Presenter's Name: Asakereh, Reza

Additional Author(s): Shooshtari P, Zhang Q

Abstract Title: Integrative data analysis to uncover genes and pathways underlying the infiltrative power of lung cancer brain metastasis and the overall prognosis

Abstract:

Introduction: Recent advances in cancer diagnosis and treatment have resulted in better management of the mortality caused directly by tumors at primary sites. However, this on the other hand, has increased the proportion of observed metastatic cases among cancer patients. In particular, we observe a relatively high portion of brain metastasis with primary lung cancer. The ability to predict the prognosis of the primary tumor and the prediction of how rapidly the metastasis may occur, would result in providing better treatment decisions. As a step towards this goal, we need to identify the specific molecular mechanisms underlying fast/slow metastasis in different patients. Current studies in this area have several drawbacks, including neglecting the heterogeneity of the tumor cells, and/or focusing on a limited number of genes rather than a whole transcriptome analysis. In our study, we have addressed these drawbacks and uncovered several genes and molecular processes affecting the prognosis of the lung tumor brain metastasis.

Methods: We have collected a dataset which contains the spatially profiled gene expression of three different organs (lung, brain, and lymph nodes) in 35 patients using Nanostring Digital Spatial Profiler. We followed a data integration method based on Multi-Omics Factor Analysis (MOFA) to discover the molecular sources of heterogeneity in the patients and applied robust statistical tests to identify the factors that contribute to prognosis differences between patients. Then we identified a set of highly relevant genes and performed KEGG pathways analysis to identify the underlying pathways.
Results. Our analysis indicated that there are three major categories of pathways likely to be implicated in lung cancer patients with fast brain metastasis. They include (a) immunity-related pathways, (b) stemness related pathways such as focal adhesion and ECM receptor interaction, and (c) metabolic pathways such as Warburg effect and oxidative phosphorylation. The tumor associated immune related pathways are twofold: (1) the impaired immune system in tumor suppression such as neutrophil extracellular traps; and (2) the enhanced ability of the tumor cells to escape the immune surveillance, such as antigen processing and presentation.

Impact: Our analyses predict target pathways underlying the brain metastasis in lung cancer patients and establish a means for the prediction of the lung tumor prognosis using molecular experiments.

POSTER PRESENTATIONS 3
3A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Bagherichimeh, Sareh

Additional Author(s): Poon AFY

Abstract Title: Selection Pressure on Surface Exposed Virus Proteins

Abstract:

The mechanism of viral infection involves interaction between virus surface proteins and the cell surface components of human cells, termed 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 comprehensive picture of the evolutionary rate of surface proteins in viruses is still lacking. Systematic analysis of the evolution of virus proteins has profound implications to understanding the evolutionary arms race between viruses and hosts. I aim to conduct a systematic analysis to examine the hypothesis that Human virus genes encoding surface-exposed virus proteins are under a higher adaptation rate than other coding regions in the viral genome. 30 Human virus species from NCBI were collected. For each virus species, we sub-sampled a fixed number of 100 genomes while retaining maximum amount of variation. For each virus species, we obtained the coding sequences (CDS) in each genome by using the reference CDS as a template. The synonymous (dS) and non-synonymous (dN) substitution rates calculated by FUBAR were both averaged out over the length of each gene, and the percentage of sites under positive and negative selection were calculated as well. FUBAR also generated the evolutionary fingerprints for each gene. The distance between every pair of genes' evolutionary fingerprints was measured using Pearson Correlation and projected onto a principal component analysis (PCA). 405 genes of 30 virus species from 14 virus families were analysed, 161 of which were surface exposed. The mean dS of surface exposed virus proteins were significantly lower (p=0.00166). 23.9% of sites on surface exposed genes were under negative selection compared to 19% in non-surface exposed genes (p=0.01860). Furthermore, the PCA, the gene fingerprint analysis, did not show surface exposed proteins to be more similar to each other than to non-surface exposed proteins.

Having examined the average dN/dS scores, the percent of sites under positive and negative selection and the evolutionary fingerprints for 30 human virus genes, surface exposed proteins do not appear to have a significantly different evolutionary trajectory than other viral proteins. Furthering our understanding of virus evolution and virus-host interaction such as providing an in depth look at the difference in the rate of evolution between surface exposed and non-surface exposed virus proteins can be valuable.

