Simulated Cancer Mutational Signatures to Classify Cancer Types using an Alignment-free Machine Learning-based Approach

David Chen[1], Gurjit S. Randhawa[2], Maximillian P. M. Soltysiak[1], Lila Kari[2], Shiva M. Singh[1], Kathleen A. Hill[1]

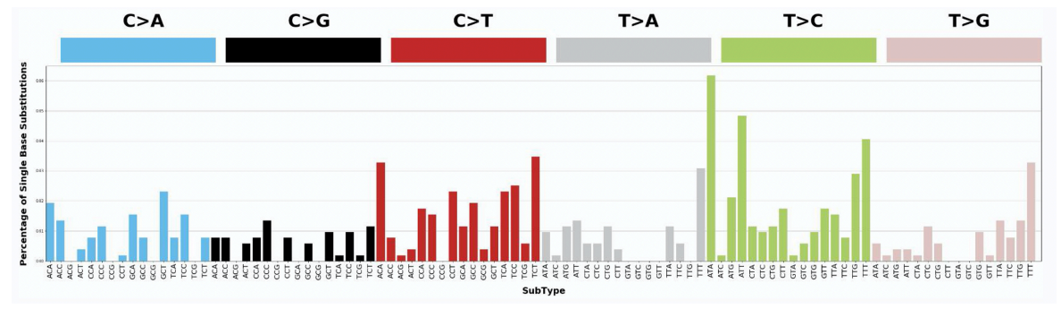
[1] Department of Biology, University of Western Ontario
[2] School of Computer Science, University of Waterloo

khill22@uwo.ca

Introduction

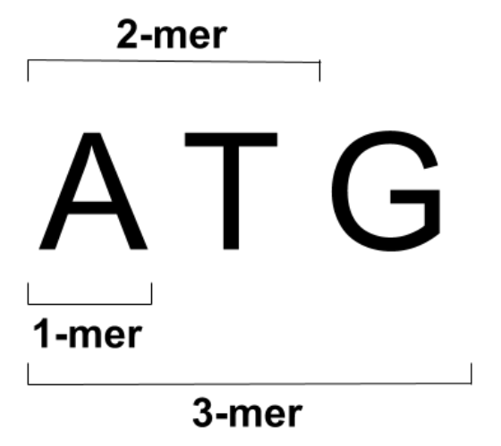
Research Goal: To simulate imposed cancer mutational signatures on a human genome sequence *in silico* and determine the accuracy that can be achieved for classification of cancer subtypes using the simulated sequences and a novel supervised machine learning classification tool.

Mutational processes in cancer genomes generate characteristic single and double base substitutions and indels that comprise multiple superimposed[1] and species-type specific[2] mutational signatures

