. 2*B*), the ratio of the correct prediction ["positive call" (PC)] to the wrong prediction (the average of "false positives"), is 10-fold (44 of 4.3). The predictions and S/N ratios for all 21 phenotypes are shown for comparative purposes in Fig. S1. These prediction accuracies are significantly higher than those predicted for single cancer prediction for a given population by *P* value-based GWAS for common SNPs (see *Comparison of Multiple Allele Assortment Model vs. PolyGenic Model*). Receiver operating characteristic curve (17), derived for the kNN model of the multiclass prediction for the inherited genomic susceptibility using one vs. all approach, shows a reasonably good prediction performance of the method (Fig. S2).

Inherited Genomic Factor vs. Environmental/Lifestyle Factor. Fig. 2*B* shows, as an example, the average prediction accuracy for the testing sets of LUSC, each consisting of 40 samples not used in the training process, obtained from the optimized kNN prediction model (see *Parameter Optimization*). The tallest white solid bar in Fig. 2*B* represents the correct LUSC prediction (PC), meaning that 44% of the testing cohort with LUSC (indicated by LUSC on the *x* axis) is predicted correctly (indicated as LUSC prediction) by the model to be susceptible for the LUSC cancer. Since our testing data contain only the genotype information but do not contain any environmental or lifestyle information, the PC (44%) corresponds to the percentage of the testing cohort of LUSC who acquired LUSC by mostly genomic components of the cancer triad. Thus, the negative calls (56%; the dotted portion above the PC bar in Fig. 2*B*) can be interpreted as (*i*) an error due to the "missed" prediction for LUSC by the model, (*ii*) the portion of the LUSC test cohort who acquired the cancer mostly from nongenomic factors of environment and lifestyle that are absent from the data, or (*iii*) a combination of *i* and *ii*. Since the model error for LUSC prediction is small (4.3%), corresponding to the average of the false positives, interpretation *ii* is likely to be correct (i.e., the negative call of LUSC prediction in Fig. 2*B* corresponds mostly to the fraction of the LUSC cohort who acquired LUSC due to nongenomic factors of environment and lifestyle). Extending this interpretation to all 21 phenotypes, Fig. 3 emphasizes that the cohort of each phenotype can be divided into two groups and that the relative fraction of the cohort who may have acquired the respective phenotype mostly by inherited genomic factors (G group) can be distinguished from those by environmental and/or lifestyle factors (L/E group). The unusually high accuracy for PCPG may be due to the high familial occurrence of pheochromocytoma or paraganglioma observed among the cohort of the phenotype (18) (see *Correlation Between kNN Predictions and Known Observations*).

Summarizing the interpretations for all 21 phenotypes, Fig. 3 and Fig. S1 show that, depending on the phenotype, (*i*) the G group ranges from 33 to 88% of the respective cohort, (*ii*) the ratio of the correct prediction to wrong prediction ranges from 17- to 40-fold, and (*iii*) the remaining portion of each cohort (67–12%) may have acquired the respective phenotype mostly by uninherited (environmental and/or lifestyle) factors.

Cohort Probability vs. Individual Probability. The percentages in previous sections refer to the population probabilities, the percentages of a given cohort who acquired the corresponding phenotype mostly by inherited factors or environment/lifestyle factors. They do not represent the individual probabilities, the probabilities of acquiring the most likely phenotype and other phenotypes for an individual. These probabilities can be estimated from the phenotypes of 10 nearest neighbors (see *Parameter Optimization*) of the individual, which can range from 100%, when all 10 nearest neighbors have the same phenotype, to lower, when minority neighbors have other phenotypes, thus providing ranked probabilities of acquiring various phenotypes, including the most likely phenotype, for the individual. For example, the G group of PCPG in Fig. 3 represents that 88% of the PCPG testing cohort is predicted to be most susceptible to PCPG among the 21 phenotypes. In addition, for each individual in the G group, we can also predict what other phenotypes the person is susceptible to with what probability. For example, Fig. 4*A* shows that, for the individual with the median probability for PCPG prediction accuracy among those in the G group of the PCPG training cohort, there are only three phenotypes found among 10 nearest neighbors as the most likely ones (PCs): eight of them (80%) with PCPG and 10% each for prostate adenocarcinoma (PRAD) and brain lower-grade glioma (LGG) (abbreviations are the same as in Fig. 3). Our analysis provides not only the statistics for each cohort with the most susceptible phenotype, but also, for each individual in a cohort, the ranked probabilities of acquiring various phenotypes other than the most likely phenotype when they are normalized by the prevalence of respective phenotypes.

