Prediction of Relative Genomic Susceptibility to 20 Major Cancers

- A Machine Learning Approach -

ByungJu Kim

The Graduate School

Yonsei University

Department of Integrated OMICS

for Biomedical Sciences

Prediction of Relative Genomic Susceptibility to 20 Major Cancers

- A Machine Learning Approach -

Thesis Advisor: Professor Gyoonhee Han

Research Advisor: Professor Sung-Hou Kim

Doctoral Dissertation submitted to
the Department of Integrated OMICS for Biomedical Sciences
and the Graduate School of Yonsei University
in partial fulfillment of the requirements
for the degree of Doctor of Science.

Byungju Kim

June 2018

This certifies that
the Doctoral Dissertation of ByungJu Kim is approved.

Thesis Supervisor: Gyoonhee Han

Research Advisor: Sung-Hou Kim

Kyoung-Tai No

Insuk Lee

Young-Joon Kim

The Graduate School
Yonsei University

June 2018

Acknowledgement

After about eight years of preparation, my journey toward the science begins now. Although there have been so many difficulties and hard times, I finally overcome most of them by the cheers and help from others. At this moment, I'd like to thanks to all the people who help me to finish my project. It would not be completed without their help. First, I am grateful to the Korean government, Yonsei University, and my mentors. The Korean government started World-Class University (WCU) project in 2009. They invited famous scholars in their field from the world and gave the students the world-class level lectures and research guidance. Admission to Yonsei University was my starting point. I've been studying genomics in a different point of view under the guidance and mentorship of my advisor, Profs. Sung-Hou Kim, and Gyoonhee Han. Their continuous support and guidance helped me all the time while doing research. I would like to thanks to my colleague, Jaejin Choi, and former colleague, Minseung Kim. They gave me a lot of ideas for my Ph.D. study. I discussed a lot of things with them. Special gratitude goes out to The Cancer Genome Atlas Consortium and The 1000 Genome Project. They provided me with the genomic data for the analysis. Last but not the least, I am grateful to everyone in my family. Their generous supporting and love always encourage me throughout all my life.

Thanks for all your encouragement!

Byung-Ju Kim.

Table of Contents

Acknowledgement iii

Table of Contents iv

Figures vii

Tables viii

Abstract ix

INTRODUCTION 1

Cancer risks and importance of prevention and early intervention 1

Triad of cancer and data availability. 3

Genome-wide Association Study vs. Machine Learning on Genomic Susceptibility. 6

k-Nearest Neighbor and SNP-Syntax 8

Objectives of the Study. 10

Note 10

Methods and Materials 11

Data Preparation. 11

Preliminary QC and Genotyping. 12

Sample Selection and Quality Control 15

Describing genomic variation and genomic comparison. 23

Parameter Optimization 27

Application of *k*NN Algorithm to Predict Genomic Susceptibility. 32

ROC and t-SNE 32

Result 34

Accuracy of an ML Prediction for Inherited Genomic Susceptibility. 34

Inherited Genomic Factor vs. Environmental/Lifestyle Factor. 41

Cohort Probability vs. Individual Probability. 44

"Multiple Allele Assortment Model" of Inherited Susceptibility for Common Cancers. 46

Discussion 48

Comparison of the Multiple Allele Assortment Model vs.the PolyGenic Model. 48

Correlation Between *k*NN Predictions and Known Observations. 48

Population Structure of the Sample 50

Systematic Bias Among Datasets of Different Phenotypes 51

Sample Size and Ethnic Diversity. 54

References 55

Abstract in Korean 61

Publication List 63

Figures

Figure 1. Breast Cancer Stages at Diagnosis vs. Treatment Costs and Its Clinical and Survival Outcome 2

Figure 2. Triad of cancer. 5

Figure 3. SNP-Syntax (SNP-Ss), the concept. 9

Figure 4. Preliminary Quality Control. 14

Figure 5. PCA analysis on Population 20

Figure 6. Concordance/Genotype Reproducibility test 22

Figure 7. Parameter Optimization 29

Figure 8. Overall Result 36

Figure 9. Schematic View on Inherited Genomic Susceptibility 39

Figure 10. Receiver operating characteristic curve analysis 40

Figure 11. Inherited Genomic factor vs. Environmental/Lifestyle factor 43

Figure 12. Relative Genomic susceptibility for an individual 45

Figure 13. t-SNE analysis. 47

Figure 14. Training versus Testing result. 53

Tables

Table 1 Sample Selection. 18

Table 2. Rabin-Karp hash table for genotype 26

Table 3. Training Contingency Table 30

Table 4. Testing Contingency Table 37

Abstract

Prediction of Relative Genomic Susceptibility to 20 Major Cancers
-A Machine Learning Approach –

ByungJu Kim

Department of Integrated OMICS for Biomedical Sciences
The Graduate School,Yonsei University

(Directed by Professor Gyoonhee Han asnd Sunghou Kim)

Cancers are caused by a complex interaction of inherited genomic susceptibility, environmental factors, and lifestyle factors. It causes economic, psychological, and physical burdens of the patients and their family. The best way to avoid and minimize the risks are prevention and early intervention. However, it requires developing a model or framework for predict the cancer susceptibility. Recent advances in technology allow us to measure inherited genomic variations quantitatively and qualitatively at a single nucleotide level. The data required to develop the model is now available with the approval. We use a *k*-Nearest Neighbor model and perform a study of multi-class classification for the genotype of 20 cancer types and one control among 5,919 white individual's genotypes from The Cancer Genome Atlas and the 1000 genome project. The prediction accuracies from

the genomic variation profile are average 45% ranging from 33% to 88% depending on the cancer type and the counterpart of 12% to 67% is mostly from environmental and lifestyle factors.

The proportion of an individual's neighboring traits also could be used as the informative inherited genomic susceptibilities with actionable predictive values which help people for the prevention and early intervention of cancer. Furthermore, we also addressed a different concept of the Multiple Allele Assortment Model. Genomic susceptibility could be obtained by multiple assortments of different sets of genomic elements independently.

Key Words : cancer, genomic susceptibility, prediction, kNN, machine learning

INTRODUCTION

Cancer risks and importance of prevention and early intervention.

Cancers are fatal diseases which are ranked the second leading cause of human death worldwide[1,2]. It was reported that cancer causes 1 in 6 global death, nearly 8.8 millions of deaths, in 2015[3]. The onset of cancer causes an economic, psychological, and physical load of the patient and their family. The more a cancer progresses at diagnosis, the higher the burdens and risks are. Recent studies showed that the treatment cost, calculated by the cost per patient allowed by the insurance company, increases[4] and its effectiveness corresponding to the 5-year relative survival rate decreases[5] as breast cancer progresses at diagnosis (**Fig 1**). The patients diagnosed with the metastasis (stage IV) cost approximately average 2.5-fold higher than those who diagnosed with non-invasive cancer (stage 0). Early diagnosed patients at the stage 0 also showed about 5-fold greater 5-year relative survival rate than those who diagnosed at stage 4. Thus, prevention and early intervention are essential and useful ways to reduce and minimize the caregiver's burden.

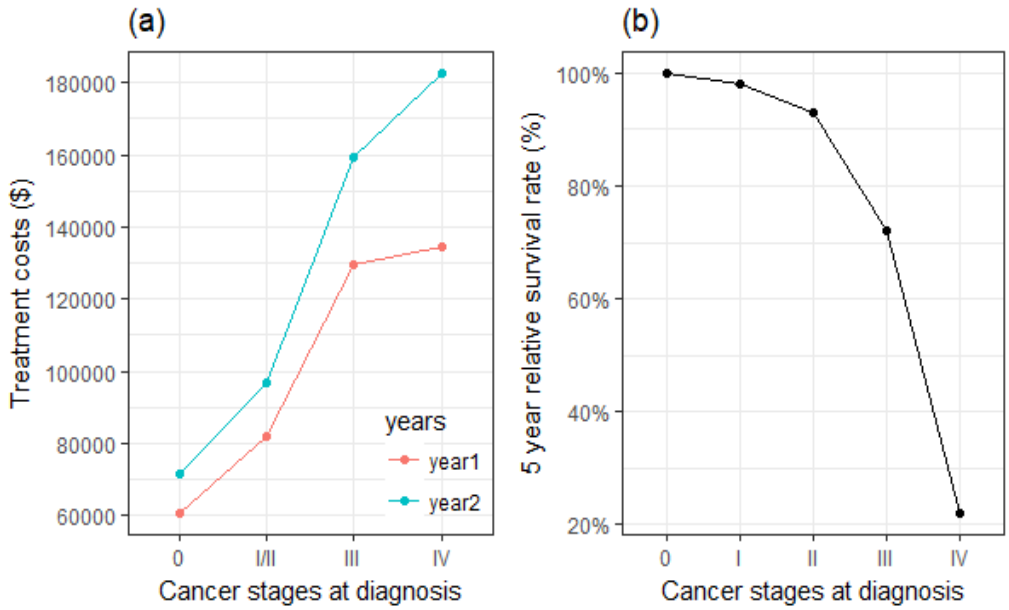

Figure 1. Breast Cancer Stages at Diagnosis vs. Treatment Costs and Its Clinical and Survival Outcome

The relationship between breast cancer stages at diagnosis and treatment costs and its clinical and survival outcome. (a) Treatment cost versus cancer stages at diagnosis. The treatment cost per patient was the average cost allowed by the insurance company during the years after diagnosis. The unit is U.S. dollars. (b) The percentage of the persons diagnosed earlier who survived more than 5 years compared to those whose breast cancer was diagnosed later. The survival rate dramatically decreases depending on the progress of cancer at diagnosis. The 5-year relative survival rate is the fraction of patients who are alive 5 years after the diagnosis of the disease, reflecting the percentage of the general population of corresponding sex and age.

Triad of cancer and data availability.

To achieve the prevention and early intervention, we should understand and predict a person's susceptibility to cancer. Cancers are complex diseases caused by an interaction of the triad, genomic, environmental, and lifestyle factors (**Fig 2**). However, our current knowledge about predicting susceptibility is insufficient. Mostly well-known genetic factors are inherited BRCA1/2 mutations causing breast cancers in the Jewish population[6–8]. Currently, only 15% to 20% of all familial breast cancers are supposed to be inherited by the BRCA1/2 mutation which accounts for 5% to 10% of all women cancers and 5% to 20% of all male breast cancers[9].

The widely accepted model for carcinogenesis is the accumulation of somatic mutations along an individual's lifetime from lifestyle and environmental factors. Tobacco smoking[10], alcohol drinking[11], obesity[12], and occupational exposures to carcinogens, such as cadmium[13], cause cancer. However, it is almost impossible to measure and gather the factors which that accumulate in one's lifetime. An alternative way to obtain the information might be electronic medical/health records which have lifestyle and phenotype information reflecting the combination of the triad at the time of the records were made. However, the clinical information obtained from TCGA is cancer type-specific, meaning the measured variables are unbalanced between different cancer types. For example, lung cancer clinical data includes smoking habits, but the records from the other cancer types do not have any information about it. Thus, the data is not eligible for multi-class classification.