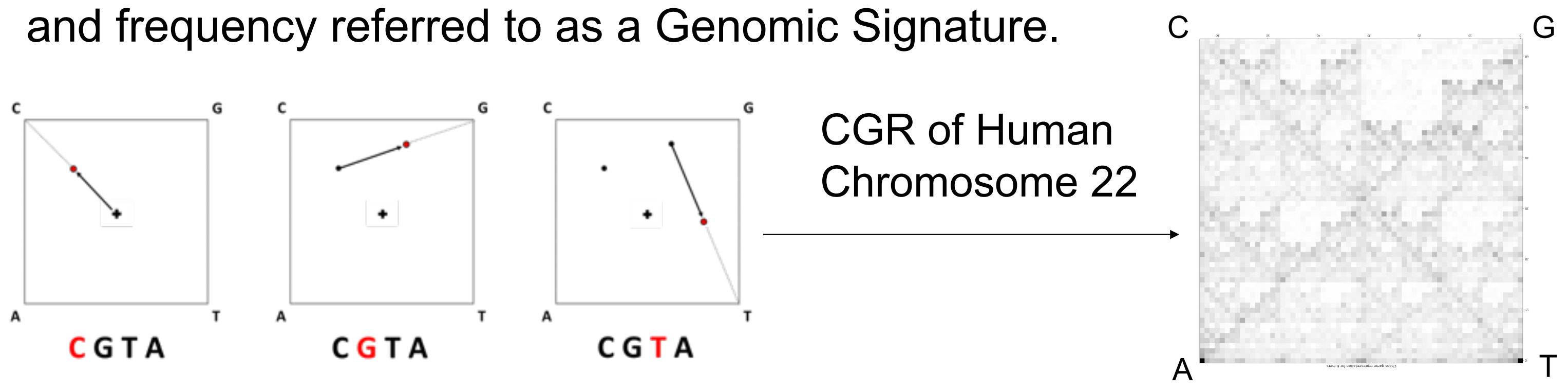


[1] Alexandrov et al. *Nature.* 2020 Feb. 2; 578(7793):94-101.
[2] Kari et al. PLoS One. 2015 May 22;10(5):e0119815.

K-mer is defined as a fragment of genomic sequence of k length. Sequence dissimilarity between genomic sequences can be computed based on the set of k-mers.



Chaos Game Representation: 2-D graphic representation of genomic sequence composition portrays non-random biases in k-mer composition and frequency referred to as a Genomic Signature.



Machine Learning with Digital Signal Processing: Alignment-free supervised machine learning tool can use CGR for ultra-fast and scalable classification of thousands of bacterial[3], mitochondrial[4] and viral[5] genomes.

[3] Randhawa et al. *BMC Genomics.* 2019 Apr 3;20(1):267.
[4] Randhawa et al. *Bioinformatics.* 2019 Dec 13; 36(7):2258-2259.
[5] Randhawa et al. *PLoS ONE.* 2020 Apr 24;15(4):e0232391.

Motivation: Rapid, accurate, and scalable classification of new cancer genome sequences supports clinical diagnosis and prognosis.

