Machine-Learning on Prediction of Inherited Genomic Susceptibility for 20 Major Cancers

Sung-Hou Kim
University of California
Berkeley, CA

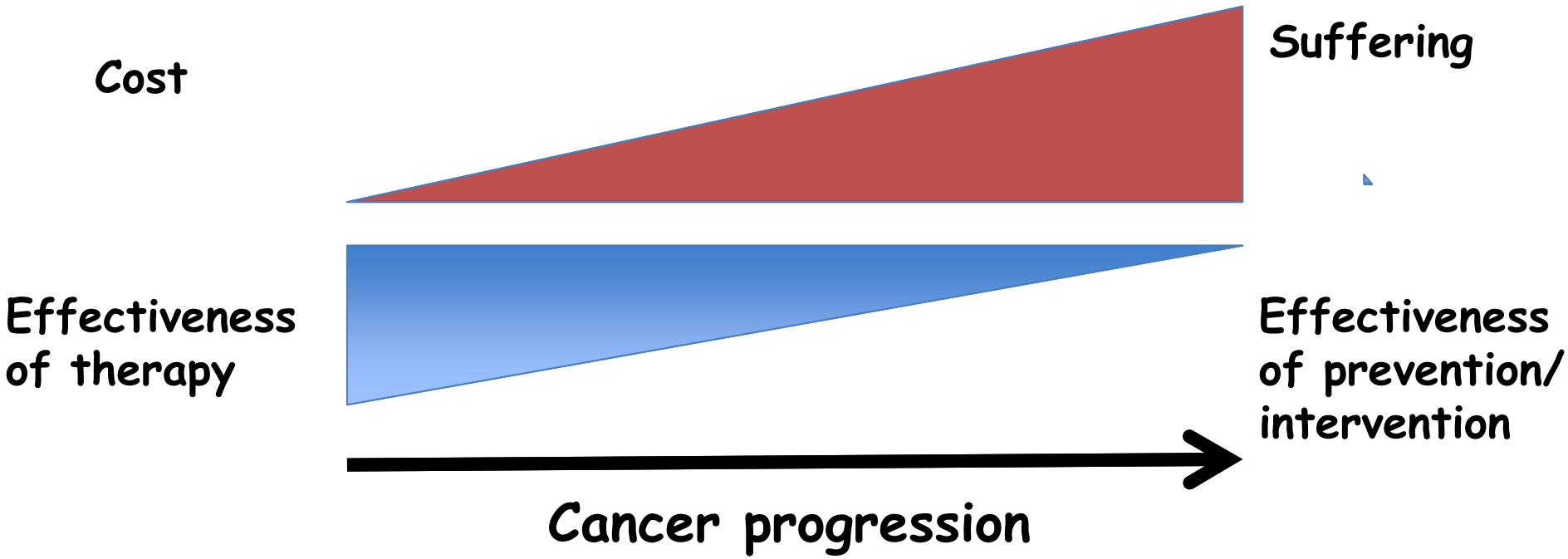
Global Bio Conference 2017
MFDS, Seoul, Korea
June 28 , 2017

Cancer "Facts"

- ~100 human cancer types are known
- 5 – 10% ("rare" cancers) are due to one or a few inherited genomic elements
- 90 – 95% ("common" cancers) are due to a large number of inherited and acquired genomic elements
- Cancer triad of genomic susceptibility, environment and life style

