Presenter's Name: Bagherichimeh, Sareh

Additional Authors: Poon AFY

Abstract Title: Selection Pressure on Surface Exposed Virus Proteins

Viruses infect human cells through interaction between virus surface proteins and the cell surface components of human cells, termed the viral receptors. This interaction influences the evolution rate of both the viral receptors in the human host and the invading surface exposed virus proteins. A systematic study in 2020 analysed the evolution of numerous viral receptors in humans and found about half of the viral receptors had undergone an elevated rate of evolution. Their results suggest that host-virus arms race drives accelerated adaptive evolution in viral receptors.1

In viruses, the surface proteins of the influenza A virus Hemagglutinin and Neuraminidase are best known for undergoing positive selection evident by the elevated rate of non-synonymous substitution in genes encoding these two surface proteins compared to the rest of the genome.2, 3 A similar increased rate of evolution in surface exposed proteins has been observed in a number of viruses, including the human immunodeficiency virus and the human papilloma virus. However, a comprehensive picture of the evolutionary rate of surface proteins is still lacking. Systematic analysis of the evolution of virus proteins has profound implications in understanding the evolutionary arms race between viruses and hosts. As such I aim to conduct a systematic analysis of the evolutionary pattern of human viruses by investigating the selective pressures acting on protein coding genes and comparing the rate of adaptation of surface exposed proteins to other virus proteins. We hypothesis that human virus genes encoding surface-exposed virus proteins are under a higher adaptation than other coding regions in the viral genome.

We retrieved all of the full genome sequences for 49 human virus species available on the National Center for Biotechnology Information. We selected human viruses with 3 or more protein coding genes with more than 100 full genome sequences. For each virus species, we aligned all of the sequences and constructed a phylogenetic tree that was used for PruneTree to subsample 100 sequences. We then eliminated the overlapping nucleotide sequences of overlapping protein coding genes. Next, we will gather meta data of virus proteins to distinguish surface-exposed and non-surface-exposed proteins. Gene ontology knowledgebase and protein function predictors will be used to supplement the protein annotations for accurate labelling. The nonsynonymous (dN) to synonymous (dS) substitution ratio in protein-coding genes is used to measure the strength of natural selection. The dN/dS ratio will be calculated for each protein coding gene using the Fast Unconstrained Bayesian Approximation (FUBAR)4. The unique fingerprint of dN/dS ratio for each site of a surface exposed protein gene would be compared with a non-surface exposed protein in a virus species.

POSTER PRESENTATIONS 2
2A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Bhai, Pratibha

Additional Authors: Bhai P, Hsia CC, Schenkel LC, Hedley BD, Levy MA, Kerkhof J, Santos S, Stuart A, Lin H, Broadbent R, Nan S, Yang P, Xenocostas A, Chin-Yee I, Sadikovic B

Abstract Title: Clinical and diagnostic impact of implementing a tier-one NGS DNA and RNA fusion panel testing in patients with suspected myeloid malignancies

Background: The use of molecular genetic biomarkers is rapidly advancing in diagnosis, prognosis, and clinical management of leukemia. We have implemented a next generation sequencing (NGS) panel for detection of gene mutations and gene fusions as a tier-one test in patients with suspected hematological malignancies. In this study, we summarize the findings and assess the clinical impact in the first 2023 patients tested.

Methods: We assessed patients for variants using NGS based Oncomine Myeloid Research Assay (ThermoFisher) including 40 genes (17 full genes and 23 genes with clinically relevant "hotspot" regions), along with a panel of 29 fusion driver genes (including over fusion 600 partners)

Results: The key findings include increased diagnostic yield and evidence of independent clinical impact of NGS on management of compared to the classical diagnostic techniques across the myeloid malignancy subtypes.

Conclusions: This study involves the largest reported cohort of patients with suspected myeloid malignancies with comprehensive NGS molecular testing in a clinical laboratory setting. Our findings demonstrate the clinical utility and feasibility of integrating NGS-based DNA and gene fusion testing as a tier one diagnostic test in patients with hematologic malignancies. Overlap with cytogenetic test results provides opportunity for testing reduction and streamlining.