In contrast to the lifestyle and environmental factors, inherited genomic factors are still worth investigating. Since the evidence of hereditary retinoblastoma was

first documented, epidemiological studies have shown correlations between carcinogenesis and genomic contribution. A recent long-term follow-up cancer incidence study of monozygotic and dizygotic twins with 200,000 cohorts showed that monozygotic twins have higher familiar cancer risk within a shared environmental exposure[14]. Women with a family history of breast cancer in first-degree relatives have three- to fourfold greater risks[15]. The genomic factor is researchable due to the completion of the human genome project[16]. Completion of the project means that we can obtain whole genomic information experimentally. Advances in sequencing and microarray technology make the price cheaper and getting an affordable amount of germline genotype data from the patients with various cancer types[17] and control population[18] for a statistical analysis across the entire genome effectively.

We used the Single Nucleotide Polymorphism (SNP, currently referred to Single Nucleotide Variation, SNV) data obtained from the germline cells, most commonly from blood leukocytes or untransformed somatic cells, from cancer patients. Below are the benefits of using them: (*i*) The germline SNP genotypes are almost constant throughout one's life and represent inherited genomic factors since they occupy 90% of the total human genomic variation and are caused by mutation at the single nucleotide level in the ancestor's germline genome. (*ii*) Computational efficiency. The human genome is a diploid consist of six billion base pairs which are approximately 99.9% identical between any unrelated two individuals[18]. Omitting these redundant proportions reduces computation time dramatically.

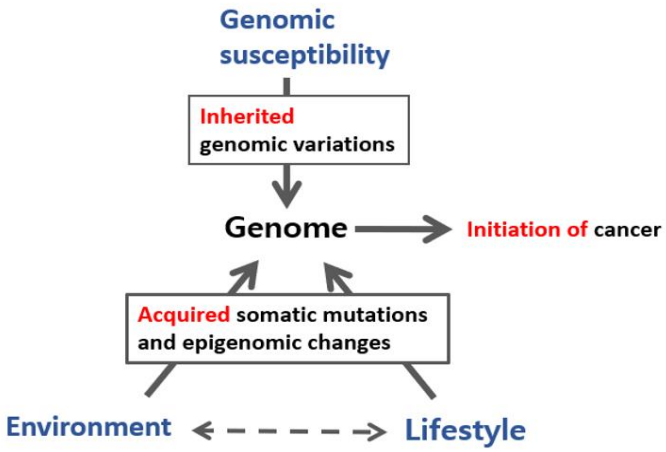

Figure 2. Triad of cancer.

Cancer is caused by the interaction of genomic susceptibility and environmental and lifestyle factors. The genome can change due to somatic mutation or epigenomic change during the lifetime. The genome changes initiate carcinogenesis.

Genome-wide Association Study vs. Machine Learning on Genomic Susceptibility.

The Genome-wide association study (GWAS) is a *de facto* analysis finding disease-associated alleles based on the statistical significance threshold[19]. As of February 6, 2018, the GWAS Catalog had been reported 2,813 unique SNPs from over 150 Cancer and cancer-related traits (search terms for disease/traits are a neoplasm, carcinoma, and cancer)[20]. However, the associated SNPs so far explained only a small proportion of heritability. Thus, the approach has limited successes in predicting disease susceptibility[19,21], e.g., the breast cancer GWAS explains so far only 15% of the familial relative risk for the breast cancer[21]

There could be some reasons. Although cancers share common characteristics, such as unregulated cell growth, apoptosis, invasion, and metastasis[22], it is uncertain which portions and how many genomic variations are associated with cellular phenotypes depending on the cancer type. If the cause of cancer is affected by different allele combinations patient by patient, then the magnitude of effect for one disease-causing allele from the case-control study could be lower than expected and the alleles could be not that statistically significant. Furthermore, healthy people also may have the same disease-causing allele, but they remain healthy since the effects of the interactions are below the threshold for the cancer incidence. Recent replication study of genotype-disease association, designed as control versus healthy people with a family history of the disease, indirectly supports the idea[23].

For those reasons, we tried to solve multi-class classification problem at the genomic level using a data-driven approach. Machine Learning (ML) algorithm

learns the pattern of the observed data we have, training set, to predict unknown data, testing set. It has shown advantages on such multi-class classification problems of complex systems, such as handwritten digits, images, sound, and others[24]. Our Prediction model was built to predict genomic susceptibility to the cancer types under a competing condition with other cancer types using the machine-learning algorithm. We expected that samples having same traits clustered a little more closely due to common susceptibility, while others would not since they randomly shared rare genomic elements.

In our previous feasibility study, we tested four prediction models, using two genotype descriptors and two machine learning algorithms, from 8 cancer types and one European control of HapMap phase 3 genotype data[25]. *k*-Nearest Neighbor (*k*NN) and Support Vector Machine (SVM) were implemented to learn genomic variation patterns from two descriptors, SNP genotypes and SNP-Syntax (collection of linked ordered genotypes at given length). The former descriptor considered each genotype was independent to its neighboring SNPs and the latter assumed that they were related.

The best prediction of 66% for an inherited genomic susceptibility was obtained from the *k*NN algorithm and SNP-Syntax descriptor combination using different parameters, length of the SNP-Syntax (l), and upper-frequency threshold (f), and k for the number of neighbors. Although the study did not test all the cancer types trained due to limited cohort size, it showed better performances than GWAS. A similar study on type I diabetes also supported that the ML algorithm, support vector machine in the study, is better for the risk assessment[26].

k-Nearest Neighbor and SNP-Syntax.

The ML algorithm and descriptors we used were *k*NN and SNP-Syntax since our previous study showed the best results when we used *k*NN and SNP-Syntaxes. *k*NN is a simple supervised learning method which learns observed labeled data and predicts/classify the label of unlabeled data where the parameter *k* is the number of neighboring samples. It is easy to interpret and understand the output. It also has a short calculation time comparing to the novel powerful but complex ML algorithms, such as random forest and neural network. The algorithm divided into two steps. (*i*) Distance calculation among the samples and (*ii*) Assigning the label to unknown samples by a majority vote of label counting of closest *k* samples from the unknown data point (see the Method section for more details)[27].

The chosen SNP-Syntax (**Fig** 3) is the consecutive SNP genotypes of given length along the whole genome coordinates based on the observation of human inheritance pattern, linkage disequilibrium (LD). The LD pattern is associated with several factors such as population structure, inbreeding, genetic drift, and others. To exclude those other factors in our model, we determined to use it as a parameter of machine learning to determine the optimal length of the SNP-syntax which describes the genomic susceptibility mostly.

Figure 3. SNP-Syntax (SNP-Ss), the concept. SNP-Ss are linked and ordered collections of SNPs at given lengths (l)

Objectives of the Study.

As more data is available, it allows us to build to build and test more reliable prediction model than before. The primary purposes of the study are: (*i*) building more reliable prediction model than before due to the increased data size and estimating the relative proportion of the susceptibility which comes from the inherited genome and environmental/lifestyle. (*ii*) finding a more practical way of using the genomic susceptibility at an individual level. (*iii*) studying on the limitation of the previous study which is not sufficiently addressed.

Note

Many parts of the dissertation (especially Results and Discussions) are from recently published my work[28]. However, additional information what we not fully addressed in the published paper is described in more details in introduction and Methods and Materials.

Methods and Materials

Data Preparation.

The Cancer Genome Atlas (TCGA)[17] and One Thousand Genome Project (G1K)[18], were used for the data sources. For the cancer data, we downloaded level 2 genotypes of 34 cancer types on Affymetrix Genome-Wide Human SNP 6.0 (AFFY6) in Feb 2015 under TCGA and dbGaP approval. All the downloaded genotypes were from germline cell (leukocyte, and untransformed solid tissue). Of the samples, we selected the samples with White ancestry for further analysis due to unbalanced cohort sizes between cancer classes. Currently, the data are available from GDC-legacy archive using GDC-client[29].

- Code to download birdseed output from GDC-legacy archive.

```
gdc-client download -t <:token:> -m <:metadata:>
```

Where the token is given encrypted file after getting the approval and metadata is downloadable after selecting samples from GDC-LEGACY Archive

For control group, since there is no "true" control data available, we used G1K European population as the "surrogate" for the control group. Two type of G1K genotype data formats, *VCF* files and raw *CEL* files, had been downloaded from NCBI repository for the G1K project. *VCF* is the variant-call-format which contains variants such as insertion-and-deletion, copy-number-variations, SNVs. Raw CEL files, which are intensity files for the SNP genotypes, generated by Coriell Institute, are in a subfolder of NCBI repository.

- Code to download *CEL* file

```
wget ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/hd_genotype_chip/coriell_affy6_intensities/Affy60_Coriell_CEL_files.tar.gz
```

- code to download *VCF* files

```
wget ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/*.vcf
```

To minimize the effect of suspicious platform and technology bias, we concluded to use genotypes from the same platform which also used by TCGA. After the genotype calling from raw *CEL* files, we validated every marker's quality by comparing every genotype we called to that of the same person in G1K *VCF* files and removed unreliable markers (see Sample Selection and quality control).

Preliminary QC and Genotyping.

Before Genotyping, we excluded some low-quality samples to improve genotyping quality. Preliminary quality control was performed within the whole G1K European population by using *apt-geno-qc* program version 1.18.

```
[:scriptdir:]/apt-geno-qc \
    --cdf-file  [:path to Affy6 library cdf file:] \
    --qcc-file  [:path to Affy6 library qcc file:] \
    --qca-file  [:path to Affy6 library qca file:] \
    --cel-files [:path to CEL file list, one column text file:] \
    --out-file  [:outfile:]  \
    --log-file  [:logfile:]
```

Where *CEL-file-list* should have a column name of *cel_files* at the first line of the file. Other library files are provided by Affymetrix[30]. Of 491 samples, 40 samples with Call Rate (CR) < 0.95 and/or Contrast QC (CQC) < 0.4 are excluded and 451 samples passed the initial quality control process (**Fig 4**). CR is the ratio of the number of samples assigned to any genotype calls. It is an indicator for data completeness and genotype cluster quality. CQC is a metric which measures the separation of allele intensities into 3 clusters in contrast space so that represents the clarity of the signal (see the Affymetrix white paper[30]). We called the genotype on the samples passed the preliminary quality control using *apt-probeset-genotype* program version 1.18 with birdseed v2 algorithm.

The G1K genotypes and downloaded TCGA genotypes were annotated by Affymetrix NetAffx annotation (Release 35) which based on GRCh37 and dbSNP141. G1K VCF files are generated based on the same genome version and dbSNP release. Called genotypes merged to one binary plink file for further analysis.