A. Cost and Effectiveness of Cancer Treatment

B. Psychological, Physical and Financial Suffering



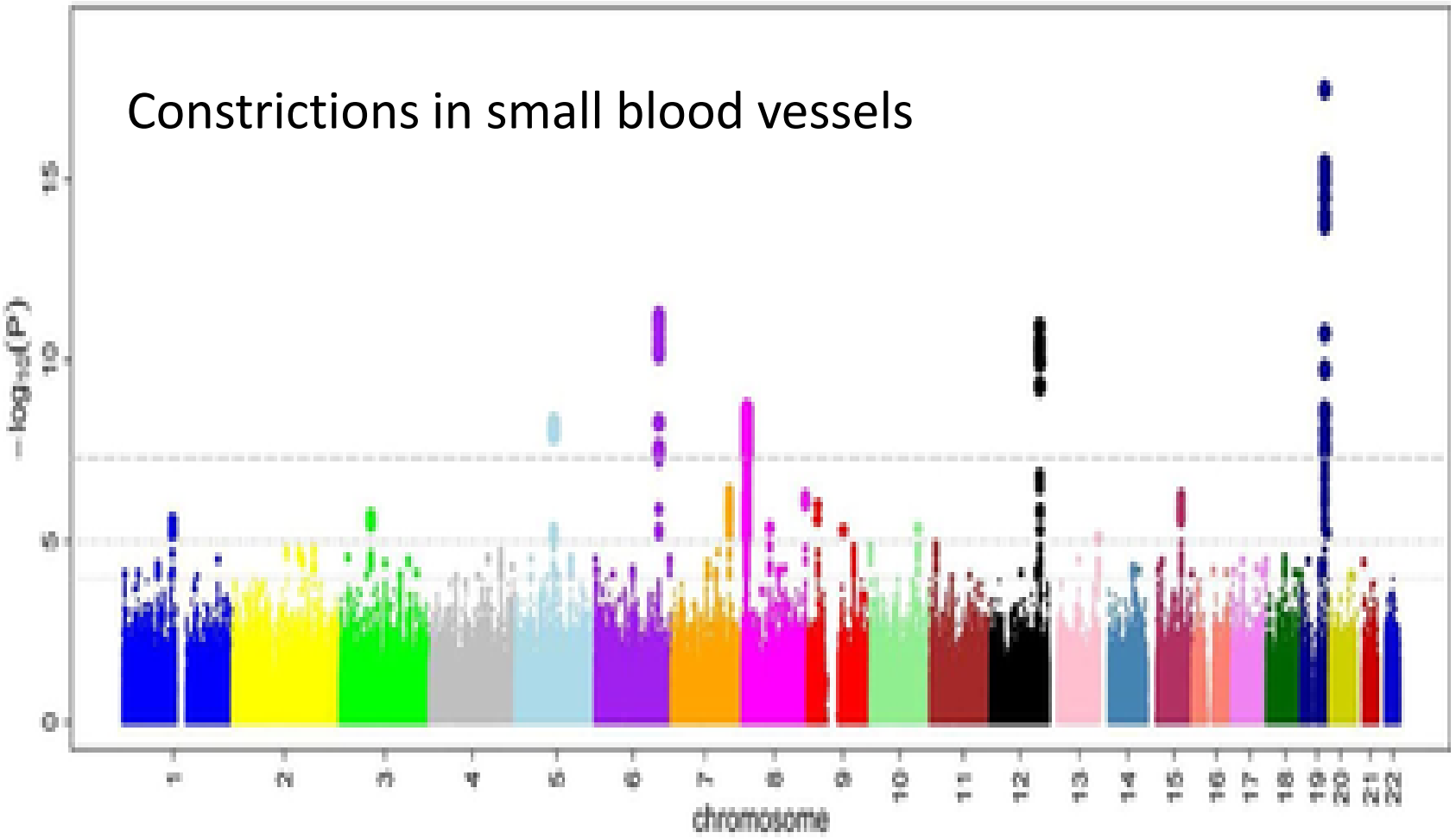
Best strategy: Prevention and Early Intervention

Objective

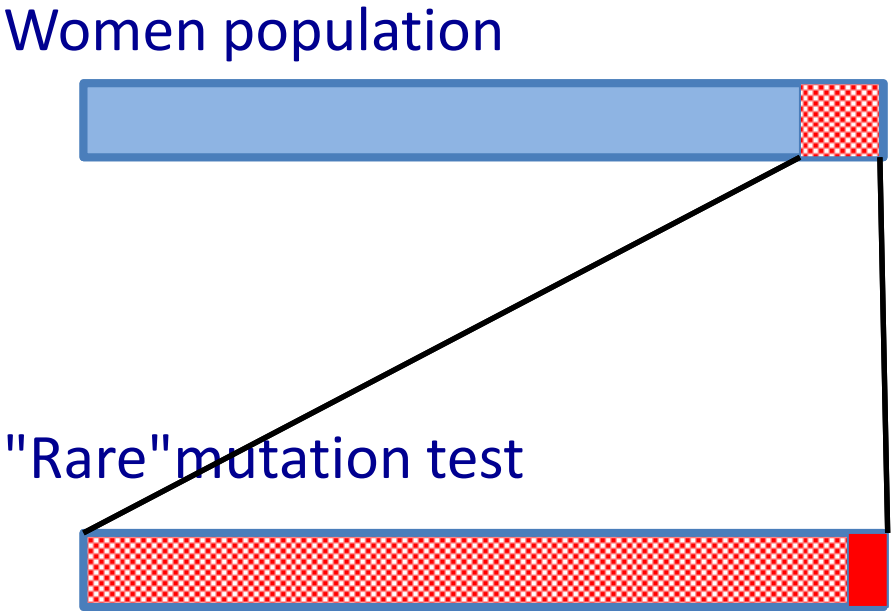
Predict inherited genomic susceptibility
to Cancer
from
Personal Genomic variations of
"Germ-line" (un-transformed) cells

