Identifying Carcinogenic Multi-hit Combinations using Weighted Set Cover Algorithm

Sajal Dash[1], Nick Kinney[2], Robin Varghese[2], Harold Garner[2], Wu-chun Feng[1], and Ramu Anandakrishnan[2]
[1]Department of Computer Science, Virginia Tech, [2] Biomedical Sciences, Edward Via College of Osteopathic Medicine
Blacksburg, VA, 24060

Abstract

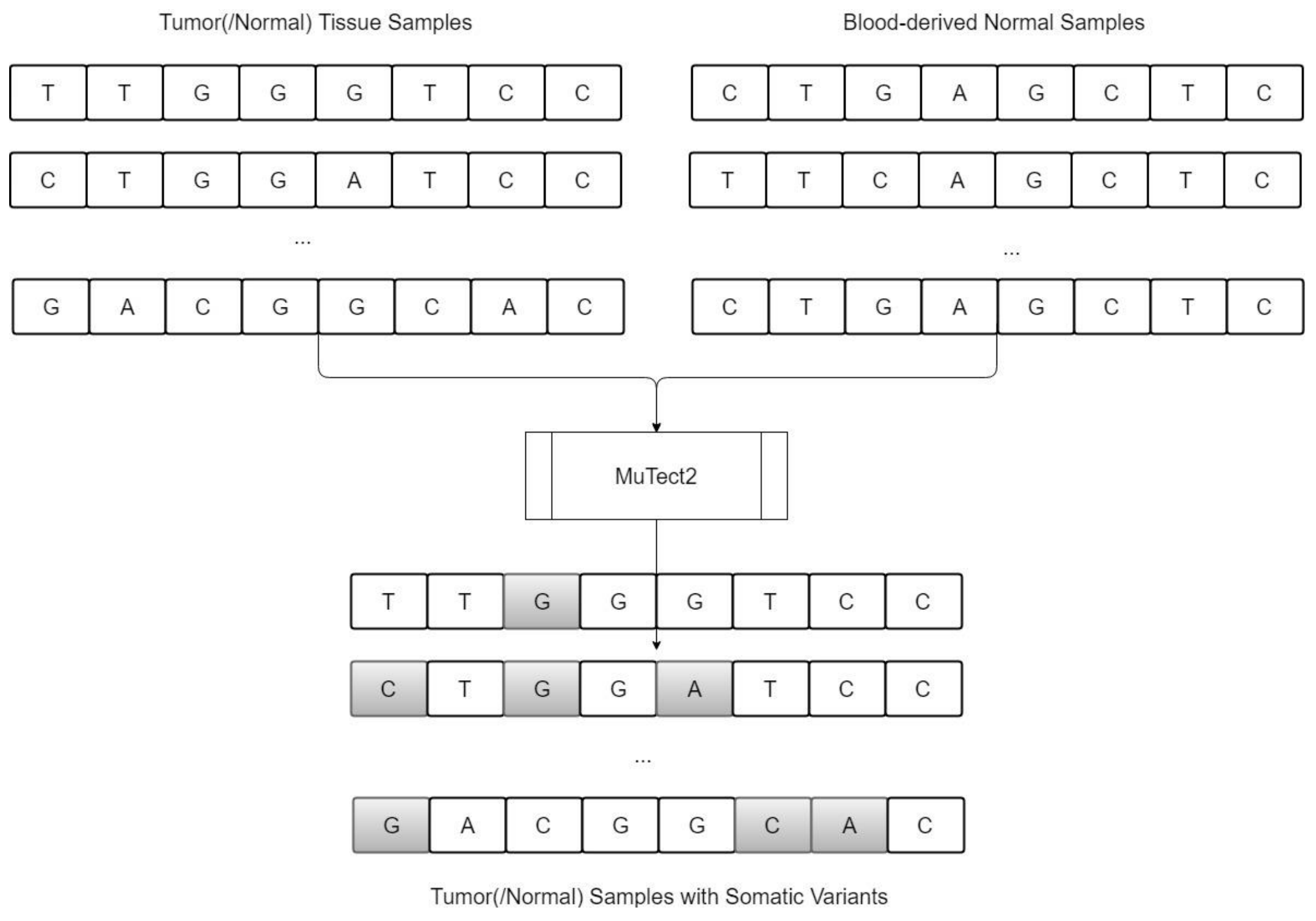
Disruptions in certain molecular pathways due to combinations of genetic mutations (hits) are known to cause cancer. Although different combinations of just two to eight hits are estimated to be required for tumorigenesis, the specific combinations of mutations responsible for the vast majority of cancers have not been identified. Due to a large number of mutations present in tumor cells, experimentally identifying these combinations is not possible except in very rare cases. Individually these driver mutations may increase the risk of cancer; however, they generally do not result in carcinogenesis, without specific additional mutations. We map the problem to weighted set cover problem(WSC) and using an approximate algorithm for WSC we identify sets of 2-hit combinations that can distinguish between cancer and normal samples with more than 90% specificity and sensitivity.