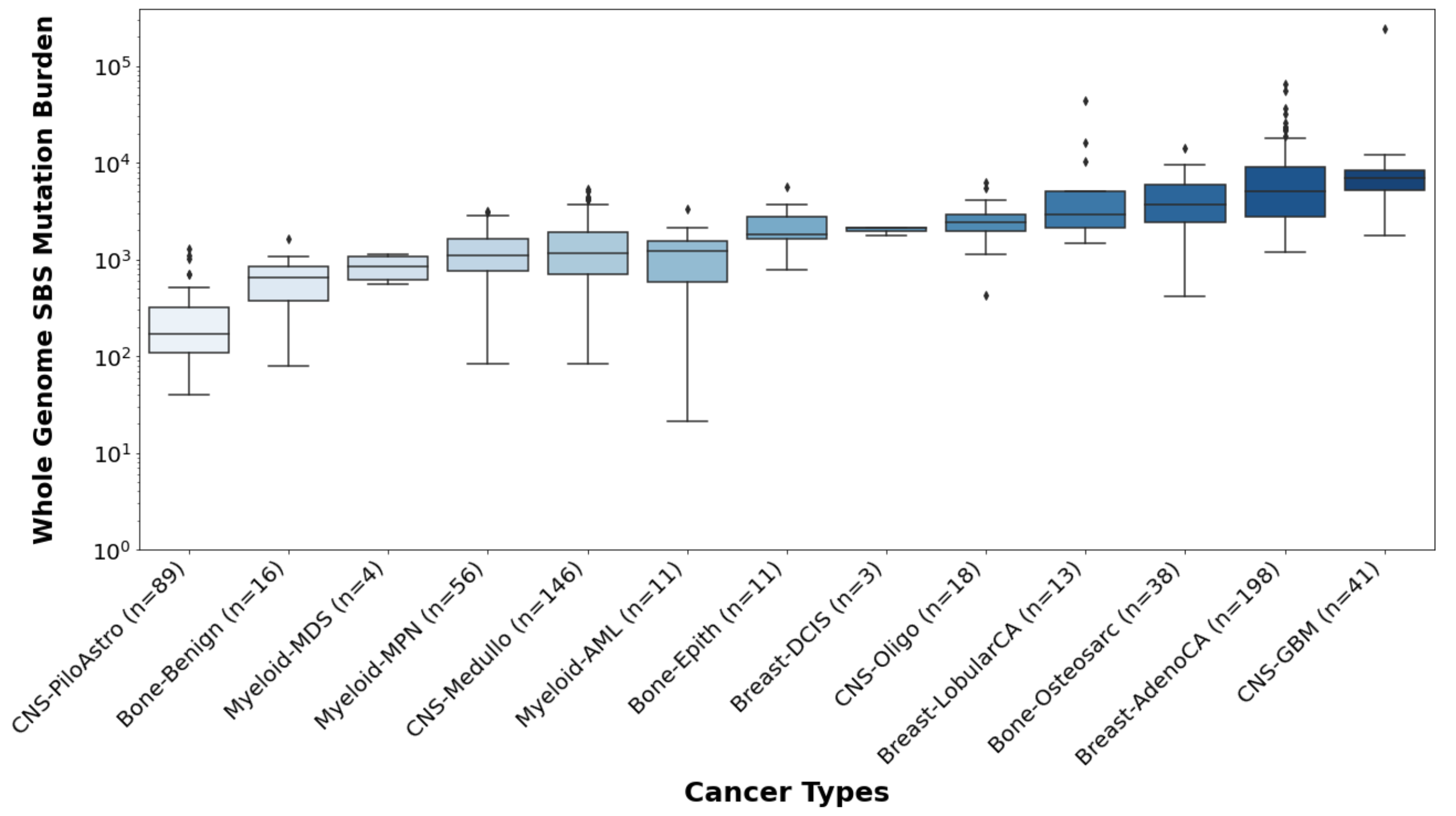
Mutation Types

- Characteristic mutation types can differentiate between different cancer genomic sequences.
- *Cosine similarity metric* from 0 (independent) to 1 (identical) measures the similarity of mutation type profiles seen in cancer whole genome sequences.