Presenter's Name: Bhai, Pratibha

Additional Author(s): Chin-Yee B, Cheong I, Matyashin M, Levy MA, Ho J, Foroutan A, Lazo-Langner A, Stuart A, Hsia C, Lin H, Chin-Yee I, Sadikovic B

Abstract Title: Spectrum of Myeloid Mutations in Patients with Elevated Hemoglobin

Abstract:

Background: JAK2 V617F and exon 12 mutations are the characteristic driver mutations in polycythemia vera (PV), identified in more than 95% of patients. In addition, other genetic mutations have previously been described in JAK2-positive PV that appear to have prognostic significance (Tefferi et al., Blood 2016). The incidence of other driver mutations in unselected patients referred for elevated hemoglobin is less well studied. This study aims to characterize the genetic mutational landscape in a real-world population of patients referred for elevated hemoglobin using a targeted Next-Generation Sequencing (NGS)-based assay.

Method: All patients referred for elevated hemoglobin levels (>160 g/L in females or >165 g/L in males) between 2018 and 2020 to hematology clinics at London Health Sciences Centre in Southwestern Ontario, Canada were assessed for genetic variants using the NGS-based Oncomine Myeloid Research Assay (ThermoFisher Scientific, MA, USA). This assay targets 40 key genes with diagnostic and prognostic implications in several myeloid malignancies (17 full and 23 genes genes with "hotspot" regions) and 29 fusion driver genes (>600 fusion partners). Patient demographics, and laboratory data were extracted from the electronic medical record. For all patients with genetic mutations, clinical diagnosis was confirmed by three independent reviewers.

Results: A total of 529 patients underwent genetic testing for elevated hemoglobin levels: 389 (73.5%) were males (mean age 58; range 18-95) and 140 (26.5%) were female (mean age 60; range 24-85). JAK2 mutations were detected in 10.9% (58/529) of patients and a diagnosis of PV was confirmed. The majority of JAK2-mutated PV patients (n=57) were positive for JAK2 V617F, while one patient had an exon 12 mutation. An additional single myeloid mutation was detected in 34.5% (20/58) of JAK2-positive patients and involved the following genes: TET2 (11; 19%), DNMT3A (2; 3.4%), ASXL1 (2; 3.4%), SRSF2 (2; 3.4%), BCOR (1; 1.7%), TP53 (1; 1.7%) and ZRSR2 (1; 1.7%) (Figure 1A). JAK2 mutations were not detected in 89.0% (471/529) of our cohort. A diagnosis of PV was confirmed in 2 JAK2-negative patients based on clinical features and myeloid mutations were detected in both: SRSF2 and TET2 gene mutations in 1 patient and SRSF2, IDH2, ASXL1 gene mutations in the other patient. Three JAK2-negative patients tested positive for the BCR-ABL fusion and were diagnosed with chronic myeloid leukemia. The remaining 466 JAK2-n

POSTER PRESENTATIONS 3
3A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Brintnell, Erin

Additional Author(s): Poon AFY

Abstract Title: Estimation of SARS-CoV-2 infection rates using phylogenetic summary statistics

Abstract:

Background: Throughout the SARS-CoV-2 pandemic, tracking of case counts has been instrumental in guiding public health decision making. Unfortunately, case count numbers are impacted by testing availability, voluntary participation in testing and asymptomatic infection. As a result, the true number of SARS-CoV-2 infections is likely higher than reported, with widespread variation between regions.

Aims: In response to the underreporting of SARS-CoV-2 infections, we aim to estimate infection counts from summary statistics (i.e. features) extracted from phylogenetic trees using a simulation approach.

Methods: Simulations of pandemic spread were performed using TiPS, a tree-based simulator of infection, run over 1000 replicates. Summary statistics were extracted from phylogenetic trees reconstructed from simulation output and correlation analysis was performed. A general linear model of infection counts was developed from significant summary statistics and simulation output. Once refined, the general linear model will be applied to SARS-CoV-2 phylogenetic trees.

Results: The number of unsampled lineages and Simpson's diversity (two phylogenetic summary statistics) showed significant correlation with the infectious individuals in a population. A general linear model based in the exponential distribution with a log-based linkage function, appears to capture the relationship between these two summary statistics and the number of infectious individuals, however model refinement is still underway.

Discussion: To date, our results suggest that there is a relationship between the number of unsampled lineages and the Simpson's diversity of a phylogenetic tree and the number of infectious individuals the phylogenetic tree was extracted from. Additionally, it appears that a general linear model may be developed to predict infection counts from these summary statistics. However, model development is still preliminary, and we cannot yet say whether the model will be useful in the prediction of SARS-CoV-2 infection counts. Once our model is refined and proven to work on SARS-CoV-2 data, we will apply to the model to SARS-CoV-2 lineages contained within Covizu.