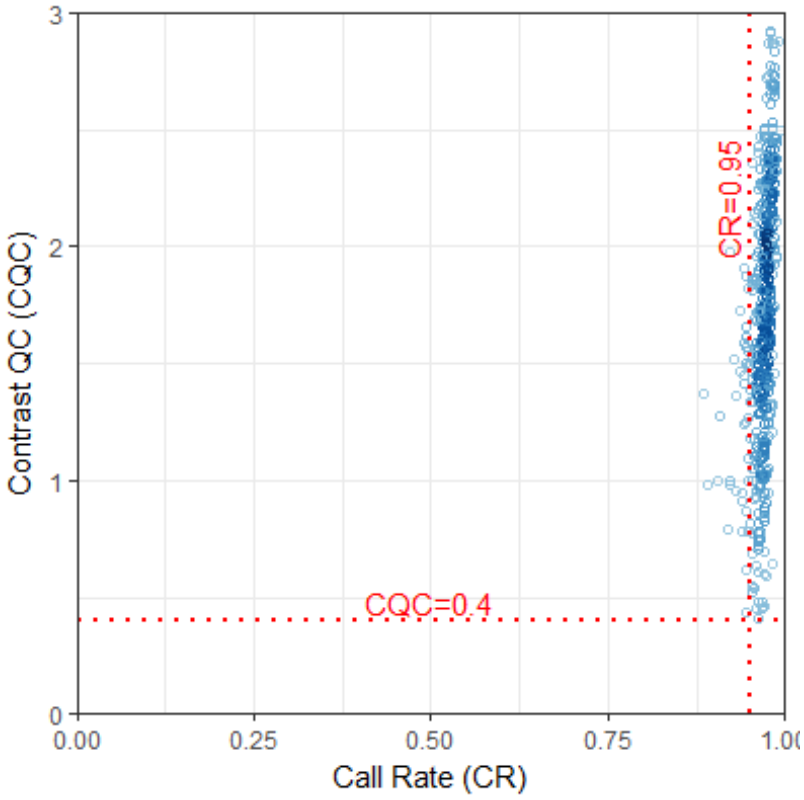

Figure 4. Preliminary Quality Control.

Preliminary quality control. The APT-GENO-QC program version 1.18 measures the quality control metrics, Contrast QC, and call rate. The x-axis is call rate, and the y-axis is Contrast QC. Forty-two samples had not met the criteria of Contrast QC >0.4 and CR >0.95. and were removed to improve genotyping quality.

Sample Selection and Quality Control

A summary of sample selection is in **Table 1.** We chose 9,704 samples from 20 cancer types and one control for which cohort size is equal to/or greater than 180 individuals. Due to the diversity of sample collecting sites and cancer types in TCGA project, investigating sample relatedness helps to evade and minimize the possibility of spurious signals generated by sample overlaps and kinship relationship between cancer cohorts. The sample relatedness is calculated by the proportion of allele sharing patterns, referred as Identity by descent(IBD)/Identity by state (IBS), between two individuals[31] since the algorithm of KING software[32] outperforms others, we excluded samples that are duplicated and have less and equal than third degree of kinship by getting those sample names from following command :

```
king -b <Input file> --related --degree 3
```

Where the input file is a binary plink file with the suffix *"BED"*. Among 9704 samples, 740 samples including 3 G1K control individuals were excluded.

Besides the kinship, PCA outliers, defined as the samples located on the outside of 3 times the interquartile range from first- and third quantile, were also removed (**Fig 5 (b)**). Since the first two eigenvectors reflect the geographic distribution of the sample collected and ethnicity[33], the outliers could be mislabeled samples or utterly different from rest of the observations. Indeed, human migration and reproductive isolation generate human population structure. Any descendants having different ancestry from the others could have different genotype frequencies and phenotypes. Involving those samples in the study population could yield false-positive signals due to unexpected genotype and phenotype differences rather than

expected ones. We first performed the PCA analysis on the whole population (**Fig 5 (a)**) and the self-reported white population (**Fig 5 (b)**). About 10% of the White population seemed mislabeled or genetically different from others Considering both labeled ethnicity and genetic information represented in the PCA plot, we removed all the outliers and finally obtained 5,919 genetically and reported Whites including the 362 G1K population. 86 Finnish in the G1K population were removed during the process (**Table 1**). We also checked if there is any significant correlation between geographical distribution and cancer types (**Fig 5 (c) and Fig 5 (d)**). Samples were clearly segregated by the country but not by the cancer types. one concern is the deviation of the g1k European (EUR) and stomach adenocarcinoma (STAD) population from the center point on PC2 axis. we will address this issue in a discussion section. For genotype or marker quality control, we first extracted autosomal 868,023 loci from 906,600 AFFY6 markers and then performed the Hardy-Weinberg Equilibrium (HWE) exact test on all autosomal markers along the entire genome. In genotyping, each genotype at a given locus is obtained from the intensities of two probes corresponding to the A and B allele, respectively. An intensity scatter plot is drawn on 2-dimensional space where each point denotes a sample. A clustering algorithm is applied to the plot, and the samples are divided at a given locus into three groups corresponding to AA, AB, and BB genotypes. If some points are located on the center position between neighboring two clusters, it is hard to call the genotypes correctly. It results in differences in the frequency of heterozygous (AB) genotypes. The HWE exact test evaluates whether observed genotype frequencies are possible by comparing it to theoretical probability at given observed allele frequencies. Any SNPs deviated from HWE (HWE exact test $p < 1.0^{-6}$) are considered as possible genotyping errors and excluded in the analysis.

We also calculated genotype reproducibility/concordance ratio between genotype calls on G1K samples and corresponding publicly available genotype of the same person in VCF files[34]. Investigating the concordance ratio is valuable for marker quality assurance and technical error detection (**Fig 6**). 622,130 SNPs were 100% reproduced (not shown in the figure), and 195,920 SNPs showed less than 1% discrepancy. Unreliable SNPs (> 1% discrepancy) were removed. However, genotype calls of 228 SNPs are 100% identical at reverse strand to those of publicly available G1K genotypes in VCF file. The 228 exceptional SNPs, which are considered as simple mapping mistakes, were not removed since the genotypes were fully reproduced experimentally. We acquired 818,278 SNP loci for further analysis by removing the unreliable markers.

5,919 samples were divided into training and testing set. A training set consisted of randomly chosen 100 samples per the cancer type, so the total number of the set (*m*) was 2100. The others, which were not training set, were divided into multiple, non-overlapping testing sets of 40 individuals for each cancer type (**Table 1**)

Table 1 Sample Selection.

Class*	Number of downloaded Samples	Filter out of 3rd degree of Kinship†	White population (TCGA clinical information)	After removing PCA outliers‡	Number of Tissue Source Sites §	Sample Size of Training set	Sample size of total testing set¶	number of testing sets
BLCA	395	381	302	285	30	100	160	4
BRCA	1094	1017	697	635	29	100	200	5
CESC	290	276	187	142	23	100	40	1
COAD	467	415	205	198	12	100	80	2
GBM	521	488	395	365	16	100	200	5
HNSC	567	506	433	400	25	100	200	5
KIRC	531	457	400	376	13	100	200	5
KIRP	304	271	196	185	27	100	80	2
LGG	486	482	445	394	25	100	200	5
LIHC	388	346	168	146	27	100	40	1
LUAD	578	487	373	356	27	100	200	5
LUSC	535	475	320	303	28	100	160	4
OV	518	510	397	377	14	100	200	5
PAAD	184	179	157	147	25	100	40	1
PCPG	180	175	144	140	16	100	40	1
PRAD	536	476	138	136	6	100	36	1
SARC	255	249	217	198	28	100	80	2
STAD	464	417	259	201	17	100	80	2
THCA	418	391	271	236	17	100	120	3
UCEC	542	518	355	337	18	100	200	5
EUR	451	448	448	362	5	100	200	5
Total	9704	8964	6507	5919		2100	2756	

To obtain a set of good quality cohorts, several filtering steps were taken. For a given phenotype, any duplicated samples and the “third degree kinship” related samples were removed. Only self-reported white samples under the race classification category of TCGA are selected; and the outliers from the PCA analysis were removed. 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.

*Phenotype name.

†The number of samples after the filtering.

‡The number of samples after removing outliers.

§The number of the locations where the samples were collected.

¶The number of the test samples after all the quality control steps.

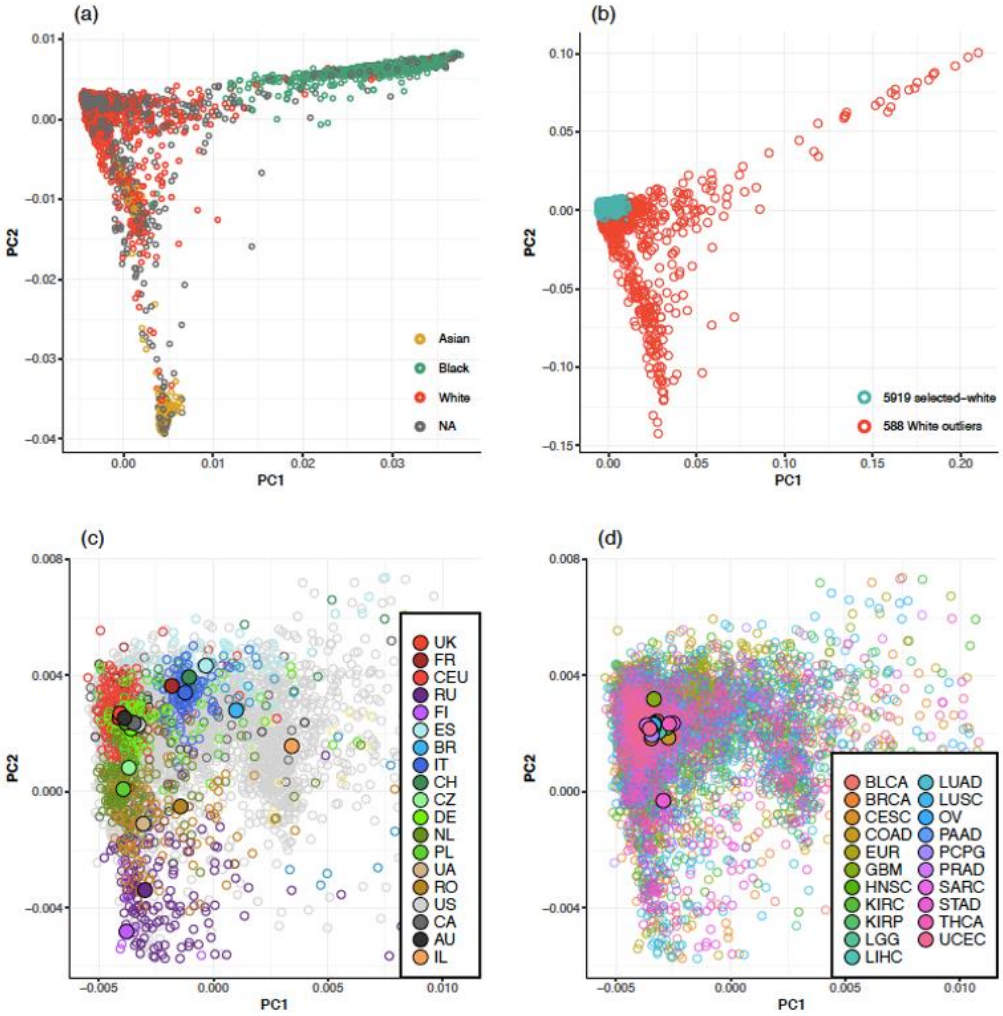

Figure 5. PCA analysis on Population

PCA analysis on the population. (a) PCA plot for all the population downloaded including Asian, Black, and White ethnicity. (b) PCA plot for self-reported "White" population. Ten percent of White samples are not genetically white. We only selected samples that are self-reported white and genetically white proved by PCA. (c) Zoomed PCA plot colored by country from figure 4 (b). The big circle indicates the median point on PC1 and PC2 for

each country. ISO Alpha-2 county code used for the labels. (d) The PCA plot is colored based on the cancer type. This time, the big circle represents the median point of the phenotypes. Except for two trait types, European (EUR) and Stomach adenocarcinoma (STAD), the center points of each phenotype on PC1 and PC2 clustered. Removing PCA outliers solved the population structure problem mostly. TCGA study abbreviation table is used for the labels. UK, United Kingdom; FR, France; CEU, Utah Resident CEPH population; RU, Russia; FI, Finland and FIN population of G1K;ES, Spain and IBS population of G1K; BR, Brazil; IT, Italy and TSI population; CH, Switzerland; CZ, Czech Republic; DE, Germany; NL, Netherlands; PL, Poland; UA, Ukraine; RO, Romania; US, USA ; CA, Canada ; AU, Australia; IL, Israel. BLCA, bladder urothelial carcinoma; BRCA, breast invasive 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; LGG, low grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, uterine carcinosarcoma.