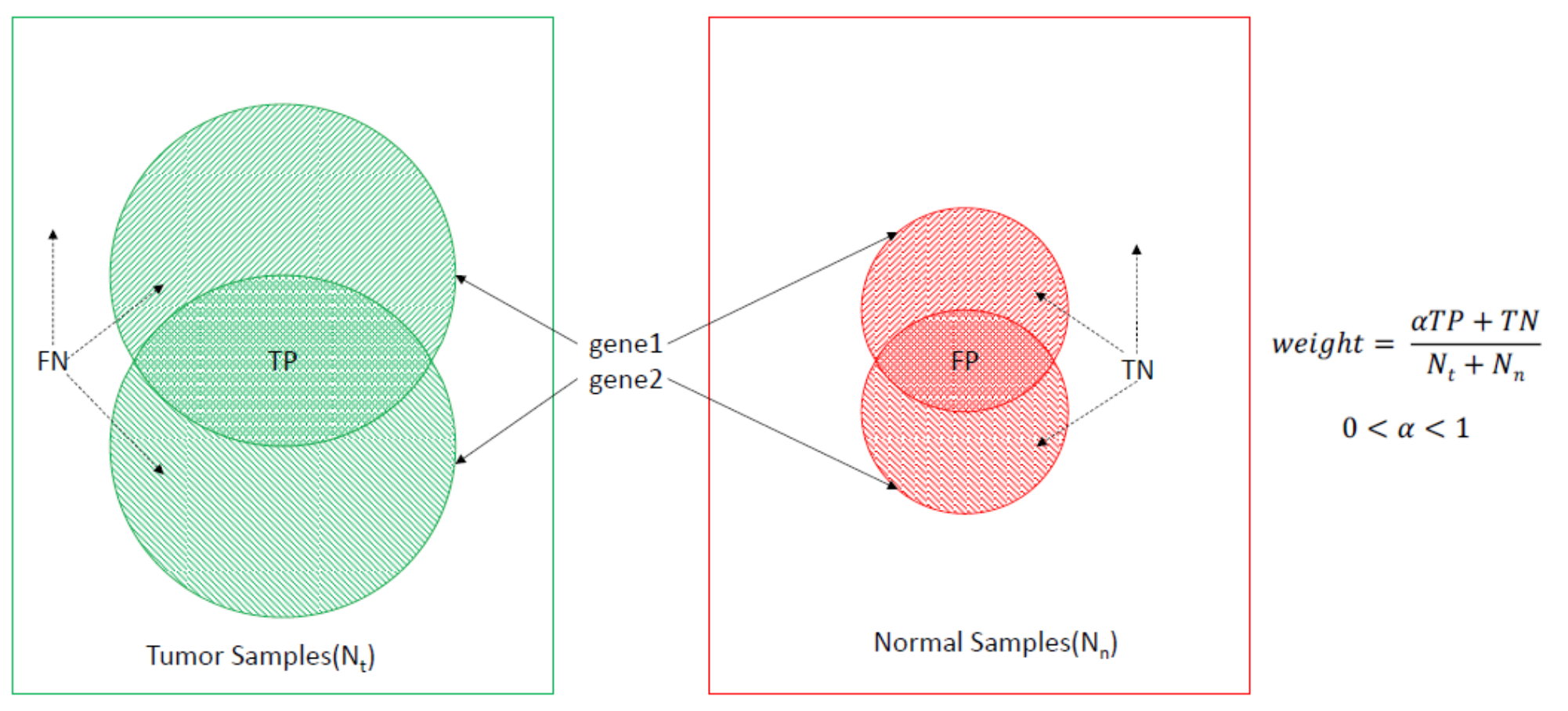
Methods

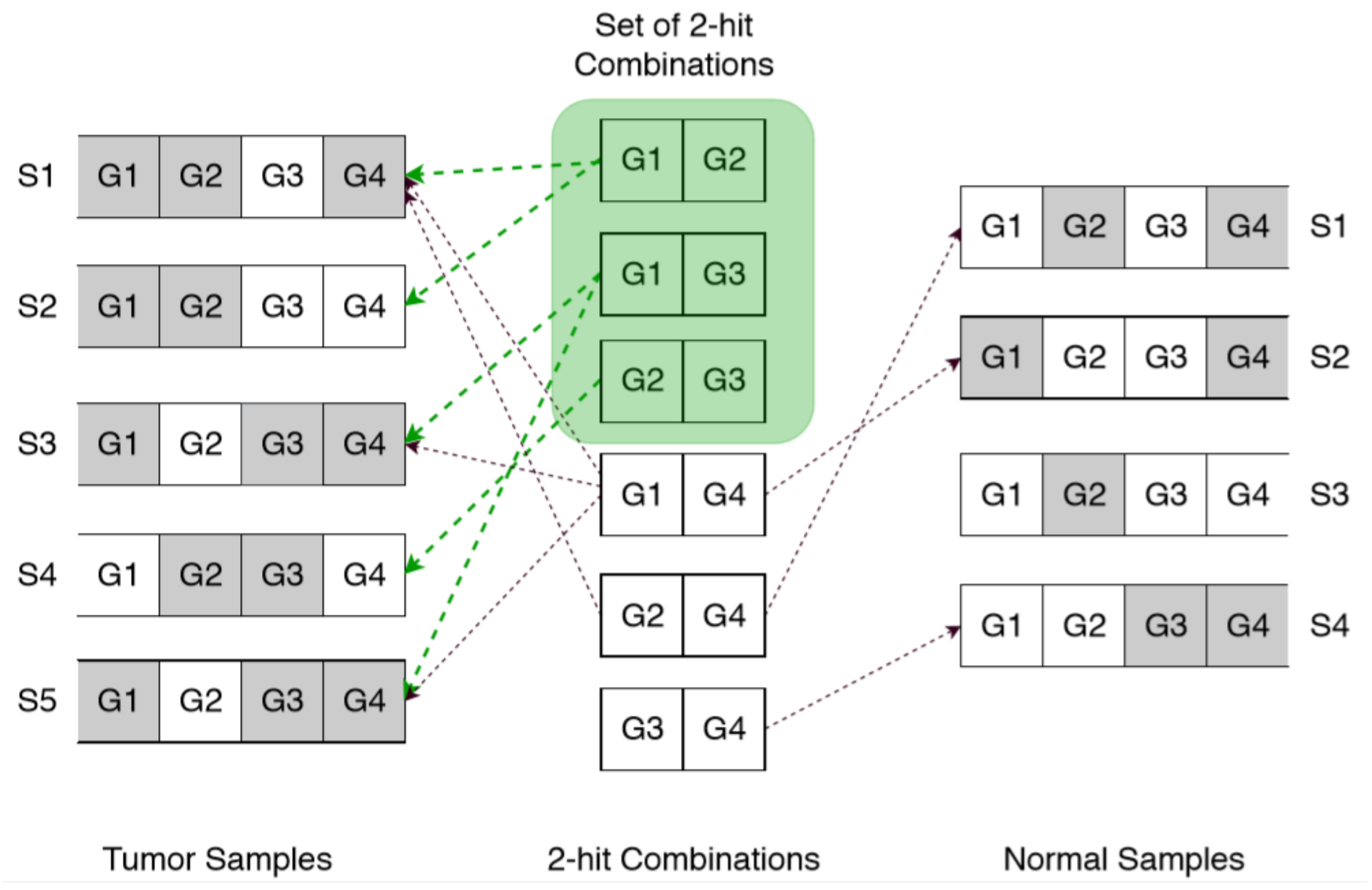
1. Identify somatic variants

2. Assign weight to combinations
weight(combination) = $\frac{\alpha TP+TN}{N_T+N_N}$

3. Map to WSC and solve using approximation algorithm

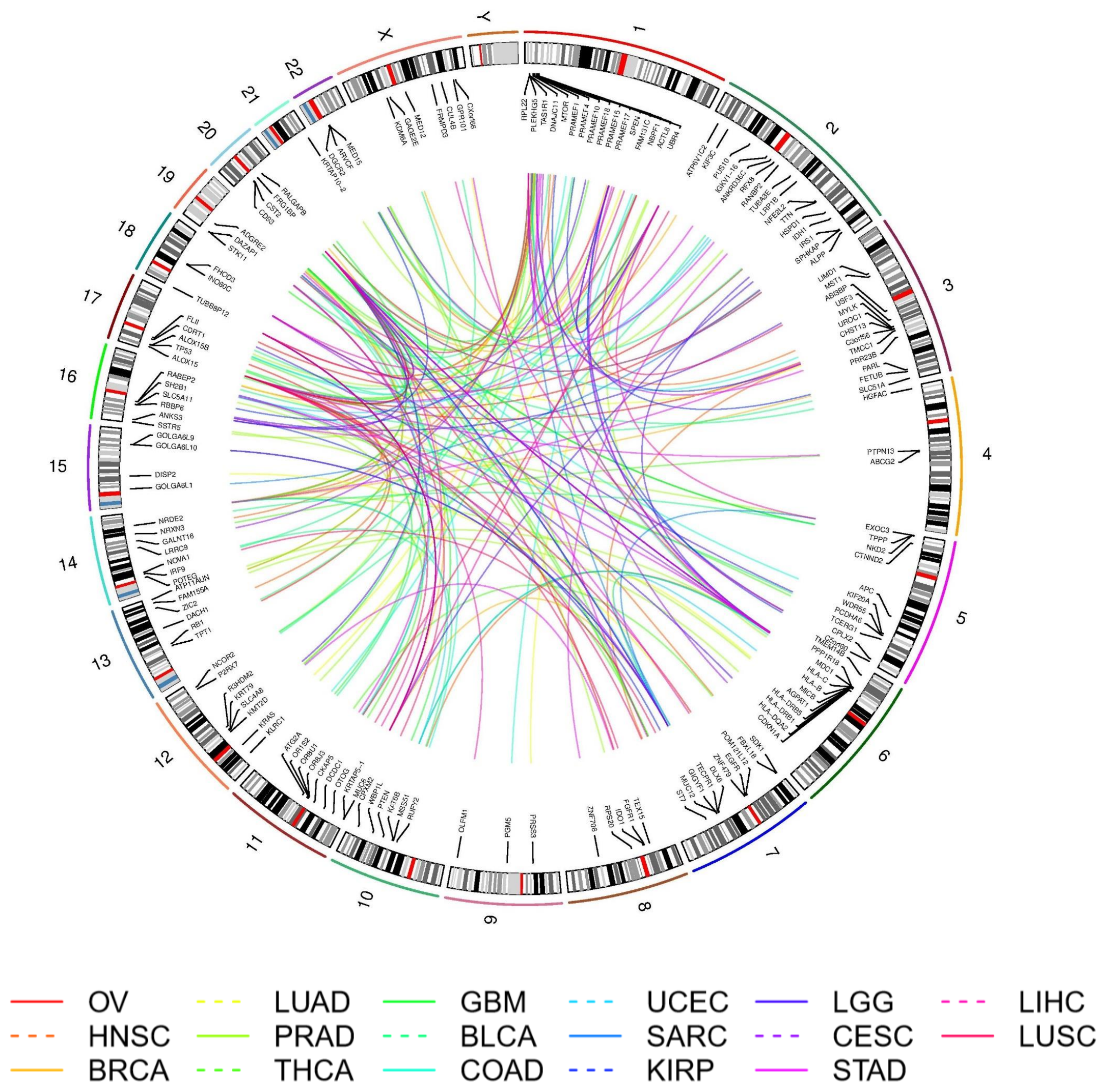