~ 4-5 M variation loci per genome;
> 90% of variations are SNPs

Current Method:
Genome-wide Association Study (GWAS)



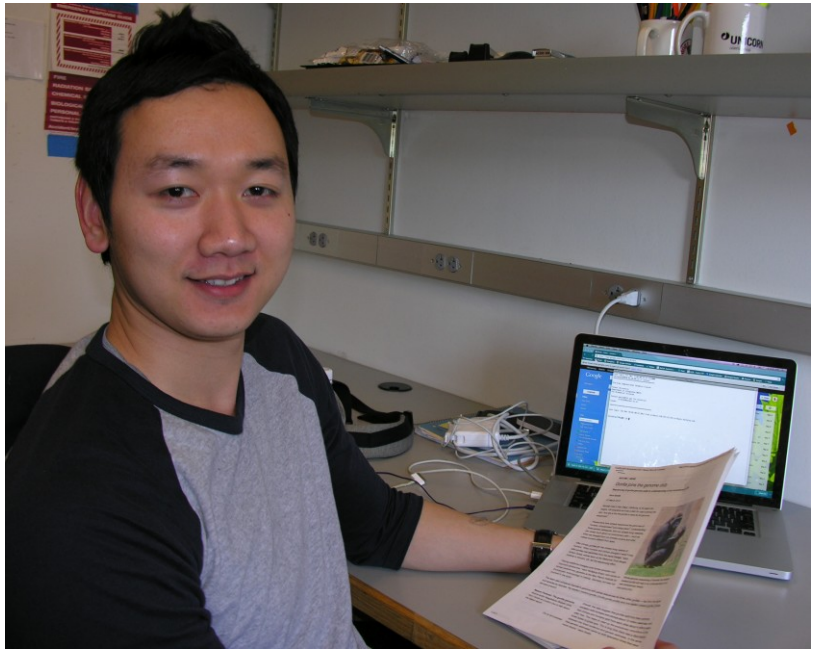
Current Prediction Status: Breast Cancer in USA



13% of women have cumulative life-time risk

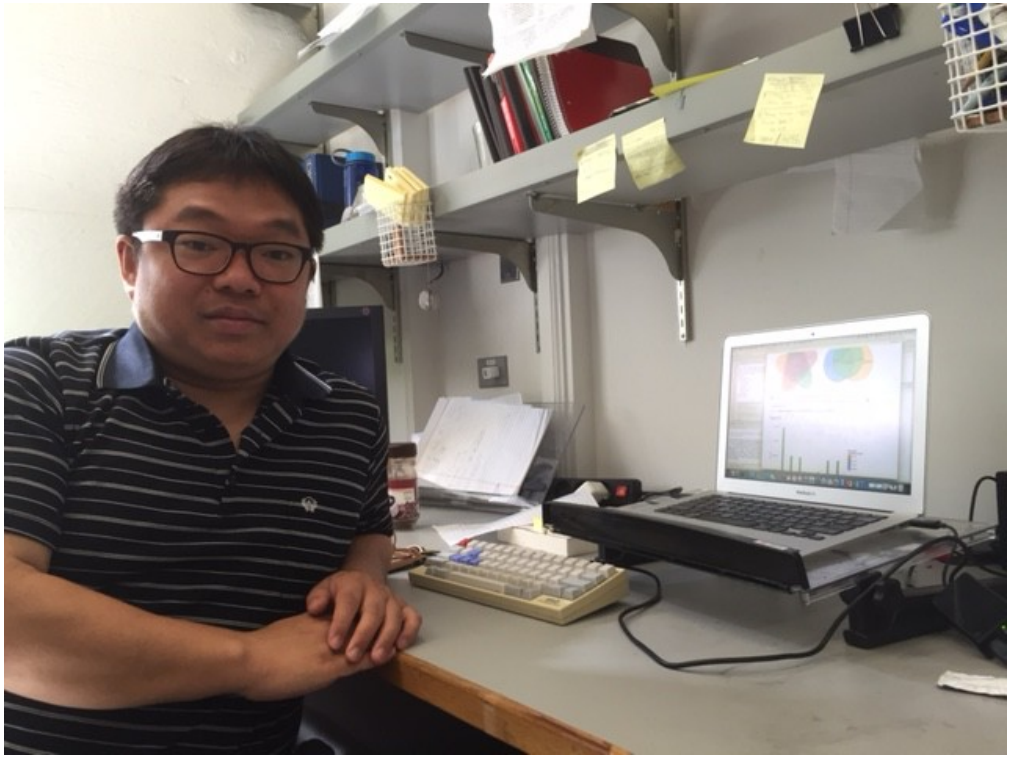
5 – 10% of cases have BRCA1/2 or other "rare" mutations*
(<80% high risk; >20% no risk)