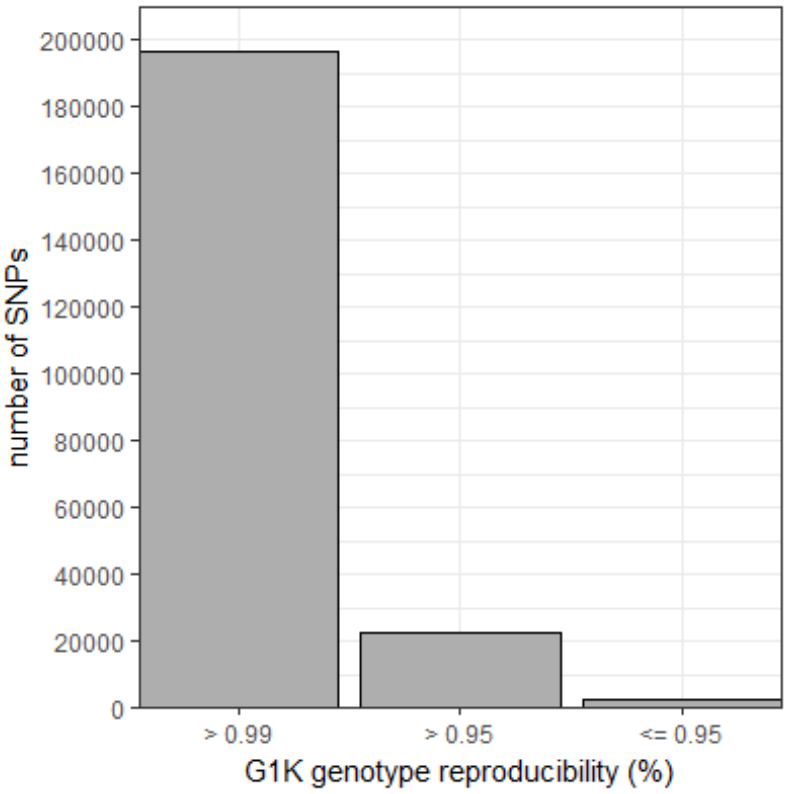

Figure 6. Concordance/Genotype Reproducibility test

G1K genotype reproducibility test result. We compared genotypes on G1K samples to those of the same person in the G1K VCF files. The x-axis is the concordance ratio. About 95% SNPs on the microarray were reproduced with more than 99% concordance ratio. The numbers on the x-axis denote the lower boundaries. 100% reproduced SNPs were excluded from the graph for the visualization.

Describing genomic variation and genomic comparison.

Genotypes (g_{ij}) of whole samples (M) are represented on 2-dimensional space of matrix consisting of *m* samples by *n*-allele loci such as

$$M = \begin{bmatrix} g_{1,1} & g_{1,2} & g_{1,3} & \cdots & g_{1,n} \\ g_{2,1} & g_{2,2} & g_{2,3} & \cdots & g_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{m,1} & g_{m,2} & g_{m,3} & \cdots & g_{m,n} \end{bmatrix}$$

To describe genomic variation at whole genome level of an individual (*i*), we use a sliding window of given length (*l*) across the whole genome-wide SNP from position 1 to end of the genome (n - l +1, where n is the whole length of the position). These l-length SNP words are a SNP-syntaxes (SNP-Ss). We made SNP-Ss profiles for each individual using the Rabin-Karp Hash algorithm. The algorithm covert genotype ($g_{i,j}$) into numeric values ($d_{i,j}$) using the hash table shown in **Table 2**. *Parental A allele + maternal G allele* and *parental G allele + maternal A allele* genotypes are considered as the same one in the analysis. Thus, 16 genotypes are converted into 10 digits, 0 to 9. the total individual genomic variations ($SS_{i,l}$) described by *l*-length SNP-Ss are:

$$SS_{i,l} \ni \{S_{i,1,l} \ldots S_{i,j,j+l-1} \ldots S_{i,n-l+1,n}\}$$

where *i* is an index for a sample, *n* for the total number of loci, j for relative genomic coordinate, and $S_{i,j,j+l-1}$ is the SNP-Ss written as

$$S_{i,j,j+l-1} = j \times 10^{l} + d_{i,j} \times 10^{l-1} + d_{i,j+1} \times 10^{l-2} \ldots d_{i,j+l-1} \times 10^{l-l}$$

The SNP-syntax ($S_{i,j,j+l-1}$) is a numeric value consisting of given *l* length numeric genotypes starting at locus *j*. The next syntax for the sample *i* is simply calculated by

$$S_{i,j+1,j+1+l-1} = (j+1) \times 10^{l} + (S_{i,j,j+l-1} \mathrm{mod} 10^{l-1}) \times 10 + d_{i,j+l}$$

We counted SNP-Ss within all the training samples and then divided each count by the cohort size of the training set to make the frequency vector consisting of SNP-Ss and its frequency. The given filtering parameter (*f*) is applied to the frequency vectors to make the list of filtered SNP-Ss at the given *l* and *f*. Then, the genomic variation vector per a sample was reconstructed by picking SNP-Ss in the list from individual's original SNP-Ss profile. The new filtered SNP-Ss profile consisted of the filtered SNP-Ss and its relative frequency within the sample.

Finally, Jensen-Shannon (JS) divergence was calculated between two newly processed filtered SNP-Ss sets (P_l and Q_l) to build a training cohort size by training cohort size distance matrix by followings [35,36]:

$$JS_l(P_l, Q_l) = \frac{1}{2} KL(P_l, M_l) + \frac{1}{2} KL(Q_l, M_l)$$

Where $M_l = \frac{(P_l + Q_l)}{2}$ and *KL* is the Kullback-Leibler divergence,

$$KL(P_l, M_l) = \sum_{j=1}^{K} (P_{l,j} \times log_2 \frac{p_{l,j}}{m_{l,j}})$$

where *K* is the total number of SNP-syntax in the individual's filtered profile at given *f* and *l*, and *j* is the index for the element of the vector. The output of Jensen-Shannon divergence value ranges from 0 to 1. The code for JS divergence

calculation was implemented from Feature Frequency Profile program[36]

Table 2. Rabin-Karp hash table for genotype

Parental Allele \ Maternal Allele	A	C	G	T
A	0	4	5	6
C	4	1	7	8
G	5	7	2	9
T	6	8	9	3

The Rabin-Karp hash algorithm converts each genotype into a numeric value based on the hash table and calculates the next SNP-Ss by reading one genotype rather than the extracting the *l*-length word every time

Parameter Optimization

Parameter optimization step empirically seeks the best parameter combination which gives the best prediction during training and validation. In our analysis, three parameters were optimized by repeating *k*NN with different parameters settings: the length of SNP-Ss (*l*), upper-frequency threshold (*f*), and the number of neighbors (*k*). In details, we first made the distance matrix among training set at given *l*, *f*, and *k* combination (Describing genomic variation and Genomic comparison) and selected the nearest *k* samples except self by rows from the distance matrix. The cancer type of the selected sample was determined based on the majority votes of its *k* neighbors. When the vote was tied, we chose the cancer type that had shorter average distance from the chosen one. **Figure 7** shows the parameter optimization result. The performance of *l*=6, 8 and, 10 were not significantly different when the parameters *f* =1 and *k*=10 were used. *l*=8 is the most stable point. Thus, we picked the optimized parameters of *l*=8, *f*=1, and, *k*=10. The training results from the optimized parameters are shown in **Table 3**.

- Pseudocode for *k*NN. We repeated the procedure several times with different *l* and *f* combination.

```
1. read the distance/divergence matrix
2. Initialize the value of k
3. To obtain the best prediction and optimize k, iterate
   from k=1 to total number of training data points
   ## for testing, only optimized one k value (optimized)
      is used.
        3-1. Calculate the distance between training/test data
             and sort them in ascending order.
```

```
        3-3. Extract top k rows from the sorted array
             (calculate the average distance per class
              during sorting process)
        3.4. Get the most voted class of these rows
             (if there is a tie, use the average distance
              calculated in step 3-3 for the tie break.)
        3.5. Return the predicted class
4. Return to step 3 with different k (e.g. k+1)
   until best prediction value is observed.
```

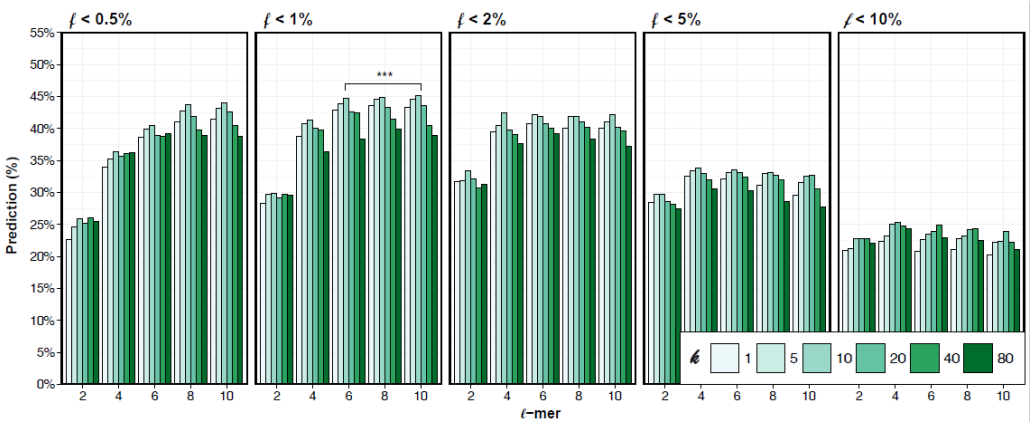

Figure 7. Parameter Optimization

Parameter optimization plot. We optimized three parameters, length (*l*), frequency threshold (*f*), and the number of neighboring samples (*k*), by repeated *k*NN at the different parameter combinations. The best parameters are 8, 1, and 10, respectively