"Multiple Allele Assortment Model" of Inherited Susceptibility for Common Cancers. Identifying the portion of the cohort of LUSC as PCs (Fig. 2*B*) that corresponds to the G group who acquired the phenotype primarily by their inherited genomic factors provides an opportunity to analyze the population structure within this portion of the cohort. For all G groups of 21 cohorts combined, our analysis consists of two steps. Since each individual is described by a vast dimensional vector of SNP-Ss, we first reduce the dimensionality by the principal component analysis (PCA) method (*Selection of Study Cohorts, SNP Loci, and Genotype Conversion*). Then, we use an unsupervised clustering algorithm, *t*-distributed stochastic neighbor embedding (t-SNE) (19), to cluster all populations of the PCs (i.e., the individuals in the G groups of the 21 testing cohorts together) (see *Inherited Genomic*

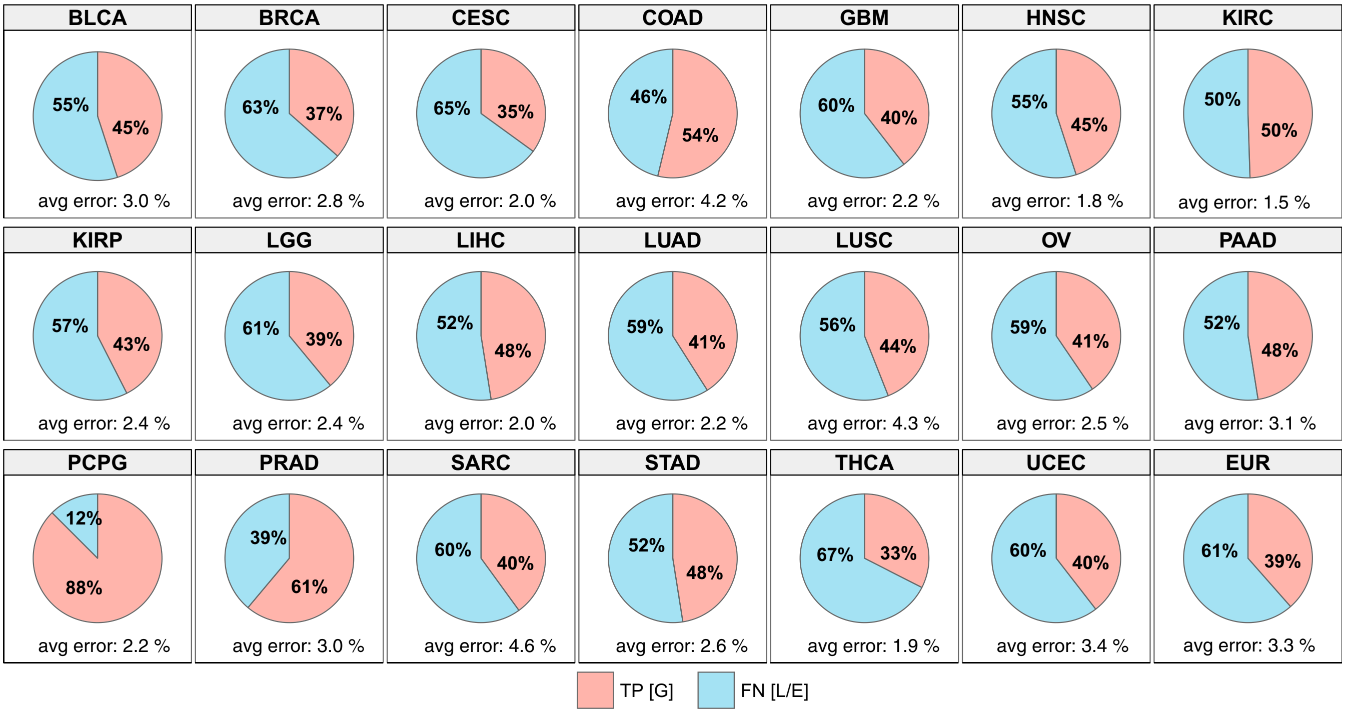

Fig. 3. The relative fraction of the cohort of each phenotype that may have acquired the phenotype mostly by inherited genomic factors (G; in pink) vs. that by environmental and/or lifestyle factors (L/E; in cyan). TP and FN stand for true positive (positive call) and false negative (negative call), respectively, as in the Fig. 2 legend. BLCA, bladder urothelial carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; COAD, colon adenocarcinoma; EUR, European white population; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; SARC, sarcoma; STAD, stomach adenocarcinoma; UCEC, uterine carcinosarcoma.

Factor vs. Environmental/Lifestyle Factor). Fig. 5 shows unsupervised clustering of all of the G group members predicted by our kNN method for the 21 phenotype testing cohorts, but for visual simplicity, only 3 cancer types (PCPG, LUSC, and THCA) of 21 types are made visible. For these three cancer types, each cancer type consists of multiple clusters of individuals, each represented by many different kinds of "features," which are SNP-Ss in this study. Furthermore, some SNP-Ss are present in one cluster but not in other clusters in the same cancer type. This observation reveals a need for a fundamentally different concept to predict the inherited susceptibility than that of the polygenic model (7) used in most GWAS (see *Comparison of Multiple Allele Assortment Model vs. PolyGenic Model*).

Discussion