Mutation Burden

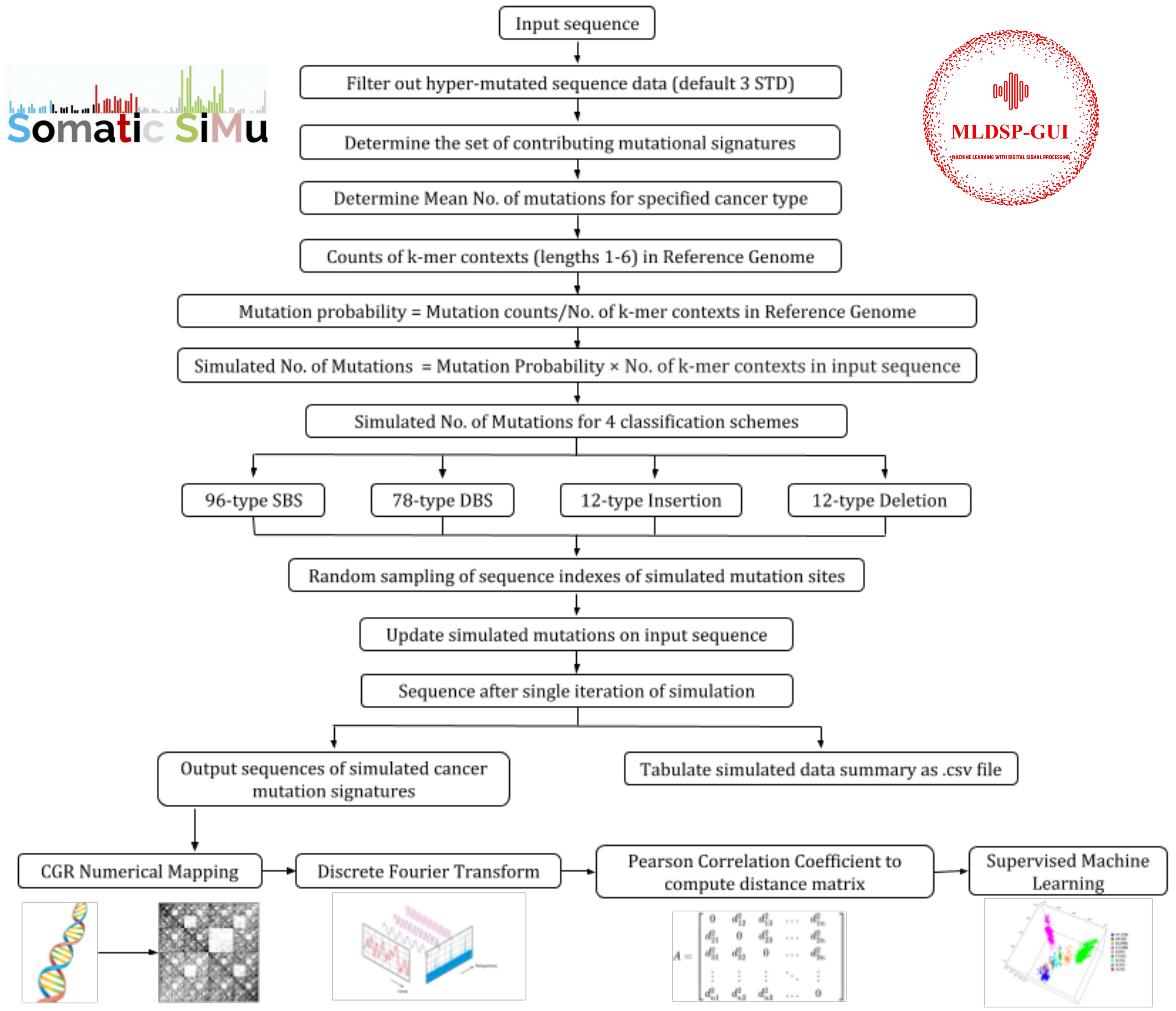
- The *whole sequence mutation burden* represents the product of multiple accumulated mutational processes found in cancer and varies by orders of magnitude between cancer types.