Table 3. Training Contingency Table

		Predicted Phenotype																				
		BLCA	BRCA	CESC	COAD	GBM	HNSC	KIRC	KIRP	LGG	LIHC	LUAD	LUSC	OV	PAAD	PCPG	PRAD	SARC	STAD	THCA	UCEC	EUR
Ground Truth	BLCA	45.00%	1.00%	7.00%	5.00%	2.00%	1.00%	0.00%	1.00%	0.00%	5.00%	0.00%	3.00%	3.00%	3.00%	1.00%	2.00%	9.00%	8.00%	2.00%	1.00%	1.00%
	BRCA	2.00%	36.00%	5.00%	4.00%	2.00%	4.00%	3.00%	1.00%	1.00%	3.00%	4.00%	2.00%	1.00%	4.00%	2.00%	3.00%	1.00%	2.00%	5.00%	11.00%	4.00%
	CESC	12.00%	1.00%	49.00%	0.00%	0.00%	0.00%	3.00%	5.00%	0.00%	0.00%	3.00%	6.00%	0.00%	2.00%	2.00%	1.00%	9.00%	0.00%	0.00%	4.00%	3.00%
	COAD	2.00%	5.00%	0.00%	51.00%	3.00%	3.00%	3.00%	1.00%	1.00%	2.00%	2.00%	1.00%	5.00%	0.00%	2.00%	4.00%	5.00%	4.00%	3.00%	1.00%	2.00%
	GBM	4.00%	2.00%	2.00%	5.00%	37.00%	0.00%	1.00%	3.00%	2.00%	3.00%	6.00%	2.00%	6.00%	5.00%	4.00%	8.00%	4.00%	2.00%	1.00%	1.00%	2.00%
	HNSC	0.00%	4.00%	0.00%	7.00%	2.00%	42.00%	0.00%	3.00%	5.00%	0.00%	0.00%	3.00%	3.00%	3.00%	4.00%	3.00%	6.00%	1.00%	4.00%	4.00%	6.00%
	KIRC	2.00%	2.00%	3.00%	2.00%	3.00%	1.00%	42.00%	3.00%	3.00%	1.00%	2.00%	4.00%	3.00%	5.00%	3.00%	3.00%	6.00%	1.00%	3.00%	4.00%	4.00%
	KIRP	2.00%	1.00%	4.00%	2.00%	0.00%	2.00%	0.00%	44.00%	1.00%	6.00%	2.00%	2.00%	2.00%	5.00%	4.00%	0.00%	12.00%	2.00%	2.00%	3.00%	4.00%
	LGG	2.00%	1.00%	1.00%	3.00%	3.00%	3.00%	0.00%	1.00%	44.00%	3.00%	1.00%	2.00%	0.00%	2.00%	3.00%	1.00%	6.00%	5.00%	5.00%	5.00%	9.00%
	LIHC	3.00%	6.00%	2.00%	6.00%	2.00%	7.00%	1.00%	5.00%	2.00%	39.00%	1.00%	1.00%	1.00%	3.00%	3.00%	2.00%	7.00%	0.00%	2.00%	3.00%	4.00%
	LUAD	0.00%	1.00%	1.00%	4.00%	6.00%	2.00%	2.00%	2.00%	1.00%	0.00%	43.00%	12.00%	7.00%	0.00%	2.00%	3.00%	2.00%	1.00%	2.00%	3.00%	6.00%
	LUSC	0.00%	1.00%	5.00%	4.00%	2.00%	1.00%	2.00%	1.00%	0.00%	1.00%	6.00%	45.00%	4.00%	3.00%	3.00%	5.00%	2.00%	3.00%	4.00%	3.00%	5.00%
	OV	2.00%	2.00%	0.00%	5.00%	3.00%	1.00%	2.00%	2.00%	2.00%	1.00%	7.00%	2.00%	41.00%	5.00%	2.00%	4.00%	3.00%	1.00%	2.00%	3.00%	10.00%
	PAAD	5.00%	1.00%	4.00%	3.00%	0.00%	1.00%	0.00%	4.00%	3.00%	2.00%	4.00%	0.00%	4.00%	41.00%	3.00%	4.00%	6.00%	7.00%	2.00%	3.00%	3.00%
	PCPG	1.00%	1.00%	1.00%	1.00%	2.00%	1.00%	0.00%	2.00%	0.00%	1.00%	1.00%	1.00%	2.00%	1.00%	77.00%	3.00%	2.00%	0.00%	1.00%	1.00%	1.00%
	PRAD	0.00%	1.00%	4.00%	3.00%	5.00%	2.00%	2.00%	1.00%	1.00%	3.00%	1.00%	3.00%	1.00%	3.00%	3.00%	51.00%	5.00%	3.00%	0.00%	3.00%	5.00%
	SARC	4.00%	0.00%	6.00%	2.00%	0.00%	2.00%	1.00%	9.00%	1.00%	6.00%	0.00%	6.00%	2.00%	6.00%	4.00%	3.00%	32.00%	5.00%	7.00%	1.00%	3.00%
	STAD	3.00%	1.00%	1.00%	7.00%	0.00%	0.00%	0.00%	2.00%	3.00%	4.00%	1.00%	6.00%	2.00%	3.00%	1.00%	2.00%	3.00%	51.00%	5.00%	3.00%	2.00%
	THCA	4.00%	4.00%	2.00%	5.00%	2.00%	3.00%	1.00%	5.00%	1.00%	4.00%	0.00%	4.00%	1.00%	0.00%	0.00%	3.00%	9.00%	4.00%	39.00%	6.00%	3.00%
	UCEC	5.00%	8.00%	1.00%	1.00%	1.00%	1.00%	5.00%	2.00%	2.00%	1.00%	4.00%	3.00%	2.00%	0.00%	4.00%	5.00%	2.00%	6.00%	3.00%	38.00%	6.00%
	EUR	1.00%	1.00%	4.00%	3.00%	2.00%	1.00%	1.00%	2.00%	2.00%	2.00%	1.00%	1.00%	2.00%	3.00%	0.00%	4.00%	4.00%	1.00%	3.00%	7.00%	55.00%

Contingency table for the training set. Training performance of the *k*NN algorithm applied to the profiles of SNP-Ss. Observed (ground truth) phenotypes are listed on the y axis, and predicted phenotypes are on the x axis. BLCA, bladder urothelial carcinoma; BRCA, breast invasive 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; LGG, low grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, uterine carcinosarcoma

Application of *k*NN Algorithm to Predict Genomic Susceptibility.

For the testing, the optimized parameters were applied to all the testing samples. We extracted SNP-Ss which included in the filtered SNP-Ss set at given l= 8 and f = 1, and then calculate relative frequency to generate genomic variation vectors for testing samples. JS divergences between all testing and training samples were calculated. Finally, the *k*NN with optimized *k* parameter was applied to the matrix to predict phenotypes of unknown testing samples (see result**)**.

ROC and t-SNE

The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are designed to measure and visualize the performance of a binary classifier. The ROC curve is a plot about sensitivity (true positive rate) versus 1-specificity (false positive rate) of the classifier. The AUC is a summary of the performance. It ranges from 0 to 1: 0.5 is a random level prediction, and 1 is the perfect test. We also used ROC curve for the performance measurement. Since our classifier is a multi-class classifier, we used the one-versus-all approach to calculate this[37]. The package we used was pROC implemented in R software[38].

t-distributed stochastic neighbor embedding (t-SNE) is the unsupervised machine learning method developed for the data visualization[39]. The algorithm brings a set of points from high-dimensional space to the low-dimensional space (usually, 2- or 3-dimensional space) and makes a faithful representation of those

points without breaking the order of data point as much as possible. In briefly, it makes similar objects more closely and distant objects more distantly but keep the relative rank of distance. We used *tsne* package[40] implemented in R for the analysis. given that our input data is high-dimensional data contains considerable noises (average ~80,000 SNP-Ss per individual), We first reduced the dimension of the data using PCA and performed t-SNE.

Result

Accuracy of an ML Prediction for Inherited Genomic Susceptibility.

The results from the ML method of *k*NN suggest that, depending on the phenotypes of 5,919 individuals of "white" ethnic population in this study representing 20 cancer types and one control phenotype considered as healthy, the prediction accuracy for each cohort ranges from about 33–88% (**Fig. 8, Table 4**). **Fig. 9** highlights the multi-class prediction 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 accuracie s of the 21 phenotypes. The signal-to-noise (S/N) ratio, for LUSC (**Fig. 9 (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 8**. 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[37], derived for the *k*NN model of the multi-class prediction for the inherited genomic susceptibility shows a reasonably good prediction performance (AUC: 0.75~0.96) of the method (**Fig. 10**).