Presenter's Name: Chao, Wan-yu

Additional Authors: Poon AFY

Abstract Title: Genetic clustering for real-time surveillance of HIV outbreaks

Introduction: Genetic clustering identifies sub-populations ("clusters") where members are more genetically similar to each-other than to the rest of the population. In HIV epidemics, short rapid molecular evolution enables the inference of transmission links from genetic clusters and therefore the development of real-time HIV prevention monitoring programs. For genetic distance based methods the threshold value of 0.015 is often recommended. However, certain factors may favour different threshold values. In our study, we demonstrate that at the current threshold values, these algorithms often identify "giant clusters" that consist of the majority of the population, which prevent the goal of focusing the public health response. We also investigate epidemic factors that may reduce predictive accuracy of these algorithms. Lastly, we attempted to improve the predictive accuracy of these genetic clustering methods.

Hypothesis: We hypothesize that by distorting the phylogenetic tree of the epidemic, we can compensate for epidemic growth factors that affect the shape of the tree, and thereby improve predictive accuracy.

Methods: We first established evidence of giant clusters in both real and simulated dataset of HIV-1 epidemics. Datasets were retrieved from NCBI PopSet and separated by both subtype and country of origin. Epidemic simulation of the epidemic was completed with TreeWithinTrees (twt) and sequence simulation with INDELible. Subsequent analysis and tree manipulation was performed in R.

Results: Both datasets from real and simulated datasets show the appearance of the giant clusters. Variation in the sampling time of the epidemic had a strong effect on predictive accuracy, however rescaling the tree did not have a positive effect on improving the predictive accuracy.

POSTER PRESENTATIONS 2
2A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Christensen, Erik

Additional Authors: Luo P, Husic M, Mahalanabis A, Diaz-Mejia J, Turinsky A, Rumani A, Brudno M, Pugh T, Shooshtari P

Abstract Title: An Analysis of Automated Cell-Type Labelling Methods

Single cell RNA sequencing (scRNA-seq) allows researchers to precisely define complex cell populations through the examination of gene expression in individual cells from a sample. This is particularly useful in cancer research, as the tumour microenvironment (TME) can be incredibly heterogeneous, containing malignant cells, cells of the tissue of origin, immune cells, fibroblasts, and more. This heterogeneity often affects how cancer progresses and responds to treatment, making it important to accurately identify all cells within the TME. Current standard scRNA-seq pipelines involve clustering the cells by differentially expressed genes and then assigning cell type labels to each cluster based on gene sets. The assignment of cell labels is typically a manual task, which can rely on subjective decisions that may yield inconsistent results. Several automated techniques for identifying cell types exist, and have been tested on healthy scRNA-seq samples, but their performance on TME data is unknown.

Here we evaluate 26 methods to automate the process of assigning labels to TME cells - 19 supervised cell-based methods and 7 semi-supervised cluster-based methods. In order to evaluate these methods on the TME, we selected 8 datasets from various cancers. Raw counts expression data were obtained from Gene Expression Omnibus, ArrayExpress, or Genome Sequence Archive and normalised, after which cell-type predictions were generated using each method. For the 7 cluster-based methods, cell clusters were obtained using Seurat. To measure the performance of each method we used a bootstrapped approach, where each set of predictions was subsampled with replacement 10000 times and assessed with F1 Score. We then took the mean of these 10000 F1 Scores as the performance for that method. We further quantified the performance by per-class F1 Score, generating a score for each algorithm's ability to predict a specific cell type. These scores allowed us to determine which methods perform better on malignant cells and identify specific weaknesses in certain algorithms.

Our study shows that cell-based algorithms outperform cluster-based algorithms overall. We reasoned that could be because the cell-based algorithms are fully supervised, learning from each individual cell in the training set instead of only examining select genes within a group of cells. Overall CaSTLe, scVI, and SVM are the top-performing algorithms each with a median F1 Score of 0.96. Of these three algorithms, SVM is the fastest running. Among cluster-based methods, GSVA ranked highest with a median F1 Score of 0.56.