Presenter's Name: Khosravifar, Ojan

Additional Author(s): Shin A, Zhang L, Asfaha S

Abstract Title: Single-Cell Analyses of Mouse Models of Colitis

Abstract:

Introduction: Inflammatory bowel disease (IBD), is a chronic inflammatory disease of the gastrointestinal tract associated with an increased risk of colorectal cancer, known as colitis-associated colorectal cancer (CAC). Over the last 30 years, mouse models of colitis have served as important avenues for pinpointing mechanisms that contribute to human disease. Nonetheless, we lack a thorough understanding of how these models work, and it is unclear which model best resembles IBD. Investigating the mechanisms of these models, our lab made the surprising finding that only the dextran sulfate sodium (DSS) model of colitis leads to CAC in mice. Follow-up studies revealed elevated levels of Ly6Chigh macrophages in the DSS model may play an important role in tumor formation.

Aims: Using single-cell RNA-sequencing (scRNA-seq), we aim to compare and better characterize the immune cell response of the mouse models of colitis, with a focus on myeloid cells.

Methods: Colitis was induced in 6-week-old C57BL/6 mice using DSS and oxazolone, the two most important colitis models. Whole colon dissociation was conducted, followed by immune cell isolation using Percoll density centrifugation and myeloid cell purification. ScRNA-seq was performed using the 10x Genomics Chromium platform. Read alignment was done using Cell Ranger; quality control, and downstream analyses were completed using the Seurat pipeline on R.

Results: Our results show elevation of myeloid precursors in DSS-induced colitis that are Ly6G+Ly6C+ double positive. Furthermore, a Ly6Chigh macrophage population unique to DSS was also observed. In contrast to Ly6Clow-int macrophages found in oxazolone, Ly6Chigh macrophages have a unique expression profile with elevated expression of inflammatory factors. Due to high levels of non-immune cell contamination in our data set, we redeveloped an optimized colonic single-cell isolation protocol using FACS.

Discussion: These findings show that the immune response in DSS-induced colitis is unique to that of oxazolone colitis. Moreover, Ly6Chigh macrophages may play an important role in the inflammatory response in DSS. In the near future, to assess how well the colitis models resemble IBD, this data set will be compared to analogous single-cell data sets from IBD patients. This project will also enable future characterizations of the mouse colon at a single-cell level, with the use of our optimized single-cell isolation protocol.

POSTER PRESENTATIONS 3
3A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Liu, Mo

Additional Author(s): Poon AFY

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

Abstract:

The identification of groups of epidemiologically-related infections in a population is a common problem in the molecular surveillance of viruses. A popular method for generating clusters is pairwise distance clustering (e.g., HIV-TRACE), which connects pairs of sequences with genetic distances below a given threshold. The result can be represented as network of infections separated into connected components. A connected component is a set of interconnected nodes (infections) that are not connected to any other node in the network. A limitation of pairwise clustering is that conventional thresholds yield components that exclude large numbers of infections, i.e., unconnected nodes.
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.

We hypothesize that community detection methods can resolve this problem. A community is a set of densely connected nodes in the network. We calculated pairwise Tamura-Nei (TN93) distances for 2915 HIV-1 genotypes sequences from Vanderbilt Comprehensive Care Clinic (VCCC) in Tennessee (Genbank accessions MH352627–MH355541). Next, we generated networks under varying TN93 thresholds, and produced clusters as either connected components or with a community detection algorithm (Markov clustering). We fit Poisson regression models to the distribution of cases in the most recent year among clusters of previous years (count outcome Y), with cluster size (n) and sample collection dates (t) as predictors. The optimal threshold maximizes the difference in AIC between the null (log Y ~ α+βn) and alternate (log Y ~α+β 1 n+β 2 t) models.

For connected components, the optimal TN93 threshold was 0.015–0.025, where 77%–85% of new cases are connected to clusters. Markov Clustering selected higher thresholds 0.025–0.035, expanding coverage to 85%–90% of new cases. In other words, using a more flexible clustering method enables us to predict a greater proportion of HIV incidence.

Presenter's Name: Pranckeviciene, Erinija

Additional Author(s): Sadikovic B

Abstract Title: Preprocessing strategy for methylation dataset in a small sample size – high dimensionality scenario: simultaneous classification and feature selection

Abstract:

Microarray data usually is characterized by a very high data dimensionality and a small sample size. While established methods such as limma - linear modelling of microarray expression data analysis to identify significantly differentially expressed probes– is widely used, overall analysis of the microarrays may benefit from a preprocessing method that simultaneously performs feature selection and linear classifier training. This class of linear models comprises Lasso and a linear programming support vector machine in which a regularization term controls a sparsity of a solution. We present an early experiment of a preprocessing of a very heterogeneous sample of public methylation datasets from NCBI Gene Expression Omnibus (GEO) database aiming at identification and validation of a significantly differentially expressed probes.