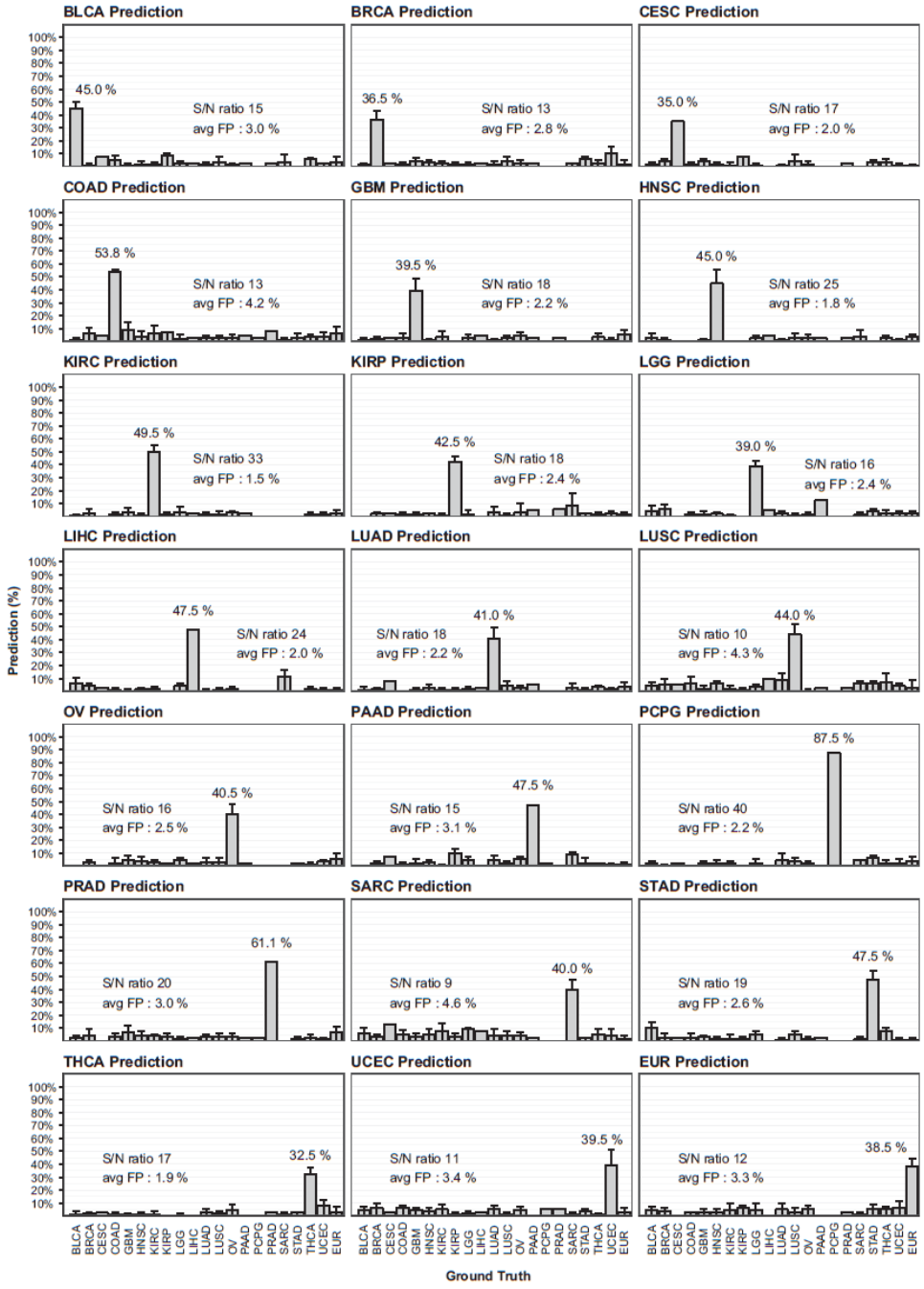

Figure 8. Overall Result

Inherited genomic susceptibility prediction for 21 traits (20 cancer and one control, European considered as healthy). The ground truth and prediction percentage (%) are on the X-axis and Y-axis, respectively. The predicted traits are at the top of each panel. Every panel contains a signal-to-noise ratio (S/N) and average false positive (FP) and true positive (TP) rates. S/N is calculated by a ratio of TPs to average FPs. CESC, PAAD, PCPG, and PRAD do not contain SD since the testing set contains only one testing set due to cohort size limitations. The bars without SDs are the phenotype group with 1 or 2 testing set. BLCA, bladder urothelial carcinoma; BRCA, breast invasive 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; LGG, low grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, uterine carcinosarcoma

Table 4. Testing Contingency Table

		Predicted phenotype																			
		BLCA	BRCA	CESC	COAD	GBM	HNSC	KIRC	KIRP	LGG	LIHC	LUAD	LUSC	OV	PAAD	PCPG	PRAD	SARC	STAD	THCA	UCEC	EUR
Ground truth	BLCA	45.00%	1.25%	1.25%	1.25%	1.25%	3.13%	0.63%	0.00%	3.75%	6.25%	1.25%	4.38%	0.00%	0.00%	1.88%	2.50%	5.63%	10.63%	1.25%	4.38%	4.38%
	BRCA	1.00%	36.50%	4.50%	6.50%	2.00%	1.00%	2.50%	2.00%	5.50%	4.00%	1.50%	5.50%	3.00%	2.50%	0.50%	4.50%	3.00%	2.50%	2.00%	6.00%	3.50%
	CESC	7.50%	2.50%	35.00%	5.00%	2.50%	0.00%	0.00%	2.50%	0.00%	2.50%	7.50%	5.00%	0.00%	7.50%	2.50%	0.00%	12.50%	2.50%	2.50%	2.50%	0.00%
	COAD	5.00%	1.25%	1.25%	53.75%	2.50%	0.00%	1.25%	1.25%	1.25%	1.25%	0.00%	6.25%	2.50%	1.25%	0.00%	3.75%	5.00%	2.50%	1.25%	6.25%	2.50%
	GBM	1.00%	4.50%	4.00%	9.00%	39.50%	1.00%	3.50%	1.50%	1.50%	0.50%	1.00%	2.00%	5.00%	2.00%	2.50%	6.50%	3.50%	3.00%	1.00%	4.50%	3.00%
	HNSC	2.00%	3.50%	1.50%	3.50%	1.00%	45.00%	1.00%	2.50%	2.00%	1.50%	3.00%	6.00%	4.00%	3.00%	2.50%	4.00%	5.00%	1.50%	1.50%	3.50%	2.50%
	KIRC	1.50%	2.50%	1.00%	6.00%	3.50%	0.00%	49.50%	1.00%	0.50%	1.50%	1.00%	2.00%	3.00%	0.50%	2.00%	4.00%	8.00%	1.50%	1.50%	5.00%	4.50%
	KIRP	8.75%	1.25%	7.50%	7.50%	0.00%	0.00%	1.25%	42.50%	0.00%	0.00%	1.25%	1.25%	2.50%	10.00%	0.00%	3.75%	3.75%	1.25%	0.00%	1.25%	6.25%
	LGG	2.50%	2.00%	1.00%	2.00%	3.00%	2.50%	3.00%	1.50%	39.00%	4.00%	1.50%	3.50%	5.00%	4.50%	2.50%	1.50%	9.00%	5.00%	0.50%	2.00%	4.50%
	LIHC	2.50%	2.50%	0.00%	2.50%	5.00%	5.00%	2.50%	0.00%	5.00%	47.50%	2.50%	10.00%	2.50%	0.00%	0.00%	2.50%	7.50%	0.00%	0.00%	2.50%	0.00%
	LUAD	2.00%	2.00%	0.50%	2.50%	1.00%	1.00%	1.50%	3.50%	2.00%	0.50%	41.00%	8.50%	3.00%	5.00%	4.50%	3.50%	4.00%	1.00%	2.50%	5.50%	5.00%
	LUSC	3.50%	4.50%	4.50%	2.50%	2.50%	3.00%	1.50%	1.00%	1.00%	1.00%	4.00%	44.00%	3.00%	2.00%	3.50%	3.00%	4.00%	5.00%	2.00%	1.50%	3.00%
	OV	1.50%	2.50%	2.00%	2.50%	4.50%	2.50%	3.00%	3.00%	1.50%	1.50%	3.00%	0.50%	40.50%	5.50%	2.00%	3.50%	4.50%	1.50%	4.50%	4.50%	5.50%
	PAAD	2.50%	2.50%	0.00%	5.00%	2.50%	2.50%	2.50%	5.00%	12.50%	0.00%	5.00%	2.50%	2.50%	47.50%	0.00%	2.50%	2.50%	2.50%	0.00%	0.00%	0.00%
	PCPG	0.00%	0.00%	0.00%	2.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.50%	87.50%	2.50%	0.00%	0.00%	0.00%	5.00%	0.00%
	PRAD	2.78%	0.00%	2.78%	8.33%	2.78%	2.78%	0.00%	5.56%	0.00%	0.00%	0.00%	2.78%	0.00%	0.00%	0.00%	61.11%	0.00%	0.00%	2.78%	5.56%	2.78%
	SARC	3.75%	2.50%	0.00%	1.25%	0.00%	3.75%	0.00%	8.75%	1.25%	11.25%	2.50%	6.25%	0.00%	8.75%	5.00%	0.00%	40.00%	1.25%	1.25%	1.25%	1.25%
	STAD	0.00%	6.25%	3.75%	2.50%	0.00%	0.00%	0.00%	2.50%	3.75%	0.00%	1.25%	6.25%	2.50%	2.50%	6.25%	1.25%	2.50%	47.50%	2.50%	3.75%	5.00%
	THCA	5.83%	2.50%	3.33%	4.17%	3.33%	2.50%	1.67%	1.67%	2.50%	1.67%	3.33%	6.67%	1.67%	2.50%	2.50%	2.50%	5.00%	7.50%	32.50%	0.83%	5.83%
	UCEC	2.50%	10.50%	1.00%	3.50%	1.50%	1.00%	1.50%	2.50%	2.00%	1.00%	1.50%	4.00%	3.50%	1.00%	2.00%	1.50%	4.50%	1.00%	8.00%	39.50%	6.50%
	EUR	3.50%	2.00%	0.50%	6.50%	5.50%	4.00%	2.50%	1.50%	2.50%	1.00%	3.50%	3.00%	5.50%	1.50%	3.50%	6.50%	2.00%	1.00%	2.50%	3.00%	38.50%

Contingency table for testing sets. Testing performance of the kNN algorithm applied to profiles of SNP-Ss. Observed (ground truth) phenotypes are listed on the y axis, and predicted phenotypes are on the X-axis. BLCA, bladder urothelial carcinoma; BRCA, breast invasive 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; LGG, low grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, uterine carcinosarcoma.

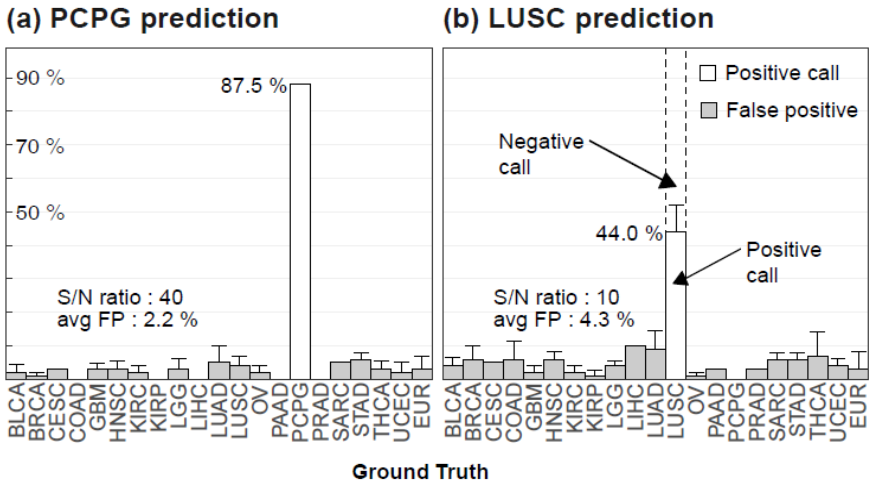

Figure 9. Schematic View on Inherited Genomic Susceptibility

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 *k*NN 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.