Whole Genome Single Base Substitution Mutation Burden in 13 Cancer Types (n=644)



SomaticSiMu Simulation and ML-DSP Classification For Cancer Subtype Classification

SomaticSimu: A novel *in silico* software tool to simulate genome evolution using known cancer mutation signatures imposed on Human GRCh38 Chr. 22 to produce a training dataset for benchmarking ML performance. Four classification tests within 3 to 4 cancer-type simulated sequences of Bone (n=400), Central Nervous System (n=500), Breast (n=400), and Myeloid (n=400) cancer was conducted using ML-DSP.



Input sequence

Filter out hyper-mutated sequence data (default 3 STD)

Determine the set of contributing mutational signatures

Determine Mean No. of mutations for specified cancer type

Counts of k-mer contexts (lengths 1-6) in Reference Genome

Mutation probability = Mutation counts/No. of k-mer contexts in Reference Genome

Simulated No. of Mutations = Mutation Probability × No. of k-mer contexts in input sequence

Simulated No. of Mutations for 4 classification schemes

96-type SBS | 78-type DBS | 12-type Insertion | 12-type Deletion

Random sampling of sequence indexes of simulated mutation sites

Update simulated mutations on input sequence

Sequence after single iteration of simulation

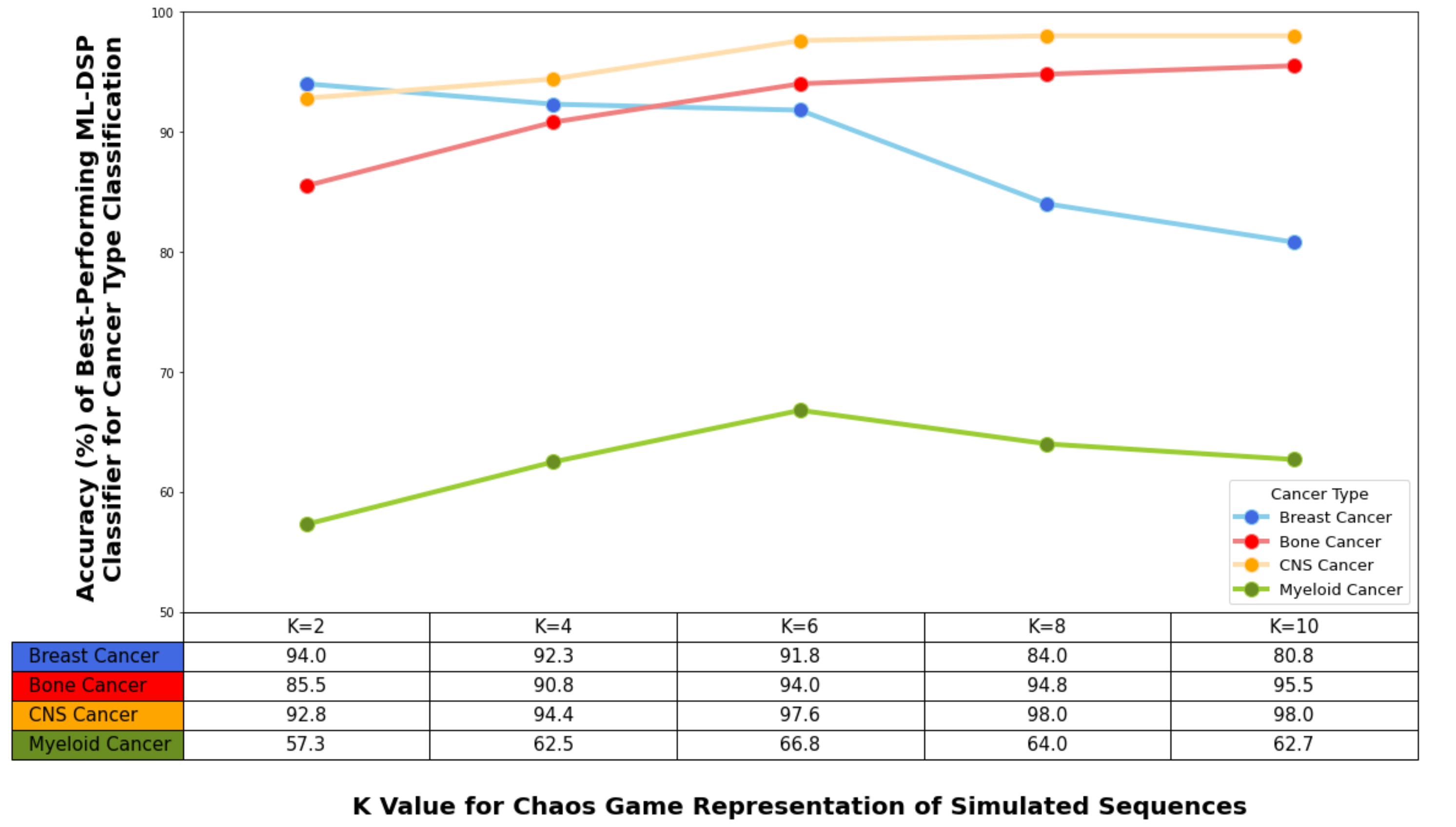
Output sequences of simulated cancer mutation signatures | Tabulate simulated data summary as .csv file

CGR Numerical Mapping → Discrete Fourier Transform → Pearson Correlation Coefficient to compute distance matrix → Supervised Machine Learning

ML-DSP Classification Performance

- Achieved up to 100% accuracy of cancer-type simulated sequences vs. non-cancer sequences (Quadratic SVM classifier).
- High accuracy in classification within subtypes of Bone (94.8%), Breast (94%), CNS (98%), and Myeloid (66.8%) cancer sequences.