"Missing Heritability" of GWAS



Columbia U. NY, USA
Yonsei U., Korea
UC Davis, USA

UC Berkeley, USA
Yonsei U. Korea



Financial Supports:

WCU grant, Ministry of Education, Science and Technology, Korea;
Gift grant to the University of California, Berkeley, CA, USA

Machine-Learning (ML) Approach:
A fundamentally different from GWAS

Descriptors and Analysis methods

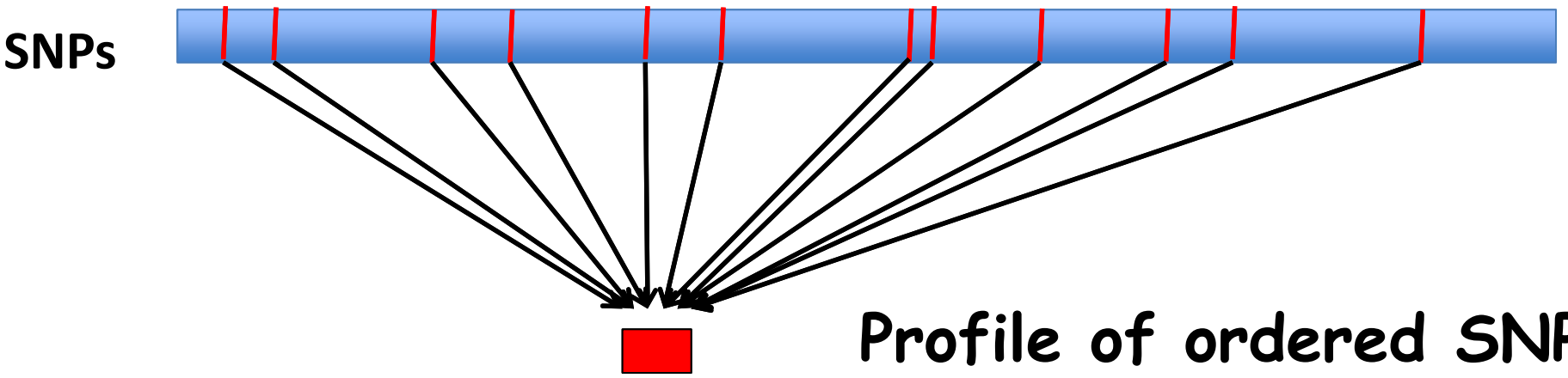
A. **Descriptors for whole genome SNPs as:**

A "book" without spaces in computational analysis of natural language or plagiarism