Cluster-based methods are often unable to identify malignant cells within a dataset due to lack of appropriate gene signatures. Cell-based algorithms, however, tended to perform better on the malignant cell categories – an interesting result as the algorithms learn from the data they are provided and most methods are not tailored to TME data specifically. This may be due to a higher proportion of malignant cells present in the datasets providing a better source for the algorithms to learn what constitutes a malignant cell profile. In summary, our analysis presents a set of guidelines for researchers to understand the various cell type identification methods available and select the best one for their study.

Presenter's Name: Foroutan, Aidin

Additional Authors: Sadikovic B

Abstract Title: Childhood-onset dystonia 28 episignatures associated with pathogenic mutations in KMT2B gene

Introduction: Childhood-onset dystonia is a rare neurological hyperkinetic movement condition characterized by sustained or intermittent muscle contractions causing abnormal movements and/or postures. Dystonias, which are usually genetically-determined and can be characterized by other additional neuropsychiatric and systemic features, pose a significant diagnostic challenge for clinicians. The advent of parallel sequencing has revolutionized the landscape of dystonias' genomic variants, enabling identification of several novel pathogenic variants in the last decade, most of them causing complex forms. Recently, heterozygous mutations in lysine methyltransferase 2B (KMT2B), was reported to be associated with a childhood-onset, progressive and complex form of dystonia named childhood-onset Dystonia 28 (DYT28). The encoded protein of KMT2B belongs to Set1-Trithorax type H3K4methyltransferases, which tri-methylates the 4th lysine residue of the histone H3 protein and this mark occurs at the transcription start sites (TSS) of actively transcribed genes to induce gene activation. Pathogenic KMT2B variants generally are de novo events and result in loss-of-function of the methyltransferase, indicating haploinsufficiency as the mechanism of disease. Recently, the development of new accurate machine learning-based classification strategies has allowed identification of disease-specific DNA methylation patterns (known as "episignatures"), for an increasing number of neurodevelopmental disorders. Therefore, this study aimed to define a disease-specific episignature for DYT28 syndrome using a cohort of clinically and genetically confirmed DYT28 patients.

Methods: Genome-wide DNA methylation of peripheral blood from a cohort of 18 individuals, of which 8 and 10 carried pathogenic and VUS mutations in KMT2B gene respectively, was performed using the Illumina EPIC arrays. Only those 8 samples that had pathogenic mutations in the KMT2B gene (P-SAMPLES) were used for finding the significant differentially methylated probe (DMP) signature and the other 10 VUS were later mapped to the discovered methylation signature. DMPs were identified using R (V. 4.0.2), comparing the 8 P-SAMPLES with 56 age and sex matched controls from our EpiSign Knowledge Database (EKD). The robustness and sensitivity of the selected DMPs was tested using unsupervised machine learning algorithms, including hierarchical clustering and multidimensional scaling, as well as 8 rounds of leave-1-out cross validation. To further check the sensitivity and specificity of the selected DMPs, the supervised support vector machine (SVM) learning model was constructed to classify the DYT28 episignature distinct from 38 other neurodevelopmental disorders and congenital anomalies (ND/CAs) available in our EKD.

Results and Discussion: In total, 196 DMPs were identified of which 96% (189 out of 196) exhibited relative hypermethylation. The results demonstrated the evidence of a unique and highly specific episignature in 8 samples that had pathogenic mutations in the KMT2B gene. Interestingly, amongst the 10 VUS samples, 2 were grouped with the DYT28 cohort, 7 grouped with controls, and 1 showed an intermediate position. Regarding the 1 VUS showing intermediate position, the father was unaffected but had the same variant as his affected child, so the difference in the disease presentation in the child suggests a partial penetrance associated with this variant.

POSTER PRESENTATIONS 2
2A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Haghshenas, Sadegheh

Additional Authors: Sadikovic B

Abstract Title: Identification of a DNA Methylation Signature for the Renpenning Syndrome 1

Introduction: Mendelian neurodevelopmental disorders are typically difficult to diagnose due to the high degree of overlap between their clinical representations, as well as the occasional ambiguity of traditional genetic testing results. It is well-established that for many of these syndromes, a mutation in the underlying gene causes alterations in methylation throughout the genome, called an episignature. Genome-wide DNA methylation analysis has been utilized in diagnosing rare genetic disorders for a few years. Here, we identify the methylation signature specific to patients with Renpenning syndrome 1 (RENS1), which is an X-linked recessive neurodevelopmental disorder, caused by variants in polyglutamine-binding protein 1 (PQBP1). Using the detected episignature, we then construct a highly sensitive and specific classification model that is capable of predicting whether a new patient has the syndrome, based on their methylation profile.