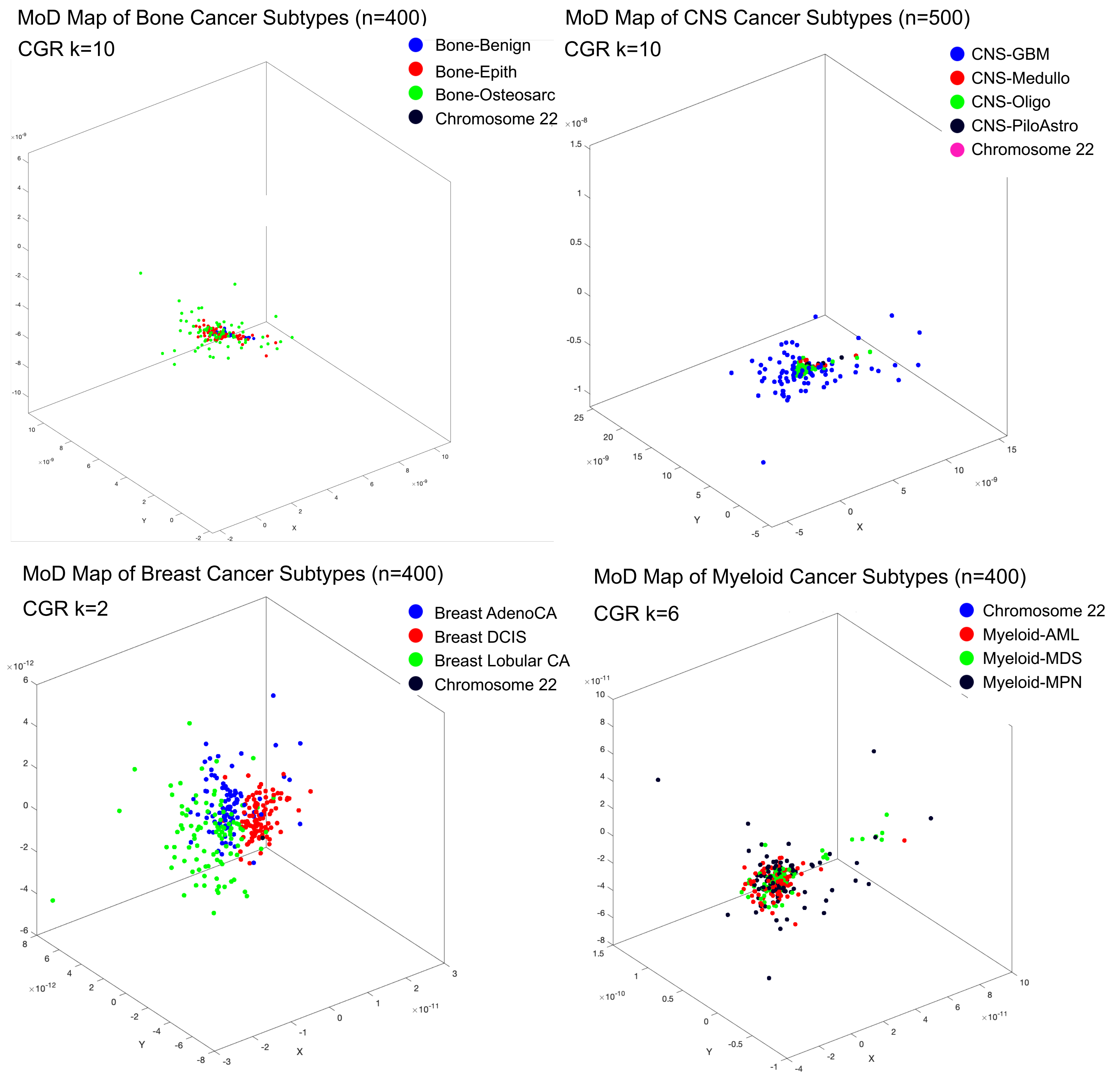
ML-DSP Classifier 10-fold Cross Validated Classification Performance on 4 Cancer Types (n=400) Simulated Sequences



	K=2	K=4	K=6	K=8	K=10
Breast Cancer	94.0	92.3	91.8	84.0	80.8
Bone Cancer	85.5	90.8	94.0	94.8	95.5
CNS Cancer	92.8	94.4	97.6	98.0	98.0
Myeloid Cancer	57.3	62.5	66.8	64.0	62.7

Molecular Distance Map of Supervised ML Clustering

3-D Molecular Distance Map of sequence representations portrays clusters of classes and visualizes more defined clusters with higher classification accuracy.



Discussion

- Quadratic SVM classifier achieved above **94% accuracy** for (Breast, Bone, CNS) cancer subtype sequence sequences for tests with **high sequence mutation burden.**
- Quadratic SVM classifier achieved up to **66.8% accuracy** for (Myeloid) cancer subtype sequences for tests with **high signature cosine similarity** (>0.75) and **low sequence mutation burden.**
- Potential for early classification of novel cancer sequences using known mutation signatures and supports clinical diagnoses

Conclusion

SomaticSiMu simulates cancer sequences with imposed mutational signatures *in silico* to benchmark Machine Learning with Digital Signal Processing classifier performance and facilitates study of cancer subtype classification using mutation signatures as biomarkers

Publication QR Codes:

MLDSP-GUI

Randhawa et al. BMC Genomics

Randhawa et al. Bioinformatics

Acknowledgements: Windsor Cancer Research Group, NSERC CRSNG, we spark health institute, Western Biology