Comparison of the Multiple Allele Assortment Model vs. the PolyGenic Model. Many GWAS have been performed to predict inherited cancer susceptibility with limited success (9, 10). Of these, BRCA has been one of the most studied cancers by GWAS. For example, a multiplicative polygenic model (7) applied on 76 BRCA-associated SNP genotypes showed 15% prediction accuracy (10) compared with 37% in this study (Fig. S1). In general, there are two major differences in the processes and results between GWAS and kNN. (*i*) All GWAS have been performed for a binary prediction between case and control for each cancer phenotype separately, while in our kNN approach, the prediction was made by a multitype classification process under competing conditions of 20 major cancer types sharing common basic mechanisms of cancer and 1 control type. (*ii*) In the GWAS, the prediction of the inherited susceptibility was made by applying one set of a small number of the *P* value-selected genotypes, usually fewer than 100 SNPs, to a single PolyGenic Model, but in the kNN approach, a very large pool of low-frequency SNP-Ss (on average, about 80,000) is selected; then, an assortment of some of them is primarily associated with one of the multiple clusters and other assortments with other clusters (multiple allele assortment model) (Fig. 5).

Correlation Between kNN Predictions and Known Observations. As shown in Fig. S1, the PCPG cohort shows the highest accuracy for PC prediction (88%), which corresponds to having the largest

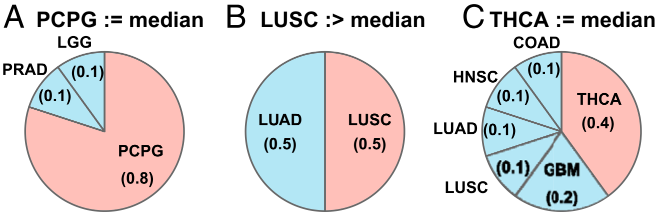

Fig. 4. Individual's probabilities of susceptibility to various phenotypes. The G group of PCPG in Fig. 3 represents the individuals in the PCPG testing cohort who have inherited genomic susceptibility to PCPG as the most likely phenotype among the 21 phenotypes. Within the group, each individual has a different degree of probability for the PCPG susceptibility, although all in the group are predicted to have PCPG as the most likely phenotype to acquire. *A* shows the probabilities of predicting PCPG and other phenotypes (for an individual who has the median probability for PCPG within the group population): 80% of its nearest neighbors have PCPG, and 10% each have PRAD and LGG. *B* and *C* are the probabilities of predicting various phenotypes for a member of LUSC or THCA cohorts, respectively. In general, these probabilities come from the phenotypes of the 10 nearest neighbors in each training set, which can range from 100%, when all nearest neighbors have the same phenotype as the predicted phenotype of the testing individual, to lower percentages when minority neighbors have other phenotypes, thus providing ranked probabilities of acquiring various phenotypes, including the most likely phenotype for the individual. COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; LUAD, lung adenocarcinoma.

