Machine Learning with Digital Signal Processing for Classification of Mouse Genotypes

Kathleen Hill[1], Gurjit Randhawa[2], Nicolas Boehler[1], Hailie Pavanel[1], Ali Coyle[1], Pok Wan[1], Lila Kari[3]

[1]Department of Biology, University of Western Ontario, London ON, Canada; [2]Department of Statistical and Actuarial Sciences, University of Western Ontario, London ON, Canada;
[3]Department of Statistical and Actuarial Sciences, University of Waterloo, Waterloo ON, Canada

khill22@uwo.ca

Machine Learning (ML) Applied to the Classification of Genomic Information Uncovers a Wealth of Information

Genomes have individual, pervasive signatures of primary sequence organization.
Kari et al. *PLoS One.* 2015 May 22;10(5):e0119815

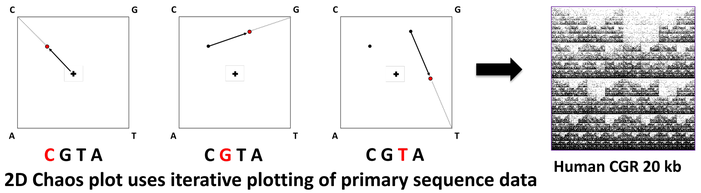
ML is a powerful tool for classification of thousands of genomic signatures
Randhawa et al. *BMC Genomics.* 2019 Apr 3;20(1):267

Here, classification of genomic signatures using ML is extended for the first time to single nucleotide polymorphic (SNP) genotypes to test accuracy in classification of genetic relatedness across a spectrum of origins and breeding histories.

The wealth of diversity in mouse genetic backgrounds provides benchmark testing for the power of ML and SNP genotypes to classify and identify different mouse genetic backgrounds.

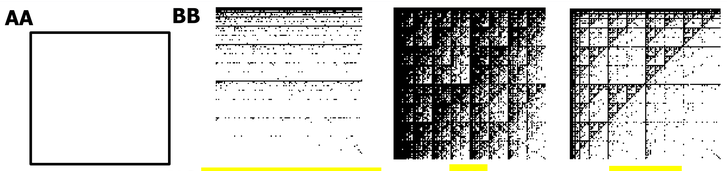
ML applied to SNP genotypes is predicted to find relevance in classifying phenotypes like cancer and inherited diseases, and in distinguishing mutation signatures of endogenous mutations and environmental mutagen exposures.

Chaos Game Representation (CGR): Two-dimensional Representation of DNA Sequence Organization in Genomic Signatures



A 2D Chaos plot uses iterative plotting of primary sequence data

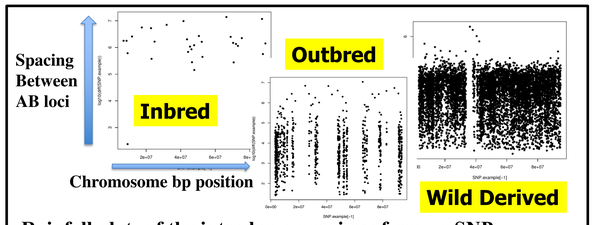
Two-dimensional CGR Representations of Mouse SNP Genotypes



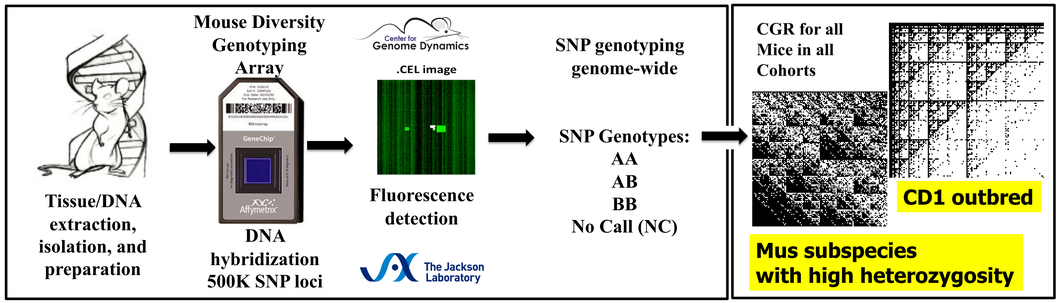
- **Chaos plot vertices are adapted replacing single nucleotides at the vertices with AA, BB, AB and no call (NC) SNP genotypes plotted for 500K SNP loci assayed by array technology.**
- **Plotting rules are unchanged.**
- **CGR patterns for different mouse SNP genotypes are as predicted.**