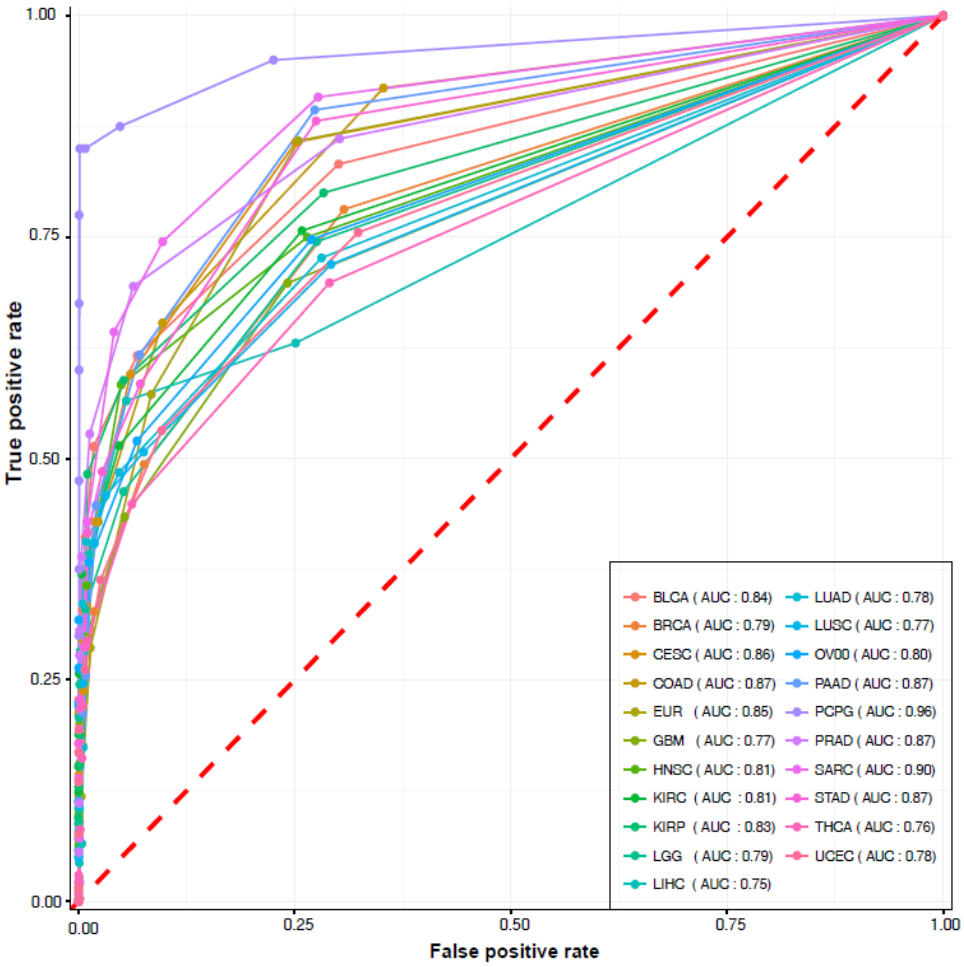

Figure 10. Receiver operating characteristic curve analysis

We accessed the performance of the *k*NN classifier with the receiver operating characteristic (ROC) curve through a one-versus-all approach. The fraction of correct calls among neighboring k samples was used for the score for each classification. The AUC ranged from 0.75 to 0.96.

Inherited Genomic Factor vs. Environmental/Lifestyle Factor.

Fig. 9 (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 *k*NN prediction model (see Parameter Optimization). The tallest white solid bar in **Fig. 9 (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. 9 (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. 9 (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. 11** 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[41] (see Correlation Between *k*NN Predictions and Known Observations).

Summarizing the interpretations for all 21 phenotypes, **Fig. 11** and **Fig. 8** 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 predictions to wrong predictions 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.

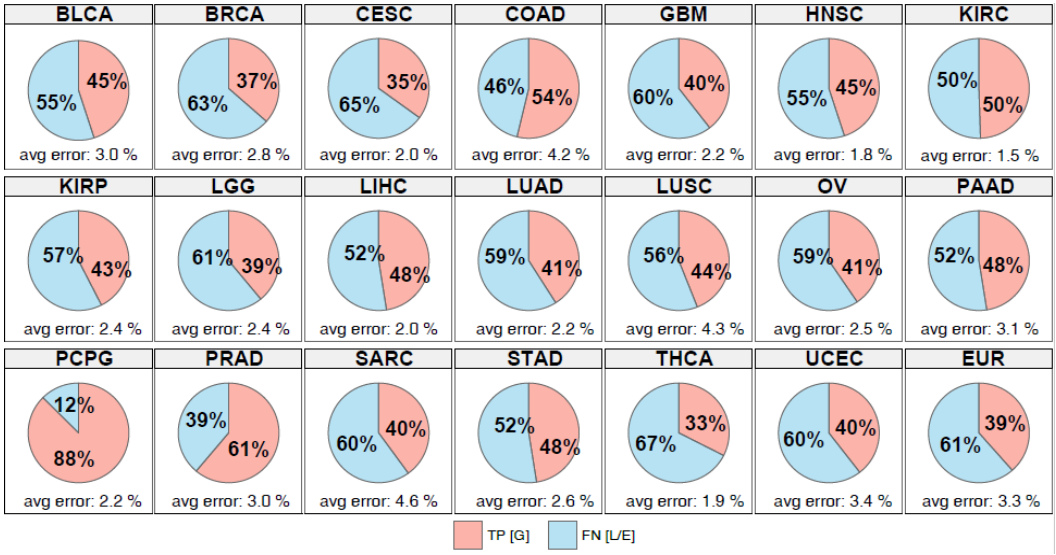

Figure 11. Inherited Genomic factor vs. Environmental/Lifestyle factor

The relative proportion of genomic contribution (G; in pink) and environmental or lifestyle contribution (L/E; in cyan) for acquiring that phenotype in testing set. It is a figure 8 summarization. TP denotes the portion of the true positive (correct prediction) and FN stands for the false negative (wrong prediction). BLCA, bladder urothelial carcinoma; BRCA, breast invasive 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; LGG, low grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine carcinosarcoma.

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. It might not be useful information for an individual for practical purpose. since 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. 11** 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. 12 (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) 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.

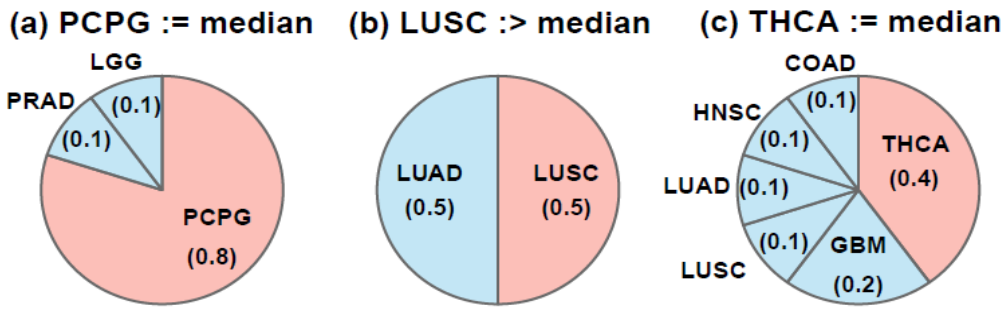

Figure 12. Relative Genomic susceptibility for an individual

Probability distributions of phenotypes among neighboring 410 samples for an individual. Each selected individual in the graph is a median sample of correctly predicted samples when sorting by the number of positive calls. Each slice and given probability may represent the relative genomic susceptibility to the phenotype and its degree. (a) the selected sample has genomic susceptibilities of 80% for PCPG and each 10% for PRAD or LGG. (b) and (c) are the probabilities of obtaining various phenotypes for a member of LUSC or THCA cohorts, respectively.

COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; LGG, low grade glioma; PRAD, Prostate adenocarcinoma; THCA, thyroid carcinoma.

"Multiple Allele Assortment Model" of Inherited Susceptibility for Common Cancers.

Identifying the portion of the cohort of LUSC as PCs (**Fig. 9 (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 (Sample Selection and Quality Control; Describing genomic variation and genomic comparison). Then, we use an unsupervised clustering algorithm, t-distributed stochastic neighbor embedding (t-SNE)[39], to cluster all populations of the PCs (i.e., the individuals in the G groups of the 21 testing cohorts together) (see Inherited Genomic Factor vs. Environmental/Lifestyle Factor). **Fig. 13** shows unsupervised clustering of all of the G group members predicted by our *k*NN 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[42] used in most GWAS (see Comparison of Multiple Allele Assortment Model vs. PolyGenic Model).

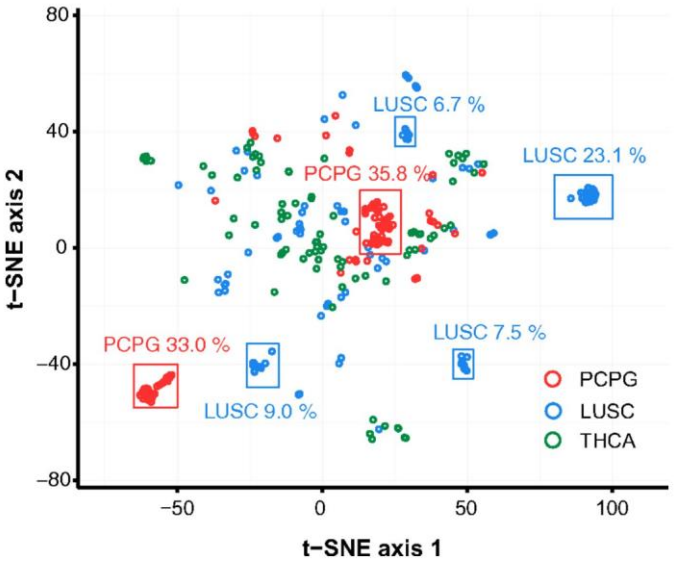

Figure 13. t-SNE analysis.

Unsupervised clustering of correctly predicted samples shown in Fig 8 by t-SNE algorithm[39]. Only 3 cancer types, PCPG, LUSC, and THCA are shown among 21 cancer types used in the clustering for the visualization. Each cluster is a multiple allele assortment model represented by the different assortment of SNP-Syntaxes. For example, PCPG shows two dense and tight cluster which occupy 33.0 % and 35.8 % of total positive called samples, respectively. LUSC shows 4 small dense clusters which account for about 46% of total positive called samples. In the case of THCA, mostly sparse and little clusters are formed. PCPG, Pheochromocytoma and Paraganglioma; LUSC, lung squamous cell carcinoma; THCA, Thyroid carcinoma.

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[19,21]. Of these, BRCA has been one of the most studied cancers by GWAS. For example, a multiplicative polygenic model[42] applied on 76 BRCA-associated SNP genotypes showed 15% prediction accuracy[21] compared with 37% in this study (**Fig. 8**). In general, there are two major differences in the processes and results between GWAS and *k*NN. (*i*) All GWAS have been performed for a binary prediction between cases and controls for each cancer phenotype separately, while in our *k*NN 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 *k*NN 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. 13**).

Correlation Between *k*NN Predictions and Known Observations.

As shown in **Fig. 11,** the PCPG cohort shows the highest accuracy for PC prediction (88%), which corresponds to having the largest G group (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[41]. Germline and somatic mutations of these gene cause approximately 60% of PCPGs. Among these genes, germline mutations were found in all except one. The genetic mutations were classified into two categories[43]. These two categories might be equivalent to the two big clusters of the PCPG population on a t-SNE plot (**Fig 13**). 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 exome 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[44]. 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[33]. 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 10% of the samples as “outliers” (**Fig. 4 (a)** and **Fig 4 (b)**). (**iii**) The PCA-selected samples (90%) were subjected to a second PCA to see if there is any significant correlation between geographical populations and cancer types. Comparison in **Fig. 4 (c)** and **(d)** shows that the

geographical populations segregate reasonably well by the second PCA (**Fig. 4 (c)**), but cohorts of different phenotypes do not cluster but are distributed broadly on the same PCA plot space (**Fig. 4 (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**) We performed HWE test on whole sample level and removed variants which might be deviated from HWE by population structure difference or unreliable genotypes ($p < 1,0e-6$). (v) 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.