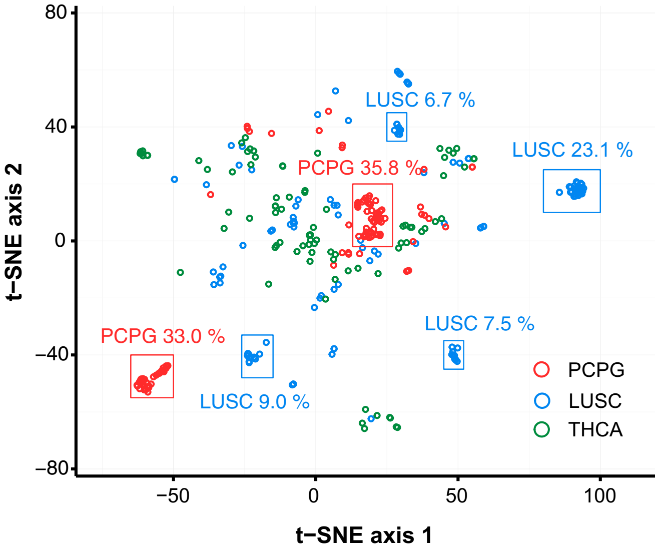

Fig. 5. Unsupervised clustering of the PCs shown in Fig. S1 by the *tsne* program package, which has an implementation of t-SNE (19) in R. For visual simplicity, only 3 cancer types, PCPG, LUSC, and THCA, of 21 types used in the clustering are made visible to illustrate the multiple allele assortment model for cancer susceptibility. There are multiple clusters, each represented by a different assortment of SNP-S features. For PCPG, which shows the highest prediction accuracy, there are only two dense and tight clusters that account for about 69% of the PCs of the combined set of training and testing, and the rest (31%) form many loose small clusters. However, about 46% of LUSC PCs form four dense and small clusters, and the rest (54%) form many small loose clusters.

G group (Fig. 3) (i.e., the most PCPG cohort acquired PCPG by inherited genomic factors). This prediction is consistent with the observation that the PCPG cohort has a very high familial occurrence, suggesting a high inherited genomic susceptibility for PCPG. The germ-line pathogenic mutation of 1 of 14 genes so far discovered accounts for about 30–40% of PCPG, and the mutations in these genes are mostly inherited in autosomal dominant fashion (18). Most likely, more pathogenic genes will be discovered in the future.

Another interesting observation is about inherited vs. somatically altered *BRCA* mutants (a subject not covered in this study). Since the BRCA phenotype has one of the largest cohorts in TCGA database of common cancers, it provides an opportunity to inquire about the contribution of these genes toward acquiring common BRCA, which is caused by the mutation of many genes of low penetrance. An examination of the exom sequence data of the BRCA cohort in TCGA reveals that, of our test set of 200 common BRCA cohorts, only 10 members (5% of the cohort) have somatic mutations of *BRCA-1/2* genes of mostly unidentified penetrance, which agrees with the lower bound of an earlier observation of 5–10% for all breast cancers (20). This observation suggests that, although the pathogenic *BRCA-1* or *-2* genes have been found to have high penetrance for BRCA and account for 20–25% of inherited BRCA, these somatic mutations participate (in collaboration with the mutations of many other genes or genomic elements as expected for a common cancer) in initiating common BRCA only in a very small fraction of the cohort. Furthermore, of the 10 members, 8 members belong to the "E/L group" (the subcohort that acquired BRCA mostly due to environmental/lifestyle factors) of the BRCA testing cohort, suggesting that environmental and lifestyle factors had more influence in triggering somatic *BRCA-1/2* mutations for this subgroup than inherited genomic factors, information useful for the close relatives of the carriers of the somatic mutants of *BRCA* genes. The multiple allele assortment model (see *"Multiple Allele Assortment Model" for Inherited Susceptibility for Common Cancers*) can provide possible explanations for both observations.