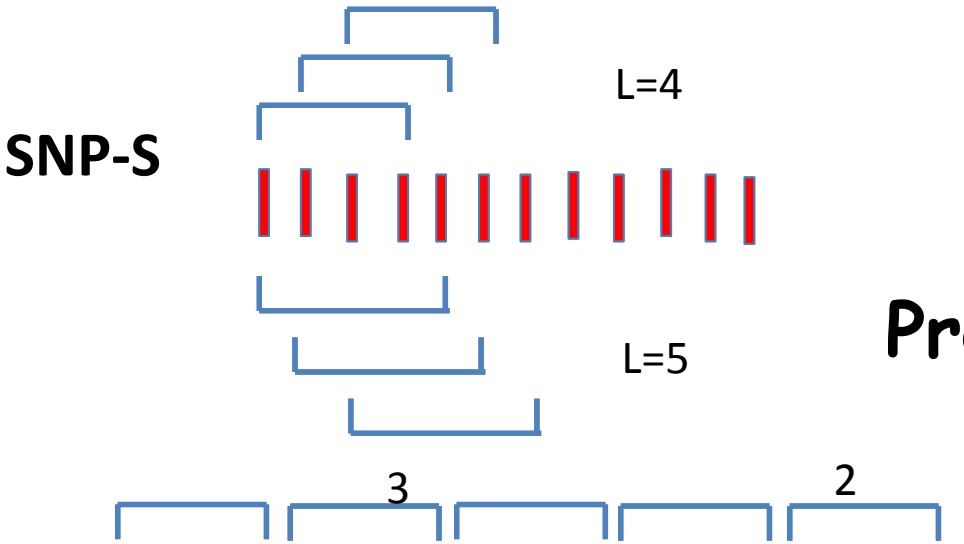
B. **Analysis methods:**

Supervised machine learning algorithms used in recognition of complex system, such as images

A. Descriptors: SNP and SNP-S(yntax)



Profile of ordered SNPs
("letters")



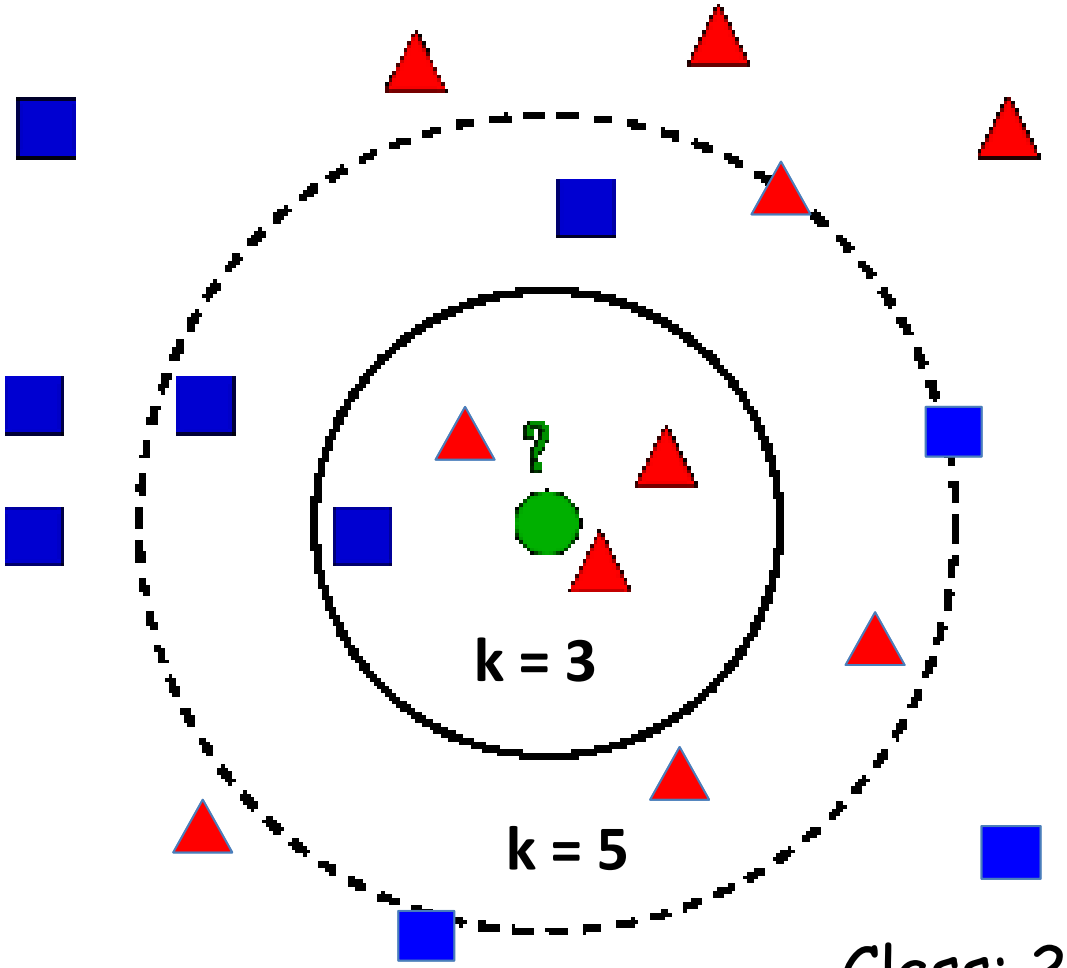
Profile of ordered SNP-Ss
("words") frequency

B. Analysis Algorithms:
Supervised Machine-Learning

1. *k* Nearest Neighbor algorithm (*k*NN)
2. Support Vector Machine algorithm (SVM)
3. Linear regression
4. Logistic regression
5. Naïve Bayes
6. Neural networks (AI)

etc.