To avoid and minimize systemic bias generated from data processing, same genotyping microarray platform and annotation file were used. Indeed, we investigated the genotype reproducibility of whole G1K European samples, and performed quality control steps such as PCA outlier removal.

Despite these efforts, there may be some variations in experimental biases among the datasets of different phenotypes that could influence the training process of the *k*NN model and thus, the prediction accuracy. The curated data used in this study (Sample Selection and Quality Control; Describing genomic variation and

genomic comparison) do not seem to have significant bias as indicated by the relatively small differences between the training and testing accuracies for all the phenotype datasets, except for the cervical squamous cell carcinoma and European white population datasets, which have slightly larger differences than the average (**Fig. 14**).

EUR's bias could be notable. In contrast to the others, EUR has two peaks on the density plot. Half of the EUR are collections of TSI and IBS population. They affect the median point on the PC2. Since our filtering parameter is a simple frequency filtering, some common SNP-Syntaxes of the deviated population could be less than the frequency filtering level. Those could act as artifact and might generate overfitting. To address this issue, we also performed *k*NN with f=100%. the other parameters and sample set were same as optimized ones. While most of the phenotypes except EUR and STAD showed the random level accuracy, ranges from 3% to 9%, Both of EUR and STAD shows more than 20% accuracy. thus, our method could not be guaranteed when there is some deviated population. It should be controlled more sophisticatedly in the future.

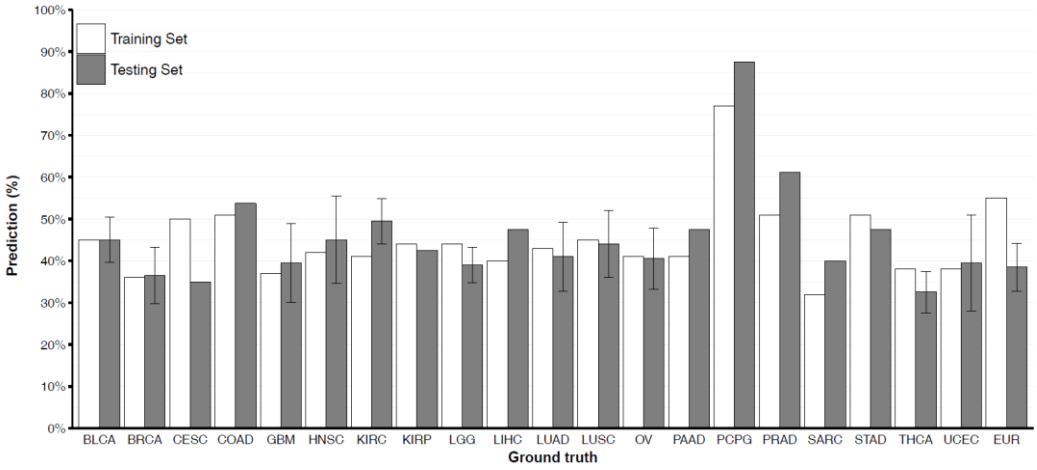

Figure 14. Training versus Testing result.

Prediction performance comparison between the training set and the testing sets. The x-axis is the ground truth, and the y-axis is the percentage of the correctly predicted by the genomic contribution. For each phenotype, "I bar" represents 2 x standard deviation for the multiple testing sets. We only drew "I bar" when the testing sets are at least 3 since there are cohort size limitations. BLCA, bladder urothelial carcinoma; BRCA, breast invasive 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; LGG, low grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, uterine carcinosarcoma.

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. 11** and **Fig. 8**). 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. 8**), as expected for only 100 samples per training set. They are expected to improve as the sample size 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.

References

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, *et al*. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer* **49,** 1374–1403 (2013).

2. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. *CA: a cancer journal for clinicians* **67,** 7–30 (2017).

3. World Health Organization. cancer. (2018). Available at: http://www.who.int/mediacentre/factsheets/fs297/en/.

4. Blumen H, Fitch K, Polkus V. Comparison of Treatment Costs for Breast Cancer, by Tumor Stage and Type of Service. *American health & drug benefits* **9,** 23–32 (2016).

5. Howlader NKM, Noone AM. SEER cancer statistics review (csr) 1975-2014 updated april 2, 2018. *Surveillance, Epidemiology, and End Results Program*

6. Metcalfe KA, Poll A, Royer R, Llacuachaqui M, Tulman A, Sun P, *et al*. Screening for founder mutations in BRCA1 and BRCA2 in unselected Jewish women. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **28,** 387–391 (2010).

7. Gabai-Kapara E, Lahad A, Kaufman B, Friedman E, Segev S, Renbaum P, *et al.* Population-based screening for breast and ovarian cancer risk due to BRCA1 and BRCA2. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 14205–14210 (2014).

8. Manchanda R, Loggenberg K, Sanderson S, Burnell M, Wardle J, Gessler S, *et al.* Population testing for cancer predisposing BRCA1/BRCA2 mutations in the Ashkenazi-Jewish community: a randomized controlled trial. *Journal of the National Cancer Institute* **107,** 379 (2015).

9. Carol E, DeSantis AGS, Jiemin Ma. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA Cancer J Clin* **67,** 439–448 (2017).

10. Hecht SS. Environmental tobacco smoke and lung cancer: the emerging role of carcinogen biomarkers and molecular epidemiology. *Journal of the National Cancer Institute* **86,** 1369–1370 (1994).

11. Seitz HK, Stickel F. Molecular mechanisms of alcohol-mediated carcinogenesis. *Nature reviews Cancer* **7,** 599–612 (2007).

12. Basen-Engquist K, Chang M. Obesity and cancer risk: recent review and evidence. *Current oncology reports* **13,** 71–76 (2011).

13. Joseph P. Mechanisms of cadmium carcinogenesis. *Toxicology and applied pharmacology* **238,** 272–279 (2009).

14. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *Journal of the American Medical Association* **315,** 68–76 (2016).

15. Singletary SE. Rating the risk factors for breast cancer. *Annals of surgery* **237,** 474–482 (2003).

16. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

17. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45,** 1113–1120 (2013).

18. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

19. Manolio TA. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine* **363,** 166–176 (2010).

20. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45,** D896–D901 (2017).

21. Bahcall O. Common variation and heritability estimates for breast, ovarian and prostate cancers. *Nature Genetics* (2013). doi:10.1038/ngicogs.1

22. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).

23. Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. *Nature genetics* **49,** 325–331 (2017).

24. Mohri M, Rostamizadeh A, Talwalkar. Introduction. *A. Foundations of machine learning* (MIT Press, Cambridge, MA) pp 1-9 (2012).

25. Kim M, Kim S-H. Empirical prediction of genomic susceptibilities for multiple cancer classes. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 1921–1926 (2014).

26. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics* **5,** e1000678 (2009).

27. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine* **4,** 218 (2016).

28. Kim B, Kim S-H. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. *Proceedings of the National Academy of Sciences of the United States of America* 115, 1322-1327 (2018).

29. Grossman RL, Health AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, *et al.* Toward a Shared Vision for Cancer Genomic Data. *The New England journal of medicine* **375,** 1109–1112 (2016).

30. Affymetrix. Support by product. *Thermo Fisher Scientific.* Available at http://www.affymetrix.com/support/technical/byproduct.affx?product=genomewidesnp_6

31. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic epidemiology* **34,** 591–602 (2010).

32. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

33. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).

34. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, *et al.* Quality Control Procedures for Genome-Wide Association Studies. **177,** *Current Protocol Human Genetics* **68**:1–18.

35. Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* **37,** 145–151 (1991).

36. Sims GE, Jun SR., Wu GA, Kim S-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 2677–2682 (2009).

37. Majnik M, Bosnić, Z. ROC analysis of classifiers in machine learning: A survey. *Intelligent data Analysis* 531–558 (2013).

38. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).

39. Maaten LJP, Hinton GE. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9,** 2579-2605 (2008).

40. Donaldson J. *Tsne: T-distributed stochastic neighbor embedding for r (t-sne)* (2016). Available at https://cran.r-project.org/web/packages/tsne

41. Lalloo F. Diagnosis and Management of Hereditary Phaeochromocytoma and Paraganglioma. *Recent results in cancer research.* **205,** 105–124 (2016).

42. Witte JS, Hofmann TJ. Polygenic modeling of genome-wide association studies: an application to prostate and breast cancer. *OMICS* **15**:393-398

43. Pillai S, Gopalan, V, Smith RA, Lam, AK-Y. Updates on the genetics and the clinical impacts on phaeochromocytoma and paraganglioma in the new era. *Critical reviews in oncology/hematology* **100,** 190–208 (2016).

44. National Cancer Institute. 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.

Abstract in Korean

기계학습을 통한 20 종의 주요 암에 대한

상대적인 유전 감수성 예측

연세대학교

융합오믹스 의생명과학과

김병주

암의 발병은 유전적 요인, 환경 그리고 생활습관과 같은 요인들의 복잡한 상호작용을 통해서 일어납니다. 그리고 그 결과는 환자와 그를 부양하는 가족들에게 경제적, 정신적 그리고 신체적인 부담으로 나타나게 됩니다. 이러한 부담을 피하거나 혹은 최소화시키기 위한 가장 좋은 방법은 암을 예방하거나 조기에 발견, 치료하는 것입니다. 하지만 이를 위해서는 암에 걸릴 확률을 예측하는 모델을 만드는 것이 요구됩니다. 최근의 진보된 기술로 인해서 게놈 다양성을 단일 뉴클레오타이드 수준에서 정량적, 정성적인 측정을 가능하게 되었고, 이러한 모델을 만들기 위해 필요한 수많은 데이터들은 현재 허가를 받으면 사용 가능하게 되었습니다. 우리는 총 5,919 명의 백인들의 유전 데이터를 The Cancer Genome Atlas 와 1000 genome project 로부터 얻어서 지도학습 방법 중 하나인 *k*-Nearest Neighbor 을 적용하여 20 종류의 암과 건강한 대조군 형질을 분류하는 연구를 수행하였습니다. 그 결과, 유전정보를 이용하여 어떤 암인지에 따라 33% 에서 88%까지 암의 종류를 예측할 수 있었습니다. 예측에 실패한 12%에서 67%는 환경요인이나 생활습관 등에 의해 암이 발병된 것으로

여겨집니다. *k*NN 모델의 각 개인의 이웃하고 있는 형질이 차지하고 있는 비율 또한 암의 예방 및 조기 치료를 돕는 예측 가치를 지닌 유용한 유전 감수성 정보로 활용될 수 있을 것입니다. 또한 유전적으로 질병에 대한 감수성을 가지게 하는 Multiple Allele Assortment 모델에 대해서도 이야기를 했는데, 많은 다양한 genomic element 조합을 이루어서 특정한 특정 질병에 대한 질병 감수성을 형성한다는 모델입니다.

핵심 단어 : 암, 유전 감수성, 기계학습, k-NN

Publication List

Kim, B. & Kim, S.-H. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. *Proceedings of the National Academy of Sciences of the United States of America* 115, 1322-1327 (2018)