A similar interpretation can be made for the role of somatic mutations of *BRCA-1/2* in the common ovarian serous cystadenocarcinoma.

Population Structure of the Sample. Population stratification of genomic variations is implicated in the polymorphic genotype variants as well as the number of variants among 26 geographic populations of the world, suggesting the presence of systematic difference in the variant alleles between the subpopulations in the human population, possibly due to different ancestry (21). To minimize the effect of such stratification in our study, we selected our study samples under the following four considerations. (*i*) We selected only the samples self-reported as white under the race classification category in TCGA, which account for the majority of TCGA data. (*ii*) Since self-reporting of race classification is not always reliable, we used PCA on all samples of the white population and removed about 14% of the samples as "outliers" (Fig. S5 *A* and *B*). (*iii*) The PCA-selected samples (86%) were subjected to a second PCA to see if there is any significant correlation between geographical populations and cancer types. Comparison in Fig. S5 *C* and *D* shows that the geographical populations segregate reasonably well by the second PCA (Fig. S5*C*), but cohorts of different phenotypes do not cluster but are distributed broadly on the same PCA plot space (Fig. S5*D*): two different coloring schemes (for geographical groups and cancer phenotype groups) show no significant correlation between ethnic group identity and the 21 phenotypes as revealed by the distinctly different distribution pattern of each population as well as the positions of the medians of the groups of phenotypes and those of ethnic groups. (*iv*) Finally, we optimized the parameters of the descriptor, SNP-Ss of individual genomic variations, that provides the description of the variation sensitive to the cancer type but not any other factors, such as ancestry, geography, etc. (see *Parameter Optimization*). Similar types of analysis can be easily performed after a large body of genomic variation data becomes available for any ethnic/geographic population or a collection of related such populations.

Systematic Bias Among Datasets of Different Phenotypes. Although all datasets used in this study were obtained from the same genotyping microarray platform, there may be some variations in experimental biases among the datasets of different phenotypes that could influence the training process of the kNN model and thus, the prediction accuracy. The curated data used in this study (*Selection of Study Cohorts, SNP Loci, and Genotype Conversion*) do not seem to have significant bias as indicated by the relatively small differences between the training and testing accuracies for all of the phenotype datasets, except for the cervical squamous cell carcinoma and European white population datasets, which have slightly larger differences than the average (Fig. S3).

Sample Size and Ethnic Diversity. It is surprising that the training sample size as small as 100 for each phenotype could produce a model that can clearly predict the most likely phenotype to which an individual is susceptible with an S/N ratio ranging from 9- to 40-fold (Fig. 3 and Fig. S1). However, it is noticeable that some of the SDs of the predictions, when more than one testing sample is available, are relatively high, ranging from a few to 10% (Fig. S1), as expected for only 100 samples per training set. They are expected to improve as the sample size and diversity for each cohort increase in future studies. As for the applicability of the method to other ethnic populations, similar studies are needed when sufficient data for "nonwhite" ethnic populations become available in future.

Materials and Methods

Data Source of the Germ-Line SNP Genotypes. Thirty-four germ-line (from leukocytes and/or untransformed normal solid tissues) SNP datasets from TCGA (deposited in February 2015) were downloaded with approval. Data downloaded were all at level 2. For the genotypes of control phenotypes, we used those from The 1000 Genome (G1K) Project (3), which used Affymetrix Genome-Wide Human SNP (GHS) Array 6.0 (which is designed to determine about 1 million "tag" SNP genotypes), the same as that used by all 20 groups of TCGA data that we selected for our study (*Selection of Study Cohorts, SNP Loci, and Genotype Conversion*). The genotypes were converted for achieving strand consistency (22) between TCGA and G1K Project data by using GHS Array 6.0 Netaffx release 35 annotation.