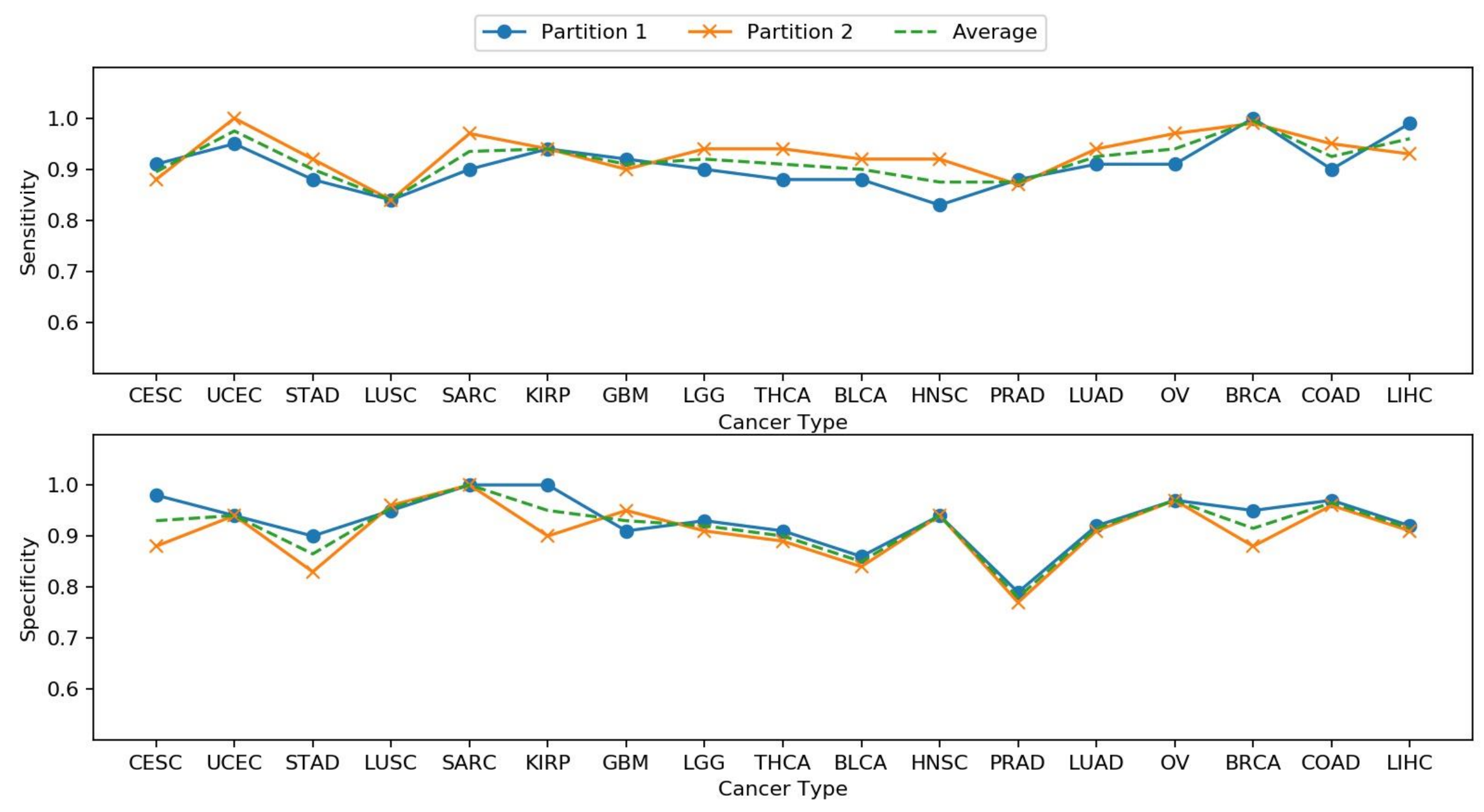


Results

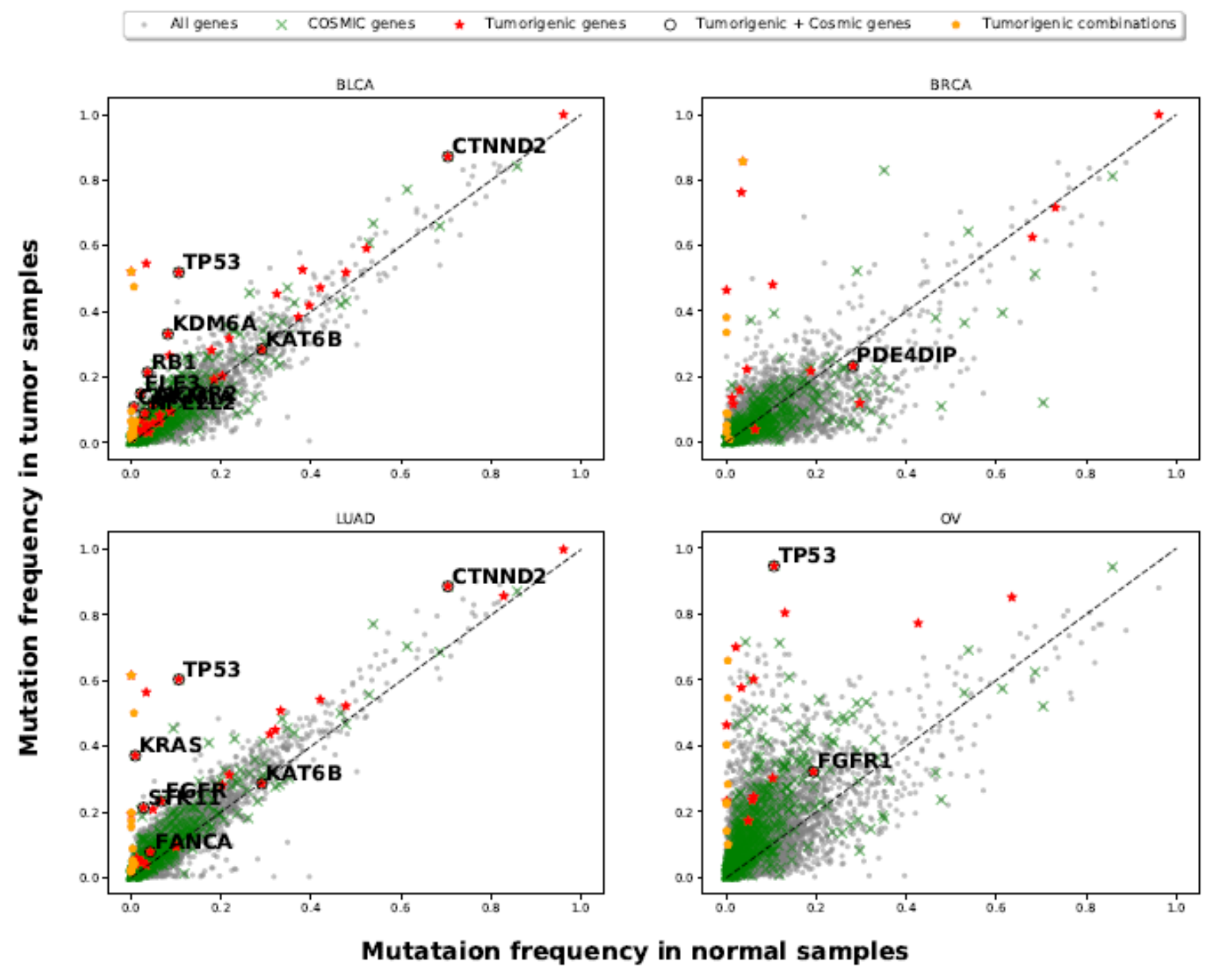
1. We have identified 197 2-hit Combinations for 17 Cancer types



2. These combinations can differentiate cancer samples from normal samples with more than 90% specificity and sensitivity. These performance Is robust across different training and test sets.



3. Identified genes and combinations are more frequently mutated in tumor samples than in normal samples



Beyond 2-hits: GPU Acceleration

I. For 2-8 hits, number of possible combinations, M = 6 x 10^{29} . Number of possible sets is 2^M . Searching through these vast solution space is impractical using sequential programming.

II. We are developing a GPU-parallel algorithm to speedup intermediate steps of WSC algorithm. Initially, we parallelized finding best combination using parallel reduction.

III. We have mapped some sub-problems as solving set operations using sparse linear algebra. We are designing efficient data structures to ensure regular memory access while reading unstructured genomics data.

SyNeRGy
synergy.cs.vt.edu

sajal@cs.vt.edu

VIRGINIA TECH