Method: *k* Nearest-Neighbor Algorithm
(e.g., two classes)



Class: 21 phenotypes
Samples: ~6,000
Dimension: ~850K/prs

Genomic Variation Space

20 Major Cancers + 1 "Control" (cohort size >= 180)

BLCA:	Bladder urothelial carcinoma.
BRCA:	Breast invasive carcinoma
CESC:	Cervical squamous cell carcinoma and endocervical adenocarcinoma
COAD:	Colon adenocarcinoma
GBM:	Glioblastoma multiforme
HNSC:	Head and neck squamous cell carcinoma
KIRC:	Kidney renal clear cell carcinoma
KIRP:	Kidney renal papillary cell carcinoma
LGG:	Brain lower 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 (neuroendocrine tissues)
PRAD:	Prostate adenocarcinoma
SARC:	Sarcoma (mesenchymal cells: bone, cartilage, vascular, hematopoietic tissues)
STAD:	Stomach adenocarcinoma
THCA:	Thyroid carcinoma
UCEC:	Uterine Carcinosarcoma

TCGA

EUR:	"Control"

G1K

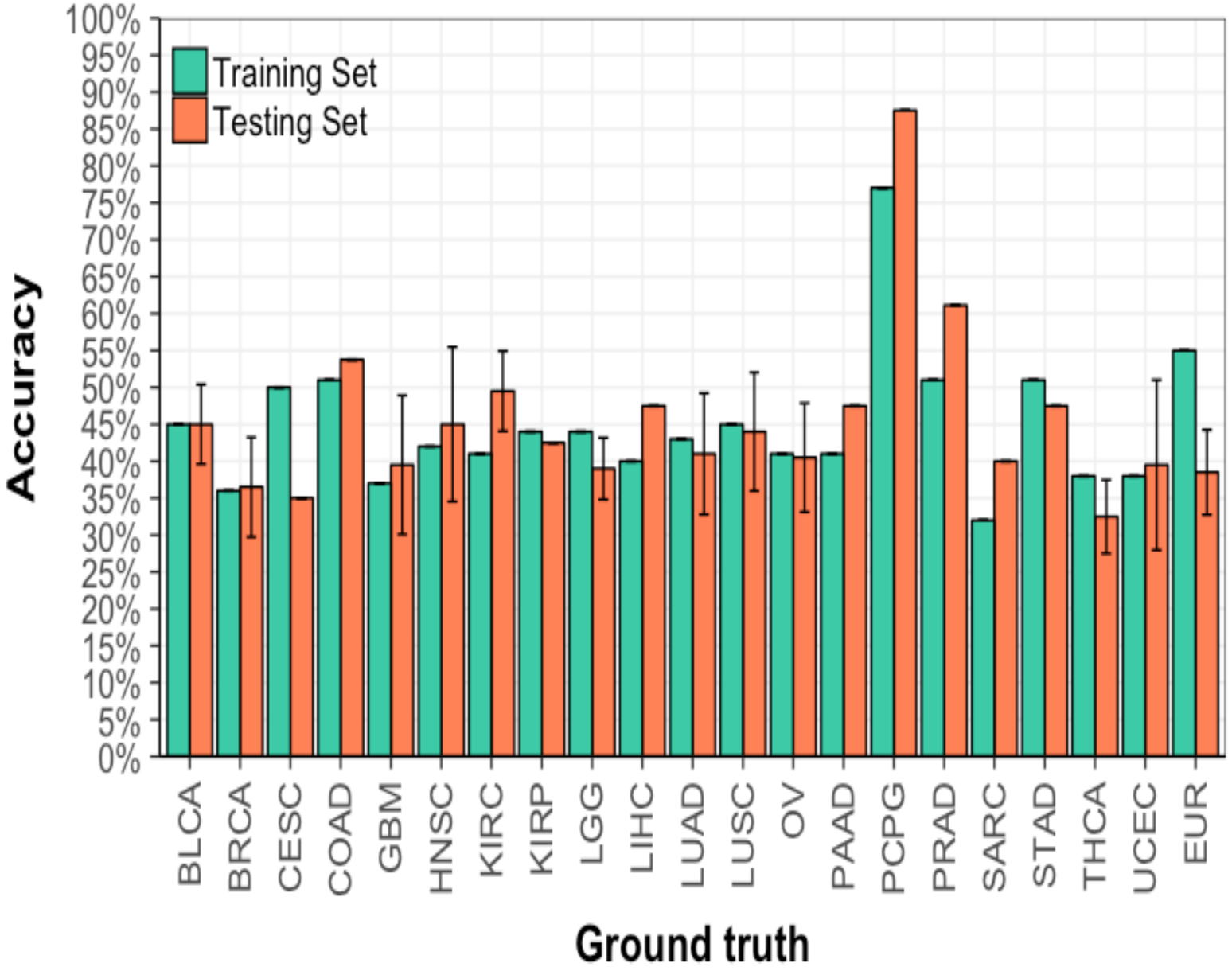
Genomic Data from Publicly Available Databases