Selection of Study Cohorts, SNP Loci, and Genotype Conversion. Of 34 TCGA datasets, we chose the SNP genotype data for 20 cancer phenotypes for which the cohort size is equal to or greater than 180 individuals. A summary of data selection is shown in Table S1. In the selection process, we removed those samples that are duplicated and have third degree kinship (23), selected self-reported white individuals under the race classification category, and removed outliers based on PCA (24, 25). For the control group from the G1K Project data, we chose the samples from five countries: Utah residents with northern and western European ancestries, British in England and Scotland, Toscani in Italy, Finnish in Finland, and Iberian population. The total sample size used for our study was 5,919 individuals. As for the SNP loci selected for the study of 906,600 SNP loci designed to probe in the Affymtrix GHS Array 6.0, 868,023 loci were selected for autosomal SNP loci. We further excluded the loci having Hardy–Weinberg equilibrium test P value $<$ 1.0e-6 and genotype inconsistency ratio $>$1% between reported G1K Project genotypes (3) and genotypes called by Affymetrix Power Tools with default parameters to arrive at the final 818,278 SNP loci, for which all experimentally determined SNP genotypes were downloaded.

Training Set and Testing Sets. The training set contains 2,100 samples, the sum of 100 randomly selected individuals from each of 21 phenotype cohorts. For testing sets for each phenotype, the remaining cohorts that were not selected for training set were divided into multiple, nonoverlapping testing sets of 40 individuals each (Table S1). The training set was used to train the kNN model (*kNN Method*). The contingency table for the training performance, which reveals the validation of the prediction model for the training set, is shown in Table S2, and that for the testing sets is shown in Table S3. The difference between training performance and testing performance is small, in general, as shown in Fig. S3, which validates that the extent of "overfitting" the training data is relatively small; thus, the training sample size is large and random enough to produce a reasonably good prediction model for the size of the testing set.

kNN Method. The details of the application of the kNN method for this study have been published in our feasibility study (11). Conceptually, all training samples (individuals) from 21 phenotypes occupy different positions in a very high-dimensional space (818,278 dimensions), where the position of each sample is described by its high-dimensional vector of ordered SNP-Ss. Then, the method assumes that, for an individual, the majority of its k nearest neighbors has the same phenotype as that of the individual. The optimal parameter k is empirically determined by using all of the training set (*Parameter Optimization*). The nearest neighbors are selected by the distance between the testing sample and all of the training samples, where the distance between two multidimensional vectors is calculated by Jensen–Shannon Divergence (26). Each vector containing the genomic variation information of an individual is described below in *SNP-S*.

Parameter Optimization. Using the data of the training set consisting of 100 samples for each phenotype (*Training Set and Testing Sets*), the optimal parameters are (*i*) the length of SNP-Ss (*SNP-S*) l; (*ii*) the frequency limit f, where the rarest SNP-S among all cohorts has its frequency of f or below; and (*iii*) the number of nearest neighbors k, where the majority of the kNNs has the correct phenotype for the testing individual and gives the best classification among the 21 phenotypes. The result of simultaneous optimization of the three parameters is shown in Fig. S2, where the optimal parameters are empirically determined to be: $l = 8$ SNP genotypes, $f = 1\%$ frequency, and $k = 10$ nearest neighbors. In the training process to find the best parameters, we used the training set of data of all 21 groups. For testing, each individual of the testing sets is applied to the kNN model (*kNN Method*) to find predicted phenotype.

SNP-S. Another assumption that we made is that each SNP locus used in this study may not be independent but linked to its neighbor loci to an unknown extent beyond what the linkage disequilibrium coefficient (12) suggests. Thus, we replace each SNP genotype with an SNP-S, which is simply a string of eight ordered SNP genotypes starting with the starting SNP genotype, where the number eight was empirically determined by optimization to get the best classification of 21 phenotypes by multiclass classification in kNN (*kNN Method*).