Methods: The DNA methylation analysis is performed by supplying DNA extracted from the peripheral blood of patients into Illumina Infinium Methylation EPIC array BeadChips after bisulfite conversion, and the methylation level at each probe was measured as the ratio of methylated signal intensity to the sum of methylated and unmethylated signal intensities. The analysis was performed in R. Using limma package, the methylation difference between the 8 case samples and the 40 age, sex, and array type-matched control samples, along with the p-values were calculated. 1000 probes with the highest methylation difference and lowest p-values (corrected using the Benjamini-Hochberg method) were selected. Next, we selected 400 probes with the highest area under the receiver operating characteristic (ROC) curve, and finally, removed the probes that had a correlation above 90%. The methylation levels at the remaining 214 probes were considered as the episignature of the syndrome. The robustness of this episignature was verified using hierarchical clustering and multidimensional scaling. The reproducibility of the episignature was assessed by performing an 8-fold cross-validation, where at each round, 7 case samples were used for training and 1 for testing. Finally, a highly sensitive and specific support vector machine (SVM) classifier was constructed using the selected probes.

Results: Our findings demonstrate the existence of a highly robust episignature for RENS1. We also illustrated the high specificity of the RENS1 episignature by comparing it to over 1000 samples from patients with episignatures in over 40 genes associated with 38 other neurodevelopmental syndromes.

Discussion: The binary classification model constructed based on the identified episignature, along with the unsupervised models (hierarchical clustering and multidimensional scaling) can be applied for future classification of new suspects of RENS1 with an inconclusive clinical diagnosis and uncertain genetic testing result.

Presenter's Name: Christensen, Erik

Additional Authors: Naidas A, Shooshtari P, Husic M

Abstract Title: TMExplorer: A Tumour Microenvironment (TME) Single-cell RNAseq Database and Search Tool

Tumour microenvironments (TME) contain a variety of immune cells, malignant cells, and cells of the tissues of origin, whose interactions likely influence tumour behaviour and response to cancer treatment. Single-cell RNA sequencing (scRNAseq) enables us to measure expression profiles of individual cells. This technology has advanced rapidly in recent years, and several scRNAseq TME datasets from a variety of cancer types have been published. Despite this, we still lack a comprehensive database collecting the TME data from these studies and presenting them in an easily accessible format. Additionally, despite the development of several computational methods to aid in the analysis of scRNAseq data, there is no unbiased evaluation of their performance, generalisability, and scalability when applied to TME datasets. This is partially due to the lack of a standardised TME scRNAseq database to use in benchmarks.

In order to address these problems, we have built a database of TME scRNAseq data. Our database contains 21 TME scRNAseq datasets comprising eight different cancer types. The number of cells and genes contained within each dataset varies between [71–57,530] cells and [5,796–57,915] genes. We have created an R package called TMExplorer, which provides an interface to easily search and access all available datasets and their metadata. Data and metadata are kept in a consistent format across all datasets, with multiple expression formats available depending on the user's experience level and use case. Users can view a table of metadata and select individual datasets or filter them by characteristics shown in the metadata. For instance, users may choose to select only those datasets with gene markers or cell annotations available in order to evaluate the performance of a new algorithm against a known gold standard. Users may also select a specific type of cancer in order to study it using all of the published scRNAseq data available in our database for that particular cancer. Users are also provided with an option to save the data in multiple formats in order to view or process the data outside of R.

With this database, we improve the ease of studying the tumour microenvironment with single-cell sequencing. A multitude of datasets are easily available with a common interface to suit both bioinformaticians looking to validate their algorithms and clinicians or basic scientists looking to study cancer more closely. We aim to add datasets as they appear in the literature and provide a way for scientists to submit their own data for inclusion. By improving the ease of accessing scRNAseq TME data, TMExplorer increases the effectiveness of single-cell research.