Number of individuals : 9,704 → 5.919 from "White" ethnic group

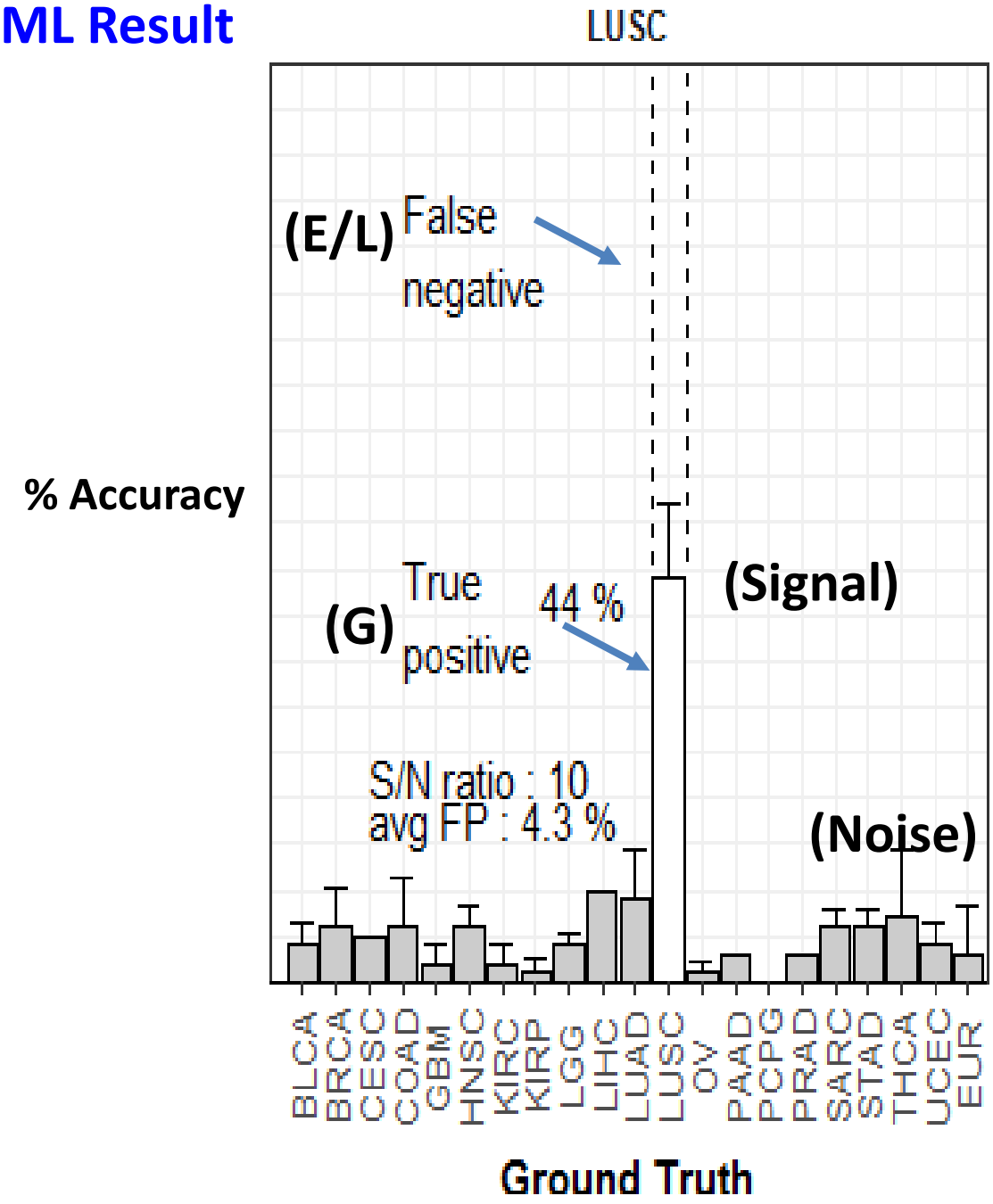
Number of SNP genotypes per person: 906,600 → 818,278 after QC and >1% disconcordance between G1K/BJKcalls