POSTER PRESENTATIONS 3
3A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Qian, Brian

Additional Author(s): Naghavi NH, Shooshtari P

Abstract Title: Collection of associations between cell types and complex disease including both bulk and single-cell open chromatin region data

Abstract:

Introduction: Several disease risk variants reside on non-coding regions of DNA, and particularly on open chromatin regions (OCR) of specific cell types. This suggests that disease risk may be driven by gene regulation rather than changing the coding sequences of protein coding genes, and therefore a combination of OCR data and genetic association data can help identify mechanisms of complex diseases. For most complex diseases, the known relevant cell types are highly heterogeneous; thus, further subsets of these cell types remain unexplored, and are potentially highly informative. In this study, I create a collection of associations between combinations of different cell types and an array of complex diseases, and in addition, use single-cell sequencing data to further expand these associations through potentially undiscovered cell subtypes.

Methods: I prepared open chromatin region (OCR) data from two curated databases: OCHROdb and scATAC.Explorer. OCHROdb is a quality-checked database of open chromatin regions gathered from multiple large-scale consortia-based projects. scATAC. Explorer is a curated collection of single-cell ATAC-seq datasets available in a standardized format. I integrated the OCR data and disease GWAS summary statistics and perform LD score regression analysis to estimate the amount of disease heritability attributing to OCR of each cell type. This resulted in prioritizing cell types that are likely to be relevant to complex diseases.

Results: I applied my method to GWAS of 27 diseases and the bulk OCR data, and found significant results (FDR < 0.05) for at least one cell type in eight complex diseases, including strong associations between immune cell types with rheumatoid arthritis and multiple sclerosis. With the integration of single-cell ATAC-seq data, there were significant correlations found in similar disease types, along with other diseases that showed little to no significance when initially integrated with bulk OCR data. This includes type 1 diabetes, primary biliary cirrhosis, and lupus, as these diseases were found to be significant with cell subtypes found in peripheral blood mononuclear cells.

Discussion: GWAS can be used to uncover associations between cell types and disease phenotypes, when coupled with OCR data. Furthermore, my results also suggest that single-cell data can further uncover new correlations through undiscovered cell-subtypes, providing more informative results versus bulk sequencing data.

POSTER PRESENTATIONS 3
3A: BIOINFORMATICS AND DATA SCIENCE

Presenter's Name: Win, Phyo

Additional Author(s): Singh SM, Castellani CA

Abstract Title: Mitochondrial DNA Copy Number and Heteroplasmy in Monozygotic Twins Discordant for Schizophrenia.

Abstract:

Introduction: Schizophrenia (SZ) is a severe mental disorder with highly heritability (~80%) and a complex polygenic etiology influenced by environmental factors. A delayed age of onset and high discordancy (~50%) in monozygotic twins (MZ) points to a role for postzygotic somatic changes that may include changes in mitochondrial DNA (mtDNA). In fact, mtDNA Copy Number (mtDNA-CN) and Heteroplasmy (HP) have been implicated in several neurological diseases. The aim of this study is to investigate potential involvement of mtDNA-CN and HP in SZ using MZ twins. We have used the unique relationship of MZ twin pairs to assess the reliability of mtDNA calling algorithms.

Methods: Genomic DNA from blood was extracted from six pairs of MZ twins discordant for SZ and two sets of parents. Affymetrix Human SNP Array 6.0 (Affy) was performed for all samples (N=16). mtDNA-CN was defined as the ratio of total autosomal reads to total mtDNA reads. Genvisis, was used to generate array mtDNA-CN estimates adjusted for principal components, age, and sex. Whole genome sequencing (WGS) was performed for two of the twin pairs and one set of parents (N=6). fastMitoCalc and Mutserve were used to extract mtDNA-CN and HP from WGS, respectively.

Results: Estimates of mtDNA-CN generated by Affy, for MZ within-pair differences were smaller than twin-father and unrelated individual comparisons, as expected (N=16, p=0.03, effect size=0.60). Further, mother-twin comparisons displayed greater similarity in mtDNA-CN estimates compared to MZ twins (p=0.23, effect size=0.81). A similar trend was also observed from WGS analysis. No discordant HP calls were detected between MZ twins. Furthermore, both mtDNA-CN estimates from Affy and WGS produced similar results, with WGS performing slightly better based on expected relatedness (WGS R2=0.15, Affy R2=0.07), and age (WGS R2=0.21, Affy R2=0.14) trends.

Discussion: Our findings suggest stronger concordance in mother-twin mtDNA-CN estimates for each twin compared to within-twin differences, this may suggest SZ discordance as a source of mtDNA variability in MZ twins. Throughout the lifespan of the SZ twin, environmental and random genetic factors could accumulate with inherited familial risk leading to the discordant SZ phenotype. No clear evidence was found to support HP involvement in the discordance of SZ. Although both technologies provide a reliable measure of mtDNA-CN, our results favour WGS over array technology for mtDNA-CN estimation.