POSTER PRESENTATIONS 2
2A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Liu, Mo

Additional Authors: Poon AFY

Abstract Title: Outbreak Detection from Virus Genetic Sequence Variation by Community Detection

Introduction: The identification of groups of related individuals within a population is a ubiquitous problem in virus study. Previous studies map the distance matrix generated from the Tamura-Nei (1993, TN93) pairwise genetic distances to an adjacency matrix by applying some threshold and identify the connected components in the resulting graph as clusters. However, due to the relatively slow rate of molecular evolution and rapid transmission rate of this virus, it is difficult to identify meaningful clustering criteria on GISAID SARS-COV-2 data. Here, we will construct graphs on varying TN93 thresholds and apply community detection methods to extract clusters. We hypothesize that community detection method is a better way to extract clusters from the global SARS-CoV-2 genome data set than connected component-based clustering.

Research Objectives: The two objectives of this project are (1) to investigate the use of community detection methods as a means of identifying epidemiological clusters from HIV-1 data sets (2) to investigate the use of community detection methods for identifying epidemiological clusters in SARS-CoV-2 data sets, using either TN93 or more rapid Manhattan distance-based criteria.

Method: We set the most recent year's data be testing data, and rest of them be training data. Base on the TN93 matrix from training data, we will set use varying thresholds to generate a series of edge lists. Community detection methods will be used to partition each graph into a set of clusters. To test the result, we first fit a Poisson regression model whose outcome is the number of new cases in the testing set that are linked by edges to clusters in the training data, based on one or more cluster-level predictor variables. Then we use "generalized Akaike information criterion" to compare two Poisson regression models, one without any covariates, and a second with covariates.

Presenter's Name: Reilly, Jack

Additional Authors: Haghshenas S, Levy M, Foroutan A, Relator R, Sadikovic B, Kerkhof J

Abstract Title: Identification of DNA Methylation Episigns for Diagnosis and Phenotype Correlations in Neurodevelopmental Disorders.

Introduction: Neurodevelopmental syndromes have a significant amount of overlap in regards to clinical features, as well as frequently negative or ambiguous genetic findings which can make accurate diagnosis and management difficult. Within this subset of diseases however, a growing number of unique DNA methylation patterns or "Episignatures" have been identified. These episignatures are derived from the assessment of epigenetic methylation tags throughout the genome which modulate expression through adjustment of DNA binding, chromatin structures and accessibility to protein machinery. Peripheral blood Episignatures have been demonstrated to be sensitive and specific tools for diagnosis, as well as the interpretation of ambiguous genetic test results. With our new national clinical trial (EpiSign-CAN) we hope to improve the assessment of rare diseases in Canada using this approach to DNA methylation assessment.

Hypothesis: We postulate the existence of additional neurodevelopmental mendelian conditions that exhibit highly specific and sensitive genomic DNA methylation patterns. Such Episignatures can be used for classification and diagnosis of these disorders. Further substratification of disorders Episignatures based on presentation of specific phenotypic characteristics may allow for further focusing of these Episignatures.

Methods: Patients' epigenetic profile will be assessed via Illumina microarray, which target around 850,000 CpG probes across the patient's genome, using peripheral blood samples. Methylation signal intensity values will be assessed in patients with confirmed or highly suspected diagnosis of the disorder in question. These values are then compared to unaffected control samples to determine probes which exhibit a high degree of methylation difference across a statistically significant number of samples. Probe selection is filtered through a number of quality control steps, including accounting for batch effect, blood cell type composition, sensitivity and specificity, ultimately deriving an episignature probe set between 200-500 probes, which are then used for the visualization and assessment of trends within the data. To further demonstrate the sensitivity and specificity of this classifying tool, samples are then provided to a support vector machine powered classification tool which scores each sample for their similarity to a number of different disorders with confirmed Episignatures.

Results: Several datasets (MRD32-KAT6A, ADNP/BISS-SMARCA2) have provided evidence of highly robust episignatures and indicate that phenotype-based correlations can improve diagnosis and understanding of disorders molecular and phenotypic presentation.

POSTER PRESENTATIONS 2
2A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Rooney, Kathleen

Additional Authors: Sadikovic B

Abstract Title: Identification of a Specific Episignature in the 22q11.2 Deletion Syndrome.