Database: The Cancer Genome Atlas (TCGA)
The 1000 Genome Project (G1K)

Validation of Mehtod

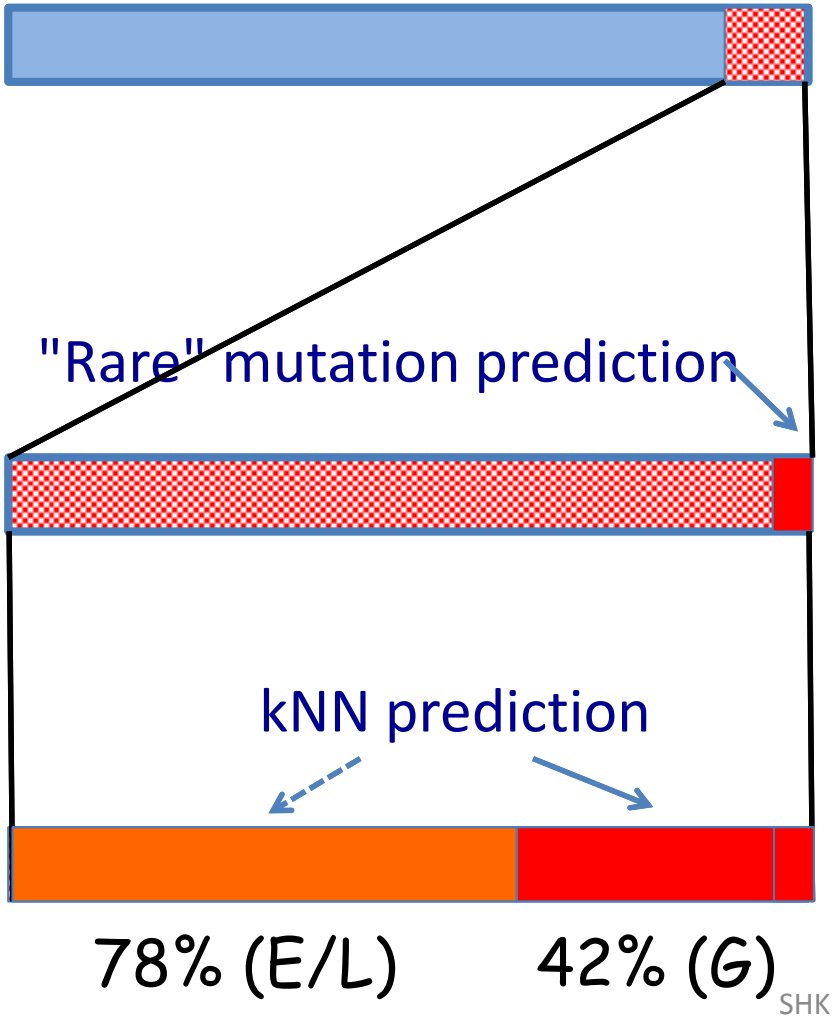


kNN ML Result



ML Example: Breast Cancer Prediction

Women population (US)



13% of women have cumulative Life time risk

5 – 10% of cases have BRCA1/2 or other "rare" mutations*
(<80% high risk; >20% no risk)

SHK 2017

* Tumor suppressor genes related to cell growth, DNA repair, or apoptosis

Summary

- Power of **Machine-Learning** algorithms for analysis of vast genomic data
- Importance of the genomic differences among **ethnic groups**
- **Practically** useful health information for individuals, health professionals, health decisions policy