Mouse Genetic Backgrounds Provide an Excellent Benchmark for a Proof-of-Concept Machine Learning Approach to Classifying Mice

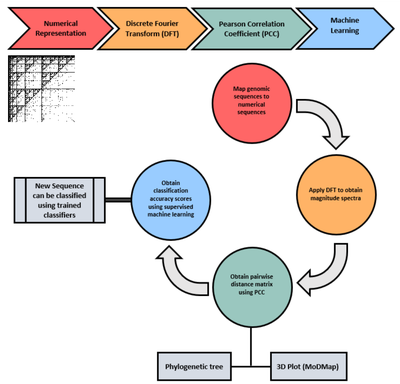
Cohorts of Mice	Heterozygosity Determined
Classical Inbred	0.06 – 1.5 %
Recombinant Inbred	0.09 – 3.9 %
Wild-Derived	0.5 – 15.8 %
NMRI Outbred	5 – 8 %
CD1 Outbred	10 – 12 %
F1	19 – 46 %



Rainfall plots of the inter-locus spacing of mouse SNP heterozygosity across the chromosome landscape show the striking pattern diversity



ML-DSP methodology



Molecular Distance Map (MoDMap)

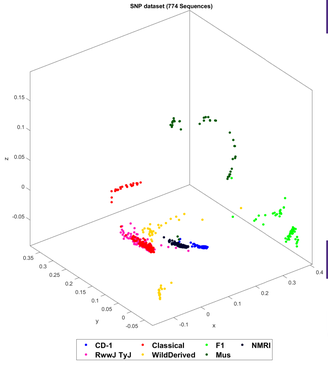
Total sequences: 774

Length (# SNP loci): 493290

Clusters (# genotypes):
CD-1 Outbred (100);
Classical Inbred (126);
F1 (59);
NMRI Outbred (287);

Recombinant Inbred RwwJ_TyJ(112);

Wild Derived (50);
Mus subspecies (40);



ML-DSP Excels in Classification Accuracy for SNP Genotyping

Classification Accuracy (%)

Linear Discriminant	97.7
Linear SVM	93.9
Quadratic SVM	96.8
Fine KNN	94.4
Subspace Discriminant	97.5
Subspace KNN	94.1
Average Accuracy	95.7

Confusion Matrix Results

TrueClass\PredictedClass	CD-1	Classical	F1	NMRI	RwwJ_TyJ	WildDerived	Mus
CD-1	100	0	0	0	0	0	0
Classical	0	115	0	0	10	1	0
F1	0	0	58	0	0	0	1
NMRI	0	0	0	287	0	0	0
RwwJ_TyJ	0	4	0	0	108	0	0
WildDerived	0	1	1	0	0	48	0
Mus	0	0	0	0	0	0	40

Inter-cluster distances

True_Predictor	CD-1	Classical	F1	NMRI	RwwJ_TyJ	Wild Derived	Mus subspecies
CD-1	0	0.1353	0.2247	0.0206	0.0909	0.0956	0.1912
Classical	0.1353	0	0.4413	0.0950	0.0781	0.1186	0.2856
F1	0.2247	0.4413	0	0.3312	0.4291	0.4303	0.2544
NMRI	0.0206	0.0950	0.3312	0	0.0476	0.0539	0.2261
RwwJ_TyJ	0.0909	0.0781	0.4291	0.0476	0	0.0736	0.2612
Wild Derived	0.0956	0.1186	0.4303	0.0539	0.0736	0	0.2835
Mus subspecies	0.1912	0.2856	0.2544	0.2261	0.2612	0.2835	0

Conclusions

ML-DSP excels in classification accuracy, speed and scalability of large datasets.

ML-DSP is relevant in classifying new, unknown and wild-caught mouse genetic backgrounds using SNP genotype data.

By extension, Human SNP genotype data can be classified to gain new insight into genotype associations with phenotypes of cancer and inherited disease.

QR codes for relevant publications

ML-DSP

MLDSP-GUI

MLDSP-GUI (Software)

Acknowledgements

Western Science
NSERC CRSNG
London Regional Genomics Centre
Western Biology
Center for Genome Dynamics
The Jackson Laboratory