Good afternoon everyone. Foremost, I would like to thank the organizers of the WCRG's International Cancer Research Conference for giving me the opportunity to present my research today. I now would like to introduce my project, Simulated Cancer Mutational Signatures to Classify Cancer Types using an Alignment-free Machine Learning-based Approach.

To begin, my research goal was 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 as a large training dataset for a novel alignment-free supervised machine learning tool.

Previous studies have shown that somatic mutations in cancer genomes can have exogenous and endogenous origins. Characteristic mutation types contribute to a mathematical approximation of the mutation landscape known as a signature. This poses the potential for classification based on signature types and sequence burdens. The figure shows an example of a mutation signature plot, a visual representation where each peak represents a single mutation type and the cumulative array of peaks wholly describes a mutation signature.

Mutation signatures are often described as certain single base substitutions, double base substitutions and small indels within a k-mer context. K-mer is defined as a fragment of genomic sequence of k length. Sequence dissimilarity between genome sequences can be computed through methods based on the set of k-mers. Kmer figure

Alignment-free methods of sequence comparison require orders of magnitude less computational resources and time with comparable accuracy to traditional alignment-based methods of sequence comparison. The Chaos Game Representation is a 2 dimensional graphic representation of genome sequence composition that can portray non-random biases in k-mer composition and frequency referred to as a genomic signature. A CGR of a DNA sequence is plotted in a unit square with the four vertices labelled by the nucleotides A, C, T, and G. The plotting procedure starts with the first nucleotide of the sequence being plotted halfway between the centre of the square and the vertex representing this nucleotide; successive nucleotides in the sequence are plotted halfway between the previous plotted point and the vertex representing the nucleotide being plotted. Increasing the K value of a CGR increases length of the k-mer and leads to an increase in the resolution of the graphic representation.

We have seen how genomic signatures can be useful as a method of classification and that alignment-free machine learning approaches can use chaos game representations of sequences to perform ultra-fast and scalable classification of thousands of bacterial, mitochondrial and viral genomes. I would like to investigate to what degree of accuracy and sensitivity can the alignment-free machine learning approach, Machine Learning with Digital Signal Processing (MLDSP), can classify cancer subtype whole genomic signatures.

This is important for the development of a pipeline for a novel, accurate and scalable classification of new cancer genome sequences to support clinical diagnosis and prognosis.

Based on a priori knowledge, one biomarker that machine learning-based classification models can use as a means to differentiate between cancer types are characteristic mutation types. Cosine similarity is a metric of similarity from 0 to 1 between two vectors of an inner product space, where a higher score indicates more similar mutation types between cancer classes and a lower score indicates more unique mutation types that delimit between cancer classes. A second biomarker that machine learning-based classification approaches can use as a means to differentiate between cancer types is mutation burden. Some cancer types such as Skin Melanoma are known to have whole genome mutation burdens on orders of magnitude greater compared to juvenile cancer types such as Pilo Astrocytoma.

To benchmark MLDSP performance, I have developed a novel simulation program, Somatic SiMu, based on known signature data from 2780 whole cancer genomes, comprising 37 cancer types and 4 signature types sourced from the Pan Cancer Analysis of Whole Genomes. Somatic SiMu can generate simulated sequences with biologically representative frequencies of single base substitutions, double base substitutions and single base indels at the genomic signature level. I use Somatic SiMu as a part of an end-to-end bioinformatics pipeline for simulating sequences in silico, where I can artificially upsample minority classes to obtain a more balanced class distribution recommended for supervised machine learning. I performed 4 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 using ML-DSP.

Using MLDSP, we were able to achieve up to 100% accuracy of cancer-type simulated sequences vs non-cancer sequences with the Quadratic Support Vector Machine classifier. Within the 4 tests conducted, we were also able to achieve high accuracy in classification within cancer types, up to 98% accuracy in differentiating between different subtypes of simulated sequences. We also tested across increasing k value of the Chaos Game Representation, which increases resolution of the CGR and provides more fine-tuned input to supervised machine learning approach, to maximize classifier performance.

On the third column, we produced molecular distance map of supervised machine learning clustering using MLDSP. The 3 dimensional molecular distance map visualizes each cancer subtype cluster as one colour and visualizes more defined clusters with higher classification accuracy. For example, looking at classification performance between the 3 breast cancer subtypes using ML-DSP, we achieved up to 94% accuracy and we can see well-defined clusters through the molecular distance map.

To sum up our results, we see that the best-performing Quadratic SVM classifier achieved above 94% for Breast, Bone, and CNS cancer-type simulated sequences with high sequence mutation burden. Conversely, the Quadratic SVM classifier achieved up to 67% accuracy for Myeloid cancer subtype sequences for tests with similar characteristic mutation types, as measured by a high signature cosine similarity, and low sequence mutation burden. We see that by finetuning hyperparameters for MLDSP model such as the k value for the input CGR, we can substantially improve accuracy up to 98% for certain cancer subtypes based on known mutational signatures .This poses great potential for the ultra-fast and scalable alignment-free supervised machine learning approach to classify novel cancer sequences using known mutation signatures within hours and can objectively support clinical diagnoses, especially in more ambiguous cases.

In 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.

I would like to sincerely thank the organizers of the WCRG's International Cancer Research Conference for this opportunity, my co-supervisors Dr. Hill and Dr. Singh, the co-authors and support from the Department of Biology at the University of Western Ontario.