ACKNOWLEDGMENTS. We acknowledge the permission to access the database of TCGA and thank all scientists who contributed their genomic data to the database. We also thank Minseung Kim (Department of Computational Sciences, University of California, Davis), who performed the feasibility study with S.-H.K. (11), for his advice and some of the programs used in this study; S. H. Seo and Prof. J. M. Kim (Korea Advanced Institute of Science and Technology) for informing us about the utility of the t-SNE unsupervised clustering method (19); and Profs. G. H. Hahn and G. W. Cho (Yonsei University) for their administrative supervision of B.-J.K. during this study. This research was partly supported, in its early stage, by a grant (to S.-H.K.) from the World Class University Project, Ministry of Education, Science and Technology, Seoul, Republic of Korea, and the rest has been supported by a private gift grant to the University of California, Berkeley (to B.-J.K.). A portion of the work was initiated while Minseung Kim, B.-J.K., and S.-H.K. were affiliated with the Department of Integrated Omics for Biomedical Sciences, The Graduate School, Yonsei University: Minseung Kim as a research associate, B.-J.K. as a graduate student, and S.-H.K. as a visiting faculty member.

1. Cancer Research UK (2016) *Cancer Risk Statistics*. Available at www.cancerresearchuk.org/health-professional/cancer-statistics/risk. Accessed May 5, 2017.
2. National Cancer Institute (2017) *Cancer Prevention Overview*. Available at https://www.cancer.gov/about-cancer/causes-prevention/patient-prevention-overview-pdq. Accessed May 5, 2017.
3. 1000 Genomes Project Consortium; Auton A, et al. (2015) A global reference for human genetic variation. *Nature* 526:68–74.
4. Cancer Genome Atlas Research Network; Weinstein JN, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45:1113–1120.
5. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: The next generation. *Cell* 144: 646–674.
6. Mohri M, Rostamizadeh A, Talwalkar A (2012) Introduction. *Foundations of Machine Learning* (MIT Press, Cambridge, MA), pp 1–9.
7. Witte JS, Hoffmann TJ (2011) Polygenic modeling of genome-wide association studies: An application to prostate and breast cancer. *OMICS* 15:393–398.
8. Anderson CA, et al. (2010) Data quality control in genetic case-control association studies. *Nat Protoc* 5:1564–1573.
9. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 363:166–176.
10. Bahcall OG (March 27, 2013) Common variation and heritability estimates for breast, ovarian and prostate cancers. *Nat Genet*, 10.1038/ngicogs.1.
11. Kim M, Kim S-H (2014) Empirical prediction of genomic susceptibilities for multiple cancer classes. *Proc Natl Acad Sci USA* 111:1921–1926.
12. Slatkin M (2008) Linkage disequilibrium–Understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9:477–485.
13. Steinbach M, Tan P-N (2009) kNN: k nearest neighbors. *The Top Ten Algorithms in Data Mining*, eds Wu X, Kumar V (Chapman and Hall/CRC, Boca Raton, FL), pp 151–162.
14. Theodoridis S, Koutroumbas K (2009) Introduction. *Pattern Recognition* (Academic, New York), 4th Ed, pp 1–12.
15. Nielsen MA (2015) Using neural nets to recognize handwritten digits. *Neural Networks and Deep Learning*. Available at neuralnetworksanddeeplearning.com. Accessed November 12, 2017.
16. Jemal A, et al. (2008) Cancer statistics, 2008. *CA Cancer J Clin* 58:71–96.
17. Majnik M, Bosnic Z (2013) ROC analysis of classifiers in machine learning: A Survey. *Intell Data Anal* 17:531–558.
18. Lalloo F (2016) Diagnosis and management of hereditary phaeochromocytoma and paraganglioma. *Recent Results Cancer Res* 205:105–124.
19. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9: 2579–2605.
20. National Cancer Institute (2015) *BRCA1* and *BRCA2: Cancer Risk and Genetic Testing.* Available at https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet. Accessed May 5, 2017.
21. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456:98–101.
22. Turner S, et al. (2011) Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 68:1–18.
23. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873.
24. Chang CC, et al. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4:7.
25. Diez DM, Barr CD, Cetinkaya-Rundel M (2015) Introduction to DATA. *Openintro Statistics*, 3rd Ed, pp 35–36. Available at www.openintro.org. Accessed November 12, 2017.
26. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37:145–151.