Introduction: The 22q11.2 deletion syndrome (22q11.2DS) is the most common genomic disorder in humans and is the result of a recurrent 1.5 to 3 Mb deletion on the long arm of chromosome 22, encompassing approximately 20-40 genes respectively. Clinical manifestations include Velo-Cardio-Facial syndrome, Di George syndrome, Opitz G/BBB syndrome and Conotruncal anomaly face syndrome. The mechanisms and genes involved in this disorder are not completely understood. An expanding number of genetic disorders are shown to exhibit consequent disruptions of epigenomic profiles in patient's peripheral blood, referred to as episignatures. These episignatures are highly consistent and are similar among patients affected by the same underlying genetic etiology. To date, DNA methylation profiles in patients with copy number variants such as deletions of the 22q11.2 region have not been systematically studied. We hypothesize that individuals with 22q11.2DS will exhibit a unique diagnostic methylation signature associated with haploinsufficiency of the 22q11.2 region.

Methods: To test this hypothesis, we performed analysis of genome-wide DNA methylation of peripheral blood from a cohort of 11 patients with 22q11.2DS using the Illumina 850K BeadChip array. Computational analysis in R was applied to compare the cohort against a subset of age and sex matched controls from our EpiSign Knowledge Database (EKD). Methylation levels for each CpG probe were measured and 424 differentially methylated probes were used for signature discovery. The robustness and sensitivity of the selected probes was tested using Euclidean hierarchical clustering and multidimensional scaling as well as rounds of leave-1-out cross validation. A support vector machine learning model was constructed to confirm the specificity for the episignature to classify the disorder distinct from known episignatures in other neurodevelopmental disorders and determine the suitability for inclusion in our EpiSign diagnostic test.

Results: Our results demonstrate the evidence of a unique and highly specific episignature in patients with 22q11.2DS. The sensitivity and specificity of this signature was further confirmed by comparing it to over 1000 patients with other neurodevelopmental disorders.

Discussion: These findings may provide a better understanding of the pathogenesis of this disorder by highlighting regions of differential methylation across the genome. This highly specific episignature can be used for molecular diagnosis of 22q11.2DS.

POSTER PRESENTATIONS 2
2A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Wong, Emmanuel

Additional Authors: Poon AFY

Abstract Title: Capturing within-host HIV-1 evolution dynamics using in-silica method

Introduction: Human Immunodeficiency Virus (HIV) is a disease with significant human and economic cost. It is a disease without a practical cure but can be treated with Anti-Retroviral Therapy (ART). Long term persistence of HIV within an individual is due to the latent reservoir, a source for reinfection that is un-affected by ART. The latent reservoir is difficult to study for various practical and ethical reasons in-vivo. Although the latent reservoir is a necessary target for HIV cures, its origin and persistence remain a contentious subject with contradictory explanations.

Objective: The overall goal of this project is to provide a simulation model that can capture the within-host dynamics, and to use the model to address the hypotheses for the origin and persistence of the within-host HIV latent reservoir. The immediate objective is to tune the simulation model to produce
data that is comparable to real-world data.

Hypothesis: The structure of the Latent Reservoir is dependent on the viral interactions occurring within an individual.

Methods: We start by defining the within-host model using the host-pathogen simulation program trees within trees (twt). twt is a R program that allows users to define unique compartments and the interactions between infectious and susceptible cells within those compartments. Model parameters include the four compartments: active, latent, replenishment and death; and the interactions between the compartments: Infections between cells, transitions between compartments, regeneration from the replenishment compartment and the death of cells. We then draw a distribution of representative samples to evaluate the effects of varying these model parameters.

Results: The presented work is a work in progress. twt can be used to simulate HIV dynamics in a within host model. Simulations reveal insights into the parameter space: certain combinations of parameters are non-sensical. These results provide direction to future work that will analyze the population structure of the latent reservoir.

Significance: Addressing the latent reservoir is a necessary step in providing a cure for HIV. Studying the Latent Reservoir is difficult and expensive in-vivo, this model will provide a method to test competing hypotheses. This project will also provide insights on the origin and persistence of the latent reservoir.